Génomique comparative et évolutive au sein du complexe d'espèces Leptosphaeria maculans-Leptosphaeria biglobosa

Jonathan Grandaubert

▶ **To cite this version:**

Jonathan Grandaubert. Génomique comparative et évolutive au sein du complexe d'espèces Leptosphaeria maculans-Leptosphaeria biglobosa. Sciences agricoles. Université Paris Sud - Paris XI, 2013. Français. NNT : 2013PA112230 . tel-01124142

HAL Id: tel-01124142

https://theses.hal.science/tel-01124142

Submitted on 6 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE : SCIENCES DU VÉGÉTAL

Laboratoire : UR1290 INRA BIOGER-CPP

THÈSE DE DOCTORAT

DISCIPLINE : BIOLOGIE

par

Jonathan GRANDAUBERT

Génomique comparative et évolutive au sein du complexe d'espèces *Leptosphaeria maculans* – *Leptosphaeria biglobosa*.

Soutenue le **22 octobre 2013** devant le jury composé de :

Mme Cristina VIEIRA	CNRS, Lyon	Rapportrice
M. Sébastien DUPLESSIS	INRA, Nancy	Rapporteur
M. Julien DUTHEIL	MPI, Marburg	Examinateur
M. Juergen KROYMANN	Université Paris-Sud XI	Président du jury
M. Thierry ROUXEL	INRA, Thiverval-Grignon	Directeur de thèse

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE : SCIENCES DU VÉGÉTAL

Laboratoire : UR1290 INRA BIOGER-CPP

THÈSE DE DOCTORAT

DISCIPLINE : BIOLOGIE

par

Jonathan GRANDAUBERT

Génomique comparative et évolutive au sein du complexe d'espèces *Leptosphaeria maculans* – *Leptosphaeria biglobosa*.

Soutenue le **22 octobre 2013** devant le jury composé de :

Mme Cristina VIEIRA	CNRS, Lyon	Rapportrice
M. Sébastien DUPLESSIS	INRA, Nancy	Rapporteur
M. Julien DUTHEIL	MPI, Marburg	Examinateur
M. Juergen KROYMANN	Université Paris-Sud XI	Président du jury
M. Thierry ROUXEL	INRA, Thiverval-Grignon	Directeur de thèse

Remerciements

Cette thèse a été réalisée au sein de l'équipe « Effecteurs et pathogenèse chez *Leptosphaeria maculans* », Lepto pour les intimes, de l'unité de recherche BIOGER-CPP de l'INRA de Versailles-Grignon et a débuté le 1er avril 2010 (ce n'est pas une blague).

Je tiens tout d'abord à remercier les chefs de l'équipe Lepto, **Thierry Rouxel** et **Mylène Balesdent**, pour m'avoir accueilli dans leur équipe pendant si longtemps (depuis janvier 2008 quand même). Je leur dis un grand merci pour la confiance qu'ils m'ont accordée tout au long de mes recherches : après mon stage de M2, j'ai pu continuer à travailler avec eux comme ingénieur d'étude pendant 1 an, ce qui m'a permis d'obtenir le record national de signature de CDD sur une si courte période (au moins un par mois), puis ils m'ont cru capable de réaliser une thèse, ce qui n'était pas mon cas, et ont réussi à m'embarquer dans ~~ce traquenard~~ cette aventure. J'espère vraiment ne pas les avoir déçus. Je tiens également à les remercier pour leur ouverture d'esprit, ce qui a permis la mise en place d'une bonne ambiance au sein de l'équipe tout au long de ces années.

Je remercie **Pascal « Paquito » Bally**, lui aussi passé du côté obscur depuis, qui avec moi formait le duo comique « Plic & Ploc » (on a jamais su qui était qui) célèbre dans le monde entier (à peu de choses près) pour leur humour noir, très noir, voire trop noir pour certains ! Cette longueur d'onde humoristique partagée m'a permis de m'intégrer plus facilement dans l'équipe, moi qui suis réservé à la base (et ailleurs aussi). Cette complicité se partageait également dans nos recherches puisque ensemble nous avons publié un outil bioinformatique, qui à la base devait lui permettre de travailler beaucoup moins tout en gagnant autant. Merci aussi pour le partage de tes goûts musicaux, qu'ils fussent bons (Thomas Fersen, Oldelaf et Monsieur D) ou un peu moins (Anal Cunt, Ultra Vomit). Bonne chance à toi pour ta thèse, tu verras c'est facile ;)

Je remercie **Salim Bourras** pour les photocopies couleurs. Nous sommes arrivés dans l'équipe en même temps, mais il a été libéré plus tôt que moi pour bonne conduite. Merci pour toutes les barres de rires exceptionnelles qu'on s'est prises ensemble. Merci d'avoir inventé le nom de John Bambou, qui avec un peu de ma touche perso, a fait rire beaucoup plus que nous deux. Merci pour le partage de tes aventures, réelles (« Salim à la préfecture », « Salim et la grosse femme dans l'avion » ou encore « Salim et la douane

américaine ») ou oniriques (« Salim est poursuivi par Thierry déguisé en samouraï »). Et puis, faut pas croire, mais on a quand même bien bossé et ça m'a fait très plaisir d'être sous tes ordres (« Oui Maître, tout de suite Maître ») pour réaliser les analyses bioinfo de ton article.

Je remercie également toutes les personnes qui sont toujours, ou ont été, membres de l'équipe Lepto depuis mon arrivée, et plus particulièrement **Azita Dilmaghani** (ne t'inquiète pas, les pandas vont bien), **Alexandre Degrave**, **Guillaume Daverdin**, **Juliette Linglin** (Jingle Bell, Jingle Bell) et **Bénédicte Ollivier**. Je passe le bonjour à **Michel Meyer**, 50 % chercheur - 50 % stagiaire - 100 % ailleurs.

Je remercie toutes les personnes que j'ai pu côtoyer de près, de loin, ou d'encore plus loin à BIOGER, qui ont fait que j'ai passé de bons moments ici. Je pense notamment à **Romain Valade**, **Guillaume Morgant**, **Saad Azzedine**, **David Morais**, **Johann Confais**, **Elisabetta Marchegiani**, **Aurélie Ducasse**, **Bérengère Dalmais**...

Je remercie **Joelle Amselem**, qui m'a présenté à l'équipe de Thierry, **Nicolas Lapalu** et **Jonathan Kreplak** de l'URGI à Versailles.

Je remercie **Catherine Etchebest**, ma responsable de Master 2, pour son aide et sa confiance lors de cette dernière année universitaire.

Je remercie **Cristina Vieira** et **Sébastien Duplessis** d'avoir accepté d'être les rapporteurs de mes travaux de thèse, ainsi que **Juergen Kroymann** et **Julien Dutheil** pour leur participation à mon jury de thèse. Merci également à **Eva Stukenbrock**, **Emmanuelle Lerat** et **Aurélie Hua-Van** pour leur participation à mes comités de thèse.

Je remercie **mes parents, mon frère et Alain** ainsi qu'**Iris** pour leur soutien lors de ma décision de m'engager dans ~~ce traquenard~~ tout ça. Merci aussi à **Thomas** et **Marie** pour tous les bons moments passés au cours de ces dernières années.

Enfin, pour terminer, je voudrais remercier **Jessica**, la plus belle découverte de ma thèse, pour son soutien, sa patience, son amitié et son amour durant ces deux dernières années.

« Maman quand j’s’rais grand, j’voudrais pas être étudiant.

Ben alors qu’est-ce tu veux être ?

Je sais pas moi, poète ? »

Renaud

Liste des figures

Introduction

Figure 1. Exemples de champignons ou de symptômes associés à des maladies fongiques des plantes.

Figure 2. Polyphylétisme du pouvoir pathogène chez les champignons.

Figure 3. Classification des éléments transposables proposée par Wicker *et al.* en 2007.

Figure 4. Éléments transposables et interactions génome-environnement.

Figure 5. Symptômes et stades d'infection par *Leptosphaeria maculans*.

Figure 6. Cycle biologique de *Leptosphaeria maculans* en Europe.

Figure 7. Phylogénie du complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa* et d'espèces *Leptosphaeria* proches.

Chapitre 1

Figure 1. Phylogenetic relationships between *Dothideomycetes* including an example of microsynteny between *Leptosphaeria maculans* and related species.

Figure 2. Main features of the *Leptosphaeria maculans* genome as exemplified by SuperContig_1 (3.38 Mb), which corresponds to a complete chromosome.

Figure 3. Repeat-induced Point mutation in ribosomal DNA of *Leptosphaeria maculans*, shown as RIPCAL output.

Figure 4. Dynamics of transposable elements in the *Leptosphaeria maculans* genome.

Figure S1. Strategies to validate and improve the *Leptosphaeria maculans* genome assembly.

Figure S2. Mesosynteny analyses to validate the *Leptosphaeria maculans* genome assembly by comparison to the closely related Dothideomycete, *Phaeosphaeria nodorum*.

Figure S3. Relationship between size of SuperContigs of *Leptosphaeria maculans* and their content of AT-rich regions or transposable elements.

Figure S4. Analysis of the Repeat-Induced Point mutation degeneracy gradient at the borders and within AT-rich genomic regions.

Figure S5. An example of Repeat-Induced Point mutation degeneracy in the DTM_*Sahana* transposable element family.

Figure S6. Transposable element distribution along the main SuperContigs of the *Leptosphaeria maculans* genome, and general features of SuperContigs.

Figure S7. Comparisons of recombination frequencies between contrasting genomic regions in *Leptosphaeria maculans* and relationship between physical and genetic distances.

Figure S8. Schematic representation of four telomeres in *Leptosphaeria maculans*.

Figure S9. Size distribution of AT-rich genomic regions in the final assembly.

Figure S10. RIPCAL analysis of the rDNA repeats of *Leptosphaeria maculans*.

Figure S11. Comparative Gene Ontology analysis of genes occurring in GC- and AT-blocks of the *Leptosphaeria maculans* genome.

Figure S12. Genomic location of putative effector genes along chromosomes of *Leptosphaeria maculans*: the example of SuperContig_0.

Figure S13. Relationship between transposable elements and Small Secreted Protein-encoding genes in the *Leptosphaeria maculans* genome.

Figure S14. Q-RT-PCR analysis of *in planta* expression of selected Small Secreted Protein-encoding genes as a function of their genomic environment.

Figure S15. Box plot representation of the TpA/ApT RIP indices in four sets of genes.

Chapitre 2 – 1ère partie

Figure 1. Chronogram of major classes in *Ascomycota*, with a focus on *Dothideomycetes*.

Figure 2. Comparative chromosome structure in *L. maculans* 'brassicae' (Lmb), *L. maculans* 'lepidii' (Lml) and *L. biglobosa* 'thlaspii' (Lbt): the example of chromosome 2.

Figure 3. Genome alignment and synteny analyses between *L. maculans* 'brassicae' (Lmb) and *L. maculans* 'lepidii' (Lml).

Figure 4. Distribution of selected Transposable Elements (TEs) within the dothideomycete phylogeny.

Figure 5. Conservation of protein-encoding genes in the *Leptosphaeria maculans-L. biglobosa* species complex.

Figure 6. Conservation of secondary metabolite gene clusters in the *Leptosphaeria maculans-L. biglobosa* evolutionary series.

Figure 7. Organisation of the genes surrounding the polyketide synthase, PKS21, of *L. biglobosa* 'brassicae'.

Figure 8. Effect of presence of repetitive element adjacent to promoter of orthologs on gene expression *in planta* compared to in axenic culture.

Figure S1. Representative electrokaryotypes and presence of transposable elements in the genomes of isolates of the *L. maculans-L. biglobosa* species complex.

Figure S2. Expanded chronogram of major classes in *Ascomycota*, with a focus on *Dothideomycetes*

Figure S3. Whole genome DNA comparison of *Leptosphaeria maculans* 'brassicae' v23.1.3 to progressively more distantly related members of the species complex.

Figure S4. Genomic DNA of isolates of *Leptosphaeria* species digested with restriction enzymes *Bam*HI (RLC_*Pholy*) or *Hind*III (RLG_*Rolly*) and hybridised with probes of transposable elements abundant in *L. maculans* 'brassicae'.

Figure S5. Circos representation of chromosome-by-chromosome genome conservation between *L. maculans* 'brassicae' and *L. maculans* 'lepidii'.

Figure S6. Conservation of cysteine spacing in a series of orthologs of avirulence genes of *L. maculans* 'brassicae'.

Chapitre 2 – 2ème partie

Figure 1. Distribution du contenu en bases GC dans les génomes de quatre souches de *L. maculans* 'brassicae'.

Figure 2. Alignements du génome de v23.1.3 avec ceux des autres souches.

Figure 3. Distribution des SNPs détectés entre v23.1.3 et les trois autres souches.

Figure 4. Répartition des PPS dans les génomes des 4 souches de *L. maculans* 'brassicae'.

Figure 5. Pseudogénisation d'AvrLm4-7 dans la souche Nz-T4.

Discussion

Figure 1. Relations phylogénétiques et estimation des temps de divergence au sein des classes majeures des Ascomycètes.

Figure 2. Synténie entre les génomes de *L. maculans* 'brassicae' et *L. maculans* 'lepidii', exemple du Chromosome 3.

Liste des tableaux

Introduction

Tableau 1. Proportion en éléments transposables (ET) et taille de génome chez les eucaryotes.

Chapitre 1

Table 1. Assembly statistics for the *L. maculans* genome.

Table 2. Features of genomes of *L. maculans* and other closely related Dothideomycetes.

Table 3. Comparative features of SSP-encoding genes occuring in diverse genome environments.

Table 4. Main families and characteristics of transposable elements and other repeats in the *L. maculans* genome.

Table S1. Taxa and sequences used in phylogenetic analyses.

Table S2. Characteristics of the main SuperContigs of the *Leptosphaeria maculans* genome and their assembly into complete chromosomes using a combination of approaches.

Table S3. Gene model statistics for the *Leptosphaeria maculans* genome.

Table S4. Automated functional annotation of *Leptosphaeria maculans* genes.

Table S5. Comparative analyses of intergenic distances between *Leptosphaeria maculans* and the closely-related Dothideomycete, *Phaeosphaeria nodorum*.

Table S6. Orthologues of the *Neurospora crassa* factors necessary for gene silencing identified in the genome of *Leptosphaeria maculans*.

Table S7. Occurrence of AT-blocks in SuperContigs and size of the borders between AT-rich and GC-equilibrated genomic regions.

Table S8. A data matrix of individual nesting of Transposable Elements in the *Leptosphaeria maculans* genome.

Table S9. Non-ribosomal peptide synthetases of *Leptosphaeria maculans*: gene location, orthologs, and putative function.

Table S10. Polyketide Synthases of *Leptosphaeria maculans*: gene location, orthologs, and putative function.

Table S11. Favoured codon usage for Small Secreted Proteins located within AT-rich genomic regions.

Table S12. Amino acid favoured usage (% per protein) in different sets of predicted proteins of *Leptosphaeria maculans* compared to proteins referenced in SWISSPROT database.

Table S13. Summary statistics for the shotgun sequencing of *Leptosphaeria maculans* isolate v23.1.3.

Table S14. cDNA libraries of *Leptosphaeria maculans* isolates generated and used here.

Chapitre 2 – 1ère partie

Table 1. Sequencing statistics and genome facts for six members of the *Leptosphaeria maculans-L. biglobosa* species complex.

Table 2. Alignment length and SNP counts between the different genomes of members of the *L. maculans-L. biglobosa* species complex.

Table 3. Transposable element content in genomes of isolates of the *L. maculans-L. biglobosa* species complex.

Table 4. Predicted genes conservation between the genomes of isolates of the *L. maculans-L. biglobosa* species complex.

Table 5. Characteristics of SSP genes and proteins between isolates of the *L. maculans-L. biglobosa* species complex.

Table 6. The PKS21 gene cluster of *Leptosphaeria biglobosa* 'brassicae' and homologs in *Arthroderma otae*.

Table S1. Characteristics of selected genomes of *Dothideomycetes*.

Table S2. Orthologues of the *Neurospora crassa* factors necessary for gene silencing identified in the genome of isolates of the *Leptosphaeria maculans-L. biglobosa* species complex.

Table S3. Gene model statistics for the *Leptosphaeria* spp. genomes.

Table S4. Chromosomal assignment and correspondance between *L. maculans* 'brassicae' and *L. maculans* 'lepidii'.

Table S5. Location and distance from TEs of the 30 intrachromosomal inversions in the chromosomes of *L. maculans* 'brassicae'.

Table S6. Characteristics of Class I (Retrotransposons) Transposable Elements identified in the *Leptosphaeria maculans-L. biglobosa* species complex.

Table S7. Characteristics of Class II (DNA transposons) Transposable Elements identified in the *Leptosphaeria maculans-L. biglobosa* species complex.

Table S8. Insertion polymorphism of Transposable Elements between the two isolates of *L. maculans* 'brassicae'.

Table S9. Predicted SSP-encoding genes conservation between the genomes of isolates of the *L. maculans-L. biglobosa* species complex.

Table S10. Conservation of genes encoding avirulence effector in members of the *L. maculans-L. biglobosa* species complex and other fungal species.

Table S11. Non-Ribosomal Peptide Synthases (NPS) genes of the *L. maculans-L. biglobosa* species complex.

Table S12. Polyketide Synthases (PKS) genes of the *L. maculans-L. biglobosa* species complex.

Chapitre 2 – 2ème partie

Tableau 1. Données de séquençage et d'annotation des génomes des 4 souches de *L. maculans* 'brassicae'.

Tableau 2. Nombre de SNPs détectés entre chaque paire de génomes alignés.

Discussion

Tableau 1. Données de séquençage et d'annotation des génomes du complexe d'espèces *L. maculans-L. biglobosa*.

Liste des abréviations

aa	Acide aminé
ADN(r)	Acide désoxyribonucléique (ribosomique)
ARN	Acide ribnonucléique
AT	Adénine-Thymine
BAC	Bacterial artificial chromosome
bp/pb	Paire de bases
CAZyme	Carbohydrate active enzyme
dpi	Days post inoculation
ET	Élément transposable
gBGC	GC-biased gene conversion
Gb	Gigabase
GC	Guanine-Cytosine
H(G)T	Horizontal (gene) transfer
INRA	Institut National de la Recherche Agronomique
ITS	Internal transcribed spacer
JGI	Joint Genome Institute
kb	Kilobase
Lbb	*Leptosphaeria biglobosa* ‘brassicae’
Lbc	*Leptosphaeria biglobosa* ‘canadensis’
Lbt	*Leptosphaeria biglobosa* ‘thlaspii’
Lmb	*Leptosphaeria maculans* ‘brassicae’
Lml	*Leptosphaeria maculans* ‘lepidii’
LTR	Long terminal repeat
Mb	Mégabase
NGS	Next generation sequencing
N(R)PS	Non-ribosomal peptide synthase
PKS	Polyketide synthase
PPS	Petite protéine sécrétée
p. ex	Par exemple
ORF	Open reading frame
RIP	Repeat-induced point mutation
SNP	Single nucleotide polymorphism
spp.	Espèces
TIR	Terminal inverted repeat

Sommaire

Introduction *15*

1. Les champignons *17*
   Styles de vie des champignons phytopathogènes *19*
2. La génomique des champignons *21*
   La génomique fongique *21*
   Structure des génomes *22*
   Génomique et pouvoir pathogène *23*
   Les effecteurs *24*
3. Les éléments transposables *27*
   Classification des ET *27*
   Abondance des ET dans les génomes *29*
   Distribution des ET dans les génomes *31*
   Incidence des ET sur les génomes *33*
   Contrôle des ET par les génomes *34*
4. *Leptosphaeria maculans* *37*
   Cycle biologique de *L. maculans* *39*
   Interactions *B. napus – L. maculans* *39*
   Le complexe d'espèces *L. maculans-L. biglobosa* *40*
   Le pouvoir pathogène de *L. maculans* *43*
5. Projet de thèse *45*

Chapitre 1 : Le génome de *Leptosphaeria maculans* 'brassicae' *49*

Introduction *50*

Article 1 : Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations *52*

Chapitre 2 : Évolution et adaptation dans le complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa* *145*

Introduction *146*

Article 2 : Transposable Element-assisted evolution and adaptation within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal plant pathogens *148*

Dynamique évolutive du génome et des effecteurs chez *L. maculans* 'brassicae' *243*

Discussion et perspectives ***253***

Article 3 : Incidence des Éléments Transposables sur l'évolution des génomes des champignons phytopathogènes et leur potentiel adaptatif
256

Perspectives *285*

Annexes ***295***

Articles *297*

Article 4 : FONZIE: An optimized pipeline for minisatellite marker discovery and primer design from large sequence data sets
298

Article 5 : Incidence of genome structure, DNA asymmetry, and cell physiology on T-DNA integration in chromosomes of the phytopathogenic fungus *Leptosphaeria maculans*
308

Article 6 : The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*
324

Article 7 : Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*
337

Article 8 : Deciphering the regulome of the causal agent of the stem canker agent of oilseed rape, *Leptosphaeria maculans* 'brassicae' to select putative regulators of pathogenicity
338

Bilan d'activité *339*

Posters *343*

Références ***353***

INTRODUCTION

Figure 1. Exemples de champignons ou de symptômes associés à des maladies fongiques des plantes. (**A**) *Aspergillus niger*, producteur industriel d'acide citrique (E330) ; (**B**) *Candida albicans*, responsable de candidoses chez l'homme ; (**C**) *Saccharomyces cerevisiae*, levure de boulanger et de bière ; (**D**) *Xanthoria parietina*, champignon lichénisé ; (**E**) *Penicillium roqueforti*, utilisé dans la production de fromages à pâte persillée ; (**F**) *Tuber melanosporum*, la truffe noire ; (**G**) *Agaricus bipolaris*, le champignon de Paris ; (**H**) *Laccaria bicolor*, champignon ectomycorhizien [source : Lawrence Livermore National Laboratory] ; (**I**) *Zymoseptoria tritici*, responsable de la septoriose du blé ; (**J**) *Botrytis cinerea*, la pourriture grise de la vigne ; (**K**) *Ustilago maydis*, responsable du charbon du maïs [credit : H Zell] ; (**L**) *Leptosphaeria maculans*, responsable de la nécrose du collet du colza [credit : MH Balesdent].

1. Les champignons

Tout le monde a sa propre idée de ce qu'est un champignon, que l'on soit gastronome, agriculteur, médecin ou *retrogamer*. En biologie, le terme « champignon » est utilisé pour décrire les organismes appartenant au règne *Fungi* (Eumycètes *i.e.* vrais champignons), c'est à dire des eucaryotes pluricellulaires ou unicellulaires dépourvus de chlorophylle incapables de synthétiser les substances organiques nécessaires à leur croissance et qui dépendent donc des molécules fabriquées par d'autres organismes. Ce règne regroupe des organismes d'une grande diversité, estimée à 1,5-3 millions d'espèces (Hawksworth, 2012a) voire 10 millions selon certains auteurs (Cannon, 1997) et qui ont réussi à se développer dans la plupart des niches écologiques (Kubicek & Druzhinina, 2007) en mettant en place différents styles de vie (Figure 1). Leur impact sur l'écosystème et le développement de notre société est considérable. Les champignons sont les agents principaux responsables, aux côtés des bactéries, de la dégradation et du recyclage de la matière organique morte (saprophytisme) et ce sont les seuls organismes capables de dégrader la lignine, constituant essentiel du bois. De nombreux champignons contribuent à différents processus industriels, d'autres sont une source de nourriture ou participent à l'élaboration de produits chimiques et pharmaceutiques utilisés quotidiennement par l'homme. Ces organismes peuvent aussi avoir des effets délétères sur l'économie ou sur la santé quand ils agissent comme agents pathogènes, causant des maladies extrêmement difficiles à contrôler chez les animaux ou les plantes (Desprez-Loustau *et al.*, 2007).

Au cours du temps, les champignons ont développé différents types d'interactions avec d'autres organismes, ce qui a permis leur évolution mais aussi celle de leurs partenaires. Des lignées d'espèces mutualistes (symbiontes, endophytes, lichénisés) et parasites ont ainsi émergé de multiples fois et de manière indépendante dans des lignées saprophytes (Figure 2) (Hock, 2001 ; Pöggeler & Wöstemeyer, 2011). Pour des raisons qui semblent principalement liées à leur faible tolérance aux températures élevées, les champignons sont plutôt associés aux plantes qu'aux animaux. En effet, sur les 100 000 espèces de champignons décrites, 10 % sont phytopathogènes alors qu'en comparaison, seulement 0,05 % sont connues comme causant des maladies chez l'homme ou les animaux (Agrios, 2005). Chez l'homme, il s'agit le plus souvent de maladies opportunistes comme par exemple l'aspergillose ou la candidose chez des patients immunodéprimés.

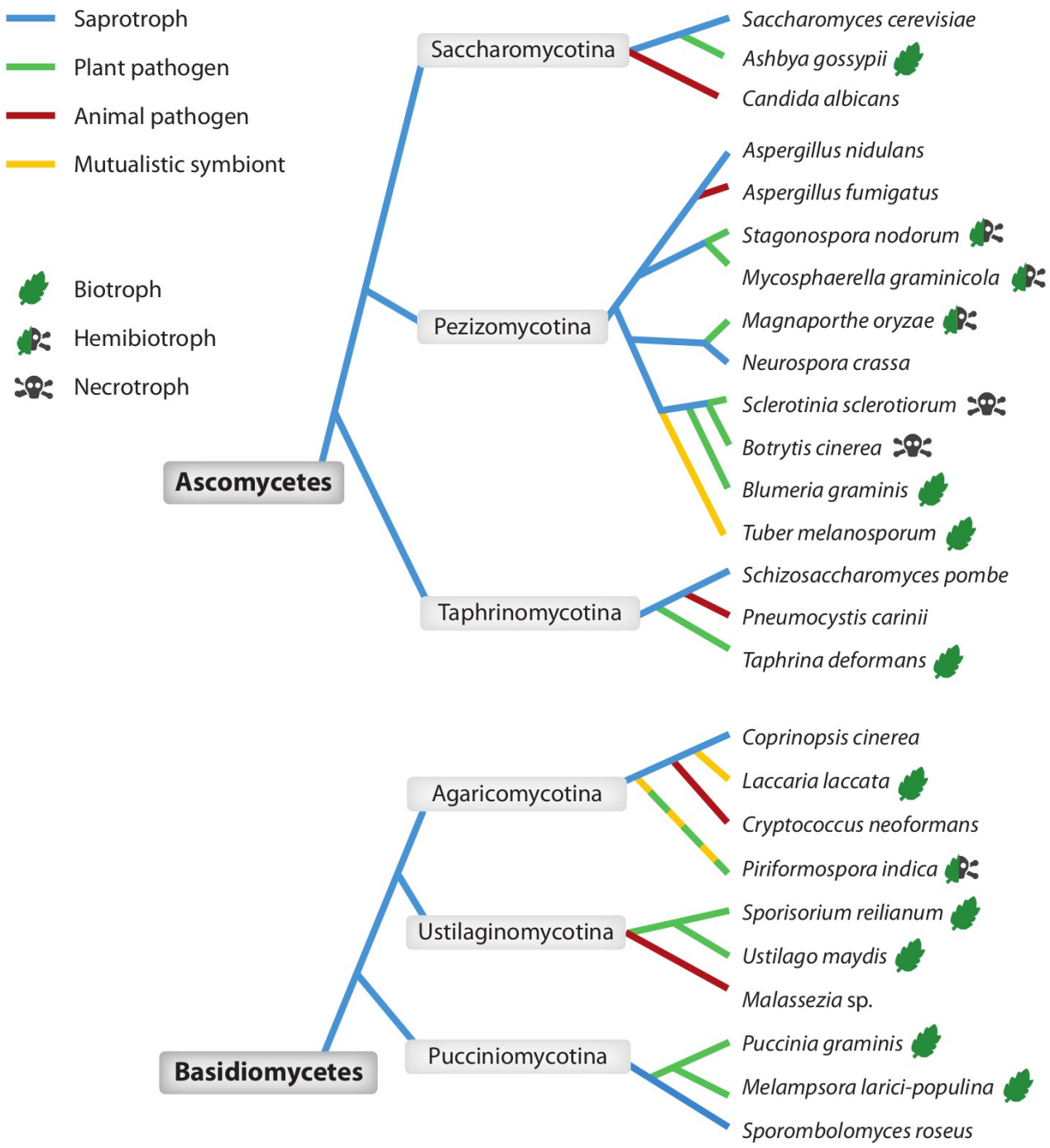


Figure 2. Polyphylétisme du pouvoir pathogène chez les champignons. Ces phylogénies montrent comment des agents pathogènes de plantes (vert) et d'animaux (rouge) ont évolué indépendamment à différents temps à partir de lignées saprophytes (bleu). Adapté de Spanu (2012).

Les champignons phytopathogènes sont les micro-organismes les plus dommageables aux cultures dans le monde entier, causant d'énormes pertes de rendement (Birren *et al.*, 2002). On estime que 50 % des maladies des plantes cultivées sont dues à des champignons (Porta-Puglia & Vannacci, 2011). Selon l'ONU, la population mondiale devrait augmenter de plus de 30 % dans les quarante prochaines années (www.un.org) et dans un tel contexte, la lutte contre les pertes de denrées alimentaires provoquées par les champignons phytopathogènes constitue un réel enjeu mondial en agriculture. Différents moyens de lutte ont donc été mis en place afin de contrôler ces agents pathogènes, soit en essayant de les supprimer grâce à l'utilisation de produits chimiques (fongicides), soit en essayant de diminuer l'impact des maladies par des changements de pratiques culturales (rotation des cultures, enfouissement des résidus de cultures, sélection de variétés résistantes aux maladies). Si l'on veut développer des mesures de lutte efficaces et respectueuses de l'environnement, basées sur la résistance génétique, une meilleure compréhension des différentes interactions établies entre le champignon et son hôte est nécessaire. Cependant, les modes de vie et de parasitisme étant très divers d'une espèce à l'autre, il est difficile d'obtenir des informations génériques concernant ces interactions.

1.1 Styles de vie des champignons phytopathogènes.

Au cours de l'évolution, les champignons pathogènes de plantes ont montré une extrême plasticité concernant leur biologie, leur cycle vital et parasitaire, ou leurs modes de survie et de dispersion (Glawe, 2008). Ils ont ainsi développé différents modes de vie que l'on catégorise comme biotrophes, nécrotrophes ou hémibiotrophes. Cette classification est toutefois très subjective et une même espèce peut être considérée différemment selon les auteurs.

Les biotrophes stricts sont des parasites obligatoires ayant une très grande spécificité d'hôte, dont ils dépendent fortement, et qui, pour se nourrir et compléter leur cycle de vie, doivent maintenir vivantes les cellules de l'hôte (Mendgen & Hahn, 2002 ; Ridout, 2009 ; Voegele *et al.*, 2009). Les membres les plus connus et les plus étudiés de ce groupe sont les rouilles (p. ex. *Puccinia graminis*) et les oïdiums (p. ex. *Blumeria graminis*) (Staples, 2000 ; Glawe, 2008). L'absorption des nutriments est réalisée via des structures spécialisées appelées haustoria. Mais ces structures ne servent pas uniquement à la nutrition, elles sont aussi capables d'exporter des molécules ayant pour

fonction de neutraliser le système immunitaire de la plante et de maintenir la biotrophie (Stergiopoulos & de Wit, 2009 ; Kloppholz *et al.*, 2011). D'un point de vue évolutif, la biotrophie stricte peut être considérée comme une spécialisation irréversible se traduisant par la perte d'un grand nombre de gènes utilisés par d'autres agents pathogènes durant des phases de saprotrophie ou de nécrotrophie (Spanu, 2012).

Les nécrotrophes détruisent les cellules de l'hôte en sécrétant des molécules phytotoxiques et des enzymes lytiques qui vont tuer et décomposer les tissus de la plante, ce qui permettra au champignon de se développer de façon saprophyte dans les tissus morts (Horbach *et al.*, 2011). Si les molécules toxiques présentent une activité spécifique à une ou quelques espèces végétales, l'agent pathogène aura une gamme d'hôte réduite (p. ex. *Cochliobolus* et *Alternaria* spp.). Dans le cas contraire, la gamme d'hôte peut être très large et comprendre plus de 200 espèces de dicotylédones comme pour *Botrytis cinerea* (van Kan, 2006).

Entre ces deux extrêmes, on trouve la grande majorité des agents pathogènes de plantes classifiés de façon plus ou moins vague comme hémibiotrophes. Les hémibiotrophes présentent un cycle de vie plus complexe que les nécrotrophes ou les biotrophes stricts puisqu'ils sont capables d'alterner différents comportements nutritionnels au cours de celui-ci. Ils montrent ainsi une extrême diversité de mode de vie et de colonisation des tissus. Certains champignons, comme *Leptosphaeria maculans*, établissent initialement une phase de biotrophie durant laquelle ils se développent de façon asymptomatique dans l'apoplaste de l'hôte avant de devenir nécrotrophes en détruisant les tissus de la plante sur lesquels ils vont finir leur cycle de vie comme saprophytes (Tyler & Rouxel, 2013). D'autres champignons, comme *Magnaporthe oryzae* ou *Colletotrichum* spp., vont pénétrer les cellules épidermiques de l'hôte de façon biotrophique via une structure spécialisée appelée appressorium et les coloniser une à une avant d'entamer une phase nécrotrophe. La plasticité nutritionnelle des hémibiotrophes suggère une régulation complexe des mécanismes de la pathogenèse. La compréhension de la mise en place des différents programmes parasitaires nécessite une analyse de génomique et transcriptomique globale pour identifier les déterminants de la pathogenèse mais aussi une dissection fine d'un cycle vital complexe en interaction avec la plante hôte (Tyler & Rouxel, 2013).

2. Les génomes des champignons

Les champignons, en tant qu'organismes eucaryotes, partagent un grand nombre de fonctions biologiques, de traits de vie ou de structures communes avec les animaux et les plantes. Cela concerne par exemple le développement et la différenciation cellulaire, les cycles circadiens, la reproduction sexuée, la signalisation intercellulaire, la méthylation de l'ADN ou les modifications de la chromatine (Galagan *et al.*, 2005a). Les gènes impliqués dans ces fonctions biologiques fondamentales ont une origine commune à tous les eucaryotes. Ainsi grâce à leurs génomes plus petits et généralement moins complexes que ceux des eucaryotes « supérieurs », les champignons sont considérés comme des organismes modèles pour comprendre les grandes fonctions biologiques du vivant. De plus, ce sont des organismes souvent facilement manipulables expérimentalement grâce au développement d'outils génétiques, moléculaires ou cytologiques. De la même façon que les génomes fongiques seraient explicatifs de la biologie des eucaryotes, l'étude de leur évolution devrait permettre d'accroître nos connaissances sur l'histoire de la vie sur Terre et l'évolution des écosystèmes.

2.1 La génomique fongique.

L'ère de la génomique fongique a été amorcée en 1996 avec la publication du premier génome eucaryote séquencé, celui de la levure *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996). Ces données ont permis les premières analyses fonctionnelles des gènes eucaryotes ainsi que de leur expression. Cela a aussi ouvert la porte à la génomique comparative et Botstein *et al.* ont montré en 1997 que 31% des 6000 gènes de la levure étaient partagés avec les mammifères. Malgré les formidables avancées technologiques et biologiques fournies par ce modèle, il faudra attendre 6 ans et le séquençage du génome de *Schizosaccharomyces pombe* (Wood *et al.*, 2002) pour voir la génomique fongique continuer sa marche. Cependant, les génomes de levures permettent d'appréhender seulement une infime partie de la diversité biologique du règne des champignons. En 2003 puis en 2005, les génomes des champignons filamenteux *Neurospora crassa* (Galagan *et al.*, 2003) et *Aspergillus nidulans* (Galagan *et al.*, 2005b) sont publiés et apportent de nouvelles données, avec des génomes près de trois fois plus grands que ceux des levures et comportant deux fois plus de gènes. En 2008, au commencement de mon travail de recherche sur *Leptosphaeria maculans*, plus de 40

génomes de champignons étaient séquencés et disponibles dans des bases de données privées ou publiques. Parmi ces génomes séquencés, une douzaine provenaient d'agents pathogènes de plantes (Soanes *et al.*, 2007) mais quatre seulement étaient publiés : ceux des ascomycètes *Magnaporthe oryzae*, *Fusarium graminearum* et *Phaeosphaeria nodorum* (Dean *et al.*, 2005 ; Cuomo *et al.*, 2007 ; Hane *et al.*, 2007) et du basidiomycète *Ustilago maydis* (Kämper *et al.*, 2006). L'exploitation de chacun de ces génomes puis la comparaison de leur séquences codantes étaient censés faciliter l'identification des déterminants potentiellement impliqués dans leurs différents styles de vie ou dans les interactions avec leur hôte et fournir ainsi un aperçu des adaptations évolutives requises par le champignon pour causer des maladies. Mais la plasticité des génomes fongiques ainsi que le manque de ressources nécessaires permettant l'attribution d'une fonction à un gène, plus ou moins prédit et/ou annoté, ont grandement complexifié le postulat de départ.

2.2 Structure des génomes.

Malgré des similitudes morphologiques et physiologiques, les champignons diffèrent grandement au niveau génomique. Les temps de divergence très importants entre les espèces aux génomes séquencés font qu'il est difficile de détecter une conservation de l'ordre des gènes, ou synténie, entre espèces de genres différents (Dean *et al.*, 2005 ; Hane *et al.*, 2007), les notions de genre et d'espèce étant elles-mêmes relativement mal définies chez les champignon. Des analyses comparatives effectuées entre trois espèces de Sordariomycètes, *M. oryzae*, *F. graminearum* et *N. crassa,* révèlent que les régions synténiques entre ces champignons sont très réduites, ne comportant que 3 à 20 gènes (Xu *et al.*, 2006). Même les membres d'un même genre peuvent avoir des génomes très divergents. Le pourcentage moyen d'identité en acides aminés observé entre les génomes de trois espèces d'*Aspergillus* (*A. nidulans*, *A. fumigatus* et *A. oryzae*) est comparable à celui observé entre les mammifères et les poissons (Galagan *et al.*, 2005b). Ces trois génomes sont globalement conservés entre eux, mais les régions synténiques ont subi de nombreux réarrangements tels que des inversions, des translocations, des délétions ou des duplications (Galagan *et al.*, 2005b). Ces réarrangements sont particulièrement observés près des régions télomériques et sont souvent associés à des éléments répétés, à l'instar de ce qui avait déjà été décrit chez d'autres eucaryotes (Galagan *et al.*, 2005a). Bien que les génomes des champignons filamenteux soient organisés différemment d'une espèce à l'autre, leur séquençage a tout de même permis d'esquisser des caractéristiques

génomiques communes : les génomes sont plutôt compacts, pour des génomes eucaryotes, avec une taille moyenne d'environ 30 Mb ; ils contiennent peu d'éléments répétés tels que les éléments transposables (ET) ; le nombre de gènes prédits est comparable, entre 10 000 et 15 000. Ces études ont mis en avant le dynamisme des génomes fongiques, et eucaryotes en général, et permettaient de mesurer la complexité à laquelle on se heurtait lorsqu'on s'engageait dans la génomique et la génomique comparative fongique.

2.3 Génomique et pouvoir pathogène.

Avant la disponibilité des génomes et des différentes approches haut-débit permettant l'étude fonctionnelle des gènes, l'étude de la pathogenèse fongique se focalisait sur la recherche de gènes essentiels à l'infection de la plante mais dispensable lors de la phase de croissance saprophyte. Cependant, la majorité des déterminants du pouvoir pathogène identifiés expérimentalement était impliquée dans des cascades de signalisation et des voies métaboliques conservées chez la plupart des espèces fongiques, dont des saprophytes et des symbiotes (Idnurm & Howlett, 2001). Avec l'avènement de la génomique, la comparaison des répertoires de gènes prédits dans des champignons pathogènes et non pathogènes offrait donc un moyen direct d'obtenir de nouvelles informations sur les mécanismes impliqués dans la pathogenèse fongique. En 2008, Soanes *et al*. ont suivi cette démarche et comparé les gènes prédits de 36 espèces de champignons, incluant des agents pathogènes des plantes et des saprophytes phylogénétiquement proches. Leur étude a montré qu'il n'existait vraisemblablement pas de déterminants du pouvoir pathogène conservés chez, et spécifiques de, toutes les espèces phytopathogènes. Les différences observées entre les espèces pouvaient être dues à l'expansion de certaines familles de gènes associées à des fonctions nécessaires à la mise en place de la pathogenèse (métabolites secondaires, protéases, CAZymes, transporteurs, facteurs de transcription). Cependant, Soanes *et al.* (2008) identifiaient certains gènes comme étant exclusivement présents chez cinq espèces phytopathogènes (*Fusarium graminearum*, *Botrytis cinerea*, *Sclerotinia sclerotorium*, *Magnaporthe oryzae* et *Phaeosphaeria nodorum*). Les produits de ces gènes ne possèdent pas d'homologues dans les bases de données, ni de domaines protéiques connus et la plupart sont prédits comme étant sécrétés (Soanes *et al.*, 2008). Ces protéines, appelées ensuite effecteurs, ont d'emblée été considérées comme des éléments clés de la pathogénie chez les

champignons phytopathogènes et jouent un rôle critique dans l'établissement de l'interaction et la progression de l'infection (Kamoun, 2007).

2.4 Les effecteurs

La notion d'effecteur chez les champignons phytopathogènes est très récente et a été théorisée peu de temps avant le début de mes travaux de recherche sur *L. maculans*. Historiquement, les gènes codant des effecteurs ont d'abord été identifiés par des approches génétiques et protéomiques. Ces approches, effectuées chez les champignons et les oomycètes (organismes distants phylogénétiquement mais montrant des convergences adaptatives avec les champignons phytopathogènes), ont permis de définir les caractéristiques structurales communes aux gènes codant des effecteurs : (i) les protéines qu'ils codent sont de petite taille, inférieure à 300 acides aminés, (ii) elles possèdent souvent un peptide signal prédit, et sont donc aussi nommées PPS (Petites Protéines Sécrétées), (iii) elles possèdent souvent de nombreux résidus cystéine potentiellement impliqués dans la création de ponts disulfures, ce qui permettrait de stabiliser leur structure tridimensionnelle et favoriser leur résistance aux protéases de la plante, (iv) elles n'ont pas ou peu d'homologues dans les bases de données et (v) leur fonction biochimique demeure inconnue (Stergiopoulos & de Wit, 2009). Généralement, les effecteurs sont définis comme des molécules qui ciblent spécifiquement des processus physiologiques de la plante, mais de façon limitée sans provoquer la mort cellulaire, dans le but de neutraliser, d'inhiber les défenses de la plante afin de faciliter l'infection (Tyler et Rouxel, 2013). Ces connaissances, essentiellement structurales, couplées à l'avènement des méthodes de séquençage haut-débit et aux méthodes de prédictions *in silico*, ont permis de générer le répertoire complet des effecteurs et des protéines sécrétées (sécrétome) d'un agent pathogène donné (Kämper *et al.*, 2006 ; Haas *et al.*, 2009). L'obtention de la séquence génomique a aussi apporté des informations quant à la localisation des gènes codant des effecteurs, ce qui a permis de voir que beaucoup de ces gènes se trouvaient dans des régions génomiques très dynamiques telles que des télomères et des régions riches en éléments transposables, augmentant ainsi la variabilité génétique de ces gènes, leur conférant potentiellement une évolution accélérée (Stergiopoulos & de Wit, 2009). C'est le cas de *AvrLm1* et *AvrLm6* de *L. maculans*, deux effecteurs induisant une reconnaissance spécifique par les gènes de résistance de la plante (aussi appelés gènes d'avirulence) se trouvant chacun dans de grandes régions

(150-270 kb) composées de mosaïques d'éléments transposables dégénérés (Gout *et al.*, 2006 ; Fudal *et al*., 2007). A cette époque, *L. maculans* est le premier champignon filamenteux montrant des signes d'une invasion de son génome par des éléments transposables, en faisant ainsi un modèle original pour l'étude de l'évolution des génomes et du pouvoir pathogène des champignons phytopathogènes.

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

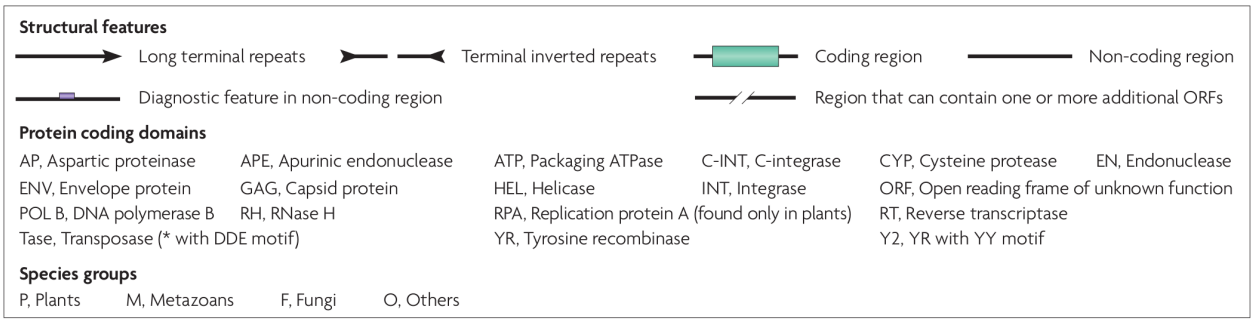


Figure 3. Classification des éléments transposables proposée par Wicker *et al.* en 2007. Les éléments sont divisés en deux classes en fonction de leur intermédiaire de transposition, ARN ou ADN. Ces classes sont divisées à leur tour en sous-classes, ordres et super-familles en fonction de la structure des éléments. Pour faciliter l'identification des éléments, Wicker *et al.* ont proposé un code à trois lettres qui est ajouté au nom de famille de l'élément.

3 - Les éléments transposables.

En associant des mutations observées chez le maïs avec des translocations non aléatoires spontanées de fragments de chromosomes à l'intérieur du génome, Barbara McClintock (1950) découvrait les éléments transposables (ET). Les ET sont définis comme des séquences d'ADN mobiles codant les instructions nécessaires à leur déplacement dans le génome (phénomène de transposition). Ils sont présents dans la plupart des organismes étudiés à ce jour, procaryotes et eucaryotes, et peuvent représenter une fraction non négligeable du génome, par exemple 85 % du génome du maïs (Schnable *et al.*, 2009) et près de 45 % du génome humain (Lander *et al.*, 2001). Les ET ont été initialement considérés comme de l'ADN poubelle (*junk DNA*, Ohno, 1972), de l'ADN parasite ou égoïste (*selfish DNA*, Orgel & Crick, 1980) mais il a été montré qu'ils constituaient une source d'innovation génétique pour l'organisme grâce à leur pouvoir mutationel ou en se comportant comme des gènes (domestication des ET) ou des éléments régulateurs de gènes (Biémont & Vieira, 2006).

3.1 Classification des ET.

En 1989, Finnegan a proposé une première classification des ET basée sur leur intermédiaire de transposition (Finnegan, 1989). On distingue ainsi deux classes : la classe I est composée de rétrotransposons qui se répliquent via un intermédiaire ARN par un mécanisme de transposition dit de « copier-coller » (*copy-and-paste*), et la classe II est composée de transposons qui se répliquent via un intermédiaire ADN par un mécanisme de « couper-coller » (*cut-and-paste*). Par la suite, la découverte de nouveaux ET capables de transposer par « copier-coller » sans intermédiaire ARN ou d'éléments non-autonomes a complexifié cette dichotomie et suggéré l'existence de nouvelles classes d'éléments. En 2007, Wicker *et al.* ont proposé une nouvelle classification des ET (Figure 3) basée sur les mécanismes de transposition, les similitudes de séquence et les relations structurales afin d'inclure les nouvelles classes d'ET identifiées. Cette classification hiérarchisée permet de conserver la séparation des éléments dans les deux classes originelles tout en précisant l'existence de sous-classes et de familles bien définies d'ET. La nomenclature servant à l'identification des ET proposée par Wicker *et al.* a été utilisée lors de l'annotation des ET des génomes des membres du complexe d'espèces *Leptosphaeria maculans-L. biglobosa* (Chapitre 1 & 2).

Les rétrotransposons ou éléments de classe I (Figure 3) commencent leur cycle de réplication en étant traduit sous forme d'ARN qui, grâce à la machinerie enzymatique qu'ils codent, va ensuite être rétro-transcrit en une molécule d'ADN qui va s'insérer à un autre endroit du génome. Chaque cycle complet produit ainsi une nouvelle copie de l'élément. Il existe deux groupes de rétrotransposons en fonction de la présence ou non de répétitions directes, les LTR (*long terminal repeats*), à leurs extrémités. Les rétrotransposons à LTR contiennent les ORFs codant pour une protéine de structure virale (*gag*) et pour une polyprotéine (*pol*) contenant une protéase, une reverse transcriptase, une RNAse et une intégrase. L'agencement de ces modules fonctionnels dans l'élément permet de discriminer différentes familles (*Gypsy*, *Copia*). Cette organisation est très proche de celle des rétrovirus, qui pourraient avoir évolué à partir des rétrotransposons à LTR grâce à l'acquisition d'une protéine d'enveloppe (*env*) (Wicker *et al.*, 2007 ; Biémont & Vieira, 2006). Les rétrotransposons sans LTR regroupent les LINE (*long interspersed elements*) et les SINE (*short interspersed elements*), ces derniers étant des éléments de petite taille non-autonomes, c'est à dire qui ne contiennent pas de séquences codantes et qui utilisent donc les enzymes codées par les éléments autonomes de leur famille d'origine ou non pour se répliquer et transposer.

Les transposons ou éléments de classe II (figure 3) transposent via un intermédiaire ADN. Ces éléments sont séparés en deux sous-classes en fonction du nombre de brins d'ADN cassés lors de la transposition. Les éléments de la sous-classe 1 codent une transposase responsable d'une cassure double-brin de l'ADN lors de l'excision de l'élément qui va ensuite s'insérer ailleurs dans le génome. Ils sont caractérisés par la présence de séquences répétées et inversées, les TIR (*terminal inverted repeats*), à leurs extrémités. On retrouve aussi dans cette sous-classe des éléments non-autonomes, les MITE (*miniature inverted-repeats transposable elements*). La deuxième sous-classe est composée des hélitrons et des polintons (*Mavericks*). Leur mécanisme de transposition diffère de celui des éléments de la sous-classe 1. Les hélitrons contiennent un domaine composé d'une hélicase et utilisent un mécanisme de transposition appelé « rolling-circle » (Kapitonov & Jurka, 2007). Les polintons (*Mavericks*) sont des éléments de grande taille (9-22 kb (Pritham, 2009)) bordés par des TIR, qui codent plusieurs protéines dont une ADN polymérase B et une intégrase laissant penser qu'il y a bien utilisation d'un intermédiaire ADN dans leur cycle de réplication. Cependant leur mécanisme de transposition reste encore indéterminé.

3.2 Abondance des ET dans les génomes.

La taille d'un génome n'est pas liée à la complexité de l'organisme ni généralement au nombre de gènes que comprend ce génome. En effet, les différences de taille de génome observées entre les espèces sont principalement dues aux parties non-codantes, comprenant les ET et autres types de séquences répétées (Biémont & Vieira, 2006). Chez les espèces eucaryotes, la taille du génome est fortement liée à la proportion en ET alors que chez les procaryotes la plupart des variations de taille des génomes résultent d'expansions géniques souvent dues à des transferts horizontaux de gènes (Frost *et al.*, 2005).

Les génomes des plantes ont un spectre de taille extrêmement étendu, allant de 64 Mb pour les espèces du genre *Genlisea* (Greilhuber *et al.*, 2006) jusqu'à 127 000 Mb pour les espèces du genre *Fritillaria* (Ambrozová *et al.*, 2011). Ces énormes variations sont majoritairement dues à la polyploïdisation et à l'expansion ou l'élimination des ET (Bennetzen *et al.*, 2005). Chez les plantes, les taux d'ET dans les génomes sont variables et peuvent être de 10 à 20 % pour les espèces à petit génome telles que *Brachypodium distachyon*, mais aussi atteindre 85 % pour des espèces possédant un génome de grande taille telles que le maïs ou l'orge (Tableau 1). Chez les mammifères, dont la taille de génome moyenne est de 3300 Mb (Gregory, 2013), la proportion en ET est assez constante et représente près de la moitié du génome (Tableau 1). Chez les vertébrés non-mammifères, on note une plus grande variation de la proportion des ET dans les génomes avec 9 % chez le poulet ou 77 % chez la grenouille verte (Tableau 1). Les micro-organismes eucaryotes tels que les champignons ont globalement des génomes de petites tailles comprenant peu d'ET, mais les nombreux séquençages récents ont permis de mettre en avant des cas « d'obésité » de certains génomes comme celui de *Blumeria graminis* ou de l'oomycète *Phytophthora infestans* (Tableau 1).

Cependant la présence des ET dans les génomes eucaryotes n'est pas systématique. Plusieurs génomes d'eucaryotes unicellulaires tels que l'algue rouge *Cyanidioschyzon merolae* (Misumi *et al.*, 2005) ou les Apicomplexa *Cryptosporidium hominis* (Xu *et al.*, 2004) et *Plasmodium falciparum* (Gardner *et al.*, 2002) ne présentent pas d'ET dans leur génome. L’absence d'ET dans ces génomes peut toutefois être due aux limites de l'identification des ET basée sur des homologies de séquence avec des ET connus définis dans des espèces phylogénétiquement distantes (Pritham, 2009).

Tableau 1. Proportion en éléments transposables (ET) et taille de génome chez les eucaryotes.

Organismes		%ET	Taille du génome (Mb)	Références
Plantes				
Arabidopsis thaliana		10	125	Initiative, T.A.G., 2000
Brachypodium dystachion		21	272	Initiative, T.I.B., 2010
Hordeum vulgare	Orge	84	5100	I.B.G.S.C., 2012
Zea mays	Maïs	85	2300	Schnable *et al.*, 2009
Mammifères				
Mus musculus	Rongeur	40	3300	Biémont & Vieira, 2006
Homo sapiens	Homme	45	3300	Biémont & Vieira, 2006
Monodelphis domestica	Marsupial	52	3600	Gentles *et al.*, 2007
Vertébrés (non mammifères)				
Gallus gallus	Oiseau	9	1050	I. C. G.S.C., 2004
Xenopus tropicalis	Amphibien	35	1700	Hellsten *et al.*, 2010
Dano rerio	Poisson	52	1412	Howe *et al.*, 2013
Rana esculenta	Amphibien	77	5800	Biémont & Vieira, 2006
Invertébrés				
Drosophila melanogaster	Insecte	10	140	Drosophila 12 genomes consortium, 2007
Caenorhabditis elegans	Nématode	12	97	Biémont & Vieira, 2006
Micro-organismes				
Ustilago maydis	Basidiomycète	1	21	Kämper *et al.*, 2006
Magnaporthe oryzea	Ascomycète	10	40	Dean *et al.*, 2005
Leptosphaeria maculans	Ascomycète	33	45	Rouxel *et al.*, 2011
Blumeria graminis	Ascomycète	64	120	Spanu *et al.*, 2010
Phytophthora infestans	Oomycète	74	240	Haas *et al.*, 2009

3.3 Distribution des ET dans les génomes.

D'un génome à un autre, les différentes classes d'ET ne sont pas représentées de la même façon en terme de nombre, de composition ou de localisation. Les rétrotransposons, grâce à leur mécanisme de réplication et à leur taille, sont souvent les principaux constituant de l'ADN répété dans les génomes eucaryotes. Par exemple, ce sont les seuls ET présents dans les génomes pauvres en ET de *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* et *Entamoeba dispar* (Pritham, 2009). On les retrouve aussi en grande proportion (90 %) dans les génomes de l'homme et de la souris (Pritham, 2009). Les transposons à ADN contribuent dans une moindre mesure à la variabilité de la taille des génomes puisqu'ils représentent souvent une faible fraction de ces derniers. Dans certains cas, ils en sont les seuls représentants comme dans les génomes de *Trichomonas vaginalis* ou *Entamoeba invadens* (Pritham, 2009).

Au sein d'un même génome, on constate le plus souvent la domination d'une famille ou d'une sous-famille. Cette domination n'est pas obligatoirement corrélée avec une grande diversité des familles. Par exemple, les LINE *L1* et les SINE *Alu* sont dominants dans le génome humain et sont représentés par peu de sous-familles. Au contraire, les rétrotransposons à LTR ont un nombre de copie relativement faible et sont représentés par une douzaine de familles différentes (Lander *et al.*, 2001). La diversité des ET et leur abondance dans le génome est très variable entre les espèces et reflète l'histoire évolutive des ET spécifique à chacun de ces génomes. Cette histoire évolutive est aussi liée à la localisation des ET dans les génomes. En effet, les ET sont fréquemment identifiés dans des régions télomériques ou péricentromériques (régions hétérochromatiniennes), des régions à faible densité génique ou dans d'autres ET. Cependant, ils sont aussi retrouvés dans des régions chromatiniennes à proximité de gènes ou de régions régulatrices (Hua-van *et al.*, 2011). Il reste à déterminer si cette distribution est le résultat d'une préférence d'insertion des éléments ou d'une sélection.

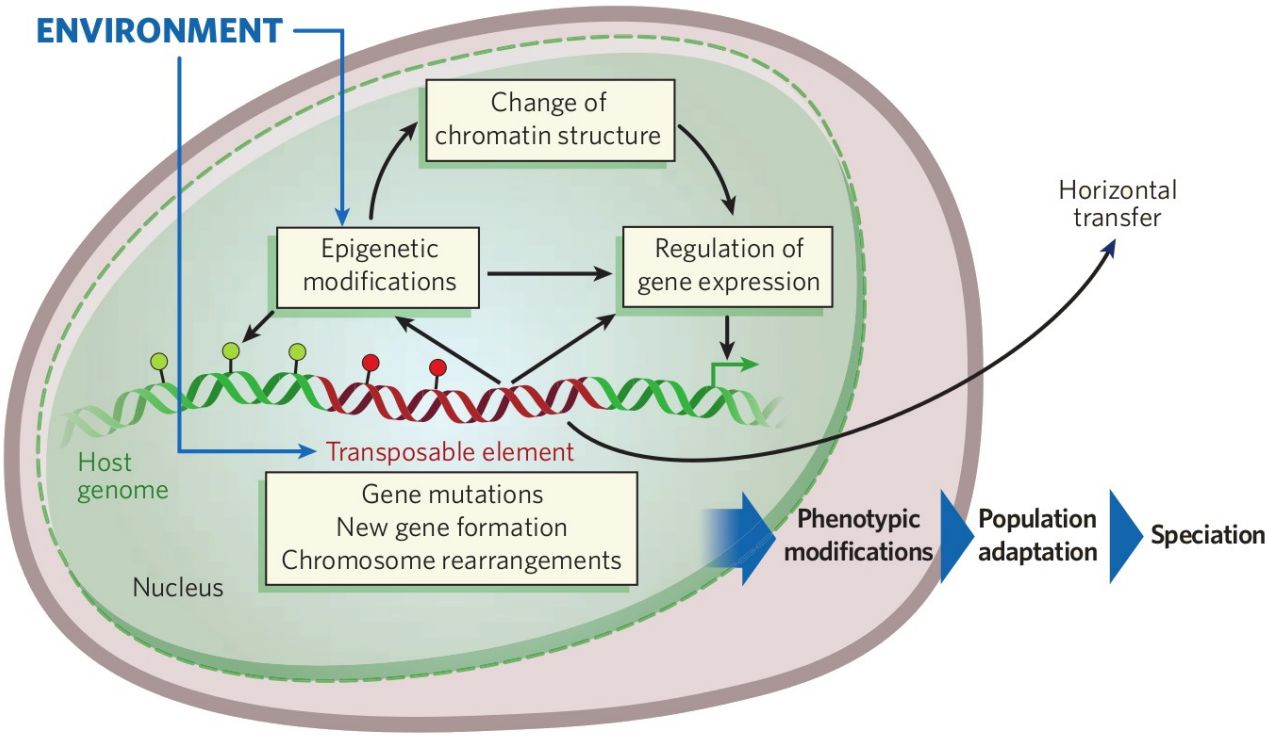


Figure 4. Éléments transposables et interactions génome-environnement. Sous l'action de l'environnement, les éléments transposables vont entraîner des modifications structurales et fonctionnelles dans le génome hôte. Les mutations peuvent être responsables de modifications phénotypiques, de changement du potentiel d'adaptation de l'organisme menant éventuellement à la spéciation. Les éléments transposables peuvent échapper aux différents mécanismes de régulation imposés par le génome hôte en se déplaçant vers les génomes d'autres organismes par transfert horizontal (Biémont & Vieira, 2006).

3.4 Incidence des ET sur les génomes.

Pendant très longtemps, les ET ont été considérés comme des éléments parasites et égoïstes, sans fonction ni rôle et n'apportant rien de bénéfique au génome hôte. Soixante ans après leur découverte (McClintock, 1950), les ET sont maintenant reconnus comme des acteurs majeurs de l'évolution des génomes. La cohabitation prolongée des ET et de leur génome hôte a conduit à des interactions multiples tant au niveau structural, avec l'induction de réarrangements chromosomiques ou la formation d'hétérochromatine, qu'au niveau fonctionnel avec la modification de l'activité des gènes ou la domestication d'ET.

A l'origine, les ET ont été identifiés à cause des (ou grâce aux) réarrangements chromosomiques qu'ils induisaient dans le génome du maïs. Les ET peuvent ainsi être à la base de recombinaisons ectopiques entre copies homologues, induisant des délétions, des translocations ou des inversions chromosomiques. Par exemple, chez les mammifères, Zhao & Bourque (2009) ont montré que les points de cassure au niveau d'inversions chromosomiques étaient enrichis en élément *L1*, un LINE qui représente près de 20% du génome humain (Lander *et al.*, 2001). Chez *Drosophila buzzatii*, Delprat *et al.* (2009) ont mis en évidence le rôle de la recombinaison ectopique due au transposon *Galileo* dans la génération de deux inversions chromosomiques. Ces recombinaisons peuvent se produire même si les ET sont inactivés (Oliver & Greene, 2012), ce qui veut dire que les ET peuvent influer sur la structure du génome longtemps après la perte de leur capacité à se déplacer dans le génome.

Les ET jouent aussi un rôle dans l'organisation de la chromatine à l'intérieur du noyau. Ils sont particulièrement abondants dans l'hétérochromatine constitutive retrouvée principalement au niveau des télomères et des centromères. Ils sont aussi fréquemment trouvés dans les régions péricentromériques (Hua-Van *et al.*, 2011) et peuvent intervenir dans la formation de l'hétérochromatine facultative en étant inactivés puis méthylés par des mécanismes épigénétiques (Lewis *et al.*, 2008).

En induisant des modifications structurales dans le génome, les ET peuvent aussi être responsables de modifications fonctionnelles. En s'insérant dans une région codante, les ET peuvent altérer ou supprimer l'activité d'un gène en modifiant sa structure (terminaison prématurée de la séquence, épissage alternatif). Le phénotype « ridé » des graines de petits pois utilisés par Mendel pour réaliser ses expériences est en fait le résultat de l'insertion d'un ET dans un gène codant pour une enzyme de ramification de l'amidon (Bhattacharyya *et al.*, 1990).

Les ET, porteurs d'éléments *cis*-régulateurs, peuvent aussi contribuer directement à l'expression des gènes en s'insérant dans des régions promotrices ou amplificatrices. Chez la vigne, l'insertion d'un rétrotransposon dans la région promotrice d'un gène impliqué dans la biosynthèse de pigment rouge serait à l'origine de l'apparition de cultivars blancs (Kobayashi *et al.*, 2003). Dans le génome humain, 4 % des gènes et près de 25 % des régions promotrices contiennent des séquences dérivées d'ET (Nekrutenko & Li, 2001 ; van de Lagemaat *et al.*, 2003).

Les fonctions des ET peuvent aussi être recrutées par le génome hôte via un processus évolutif appelé domestication. Ce phénomène a d'abord été décrit chez la drosophile avec l'élément *P* (Miller *et al.*, 1992) puis a été généralisé à d'autres organismes comme les plantes ou les animaux (Sinzelle *et al.*, 2009). Dans le génome humain, 47 gènes ont été identifiés comme dérivant, entièrement ou partiellement, de séquences codantes d'ET, principalement des transposases (Feschotte, 2008). Un des exemples les plus connus de domestication d'ET est le mécanisme de recombinaison V(D)J du système immunitaire des vertébrés. Ce mécanisme dépend de deux gènes dérivant d'un élément de classe II, *Transib* (Kapitonov & Jurka, 2005).

Les ET, en modifiant la structure du génome et en contrôlant l'activité des gènes, pourraient non seulement générer du polymorphisme génétique favorisant l'adaptation de populations mais aussi être à l'origine de phénomènes de spéciation (Figure 4). Cependant, il est difficile d'évaluer si des modifications de nombre ou d'activité d'ET au cours d'une période évolutive spécifique sont les conséquences ou les causes du processus de spéciation (Biémont, 2010).

3.5 Contrôle des ET par les génomes.

Chez *Drosophila melanogaster*, le taux d'insertion des copies d'ET est beaucoup plus important que le taux d'excision (Maside *et al.*, 2001), on devrait donc s'attendre à une accumulation d'ET qui aurait pour conséquence une augmentation constante de la taille du génome. Or ce n'est pas ce que l'on observe, il y a donc régulation du nombre de copies d'ET. Cette régulation peut s'effectuer via la sélection naturelle lorsque des insertions délétères sont contre-sélectionnées (Biémont *et al.*, 1997). Elle peut aussi s'effectuer via des mécanismes de contrôle du génome hôte.

Par exemple, l'inactivation des ET peut être exercée grâce à des mutations dirigées contre ces éléments. Un tel mécanisme d'inactivation des ET a été décrit chez le champignon filamenteux *Neurospora crassa*, il s'agit du RIP (*Repeat-Induced Point Mutations*) (Galagan *et al.*, 2003). Ce mécanisme se déroule durant la phase sexuée du champignon et va induire des mutations de type C:G vers T:A sur les deux copies de l'élément dupliqué. Le RIP va ainsi inactiver irréversiblement les ET et induire de la divergence entre les copies d'ET empêchant potentiellement toute recombinaison ultérieure (Galagan & Selker, 2004). Ce mécanisme semble très efficace puisque aucune copie d'ET est intacte dans le génome de *N. crassa* et aucune activité de transposition n'a été détectée (Galagan *et al.*, 2003).

Le génome peut aussi inactiver les ET sans altération de leur séquence en les soumettant à un contrôle épigénétique. Ce processus efficace a les avantages d'être à la fois transmissible à la descendance et réversible. Grâce à ce système, les ET représentent donc un réservoir potentiel de variabilité et peuvent être occasionnellement réactivés (Hua-Van *et al.*, 2011). Des explosions rapides d'amplification des ET ont été décrites dans plusieurs espèces, suggérant une perte temporaire du contrôle épigénétique de ces éléments (Rebollo *et al.*, 2010).

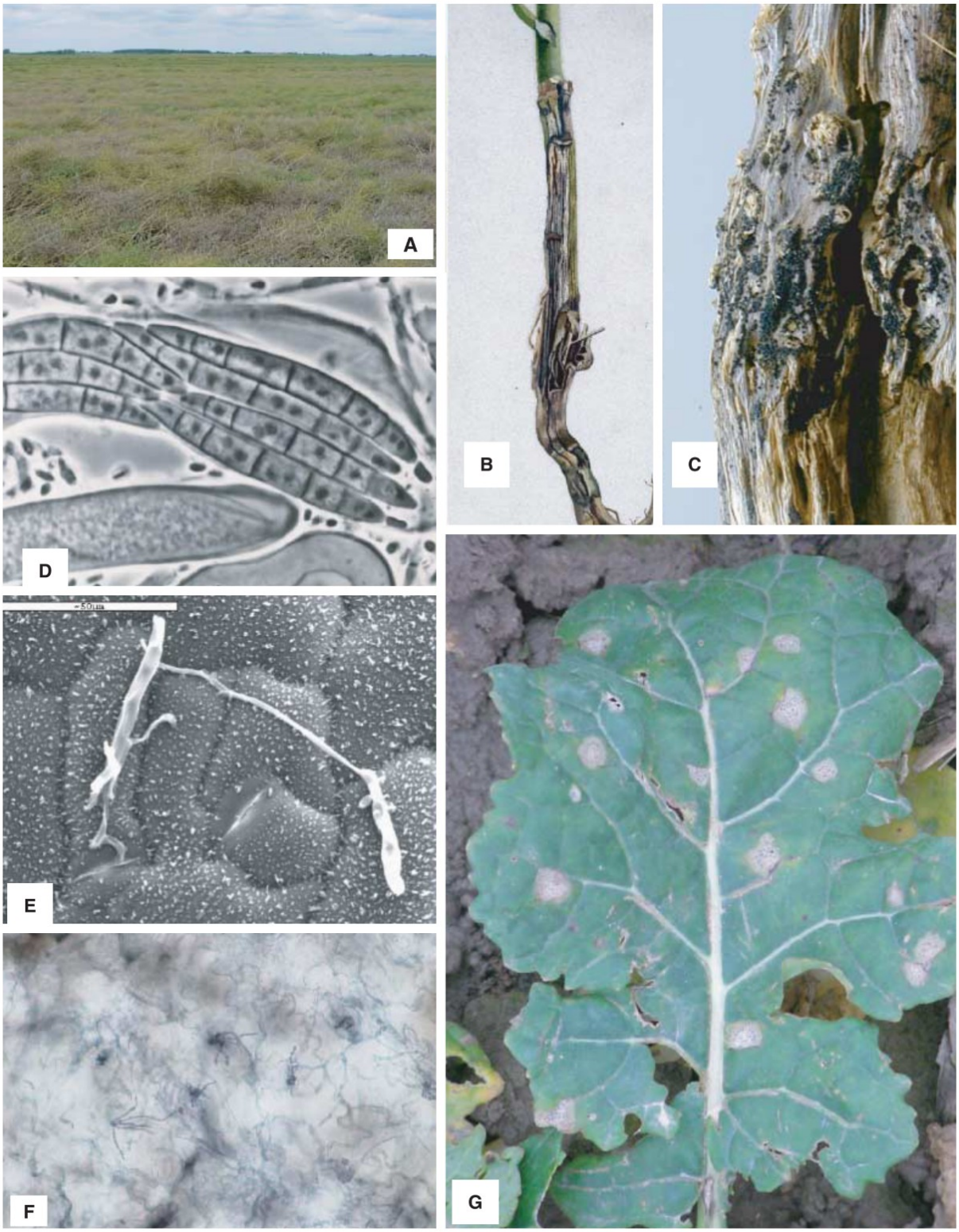


Figure 5. Symptômes et stades d'infection par *Leptosphaeria maculans*. (**A**) Verse parasitaire d'une culture de colza due à la nécrose du collet ; (**B,C**) Nécrose des tissus du collet et développement de périthèces sur des résidus de culture 9 mois après la récolte ; (**D**) Observation au microscope d'asques typiques contenant 8 ascospores ; (**E**) Ascospores germant sur une feuille de *Brassica napus* (microscope électronique à balayage) ; (**F**) Colonisation des espaces inter-cellulaires du mésophylle par du mycélium (coloration au bleu de trypan) ; (**G**) Macules foliaires sur feuille de *Brassica napus* après infection naturelle au champ. Adapté de Rouxel & Balesdent, 2005.

4 - *Leptosphaeria maculans*

Leptosphaeria maculans, forme parfaite de *Phoma lingam*, est un champignon filamenteux ascomycète de la classe des Dothidéomycètes. Cette classe est de loin la plus importante et la plus diversifiée au sein des Ascomycètes avec plus de 19 000 espèces répertoriées réparties dans 1300 genres (Kirk *et al.*, 2008). Les membres de cette classe couvrent tous les styles de vie et sont présents dans toutes les niches écologiques accessibles aux champignons. Certaines espèces existent sous formes lichénisées, certaines sont aquatiques, certaines sont des parasites de l'homme ou d'autres espèces du règne *Animalia.* Néanmoins la plupart d'entre elles sont des saprophytes terrestres. Les espèces les plus connues de cette classe font parties des genres *Cochliobolus* (syn. *Bipolaris*), *Phaeosphaeria*, *Zymoseptoria*, *Pyrenophora*, *Venturia*, *Pseudocercospora* et *Leptosphaeria* et sont des agents pathogènes de plantes cultivées (maïs, blé, orge, pommier, bananier, colza, etc.) occasionnant d'importantes pertes de rendement ou de qualité lors des récoltes (Schoch *et al.*, 2009b).

L. maculans est un agent pathogène des crucifères (*Brassicaceae*) et plus particulièrement des plantes du genre *Brassica* telles que *Brassica napus* (colza), *Brassica rapa* (navet), *Brassica juncea* (moutarde) ou *Brassica oleracea* (choux) (Rouxel & Balesdent, 2005). Ce champignon est responsable de la maladie la plus dommageable au colza : la nécrose du collet, qui peut entraîner localement la destruction de parcelles entières et des pertes de 5 à 20 % de la production au niveau national (Fitt *et al.*, 2006). Cette maladie est mondialement répandue et touche tous les pays producteurs de colza à l'exception de la Chine et de certains pays d'Europe de l'Est. L'ouverture des marchés à l'international et la globalisation de la sélection du colza sont pour ces pays sources de craintes quant au futur développement de *L. maculans* sur leurs territoires.

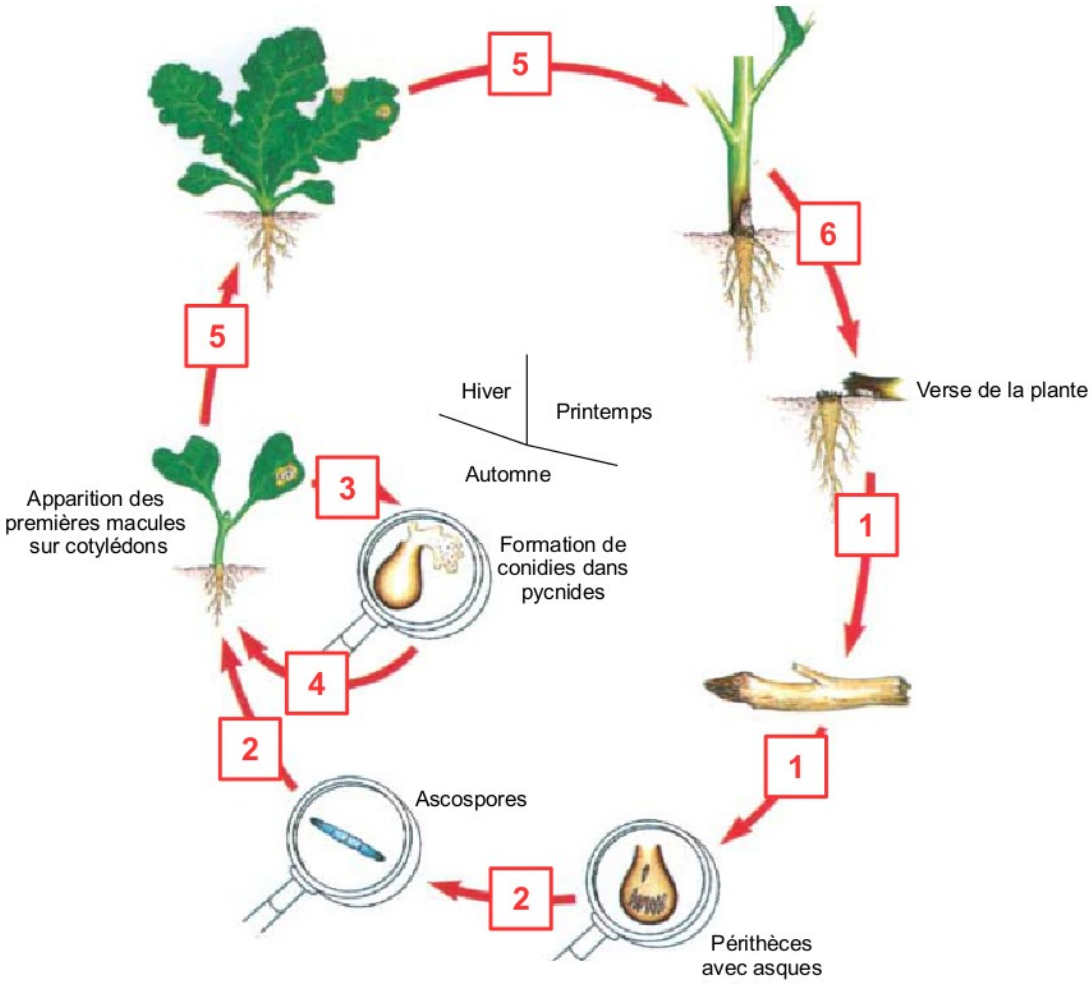


Figure 6. Cycle biologique de *Leptosphaeria maculans* en Europe. (**1**) phase saprophyte sur résidus de culture ; (**2**) reproduction sexuée et génération de l'inoculum primaire ; (**3,4**) courte phase nécrotrophe suivant l'infection primaire, apparition de macules foliaires, multiplication asexuée ; (**5**) phase endophyte avec colonisation systémique de la tige et du collet ; (**6**) phase nécrotrophe à l'origine de la verse de la plante. Adapté de CETIOM (1996).

4.1 Cycle biologique de *L. maculans*.

L. maculans est un agent phytopathogène hémibiotrophe présentant un cycle de vie complexe au cours duquel il va alterner plusieurs modes de nutrition, passant de saprophyte à nécrotrophe puis à biotrophe pour finir par être de nouveau nécrotrophe (Rouxel & Balesdent, 2005) (Figure 6). Son cycle de vie est étroitement lié à celui de sa plante hôte, le colza (*B. napus*). Le cycle est initié hors saison de culture lorsque *L. maculans* se développe de façon saprophytique sur les résidus de tiges infectées de colza récolté lors de la saison précédente (Figure 5C). Durant cette étape importante, le champignon effectue sa reproduction sexuée et produit des ascospores (Figure 5D), l'inoculum primaire permettant sa dispersion. Les ascospores sont ensuite libérées et disséminées par le vent après des baisses de températures et des précipitations, ce qui correspond à la période des semis et aux premières semaines de croissance des plants de colza en Europe. Ainsi, de jeunes plantules vont être les cibles principales sur lesquelles les ascospores vont germer et infecter la feuille (ou le cotylédon) en pénétrant les tissus via des ouvertures naturelles telles que les stomates ou des blessures (Figure 5E). Une fois à l'intérieur de son hôte, *L. maculans* va coloniser les tissus sans expression de symptômes visibles et croître exclusivement dans l'apoplaste pendant quelques jours avant de devenir nécrotrophe (Figure 5F). A ce stade, la mort des cellules infectées va former des macules foliaires (Figure 5G), non préjudiciables pour le rendement final, qui vont être le lieu de la reproduction asexuée et de la production de conidies (pycnidiospores, inoculum secondaire). Suivant cette courte phase nécrotrophe dans les feuilles, le champignon va entamer une phase endophytique pouvant durer jusqu'à 9 mois et ainsi coloniser de façon systémique la plante jusqu'à atteindre la tige. En fin de période culturale, *L. maculans* redevient nécrotrophe et détruit les tissus du collet (Figure 5B). Cette nécrose, en fragilisant la base de la plante, peut mener à la verse du plant de colza (Figure 5A) (West *et al.*, 2001) et aux pertes de rendement mentionnées ci-dessus.

4.2 Interactions *B. napus – L. maculans*.

Deux types de résistances sont décrits chez *B. napus* : la résistance quantitative (résistance générale) et la résistance qualitative (résistance spécifique). La résistance quantitative se met en place une fois l'infection établie et va conduire à un ralentissement de l'expression de la maladie (Delourme *et al.*, 2006). Cependant, les mécanismes à la base de la mise en place de cette (ou de ces) résistance(s) ne sont pas encore résolus et

le fait qu'ils impliquent probablement plusieurs gènes chez chaque partenaire compliquent grandement leur étude (Delourme *et al.*, 2006). La résistance spécifique dans le couple *L. maculans* / *B. napus* suit le modèle gène-pour-gène décrit par Flor (1955) (Ansan-Melayah *et al.*, 1995), ce qui implique la présence de gènes de résistance spécifiques (*R*) chez la plante et de gènes d'avirulence (*Avr*) correspondants chez l'agent pathogène. Si une plante possédant un gène de résistance est attaquée par un agent pathogène possédant le gène d'avirulence correspondant, on parle d'interaction incompatible, la plante est résistante et sera ainsi capable de se défendre contre son agresseur. Dans tous les autres cas, l'interaction est dite compatible, la plante est dite sensible et la maladie se propagera. Dans le couple *L. maculans* / *B. napus*, les gènes d'avirulence du champignon sont appelés *AvrLm* et les gènes de résistance correspondant chez la plante *Rlm*. A ce jour, 11 gènes d'avirulence ont été identifiés génétiquement chez *L. maculans* parmi lesquels 4 ont été clonés : *AvrLm1* (Gout *et al.*, 2006), *AvrLm6* (Fudal *et al.*, 2007), *AvrLm4-7* (Parlange *et al.*, 2009) et *AvrLm11* (Balesdent et *al.*, 2013).

4.3 Le complexe d'espèces *L. maculans-L. biglobosa*.

L. maculans fait partie d'un complexe comprenant au moins deux groupes génétiquement distincts initialement décrits comme virulent et avirulent (McGee & Petrie, 1978), agressif et non agressif (Hammond & Lewis, 1987), Tox^{+} et Tox^{0} (Balesdent *et al.*, 1992) ou groupe A et groupe B (Johnson & Lewis, 1994) en fonction des méthodes utilisées pour les discriminer. Grâce à l'avènement des outils moléculaires et à la description plus précise de différences morphologiques, en particulier concernant les périthèces, ces deux groupes ont été reconnus comme étant deux espèces différentes : *L. maculans* et *L. biglobosa* (Shoemaker & Brun, 2001). Sur colza, ces deux espèces ont le même cycle de vie, se développent sur les mêmes hôtes et peuvent être retrouvées simultanément sur la même plante. Elles sont toutes les deux capables de provoquer des lésions nécrotiques et de coloniser les tissus des plantes hôtes mais leurs impacts sur les cultures ne sont pas similaires : seule *L. maculans* est susceptible de provoquer la verse parasitaire et donc d'être dommageable pour les cultures (Williams & Fitt, 1999; West *et al.*, 2002; Rouxel *et al.*, 2004), *L. biglobosa*, quant à elle, va causer des lésions de la tige ne conduisant qu'à des dégâts mineurs à la culture. Chez *L. maculans*, de nombreuses interactions spécifiques (interaction gène-pour-gène) ont été montrées avec ses différents hôtes alors que pour le moment aucune interaction de ce type n'a été décrite chez *L. biglobosa*

(Vincenot *et al.*, 2008).

Des relations phylogénétiques entre les membres du complexe d'espèces ont été établies sur la base d'analyses de plusieurs séquences nucléiques (ITS, actine, β-tubuline), révélant un niveau de complexité inattendu en comparaison à la simple séparation en deux espèces (Mendes-Pereira *et al.*, 2003 ; Voigt *et al.*, 2005 ; Vincenot *et al.*, 2008). En effet, ces résultats montrent que *L. maculans* est une espèce très monomorphe avec uniquement une séparation en deux sous-clades dont un, *L. maculans* 'lepidii', est très rare et représenté seulement par quelques individus isolés sur *Lepidium* sp., une crucifère adventice. *L. biglobosa*, quant à elle, est une espèce très polymorphe avec une séparation en six sous-clades correspondant à des spécificités d'hôtes ou à différentes origines géographiques : *L. biglobosa* 'brassicae' présent sur plusieurs espèces de *Brassica* en Eurasie et aux Etats-Unis, *L. biglobosa* 'canadensis' retrouvé principalement sur les cultures de colza en Amérique et en Australie, *L. biglobosa* 'occiaustralensis' présent sur colza et chou en Australie et au Mexique, *L. biglobosa* 'australensis' présent sur colza en Australie, *L. biglobosa* 'thlaspii' isolé sur *Thlaspi arvense* au Canada et le très rare *L. biglobosa* 'erysimii' isolé sur *Erysimum* spp. au Canada (Figure 7) (Mendes-Pereira *et al.*, 2003 ; Voigt *et al.*, 2005 ; Vincenot *et al.*, 2008 ; Van de Wouw *et al.*, 2009 ; Dilmaghani *et al.*, 2009, 2012).

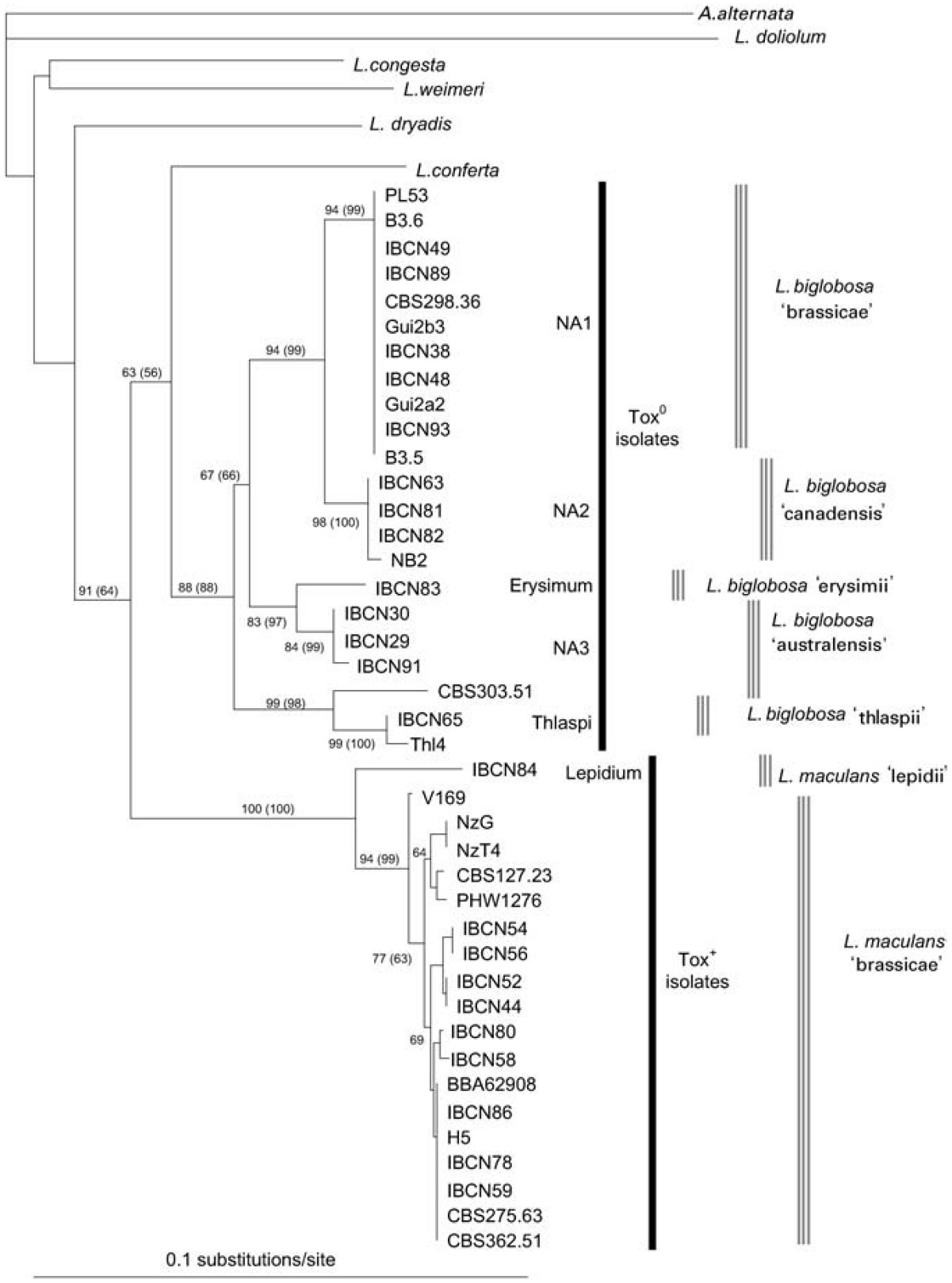


Figure 7. Phylogénie du complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa* et d'espèces *Leptosphaeria* proches. Arbre de distance réalisé à partir des séquences ITS1-5.8S et ITS2 du rDNA (Mendes-Pereira *et al.*, 2003). Un sous-clade supplémentaire de *L. biglobosa*, *L. biglobosa* 'occiaustralensis' proche de *L. biglobosa* 'canadensis' a été identifié par Vincenot *et al.* (2008).

4.4 Le pouvoir pathogène de *L. maculans*.

Chez *L. maculans*, très peu de gènes ont été décrits comme impliqués dans le pouvoir pathogène. Pour identifier ces gènes, une première approche consistait à rechercher des candidats en se basant sur les homologies de séquences avec ce qui était déjà connus chez d'autres espèces et sur des données d'expression *in planta* lors de l'interaction. Des gènes codant pour des ABC transporteurs, des enzymes de dégradation de la paroi cellulaire et des enzymes de détoxication ont été identifiés par cette approche mais aucune validation fonctionnelle de leur rôle dans la pathogenèse n'a été réalisée (Sexton *et al.*, 2000 ; Sexton & Howlett, 2000). Avec le développement des méthodes de transformation génétique et d'extinction des gènes chez *L. maculans*, certains gènes candidats ont pu être éteints mais aucun défaut dans le pouvoir pathogène n'a été observé chez ces transformants (Wilson *et al.*, 2002 ; Idnurm *et al.*, 2003a, 2003b). D'autres gènes candidats ont permis l'identification de clusters de gènes codant des métabolites secondaires, et notamment celui de la sirodesmine PL. La sirodesmine PL est une phytotoxine produite par *L. maculans* mais pas *L. biglobosa*. Elle appartient à une classe de métabolites secondaires, les epipolythiodioxopiperazine (ETP), produits par des champignons parmi lesquels on retrouve des agents pathogènes de plantes et d'animaux (Gardiner *et al.*, 2004). Cette toxine possède des propriétés antivirales et antibactériennes et est essentielle pour le pouvoir pathogène de *L. maculans* sur les tiges de *B. napus* (Elliott et *al.*, 2007). Une autre phytotoxine, la phomalide, a été identifiée mais sa production ne semble pas essentielle au pouvoir pathogène de *L. maculans* (Elliott *et al.*, 2013).

Une autre stratégie très intéressante pour identifier de nouveaux déterminants du pouvoir pathogène chez *L. maculans* a été la production d'une large collection de mutants par ATMT (*Agrobacterium tumefaciens*-mediated transformation) associée à du génotypage et du phénotypage. Les mutants ainsi générés montrant une défaillance dans leur pouvoir pathogène ont permis d'identifier des gènes, dont certains sont impliqués dans le métabolisme primaire du champignon (Idnurm & Howlett, 2002 ; Elliott & Howlett, 2006 ; Rémy *et al.*, 2008a, 2008b ; Van de Wouw *et al.*, 2009).

Une autre approche, le clonage positionnel, a permis l'identification et la caractérisation de quatre gènes d'avirulence, *AvrLm1*, *AvrLm4-7*, *AvrLm6* et *AvrLm11*, codant des PPS dont les fonctions moléculaires n'ont pas encore été définies (Gout *et al.*, 2006 ; Fudal *et al.*, 2007 ; Parlange *et al.*, 2009 ; Balesdent et *al.*, 2013). Ces gènes ont tous la particularité d'être localisés dans des compartiments génomiques riches en bases

AT, composées principalement d'ET. Cette localisation, alors inédite, ainsi que l'importance du rôle du produit de ces gènes lors de la mise en place de la pathogenèse sont les bases qui ont mené au séquençage du génome de *L. maculans*.

5. Projet de thèse.

Aujourd'hui, avec les nouvelles générations de technologies de séquençage et d'assemblage, obtenir la séquence d'un génome est devenu possible rapidement et à moindre coût, ce qui permet le séquençage en masse de plusieurs dizaines de génomes simultanément. Toutefois, si l'on retourne un peu moins de 10 ans en arrière, produire le génome d'un organisme demandait beaucoup plus d'investissement en temps et en argent, et le choix de l'organisme à séquencer était très important et conduisait à des débats souvent passionnés au sein des instituts tels que l'INRA ou de la communauté scientifique internationale. Ainsi, si le séquençage systématique des virus puis des bactéries phytopathogènes a rapidement été établi de façon routinière, celui des champignons phytopathogènes a été lent à s'initier malgré les problèmes importants qu'ils causent en agriculture (et en particulier en grandes cultures).

Ainsi, à l'initiative des équipes concernées et grâce à la mise en place du centre national de séquençage français, le Genoscope, deux modèles étudiés au sein de l'INRA ont été initialement mis en avant : *Botrytis cinerea* et *Leptosphaeria maculans.* Si *B. cinerea* résultait en partie d'un choix culturel en tant qu'agent pathogène de la vigne, le choix de *L. maculans,* fortement soutenu par une vaste communauté internationale et nationale, tant scientifique qu'appliquée, était plus lié à son importance agronomique et à ses capacités adaptatives conduisant à de rapides cycles de *boom-and-bust*. *L. maculans* étant un agent pathogène responsable de pertes massives de rendement dans les cultures de colza, on espérait tout d'abord du séquençage de son génome une identification plus aisée des gènes dont les produits jouent un rôle direct dans les interactions avec son hôte. En effet, la protection des cultures de colza face à *L. maculans* repose principalement sur l'utilisation de gènes de résistance spécifiques ciblant des gènes d'avirulence. La séquence du génome et la mise en place d'outils d'annotation dédiés devaient ainsi permettre l'identification de tous les gènes d'avirulence correspondant aux nouvelles résistances. Ce postulat initial a d'ailleurs été validé au cours de ma thèse par l'équipe d'accueil et le clonage positionnel de gènes d'avirulence fortement facilité par mes travaux. D'un point de vue biologique, l'intérêt était tout aussi important. *L. maculans* est un parfait représentant des stratégies d'infection mises en place par les Dothidéomycètes et a montré au cours du temps de grandes capacités

d'adaptation, notamment grâce à certaines caractéristiques comme par exemple son cycle de vie très complexe alternant différents types de nutrition et de reproduction. De plus, les premières données génomiques (séquençage de clones BACs comprenant trois gènes d'avirulence cibles) montraient deux aspects très importants et potentiellement spécifiques de *L. maculans* : (i) tous les gènes d'avirulence caractérisés étaient localisés dans de grandes régions riches en bases AT et composés d'ET, (ii) ces régions riches en AT préfiguraient une structure génomique particulière, qui, si elle se généralisait à l'ensemble du génome, aurait été totalement inédite chez un micro-organisme eucaryote.

Mes travaux de thèse s'inscrivent dans un contexte d'essor du séquençage des génomes fongiques amorcé quelques années auparavant avec les publications des premiers génomes de champignons filamenteux (*Neurospora crassa*, *Aspergillus nidulans*) et phytopathogènes (*Magnaporthe oryzae*, *Phaeosphaeria nodorum*). Mes travaux de Master 2 puis ceux de thèse ont eu lieu à une période charnière du séquençage des champignons filamenteux, au moment où les premiers génomes étaient obtenus, mais très peu encore publiés, et où de nombreux outils étaient en cours d'élaboration pour l'assemblage, l'annotation et l'exploitation des données génomiques. Grâce à l'obtention du génome de référence de *L. maculans*, puis du séquençage de plusieurs espèces proches formant une série évolutive, l'objectif de ma thèse était de mettre en avant, à l'aide d'une approche de génomique puis de génomique comparative, les relations entre l'évolution de la structure et du fonctionnement du génome des membres du complexe d'espèces *L. maculans-L. biglobosa,* et leur pouvoir pathogène ou leur adaptabilité à l'hôte.

Suite à l'introduction présentant les champignons et le développement des approches génomique sur ces organismes, puis les ET et leurs rôles dans l'évolution des génomes puis enfin le modèle d'étude sur lequel je travaille, mon manuscrit est composé de trois parties : (i) les résultats obtenus au cours de ma thèse qui sont présentés dans les deux premiers chapitres, (ii) une discussion générale et perspectives, (iii) des annexes comportant les articles sur des sujets liés mais extérieurs à mon projet de thèse auxquels j'ai participé, ainsi qu'un bilan d'activité listant les différentes présentations effectuées au cours de ma thèse.

Le chapitre 1 présente le génome de *Leptosphaeria maculans* en se focalisant plus particulièrement sur sa structure et l'impact potentiel de celle-ci sur la diversification et l'évolution de protéines jouant un rôle essentiel lors de l'interaction agent pathogène-plante, les effecteurs. Ce chapitre est composé de l'article « Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations » publié le 15 février 2011 dans le journal en ligne *Nature Communications*.

Le chapitre 2 présente tout d'abord une étude de génomique comparative et évolutive au sein du complexe d'espèces *L. maculans-L. biglobosa* en se focalisant sur les relations entre les différents niveaux d'adaptation au colza de chacun de ses membres et l'invasion des génomes par des éléments transposables. Cette partie est composé de l'article «Transposable Element-assisted evolution and adaptation within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal plant pathogens» en cours de finalisation pour publication dans la revue scientifique *Genome Research*. La seconde partie de ce chapitre présente une analyse comparative des génomes de quatre isolats diversifiés de *L. maculans* 'brassicae', dont le but est d'évaluer les divergences génétiques et leurs conséquences sur les gènes impliqués dans la pathogenèse en se focalisant sur le répertoire d'effecteurs de chaque génome.

La discussion-conclusion est organisée sous la forme d'une publication de synthèse illustrant l'incidence que peuvent avoir les ET sur l'évolution du génome d'un champignon phytopathogène. Cette synthèse fait l'objet d'un article intitulé « Incidence des Éléments Transposables sur l'évolution des génomes des champignons phytopathogènes et leur potentiel adaptatif » soumis dans la revue scientifique française *Biologie Aujourd'hui*. Ce chapitre comporte aussi les perspectives découlant des différents résultats et données accumulés lors de mes recherches, ainsi que mon projet post-doctoral.

CHAPITRE 1

Le génome de *Leptosphaeria maculans* ‘brassicae’

Ce chapitre est composé de l'article « Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations » publié le 15 février 2011 dans la revue scientifique en ligne *Nature Communications*. Cet article décrit le génome de *L. maculans* en se focalisant plus particulièrement sur sa structure et l'impact potentiel de celle-ci sur la diversification et l'évolution de protéines jouant un rôle essentiel lors de l'interaction agent pathogène-plante, les effecteurs.

Chez *L. maculans*, la caractérisation de gènes codant des effecteurs correspondant à trois gènes d'avirulence *AvrLm1*, *AvrLm6* et *AvrLm4-7* avait mis en évidence une localisation inédite de ces derniers dans de grandes régions riches en bases AT. La première analyse de génomique utilisant le génome complet de *L. maculans* consistait donc en l'identification de l'ensemble des régions riches en AT puis visait à prédire dans ces régions des gènes codant des protéines aux caractéristiques proches des effecteurs connus, c'est à dire des protéines de petite taille potentiellement sécrétées (PPS). Cette analyse a permis de généraliser la présence de ces régions riches en AT dans l'ensemble du génome, faisant ainsi de *L. maculans* le premier micro-organisme à avoir une telle structure de génome. C'est aussi le premier champignon Ascomycète dont le répertoire des effecteurs potentiels a été publié. D'ailleurs, la structure du génome et ce répertoire sont très liés puisque nous montrons que les régions riches en AT sont enrichies en gènes codant des PPS.

Si le génome de *L. maculans* a cette structure particulière, c'est parce qu'il a été envahi par des ET qui sont les principaux composants des régions riches en AT. Au cours de l'évolution, les champignons Ascomycètes ont développé un mécanisme de défense, le RIP (*Repeat-Induced Point mutations*), contre ce type d'éléments répétés considérés comme parasites par la cellule. Chez *L. maculans*, nous montrons que les séquences codantes identifiées au sein des régions riches en AT, donc riches en ET, présentent aussi des signatures de RIP. Ce qui signifie que dans le génome de *L. maculans*, les ET sont une source, à travers le mécanisme du RIP, de diversification pour les gènes proches de ces éléments. Le fait de garder des gènes codant des protéines importantes pour le pouvoir pathogène dans des régions hautement dynamiques (au niveau mutationel) pourrait permettre à *L. maculans* de s'adapter plus rapidement aux différentes pressions de sélection exercées contre lui.

Pour cet article, dont je suis co-premier auteur, j'ai contribué à (i) l'identification des régions riches en AT dans l'ensemble du génome, (ii) à la prédiction de gènes codant des effecteurs dans ces régions, (iii) à l'analyse de ces gènes (indices de RIP, usage des codons) et de leur produit (composition, annotation fonctionnelle, recherche de motifs de translocation), (iv) à l'identification des gènes codant des effecteurs dans les autres régions du génome et à ainsi établir le répertoire complet des effecteurs de la souche v23.1.3 *L. maculans* 'brassicae', (v) à l'annotation de l'ADN ribosomique, et (vi) à une évaluation de la dynamique des ET et à la datation des événements de transposition.

Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations

Thierry Rouxel[1*+], Jonathan Grandaubert[1+], James K. Hane [2], Claire Hoede[3], Angela P. van de Wouw[4], Arnaud Couloux[5], Victoria Dominguez[3], Véronique Anthouard[5], Pascal Bally[1], Salim Bourras[1], Anton J. Cozijnsen[4], Lynda M. Ciuffetti[6], Alexandre Degrave[1], Azita Dilmaghani[1], Laurent Duret[7], Isabelle Fudal[1], Stephen B. Goodwin[8], Lilian Gout[1], Nicolas Glaser[1], Juliette Linglin[1], Gert H. J. Kema[9], Nicolas Lapalu[3], Christopher B. Lawrence[10], Kim May[4], Michel Meyer[1], Bénédicte Ollivier[1], Julie Poulain[5], Conrad L. Schoch[11], Adeline Simon[1], Joseph W. Spatafora[6], Anna Stachowiak[12], B. Gillian Turgeon[13], Brett M. Tyler[10], Delphine Vincent[14], Jean Weissenbach[5], Joëlle Amselem[3], Hadi Quesneville[3], Richard P. Oliver[15], Patrick Wincker[5], Marie-Hélène Balesdent[1], Barbara J. Howlett[4]

Published in *Nature Communications* **2:**202

---

[1] INRA-Bioger, UR1290, Avenue Lucien Brétignières, BP 01, F-78850 Thiverval-Grignon, France; [2] Murdoch University, South Street, Murdoch, WA 6150, Australia; [3] INRA-URGI, Route de Saint Cyr, 78026 Versailles Cedex, France; [4] School of Botany, The University of Melbourne, Vic 3010, Australia; [5] GENOSCOPE, Centre National de Séquençage, Institut de Génomique CEA/DSV, 2, rue Gaston Crémieux, CP 5706, F-91057 Evry Cedex, France; [6] Department of Botany and Plant Pathology, Cordley Hall 2082, Oregon State University, Corvallis, OR 97331-2902, USA; [7] Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, 43 Bld du 11 Novembre 1918, F-69622 Villeurbanne cedex, France, [8] USDA-ARS, Crop Production and Pest Control Research Unit, Purdue University, 915 West State Street, West Lafayette, IN 47907-2054, USA; [9] Wageningen UR, Plant Research International, Dept. Biointeractions and Plant Health, P.O. Box 69, 6700 AB Wageningen, the Netherlands; [10] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0477, USA; [11] NIH/NLM/NCBI, 45 Center Drive, MSC 6510, Bethesda, MD 20892-6510, USA; [12] Institute of Plant Genetics, Polish Academy of Sciences, Strzeszynska 34, PL-60479, Poznan, Poland; [13] Dept. of Plant Pathology & Plant-Microbe Biology, Cornell University, Ithaca NY 14853 USA, [14] INRA, UMR1202 BIOGECO, 69 Route d'Arcachon, F-33612 Cestas, France, [15] Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, WA 6845, Australia

* to whom correspondence should be addressed: E-mail: rouxel@versailles.inra.fr
[+] These authors contributed equally to this work

Abstract

Fungi are of primary ecological, biotechnological and economic importance. Many fundamental biological processes shared by animals and fungi, are studied in fungi due to their experimental tractability. Many fungi are pathogens or mutualists. They are model systems to analyse effector genes and their mechanisms of diversification. Here we report the genome sequence of the phytopathogenic ascomycete *Leptosphaeria maculans* and characterize its repertoire of protein effectors. The *L. maculans* genome has an unusual bipartite structure – alternating distinct GC-equilibrated and AT-rich blocks of homogenous nucleotide composition. The AT-rich blocks comprise one third of the genome and contain effector genes and families of transposable elements, both of which are affected by Repeat Induced Point mutation (RIP), a fungal-specific genome defense mechanism. This genomic environment for effectors promotes rapid sequence diversification and underpins the evolutionary potential of the fungus to adapt rapidly to novel host-derived constraints.

Introduction

Fungi are the most important pathogens of cultivated plants, causing about 20% yield losses worldwide. Such diseases are a major cause of malnutrition worldwide[1]. Their phenotypic diversity and genotypic plasticity enables fungi to adapt to new host species and farming systems and to overcome new resistance genes or chemical treatments deployed in attempts to limit losses to crop yields[2]. Along with such genotypic plasticity, natural or anthropogenic long-distance dispersal of fungi allows the emergence of novel, better-adapted phytopathogens and more damaging diseases. These processes of adaptation are exemplified by *Leptosphaeria maculans* 'brassicae' (Phyllum Ascomycota, class Dothideomycetes), which causes stem canker (blackleg) of oilseed rape (*Brassica napus*) and other crucifers. This fungus has been recorded on crucifers (mainly cabbages) since 1791, but only began to cause important damages to broad acre Brassica species and spread around the world in the last four decades[3]. Other phytopathogens often rapidly cause lesions on plants to ensure asexual reproduction. In contrast, *L. maculans* shows an unusually complex parasitic cycle with alternating saprotrophy associated with sexual reproduction on stem debris, necrotrophy and asexual sporulation on leaf lesions, endophytic and symptomless systemic growth, and a final necrotrophic stage at the stem base[3].

Some features of filamentous fungal genomes are remarkably constant; for instance size (20-60 Mb typically about 34 Mb), gene number (10,000 to 13,000), gene content, intron size and number, and the low content of repeated sequences[4]. Comparative genomic approaches have shown that most of candidate "pathogenicity genes" (*e.g.* those encoding hydrolytic enzymes that can degrade plant cell walls, or involved in formation of infection structures) analyzed in the last decade in a gene-by-gene approach are shared by saprobes and pathogens[4]. These genes were probably recruited as pathogenicity factors when phytopathogens evolved from saprobes, but they do not account for host range or host specificity of phytopathogens. Such roles are played by 'effector' proteins, which modulate host innate immunity, enable parasitic infection, and are generally genus, species, or even isolate-specific[5,6]. Such effector genes include those with primary function as avirulence genes or encoding toxins or suppressors of plant defense. While bacteria produce few effectors (typically less than 30), which mostly appear to suppress plant innate immunity[7], hundreds of candidate effectors have been identified in oomycetes[8-10]. In fungi, in contrast, such a catalogue of effectors has only been established to-date in the hemibasidiomycete pathogen of maize, *Ustilago maydis*, in which many of the effector

genes are organized as gene clusters[11].

In *L. maculans*, the only characterized effectors include a toxic secondary metabolite, sirodesmin PL[12] and the products of three avirulence genes, *AvrLm1*, *AvrLm6* and *AvrLm4-7*, of which at least one, *AvrLm4-7*, is implicated in fungal fitness[13-15]. These three avirulence genes show typical features of effector genes, *i.e.*, they are predominantly expressed early in infection, encode small proteins predicted to be secreted (SSPs) into the plant apoplast and have no or few matches in databases. Intriguingly, all three are located within large AT-rich, heterochromatin-like regions that are mostly devoid of other coding sequences[13,15].

In this paper we describe the genome of *L. maculans*. We speculate how the genome, characterized by a distinct division into GC-equilibrated and AT-rich blocks of homogenous nucleotide composition, has been reshaped following massive invasion by and subsequent degeneration of transposable elements (TEs). We also predict the repertoire of pathogenicity effectors for the first time in an ascomycete genome and we propose how the unusual genome structure may have led to the diversification and evolution of effectors.

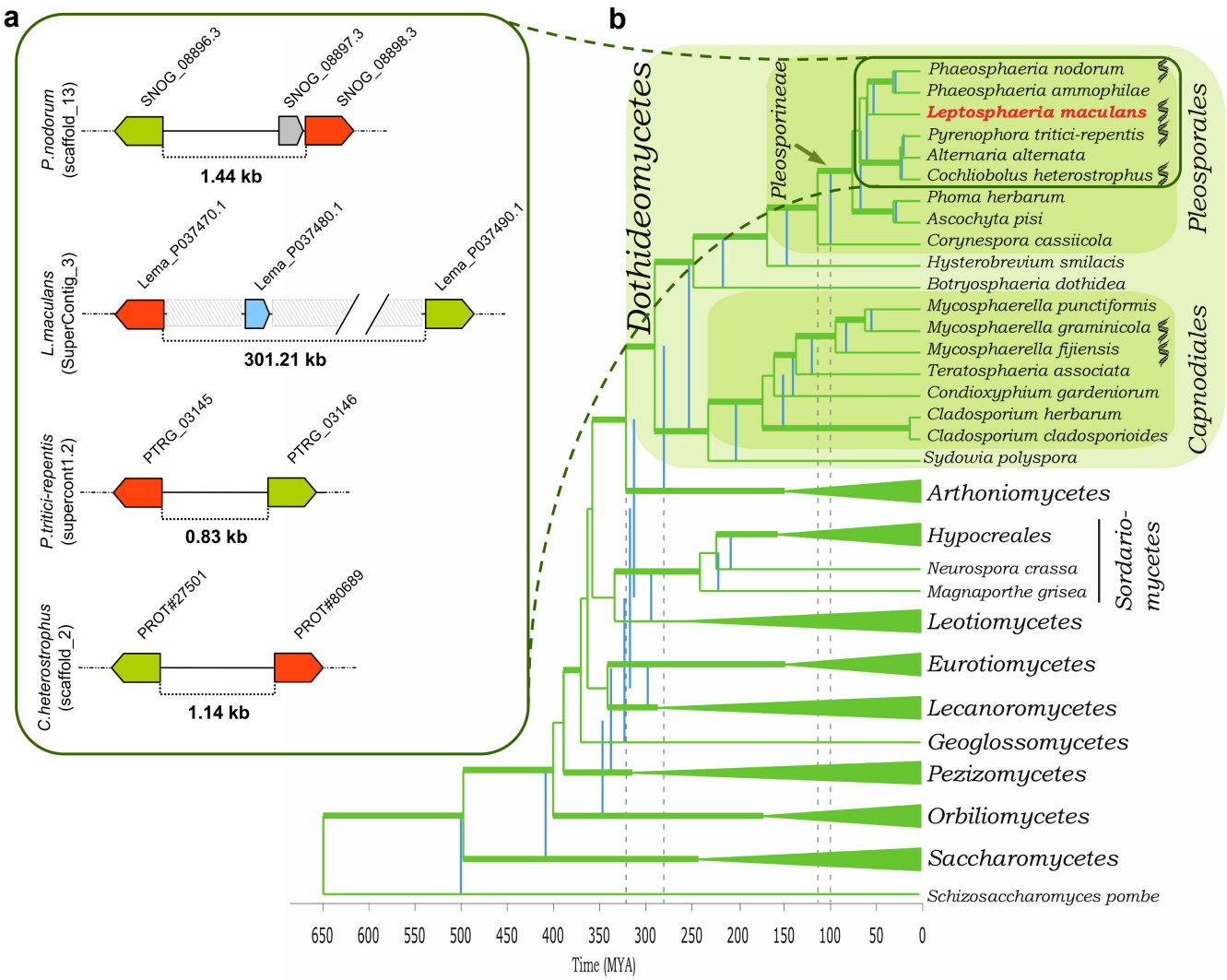


Figure 1. Phylogenetic relationships between Dothideomycetes including an example of microsynteny between *Leptosphaeria maculans* and related species. (**a**) An example of microsynteny between *Leptosphaeria maculans* and closely related Dothideomycetes, *Phaeosphaeria nodorum*, *Cochliobolus heterostrophus* and *Pyrenophora tritici-repentis*, showing the integration of an AT-rich genomic region (grey boxes) between two orthologous genes encoding for fungal transcription factors (red and green arrows) of the three other species, along with generation of one novel Small Secreted Protein-encoding gene (blue arrow) in *L. maculans* only. Grey arrow, *P. nodorum* predicted gene. The ID of each gene in the corresponding genome sequence is indicated. The intergenic distance (expressed in kb) is shown. (**b**) A phylogenetic tree and estimated time divergences of major lineages in Ascomycota with a selection of plant pathogenic lineages in Dothideomycetes. The phylogenetic analysis was done using RaxML[44] and the chronogram, calibrated using recent data from the literature and fossil dates, produced using r8s[45]. Classes outside of the Dothideomycetes were collapsed in TreeDyn except for Sordariomycetes where the order *Hypocreales* represented an important calibration point. The blue vertical lines correlate with divergence times when the root of the tree was fixed at 500 MYA while the green lines of the tree represent a fixed root of 650 MYA. The range of dates for the emergence of Dothideomycetes and *Pleosporineae* are highlighted with stippled lines. Thickened branches on the tree represents nodes that had more than 70% bootstrap values in a RAxML run. Species with genome data are marked with a DNA logo.

Results

General features of the *Leptosphaeria maculans* genome

The haploid genome of strain v23.1.3 of *L. maculans* 'brassicae' was sequenced using a whole-genome shotgun strategy. This fungus is closely related to *Phaeosphaeria (Stagonospora) nodorum*, *Pyrenophora tritici-repentis* and *Cochliobolus heterostrophus,* as seen in the phylogeny based on sequence analysis of a range of genes (Supplementary Table S1; Figure 1a). The genome assembly had a total size of 45.12 Mb, scaffolded into 76 SuperContigs (SCs) (30 large SCs > 143 kb) (Tables 1, 2; Supplementary Table S2). The correspondence of SCs to chromosomes was inferred by a combination of approaches (Figure 2; Supplementary Figures S1, S2). Conglomerated data are consistent with the presence of 17 or 18 chromosomes, ten of which correspond to single SCs (Supplementary Figure S1; Supplementary Table S2).

Table 1. Assembly statistics for the *L. maculans* genome.

	SuperContigs	Contigs
Number	76	1743
Size (Mb)	45,12	43,76
N50 (kb)	1770	61
Min / Max size (kb)	0.49 / 4258.57	0.22 / 395.37
Mean size (kb)	594	26
Median size (kb)	29	11

Gene models were identified using the EuGene prediction pipeline (Supplementary Tables S3, S4), and the resultant total of 12,469 genes is consistent with that in other Dothideomycetes (Table 2). Expression of 84.4% of predicted genes was detected using NimbleGen custom-oligoarrays in free-living mycelium or during early stages of oilseed rape infection (Table 3). About 10% of the genes were significantly over-expressed during infection (Table 3). Taking into account EST, transcriptomic, and proteomic support, 84.8% of the gene models were biologically validated (Table 3). The genes are shorter than those in the other Dothideomycetes whose genomes have been sequenced (Table 2). Intergenic distances are shorter than those of *P. nodorum*, the closest relative to have been sequenced, and bi-directional promoters are common (Supplementary Table S5).

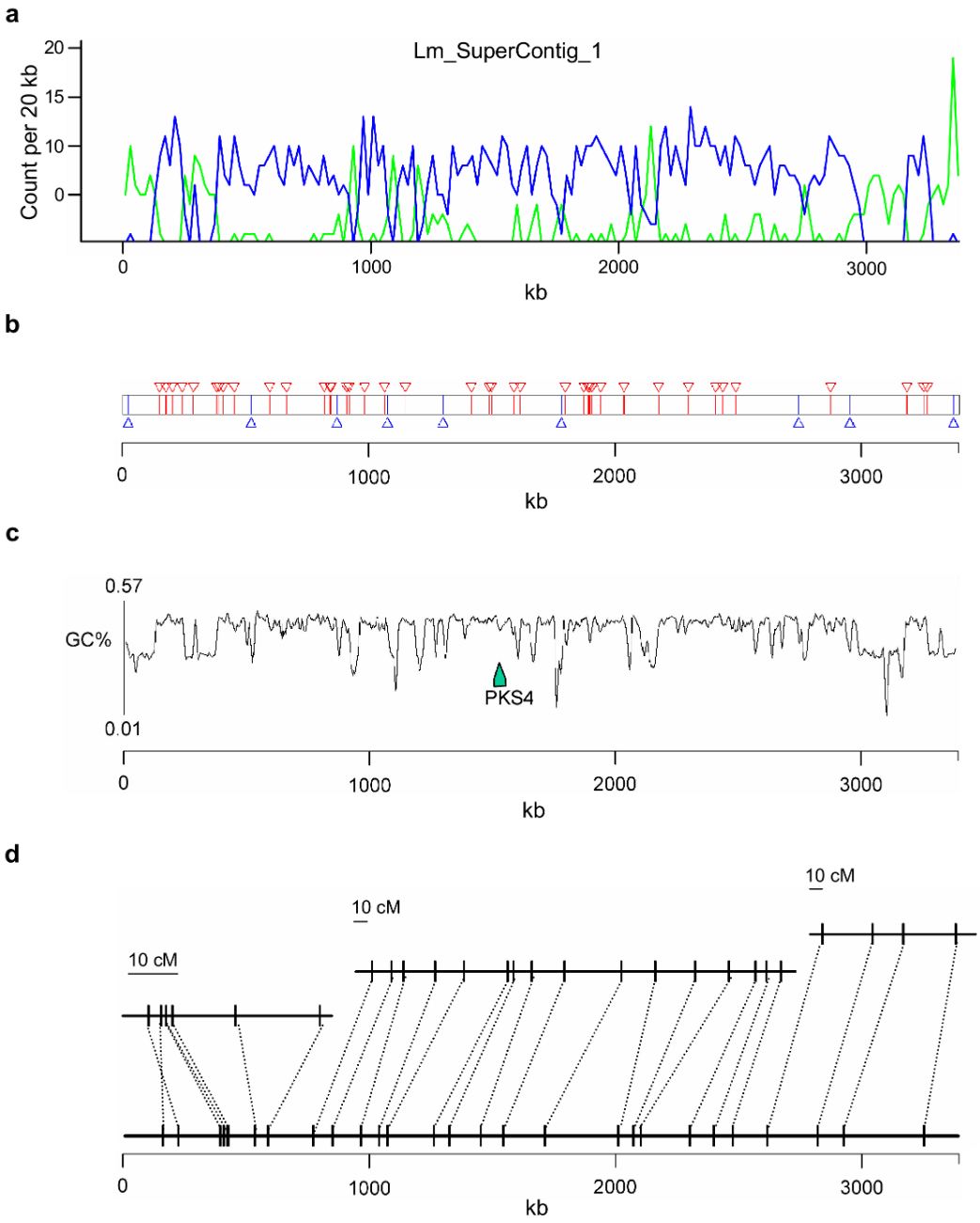


Figure 2. Main features of the *Leptosphaeria maculans* genome as exemplified by SuperContig_1 (3.38 Mb), which corresponds to a complete chromosome. (**a**) Transposable elements (TE) distribution and gene density along the SC. TE density is drawn in green and gene density is in blue. (**b**) location of SSPs (small-secreted protein encoding genes). Blue arrowheads, SSP in AT-blocks, corresponding to TE-rich regions in a; red arrowheads, SSP in GC-blocks, corresponding to gene-rich genome regions in a. (**c**) GC content along the SC showing alternating GC-equilibrated and AT-rich regions, with location of a polyketide synthase-encoding gene, PKS4. (**d**) genetic (upper part, expressed in centiMorgan-cM) to physical (expressed in kb) distance relationship as a function of the isochore-like structure. Lower part: physical location of genetic markers. Upper part of the panel: genetic map using MapMaker/Exp 3.0 with parameter set at LOD value > 3.0 and minimum distance = 20 cM. Only markers drawn from the sequence data are represented.

Automated finding and annotation of repeated elements in the genome using the REPET pipe-line showed they comprise one third of the genome compared to 7% in *P. nodorum* (Table 2). Although most of the repeat elements are truncated and occur as mosaics of multiple families, their origin as TEs is clear. Class I elements (see [16] and Table 4 for classification of TEs), dominate with nine families comprising 80% of the repeated elements (Table 4). Of these, just four families comprise 11.37 Mb, which is 25% of the genome assembly. Very few, if any, of the TEs are transcribed, as shown by EST inspection and transcriptomic analysis. TEs are clustered in blocks distributed across SCs and the number of TE copies per SC correlates with size of the SC (R^2 = 0.86) (Supplementary Figure S3a).

Table 2. Features of genomes of *L. maculans* and other closely related Dothideomycetes.

	L. maculans [a]	*P. (Stagonospora) nodorum* [a]	*P. tritici-repentis* [a]	*C. heterostrophus* [a]	*A. brassicicola* [a]	*M. graminicola* [a]
No. of chromosomes	17-18	19	11	15-16	9-11	21
Genome size (Mb)	45,1	36,6	37,8	34,9	30,3	39,7
No. of contigs	1743	496	703	400	4039	21
No. of SuperContigs (SCs)	76	107	47	89	838	21
SC N50 (Mb)	1,8	1,1	1,9	1,3	2,4	n/a [b]
Gaps (%)	2,5	0,4	1,7	1,1	5,4	0,01
No. of predicted genes	12469	10762	12141	9633	10688	10952
Average gene length (bp)	1323	1326	1618	1836	1523	1600
GC content (%)	44,1	50,3	50,4	52-54	50,5	55
Repeat content (%)	34,2	7,1	16	7	9	18
'Core' genome size (Mb) [c]	29,7	34,5	31,7	32,5	27,6	32,6
Gene density/core genome (no. of gene per 10 kb)	4,2	3,1	3,8	3	3,9	3,4

[a] References for the genomes as follows: *L. maculans*[54], *P. nodorum*[55], *P. tritici-repentis*[56], *C. heterostrophus*[57], *A. brassicicola*[58], *M. graminicola*[59]. Unpublished reannotation of *P. nodorum* genome was provided by J. Hane and R. Oliver

[b] Not applicable since the *M. graminicola* genome is finished *i.e.*, each SC corresponds to a chromosome

[c] 'Core' genome excluding the repeated elements, but including the gaps in the genome sequence

The Transposable Elements are RIP-affected

Alignment and comparison of repeat families also showed a pattern of nucleotide substitution consisting mainly of C-to-T and G-to-A changes, suggesting the presence of Repeat-Induced Point mutation (RIP). RIP is a premeiotic repeat-inactivation mechanism specific to fungi and has been previously experimentally identified in *L. maculans*[17]. The *L. maculans* genome possesses orthologs of all the *Neurospora crassa* genes currently postulated to be necessary for RIP[18] (Supplementary Table S6). Analysis using RIPCAL, a quantitative alignment-based method[19], indicated that C bases within CpA dinucleotides were mutated to T more frequently than the sum of CpC, CpG and CpT dinucleotides confirming the action of RIP on all of the TEs (Supplementary Figures S4, S5).

The compartmentalized genome of *Leptosphaeria maculans*

The *L. maculans* genome is larger and has a lower overall GC content (44.1% GC) than those of the related Dothideomycetes *P. nodorum*, *Alternaria brassicicola*, *C. heterostrophus*, *P. tritici-repentis* or the more divergent species *Mycosphaerella graminicola* (Table 2). As previously reported for a broader range of fungi[20], the larger size is consistent with the genome having been extensively invaded by TEs. The GC content of ESTs and other known coding sequences is 50.5% and the low genome GC content is due to the compartmentalized structure of the genome into GC-equilibrated regions (51.0% GC content, sizes between 1 to 500 kb, average 70.4 kb) (henceforth denoted as GC-blocks) alternating with AT-rich regions (henceforth denoted as AT-blocks) (averaging 33.9% GC content; with sizes between 1 and 320 kb, average of 38.6 kb). Whole-genome analysis identified 413 AT-blocks and 399 GC-blocks (Supplementary Table S7). The AT-blocks cover 36% of the genome and are distributed within the large SCs, comprising between 23.1 and 49.2% (Figure 2c, Supplementary Figures S2b, S3b, S6; Table S7). SC22, corresponding to a minichromosome[21] contains nine AT-blocks amounting to 92.5% of the SC (Supplementary Table S7).

As well as differences in GC content, the two types of genomic regions are dissimilar in terms of recombination frequency and gene content. The number of crossovers (CO) along a chromosome ranges between 1.16 to 3.31, depending on size of the chromosome, with one CO every 820 kb on average. The recombination frequency is significantly higher between marker pairs located within GC-blocks than those located on each side of one AT-block (F Fisher = 5.873, P=0.019) (Figure 2d, Supplementary Figure

S7).

GC-blocks contain 95% of the predicted genes of the genome, at a higher density (4.2 per 10 kb) than in other Dothideomycetes (Table 2) and are mostly devoid of TEs. In contrast, AT-blocks are gene-poor, comprising only 5.0% of the predicted coding sequences, and mainly contain mosaics of TEs mutated by RIP, thus resulting in a low GC content of TEs. There are three categories of AT-blocks: telomeres, which include a *Penelope* retroelement[22] (Supplementary Figure S8); large AT-blocks (216 sized 13-325 kb); and mid-sized AT-blocks (197 sized 1-13 kb) (Supplementary Figure S9), mostly corresponding to single integrations of only two families of DNA-transposons (Supplementary Table S8).

In almost half of the cases where pairs of orthologs are on the same SC, the genes flanking AT-blocks in *L. maculans* have orthologs in *P. nodorum* that are either two consecutive genes or genes separated by only a few others (Figure 1b). A similar pattern was observed for *C. heterostrophus* and *P. tritici-repentis*, suggesting the TEs invaded the genome after the separation of *Leptosphaeria* from other species of suborder *Pleosporineae* 50-57 million years ago (MYA) (Figure 1a).

The ribosomal DNA repeat is extensively affected by RIP

In eukaryotes, the ribosomal DNA (rDNA) comprises a multigene family organized as large arrays of tandem repeats. The core unit is a single transcription unit that includes the 18S or Small Subunit (SSU), 5.8S, and 28S or Large Subunit (LSU) separated by internal transcribed spacers (ITS1 and ITS2). Each transcription unit is separated by the Intergenic Spacer (IGS) (Figure 3a). Although essential duplicated regions would be expected to be protected from RIP mutations, the rDNA repeats in *L. maculans* are in part affected by RIP (Figure 3b, 3c, Supplementary Figure S10). The number of rDNA repeats ranges between 56 and 225 in different *L. maculans* isolates[23]. The assembly of strain v23.1.3 has > 150 repeats, only two of which are highly similar (99.6% identity) and are not affected by RIP. Fifty complete rDNA units and 107 incomplete units are present, and most of them are on extreme ends of SC2 and SC19, which are not complete chromosomes. Many of these repeats are severely affected by RIP (Figure 3, Supplementary Figure S10). Selker[24] has suggested that rDNA repeats in the nucleolus organizer region are protected from RIP. Our data indicate that this is not the case in *L. maculans*, at least for a part of the array of tandem repeats.

AT-blocks as niches for effectors

As described above, AT-blocks have few genes. Furthermore, 76% of these genes are located close to the borders with GC-blocks; only 24% (148 genes) are located within AT-blocks (Table 3). Protein comparisons and Gene Ontology (GO) analysis indicate that AT-blocks are enriched in genes likely to have a role in pathogenicity (Supplementary Figure S11). These include orphan genes such as those encoding SSPs, genes involved in response to chemical or biotic stimuli (Supplementary Figure S11), as well as non-ribosomal peptide synthetases and polyketide synthases, which encode enzymes involved in biosynthesis of secondary metabolites (Supplementary Tables S9, S10; Supplementary Figure S12).

Table 3. Comparative features of SSP-encoding genes occuring in diverse genome environments.

	All predicted genes	SSPs in GC-equilibrated regions	Non-SSPs in borders[a]	Non-SSPs within AT-rich regions	SSPs in borders[a]	SSPs within AT-rich regions
No.	12469	529 (4.2%)	407 (3.3%)	91 (0.7%)	65 (0.5%)	57 (0.5%)
BLAST hits (%)[b]	71,3	48,4	60,2	34,1	15,4	8,8
GC content (%)	54,1	54,6	52,9	48,9	51,1	48,2
TpA/ApT	1,04	1,2	1,12	1,49	1,19	1,44
TpA/ApT > 1.5 (%)[c]	6,9	16,4	11,5	36,3	20	38,6
EST, transcriptomic or proteomic support (%)	84,8	77,1	73,7	54,9	56,9	60
No. of genes present on the NimbleGen array, and with transcriptomic support	10524	396	298	47	35	33
Genes over-expressed in planta 7dpi[d] (%)	9,9	19,1	11,1	36,2	13,9	72,7
Genes over-expressed in planta 14 dpi[d] (%)	11	15,4	11,8	8,5	22,2	24,2
Average protein size (amino acid)	418,4	167,7	396,1	192,4	111,6	98,6
% Cysteines in the predicted protein	1,7	2,9	1,9	2,1	3,8	4,5

[a] "Borders" refer to 859 ± 385 bp transition regions between AT-rich and GC-equilibrated genomic regions.
[b] BLAST to nr cutoff = $1.e^{-10}$
[c] Percentage of genes showing a TpA/ApT Repeat Induced Point mutation (RIP) index above 1.5. This cutoff corresponds to that observed in the majority of RIP-inactivated *AvrLm6* alleles[39]
[d] Genes with more than 1.5 fold change in transcript level and an associated *P*-value lower than 0.05 were considered as significantly differentially expressed during infection (7 or 14 dpi) compared to growth *in vitro*; expressed as a percent of genes with transcriptomic support. Similar genes may be overexpressed at 7 and 14 dpi.

One hundred and twenty-two (*ca.* 20%) of the genes located in AT-blocks encode putative SSPs (Table 3). Only 4.2% of the genes in the GC-blocks encode SSPs (529 genes), and these lack many features of known effectors of *L. maculans* (Table 3). In contrast, the SSPs encoded in AT-blocks have features indicative of effectors such as low EST support in *in vitro* grown cultures, low abundance in *in vitro* secretome samples, increased expression upon plant infection, lack of recognizable domains or homologs in other fungi, and high cysteine content (Table 3). Three TEs, the retrotransposon, RLx_*Ayoly*, and Two DNA transposons, DTF_*Elwe* and DTx_*Gimli* are significantly over-represented in the immediate vicinity of SSPs (Supplementary Figure S13). Although SSPs are never embedded within a single TE, four SSPs are inserted between two tandemly repeated copies of the DNA transposon DTM_*Sahana*.

As well as the avirulence genes, two SSPs, LmCys1 and LmCys2, have been functionally analysed. LmCys1 contributes to fungal growth *in planta*, whilst LmCys2 contributes to suppression of plant defence responses, reflecting their roles as effectors (I. Fudal, unpublished data). Expression of 70.2% of the SSP-encoding genes was detected (Table 3). Of these, 72.7% of the SSP-encoding genes located within AT blocks (compared to 19.1 to 22.2% in GC-blocks) were over-expressed at early stages of infection of cotyledons compared to *in vitro* mycelium growth (Table 3; Supplementary Figure S14). Accordingly these are postulated to be effectors. In addition, 45% of the predicted SSPs in AT-blocks show a presence/absence polymorphism in field populations, as is the case for avirulence genes in *L. maculans* and other fungi[25]. The SSPs in GC-blocks include 110 (20.8%) with best BLAST hits to hypothetical proteins from *P. nodorum*. In contrast, very few SSPs in AT-blocks have identifiable orthologs; only two (1.8%) had a best match to a predicted protein of *P. nodorum*. In addition to their lack of orthologs, SSPs in AT-blocks also lack paralogs; only seven genes belong to gene families comprising one to four paralogs. Biases in codon usage occur: in GC-blocks, the preferred codon for each of the 20 amino acids ends with a C or a G and the preferred stop codon is TGA, whilst in SSP genes located in AT-blocks, the preferred codon ends with an A or T for 13 amino acids and the preferred stop codon is TAA (Supplementary Table S11). This, however only has a limited impact on amino acid favoured usage by SSPs (Supplementary Table S12).

Table 4. Main families and characteristics of transposable elements and other repeats in the *L. maculans* genome.

	Genome coverage	Size (bp)	LTR/TIR size (bp)	Number of copies	Number of complete copies	Complete/ incomplete copies	SuperFamily
Class I (retrotransposons) [a]							
LTR : 9 families.	12.30 Mb (27.26%)						
RLG_*Olly*	3.06 Mb	7246	250	1085	187	0.172	*Ty3/Gypsy*
RLG_*Polly*	2.97 Mb	6928	179	1014	164	0.162	*Ty3/Gypsy*
RLG_*Rolly*	2.24 Mb	11875	235	594	46	0.077	*Ty3/Gypsy*
RLC_*Pholy*	3.10 Mb	6981	281	1020	83	0.081	*Ty1/Copia*
RLG_*Dolly*	0.30 Mb	6620	228	85	30	0.353	*Ty3/Gypsy*
RLC_*Zolly-1* & *-2*	0.16 Mb	5306	177	97	14	0.144	*Ty1/Copia*
RLx_*Jolly*	0.02 Mb	803	259	57	5	0.088	Unknown
RLx_*Ayoly*	0.40 Mb	10397	217	164	8	0.049	Unknown
RLG_*Brawly*	0.05 Mb	7289	None	22	3	0.136	*Ty3/Gypsy*
Class II (DNA transposons)							
TIR : 10 families	1.19 Mb (2.64%)						
DTF_*Elwe*	199.3 kb	2173	57	158	54	0.342	*Fot1-Pogo*
DTM_*Lenwe*	25.6 kb	3489	49	36	3	0.083	*Mutator*
DTx_*Olwe*	5.3 kb	866	49	15	4	0.267	Unknown
DTx_*Valwe*	11.5 kb	1793	37	73	2	0.027	Unknown
DTT_*Finwe-1*	2.7 kb	529	29	7	4	0.571	*Tc1-Mariner*
DTT_ *Finwe-2*	10.8 kb	523	29	31	11	0.355	*Tc1-Mariner*
DTT_ *Finwe-3*	7.4 kb	806	29	15	4	0.267	*Tc1-Mariner*
DTM_*Sahana*	782.8 kb	5992	none	873	49	0.056	*Mutator*
DTx_*Gimli*	112.9 kb	606	none	279	51	0.183	Unknown
DTM_*Ingwe*	33.4 kb	3582	37	48	1	0.021	*Mutator*
Uncharacterized repeats (11 families)	159.9 kb (0.35%)						
rDNA repeats [b]	767 kb (1.70%)	7800 [d]		> 100	50		
Telomeric repeats [c]	935.0 kb (2.07%)						

[a] Classification of TEs according to Wicker et al.[16]: the three-letter code refers to class (R, retrotransposon; D, DNA transposon), order (L, Long Terminal Repeat –LTR-; T, Terminal Inverted Repeat –TIR-; P, Pelelope Like Element –PLE) and superfamily (G, *Gypsy*; C, *Copia*; P, *Penelope*; F, *Fot1-Pogo*; T, *Tc1-Mariner*; M, *Mutator*, x, unknown superfamily) followed by the family (or subfamily) name italicized.
[b] Including a rDNA-specific LINE element
[c] Including Telomere Associated Penelope-like retroelement RPP-*Circe* and RecQ telomere-linked Helicase
[d] Excluding variable length short-tandem repeats flanking almost every rDNA repeat.

Motifs resembling the RxLR translocation motifs of oomycetes were sought[26] following the validation that one such motif, RYWT, present in the N-terminal part of AvrLm6 allows translocation into plant and animal cells[26]. Searches for <[RKH] X [LMIFYW] X> or <[RKH] [LMIFYW] X [RKH]> showed that up to 60% of SSPs in AT-blocks and up to 73% of SSP in GC-blocks have putative “RxLR-like” motifs, implicating these SSPs as candidate effectors that enter plant cells.

History of genome invasion by Transposable Elements

A range of 278-320 MYA is estimated for the origin of the *Dothideomycetes* with the crown radiation of the class during the Permian (251-289 MYA) (Figure 1a). The origins of the plant pathogenic *Pleosporineae* is determined at 97-112 MYA, placing it in the Cretaceous at a time when flowering plants were beginning to become widespread and eudicots were emerging, during the late Cretaceous and Paleocene. *Leptosphaeria* likely diverged from the other species analysed between 50 and 57 MYA (Figure 1a). Phylogenetic analyses suggest three main features of genome invasion by TEs: transposition bursts mostly after separation of *L. maculans* from other species of suborder *Pleosporineae* as indicated by a “recent” divergence of the TE families, estimated to 4-20 MYA (Figure 4a); a single or few wave(s) of massive transposition(s) followed by a “rapid” decay, with some cases like DTM_*Sahana* where divergence between copies is extremely low; and no on-going waves of genome invasion by TEs (Figure 4b). Like other organisms with a high density of TEs, the *L. maculans* genome exhibits ‘nesting’, where repeats occur within previously inserted TEs. In this fungus TEs are commonly invaded by other TEs generating a complex ‘nesting network’. Eighty-five% of these cases correspond to TEs invading one other TE (primary nesting relationship). Most of the retrotransposon families investigated can invade or be invaded to similar extents (Supplementary Table S8). They also can invade TEs from the same family (self-nests) but usually at a very low frequency compared to invasion of retrotransposons from other families. In contrast, the DNA transposons are more commonly invaded (23.3% of the cases) than acting as invaders (3.5% of the cases; Supplementary Table S8). In accordance with overlapping divergence time estimates (Figures 1a, 4), these data indicate periods of overlapping transpositional activity for the LTR-retrotransposons that form the major part of AT-blocks. In such a scenario, the later insertions would be preferentially tolerated in existing

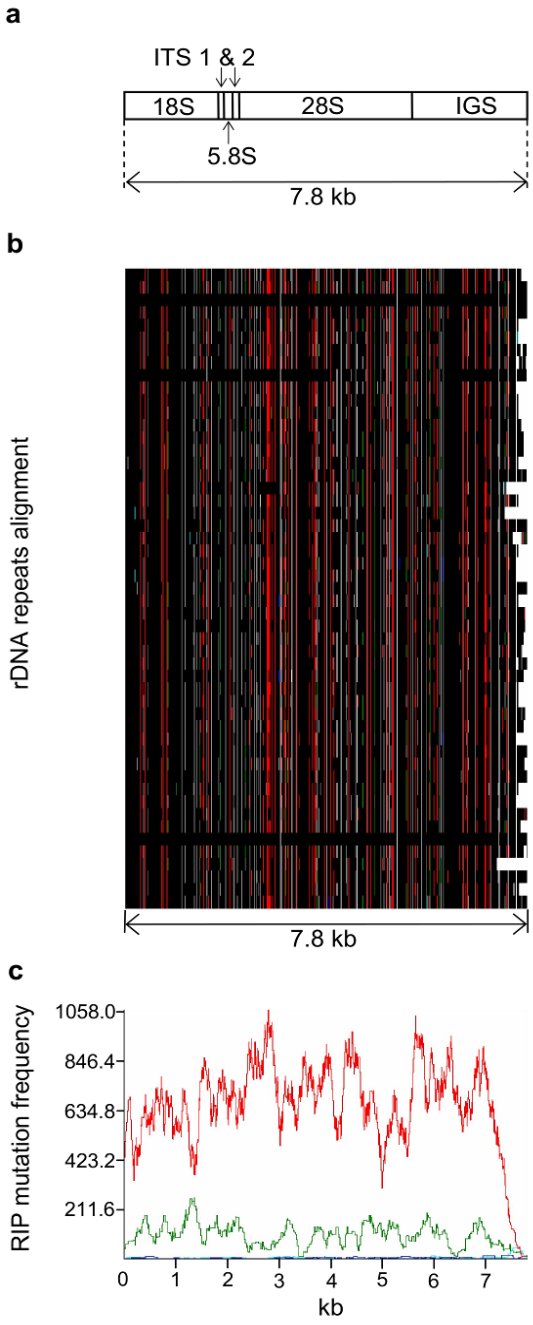


Figure 3. Repeat-induced Point mutation in ribosomal DNA of *Leptosphaeria maculans,* shown as RIPCAL output. (**a**) schematic representation of the rDNA unit in *L. maculans* (ITS, internal transcribed spacers; IGS, intergenic spacer); (**b**) a schematic multiple alignment of the 7.8 kb "complete" ribosomal DNA (rDNA) units occurring in SuperContigs 2 and 19. Polymorphic nucleotides are coloured as a function of the type of Repeat-induced Point (RIP) mutation observed with black, invariant nucleotide, red, CpA ←→ TpA or TpG ←→ TpA mutations; dark blue, CpC ←→ TpC or GpG ←→ GpA mutations; pale blue, CpT ←→ TpT or ApG ←→ ApA mutations; green, CpG ←→ TpG or CpG ←→ CpA mutations; (**c**) RIP mutation frequency plot over a rolling sequence window, corresponding to the multiple alignment directly above. Nucleotide polymorphisms (against the alignment consensus, which is also the highest GC-content sequence) mostly correspond to CpA ←→ TpA or TpG ←→ TpA (red curve) and CpG ←→ TpG or CpG ←→ CpA (green curve).

decayed transposons. These TEs having undergone RIP in their turn, would initiate a positive reinforcement loop that would create large AT-rich and gene-poor blocks of homogeneous nucleotide composition.

Discussion

The peculiar genomic structure of the *L. maculans* genome is reminiscent of that discovered in mammals and some other vertebrates: the base composition (GC-content) varies widely along chromosomes, but locally, base composition is relatively homogenous. Such structural features have been termed 'isochores'[27]. In *L. maculans*, AT-blocks are gene-poor, rich in TEs and deficient in recombination compared to GC-blocks, as in mammals[27]. However, despite these similarities, these genomic landscapes appear to result from different mechanisms. In mammals, the evolution of GC-rich isochores is most likely driven by recombination: genomic regions sized between 100 kb to several Mb with a high recombination rate tend to increase in GC content relative to the rest of the genome. This pattern is not due to a mutational effect of recombination, but most probably due to biased gene conversion[28]. In *L. maculans*, variations in base composition occur at a much finer scale (the isochore-like blocks are about 10-20 times smaller than in mammals), and it is unknown whether biased gene conversion contributes to increase the GC content of GC-blocks. Conversely, *L. maculans* isochores can be attributed to the AT-biased mutational pattern induced by RIP mutation of TEs and their flanking regions, thus leading to the evolution of AT-rich isochores.

Although the evolutionary forces we postulate shaped the *L. maculans* genome are common to many species, no fungal genome characterized so far has a similar isochore-like structure. This structure reflects extensive genome invasion by TEs that are nonetheless tolerated by the pathogen, existence of an active RIP machinery (Supplementary Table S6), so far restricted to the *Pezizomycotina* subphylum of the *Ascomycota* and maintenance of sexual reproduction (necessary for RIP). Whereas many species seem to have maintained an active RIP machinery, most of the sequenced fungal genomes are poor in TEs, indicating that run-away genome expansion is normally deleterious. Also, many fungal species have lost the ability to cross in nature (*e.g.*, *F. oxysporum*, *Magnaporthe oryzae*) and no case of large-scale sculpting of repeat-rich regions is found in these species, only some ancient signatures of RIP are[29].

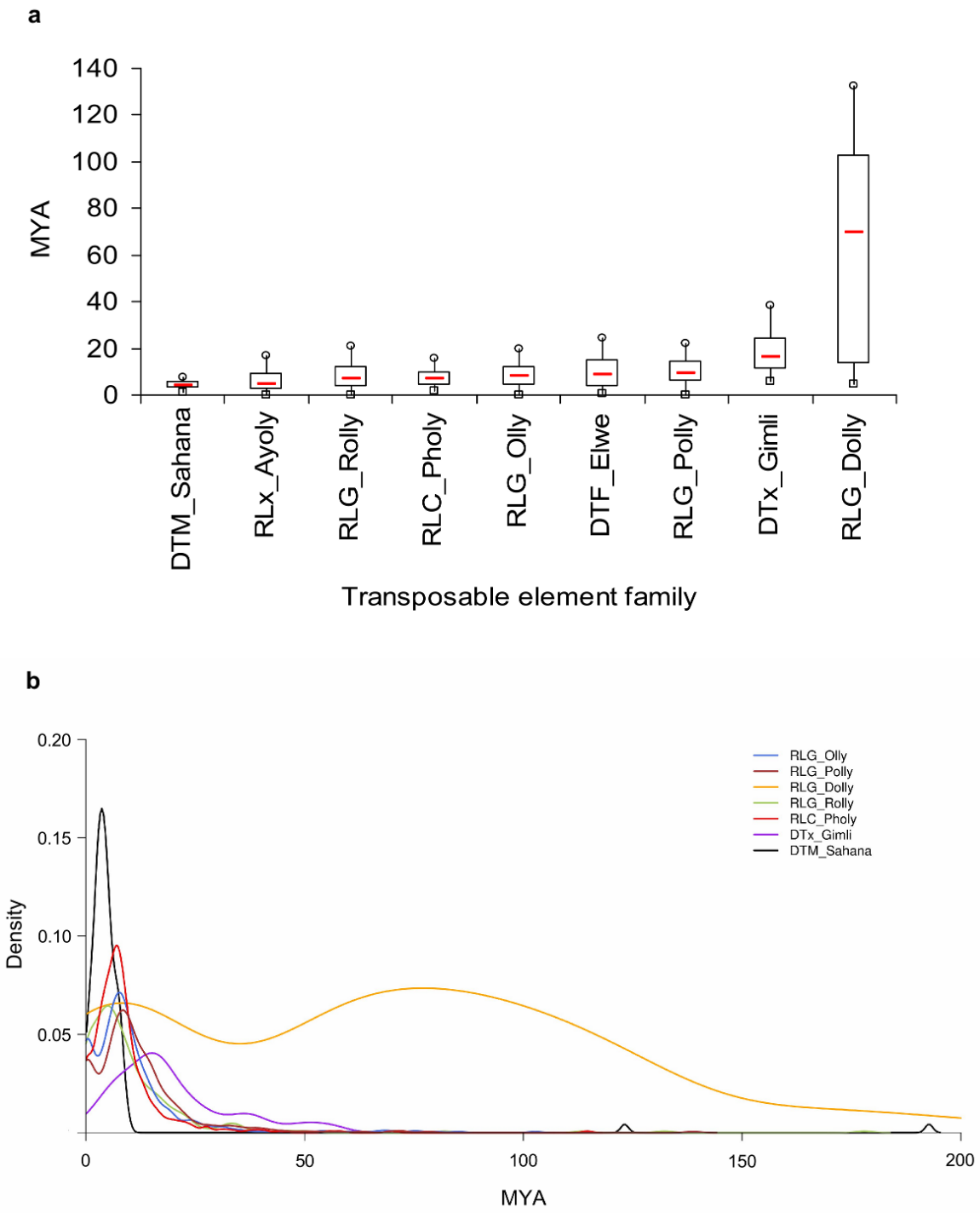


Figure 4. Dynamics of transposable elements in the *Leptosphaeria maculans* genome. A phylogenetic analysis was used to retrace the evolutionary history of each transposable element (TE) family after elimination of mutations due to Repeat-induced Point mutations. Terminal fork branch lengths were assumed to correspond to an evolutionary distance used to estimate the age of the last transposition activity. The divergence values were converted to estimated divergence time using a substitution rate of $1.05\ 10^{-9}$ substitution per location per year[52,53] (expressed as "million years ago" MYA). (**a**) Box plot graph of divergence times. The red line represents the median value; the boxes include values between the first and the third quartile of the distribution; squares and circles, 1st and 9th decile, respectively. (**b**) Kernel density of divergence plots. A R script was written to plot a histogram of the terminal fork branch length with kernel density estimate for each family.

On the basis of the characteristics of avirulence genes in *L. maculans*, we have described a comprehensive repertoire of putative effectors, which has not previously been done for an Ascomycete. In *L. maculans*, AT-rich blocks are enriched in effector-like sequences. Location of effector genes has been investigated in only some eukaryotic genomes. A few of the effectors of *M. oryzae* are subtelomeric[30], as are those in protozoan parasites of animals, such as *Plasmodium* and *Trypanosoma* [31]. Genomes of many *Fusarium* species contain supernumerary "B" chromosomes enriched in strain-specific effectors and accounting for the host range of each "*forma specialis*"[32,33]. The genome of the oomycete *Phytophthora infestans* has a plethora of effector candidates embedded in repetitive DNA and diversification of these effectors is postulated to occur via segmental duplication and variation in intra-specific copy number resulting in rapidly diverging multigene families[8]. The association between one family of effectors and a LINE in *Blumeria graminis*, the barley powdery mildew fungus, is proposed to provide a mechanism for amplifying and diversifying effectors[34]. Diversification of effectors in the species mentioned above is postulated to be associated with TE-driven gene duplication and generation of multigene families. In the *L. maculans* genome, SSP-encoding genes are associated with only a few TE families, which may indicate the ability of TEs to "pickup and move" effectors. In contrast to the above examples, duplicated effector genes are not present in *L. maculans*, a finding consistent with the steady inactivation of TEs by RIP and with ancient transposition activity prior to underdoing RIP.

The origins of some effector genes might be at least partially ascribed to lateral gene transfer (LGT), a specialty of species within the Pezizomycotina[35-37]. Regardless of the origin of the effector genes, our data suggest that RIP is an important mechanism for generating diversity for genes occurring within AT-blocks of the genome of *L. maculans*, in a manner not previously documented in any other species. RIP has previously been reported to be restricted to duplicated DNA but most SSPs or other genes in AT-blocks are present in single copies. How, then, can RIP act on SSP-encoding genes? Studies in *N. crassa* indicated that the RIP machinery can occasionally overrun the repeated region into adjacent single-copy genes[38]. The embedding of SSPs within RIP-degenerated TEs would then favour such RIP leakage (Supplementary Figure S4c) while selection pressure to maintain functional effectors would prevent them from becoming extinct due to an excessive degree of RIP. This would result in extensive mutation of the affected gene and could account for the mutation rate required for diversifying selection. In contrast, effector genes that became detrimental to pathogen fitness, such as avirulence genes subjected to

resistance gene selection, would be lost rapidly as alleles that have undergone extensive RIP are selected for[39]. Evidence for this scenario is provided by examination of RIP indices and in alignment-based studies of alleles of SSPs[39]. The genes (including SSPs) within AT-blocks had higher TpA/ApT indices than those in GC-blocks (Table 3; Supplementary Figure S15) consistent with former genes having been RIP-affected. RIP indices for the effectors located within AT-blocks thus would be a compromise between values leading to complete degeneration of the sequence and values enabling sequence diversification while retaining functionality. In plant-pathogen systems, diversifying selection operates on effector genes whose products interact with host proteins[25]. This has been demonstrated for both resistance and avirulence genes, but mechanisms for the diversifying selection of effectors have not been proposed. RIP is shown here to be a potential factor to create the genetic (hyper)variation needed for selection to occur in *L. maculans* and this process may also act on effectors in other fungi[40].

These findings allow speculation about an evolutionary scenario for birth of isochore-like structures in the *L. maculans* genome and its incidence on effector diversification. Firstly, the genome was invaded by a few families of TEs over a (relatively) short time period, mostly after the separation of *L. maculans* from other related fungi. This TE invasion is unlikely to have been targeted to pre-existing effector-rich genome regions as seen in microsynteny analyses (Figure 1b) and the fact that the most recent invader, DTM_*Sahana*, is not specifically associated with SSPs. Secondly, waves of overlapping transposition occurred with probable transduction, translation or duplication of genes, resulting in large amplification of a few families. Such transpositions were primarily targeted to other TEs as shown by the nesting of retrotransposons within other TEs. In parallel, duplicated copies of TEs and genes (either duplicated or not) hosted within TE-rich regions underwent RIP either to extinction for TEs or to generate gene diversity in cases where a strong selection pressure to retain genes was exerted. This eventually resulted in complete inactivation of transposition events, and the sculpting of the genome in an isochore-like structure. Effector genes were maintained in AT-blocks to favor rapid response to selection pressure[39,41] and probable epigenetic concerted regulation of their expression (Supplementary Figure S14b). *L. maculans* shows intriguing evolutionary convergence with both higher eukaryotes in terms of an isochore-like genome structure, and with oomycetes in terms of hosting effectors in highly dynamic “plastic” regions of the genome[8]. It differs in exploiting a RIP-based mechanism for diversification and inactivation of effector genes.

The sequencing of genomes of several species or sub-species of the recent and more ancient outgroups that derived from a common ancestor with *L. maculans* will provide more information on origin of effectors, genome invasion by TEs and the subsequent effect on generation/diversification of effectors, and thus test the validity of the proposed evolutionary scenario.

Methods

Phylogenetic analysis

A taxon set containing representatives of most classes in Ascomycota was selected from the data matrices produced in two previous papers[42,43]. Sequences were concatenated from the SSU and LSU of the nuclear ribosomal RNA genes and three protein coding genes, namely the translation elongation factor-1 alpha and the largest and second largest subunits of RNA polymerase II (Supplementary Table S1). A phylogenetic analysis was performed using RAxML v. 7.0.4[44] applying unique model parameters for each gene and codon. A combined bootstrap and maximum likelihood (ML) tree search was done in RAxML with 500 pseudo replicates. The best scoring ML tree was analysed in the program R8sv1.7[45] in order to produce a chronogram (Figure 1a).

Sequencing and assembly

L. maculans 'brassicae' isolate v23.1.3. was sequenced because it harbours numerous avirulence genes, three of which have been cloned by a map-based strategy involving large-scale sequencing of surrounding genomic regions[13,15,41]. Isolate v23.1.3 results from a series of *in vitro* crosses between European field isolates[46] and is representative of the populations of the pathogen prevalent in the EU in the mid 1990s.

DNA was provided as agarose plugs containing partly digested conidia[21]. Whole genome shotgun sequencing of three types of libraries (high-copy-number plasmids with 3.3 kb inserts; low-copy-number plasmids with 10 kb inserts and fosmids with inserts 35 or 40 kb) was performed, and also six cDNA libraries, including ones derived from infected plants were sequenced (Supplementary Table S13). Sequencing reads were assembled using Arachne[47] (Table 1) and the correspondence of SCs to chromosome was inferred by aligning the genetic map to the genome sequence, hybridization of single-copy markers to chromosomal DNA separated by pulsed-field gel electrophoresis, identification of telomere-specific repeats, and by mesosynteny analyses (conserved gene content) with genomes of other Dothideomycetes (Supplementary Table S2).

Leptosphaeria maculans genome annotation

Automated structural annotation of the genome was performed using the URGI genomic annotation platform including pipelines, databases and interfaces, developed or locally set up for fungi. The EuGene prediction pipeline v. 3.5a[48], which integrates *ab initio*

(Eugene_IMM, SpliceMachine and Fgenesh 2.6 (www.softberry.com)) and similarity methods (BLASTn, GenomeThreader, BLASTx) was used to predict gene models. The functional annotation pipeline was run using InterProScan[49]. Genome assembly and annotations are available at INRA (http://urgi.versailles.inra.fr/index.php/urgi/Species/Leptosphaeria).

Genome assemblies together with predicted gene models and annotations were deposited at DNA European Molecular Biology Laboratory/GenBank under the accession numbers FP929064-FP929139 (SC assembly and annotations). ESTs were submitted to dbEST under accession no FQ032836-FQ073829.

Full description and associated references for sequencing, assembly and gene annotation are provided as Supplementary Methods.

Annotation and analysis of repeated elements

TEs[16] were identified and annotated using the "REPET" pipeline (http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET), optimized to better annotate nested and fragmented TEs. Repeats were searched with BLASTER for an all-by-all BLASTn genome comparison, clustered with GROUPER, RECON and PILER, and consensuses built with the MAP multiple sequence alignment program. Consensuses were classified with BLASTER matches, using tBLASTx and BLASTx against the Repbase Update databank[50] and by identification of structural features such as long terminal repeats (LTR), terminal inverted repeats (TIR)[16] etc. Additional steps of clustering and manual curation of data were performed resulting in a series of consensuses used as an input for the REPET annotation pipeline part, comprising the TE detection software BLASTER, RepeatMasker and Censor, and the satellite detection softwares RepeatMasker, TRF and Mreps.

Analysis of the dynamics of genome invasion by TEs was firstly based on phylogenetic analysis of each family of repeats, retracing the evolutionary history regardless of truncation, insertion in other TEs and deletion events[51]. After elimination of all RIP targets, the tree topology was used to retrace the dynamics and demography of TE invasion in the genome. Terminal forks branch lengths from the trees were used to calculate the age of the last transposition events of the copies in the genome. The divergence values were converted in estimated divergence time using a substitution rate of 1.05×10^{-9} nucleotide per site per year as applied to fungi[52,53].

Dynamics of TE aggregation over time was also analysed by a visual analysis of nesting relationships between TEs. Following the long join annotation, mosaics of TEs were visualised using Artemis v. 12.0 (http://www.sanger.ac.uk/Software/Artemis/) in SC0-22 and a data matrix recording the frequency with which a given TE family was inserted into another one (invader) and the frequency with which one given TE was recipient of an insertion from one or multiple other TEs (invaded TE) was generated (Supplementary Table S8). The statistical identification and significance of the favoured invasion of other TE families as compared to random association was evaluated with a Chi-squared test for given probabilities with simulated *P* values, based on 20,000 replicates, as implemented in R. In addition a Python script was written to identify, quantify and visualise nesting relationships.

RIP and DeRIP analyses

Automated analysis of RIP in *L. maculans* genomic DNA repeats was performed using RIPCAL (http://www.sourceforge.net/projects/ripcal), a software tool that performs both RIP index and alignment-based analyses[19]. In addition, RIP indices such as TpA/ApT and (CpA+TpG)/(ApC+GpT) were used to evaluate the effect of RIP on genes or genome regions for which multiple alignments could not be generated. DeRIP analyses, which predict putative ancient pre-RIP sequences, were perfomed using an updated version of RIPCAL, including the Perl script “deripcal” and ripcal_summarise.

Analysis of AT-blocks

AT- and GC-blocks were manually discriminated from each scaffold using Artemis (http://www.sanger.ac.uk/Software/Artemis/), and a Python script was used to extract sequences and features of AT-blocks. TE content of AT- and GC-blocks was analyzed using the REPET pipeline. Size distribution of AT-blocks, occurrence of AT-blocks on chromosomes and relationship between AT-blocks, TE content and chromosome length were calculated.

To evaluate meiotic recombination differences between AT- and GC-blocks, micro and minisatellites located either in GC-blocks or located on both sides of a single AT-block were mapped in a reference cross and the number of cross-overs between two consecutive markers was calculated. The recombination frequency between two successive markers was calculated, plotted against the physical distance between the two

markers and subjected to an ANOVA and a non-parametric test (Mann-Whitney test) using XLStat, to compare recombination frequencies between and within GC-blocks.

Intergenic distances were compared between AT- and GC-blocks in *L. maculans*, and also compared to those of the closely-related Dothideomycete, *P. nodorum* (Supplementary Table S5).

Gene Ontology annotations were compared between genes occurring in AT- and GC-blocks using the blast2GO program.

Identification and features of Small Secreted Proteins (SSP)

Non-repeated regions within AT-blocks were identified following masking of TEs with RepeatMasker. The EMBOSS:GETORF program was used on these genomic regions to refine the identification of genes encoding SSPs with a size limit set at 600 amino acids (lower limit: 60 amino acids). A dedicated script combined the outputs of GETORF, FgeneSH and EuGene and a pipeline written in Python screened the predicted proteins according to their size and the presence of signal peptide and transmembrane domains (SignalP 3.0, TargetP and TMHMM). Base composition of the genes encoding SSPs (percent of each base in the sequence, GC content and GC3 content) and amino acid count of the SSPs (as% of each amino acid in the protein) were calculated by custom Python scripts. Statistical bias in amino acid occurrence was evaluated by an F-test to determine if the variances were equal in both sets, followed by a Student t test (95% confidence level) to compare the mean use of each amino acid in each set of predicted proteins. Biases in codon usage were evaluated using EMBOSS:CHIPS. A Chi-squared test for given probabilities with simulated values (20,000 replicates) as implemented in R was performed to test random association of SSP-encoding genes in AT-blocks with specific TEs. Motifs similar to the RxLR motif necessary for oomycete effectors to be translocated within plant cells were sought in predicted SSPs using a Python script aiming at identification of motifs (<[RKH] X [LMIFYW] X> or <[RKH] [LMIFYW] X [RKH]>).

Analysis of expression patterns of SSP-encoding genes were compared between *in vitro* (mycelium grown in axenic medium) and *in planta* (3, 7 and 14 days after inoculation of oilseed rape cotyledons), either using the *L. maculans* whole-genome expression array (manufactured by NimbleGen Systems Limited) or by qRT-PCR on a selected subset of SSP-encoding genes.

Acknowledgements

The authors acknowledge Marc-Henri Lebrun (INRA-Bioger) and Francis Martin (INRA, Interactions arbres/micro-organismes, Champenoux, France) for support and fruitful advice. The genome sequencing of *Leptosphaeria maculans* was funded by the Genoscope, Institut de Génomique, CEA, France. The establishment of databases and interfaces was funded by Agence Nationale de la Recherche (GnpAnnot project; ANR-07-GPLA-051G). Whole-genome effector analysis was funded by ANR (FungEffector project; ANR-06-BLAN-0399). Recombination analysis and P.B. were funded by ANR (AvirLep project; ANR-07-GPLA-015). B.J.H and R.P.O. thank the Australian Grains Research and Development Corporation for funding. C.L.S was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. J.W.S acknowledges support from the U.S. National Science Foundation, grant number DEB-0717476. B.M.T. was supported in part by the U.S. National Science Foundation, grant number IOS-0924861. Special thanks are due to INRA-SPE department, the "*Leptosphaeria maculans*" scientific and applied community, and the "Dothideomycete" community for strong support of the *L. maculans* genome initiative.

Author Contributions

T.R. and J.G. made equivalent contributions and should be considered joint first authors; J.K.H., C.H., A.P.vdW, A.C. and V.D. contributed equally to this work as second authors. J.A., H.Q., R.P.O., P.W., M.H.B. and B.J.H. coordinated genome sequencing, annotation and data analyses, and made equivalent contributions as senior authors. Individual contributions were as follows. T.R., M.H.B. and B.J.H. initiated the sequencing project. M.H.B was responsible for DNA extraction and purification; P.W. and J.W. coordinated the sequencing phase at Genoscope. J.P. and sequencing staff at Genoscope performed the shotgun sequencing. A.C., V.A. and his staff at Genoscope assembled the genome. J.A. was responsible for structural and functional annotation pipelines, databases and interfaces set up and maintenance at INRA-URGI. V.D., C.H., and J.A. did the ab initio annotation of protein-coding gene models. M.M., A.J.C., and B.J.H. provided EST/cDNA information for the ab initio and manual annotation. V.D. and C.H. did the cDNA clustering and the training set of full-length cDNA/Genomic for *ab initio* gene finder training steps. V.D., C.H. and J.A. inserted annotation data in the database. Genome statistics were performed by J.A., H.Q. and J.G. J.K.H., R.P.O. and J.G. did the genome synteny

analyses; M.H.B., P.B., L.G., A.J.C., A.P.vdW. A. St. and J.G. identified and designed mini- and microsatellite markers; M.H.B and A.P.vdW. built the genetic maps; J.K.H. and R.P.O. performed mesosyteny and RIPCAL analyses, A.J.C. and K.M. hybridized electrokaryotypes and annotated NRPSs and PKSs; H.Q., V.D., L.G. and T.R. analysed transposable elements; V.D. and J.G. performed phylogenetic analyses and time estimates for TE transposition events; J.G., T.R. and M.H.B. perfomed analysis of TE nesting; N.L. performed automated functional analysis and set up of functional databases and interfaces; N.L. and S.B. performed GO analyses; B.M.T. and J.G. did RXLR analysis of effector candidates; I.F. designed micro-arrays, B.O. and J.L. obtained RNA for transcriptomics approaches, A.S. and J.G. validated design of microarrays, A.S., I.F. and J. L. analysed microarray data, B.O., N.G. and I.F. performed expression analysis of effectors by RT-PCR and qRT-PCR; A.D. analysed polymorphism of effectors in field populations; D.V. did proteomic and secretomic analyses. L.M.C., S.B.G., C.B.L., G.H.J.K. and B.G.T. contributed to comparative genomics approaches and by providing unpublished information on genomes of *M. graminicola*, *P. tritici-repentis*, *A. brassicicola* and *C. heterostrophus*. L.D. analysed isochores-like blocks and contributed to comparative analysis with those of mammals. C.L.S. and J.W.S. performed phylogenetic analyses and estimated divergence time. T.R. organized co-ordination between different groups. T.R. wrote and edited the paper with major input from B.J.H. and R.P.O. Final editing of the text, Tables and Figures was done by S.B., J.G., M.H.B., B.J.H. and T.R.

References

1. Skamniotia, P. & Gurr, S. J. Against the grain: safeguarding rice from rice blast disease. *Trends Biotechnol.* **27,** 141-150 (2009)
2. Oliver, R.P. & Solomon, P.S. Recent fungal diseases of crop plants: is lateral gene transfer a common theme? *Mol. Plant-Microbe Interact.* **21,** 287-293 (2008).
3. Rouxel, T. & Balesdent, M. H. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Mol. Plant Pathol.* **6,** 225-241 (2005)
4. Soanes, D. N., *et al.* Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *Plos One* **3,** 1-15 (2008)
5. Stergiopoulos, I. & de Wit, P.J.G.M. Fungal effector proteins. *Annu. Rev. Phytopathol.* **47,** 233-63 (2009)
6. Rouxel, T. & Balesdent, M. H. Avirulence Genes. In: *Encyclopedia of Life Sciences (ELS)*. John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a00212672010 (2010)
7. Alfano, J.R. Roadmap for future research on plant pathogen effectors. *Mol. Plant Pathol.* **10,** 805-813 (2009)
8. Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans. Nature* **461,** 393-398 (2009)
9. Jiang, R. H. Y., Tripathy, S., Govers, F. & Tyler, B. M. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving super-family with more than 700 members. *Proc. Natl. Acad. Sci. USA* **105**, 4874-4879 (2008)
10. Tyler, B. M. *et al. Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-1266 (2006)
11. Kämper, J. *et al.* Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis. Nature* **444,** 97-101 (2006)
12. Elliott, C.E., Gardiner, D.M., Thomas, G., Cozijnsen, A.J., van de Wouw, A. & Howlett, B.J. Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. *Mol. Plant Pathol.* **8,** 791-802 (2007)
13. Fudal, I., *et al.* Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. *Mol. Plant-Microbe Interact.* **20,** 459-470 (2007)
14. Huang, Y.J., Li, Z.Q., Evans, N., Rouxel, T., Fitt, B.D.L. & Balesdent, M.H. Fitness cost associated with loss of the *AvrLm4* function in *Leptosphaeria maculans*

(Phoma stem canker of oilseed rape). *Eur. J. Plant Pathol.* **114,** 77-89 (2006)

15. Parlange, F., *et al*. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. *Mol. Microbiol.* **71,** 851-863 (2009)
16. Wicker, T., *et al.* A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* **8,** 973-982 (2007)
17. Idnurm, A. & Howlett, B.J. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genet. Biol*. **39,** 31-37 (2003)
18. Espagne, E. *et al.* The genome sequence of the model ascomycete fungus *Podospora anserina. Genome Biol.* **9,** R77 (2008)
19. Hane, J.K. & Oliver, R.P. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics* **9**, 478 (2008)
20. Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L. & Levine, D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in Eukaryotic genomes. *Genome Biol. Evol.* **1**, 205-220 (2009)
21. Leclair, S., Ansan-Melayah, D., Rouxel, T. & Balesdent, M. H. Meiotic behaviour of the minichromosome in the phytopathogenic ascomycete *Leptosphaeria maculans. Curr. Genet.* **30,** 541-548 (1996)
22. Gladyshev, E.A. & Arkhipova, I.R. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. *Proc. Natl. Acad. Sci USA* **104,** 9352-9357 (2007)
23. Howlett, B.J., Cozijnsen, A.J. & Rolls, B.D. Organisation of ribosomal DNA in the ascomycete *Leptosphaeria maculans. Microbiol. Res.* **152,** 1-7 (1997)
24. Selker, E. U. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* **24,** 579-613 (1990)
25. Stukenbrock, E.H. & McDonald, B.A. Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol. Plant-Microbe Interact*. **22,** 371-380 (2009)
26. Kale, S.D. *et al.* External lipid PI-3-P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell*, **142,** 284–295 (2010)

27. Eyre-Walker, A. & Hurst, L.D., The evolution of isochores. *Nat. Rev. Genet.* **2**, 549 (2001)
28. Duret, L. & Galtier, N., Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285-311 (2009)
29. Ikeda, K., *et al*. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. *Mol. Microbiol.* **45,** 1355-1364 (2002)
30. Farman, M.L. Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it. *FEMS Microbiol. Lett.* **273,** 125-132 (2007)
31. Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi. Nature* **455,** 799-803 (2008)
32. Ma, L. J. *et al*. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium oxysporum. Nature* **464,** 367-373 (2010)
33. Coleman, J.J. *et al.* The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PloS Genet.* **5,** e1000618 (2009)
34. Sacristan, S. *et al.* Coevolution between a family of parasite virulence effectors and a class of LINE-1 retrotransposons. *PloS One* **4,** e7463 (2009).
35. Friesen, *et al*. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet*. **38,** 953-956 (2006)
36. Marcet-Houben M. & Gabaldón T. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet.* **26,** 5-8 (2010)
37. Khaldi, N. & Wolfe, K. H. Elusive origins of the extra genes in *Aspergillus oryzae. Plos One* **3,** e3036 (2008)
38. Irelan, J.T., Hagemann, A.T. & Selker, E.U. High Frequency Repeat-Induced Point mutation (RIP) is not associated with efficient recombination in *Neurospora. Genetics* **138,** 1093-1103 (1994)
39. Fudal, I., *et al*. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution towards virulence in *Leptosphaeria maculans*. *Mol. Plant-Microbe Interact*. **22,** 932-941 (2009)
40. Stergiopoulos, I., De Kock, M.J.D., Lindhout, P. & de Wit, P.J.G.M. Allelic variation in the effector genes of the tomato pathogen *Cladosporium fulvum* reveals different modes of adaptive evolution. *Mol. Plant-Microbe Interact.* **20,** 1271-1283 (2007)
41. Gout, L., *et al.* Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen *Leptosphaeria maculans. Environ. Microbiol.*

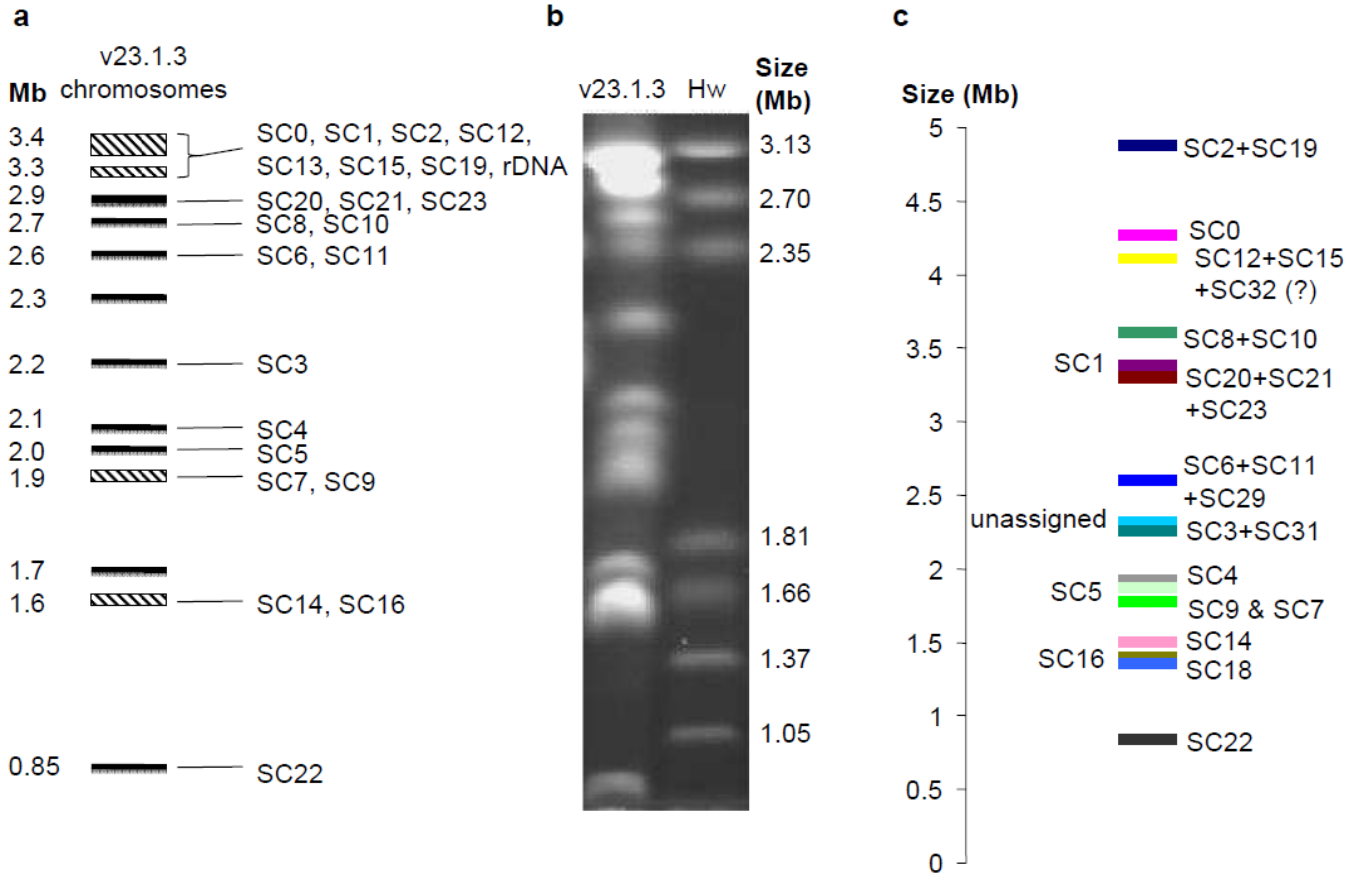
9, 2978-2992 (2007)

42. Schoch, C.L., *et al.* A class-wide phylogenetic assessment of *Dothideomycetes*. *Stud. Mycol.* **64,** 1-15S10 (2009)
43. Schoch, C.L., *et al*. The *Ascomycota* Tree of Life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst. Biol.* **58,** 224-239 (2009)
44. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688-2690 (2006)
45. Sanderson, M.J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301-302 (2003)
46. Balesdent, M.H., Attard, A., Ansan-Melayah, D., Delourme, R., Renard, M. & Rouxel, T. Genetic control and host range of avirulence toward *Brassica napus* cultivars Quinta and Jet Neuf in *Leptosphaeria maculans. Phytopathology* **91,** 70-76 (2001).
47. Jaffe, D. B., *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91-96 (2003)
48. Foissac, S., *et al.* Genome annotation in plants and fungi : EuGene as a model platform. *Curr. Bioinformatics* **3,** 87-97 (2008).
49. Quevillon, E., *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33 (suppl. 2),** W116–W120 (2005)
50. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462-467 (2005)
51. Fiston-Lavier, A.S. Etude de la dynamique des répétitions dans les génomes eucaryotes : de leur formation à leur élimination. PhD Thesis, University Pierre et Marie Curie, Paris, France (2008)
52. Berbee, M.L. & Taylor, J.W. Dating the molecular clock in fungi – how close are we? *Fungal Biol. Rev.*, doi:10.1016/j.fbr.2010.03.001 (2010)
53. Kasuga, T., White, T.J. & Taylor, J.W. Estimation of nucleotide substitution rates in eurotiomycete fungi. *Mol. Biol. Evol.* **19,** 2318–2324 (2002)
54. Rouxel, T., Balesdent, M.H., Amselem, J. & Howlett, B.J. GnpGenome: a Genome Browser for *Leptosphaeria maculans* structural annotation (2010) http://urgi.versailles.inra.fr/index.php/urgi/Species/Leptosphaeria
55. Hane, J.K., *et al.* Dothideomycete plant interactions illuminated by genome

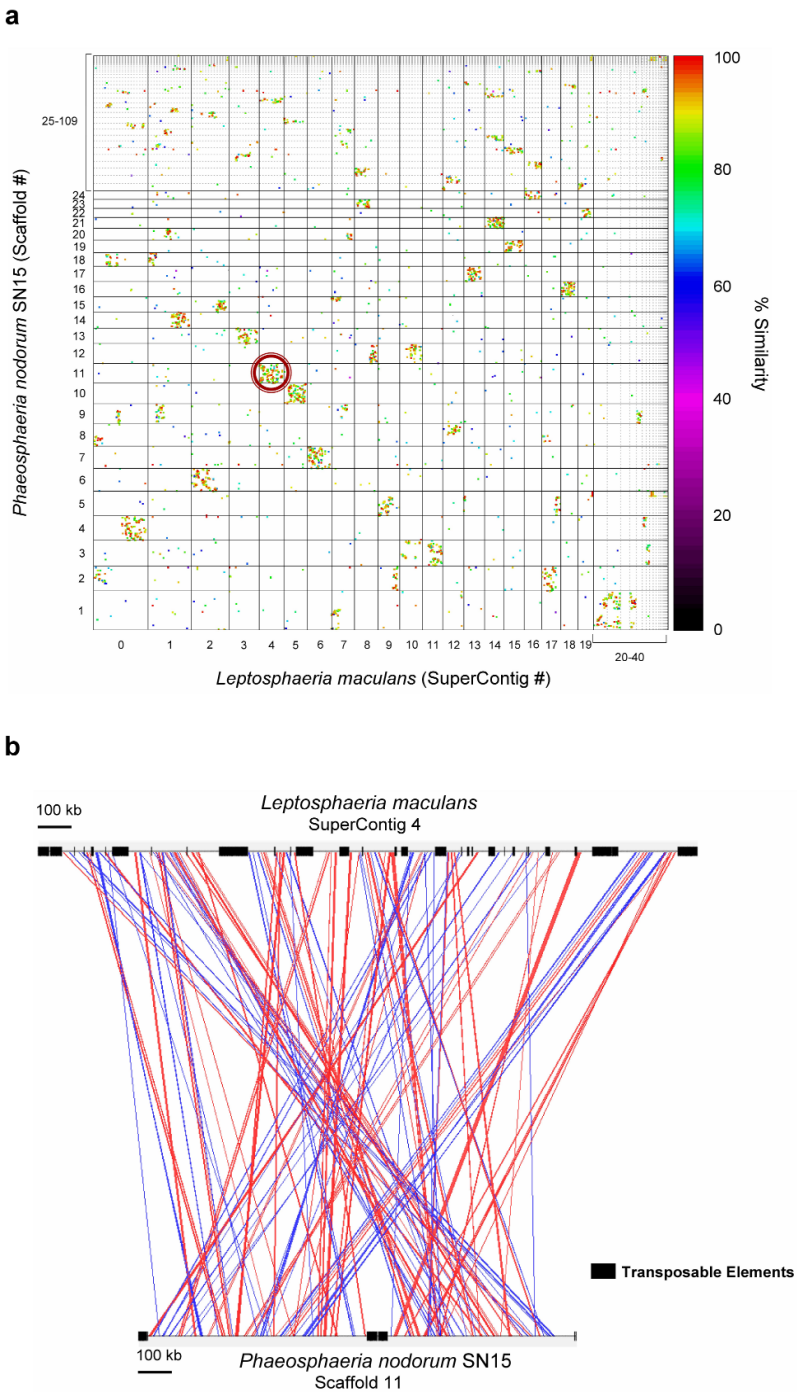
sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell* **19,** 3347-3368 (2007)

56. Cuifetti, L.M. *Pyrenophora tritici-repentis* database (2008) http://www.broadinstitute.org/annotation/genome/pyrenophora_tritici_repentis/Home.html
57. Turgeon, B.G. *Cochliobolus heterostrophus* C5 whole genome project (2008) http://genome.jgi-psf.org/CocheC5_1/CocheC5_1.home.html
58. Lawrence, C.B. *Alternaria brassicicola* whole genome project (2006) http://genome.jgi-psf.org/Altbr1/Altbr1.home.html
59. Goodwin, S.B. & Kema, G.H.J. *Mycosphaerella graminicola* whole genome project (2008) http://genome.jgi-psf.org/Mycgr3/Mycgr3.home.html

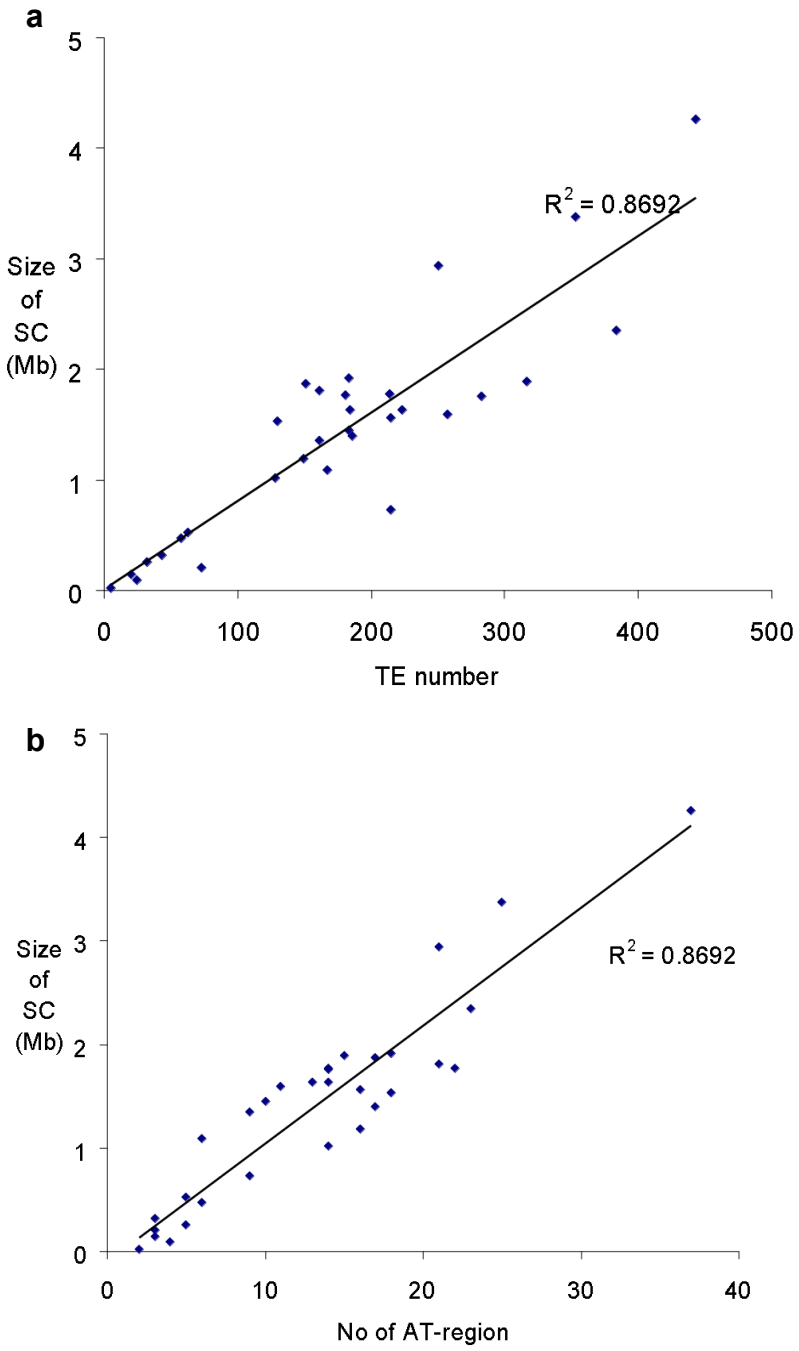
Supplementary figures



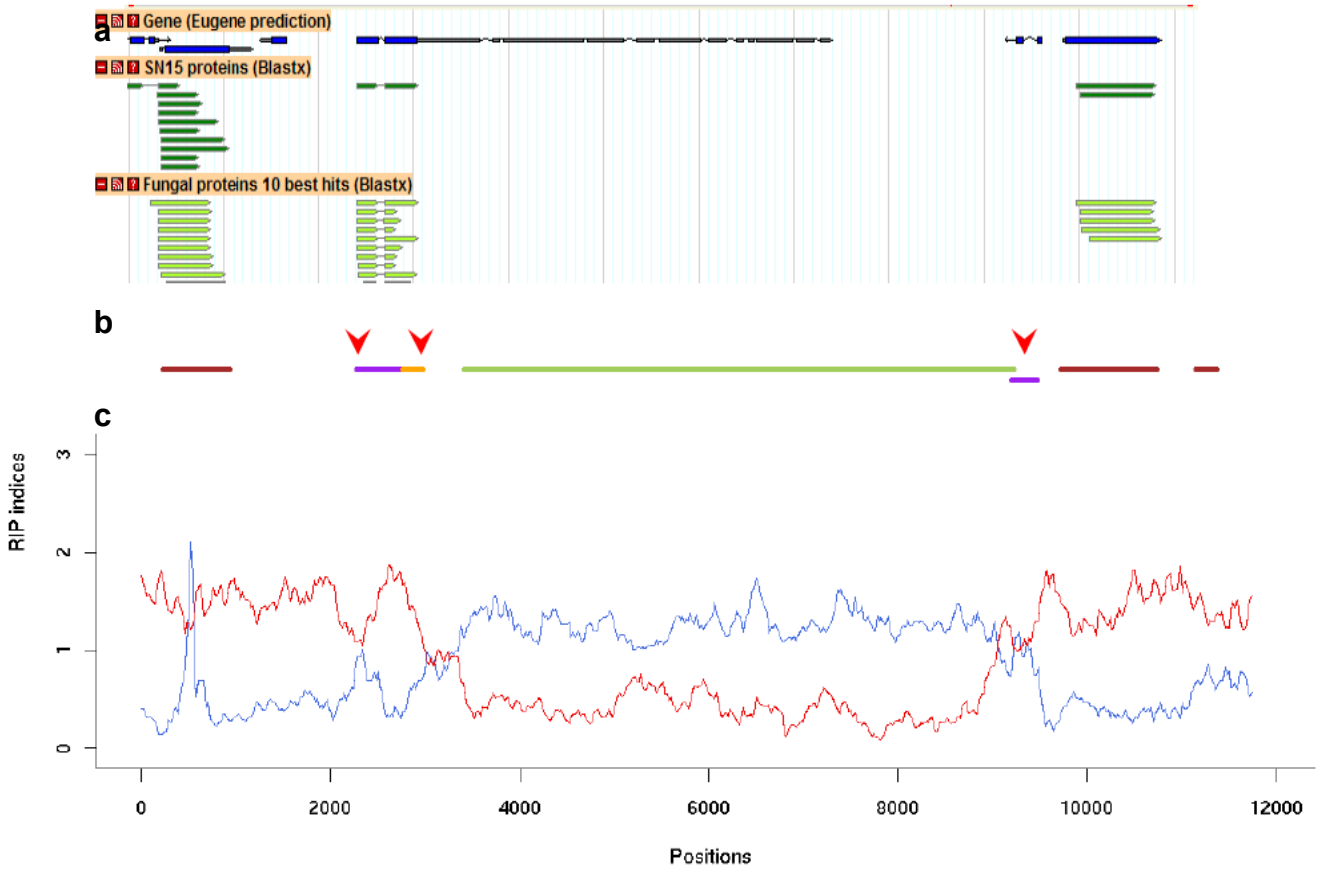
Supplementary Figure S1. Strategies to validate and improve the *Leptosphaeria maculans* genome assembly. The example of hybridisations of chromosomal DNA of isolate v23.1.3 separated by Contour-Clamped Homogeneous Field (CHEF) analysis with SuperContig (SC)-specific single copy probes. (**a**) Schematic representation of chromosomes and corresponding SCs based on a composite of gels run under various conditions that resolved DNA in different size ranges, and Southern blotting with single-copy probes. Hashed bands represent multiple chromosomes whilst black bands represent single chromosomes. (**b**) Chromosomal DNA resolved on 0.85% low-melting agarose in 0.5 X TBE; 3V/cm (100V); switching times: 500 sec for 70h followed by 420 sec for 48 h. Note that chromosomes above 3 Mb correspond to a compression zone and are usually difficult to separate by CHEF electrophoresis. Hw, *Hansenula wingei* chromosomes as molecular size markers. (**c**) *L. maculans* karyotype as generated by the combination of hybridisation, genetic mapping, telomere annotation and mesosynteny; in **c**, + between SCs indicates SCs joined to make up a chromosome.



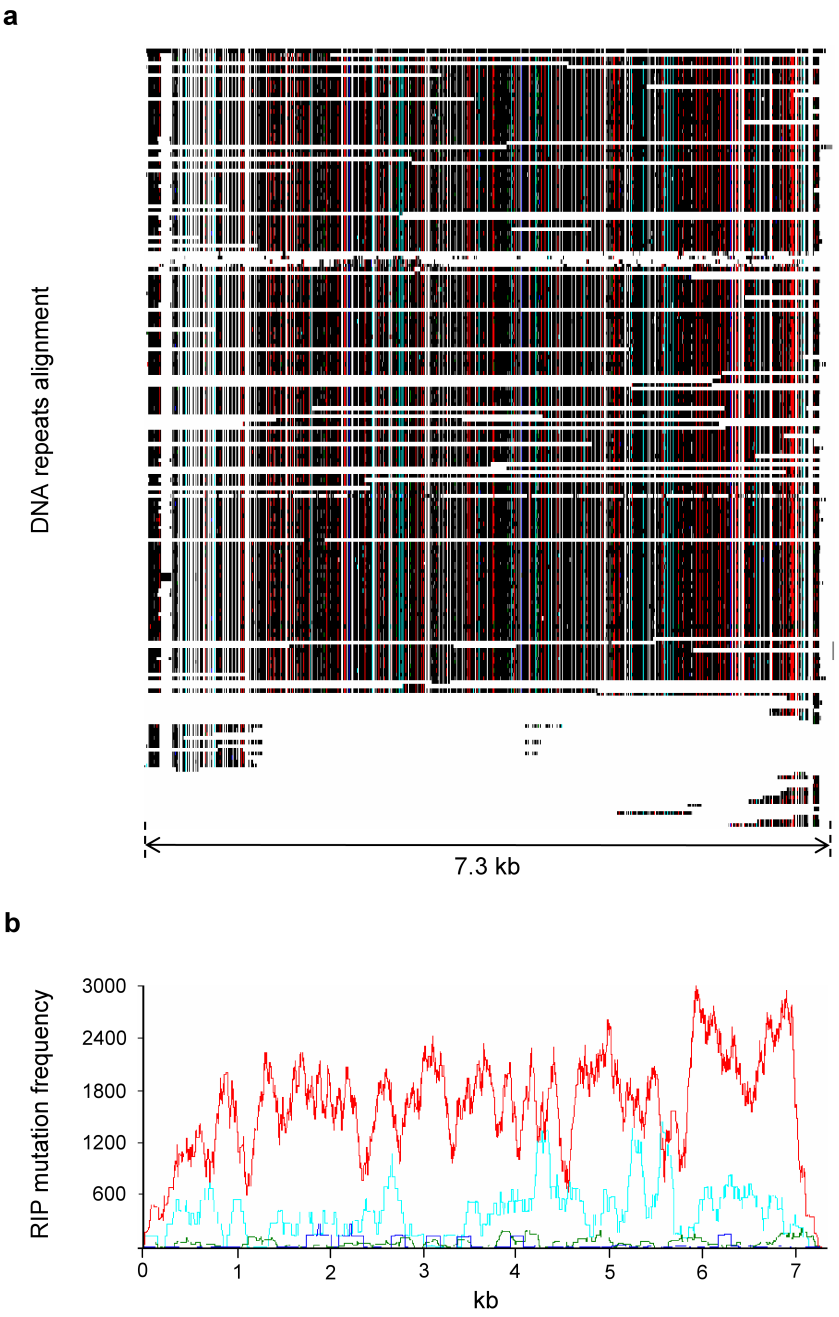
Supplementary Figure S2. **Mesosynteny analyses to validate the *Leptosphaeria maculans* genome assembly by comparison to the closely related Dothideomycete, *Phaeosphaeria nodorum*.** (**a**) A MUMMER-generated dotplot comparing SuperContigs/Scaffolds (SCs) 1-40 from *L. maculans* (x-axis) and SCs 1-109 from *P. nodorum* (y-axis). Dots represent matching regions (similar to the lines in **b**), between translated SC sequences. Presented as a dotplot, mesosyntenic regions appear as rectangular 'blocks' comprised of many dots (eg. red circle). (**b**) SC4 of *L. maculans* and SC11 of *P. nodorum* share multiple regions of similarity. Lines drawn between the sequence pair represent regions matching by tBLASTx with >75% identity over >500 bp length. Red lines indicate matches in the same orientation, whereas blue lines are matches in opposing orientations. In **b**, black boxes along SCs indicate transposable element-rich genomic regions.



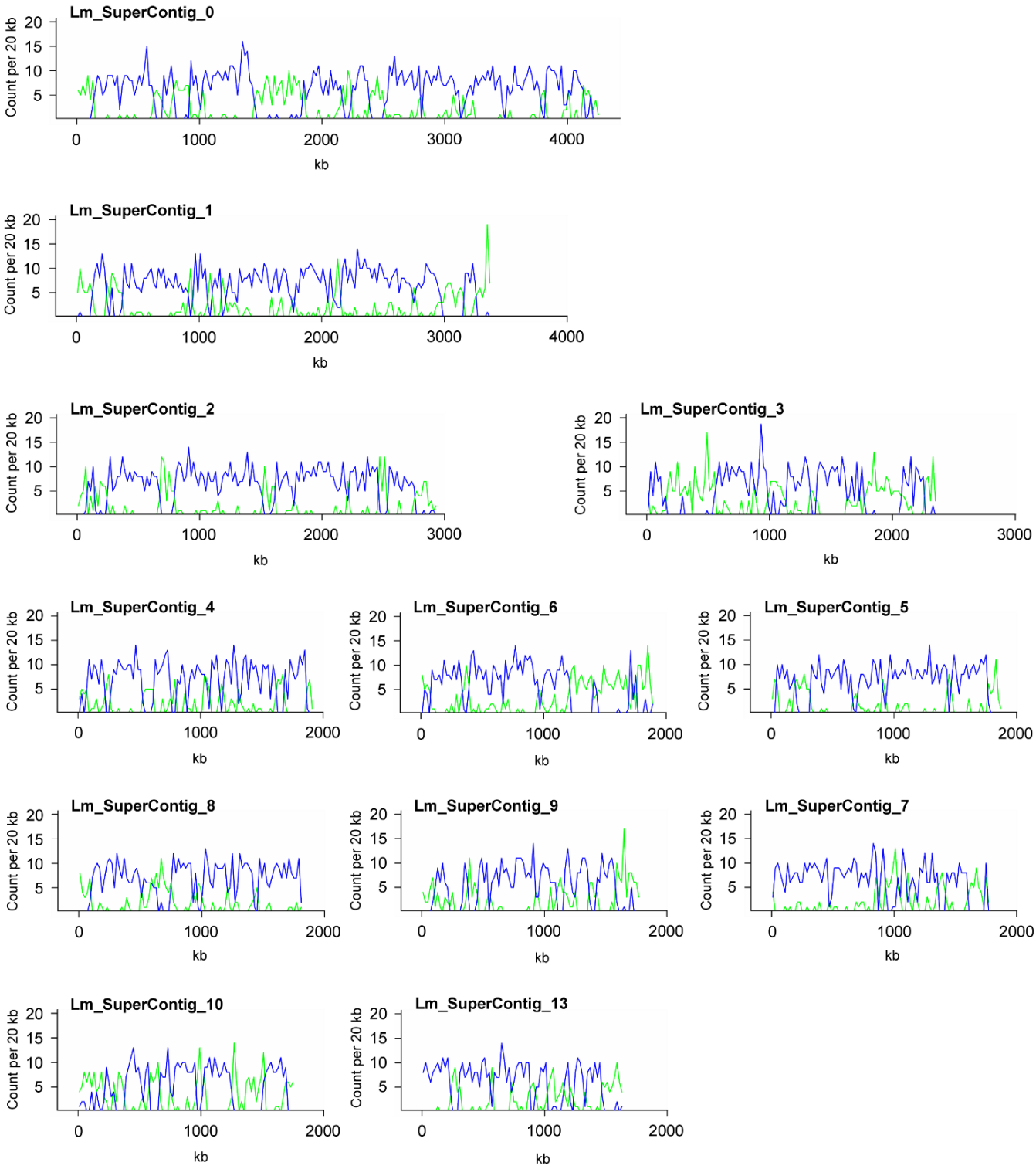
Supplementary Figure S3. Relationship between size of SuperContigs of *Leptosphaeria maculans* and their content of AT-rich regions or transposable elements. This analysis excludes SuperContigs (SCs) only made of one AT-rich region and SCs corresponding to extra-chromosomal DNA. (**a**) Relationship between size of SCs and the number of transposable element (TE) copies they host; (**b**) Relationship between size of SCs and their content in AT-rich genomic regions.



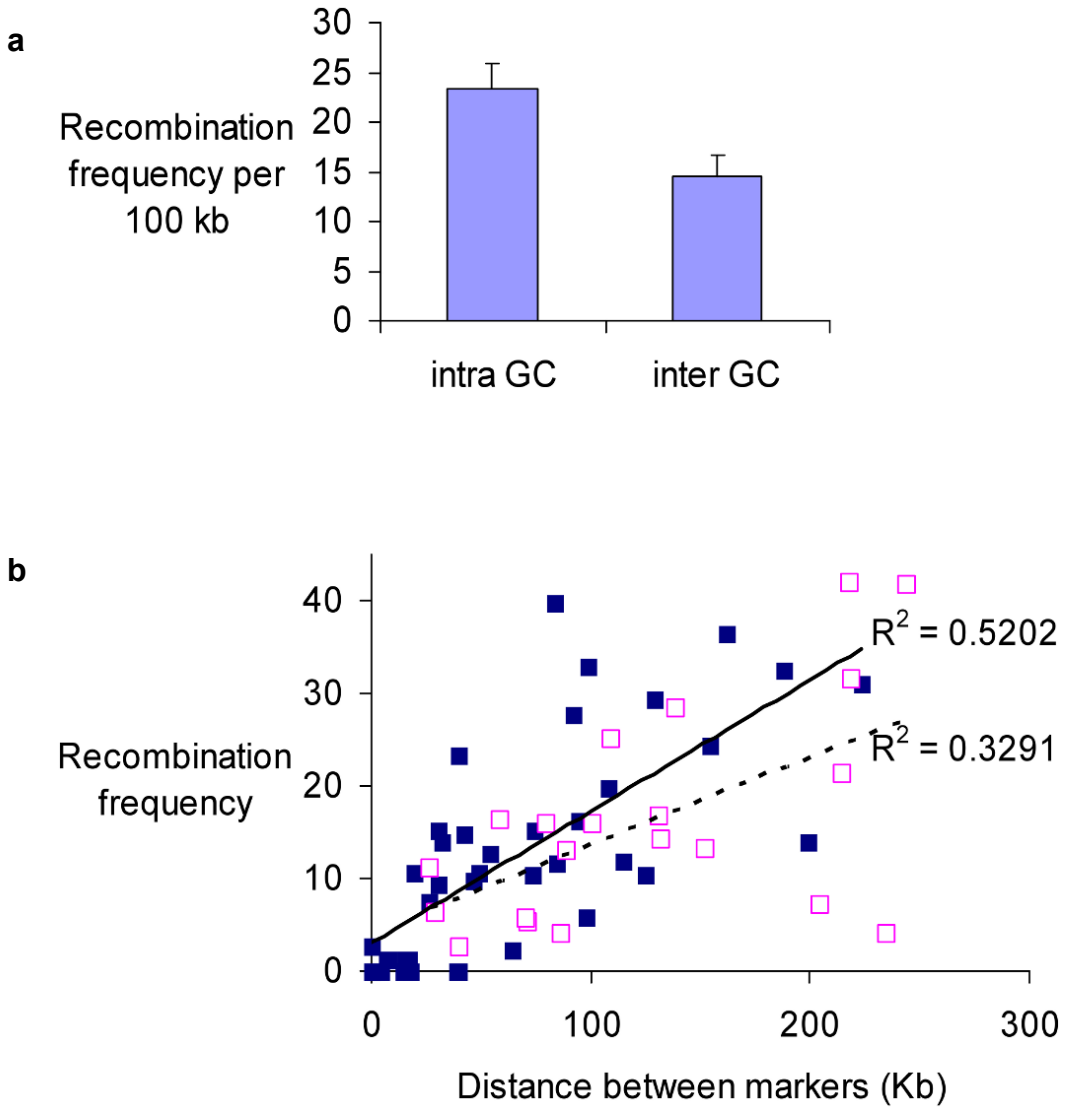
Supplementary Figure S4. Analysis of the Repeat-Induced Point mutation degeneracy gradient at the borders and within AT-rich genomic regions. The example of AT-block 02 in SuperContig_0, sized 7 kb, including 2.5 kb of flanking GC-equilibrated genomic region. (**a**) a snapshot of the *Leptosphaeria maculans* Gbrowser indicating automated EuGene gene prediction along with best blast hit of predicted genes with those of *Phaeosphaeria nodorum* isolate SN15 and other fungal proteins. (**b**) a synthetic representation of genes occurring in the region following manual reannotation. Color code as follows: green, transposable elements; orange, Small-Secreted Protein-encoding gene present in the AT-block; purple, other protein-encoding genes present in the AT-block; red, protein-encoding genes present in GC-block. The red arrowheads point to regions of the genes with altered Repeat induced Point (RIP) mutation indices. (**c**) Analyses of RIP indices TpA/ApT (blue line) and CpA+TpG/ApC+GpT (red line) computed on a 300-bp sliding window.



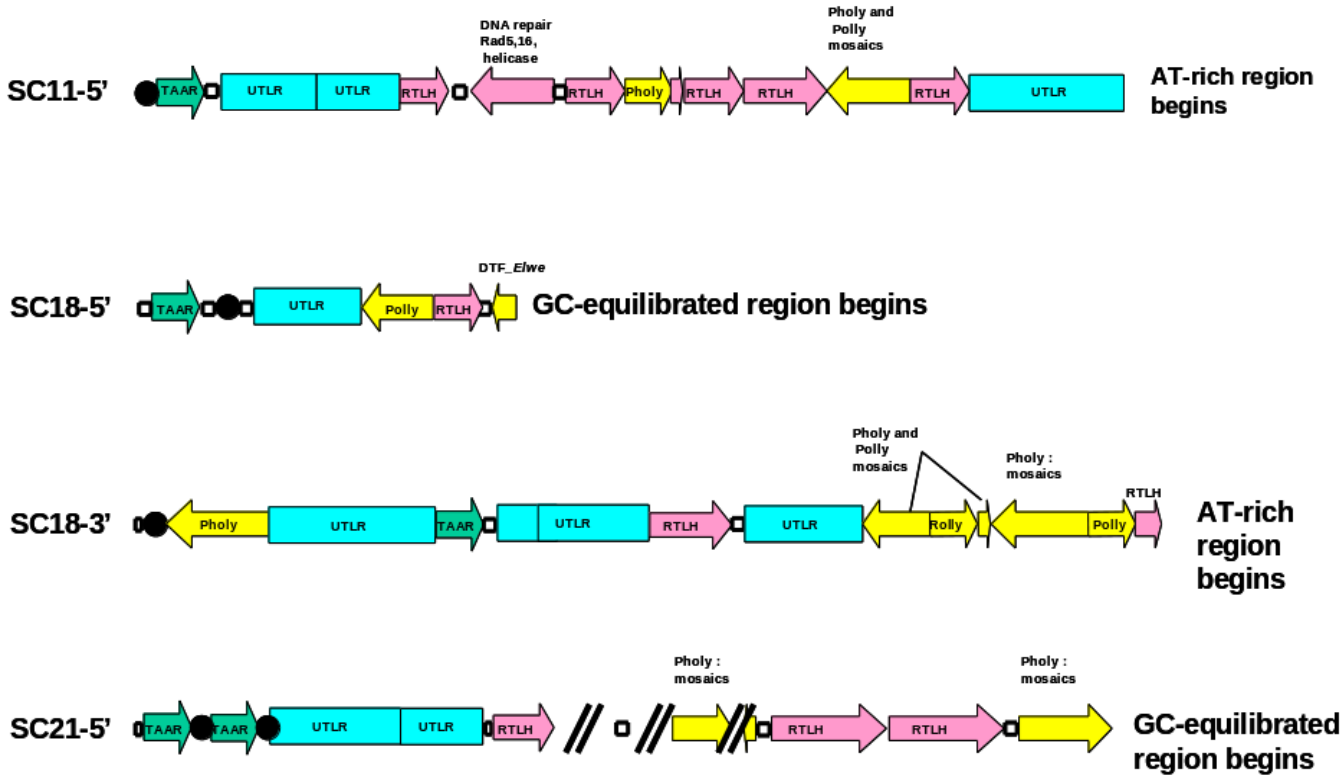
Supplementary Figure S5. **An example of Repeat-Induced Point mutation degeneracy in the DTM_*Sahana* transposable element family.** DTM_*Sahana* is a transposable element (TE) characterized by its common occurrence as a single TE inserted within GC-equilibrated, gene-rich genomic regions. (**a**) Multiple alignment of genome regions corresponding to repeat units of the DTM_*Sahana* family. Polymorphic nucleotides are coloured as a function of the type of Repeat-induced Point (RIP) mutation observed with black, invariant nucleotide, red, CpA ←→ TpA or TpG ←→ TpA mutations; dark blue, CpC ←→ TpC or GpG ←→ GpA mutations; pale blue, CpT ←→ TpT or ApG ←→ ApA mutations; green, CpG ←→ TpG or CpG ←→ CpA mutations. (**b**) RIP mutation frequency graph corresponding to the alignment directly above in **a**, and colour-coded as in **a**.



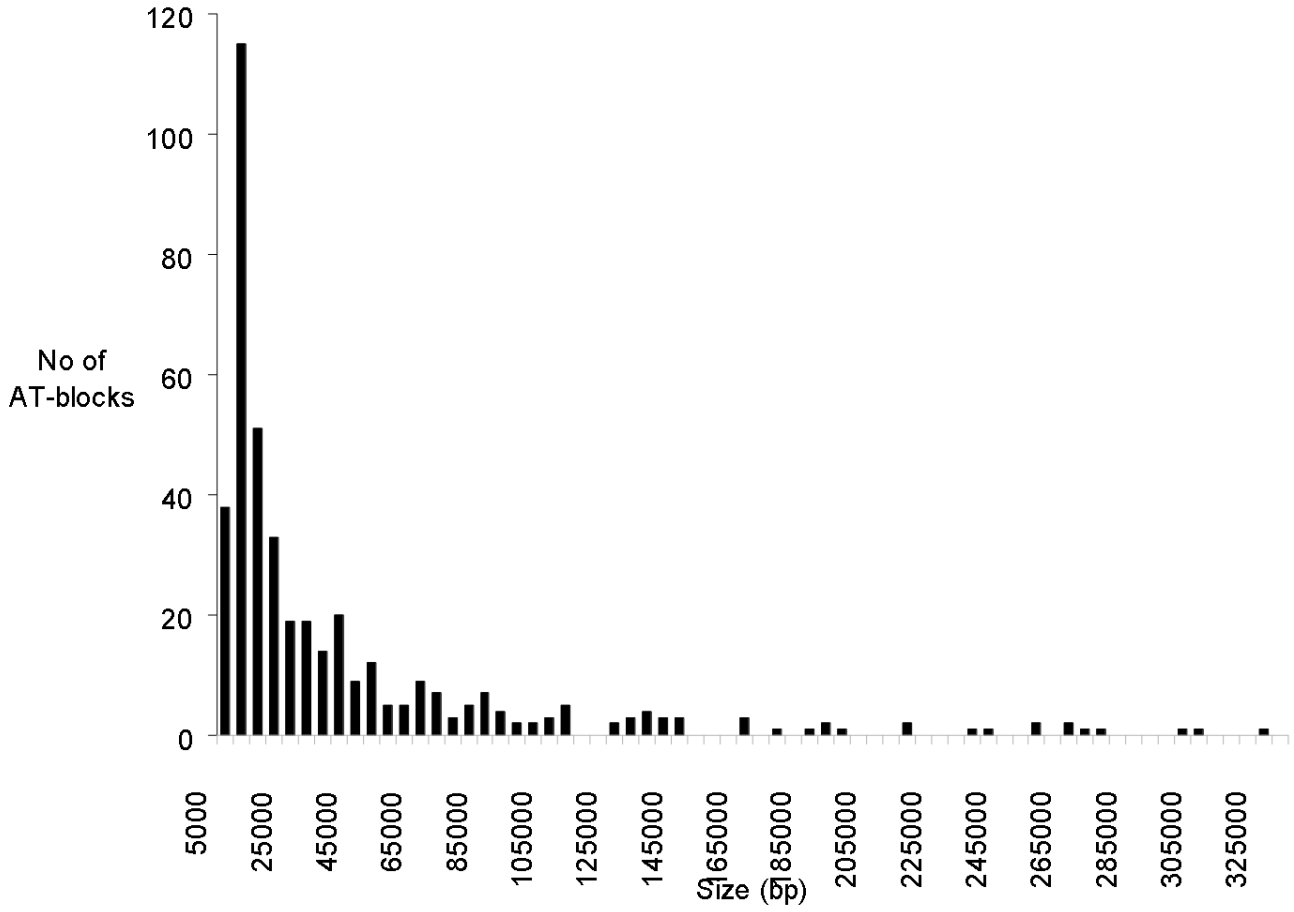
Supplementary Figure S6. Transposable element distribution along the main SuperContigs of the *Leptosphaeria maculans* genome, and general features of SuperContigs. Transposable element density is drawn in green and gene density is in blue.



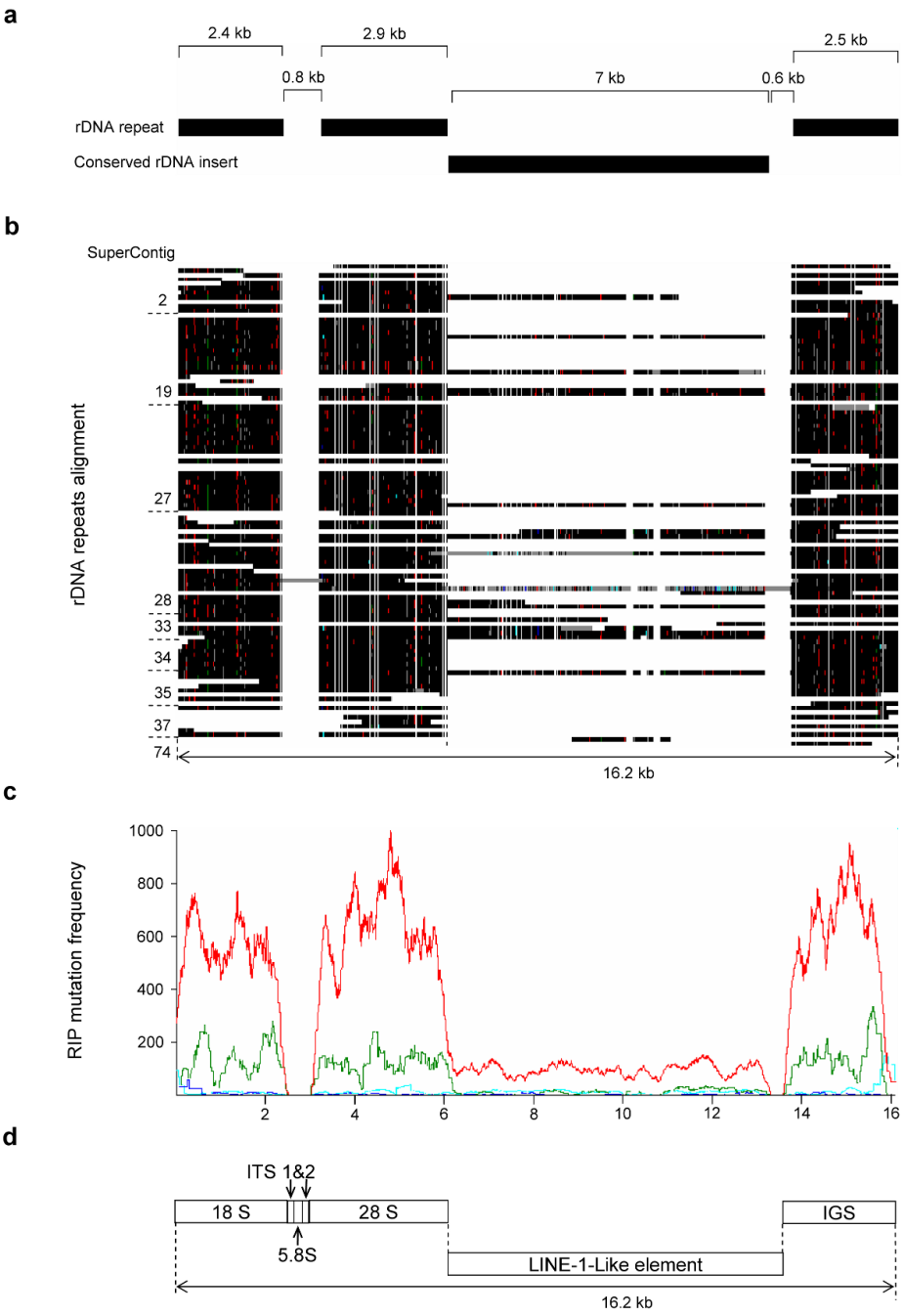
Supplementary Figure S7. Comparisons of recombination frequencies between contrasting genomic regions in *Leptosphaeria maculans* and relationship between physical and genetic distances. Data are micro-or minisatellites distributed along 7 SuperContigs (SCs) and include 41 pairs of markers located within the same GC-equilibrated genomic region (dark blue squares in **b**) and 21 pairs of markers located on both sides of one AT-rich region (pink empty squares in **b**). (**a**) Recombination frequencies per 100 kb between successive markers located within the same GC-equilibrated genomic region (intra GC) or spanning one AT-rich region (inter GC). (**b**) Relationship between physical and genetic distances. Plain line, linear regression curve for GC-located marker pairs; dotted line, linear regression curve for marker pairs spanning AT-blocks.



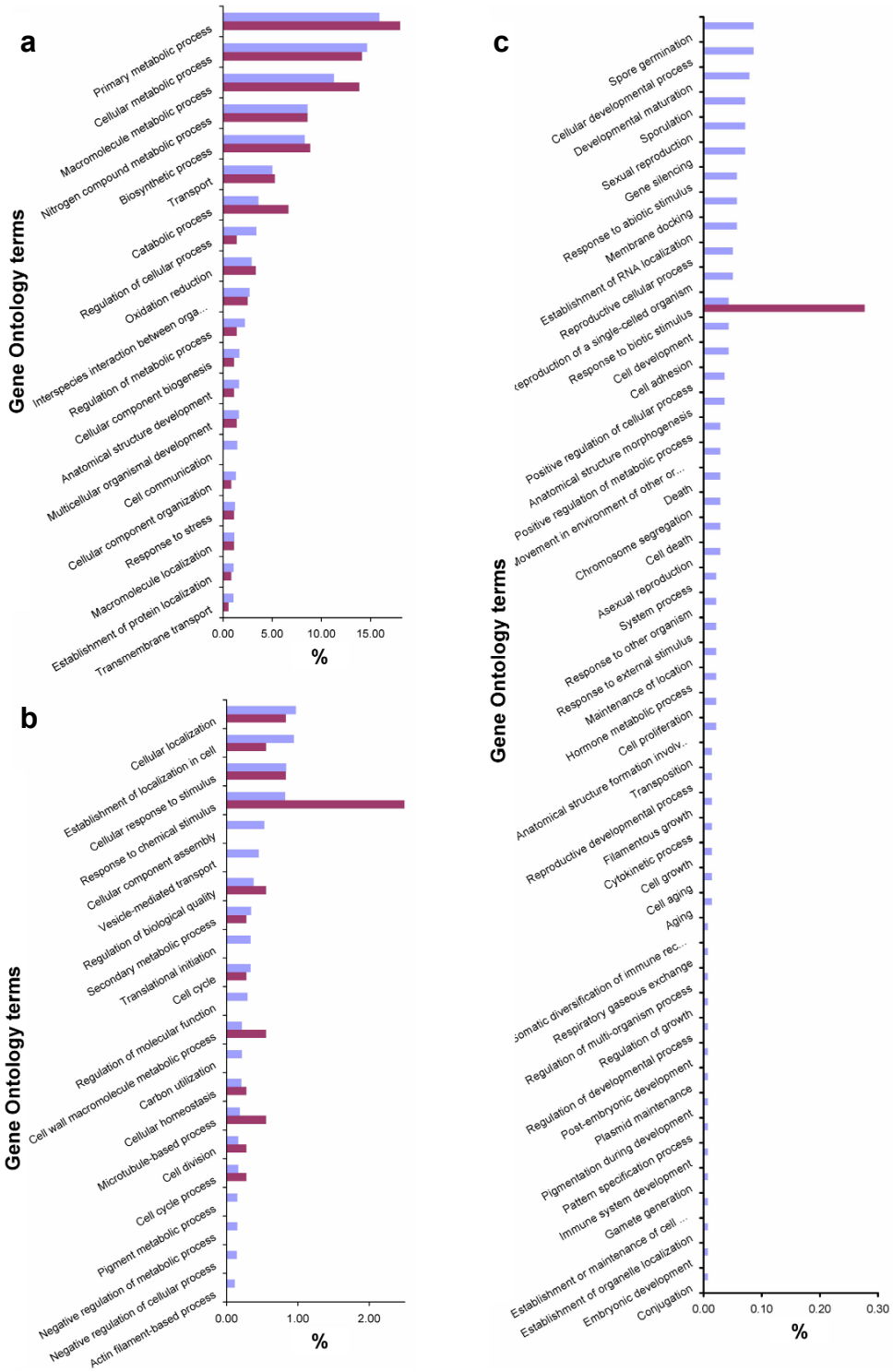
Supplementary Figure S8. Schematic representation of four telomeres in *Leptosphaeria maculans*. Telomeres typically encompass (i) $TTAGGG_n$ (represented as black circles; *e.g.*, $TTAGGG_{28}$ in the case of the 5' telomere of SuperContig (SC) 11) or related tandem repeats (represented as empty squares), (ii) one or two copies of RPP_*Circe*, a Telomere Associated *Athena* Retrotransposon (TAAR; green arrows), usually in the most distal part of the telomere, (iii) an ill-defined repeat, termed Unclassified Telomere Linked Repeat (UTLR, blue box), made up of numerous large tandem repeats and putative remnants of a helicase domain, and (iv) numerous copies of a RecQ Telomere Linked Helicase (RTLH, pink arrows). Non telomere-specific TEs may be interspersed with telomere-specific repeats and are represented as yellow arrows. Note All components of the telomeres depicted are severely degenerated by RIP and many TAAR and RTLH are truncated to various extents. The slash marks in the 5' telomere of SC21 represent gaps in the assembly.



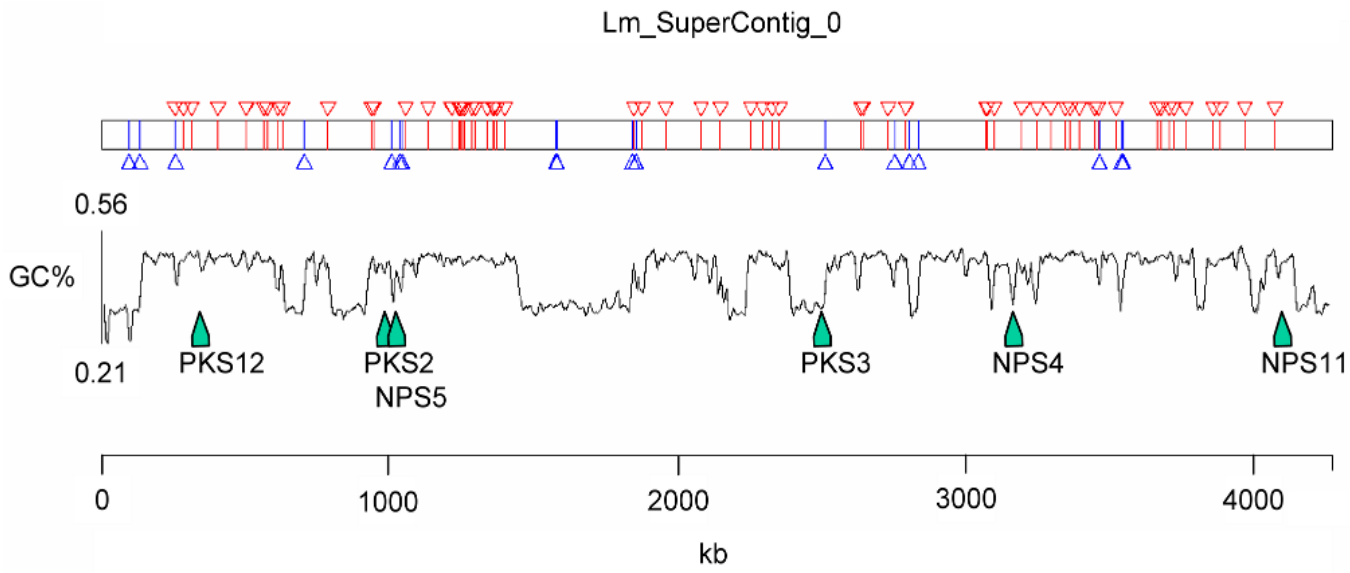
Supplementary Figure S9. Size distribution of AT-rich genomic regions in the final assembly. Four hundred and thirteen AT-blocks were identified from the SuperContig sequences and their size distribution plotted using 5000 bp windows.



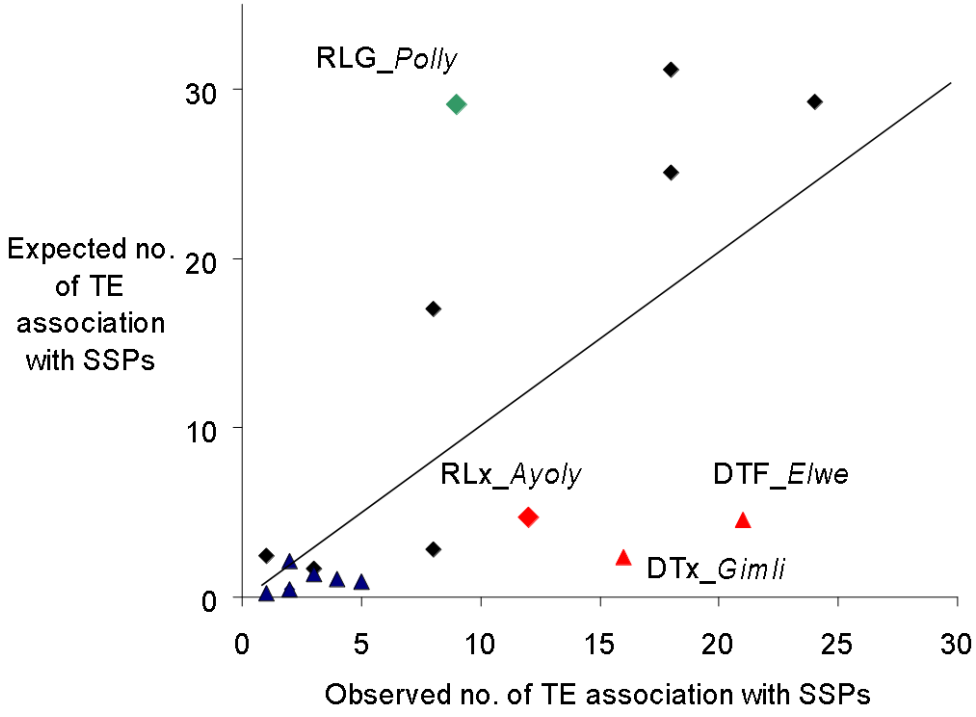
Supplementary Figure S10. RIPCAL analysis of the rDNA repeats of *Leptosphaeria maculans*. (**a**) Schematic representation of the rDNA repeat unit including size of the different parts and of the occasionally inserted sequences. (**b**) Multiple alignments of the genomic regions corresponding to the rDNA repeat unit and the SuperContig (SC) on which they are located. Repeat-induced Point (RIP) mutations (against the alignment consensus) are indicated by colour with black, invariant nucleotide; red, CpA ←→ TpA or TpG ←→ TpA mutations; dark blue, CpC ←→ TpC or GpG ←→ GpA mutations; pale blue, CpT ←→ TpT or ApG ←→ ApA mutations; green, CpG ←→ TpG or CpG ←→ CpA mutations. (**c**) RIP mutation frequency graph corresponding to the alignment directly above in **b** and colour-coded as in **b**. (**d**) Identification of the different components of the rDNA unit with a 7-kb insertion of the degenerated LINE element within the 3' end of the 28S gene.



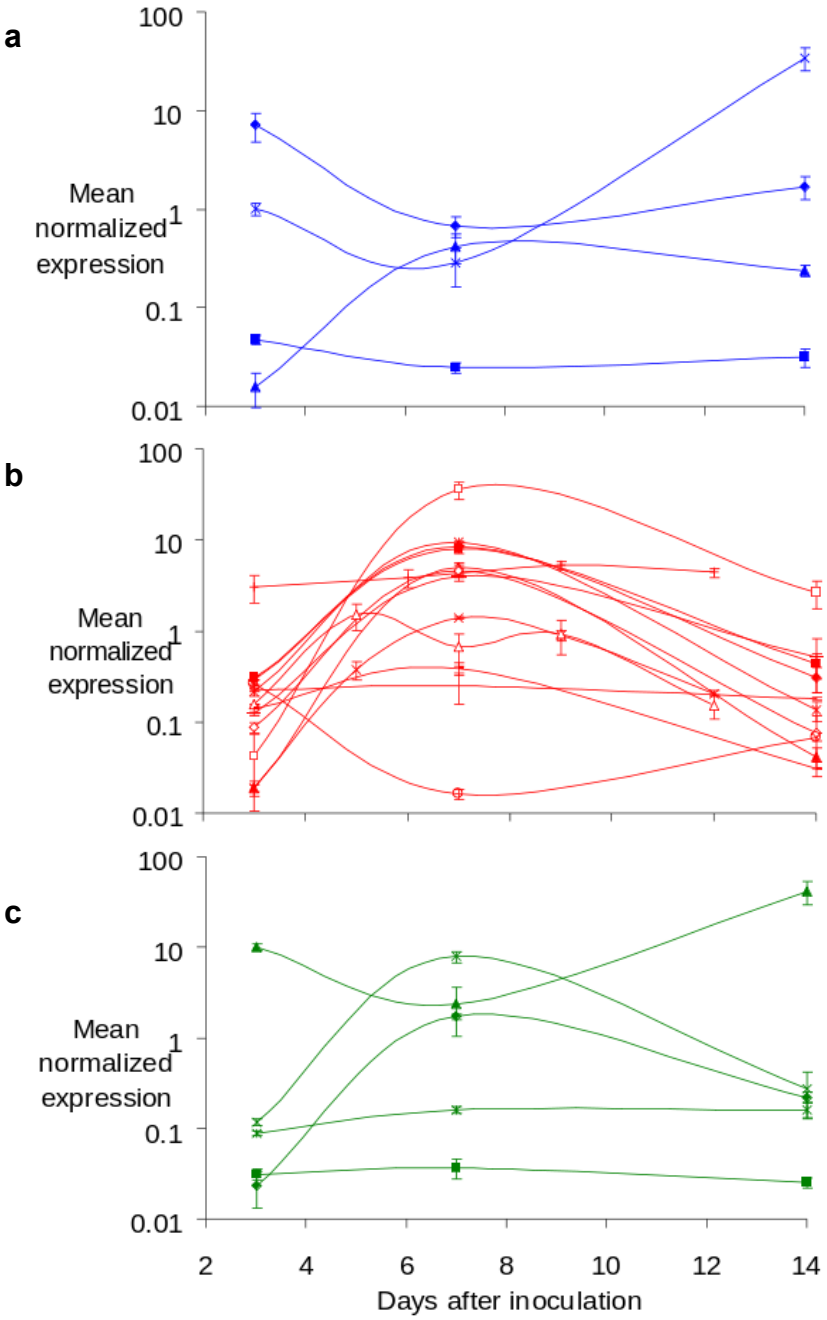
Supplementary Figure S11. Comparative Gene Ontology analysis of genes occurring in GC- and AT-blocks of the *Leptosphaeria maculans* genome. 5528 proteins (44.3% of the sequences) of the whole-genome set and 177 (35.5% of the sequences) of the AT-block set were annotated using Blast2GO and compared for Gene Ontology (GO) term occurrence, expressed as a percent of occurrence within the set. A figure is provided for each main branches of the GO: (**a**) Biological Process; (**b**) Cellular Component; (**c**) Molecular Function. Blue bars, occurrence within the whole genome; purple bars, occurrence within AT-blocks. Note that the scale differs between each panel.



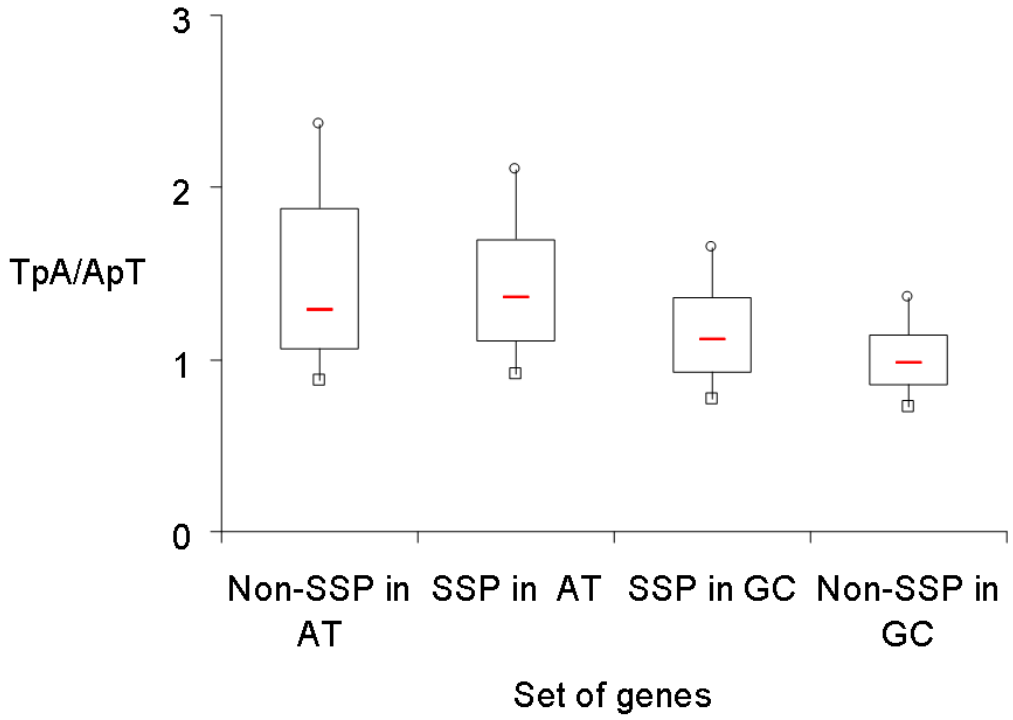
Supplementary Figure S12. Genomic location of putative effector genes along chromosomes of *Leptosphaeria maculans*: the example of SuperContig_0. Upper panel, location of genes encoding Small-Secreted Proteins (SSPs) along SuperContig (SC)0 with blue arrows, SSP-encoding genes in AT-blocks; red arrows, SSP-encoding genes in GC-blocks. Lower panel, GC content variation along SC0 defining the typical *L. maculans* genome isochore-like structure, and location of genes encoding for Non-ribosomal Peptide Synthases (NPS) and Polyketide Synthases (PKS).



Supplementary Figure S13. Relationship between transposable elements and Small Secreted Protein-encoding genes in the *Leptosphaeria maculans* genome. Diamonds, retrotransposons; triangles, DNA transposons. Large-sized diamonds and triangles indicate those elements that are significantly over-represented (red) or under represented (green) in the immediate vicinity of Small Secreted Protein (SSP)-encoding genes, according to a Chi-squared test for given probabilities. Black diamonds and triangles correspond to transposable elements, (TE) families which are not significantly over-represented in the vicinity of SSP-encoding genes.



Supplementary Figure S14. Q-RT-PCR analysis of *in planta* expression of selected Small Secreted Protein-encoding genes as a function of their genomic environment. (**a**) Blue curves: Small Secreted Protein-encoding genes (SSPs) in GC-equilibrated genomic regions; (**b**) red curves: SSPs in AT-rich regions; (**c**) green curves: SSPs located at the border between GC- and AT-blocks. Gene expression levels are relative to *β–tubulin*. Each data point is the average of three biological repeats (three extractions from different biological material) and two technical repeats. Standard error of the Mean Normalized Expression level is indicated by error bars.



Supplementary Figure S15. Box plot representation of the TpA/ApT RIP indices in four sets of genes. The four sets of genes are as follows: Small secreted protein (SSP)-encoding genes in GC-blocks (529 genes: SSP in GC); SSP-encoding genes within AT-blocks (57 predicted genes, SSP in AT); genes encoding other proteins (non-SSPs) in AT-blocks (498 genes, non-SSP in AT); and genes encoding other proteins in GC-blocks (11,394 genes) (non-SSP in GC). The red line represents the median value; the box includes values between the first and the third quartile of the distribution; squares and circles, 1st and 9th decile, respectively.

Supplementary tables

Supplementary Table S1. Taxa and sequences used in phylogenetic analyses.

Taxon	SSU[a]	LSU[a]	*RPB1*[a]	*RPB2*[a]	*TEF1*[a]
Ascochyta pisi	DQ678018	DQ678070		DQ677967	DQ677913
Ajellomyces capsulatum	Genome	Genome	Genome	Genome	Genome
Alternaria alternata	DQ678031	DQ678082		DQ677980	DQ677927
Arthrobotrys elegans	FJ176810	FJ176864		FJ238349	FJ238395
Botryosphaeria dothidea	DQ677998	DQ678051	EU186063	DQ677944	DQ767637
Botryotinia fuckeliana	AY544695	AY544651	DQ471116	DQ247786	DQ471045
Cladosporium cladosporioides	DQ678004	DQ678057	GU357790	DQ677952	DQ677898
Cladosporium herbarum	DQ678022	DQ678074	GU357793	DQ677971	DQ677918
Coccidioides immitis	Genome	Genome	Genome	Genome	Genome
Cochliobolus heterostrophus	AY544727	AY544645		AY544737	DQ247790
Conidioxyphium gardeniorum	GU296143	GU301807	GU357774	GU371743	GU349054
Corynespora cassiicola	GU296144	GU301808	GU357772	GU371742	GU349052
Debaryomyces hansenii	DHA508273	AF485980	XM_456921	CR382139	Genome
Geoglossum nigritum	AY544694	AY544650	DQ471115	DQ470879	DQ471044
Hypocrea lutea	AF543768	AF543791	AY489662	DQ522446	AF543781
Hysterobrevium smilacis	FJ161135	FJ161174	GU357806	FJ161114	FJ161091
Icmadophila ericetorum	DQ883704	DQ883694	DQ883723	DQ883711	DQ883730
Leotia lubrica	AY544687	AY544644	DQ471113	DQ470876	DQ471041
Leptosphaeria maculans	DQ470993	DQ470946	DQ471136	DQ470894	DQ471062
Magnaporthe grisea	AB026819	AB026819	Genome	Genome	Genome
Mycosphaerella fijiensis (anamorph *Pseudocercospora fijiensis*)	DQ767652	DQ678098	Genome	DQ677993	
Mycosphaerella graminicola	DQ678033	DQ678084	Genome	DQ677982	
Mycosphaerella punctiformis (anamorph *Ramularia endophylla*)	DQ471017	DQ470968	DQ471165	DQ470920	DQ471092
Nectria cinnabarina	U32412	U00748	AY489666	DQ522456	AF543785
Neurospora crassa	X04971	AF286411	Genome	Genome	Genome
Orbilia vinosa	DQ471000	DQ470952	DQ471145		DQ471071
Peltigera degenii	AY584681	AY584657	DQ782826	AY584688	DQ782897
Peziza vesiculosa	DQ470995	DQ470948	DQ471140	DQ470898	DQ471066
Phaeosphaeria ammophilae	GU296185	GU301859	GU357746	GU371724	GU349035
Phaeosphaeria nodorum	Genome	Genome	Genome	Genome	Genome
Phoma exigua	EU754084	EU754183	GU357813	GU371780	GU349080
Phoma herbarum	DQ678014	DQ678066	GU357792	DQ677962	DQ677909
Pyrenophora tritici-repentis		AY544672	Genome	Genome	DQ677882
Pyronema domesticum	DQ247813	DQ247805	DQ471166	DQ247795	DQ471093
Roccellographa cretacea	DQ883705	DQ883696		DQ883713	DQ883733
Saccharomyces cerevisiae	SCYLR154C	SCYLR154C	X96876	SCYOR151C	Genome
Schismatomma decolorans	AY548809	AY548815		DQ883715	DQ883725
Schizosaccharomyces pombe	X54866	Z19136	X56564	D13337	Genome
Sydowia polyspora	DQ678005	DQ678058	GU357791	DQ677953	DQ677899
Teratosphaeria associata	GU296200	GU301874	GU357744	GU371723	GU349025

[a] SSU, small subunit of the nuclear ribosomal RNA gene; LSU, large subunit of the nuclear ribosomal RNA gene; *RPB1* and *RPB2*, first and second largest subunits of RNA polymerase II, respectively; *TEF1*, translation elongation factor-1 alpha. When individual sequences are deposited in databases, their GenBank accession number is indicated.

Supplementary Table S2. Characteristics of the main SuperContigs of the *Leptosphaeria maculans* genome and their assembly into complete chromosomes using a combination of approaches (1/2).

Name	Size (bp)	GC%	N%	Telomere 5' size (kb)	Telomere 3' size (kb)	Validation of assembly	Validation of linkage to another SC	Assembled chromosome (Mb)	Linked with	Comments
Chromosomes										
SC0	4,258,568	45.70	0.99	43	61	Mesosynteny; Telomeres; Genetic map ; Hybridization		4.26		Complete chromosome
SC1	3,378,610	45.30	3.50	23	20	Mesosynteny; Telomeres; Genetic map ; Hybridization		3.38		Complete chromosome
SC2	2,939,989	45.00	4.75	76	none	Mesosynteny; Genetic map	Presence rDNA; Hybridization	>3.43	SC19?	Each SC hybridises to a band sized > 3.43 Mb, which may comprise a single chromosome sized 4.87 Mb or two chromosomes sized > 3.43Mb. The exact size is difficult to establish due to the lack of one telomere and presence of rDNA repeats in small SCs
SC19	1,186,800	43.20	1.76	none	none		Presence rDNA; Hybridization	>3.43	SC2?	
SC3	2,348,246	42.30	2.00	14	13	Mesosynteny; Genetic map ; Hybridization	Mesosynteny; Genetic map	2.49	SC31	
SC31	143,268	44.20	0.20	16	none				SC3	
SC4	1,918,205	46.60	0.59	30	19	Mesosynteny; Telomeres; Genetic map ; Hybridization		1.92		Complete chromosome
SC5	1,869,450	47.20	0.52	26	36	Mesosynteny; Telomeres; Genetic map ; Hybridization		1.87		Complete chromosome
SC6	1,888,674	42.80	2.46	32	none	Sequencing of a large genomic region with finishing[13]	Linkage between SC6 and SC11 validated by RFLP hybridization	2.60	SC29 and SC11	Reassembled chromosome is 1.1 Mb larger than expected on the basis of CHEF gel hybridisation. Comparisons between finished sequence data and the reassembly suggest SC11 is a chimera where one part is linked with SC6 and SC29 to make up a complete chromosome sized 2.60 Mb. Another 1 Mb is part of other uncharacterised chromosome(s).
SC29	200,94	33.60	5.98	none	none				SC6 and SC11	
SC11	1,590,160	42.70	0.45	78	13				SC6 and SC29	
SC7	1,769,547	46.00	0.32	7.5	9.9	Mesosynteny; Telomeres; Genetic map ; Hybridization		1.77		Complete chromosome
SC8	1,809,296	46.10	2.75	36	none	Mesosynteny; Genetic map	Hybridization	> 3.6	SC10	One telomere missing. Size consistent with lack of resolution of chromosomal DNA > 3.6 Mb in CHEF gels
SC10	1,758,670	40.20	9.47	none	none				SC8	
SC9	1,772,623	45.10	0.81	93	68	Mesosynteny; Telomeres; Genetic map ; Hybridization		1.77		Complete chromosome
SC12	1,631,710	43.70	3.23	2	42	Mesosynteny; Genetic map	Hybridization	4.10	SC15 SC32?	Dubious reassembly, so uncertain
SC15	1,560,629	43.70	2.62	none	none				SC12 and SC 32?	
SC32	87,679	37.50	0.00	40	none				SC12 and SC15?	
SC13	1,634,580	43.90	3.67	none	11	Hybridization to one chromosome sized > 3Mb	One telomeric end missing. No genetic map evidence	> 1.6	?	Unassembled SC
SC14	1,533,332	47.40	0.44	44	19	Telomeres; Hybridization		1.50		Complete chromosome

Supplementary Table S2. Characteristics of the main SuperContigs of the *Leptosphaeria maculans* genome and their assembly into complete chromosomes using a combination of approaches (2/2).

Name	Size (bp)	GC%	N%	Telomere 5' size (kb)	Telomere 3' size (kb)	Validation of assembly	Validation of linkage to another SC	Assembled chromosome (Mb)	Linked with	Comments
Chromosomes										
SC16	1,397,653	44.30	0.15	52	38	Mesosynteny; Telomeres; Hybridization		1.40		Complete chromosome
SC17	1,445,693	43.70	5.05	none	none	Mesosynteny		>1.44		Lack of telomeres. Unassembled SC
SC18	1,351,976	44.70	0.89	29	85.5	Telomeres	Mesosynteny	1.35		Complete chromosome
SC20	1,087,932	44.60	0.47	57	none	Genetic map	Mesosynteny; Hybridization	3.30	SC21 and SC23	
SC21	1,020,521	44.40	2.33	57	none				SC20 and SC23	Complete chromosome comprising SC20, SC21 and SC23
SC23	521,426	45.20	0.74	none	none				SC20 and SC21	
SC22	731,443	35.30	0.12	none	15.5	Hybridization		0.84		Lacks one telomere but size almost that of smallest chromosomal band. Probably minichromosome
Unassembled telomeric sequences										
SC24	475,869	44.50	0.37	none	51kb					
SC25	318,058	46.80	0.60	none	5.8kb					
SC26	261,54	45.80	0.29	32kb	none					
SC39	22,454									Telomeric repeats only
SC43	10,063									Telomeric repeats only
SC45	9396									Telomeric repeats only
SC53	7218									Telomeric repeats only
SC57	6329									Telomeric repeats only
SC64	5236									Telomeric repeats only
Unassembled rDNA sequences										
SC27	250,629	28.12	21.1							RIPped rDNA copies
SC28	236,098	28.90	18.0							RIPped rDNA copies
SC33	65,326	33.80	1.57							RIPped rDNA copies
SC34	58,596									RIPped rDNA copies
SC35	79,158									RIPped rDNA copies
SC37	52,193									RIPped rDNA copies
SC74	3630									RIPped rDNA copies
SC75	491									RIPped rDNA copies
Non chromosomal DNA										
SC30	154863	30.00	0.09							Mitochondrial DNA
SC50	7683									Linear plasmid pLm10
SC52	7314									Linear plasmid pLm9
SC60	6199									Linear plasmid

Supplementary Table S3. Gene model statistics for the *Leptosphaeria maculans* genome.

	All predictions	"Reliable" predictions[a]	"Less reliable" predictions[a]
Number of genes	12469	11561	908
Gene length median (bp)	1162	1254	250
Gene length mean (bp)	1323	1535	312
Coding gene length median (bp)	1002	1083	180
Exon length median (bp)	214	229	89
Intron length median (bp)	63	63	113
Gene with introns (%)	76	77	63
Introns/gene average	1.8	1.9	0.8
Density of genes (nb of genes/kb)	0.28	-	-

[a] "Reliable" predictions correspond to those whose coding sequence is > 300 nt or less than 300 nt with EST, domain and/or motif associated with the prediction; "less reliable" corresponds to coding sequences < 300 nt and without EST, domain, or motif support

Supplementary Table S4. Automated functional annotation of *Leptosphaeria maculans* genes.

Programs	No. of domains/motifs	No. of genes with domain/motif	
		Total	% genes
InterPro entry	**34,387**	**6577**	**52.75**
Superfamily	7330	4990	40.02
HMMPfam	9145	5959	47.79
HMMPanther	8212	5051	40.51
HMMSmart	3791	1712	13.73
HMMPIR	383	382	3.06
HMMTigr	861	710	5.69
PatternScan	2546	1936	15.53
BlastProDom	208	163	1.31
ProfileScan	3663	2165	17.36
FPrintScan	6021	1199	9.62
Gene3D	5325	3737	29.97
Seg	26,243	8410	67.45
Coil	2888	1770	14.20
SignalPHMM	2400	2399	19.24
TMHMM	8892	2111	16.93
Genes with at least one domain/motif identified	**78,61**	**10,434**	**84.73**

Supplementary Table S5. Comparative analyses of intergenic distances between *Leptosphaeria maculans* and the closely-related *Dothideomycete*, *Phaeosphaeria nodorum*.

	Intergenic distance (bp) [a]		
	L. maculans **AT-blocks**	***L. maculans*** **GC-blocks**	***P. nodorum*** **whole genome**
TYPE A: two genes in the same orientation, ie. ----> ---->			
Mean	6006.83	1197.15	1854.98
Median	2489	990	1083
Count	344	3495	5119
TYPE B: two genes in the opposite orientation, meeting at terminators, **ie. ----> <-----**			
Mean	4731.03	910.37	1577.96
Median	2006	670	670.5
Count	172	2623	2758
TYPE C: two genes in the opposite orientation, meeting at promoters, **ie. <----- ---->**			
Mean	5874.63	997.26	1512.43
Median	1991	781	970
Count	180	2599	2789

[a] Genes separated by gaps in the assembly were not analysed

Supplementary Table S6. Orthologs of the *Neurospora crassa* factors necessary for gene silencing[18] identified in the genome of *Leptosphaeria maculans.*

	Function	*L. maculans*[a]	*N. crassa*[a]
RIP[b]			
RID	Putative DMT, essential for RIP and MIP	Lema_P040230.1	NCU02034.3
Dim-5	H3 3mK9 HMT essential for RIP	Lema_P050470.1	NCU04402.3
Quelling			
QDE-1	RdRP, essential for quelling	Lema_P088990.1	NCU07534.3
QDE-2	Argonaute-like protein, essential for quelling	Lema_P015760.1	NCU04730.3
QDE-3	RecQ helicase, essential for quelling	Lema_P098920.1	NCU08598.3
DCL1	Dicer-like protein, involved in quelling	Lema_P036310.1	NCU08270.3
DCL2	Dicer-like protein involved in quelling	Lema_P041660.1	NCU06766.3
QIP	Putative exonuclease protein, involved in quelling	Lema_P099110.1	NCU00076.3
MSUD			
SAD-1	RdRP essential for MSUD	Lema_P117350.1	NCU02178.3
SAD-2	Essential for MSUD	None	NCU04294.3
DNA methylation			
Dim-2	*De novo* CpN DMT / maintenance CpG DMT	Lema_P122420.1	NCU02247.3
HP1	Heterochromatin factor, essential for CpN methylation	Lema_P110410.1	NCU04017.3
Chromatin remodelling factors			
HDA6	Histone deacetylase involved in CpG methylation	Lema_P067960.1	NCU00824.3
SIR2	NAD-dependant histone deacetylase involved in TGS	Lema_P055610.1	NCU04737.3
DDM1	SW12 / SNF2-like protein involved in CpN methylation	Lema_P071270.1	NCU03875.3

[a] GenBank accession number for *N. crassa* genes and ID of their genome ortholog in *L. maculans*.
[b] Abbreviations: MIP, Methylation Induced Premeiotically; RIP, Repeat Induced Point mutation; MSUD, Meiotic Silencing of Unpaired DNA; DMT: DNA Methyltransferase, HMT: Histone Methyltransferase, RdRP : RNA dependent RNA polymerase, TGS: transcriptional gene silencing.

Supplementary Table S7. Occurrence of AT-blocks in SuperContigs and size of the borders between AT-rich and GC-equilibrated genomic regions.

SuperContig [a]	Size (bp)	Number of AT-rich regions	Total size of AT-rich regions in the SC (bp)	SC coverage (%)	Left border size (bp)	Right border size (bp)
SC0	4,258,568	37	1,359,311	31.9	1087	1065
SC1	3,378,610	25	907,84	26,9	750	1036
SC2	2,939,989	21	799,223	27,2	680	674
SC3	2,348,246	23	1,122,437	47.8	1530	879
SC4	1,918,205	18	535,669	27,9	1316	1430
SC5	1,869,450	15	484,191	25,9	1160	1282
SC6	1,888,674	17	839,691	44.4	973	863
SC7	1,769,547	21	532,22	30,1	528	567
SC8	1,809,296	22	437,727	24,2	1113	945
SC9	1,772,623	14	610,287	34.4	750	946
SC10	1,758,670	14	789,897	44.9	740	960
SC11	1,590,160	14	783,033	49.2	1046	1191
SC12	1,631,710	13	630,489	38.6	1686	1076
SC13	1,634,580	11	551,321	33.7	750	889
SC14	1,533,332	16	377,531	24,6	1161	855
SC15	1,560,629	18	586,22	37.6	705	807
SC16	1,397,653	10	569,872	40.8	893	939
SC17	1,445,693	17	449,508	31,1	643	444
SC18	1,351,976	9	527,452	39.0	890	790
SC19	1,186,800	16	510,07	43.0	708	809
SC20	1,087,932	6	462,635	42.5	1491	1998
SC21	1,020,521	14	359,32	35.2	617	744
SC22	731,443	9	676,652	92.5	702	783
SC23	521,426	5	177,245	34.0	517	766
SC24	475,869	6	198,348	41.7	913	756
SC25	318,058	3	73,603	23,1	299	673
SC26	261,54	5	86,848	33.2	654	615
SC29	200,94	3	189,534	94.3	999	2239
SC31	143,268	3	59,241	41.3	771	423
SC32	87,679	4	73,636	84.0	402	356
SC39	22,454	2	10,513	46.8	252	310
SC45	9396	2	5958	63.4	256	301

[a] Excluding SuperContigs (SCs) only made of one AT-rich region and SCs corresponding to extra-chromosomal DNA; note that in the sequence databases, 'SCx' is referred to as 'lm_SuperContig_x_v2'.

Supplementary Table S8. A data matrix of individual nesting of Transposable Elements in the *Leptosphaeria maculans* genome.

	Invader																	
	RLC_ *Pholy*	RLG_ *Olly*	RLG_ *Polly*	RLG_ *Rolly*	DTF_ *Elwe*	RLx_ *Ayoly*	DTM_ *Sahana*	DTx_ *Gimli*	DTM_ *Lenwe*	RLx_ *Jolly*	RLC_ *Zolly-2*	RLG_ *Dolly*	DTM_ *Ingwe*	RLC_ *Zolly-1*	DTT_ *Finwe-1&-2*	RLG_ *Brawly*	**TOTAL invaded**	**Invaded/ invader**
Invaded																		
RLC_*Pholy*	*12* *[a]	26	46	24	3	1	0	1	0	8	0	4	0	0	0	2	127	0.76
RLG_*Olly*	32	21	35	18	1	1	0	1	0	7	0	2	0	0	0	1	119	0.92
RLG_*Polly*	42	22	*12* *	22	1	0	0	0	0	5	0	3	0	0	0	0	107	0.63
RLG_*Rolly*	26	17	19	*3* *	0	0	0	2	0	4	0	0	0	1	0	0	72	0.86
DTF_*Elwe*	**13** *	**21** *	**21** *	2	2	**3** *	0	0	0	**6** *	0	**2** *	0	0	0	0	70	7.00
RLx_*Ayoly*	**20** *	1	12	**10** *	1	0	1 *	1	0	0	0	2	0	0	0	0	48	8.00
DTM_*Sahana*	5	8	10	5	**2** *	0	0	0	0	0	0	1	0	0	0	0	31	10.33
DTx_*Gimli*	2	2	2	0	0	**1** *	0	**2** *	0	1	0	**3** *	0	0	0	0	13	1.86
DTM_*Lenwe*	**2** *	1	**5** *	0	0	0	0	0	0	0	0	**1** *	0	0	0	0	9	n/a
RLx_*Jolly*	**2** *	0	1	0	0	0	**2** *	0	0	**1** *	0	**1** *	1	0	0	0	8	0.24
RLC_*Zolly-2*	3 *	1	**3** *	0	0	0	0	0	0	0	0	0	0	0	0	0	7	n/a
RLG_*Dolly*	**3**	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0.32
DTM_*Ingwe*	3 *	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	6	6.00
RLC_*Zolly-1*	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4.00
DTT_*Finwe-1 &-2*	1	**4** *	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	n/a
RLG_*Brawly*	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0.67
TOTAL invader	167	130	171	84	10	6	3	7	0	33	0	19	1	1	0	3		

[a] Numbers followed by a * correspond to those families behaving significantly more (bolded numbers) or less (italicized numbers) as an invader (P = 0.05) as compared to what is expected in the case of random invasion (according to a chi-squared test for given probabilities with simulated P values).

Supplementary Table S9. Non-ribosomal peptide synthetases of *Leptosphaeria maculans*: gene location, orthologs, and putative function.

NRPS	SC location	Np. of modules and size (aa)	Closest matches	% similarity	Synteny with *Phaeosphaeria nodorum*	Putative function	Gene cluster	Location
SirP	13 (1399536-1391626)	2 2176	*Aspergillus fumigatus* GliP-like XP_754329 *Neosartorya fischeri* GliP2 XP_001263173 *A. fumigatus* GliP2 EDP52461	53.6 53.1 52.0	No homolog	Sirodesmin biosynthesis	Yes	GC-equilibrated region
maa1	5 (436833- 440678)	1 1282	*Pyrenophora tritici-repentis* EDU39885 *Cochliobolus heterostrophus* NPS10 AAX09992 SNOG_07126 XP_001797479	94.9 94.5 91.1	Limited	unknown	No	GC-equilibrated region
NPS1	12 (681838-685052)	1 1038	*Gibberella zeae* XP_382427 *Penicillium chrysogenum* CAP85555.1 *A. nidulans* XP_662431 *Botryotinia fuckeliana* XP_001548852 SNOG_01846 XP_001792471	76.1 64.7 62.2 59.9 48.5	None	unknown	No	AT-rich region upstream
NPS2	10 (271464-253889)	4 5592	*C. heterostrophus* NPS2 AAX09984 SNOG_02134 XP_001792752 *Alternaria brassicicola* NPS2 ABU42595	68.1 51.7 25.7	Limited	Ferricrocin (intracellular siderophore) biosynthesis	Yes	AT-rich region upstream
NPS3	9 (490591-496848)	1 1579	*P. tritici-repentis* EDU45320 SNOG_03771 XP_001794318	53.7 53.7	Yes	unknown	Yes	Flanked by AT-rich region
NPS4	0 (3115396-3137646)	4 7187	*A. brassicae* NPS2 AAP78735 NRPS1 *C. heterostrophus* NPS4 AAX09986 *P. tritici-repentis* EDU41238 SNOG_14834 XP_001805009 SNOG_04863 XP_001795276	82.0 80.0 79.7 77.2 72.0	Limited	Role in virulence ? HC-toxin synthetases *(P. tritici repentis*)	No	AT-rich region downstream
NPS5	0 (1015484-1021453)	1 1350	*Talaromyces stipitatus* EED13476 *A. terreus* XP_001212754 SNOG_09488 XP_001799780	51.6 15.6 11.2	Limited	Unknown	No	Flanked by AT-rich region
NPS6	7 (898191-890963)	1 2122	*C. miyabeanus* ABI51982 *C. heterostrophus* NPS6 AAX09988 SNOG_14368 XP_001804557	73.8 73.7 73.9	Yes	Coprogen (extracellular siderophore) biosynthesis	Yes	In AT-rich region
NPS7	9 (1709321-1712394)	1 971	SNOG_03620 XP_001794175 *A. niger* XP_001393822 *P.marneffei* XP_002146735.1	84.0 52.0 52.0	Limited	unknown	Yes	In AT-rich region
NPS8	10 (21848-3955)	5 5868	SNOG_09081 XP_001799383 *A. nidulans* XP_660149 *A. nidulans* XP_681153	52.1 49.9 51.6	None	Cyclic peptide/ phomalide biosynthesis?	No	Flanked by AT-rich region
Lys2	4 (428379-424738)	1 1179	*P. tritici-repentis* SNOG_07784 XP_001798111 *A. nidulans* XP_663214	95.8 93.5 64.4	Limited	α-aminoadipate reductase (lysine biosynthesis)	No	GC-equilibrated region
NPS10	11 (1553383- 1550441)	1 981	*P.tritici-repentis* XP_001931634 *Chaetomium globosum* XP_001228818 *Podospora anserina* XP_001903778 *P. tritici-repentis* EDU40739	94.0 64.2 60.1 58.3	No homolog	Unknown	Yes	AT-rich region downstream
NPS11	0 (4083398–4088008)	1 1256	*T. stipitatus* EED16026.1 *G. zeae* XP_383912 *N. fischeri* XP_001263102	76.0 67.8 61.5	None	Unknown	?	In AT-rich region

Supplementary Table S10. Polyketide Synthases of *Leptosphaeria maculans*: gene location, orthologs, and putative function.

PKS	SC location	Closest matches	% similarity	Synteny with *Phaeosphaeria nodorum*	Putative function	Gene cluster
PKS1	13 (1417543-1424481)	*Aspergillus flavus* PksA AAS89999 *A. oryzae* XP_001821511 *A. parasiticus* Q12053	67.5 67.5 67.3	SNOG_08614 46.2%	Aflatoxin	Yes
PKS2	0 (988418–996463)	*Phaeosphaeria nodorum* XP_001795280 *Podospora anserina* XP_001904919 *Alternaria solani* BAD83684.1	93.8 71.3 50.0	SNOG_04868	Type I PKS Alternapyrone in *Alternaria solani*	Yes
PKS3	0 (2503452-2513155)	*Botryotinia fuckeliana* AAR90244 *Chaetomium globosum* XP_001226673 *Magnaporthe grisea* XP_360215	50.6 50.5 50.1	SNOG_11066 52.1%	Type I PKS	Yes
PKS4	1 (1511847–1521828)	*A. solani* BAE80697 *Cochliobolus heterostrophus* AAR90264 *Pyrenophora tritici-repentis* XP_001930443	55.0 55.1 54.7	None	Type 1 PKS Phenolpthiocerol ppsA	No
PKS5	11 (1300692-1306965)	*Penicillium chrysogenum* CAP95290 *P. tritici-repentis* XP_001942484 *A. clavatus* XP_001273596	53.4 51.8 50.9	None	Type 1 PKS Lovastatin synthesis	No
PKS6	11 (273779-282201)	*Acremonium strictum* CAN87161 *Coccidioides immitis* XP_001243185 *A. niger* XP_001393501	65.6 63.2 59.5	SNOG_06682 47%	Methylorcinaldehyde synthase Citrinin toxin	No
PKS7	12 (88853-96821)	*P. tritici-repentis* XP_001937136 *A. clavatus* XP_001270321 *P. anserina* XP_00190896	79.7 55.9 53.7	SNOG_05791 45.5%	Phenolpthiocerol ppsA Alternapyrone	Yes
PKS8	14 (1277125-1284521)	*Gibberella zeae* ABB90283 *Hypomyces subiculosus* ACD39758 *H.subiculosus* ACD39767	60.0 58.0 58.0	None	Reducing PKS for hypothemycin	No
PKS9	14 (1285013-1290007)	*Pochonia chlamydosporia* ACD39770 *G. zeae* ABB90282 *H. subiculosus* ACD39762	24.4 22.7 22.9	None	Non-reducing PKS: Partial gene – only PP binding domain and thioesterase domain	No
PKS10	14 (1410771-1417375)	*Bipolaris oryzae* BAD22832 *C. heterostrophus* AAR90272 *Elsinoe fawcettii* ABU63483	91.8 89.4 75.5	SNOG_11981 89.2%	Melanin Elsinochrome	No
PKS11	21 (723403-733783)	*P. tritici-repentis* XP_001936632 *G. zeae* XP_385970 *Sclerotinia sclerotiorum* XP_001592682	74.1 65.7 45.8	SNOG_04868 31.9%	Type 1 PKS Mycocerosic acid	Yes
PKS12	0 (348658-356839)	*A. clavatus* (Ace1) XP_001270543 *P. chrysogenum* CAP74149 *Coccidiodides immitis* XP_001242733	37.3 37.2 37.2	SNOG_00308 57%	Ace1 homolog but lacks NRPS modules	Yes

Supplementary Table S11. Favoured codon usage for Small Secreted Proteins located within AT-rich genomic regions.

Amino acid	All predicted proteins[a]	SSPs in borders[b]	SSPs within AT-blocks[c]
*	TGA	TGA	TAA
A	GCC	GCC	GCT
C	TGC	TGC	TGC
D	GAC	GAC	GAT
E	GAG	GAG	GAA
F	TTC	TTC	TTC
G	GGC	GGC	GGC
H	CAC	CAT	CAT
I	ATC	ATC	ATT
K	AAG	AAG	AAA
L	CTC	CTC	CTT
M	ATG	ATG	ATG
N	AAC	AAC	AAC
P	CCC	CCT	CCT
Q	CAG	CAG	CAA
R	CGC	CGC	AGA
S	AGC	TCT	TCT
T	ACC	ACC	ACC
V	GTC	GTC	GTT
W	TGG	TGG	TGG
Y	TAC	TAC	TAT
Number ending with G or C	**20**	**17**	**7**

[a] Favoured usage in all predicted proteins of the genome
[b] Favoured usage in Small Secreted Protein (SSPs) located at the borders of AT-isochores
[c] Favoured usage in SSPs located within AT-isochores

Supplementary Table S12. Amino acid favoured usage (% per protein) in different sets of predicted proteins of *Leptosphaeria maculans* compared to proteins referenced in SWISSPROT database.

Amino acid	SWISSPROT	All predicted proteins[a]	SSPs in GC-blocks[b]	Non-SSP in AT-blocks[c]	SSPs in AT-blocks[d]	SSPs within AT- blocks[e]
Alanine	8,14	8,83	9,91	8,53	8,26	8,9
Cysteine	1,41	1,65	2,7	1,8	4,28	4,65
Aspartic acid	5,41	5,13	4,18	5,32	3,4	3,51
Glutamic acid	6,73	5,6	4,06	5,56	3,37	3,41
Phenylalanine	3,87	3,5	3,92	3,74	4,17	4,6
Glycine	7,04	6,97	7,87	6,63	6,28	6,03
Histidine	2,28	2,75	2,21	2,77	2,88	2,5
Isoleucine	5,92	4,65	4,6	4,88	4,83	4,86
Lysine	5,88	4,83	3,9	4,67	3,73	4,47
Leucine	9,67	8,59	9,67	8,84	10,76	10,61
Methionine	2,4	2,51	2,52	2,44	3,02	2,75
Asparagine	4,06	3,53	3,61	3,56	3,44	3,66
Proline	4,77	6,38	6,44	5,89	5,59	5,38
Glutamine	3,96	4,05	3,61	3,96	3,73	4,02
Arginine	5,5	6,48	4,68	6,3	5,42	4,7
Serine	6,66	8,21	8,44	8,22	8,24	7,35
Threonine	5,36	6,24	7,01	6,33	6,44	7,05
Valine	6,81	5,93	6,42	6,07	6,89	6,65
Tryptophane	1,09	1,49	1,63	1,68	1,97	1,54
Tyrosine	2,92	2,68	2,63	2,8	3,28	3,35

[a] Favoured usage in all predicted proteins of the genome
[b] Favoured usage in Small Secreted proteins (SSPs) located within GC-blocks
[c] Favoured usage in other proteins encoded by genes located within AT-blocks
[d] Favoured usage in SSPs whose genes are located in AT-blocks
[e] Favoured usage in SSPs located within AT-blocks (excluding those located at the border between AT- and GC-blocks)

Supplementary Table S13. Summary statistics for the shotgun sequencing of *Leptosphaeria maculans* isolate v23.1.3.

Average insert size	Vector type [a]	No. of reads	Average trimmed read length (bp)	Fraction paired (%)	Genomic coverage
3.3 kb	HC plasmid	200,566	721	97.7	3.20x
10 kb	LC plasmid	255,592	697	96.5	3.95x
79 kb	BAC	11,703	535	94.8	0.14x
42 kb	BAC	10,603	586	97.5	0.14x
39.8 kb	Fosmid	57,105	657	84.0	0.88x
Total		**535,649**	**696**	**95.6**	**8.31x**

[a] HC, high copy number; LC, low copy number; BAC, Bacterial Artificial Chromosome

Supplementary Table S14. cDNA libraries of *Leptosphaeria maculans* isolates generated and used here.

Isolate	Library name [a]	No. of ESTs	Description
IBCN18 (M1) n = 15627	BG	119	V8-grown mycelium (5' sequencing)
	DT	313	Mycelium submitted to starvation stress
	RS0AAB	8355	Mycelium grown in oilseed rape-juice for 5 days and 8 hours
	RS0AAC	6628	Mycelium in V8 for 5 days and minimal medium for 8 hours
	RS0AAD	212	Conidia germinating in V8 juice for 28h
PL86 n = 783	EB	760	Mycelium in V8 juice
	EC	23	*Brassica napus* leaf tissue eight days post infection (dpi)
v23.1.3 n = 25812	RS0AAA	5' 1644 3' 1224	Full-length-enriched mycelium grown for 48h in Fries liquid medium
	RS0AAE	3528	Conidia germinating in Fries liquid medium for 48 h (inserts < 1kb)
	RS0AAF	2934	Conidia germinating in Fries liquid medium for 48 h (inserts 1 to 2kb)
	RS0AAG	3644	Ungerminated conidia (V8 agar) (inserts up to 1kb)
	RS0AAH	3249	Ungerminated conidia (V8 agar) (inserts 1 to 2kb)
	RS0ABA	3686	Cotyledon-pathogen interaction 14 dpi (inserts 0.8 to 1.2kb)
	RS0ABB	3364	Cotyledon-pathogen interaction 14 dpi (insert 1.2 to 2kb)
	RS0ABC	2539	Cotyledon-pathogen interaction 14 dpi (insert larger than 2kb)

[a] BG, DT, EB, and EC series of sequences were retrieved from public databases; RS0A* sequences were sequenced by Genoscope (this study)

Supplementary Methods

Phylogenetic analyses and divergence time estimates

A taxon set containing representatives of most classes in *Ascomycota* was selected from the data matrices produced in two previous papers[42,43], both available from TreeBASE. Published sequence data were concatenated from the small and large subunits of the nuclear ribosomal RNA genes (SSU, LSU) and three protein coding genes, namely the translation elongation factor-1 alpha (*TEF1*) and the largest and second largest subunits of RNA polymerase II (*RPB1, RPB2*) (Supplementary Table S1). Where required (protein coding genes), DNA was translated into amino acids and single gene alignments were merged with MAFFT v6.713[60]. The combined exclusion sets of both alignments were used to exclude variable regions. The resulting matrix consisted of 40 taxa with 2967 characters of which 10% was coded as missing or undetermined. Full details on the GenBank accessions used in this analysis are provided in Supplementary Table S1 and this alignment was submitted to TreeBASE. Phylogenetic analyses were performed using RAxML v. 7.0.4[44,61] applying unique model parameters for each gene and amino acid. The dataset was divided in 5 partitions as previously described[43]. For the DNA sequences a general time reversible model was applied with a proportion of invariant sites and discrete gamma distribution for four rate classes following procedures laid out previously[43]. The three protein sequences were individually subjected to model testing with ProtTest v.1.2.6[62] under the same criteria resulting in a choice of models for the following partitions: RPB1 WAG+I+G and RPB2 and TEF1 RTREV+I+G+F. A combined bootstrap and maximum likelihood (ML) tree search was performed in RAxML with 500 pseudo replicates. The resulting trees were exported and manipulated (*e.g.*, collapsing of nontarget clades such as Eurotiomycetes) in TreeDyn[63].

In order to produce temporal estimates of the major phylogenetic divergences of the Dothideomycetes, the best scoring ML tree obtained from RAxML was analysed in the program r8sv1.7[45] (Figure 1b). Following dates suggested in more comprehensive analyses[64] the root of the tree was set to an age of 500 million years ago (MYA). Analyses were further constrained with minimum and maximum age constrictions based on dates presented in Lücking *et al*.[64] and Sung *et al*.[65], with the following ranges for a number of major nodes (crown *Pezizomycotina*, minimum age – 320 MYA, maximum age 400 MYA; base *Pezizomycotina* minimum age – 400 MYA, maximum age 520 MYA; base *Sordariomycetes* minimum age – 290 MYA, maximum age 380 MYA; base *Eurotiomycetes*

minimum age – 280 MYA, maximum age 330 MYA; crown *Hypocreales* minimum age – 155 MYA, maximum age 232 MYA). The Langley-Fitch method and a truncated Newton method with bound constraints were applied following Taylor and Berbee[66]. A second analysis was performed under the same conditions with the root fixed to 650 MYA following the range in Lücking *et al*.[64].

Sequencing and assembly

For whole-genome shotgun sequencing, DNA of isolate v23.1.3 was provided as plugs containing partly digested conidia[21] (Supplementary Figure S1b). Conidia were collected from highly sporulating 12-day-old cultures[67] in water containing 0.05% Tween 80 and gently centrifuged at 5000 *g* for 20 min. The pellet was rinsed by suspension in water and centrifuged gently again, followed by recovery of the pellet in Tris/sorbitol/EDTA (TSE) buffer. The conidia suspension was adjusted to $6x10^9$ conidia ml^{-1} and maintained at 37°C before being mixed with an equal volume of 2.5% low melting point agarose (SeaPlaque GTG, FMC) in TSE, also maintained at 37°C. The resultant plugs were then equilibrated in EDTA (0.5M, pH 8.0) for 1 h at room temperature, then incubated for 20 h with pronase E (5.5 units ml^{-1}; Sigma) in SDS/EDTA at 50°C. Plugs were rinsed four times in EDTA (0.5M, pH 8.0) at 50°C and then stored in EDTA (0.5M, pH 8.0) at 4°C until used.

In addition to genomic DNA, 1215 cDNA sequences available from public databases (Supplementary Table S14), cDNA libraries sequenced by Genoscope were made from mRNA from two isolates, v23.1.3 and an Australian isolate, IBCN18, grown under a range of conditions (Supplementary Table S14):

(i) Mycelium submitted to starvation. Isolate IBCN18 was grown in still culture in 10% Campbell's V8 juice for five days at 22°C then transferred to minimal medium[68] lacking carbon or nitrogen sources, and grown for an additional 8 h at 22°C.

(ii) Mycelium grown in plant-derived culture medium. Isolate IBCN18 was grown in still culture in *B. napus* leaf extract (soluble extract after boiling 200 g of chopped *B. napus* leaves in water; pH 6.0) for 128 h at 22°C.

(iii) Germinating conidia in complete medium. Isolate v23.1.3 was grown in Fries liquid medium for two days at 25° in flasks with shaking at 100 rpm.

(v) Ungerminated conidia of isolate v23.1.3 were frozen in liquid nitrogen.

(vi) Cotyledons of *B. napus* were infected with conidia of isolate v23.1.3 as previously described[67]. Infected plant tissue was harvested 14 days post inoculation.

For libraries i and ii, total RNA was extracted using TRIzol reagent (Invitrogen) and mRNA was purified using the PolyA Tract® mRNA Isolation System (Promega). cDNA was then prepared and cloned into pDONR 222 using the CloneMiner cDNA library construction kit (Invitrogen). For libraries iii-vi, total RNA was extracted using TRIzol reagent (Invitrogen) according to Fudal *et al.*[13] and cDNA libraries were constructed in plasmid pDNR-LIB using the Creator SMART cDNA Library Construction kit (Takara Bio Europe/Clontech, France). 42,222 ESTs sequences were obtained, resulting into 6190 unisequences mapped on the genome.

In case of genomic DNA, the sequencing reads were assembled using Arachne[47] resulting in an assembly of total size of 45.12 Mb organized into 76 SuperContigs (SCs) with a N50 of 1.77 Mb. There were 1743 contigs constituting these 76 SCs (Table 1). Of the 76 SCs, eight sized from 3.6 kb to 236.1 kb contained only unassembled rDNA repeats and 28 sized from 3.6 to 35.4 kb were only made of mosaics of repeated elements, with no predicted coding sequences (total size = 244 kb) (Supplementary Table S2). The mitochondrial genome was represented as a 154.7 kb single SC, and four small SCs containing RNA or DNA polymerases which corresponded to linear plasmids such as the previously identified pLm9 and pLm10[69] (Supplementary Table S2). Excluding these, the final nuclear genome thus comprised 30 SC > 143 kb. Eight additional small SCs (less than 23 kb) corresponded to unassembled putative telomere portions, and another eight SCs corresponded to unassembled rDNA repeats (Supplementary Table S2).

The assembly was validated and improved by joining SCs to make up whole chromosomes using a range of approaches:

(i) exploitation of sequence data to generate single-copy genetic markers (minisatellites and microsatellites) to co-linearize the available genetic maps to the actual genome sequence[70] (Figure 2d). Using Tandem Repeat Finder[71] and the “FONZIE” pipeline[72], 248 new microsatellites and 216 new polymorphic minisatellites were designed and mapped in reference genetic maps. A total of 234 micro- and minisatellites were mapped in the genetic map derived from a cross between isolates a.2 and H5[70]. Linkage between markers was analysed using Mapmaker/Exp 3.0 software[73] whereby data were analysed as a F_2 backcross segregating population. Markers with log of the likelihood ratio (LOD) values of > 3.0, and maximum distances of 20 cM apart were defined as linked. Similarly, using the web-based program, WebSat[74] 248 microsatellites were mapped from crosses between Australian isolates IBCN18 and 691, and isolates 535 and 691. Linkage between markers was analysed using MapManager QTX software whereby data were analysed as

a backcross segregating population. Markers with LOD scores of > 3.0 and maximum distance of 30 cM were defined as linked. Apart from joining previously unassembled SCs (*e.g.*, linkage of SCs 20, 21 and 23; Supplementary Table S2), this strategy also confirmed the assembly within SCs, and generally showed conservation of the order of markers between the physical map (*i.e.*, the SC sequence) and the genetic map (Figure 2d);

(ii) hybridisation of probes for single-copy genes or non-coding sequences specific to 24 individual SCs to chromosomal DNA separated by pulsed-field gel electrophoresis allowing size comparison of SCs and the hybridized chromosome. Chromosomal DNA was electrophoresed, transferred to nylon membranes and hybridised as previously described[75] (Supplementary Figure S1). To determine the organisation of ribosomal DNA (chromosomal location and arrangement of blocks of repeats), chromosomal DNA was digested with either *Sal*1 or *Xho*1 overnight, enzymes that cut outside the rDNA repeat, and electrophoresed as described in Howlett *et al.*[23]. Blots were probed with a fragment of 18S rDNA, or specific sequences from SC2 or SC19 that were PCR-amplified using primers, whose sequences are available upon request.

(iii) identification of telomere-specific repeats (see below),

(iv) mesosynteny with genomes of closely related dothideomycete species (Supplementary Figure S2). Since some SCs of *L. maculans* represent incompletely assembled sub-regions of chromosomes, comparison with large scaffolds from a genome of a related fungus was used to predict which scaffolds resided on the same chromosome. Matching regions between the genomes of *L. maculans* and *P. nodorum* were identified via MUMmer[76], using the promer algorithm. *L. maculans* SCs were defined as mesosyntenic (*i.e.*, sharing the same chromosomal lineage) with *P. nodorum* scaffolds according to the percent of their sequence covered by matching regions. A binomial test was employed to determine if the level of percent coverage for a given pair of SC was significant:

$$P_{Mesosyntenic} = F(x,p,n) = \sum_{i=0}^{x} \binom{n}{i} (p)^{i} (1-p)^{n-1}$$

[S1]

Where x = percent coverage,

n = 100,

p = linkage probability

The linkage probability (p) represented the chance that any pair of chromosomes from *L. maculans* and *P. nodorum* is mesosyntenic. This value was 1/(17x19) or 3.09 x 10^{-3}.

Chromosome numbers estimated in *L. maculans* (excluding the minichromosome; Table 3) and *P. nodorum*[77] were 17 and 19, respectively. Percent coverages were tested for SC pairs relative to the *L. maculans* SC and again relative to the *P. nodorum* SC. Pairs of SCs were predicted to be mesosyntenic when $P_{Mesosyntenic}$ was ≥ 0.99 for either test and both sequences were greater than 5 kb in length.

Using this suite of approaches, only two SCs, SC13 and SC17, remained unassembled and could not be assigned to chromosomes. In addition, the assignment of a few reassembled chromosomes was dubious (Supplementary Table S2). A maximum number of 18 chromosomes was discriminated (Supplementary Figure S1); ten of these corresponded to single SCs following the Arachne assembly, which can thus be considered as a high-standard assembly.

Leptosphaeria maculans genome annotation

The EuGene prediction pipeline v. 3.5a[48], which integrates several *in silico* (*ab initio* and similarity) data was used. *Ab initio* (intrinsic methods) gene finding software used were EuGene_IMM[78], which exploits probabilistic models for discriminating coding from non-coding sequences, SpliceMachine[79], for predicting Start and splicing sites, and the *ab initio* gene predictor Fgenesh 2.6 (www.softberry.com). Similarity methods such as BLASTn or GenomeThreader were used to compare genome data with ESTs of *L. maculans*, or of other fungi such as *Botrytis cinerea* and *Sclerotinia sclerotiorum*. BLASTx was performed against Uniprot and fungal protein databases. Information on location of transposable elements (TEs) was also provided to EuGene to avoid including TEs in coding exon prediction. The TE databank was provided by the REPET TE *de novo* pipeline (see below) and compared to predicted genes by RepeatMasker. All results were then gathered and integrated by EuGene to predict gene models.

The three *ab initio* gene finders were firstly trained using a set of manually annotated genes: FgeneSH was trained by Softberry Cie, EuGene-IMM and SpliceMachine were trained at URGI using a set of 290 pairs of *L. maculans* full-coding cDNAs and their genome counterparts. One third of the set was used to train *ab initio* methods, one third for optimization of weighting parameters and the last set was used to evaluate the accuracy of EuGene to generate adequate gene models. The final sensitivity of the integrator prediction was 93.4% and 83.5% (true positive / (true positive + false negative)) in the case of exons and genes, respectively. The specificity of the gene model

generated was 93.8% and 83.5% (true negative / (true negative / false positive)) for exons and genes, respectively.

The functional annotation pipeline was run using InterProScan[49] version 4.4. It includes a set of different methods for pattern matching, motif and domain recognition against well known databases (*i.e.* SuperFamily, PFAM, Panther, SMART, PIR, Tigr, Prosite, PRODOM, Gene3D); it also runs automated assignation of Interpro entries and associated GO terms. The pipeline operates with its own version of databases, which can slightly differ from public releases. Interproscan has been used with default parameters for each program (Superfamily: evalue=1e-5; HMMPfam: evalue=1000; HMMPanther: evalue=1e-3; HMMSmart: evalue=0.01; HMMTigr: evalue=20; fprintscan: evalue=0.0001; Gene3D: evalue=59.5). Targeting detection and cellular localization (SignalP, TMHMM) were also launched through InterProScan.

The genomic annotation system relies on the international open source project Generic Model Organism Database (GMOD; http://www.gmod.org/wiki) project: database (chado model), Gbrowse (visualization interface)[80] and Apollo (editing interface)[81] setup for the *L. maculans* genome project. The structural annotation database includes *L. maculans* genome sequences (SuperContigs and contigs) and features mapped on SC and contigs: gene predictions, proteins from Uniprot and local fungal protein databases, ESTs from *L. maculans* cDNA libraries (6190 mapped unisequences), predicted genes from *P. nodorum*, and repeats (TEs, SSRs, low-complexity regions). In addition, 94 transfer RNA genes were predicted by tRNAscan-SE[82] and are included in the Gbrowser.

EuGene has generated 12,469 gene models > 100 bases (Supplementary Table S3). The gene models were separated into two subsets: one main set comprising 11,561 genes with coding sequence > 300 bases, or < 300 bases but with at least one EST evidence or domain/motif evidence (as provided by the InterProScan analysis); and a second subset of 908 more dubious genes with a coding sequence smaller than 300 bases and without any EST or domain/motif support. Genes in the second category such as those encoding small-secreted proteins (SSPs), often were inadequately predicted by EuGene, and necessitated an additional round of manual annotation. The InterProScan pipeline was launched on the 12,469 genes predicted by EuGene and showed domains and/or motifs for 84.73% of genes; 52.75% of them have an InterProEntry with ontology associated (Supplementary Table S4).

Orthologs of *Neurospora crassa* genes postulated to be involved in gene silencing were sought in the genome of *L. maculans* by the Best Reciprocal Hits (BRH) method and

showed all but one of the orthologs involved in RIP, vegetative quelling (a post transcriptional gene silencing mechanism analogous to RNA interference) and Meiotic Silencing of unpaired DNA (MSUD) were present[83,84] (Supplementary Table S6). The only exception was the SAD-2 ortholog, essential for meiotic MSUD. The presence of these orthologs supports the idea that RIP and quelling are active in *L. maculans*. Quelling was previously described to be particularly efficient in *L. maculans* as a tool to silence endogenous genes[13]. The DIM-2 DNA methyltransferase[85], the RID DNA methyltransferase-related protein[86] and the HP1 homolog necessary for DNA methylation[87] are present in *L. maculans* and are likely to be functional, since their catalytic domains are conserved.

Repeat-induced point mutation analysis

Repeat-induced point mutation (RIP) is a fungal-specific genome defence mechanism that alters the sequences of repetitive DNA. Repeated DNA sequences align between mating and meiosis and both sequences undergo C:G to T:A transitions. In most fungi these transitions preferentially affect CpA di-nucleotides thus altering the frequency of certain di-nucleotides in the affected sequences, and generally resulting in the generation of STOP codons, thereby inactivating duplicated coding genes. An easy way to identify RIP is to compare the ratios of pre- and post-RIP di-nucleotides in putatively RIP-affected sequences using a series of RIP indices, initially identified in the model fungus *N. crassa*[88]. TpA/ApT is the simplest index and measures the frequency of TpA RIP products with correction for false positives due to AT-rich regions. Higher values of TpA/ApT indicate a higher degree of RIP. According to Galagan *et al.*[88], TpA/ApT > 2.0 indicates sequences that are strongly degenerated and inactivated following RIP. The index (CpA+TpG)/(ApC+GpT) is similar in principle to TpA/ApT but measures the depletion of the RIP targets CpA and TpG. In this case lower values of (CpA+TpG)/(ApC+GpT) are indicative of a higher degree of RIP[88]. Such RIP indices were used to evaluate the RIP effect on genes or genome regions for which multiple alignments could not be generated. For example, in the case of RIP gradients at the borders between AT-rich and GC-equilibrated blocks (see below), the borders were extracted manually following masking of TEs and a Python script which calculates %GC and length along with the two above-mentioned RIP indices in sliding windows was exploited (Supplementary Figure S4). The results were then displayed as graphics created by an R script. Otherwise, automated analysis of RIP in *L.*

maculans genomic DNA repeats was performed using RIPCAL (http://www.sourceforge.net/projects/ripcal), a software tool that performs both RIP index and alignment-based analyses[19] (Figure 3; Supplementary Figures S5, S10).

Annotation and analysis of repeated elements

Transposable Elements (TEs) consensuses were predicted *ab initio* in three consecutive steps: first, repeats were searched with BLASTER for an all-by-all BLASTn[89] genome comparison. Second, the results were clustered by the three methods GROUPER[90], RECON[91], and PILER[92] with default parameters. Third, one consensus per group was built with the MAP multiple sequence alignment program[93] and each consensus classified with (i) BLASTER matches, using TBLASTx and BLASTx[89] against the entire Repbase Update databank[50] and (ii) structural features such as terminal repeats (TIR, LTR, and polyA or SSR tails). For example, a consensus is classified as MITE if: (i) it carries TIRs; (ii) doesn't match via TBLASTx or BLASTx with known TEs; (iii) and the length without TIRs is less than 500 bp. Then the set of consensus TE sequences was analyzed by an all-by-all BLASTER procedure (parameters set at 95% identity threshold and 98% length threshold) to remove redundancies, *i.e.*, when a consensus sequence is included into another.

A total of 1850 consensus sequences representing ancestral copies of TE families was obtained. Of these, all SSR (single-sequence repeats) and unclassified TEs containing fewer than 10 sequences were removed, ending up with 472 consensus families. These sequences were clustered into groups for identification of TE families using the GROUPER clustering method, resulting in 15 groups. Each family was identified assuming that the most populated, well characterized TE category in a group of consensus sequences can define the order of the TE family to whom it belongs. Forty-three families containing at least two TE consensus sequences were then manually curated using multiple sequence alignments and phylogenies. This close examination allowed confirmation of groupings and deciphering of specific features like chimeric TE families or sub-families. A final consensus was derived from each curated multiple alignment. Seventy four TE consensuses representing TE families showing less than 90% identity at the nucleotide level with other families were generated. The set of sequences from the original “confused” category was then clustered with the 74 TE families to eliminate “confused” sequences sharing more than 90% identity with one of the 74 families. A total of 199 “confused” TE families could not be removed in this way; thus a set of 134 TE

consensuses had matches on the genome that were the input of the REPET annotation pipeline part. This pipeline is composed of the TE detection software BLASTER[90], RepeatMasker[94] and Censor[95] and the satellite detection softwares RepeatMasker, TRF[71] and Mreps[96]. The genome was segmented into 200 kb fragments overlapping by 10 kb, which were independently analyzed by the different programs. Simple repeats were used to filter out spurious hits. TE or repeat copies fewer than 20 bp after removing simple repeat regions were discarded.

Since TEs often insert into other TEs causing fragmentation, a “long join” annotation procedure was performed, using age estimates of repeat fragments based on number of mutations not attributable to RIP to correctly identify fragments from the same repeat. The percent identity between a fragment and its reference TE/repeat consensus can be used to estimate the age of TE fragments. Consecutive fragments on both the genome and the same reference repeat consensus were automatically joined (i) if the difference between their percent identities was less than 2% (the two fragments had approximately the same age) and (ii) if they were separated by a gap of less than 5000 bp and/or by a mismatch region of less than 500 nucleotides, or (iii) if there were nested repeats: the fragments were separated by a sequence of which more than 95% consisted of other repeat insertions, all inserts having a higher identity compared to their respective consensus. Fragments separated by more than 100 kb were not joined. Finally nested repeats were split if inner repeat fragments were longer than outer joined fragments.

The set of 134 TE consensuses was inspected visually for redundancy, taking into account the high level of nucleotide divergence introduced by RIP mutation and, whenever possible, “type” TE reconstructed by identification of bordering LTRs (including the canonical TGT-ATA/ACA borders of LTRs) and TIRs.

All these steps led to the identification of only 21 TE families in the genome encompassing 9 LTR retrotransposons, 10 TIR DNA transposons (Table 5), one rDNA-specific LINE (Supplementary Figure S10) and one *Penelope*-like non-LTR retrotransposon, RPP_*Circe* (SuperFamily: *Athena*; Supplementary Figure S8)[22]. In addition, 11 “confused” and uncategorized repeats could not be related to any of the type TEs and only covered 160 kb. As is often the case in fungi, LTR retrotransposons largely predominated among TEs, representing 27.3% of the genome. However, this large prevalence was mostly due to only four TEs, three *Gypsy* retroelements, RLG_*Olly*, RLG_*Polly* and RLG_*Rolly*, each covering 2.24-3.06 Mb of the genome, and one *Copia* retrotransposon, RLC_*Pholy*, covering 3.10 Mb of the genome (Table 5). All the nine LTR

elements, along with the DNA transposons DTM_*Sahana* and DTx_*Gimli,* were subdivided into numerous subfamilies by REPET, corresponding to sequence variants of the type element, truncated copies, mosaics composed of only one TE or, more rarely, mosaics comprising parts of more than one element. Some of the components of these mosaics of TEs, such as one of the most common of the TE in the genome, RLG_*Olly*, had a predicted chromodomain[97], which can be indicative of a genome invasion specifically targeted at TEs or TE-rich regions of the genome behaving as heterochromatin.

All TEs, as well as telomeric repeats and part of the rDNA, were severely mutated by RIP with nucleotidic identity between members of a super-family as low as 60% (Figure 3, Supplementary Figure S5). This contributed to the artificial splitting of type families into sub-families showing intense nucleotidic polymorphism due to RIP. RIP degeneracy regarded not only mosaics of TEs in large AT-blocks but also AT-blocks corresponding to single TEs and even small-sized elements (*e.g.*, DTx_*Gimli*) and remnants of TEs occurring within GC-blocks. Very few TEs showed evidence of being still transcribed into mRNA and thus being still potentially active transposable elements: of 42,222 EST sequences, 19 had significant hits with only 6 families of repeated elements, including the uncategorized telomeric repeats LmTelo1 and LmTelo2, encompassing helicase domains. However, none of these showed evidence of expression in mycelium or during plant infection.

As a final step, type repeats were deRIPped using the RIPCAL pipeline (see above) to facilitate the annotation and categorization of TEs. For “deRIP” of TE families, an updated version of RIPCAL[19] was used, including the Perl script “deripcal” (which predicts ancient pre-RIP sequences) and ripcal_summarise (which provides summary RIP statistics over a whole alignment). DeRIP was performed as follows: repeat families were aligned with ClustalW2[98] (window=50, ktuple=6, quicktree) and analysed for RIP with RIPCAL. The alignments were then compared to three model sequences: (i) highest total G+C count (RIP depletes G and C nucleotides, therefore highest G + C count should indicate the least RIP-mutated sequence). However, G+C count models may not represent the least RIP-affected sequence if there is a particularly long sequence among several short ones, and may not cover the total alignment length, rendering the deRIP uneven along the whole of the sequence; (ii) alignment consensus (most common base at alignment position, using degenerate nucleotide code for ambiguous base counts). Consensus models will not detect RIP mutation if RIP has saturated the alignment, i. e., all CpNs have been converted to TpNs at a particular alignment position. However consensus models will

usually span the total alignment length, except where there is only 1 sequence; (iii) deRIP consensus (similar to consensus, however where CpN-->TpN mutations are detected within the alignment the consensus sequence is automatically set to CpN; this includes the reverse complement of CpN->TpN mutations) using “deripcal”. This is similar to the consensus method, however, where more than one type of CpN-->TpN mutation occurred at a particular alignment position, the most frequent type was used to generate the deRIP consensus. As with consensus models, this model also relies on RIP not having saturated the alignment.

The TEannot consensus families and the deRIP consensus sequences generated were compared via BLAST to the NCBI NR protein database. However, since deRIP consensus sequences may contain some degenerate code, these could not be processed by BLASTn/BLASTx, thus were translated in six frames via virtual ribosome (reading through stops) and then compared via BLASTp. Both deRIP consensus and TEannot consensus sequences were compared to the REPBASE database (version 13.12) via both BLASTn and tBLASTx. The 'strength' of BLAST hits to NR/REPBASE was compared by cross-referencing both datasets for both repeat class and hit name. All of these processes generated new, rebuilt consensus TEs for each family, usually these better matched functional protein sequences than the original consensus sequences did, and facilitated assignment to referenced families whenever RIP degeneracy was too intense for their identification. This was the case in particular for a LINE-1-like insertion in some copies of the rDNA units (Supplementary Figure S10).

Dynamics and demography of TEs in the genome

Analysis of dynamics of genome invasion by TEs was firstly based on phylogenetic analysis of each family of repeats, retracing the evolutionary history of each family regardless of truncation, insertion in other TEs and deletion events[51]. In such an approach, terminal fork branch length of one copy corresponds to an evolutionary distance used to estimate the age of the last transposition activity, assuming that each transposition event will generate copies independently submitted to mutation. The terminal branch distance corresponds to the number of substitutions per mutable positions and, assuming a suitable molecular clock is available, the tree of TE families can be used to date the dynamics and demography of TE invasion in the genome. The sequences were firstly aligned using REFALIGN from the REPET package, by doing a master-slave multiple alignment, using

the consensus sequence as master to align TE copies. However, this approach was complicated by the high rate of mutation due to RIP, which was likely to lead to major biases in time estimates. To address this point, a Perl script was developed and used to eliminate all possible RIPed dinucleotides in the alignments by replacing TA by NN. The resulting alignments were then used as an input in the PhyML program[99], allowing phylogenies to be built from DNA sequences using maximum likelihood. A BIONJ distance-based tree is used as starting tree[100]. Topology, branch lengths, and rate parameters were optimised. The transition/transversion ratio default value was 4.0. The gamma distribution parameter is estimated by maximizing the likelihood of the phylogeny. Data sets were analyzed under the HKY85 evolution model[101]. Terminal forks branch lengths were extracted from these Newick files to calculate the age of the last transposition events of the copies in the genome sequences using a script called ITermForks[51]. This was done both with trees obtained from the complete alignment containing NN columns and in trees from alignments in which all columns containing Ns were deleted. However, in this latter case, only those alignments in which a minimum size of 200 resolved aligned nucleotides were analysed. Seventy or more sequences could be aligned for the following TEs: RLG_*Dolly* (148 sequences), DTx_*Gimli* (70 sequences), RLG_*Polly* (422 sequences), DTF_*Elwe* (70 sequences), RLG_*Olly* (488 sequences), RLC_*Pholy* (462 sequences), RLG_*Rolly* (288 sequences), RLx_*Ayoly* (76 sequences) and DTM_*Sahana* (106 sequences). For final graphical visualisation an R statistical package script with a kernel_density procedure was used. An R script was written to plot a histogram of the terminal fork branch length with kernel density estimate for each family (Figure 4b). Lastly the divergence values were converted in estimated divergence time using the 1.05×10^{-9} nucleotide per site and per year substitution rate currently estimated for protein-encoding genes in fungi[53,102]. The distribution of the values was displayed as a boxplot graph (Figure 4b). Using these approaches very low divergence rates for *L. maculans* TEs, typically ranging between an average of 0.007 and 0.017 substitution per site per year were found (Figure 4).

Dynamics of TE aggregation over time was also analysed by a visual and graphical analysis of nesting relationships between TEs. Following the long join annotation, mosaics of TEs were visualized using Artemis v. 12.0 (http://www.sanger.ac.uk/Software/Artemis/) in SC0-22. To avoid erroneous interpretations due to errors in assembly, TEs were considered nested if (i) an uninterrupted mosaic was present (*i.e.*, those mosaics in which gaps in the sequences are present were excluded from the analysis), (ii) a continuous

element could be rebuilt once invaders were omitted (*i.e.*, the integration within another TE was not accompanied by deletions or sequence inversion of the invaded TE). Also (iii) portion of TEs (*e.g.*, solo-LTRs) that could not be joined to form a complete element were not classified as invaders or invaded. Five hundred and seventeen mosaics were identified and a data matrix recording the number of integrations of one given TE family into another one (invader) and the number of cases where one given TE was recipient of an insertion from one or multiple other TEs (invaded TE) was generated (Supplementary Table S8). To evaluate whether the observed values differ from those expected in case of random integration, data were submitted to a Chi-squared test for given probabilities with simulated *P* values (based on 20,000 replicates) as implemented in R.

This analysis indicated that a major burst of transposition involving all major families of retrotransposons occurred simultaneously, and that only few waves of transposition occurred for these, as shown by the divergence time estimates (Figure 4), the significant under-representation of self-nests (TEs invading a TE from the same family) for RLC_*Pholy*, RLG_*Polly*, RLG_*Rolly* and RLx_*Jolly* (Supplementary Table S8) and the over representation of primary nesting relationships compared to more complex secondary or tertiary invasions. The scarcity of invaded DTM_*Sahana*, its low sequence divergence and its common occurrence as a single TE in the middle of GC-blocks indicates that DTM_*Sahana* expanded more recently than other TEs. Interestingly, these analyses show a large disequilibrium in source/sink relationships between retrotransposons and DNA-transposons. Whereas all retrotransposons are invaders or have been invaded to a similar extent, and DNA transposons such as DTF_*Elwe*, DTM_*Sahana*, DTx_*Gimli*, DTM_*Ingwe* are mostly invaded and rarely invaders (Supplementary Table S8).

Identification of telomeres

In most organisms, telomeres consist of short repeated motifs, and the sequences adjacent to the telomere repeats are often duplicated at multiple chromosome ends, thus defining a specific subtelomeric region[30]. The mining of the genome of *L. maculans* for annotated telomere-linked helicase, $TTAGGG_n$ telomeric repeats and the examination of extremities of SCs for the occurrence of TEs identified by REPET showed that telomeres are characterized by the alternation of three repeated elements specific to chromosome extremities; (i) RPP_*Circe*, a non-LTR retrotransposon, separated by $TTAGGG_n$ and other SSRs, (ii) *LmTelo1* and (ii) *LmTelo2*. Even though it was degenerated by RIP, RPP_*Circe*

had a match to a reverse transcriptase, but not of the GIY-YIG endonuclease typical of *Penelope* elements. It thus was extremely similar to the *Athena* telomere-associated retroelements described in a series of Eukaryotes, including Basidiomycetes, but not Ascomycetes[22]. As described for other *Athena* elements, the 5' part of RPP_*Circe* was telomere-oriented and capped with typical TTAGGG repeats, whereas its 3' ends contained more complex tandem repeats of which some were modified TTAGGG repeats (Supplementary Figure S8). *LmTelo1* and *LmTelo2* contained a predicted RecQ helicase (Supplementary Figure S8 for an analysis of type telomeres in *L. maculans*). The distal portion of telomeres is highly prone to truncation and probably poorly assembled and, whereas *LmTelo1* and/or *LmTelo2* are always present in the telomeric AT-rich region, the presence of the most distal element, RPP_*Circe* was much more variable. These three repeated elements were the core part of telomeres but they also occurred as truncated copies intermingled with various other TEs (Supplementary Figure S8).

Chromosomal location and organization of ribosomal DNA

We firstly reconstructed the complete sequence of the rDNA unit by identifying the SC containing rDNA repeats. Of these, the extremity of SC19 showed perfect matches (99.6% identity or more) with available Small Subunit (SSU), internal transcribed spacer 1 (ITS1), 5.8S, ITS2 and partial Large Subunit (LSU) sequences of *L. maculans*. The genomic region separating two repeats was postulated to be the missing LSU (as validated by BLAST against the nt database) and Intergenic Spacer (IGS) sequences. The final canonical rDNA copy was 7.8 kb in length with 45% GC content and encompassed the following: the SSU, sized 1798 bp, ITS1 sized 111 bp, the 5.8S sized 211 bp, ITS2 sized 148 bp, the LSU sized 3327 bp and the IGS sized approximately 2 kb (Figure 3). However, a standard sequence for the IGS was difficult to define due to the presence of variable-length short tandem repeats flanking almost every rDNA repeat, missing parts in most of the assembly and probable size/sequence variations from one repeat to the other. Using REPET, the rDNA repeats were identified mostly as "confused" TEs corresponding to fragments of the rDNA unit. Both the canonical rDNA copy and the REPET outputs were used to mine the genome and indicated that nine other SCs contained parts of the rDNA. One hundred and seven complete and incomplete copies of 18S rDNA were present in the assembly which is within the range of the 56-225 copy number estimated previously[23] but may be an underestimate, due to loss of perfectly matching copies in the course of the

assembly. In addition, 50 complete copies of the transcription unit were identified in the whole assembly, with up to 16 tandem copies assembled in SCs only consisting of rDNA tandem repeats. In contrast to those in SC19, the matches showed between 85% and 88% sequence identity with the canonical unit.

Southern hybridization of CHEF chromosomal blots with 18S rDNA and ORFs from SC2 and SC19 showed binding to a large unresolved chromosomal DNA band (>3.43 Mb). This suggested that rDNA is on a single or on two large chromosomes. To determine whether rDNA was in a single tandem array, chromosomal DNA was digested with either *Sal*I or *Xho*I, enzymes that cut outside the rDNA repeat unit. Two large bands 1.05 and 0.61 Mb hybridized to an rDNA probe. The 1.05-Mb band hybridized to a single-copy sequence between the *Xho*1 site and the first rDNA unit on SC19, whilst the 0.61 band hybridized to a single copy sequence between the *Xho*1 site and the first rDNA unit on SC2. This shows that rDNA is arranged in two blocks of tandem repeats. Enzymes that cut once within the rDNA repeat unit, *Bam*HI and *Hin*dIII, produced several bands, thus confirming size variation of the IGS as shown in Supplementary Figure S10. Sequence analysis of BAC ends showed that 1.3% of the clones contained only rDNA repeats and 0.3% (*ca.* 3-x genome coverage) contained rDNA on one extremity, whereas the other consisted of diverse TEs. Interestingly, one third of the clones containing only rDNA showed sequence polymorphism indicative of RIP whereas a few BAC clones (corresponding to a 1.5-x genome coverage) contained rDNA copies showing RIP degeneracy on one extremity but not the other. All of these data strongly suggest that the rDNA in v23.1.3 is organized as two arrays of tandem repeats located on either one or two large chromosomes. These arrays are flanked by mosaics of TEs, and one part of the array is protected from RIP whilst another part is affected by RIP mutations. Finally, some copies showed an additional level of degeneracy with the insertion of a 7-kb RIPped LINE element in the 3' part of the 28S rDNA gene (Supplementary Figure S10). This feature, reminiscent of what has been observed in some rDNA units of arthropods[103] has to our knowledge never been described in fungal rDNA.

Analysis of AT-rich genomic regions

A sharp contrast between GC-equilibrated regions (henceforward GC-blocks) and AT-rich regions (henceforward AT-blocks) was evident using genome viewer tools such as Artemis (Figure 2c; Supplementary Figure S12). Sharp transitions between the blocks resulted in

small-sized transition regions. AT-rich and GC-equilibrated regions were manually extracted from each SC using Artemis. Two sets of location data were generated for each AT-block: one exclusively encompassing the AT-block stopping at the GC-content rise, and the second including the transitional regions between the AT-blocks and the GC-blocks (hereafter called “borders”).

The sequences of the AT-blocks were then extracted using a Python script to generate a file containing (i) the name/number of the AT-block, (ii) the SC number in which it had been identified, (iii) the start and (iv) end positions on the SC. AT-blocks could be discriminated into three categories (i) telomeres, which are up to 93 kb (Supplementary Figure S8); (ii) large AT-blocks (216 sized 13-325 kb) (Supplementary Figure S9), usually composed of mosaics of more or less truncated TEs. Previous sequencing and finishing of a large genomic region showed larger AT-blocks (up to 450 kb)[13], thus suggesting that very large AT-blocks may be poorly assembled, resulting in many unassembled smaller AT-rich SCs comprising only one AT-block (Supplementary Table S2); (iii) mid-sized AT-blocks (197 sized 1-13 kb) (Supplementary Figure S9), corresponding mostly to a single TE family member. Single-TE AT-blocks were mostly due to only two DNA transposons, DTM_*Sahana* (62.4% of the repeats occurring as single TE within GC-blocks) and DTF_*Elwe* (19.3% of the repeats occurring as a single TE within GC-blocks). TE content of AT- and GC-blocks was analyzed using the REPET pipeline (see above). Size distribution of AT-blocks, occurrence of AT-blocks on chromosomes and relationship between AT-block, TE content and chromosome length were calculated (Supplementary Figure S3). The number of AT-blocks was linearly correlated to the size of the SC ($R^2 = 0.87$), as was the total size of the AT- blocks within a SC ($R^2 = 0.77$) (Supplementary Figure S3). In all cases, the transition between AT- and GC-blocks was extremely abrupt (859 ± 385 bp) (Supplementary Table S7).

Analysis of the distribution along the SC/chromosomes indicated clustering of 99.8% of the annotated repeat families within AT-blocks. TEs identified in GC-blocks only were small-sized portions of TEs (287 bp on average) or the small-sized DTx_*Gimli.* The total of 461 small TEs within GC-blocks comprised only 131.8 kb (0.2% of the genome). These portions of TEs showed a GC bias similar to that of larger TEs lying within AT-blocks (average GC% = 40%).

Micro and minisatellites were systematically searched for along selected SCs (SC0, SC1, SC4, SC7, SCs14-16), mostly corresponding to complete chromosomes (Supplementary Figure S1). These markers were exclusively found in GC-blocks and their

position along each SC recorded. Ninety-four micro- and minisatellite polymorphic markers between two parental isolates (a.2 and H5) of a cross described by Kuhn *et al*.[70] were mapped in the random progeny. For each progeny the number of cross-over (CO) between two consecutive markers was calculated along each SC. The recombination frequency between two successive markers (number of progeny with one recombination event between the two markers/nb of isolates in the progeny) was calculated and plotted against the physical distance between the two markers.

Two sets of data were compared: a set encompassing recombination frequencies between successive markers located within the same GC-block, and a set comprising the frequencies between two markers located on both sides of a single AT-block. When more than one AT-block or GC-block separated two successive markers, the recombination frequencies between them were excluded from the statistical analysis. Forty-one pairs of markers belonged to the same GC-block, and 21 pairs corresponded to markers spanning one single AT-block. These datasets were subjected to an ANOVA and a non-parametric test (Mann-Whitney test) using XLStat, to compare recombination frequencies between and within GC-blocks. The recombination frequency differed significantly between these two groups of marker pairs (Supplementary Figure S7a), with a higher recombination frequency within GC-blocks (F Fisher = 5.873, P=0.19). Within GC-blocks, the recombination frequency correlated significantly with the distance between the markers (Pearson's correlation coefficient r=0.721, P<0.001), whereas the correlation was lower when the markers were located on both sides of an AT-block (r =0.574, P=0.007) (Supplementary Figure S7b).

The databank of TEs provided by the REPET pipeline (see above) was used to mask the TE content of AT-blocks using RepeatMasker, and thus specifically identify non-repeated genomic regions within AT-blocks. Gene prediction in masked AT-blocks firstly relied on FgeneSH and final EuGene prediction. However, these predictions were sometimes inaccurate, and in a number of cases, small putative ORFs were not predicted. For example, in the final set of 122 Small Secreted Protein (SSP)-encoding genes, 63 were not predicted by either EuGene, or FgeneSH alone. For a better identification of putative SSP-encoding genes, the EMBOSS:GETORF program was used with a size limit set at 600 amino acids (lower limit : 60 amino acids). In a second step, outputs of each predictor were compared to avoid redundancy. If a GETORF prediction was included or partly redundant with a EuGene/FgeneSH prediction, the former was discarded.

The Gene Ontology (GO) project standardizes representation of gene and gene

product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data (http://www.geneontology.org/). The software Blast2GO[104] was chosen to annotate all predicted *L. maculans* ORFs with GO. The first step of the process was a BLAST against a chosen databank to get similarity with other known sequences. BLAST v. 2.2.21 has been performed on URGI cluster Sauron with all *L. maculans* predicted proteins (12,469) and the 498 non-SSP proteins present in AT-blocks (including border sequences) against "nr" version 16th october 2009. SSPs occurring in AT-blocks were excluded from the analysis due to the low number of BLAST hits and association to GO terms. The second step was the mapping by querying a database with Gene Ontology, geneinfo, gene2accesion and PIR data. GO terms associated with a particular hit were transferred to the sequence. Blast2GO mapped 12,455 resources for *L. maculans* proteins. When blast2GO performs mapping, it keeps the source of annotations and their associated Evidence Code (EC). Blast2GO implements an Annotation Rule (AR), which takes into account the source of the GO (Electronic or Manual Inference) for each GO mapped and modulates it with Blast Hit e-values. Annotation was firstly performed with BLAST, using default parameters provided by Blast2GO (Blast Threshold evalue=1.e-6, annotation.goweigth=5, annotation.cutoff=55 and default weights for each evidence code (IDA=1, IPI=1, IMP=1, IGI=1, IEP=1, EXP=1, ISS=0.8, ISO=0.8, ISA=0.8, ISM=0.8, IGC=0.7, RCA=0.9, TAS=0.9, NAS=0.8, IC=0.9, ND=0.8, IEA=0.7, NR=0)). Then annotation was expanded with “Annex”, which increased by 13.5% the number of annotations so that the mean of GO-level gains 0.05, and lastly with InterProScan v. 4.4 (runs performed on Sauron cluster). Interpro entries are often linked with one or several GO terms. According to Gotz *et al.*[104], InterProScan results significantly increase annotations, especially in a *de novo* analysis. Here, this third annotation increased the number of annotations by 33.7% as compared to the BLAST + Annex annotation. In total, 5528 proteins (44.3% of the sequences) of the whole-genome set and 177 (35.5% of the sequences) of the AT-block set were annotated and compared for GO term occurrence (Supplementary Figure S11).

Identification and characterization of clusters of genes encoding secondary metabolites

The complete sequence, as well as adenylation domains of non-ribosomal peptide

synthetase (NPS) genes from other filamentous fungi were blasted against the *L. maculans* genome and homologs were identified. Regions of the genome (minimum of 50 kb) with a NPS homolog were analysed by FgeneSH (Softberry.com) and neighbouring genes were identified. Also entire SCs were analyzed using FgeneSH/Softberry. NPS and polyketide synthases (PKS) genes also were identified using domain searches in NCBI and the PKS/NRPS Analysis website (http://www.tigr.org/jravel/nrps/). Using these approaches 13 NPS genes were identified, including the previously identified ones *SirP*, involved in sirodesmin biosynthesis[105], *maa1*, and *NPS9*, a gene with high sequence similarity to *Lys2* of *Saccharomyces cerevisiae*[106] (Supplementary Table S9). Twelve PKS genes were identified, including *PKS1*, which is located close to the sirodesmin gene cluster and *PKS12*, the ortholog of *Aspergillus clavatus Ace1* (Supplementary Table S10). This latter PKS, unlike that in *Magnaporthe grisea* and *P. nodorum*, lacks a NPS module. Seven of the 12 PKSs had close matches to PKSs in *P. nodorum*. Only five of the NPS genes had multiple modules while the remaining eight were monomodular. Four of the monomodular proteins had a condensation domain as the last predicted domain, while reductase and thioesterase domains were predicted for two each of the other NPS genes. Domains were also analysed in the PKS genes.

NCBI BLAST searches identified orthologs of the NRPS and PKS genes. The percentage of similarity between ortholog pairs was determined via NEEDLE[107] (Supplementary Table S9). These results were then used to suggest putative functions for the individual *L. maculans* genes (Supplementary Tables S9, S10). *NPS2* and *NPS6* had close matches to siderophore genes[108-110]. Based on the expected number of domains and predicted matches of adenylation domains, NPS8 may be involved in biosynthesis of the depsipeptide, phomalide, postulated to be a host-selective toxin[111]. Functions for the other NPS genes could not be proposed.

Sequence (~50 kb) surrounding each NPS gene was analyzed for gene content using FgeneSH. The function of predicted proteins was determined by blasting them against the NCBI database. Classes of proteins with roles in the biosynthesis of secondary metabolites (cytochrome P450 monooxygenases, methyltransferases, prenyl transferases, transmembrane transporters and transcription factors) were sought. On this basis, eight of the 13 NPS genes were predicted to be part of gene clusters (Supplementary Table S9), including *SirP*, part of the sirodesmin biosynthetic cluster. The best matches of genes flanking *NPS2* and *NPS6* supported the hypothesis that these were part of siderophore gene clusters[108-110]. Similarly, best matches of genes flanking *NPS8* (an

aldoketoreductase), suggested its product may be phomalide. *NPS3* and *NPS7* were in clusters, but their products cannot be predicted. The NRPS4 gene had ATP Binding Cassette (ABC) transporter and Multi Facilitator Superfamily (MFS) transporter genes nearby but best matches of the remaining genes have no predicted roles in secondary metabolism. Likewise, *NPS1* has an MFS transporter gene immediately upstream; however, surrounding genes have no known roles in secondary metabolism. While *NPS5* is surrounded by genes of unknown function, about 15 kb downstream from it is a predicted PKS gene (*PKS2*) whose end product cannot be predicted. The NPS9 and maa1 genes are surrounded by genes, but not of the classes described above.

Of the 13 *L. maculans* NPSs, ten are located in or close to AT-blocks, whilst three (including *lys2* involved in lysine biosynthesis) are in GC-blocks (Figure 2b; Supplementary Tables S9, S10; Supplementary Figure S12). Some of these NPSs in AT-rich regions are sub-telomeric (NPS7, NPS10, NPS11), as are secondary metabolite gene clusters in other ascomycetes[112,113]. In contrast, the nine NPSs of *P. nodorum* are not located in AT-rich regions; for instance, *NPS7* is in an AT-repetitive rich, sub-telomeric region in *L. maculans*, but is not associated with any repetitive elements in *P. nodorum.* An exception is *NPS5*, which is flanked by repetitive elements in both species. All *L. maculans* NPSs had either EST support or were transcribed in complete media, as shown by RT-PCR. Monomodular NRPSs including *lys2* have more ancient origins and more conserved domain architectures than most multimodular NPSs. They also are predicted to play more pivotal roles in cellular metabolism than products of multimodular NPSs. In contrast, multimodular subfamilies of NPSs are of more recent origin, are restricted to fungi, are more variable, and biosynthesise metabolites that perform more niche-specific functions than monomodular NPS products[114]. Whereas the location of the monomodular *lys2* or *maa1* within a GC-block is consistent with their being protected from hypermutation due to RIP, this is not the case for the conserved monomodular NPS7 (Supplementary Table S9) and poses questions about the influence of genome environment on evolution of NPS genes in *L. maculans.*

L. maculans has fewer *PKSs* (12) than *P. nodorum* (19) or *C. heterostrophus* (24)[55]. Several *PKS* are organized into gene clusters (Supplementary Table S10), but few polyketide-derived molecules have been described in *L. maculans* so little is known about their products. Like the *NPS*, several of the *PKS* are located in or close to AT-rich regions, some of which are sub-telomeric (e. g., *PKS4*, *PKS6*, *PKS7*, *PKS10*).

Synteny between *L. maculans* and related *Dothideomycetes* in the AT-block surroundings

Predicted proteins of *L. maculans* were reciprocally blasted against the predicted proteins of *P. nodorum*. Best reciprocal hits (BRH) genes flanking an AT-block (one in 5', one in 3') in *L. maculans* were extracted and the two closest BRH in *P. nodorum* were identified. The distance between each pair of genes and the ratio of the lengths were calculated in both organisms (Figure 1b). Two hundred and fifteen pairs of genes separated by an AT-block were identified in *L. maculans* with an orthologous pair of genes in *P. nodorum*. In 29% of the cases, the pair of genes of *P. nodorum* was located on different SuperContigs and could not be analysed. In 48% of the remaining cases, the intergenic distance between the pair of orthologs in *P. nodorum* was less than 10 kb (average 4 kb) and contained 0 to 3 genes (average = 0.9). The corresponding pair of genes in *L. maculans* was separated by AT-blocks sized 4.4-213 kb (average 30 kb) and contained 0-6 genes (average 2) (Figure 1b). In addition, in 31 cases, consecutive genes in *P. nodorum* were separated by an AT-block in *L. maculans* with an average enrichment of 2 genes in *L. maculans* compared to *P. nodorum*. In contrast, the 80 cases where the pair of *P. nodorum* orthologs was separated by more than 10 kb usually represented large breaks in synteny, with an average of 189 genes present in the interval. To validate the lack of large TE-rich regions in related species, the pair of orthologs was then investigated in two other species of the suborder *Pleosporineae*, *C. heterostrophus* and *P. tritici-repentis*.

Identification and features of Small Secreted Proteins (SSPs)

To define whether a protein can be considered as a putative SSP, a pipeline using different prediction programs was written in Python. The predicted protein was firstly screened according to its size with a length limit set at 330 amino acids (lower limit: 30 amino acids). Each putative protein was then submitted to a signal peptide prediction program, SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/; set for eukaryotes), a target location prediction program, TargetP (using non-plant networks, including cleavage site prediction), and a transmembrane domain prediction program, TMHMM. A predicted protein was considered as a putative SSP if both methods of SignalP (Neural Networks and Hidden Markov Model) were consistent, if TargetTP located it in the secretory pathway, and if TMHMM predicted 0 or 1 transmembrane domains. In this latter case, the transmembrane domain had to be included in the signal peptide.

For structural comparison purposes, the following sets of genes encoding SSPs were constructed: (i) SSPs in GC-blocks containing 529 genes; (ii) SSPs in AT-blocks containing 122 predicted genes and further sub-divided in SSPs located at the borders (65 predicted genes) and SSPs within AT-blocks (57 predicted genes); (iii) genes encoding other proteins (non-SSPs) in AT-blocks (498 predicted genes) and further sub-divided in non-SSPs located at the borders (91 predicted genes) and non-SSPs within AT-blocks (407 predicted genes); (iv) genes encoding other proteins (non-SSPs) in GC-blocks (11,394 predicted genes) and (v) all 12,469 genes predicted by EuGene in the genome (Table 4).

Base composition of the genes encoding SSPs (percent of each base in the sequence, GC content and GC3 content) and amino acid count of the SSPs (as% of each amino acid in the protein) were calculated by custom Python scripts. Statistical bias in amino acid occurrence was evaluated by an F-test to determine if the variances were equal in both sets, followed by a Student t test (95% confidence level) to compare the mean use of each amino acid in each set of predicted proteins (Supplementary Table S12).

RIP indices were calculated for each set of genes, compared using the non parametric Mann-Whitney test as implemented in ExcelStat v.2010.3.02, and the distribution of the values was displayed as a boxplot graph (Supplementary Figure S15). TpA/ApT indices were significantly higher ($P < 0.0001$) for all genes located within AT-blocks as compared to those located within GC-blocks (Table 4; Supplementary Figure S15). Similarly, TpA/ApT indices were significantly higher ($P < 0.0001$) for SSP-encoding genes located within AT-blocks as compared to SSP-encoding genes located in GC-blocks (Table 4; Supplementary Figure S15). In contrast SSP and other genes in AT-blocks did not show significantly different TpA/ApT indices ($P = 0.704$, Table 4; Supplementary Figure S15).

Biases in codon usage were evaluated using EMBOSS:CHIPS. The proportion of usage of a given codon among the set of codons that code for this codon's amino acid was calculated for each set of predicted proteins. The Relative Synonymous Codon Usage (RSCU) was then calculated by dividing the fraction of a codon corresponding to an amino acid by the number of synonymous codons. For a given amino acid, the codon with the highest RSCU value was determined as being the preferential codon (Supplementary Table S11). Whereas biases in codon usage were observed, this only had limited consequences on amino acid usage with only four amino acids showing significant differences when comparing the set of SSPs in AT-blocks with other predicted proteins or

SSPs in GC-blocks: two, C and F were more common, and two, D and E, were less commonly used. In addition, all SSPs, whatever the genome location of their encoding gene, were depleted in K and R as compared to other predicted proteins in the genome (Supplementary Table S12).

Presence and identity of TEs in the vicinity of SSPs (up to 5000 bp) was sought using a Python script, and validated by visual inspection. A Chi-squared test for given probabilities with simulated values (20,000 replicates) as implemented in R was performed to test random association of SSPs in AT-blocks with specific TEs. The analysis revealed that this null hypothesis should be rejected (Chi-squared = 232.1, $P < 0.0001$). One of the most common retrotransposon in the genome, RLG_*Polly*, was significantly under-represented in the vicinity of SSPs (Supplementary Figure S13). In contrast, three TEs were significantly over-represented in the proximity of SSPs: an uncharacterized minor retrotransposon, RLx_*Ayoly*, and two DNA transposons, DTF_*Elwe* and DTx_*Gimli* (Supplementary Figure S13).

Multigene families of SSPs were sought within the predicted proteins of the genome by PSI-BLAST with variable E-value cut-off and iteration numbers. Output was analysed by a Python script and families built with the following criteria: coverage > 75% of the query's length, and percentage of similarity > 20%. Putative functions were investigated by additional BLAST searches against the NR database.

Oomycete phytopathogens produce a wide variety of effectors that are delivered within the plant cell using translocation signals composed of an RXLR motif and a nearby acidic motif, dEER[9]. The fungal effectors known to date contain no obvious RXLR or dEER motifs, but an experimental analysis of the range of residues within the RXLR motif revealed that lysine (K) or histidine (H) but not glutamine (Q) could replace the arginine at position 1 in the motif, that any large hydrophobic residue (isoleucine, I; methionine, M; phenylalanine, F; tyrosine, Y) could replace the leucine (L) at position 3, albeit with varying efficiencies, but valine (V) and alanine (A) could not. At position 4, K, Q and G allowed function. Furthermore, the presence of either an L or M residue at position 2 could substitute for a large hydrophobic residue at position 3[115]. Interestingly, *L. maculans AvrLm6* possesses two such putative translocation motifs in its N terminus (RTLK and RYWT), of which the RYWT, but not the RTLK was found to be functional[26]. A Python script to identify putative translocation motifs in *L. maculans* SSPs was deployed. The pipeline encompassed six steps, and its final output is a 0-5 score; the lower the score, the higher the probability of an RxLR-like motif. Step 1 is harmonisation of the cleavage position of

the signal peptide when SignalP3.0 and TMHMM did not agree on its location (in this case a + 0.5 penalty was applied), step 2 is the search for motifs using EMBOSS:Fuzzpro (<[RKH] X [LMIFYW] X> or <[RKH] [LMIFYW] X [RKH]>). This created a set of protein sequences including the motifs and their location from the start of the sequence, of which motifs identified in the signal peptide were excluded. Score increases were + 4.0 (with exclusion from subsequent steps) if no motif was found, and + 2.0 if one or more motifs were found. The third step analysed the location of the motif within the protein with a + 2.0 penalty if (i) motifs were located within the first 10 amino acids or (ii) if they were farther away than the first 100 amino acids, or (iii) they are located within the last 20% of the amino acids of the protein. The fourth step analysed the composition of the motifs with scores decreased by 0.5 if there was a [LMIFYW] motif at positions 2 or 3 of the four-amino acid motif, and also decreased by 0.5 if there was a [RKH] motif at positions 2 or 4. The fifth step evaluated the distance between motifs when more than one was present: a decreased score of 0.2 was applied if the two motifs were separated by 1-10 amino acids. The sixth step evaluated the distance of the motif from the C-terminus. A +0.5 penalty was applied if the motif was within the 20 last amino acids of the protein (which was redundant with step 3 for proteins smaller than 100 AA).

Since avirulence genes of *L. maculans* evolve mainly via deletion when submitted to selection pressure by extensive sowing of Brassica cultivars with particular resistance genes[39,41], presence/absence PCR-based assays were used to determine the dispensability of SSP-encoding genes in a wide sample of 1043 isolates collected world-wide[116]. Of 36 predicted SSP-encoding genes analysed (all located within AT-blocks), 44.5% showed a presence/absence polymorphism with wide variation in the number of polymorphic SSPs from one population to the other (A. Dilmaghani, unpublished data). Primers for these SSP-encoding genes are available on request.

Whole-genome oligoarray and QRT-PCR analyses of expression of genes encoding SSPs

The *L. maculans* whole-genome expression array was manufactured by NimbleGen Systems Limited (Madison, WI). It contains fourteen independent, non-identical, 60-mer probes per gene model, each being duplicated on the array. The considered gene models were 12,396 EuGene-predicted gene models, 63 SSP-encoding genes not included in the EuGene gene models, 1316 clustered ESTs that did not match the gene models, 8651

random 60-mer control probes and labelling controls. Total RNA was extracted from mycelia grown during one week in Fries liquid medium and from oilseed rape-infected leaves (7 and 14 days post inoculation), using TRIzol reagent (Invitrogen) according to the manufacturer's protocol. Total RNA was treated with DNase I RNase-Free (New England Biolabs). Total RNA preparations (three biological replicates for each sample) were amplified by NimbleGen using the SMART PCR cDNA Synthesis Kit (Invitrogen) according to the manufacturer's instructions. Single dye labeling of samples, hybridization procedures, data acquisition, background correction and normalization were performed at the NimbleGen facilities (NimbleGen Systems, Reykjavik, Iceland) following their standard protocol[117,118]. Average expression levels were calculated for each gene from the independent probes on the array and were further analysed.

Gene-normalized data were subjected to Analysis of NimbleGen Array Interface Suite[119] (ANAIS; http://anais.versailles.inra.fr). ANAIS performs an ANOVA test on log-10 transformed data to identify statistically differentially expressed genes. This test uses the observed variance of gene measurements across the three replicated experiments. To deal with multiple testings, the ANOVA *P*-values are further subjected to the Bonferroni correction. Transcripts with a *P*-value lower than 0.05 and more than 1.5 change in transcript level were considered as significantly differentially expressed during infection compared to mycelial growth.

To estimate the signal-to-noise threshold (signal background), ANAIS calculates the median of the intensity of all of the random probes present on the microarray, and provides adjustable cut-off levels relative to that value. Gene models with an expression higher than three-times the median of random probe intensities in at least two of three biological replicates were considered as transcribed. Among the 13,779 gene models and ESTs included in the oligoarray, 11,687 (84.8%, corresponding to 84.4% of the EuGene-predicted gene models, 51.0% of the additional SSP-encoding genes and to 90.8% of the clustered ESTs not matching with the gene models) were expressed above background level in at least one of the analysed conditions. Taking into account transcriptomic, proteomic and EST support, the existence of 84.8% of the gene models was validated (Table 2). It suggests other gene models will be validated when analysing other biological conditions and developmental stages, such as dormant conidia (with a few cases of EST evidence for SSPs in the absence of transcriptomic evidence), development of fruiting bodies, late stages of plant infection or saprophytic life. From the 643 SSP-encoding genes represented in the oligoarray, 464 (70.2%) were expressed above background level

in at least one of the conditions. SSP-encoding genes located within AT-blocks were specifically or highly expressed during the primary plant infection, as already found for *AvrLm1*, *AvrLm4-7* or *AvrLm6*[15], and could play the role of effectors at this stage of infection. Whether their expression is controlled by a common transcription factor or by their particular genomic location is an intriguing question.

Primers for qRT-PCR experiments were designed for 22 of the genes encoding for SSPs, previously shown as being expressed in at least one of the conditions tested and located either in GC-blocks (4 genes), AT-blocks (13 genes) or in the bordering zones between the two (5 genes). Total RNA prepared for oligoarray experiments was used and extraction of total RNA from infected leaves 3 days post inoculation was also performed. All samples, either mycelia or infected plant tissues, were adjusted to 4 µg of RNA and single-strand cDNA was generated using oligo-dT-primed reverse transcription (RT) with PrimeScript Reverse Transcriptase (Clontech, CA) according to the manufacturer's protocol. For each condition tested, three RNA extractions from different biological samples were performed and two technical repeats were analyzed. Water was used as a negative control. qRT-PCR was performed using 7700 real-time PCR equipment (Applied Biosystems) and ABsolute SYBR Green ROX dUTP Mix (ABgene, Courtaboeuf, France), as previously described[13]. Ct values were analyzed as described by Muller *et al*.[120] for expression kinetic analysis. β-*tubulin* was used as a constitutive reference gene (Supplementar Figure S14). Primers used for qRT-PCR experiments are available upon request.

Proteomics and secretomics

To validate predicted secretion signals, mycelia and filtrates of seven-day old cultures of isolate v23.1.3 grown in liquid V-8 medium were subjected to proteomic analyses, using liquid-phase isoelectric focusing prior to high-resolution 2-D electrophoresis and shotgun proteomics (1-DE followed by liquid chromatography-mass spectrometry (LC-MS/MS))[121]. Up to approximately 2000 2-D spots were resolved in the culture filtrate (secretome) and MS identified major secretome markers such as endopolygalacturonases, β-glucanosyltransferases, pectate lyases and endoglucanases. Shotgun proteomic experiments showed the enrichment of secreted proteins within the culture filtrate with 83% of the proteins containing a predicted signal peptide, as expected[121]. These data were reanalysed to identify SSPs that could be identified in the *L. maculans* secretome. Thirty-

nine SSP were found in the secretome of which only two were from AT-blocks borders and none from within AT-blocks. The scarcity of SSP from AT-blocks in the secretome of *in vitro* grown mycelia is consistent with transcriptomics and qRT-PCR data indicating overexpression *in planta* but low or no expression in axenic culture.

Supplementary references

60. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537,** 39-64 (2009)
61. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **57,** 758-771 (2008)
62. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104-2105 (2005)
63. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7,** 439 (2006)
64. Lücking, R., Huhndorf, S., Pfister, D.H., Plata, E.R. & Lumbsch, H.T. Fungi evolved right on track. *Mycologia* **101,** 810-822 (2009)
65. Sung, G.H., Poinar, G.O. & Spatafora, J.W. The oldest fossil evidence of animal parasitism by fungi supports a Cretaceous diversification of fungal-arthropod symbioses. *Mol. Phylog. Evol.* **49,** 495-502 (2008)
66. Taylor, J.W. & Berbee, M.L. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* **98,** 838-849 (2006)
67. Ansan-Melayah, D., Balesdent, M.-H., Buée, M. & Rouxel, T. Genetic characterization of *AvrLm1*, the first avirulence gene of *Leptosphaeria maculans*. *Phytopathology* **85,** 1525-1529 (1995).
68. Newton, A. C. & Caten, C. E. Auxotrophic mutants of *Septoria nodorum* isolated by direct screening and by selection for resistance to chlorate. *Trans. British Mycol. Soc.* **90,** 199-207 (1988)
69. Lim, L. & Howlett, B. J. Linear plasmids, pLm9 and pLm10, can be isolated from the phytopathogenic ascomycete *Leptosphaeria maculans* by pulsed-field gel electrophoresis. *Curr. Genet*. **26,** 276-280 (1994).
70. Kuhn, M. L., *et al.* Genetic linkage maps and genomic organization in *Leptosphaeria maculans. Eur. J. Plant Pathol.* **114,** 17-31 (2006)
71. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573-580 (1999).
72. Bally, P., Grandaubert, J., Rouxel, T. & Balesdent, M.H. FONZIE: An optimized pipeline for minisatellite marker discovery and primer design departing from large sequence data sets. *BMC Research Notes* (in press)

73. Lander, E. S., *et al.* Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations.*Genomics* **1,** 174–181 (1987)
74. Martins, W. S., Lucas, D. C. S., Neves, K. F. S. & Bertioli, D. J. WebSat - A Web software for microsatellite marker development. *Bioinformation* **3,** 282-283 (2009).
75. Cozijnsen, A. J., Popa, K. M., Purwantara, A., Rolls, B. D. & Howlett, B. J. Genome analysis of the plant pathogenic ascomycete *Leptosphaeria maculans*; mapping mating type and host specificity loci. *Mol. Plant Pathol.* **1,** 293–302 (2000)
76. Kurtz, S., *et al*.Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004)
77. Caten, C.E. & Newton, A.C. Variation in cultural characteristics, pathogenicity, vegetative compatibility and electrophoretic karyotype within field populations of *Stagonospora nodorum*. *Plant Pathol.* **49,** 219–226 (2000)
78. Schiex, T., Moisan, A. & Rouzé, P. EuGene: An Eucaryotic gene finder that combines several sources of evidence. *Lecture Notes Comp. Sci.* **2066,** 111-125 (2001).
79. Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **15,** 1332-1338 (2005)
80. Stein, L. D., *et al*. The generic genome browser: a building block for a model organism system database. *Genome Res.* **12,** 1599-1610 (2002).
81. Lewis, S. E., *et al*. Apollo: a sequence annotation editor. *Genome Biol.* **3,** research0082 (2002)
82. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. **25,** 955-964 (1997)
83. Shiu, P.K., Raju, N.B., Zickler, D. & Metzenberg, R.L.: Meiotic silencing by unpaired DNA. *Cell* **107,** 905-916. (2001)
84. Fulci, V. & Macino, G. Quelling: post-transcriptional gene silencing guided by small RNAs in *Neurospora crassa. Curr. Opin. Microbiol.* **10,** 199-203 (2007)
85. Kouzminova, E. & Selker, E.U. Dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*. *EMBO J.* **20,** 4309-4323 (2001)
86. Freitag, M., Williams, R.L., Kothe, G.O. & Selker, E.U. A cytosine methyltransferase homolog is essential for repeat induced point mutation in *Neurospora crassa*. *Proc.*

Natl. Acad. Sci. USA **99,** 8802-8807 (2002)

87. Freitag, M., Hickey, P.C., Khlafallah, T.K., Read, N.D. & Selker, E.U. HP1 is essential for DNA methylation in *Neurospora*. *Mol. Cell* **13,** 427-434 (2004)
88. Galagan, J.E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa. Nature* **422,** 859-868 (2003).
89. Altschul, S. F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402 (1997).
90. Quesneville, H., *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1,** 166-175 (2005).
91. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*. **12,** 1269-1276 (2002).
92. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1,** i152-158 (2005).
93. Huang, X. On global sequence alignment. *Comput. Appl. Biosci.* **10,** 227-35 (1994).
94. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0.1996-2004, http://www.repeatmasker.org.
95. Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. CENSOR-a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20,** 119-121 (1996).
96. Kolpakov, R., Bana, G. & Kucherov, G. Mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **31,** 3672-3678 (2003).
97. Gao, X., Hou, Y. Ebina, H., Levin, H.L. & Voytas, D.F. Chromodomains direct integration of retrotransposons to hétérochromatine. *Genome Res.* **18,** 359-369 (2008)
98. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-4680 (1994)
99. Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33 (Web Server issue),** W557–W559 (2005)
100. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14,** 685-695 (1997)
101. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a

molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160-174 (1985)

102. Aveskamp, M.M., de Gruyter, J., Woudenberg, J.H., Verkley, G.J. & Crous, P.W. Highlights of the *Didymellaceae*: A polyphasic approach to characterise *Phoma* and related pleosporalean genera. *Stud. Mycol.* **65,** 1-60 (2010)
103. Pérez-Gonzales, C.E. & Eickbush, T.H. Dynamics of R1 and R2 elements in the rDNA locus of *Drosophila simulans. Genetics* **158,** 1557-1567 (2001).
104. Götz, S., *et al.* High-throughput functional annotation and data mining with the blast2GO suite. *Nucleic Acids Res.* **36,** 3420-3435 (2008)
105. Gardiner, D. M., Cozijnsen, A. J., Wilson, L. M., Pedras, M. S. C. & Howlett, B. J. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans. Mol. Microbiol.* **53,** 1307-1318 (2004)
106. Idnurm, A., Taylor, J. L., Pedras, M. S. C. & Howlett, B. J. Small-scale functional genomics of the blackleg fungus, *Leptosphaeria maculans*: analysis of a 38 kb region. *Aust. Plant Pathol.* **32,** 511-519 (2003)
107. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443-453 (1970).
108. Bushley, K.E., Ripoll, D.R. & Turgeon, B.G. Module evolution and substrate specificity of fungal nonribosomal peptide synthetases involved in siderophore biosynthesis. *BMC Evol. Biol.* **8,** 328 (2008)
109. Oide, S., Krasnoff, S.B., Gibson, D.B. & Turgeon, B.G. Intracellular siderophores are essential for Ascomycete sexual development in heterothallic *Cochliobolus heterostrophus* and homothallic *Gibberella zeae. Euk. Cell* **6,** 1339-1353 (2007).
110. Oide, S., Moeder, W., Krasnoff, S., Gibson, D., Haas, H., Yoshioka, K. & Turgeon, B.G. NPS6, encoding a nonribosomal peptide synthetase involved in siderophore-mediated iron metabolism, is a conserved virulence determinant of plant pathogenic ascomycetes. *Plant Cell* **18,** 2836-2853 (2006)
111. Pedras, M.S.C., Taylor, J.L. & Nakashima, T.T. A novel chemical signal from the blackleg fungus; beyond phytotoxins and phytoalexins. *J. Org. Chem*. **58,** 4778-4780 (1997)
112. Fedorova, N. D. *et al.* Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PloS Genet.* **4,** e1000046 (2008)
113. Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism : from biochemistry to genomics. *Nature Rev. Microbiol.* **3,** 937-947 (2005)

114. Bushley, K.E. & Turgeon, B.G. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol. Biol.* **10,** 26 (2010)
115. Dou, D., *et al.* RXLR-mediated entry of *Phytophthora sojae* effector Avr1b into soybean cells does not require pathogen-encoded machinery. *Plant Cell* **20,** 1930-1947 (2008).
116. Dilmaghani, A., *et al.* The *Leptosphaeria maculans–Leptosphaeria biglobosa* species complex in the American continent. *Plant Pathol.* **58,** 1044-1058 (2009)
117. Bolstad, B. M., Iizarry, R. A., Anstrand, M. & Speed, T. P. A comparison of normalisation methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185-193 (2003).
118. Irizarry, R. A., *et al.* Exploration, normalisation, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249-264 (2003).
119. Simon, A. & Biot, E. ANAIS: Analysis of NimbleGen Arrays Interface. *Bioinformatics* **26,** 2468-2469 (2010)
120. Muller, P.Y., Janovjak, H., Miserez, A.R. & Dobbie, Z. Processing of gene expression data generated by quantitative real-time RT-PCR. *Biotechniques* **32,** 1372-1378 (2002)
121. Vincent, D., *et al.* Hunting down fungal secretomes using liquid-phase IEF prior to high resolution 2-DE. *Electrophoresis* **30,** 4118-4136 (2009).

CHAPITRE 2

Évolution et adaptation dans le complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa*

La première partie de ce chapitre est composée de l'article «Transposable Element-assisted evolution and adaptation within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal plant pathogens» prévu d'être soumis dans la revue scientifique *Genome Research*. Il décrit une étude de génomique comparative et évolutive au sein du complexe d'espèces *L. maculans-L. biglobosa* en se focalisant sur les relations entre les différents niveaux d'adaptation au colza de chacun de ses membres et l'invasion des génomes par des éléments transposables.

Suite à l'analyse du génome de *L. maculans* 'brassicae' (Lmb) (Chapitre 1), dans le but d'obtenir des informations supplémentaires sur l'invasion du génome par des ET et son incidence au niveau structural et fonctionnel, nous avons initié une étude de génomique comparative entre Lmb et les espèces les plus proches phylogénétiquement : les membres du complexe d'espèces *L. maculans-L. biglobosa*. Ce complexe, plus ou moins bien défini, est composé de deux espèces et plusieurs « sous-espèces » dont les membres sont des agents pathogènes des crucifères. Deux données importantes nous ont poussé à réaliser cette étude : (i) des analyses préliminaires effectuées sur les autres membres du complexe (électrocaryotypes et hybridation avec des ET connus) suggéraient qu'ils possédaient un génome de plus petite taille et pauvre en ET, (ii) les membres du complexe ne possèdent pas les même capacités infectieuses vis-à-vis du colza.

Nous avons donc, grâce aux technologies NGS (*Next Generation Sequencing*), séquencés cinq autres membres du complexe d'espèces. Le séquençage et l'annotation des génomes ont permis de confirmer que seul le génome de Lmb a été envahi par des ET, les autres membres ayant un génome de taille comparable à celle de la plupart des autres champignons ascomycètes filamenteux. Dans cet article, une analyse phylogénétique des différentes souches séquencées du complexe permet de mettre en évidence des temps de divergence entre les différents membres suggérant l'existence d'espèces distinctes dans les clades *L. maculans* et *L. biglobosa*, et d'affiner la date d'invasion du génome de Lmb par des ET. Cet article présente également une histoire évolutive des ET dans la lignée des Dothidéomycètes et l'incidence qu'ils ont pu avoir à l'intérieur du complexe d'espèces (i) sur l'évolution de la structure du génome en générant de la mésosynténie, (ii) sur la génération de gènes spécifiques ou (iii) sur le déplacement de gènes jouant un rôle important lors de la pathogenèse.

Pour cet article, j'ai contribué (i) à l'annotation automatique des gènes et de leur fonction pour chaque génome nouvellement séquencé, (ii) à la construction de la phylogénie des membres du complexe, (iii) à l'annotation automatique et manuelle des familles d'éléments répétés et l'étude de leur distribution dans le complexe et les espèces Dothidéomycètes proches, (iv) aux analyses de synténie et à la caractérisation des inversions intra-chromosomiques, (v) à la comparaison des protéomes et la mise en évidence des séquences spécifiques, (vi) à l'étude du comportement des gènes présents au sein des isochores AT conservés dans les autres génomes.

La deuxième partie de ce chapitre présente une étude de génomique comparative au sein de l'espèce *L. maculans* 'brassicae'. Deux nouvelles souches potentiellement divergentes des deux souches précédemment séquencées (v23.1.3 et WA74), ont été séquencées dans le but d'évaluer le polymorphisme des gènes codant des effecteurs ou impliqués dans la pathogenèse. Cette étude montre une divergence génomique très limitée entre les souches analysées. La principale différence entre ces souches est leur contenu en gènes, et en particulier en gènes codant pour des effecteurs, qui peuvent représenter jusqu'à 45 % des gènes souche-spécifiques. Cela a permis d'obtenir un répertoire exhaustif des PPS de l'espèce *L. maculans* 'brassicae' contenant 1177 séquences.

Transposable Element-assisted evolution and adaptation within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal plant pathogens

Jonathan Grandaubert[1], Rohan G. T. Lowe[2], Jessica L. Soyer[1], Conrad L. Schoch[3], Angela P. Van de Wouw[2], Barbara Robbertse[3], Nicolas Lapalu[4], Matthew G. Links[5], Bénédicte Ollivier[1], Valérie Barbe[6], Sophie Mangenot[6], Corinne Cruaud[6], Hossein Borhan[5], Barbara J. Howlett[2], Marie-Hélène Balesdent[1], Thierry Rouxel[1]

Planned to be submitted in *Genome Research*

---

[1] INRA-Bioger, UR1290, Avenue Lucien Brétignières, BP 01, 78850 Thiverval-Grignon, France; [2] School of Botany, The University of Melbourne, Vic. 3010, Australia; [3] NIH/NLM/NCBI, 45 Center Drive, MSC 6510, Bethesda, Maryland 20892-6510, USA; [4] INRA-URGI, Route de Saint Cyr, 78026 Versailles Cedex, France; [5] Agriculture and Agri-Food Canada/Agriculture et Agroalimentaire Canada, Saskatoon Research Centre, 107 Science Place Saskatoon, SK, S7N OX2, Canada; [6] GENOSCOPE, Centre National de Séquençage, Institut de Génomique CEA/DSV, 2, rue Gaston Crémieux, CP 5706, 91057 Evry Cedex, France

Abstract

Transposable Elements (TEs) are often considered as genome shapers, and whole genome sequencing sometimes shows that fungal phytopathogens develop “two-speed” genomes in which TE-enriched regions also are enriched in genes involved in niche adaptation. We have investigated when and how TE-mediated genome expansion took place in the *Leptosphaeria maculans-L. biglobosa* species complex, and the consequences it had on genome structure, adaptability and pathogenicity. The genomes of five members of this species complex, which have different host ranges and abilities to infect cruciferous plants, were sequenced. Compared to the 45-Mb *Leptosphaeria maculans* ‘brassicae’ reference genome, which has an unusual bipartite structure, in which large AT-isochores are made of degenerated and truncated TEs and are enriched in effector genes, all the other genomes were compact (30-32-Mb), with very few TEs (<4%). Forty five TE families, all affected by Repeat Induced Point mutations (RIP) were identified. Some TE families were lineage-specific, while others had been present in dothideomycete fungi, to which *Leptosphaeria* belongs, for more than 90 MYA. Phylogeny analyses indicate that *L. maculans* ‘brassicae’ and *L. maculans* ‘lepidii’ diverged 5.1 MYA, which coincided with a major burst of TE transposition in the genome of *L. maculans* ‘brassicae’. A nearly perfect synteny was observed at the chomosomal level between *L. maculans* ‘brassicae’ and *L. maculans* ‘lepidii’, but 30 intrachromosomal inversions were present and usually bordered by TEs. A similar gene number was predicted in each genome (~11,000), but 5-11% of the genes were species-specific. Effector genes and genes encoding secondary metabolites essentially showed either species-specific occurrence or inconsistent phylogenetic distribution. For these, examples of translocations or duplications were described in *L. maculans* ‘brassicae’ genome. In addition, presence of repeated element adjacent to the promoter of genes correlated with increased expression of these genes during plant colonisation.

Introduction

Fungi (and fungal-like oomycetes) are an incredibly diverse and adaptable group of organisms which colonise all habitats on Earth. Some are plant and animal pathogens of major economic importance and as stressed by Kupferschmidt (2012) “Fungi have now become a greater global threat to crops, forests, and wild animals than ever before. They have killed countless amphibians, pushing some species to extinction, and they are threatening the food supply for billions of people.” The sequencing of fungal genomes (the yeast *Saccharomyces cerevisiae*, followed by filamentous ascomycetes such as *Neurospora crassa*, and later plant pathogenic fungi such as *Ustilago maydis*, *Magnaporthe oryzae*, *Fusarium graminearum* and *Phaeosphaeria nodorum* (Galagan *et al.*, 2003; Dean *et al.*, 2005; Kämper *et al.*, 2006; Cuomo *et al.*, 2007; Hane *et al.*, 2007) initially indicated that genomes of filamentous ascomycetes are relatively small (21-39 Mb), with few repetitive elements (typically less than 10%) and they contain 10,000-15,000 protein-encoding genes. Unlike symbiotic and parasitic bacteria, which usually have smaller genomes than their free-living relatives (Raffaele & Kamoun, 2012), the first genome data from filamentous fungi did not indicate differences between genome size of fungal pathogens and that of saprophytes. However, contrasting to what is observed in bacteria, recent sequence data indicate that some phytopathogens have drastically expanded genomes, mostly due to massive invasion by Transposable Elements (TEs). The 45-Mb genome of *Leptosphaeria maculans*, the ascomycete that causes stem canker of crucifers contains 33% TEs, the 120-Mb genome of cereal downy mildew, *Blumeria graminis*, contains 64% TEs (while other closely related downy mildew pathogens have even larger genomes), and that of the 240-Mb oomycete *Phytophthora infestans* contains 74% TEs (Raffaele & Kamoun, 2012). These data, suggesting convergent evolution towards bigger genomes in widely divergent fungal and oomycete species, are intriguing and the counterbalance between selective advantage conferred to the pathogen and the cost of maintaining this amount of “parasitic” TE DNA is unknown.

Analysis of *L. maculans* ‘brassicae’ (Lmb) genome (Rouxel *et al.*, 2011) revealed a genomic structure at this time only observed in mammals and other vertebrates: the base composition (GC-content) varied widely along the chromosomes, but locally, was relatively homogeneous. Such structural features of chromosomes are termed “isochores” (Eyre-Walker & Hurst, 2001). In Lmb, the high TE proportion in conjunction with the RIP (Repeat-Induced Point mutation) mechanism (Galagan & Selker, 2004) are responsible for the generation of large AT-rich regions (33.9% GC-content), called AT-isochores. These AT-

isochores are scattered along the genome alternating with large GC-equilibrated regions (51% GC-content), GC-isochores, containing 95% of the predicted genes and mostly devoid of TEs. AT-isochores cover 36% of *L. maculans* genome, are mainly composed of mosaics of truncated and RIP-degenerated TEs and only contain 5% of the predicted genes, but 20% of those genes encode small secreted proteins (SSPs) considered as putative effectors of pathogenicity (Rouxel *et al.*, 2011). As only 4% of the genes located in GC-isochores encode SSPs, we hypothesized that AT-isochores were niches for effectors. This was corroborated by the presence genes encoding effectors in *L. maculans,* including four avirulence genes (*i.e.* encoding proteins recognised by plant resistance genes), *AvrLm1* (Gout *et al.*, 2006), *AvrLm6* (Fudal *et al.*, 2007), *AvrLm4-7* (Parlange *et al.*, 2009) and *AvrLm11* (Balesdent *et al.*, 2013) within these AT-isochores. Moreover, genes within AT-isochores are affected by RIP, which can occasionally overrun the repeated region into adjacent single-copy genes, resulting in extensive mutation of the affected genes. While selection pressure to maintain genes beneficial to the fungus would prevent their extinction due to an extensive degree of RIP (Fudal *et al.*, 2009; Van de Wouw *et al.*, 2010; Daverdin *et al.*, 2012). This genome environment allows a rapid response to selection pressure, for instance effector genes acting as avirulence genes can be inactivated in a single sexual cycle due to RIP and large-scale genome rearrangements in AT-isochores (Gout *et al.*, 2007; Fudal *et al.*, 2009; Daverdin *et al.*, 2012). In addition to genes encoding effectors and genes with no predicted function, AT-isochores are also enriched in gene clusters responsible for the biosynthesis of secondary metabolites (Rouxel *et al.*, 2011). Both effector genes and secondary metabolite gene clusters are distributed discontinuously among ascomycetes and/or show species-specific repartition (Patron *et al.*, 2007; Rouxel *et al.*, 2011). We have postulated that genome invasion by TEs, followed by their degeneracy by RIP shaping large AT-isochores, is a recent evolutionary event that has contributed to the rise of a better-adapted new species, and that maintaining large TE-rich regions hosting pathogenicity determinants has favoured adaptation to new host plants along with generation of new virulence specificities (Rouxel *et al.*, 2011).

This hypothesis can be now tested by exploiting comparative genomics between Lmb and other members of the *Leptosphaeria* species complex, for which preliminary electrokaryotype and Southern blot analyses suggested only limited invasion of the genome by TEs (M.R. Eckert, unpublished data; Supplementary Figure S1). *L. maculans* and *L. biglobosa* are dothideomycete phytopathogens specialized on crucifers and encompassing a series of ill-defined entities more or less adapted to oilseed rape/canola

(*Brassica napus*). Some of the species of the complex can attack oilseed rape and are responsible for the major disease (blackleg or Phoma stem canker) of oilseed rape (Mendes-Pereira *et al.*, 2003; Voigt *et al.*, 2005; Fitt *et al.*, 2006b). The complex also encompasses related lineages only found on cruciferous weeds. These include *L. maculans* 'lepidii' (Lml) isolated from *Lepidium* spp. and *L. biglobosa* 'thlaspii' (Lbt) isolated from *Thlaspi arvense*, but also found occasionally on *Brassica* species in Canada (Mendes-Pereira *et al.*, 2003; Voigt *et al.*, 2005). The *L. maculans* and *L. biglobosa* lineages infecting *B. napus* (Lmb, *L. biglobosa* 'brassicae' (Lbb) and *L. biglobosa* 'canadensis' (Lbc)) share morphological traits, epidemiology, infection strategies, ecological niches and regardless of geographic distribution, are often found together in tissues of individual infected plants (West *et al.*, 2002).

In this paper we investigate when and how genome expansion took place in the *L. maculans-L. biglobosa* species complex, and the consequences it had on genome structure, adaptability and pathogenicity. The genomes of five members of the species complex were sequenced: (i) an isolate of Lmb obtained from *B. napus* in Australia (the reference v23.1.3 was from Europe (Rouxel *et al.*, 2011), (ii) an isolate of Lml isolated from *Lepidium* spp. in Canada, (iii) an isolate of Lbb isolated from *Brassica juncea* in Europe, (iv) an isolate of Lbt isolated from *T. arvense* in Canada and (v) an isolate of Lbc isolated from *B. juncea* in Australia. The genomic data were analysed and compared to those from other dothideomycete species to address the following questions: (i) How divergent are the different isolates, when did they emerge and do they constitute separate biological species? This question also aims at knowing whether Lmb actually is a recently emerged species; (ii) when did TEs invade the genome(s) in the *L. maculans-L. biglobosa* lineage and is genome expansion a recent trait or an ancestral one lost in recently differentiated species?; (iii) did TE invasion (and how did it) contribute to the rise of a better adapted and more adaptable species: what was the incidence of TE invasion on the generation of novel pathogenicity determinants (including secondary metabolite gene clusters), the rise of a species better adapted to oilseed rape and adaptability in terms of co-evolution (constant production of new virulent phenotypes and occasional host-jumps)?

Results

Phylogeny and divergence time estimates

Dothideomycetes is the largest and the most phylogenetically diverse class within the largest fungal subphylum, the *Pezizomycotina* (filamentous ascomycetes), and encompasses numerous plant pathogens causing serious crop losses such as species in the genera *Cochliobolus*, *Phaeosphaeria*, *Pyrenophora*, *Mycosphaerella*, *Zymoseptoria* and *Venturia* (Schoch *et al.*, 2009a; Ohm *et al.*, 2012) (Figure 1). *L. maculans* and *L. biglobosa* belong to order Pleosporales, in class *Dothideomycetes*.

Alignments of 19 proteins were used for phylogeny analyses and divergence time estimates. The chronogram presented in Figure 1 (for a more detailed analysis see Supplementary Figure S2) has representatives from three of the four main classes in the filamentous *Ascomycota*, but the main focus is on *Dothideomycetes*. All nodes had bootstrap values above 70% in a separate RAxML analysis with only the placement of *Alternaria brassicicola* poorly resolved. The phylogeny is congruent to other more complete analyses (Schoch *et al.*, 2009a, 2009b) and indicates that the *Leptosphaeria* species analysed diverged from the sampled plant pathogens of Pleosporaceae (*Cochliobolus*, *Pyrenophora* and *Alternaria*) at approximately 73 MYA. These analyses relied on previously published data incorporating numerous fossil calibrations, but the divergence dates agree with placements of the most detailed *Ascomycota* fossils (Berbee & Taylor, 2010). The mean divergence date of the two *Pyrenophora* species analysed (7 MYA) also compares well with a recent estimate of 8 MYA for these species (Ellwood *et al.*, 2012). This latter study used a different approach: non coding DNA regions of the genome were analysed and divergence was determined based on a commonly used average substitution rate of 8.8×10^{-9} per site per year (Kasuga *et al.*, 2002). Based on morphological features, *L. maculans* and *L. biglobosa* were considered as two pathotypes of the same species for more than 70 years (Cunningham, 1927; Pound, 1947) until formal renaming in 2001 (Shoemaker & Brun, 2001). Our data estimate divergence time of 22 MYA between *L. maculans* and *L. biglobosa* (Figure 1).

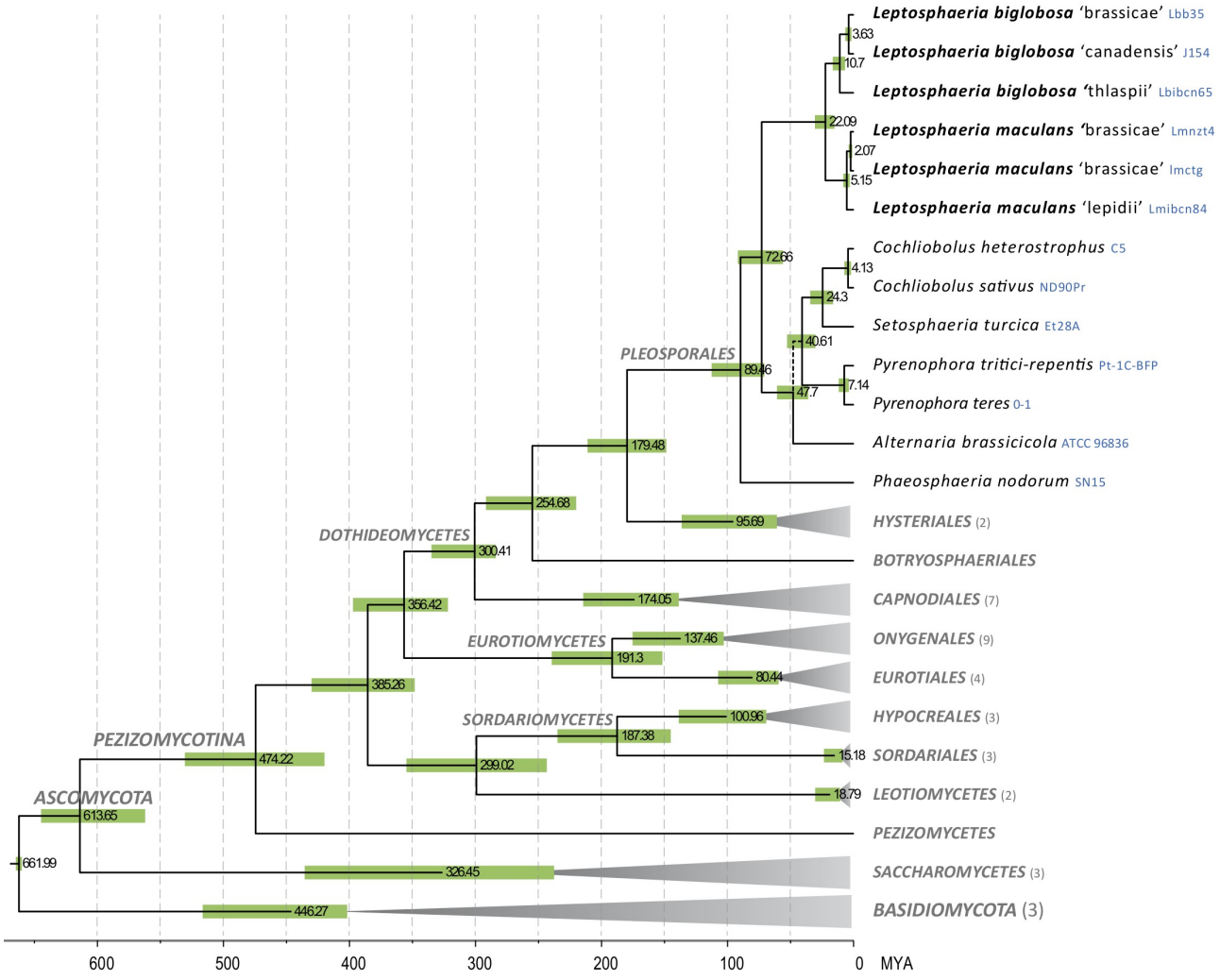


Figure 1. Chronogram of major classes in *Ascomycota*, with a focus on *Dothideomycetes*. The chronogram produced with BEAST from a data set of 19 truncated proteins. Branches with terminal grey triangles were collapsed and their number of leaves indicated in brackets after taxon labels. Numbers at nodes indicate mean node ages in millions of years and green bars indicate their 95% highest posterior density intervals.

Within the two separate branches, noticeable divergence times were also identified for the *L. maculans* and *L. biglobosa* lineages. Using three protein-encoding genes, Voigt *et al.* (2005), suggested that Lbt was intermediate between *L. maculans* and *L. biglobosa*. However our current data clearly support Lbt belonging to the *L. biglobosa* clade but being divergent by *ca.* 11 MYA from the terminal Lbb or Lbc (Figure 1; Supplementary Figure S2). The divergence time between Lmb and Lml (5.1 MYA) is in the same range (3.6 MYA) as that of Lbb and Lbc (Figure 1).

Genome statistics

The reference genome of Lmb (of strain v23.1.3), was sequenced in 2007 using a whole genome shotgun strategy and Sanger technology, assembled using an improved version of Arachne (Jaffe *et al.*, 2003) and described in Rouxel *et al.* (2011). In this study, we used Next Generation Sequencing (NGS) technologies (454 and/or Illumina) to sequence the genomes of five members of the species complex, including another isolate of Lmb. The sequences were assembled using Newbler (Margulies *et al.*, 2005), except for the Lbc isolate (J154) whose assembly was obtained using Velvet.

Despite different sequencing and assembly strategies, the resequenced isolate of Lmb (WA74) had a similar genome size and proportion of TEs (25.8%) compared to v23.1.3 (32.5% of TEs) (Table 1). However, the 454/Illumina assembly resulted in a more fragmented genome with a high number of scaffolds and a low N50 value (Table 1). Not unexpectedly, the Lbc isolate showed an even more fragmented genome (Table 1), preventing comparison at the chromosomal scale or detailed analysis of TE content. The Lml and to a lesser extent the Lbt isolates had a high-quality assembly with number of scaffolds and N50 data similar to those obtained for the reference genome (Table 1; Figure 2; Supplementary Figure S3).

Compared to the Lmb isolates, all the other isolates of the species complex had smaller genomes, ranging from 30.2 to 32.1 Mb and only comprising from 2.7% to 4.0% of TEs. These sequencing data corroborate previous electrokaryotypes done using Pulsed Field Gel Electrophoresis (PFGE) and results of DNA hybridization with known TEs (Supplementary Figures S1, S4).

Table 1. Sequencing statistics and genome facts for six members of the *Leptosphaeria maculans-L. biglobosa* species complex.

	L. maculans 'brassicae'		*L. maculans* 'lepidii'	*L. biglobosa* 'brassicae'	*L. biglobosa* 'canadensis'	*L. biglobosa* 'thlaspii'
	v23.1.3	**WA74**	**IBCN84**	**B3.5**	**J154**	**IBCN65**
Genome size (Mb)	45,1	44,2	31,5	31,8	30,2	32,1
Contig number	1743	3765	2802	2533	7124	3506
Scaffold number	76	986	123	606	6748	237
Scaffold N50 (kb)	1770	263	1356	779	245	715
Gaps (%)	2,5	9,6	7,1	7,4	0,1	8,7
Repeats (%)	35,5	27,5	4,0	4,4	3,9	5,1
TEs (%)	32,5	25,8	2,7	3,2	2,9	4,0
'No repeats' genome size (Mb)	29,3	28,5	28,4	28,4	29,3	28,0
GC genome (%)	45,2	46,5	50,9	51,4	51,1	51,4
GC 'no repeats' genome (%)	51,6	51,6	51,6	52,0	51,6	52,1
GC TEs (%)	34,3	34,5	32,3	36,6	34,7	36,6
Predicted gene number	12543	10624	11272	11390	11068	11691

The overall GC-content of the Lmb reference genome was similar to that of Lmb isolate WA74 (Table 1). In contrast, all other isolates of the species complex had genome GC-contents (around 51%) that related more to those of other *Dothideomycetes* (50-52%, Supplementary Table S1) (De Wit *et al.*, 2012; Ohm *et al.*, 2012; Manning *et al.*, 2013) than to that of Lmb (Table 1) (Rouxel *et al.*, 2011). As previously observed in Lmb, all TEs were affected by RIP (data not shown), explaining their similarly low GC-content in all isolates of the species complex. This was further substantiated by the finding that all the genomes contain orthologs of the *Neurospora crassa* genes currently postulated to be necessary for RIP (Espagne *et al.*, 2008) (Supplementary Table S2). Since the GC-content of nonrepetitive sequences (51.6-52.1%) and that of TEs that had been mutated by RIP (32.3-36.6%) were equivalent in all isolates sequenced here (Table 1), the overall TE-content of each genome directly influenced the overall GC-content of the genomes. These data again indicate that Lmb is an exception in terms of TE-content compared to related *Dothideomycetes* but also to even more closely related species of the species complex, some of which diverged from Lmb only 5.1 MYA (Figure 1). Actually, only very distant dothideomycete species of order Capnodiales such as *Pseudocercospora fijiensis* and *Cladosporium fulvum* (Figure 1; Supplementary Table S1) have genomes enriched in TEs, resulting in low overall GC-content of the genome (de Wit *et al.*, 2012; Ohm *et al.*, 2012) (Supplementary Table S1).

In all genomes, the size of the core genome, excluding repeated elements was very similar at 28-29 Mb (Table 1). This may indicate that the size differences between genomes of the members of the species complex are essentially due to a difference in TE amount and not due to expansion or loss of gene families as found in other fungi (Spanu *et al.*, 2010; Duplessis *et al.*, 2011). Gene annotation predicted a comparable number (10624 to 11691) of protein-encoding genes in all genomes (Table 1). For each genome, the gene features were also very similar in terms of mean gene length, mean coding sequence length or proportion of genes with introns (Supplementary Table S3).

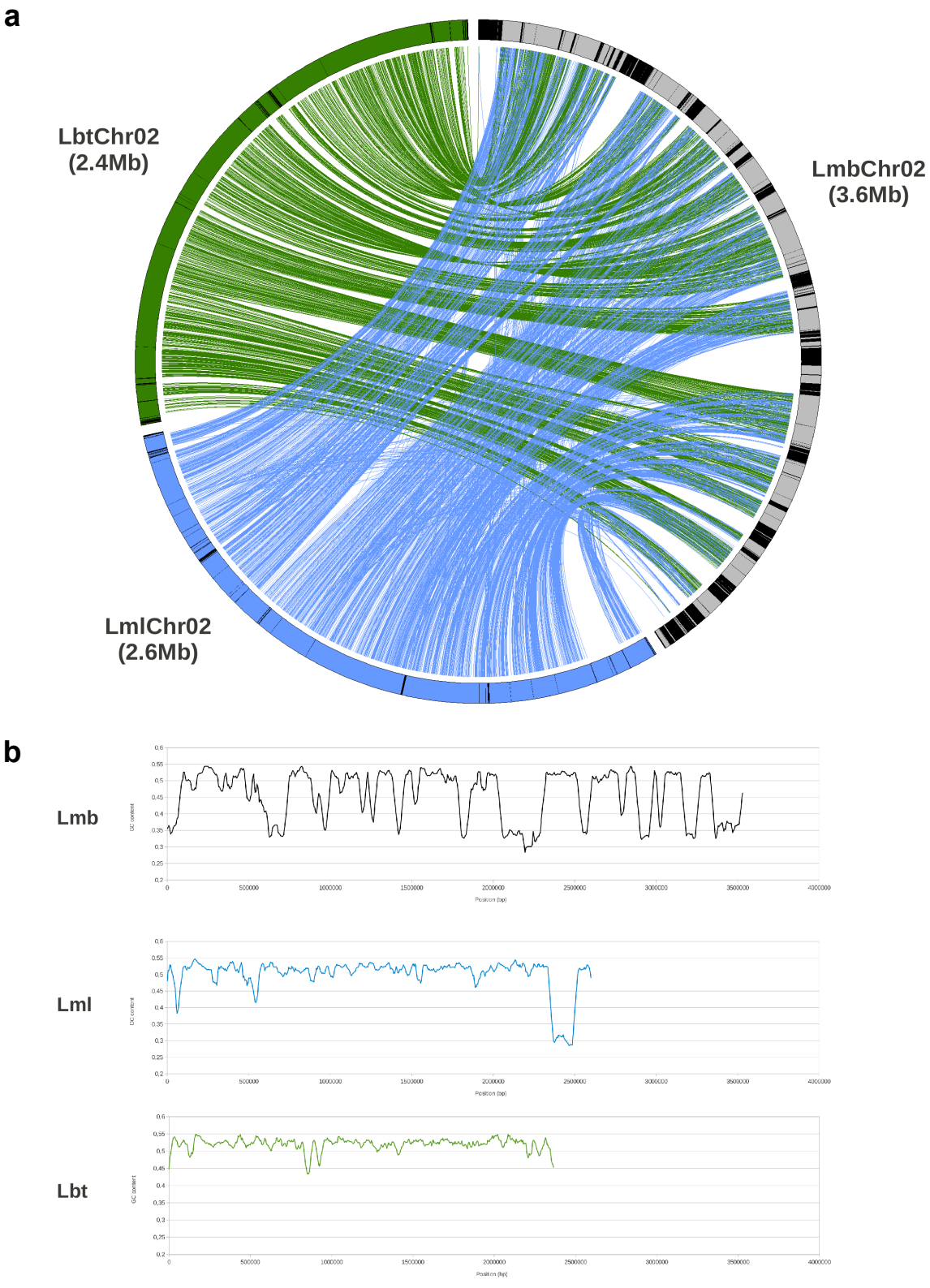


Figure 2. Comparative chromosome structure in *L. maculans* 'brassicae' (Lmb), *L. maculans* 'lepidii' (Lml) and *L. biglobosa* 'thlaspii' (Lbt): the example of chromosome 2. (**a**) Circos representation of synteny between the homologous chromosomes. The black parts represent the TE-rich, AT-rich chromosomal regions; (**b**) GC content changes along the chromosomes showing the typical isochore structure of the *L. maculans* 'brassicae' chromosomes, absent form the other species.

Chromosomal organisation and synteny

Lmb was estimated to have 17-18 chromosomes by combination of analysis of electrokaryotypes developed by PFGE (Supplementary Figure S1) and sequencing of the reference genome (Rouxel *et al.*, 2011). The new genomic data described here, especially those of the well-assembled genome of Lml (123 scaffolds, N50=1.35 Mb) (Supplementary Figure S3), allowed this estimate to be refined by considering scaffolding errors on the reference assembly. With the support of genetic mapping, the chromosome count was updated to 19 chromosomes (Supplementary Table S4), including one dispensable chromosome (Balesdent *et al.*, 2013) only present in the two Lmb isolates.

In addition to a direct consequence on the overall GC-content of the genomes, the limited amount of TEs in Lml, Lbb and Lbt resulted in a major difference in terms of chromosomal structure: the AT-rich landscapes were mostly restricted to chromosome ends in Lml, Lbb or Lbt (Figure 2). Thus, within the species complex, only the Lmb genome had an isochore structure alternating large AT-rich and GC-equilibrated regions.

Repeat-masked nucleotide sequences of each genome were aligned with the others using MUMmer (Kurtz *et al.*, 2004) and dot plots based on alignment data were used to compare genome organisation. The alignments between the two Lmb isolates showed a perfect conservation of macrosynteny (Supplementary Figure S3a), as also observed by Ohm *et al.* (2012) for isolates of *Cochliobolus heterostrophus*. At the largest divergence time included in our study, comparisons between Lmb and Lbb genomes, or between Lmb and Lbt showed a globally well-conserved synteny with many intrachromosomal inversions (Supplementary Figures S3c-d). Of particular interest, the alignment of the most closely related species Lmb and Lml showed a highly conserved macrosynteny pattern with only few major genomic rearrangements (Figure 3a; Supplementary Figures S3b, S5). The excellent quality of the assembly for v23.1.3 (Lmb) and IBCN84 (Lml) allowed more precise alignment at the chromosomal level. While large scale translocations were not seen, 30 intrachromosomal sequence inversions sized from 1.3 kb to 355 kb were identified (median size: 9.5 kb). These inversions were scattered along all the chromosomes and encompassed one to more than a hundred genes (Figure 3). The inversions within the chromosomes of the two isolates of Lmb analysed were bordered by TEs in 70% of the cases with TEs located at 82 bp on average (1-555 bp) from the inversion borders (Figure 3b; Supplementary Table S5).

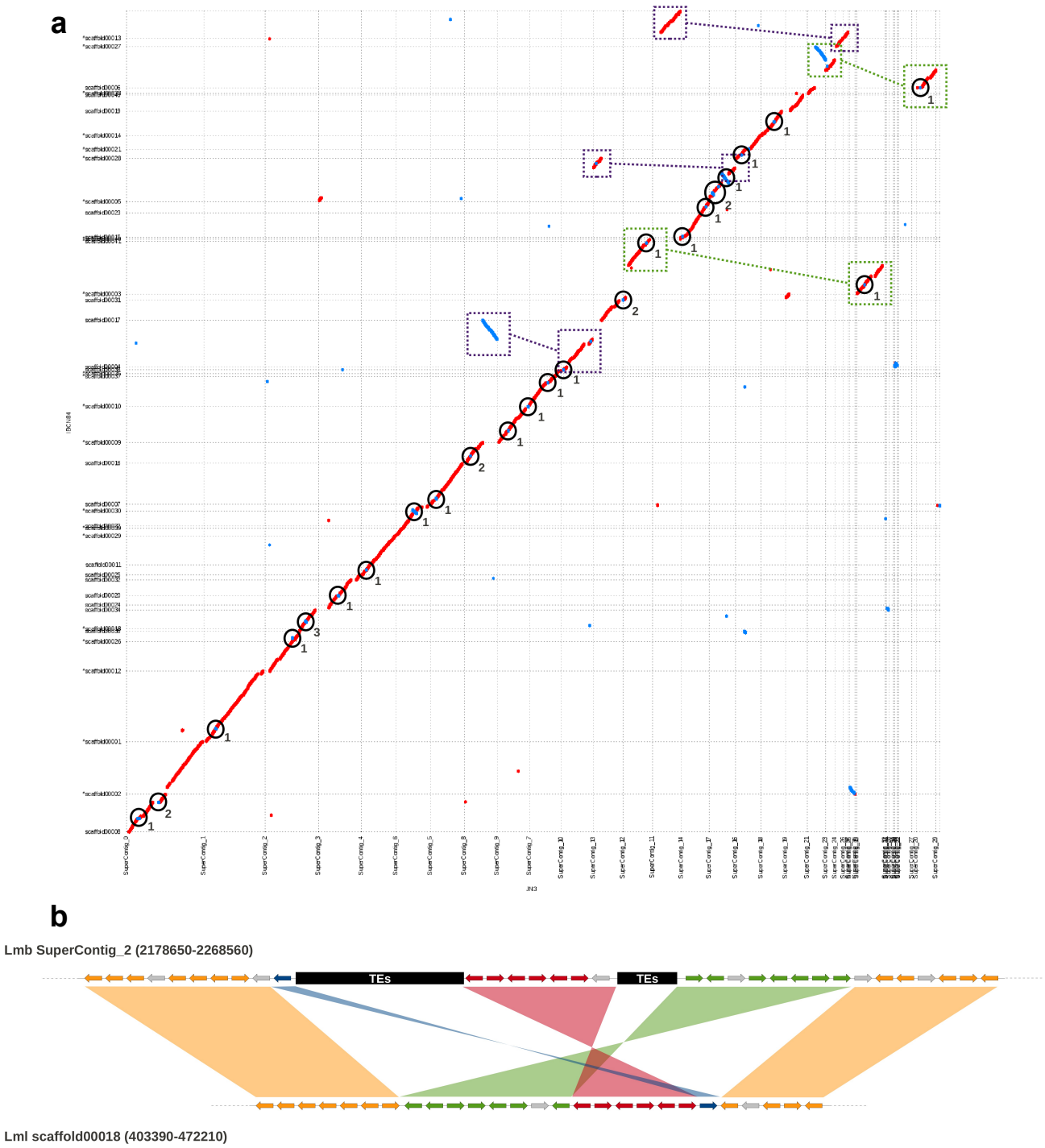


Figure 3. Genome alignment and synteny analyses between *L. maculans* 'brassicae' (Lmb) and *L. maculans* 'lepidii' (Lml). (**a**) Whole genome dotplot showing intrachromosomal inversions occurring in the genome of *L. maculans* 'brassicae' (horizontal axis) compared to that of *L. maculans* 'lepidii' (vertical axis). Inverted regions are circled in black and the number of inversions is mentioned close to the circle. Regions that are boxed and linked with another regions correspond to assembly errors in the Lmb genome. (**b**) Schematic zoomed representation of one SuperContig_2 region (in Lmb) and its syntenic region in Lml. The region contains three inversions (in blue, red and green) and the location of bordering transposable elements in Lmb are indicated as black boxes. The arrows represent genes and their orientation. Genes with no orthologs are colored in grey.

The two Lmb genomes were almost identical since 99.5% (27.9 Mb) of the non repetitive sequences were included in the alignment. A total of 20,404 SNPs was identified from aligned regions, corresponding to a frequency of 1 SNP every 1367 bases (Table 2). While there was a very high alignment coverage between recently diverged species of *L. maculans* and *L. biglobosa*, *i.e.* 86.4% (24.2 Mb) between Lmb and Lml and 92.5% of their nonrepetitive sequences between Lbb and Lbc, the divergence between the isolates was evident at the SNP count and density levels, *i.e.* 1,598,215 SNPs (1 SNP every 15 bases) between Lmb and Lml, and 1,459,160 SNPs (1 SNP every 18 bases) between Lbb and Lbc (Table 2). Comparison of more distant species, revealed differences not in the SNP density but in the total length of the aligned sequences: Lbt only shared 57% sequence coverage with Lbb and Lbc (15.6-16.3 Mb) with an average of 1 SNP every 9 bases of the aligned sequences and alignments between sequences of the two clades were very poorly aligned with a coverage length of 8.4 Mb (29.9%) with 1 SNP every 8 bases when comparing all *L. maculans* and *L. biglobosa* sequences. Alignments of translated sequences increased the coverage length to 16 Mb (57%). When sequences of the *Leptosphaeria* species complex were aligned to those of *Pyrenophora tritici-repentis* and *Phaeosphaeria nodorum*, coverage lengths only represented 2% of the nonrepetitive sequences, but 43% of the protein-encoding sequences (data not shown).

Table 2. Alignment length and SNP counts between the different genomes of members of the *L. maculans-L. biglobosa* species complex.

		***L. maculans* 'brassicae'**	***L. maculans* 'lepidii'**	***L. biglobosa* 'brassicae'**	***L. biglobosa* 'canadensis'**	***L. biglobosa* 'thlaspii'**
		Genome alignment length (Mb)				
***L. maculans* 'brassicae'**	No. of SNP	20404	24.2	8.1	8.4	8.6
***L. maculans* 'lepidii'**		1598215	-	8.2	8.6	8.7
***L. biglobosa* 'brassicae'**		1095420	1104219	-	25.9	15.6
***L. biglobosa* 'canadensis'**		1134785	1153336	1459160	-	16.3
***L. biglobosa* 'thlaspii'**		1147859	1163826	1739473	1794939	-

Table 3. Transposable element content in genomes of isolates of the *L. maculans-L. biglobosa* species complex.

		L. maculans 'brassicae'		*L. maculans* 'lepidii'		*L. biglobosa* 'brassicae'		*L. biglobosa* 'canadensis'		*L. biglobosa* 'thlaspii'	
TE class	**TE order**	**Bases covered**	**% of repetitive fraction**	**Bases covered**	**% of repetitive fraction**	**Bases covered**	**% of repetitive fraction**	**Bases covered**	**% of repetitive fraction**	**Bases covered**	**% of repetitive fraction**
Class I	Ty1-Copia LTR	3372500	22,11	77794	7,60	166805	14,89	34236	3,81	141432	9,95
	Ty3-Gypsy LTR	8749750	57,35	426573	41,69	368819	32,91	399244	44,39	365617	25,72
	LINE	6996	0,05	6523	0,64	15680	1,40	4535	0,50	54520	3,84
	Misc.	586134	3,84	1552	0,15	0	0,00	0	0,00	0	0,00
	Sub-total :	**12715380**	**83,34**	**512442**	**50,08**	**551304**	**49,20**	**438015**	**48,70**	**561569**	**39,51**
Class II	Tc1-Mariner	36075	0,24	193492	18,91	157676	14,07	76864	8,55	136579	9,61
	hAT	79154	0,52	1845	0,18	71796	6,41	21781	2,42	16950	1,19
	Mutator	890076	5,83	22524	2,20	42470	3,79	1344	0,15	3957	0,28
	MITE	0	0,00	11065	1,08	462	0,04	0	0,00	0	0,00
	Misc.	336819	2,21	19579	1,91	81576	7,28	30251	3,36	8034	0,57
	Sub-total :	**1342124**	**8,80**	**248505**	**24,29**	**353980**	**31,59**	**130240**	**14,48**	**165520**	**11,64**
Non-categorized	Non-categorized	**1199008**	**7,86**	**262241**	**25,63**	**215260**	**19,21**	**331133**	**36,82**	**694332**	**48,85**
	Total :	**15256512**		**1023188**		**1120544**		**899388**		**1421421**	

Transposable Elements and their evolutionary dynamics

Even though TEs are highly degenerated by RIP, the use of the REPET pipeline followed by extensive manual annotation allowed identification and classification of a large proportion of the TEs present in the genomes of Lmb, Lml, Lbb and Lbt (Supplementary Tables S6, S7; Supplementary Data 1).

Many types of TEs are represented within the genomes of the *Leptosphaeria* species including long interspersed elements (LINEs), non-long-terminal repeats (non-LTRs), Penelope-like elements (PLEs) and long-terminal repeats (LTRs) retrotransposons, terminal-inverted repeats (TIRs) and minitransposable elements (MITEs) DNA transposons (Table 3). However, LTR retrotransposons and TIR DNA transposons were largely prevalent.

TEs represented *ca.* 30% of the Lmb genome but only 4% and less in the other genomes (Table 1). Class I elements were more abundant in all genomes (39.5 to 50.1% of the repeated fraction in Lml, Lbb, Lbc and Lbt genomes) and up to 83% of the repetitive fraction in the genome of Lmb isolates. In contrast Class II elements were relatively more represented in genomes of Lml, Lbb, Lbc and Lbt (11.6 to 36.8% of the repetitive fraction) compared to those of Lmb isolates (8.8%) (Table 3). In spite of this relative difference, the total coverage of DNA transposons in the assemblies was higher in Lmb (3% of the genome) compared to that observed in the other isolates (0.8 to 1.1% in the assembly of Lml, Lbb or Lbt). These data also have to take into account the high proportion of non-categorized repeats in the genomes of Lml, Lbb, Lbc and Lbt (19.2-48.8%) and higher proportion of unresolved nucleotides in the assemblies compared to the reference Lmb v23.1.3 (Table 1) which probably result from difficulties in assembly of repeated reads from NGS technologies. Regardless of these uncertainties, and consistent with the lower GC-content of Lmb compared to that of other isolates, these data support the fact that Lmb is an exception in terms of TEs and further indicate that expansion of Class I TEs and mainly LTR retrotransposons is remarkable in this species compared to related ones (27.3% of the genome *vs.* only 1.6% for Lml; Table 3).

The repertoire of TEs in all the *Leptosphaeria* isolates sequenced here comprised 121 consensus sequences of which 57% were classified as Class I or Class II elements (Supplementary Tables S6, S7). Of these, 40 represented copy variants of a same TE family identified in several genomes and were grouped into 16 larger families, while the 29 remaining families were invariant between species (regardless of presence of RIP

mutations). Importantly, despite the major differences in genome coverage between Lmb and related species, a similar number of TE families (~30) were identified in each of the annotated genomes (Supplementary Tables S6, S7).

Previously, using an alignment-based phylogenetic approach with TE sequences deprived of the RIP-affected sites, we postulated that TE expansion in Lmb genome took place between 4 and 20 MYA for TEs such as DTM_*Sahana*, RLG_*Rolly*, RLG_*Olly*, RLG_*Rolly* and RLC_*Pholy* while others such as RLG_*Dolly* had been resident and maintained in the genomes for very long times (> 100 MYA) (Rouxel *et al.*, 2011). With the newly acquired data on the repertoire of TE families in the *Leptosphaeria* species complex along with a more accurate estimate of divergence time between its members and other dothideomycete species, we sought presence or absence of these TE families in the entire *Ascomycota* phylum to infer dates of invasion and patterns of gain or loss during evolution. For this purpose, the most GC-rich copies (*i.e.* one with the least degrees of RIP mutation) of the different TE families were used to search for homologous sequences in the *Ascomycota* (Supplementary Tables S6, S7). However, homologous sequences were only found in class *Dothideomycetes* and were restricted to the order Pleosporales, with two exceptions, RLG_*Dolly* and DTT_*Krilin*, found in *Pseudocercospora fijiensis* (syn. *Mycosphaerella fijiensis*) a member of order Capnodiales (Figure 4; Supplementary Data 2).

These two might have been present in the dothideomycete lineage for at least 300 MYA. This assumption is consistent with previous data that suggested RLG_*Dolly* as the most ancient TE present in the Lmb genome (Rouxel *et al.*, 2011). Element DTT_*Krilin* was conserved in most Pleosporales while RLG_*Dolly* had a more sporadic distribution in the Pleosporales, a trait that seems to be common to many retrotransposons. However it remains uncertain whether these two TE families invaded the fungal genomes before the Capnodiales-Pleosporales separation and then were widely lost in the Capnodiales or if their presence in *P. fijiensis* is incongruent with vertical inheritance and due to other phenomena such as horizontal transfer (HT). It should be noted that genome sampling still only covers a small number of the order level variation (comparable to divergences more than 100 MYA) within *Dothideomycetes*. The remaining TE families are likely to represent surges of genome invasion at different times before and after the divergence of the Pleosporales: (i) eighteen families were present in early diverging lineages of the analysed Pleosporales indicating a probable invasion date of the lineage before 90 MYA; (ii) thirteen families were present in the common ancestor of *Pleosporaceae* and *Leptosphaeriaceae*

indicating a probable invasion date of the lineage before 73 MYA; (iii) six families (including three unclassified repeats) were specific to the *L. maculans-L. biglobosa* species complex indicating an invasion date of the lineage before 22 MYA; (iv) fourteen families (including three unclassified repeats) were specific to either *L. maculans* or *L. biglobosa* isolates indicating an invasion date of the lineages before 11-5 MYA; (v) lastly, the remaining families, including forty unclassified repeats were specific to one or the other of the members of the *L. maculans-L. biglobosa* species complex (Figure 4; Supplementary Data 2). The Class II elements usually were better conserved amongst the *Pleosporales* than the Class I elements, which more often showed a patchy distribution and presence/absence patterns in terminal nodes or terminal branches of the phylogeny. In addition complete copies of DNA transposons were often maintained in all species in which they were present whilst only truncated copies or small-sized remnants were present for many Class I elements (Supplementary Data 2). For example, DTT_*Bulma* was present in all species of the Pleosporales investigated, but absent from Lbt and Lmb (Supplementary Data 2). Similarly, DTT_*Molly* (previously identified in *Phaeosphaeria nodorum* (Hane & Oliver, 2010)) was present in all species of Pleosporales except the *L. maculans* lineage (Figure 4). In many cases, two closely related species were missing one well-conserved TE family, further exemplifying the highly dynamic gain and loss patterns of TEs. This was the case for the DTT_*Finwe* super-family, absent from Lbt, but present in other members of the species complex, or DTA_*Kami* absent from *C. sativus*, but present in *C. heterostrophus* (Supplementary Data 2). Lastly, some TE families (*e.g.* DTT_*Yamcha* in Lml and *C. heterostrophus*, DTF_*Elwe* in Lmb and *P. tritici-repentis*, RLG_*Piccolo* in Lbb and *A. brassicicola* or RLG_*Shu* in Lml and *S. turcica*) were only found in distantly related species of the *Leptosphaeriaceae* and *Pleosporaceae*. This indicates multiple losses, independent invasion of the genomes or horizontal gene transfer events (Supplementary Data 2). In the cases of DTT_*Yamcha* and DTF_*Elwe*, the existence of truncated copies in other Pleosporales would suggest secondary losses.

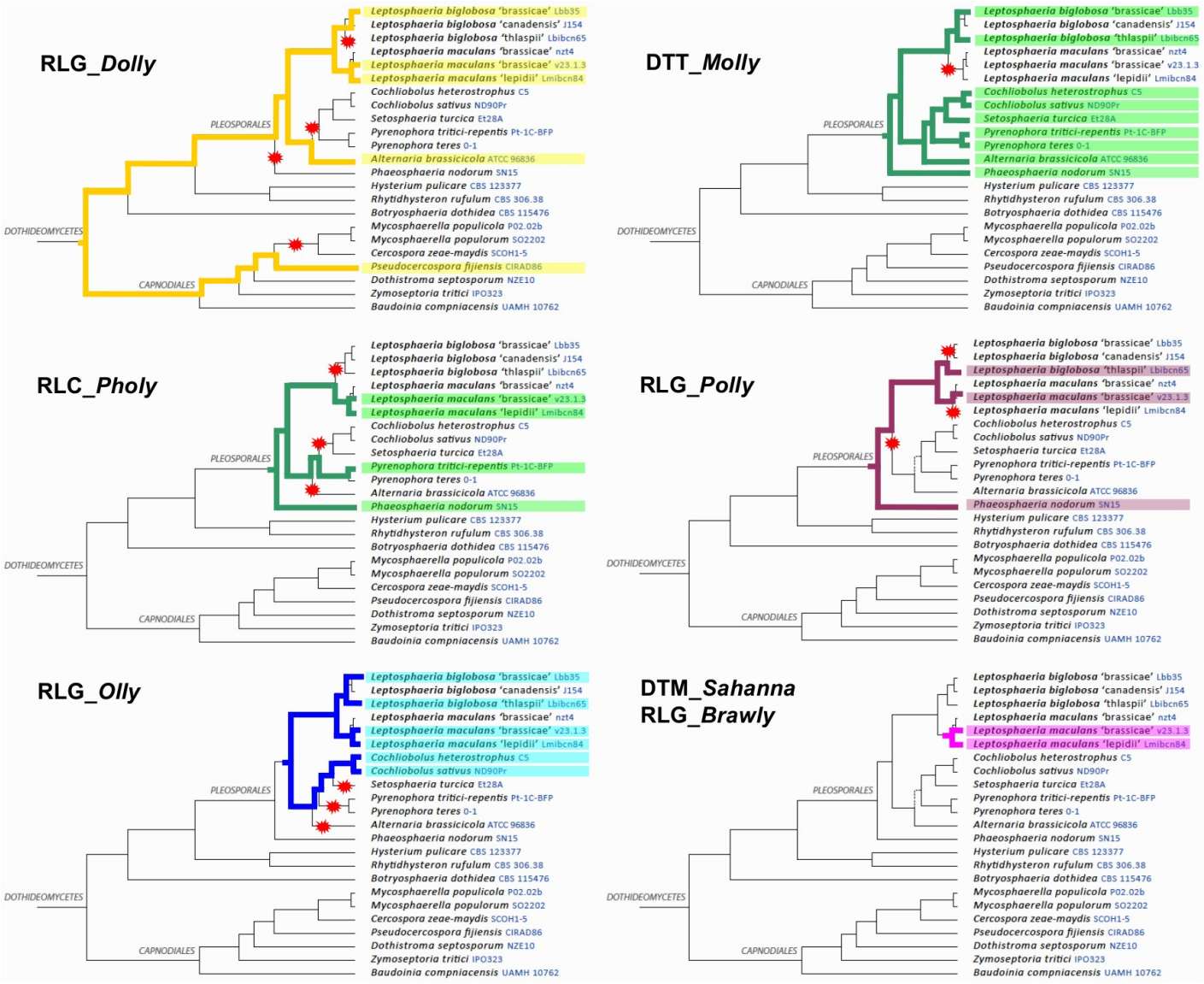


Figure 4. Distribution of selected Transposable Elements (TEs) within the dothideomycete phylogeny. The phylogeny is a simplified version of that in Figure 1. TE families are classified according to Wicker *et al.* (2007) with Rxx corresponding to retrotransposons and Dxx corresponding to DNA transposons. The species in which the family has been identified are highlighted along with the corresponding branches. The red stars indicate these branches of the phylogeny in which the TE family is postulated to have been lost.

In Rouxel *et al.* (2011), we postulated that massive invasion of the genome of Lmb was recent (4-20 MYA), after divergence from *P. nodorum* and could account for genome size expansion in this species. Fifty one families, mainly corresponding to non-categorized repeats, but also encompassing RLC_*Gohan* in Lbb and DTM_*Ingwe* in Lmb were specific to the species complex. Interestingly the number of these families varied from *ca.* 10 for the TE-poor genomes to 18 for the TE-rich Lmb genome, thus substantiating recent invasion of the Lmb genome by families of TE absent from other members of the species complex. DTM_*Sahana*, present only in Lmb and Lml had 195 copies covering 783 kb in the reference Lmb isolate *vs.* 17 incomplete copies covering 21 kb in Lml. Similarly, and consistent with our postulate, the Lmb-specific TE families DTx_*Gimli* (279 copies covering 113 kb), RLx_*Ayoly* (164 copies covering 400kb) and RLG_*Rolly* (594 copies covering 2.2 Mb of the genome of Lmb v23.1.3) (Rouxel *et al.*, 2011) (Figure 4; Supplementary Figures S1, S4) clearly indicate novel, recent and massive invasion of the Lmb genome. However, when comparing the insertion sites between the two Lmb isolates, 61.5% of the TE families or mosaics of TEs showed a conserved insertion pattern between v23.1.3 and WA74 indicating transposition events before or at the onset of intraspecies divergence. In contrast, DTM_*Sahana* showed highly diverse insertion sites with only 12% of identical locations in the two Lmb genomes (Supplementary Table S8). This strongly suggests, as previously postulated from phylogenetic analysis (Rouxel *et al.*, 2011) that DTM_*Sahana* is one of the most recent genome invaders and that waves of transposition activity took place after the separation between Lmb and Lml.

This simple picture indicating coincidence between genome invasion by new TE families and speciation becomes less straightforward when considering the three other retrotransposons that account for most of the v23.1.3 genome expansion along with RLG_*Rolly*: RLC_*Pholy* (1020 copies covering 3.1 Mb), RLG_*Polly* (1014 copies covering 3 Mb) and RLG_*Olly* (1085 copies covering over 3Mb) (Rouxel *et al.*, 2011). These retrotransposons are occasionally found outside of the *Leptosphaeriaceae.* RLG_*Polly* and RLC_*Pholy* were present in the distantly related *P. nodorum* but showed an extremely patchy distribution within the Pleosporales, with RLG_*Polly* being completely absent from the *Pleosporaceae* and RLC_*Pholy* only present in *C. sativus* (Figure 4; Supplementary Data 2).

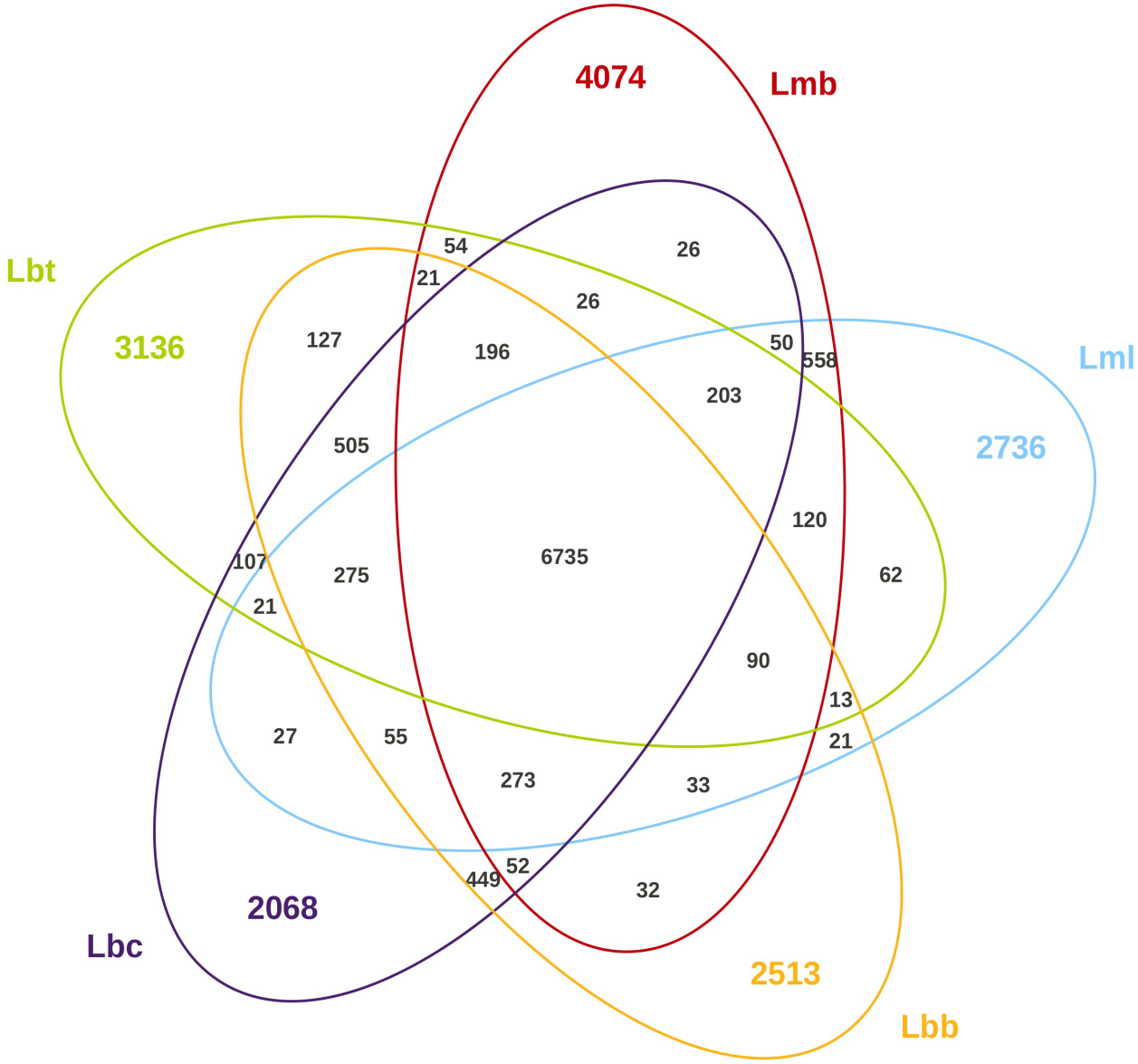


Figure 5. Conservation of protein-encoding genes in the *Leptosphaeria maculans-L. biglobosa* species complex. Lmb, *L. maculans* 'brassicae'; Lml, *L. maculans* 'lepidii'; Lbb, *L. biglobosa* 'brassicae'; Lbt, *L. biglobosa* 'thlaspii'; Lbc, *L. biglobosa* 'canadensis'.

RLG_*Olly* was present in all species of the *L. maculans-L. biglobosa* species complex and all *Pleosporaceae* which may indicate a more recent invasion than by RLC_*Pholy* and RLG_*Polly* but still well before differentiation of the *L. maculans-L. biglobosa* species complex. However, as in the previous case, RLG_*Olly* was only present in species of the *Cochliobolus/Setosphaeria* clade (Figure 4; Supplementary Data 2). This low conservation during evolution may indicate common elimination from most lineages as it may also represent series of invasions of the lineages at different times. Again a species-specific expansion of these three TEs was evident in Lmb compared to the other members of the species complex. For example, 70 incomplete copies of RLG_*Olly* covered only 43 kb in Lml, only nine incomplete copies of RLC_*Pholy* covered 11 kb in Lml while it was completely absent from *L. biglobosa* isolates. For all these three families, the homologies and/or coverage of homologous sequences were low in other species. Alternatively, as exemplified by RLG_*Olly*, variants of the family were in other members of the species complex which had much higher homologies to the corresponding family in other dothideomycete species that the Lmb variant had (Supplementary Data 2).

Chromosomal inversions between Lmb and Lml were bordered by TEs in 70% of cases (Supplementary Table S5). Classification of the TEs in families and phylogenetic analyses further indicated that these TEs were mainly (67%) *L. maculans*- and even Lmb-specific TEs, such as DTM_*Sahana*, DTM_*Ingwe,* DTx_*Gimli,* or RLx_*Ayoly* (Supplementary Table S5).

Gene conservation

Of the 57,964 proteins predicted in the *Leptosphaeria* species complex, 48,013 of them were grouped into 10,916 families using orthoMCL (Li *et al.*, 2003). For the initial global comparison of gene conservation between species, protein families containing paralogs present in only one species and families containing more sequences than species (protein showing expansions in one or more species of the species complex) were excluded. A total of 10,131 families was obtained encompassing 75% of the repertoire of predicted proteins. Of these, 6735 families with one highly conserved single-copy gene per genome were established as the core proteome of the species complex (Figure 5). Consistent with phylogeny, the core proteome was similar between the different *L. biglobosa* isolates (7962 to 8068 common families) and between Lmb and Lml (8096 proteins) and higher than between *L. biglobosa* and *L. maculans* isolates (7435-7639 families) (Figure 5).

Table 4. Predicted genes conservation between the genomes of isolates of the *L. maculans-L. biglobosa* species complex.

		L. maculans 'brassicae'	*L. maculans* 'lepidii'	*L. biglobosa* 'brassicae'	*L. biglobosa* 'canadensis'	*L. biglobosa* 'thlaspii'
Sequences clusterised using orthoMCL	No. of orthologs in all species (Core genome)	6735	6735	6735	6735	6735
	No. of orthologs in at least one other species	1734	1801	2142	2265	1820
	sub-total :	**8469**	**8536**	**8877**	**9000**	**8555**
Unclusterised sequences	No. of sequences absent in all other species	1337	734	595	536	1199
	No. of sequences in at least one other species	2123	1724	1738	1372	1601
	No. of sequences with unresolved absence or presence in other species	614	278	180	160	336
	sub-total :	**4074**	**2736**	**2513**	**2068**	**3136**
	Total :	**12543**	**11272**	**11390**	**11068**	**11691**
	No. of pseudogenes in unclusterised sequences [a]	1171	734	858	581	813

[a] Refers to sequence which appear at least once as pseudogenes in other species.

The greater divergence time between Lbt and the other *L. biglobosa* isolates than between Lbb and Lbc was not evident at this scale with a comparable core proteome when comparing pairwise *L. biglobosa* isolates (Figure 5).

To identify species-specific proteins, all the proteins that were not grouped into the 10,131 families were compared by BLAST at the nucleotide (BLASTn) and protein level (tBLASTn) against the other genomes and at the protein level (BLASTp) against the set of ungrouped proteins of the other species. The analysis allowed us to discriminate three classes of protein sequences: (i) species-specific sequences, (ii) sequences present in at least one other species and (iii) sequences for which presence or absence remained unresolved with our criteria (Table 4). In addition, the occurrence of pseudogenes was investigated when lack of matching predicted protein sequence was associated with BLASTn or tBLASTn hits (Table 4).

In average, each genome shared 10,400 proteins with (at least one of) the other ones, which represents 84.4-93.7% of the predicted proteins depending on the species. Usually, these sequences were conserved in organisms outside of the species complex since 74.3-85.5% of them have homologies in the NR database or harbour known protein domains.

The remaining species-specific proteins which represent 4.8-10.7% of the genomes (Table 4) had very few putative functions since 93.6-97.8% of them had no automated functional annotations or known protein domain. As a consequence, the predictive value of species-specific genes to explain species-specific biological or pathogenicity traits was not straightforward and now requires gene-by-gene annotation to uncover biological function. However, two categories of species-specific genes could be identified and are analysed in more details below: genes encoding putative effectors and cluster of genes encoding secondary metabolites.

Interestingly, the Lmb genome was more enriched in orphan sequences than all other members of the species complex, including both sequences absent in other species and sequences corresponding to pseudogenes in at least one of the other species (Table 4). Species-specific genes in the Lmb genomes were located in AT-isochores: 25.6% of the 620 genes of Lmb located in AT-isochores are species-specific, whereas only 9.8% of the genes located in GC-isochores were (data not shown).

Pathogenicity gene analysis

Candidate effectors

The number of Small Secreted Proteins (SSPs) amongst the whole predicted proteome of each genome was similar in all the sequenced *Leptosphaeria* genomes, ranging from 621 in Lbc to 737 in Lml. They represented *ca*. 6% of the predicted proteins and 60% of the predicted secretome of each genome. Their features (153 amino acids on average; 2.8 times as rich in cysteine residues as other proteins in the genomes) were very similar in all *Leptosphaeria* isolates (Table 5). Their encoding genes were scattered along the chromosomes and had a similar GC-content (~52%) not different from the other predicted genes of the genomes (Table 5). In contrast to what is observed for secondary metabolite biosynthetic genes and what is described in fungi such as *Ustilago maydis* (Kämper *et al.*, 2006), SSP-encoding genes were not organized in dedicated clusters in any member of the species complex investigated.

Table 5. Characteristics of SSP genes and proteins between isolates of the *L. maculans*-*L. biglobosa* species complex.

	L. maculans **'brassicae'**	***L. maculans*** **'lepidii'**	***L. biglobosa*** **'brassicae'**	***L. biglobosa*** **'canadensis'**	***L. biglobosa*** **'thlaspii'**
No. of SSPs	651	737	665	621	676
SSPs mean size (aa)	**155.9**	**144.8**	**153.8**	**157.4**	**152.9**
All predicted proteins mean size (aa)	416.5	427.4	434.6	464.2	421.0
%Cys in SSPs	**3.1**	**2.7**	**2.6**	**2.9**	**2.8**
%Cys in all predicted proteins	1.7	1.5	1.4	1.4	1.5
%GC in SSP genes	**53**	**51**	**53**	**52**	**53**
%GC in all predicted genes	53	53	54	53	54
RIP index in SSP genes [a]	**1.22**	**1.23**	**1.15**	**1.04**	**1.16**
RIP index in all predicted genes	1.04	1.02	1.01	0.89	1.01

[a] RIP index=TpA/ApT

The repertoire of SSP-encoding genes appeared to be in part extremely plastic and its conservation was consistent with phylogenetic distance with 73.8% of the SSP-encoding genes of v23.1.3 conserved in the other Lmb isolate, WA74, 54.2% conserved in Lml, and 42.5 to 44.0% conserved in *L. biglobosa* isolates (data not shown). For each genome, the conservation of SSP-encoding genes was similar with around 30% of these in the core proteome, 40-50% present in at least one other member of the species complex and 10-25% of species-specific sequences (Supplementary Table S9). The finding that 40-50% of SSP-encoding genes were present in at least one other species was mostly consistent with the phylogeny since more than 42% of this category of SSP gene in Lmb are only present in Lml (data not shown). However, the remaining cases corresponded to all possible presence/absence patterns within the species complex with, for example, 34.7% of the cases in which the v23.1.3 orthologs were missing in only one other member of the species complex (data not shown), possibly indicating secondary loss of an anciently present gene. In the core proteome, 20% of SSPs have a predicted function, and this may relate to factors linked with pathogenicity such as carbohydrate-degrading enzymes (CAZymes). However, the greater the specificity, the lesser obvious was the potential function of SSP-encoding genes: only *ca.* 5% of SSP sequences present in at least two members of the complex had a predicted function *vs.* none in the species-specific SSP sets.

Species-specific orphan SSP sequences were twice as numerous in Lml, Lbt and, to a lesser extent Lmb than they were in Lbc or Lbt (Supplementary Table 9). The most likely explanation for that would reside in the greatest proximity between Lbc and Lbb while other members of the complex would be more divergent (Figure 1). While a similar number of species-specific SSP-encoding genes was found in Lmb, Lml and Lbt, Lmb showed a specific pattern of genome location for these genes since 41% of the SSP-encoding genes in AT-isochores were specific to Lmb against 14% for the SSP-encoding genes in GC-isochores. Such a genome location was only occasionally found for other members of the species complex.

SSP-encoding genes did not generally occur as multigene families in the *L. maculans-L. biglobosa* complex, but few of them (less than ten) had paralogs, usually conserved in all genomes. They correspond to proteins usually belonging to multigenic families such as CAZymes and are mainly located in GC-isochores in the Lmb genome. Only one effector gene with avirulence activity had a paralog in the Lmb genome: the *AvrLm4-7* SSP-encoding gene located within an AT-isochore had a paralog located 30 kb

upstream on the same chromosome.

Avirulence effectors

Six SSP genes conferring an avirulence phenotype on a series of genotypes of *Brassica* spp., *i.e.* interacting in a gene-for-gene fashion with the plant surveillance machinery, have been cloned in Lmb. All are located in AT-isochores and BLAST hits usually show a high level of diversification compared to their closest relatives (rarely above 30% sequence identity at the amino acid level), while usually maintaining a specific pattern of cysteine spacing (Supplementary Figure S6). They mostly show a patchy distribution along the phylogenies or are specific of a few members of the species complex.

AvrLmMex (A. Degrave, unpublished data) and *AvrLm11* (Balesdent *et al.*, 2013) were specific to the *Leptosphaeriacae*, with AvrLmMex having homologs in every member of the species complex except Lml. In all cases the identity at the protein level was *ca.* 35% over 98% of the protein length (Supplementary Table S10). The AvrLm11 encoding gene is located on a conditionally dispensable chromosome (CDC) present in some isolates of Lmb and absent from other isolates of the species complex (Balesdent *et al.*, 2013). This CDC contains 35 additional genes with no predicted and with no orthologs in Lml or Lbc. AvrLm11, however, showed homologies with a SSP predicted in the Lbt genome (Lb_ibcn65_P001030, Supplementary Table S10) while three additional genes of the CDC also had orthologs in the genome of Lbt (Lb_ibcn65_P009771, Lb_ibcn65_P009770, and Lb_ibcn65_P009768). These three genes were grouped in the Lbt genome at a different location to that of gene Lb_ibcn65_P001030.

Sequences related to AvrLm1 (Gout *et al.*, 2006) and AvrLm4-7 (Parlange *et al.*, 2009) were in a few species of class *Dothideomycetes* only. An AvrLm1 homolog was found in the set of predicted proteins of Lbt but not in other members of the species complex (Lb_ibcn65_P011530, Supplementary Table S10). Interestingly, the orthologous gene was also located in a large but poorly assembled AT-rich region, devoid of other predicted genes. Both AvrLm1 and Lb_ibcn65_P011530 showed homology with two SSPs, with unknown function from the Pleosporales species *Pyrenophora teres f. teres* (Supplementary Table S10).

As described above, *AvrLm4-7* had a paralog in the Lmb genomes (*LmCDS2*, 65.5% of identity at the nucleotide level) that was located 30 kb away at the border between an AT- and a GC-isochore (F. Parlange, unpublished data). While AvrLm4-7 or LmCDS2 had SSP homologs in all members of the species complex (Supplementary Table

S10) none were duplicated in Lml or *L. biglobosa* meaning the orthology relationships are difficult to sort out. Outside of the species complex, AvrLm4-7 or *LmCDS2* had homologies with two SSPs with no function of the botryosphaeriales species *Macrophomina phaseolina* (Supplementary Table S10).

In contrast to the previous examples, AvrLm6 (Fudal *et al.*, 2007) and Lema_P086540.1 (conferring avirulence towards a mustard species; A. Degrave, unpublished data) had homologs outside of the *Dothideomycetes*, but only in a few species of class *Sordariomycetes*. Both AvrLm6 and Lema_P086540.1 were homologous to SSPs produced by Lml and Lbt (Supplementary Table S10). While it did not match to other proteins of the *Dothideomycetes*, AvrLm6 also matched two SSPs in the *Glomerellales* species *Colletotrichum gloeosporioides* and *Colletotrichum higginsianum* (Supplementary Table S10). Lema_P086540.1 had homologs in other dothideomycete species, with one SSP of *C. heterostrophus*, but it also had homologs in two sordariomycetes species *C. gloeosporioides* and *Fusarium oxysporum* in which it matched with two proteins (Supplementary Table S10). One of them is not secreted and the other one correspond to the SSP SIX5, which play an important role during infection of tomato by *F. oxysporum* f. sp. *lycopersici* (Lievens *et al.*, 2009). All these proteins and their homologs display a good conservation of the cysteine spacing (Supplementary Figure S6). Another candidate effector had a similarly patchy phylogenetic distribution: Lema_uP022890.1 is homologous to SSPs in every member of the species complex and homologies were found with one SSP of the dothideomycete *M. phaseolina* and three SSPs of the sordariomycete *C. gloeosporioides*.

Secondary metabolite gene clusters

Similar to genes encoding effectors, secondary metabolite biosynthetic genes such as Non-ribosomal Peptide Synthase (NPS) and Polyketide Synthase (PKS) genes showed examples of extreme specificity of occurrence and complete conservation within the species complex. Seventeen NPS genes were identified across the five species (Figure 6a; Supplementary Table S11) and most (10) were shared by all five species, but some were unique to particular species or found only in the *L. biglobosa* or the *L. maculans* clade. Three genes, NPS13, NPS14, NPS15 were absent from the *L. maculans* clade. NPS14 and NPS15 were identified only in Lbt, while NPS13 was present in Lbc and Lbb. NPS8 was only identified in Lmb.

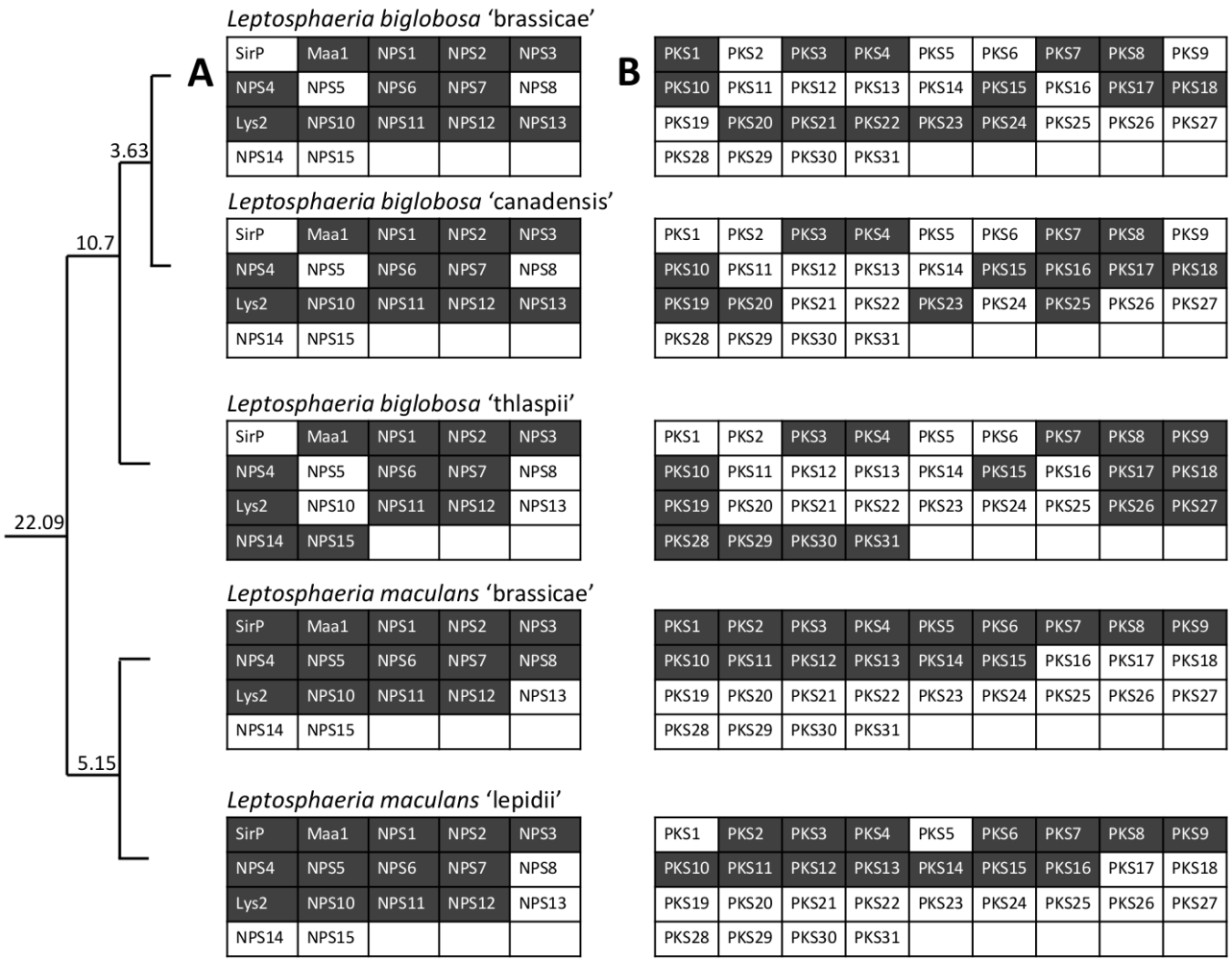


Figure 6. Conservation of secondary metabolite gene clusters in the *Leptosphaeria maculans-L. biglobosa* evolutionary series. (**a**) Non-ribosomal peptide synthases (NPS); (**b**) Polyketide synthase (PKS) genes. The colour of shading indicates those genes present (dark shading) or absent (white) in the corresponding species. The phylogenetic tree on the left is a simplified version of that in Figure 1, and the calculated divergence times (in MYA) are indicated.

SirP, the NPS involved in the production of the epipolythiodioxopiperazine toxin, sirodesmin PL (Gardiner *et al.*, 2004), was only present in Lmb and Lml as was NPS5. This latter gene had 37% sequence similarity to a putative aminoadipate semialdehyde dehydrogenase from the sordariomycete insect pathogen *Metarhizium acridum.*

In summary *L. maculans* species shared all but one NPS. Lbb and Lbc had similar homologs, and Lbt had an intermediate pattern of NPS homologs. Thus the sequence similarity of NPS genes was generally consistent with the phylogeny of the 'species' as described by Mendes-Pereira *et al.* (2003) and Voigt *et al.* (2005). A few of these genes only were species-specific (NPS14 and NPS15 in Lbt, NPS13 in Lbb and Lbc, SirP and NPS5 in Lmb and Lml, NPS8 in Lmb) and may represent specific adaptation or newborn pathogenicity determinant as illustrated by NPS8. NPS15 had close relatives in the distantly related *Dothideomycetes P. nodorum* and *M. populorum*, NPS13 in *C. sativus* (syn. *Bipolaris sorokiniana)*, and NPS5 in *P. teres f. teres* and *P. nodorum* (Supplementary Data 3). In contrast, SirP, NPS8 and NPS14 did not have any close relatives in other *Dothideomycetes*, and all other closely related sequences of NPS15, NPS13 and NPS5 were in distantly related fungal species outside of the *Dothideomycetes* (Supplementary Data 3). In the opposite, some NPS conserved in all species of the species complex had no close relatives in other dothideomycete fungi (examples of NPS1 and NPS11), or even were not found at all in other fungal species (example of NPS3) (Supplementary Data 3).

Unlike the NPS genes, there was a high degree of diversity in number and types of PKS genes identified across the five members of the species complex (Figure 6b; Supplementary Table S12) and the complements of PKS genes did not follow the expected phylogeny of the species. In addition, and except for the four cases mentioned below, homologs of PKSs in the species complex were rare or absent in other *Dothideomycetes* and often had closest relatives in *Sordariomycetes* such as *Fusarium* spp. or *Colletotrichum* spp., or in *Aspergilli* (Supplementary Data 3). A total of 31 PKS genes were identified; of which only six were found in all species. Four of these six are largely conserved in the dothideomycete phylogeny, while the two others, PKS3 and PKS8 are essentially found in sordariomycetes species.

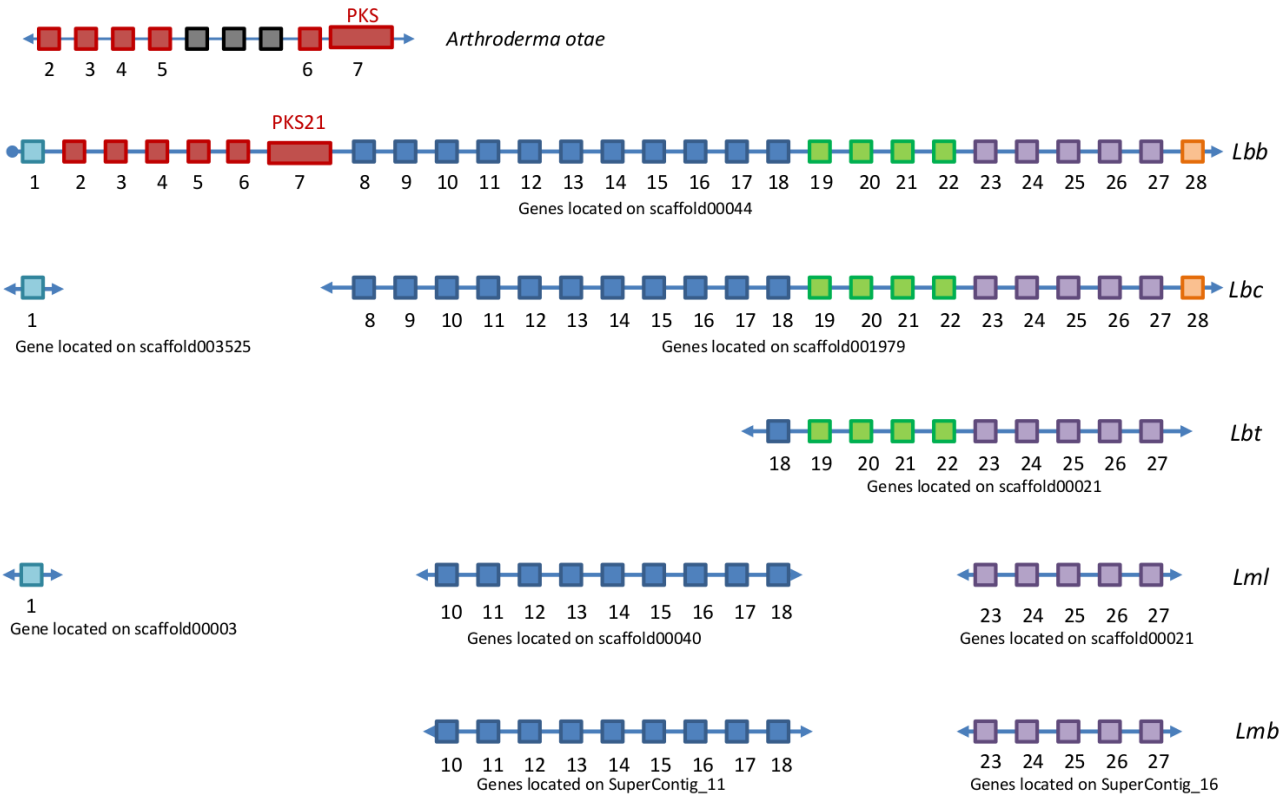


Figure 7. Organisation of the genes surrounding the polyketide synthase, PKS21, of *L. biglobosa* 'brassicae'. The PKS and upstream genes show > 88% identity with genes from *Anthroderma otae*. Each square represents a single gene (not to scale). Lmb, *L. maculans* 'brassicae'; Lml, *L. maculans* 'lepidii'; Lbb, *L. biglobosa* 'brassicae'; Lbt, *L. biglobosa* 'thlaspii'; Lbc, *L. biglobosa* 'canadensis'. Genes that are present in different species are highlighted in the same colour. Ends of contigs are represented with a circle whilst arrows indicate that the contig continues in that species. Genes common to Lbb and *A.otae* have the following predicted functions; gene 2 has a retinol dehydrogenase (short chain dehydrogenase) motif; genes 3 and 5 have a methyltransferase motif, gene 4 has no conserved domains and gene 6 is a dimethyl allyl transferase.

Lmb and Lml had a very similar complement of PKS genes with 15 PKS in Lmb and only two missing in Lml: In Lml PKS1 was truncated and PKS5 was absent. Twenty-three different PKSs were found in *L. biglobosa* species and many of these were species-specific, with three only found in Lbb, one only found in Lbc and six only found in Lbt, clearly diverging from all other members of the species complex in terms of PKS complement (Figure 6b). Excluding these six PKS genes common to all members of the species complex, Lbc and Lbb only had five PKS genes in common while Lbt had only three in common with Lbc and two in common with Lbb. Intriguingly, two genes were common to one *L. biglobosa* (but absent from the other *L. biglobosa* species) and one (of the two) *L. maculans* species. For example, PKS1 was present between Lbb and Lmb, and PKS9 was in Lmb, Lml and Lbt (Figure 6). In both cases, most if not all homologous genes were found in fungal species outside of the *Dothideomycetes* (Supplementary Data 3).

The PKS-NPS gene clusters conserved between Lmb-Lml and between Lbb-Lbc displayed a high level of synteny: of the 16 clusters conserved in all the sequenced members of the species complex, eight were embedded in highly syntenic regions. This included genes widely conserved in the *Dothideomycetes* (PKS4, PKS7, PKS10, NPS4, NPS6, and Lys2), but also NPS1 and NPS11 conserved in the species complex but absent from other dothideomycete species. In eight additional cases, secondary metabolite clusters showed microsynteny in the *L. maculans* isolates while the cluster was reorganised in Lbb and Lbc. Interestingly, for these cases, Lbt showed an intermediate genomic pattern with PKS3, PKS15, NPS2 and NPS3 clusters being syntenic with those in Lmb and Lml, and PKS8, PKS9, NPS7 and Maa12 clusters showing microsynteny with Lbb and Lbc (data not shown).

PKS21 was only present in Lbb. Furthermore it it had > 90% identity at both the nucleotide and protein level with a PKS from the eurotiomycete, *Arthroderma otae*, which is a dermatophyte (Figure 7). This PKS is a hybrid with an incomplete NPS module that had a single adenylation and thiolation domains, but lacked a condensation domain. The five upstream genes showed high sequence similarity (> 88% amino acid identity and 83% nucleotide identity) to genes upstream of the *A. otae* PKS, although this microsynteny was interrupted by the presence of three genes in *A. otae*. Several of the homologous genes had predicted functions consistent with biosynthesis of a secondary metabolite.

Table 6. The PKS21 gene cluster of *Leptosphaeria biglobosa* 'brassicae' and homologs in *Arthroderma otae*.

L. biglobosa 'brassicae'	Location	Domains/Motif	*A. otae* gene	Location	Identity (%)	*L. biglobosa* 'canadensis'	Location	Identity (%)	Other Dermatophyte matches	Identity (%)
Lbb_b35_P010265	Sc44	FtsJ-like methyltransferase	MYCG_01736	Sc2	42	Lb_j154_P004832	Node_3525	93	*M. gypseum*	42
Lbb_b35_P010266	Sc44	Retinol dehydrogenase	MYCG_04899	Sc4	88	Lb_j154_P005074	Node_3836	51	*T. tonsurans*	47
Lbb_b35_P010267	Sc44	Methyltransferase	MYCG_04900	Sc4	93	Lb_j154_P005876	Node_5114	38	*M. gypseum*	29
Lbb_b35_P010268	Sc44	No conserved domains	MYCG_04901	Sc4	97	No match			*M. gypseum*	47
Lbb_b35_P010269	Sc44	Methyltransferase	MYCG_04902	Sc4	91	No match			*No match*	
No match		C4-methylsterol; Fatty acid hydroxylase superfamily/sterol desaturase	MYCG_04903	Sc4		No match			*M. gypseum*	84
No match		No conserved domains	MYCG_04904	Sc4		No match			*No other matches*	
No match		No conserved domains	MYCG_04905	Sc4		No match			*No other matches*	
Lbb_b35_P010270	Sc44	DMAT	MYCG_04906	Sc4	89	No match			*M. gypseum*	36
Lbb_b35_P010271 + Lbb_b35_P010272	Sc44	Polyketide synthase	MYCG_04907	Sc4	94	Lb_j154_P007539	Node_10192	36	*T. tonsurans*	29
Lbb_b35_P010273	Sc44	Enoyl reductase of polyketide synthase	MYCG_07786	Sc7	34	Lb_j154_P003162	Node_1929	79	*M. gypseum*	34
Lbb_b35_P010275 + Lbb_b35_P010276	Sc44	CypX superfamily	MYCG_02371	Sc2	42	Lb_j154_P003162	Node_1929	87	*A. benhamiae*	44
Lbb_b35_P010277	Sc44	CypX superfamily	MYCG_03969	Sc3	51	Lb_j154_P003173	Node_1929	90	*T. equinum*	39
Lbb_b35_P010278	Sc44	MFS superfamily	MYCG_03971	Sc3	56	Lb_j154_P003172	Node_1929	97	*T. rubrum*	56

This gene cluster is lacking in six other dermatophytes sequenced by the Broad Institute (Dermatophyte comparative sequencing project, Broad Institute, http://www.broadinsitute.org) (Table 6). Genes downstream of PKS21 in Lbb were also present in genomes of some of the other *Leptosphaeria* species (Figure 7). These data strongly suggest that Lbb has gained this gene cluster through horizontal gene transfer (HGT).

Dimethyl allyl transferases (DMATS) add prenyl groups to aromatic moieties such as tryptophan or tyrosine. A DMATS gene, SirD, is located in the sirodesmin gene cluster of *L. maculans* species, and prenylates a tyrosine in the first step in sirodesmin biosynthesis (Kremer & Li, 2010). No other DMATS were present in the two *L. maculans* genomes, but one was present in Lbb in the PKS21 gene cluster described above. This gene has 35% amino acid identity with the homolog in the sirodesmin gene cluster, and >90% amino acid identity with the homolog in the PKS cluster of *A. otae*.

TEs and pathogenicity genes

Of the 620 genes located in AT-isochores in the Lmb genome, 148 genes (24%) were surrounded by TEs (presence of TEs on both 5' and 3' side of the genes), whereas the others were located at the transitional regions between AT- and GC-isochores (presence of TEs either on 5' or 3' side of the genes). A large proportion of the former (64 genes) were specific to Lmb, with no homolog in other genome either as a complete sequence or as a pseudogene. In these cases and that of genes that have probable orthologs that were not predicted in the other species (39 genes), the incidence of TEs on translocation or HGT could not be deduced. The remaining 45 genes were firstly mined to exclude those for which the assembly in one or the other genome did not allow to conclude on conservation of microsynteny (6 cases). We then investigated whether the gene of interest and its closest flanking genes were syntenic in the other genomes, or if orthologs of the flanking genes were syntenic in the other genomes but the ortholog of the gene of interest is not (located in another genomic region or elsewhere in the same genomic region). Among the remaining 39 genes, 26 did not encode SSPs. Most of them (80.8%) showed a conserved microsynteny in the other genomes of the species complex suggesting the TE expansion did not influence gene order or orientation in these cases.

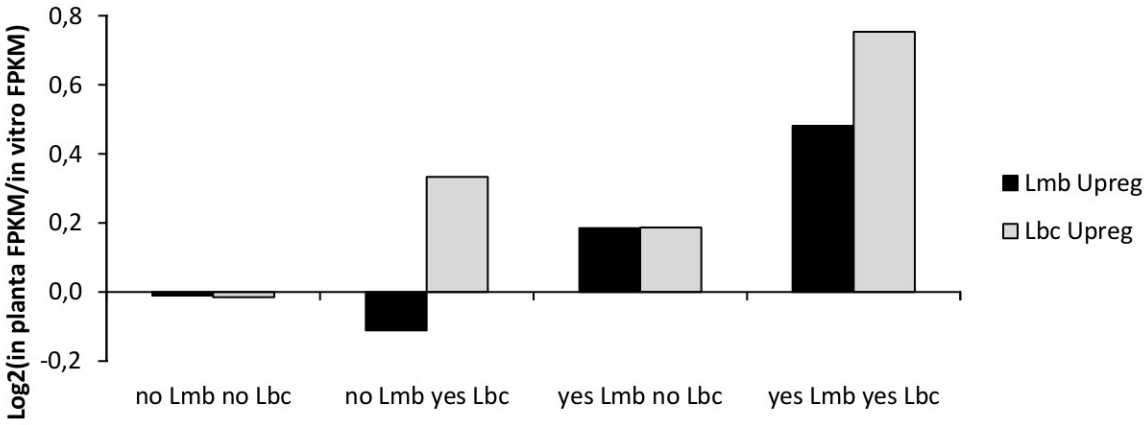


Figure 8. Effect of presence of repetitive element adjacent to promoter of orthologs on gene expression *in planta* compared to in axenic culture. *In planta* upregulation ratio (Log2(*in planta* FPKM/*in vitro* FPKM) calculated from RNA-seq data for orthologs in *L. maculans* 'brassicae' (Lmb) and *L. biglobosa* 'canadensis' (Lbc) during growth *in planta* and *in vitro*. Orthologs pairs were categorised according to the presence or absence of repeat elements adjacent to the promoter region of both genes. The distance to the repeat was not included in this analysis. Positive up-regulation ratios indicate higher expression during growth *in planta*. The categories were: no Lmb no Lbc (n=6959), no Lmb yes Lbc (n=60), yes Lmb no Lbc (n=382), yes Lmb yes Lbc (n=14).

The five remaining genes, including one phospholipase and two non-annotated genes, showed a break in microsynteny and were not colocalized with their flanking genes, suggesting they were translocated in other places of the genomes. Thirteen of the 39 genes were encoding SSPs. Of these, seven showed a conserved microsynteny pattern while six were translocated within the genome. Interestingly, five of the characterized avirulence genes (*AvrLm6*, *AvrLm4-7*, *AvrLm11*, *AvrLmMex* and Lema_P086540.1) were subjected to translocation, whereas the sixth one, *AvrLm1* could not be studied due to poor assembly of this region in Lbt where it is also likely to be hosted in an AT-rich genome environment.

In GC-isochores, there was no preferential association between TEs and secondary metabolite gene clusters: 11% had TEs in their vicinity (*i.e.* in a 2 kb maximum distance) which was not different what is observed for other genes in GC-isochores (average of 13% of genes with TEs in their immediate vicinity). Of the 35 genes encoding PKSs and NPSs in the Lmb genome, seven were located in AT-isochores. Globally, the high level of synteny of these gene clusters between Lmb and Lml showed that the TE invasion did not incidence the organisation of these regions. Interestingly, PKS5 and NPS8 were located within AT-isochores and surrounded by TEs.

The effect of proximity of repeat elements on gene expression was determined. The RNA-seq gene expression values for Lmb and Lbc orthologs were coupled with locations of repeat sequences. Excluding genes that had low levels of expression *in vitro* or *in planta*, 7415 ortholog pairs were identified. Orthologs pairs were grouped into four categories, on the basis of repeat presence in their promoter regions (Figure 8). An expression ratio was calculated from the log2 (*in planta* expression*/in vitro* expression) for each gene. Repetitive elements located adjacent to a gene promoter in both Lmb and Lbc caused general up-regulation during *in planta* growth compared to *in vitro* growth (Figure 8). This differential effect was more obvious when Lbc orthologs had a repeat element nearby and the Lmb ortholog did not: only the Lbc orthologs showed an *in planta* up-regulation pattern (Figure 8). The opposite situation, where repetitive elements were adjacent to Lmb orthologs but not to Lbc promoter orthologs, showed both genes were upregulated *in planta* (Figure 8). In this case, the extremely fragmented assembly of Lbc may have prevented accurate identification of repetitive elements.

Discussion

Divergence and speciation

Fungi are estimated to be represented by over 1.5 millions species. Their description and classification have been long based on macro- or microscopic criteria that were considered to be informative, such as morphology or reproduction style. The availability of genomic data for a large number of fungal species allowed the inference of new phylogenetic relationships between close or even unrelated taxa previously aggregated in the same genera, which can be exemplified by the *Mycosphaerella* genera and its new classification (Crous *et al*., 2009; Quaedvlieg *et al*., 2011). In this paper, we firstly reinvestigated speciation issues in the *L. maculans-L. biglobosa* species complex based on sequence data analyses. The estimated divergence times between the terminal clades (Lmb-Lml: 5.1 MYA and Lbb-Lbc: 3.6 MYA) are consistent with speciation dates between closely related species in other genera (*Cochliobolus sativus / C. heterostrophus:* 4.3 MYA, *Pyrenophora tritici-repentis / P. teres*: 7.5 MYA, *Mycosphaerella populicola / M. populorum*: 4.7 MYA, *Coccidioides posadasii / C. immitis* 5.1 MYA (Sharpton *et al.*, 2009)) (Supplementary Fig S2) and strongly indicate all *Leptosphaeria* isolates analysed here belong to different species. This is further substantiated by analysis of genome alignment, SNP density, and chromosomal organization. Pairwise genome alignments at the nucleotide level allowed the identification of SNPs between the members of the species complex. The amount of SNPs between each lineage of the species complex is around 1.5 millions. These figures are in the same order of magnitude than those observed between different species of a genus closely related to *Leptosphaeria*, *Cochliobolus* (Condon *et al*., 2013). Analysis of chromosomal organization between the different isolates supports the mesosynteny postulate and illustrates the first steps toward speciation. Mesosynteny, *i.e*., the conservation within chromosomes of gene content but not order or orientation, was recently postulated to be a mode of chromosomal evolution specific to fungi, and more specifically of *Dothideomycetes* (Hane *et al.*, 2011). Comparative analyses of 18 dothideomycete species, and a simulation-based approach indicated that serial random inversions within the chromosome lead with time to intense reshuffling of gene order within homologous chromosomes from one species to a distantly related one (Ohm *et al.*, 2012). Focusing on relatively short divergence times, our analyses are in perfect accordance with this modelling-based hypothesis. The most illuminating data comes from the comparison of chromosome organization between Lmb and Lml, showing that the number of genome

rearrangements is much fewer than what has been reported by Ohm *et al.* (2012), *e.g.* 30 inversions in a single scaffold between *C. sativus* and *C. heterostrophus*, and indeed may represent a first step towards mesosynteny. Of special interest, both divergence time estimates (*ca.* 22 MYA) and the intensity of intrachromosomal inversion between *L. maculans* and *L. biglobosa* relate to those observed between *C. heterostrophus* and *Setosphaeria turcica*, belonging to different genera (Ohm *et al.*, 2012). This may indicate that not only the different *Leptosphaeria* isolates belong to different species, but that *L. maculans* and *L. biglobosa* may indeed belong to two different genera.

Using population studies and phylogeography analyses, we recently postulated that *L. maculans* is a recently emerged species. The data obtained here both comfort and contradict this postulate. On the one hand the divergence time estimates between Lmb and Lml, and the postulated times of genome invasions by TEs indicate that the rise of the Lmb species occurred *ca.* 5.1 MYA. On the other hand the perfect genome alignment and low density of SNP between one Lmb isolate from Europe and one from Australia, lower from what is known for *Z. tritici*, suggested to have emerged 10,000 years ago (Stukenbrock *et al*. 2007) is in favor of a recently emerged species as proposed by Dilmaghani *et al.* (2012). One explanation would be that Lmb was submitted to a bottleneck, during speciation or more recently, due to adaptation to oilseed rape and that only a sub-population of a more ancient species was selected, leading to a rapid world-wide expansion of a species with a low level of polymorphism. By comparison, the *Z. tritici* species has been suggested to arise without being submitted to a bottleneck, thus maintaining a large part of its previous biodiversity (Stukenbrock *et al*., 2010).

TE invasion in Dothideomycetes: how fungal genomes control them and their incidence on speciation events

For the first time, a species complex of ascomycete phytopathogens has been used for evolutionary studies combining phylogenetic analyses of divergence time and analysis of TE occurrences in their genomes, and relating these data with previously established dating of transposition events (Rouxel *et al.*, 2011). The identification of the repeated elements in the genomes of the *L. maculans-L. biglobosa* species complex followed by the annotation of many TE families and the analysis of their distribution allowed us to get insights into the relationships between TEs and speciation events.

Consistently with preliminary data, the sequencing of several members of the species complex allowed us to confirm that the genome of Lmb contains a larger proportion of TEs than those of the other members of the species complex, but also those of other Pleosporales species. Such a massive invasion by TEs, along with the recent expansion of four families of LTR retrotransposons and the patterns of insertion of the DNA transposon DTM_*Sahana*, suggest that this is more likely the result of a recent TE expansion in the Lmb genome than the result of a loss of TEs in all other genomes.

In ascomycetes such as *Blumeria* spp., or in oomycetes (Raffaele & Kamoun, 2012; Wicker *et al*., 2013), heavily TE-invaded genomes suggest the absence of defense systems able to regulate the TE activity and that only the fitness cost linked with genome obesity will eventually counter-select species for which the fitness deficit is higher than the selective advantage conferred by enhanced genome plasticity. Amongst these examples, many related species show a heavily TE-invaded genome (Raffaele & Kamoun, 2012; Wicker *et al*., 2013). In Pleosporales, we observe a different trend since TE expansion was not observed in species (that we currently are aware of) other than Lmb. An intermediate situation is found in other fungi (*e.g. Cladosporium fulvum*, *Pseudocercospora fijiensis*) (De Wit *et al*., 2012; Ohm *et al*., 2012) that have close relatives species with TE-poor genomes while TE-invaded genomes are less uncommon than in the Pleosporales. Compared to the *Blumeria* or *Phytophthora* examples above, we can also notice that the expansion of TEs in the genome of Lmb remains limited (one third of the genome for Lmb *vs.* 64% of the genome for *B. graminis* and 74% for *P. infestans* (Raffaele & Kamoun, 2012)).

The analysis of TE distribution within the genomes of the sequenced members of the *L. maculans-L. biglobosa* species complex but also within the available genomes of the phylum *Ascomycota* showed that 34% of TE families identified in the genomes of the *L.maculans-L. biglobosa* species complex were also present outside of this complex, mainly in Pleosporales species. In the Lmb genome, the TE expansion that occurred *ca.* 4-5 MYA is mainly due to a burst of transposition of four LTR retrotransposon families (Rouxel *et al*., 2011) of which three have been present in the dothideomycete lineage for more than 90 million years. These data question about how TEs were controlled through time by Lmb and, in general, by ascomycete species.

RIP is likely the most efficient genome defense mechanism against TEs. Its current activity has been experimentally demonstrated in *L. maculans* (Idnurm & Howlett, 2003), and evidences of RIP have been reported in most ascomycete species (Clutterbuck,

2011). These observations are consistent with the postulate that RIP is an ancestral mechanism common to (at least) ascomycete fungi (Clutterbuck, 2011), with occasional patterns of secondary losses like observed in *Blumeria* spp. (Spanu *et al*., 2010). The common incidence of RIP and its current activity in Lmb is consistent with a limited invasion of Pleosporales genomes by TEs, but it does not allow us to understand how massive transposition activity could have taken place in Lmb for four LTR retrotransposon families, since RIP induces irreversible changes making TE reactivation impossible.

Actually our data indicate a “recent” transposition activity for TE families potentially present in the Pleosporales lineage for several tens of millions of years, suggesting that they were not inactivated by RIP or that they were reactivated. Three main hypotheses could explain these data: (i) loss and regain of RIP activity, but involvement of another epigenetic mechanism to control TE transposition events, (ii) biological traits preventing the set up of RIP, *i.e.* absence of sexual reproduction and (iii) recent events of HT.

(i) Bursts of transposition have been observed in many species and thought to be the result of a breakdown of the genome defenses (Rebollo *et al*. 2010). Thus, we can hypothesize that *ca.* 4-5 MYA a stress-inducing event was responsible for the breakdown of the control of TEs in the Lmb genome, resulting in the unleashing of transposition activity of these elements. However, inactivation by RIP should prevent the reviving of heavily mutated TEs. Other inactivation mechanisms have been described in fungi such as methylation induced premeiotically (MIP) in the ascomycete *Ascobolus immersus* (Goyon *et al*., 1996) and in a basidiomycete *Coprinus cinereus* (Irelan & Selker, 1996). MIP is an epigenetic mechanism which methylates but does not irreversibly mutate duplicated sequences. The fact that it is conserved between ascomycetes and basidiomycetes, raises the possibility that RIP evolved from MIP or from a similar silencing process (Galagan & Selker, 2004). One can then hypothesize that MIP was previously active and relieved under a stress-inducing event at the time of speciation, and then replaced by RIP as an irreversible control mechanism.

(ii) RIP is only active during meiosis, therefore prolonged growth in absence of sex would prevent RIP inactivation of TEs. Lmb is currently known to have an obligate sexual reproduction during its life cycle (Gout *et al*., 2006). However, some recent findings indicate that fully asexual populations exist either adapted to other crucifers such as cabbage or present in environments where short growth season of the host plant prevent the fungus from completing its sexual cycle. (Dilmaghani *et al*., 2012, 2013). Asexual growth could also be imposed within a founder population with only one mating type. One

can thus hypothesize that the rise of the Lmb species was followed by a long phase of asexual behaviour in order to propagate rapidly the newly-born species, without preventing expansion of TEs until sexual reproduction became prevalent with the need to diversify genotypes to adapt changing environmental conditions.

(iii) TEs are known to be transmitted by HT (Silva *et al*., 2004; Walsh *et al*., 2013) and vectors like arthropods are likely to be main actors allowing HT between animals species (Walsh *et al*., 2013). Our phylogenetic analyses did not allow us to establish whether such HT events occurred in the dothideomycete phylogeny, but multiple cases of patchy distribution do not contradict this hypothesis. In addition, even with RIP being active and sexual reproduction used in the fungal life cycle, one can imagine constant donation of TEs from a donor that was present over an extended period of many million years, such as a virus or an intracellular symbiotic bacteria (Loreto *et al*., 2008). The donor TEs would not be affected by RIP, as it is protected in the vector, but each new introduction would be RIPed and inactivated as it arrived. Long term addition could result in the establishment and the expansion of TE-rich regions. This hypothesis is extremely far-fetched as we currently cannot distinguish introduction and expansion of TEs from multiple introductions, due to extensive sequence degeneracy following RIP.

Evolutionary incidence of TEs is now established in many lineages and links between transposition bursts and speciation are observed without possibility to determine whether the burst was responsible for speciation or if it was a consequence of it (Rebollo *et al*., 2010). In this study, we observed two events that are estimated to have taken place during the same period of time than the burst of transposition: (i) the speciation event separating Lmb from Lml and (ii) the introduction in the Lmb genome of new TE families that took place after the divergence from Lml. The effects of invasion by species-specific TEs on speciation events are difficult to sort out since we observed introduction of new TE families several times during evolution within the dothideomycete lineage while bursts of transposition did not occur in any other Pleosporales species. However, the analysis of TE distribution was based only on TE identified in *Leptosphaeria* spp. which did not allow us to infer relationships between TEs and speciation in the other species that were not investigated for species-specific TEs. Moreover, as underlined before, the patchy distribution of some TE families, especially LTR retrotransposons, could be interpreted as an evidence of HT of TE between species (Daboussi & Capy, 2003) and thus interfere with the evolutionary history of these elements. Of special relevance for this question, we observed that in Lmb 70% of the intrachromosomal inversions were bordered by TEs

specific to the *L. maculans* clade. In the Ohm *et al.* (2012) paper, inversion breakpoints in *Dothideomycetes* genomes were suggested to be associated with Simple Sequence Repeats (SSR), a feature that was not apparent when comparing the Lmb and Lml genomes. These data may indicate a major role for TEs rather than SSRs in genome reshaping and reshuffling at the chromosomal level, and thus generating mesosynteny in chromosomes of *Dothideomycetes*. This first step towards mesosynteny clearly ascribes a role to species-specific TEs in this evolutionary mechanism that will eventually generate non-homologous chromosome sections and isolate part of the genome from meiotic recombination.

Can genome data explain biological and phytopathological specificities?

The *L. maculans-L. biglobosa* species complex provided us a biologically relevant model to investigate the evolution of pathogenicity and adaptation to host. While being morphologically very similar in all aspects (Petrie, 1969) and adapted to crucifers, all the species show very divergent host range, pathogenic abilities and infection strategies. Interestingly, two of the most closely related species, Lmb and Lml drastically differ in pathogenic abilities whereas more divergent species like Lmb and Lbb-Lbc have very similar host range and pathogenic behaviour.

Our knowledge regarding Lml and Lbt pathogenicity is limited since their natural hosts are crucifer weeds (*Lepidium* sp. and *T. arvense*, respectively) which are difficult to get, to grow or to infect in controlled conditions. However, recent data concur to indicate that Lml is nonpathogenic to *Brassica* spp. while Lbt is likely to have a broader host range, substantiated by the fact that it can be isolated from *Brassica* plants grown in the field such as *Brassica juncea* (Mendes-Pereira *et al*., 2003). In contrast, data on Lmb, Lbb and Lbc are more frequent and consistent since they infect *B. napus*, a major oilseed crop. On *B. napus*, Lmb, Lbb and Lbc have a very similar pathogenic behaviour with similar ability to develop a lengthy endophytic life in the tissues following the initial penetration step (Fitt *et al*., 2006a). The main difference consists in the ability to cause a stem basis canker which is a specificity of Lmb, while Lbc or Lbb cause upper stem lesions. In summary the different species differ by (i) their main host, (ii) their host range, broad for Lbt *vs.* narrow for all the others, and their ability or not (Lml) to infect *B. napus* (Lmb, Lbb, Lbc and Lbt), (iii) their ability to colonise the stem basis of *B. napus* plants and develop the late stem canker (Lmb *vs.* Lbb-Lbc), and (iv) their ability to develop very specific gene-for-gene type

interactions with *B. napus* (Lmb only).

Our initial postulate that the genome of Lmb drastically differed from that of other members of the species complex, in terms of size, TE content and chromosomal structure, was fully established in this paper. These data provided possible genomic bases to explain mechanisms leading to the rise and the evolution of new pathogenicity determinants and infection strategies. On these bases, our objective was to try to relate biology to genome data and infer evolutionary traits explaining differences in pathogenic abilities. This was investigated with two options: (i) the incidence of gene contents on adaptation to host and/or development of new pathogenic abilities and (ii) the incidence of genome structure and TEs.

Gene content data were very frustrating since most of genes which are species-specific, and thus putatively involved in species-specific life traits, did not have any attributed function. In contrast, genes with possible function in pathogenicity (CAZymes, MAP kinases, proteases, etc..) were conserved between the different members of the species complex. The study of secondary metabolites showed that PKS gene profiles were too divergent from one species to another to easily relate any of these with pathogenic abilities. Though, the possibility that the expanded PKS content in Lbt genome might be linked with its broad host range would deserve to be investigated. Regarding the content of putative effector-encoding genes, the number of those that were species-specific was comparable between members of the species complex.

Interestingly, Lml was the only species with a notably higher number of SSPs than the others. In *M. oryzae*, the importance of the effector repertoire in host range expansion has been established (Kang *et al*., 1995; Couch *et al*., 2005). For example, host range expansion from foxtail millet to rice is assumed to be due to the loss of the single effector AVR-Co39 which allowed the fungus to colonize rice (Couch *et al*., 2005). On this basis, one can question whether this expanded repertoire of effectors in Lml may contribute to its narrower host range compared to other members of the species complex.

While the amount of effector genes is comparable between genomes, their patterns of expression seems to differ between Lmb and Lbb-Lbc. Amongst the top 100 genes expressed at seven dpi, 45% encode SSPs in Lmb genome, whereas this proportion decreases to 21% for Lbb and 13% for Lbc. Of these expressed SSP-encoding genes, 30% represent species-specific genes in Lmb genome, when very few (or none) are species-specific genes in Lbb and Lbc. These figures suggest that species-specific SSPs in Lmb play a major role at the onset of the plant colonization and may contribute to the

development of the typical grayish-green tissue collapse on leaves, during which the plant machinery is not alerted. This contrasts with the dark symptoms expressed by Lbb or Lbc (Vincenot *et al.*, 2008), suggesting that the plant is alerted but cannot efficiently cope with the fungal expansion and tissue colonization. In addition, this larger amount of species-specific SSP-encoding genes is consistent with the existence of gene-for-gene interactions with *B. napus* described only for Lmb and with the need for a more plastic repertoire of effectors produced *in planta* to counteract plant major resistance genes.

NPS secondary metabolite cluster occurrence data were also difficult to exploit in terms of biological relevance. For example, the best known cluster in Lmb is the sirodesmin cluster, which produces a toxin acting as a virulence factor in stems of *B. napus* during colonization (Elliott *et al*. 2007). This cluster is absent from the *L. biglobosa* isolates, consistently with their inability to cause severe cankers in stems. In contrast, it is present in Lml which cannot colonize *B. napus*, suggesting the cluster was acquired before Lml or Lmb became strictly adapted to their host. The only difference between Lmb and Lml in NPS clusters regarded NPS8. This five-module NPS is thought to be involved in the production of the transiently produced depsipeptide, phomalide, which has been postulated to be a host-selective toxin (HST) since it elicits a hypersensitive response on leaves of *Brassica juncea*, but not on those of *B. napus* (Pedras *et al.*, 1993). These data are currently debated, and the biological role of phomalide has not been substantiated on other *Brassica* species or genotypes, but it would make sense that the only divergent NPS between Lmb and Lml is produced at the onset of tissue penetration and is postulated to be a HST.

Apart from being genome shapers, TEs are shown to be associated with effectors genes of fungal phytopathogens, and also known to be involved in inactivation of fungal avirulence genes of importance when the fungus is faced to the corresponding plant resistance (Fudal *et al*., 2005; Daverdin *et al*., 2012) and in deletion leading to loss of entire genomic regions encompassing avirulence genes. In Lmb, TE-rich isochores promote accelerated mutations for the (effector) genes included in these regions through RIP (Fudal *et al*., 2009; Daverdin *et al*., 2012) and create heterochromatic landscapes responsible for an epigenetic control of the effector gene expression (Soyer *et al*., under review). TEs are shown to be involved in adaptation of phytopathogens to new hosts by promoting gene duplication and translocation (*e.g.* Chuma *et al*., 2011). TEs can also directly contribute to the “birth” of new genes through mechanisms such as retroposition (also known as retroduplication) (Kaessman *et al*., 2009). Such processes appear to have

occurred frequently in the genome of *B. graminis*. This genome has experienced a massive proliferation of retrotransposons associated with loss of RIP, which has probably contributed to the extensive gene losses, expansions and reshuffles of effector genes (Raffaele & Kamoun, 2012). In previous studies we suggested that effector genes in *L. maculans* do not belong to multigene families and we postulated that duplications were followed by a rapid sequence diversification preventing us to recognize gene families (Rouxel *et al*., 2011). While this postulate still holds true, we found here one case where a paralog of *AvrLm4-7* could be identified (with 65.5% identity at the nucleotide level). In addition, we found that five effector genes of Lmb were translocated to TE-rich regions compared to the other members of the species complex. This is reminiscent at a smaller scale of what is observed for the *Avr-Pita* effector gene in *M. oryzae* (Chuma *et al*., 2011). In this species a paralog of *Avr-Pita* (termed *Avr-Pita3*, with 71% identity with the reference) is found at an invariant genome location, while *Avr-Pita* and other variants may be fully absent from the genome or present at different chromosomal locations, or on different chromosomes, depending on the isolate (Chuma *et al*., 2011). In this species, both patterns of complete loss of the gene and patterns of multiple translocations were ascribed to the close association of *Avr-Pita* with a retrotransposon (Chuma *et al*., 2011). Lastly, the authors postulated that other avirulence genes of *M. oryzae* may have the same behaviour and also submitted to multiple translocations in the genomes. Having only compared two Lmb isolates in the present study, we have not such evidence of possible multiple translocations in different isolates of Lmb. However, translocations are found in Lmb compared to other members of the species complex, suggesting that at least at the time of speciation, effector genes were submitted to TE-mediated translocations in the Lmb genome, resulting in many cases in their isolation within large AT-rich isochores. The adaptive advantage of these multiple translocations was considered by Chuma *et al*. (2011) to maintain a pool of isolates with the avirulence gene when selection is exerted so that it can be easily disseminated in populations when the selection is no more present. In Lmb, one additional adaptive advantage could be to move effector genes to genomic regions submitted to an epigenetic control so that these genes are only expressed during infection (Soyer *et al*., under review).

Fungal genomes can also be “invaded” through HGT, and TEs are key drivers to HGT in a series of species. For example, Richards *et al*. (2011) suggested that HGT played an important part in the evolution of plant parasitism in the oomycetes. Also, the HST ToxA has been demonstrated to have been transmitted from *P. nodorum* to *P. tritici-*

repentis in the 1940's while the proximity of the *ToxA* gene to a hAT family TE is suggested to have contributed to the transfer and integration of this gene into the *P. tritici-repentis* genome (Friesen *et al*., 2006). Clusters of genes encoding enzymes responsible for biosynthesis of secondary metabolites, molecules with a diverse range of structure and metabolic activities are known to be readily transmitted by HGT. Many biosynthetic gene clusters including those for the class of secondary metabolites that includes sirodesmin, are distributed discontinuously among ascomycetes and appear to have a single origin and to have been inherited relatively intact rather than independently in the different ascomycete lineages (Patron *et al*., 2007). However, gene clusters for polyketides such as dothistromin are often fragmented in ascomycetes and their evolution is more complicated (Bradshaw *et al*., 2013). Another example of HGT, that we present here, regards PKS21 in Lbb genome which appears to have been obtained via HGT from a dermatophyte species, *Arthroderma otae*. PKS5 and NPS8, that are specific to Lmb genome within the species complex, might also be the result of HGT since they are located within AT-isochores and surrounded by TEs in Lmb genome and rarely present in other dothideomycete species. Effector genes hosted in AT-isochores that are conserved in several species often display a patchy distribution along the phylogeny. Because of the low identity of sequence, it remains difficult to ascribe this to accelerated sequence evolution, regular gain and loss patterns or TE-mediated HGT as shown for *ToxA*. Intriguingly, both in the cases of secondary metabolite genes and effector genes, the relatives, if HGT are postulated, were often Sordariomycetes (and mostly *Fusarium* or *Colletotrichum* species) or *Aspergilli*. It is not clear whether these data have an ecological or evolutionary relevance or if this is the result of a bias due to currently available whole genome sequences in databases.

All these data indicate a hypothetical evolutionary scenario: (i) Invasion by new TE families, eventually generating mesosyteny and reproductive isolation, followed by (ii) a massive accumulation of TE-rich regions rapidly inactivated by RIP along with transposition of new invaders generating isolate-specific chromosomal landscapes and (iii) the generation of new genes, new specificities and new regulation in these TE-rich landscapes contributing to a better adapted species towards crucifers.

Materials & Methods

Biological material

Isolates IBCN84 (*L. maculans* 'lepidii') and IBCN65 (*L. biglobosa* 'thlaspii') are part of the International Blackleg of Crucifer Network collection maintained at INRA-Grignon and AAF Saskatoon. Both isolates were obtained as hyphal tip-purified isolates in Saskatchewan either from stinkweed (*Thlaspi arvense*) (IBCN65) or from peppergrass (*Lepidium sp.*) (IBCN84) (Mendes-Pereira *et al.*, 2003). Isolate B3.5 (*L. biglobosa* 'brassicae') is one of two isolates that were used to formally establish the new *L. biglobosa* species name (Shoemaker & Brun, 2001). It was obtained as a single-ascospore isolate ejected from a pseudothecia formed on *Brassica juncea* cv. Picra in an experimental field at INRA-Le Rheu (Shoemaker & Brun, 2001).

The *L. biglobosa* 'canadensis' isolate 06VTJ154 (hereafter named J154) was cultured from ascospores released from sexual fruiting bodies on *B. juncea* stubble collected from Burren Junction, New South Wales, Australia as described by Van de Wouw *et al.* (2008).

The *L. maculans* isolate WA74 was originally purified from infected stubble of oilseed rape collected from Western Australia as described by Mengistu *et al.* (1993). It is included in the IBCN collection under ref IBCN76 (Mendes-Pereira *et al.*, 2003).

Sequencing and assembly

For isolates IBCN84 (*L. maculans* 'lepidii'), IBCN65 (*L. biglobosa* 'thlaspii') and B3.5 (*L. biglobosa* 'brassicae'), a Paired-End 8 Kb library was constructed according to the protocol 454. A run and a half was done on the Titanium version, generating respectively 580 Mb, 511 Mb and 614Mb of raw data. An additional Paired-End 20-kb library was built for B3.5. An additional run was done on the Titanium version, generating 380 Mb of raw data. Additional libraries were constructed according to the Illumina protocol with an average size of 250 bp inserts for IBCN84 and IBCN65 and 350 bp for B3.5. Libraries were sequenced on three lanes (for IBCN84 and B3.5) and 2 lanes for IBCN65 on a GAIIx, 76 bp in single sequencing, generating 7.8 Gb of raw data for IBCN84, 6.08 Gb for B3.5 and 4.9 Gb for IBCN65. The Titanium sequences were assembled by Newbler and the sequences of the scaffolds were corrected using the Illumina sequences.

L. biglobosa 'canadensis' isolate J154 was grown in 10% Campbell's V8 juice for five days and genomic DNA was extracted from mycelia as described previously (Sexton &

Howlett, 2000). Genomic DNA was sequenced by the Australian Genome Research Facility, Melbourne, Australia using Illumina HiSeq2000. Libraries were constructed with a 250 bp insert size. A single lane of 100 bp pair-end sequencing was used to generate 32 Gb of raw data. Reads were trimmed to a minimum quality of phred 28 using FastX-toolkit software, producing 310 million trimmed reads. Trimmed reads were assembled de novo using the Velvet short read assembler v1.1.06 (Zerbino & Birney, 2008) with the Velvet Optimiser Perl script to select the k-mer (49), expected coverage (402) and coverage cutoff (43) values. The final *L. biglobosa* 'canadensis' J154 assembly produced 29.48 Mb of scaffolds > 200 bp. WA74 DNA was extracted (CTAB extraction protocol) from mycelia grown on V8 agar plates. Sequencing was done using Roche 454 Titanium platform both for shotgun sequencing (1 plate) and sequencing of an 8 kb mate-paired library (half plate).

Bioinformatics

Gene annotation.

The automated gene annotation was carried out using a combination of two *ab initio* gene predictors, Fgenesh (Salamov & Solovyev, 2000) and Genemark (Lomsadze *et al.*, 2005). Fgenesh has been previously used as a part of the EuGene pipeline (Foissac *et al.*, 2008) for gene annotation of the *L. maculans* 'brassicae' v23.1.3 genome (Rouxel *et al.*, 2011). Genemark, a free and easy-to-use software, was trained on the twenty largest SuperContigs of *L. maculans* 'brassicae' v23.1.3 with the repeated sequences masked using RepeatMasker (Smit, 1996) in order to avoid the RIP bias in basis usage on the repeated elements. Both Fgenesh and Genemark were firstly benchmarked on one SC of v23.1.3 and showed that Genemark gene models were better defined than Fgenesh and EuGene gene models. As a consequence, Genemark predictions were always prioritized over Fgenesh prediction in case of inconsistent annotation between the two predictors. Both Fgenesh and Genemark were run on repeat-masked genomic sequences and the results from the two gene callers were combined. Gene models encoding proteins greater than 30 amino acids were compared and a decision made as follows: (i) if a similar gene model was predicted at the same locus, or if it was predicted by one or the other of the predictors, this gene model was kept; (ii) if two different gene models were predicted at the same locus, we chose to prioritize the Genemark prediction; (iii) if a predicted gene model corresponded to two or more gene models from the other predictor, the latter was kept. All

validated genes were then translated into proteins and their features were written in a GFF3 formatted file which was used for visualisation (*e.g.* with Artemis (Berriman & Rutherford, 2003)).

Functional annotation.

Automated functional annotation of the predicted proteins was performed by using a combination of BLAST (Altschul *et al.*, 1990) and InterProScan (Quevillon *et al.*, 2005). First, each protein was compared by BLASTp to the NR database, and only the hit results matching the following criteria were kept for further analysis: (i) the e-value should less than 1e-06, (ii) the similarity percentage should be over 30%, (iii) at least 70% of the query sequence should be covered by the alignment length, (iv) the subject (hit) sequence length should represent between 75% and 125% of the query sequence length. Then, all validated results were pooled and average values were calculated for each of the above-mentioned criteria. A consensus description was then obtained by text-mining. According to these different features, proteins were divided into three classes: (i) proteins with no BLAST hit result or BLAST hit results with a mean percentage of similarity lesser than 40%, a mean percentage of coverage of the query sequence length by the alignment length lesser than 80% and a mean percentage of coverage of the query sequence length by the subjects length lesser than 85% or greater than 115%, and (ii) proteins with no protein domain identified by InterProScan. This class is classified as « predicted protein », i.e., predictions with no functional support. It usually contained species-specific sequence with no known domains. The second class, termed « hypothetical proteins », included predictions that (i) had at least one domain identified by InterProScan in the InterPro database (Hunter *et al.*, 2011) and fulfill the above-mentioned BLAST result criteria but for which text mining indicated « hypothetical protein » in more than 90% of the hits, or (ii) fulfill the BLAST results conditions with a consensus description corresponding to a function but with no protein domain identified, or (iii) with no BLAST results but with at least one domain identified from the InterPro database. Globally, this class contained predictions with slight functional support that might correspond to conserved proteins among several organisms but with no defined functions, or to splitted/merged predictions. The third class included predictions that fulfill the BLAST result conditions, with at least one domain from the InterPro database and with consensus hit description corroborated by the protein domains identified with InterProScan. This class was termed the « similar to *function* » protein and included well conserved proteins with defined functions among

several organisms. All parameter values used in this section were optimized values based on comparison between results of automated *in silico* and « manual » *in silico* functional annotation.

Identification and annotation of genes encoding SSPs and secondary metabolite clusters

A protein was classified as a SSP (i) if a signal peptide was predicted by both Neural-Network and Hidden Markov Model methods of SignalP 3.0 (Bendtsen *et al.*, 2004), (ii) if TargetP 1.1 (Emmanuelson *et al.*, 2000) predicted the protein being in the secretion pathway, (iii) if TMHMM 2.0 (Krogh *et al.*, 2001) detected none or one transmembrane domain if this latter is at least at 30% included in the signal peptide, (iv) and if its length was lower than or equal to 300 amino acids. SSP identification resulted from the merging of two sets of predicted proteins. The first one resulted from a step of gene annotation as described above. After this step, the validated gene models were masked on the genomic sequences and another round of gene annotation was carried out. From this second gene model set, we exclusively kept the ones encoding SSPs.

Genes encoding key enzymes in fungal secondary metabolism were sought. Non-ribosomal peptide synthases (NRPS), polyketide synthase (PKS) and dimethyl tryptophan synthases (DMATS) were identified by searching both the predicted proteins and genome assemblies of each species with previously characterised NPS, PKS and DMATS proteins from *L. maculans* 'brassicae' isolate v23.1.3 (Rouxel *et al.*, 2011). Both BLASTp and tBLASTn algorithms were used. Any match with greater than 35% sequence similarity to the previously identified proteins was compared by BLAST against the NCBI database to detect indicative domains, as well as best matches in other species. Genes encoding geranyl geranyl diphosphate (GGPP) synthases and terpene cyclases (TC) were identified by tBLASTn and BLASTp searches of *Leptosphaeria* genomes and predicted protein sets using verified references sequences as the query. GGPP synthase reference sequences were *Penicillium paxilli* GGS1 (AAK11525) and PAXG (AAK11531). Terpene cyclase reference sequences were *Saccharomyces cerevisiae* ERG7 (P38604), and *Fusarium graminearum* TRI5 (XP_383713).

Analysis of gene conservation

The 57,964 proteins predicted in the *Leptosphaeria* species complex were grouped using orthoMCL (with default parameters). OrthoMCL created 10,916 families that contained 48,013 sequences. In order to obtain 1:1 orthologous relationships between the

sequences, families made of proteins from only one taxon or containing a number of protein larger than that of taxa, were excluded. After this screening step, 10,131 families containing 43,437 sequences (74.9% of the species complex proteome) remained. These data were used to draw a Venn diagram (Figure 5).

The presence of proteins that were not included within the 10,131 families, was investigated in genomes of the other members of the species complex using different programs of BLAST with an e-value cut-off of $1e^{-5}$ and a request for Best Hits only: (i) BLASTn; (ii) tBLASTn on the 6 frame-translated genome sequences; (iii) BLASTp on the ungrouped proteins of the other members. All BLAST hits with coverage lower than 50% of the query sequence were considered as negative. According to the different combination of results, each protein was classified as: (i) species-specific, *i.e.* sequences that were not detected at the nucleotide level in the other assemblies or were pseudogenized; (ii) specific of a few members of the species complex, *i.e.* present in at least another member of the species complex, which can or not correspond to non-predicted genes in the other assemblies; (iii) unresolved, corresponding to sequences for we could not decide presence or absence status. Pseudogenes were hypothetized to occur when BLASTn and tBLASTn, but not BLASTp, provided significant results on other predicted proteins. To ascertain the reality of pseudogenes, the tBLASTn alignment was analyzed and the subject sequence was investigated for occurrence of stop codons or mutations in the start codon.

Structural and synteny analysis of chromosomes

Chromosomes of *L. maculans* 'brassicae' v23.1.3 were defined in Rouxel *et al.* (2011) by a combination of CHEF analysis, identification of telomeres and genetic mapping. The chromosome number was estimated to be 17-18. Here, the assemblies of *L. maculans* 'brassicae' WA74 and *L. maculans* ‘lepidii’ IBCN84 were aligned to that of Lmb v23.1.3 using the program *nucmer* of the MUMmer package. Alignments were then visualized using *mummerplot* of the MUMmer package which allowed to confirm postulated assembly errors in Lmb v23.1.3 using other approaches and to highlight a few previously unidentified assembly errors. These computational data coupled with biological data allowed us to infer a new number of chromosomes, now estimated to 19 (Supplementary Table S4). SNPs between aligned nonrepetitive sequences were identified using the program *show-snps* of the MUMmer package.

Transposable Elements annotation

Transposable Elements were identified and annotated using the REPET pipeline (Flutre *et al.*, 2011) (http://urgi.versailles.inra.fr/Tools/REPET). The TEdenovo pipeline detects repeat copies, clusters them into families and generates a consensus sequence for each family. Then these sequences are classified (TEclassifier.py) using tBLASTx and BLASTx against the Repbase Update database (Jurka *et al.*, 2005) and by identification of structural features such as long terminal repeats (LTRs) or terminal inverted repeats (TIRs), but due to the difficulties for Newbler to assemble correctly repeated regions, the majority of the consensus were not categorized into known TE families (Wicker *et al.*, 2007; Kapitonov & Jurka, 2008). Thus, manual annotation was necessary. Consensus sequences were clustered based on homologies research by BLAST, then aligned with ClustalX2 (Larkin *et al.*, 2007) and a new consensus was created. These steps were repeated until there were no more alignments by BLAST between the sequences, then consensuses were submitted to TEclassifier.py from the TEdenovo pipeline. The sequences were also translated into the six frames using Transeq from the EMBOSS package (Rice *et al.*, 2000) in order to carry out a protein domain research on the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2011) using RPS-BLAST. TE families of each strains were classified and named according to Wicker *et al.* (2007). These families constitute the TE repertoire of the *L. maculans – L. biglobosa* species complex which was used afterwards to reannotate each genome of the complex and to retrieve similar sequences among genomes of other Ascomycete species, mined on the JGI MycoCosm portal (http://genome.jgi.doe.gov/programs/fungi/index.jsf). RIP analysis was performed automatically on multiple alignment of sequences of each TE family using the RIPCAL software (Hane *et al.*, 2008).

Setting up of the synteny browser

A GBrowse-based synteny browser, GBrowse-syn (McKay *et al.*, 2010), was set up to display the synteny between genomes of the *Leptosphaeria* complex. Genome assemblies were aligned with Mercator and MAVID softwares (Dewey, 2007). Using Blat, Mercator aligned the CDS of all assemblies, providing constraints for genomic alignments with MAVID. For the *L. maculans* isolates, Mercator identified at least 94% of CDS as anchors for alignments (Lmb v23.1.3: 95%, Lmb WA74: 95%, Lml IBCN84: 94.3%), while it was lowered to 88% for *L. biglobosa* isolates (Lbt IBCN65: 89.5%, Lbb b3.5: 87.3%). The Lbc isolate was not included due to important genome fragmentation. Where the number of

aligned CDS was consistent for *L. maculans* isolates, the orthology map inferred from anchors revealed that the orthologous segments covered 64% and 66% of WA74 and v23.1.3 genomes, respectively. This seemingly low coverage is due to the importance of repeated region in the genomes of Lmb isolates. The other genomes were covered at 86.5%, 89.5% and 87.3% for IBCN84, IBCN65 and b3.5, respectively. Genomic alignments were performed with MAVID and loaded with genome annotations in MySQL databases. GBrowse-syn was configured to display the blocks of synteny (http://urgi.versailles.inra.fr/gb2/gbrowse_syn/leptosphaeria_synteny/).

Phylogeny and divergence time estimates

Cluster analyses.

An aggregative hierarchical clustering procedure, similar to the procedure used to generate protein clusters in the protein cluster database at NCBI (ProtClustDB; was used to identify orthologs. All versus all BLASTp results from protein sequences of 80 annotated genomes at NCBI. This represented a broad sampling of *Ascomycota*, *Basidiomycota* and *Chytridiomycota*. The following filters were applied: cluster members were required to have compositional BLAST hits covering at least 70% of each protein length and a pairwise score between cluster members was required to be at least 20% of the largest of the self-scores. Clusters were selected that contained one protein per taxon with at least 75 taxa present. Protein sequences from these clusters were subsequently used as queries in BLASTp searches in order to extract orthologs from annotated dothideomycetes genomes at JGI (http://genome.jgi.doe.gov/programs/fungi/index.jsf) and those generated for this study. Individual protein alignments were used to generate phylogenetic trees in FastTree (Price *et al.*, 2010). The phylogenies were then manually inspected for contradictory placement of taxa resulting from paralogs, or poor annotations. Taxa were judged to be contradictory when their placement above order level, in 70% bootstrap resamplings, contradicted accepted phylogenetic placements from recent broad analyses (Schoch *et al.*, 2009a, 2009b). Finally, 19 selected protein alignments were edited with Gblocks v.0.91b (Talavera and Castresana 2007) with the following parameters b3=8 b4=5 b5=h. The final concatenated alignment consisted of the following 19 proteins: Cct3p, Chc1p, Frs2p, Hsp60p, Imp3p, Kre33p, Lys1p, Pol3p, Pro2p, Pup1p, Ret1p, Rpo21p, Rpt2p, Rpt5p, Rrb1p, Rvb1p, Rvb2p, Sec26p and Tcp1p (following *Saccharomyces cerevisiae* nomenclature). The fungal taxa were trimmed to 51 in order to allow a relatively small and

focused data set for analysis. This resulted in a data matrix of 11694 amino acids (aa) with 3.31% consisting of gaps and completely undetermined characters.

Divergence time estimates.

The data set was analysed using Bayesian relaxed molecular clock approaches in BEAST v1.7.3. An xml file was prepared with the aid of BEAUTi v.1.7.3 (Drummond *et al.*, 2012). This included a starting tree generated in RAxML and calibrated with r8s (Sanderson, 2003). The same phylogeny was also used to determine which nodes could be constrained. Bayesian relaxed uncorrelated lognormal clock analyses with a birth-death tree prior were specified under the WAG substitution model (Gamma+invariant sites). Secondary time calibrations were used from recent publications: A normal distributed prior of 662 MYA was set for the root of the tree with uniform prior minimums of 417 MYA and 402 MYA for *Ascomycota* and *Basidiomycota* respectively following Floudas *et al.* (2012). Maximum dates were set to 644 and 664. Similar restrictions were placed on *Dothideomycetes* with a minimum age of 284 and a maximum of 394 following Gueidan *et al.* (2011). Five independent BEAST MCMC chains were run for 15 million generations sampling data every 1000th generation using the XSEDE computing infrastructure through the CIPRES Science Gateway webportal (Miller *et al.*, 2010). The resulting log files were combined using LogCombiner v1.7.3 and inspected with Tracer v1.5 to confirm the estimated sample sizes (ESS).The ESS values from six combined Bayesian relaxed clock uncorrelated lognormal clock analyses are indicated in Supplementary Table S13. The runs converged to stable likelihood values independently and the first 1 500 trees were discarded. The remaining 13 500 ultrametric trees from each run were combined and analyzed using TreeAnnotator v1.7.3 to estimate the 95% highest posterior densities (HPD). The consensus chronogram including the 95% HPD and the mean age estimates were visualised in Figtree v1.3.1 and certain clades were collapsed for clarity in Figure 1. The full figure is available as Supplementary Figure S2.

RNA-seq data collection and processing

Mycelia of *L.maculans* brassicae IBCN18 and of *L.biglobosa* canadensis J154 were harvested on Miracloth after 7 days growth in oilseed rape liquid medium. Infected cotyledonary tissue (0.5 mm diameter) was harvested from 32 *B. napus* cv. Westar seedlings 7 days post inoculation (dpi) (Van de Wouw *et al.* 2009). RNA was extracted

using Trizol reagent (Life technologies) and subsequently DNaseI-treated (Promega). Two biological replicates of RNA, with an RNA integrity number (RIN) > 6, were sequenced with Illumina TruSeq version 3 chemistry on an Illumina HiSeq2000 sequencer at the Australian Genome Research Facility (AGRF). *In vitro* derived RNA was sequenced with 100 bp paired-end reads in order to aid gene annotation, while *in planta* derived RNA was sequenced with 100 bp single-end reads. A total of 15.5 Gb sequence was generated from two i*n vitro* libraries (7.75 Gb per sample), and 72 Gb sequence was generated from 12 *in planta* libraries (6 Gb per sample). Reads were trimmed to a minimum phred score of 20 using Nesoni sequence software and remaining Illumina Tru-seq adaptor sequences were removed. Trimmed reads were aligned to reference genome sequences with Tophat v1.4.1 splice-junction mapper (Trapnell *et al.*, 2009). Reference genomes were *L. maculans* 'brassicae' v23.1.3 (Rouxel *et al.*, 2011) *L. biglobosa* 'canadensis' isolate J154 (this study). Aligned reads were quantified using Cufflinks v1.0.3 transcript assembly and quantification software (Trapnell *et al.*, 2010).

Gene models for Lbc isolate J154 were validated with the RNA-seq sequence data. Of the 11068 gene models, 8521 had >10 x coverage, 9979 had >1x coverage and 530 did not have a detectable transcript. Of the 11068 gene models, 8948 had a FPKM (fragments per kb of exon per million mapped reads) value >1; this value is an indication of significant expression.

Incidence of repeats in the in planta *expression of orthologs.*

We firstly established the 7561 orthologous pairs between Lmb M1 and Lbc J154 and recorded RNA-seq gene expression values (FPKM) for each gene in two conditions, *in planta* (infected *B. napus* cv Westar at 7dpi), and *in vitro* (oilseed rape medium at 7 dpi). Genes without sufficient expression data were excluded, leaving 7415 ortholog pairs. Quantile Normalisation was applied on the four gene expression datasets so that they could be directly compared. The repeat category for each ortholog pair was determined: the RepeatMasker output based on TE families described above defined the repeat content of each genome, and all repeats were included in the analysis. Either another gene or a repeat is directly adjacent to each gene considered; if a gene was adjacent to the end of a scaffold it was categorised as "no repeat adjacent". Sixteen repeat categories are possible based on the promoter and terminator repeat status for both orthologous genes. The four categories based on only the promoter region were analysed. An expression ratio was calculated from the log2 (*in planta* expression /in vitro expression) for

each gene. A positive value means the gene is more highly expressed *in planta*; a negative value means the gene is more highly expressed *in vitro* culture.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:**403-410.

Balesdent MH, Fudal I, Ollivier B, Bally P, Grandaubert J, Eber F, Chèvre AM, Leflon M, Rouxel T. 2013. The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytol* **198:**887-898.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340:**783-795.

Berbee ML and Taylor JW. 2010. Dating the molecular clock in fungi -- how close are we? *Fungal Biol Rev* **24:**1-16.

Berriman M and Rutherford K. 2003. Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4:**124-132.

Bradshaw RE, Slot JC, Moore GG, Chettri P, de Wit PJ, Ehrlich KC, Ganley AR, Olson MA, Rokas A, Carbone I *et al*. 2013. Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. *New Phytol* **198:**525-535.

Chuma I, Isobe C, Hotta Y, Ibaragi K, Futamata N, Kusaba M, Yoshida K, Terauchi R, Fujita Y, Nakayashiki H *et al*. 2011. Multiple translocation of the *AVR-Pita* effector gene among chromosomes of the rice blast fungus *Magnaporthe oryzae* and related species. *PLoS Pathog* **7:**e1002147.

Clutterbuck AJ. 2011. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. *Fungal Genet Biol* **48:**306-326.

Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otillar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N *et al*. 2013. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genet* **9:**e1003233.

Couch BC, Fudal I, Lebrun MH, Tharreau D, Valent B, van Kim P, Nottéghem JL, Kohn LM. 2005. Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. *Genetics* **170:**613-630.

Crous PW, Summerell BA, Carnegie AJ, Wingfield MJ, Hunter GC, Burgess TI, Andjic V, Barber PA, Groenewald JZ. 2009. Unravelling *Mycosphaerella*: do you believe in genera? *Persoonia* **23:**99-118.

Cunningham GH. 1927. Dry-rot of Swedes and turnips: its cause and control. *N-Z Dep Agric* **Bull 133**.

Cuomo CA, Güldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M *et al*. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317:**1400-1402.

Daboussi MJ, Capy P. 2003. Transposable elements in filamentous fungi. *Annu Rev Microbiol* **57:**275-299.

Daverdin G, Rouxel T, Gout L, Aubertot JN, Fudal I, Meyer M, Parlange F, Carpezat J, Balesdent MH. 2012. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PloS Pathog* **8:**e1003020.

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H *et al*. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434:**980-986.

Dewey CN. 2007. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Bio.* **395:**221-236.

De Wit PJ, van der Burgt A, Ökmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP *et al*. 2012. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PloS Genet* **8:**e1003088.

Dilmaghani A, Gladieux P, Gout L, Giraud T, Brunner PC, Stachowiak A, Balesdent MH, Rouxel T. 2012. Migration patterns and changes in population biology associated with the worldwide spread of the oilseed rape pathogen *Leptosphaeria maculans*. *Mol Ecol* **21:**2519-2533.

Dilmaghani A, Gout L, Moreno-Rico O, Dias JS, Coudard L, Castillo-Torres N, Balesdent MH, Rouxel T. 2013. Clonal populations of *Leptosphaeria maculans* contaminating cabbage in Mexico. *Plant Pathology* **62:**520-532.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29:**1969-1973.

Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL *et al*. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci USA* **108:**9166-9171.

Elliott CE, Gardiner DM, Thomas G, Cozijnsen AJ, van de Wouw A, Howlett BJ. 2007. Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. *Mol Plant Pathol* **8:**791-802.

Ellwood SR, Syme RA, Moffat CS, Oliver RP. 2012. Evolution of three *Pyrenophora* cereal pathogens: recent divergence, speciation and evolution of non-coding DNA. *Fungal Genet Biol* **49:**825-829.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300:**1005-1016.

Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, Couloux A, Aury

JM, Ségurens B, Poulain J *et al*. 2008. The genome sequence of the model ascomycete fungus *Podospora anserina*. *Genome Biol* **9:**R77.

Eyre-Walker A and Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet* **2:**549-555.

Fitt BDL, Huang YJ, van den Bosch F, West JS. 2006a. Coexistence of related pathogen species on arable crops in space and time. *Annu Rev Phytopathol* **44:**163-182.

Fitt BDL, Brun H, Barbetti MJ, Rimmer SR. 2006b. Worldwide importance of phoma stem canker (*Leptosphaeria maculans* and *L. biglobosa*) on oilseed rape (*Brassica napus*). *Eur J Plant Pathol* **114**:3-15.

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS *et al*. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336:**1715-1719.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in *de novo* annotation approaches. *Plos One* **6:**e16526.

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, V. de Peer Y, Rouze P, Schiex T. 2008. Genome Annotation in Plants and Fungi: EuGene as a model platform. *Current Bioinformatics* **3:**87-97.

Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP. 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* **38:**953-956.

Fudal I, Böhnert HU, Tharreau D, Lebrun MH. 2005. Transposition of MINE, a composite retrotransposon, in the avirulence gene *ACE1* of the rice blast fungus *Magnaporthe grisea*. *Fungal Genet Biol* **42:**761-772.

Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, Eckert MR, Cattolico L, Bernard-Samain S, Balesdent MH, Rouxel T. 2007. Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. *Mol Plant-Microbe Interact* **20:**459-470.

Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, Balesdent MH, Rouxel T. 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. *Mol Plant Microbe Interact* **22:**932-941.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422:**859-868.

Galagan JE and Selker EU. 2004. RIP: the evolutionary cost of genome defense. *Trends Genet* **20:**417-423.

Gardiner DM, Cozijnsen AJ, Wilson LM, Pedras MS, Howlett BJ. 2004. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans*. *Mol Microbiol* **53:**1307-1318.

Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, Balesdent MH, Rouxel T. 2006. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Mol Microbiol* **60:**97-80.

Gout L, Kuhn ML, Vincenot L, Bernard-Samain S, Cattolico L, Barbetti M, Moreno-Rico O, Balesdent MH, Rouxel T. 2007. Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. *Environ Microbiol* **9:**2978-2992.

Goyon C, Barry C, Grégoire A, Faugeron G, Rossignol JL. 1996. Methylation of DNA repeats of decreasing sizes in *Ascobolus immersus*. *Mol Cell Biol* **16:**3054-3065.

Gueidan C, Ruibal C, de Hoog GS, Schneider H. 2011. Rock-inhabiting fungi originated during periods of dry climate in the late Devonian and middle Triassic. *Fungal Biol* **115:**987-996.

Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE *et al*. 2007. Dothideomycete–plant Interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum. Plant Cell* **19:**3347-3368.

Hane JK and Oliver RP. 2008. RIPCAL: a tool for alignment-based analysis of repeat-induce point mutations in fungal genomic sequences. *BMC Bioinformatics* **9:**478.

Hane JK, Oliver RP. 2010. In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. *BMC Genomics* **11:**655.

Hane JK, Rouxel T, Howlett BJ, Kema GH, Goodwin SB, Oliver RP. 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol* **12:**R45.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S *et al*. 2011. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* **40:**D306-D312.

Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genet Biol* **39:**31-37.

Irelan JT, Selker EU. 1996. Gene silencing in filamentous fungi: RIP, MIP and quelling. *J Genet* **75:**313-324.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13:**91-96.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:**462-467.

Kaessmann H, Vinckenbosch N, Long M. 2011. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10:**19-31

Kämper J, Kahmann R, Bölker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O *et al.* 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444:**97-101.

Kang S, Sweigard JA, Valent B. 1995. The PWL host specificity gene family in the blast fungus *Magnaporthe grisea*. *Mol Plant Microbe Interact* **8:**939-948.

Kapitonov VV and Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* **9:**411-412.

Kasuga T, White TJ, Taylor JW. 2002. Estimation of nucleotide substitution rates in Eurotiomycete fungi. *Mol Biol Evol* **19:**2318-2324.

Kremer A, Li SM. 2010. A tyrosine O-prenyltransferase catalyses the first pathway-specific step in the biosynthesis of sirodesmin PL. *Microbiology* **156:**278-286.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J Mol Biol* **305:**567-580.

Kupferschmidt K. 2012. Mycology. Attack of the clones. *Science* **337:**636-638.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:**R12.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:**2947-2948.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:**2178-89.

Lievens B, Houterman PM, Rep M. 2009. Effector gene screening allows unambiguous identification of *Fusarium oxysporum* f. sp. *lycopersici* races and discrimination from other *formae speciales*. *FEMS Microbiol Lett* **300:**201-215.

Lomsadze A, Ter-Hovhannisyan V, Chernoff Y, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33:**6494-6506.

Loreto EL, Carareto CM, Capy P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. *Heredity* **100:**545-554.

Manning VA, Pandelova I, Dhillon B, Wilhelm LJ, Goodwin SB, Berlin AM, Figueroa M, Freitag M, Hane JK, Henrissat B *et al*. 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3* **3:**41-63.

Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ *et al*. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39:**D225-D229.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:**376-380.

McKay SJ, Vergara IA, Stajich J. 2010. "Using the Generic Synteny Browser (Gbrowse_syn). *Curr Protoc Bioinfo* **9:**Unit 9.12.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)* **New Orleans 2010:**1-8.

Mendes-Pereira E, Balesdent MH, Brun H, Rouxel T. 2003. Molecular phylogeny of the *Leptosphaeria maculans-L. biglobosa* species complex. *Mycol Res* **107:**1287-1304.

Mengistu A, Rimmer RS, Williams PH. 1993. Protocols for in vitro sporulation, ascospore release. sexual mating and fertility in crosses of *Leptosphaeria maculans*. *Plant Disease* **77:**538-540.

Morales VM, Séguin-Swartz G, Taylor JL. 1993. Chromosome size polymorphism in *Leptosphaeria maculans*. *Phytopathology* **83:**503-509.

Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F *et al*. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PloS Pathog* **8:**e1003037.

Patron NJ, Waller RF, Cozijnsen AJ, Straney DC, Gardiner DM, Nierman WC, Howlett BJ. 2007. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. *BMC Evol Biol* **7:**e174.

Parlange F, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, Blaise F, Grezes-Besset B, Rouxel T. 2009. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents Rlm4-mediated recognition through a single amino acid change. *Mol Microbiol* **71:**851-863.

Pedras MSC, Taylor JL, Nakashima TT. 1993. A novel chemical signal from the “blackleg” fungus: beyond phytotoxins and phytoalexins. *J Org Chem* **58:**4778-4780.

Petrie GA. 1969. Variability in *Leptosphaeria maculans* (Desm.) Ces. et De Not., the cause of blackleg of rape. PhD thesis, University of Saskatchewan.

Pound GS. 1947. Variability in *Phoma lingam. J Agric Res* **75:**113-133.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5:**e9490.

Quaedvlieg W, Kema GH, Groenewald JZ, Verkley GJ, Seifbarghi S, Razavi M, Mirzadi Gohari A, Mehrabi R, Crous PW. 2011. *Zymoseptoria* gen. nov.: a new genus to accommodate *Septoria*-like species occurring on graminicolous hosts. *Persoonia* **26:**57-69.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res* **33:**W116-W120.

Raffaele S and Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol* **10:**417-430.

Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: Towards new species. *Gene* **454:**1-7.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16:**276-277.

Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc Natl Acad Sci U S A* **108:**15258-15263.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S *et al*. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point Mutations. *Nat Commun* **2:**202.

Salamov AA and Solovyev VV. 2000. *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res* **10:**516-522.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19:**301-302.

Schoch CL, Sung GH, López-Giráldez F, Townsend JP, Miadlikowska J, Hofstetter V, Robbertse B, Matheny PB, Kauff F, Wang Z *et al*. 2009a. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst Biol* **58:**224-239.

Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C *et al*. 2009b. A class-wide phylogenetic assessment of Dothideomycetes. *Stud Mycol* **64:**1-15S10.

Sexton AC and Howlett BJ. 2000. Characterisation of a cyanide hydratase gene in the phytopathogenic fungus *Leptosphaeria maculans* and its role in infection. *Mol Gen Genet* **263:**463-470.

Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q *et al*. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res* **19:**1722-1731.

Shoemaker RA and Brun H. 2001. The telemorph of the weakly aggressive segregate of *Leptosphaeria maculans*. *Can J Bot* **79:**412-419.

Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* **6:**57-71.

Smit AFA., Hubley R, Green P. 1996-2012. RepeatMasker Open-3.0. *<http://www.repeatmasker.org>.*

Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, Balesdent MH, Connolly LR, Freitag M, Rouxel T *et al*. 2013. Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. *Plos Pathog* Under Review.

Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ *et al*. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330:**1543-1546.

Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. 2007. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol Biol Evol* **24:**398-411.

Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH. 2010. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genet* **6:**e1001189.

Talavera G and Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56:**564-577

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzburg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28:** 511-515.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:**1105-1111.

Van de Wouw AP, Thomas VL, Cozijnsen AJ, Marcroft SJ, Salisbury PA, Howlett BJ. 2008. Identification of *Leptosphaeria biglobosa* 'canadensis' on *Brassica juncea* stubble from northern New South Wales, Australia. *Australasian Plant Disease Notes* **3:**124-128.

Van de Wouw AP, Marcroft SJ, Barbetti MJ, Hua Li, Salisbury PA, Gout L, Rouxel T, Howlett BJ, Balesdent MH. 2009. Dual control of avirulence in *Leptosphaeria maculans* towards a *Brassica napus* cultivar with sylvestris-derived resistance suggests involvement of two resistance genes*. Plant Pathology* 58: 305–13

Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ. 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. *PloS Pathog* **6:**e1001180.

Vincenot L, Balesdent MH, Li H, Barbetti MJ, Sivasithamparam K, Gout L, Rouxel T. 2008. Occurrence of a new subclade of *Leptosphaeria biglobosa* in Western Australia. *Phytopathology* **98:**321-329.

Voigt K, Cozijnsen AJ, Kroymann J, Pöggeler S, Howlett BJ. 2005. Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and β-tubulin sequences. *Mol Phylogenet Evol* **37:**541-557.

Walsh AM, Kortschak RD, Gardner MG, Bertozzia T, Adelsona DL. 2013. Widespread horizontal transfer of retrotransposons. *Proc Natl Acad Sci USA* **15:**1012-1016.

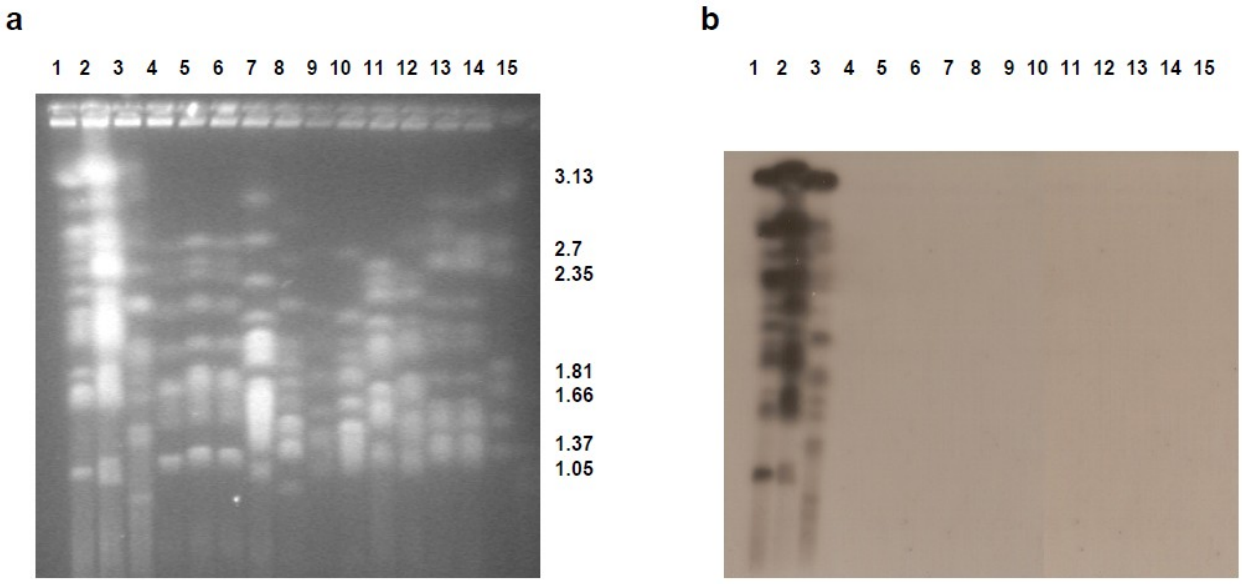
West JS, Balesdent MH, Rouxel T, Narcy JP, Huang YJ, Roux J, Steed JM, Fitt BDL, Schmit J. 2002. Colonization of winter oilseed rape tissues by A/Tox+ and B/Tox0 *Leptosphaeria maculans* (phoma stem canker) in France and England. *Plant Pathol* **51:**311-321.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O *et al*. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:**973-982.

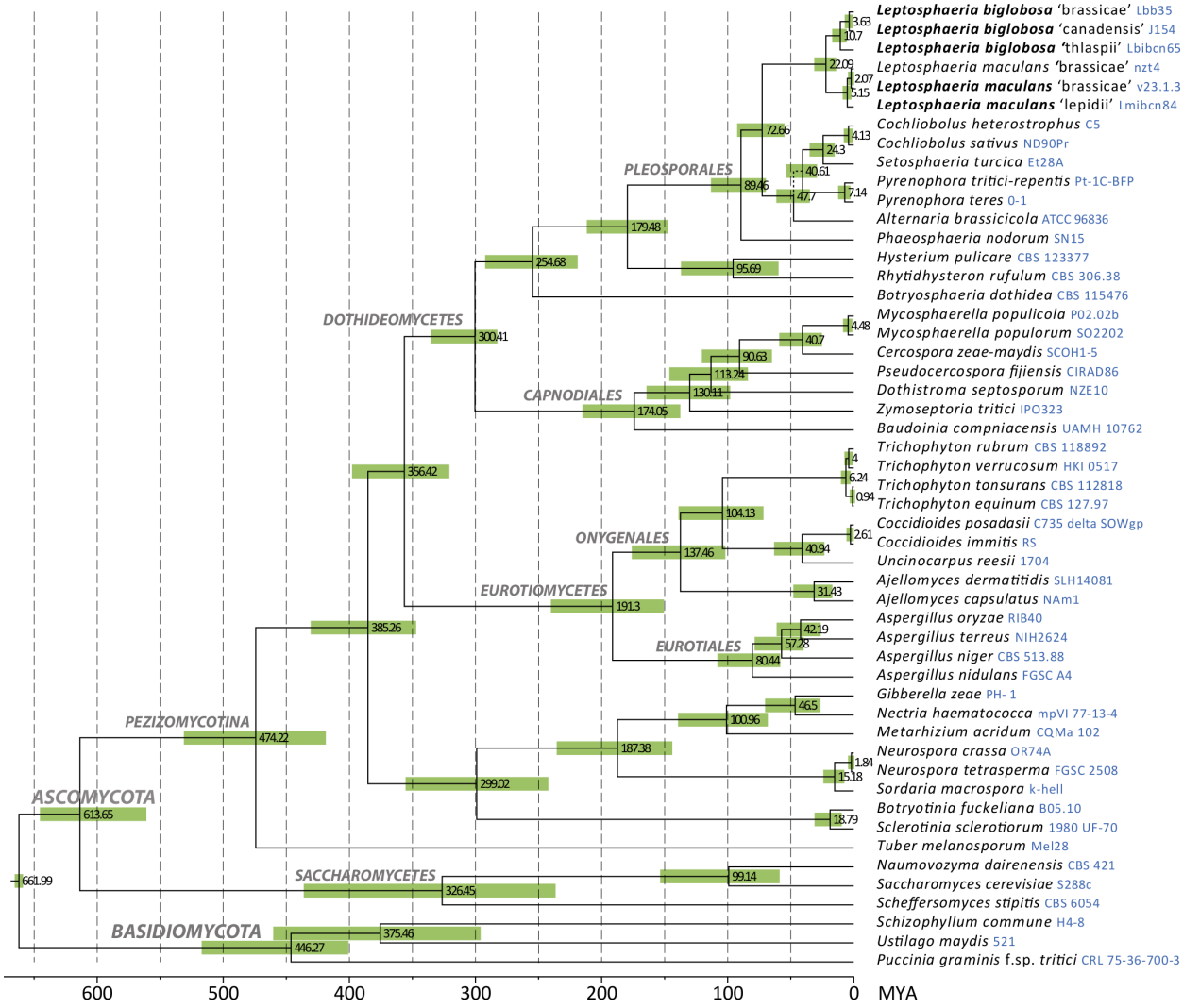
Wicker T, Oberhaensli S, Parlange F, Buchmann JP, Shatalina M, Roffler S, Ben-David R, Doležel J, Simková H, Schulze-Lefert P *et al*. 2013. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nat Genet* **45:**1092-1096.

Zerbino DR and Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18:**821-829.

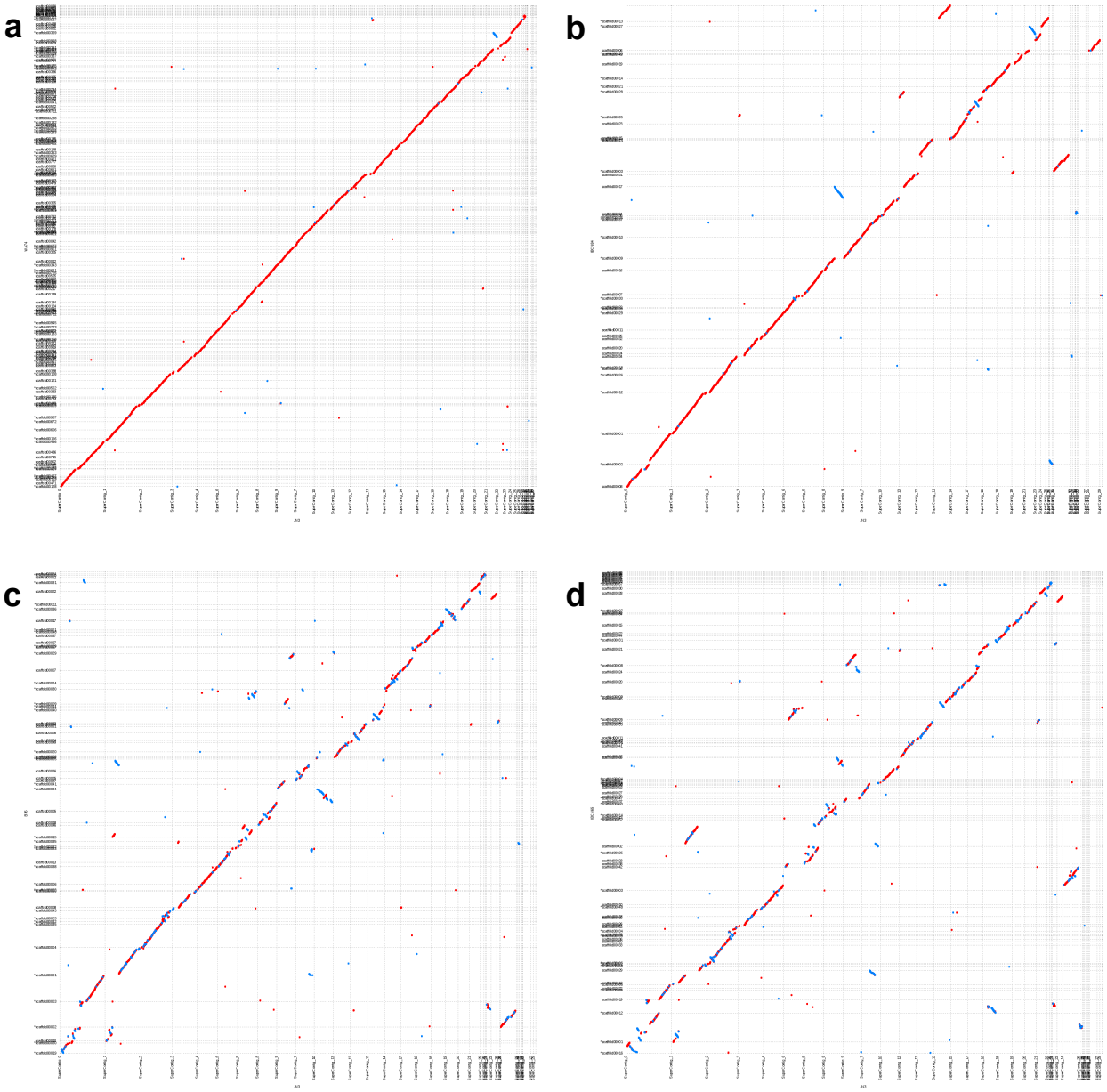
Supplementary figures



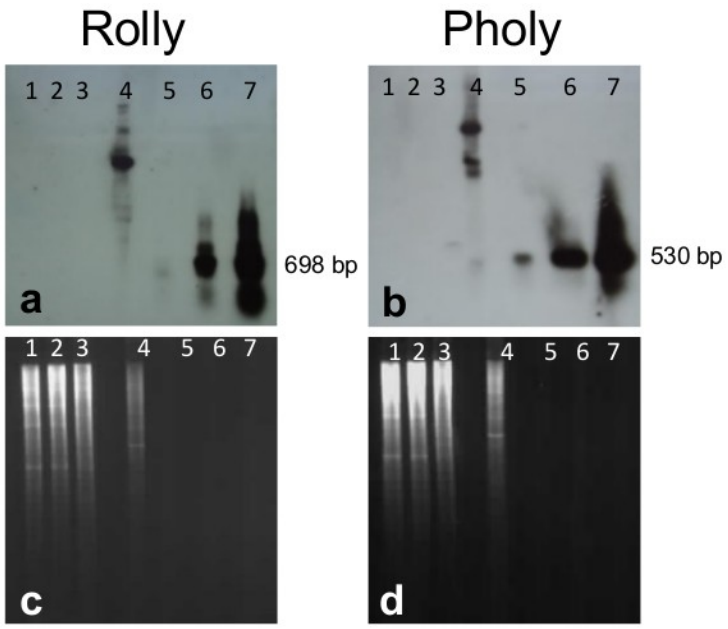
Supplementary Figure S1. Representative electrokaryotypes and presence of transposable elements in the genomes of isolates of the *L. maculans-L. biglobosa* species complex. (**a**) Electrokaryotypes were separated by Contour-clamped Homogeneous Electric Field (CHEF) electrophoresis according to protocols by Morales *et al.* (1993). (**b**) Southern blotting using one probe derived from the retrotransposon RLG_*Rolly*. Identity of the isolates is as follows *L. maculans* 'brassicae': lane 1, **v23.1.3**; lane 2, v29; lane 3, IBCN18; *L. maculans* 'lepidii': lane 4, **IBCN84**; lane 5, Lepi-1; lane 6, Lepi-2; *L. biglobosa* 'brassicae': lane 7, IBCN10; lane 8, IBCN93; *L. biglobosa* 'canadensis': lane 9, IBCN62; lane 10, IBCN82; *L. biglobosa* 'australensis': lane 12, IBCN30; lane 13, IBCN91; *L. biglobosa* 'thlaspii': lane 13, **IBCN65**; lane 14, IBCN64; lane 15, molecular weight marker *H. wingei* chromosomes. Names of isolates in bold are those sequenced here or the reference isolate previously sequenced (Rouxel *et al.*, 2011).



Supplementary Figure S2. Expanded chronogram of major classes in *Ascomycota*, with a focus on *Dothideomycetes*, produced with BEAST from a data set of 19 truncated proteins. Numbers at nodes indicate mean node ages in millions of years and light green bars indicate their 95% highest posterior density intervals.

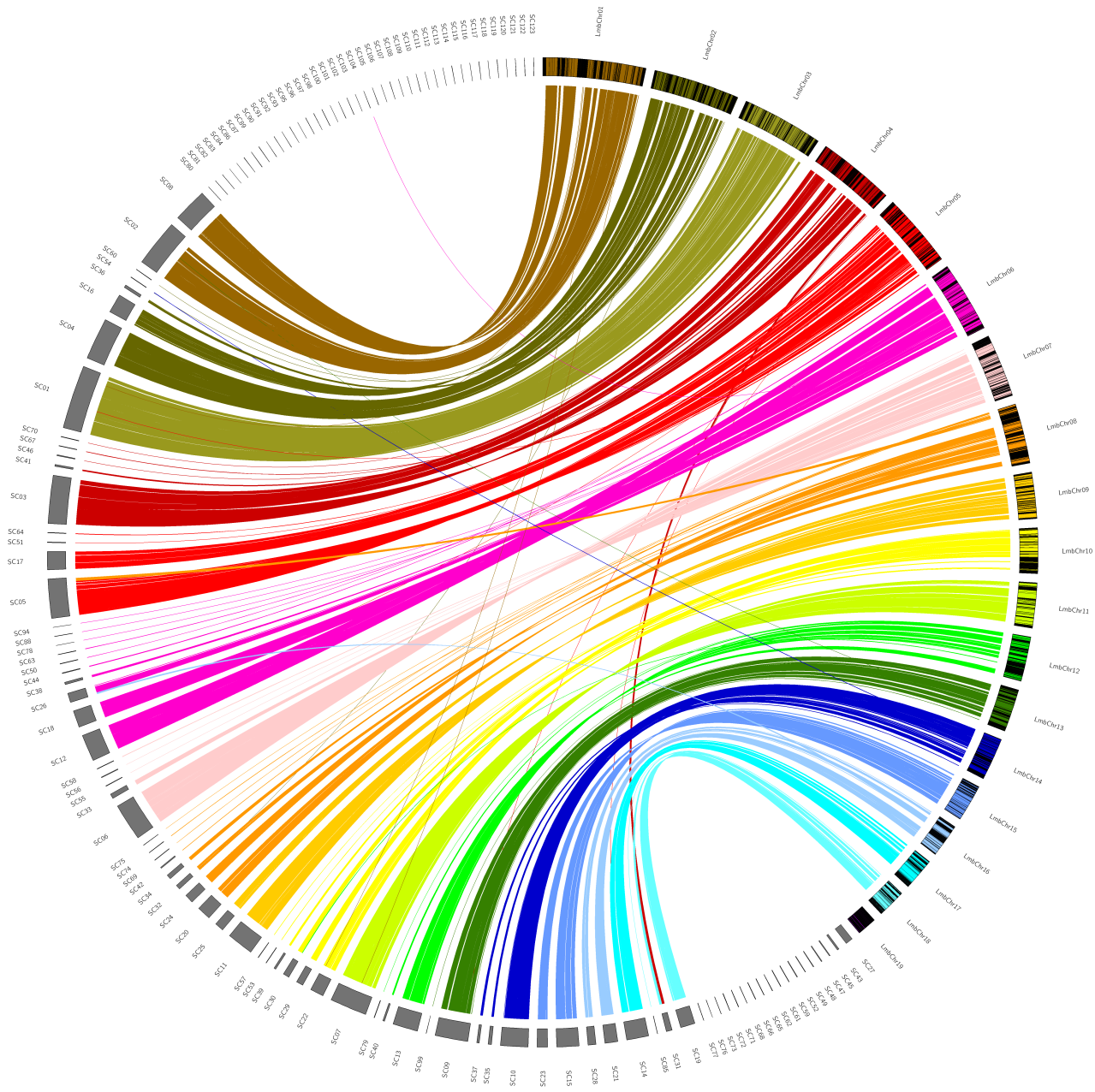


Supplementary Figure S3. Whole genome DNA comparison of *Leptosphaeria maculans* 'brassicae' v23.1.3 to progressively more distantly related members of the species complex. (**a**) comparison between two *Leptosphaeria maculans* 'brassicae' isolates v23.1.3 and WA74 showing lack of detectable chromosomal reorganisations; (**b**) comparison with *L. maculans* 'lepidii' showing extensive macrosynteny with only limited number of intrachromosomal inversions; (**c**) comparison to *Leptosphaeria biglobosa* 'brassicae' and (**d**) comparison to *Leptosphaeria biglobosa* 'thlaspii' showing many intrachromosomal inversions but no detectable large scale translocations.

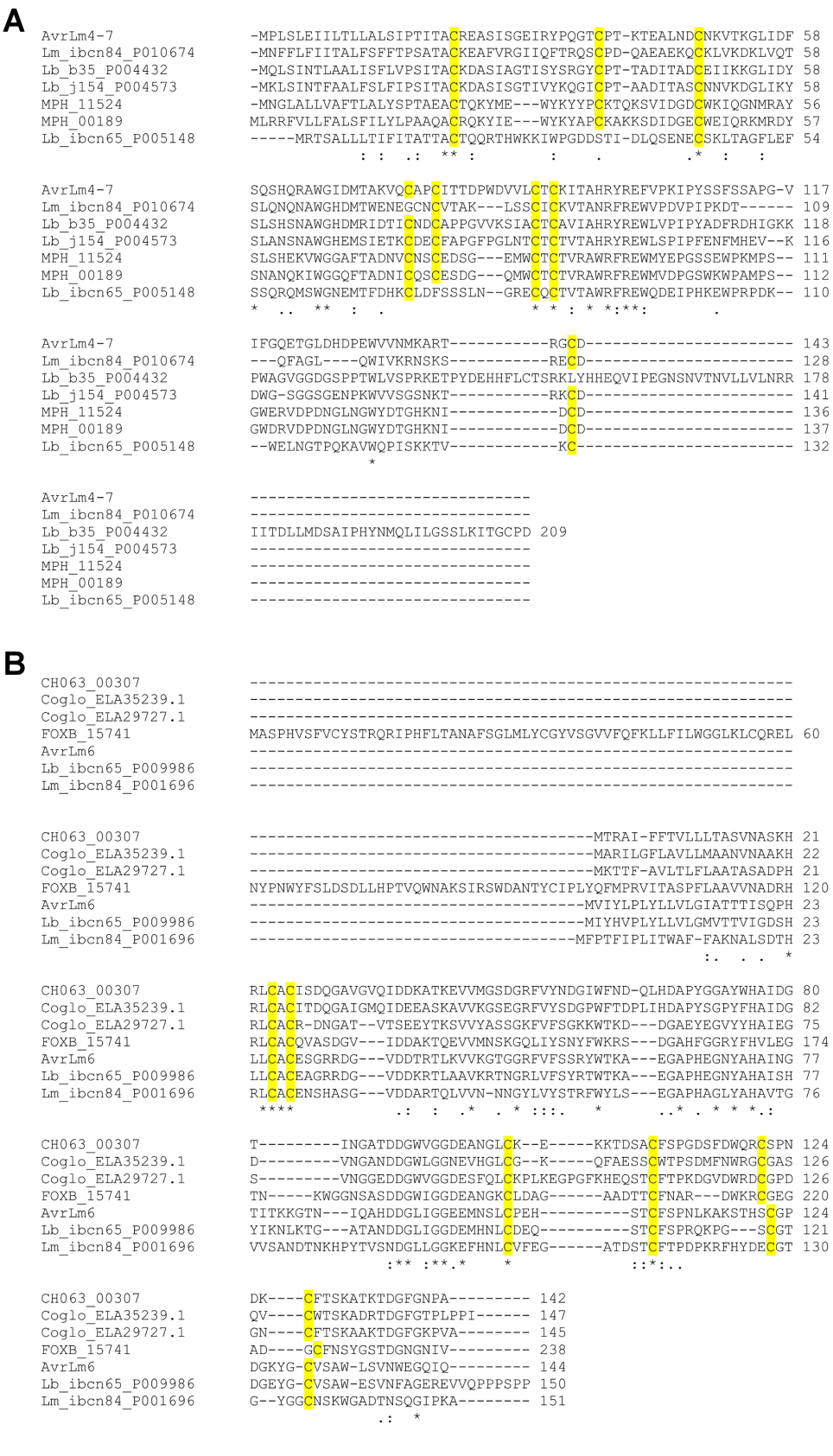


Lane	Description
1.	*L. biglobosa* 'canadensis' J154 genomic DNA
2.	*L. biglobosa* 'canadensis' J140 genomic DNA
3.	*L. biglobosa* 'occiaustralensis' MU7 genomic DNA
4.	*L. maculans* 'brassicae' v23.1.3 genomic DNA
5.	1 copy of transposable element
6.	10 copies of transposable element
7.	100 copies of transposable element

Supplementary Figure S4. Genomic DNA of isolates of *Leptosphaeria* species digested with restriction enzymes *Bam*HI (RLC_*Pholy*) or *Hind*III (RLG_*Rolly*) and hybridised with probes of transposable elements abundant in *L. maculans* 'brassicae'. (**a**, **b**) Probed with RLG_*Rolly*; RLC_*Pholy*, respectively. (**c**, **d**) Ethidium bromide stained gel of (**a**, **b**), respectively. None of the probes hybridised to DNA of *L. biglobosa* 'canadensis' or with another Australian member of the species complex (not included in the present study), *L. biglobosa* 'occiaustraliensis'.



Supplementary Figure S5. Circos representation of chromosome-by-chromosome genome conservation between *L. maculans* 'brassicae' (right part of the diagramme) and *L. maculans* 'lepidii' (left part). For *L. maculans* 'brassicae', each chromosome is represented in a different colour and black blocks within coloured bars represent AT-rich genome blocks enriched in transposable elements.



Supplementary Figure S6. Conservation of cysteine spacing in a series of orthologs of avirulence proteins of *L. maculans* 'brassicae'. (**a**) multiple alignment of orthologs of AvrLm4-7, (**b**) multiple alignment of orthologs of AvrLm6.

Supplementary tables

Supplementary Table S1. Characteristics of selected genomes of *Dothideomycetes*.

	A. brassicicola	*C. heterostrophus*	*C. sativus*	*P. tritici-repentis*	*P. teres f. teres*	*S. turcica*	*P. nodorum*	*Z. tritici*	*P. fijiensis*
	ATCC 96836	**C5**	**ND90Pr**	**Pt-1C-BFP**	**0-1**	**Et28A**	**SN15**	**IPO323**	**CIRAD86**
Genome size (Mb)	32,0	36,5	34,4	37,8	33,6	43,0	37,2	39,7	74,1
Scaffold number	838	68	157	47	6412	407	108	21	56
Scaffold N50 (kb)	2486	1842	1789	1986	36	2141	1045	2675	5902
Gaps (%)	5,3	0,4	3,5	1,7	0,0	11,1	0,4	0,0	0,6
Repeat content (%)[a]	6,15	8,64	6,14	12,24	3,08	12,96	2,88	12,26	39,50
'Core' genome size (Mb)	28,3	33,2	31,1	32,6	32,5	32,7	36,0	34,8	44,4
GC genome (%)	50,5	49,8	49,8	51,0	50,9	51,4	50,4	52,1	45,2
Predicted gene number[a]	10688	13336	12250	12141	17799	11702	12380	10933	13107

[a] data from Ohm *et al.*, 2012

Supplementary Table S2. Orthologues of the *Neurospora crassa* factors necessary for gene silencing identified in the genome of isolates of the *Leptosphaeria maculans-L. biglobosa* species complex.

	Function [a]	*N. crassa* [b]	*L. maculans* 'brassicae'	*L. maculans* 'lepidii'	*L. biglobosa* 'brassicae'	*L. biglobosa* 'canadensis'	*L. biglobosa* 'thlaspii'
RIP							
RID	Putative DMT, essential for RIP and MIP	NCU02034.5	Lema_P040230.1	Lm_ibcn84_P006192	Lb_b35_P003215	Lb_j154_P009468	Lb_ibcn65_P004900
DIM-5	H3 3mK9 HMT, essential for RIP	NCU04402.5	Lema_P050470.1	Lm_ibcn84_P004260	Lb_b35_P009656	Lb_j154_P005795	Lb_ibcn65_P008497
Quelling							
QDE-1	RdRP, essential for quelling	NCU07534.5	Lema_P088990.1	Lm_ibcn84_P007103	Lb_b35_P005332	Lb_j154_P009184	Lb_ibcn65_P002629
QDE-2	Argonaute-like protein, essential for quelling	NCU04730.5	Lema_P015760.1	Lm_ibcn84_P000699	Lb_b35_P006441	Lb_j154_P010439	Lb_ibcn65_P005933
QDE-3	RecQ helicase, essential for quelling	NCU08598.5	Lema_P098920.1	Lm_ibcn84_P009839	Lb_b35_P005693	Lb_j154_P005459	Lb_ibcn65_P001448
DCL1	Dicer-like protein, involved in quelling	NCU08270.5	Lema_P036310.1	Lm_ibcn84_P009183	Lb_b35_P003932	Lb_j154_P003181	Lb_ibcn65_P007287
DCL2	Dicer-like protein, involved in quelling	NCU06766.5	Lema_P041660.1	Lm_ibcn84_P006320	Lb_b35_P003086	Lb_j154_P008465	Lb_ibcn65_P005049
QIP	Putative exonuclease protein, involved in quelling	NCU00076.5	Lema_P099110.1	Lm_ibcn84_P003525	Lb_b35_P003275	Lb_j154_P004774	Lb_ibcn65_P007705
MSUD							
SAD-1	RdRP essential for MSUD	NCU02178.5	Lema_P117350.1	Lm_ibcn84_P004184	Lb_b35_P007773	Lb_j154_P010097	Lb_ibcn65_P009184
SAD-2	Essential for MSUD	NCU04294.5	None	None	None	None	None
DNA methylation							
DIM-2	De novo CpN DMT / maintenance CpG DMT	NCU02247.5	Lema_P122420.1	Lm_ibcn84_P001650	Lb_b35_P009122	Lb_j154_P002406	Lb_ibcn65_P000313
HP1	Heterochromatin factor, essential for CpN methylation	NCU04017.5	Lema_P110410.1	Lm_ibcn84_P010723	Lb_b35_P006752	Lb_j154_P002792	Lb_ibcn65_P006432
Chromatin remodelling factors							
HDA6	Histone deacetylase, involved in CpG methylation	NCU00824.5	Lema_P067960.1	Lm_ibcn84_P006089	Lb_b35_P008368	Lb_j154_P007151	Lb_ibcn65_P009022
SIR2	NAD-dependant histone, deacetylase involved in TGS	NCU04737.5	Lema_P055610.1	Lm_ibcn84_P004704	Lb_b35_P004176	Lb_j154_P010415	Lb_ibcn65_P001415
DDM1	SW12 / SNF2-like protein, involved in CpN methylation	NCU03875.5	Lema_P071270.1	Lm_ibcn84_P005837	Lb_b35_P006329	Lb_j154_P004702	Lb_ibcn65_P006051

[a] Abbreviations: MIP, Methylation Induced Premeiotically; RIP, Repeat Induced Point mutation; MSUD, Meiotic Silencing of Unpaired DNA; DMT: DNA Methyltransferase, HMT: Histone Methyltransferase, RdRP: RNA dependent RNA polymerase, TGS: transcriptional gene silencing.
[b] GenBank accession number of *Neurospora crassa* genes.

Supplementary Table S3. Gene model statistics for the *Leptosphaeria* spp. genomes.

	L. maculans 'brassicae'		*L. maculans* 'lepidii'	*L. biglobosa* 'brassicae'	*L. biglobosa* 'canadensis'	*L. biglobosa* 'thlaspii'
	v23.1.3	**WA74**	**IBCN84**	**B3.5**	**J154**	**IBCN65**
No. of predicted genes	12543	10624	11272	11390	11068	11691
Average gene length (bp)	1446.2	1595.6	1480.6	1500.8	1612	1445.5
Median gene length (bp)	1162	1301	1189	1217	1319	1176
Total gene length (Mb)	18.0	17.0	16.7	17.1	17.8	16.9
Average coding gene length (bp)	1258.3	1370.2	1285.5	1307.1	1395.6	1266.5
Median coding gene length (bp)	1002	1125	1035	1065	1146	1026
Total coding gene length (Mb)	15.7	14.6	14.5	14.9	15.4	14.8
No. of exons	35201	29500	29869	30388	30995	30533
Average exon length (bp)	445.7	493.5	485.1	489.9	498.4	485
Median exon length (bp)	214	261	258	267	264	265
No. of genes with introns	9453	8437	8586	8664	8766	8714
Genes with introns (%)	75.4	79.4	76.2	76.1	79.2	74.5
No. of introns	22732	18876	18597	18998	19927	18842
Average intron length (bp)	103.1	126.9	118.3	116.1	120.2	111
Median intron length (bp)	63	59	59	57	57	56
Total intron length (Mb)	2.34	2.39	2.20	2.20	2.39	2.10
Average intergenic region length (bp)	1911.8	1512.9	1276.7	1220.8	1031.2	1092.7
Median intergenic region length (bp)	639	906	878	831	817	812
Total intergenic region length (Mb)	23.8	15.4	14.1	13.3	10.7	12.5

Supplementary Table S4. Chromosomal assignment and correspondance between *L. maculans* 'brassicae' and *L. maculans* 'lepidii'.

	L. maculans 'brassicae'		*L. maculans* 'lepidii'	
Chromosome ID	**Sequence ID**	**Length (bp)**	**Sequence ID**	**Length (bp)**
Chr01	SuperContig_0	4258568	scaffold00002 ;scaffold00008	3378240
Chr02	SuperContig_8 SuperContig_10	3567966	scaffold00004 ;scaffold00016 scaffold00036 ; scaffold00060	2650413
Chr03	SuperContig_1	3378610	scaffold00001	2643178
Chr04	SuperContig_15 SuperContig_12	3192339	scaffold00003 ;scaffold00031 scaffold00041 ;scaffold00046 scaffold00067 ;scaffold00070	2353333
Chr05	SuperContig_13 SuperContig_17	3080273	scaffold00005 ;scaffold00017 scaffold00051 ;scaffold00064	2409061
Chr06	SuperContig_2	2939989	scaffold00012 ;scaffold00018 scaffold00026 ;scaffold00038 scaffold00044 ;scaffold00050 scaffold00063 ;scaffold00078 scaffold00088 ;scaffold00094	2356794
Chr07	SuperContig_20 SuperContig_23 SuperContig_21	2629879	scaffold00006 ;scaffold00033 scaffold00055 ;scaffold00056 scaffold00058	1802772
Chr08	SuperContig_31 SuperContig_3	2491514	scaffold00020 ;scaffold00024 scaffold00032 ;scaffold00034 scaffold00042 ;scaffold00069 scaffold00074 ;scaffold00075	1408145
Chr09	SuperContig_4	1918205	scaffold00011 ;scaffold00025	1466034
Chr10	SuperContig_6	1888674	scaffold00022 ;scaffold00029 scaffold00030 ;scaffold00039 scaffold00053 ;scaffold00059	1218863
Chr11	SuperContig_5	1869450	scaffold00007	1533342
Chr12	SuperContig_11 SuperContig_26	1851700	scaffold00013 ;scaffold00040 scaffold00079	1110430
Chr13	SuperContig_9	1772623	scaffold00009 ;scaffold00099	1358910
Chr14	SuperContig_7	1769547	scaffold00010 ;scaffold00035 scaffold00037	1363430
Chr15	SuperContig_14	1533332	scaffold00015 ;scaffold00023	1343880
Chr16	SuperContig_16	1397653	scaffold00021 ;scaffold00028	854532
Chr17	SuperContig_18	1351976	scaffold00014	907115
Chr18	SuperContig_19	1186800	scaffold00019 ;scaffold00031 scaffold00085	846552
Chr19	SuperContig_32 SuperContig_22	819122	-	-

Supplementary Table S5. Location and distance from TEs of the 30 intrachromosomal inversions in the chromosomes of *L. maculans* 'brassicae'.

L. maculans 'brassicae'							*L. maculans* 'lepidii'						
Sequence ID	**Start**	**End**	**Length (bp)**	**Presence of repeats in the environment**	**Minimal distance from repeats (bp)**	**Repeats Family**	**Sequence ID**	**Start**	**End**	**Length (bp)**	**Presence of repeats in the environment**	**Minimal distance from repeats (bp)**	**Repeats Family**
SuperContig_0	625568	627770	2203	Yes	1	DTX_*Gimli*	scaffold00008	507793	506474	1320	No	-	-
SuperContig_0	1738229	1740435	2207	Yes	201	RLG_*Olly*	scaffold00008	1105390	1103183	2208	No	-	-
SuperContig_0	1779322	1783761	4440	Yes	54	Lmac_Grouper_2395_3	scaffold00008	1110039	1105602	4438	No	-	-
SuperContig_1	660640	674747	14108	No	-	-	scaffold00001	2143507	2157635	14129	No	-	-
SuperContig_2	1484140	1493271	9132	No (but gap)	-	-	scaffold00026	251075	260378	9304	No	-	-
SuperContig_2	2202727	2204123	1397	Yes	555	RLX_*Ayoly*	scaffold00018	414199	415240	1042	No	-	-
SuperContig_2	2219863	2233167	13305	Yes	1	RLX_*Ayoly*	scaffold00018	416169	429356	13188	No	-	-
SuperContig_2	2238979	2257503	18525	Yes	1	DTM_*Sahana*	scaffold00018	429348	447892	18545	No	-	-
SuperContig_3	1086298	1117472	31175	Yes	65	DTX_*Gimli*	scaffold00024	368267	352536	15732	No	-	-
SuperContig_4	293086	298393	5308	No	-	-	scaffold00025	192294	183986	8309	No	-	-
SuperContig_5	320973	348004	27032	Yes	320	DTT_*Goku*	scaffold00007	222047	194749	27299	Yes	59	DTT_*Goku*
SuperContig_6	1011828	1073156	61329	Yes	9	RLX_*Ayoly*	scaffold00022	557314	495566	61749	No	-	-
SuperContig_6	1090646	1140639	49994	Yes	102	RLX_*Ayoly*	scaffold00022	495489	446386	49104	No	-	-
SuperContig_7	1026871	1034040	7170	Yes	12	DTX_*Suno*	scaffold00010	242217	249452	7236	No	-	-
SuperContig_8	357739	359460	1722	No	-	-	scaffold00016	285866	284146	1721	No	-	-
SuperContig_9	611417	623309	11893	No	-	-	scaffold00009	904844	916714	11871	No	-	-
SuperContig_9	1704866	1714625	9760	Yes	155	Lmac_Recon_56_3	scaffold00009	10690	20630	9941	Yes	451	DTX_*Upa*
SuperContig_10	98932	103662	4731	Yes	8	DTM_*Ingwe*	scaffold00036	16945	12142	4804	No	-	-
SuperContig_12	2257	5754	3498	Yes	3	RLG_*Olly*	scaffold00031	20291	16839	3453	No	-	-
SuperContig_12	7101	9515	2415	Yes	18	DTX_*Suno*	scaffold00031	16625	14207	2419	Yes	327	IBCN65-B-R84-Map5
SuperContig_12	1292059	1296156	4098	No	-	-	scaffold00003	69027	72623	3597	No	-	-
SuperContig_14	43755	70780	27026	Yes	25	DTX_*Valwe*	scaffold00015	47586	19880	27707	No	-	-
SuperContig_14	1338993	1342781	3789	No	-	-	scaffold00023	242015	238094	3922	No	-	-
SuperContig_15	431557	441932	10376	No	-	-	scaffold00003	1634669	1644142	9474	No	-	-
SuperContig_16	329025	337618	8594	Yes	13	Lmac_Grouper_284_9	scaffold00028	197212	207221	10010	Yes	32	DTX_*Suno*
SuperContig_17	163074	249435	86362	Yes	79	RLG_*Olly*	scaffold00005	1294385	1380920	86536	No	-	-
SuperContig_17	674495	697280	22786	Yes	1	DTM_*Sahana*	scaffold00005	584507	601801	17295	No	-	-
SuperContig_17	708567	1063786	355220	Yes	1	DTM_*Sahana*	scaffold00005	601793	928931	327139	No	-	-
SuperContig_18	722405	743522	21118	No	-	-	scaffold00014	391380	413537	22158	No	-	-
SuperContig_20	228029	231090	3062	Yes	94	RLG_*Olly*	scaffold00006	17451	14377	3075	No	-	-

Supplementary Table S6. Characteristics of Class I (Retrotransposons) Transposable Elements identified in the *Leptosphaeria maculans-L. biglobosa* species complex.

Order	Superfamily	Name	Size (bp)	LTR size (bp)	Identified in	Specificity
LTR	Ty1-Copia	RLC_*Chaozu-1*_Lbb	3685	-	*L. biglobosa* 'brassicae'	
LTR	Ty1-Copia	RLC_*Chaozu-2*_Lbb	2574	-	*L. biglobosa* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
LTR	Ty1-Copia	RLC_*Chaozu*_Lbt	6295	-	*L. biglobosa* 'thlaspii'	
LTR	Ty1-Copia	RLC_*Tenshin*_Lbb	5258	214	*L. biglobosa* 'brassicae'	*Pleosporales*
LTR	Ty1-Copia	RLC_*Tenshin*_Lbt	2457	-	*L. biglobosa* 'brassicae'	
LTR	Ty1-Copia	RLC_*Zolly*_Lml	4525	-	*L. maculans* 'lepidii'	
LTR	Ty1-Copia	RLC_*Zolly-1*_Lmb	5306	177	*L. maculans* 'brassicae'	*Pleosporales*
LTR	Ty1-Copia	RLC_*Zolly-2*_Lmb	5306	177	*L. maculans* 'brassicae'	
LTR	Ty1-Copia	RLC_*Gohan*_Lbb	7820	180	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
LTR	Ty1-Copia	RLC_*Lunch*_Lbt	3955	-	*L. biglobosa* 'thlaspii'	*Pleosporales*
LTR	Ty1-Copia	RLC_*Chichi*_Lml	4888	-	*L. maculans* 'lepidii'	*Pleosporales*
LTR	Ty1-Copia	RLC_*Pholy*_Lmb	6981	281	*L. maculans* 'brassicae'	*Pleosporales*
LTR	Ty3-Gypsy	RLG_*Dolly*_Lbb	4815	-	*L. biglobosa* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Dolly*_Lml	6407	231	*L. maculans* 'lepidii'	*Dothideomycetes*
LTR	Ty3-Gypsy	RLG_*Dolly*_Lmb	6620	228	*L. maculans* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Olly*_Lbb	7313	279	*L. biglobosa* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Olly*_Lbt	7375	260	*L. biglobosa* 'thlaspii'	*Leptosphaeriaceae / Pleosporaceae*
LTR	Ty3-Gypsy	RLG_*Olly*_Lml	6035	-	*L. maculans* 'lepidii'	
LTR	Ty3-Gypsy	RLG_*Olly*_Lmb	7239	250	*L. maculans* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Polly*_Lbt	7035	187	*L. biglobosa* 'thlaspii'	*Pleosporales*
LTR	Ty3-Gypsy	RLG_*Polly*_Lmb	6928	179	*L. maculans* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Pilaf*_Lbt	6906	-	*L. biglobosa* 'thlaspii'	*Pleosporales*
LTR	Ty3-Gypsy	RLG_*Pilaf*_Lml	7239	113	*L. maculans* 'lepidii'	
LTR	Ty3-Gypsy	RLG_*Brawly*_Lml	12438	150	*L. maculans* 'lepidii'	*L. maculans*
LTR	Ty3-Gypsy	RLG_*Brawly*_Lmb	7289	-	*L. maculans* 'brassicae'	
LTR	Ty3-Gypsy	RLG_*Drum*_Lbb	7058	168	*L. biglobosa* 'brassicae'	*L. biglobosa 'brassicae'*
LTR	Ty3-Gypsy	RLG_*Piccolo*_Lbb	12145	170	*L. biglobosa* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
LTR	Ty3-Gypsy	RLG_*Piano*_Lbb	2561	-	*L. biglobosa* 'brassicae'	*Pleosporales*
LTR	Ty3-Gypsy	RLG_*Oolong*_Lbt	6731	271	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
LTR	Ty3-Gypsy	RLG_*Mai*_Lml	6786	-	*L. maculans* 'lepidii'	*Leptosphaeriaceae / Pleosporaceae*
LTR	Ty3-Gypsy	RLG_*Shu*_Lml	8224	174	*L. maculans* 'lepidii'	*Leptosphaeriaceae / Pleosporaceae*
LTR	Ty3-Gypsy	RLG_*Rolly*_Lmb	11894	235	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
LINE	I	RII_*Yajirobe*_Lbt	2054	20	*L. biglobosa* 'thlaspii'	*Leptosphaeriaceae / Pleosporaceae*
LINE	I	RII_*Karin*_Lbt	6610	31	*L. biglobosa* 'thlaspii'	*Leptosphaeria* species complex
LTR	Unknown	RIx_*Ayoly*_Lmb	10397	217	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
LTR	Unknown	RLx_*Jolly*_Lmb	796	259	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
PLE	Penelope	RPP_*Circe*_Lmb	7219	-	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'

Supplementary Table S7. Characteristics of Class II (DNA transposons) Transposable Elements identified in the *Leptosphaeria maculans-L. biglobosa* species complex.

Order	Superfamily	Name	Size (bp)	TIR size (bp)	Identified in	Specificity
TIR	hAT	DTA_*Kami*_Lbt	1806	-	*L. biglobosa* 'thlaspii'	*Pleosporales*
TIR	hAT	DTA_*Kami*_Lbb	6099	-	*L. biglobosa* 'brassicae'	
TIR	Mutator	DTM_*Sahana*_Lml	5810	-	*L. maculans* 'lepidii'	*L. maculans*
TIR	Mutator	DTM_*Sahana*_Lmb	5870	-	*L. maculans* 'brassicae'	
TIR	Mutator	DTM_*Roshi*_Lbb	6574	52	*L. biglobosa* 'brassicae'	*Leptosphaeria* species complex
TIR	Mutator	DTM_*Mutaito*_Lbb	3274	-	*L. biglobosa* 'brassicae'	*L. biglobosa*
TIR	Mutator	DTM_*Lenwe*_Lmb	3489	49	*L. maculans* 'brassicae'	*Pleosporales*
TIR	Mutator	DTM_*Ingwe*_Lmb	3582	37	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
TIR	Tc1-Mariner	DTT_*Yamcha*_Lml	1926	82	*L. maculans* 'lepidii'	*Leptosphaeriaceae / Pleosporaceae*
TIR	Tc1-Mariner	DTT_*Molly*_Lbt	1865	74	*L. biglobosa* 'thlaspii'	*Pleosporales*
TIR	Tc1-Mariner	DTT_*Molly*_Lbb	1861	71	*L. biglobosa* 'brassicae'	
TIR	Tc1-Mariner	DTT_*Krilin*_Lbt	2141	49	*L. biglobosa* 'thlaspii'	*Dothideomycetes*
TIR	Tc1-Mariner	DTT_*Krilin*_Lbb	2135	55	*L. biglobosa* 'brassicae'	
TIR	Tc1-Mariner	DTT_*Goku-2*_Lbt	1112	-	*L. biglobosa* 'thlaspii'	*Pleosporales*
TIR	Tc1-Mariner	DTT_*Goku-1*_Lbt	1422	-	*L. biglobosa* 'thlaspii'	
TIR	Tc1-Mariner	DTT_*Goku*_Lml	1889	75	*L. maculans* 'lepidii'	
TIR	Tc1-Mariner	DTT_*Goku*_Lbb	1889	67	*L. biglobosa* 'brassicae'	
TIR	Tc1-Mariner	DTT_*Finwe-3*_Lml	803	-	*L. maculans* 'lepidii'	*Leptosphaeriaceae / Pleosporaceae*
TIR	Tc1-Mariner	DTT_*Finwe-3*_Lmb	806	29	*L. maculans* 'brassicae'	
TIR	Tc1-Mariner	DTT_*Finwe-2*_Lmb	523	29	*L. maculans* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
TIR	Tc1-Mariner	DTT_*Finwe-1-2*_Lml	707	-	*L. maculans* 'lepidii'	
TIR	Tc1-Mariner	DTT_*Finwe-1*_Lmb	529	29	*L. maculans* 'brassicae'	
TIR	Tc1-Mariner	DTT_*Bulma*_Lml	1859	68	*L. maculans* 'lepidii'	*Pleosporales*
TIR	Tc1-Mariner	DTT_*Bulma*_Lbb	1854	77	*L. biglobosa* 'brassicae'	
TIR	Unknown	DTX_*Suno*_Lml	2105	54	*L. maculans* 'lepidii'	*Leptosphaeriaceae / Pleosporaceae*
TIR	Unknown	DTX_*Kinto*_Lbb	5446	24	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
TIR	Unknown	DTX_*Valwe*_Lmb	1792	37	*L. maculans* 'brassicae'	*Leptosphaeria* species complex
TIR	Unknown	DTX_*Olwe*_Lmb	866	49	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
TIR	Unknown	DTX_*Gimli*_Lmb	603	-	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
TIR	Unknown	DTF_*Elwe*_Lmb	2173	57	*L. maculans* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
MITE	Unknown	DTX_*Upa*_Lml	441	127	*L. maculans* 'lepidii'	*Pleosporales*
MITE	Unknown	DTX_*Arale*_Lbb	403	83	*L. biglobosa* 'brassicae'	*L. biglobosa*

Supplementary Table S8. Insertion polymorphism of Transposable Elements between the two isolates of *L. maculans* 'brassicae'.

Family	No. of occurrences in Lmb v23.1.3 genome	No. of occurrences analyzed	No. of non polymorphic insertions	% of non polymorphic insertions	No. of polymorphic insertions	% of polymorphic insertions
DTM_*Sahana*	195	167	20	12,0	147	88,0
DTx_*Gimli*	279	114	98	86,0	16	14,0
DTF_*Elwe*	158	41	31	75,6	10	24,4
DTx_*Valwe*	73	28	24	85,7	4	14,3
RLC_*Zolly*	115	27	16	59,3	11	40,7
RLG_*Brawly*	22	16	16	100,0	0	0,0
RLx_*Ayoly*	164	14	9	64,3	5	35,7
DTx_*Olwe*	15	8	5	62,5	3	37,5
DTM_*Ingwe*	48	7	7	100,0	0	0,0
DTM_*Lenwe*	36	6	5	83,3	1	16,7
DTT_*Finwe*	53	5	4	80,0	1	20,0
RLC_*Pholy*	1020	5	1	20,0	4	80,0
RLG_*Olly*	1085	2	2	100,0	0	0,0
RLG_*Polly*	1014	2	1	50,0	1	50,0
RLG_*Rolly*	594	1	0	0,0	1	100,0
DTx_*Kinto*	-	1	1	100,0	0	0,0
RII_*Karin*	-	1	1	100,0	0	0,0
RLx_*Jolly*	57	1	1	100,0	0	0,0
RLG_*Drum*	-	1	1	100,0	0	0,0

Supplementary Table S9. Predicted SSP-encoding genes conservation between the genomes of isolates of the *L. maculans-L. biglobosa* species complex.

		L. maculans **'brassicae'**	***L. maculans*** **'lepidii'**	***L. biglobosa*** **'brassicae'**	***L. biglobosa*** **'canadensis'**	***L. biglobosa*** **'thlaspii'**
Sequences clusterised using orthoMCL	No. of orthologs in all species (Core genome)	219	212	220	211	220
	No. of orthologs in at least one other species	138	172	185	182	147
	sub-total :	**357**	**384**	**405**	**393**	**367**
Unclusterised sequences	No. of sequences absent in all other species	125	142	75	75	163
	No. of sequences in at least one other species	117	135	146	114	101
	No. of sequences with unresolved absence or presence in other species	52	76	39	39	45
	sub-total :	**294**	**353**	**260**	**228**	**309**
	Total :	**651**	**737**	**665**	**621**	**676**

Supplementary Table S10. Conservation of genes encoding avirulence effector in members of the *L. maculans-L. biglobosa* species complex and other fungal species.

ID	Specificity	Presence in	Homologs IDs	SSP	Description	% of identity	% of coverage
AvrLm11	*Leptosphaeriaceae*	*L. biglobosa* 'thlaspii'	Lb_ibcn65_P001030	Yes	Hypothetical protein	35	100
AvrLmMex	*Leptosphaeriaceae*	*L. biglobosa* 'brassicae'	Lb_b35_P010091	Yes	Hypothetical protein	33	97
		L. biglobosa 'canadensis'	Lb_j154_P003932	Yes	Predicted protein	36	97
		L. biglobosa 'thlaspii'	Lb_ibcn65_P010997	Yes	Hypothetical protein	38	98
AvrLm1	*Dothideomycetes*	*L. biglobosa* 'thlaspii'	Lb_ibcn65_P011530	Yes	Hypothetical protein	33	85
		P. teres f. teres	PTT_01492	Yes	Hypothetical protein	36	93
		P. teres f. teres	PTT_15034	Yes	Hypothetical protein	30	93
AvrLm4-7	*Dothideomycetes*	*L. maculans* 'lepidii'	Lm_ibcn84_P010674	Yes	Hypothetical protein	32	100
		L. biglobosa 'brassicae'	Lb_b35_P004432	Yes	Predicted protein	40	97
		L. biglobosa 'canadensis'	Lb_j154_P004573	Yes	Predicted protein	41	100
		L. biglobosa 'thlaspii'	Lb_ibcn65_P005148	Yes	Predicted protein	28	71
		M. phaseolina	MPH_11524	Yes	Hypothetical protein	32	73
		M. phaseolina	MPH_00189	Yes	Hypothetical protein	29	69
AvrLm6	*Leotiomycetes*	*L. maculans* 'lepidii'	Lm_ibcn84_P001696	Yes	Hypothetical protein	41	91
		L. biglobosa 'thlaspii'	Lb_ibcn65_P009986	Yes	Hypothetical protein	66	99
		C. gloeosporioides	CGGC5_9879	Yes	AvrLm6 protein	38	83
		C. gloeosporioides	CGGC5_5005	Yes	AvrLm6 protein	35	73
		C. higginsianum	CH_063_00307	Yes	Hypothetical protein	37	84
		F. oxysporum	FOXB_15741	No	Hypothetical protein	33	90
Lema_P086540.1	*Leotiomycetes*	*L. maculans* 'lepidii'	Lm_ibcn84_P000531	Yes	Hypothetical protein	52	99
		L. biglobosa 'thlaspii'	Lb_ibcn65_P011645	Yes	Predicted protein	63	51
		C. heterostrophus	COCHEDRAFT_1148544	Yes	Hypothetical protein	40	95
		F. oxysporum	FOXB_07083	No	Hypothetical protein	41	83
		F. oxysporum f. sp. lycopersici	Six5	Yes	Six5	37	91
		C. gloeosporioides	CGGC5_11832	Yes	Six5	39	98

Supplementary Table S11. Non-Ribosomal Peptide Synthases (NPS) genes of the *L. maculans-L. biglobosa* species complex.

	L. maculans 'brassicae'		*L. maculans* 'lepidii'		*L. biglobosa* 'thlaspii'		*L. biglobosa* 'canadensis'		*L. biglobosa* 'brassicae'		
NPS	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Predicted end product**
SirP	P081810	100% (Lmb)	P008587	90% (Lmb)	No match	n/a	No match	n/a	No match	n/a	Sirodesmin
Maa12	P051070	100% (Lmb)	P004314	98% (Lmb)	P008462	94% (Lmb)	P008547	94% (Lmb)	P009624	94% (Lmb)	
NPS1	P084060	100% (Lmb)	P001896	99% (Lmb)	P005290	90% (Lmb)	P006181	93% (Lmb)	P008091	93% (Lmb)	
NPS2	P073200	100% (Lmb)	P011015	90% (Lmb)	P011354	70% (Lmb)	P005017	72% (Lmb)	P002833	72% (Lmb)	siderophore
NPS3	P063290	100% (Lmb)	P005646	91% (Lmb)	P004377	75% (Lmb)	P001028	74% (Lmb)	P010129	74% (Lmb)	
NPS4	P009040	100% (Lmb)	P001325	94% (Lmb)	P007470	61% (Lmb)	P006735	55% (Lmb)	P001549	83% (Lmb)	
NPS5	P002700	100% (Lmb)	P004994	64% (Lmb)	No match	n/a	No match	n/a	No match	n/a	
NPS6	P070690	100% (Lmb)	P005868	92% (Lmb)	P006085	87% (Lmb)	P009888	86% (Lmb)	P006298	96% (Lmb)	siderophore
NPS7	P067200	100% (Lmb)	P005267	92% (Lmb)	P006945	82% (Lmb)	P007696	80% (Lmb)	P010331	92% (Lmb)	
NPS8	P072890	100% (Lmb)	No match	n/a	No match	n/a	No match	n/a	No match	n/a	Phomalide
Lys2	P039690	100% (Lmb)	P010114	95% (Lmb)	P004848	95% (Lmb)	P004295	92% (Lmb)	P008702	92% (Lmb)	Lysine
NPS10	P126000	100% (Lmb)	P011137	93% (Lmb)	No match		P003164	91% (Lmb)	P010286	91% (Lmb)	
NPS11	P012390	100% (Lmb)	P001023	94% (Lmb)	P009645	84% (Lmb)	P004216	84% (Lmb)	P001255	84% (Lmb)	
NPS12	P065030	100% (Lmb)	P005485	94% (Lmb)	P004207	79% (Lmb)	P011023	75% (Lmb)	P004357	78% (Lmb)	
NPS13	No match	n/a	No match	n/a	No match		P002577	100% (Lbc)	P006247	92% (Lbc)	
NPS14	No match	n/a	No match	n/a	P001625	100% (Lbt)	No match	n/a	No match	n/a	
NPS15	No match	n/a	No match	n/a	P000090	100% (Lbt)	No match	n/a	No match	n/a	

Supplementary Table S12. Polyketide Synthases (PKS) genes of the *L. maculans-L. biglobosa* species complex (1/2).

	L. maculans 'brassicae'		*L. maculans* 'lepidii'		*L. biglobosa* 'thlaspii'		*L. biglobosa* 'canadensis'		*L. biglobosa* 'brassicae'		
PKS	**Gene**	**Identity (spp c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Predicted end product**
PKS1	P081920	100 % (Lmb)	No match	n/a	No match	n/a	No match	n/a	P007270	89% (Lmb)	
PKS2	P002660	100 % (Lmb)	P004990	96% (Lmb)	No match	n/a	No match	n/a	No match	n/a	Phomenoic acid
PKS3	P006610	100 % (Lmb)	P001547	85% (Lmb)	P000212	62% (Lmb)	P001840	60% (Lmb)	P001772	60% (Lmb)	
PKS4	P017310	100 % (Lmb)	P000558	91% (Lmb)	P000902	75% (Lmb)	P000246	74% (Lmb)	P000540	73% (Lmb)	
PKS5	P087730	100 % (Lmb)	No match	n/a	No match	n/a	No match	n/a	No match	n/a	
PKS6	P086720	100 % (Lmb)	P010582	87% (Lmb)	No match	n/a	No match	n/a	No match	n/a	
PKS7	P082310	100 % (Lmb)	P010688	91% (Lmb)	P010608	79% (Lmb)	P001714	80% (Lmb)	P010565	81% (Lmb)	
PKS8	P098040	100 % (Lmb)	P009766	67% (Lmb)	P008727	56% (Lmb)	P007276	52% (Lmb)	P006101	55% (Lmb)	
PKS9	P098070	100 % (Lmb)	P009768	48% (Lmb)	P008724	34% (Lmb)	No match	n/a	No match	n/a	
PKS10	P098490	100 % (Lmb)	P009804	98% (Lmb)	P001467	91% (Lmb)	P009788	93% (Lmb)	P005709	93% (Lmb)	Melanin
PKS11	P117970	100 % (Lmb)	P004141	96% (Lmb)	No match	n/a	No match	n/a	No match	n/a	
PKS12	P000820	100 % (Lmb)	P004838	92% (Lmb)	No match	n/a	No match	n/a	No match	n/a	
PKS13	P057720	100 % (Lmb)	P008173	91% (Lmb)	No match	n/a	No match	n/a	No match	n/a	
PKS14	P057650	100 % (Lmb)	P008166	96 % (Lmb)	No match	n/a	No match	n/a	No match	n/a	
PKS15	P030270	100 % (Lmb)	P008813	93 % (Lmb)	P009926	65% (Lmb)	P007952	62% (Lmb)	P007989	67% (Lmb)	Chalcone-synthase-like

Supplementary Table S12. Polyketide Synthases (PKS) genes of the *L. maculans-L. biglobosa* species complex (2/2).

	L. maculans 'brassicae'		*L. maculans* 'lepidii'		*L. biglobosa* 'thlaspii'		*L. biglobosa* 'canadensis'		*L. biglobosa* 'brassicae'		
PKS	**Gene**	**Identity (spp c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Gene**	**Identity (spp. c.f.)**	**Predicted end product**
PKS16	No match	n/a	P005350	100 % (Lml)	No match	n/a	P005809	88% (Lml)	No match	n/a	
PKS17	No match	n/a	No match	n/a	P007094	93% (Lbb)	P004061	93% (Lbb)	P004717	100% (Lbb)	
PKS18	No match	n/a	No match	n/a	P006653	85% (Lbb)	P008796	94% (Lbb)	P006589	100% (Lbb)	
PKS19	No match	n/a	No match	n/a	P006652	86% (Lbc)	P008795	100% (Lbc)	No match	n/a	
PKS20	No match	n/a	No match	n/a	No match	n/a	P002736	94% (Lbb)	P010023	100% (Lbb)	
PKS21	No match	n/a	No match	n/a	No match	n/a	No match	n/a	P010271	100% (Lbb)	98% identity with *Arthroderma otae* PKS
PKS22	No match	n/a	No match	n/a	No match	n/a	No match	n/a	P010859	100% (Lbb)	
PKS23	No match	n/a	No match	n/a	No match	n/a	P007539	92% (Lbb)	P010864	100% (Lbb)	
PKS24	No match	n/a	No match	n/a	No match	n/a	No match	n/a	P010890	100% (Lbb)	
PKS25	No match	n/a	No match	n/a	No match	n/a	P003162	100% (Lbc)	No match	n/a	
PKS26	No match	n/a	No match	n/a	P001622	100% (Lbt)	No match	n/a	No match	n/a	
PKS27	No match	n/a	No match	n/a	P010112	100% (Lbt)	No match	n/a	No match	n/a	
PKS28	No match	n/a	No match	n/a	P011164	100% (Lbt)	No match	n/a	No match	n/a	
PKS29	No match	n/a	No match	n/a	P011526	100% (Lbt)	No match	n/a	No match	n/a	
PKS30	No match	n/a	No match	n/a	P011544	100% (Lbt)	No match	n/a	No match	n/a	
PKS31	No match	n/a	No match	n/a	P000089	100% (Lbt)	No match	n/a	No match	n/a	

Supplementary data

Supplementary Data 1. Unclassified repeated elements in species of the *L. maculans-L. biglobosa* species complex (1/2).

Name	Size (bp)	Identified in	Specificity
B35_element11	3139	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35_element17	7681	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35_element21	6382	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35_element22	647	*L. biglobosa* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
B35-B-R50-Map3	1989	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35-B-R56-Map3	1560	*L. biglobosa* 'brassicae'	*Leptosphaeriaceae / Pleosporaceae*
B35-B-R59-Map7	3462	*L. biglobosa* 'brassicae'	*Pleosporales*
B35-B-R6-Map8	764	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35-B-R61-Map12	5044	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35-B-R67-Map6	3538	*L. biglobosa* 'brassicae'	*L. biglobosa* 'brassicae'
B35-B-R89-Map3	1320	*L. biglobosa* 'brassicae'	*Pleosporales*
IBCN65_element10	4378	*L. biglobosa* 'thlaspii'	*L. biglobosa*
IBCN65_element13	6963	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65_element19	7538	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65_element21	1702	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65_element29	10009	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65_element33	8012	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-G206-Map3	2010	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-R33-Map5	821	*L. biglobosa* 'thlaspii'	*Leptosphaeria* species complex
IBCN65-B-R38-Map14	2304	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-R48-Map5	2909	*L. biglobosa* 'thlaspii'	*Leptosphaeria* species complex
IBCN65-B-R52-Map5	2427	*L. biglobosa* 'thlaspii'	*Pleosporales*
IBCN65-B-R59-Map5	2203	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-R80-Map4	1053	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-R84-Map5	1158	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'

Supplementary Data 1. Unclassified repeated elements in species of the *L. maculans-L. biglobosa* species complex (2/2).

Name	Size (bp)	Identified in	Specificity
IBCN65-B-R94-Map4	1281	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN65-B-R98-Map3	571	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'thlaspii'
IBCN84_element16	2263	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84_element17	3539	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R30-Map5	1737	*L. maculans* 'lepidii'	Pleosporales
IBCN84-B-R37-Map3	1049	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R41-Map20	7089	*L. maculans* 'lepidii'	*L. maculans*
IBCN84-B-R52-Map5	1022	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R58-Map4	1772	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R59-Map10	2738	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R8-Map4	2035	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R84-Map7	3595	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R87-Map7	3592	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
IBCN84-B-R9-Map7	3027	*L. maculans* 'lepidii'	*L. maculans* 'lepidii'
Lmac_Grouper_1227_4	917	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_1517_5	967	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_1829_15	1088	*L. maculans* 'brassicae'	*Leptosphaeria* species complex
Lmac_Grouper_2395_3	1425	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_284_9	674	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_498_20	743	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_595_20	776	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Grouper_69_20	616	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Recon_56_3	10880	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Recon_83_20	10940	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
Lmac_Recon_92_20	6256	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
LmTelo1_Lmb	18114	*L. maculans* 'brassicae'	*L. maculans* 'brassicae'
LmTelo2_Lmb	15981	*L. maculans* 'brassicae'	*L. maculans*

Supplementary Data 2. Presence of the repeat families in dothideomycete genomes (1/4). The values represent the percentage of coverage of the longest copy of a repeat family annotated by RepeatMasker in the genome. Cells colored in orange indicate that this percentage is >= 30% and < 50%. Cells colored in green indicate that this percentage is >= 50%.

ID	*A. brassicicola*	*C. heterostrophus*	*C. sativus*	*P. fijiensis*	*P. tritici-repentis*	*P. teres*	*S. turcica*	*P. nodorum*	*Lmb JN3*	*Lmb WA74*	*Lml IBCN84*	*Lbb B3.5*	*Lbc J154*	*Lbt IBCN65*
DTM_Roshi_Lbb			0,68	1,25				12,88	75,33	71,98		97,82	0,81	1,57
DTx_Valwe	4,97			7,09	2,4			11,1	49,27	50,06	24,27	8,43	100	4,74
IBCN65-B-R33-Map5				7,43		9,38					99,27			100
IBCN65-B-R48-Map5	6,33	29,32		3,16			5,26				54,42	1,89		100
Lmac_Grouper_1829_15			23,25			14,34			71,51	64,15	14,89			69,58
RII_Karin_Lbt	48,94	39,67	17,94	2,53	40,77	21,07	31,91	0,86	60,89	80,32	3,25	79,27	4,93	99,52
B35_element22	48,53	62,44	44,2	24,57		39,1	56,11	13,6	63,21	61,82	57,96	100	99,85	56,11
B35-B-R56-Map3		80,45	5,13	3,59	5,13	4,74	5,13	2,95				100	100	
DTF_Elwe	14,04	1,89	41,19	23,38	100	16,89	48,27	4,6	100	100				3,27
DTT_Finwe-1	33,08	36,86	58,22		67,11	40,83	58,03	22,12	99,81	99,81	43,67	32,89	58,22	
DTT_Finwe-1-2_Lml	100	99,58	99,72	5,23	100	98,16	100		99,86	99,86	100	81,47	99,86	
DTT_Finwe-2		99,81	8,99	26,96	51,43	51,82	47,42		100	100	25,05			
DTT_Finwe-3	100	99,75	97,27		99,88	98,51	99,5		100	100	30,15	86,48	86,48	9,06
DTT_Finwe-3_Lml	21,42	85,31	96,26	9,22	96,26	96,89	100	14,32	72,73	72,73	100	85,06		
DTT_Yamcha_Lml	15,73	64,43	33,64	30,06			16,41	20,46			69	1,92		1,71
DTX_Suno_Lml	80,24	2,8	18,43	4,09	100	81,33	25,32	13,59	51,83	48,08	100	4,13	37,81	4,99
RII_Yajirobe_Lbt		55,99	64,95	14,07	48,69	17,14	95,23				99,85			100
RLC_Chaozu_Lbt	1,45			2,07	46,08	32,88		1,46	1,3			92,55	1,48	87,72
RLC_Chaozu-1_Lbb				2,74	10,23	16,23						87,73	8,17	19,86
RLC_Chaozu-2_Lbb				2,95	89,16	32,32						100		
RLC_Lunch_Lbt	3,72		87,64	10,19	97,65	29,94	94,24	85,18			1,62		2,5	100
RLG_Mai_Lml	64,74	59,99	46,7	41,08	47,26	18,66	50,34	41,39	2,61	2,62	94,52	50,49	8,56	3,43
RLG_Olly	2,78	5,19	18,3	15,38	4,88	8,81	9,74	2,06	100	100	6,29	2,36	3,94	7,68
RLG_Olly_Lbb	2,73	6,03	76,9	17,42	10,12	5,73	76,47	2,09	72,38	54,7	4,68	87,27	25,91	11,8
RLG_Olly_Lbt	9,48	75,93	75,72	14,96	7,29	4,85	45,32	8,14	76,05	5,68	41,4	21,03	10,89	100
RLG_Olly_Lml	18,59	81,87	79,72	20,25	3,74	4,46	17,18	11,07		49,53	63,02	27,19	2,57	
RLG_Piccolo_Lbb	52,24	4,24	0,98	8,42	5,18	9,85	20,96	19,79	1,74	1,74	1,44	49,31	13,22	26,33
RLG_Shu_Lml	6,2	4,84	9,2	16,52	7,22	8,12	51,96	6,61	1,41	0,67	100	1,82		0,67
B35-B-R59-Map7	37,81	4,02	19,44	5,81	10,72	74,52	55,78	52,69				78,54	21,14	5,52
B35-B-R89-Map3	88,94	92,5	88,86	15,45	46,97	13,33	85,91	69,39	19,55	19,55	9,09	100	39,47	17,35
DTA_Kami_Lbb	35,69	36,87	31,12	5,07	37,74	20,04	39,78	40,92	89,1	89,15	8,46	89,21	85,72	38,24
DTA_Kami_Lbt	99,56	99,94			7,64	45,29	70,27	96,73	100	82,39	8,25		5,76	100
DTM_Lenwe		1,09		5,79		1,83	59,3	66,98	99,97	95,84	4,87	1,69	1,46	15,79
DTT_Bulma_Lbb	100	94,55	100	3,18	100	27,24	99,95	99,95			32,2	100	33,93	
DTT_Bulma_Lml	99,62	99,95	99,89	10,06	99,68	99,62	99,95	99,78		5,49	99,95		48,52	27,33

Supplementary Data 2. Presence of the repeat families in dothideomycete genomes (2/4).

ID	*A. brassicicola*	*C. heterostrophus*	*C. sativus*	*P. fijiensis*	*P. tritici-repentis*	*P. teres*	*S. turcica*	*P. nodorum*	*Lmb JN3*	*Lmb WA74*	*Lml IBCN84*	*Lbb B3.5*	*Lbc J154*	*Lbt IBCN65*
DTT_Goku_Lbb	91	99,84	95,08	23,82	99,79	99,79	99,95	99,95	60,93	60,93	61,73	99,95	27,63	99,79
DTT_Goku_Lml	93,49	68,5	99,74	4,92	99,95	74,11	97,51	68,34	35,42	98,57	99,95	36,32	32,5	91,11
DTT_Goku-1_Lbt	99,09	85,86	99,02	13,64	89,17	90,3	99,02	81,86	15,05	21,52	84,53	96,62	42,69	100
DTT_Goku-2_Lbt		19,24	29,5		29,5	29,14	26,08	28,87	9,62	9,53	23,92	60,79	21,49	57,82
DTT_Molly_Lbb	78,99	60,67	77,11	31	58,41	85,38	77,38	100			4,62	100	9,56	99,79
DTT_Molly_Lbt	85,47	19,41	100	6,22	98,39	56,51	100	99,79			2,95	99,73		99,68
DTX_Upa_Lml		12,93			13,15	99,32	22,9	100			100	11,56	11,11	12,93
IBCN65-B-R52-Map5	2,88					2,93		72,35			70,91	2,68		100
IBCN84-B-R30-Map5	7,08	11,11		3,45			25,22	94,42	78,01	95,51	100			
RLC_Chichi_Lml	3,64	1,51	29,3	8,9	54,99	9,14	38,32	55,59	2,09	2,09	57,18			31,91
RLC_Pholy	5,56		56,74	9,27	20,87	7,71	20,04	56,2	100	100	56,45	13,41	44,81	26,9
RLC_Tenshin_Lbb	6,94		74,02	7,4	83,02	54,47	46,23	82,22		9,38		100	5,61	47,85
RLC_Tenshin_Lbt	4,19		48,76	3,5	54,78	23,48		67,93				10,18	7,61	99,96
RLC_Zolly_Lml				4,22		2,87		94,5	13,79	16,88	100			
RLC_Zolly-1	1,49			1,47				27,76	100	100	99,77			
RLC_Zolly-2	0,92			2,04		2,98		6,03	100	100	22,75			11,65
RLG_Piano_Lbb	85,71	69,27	68,88	44,05	74,42	31,2	74,85	67,67			5,31	100	27,45	16,52
RLG_Pilaf_Lbt	17,71	9,09	7,04	11,18	4,85	8,24	42,48	64,28	5,37	5,37	2,2	2,48	60,67	94,85
RLG_Pilaf_Lml	15,07	8,63	7,83	14,96	6,48	1,27	45,67	56,13	4,61	4,27	91,77		0,94	
RLG_Polly	42,36	32,98	16,53	18,24	15,82	7,66	18,78	62,75	100	100	17,35		4,13	8,93
RLG_Polly_Lbt	49,27	8,4	17,97	14,94	14,39	20,14	41,15	64,52	3,41	3,41	2,03	2,83	2,57	95,21
DTT_Krilin_Lbb	99,63	46,42	32,51	50,68	100	70,68	78,59	100			35,88	100	100	89,32
DTT_Krilin_Lbt	100	54,46	46,15	38,49	95,66	55,96	92,76	95,61	3,22	3,22	9,34	99,95	99,95	100
RLG_Dolly	42,85	4,37	2,31	45,02	3,02	2,42	4,43	17,33	99,95	99,94		37,85	7,52	0,68
RLG_Dolly_Lbb	50,24	12,86	6,5	60,93	19,81	7,83	11,86	22,41	93,81	23,26	1,16	57,74	7,68	2,87
RLG_Dolly_Lml	87,5	4,1	20,06	47,25	16,48	8,44	6,04	1,59	1,79	1,55	69,81	2,23	1,67	2,06
DTM_Mutaito_Lbb	2,5			4,12	15,64		39,86	33,14	3,21	9,07	9,59	100	6,44	53,27
DTX_Arale_Lbb												100	99,26	38,46
IBCN65_element10				2,49			1,85				1,62	2,44	89,49	100
B35_element11	1,94			2,61		6,05	8,19					100	11,76	29,72
B35_element17	1,31			2,07		14,27		7,79	0,94	0,94		78,7	1,65	7,72
B35_element21		1,18	4	1,55	2,21	1,79	0,69	37,32				90,03	19,77	0,71
B35-B-R50-Map3	26,29	26,14	26,09	2,87	26,09	25,34			25,34	28,56	17,85	100	10,76	21,87
B35-B-R6-Map8				7,46								100	25,65	
B35-B-R61-Map12		1,33		1,84				1,53				98,65	19,55	

Supplementary Data 2. Presence of the repeat families in dothideomycete genomes (3/4).

ID	*A. brassicicola*	*C. heterostrophus*	*C. sativus*	*P. fijiensis*	*P. tritici-repentis*	*P. teres*	*S. turcica*	*P. nodorum*	*Lmb JN3*	*Lmb WA74*	*Lml IBCN84*	*Lbb B3.5*	*Lbc J154*	*Lbt IBCN65*
B35-B-R67-Map6	10,83	10,6	10,6	5,68	10,6	11,33	5,12	7,41	15,52	15,52	15,52	87,82	14,92	15,52
DTX_Kinto_Lbb	32,06	8,94	4,11	5,51	33,6	23,91	13,22	33,22	3,32	3,32	1,95	87,62	29,69	40,18
RLC_Gohan_Lbb			7,57	4,78	3,71	2,21	11,82	8,16	2,17	2,17	10,36	100	13,87	2,44
RLG_Drum_Lbb	6,67	9,42	5,71	19,04	9,86	5,04	9,55	13,09	19,79	12,67	3,54	100	18,42	41,19
IBCN65_element19	1,43			3,46		0,77		4,39	3,13	3,13	3,05		34,98	92,7
IBCN65_element21		20,09	14,22	15,75		18,39	18,16	5,29	9,05	7,52		17,74	15,45	100
IBCN65_element29	0,69	1,07		3,52	1,3	1,53	0,96	2,11	2,38	0,56		0,57	34,6	99,05
IBCN65_element33		2,25	0,45	1,76		3,46	4,39	3,32						62,34
IBCN65-B-G206-Map3	2,99		1,74	3,18								6,67	6,62	100
IBCN65-B-R38-Map14							3,56							91,54
IBCN65-B-R59-Map5				6,54		12,12					9,94		13,03	100
IBCN65-B-R80-Map4				3,51		12,82								100
IBCN65-B-R84-Map5			28,24						42,14	28,84	47,15	49,05	49,05	100
IBCN65-B-R94-Map4	3,43					27,24								100
IBCN65-B-R98-Map3													46,23	100
RLG_Oolong_Lbt	35,08	14,81	18,79	14,26	22,4	14,32	31,33	1,25	1,19	1,46	10,38	1,69	7,06	54,4
DTM_Sahana	1,02			2,47		1,75		2,3	100	100	50,72	2,44	2,56	
DTM_Sahana_Lml	2,75			1,94				3,41	100	100	86,49	27,66		1,48
IBCN65_element13				2,59		4,21		4,19	0,59	0,59	21,04	0,67	5,17	63,59
IBCN84-B-R41-Map20	1,34	1,23		2,29			0,78	5,54	1,4	99,44	74,14	1,51	4,15	38,5
LmTelo2	0,57	1,36	0,88	0,73		2,93	0,96	9,54	91,51	85,92	82,62	0,36	2,97	17,73
RLG_Braw ly				2,13	2,44		18,05	2,5	99,99	94,62	1,22	5,39		2,57
RLG_Braw ly_Lml		4,47		2,27		0,55	5,56	0,87	81,08	40,54	99,98		7,07	3,17
DTM_Ingw e	2,15	2,99	4,41	7,82	3,63	2,71	20,35	6,7	99,97	99,97	2,54	4,13	5	
DTx_Gimli				12,44			14,1		100	100	44,94			
DTx_Olw e									100	100	10,51			

Supplementary Data 2. Presence of the repeat families in dothideomycete genomes (4/4).

ID	*A. brassicicola*	*C. heterostrophus*	*C. sativus*	*P. fijiensis*	*P. tritici-repentis*	*P. teres*	*S. turcica*	*P. nodorum*	*Lmb JN3*	*Lmb WA74*	*Lml IBCN84*	*Lbb B3.5*	*Lbc J154*	*Lbt IBCN65*
Lmac_Grouper_1227_4	19,19			26,5			7,09		100	100			7,42	18,43
Lmac_Grouper_1517_5	11,79			11,48					73,22	100				12,2
Lmac_Grouper_2395_3			2,81	14,88					100	100			2,95	
Lmac_Grouper_284_9				11,42					99,41	99,55	13,5			11,57
Lmac_Grouper_498_20									44,15	45,22				
Lmac_Grouper_595_20									90,34	97,94				
Lmac_Grouper_69_20				9,25					100	99,19	12,01			11,36
Lmac_Recon_56_3	9,76	5,97	10,27	1,67	10,6	10,6	8,44	7,7	88,9	70,59	25,48	44,82	23,05	43,53
Lmac_Recon_83_20	29,25	1,64	0,95	1,93			1,67	0,85	53,94	51,78	2,29	4,52	1,18	2,01
Lmac_Recon_92_20	17,23	2,13	2,41	3,58	2,3	2,65	2,48	4,01	60,73	63,51	2,22	2,37	2,88	2,3
LmTelo1	0,86	0,52		0,97			0,47	17,81	99,66	92,74	1,66	0,72	1,11	
RLG_Rolly	0,92	16,1		9,12	9,29	8,91	26,64	13,69	100	100	14,07	1,16		21,36
RLx_Ayoly				1,64	0,87			1,77	99,97	99,96	2,25			
RLx_Jolly		7,41	17,84	12,94			16,96		100	100	34,67			
RPP_Circe				30,9					57,31	28,4	1,04			
IBCN84_element16	23,6	10,56		10,3			25,01	31,24	7,95	8,09	97,08		3,45	
IBCN84_element17				3,16				2,71			60,05		2,32	
IBCN84-B-R37-Map3											100		2,67	
IBCN84-B-R52-Map5		14,68		11,84	14,09	18					100			
IBCN84-B-R58-Map4				6,43		27,54		37,87			100		5,19	
IBCN84-B-R59-Map10				3,87				16,36			95,14		6,39	42
IBCN84-B-R8-Map4	3,78	4,23		6,83			2,51				100			
IBCN84-B-R84-Map7		8,4		4,37			1,92				100		1,5	
IBCN84-B-R87-Map7	1,75			1,61							100			
IBCN84-B-R9-Map7	4,76			6,21				1,88			100			

Supplementary Data 3. Examples of Top 10 BLAST hit organisms for the secondary metabolite genes (PKS and NPS) identified in members of the *L. maculans-L. biglobosa* species complex (1/3).

ID	BLASTP (10-6) Top 10 Organisms	Class	Order
PKS1	*Aspergillus flavus*	Eurotiomycetes	Eurotiales
	Aspergillus oryzae	Eurotiomycetes	Eurotiales
	Aspergillus parasiticus	Eurotiomycetes	Eurotiales
	Aspergillus nomius	Eurotiomycetes	Eurotiales
	Dothistroma septosporum	Dothideomycetes	Capnodiales
	Passalora arachidicola	Dothideomycetes	Capnodiales
	Podospora anserina	Sordariomycetes	Sordariales
	Aspergillus nidulans	Eurotiomycetes	Eurotiales
	Aspergillus ochraceoroseus	Eurotiomycetes	Eurotiales
	Aspergillus sojae	Eurotiomycetes	Eurotiales
PKS3	*Glarea lozoyensis*	Leotiomycetes	Helotiales
	Colletotrichum orbiculare	Sordariomycetes	Glomerellales
	Ophiostoma piceae	Sordariomycetes	Ophiostomatales
	Glomerella graminicola	Sordariomycetes	Glomerellales
	Gaeumannomyces graminis	Sordariomycetes	Magnaporthales
	Botryotinia fuckeliana	Leotiomycetes	Helotiales
	Sordaria macrospora	Sordariomycetes	Sordariales
	Colletotrichum gloeosporioides	Sordariomycetes	Glomerellales
	Neurospora tetrasperma	Sordariomycetes	Sordariales
	Thielavia terrestris	Sordariomycetes	Sordariales
PKS8	*Setosphaeria turcica*	Dothideomycetes	Pleosporales
	Colletotrichum higginsianum	Sordariomycetes	Glomerellales
	Colletotrichum orbiculare	Sordariomycetes	Glomerellales
	Fusarium graminearum	Sordariomycetes	Hypocreales
	Fusarium pseudograminearum	Sordariomycetes	Hypocreales
	Aspergillus terreus	Eurotiomycetes	Eurotiales
	Hypomyces subiculosus	Sordariomycetes	Hypocreales
	Pyrenophora tritici-repentis	Dothideomycetes	Pleosporales
	Glarea lozoyensis	Leotiomycetes	Helotiales
	Metacordyceps chlamydosporia	Sordariomycetes	Hypocreales
PKS21	*Arthroderma otae*	Eurotiomycetes	Onygenales
	Mycosphaerella populorum	Dothideomycetes	Capnodiales
	Talaromyces stipitatus	Eurotiomycetes	Eurotiales
	Metarhizium anisopliae	Sordariomycetes	Hypocreales
	Beauveria bassiana	Sordariomycetes	Hypocreales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Claviceps purpurea	Sordariomycetes	Hypocreales
	Podospora anserina	Sordariomycetes	Sordariales
	Colletotrichum gloeosporioides	Sordariomycetes	Glomerellales

Supplementary Data 3. Examples of Top 10 BLAST hit organisms for the secondary metabolite genes (PKS and NPS) identified in members of the *L. maculans-L. biglobosa* species complex (2/3).

ID	BLASTP (10-6) Top 10 Organisms	Class	Order
SirP	*Neosartorya fischeri*	Eurotiomycetes	Eurotiales
	Aspergillus fumigatus	Eurotiomycetes	Eurotiales
	Arthroderma gypseum	Eurotiomycetes	Onygenales
	Arthroderma otae	Eurotiomycetes	Onygenales
	Talaromyces stipitatus	Eurotiomycetes	Eurotiales
	Trichoderma virens	Sordariomycetes	Hypocreales
	Aspergillus terreus	Eurotiomycetes	Eurotiales
	Trichoderma atroviride	Sordariomycetes	Hypocreales
	Trichoderma reesei	Sordariomycetes	Hypocreales
	Trichophyton equinum	Eurotiomycetes	Onygenales
NPS5	*Pyrenophora teres f. teres*	Dothideomycetes	Pleosporales
	Phaeosphaeria nodorum	Dothideomycetes	Pleosporales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Talaromyces stipitatus	Eurotiomycetes	Eurotiales
	Eutypa lata	Sordariomycetes	Xylariales
	Metarhizium anisopliae	Sordariomycetes	Hypocreales
	Mycosphaerella populorum	Dothideomycetes	Capnodiales
	Phaeosphaeria nodorum	Dothideomycetes	Pleosporales
	Aspergillus terreus	Eurotiomycetes	Eurotiales
	Colletotrichum gloeosporioides	Sordariomycetes	Glomerellales
NPS8	*Neofusicoccum parvum*	Dothideomycetes	Botryosphaeriales
	Fusarium graminearum	Sordariomycetes	Hypocreales
	Macrophomina phaseolina	Dothideomycetes	Botryosphaeriales
	Penicillium oxalicum	Eurotiomycetes	Eurotiales
	Trichoderma virens	Sordariomycetes	Hypocreales
	Exophiala dermatitidis	Eurotiomycetes	Chaetothyriales
	Trichoderma reesei	Sordariomycetes	Hypocreales
	Cordyceps militaris	Sordariomycetes	Hypocreales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Metarhizium anisopliae	Sordariomycetes	Hypocreales
NPS11	*Metarhizium anisopliae*	Sordariomycetes	Hypocreales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Talaromyces stipitatus	Eurotiomycetes	Eurotiales
	Glarea lozoyensis	Leotiomycetes	Helotiales
	Talaromyces marneffei	Eurotiomycetes	Eurotiales
	Trichophyton verrucosum	Eurotiomycetes	Onygenales
	Arthroderma benhamiae	Eurotiomycetes	Onygenales
	Penicillium oxalicum	Eurotiomycetes	Eurotiales
	Fusarium fujikuroi	Sordariomycetes	Hypocreales
	Fusarium oxysporum	Sordariomycetes	Hypocreales

Supplementary Data 3. Examples of Top 10 BLAST hit organisms for the secondary metabolite genes (PKS and NPS) identified in members of the *L. maculans-L. biglobosa* species complex (3/3).

ID	BLASTP (10-6) Top 10 Organisms	Class	Order
Lys2	*Pyrenophora teres f. teres*	Dothideomycetes	Pleosporales
	Pyrenophora tritici-repentis	Dothideomycetes	Pleosporales
	Bipolaris sorokiniana	Dothideomycetes	Pleosporales
	Setosphaeria turcica	Dothideomycetes	Pleosporales
	Phaeosphaeria nodorum	Dothideomycetes	Pleosporales
	Bipolaris maydis	Dothideomycetes	Pleosporales
	Neofusicoccum parvum	Dothideomycetes	Botryosphaeriales
	Aspergillus nidulans	Eurotiomycetes	Eurotiales
	Coniosporium apollinis	Eurotiomycetes	Chaetothyriales
	Neosartorya fischeri	Eurotiomycetes	Eurotiales
NPS13	*Bipolaris sorokiniana*	Dothideomycetes	Pleosporales
	Aspergillus fumigatus	Eurotiomycetes	Eurotiales
	Neosartorya fischeri	Eurotiomycetes	Eurotiales
	Trichoderma virens	Sordariomycetes	Hypocreales
	Botryotinia fuckeliana	Leotiomycetes	Helotiales
	Sclerotinia sclerotiorum	Leotiomycetes	Helotiales
	Aspergillus oryzae	Eurotiomycetes	Eurotiales
	Aspergillus flavus	Eurotiomycetes	Eurotiales
	Trichophyton equinum	Eurotiomycetes	Onygenales
	Arthroderma gypseum	Eurotiomycetes	Onygenales
NPS14	*Coccidioides posadasii*	Eurotiomycetes	Onygenales
	Coccidioides immitis	Eurotiomycetes	Onygenales
	Uncinocarpus reesii	Eurotiomycetes	Onygenales
	Trichoderma virens	Sordariomycetes	Hypocreales
	Aspergillus clavatus	Eurotiomycetes	Eurotiales
	Aspergillus flavus	Eurotiomycetes	Eurotiales
	Aspergillus oryzae	Eurotiomycetes	Eurotiales
	Arthroderma otae	Eurotiomycetes	Onygenales
	Penicillium chrysogenum	Eurotiomycetes	Eurotiales
	Talaromyces stipitatus	Eurotiomycetes	Eurotiales
NPS15	*Phaeosphaeria nodorum*	Dothideomycetes	Pleosporales
	Mycosphaerella populorum	Dothideomycetes	Capnodiales
	Aspergillus oryzae	Eurotiomycetes	Eurotiales
	Aspergillus flavus	Eurotiomycetes	Eurotiales
	Metarhizium acridum	Sordariomycetes	Hypocreales
	Metarhizium anisopliae	Sordariomycetes	Hypocreales
	Claviceps purpurea	Sordariomycetes	Hypocreales
	Trichoderma virens	Sordariomycetes	Hypocreales
	Metarhizium robertsii	Sordariomycetes	Hypocreales
	Neosartorya fischeri	Eurotiomycetes	Eurotiales

Tableau 1. Données de séquençage et d'annotation des génomes des 4 souches de *L. maculans* 'brassicae'.

	v23.1.3	WA74	OMR19	Nz-T4
Taille de l'assemblage (Mb)	45.12	44.16	29.72	29.24
Nombre de scaffold	76	986	265	224
Nombre de contig	1743	3765	1391	1613
Scaffold N50 (Mb)	1769.6	263.0	244.5	316.9
Gaps (%)	2.5	9.6	3.4	3.4
Contenu en base GC (%)	45.2	46.5	51.7	51.7
Éléments transposables (%)	32.7	26.1	?	?
Nombre de gènes prédits	12543	10624	10559	10631
Nombre d'effecteurs prédits	651	632	634	637

Dynamique évolutive du génome et des effecteurs chez *L. maculans* ‘brassicae’

Introduction

Afin d'évaluer le polymorphisme des gènes codant des effecteurs ou potentiellement impliqués dans la pathogenèse, les génomes de deux souches de *L. maculans* 'brassicae' (Lmb) ont été reséquencés puis comparés à celui de la souche européenne de référence, v23.1.3 (JN3) précédemment séquencée par le Genoscope (Rouxel *et al.*, 2011), et à celui d’une souche australienne, WA74 (collaboration AAF Saskatoon, Canada). Les deux nouvelles souches qui ont été choisies sont potentiellement très divergentes des souches déjà séquencées et inféodées au colza (*Brassica napus* var. *oleifera*) :

- Nz-T4, une souche néo-zélandaise isolée de rutabaga (*Brassica napus* var. *rapifera*) et qui montre une faible agressivité sur colza.
- OMR19, une souche mexicaine adaptée au chou (*Brassica oleracea*) et avirulente sur la plupart des cultivars de colza utilisés en Europe.

Ces deux souches ont été séquencées (454 et Illumina) et assemblées (Newbler) par le Genoscope.

Statistiques de séquençage

Les génomes des isolats Nz-T4 et OMR19 de Lmb semblent avoir une taille inférieure (29,2-29,7 Mb) à ceux des souches précédemment séquencées, 45,1 Mb pour v23.1.3 et 44,2 pour WA74 (Tableau 1). Cette différence de taille est due au fait que les assemblages finaux de Nz-T4 et OMR19 ne contiennent pas les séquences répétées correspondant principalement aux ET, qui chez v23.1.3 et WA74 sont regroupés en grandes régions génomiques (isochores AT) et représentent près d'un tiers du génome. Les ET dans les génomes de v23.1.3 et WA74 sont dégénérés par le RIP, qui en mutant les bases G en A et les bases C en T, est responsable d'un taux global en bases GC de 45 %. Pour Nz-T4 et OMR19, l’absence d’assemblage des ET et autres éléments répétés conduit à un taux de GC des deux génomes artificiellement plus élevé, de l’ordre de 52 % (Tableau 1, Figure 1).

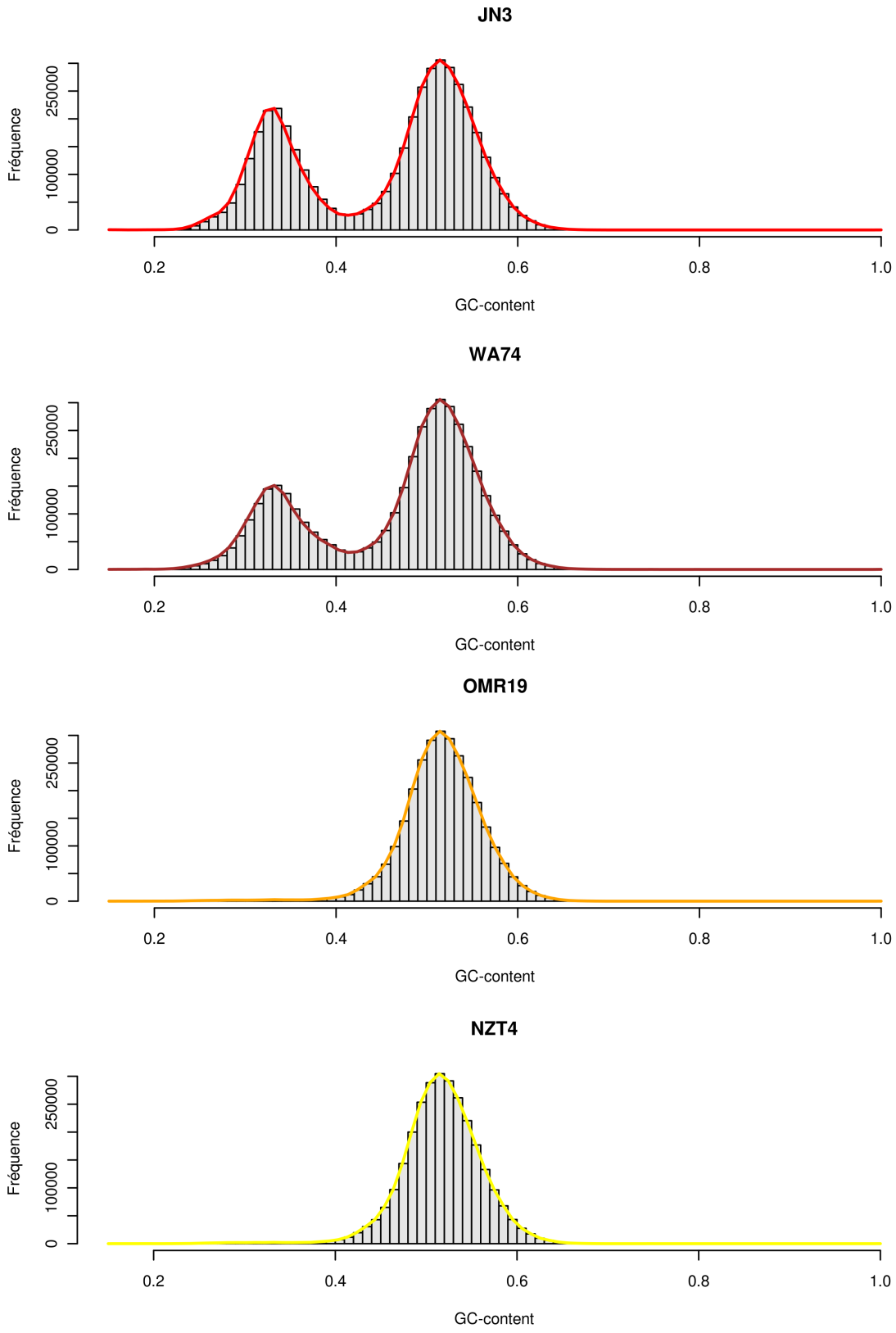


Figure 1. Distribution du contenu en bases GC dans les génomes de quatre souches de *L. maculans* 'brassicae'. Le contenu en bases GC a été calculé le long des chromosomes dans des fenêtres glissantes de 1 kb. La distribution bimodale observée chez v23.1.3 (JN3) et WA74 reflète la structure compartimentée de leur génome avec des régions génomiques pauvres en GC correspondant aux isochores AT. Cette bimodalité n'est pas observée dans les génomes de OMR19 et Nz-T4 car les éléments répétés n'ont pas été assemblés.

Cependant, nous pouvons affirmer qu'il ne s'agit là que d'un biais lié aux stratégies NGS, car un séquençage de quelques clones BAC de la souche Nz-T4 par la méthode de Sanger avait précédemment montré la présence d'isochores AT totalement absents dans le nouvel assemblage. On peut donc supposer que Nz-T4 et OMR19 ont un génome de taille similaire à celle des autres souches de Lmb et structuré de la même façon en isochores. La conservation des isochores AT, riches en ET et en gènes codant pour des effecteurs, ne peut donc pas être étudiée entre les différentes souches avec ces assemblages.

Alignements des génomes et identification des SNPs

Des alignements nucléotidiques ont été réalisés entre les séquences non répétées des différents génomes de Lmb en utilisant le programme *nucmer* issu du package MUMmer (Figure 2). Ces alignements couvrent en moyenne 97 % des génomes avec un pourcentage d'identité de 98 %, ce que l'on peut considérer comme étant une macrosynténie quasi parfaite. En se basant sur ces alignements, le nombre de SNP entre chaque paire de génome a été calculé en utilisant le programme *show-snps* du package MUMmer. Cela a permis d'identifier environ 20000 SNPs entre chaque paire de génome, soit une moyenne de 1 SNP toutes les 1400 bases (Tableau 2). Ces chiffres sont très proches de ceux obtenus par Zander et *al.* (2013) qui identifient 21000 SNPs entre deux souches australiennes de *L. maculans* dans une étude similaire à celle-ci. Nos données sont aussi du même ordre de grandeur que le nombre de SNPs identifiés entre différents isolats de *Cochliobolus heterostrophus*, soit 30000 à 50000 (Condon et *al.*, 2013). Il est intéressant de noter que l'alignement entre les génomes de WA74 et Nz-T4, souches géographiquement proches, montre un nombre de SNP plus faible (14000).

Tableau 2. Nombre de SNPs détectés entre chaque paire de génomes alignés.

	WA74	**OMR19**	**Nz-T4**
v23.1.3	19396	23953	20307
WA74	-	18712	13790
OMR19	-	-	20330

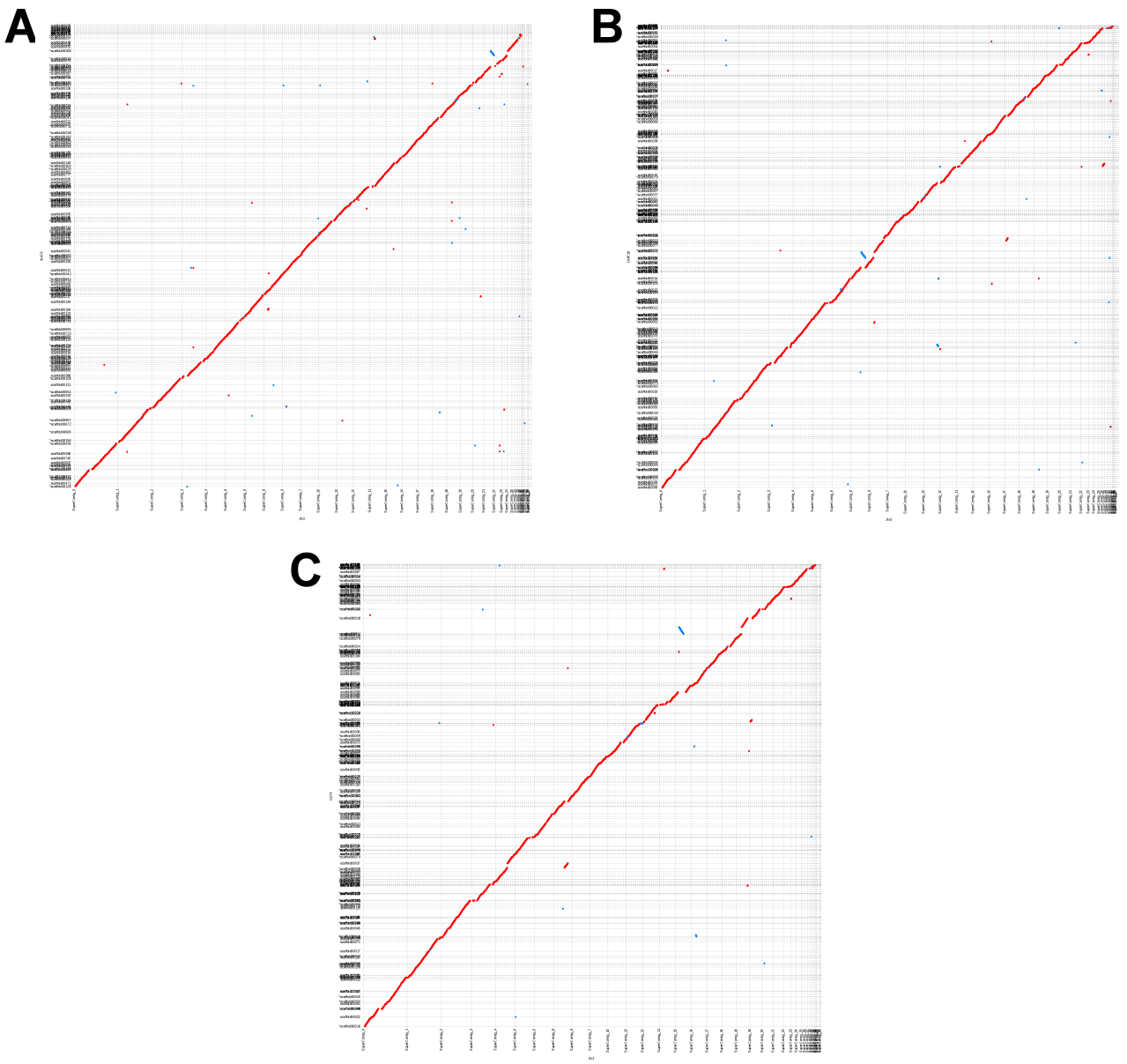


Figure 2. Alignements du génome de v23.1.3 avec ceux des autres souches. Les génomes masqués de leur éléments repétés ont été alignés au niveau nucléotidique les uns avec les autres grâce au programme *nucmer* de MUMmer. En rouge apparaissent les portions de séquence orientées similairement dans les génomes, en bleu apparaissent celles qui sont inversées. (**A**) Alignement du génome de v23.1.3 avec celui de WA74. (**B**) Alignement du génome de v23.1.3 avec celui de OMR19. (**C**) Alignement du génome de v23.1.3 avec celui de Nz-T4.

Localisation des SNPs

Pour chaque souche, le positionnement des SNPs sur le génome de référence (v23.1.3) permet de montrer une présence préférentielle de ces mutations dans les isochores GC puisque près de 90 % des SNPs sont localisés dans ces compartiments. Cependant, une analyse de la fréquence de ces mutations, dans des fenêtres glissantes de 10 kb, permet de mettre en évidence une répartition non-homogène le long des chromosomes. En effet, il semble que certaines régions génomiques soient plus affectées que d'autres par les mutations (Figure 3). En regardant la localisation de ces mutations par rapport aux régions codantes, on s'aperçoit que 34 % d'entre elles sont situées dans des régions exoniques, 10 % en régions introniques et le reste, 56 %, dans des régions intergéniques. Ce sont plus de 5500 gènes qui sont touchés par ces mutations, et bien que ces dernières soient retrouvées principalement en isochores GC, la proportion de gènes touchés dans chaque compartiment est similaire (40-45 %). Concernant les fonctions des gènes mutés, il n'y a aucun enrichissement observé d'une fonction donnée que ce soit au niveau des gènes de « ménage » ou des gènes codant des effecteurs. La proportion de gènes mutés ayant une fonction prédite *in silico* est identique à celle de l'ensemble des gènes (30 %).

Conservation et spécificité des gènes

L'annotation automatique des génomes identifie 10631 gènes chez Nz-T4 et 10559 gènes chez OMR19, un nombre très similaire aux 10624 gènes prédits chez WA74 avec le même pipeline d’annotation automatique (Tableau 1). La comparaison des séquences protéiques par BLASTP réciproques nous indique qu’environ 9000 gènes composent le *core* protéome des 4 souches de *L. maculans*. Cent soixante-douze gènes sont apparemment spécifiques de la souche OMR19 et 280 de Nz-T4.

Le nombre de gènes codant pour des effecteurs potentiels est très proche entre les deux isolats, 634 chez OMR19 et 637 chez Nz-T4, et très similaire au nombre identifié dans les autres génomes, 651 chez v23.1.3 et 632 chez WA74 (Tableau 1). Les protéines codées ont une taille moyenne proche (153 aa) et une proportion en cystéines 2,5 fois plus élevée que dans l'ensemble des protéines des génomes. Une comparaison des séquences au niveau protéique permet de définir un set commun de 290 séquences et des sets spécifiques à chaque isolat (77 séquences pour OMR19, 100 pour Nz-T4, 95 pour WA74 et 263 pour v23.1.3) (Figure 4).

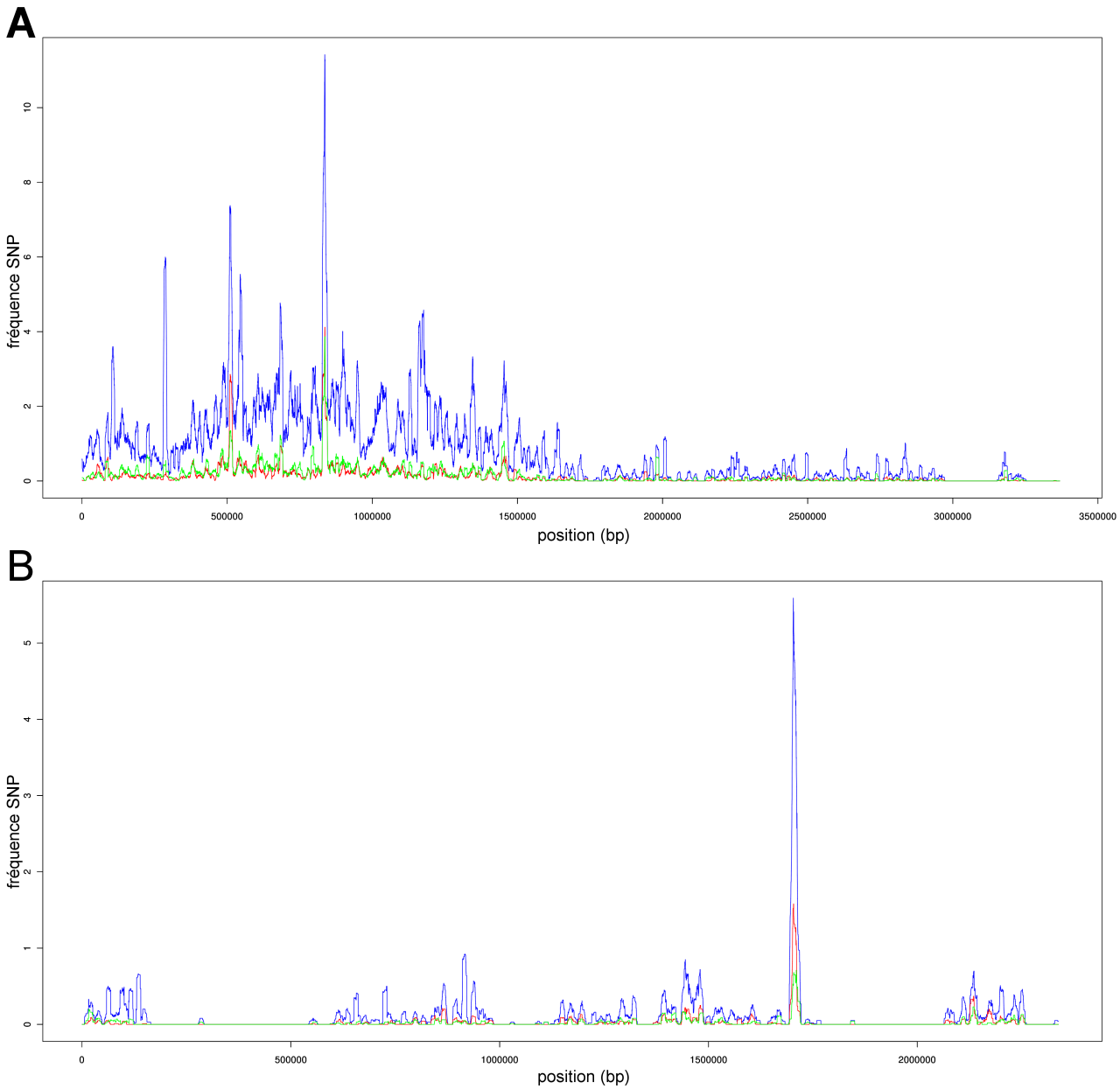


Figure 3. Distribution des SNPs détectés entre v23.1.3 et les trois autres souches. La fréquence des SNPs identifiés entre v23.1.3 et les 3 autres souches a été calculée dans des fenêtres glissantes de 10 kb (décalage de 50 bp) le long des SuperContig (SC) de v23.1.3. En bleu est indiquée la fréquence de toutes les mutations détectées, en rouge est indiquée la fréquence des mutations synonymes et en vert celle des mutations non synonymes. (**A**) Sur le SC1 du génome de v23.1.3, qui correspond à un chromosome entier, la moitié en 5' (correspondant peut être à un bras du chromosome) semble plus affectée par les mutations. (**B**) Sur le SC3, une région de 30 kb contenant 11 gènes semble être préférentiellement soumise à mutation.

Les séquences correspondantes ne sont généralement pas absentes des autres génomes et, au niveau nucléique, 95 % de ces séquences « spécifiques » d'une souche sont présentes dans au moins un des autres génomes. Dans certains cas, l'absence de prédiction est liée à une modification de la séquence (apparition de codons stop prématurés, mutation du codon start) conduisant à une inactivation du gène et donc à une pseudogénisation (Figure 5). Dans d'autres cas, dont l'importance est difficile à évaluer dans l'état actuel de l'assemblage (de nombreux gènes codant pour des effecteurs putatifs sont localisés au sein des isochores AT peu ou très mal assemblés), la protéine correspondante n'a pas été prédite. Cette étude permet tout de même de définir un set d'effecteurs potentiels propre à Lmb comprenant 1177 PPS différentes.

Discussion

Le reséquençage de deux souches de *L. maculans* 'brassicae' présentant une diversité potentiellement importante avec les souches déjà séquencées, due à un isolement géographique ou à une spécialisation d'hôte, avait pour objectif d'évaluer les divergences générées à l'échelle du génome par ces isolements. Il s'agissait, en particulier, d'établir leurs conséquences sur les gènes impliqués dans la pathogenèse fongique avec une focalisation sur le répertoire d'effecteurs que contenait chacun des génomes.

Outre les problèmes d'assemblage de régions riches en ET dont les conséquences en terme de prédiction de gènes pertinents pour la pathogenèse restent à évaluer, ce reséquençage montre une divergence génomique très limitée entre les souches analysées. En particulier, la conservation de séquence entre régions codantes est très importante et compatible avec une spéciation récente de *L. maculans*, suivie d'une dissémination à l'échelle mondiale dans les 150 dernières années (Dilmaghani *et al.*, 2012). Quoique la quantité de gènes codant pour des protéines de fonction caractérisée reste limitée chez les champignons, les quelques 1700 gènes polymorphes entre les différentes souches, ainsi que les régions génomiques plus spécifiquement affectées par des SNPs, méritent une investigation plus poussée de par leur implication potentielle dans la pathogenèse. Les régions génomiques dans lesquelles la densité des SNPs est réduite sont tout aussi intéressantes, puisqu'elles peuvent indiquer la présence de séquences fonctionnellement conservées pouvant contenir par exemple des régions codantes ou des régions régulatrices.

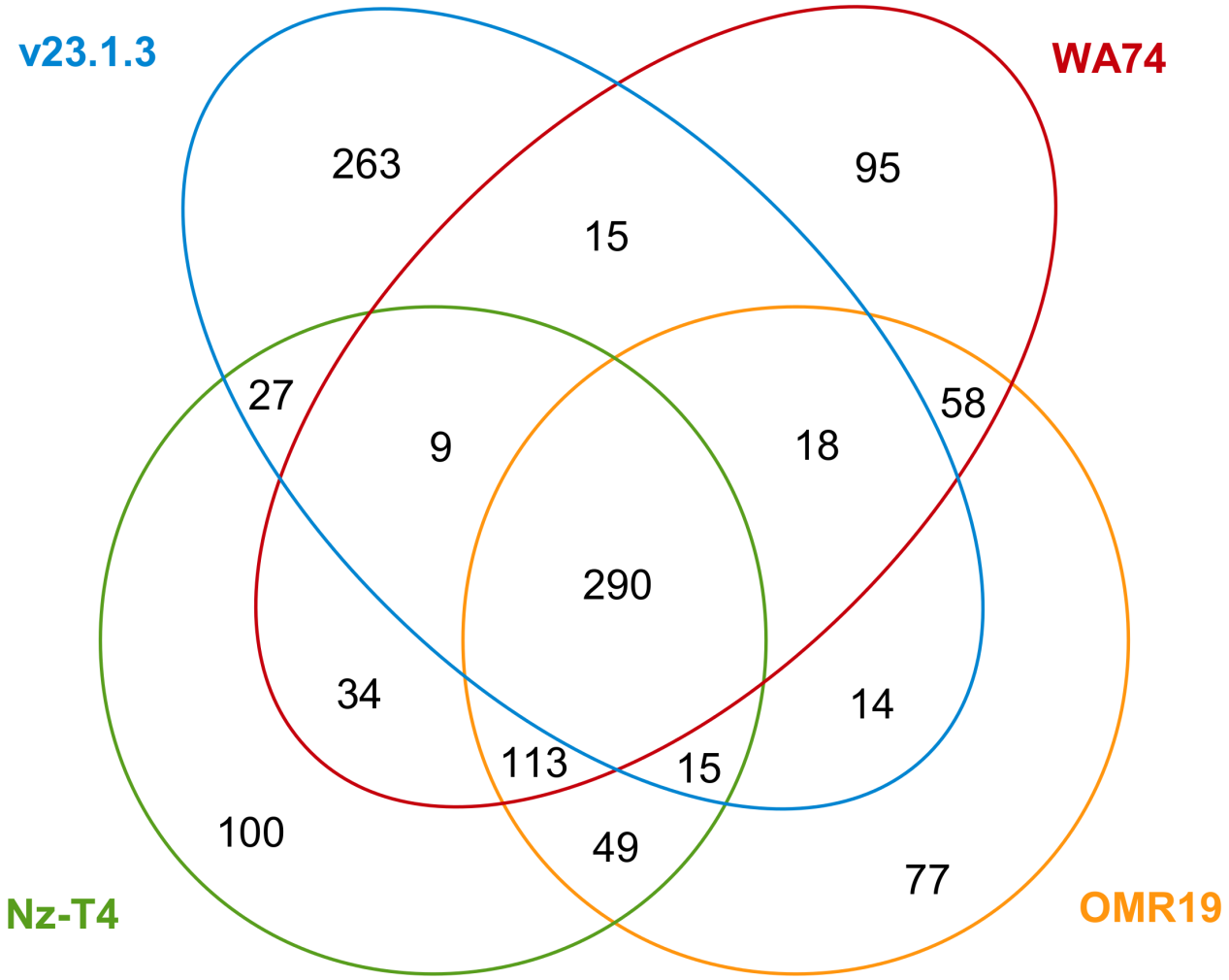


Figure 4. Répartition des PPS dans les génomes des 4 souches de *L. maculans* 'brassicae'. Les PPS de ces 4 souches forment un répertoire de 1177 PPS non redondantes.

Par la suite, il serait intéressant d'effectuer des analyses évolutives en calculant les rapports entre le nombre de mutations non synonymes et synonymes (dN/dS) présent dans les alignements de chromosomes afin de visualiser si certaines régions du génome sont soumises à des pressions de sélection.

Plus que le polymorphisme nucléotidique, ce qui semble différencier les souches est le contenu en gènes, et en particulier en gènes codant pour des effecteurs. En effet, 35 % et 45 % des gènes spécifiques de Nz-T4 et OMR19 respectivement codent pour des PPS. Sur l'ensemble des souches, entre 12 % et 40 % des gènes codant des PPS sont souche-spécifiques. La mise en évidence de pseudogénisation pour certains de ces gènes suggère qu'ils sont soumis à l'évolution accélérée typique de ce type de gènes chez *L. maculans*, lié en particulier à leur environnement génomique. Ce travail a donc permis d'enrichir de façon substantielle le catalogue d'effecteurs produit par *L. maculans* et de confirmer la plasticité de ce répertoire favorisant une adaptation rapide à des conditions environnementales changeantes. Il reste maintenant, sur la base de ces données à affiner ou développer des analyses évolutives, génétiques et fonctionnelles pour analyser l'implication de tels gènes dans l'adaptation à des hôtes spécifiques (par exemple chou *vs.* colza).

```
 Score =  211 bits (537), Expect = 1e-55,   Method: Compositional matrix adjust.
 Identities = 108/136 (79%), Positives = 119/136 (87%)
 Frame = +2


Query: 1    MPLSLEIILTLLALSIPTITACREASISGEIRYPQGTCPTKTEALNDCNKVTKGLIDFSQ 60
             MPLSL+IIL LLALSIPTI ACREA IS EIRYP GTCPTKT+ALN+CNKVTKGLI+FS
Sbjct: 1190 MPLSLKIILRLLALSIPTIIACREALISREIRYP*GTCPTKTKALNNCNKVTKGLINFS* 1369


Query: 61   SHQRAWGIDMTAKVQCAPCITTDPWDVVLCTCKITAHRYREFVPKIPYSSFSSAPGVIFG 120
             SH RA GID+TAKV CAPCITT+P DVVLCTCKITAHRY EFVPKIPYSSFS AP VIF
Sbjct: 1370 SH*RA*GIDITAKV*CAPCITTNP*DVVLCTCKITAHRY*EFVPKIPYSSFSLAPRVIFS 1549


Query: 121  QETGLDHDPEWVVNMK 136
              ET LD++P+ VVN++
Sbjct: 1550 *ETSLDYNPK*VVNVR 1597
```

Figure 5. Pseudogénisation d'AvrLm4-7 dans la souche Nz-T4. La séquence correspondant au gène *AvrLm4-7* est présente dans le génome de Nz-T4 mais des mutations ont engendré l'apparition de codons stop prématurément.

DISCUSSION ET PERSPECTIVES

DISCUSSION - CONCLUSION

Incidence des Eléments Transposables sur l'évolution des génomes des champignons phytopathogènes et sur leur potentiel adaptatif

Jonathan Grandaubert, Marie-Hélène Balesdent, Thierry Rouxel

Article écrit pour la séance « Génétique des champignons » de Biologie Aujourd'hui.

INRA-Bioger, Campus AgroParisTech, 78850 Thiverval-Grignon, France

rouxel@versailles.inra.fr

Résumé : Les champignons phytopathogènes, menace majeure pour la sécurité alimentaire mondiale, présentent une plasticité de modes de vie remarquable et une extrême capacité à s'adapter aux méthodes de luttes qui leur sont opposées par l'homme dans les agro-systèmes. La génomique et la génomique comparative sont utilisées ici pour évaluer le lien entre plasticité génomique et potentiel évolutif et adaptatif. Une série évolutive de champignons phytopathogènes inféodés aux crucifères, le complexe d'espèce *Leptosphaeria maculans-Leptosphaeria biglobosa* a été choisie ici comme modèle puisqu'elle regroupe cinq entités dont le statut con- ou hétérospécifique est peu clair mais présentant des différences de gamme d'hôte ou de pouvoir pathogène. En particulier, l'espèce la mieux adaptée sur le colza (et la plus dommageable pour l'agriculture), *L. maculans* 'brassicae', présente, par rapport aux autres membres du complexe d'espèces, un génome caractérisé par l'expansion récente, mais massive, de quelques familles d'éléments transposables (ET). Celle-ci a sans doute eu un effet encore peu clair sur la spéciation, mais a surtout contribué à la diversification de molécules de type « effecteurs », donc à l'acquisition de nouvelles spécificités parasitaires. La localisation des gènes codant pour des effecteurs dans des régions génomiques enrichies en ET a par ailleurs un effet direct sur l'adaptation aux résistances variétales en favorisant une diversité d'événements mutationnels. Ces données sont confrontées à d'autres exemples de la littérature qui tendent à généraliser l'idée que les champignons phytopathogènes ont, au cours de l'évolution, développé des « génomes à deux vitesses » incluant un compartiment plastique enrichi en ET et en gènes impliqués dans le pouvoir pathogène et l'adaptation à l'hôte.

Mots-clés : Génome fongique, éléments transposables, spéciation, effecteurs, adaptation

Abstract: Phytopathogenic fungi are a major threat for global food security and show an extreme plasticity in pathogenicity behaviours. They often have a high adaptive potential allowing them to rapidly counteract the control method used by men in agrosystems. In this paper, we evaluate the link between genome plasticity and adaptive potential using genomics and comparative genomics approaches. Our model is the evolutionary series *Leptosphaeria maculans-Leptosphaeria biglobosa*, encompassing five distinct entities, whose conspecificity or heretrospecificity status is unclear, and which all are pathogens of cruciferous plants. They however differ by their host range and pathogenicity. Compared to other species of the species complex, the species best adapted to oilseed rape, *L. maculans* 'brassicae', causing important losses in the crop, has a genome that was submitted to a recent and massive burst of transposition by a few families of transposable elements (TEs). Whether the genome invasion contributed to speciation is still unclear to-date but there is a coincidence between this burst of TEs and divergence between two species. This TE burst contributed to diversification of effector proteins and thus to generation of novel pathogenic specificities. In addition, the location of effector genes within genome regions enriched in TEs has direct consequences on adaptation to plant resistance and favours a multiplicity of mutation events allowing « breakdown » of resistance. These data are substantiated by other examples in the literature showing that fungi tend to have a « two-speed » genome, in which a plastic compartment enriched in TEs hosts genes involved in pathogenicity and adaptation to host.

Keywords: fungal genome, transposable elements, speciation, effector, adaptation to host

Introduction

Avec un nombre (sous-)estimé à 1,5 millions d'espèces, les champignons représentent une partie extrêmement importante de la biodiversité globale ; leur biomasse est considérable au sein des écosystèmes dans lesquels ils jouent des rôles fondamentaux (Silar, 2013). Avec une majorité d'espèces saprophytes, les champignons sont responsables de la dégradation de la matière organique permettant ainsi le recyclage de composés tels que la lignine qu'ils sont les seuls à pouvoir dégrader efficacement (Silar, 2013). Les champignons sont les organismes interagissant le plus souvent avec les plantes, en particulier les plantes cultivées, et leurs impacts sur l'environnement, l'économie et la santé humaine sont conséquents. Au cours de l'évolution, de nombreuses espèces ont développé des interactions très spécialisées avec les plantes, et des lignages endophytes, mutualistes ou phytopathogènes ont émergé à de multiples reprises et indépendamment au sein de clades de saprophytes. Les champignons phytopathogènes ont un impact tout particulier en termes de pertes de rendement, d'altération de la qualité des produits et de coûts économiques et environnementaux engendrés par des méthodes de lutte souvent basées sur la chimie de synthèse. Ils peuvent aussi avoir un impact plus direct sur la santé humaine via la production de mycotoxines. De plus, la lutte chimique dirigée contre ces organismes contribue fortement à des effets adverses sur la santé humaine et sur l'environnement, notamment à cause de la persistance et de la toxicité des biocides utilisés.

Les champignons, de par leur petit génome particulièrement plastique, leur cycle de vie court, la taille de leur population et leurs nombreuses formes de multiplication, montrent d'extrêmes capacités évolutives et adaptatives. Ainsi de nouvelles maladies fongiques des plantes, mais aussi des animaux et des hommes émergent constamment (Olsen *et al.*, 2011). Toutefois, il ne s'agit pas là d'une menace nouvelle puisque des épidémies dues à des champignons ou à des organismes présentant une convergence évolutive avec les champignons, les oomycètes, ont influé sur l'histoire humaine. L'exemple le plus célèbre est la Grande Famine due aux épidémies de mildiou de la pomme de terre (*Phytophthora infestans*) survenues en Irlande au milieu du 19e siècle et qui a causé la mort d'un million de personnes et entraîné l'émigration de deux millions de personnes principalement vers les États-Unis, la Grande-Bretagne, le Canada et l'Australie. Au cours du 20e siècle, et en particulier durant les deux dernières décennies, la fréquence et la vitesse d'apparition de nouvelles maladies semblent s'accroître, comme le

démontrent des rapports alarmistes récemment publiés dans des revues prestigieuses telles que *Nature* ou *Science* (Fisher *et al.*, 2012 ; Kupferschmidt *et al.*, 2012) et de nombreux articles dans les médias ou sur internet (p. ex. concernant l'encre du châtaigner -*Cryphonectria parasitica*- [1] ; la maladie du flétrissement des frênes -*Chalara fraxinea*- [2] ; la graphiose de l'orme -*Ophiostoma novo-ulmi*- [3] ; ou la rouille du caféier -*Hemileia vastatrix*- [4]). De plus en plus d'études insistent sur l'émergence de nouvelles maladies fongiques chez de nombreux organismes (chauves-souris, batraciens, tortues marines, poissons, coraux) qui peuvent être dues soit à l'apparition de nouvelles espèces fongiques, soit à l'adaptation d'une espèce préexistante à un nouvel hôte (Fisher *et al*., 2012). A ce jour, les plantes restent les organismes les plus attaqués par les champignons et ces derniers sont considérés comme une menace majeure pour la sécurité alimentaire mondiale puisqu'on estime qu'environ 10 % de la production agricole mondiale est perdue à cause de maladies fongiques (Strange & Scott, 2005 ; Pennisi, 2010; Fisher *et al.*, 2012 ; Silar, 2013). Plus de 125 millions de tonnes des produits des cinq plantes de grandes cultures les plus utilisées pour l'alimentation humaine (incluant le blé et le riz) sont ainsi détruites chaque année par des maladies fongiques (Anderson *et al.*, 2004 ; Olsen *et al*., 2011).

Dans le contexte du changement global, les champignons phytopathogènes agissent comme des espèces invasives, ce qui est favorisé par leurs caractéristiques biologiques mais aussi par la mondialisation des échanges commerciaux (Stukenbrock & McDonald, 2008 ; Stukenbrock & Bataillon, 2012). Les champignons ont souvent de fortes capacités de dispersion (principalement pour les parasites aériens) mais profitent aussi des échanges intercontinentaux des plantes qu'ils infectent. Ils possèdent aussi de grandes tailles de population efficace, des régimes reproductifs mixtes (une reproduction sexuée permettant la recombinaison des caractères et une multiplication asexuée favorisant une expansion massive des génotypes les mieux adaptés), et des capacités de survie à long terme en dehors de leur hôte (tant en initiant un mode de vie saprophyte que sous forme de spores dormantes). Toutes ces caractéristiques biologiques couplées à certaines particularités génomiques indiquent que les champignons sont des organismes aux mécanismes adaptatifs spécifiques particulièrement efficaces, comme nous le montrerons dans cet article à l'aide du modèle *Leptosphaeria maculans*, leur permettant de s'adapter rapidement aux méthodes de lutte (résistance variétale ou lutte chimique) utilisées par l'homme pour les contrôler (Giraud *et al.*, 2010).

Depuis les origines de l'agriculture, l'homme a utilisé, intentionnellement ou non, des résistances variétales basées sur des gènes majeurs de résistance montrant un déterminisme mendélien simple (Biffen, 1905) et définissant la relation « gène-pour-gène » dans laquelle la résistance est mise en place lorsque le gène de résistance de la plante « reconnaît » la présence d'un gène correspondant chez l'agent pathogène (appelé « gène d'avirulence »). Depuis l'identification de ce déterminisme simple (Biffen, 1905), les phytopathologistes ont observé dans les agro-écosystèmes, des centaines de cas dans lesquelles des populations pathogènes devenaient capables de contourner une résistance variétale qui leur était opposée. De tels effondrements de l'efficacité d'une résistance ont par exemple été décrits pour des maladies aériennes des céréales (rouilles ou oïdiums) (McDonald & Linde, 2002) et pour la nécrose du collet du colza causée par *Leptosphaeria maculan*s (Rouxel *et al*., 2003). Typiquement, un gène majeur montrant une efficacité extrême est utilisé de plus en plus massivement par les agriculteurs et sur de grandes superficies, ce qui permet, les premières années une absence visible de la maladie puisque la résistance éradique effectivement la partie « avirulente », majoritaire de la population pathogène. En général, la population pathogène s'adapte à la pression de sélection en quelques années seulement et la composante « avirulente » devient peu à peu minoritaire dans la population, rendant la résistance inefficace, et réduisant à néant des efforts de sélection variétale de plusieurs années (Rouxel *et al.*, 2003 ; Daverdin *et al.*, 2012). Les caractéristiques biologiques citées précédemment sont importantes pour de tels contournements, mais la structure et le dynamisme du génome des champignons sont essentiels pour une adaptation optimale à ce type de pression de sélection.

L'obtention des séquences génomiques de la levure, puis des premiers champignons filamenteux (Galagan *et al.*, 2003 ; Machida *et al*., 2005) a permis d'esquisser les caractéristiques génomiques des champignons : génome de petite taille présentant une forte densité génique et comportant peu de séquences répétées. Les premiers séquençages de champignons phytopathogènes, montrant des génomes de 20-40 Mb (*Ustilago maydis*, *Magnaporthe oryzae*, *Fusarium graminearum*, *Phaeosphaeria nodorum* (Dean *et al*., 2005 ; Kämper *et al* ., 2006 ; Cuomo *et al*., 2007 ; Hane *et al.*, 2007) ne dérogeaient pas à cette règle. En ce sens le génome était peu explicatif des capacités évolutives et adaptatives de ces organismes. Ce point de vue a drastiquement changé à partir de 2009, chez les oomycètes (Haas *et al.*, 2009 ; Raffaele *et al.*, 2010), puis chez les ascomycètes (Spanu *et al.*, 2010 ; Rouxel *et al.*, 2011) et les basidiomycètes (Duplessis *et al.*, 2011), lors de la

caractérisation du génome d'agents pathogènes montrant une diversité de taille et d'architecture inattendue et contradictoire avec les données issues des premiers génomes séquencés. En particulier, on observe une tendance à une augmentation de la taille des génomes due à l'invasion de ces derniers par des éléments transposables (ET), qui sculptent un génome « à deux vitesses » favorisant l'adaptabilité des organismes. C'est ce que nous allons illustrer dans la suite de cet article, principalement avec l'analyse du génome de *Leptosphaeria maculans* et de plusieurs espèces proches.

Montagnes russes dans le génome de *Leptosphaeria maculans*

Leptosphaeria maculans est un champignon phytopathogène inféodé aux crucifères, responsable de l'une des principales maladies du colza (*Brassica napus*), la nécrose du collet (*blackleg* des anglo-saxons) (Rouxel & Balesdent, 2005). Durant les années 1990 et 2000, *L. maculans* est devenu un modèle d'étude important grâce aux ressources accumulées telles que des grandes collections de souches et de grandes collections de mutants d'insertion (Blaise *et al.*, 2007) et à sa plasticité au laboratoire (incluant la possibilité d'obtenir la reproduction sexuée au laboratoire et donc d'initier des analyses de génétique formelle, chose rare pour un champignon phytopathogène). Ainsi, *L. maculans* est à ce jour l'un des modèles fongiques chez lesquels le plus de gènes d'avirulence ont été clonés et caractérisés, favorisant ainsi la connaissance des gènes de résistance majeurs correspondants (gènes *Rlm*) chez le colza ou les espèces de *Brassica* apparentées. Au milieu des années 2000, une initiative conjointe de l'INRA et de l'Université de Melbourne (Australie) a conduit au séquençage du génome d'une souche européenne de *L. maculans* 'brassicae' (v23.1.3) par le Genoscope (Rouxel *et al*., 2011). La séquence du génome, obtenue à l'aide d'une technologie de séquençage de type Sanger a été fournie en 2007 sous la forme d'un assemblage de très bonne qualité et a révélé plusieurs traits caractéristiques. Avec une taille de plus de 45 Mb il constitue en 2007 l'un des plus grands génomes séquencé jusqu'alors chez les champignons filamenteux (Rouxel *et al.*, 2011). De plus, le génome présente une composition en bases GC (44 %) globalement inférieure à celles des espèces fongiques proches taxinomiquement (50-55 %) (Rouxel *et al.*, 2011). Mais la particularité inédite de ce génome réside dans sa structuration en isochores, c'est à dire une alternance le long des chromosomes de deux compartiments génomiques, l'un équilibré en bases GC (ou isochores GC) et l'autre pauvre en bases GC (ou isochores AT), qui jusqu'ici n'avait été

observée que chez des Eucaryotes plus complexes tels que les Mammifères (Rouxel *et al.*, 2011 ; Raffaele & Kamoun, 2012). L'analyse du génome a montré que les isochores GC comprennent la majorité des gènes (95 %), avec une forte densité génique, alors que les isochores AT sont surtout composés de mosaïques plus ou moins complexes d'éléments transposables inactivés par un phénomène préméiotique d'inactivation des éléments répétés spécifique des champignons, le RIP (*Repeat-Induced Point mutations*). Ces isochores AT, correspondant à 36 % du génome, contiennent certes peu de gènes, mais ils sont enrichis en gènes codant des petites protéines sécrétées, collectivement appelées effecteurs, potentiellement impliquées dans les interactions entre *L. maculans* et son hôte lors de la mise en place de la pathogenèse (Rouxel *et al.*, 2011).

Les questions posées par la structure du génome

Une telle structure de génome pose des questions évolutives, notamment sur la contribution des ET dans l'émergence d'une nouvelle espèce et des différents mécanismes mis en œuvre lors de la spéciation. Elle nous questionne aussi sur l'incidence de cette structure sur les capacités parasitaires et adaptatives et l'on peut se demander quels sont les avantages/désavantages du maintien de cette structure pour l'organisme.

Plus spécifiquement, les questions que l'on peut envisager de traiter sur la base d'une telle structure de génome sont les suivantes : (i) Peut-on déterminer quand les ET ont envahi les génomes fongiques au cours de l'évolution et en déduire des mécanismes de spéciation ? (ii) Dans le cas de *L. maculans*, les ET ont-ils contribué à générer de nouveaux gènes ou à créer des duplications à l'origine de nouvelles fonctions ? En quoi une telle innovation génique a-t-elle favorisé l'adaptation à la plante hôte, le colza ? (iii) Quel est le lien entre la structure du génome et l'adaptation aux résistances variétales utilisées en agriculture pour contrôler la maladie ?

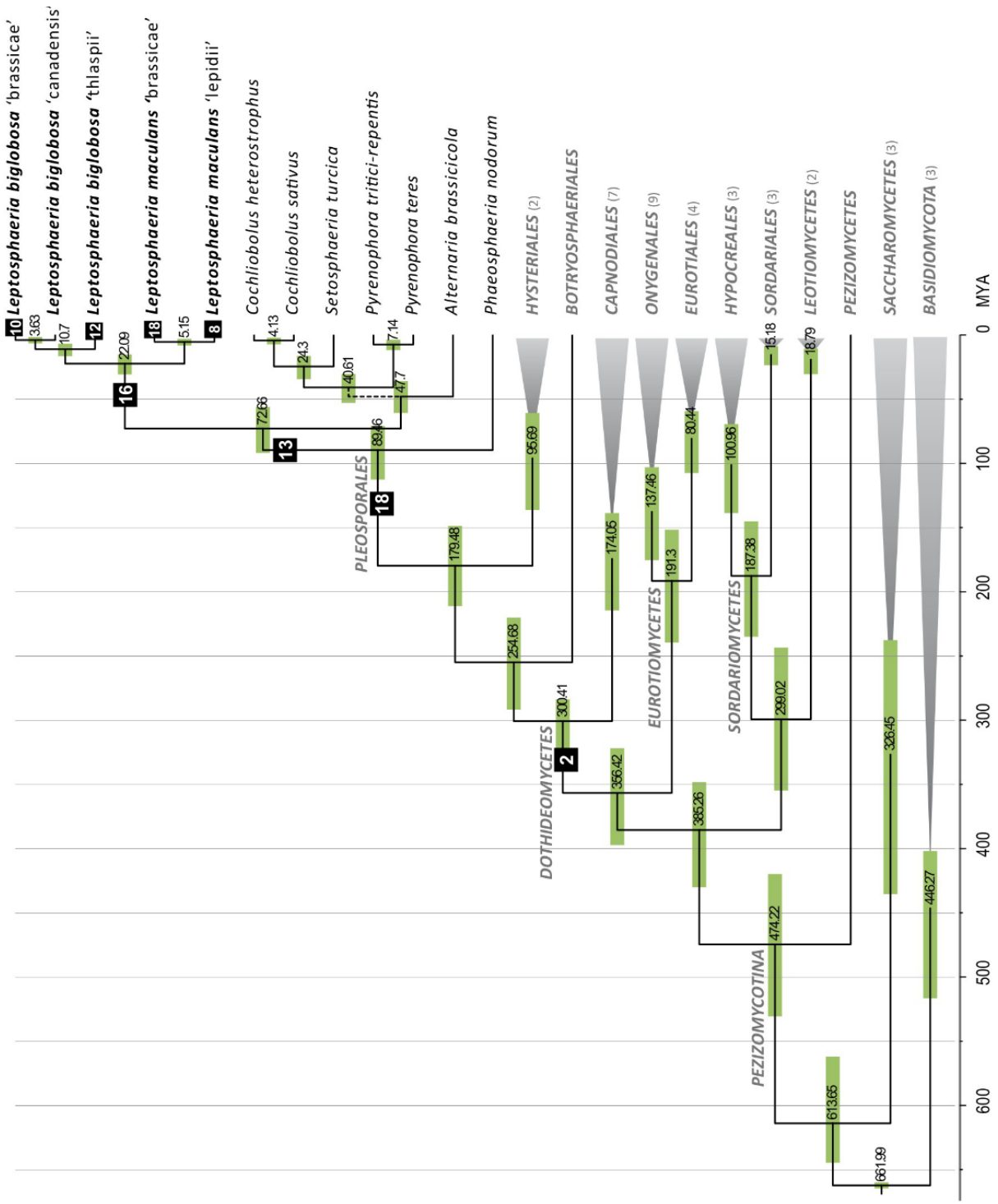


Figure 1. Relations phylogénétiques et estimation des temps de divergence au sein des classes majeures des Ascomycètes. Les chiffres indiqués dans des carrés noir représentent le nombre de familles d'éléments transposables qui ont envahi les différentes branches de l'arbre.

Un modèle pour la génomique évolutive et adaptative : le complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa*

Pour traiter ces différentes questions, nous avons initié une approche de génomique comparative prenant avantage de l'existence d'un complexe d'espèces plus ou moins bien définies, proches de *L. maculans* et parasitant les crucifères adventices ou cultivées. Ainsi, si depuis le début du 20e siècle, les mycologues estimaient que l'espèce « *L. maculans* » était extrêmement polymorphe et reconnaissaient deux formes distinctes principalement selon leur agressivité et les dégâts causés sur colza, ça n'est qu'en 2001 que les deux « formes » ont été formellement reconnues comme deux espèces distinctes dont l'une était alors renommée *Leptosphaeria biglobosa* (Shoemaker & Brun, 2001). Très proches morphologiquement, les deux espèces partagent la même niche écologique et le même style de vie et, à l'exception de quelques régions du globe, se retrouvent généralement ensemble dans les parcelles de colza et même dans les plantes infectées chez lesquelles elles peuvent toutes deux développer une longue colonisation systémique et endophyte sans expression de symptôme (West *et al.*, 2002). Cette séparation en deux espèces n'était que le premier pas vers une plus grande complexité puisque des collectes effectuées sur crucifères adventices dans les années 1960 par un mycologue canadien, J.A. Petrie, identifiaient plusieurs sous-groupes de souches inféodées à diverses espèces de crucifères tant chez *L. maculans* (appelées *L. maculans* 'lepidii', par contraste avec la souche de référence isolée de colza ou *L. maculans* 'brassicae') que chez *L. biglobosa* (*L. biglobosa* 'brassicae' et *L. biglobosa* 'thlaspii') et de sous-groupes géographiques au sein des *L. biglobosa* infectant le colza (*L. biglobosa* 'canadensis', *L. biglobosa* 'occiaustralensis') (Mendes-Pereira *et al*., 2003 ; Vincenot *et al*., 2008) (Figure 1). Pour ces différents sous-groupes des données préliminaires montraient par ailleurs des structures chromosomiques et des tailles de génome différentes de celles de la souche de référence. Il s'agissait donc tout d'abord de mieux définir les relations phylogénétiques entre ces différentes entités, puis de comparer leurs structures de génome et l'incidence des ET dans la structure du génome de chacune d'entre elles.

Tableau 1. Données de séquençage et d'annotation des génomes du complexe d'espèces *L. maculans-L. biglobosa*.

	L. maculans 'brassicae'	*L. maculans* 'lepidii'	*L. biglobosa* 'brassicae'	*L. biglobosa* 'canadensis'	*L. biglobosa* 'thlaspii'
Taille du génome (Mb)	45,1	31,5	31,8	30,2	32,1
Nombre de contig	1743	2802	2533	7124	3506
Nombre de *scaffold*	76	123	606	6748	237
Scaffold N50 (kb)	1770	1356	779	245	715
Gaps (%)	2,5	7,12	7,37	0,11	8,69
Éléments répétés (%)	35,5	3,99	4,44	3,88	5,11
Éléments transposables (%)	32,5	2,7	3,2	2,9	4,0
Contenu en GC (%)	45,2	50,9	51,4	51,1	51,4
Nombre de gènes prédits	12543	11272	11390	11068	11691
Nombre d'effecteurs prédits	651	737	665	621	676

A l'aide de stratégies NGS (*Next Generation Sequencing* : 454 Roche et HiSeq Illumina) nous avons ainsi séquencé plusieurs autres souches incluant trois autres souches de *L. maculans* 'brassicae' (Lmb), une souche de *L. maculans* 'lepidii' (Lml), une souche de *L. biglobosa* 'brassicae' (Lbb) et une souche de *L. biglobosa* 'thlaspii' (Lbt). Nos collègues australiens de l'Université de Melbourne ont, quant à eux séquencé une souche de *L. biglobosa* 'canadensis' (Lbc) à l'aide de la technologie HiSeq Illumina uniquement.

Ces données nous ont permis de préciser la phylogénie des différentes souches séquencées. Elles ont ainsi indiqué que les temps de divergence entre Lmb et Lml ou entre Lbb et Lbc sont de l'ordre de 4 à 5 millions d'années (Figure 1), suggérant dans les deux cas une divergence compatible avec l'existence d'espèces distinctes qu'il faudra formellement renommer dans le futur. L'analyse des données de séquence (Tableau 1) nous a ensuite montré que les génomes des membres du complexe d'espèce, hormis la référence Lmb, ont une taille similaire à celle du génome de la plupart des ascomycètes filamenteux (entre 30 et 32 Mb) ce qui est cohérent avec nos données préliminaires basées sur l'analyse des électrocaryotypes de ces espèces. Les données obtenues sont compatibles avec une très faible quantité de régions répétées chez les membres du complexe d'espèce autres que Lmb. Ainsi le génome de Lml ne contient que 2,7 % d'ET (Tableau 1, Figure 2). L'ensemble de ces données suggère fortement que le génome de

Lmb a été envahi par des ET il y a moins de 5 millions d'années, concomitamment ou après l'émergence de Lmb en tant qu'espèce nouvelle.

Récemment, les analyses de génomique comparative entre espèces fongiques ou oomycètes plus ou moins proches se sont multipliées (Schirawski *et al*., 2010 ; De Wit *et al*., 2012 ; Ohm *et al*., 2012 ; Gan *et al*., 2013 ; Hacquard *et al*., 2013 ; Manning *et al*., 2013 ; Schardl *et al*., 2013 ; Syme *et al*., 2013). Généralement, la structure des génomes est comparable entre espèces proches, que celles-ci montrent toutes un génome riche en ET, comme pour les comparaisons entre espèces de *Colletotrichum* ou entre *Blumeria* spp. (Spanu, 2012 ; Gan *et al*., 2013), ou au contraire un faible taux d'ET chez toutes les espèces proches, comme lors de comparaisons entre différentes espèces de *Cochliobolus* (Ohm *et al*., 2012 ; Condon *et al*., 2013). Les rares cas où des comparaisons sont réalisées entre espèces différant radicalement dans leur contenu en ET concernent surtout à ce jour des entités phylogénétiquement relativement éloignées telles que *Cladosporium fulvum* et *Dothistroma septosporum* (De Wit *et al*., 2012). La série évolutive *L. maculans-L. biglobosa* est donc un modèle tout à fait original pour étudier le déterminisme d'une invasion récente de génome par des ET et les conséquences que cela peut avoir sur la structure et le fonctionnement d'un génome fongique.

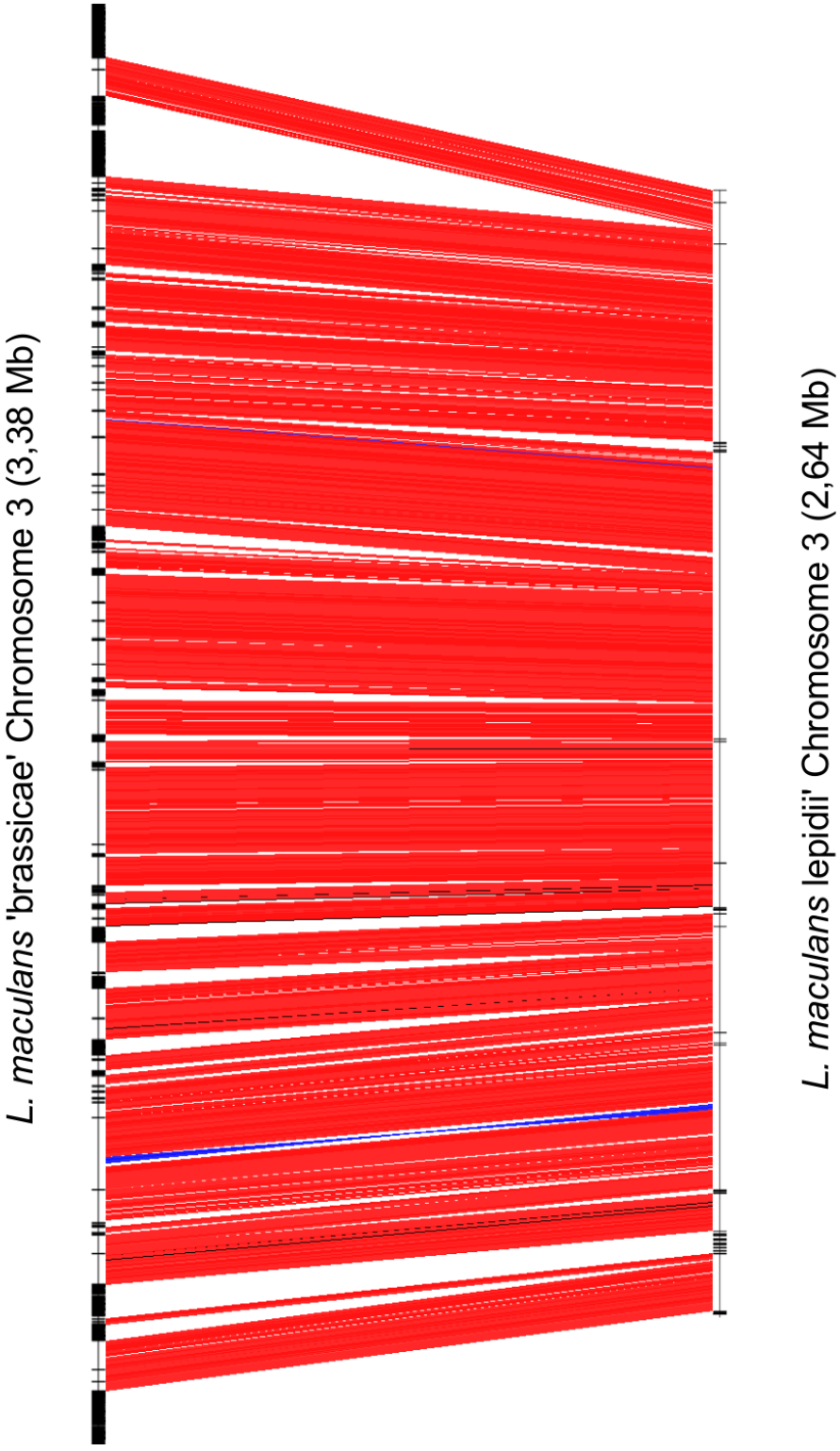


Figure 2. Synténie entre les génomes de *L. maculans* 'brassicae' et *L. maculans* 'lepidii', exemple du Chromosome 3. Le chromosome 3 de *L. maculans* est ici schématisé par une ligne sur laquelle sont représentés les ET (traits noir verticaux). Ce schéma permet de voir que les séquences se trouvant de part et d'autres de grandes régions riches en ET, les isochores AT, sont très conservées. Lorsqu'un trait rouge relie des régions synténiques entre les deux séquences cela signifie qu'elles sont orientées de la même façon, dans le cas contraire le trait est bleu. Ce schéma permet donc de visualiser deux inversions intra-chromosomiques, qui dans ce cas là ne présentent pas d'ET à leurs points de cassure ou à proximité.

Éléments transposables et spéciation

L'annotation des éléments répétés dans l'ensemble des génomes du complexe d'espèces a permis d'identifier 121 séquences réparties par la suite en 3 catégories : (i) les ET de classe I ou rétrotransposons, représentés par 24 familles, (ii) les ET de classe II ou transposons à ADN, représentés par 21 familles, et (iii) les éléments non catégorisés comprenant 52 séquences. Une étude de la proportion de ces familles dans les génomes du complexe d'espèces a mis en évidence une très forte expansion de rétrotransposons à LTR chez Lmb, comme cela est généralement le cas pour les génomes fongiques riches en ET (p. ex. Gan *et al*., 2013). Dans l'ensemble des génomes du complexe, toutes les copies d'ET sont généralement tronquées et dégénérées par le RIP, qui introduit de multiples codons stop le long de la séquence codante ; il n'y a donc actuellement plus d'activité de transposition possible pour ces éléments.

Afin d'établir la distribution de ces familles d'ET dans les génomes du complexe d'espèces et de savoir, par la même occasion, si ces éléments sont spécifiques au complexe d'espèces, leur présence a été analysée dans l'ensemble des génomes de champignons Ascomycètes disponibles dans la base de données Mycocosm du JGI [5]. En couplant ces données avec les temps de divergence calculés entre ces espèces, il est possible d'estimer la date d'invasion par les familles d'ET. Plusieurs points importants ressortent de cette étude : (i) les ET présents chez les membres du complexe d'espèce *L. maculans-L. biglobosa* sont retrouvés uniquement dans des génomes de Dothidéomycètes, et majoritairement des Pléosporales ; (ii) 66 % des familles ne sont présentes que dans le complexe *L. maculans-L. biglobosa* et sont réparties selon plusieurs niveaux de spécificité (genre, espèces, « sous-espèces ») ; (iii) trois des quatre familles d'ET les plus abondantes (rétrotransposons à LTR) chez Lmb sont présentes dans les phylogénies depuis 70 à 90 millions d'années, mais elles sont généralement très minoritaires chez les autres espèces fongiques ou chez les autres membres du complexe d'espèces (Figure 1). Des analyses phylogénétiques basées sur la divergence nucléotidique (hors nucléotides mutés par le RIP) entre les copies d'ET de chaque famille indiquent ainsi un « *burst* » de transposition pour les familles les plus représentées dans le génome de Lmb, il y a environ 5 millions d'années (Rouxel *et al.*, 2011).

Les ET sont connus pour induire des réarrangements chromosomiques tels que des translocations, des inversions ou des délétions (Zolan, 1995 ; Rebollo *et al*., 2010). Afin de savoir si cette expansion soudaine d'ET a eu des effets sur la structure du génome

de Lmb, nous avons comparé ce dernier avec le génome de Lml, pauvre en ET et phylogénétiquement proche. L'alignement obtenu présente une très grande conservation des chromosomes et de l'ordre des gènes au sein des chromosomes (macrosynténie) et l'on dénombre seulement 30 inversions intra-chromosomiques et aucune translocation entre ces deux espèces (Figure 2). Cela correspond à une image des toutes premières étapes conduisant à ce que Hane *et al*. (2011) ont appelé mésosynténie, c'est-à-dire, entre deux espèces, la conservation des chromosomes et de leur contenu en gène sans conservation de l'ordre ou de l'orientation de ces gènes au sein de ces chromosomes. Ce phénomène évolutif semble restreint aux champignons filamenteux ascomycètes et plus particulièrement aux Dothidéomycètes (Hane *et al.,* 2011 ; Ohm *et al.*, 2012). En 2012, Ohm *et al.* ont montré, à travers une simulation, que l'accumulation de séries d'inversions intra-chromosomiques génère la mésosynténie au cours de l'évolution. La comparaison de 18 génomes de Dothidéomycètes a par ailleurs mis en évidence la présence de répétitions simples (de type microsatellite) aux points de cassure des inversions et suggéré un lien mécanistique entre les remaniement intra-chromosomiques et la présence de telles répétitions (Ohm *et al.*, 2012). Chez Lmb, toutefois, de telles répétitions simples ne sont pas retrouvées et les bordures des inversions intra-chromosomiques sont, pour les deux-tiers d'entre-elles, colocalisées avec des ET ayant envahi les génomes durant ou après la spéciation Lmb-Lml.

L'importance évolutive des ET est aujourd'hui démontrée dans tous les règnes du vivant. Ils peuvent contribuer à l'innovation génique, à la perte de gènes ou à une instabilité génomique en tant que substrat pour la recombinaison méiotique conduisant à des réarrangements chromosomiques (Rebollo *et al.*, 2010). Les « *bursts* » de transposition d'éléments qui sont souvent lignée ou espèce-spécifiques sont reliés à des événement de spéciation sans qu'il soit possible de savoir s'ils en sont la cause ou la conséquence (Rebollo *et al.*, 2010). Une telle importance évolutive des ET est ainsi illustrée par nos analyses chez les Dothidéomycètes. En prenant comme référence les ET présents au sein du complexe d'espèces *L. maculans-L. biglobosa*, nous observons des « *bursts* » de transposition et des ET spécifiques d'un lignage à différents niveau de la phylogénie des Dothidéomycètes. Ainsi 18 familles sont spécifiques des Pléosporales, ayant potentiellement envahi les génomes 89 millions d'années auparavant, 13 familles sont spécifiques d'une branche postérieure au sein des Pléosporales ayant potentiellement envahi les génomes 72 millions d'années auparavant, 16 familles sont

concomitantes ou postérieures à la séparation entre *L. maculans* et *L. biglobosa* ayant potentiellement envahi les génomes 22 millions d'années auparavant (Figure 1). A chaque branche majeure de la phylogénie on note donc l'invasion par des ET spécifiques qui ont potentiellement contribué aux événements de spéciation. L'exemple le plus frappant, et sans équivalent dans la phylogénie des Dothidéomycètes concerne la séparation entre Lml et Lmb. Les deux génomes montrent une forte conservation de synténie mais diffèrent par leur contenu en ET puisque seul le génome de Lmb a subi une expansion massive d'ET il y a environ 5 millions d'années ce qui coïncide avec la date de séparation entre Lmb et Lml. Ces données sont cohérentes avec une relation causale entre « *bursts* » de transposition et spéciation. Cependant, corréler les ET et la spéciation est assez compliqué au vu des imprécisions de datation, mais aussi de l'isolement reproductif des deux nouvelles espèces permettant alors une dynamique indépendante d'invasion du génome par les ET (Rebollo *et al.*, 2010). A contrario, il semble établi dans ce système que les ET bordent, et sont sans doute la cause, de la plupart des inversions intra-chromosomiques entre les génomes de Lmb et Lml. Ces événements créent ainsi plusieurs régions génomiques (dont certaines de très grande taille contiennent plus d'une centaines de gènes) inaccessibles à la recombinaison méiotique entre les ancêtres de Lmb et Lml.

Cependant, une donnée reste obscure suite à cette étude : la transposition massive des ET dans le génome de Lmb concerne des familles ayant nouvellement envahi les génomes, comme l'on pouvait s'y attendre, mais aussi trois familles de rétrotransposons à LTR présents dans la phylogénie des Pléosporales depuis au moins 90 millions d'années, et qui sont, dans toutes les espèces, fortement dégénérées par troncatures et par RIP. Le mécanisme ayant permis une réactivation de ces familles pour leur permettre, il y a environ 5 million d'années, un « *burst* » de transposition dans le génome de Lmb seulement reste à élucider.

Naissance et mort des effecteurs

Lors d'une analyse de génomique comparative, on attend de la comparaison des protéomes la mise en évidence de différences fonctionnelles que l'on pourra ensuite relier aux différentes stratégies ou capacités parasitaires ou biologiques. Les génomes des espèces du complexe sont très proches en termes d'organisation et de séquence, il est donc normal de penser qu'une grande partie des protéines produites dans ces génomes

sera conservée au sein du complexe et c'est effectivement le cas, puisque les membres du complexe partagent près de 90 % de leurs protéines prédites entre eux. Les 10 % de séquences spécifiques de l'un ou l'autre génome représentent des pistes potentiellement explicatives de la spécialisation d'hôte ou des différentes symptomatologies sur un hôte donné. Malheureusement, comme pour beaucoup d'autres cas d'analyse de génomique comparative, ces espoirs sont en grande partie déçus puisque plus de 95 % des gènes spécifiques de l'un ou l'autre membre du complexe n'ont pas de fonctions attribuées informatiquement. Il y a donc besoin d'un énorme travail expérimental pour l'annotation fonctionnelle de ces séquences, la bioinformatique permettant uniquement de pointer du doigt les séquences à étudier.

Si la majeure partie des gènes spécifiques d'une espèce n'a pas d'annotation, deux catégories fonctionnelles semblent toutefois ressortir dans toutes les études de génomique comparative, les effecteurs peptidiques et les métabolites secondaires (Condon *et al*., 2013 ; Manning *et al*., 2013 ; Syme *et al*., 2013 ; Schardl *et al*., 2013). Depuis quelques années maintenant, l'étude de la pathogenèse fongique s'est focalisée sur un groupe de gènes spécifiques : les gènes codant pour des effecteurs. Les effecteurs sont des molécules produites par le champignon lors de l'interaction avec son hôte et jouant un rôle souvent encore obscur lors de la mise en place de la pathogénie. Le plus souvent ils correspondent à des petites protéines sécrétées (PPS). Aujourd'hui, tout champignon phytopathogène séquencé qui se respecte a son répertoire de PPS caractérisé. Ainsi, pour chaque membre du complexe *L. maculans-L. biglobosa*, le répertoire des PPS a été défini. Ces protéines varient très peu en nombre entre les génomes (650-700), elles représentent une grande partie (60 %) du sécrétome de chaque génome et leurs caractéristiques structurales sont très similaires. Pour chaque génome, 30 % de ces séquences sont conservés chez tous les membres du complexe, 40-50 % sont présents chez au moins un autre membre et 10-25 % sont spécifiques. Bien entendu plus la spécificité augmente, moins les séquences ont des fonctions associées.

Chez Lmb, les isochores AT sont enrichis en gènes codant pour des effecteurs (20 % contre 4 % dans les isochores GC). D'ailleurs, c'est dans ces compartiments qu'a été identifié l'ensemble des gènes d'avirulence de *L. maculans*. La présence de séquences codantes au sein de régions riches en ET peut être en effet bénéfique pour l'organisme, puisque des analyses ont montré que les mutations RIP pouvait « déborder » des ET vers les séquences adjacentes, leur fournissant ainsi une nouvelle source de

diversification (Rouxel *et al*, 2011). Cette association entre régions riches en ET (ou ET isolé) et effecteurs est retrouvée chez la plupart des espèces de champignons filamenteux ou d'oomycètes phytopathogènes (Orbach *et al*, 2000 ; Haas *et al*, 2009 ; Ma *et al*, 2010 ; De Wit *et al*., 2012 ; Hacquard *et al*., 2013).

Des analyses de synténie effectuées entre les génomes de Lmb et d'autres Pléosporales telles que *Phaeosphaeria nodorum* ou *Pyrenophora tritici-repentis* ont montré que les gènes directement séparés par des isochores AT (jusqu'à plusieurs centaines de kilobases) chez Lmb sont conservés et consécutifs dans les autres génomes. Par contre, certains gènes présents au sein des isochores AT, plus particulièrement codant pour des PPS, n'ont pas d'homologues dans les génomes des autres organismes. Il semblerait donc que la formation des isochores dans le génome de Lmb soit liée à la génération de nouveaux gènes. Cette hypothèse a été renforcée lors de l'étude comparative entre les génomes des membres du complexe, car chez Lmb 25 % des gènes localisés dans les isochores AT sont spécifiques, contre 10 % en isochores GC. Si l'on considère uniquement les gènes codant des effecteurs en isochores AT, cette spécificité atteint 41 %. Cette étude nous a aussi permis d'observer des « mouvements » de gènes qui pourraient être liés au « *burst* » de transposition : les orthologues des gènes d'avirulence de Lmb, lorsqu'ils existent chez un ou plusieurs membres du complexe d'espèces, sont systématiquement localisés dans des régions différentes, ou sur des chromosomes différents, alors que les gènes environnants conservent leur synténie avec leurs orthologues chez Lmb. Cette information rappelle, à une échelle moindre, la localisation chromosomique extrêmement variable dans le génome du champignon *Magnaporthe oryzae* du gène d'avirulence *AVR-Pita*, lui aussi flanqué de rétrotransposons (Chuma *et al*., 2011). Chez d'autres espèces fongiques ou d'oomycètes, on postule que les événements de transposition favorisent la duplication des gènes codant pour des effecteurs, conduisant finalement à des diversifications géniques (et à des néofonctionalisations ?) (Raffaele & Kamoun, 2012), et il est tentant de penser qu'au cours de l'évolution le déplacement de ces gènes dans des régions très dynamiques a mené à l'apparition ou l'acquisition de nouvelles capacités parasitaires chez Lmb. Les ET peuvent aussi avoir, et c'est bien connu, un rôle négatif quand ils viennent s'insérer dans une séquence codante. Ce cas permet de créer de la diversité entre deux espèces, mais aussi entre deux isolats d'une même espèce. En comparant les génomes de deux isolats de Lmb, on note du polymorphisme d'insertion d'ET relativement jeunes, c'est-à-dire

spécifiques de l'espèce *L. maculans*, pouvant engendrer de la spécificité génique.

Structure du génome et adaptation aux pressions de sélection

L'utilisation massive de résistances variétales de type « gène majeur » conduit à une forte pression de sélection contre un gène unique du champignon et à des cycles récurrents de « *boom-and-bust* » dans lesquels l'utilisation d'une nouvelle résistance variétale conduit à une valeur ajoutée pour la culture et donc à un succès massif des variétés possédant cette nouvelle résistance, suivi très rapidement par une perte d'efficacité de la résistance suite à l'apparition et à la dissémination de souches mutantes devenues virulentes (McDonald & Linde, 2002). Si ce phénomène est décrit depuis longtemps, la vitesse d'apparition et de dissémination des souches virulentes (souvent 2 à 3 ans suffisent pour qu'une résistance variétale soit contournée) et les mécanismes moléculaires permettant ce contournement étaient restés obscurs jusqu'à très récemment. En effet, si diverses études analysant (souvent rétrospectivement) des souches ou des populations fongiques devenues virulentes vis-à-vis d'une résistance décrivent de nombreux événements moléculaires conduisant généralement à l'inactivation totale de celui-ci (mutations ponctuelles du gène d'avirulence, délétion totale ou partielle, insertion d'ET, mutations dues au RIP (Kang *et al*., 2001 ; Rep *et al*., 2004 ; Fudal *et al*., 2005, 2009 ; Farman, 2007 ; Gout *et al*., 2007 ; Stergiopoulos *et al*., 2007 ; Zhou *et al*., 2007), les tous premiers événements moléculaires permettant le contournement et leur dissémination dans les populations restaient inconnus. Pour clarifier ce point nous avons mis en place un essai pluri-annuel dans lequel des populations avirulentes de Lmb étaient confrontées pour la première fois à une nouvelle résistance, *Rlm7* (Daverdin *et al*., 2012). Le gène fongique soumis à sélection, *AvrLm7*, étant localisé en isochore AT (Parlange *et al*., 2009), comme tous les autres gènes d'avirulence connus chez Lmb, cette expérience avait ainsi pour objet d'élucider le lien entre dynamique évolutive et structure du génome. Dans le cadre de cette expérience, nous avons effectivement observé un contournement de la résistance, avec des populations qui, de totalement avirulentes sont devenues virulentes pour 40 % des souches échantillonnées, et ce en trois ans seulement (Daverdin *et al*., 2012). De façon surprenante, nous avons observé, dans une petite parcelle de 0,25 ha, tous les événements moléculaires décrits dans la littérature, voire d'autres non encore décrits, pour l'inactivation ou la mutation de gènes d'avirulence, avec une sur-représentation des mutations dues au RIP lors des premières années d'expérience, et une

prévalence globale des événements de délétion partielle ou totale (Daverdin *et al*., 2012). Cette étude confirme l'importance de l'environnement génomique des gènes d'avirulence sur la capacité que montre l'agent pathogène à s'adapter à une pression ciblée en particulier lors des cycles de reproduction sexuée. Une étude récente montre toutefois que de tels événements liés à un environnement génomique riche en ET est aussi retrouvé chez des espèces majoritairement asexuées, telles que *Verticillium dahliae* (De Jonge *et al*., 2013). L'environnement génomique couplé avec un mode de vie dans lequel la reproduction sexuée est obligatoire favorise des phénomènes d'hypermutabilité tels que le RIP, dont on a vu qu'il peut déborder des régions riches en ET vers des gènes copie-unique inclus dans de grandes régions riches en ET, ou des délétions induites par des mésappariements chromosomiques entre ET lors de la méiose. Dans le cas de Lmb, mais aussi probablement dans de nombreux autres cas de champignons phytopathogènes pour lesquels les gènes codant des effecteurs et/ou des gènes d'avirulence sont localisés dans des régions génomiques riches en ET il apparaît donc qu'une telle localisation présente un double avantage adaptatif en permettant une diversification génique et pouvant générer de nouvelles fonctionnalités, mais aussi en favorisant des phénomènes d'hypermutabilité explicatifs de la rapidité avec laquelle on observe des contournements de résistance variétale au champ.

Conclusion

L'ensemble des travaux exposés ici indique la valeur d'exemple générique du complexe d'espèces *L. maculans-L. biglobosa.* Ils suggèrent ainsi fortement que, dans les phylogénies fongiques, les événements d'invasion/expansion par des ET touchent majoritairement des branches terminales et donc sont « récents », même si leur déterminisme reste peu clair en particulier pour les cas observés chez *L. maculans* de « réactivation » d'ET inactivés par des mutations massives de type RIP. Ces phénomènes d'invasion/expansion, concourent comme chez de nombreux autres organismes à des remaniements chromosomiques et potentiellement à la génération de nouvelles espèces. Dans de nombreux cas d'espèces ascomycètes, des mécanismes de protection du génome tel que le RIP permettent de limiter l'expansion des ET et d'aboutir à un équilibre entre inflation de taille de génome et maintien de régions dispensables présentant un avantage évolutif ou sélectif. Toutefois, dans certains cas extrêmes chez lesquels le RIP ou autres mécanismes de protection du génome sont inactifs (cas par exemple des

oïdiums et dans une moindre mesure des rouilles et des mildious) on assiste à une inflation de taille de génome (parfois largement supérieure à 100 Mb) liée à un envahissement massif du génome par les ET et résultant en une perte massive de gènes et un mode de vie strictement biotrophe, l'agent pathogène devenant incapable de vivre en dehors de son hôte en tant que saprophyte comme le font la plupart des champignons phytopathogènes. De façon moins drastique, l'ensemble de ces études montre que les champignons ont quasi-systématiquement généré et utilisent des génomes à deux vitesses, dont le compartiment « plastique » enrichi en ET peut prendre des formes diverses (chromosomes dispensables, amplification des régions sub-télomèriques, isochores AT). On observe une sélection pour la localisation de gènes importants pour la pathogenèse et l'adaptation dans de telles régions plastiques du génome avec des conséquences sur la diversification/expansion du répertoire d'effecteurs, des conséquences sur l'adaptation aux pressions dues à l'hôte, des capacités accrues de transfert horizontaux entre espèces partageant la même niche écologique et sans doute des conséquences sur la régulation de l'expression des gènes inclus dans de telle régions génomiques.

Références

Anderson P.K., Cunningham A.A., Patel N.G., Morales F.J., Epstein P.R., Daszak P. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Evol Ecol*, 2004, 19, 535-544.

Biffen R.H. The inheritance of sterility in the barleys. *J Agric Sci*, 1905, 1, 250-257.

Blaise F., Rémy E., Meyer M., Zhou L., Narcy J.P., Roux J., Balesdent M.H., Rouxel T. A critical assessment of *Agrobacterium tumefaciens*-mediated transformation as a tool for pathogenicity gene discovery in the phytopathogenic fungus *Leptosphaeria maculans*. *Fungal Genet Biol*, 2007, 44, 123-138.

Chuma I., Isobe C., Hotta Y., Ibaragi K., Futamata N., Kusaba M., Yoshida K., Terauchi R., Fujita Y., Nakayashiki H., Valent B., Tosa Y. Multiple translocation of the AVR-Pita effector gene among chromosomes of the rice blast fungus *Magnaporthe oryzae* and related species. *PloS Pathog*, 2011, 7, e1002147.

Condon B.J., Leng Y., Wu D., Bushley K.E., Ohm R.A., Otillar R., Martin J., Schackwitz W., Grimwood J., MohdZainudin N., Xue C., Wang R., Manning V.A., Dhillon B., Tu Z.J., Steffenson B.J., Salamov A., Sun H., Lowry S., LaButti K., Han J., Copeland A., Lindquist E., Barry K., Schmutz J., Baker S.E., Ciuffetti L.M., Grigoriev I.V., Zhong S., Turgeon B.G. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PloS Genet*, 2013, 9, e1003233.

Cuomo C.A., Güldener U., Xu J.R., Trail F., Turgeon B.G., Di Pietro A., Walton J.D., Ma L.J., Baker S.E., Rep M., Adam G., Antoniw J., Baldwin T., Calvo S., Chang Y.L., Decaprio D., Gale L.R., Gnerre S., Goswami R.S., Hammond-Kosack K., Harris L.J., Hilburn K., Kennell J.C., Kroken S., Magnuson J.K., Mannhaupt G., Mauceli E., Mewes H.W., Mitterbauer R., Muehlbauer G., Münsterkötter M., Nelson D., O'donnell K., Ouellet T., Qi W., Quesneville H., Roncero M.I., Seong K.Y., Tetko I.V., Urban M., Waalwijk C., Ward T.J., Yao J., Birren B.W., Kistler H.C. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, 2007, 317, 1400-1402.

Daverdin G., Rouxel T., Gout L., Aubertot J.N., Fudal I., Meyer M., Parlange F., Carpezat J., Balesdent M.H. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PloS Pathog*, 2012, 8, e1003020.

Dean R.A., Talbot N.J., Ebbole D.J., Farman M.L., Mitchell T.K., Orbach M.J., Thon M., Kulkarni R., Xu J.R., Pan H., Read N.D., Lee Y.H., Carbone I., Brown D., Oh Y.Y., Donofrio N., Jeong J.S., Soanes D.M., Djonovic S., Kolomiets E., Rehmeyer C., Li W., Harding M., Kim S., Lebrun M.H., Bohnert H., Coughlan S., Butler J., Calvo S., Ma L.J., Nicol R., Purcell S., Nusbaum C., Galagan J.E., Birren B.W. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, 2005, 434, 980-986.

De Jonge R., Bolton M.D., Kombrink A., van den Berg G.C., Yadeta K.A, Thomma B.P. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res*, 2013, 23, doi:10.1101/gr.152660.112

De Wit P.J., van der Burgt A., Ökmen B., Stergiopoulos I., Abd-Elsalam K.A., Aerts A.L., Bahkali A.H., Beenen H.G., Chettri P., Cox M.P., Datema E., de Vries R.P., Dhillon B., Ganley A.R., Griffiths S.A., Guo Y., Hamelin R.C., Henrissat B., Kabir M.S., Jashni M.K., Kema G., Klaubauf S., Lapidus A., Levasseur A., Lindquist E., Mehrabi R., Ohm R.A., Owen T.J., Salamov A., Schwelm A., Schijlen E., Sun H., van den Burg H.A., van Ham R.C., Zhang S., Goodwin S.B., Grigoriev I.V., Collemare J., Bradshaw R.E. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PloS Genet*, 2012, 8, e1003088

Duplessis S., Cuomo C.A., Lin Y.C., Aerts A., Tisserant E., Veneault-Fourrey C., Joly D.L., Hacquard S., Amselem J., Cantarel B.L., Chiu R., Coutinho P.M., Feau N., Field M., Frey P., Gelhaye E., Goldberg J., Grabherr M.G., Kodira C.D., Kohler A., Kües U., Lindquist E.A., Lucas S.M., Mago R., Mauceli E., Morin E., Murat C., Pangilinan J.L., Park R., Pearson M., Quesneville H., Rouhier N., Sakthikumar S., Salamov A.A., Schmutz J., Selles B., Shapiro H., Tanguay P., Tuskan G.A., Henrissat B., Van de Peer Y., Rouzé P., Ellis J.G., Dodds P.N., Schein J.E., Zhong S., Hamelin R.C., Grigoriev I.V., Szabo L.J., Martin F. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci U.S.A.*, 2011, 108, 9166-9171.

Farman M.L. Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it. *FEMS Microbiol lett*, 2007, 273, 125-132.

Fisher M.C., Henk D.A., Briggs C.J., Brownstein J.S., Madoff L.C., McCraw S.L., Gurr S.J. Emerging fungal threats to animal, plant and ecosystem health. *Nature*, 2012, 484, 186-194.

Fudal I., Böhnert H.U., Tharreau D., Lebrun M.H. Transposition of MINE, a composite retrotransposon, in the avirulence gene *ACE1* of the rice blast fungus *Magnaporthe grisea*. *Fungal Genet Biol*, 2005, 42, 761-772.

Fudal I., Ross S., Brun H., Besnard A.L., Ermel M., Kuhn M.L., Balesdent M.H., Rouxel T. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. *Mol Plant Microbe Interact*, 2009, 22, 932-941.

Galagan J.E., Calvo S.E., Borkovich K.A., Selker E.U., Read N.D., Jaffe D., FitzHugh W., Ma L.J., Smirnov S., Purcell S., Rehman B., Elkins T., Engels R., Wang S., Nielsen C.B., Butler J., Endrizzi M., Qui D., Ianakiev P., Bell-Pedersen D., Nelson M.A., Werner-Washburne M., Selitrennikoff C.P., Kinsey J.A., Braun E.L., Zelter A., Schulte U., Kothe G.O., Jedd G., Mewes W., Staben C., Marcotte E., Greenberg D., Roy A., Foley K., Naylor J., Stange-Thomann N., Barrett R., Gnerre S., Kamal M., Kamvysselis M., Mauceli E., Bielke C., Rudd S., Frishman D., Krystofova S., Rasmussen C., Metzenberg R.L., Perkins D.D., Kroken S., Cogoni C., Macino G., Catcheside D., Li W., Pratt R.J., Osmani S.A., DeSouza C.P., Glass L., Orbach M.J., Berglund J.A., Voelker R., Yarden O., Plamann M., Seiler S., Dunlap J., Radford A., Aramayo R., Natvig D.O., Alex L.A., Mannhaupt G., Ebbole D.J., Freitag M., Paulsen I., Sachs M.S., Lander E.S., Nusbaum C., Birren B.W. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, 2003, 422, 859-868.

Gan P., Ikeda K., Irieda H., Narusaka M., O'Connell R.J., Narusaka Y., Takano Y., Kubo Y., Shirasu K. Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum* fungi. *New Phytol*, 2013, 197, 1236-1249.

Giraud T., Gladieux P., Gavrilets S. Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol Evol*, 2010, 25, 387-395.

Gout L., Kuhn M.L., Vincenot L., Bernard-Samain S., Cattolico L., Barbetti M., Moreno-Rico O., Balesdent M.H., Rouxel T. Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. *Environ Microbiol*, 2007, 9, 2978-2992.

Haas B.J., Kamoun S., Zody M.C., Jiang R.H., Handsaker R.E., Cano L.M., Grabherr M., Kodira C.D., Raffaele S., Torto-Alalibo T., Bozkurt T.O., Ah-Fong A.M., Alvarado L., Anderson V.L., Armstrong M.R., Avrova A., Baxter L., Beynon J., Boevink P.C., Bollmann S.R., Bos J.I., Bulone V., Cai G., Cakir C., Carrington J.C., Chawner M., Conti L., Costanzo S., Ewan R., Fahlgren N., Fischbach M.A., Fugelstad J., Gilroy E.M., Gnerre S., Green P.J., Grenville-Briggs L.J., Griffith J., Grünwald N.J., Horn K., Horner N.R., Hu C.H., Huitema E., Jeong D.H., Jones A.M., Jones J.D., Jones R.W., Karlsson E.K., Kunjeti S.G., Lamour K., Liu Z., Ma L., Maclean D., Chibucos M.C., McDonald H., McWalters J., Meijer H.J., Morgan W., Morris P.F., Munro C.A., O'Neill K., Ospina-Giraldo M., Pinzón A., Pritchard L., Ramsahoye B., Ren Q., Restrepo S., Roy S., Sadanandom A., Savidor A., Schornack S., Schwartz D.C., Schumann U.D., Schwessinger B., Seyer L., Sharpe T., Silvar C., Song J., Studholme D.J., Sykes S., Thines M., van de Vondervoort P.J., Phuntumart V., Wawra S., Weide R., Win J., Young C., Zhou S., Fry W., Meyers B.C., van West P., Ristaino J., Govers F., Birch P.R., Whisson S.C., Judelson H.S., Nusbaum C. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, 2009, 461, 393-398.

Hane J.K., Lowe R.G.T., Solomon P.S., Tan K.C., Schoch C.L., Spatafora J.W., Crous P.W., Kodira C., Birren B.W., Galagan J.E., Torriani S.F.F., McDonald B.A., Oliver R.P. Dothideomycete–plant Interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum. Plant Cell*, 2007, 19, 3347-3368.

Hane J.K., Rouxel T., Howlett B.J., Kema G.H., Goodwin S.B., Oliver R.P. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol*, 2011, 12, R45.

Kämper J., Kahmann R., Bölker M., Ma L.J., Brefort T., Saville B.J., Banuett F., Kronstad J.W., Gold S.E., Müller O., Perlin M.H., Wösten H.A., de Vries R., Ruiz-Herrera J., Reynaga-Peña C.G., Snetselaar K., McCann M., Pérez-Martín J., Feldbrügge M., Basse CW., Steinberg G., Ibeas J.I., Holloman W., Guzman P., Farman M., Stajich J.E., Sentandreu R., González-Prieto J.M., Kennell J.C., Molina L., Schirawski J., Mendoza-Mendoza A., Greilinger D., Münch K., Rössel N., Scherer M., Vranes M., Ladendorf O., Vincon V., Fuchs U., Sandrock B., Meng S., Ho E.C., Cahill M.J., Boyce K.J., Klose J., Klosterman S.J., Deelstra H.J., Ortiz-Castellanos L., Li W., Sanchez-Alonso P., Schreier P.H., Häuser-Hahn I., Vaupel M., Koopmann E., Friedrich G., Voss H., Schlüter T., Margolis J., Platt D., Swimmer C., Gnirke A., Chen F., Vysotskaia V., Mannhaupt G., Güldener U., Münsterkötter M., Haase D., Oesterheld M., Mewes H.W., Mauceli E.W., DeCaprio D., Wade C.M., Butler J., Young S., Jaffe D.B., Calvo S., Nusbaum C., Galagan

J., Birren B.W. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature*, 2006, 444, 97-101.

Kang S., Lebrun M.H., Farrall L., Valent B. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. *Mol Plant Microbe Interact*, 2001, 14, 671-674.

Kupferschmidt K. Mycology. Attack of the clones. *Science*, 2012, 337, 636-638.

Ma L.J., van der Does H.C., Borkovich K.A., Coleman J.J., Daboussi M.J., Di Pietro A., Dufresne M., Freitag M., Grabherr M., Henrissat B., Houterman P.M., Kang S., Shim W.B., Woloshuk C., Xie X., Xu J.R., Antoniw J., Baker S.E., Bluhm B.H., Breakspear A., Brown D.W., Butchko R.A., Chapman S., Coulson R., Coutinho P.M., Danchin E.G., Diener A., Gale L.R., Gardiner D.M., Goff S., Hammond-Kosack K.E., Hilburn K., Hua-Van A., Jonkers W., Kazan K., Kodira C.D., Koehrsen M., Kumar L., Lee Y.H., Li L., Manners J.M., Miranda-Saavedra D., Mukherjee M., Park G., Park J., Park S.Y., Proctor R.H., Regev A., Ruiz-Roldan M.C., Sain D., Sakthikumar S., Sykes S., Schwartz D.C., Turgeon B.G., Wapinski I., Yoder O., Young S., Zeng Q., Zhou S., Galagan J., Cuomo C.A., Kistler H.C., Rep M. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, 2010, 464, 367-373.

Machida M., Asai K., Sano M., Tanaka T., Kumagai T., Terai G., Kusumoto K., Arima T., Akita O., Kashiwagi Y., Abe K., Gomi K., Horiuchi H., Kitamoto K., Kobayashi T., Takeuchi M., Denning D.W., Galagan J.E., Nierman W.C., Yu J., Archer D.B., Bennett J.W., Bhatnagar D., Cleveland T.E., Fedorova N.D., Gotoh O., Horikawa H., Hosoyama A., Ichinomiya M., Igarashi R., Iwashita K., Juvvadi P.R., Kato M., Kato Y., Kin T., Kokubun A., Maeda H., Maeyama N., Maruyama J., Nagasaki H., Nakajima T., Oda K., Okada K., Paulsen I., Sakamoto K., Sawano T., Takahashi M., Takase K., Terabayashi Y., Wortman J.R., Yamada O., Yamagata Y., Anazawa H., Hata Y., Koide Y., Komori T., Koyama Y., Minetoki T., Suharnan S., Tanaka A., Isono K., Kuhara S., Ogasawara N., Kikuchi H. Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, 2005, 438, 1157-1161.

Manning V.A., Pandelova I., Dhillon B., Wilhelm L.J., Goodwin S.B., Berlin A.M., Figueroa M., Freitag M., Hane J.K., Henrissat B., Holman W.H., Kodira C.D., Martin J., Oliver R.P., Robbertse B., Schackwitz W., Schwartz D.C., Spatafora J.W., Turgeon B.G., Yandava C., Young S., Zhou S., Zeng Q., Grigoriev I.V., Ma L.J., Ciuffetti L.M. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3*, 2013, 3, 41-63.

McDonald B.A., Linde C. Pathogen population genetics, evolutionary potential, and durable resistance. *Annu Rev Phytopathol*, 2002, 40, 349-379.

Mendes-Pereira E., Balesdent M.H., Brun H., Rouxel T. Molecular phylogeny of the *Leptosphaeria maculans-L. biglobosa* species complex. *Mycol Res,* 2003, 107**,** 1287-1304.

Ohm R.A., Feau N., Henrissat B., Schoch C.L., Horwitz B.A., Barry K.W., Condon B.J., Copeland A.C., Dhillon B., Glaser F., Hesse C.N., Kosti I., LaButti K., Lindquist E.A., Lucas S., Salamov A.A., Bradshaw R.E., Ciuffetti L., Hamelin R.C., Kema G.H., Lawrence C.,

Scott J.A., Spatafora J.W., Turgeon B.G., de Wit P.J., Zhong S., Goodwin S.B., Grigoriev I.V. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PloS Pathog*, 2012, 8, e1003037.

Olsen L.A., Choffnes E.R., Relman D.A., Pray L. Fungal diseases: an emerging threat to human, animal, and plant health: Workshop Summary. 2011, The National Academies Press, Washington DC, USA, http://www.nap.edu/catalog.php?record_id=13147.

Orbach M.J., Farrall L., Sweigard J.A., Chumley F.G., Valent B. A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. *Plant Cell*, 2000, 12, 2019-2032.

Parlange F., Daverdin G., Fudal I., Kuhn M.L., Balesdent M.H., Blaise F., Grezes-Besset B., Rouxel T. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents Rlm4-mediated recognition through a single amino acid change. *Mol Microbiol*, 2009, 71, 851-863.

Pennisi E. Armed and dangerous. *Science*, 2010, 327, 804-805.

Raffaele S., Win J., Cano L.M., Kamoun S. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. *BMC Genomics*, 2010, 11, 637.

Raffaele S., Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol*, 2012, 10, 417-430.

Rebollo R., Horard B., Hubert B., Vieira C. Jumping genes and epigenetics: Towards new species. *Gene*, 2010, 454, 1-7.

Rep M., Meijer M., Houterman P.M., van der Does H.C., Cornelissen B.J. A small, cysteine-rich protein secreted by *Fusarium oxysporum* during colonization of xylem vessels is required for I-3-mediated resistance in tomato. *Mol Microbiol*, 2004, 53, 1373-1383.

Rouxel T., Balesdent M.H. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Mol Plant Pathol*, 2005, 6, 225-241.
Rouxel T., Penaud A., Pinochet X., Brun H., Gout L., Delourme R., Schmit J., Balesdent, M. H. A 10-year survey of populations of *Leptosphaeria maculans* in France indicates a rapid adaptation towards the *Rlm1* resistance gene of oilseed rape. *Eur J Plant Pathol*, 2003, 109, 871-881.

Rouxel T., Grandaubert J., Hane J.K., Hoede C., van de Wouw A.P., Couloux A., Dominguez V., Anthouard V., Bally P., Bourras S., Cozijnsen A.J., Ciuffetti L.M., Degrave A., Dilmaghani A., Duret L., Fudal I., Goodwin S.B., Gout L., Glaser N., Linglin J., Kema G.H., Lapalu N., Lawrence C.B., May K., Meyer M., Ollivier B., Poulain J., Schoch C.L., Simon A., Spatafora J.W., Stachowiak A., Turgeon B.G., Tyler B.M., Vincent D., Weissenbach J., Amselem J., Quesneville H., Oliver R.P., Wincker P., Balesdent M.H., Howlett B.J. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point Mutations. *Nat Commun*, 2011, 2, 202.

Schardl C.L., Young C.A., Hesse U., Amyotte S.G., Andreeva K., Calie P.J., Fleetwood D.J., Haws D.C., Moore N., Oeser B., Panaccione D.G., Schweri K.K., Voisey C.R., Farman M.L., Jaromczyk J.W., Roe B.A., O'Sullivan D.M., Scott B., Tudzynski P., An Z., Arnaoudova E.G., Bullock C.T., Charlton N.D., Chen L., Cox M., Dinkins R.D., Florea S., Glenn A.E., Gordon A., Güldener U., Harris D.R., Hollin W., Jaromczyk J., Johnson R.D., Khan A.K., Leistner E., Leuchtmann A., Li C., Liu J., Liu J., Liu M., Mace W., Machado C., Nagabhyru P., Pan J., Schmid J., Sugawara K., Steiner U., Takach J.E., Tanaka E., Webb J.S., Wilson E.V., Wiseman J.L., Yoshida R., Zeng Z. Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the *clavicipitaceae* reveals dynamics of alkaloid loci. *PloS Genet*, 2013, 9, e1003323.

Schirawski J., Mannhaupt G., Münch K., Brefort T., Schipper K., Doehlemann G., Di Stasio M., Rössel N., Mendoza-Mendoza A., Pester D., Müller O., Winterberg B., Meyer E., Ghareeb H., Wollenberg T., Münsterkötter M., Wong P., Walter M., Stukenbrock E., Güldener U., Kahmann R. Pathogenicity determinants in smut fungi revealed by genome comparison. *Science*, 2010, 330, 1546-1548.

Shoemaker R.A., Brun H. The telemorph of the weakly aggressive segregate of *Leptosphaeria maculans*. *Can J Bot*, 2001, 79, 412-419.

Silar P. La mycologie au début du 21ème siècle : crise et renouveau. *Biol Auj*, 2013, ce numéro.

Spanu P.D. The genomics of obligate (and nonobligate) biotrophs. *Annu Rev Phytopathol*, 2012, 50, 91-109.

Spanu P.D., Abbott J.C., Amselem J., Burgis T.A., Soanes D.M., Stüber K., Ver Loren van Themaat E., Brown J.K., Butcher S.A., Gurr S.J., Lebrun M.H., Ridout C.J., Schulze-Lefert P., Talbot N.J., Ahmadinejad N., Ametz C., Barton G.R., Benjdia M., Bidzinski P., Bindschedler L.V., Both M., Brewer M.T., Cadle-Davidson L., Cadle-Davidson M.M., Collemare J., Cramer R., Frenkel O., Godfrey D., Harriman J., Hoede C., King B.C., Klages S., Kleemann J., Knoll D., Koti P.S., Kreplak J., López-Ruiz F.J., Lu X., Maekawa T., Mahanil S., Micali C., Milgroom M.G., Montana G., Noir S., O'Connell R.J., Oberhaensli S., Parlange F., Pedersen C., Quesneville H., Reinhardt R., Rott M., Sacristán S., Schmidt S.M., Schön M., Skamnioti P., Sommer H., Stephens A., Takahara H., Thordal-Christensen H., Vigouroux M., Wessling R., Wicker T., Panstruga R. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, 2010, 330, 1543-1546.

Stergiopoulos I., De Kock M.J., Lindhout P., De Wit P.J. Allelic variation in the effector genes of the tomato pathogen *Cladosporium fulvum* reveals different modes of adaptive evolution. *Mol Plant Microbe Interact*, 2007, 20, 1271-1283.

Strange R.N., Scott P.R. Plant disease: a threat to global food security. *Annu Rev Phytopathol*, 2005, 43, 83-116.

Stukenbrock E.H., McDonald B.A. The origins of plant pathogens in agro-ecosystems. *Annu Rev Phytopathol*, 2008, 46, 75-100.

Stukenbrock E.H, Bataillon T. A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PloS Pathog*, 2012, 8, e1002893.

Syme R.A., Hane J.K., Friesen T.L., Oliver R.P. Resequencing and Comparative Genomics of *Stagonospora nodorum*: Sectional Gene Absence and Effector Discovery. *G3*, 2013, 3, 959-969.

Vincenot L., Balesdent M.H., Li H., Barbetti M.J., Sivasithamparam K., Gout L., Rouxel T. Occurrence of a new subclade of *Leptosphaeria biglobosa* in Western Australia. *Phytopathology*, 2008, 98, 321-329.

West J.S., Fitt B.D.L., Leech P.K., Biddulph J.E., Balesdent M.H., Huang Y.J. Effects of timing of *Leptosphaeria maculans* ascospore release and fungicide regime on phoma leaf spot and phoma stem canker development on winter oilseed rape *Brassica napus* in southern England. *Plant Pathol,* 2002, 51, 454-463.

Zhou E., Jia Y., Singh P., Correll J.C., Lee F.N. Instability of the *Magnaporthe oryzae* avirulence gene *AVR-Pita* alters virulence. *Fungal Genet Biol*, 2007, 44, 1024-1034.

Zolan M.E. Chromosome-length polymorphism in fungi. *Microbiol Rev*, 1995, 59, 686-698.

[1] http://www.thesundaytimes.co.uk/sto/news/uk_news/Environment/article1238345.ece
[2] https://www.youtube.com/watch?v=HIll-_blL5c&list=UU_iFLW6c8oHkYuWnc8hzTqA
[3] http://news.bbc.co.uk/2/hi/programmes/newsnight/8912727.stm
[4] http://www.miamiherald.com/2013/05/19/3405435/colombias-fight-against-coffee.html
[5] http://genome.jgi.doe.gov/programs/fungi/index.jsf

PERSPECTIVES

Depuis mon arrivée dans l'équipe d'accueil lors mon stage de M2 jusqu'à la fin de cette thèse, l'exploitation du génome de *L. maculans*, et également des autres membres du complexe d'espèces *L. maculans-L. biglobosa*, a permis d'apporter de nouveaux indices sur la caractérisation de nouveaux déterminants du pouvoir pathogène nécessaires à une meilleure compréhension des organismes fongiques phytopathogènes et au développement de stratégies de lutte efficaces et durables, mais aussi, de façon plus générale, sur l'évolution des génomes et son incidence sur la biologie des organismes. Ces travaux se sont déroulés lors d'une période faste en ce qui concerne les analyses de génomique, de génomique comparative et de génomique évolutive chez les champignons, et ont permis à *L. maculans* de se maintenir comme un organisme modèle pour l'étude des champignons phytopathogènes et de devenir, qui sait, un organisme modèle pour l'étude de l'évolution des génomes chez les eucaryotes.

Le point de départ de mon travail était simple : localiser des gènes codant des PPS dans des régions riches en AT. Grâce à ce postulat de départ, j'ai développé une analyse structurelle du génome de *L. maculans* 'brassicae' mettant à jour l'existence systématique de nombreuses grandes régions riches en AT le long de tous les chromosomes et une structure en isochore du génome considérée alors comme inédite chez les champignons et jusque récemment, comme spécifique de cette espèce fongique. L'annotation des gènes situés dans ces isochores AT et dans le reste du génome a permis d'établir le répertoire, le plus exhaustif possible, des effecteurs putatifs de Lmb grâce à un pipeline informatique dédié, qui depuis a été réutilisé pour identifier les PPS chez les autres membres du complexe d'espèces, mais aussi dans d'autres espèces fongiques comme *Venturia inaequalis* (ANR FungIsochores). L'analyse comparative des différents génomes a permis d'inférer le nombre de chromosomes présents chez les espèces *Leptosphaeria* étudiées ainsi que la spécificité du chromosome dispensable à Lmb. L'identification de ces isochores AT nous a ensuite amené à étudier leur composition, ce qui s'est traduit par une annotation et une classification des ET chez Lmb puis dans l'ensemble des génomes séquencés du complexe d'espèces. Grâce à deux méthodes différentes, nous avons pu (i) dater l'apparition de ces ET dans les génomes du complexe d'espèces mais aussi dans la lignée des Dothidéomycètes et (ii) approximer la date d'expansion massive des ET dans le génome de Lmb, ce qui nous a permis de construire des hypothèses évolutives au sein des Dothidéomycètes et du complexe d'espèces. Mon travail a aussi contribué, comme d'autres (Fudal *et al.*, 2009 ; Daverdin *et al.*, 2012) à démontrer l'importance du RIP dans l'évolution des gènes codant des effecteurs et j'ai ainsi participé à la généralisation du

concept selon lequel un lien fort existe chez les champignons filamenteux phytopathogènes entre ET et gènes impliqués dans la pathogenèse ou l'adaptation à l'hôte, et de façon plus générique à l'existence chez les champignons de génome à « deux vitesses ».

Durant ma thèse, j'ai pu utiliser librement de nombreux outils informatiques qui étaient nécessaires au bon déroulement de mon travail, comme par exemple pour les alignements de génomes, l'identification des éléments répétés ou le *clustering* des séquences, mais j'ai surtout eu la chance de pouvoir développer mes propres outils qui m'ont permis de répondre plus précisément aux différentes attentes, notamment en ce qui concerne l'identification des isochores AT, l'annotation des gènes et de leur fonction, l'identification des effecteurs, la recherche de motifs de translocation ou l'analyse de la composition des séquences (indices de RIP, usage des codons). J'ai aussi pu mettre à contribution mes connaissances en informatique dans d'autres projets : (i) j'ai réalisé un pipeline permettant une génération automatique et systématique des marqueurs génétiques dans une séquence génomique, ce qui permet d'obtenir la saturation de la carte génétique et ainsi améliorer les études populationnelles et la découverte de nouveaux gènes du pouvoir pathogène. Ce travail est décrit dans l'article « FONZIE: an optimized pipeline for minisatellite markers discovery and primer design from large sequence data sets » publié dans la revue scientifique en ligne *BMC Research Notes* en novembre 2010 (Annexe 1, article 1) et (ii) j'ai participé à l'étude de l'incidence de la structure du génome sur la génération de mutants via l'intégration d'ADN-T dont les résultats ont été décrits dans l'article « The incidence of genome structure, DNA asymmetry and cell physiology on T-DNA integration in chromosomes of the phytopathogenic fungus *Leptosphaeria maculans* » publié dans la revue scientifique *G3* en août 2012 (Annexe 1, article 2).

Sur les bases de mon travail, diverses perspectives sont maintenant ouvertes, certaines concernant plus spécifiquement *L. maculans*, comme la caractérisation de nouveaux gènes d'avirulence, alors que d'autres ont une portée plus générique et peuvent être effectivement appliqués à d'autres modèles, comme la régulation de l'expression des gènes impliqués dans l'interaction ou l'évolution des génomes et de leur architecture.

Caractérisation de nouveaux gènes d'avirulence chez *L. maculans*.

Les répertoires d'effecteurs potentiels prédits au sein des génomes de *L. maculans* 'brassicae' que j'ai généré au cours de cette thèse mais aussi des autres membres du complexe d'espèces représentent une bonne base de départ pour caractériser de nouveaux gènes impliqués dans le pouvoir pathogène et en particulier de nouveaux gènes d'avirulence. En se basant sur les données de séquence, les répertoires d'effecteurs générés, la localisation génomique et les caractéristiques d'expression, il est donc possible d'obtenir rapidement les meilleurs candidats. Nous avons ainsi appliqué cette stratégie pour l'identification de plusieurs nouveaux gènes d'avirulence. Ainsi, un nouveau gène d'avirulence localisé sur le chromosome dispensable de *L. maculans* 'brassicae' a pu être caractérisé au cours de ma thèse. Cette étude est décrite dans l'article « The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa* » publié dans la revue scientifique *The New Phytologist* en février 2013 (Annexe 1, article 3). De même, mes travaux sont à la base de l'identification de candidats pour deux nouveaux gènes d'avirulence effectuée dans le cadre du post-doctorat d'Alexandre Degrave (ANR Génoplante AvirLep), dont la caractérisation fonctionnelle est toujours en cours au laboratoire. Récemment, l'apport de données transcriptomiques de type RNAseq a permis de valider l'existence de gènes codant pour des PPS non prédits par le pipeline d'annotation automatique des gènes mais que j'avais réussi à prédire grâce à une approche plus « manuelle » et centrée sur les isochores AT. De façon tout à fait originale, la convergence récente des répertoires d'effecteurs et des approches RNAseq, réalisées dans les phases tardives de l'interaction entre *L. maculans* et le colza, donne accès à de nouveaux effecteurs fongiques, dont les gènes sont généralement localisés en isochores GC et qui semblent spécifiquement exprimés à ces stades tardifs. Cette donnée d'importance fondamentale et (potentiellement) agronomique, donne actuellement lieu à un projet de transcriptomique et de métagénomique, récemment proposé aux appels d'offre « France Génomique » par l'équipe d'accueil, visant à disséquer finement le dialogue moléculaire champignon-plante-micro-organismes associés.

Structure du génome et régulation des gènes codant des effecteurs chez *L. maculans* 'brassicae'.

L. maculans 'brassicae' présente un cycle de vie très complexe au cours duquel il alterne différents modes de vie et différentes stratégies nutritionnelles. Ces changements de comportements traduisent l'existence de mécanismes de régulation complexes de l'expression des gènes, qui permettraient à l'organisme de s'adapter rapidement aux nouvelles conditions auxquelles il doit faire face. Nos travaux ont montré l'importance de la localisation génomique des effecteurs dans leur évolution et leur diversification. De plus, les gènes codant des effecteurs localisés en isochores AT présentent un comportement transcriptionnel différent de celui des ces gènes localisés en isochores GC. En effet, les gènes codant des effecteurs localisés en isochores AT sont peu ou pas exprimés pendant la croissance mycélienne alors que 55 % de ces gènes localisés en isochores GC présentent un support transcriptomique dans les mêmes conditions. Et au contraire, pendant l'infection, 72 % des gènes codant des effecteurs situés en isochores AT sont sur-exprimés à 7 jours après l'infection du colza par rapport à la croissance mycélienne, contre seulement 19 % de ces gènes en isochores GC. De même, les travaux réalisés en collaboration avec l'équipe de Barbara Howlett (Université de Melbourne, Australie) ont montré un lien entre proximité d'ET en 5' des gènes et leur sur-expression *in planta*. Durant ma thèse, un projet parallèle auquel j'ai participé avait pour objectif de caractériser le déterminisme de la co-expression des gènes codant des effecteurs chez Lmb (Annexe 1, Article 4 : Soyer *et al.*, en cours de révision). Puisque les gènes codant des effecteurs chez Lmb sont localisés dans un environnement génomique particulier et qu'ils présentent tous la même cinétique d'expression pendant l'infection, souvent différente de celle des effecteurs localisés en isochore GC, il s'agissait d'évaluer si : (i) la structure des isochores AT permet une régulation des gènes codant des effecteurs faisant intervenir un mécanisme épigénétique reposant sur la structure de la chromatine au niveau de ces isochores ? et/ou si (ii) un, ou plusieurs, régulateur(s) commun(s) sont impliqués dans la régulation de l'expression de ces gènes qui expliquerait leur co-expression pendant l'infection ?

Pour répondre à ces questions, plusieurs stratégies ont été adoptées :

- Tout d'abord, afin de déterminer si la structuration de la chromatine est impliquée dans la régulation de l'expression des gènes codant des effecteurs, l'analyse fonctionnelle de deux acteurs clés impliqués dans le remodelage de la chromatine

a été réalisée. Cette analyse a permis l'écriture d'un article « Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans.* » (Annexe 1, article 4).

- Afin d'identifier le(s) différent(s) régulateur(s) commun(s) qui pourrai(en)t être impliqué(s) dans la régulation des gènes codant des effecteurs, une analyse globale des gènes codant des facteurs de transcription chez *L. maculans* a été réalisée et fait l'objet d'un article « Deciphering the regulome of the causal agent of the stem canker agent of oilseed rape, *Leptosphaeria maculans* 'brassicae' to select putative regulators of pathogenicity. » (Annexe 1, article 5).

Ces travaux ont permis l'émergence de nouveaux questionnements (i) fondamentaux quant à la structure de la chromatine chez *L. maculans* et d'autres champignons phytopathogènes mais aussi (ii) évolutifs au sein du complexe d'espèces *Leptosphaeria*, visant à comprendre les différences de structure de génome observées au sein de ce complexe ainsi que l'influence de cette structure sur la chromatine et sur le comportement des différentes espèces au cours de l'infection du colza. Ces travaux très originaux en ce qui concerne les aspects épigénétiques nécessitent maintenant d'être amplifiés au sein du complexe *L. maculans-L. biglobosa*, qui à ce jour est le modèle le plus avancé sur cette problématique. Ces travaux méritent aussi d'être étendus à d'autres modèles en intégrant des aspect évolutifs. Par exemple, il pourrait être intéressant d'évaluer le lien entre épigénétique et expression pour des modèles dans lesquels les ET sont encore fonctionnels et actifs. Au final, de tels travaux pourraient avoir des applications agronomiques s'ils débouchent sur l'identification de facteurs issus de la plante hôte responsables d'une signalisation induisant un changement d'état de la chromatine dans les génomes fongiques. Une manipulation de ces signaux fournirait des perspectives pour de nouvelles méthodes de lutte générique contre de nombreux filamenteux phytopathogènes.

Architecture des génomes et évolution des champignons phytopathogènes.

Mes travaux ont permis, grâce à une approche originale couplant une comparaison structurelle et fonctionnelle des génomes à une étude précise des ET (annotation, datation et distribution), d'apporter des éléments de réponses (ainsi que de nouvelles questions) sur les mécanismes évolutifs chez les champignons phytopathogènes. Ce type d'approche peut maintenant être appliquée à d'autres modèles fongiques, en tenant compte bien sûr des différents points qui peuvent être améliorés et ainsi permettre une meilleure valorisation de cette approche. Pour ma part, l'expérience tirée de mes recherches sur *L. maculans* me permet de citer trois points importants à prendre en considération avant de se lancer dans ce type d'analyse :

- **Le modèle d'étude** : le complexe d'espèces *L. maculans-L.biglobosa* était un très bon modèle car ses membres étaient très diversifiés en terme de pouvoir pathogène vis-à-vis du colza et montraient des différences structurelles de génome d'après des études préliminaires telles que des électrocaryotypes. Cependant, le nombre de représentants de chaque espèce était sans doute trop faible et ne nous a pas permis de fournir une certaine significativité à nos résultats. Dans notre cas, nous étions limité par le peu de ressources biologiques disponibles concernant *L. maculans* 'lepidii'. C'est pourquoi je préconise l'utilisation de complexes d'espèces dont la diversité est bien connue, mais aussi bien échantillonnée, pour réaliser ce type d'étude.
- **L'assemblage des génomes** : c'est une partie que je n'ai suivi que passivement au cours de mes recherches, n'ayant jamais pris part à l'assemblage d'un génome. Cependant, je me suis vite rendu compte que les logiciels d'assemblage rencontraient des difficultés quand il s'agissait d'assembler des *reads* issus de séquences répétées, ce que l'on a facilement pu observer quand nous avons reséquencé deux nouveaux isolats de *L. maculans* 'brassicae' (Chapitre 2). Le fait d'avoir des régions répétées suffisamment bien assemblées est primordial si l'on veut identifier correctement les éléments qui les composent. Cela permettrait aussi de réduire le nombre d'éléments répétés non catégorisés. Il faut donc optimiser cette étape d'assemblage pour obtenir une meilleure représentation des régions répétées dans les génomes.
- **L'étude des différentes sources évolutives** : dans le complexe d'espèces *L. maculans-L. biglobosa*, nous avons montré que les ET et le RIP jouaient un rôle

important dans l'évolution des génomes du point de vue structural et fonctionnel, sans toutefois être capable de conclure quant à leur incidence dans la spéciation, indépendamment des pressions de sélection (cause ou conséquence ?) De plus, nous n'avons aucune information concernant l'origine et la dynamique des mutations identifiées dans les régions non répétitives des génomes de chaque espèce. Il serait très intéressant de connaître leur rôle dans la spéciation des membres d'un complexe et d'être capable de discerner les événements drastiques (remaniements chromosomiques) des mutations adaptatives dans la spéciation. C'est pourquoi je pense que l'approche que j'ai utilisée au cours de cette thèse doit être complétée par une analyse de génomique évolutive basée sur la composition des séquences et l'identification de signatures de pressions de sélection.

Pour mon post-doctorat, j'ai décidé d'utiliser un autre modèle fongique permettant une meilleure application de l'ensemble de ces approches. Ce projet se déroulera sur une période de deux ans et sera effectué dans l'équipe d'Eva Stukenbrock au Max Planck Institute de Marburg (Allemagne), dont la thématique de recherche est l'évolution et les processus de spéciation chez les membres du complexe d'espèces *Zymoseptoria*, qui sont des agents pathogènes d'espèces de graminées cultivées (p. ex. blé et orge) et sauvages. Mon projet s'intègre dans cette thématique puisqu'il consiste à analyser la relation entre évolution et l'architecture des génomes chez *Zymoseptoria* spp. par une approche de génomique comparative des populations. Pour cela je dispose d'un ensemble de 52 génomes représentant quatre espèces diversifiées de *Zymoseptoria* infectant des espèces cultivées et des espèces sauvages. Ce sera la première fois qu'un aussi grand nombre de génomes séquencés appartenant à un complexe d'espèces est utilisé pour des études évolutives chez les champignons phytopathogènes. Cela permettra d'observer les différents mécanismes responsables de l'évolution des génomes au sein de ces espèces dans une période de temps postulée comme beaucoup plus courte que lors de l'étude chez *Leptosphaeria* spp. De plus, grâce à leur diversité, nous pourrons étudier les impacts évolutifs apportés par l'environnement puisque certaines espèces sont issues d'un agro-écosystème contrôlé alors que d'autres proviennent d'un écosystème naturel.

Pour le moment, les 52 génomes séquencés par *NGS* ont été assemblés avec SOAPdenovo (Li *et al.*, 2010), mais cet assembleur n'a pas été optimisé pour l'assemblage des répétitions. La première étape de mon projet consiste donc à améliorer l'assemblage *de novo* des génomes des quatre espèces dans le but d'obtenir une

meilleure couverture des régions répétées dans les génomes. Pour cela une étape de filtration des *reads* sera nécessaire afin d'exclure des artefacts de séquençage. Puis, différents assembleurs seront configurés et comparés afin de produire le meilleur assemblage.

En se basant sur les nouveaux assemblages, la deuxième étape du projet consiste à identifier et à annoter les éléments répétés présents dans les génomes, dans le but d'étudier leur implication dans l'architecture des génomes des membres du complexe d'espèces *Zymoseptoria*. Chez *Z. tritici*, on sait peu de chose sur la nature des éléments répétés présents dans les chromosomes. Je me servirai donc de mon expérience dans ce domaine afin d'identifier de quels types d'éléments il s'agit et dans quelle proportion ils sont présents dans les différents génomes. Puis l'étude de leur distribution dans les 52 membres séquencés du complexe d'espèces *Zymoseptoria* devrait permettre de définir leur rôle dans l'adaptation et la spéciation. Le grand nombre de séquences par espèces devrait permettre d'améliorer la robustesse et la significativité des résultats en ce qui concerne le rôle des ET dans l'évolution des génomes de ce complexe d'espèces.

La troisième étape consiste à comparer la structure des différents génomes en effectuant des alignements inter- et intra-espèces qui permettront de mettre en évidence (i) des régions très conservées et synténiques sûrement très importante pour la vie de l'organisme, (ii) des régions ayant subi des réorganisations structurales telles des inversions ou des translocations et (iii) des régions uniques ou spécifiques susceptibles de jouer un rôle potentiel dans le pouvoir adaptatif de ces organismes. Le petit nombre de membres séquencés dans le complexe d'espèces *L. maculans-L. biglobosa* permettait d'effectuer des alignements de génomes par paire, ce qui dans le cas de ce projet ne sera pas possible au vu du grand nombre de séquences et des temps de calcul et d'analyses que cela engendrerait. Je vais donc devoir utiliser de nouveaux outils comme MultiZ (Blanchette *et al.*, 2004) permettant de réaliser des alignements multiples de génomes.

Les alignements multiples générés vont permettre de comparer la composition des séquences des génomes au niveau nucléotidique afin d'identifier des régions soumises à pression de sélection. Nous nous intéresserons particulièrement à deux mécanismes évolutifs, le RIP et le gBGC (*GC-biased gene conversion*), tous les deux responsables de l'évolution du contenu en GC des génomes (Chapitre 1 ; Duret & Galtier, 2009). Il a été montré chez plusieurs organismes, et notamment chez *L. maculans*, que le mécanisme du RIP pouvaient continuer son action mutagène sur des séquences non dupliquées se trouvant à proximité de séquences répétées, créant ainsi une diversification accélérée

dans ces régions et leur conférant ainsi des avantages évolutifs (Fudal *et al.*, 2009 ; Daverdin *et al.*, 2012). Chez *L. maculans* 'brassicae', nous avons vu que l'action du RIP sur les régions riches en ET était responsable de la structure en isochores du génome, ce qui est aussi postulé chez diverses espèces relativement proches de *Z. tritici*, telles que *Cladosporium fulvum* et *Pseudocercospora fijiensis*. Une telle structure en isochores avait déjà été observée dans les génomes de mammifères, mais elle était le résultat d'un autre mécanisme, le gBGC. Contrairement au RIP qui induit une diminution du contenu GC, le gBGC est responsable d'une augmentation de celui-ci. C'est un processus associé à la recombinaison qui a pour conséquence une fixation préférentielle d'allèles GC sur des allèles AT et qui est retrouvé dans plusieurs organismes eucaryotes y compris des champignons (Pessia *et al.*, 2012). Ce mécanisme qui affecte l'évolution des séquences codantes peut être interprété à tort comme une marque de sélection positive (Ratnakumar *et al.*, 2010). Une fois les signatures de ces deux mécanismes identifiés, des indicateurs de pression de sélection seront calculés le long des séquences alignées en se basant sur le rapport de mutations non-synonymes et synonymes (dN/dS). Ces analyses vont permettre d'identifier des régions soumises à différentes pressions de sélection et qui peuvent contenir des éléments, tels des gènes ou des séquences régulatrices, jouant un rôle important dans l'adaptation de l'organisme à son environnement.

Ce projet est la continuation directe de ma thèse et permettra d'appuyer certaines hypothèses évolutives formulées au cours de mes travaux de recherches sur *L. maculans* ou dans le cas contraire, d'en apporter des nouvelles. Tout mon travail s'inscrit dans une suite logique d'études liée à l'évolution de la génomique fongique au cours de ces dernières années et permet, je l'espère, d'apporter certains éléments de réponse aux différentes questions posées par l'évolution des champignons.

ANNEXES

ANNEXE 1 : ARTICLES

Article 1. FONZIE : An optimized pipeline for minisatellite marker discovery and primer design for large sequence data sets.

P Bally, J Grandaubert,T Rouxel & MH Balesdent.

Publié le 29 novembre 2010 dans *BMC Research Notes* **3:**322.

Bally et al. BMC Research Notes 2010, 3:322
http://www.biomedcentral.com/1756-0500/3/322

BMC Research Notes

RESEARCH ARTICLE Open Access

FONZIE: An optimized pipeline for minisatellite marker discovery and primer design from large sequence data sets

Pascal Bally, Jonathan Grandaubert, Thierry Rouxel, Marie-Hélène Balesdent*

Abstract

Background: Micro-and minisatellites are among the most powerful genetic markers known to date. They have been used as tools for a large number of applications ranging from gene mapping to phylogenetic studies and isolate typing. However, identifying micro-and minisatellite markers on large sequence data sets is often a laborious process.

Results: FONZIE was designed to successively 1) perform a search for markers via the external software Tandem Repeat Finder, 2) exclude user-defined specific genomic regions, 3) screen for the size and the percent matches of each relevant marker found by Tandem Repeat Finder, 4) evaluate marker specificity (i.e., occurrence of the marker as a single copy in the genome) using BLAST2.0, 5) design minisatellite primer pairs via the external software Primer3, and 6) check the specificity of each final PCR product by BLAST. A final file returns to users all the results required to amplify markers. A biological validation of the approach was performed using the whole genome sequence of the phytopathogenic fungus *Leptosphaeria maculans*, showing that more than 90% of the minisatellite primer pairs generated by the pipeline amplified a PCR product, 44.8% of which showed agarose-gel resolvable polymorphism between isolates. Segregation analyses confirmed that the polymorphic minisatellites corresponded to single-locus markers.

Conclusion: FONZIE is a stand-alone and user-friendly application developed to minimize tedious manual operations, reduce errors, and speed up the search for efficient minisatellite and microsatellite markers departing from whole-genome sequence data. This pipeline facilitates the integration of data and provides a set of specific primer sequences for PCR amplification of single-locus markers. FONZIE is freely downloadable at: http://www.versailles-grignon.inra.fr/bioger/equipes/leptosphaeria_maculans/outils_d_analyses/fonzie

Background

Satellite sequences are abundantly interspersed in the genome of almost all eukaryotic species studied [1] and have been analyzed extensively in animals [2-4], plants [5], and more recently in fungi [6].

Two major classes of satellites are usually defined: minisatellites (MS) and microsatellites (μS). μS are repetitive sequences of mostly 2 to 4 nucleotides with a widespread occurrence in multicellular organisms, whereas minisatellites are usually defined as the repetition in tandem of a short, 6 to 100 bp motif, spanning 100 bp to several kilobases [7].

* Correspondence: mhb@versailles.inra.fr
Institut National de la Recherche Agronomique, UMR 1290 BIOGER, BP 01, Avenue Lucien Brétignières, 78850 Thiverval-Grignon, France

μS analysis requires sophisticated separation and visualization apparatus, due to the small size of the individual repeat units. In contrast, and because of the larger size of their individual core motif, MS can be separated and visualized by conventional agarose gel electrophoresis [6]. Using a panel of single-locus minisatellite markers, isolate-specific DNA fingerprint can be produced in a PCR-based assay [8,9]. MS also provided highly polymorphic markers for linkage studies [10] making them informative genetic markers.

Genome mapping is still the primary tool for genome knowledge in a series of model organisms. Due to their small genome size and laboratory tractability, fungal models such as the ascomycete yeast *Saccharomyces cerevisiae*, and the filamentous ascomycete *Neurospora*

BioMed Central

© 2010 Balesdent et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

crassa, have been pioneers in the process of building genetic maps [11]. In this respect, the Dothideomycete *Leptosphaeria maculans* causing stem canker of oilseed rape (*Brassica napus*), is amenable to genetics in the lab and is thus an adequate and complete model for genome-wide-based functional studies of pathogenicity, and for identifying signalling and regulation processes responsible for shifts in lifestyle [12]. Although preliminary genetic maps have been developed for *L. maculans* [13-15], it is difficult to integrate these maps as the markers (RAPD and AFLP) used to generate them are not readily transferable [16]. Based on the analysis of a set of around 800 BAC-end sequence data, MS were found to be powerful for genetic mapping or population genetic studies in *L. maculans* [6]. The availability of the *L. maculans* genome now enables the generation of a large quantity of such markers, which could allow the saturation of the genetic map and improve population genetic studies and pathogenicity gene discovery [12].

Despite the availability of whole genome sequences, *in silico* MS identification and primer design is still a laborious process, requiring the use of different softwares and numerous steps of marker validation, which take time and often generate errors. Most of the currently available software or pipeline solutions are focused on the identification of μS [17-19]. Moreover, they are usually platform dependent [20], or may need several tedious pre-processing steps [19,21]. In order to avoid these constrains and to reduce time, FONZIE has been developed to automate and facilitate the design of PCR primers from large sets of sequences, that will amplify single locus MS and their flanking sequences, with the possibility of excluding some specific regions of the genome from the MS design process. This pipeline integrates various external tools such as BLAST [22], Tandem Repeat Finder (TRF) [23], and Primer3 [24]. FONZIE is able to successively perform TRF on a Fasta or Multifasta formatted sequence file, search for markers specifically located in user-defined sequence subset, as exemplified here with GC-equilibrated isochores of *L. maculans*, eliminate markers which do not satisfy user-defined criteria, eliminate markers which are not single-copy, and finally design minisatellite primer pairs and check the specificity of the PCR product. The primary targeted public of FONZIE is biologists unfamiliar with complex bioinformatics solutions. Therefore it does not require a computer science background and can run on personal computers, yielding results within an hour or less for most queries.

Implementation

A comprehensive "readme" page includes details on how to configure and run the pipeline, and how to search and display FONZIE results.

Before launching FONZIE, users must have installed the Python programming language on the computer. In a second step, users must create a database of sequences (ideally, the whole-genome sequence of the investigated organism) in order to perform BLAST during the execution of the pipeline. In our project, the database used was the pilot genome of *L. maculans* (45.12 Mb, 76 Super-Contigs, Table 1) as a Multifasta file.

Pipeline Components

FONZIE consists of three major components: a set of pipelined programs, a BLAST database and a graphical user interface (Figure 1). The FONZIE pipeline consists of Python modules allowing the execution of different external software programs.

The BLAST database was created with FORMATDB, a BLAST [22] database-related tool. The graphical user interface was implemented with a Python graphic standard library named Tkinter. FONZIE was developed under both Linux Ubuntu 7.10 and Windows XP operating systems.

The automated process consists of six steps (Figure 2):

Tandem Repeat Finder

FONZIE accepts nucleotidic sequence files in Fasta or Multifasta format and uses the Tandem Repeat Finder software [23] in order to find tandemly repeated elements (minisatellites and microsatellites) (Figure 2 step a). TRF parameters can be modified via the graphical user interface (Figure 1).

Exclusion of markers in specific regions

Isochores have been described for many eukaryotic organisms such as plants and mammals [25,26]. Isochores are very long stretches of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. In the *L. maculans* genome this structure is particularly clear, with alternation of GC-equilibrated coding regions with no or few transposable elements (TE) and AT-rich regions mainly composed of degenerated and truncated TE and encompassing very few ORFs [13,27]. Due to these particularities, most micro-or minisatellites identified in *L. maculans* AT isochores are not single locus markers and have to be excluded from the search for single locus markers. This can be easily extended to all repeat-rich genomes such as that of the Oomycete *Phytophthora infestans* containing 74% of repeats [28].

In order to exclude such specific genomic regions from the search for markers, FONZIE compares the location of the tandem repeats with that of the AT-rich isochores (or whichever other user-defined sequences). Tandem repeats fully or partially overlapping the AT-rich isochores (or other user-defined sequences) are thus automatically excluded from the analysis (Figure 2 step b). The AT-rich isochores or other user-defined sequences to be excluded must be specified in a GFF

Table 1 FONZIE results when performed on different fungal or oomycetes whole genomes.

Organism	Genome Size (Mb)	Nb of contigs or super contigs[a]	Execution Time[b]	Nb of markers identified[c]			Nb of amplification products (AP) and primers designed[d]				
				Total	Single copy	Multiple copies	Single copy	Multiple copies	No primers	No BLAST results	% single-copy AP
Aspergillus niger	34.85	24	2 min 53 sec	637	533	104	600	18	17	2	94.19
Stagonospora nodorum	37.21	108	4 min 41 sec	1288	978	310	959	154	169	6	74.46
Pyrenophora tritici repentis	37.84	47	5 min 12 sec	1141	878	263	935	195	8	3	81.94
Botrytis cinerea	42.66	588	12 min 03 sec	3648	2719	929	3339	118	176	15	91.53
Leptosphaeria maculans	45.12	76	23 min 36 sec	2606	1799	807	2405	146	49	6	92.29
Laccaria bicolor	64.88	665	45 min 39 sec	5393	1516	3877	1640	3409	338	6	30.41
Phytophthora infestans	228.54	4921	3 h 21 min 37 sec	7514	1855	5659	1718	5239	553	4	22.86

[a] Number of sequences in the Multifasta file

[b] Machine used for this test: Laptop Intel Core 2 Duo, 2.4 GHz and 3Go RAM

[c] FONZIE results after step d of the workflow (Figure 1), using the TRF default parameters (match = 2, indel = 7, mismatch = 7, pi = 10, pm = 80, minscore = 50, maxperiod = 500) and screen parameters for core motif size >3, % identity between motifs = 90%, BLAST cut-off value = 1e-10)

[d] Final FONZIE results after steps e (Primer pair design) and f (checking for the specificity of the amplification product) of the workflow shown in Figure 1, using a BLAST cut-off value of 1e-40

format file. If no isochore regions, or any other regions to be excluded from the search for MS, are specified, FONZIE analyses all the tandem repeats found by TRF.

Length and percent match screening

Markers are then screened for the size of the core motif (more than 6 bp by default), and the percent matches between core motifs (i.e., the percentage of matched bases between tandem repeats of the core motif within the MS, more than 90% by default). A minimum of 90% identity between core motifs was selected here because previous experiments showed that it was a characteristic of minisatellites showing actual sequence length polymorphism in *L. maculans* [[6]; unpublished data]. The output of this step is a Fasta formatted sequence file for each selected marker (Figure 2 step c).

Screening for specificity of the markers

This step verifies the specificity of the selected MS (Figure 2 step d). Each MS is analysed for sequence similarity using BLAST [22] against a BLAST database, created by users, with a default E-value cut-off set at 1e-10. This value can be modified via the graphical user interface. Typically, the specificity is evaluated against the whole genome sequence if available.

Using the output of the BLAST, FONZIE classifies MS into 4 categories, defined by both the number of sequences in the BLAST database matching the query sequence (i.e. with BLAST E-value < 1e-10) and the E-value obtained for the query sequence compared to that obtained for other matching sequences, as follows:

1) MS for which the BLAST results exhibit a unique match in the database, corresponding to the query sequence, hereafter referred to as "UNIQUE COPY".

2) MS for which the BLAST results exhibit several matches on several sequences, with a best hit on the query sequence, hereafter referred to as "MULTIPLE COPIES".

3) MS for which the BLAST results exhibit several matches on several sequences, with a best hit on other sequence than the query sequence, hereafter referred to as "MULTIPLE COPIES (BEST HIT ON SEQUENCE X)". X represents the sequence id. with the best alignment score. The repetition of the motifs among the whole genome can explain such results.

4) (Rare) MS for which the BLAST results exhibit no match on any sequences with the default cut-off of E-value at 1e-10, hereafter referred to as "NO BLAST RESULTS" (0.2% in the case of *L. maculans*, Table 1). The high default threshold parameter of the E-value (1e-10), the low complexity of the motifs, or the small size of the MS sequence can explain such results.

"MULTIPLE COPIES (BEST HIT ON SEQUENCE X)" and "NO BLAST RESULTS" are automatically excluded for the next steps of the analysis.

Primer design

In order to design primer pairs flanking the MS sequences, FONZIE extracts 100 bps by default on each side of the query sequence whenever possible (or whichever other user-defined length, modified via the graphical user interface, Figure 1). Otherwise, FONZIE

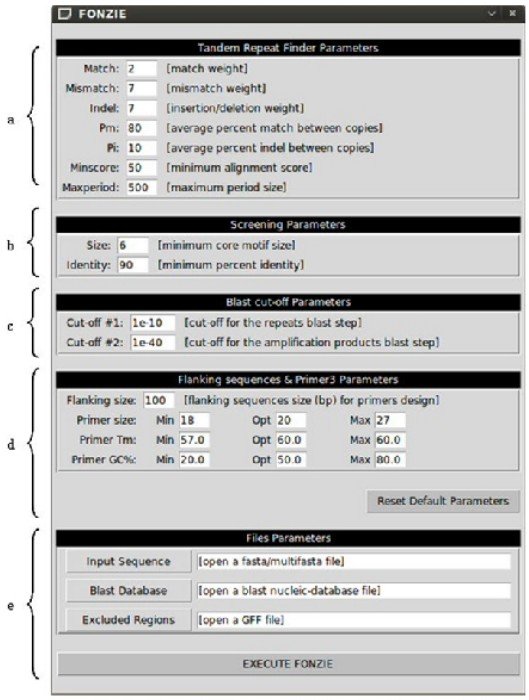


Figure 1 The graphical user interface of FONZIE. The graphical user interface is composed of five major sections. **a**. TRF parameters: All the modifiable parameters of TRF are located in this section. Users can modify each of them separately. **b**. Screening parameters. Users can modify the core motif size and the percent match between core motifs **c**. BLAST cut-off parameters: Users can modify the two E-value cut-off used during the screening for marker specificity and the virtual PCR steps **d**. Flanking sequences size and Primer3 parameters. **e**. The file field: In this section, users can open the Fasta- or Multifasta formatted sequence file, the BLAST database file, and the optional gff format file which contains the excluded regions.

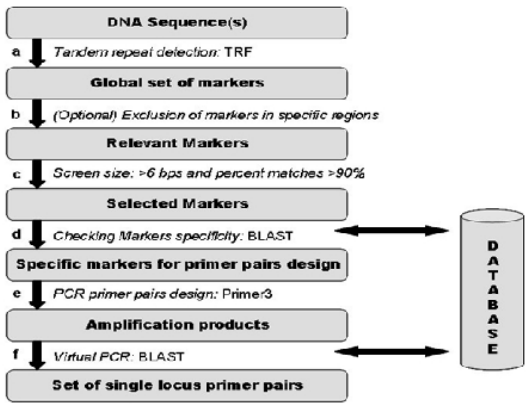


Figure 2 The FONZIE workflow: The FONZIE workflow consists of 6 steps. a Execution of TRF on a Fasta sequence or a Multifasta file, b Optional: exclusion of some user-defined specific regions, c Screening by core motif size and percent match parameters, d Screening for marker specificity with BLAST (1e-10 by default) against a database, e Primer design step using PRIMER3, f Virtual PCR by making a BLAST (1e-40 by default) against the database to check the specificity of the PCR product (single copy locus).

extracts the maximum number of nucleotides available upstream and downstream of the MS sequence.

FONZIE then uses Primer3 [24] which analyses the flanking DNA sequences in order to define suitable forward and reverse PCR primers, designed with the standard set of constraints of Primer3 (Figure 2 step e): optimal primer size at 20 bases pairs, optimal Tm at 60° C, and optimal primer GC content at 50%. These parameters can also be modified by users via the graphical user interface (Figure 1). FONZIE then calculates the final product size range as the size of the MS (= the size of the core motif × the number of repeats as defined by TRF) + 2×, or 10×, the minimal primer size (chosen by users), for the minimal and the maximal size, respectively. The first primer pair returned by Primer3 is displayed by FONZIE. If no primer pairs are found with Primer3, the result is displayed as "NO DESIGNED PRIMER" in the final text file.

Virtual PCR

The virtual PCR step validates the specificity of amplification products identified in step e (Figure 2 step f). Each amplification product is analysed for sequence similarity using BLAST against all the sequences of the BLAST database. The E-value cut-off is fixed at 1e-40 by default and may be modified by the users. The same type of analysis is implemented in the SAT software [19], but SAT only verifies the specificity of the PCR primers, whereas FONZIE analyses the specificity of the whole amplification product.

Using the outputs of the virtual PCR step, FONZIE classifies sequences in 4 categories according to the specificity of amplification products. These categories are exactly the same as for the markers' specificity screen.

Running the pipeline

Users have to provide an input sequence (Fasta or Multifasta format), and a GFF format file if they want to exclude specific regions from the analysis.

Setting parameters and pipeline execution

TRF default options are as follows: alignment parameters of 2,7,7, which correspond to the weight attributed to each match, mismatch and indel, respectively, between core motifs of the MS; minimum alignment score of 50; maximum period size of 500, pM (i.e., the average percent identity between the copies of a pattern) of 80; pI (i.e., the average percent of insertions and

deletions between the copies of a pattern) of 10 (Figure 1) [23]. Users can modify through the graphical interface all TRF parameters, but also the screening parameters (i.e., the minimum identity between core motif fixed by default at 90%, and the size of the motif, >6 bps by default) as well as the two BLAST cut-off parameters (E-values cut-off of 1e-10 and 1e-40 by default) (Figure 1).

Visualization of results

At the end of the process, users have access to all the files and directories created at each step of the pipeline. A final tab-delimited text file summarizes, for all numbered satellites found by FONZIE, the name of each satellite automatically attributed by the pipeline, the core motif of each satellite, the number of repetitions of the core motif, the BLAST results of specificity of the satellite and of the amplification product (i.e. if the marker and the amplification product are found as a unique copy or not), the sequences of the forward and reverse primers and the start and the end location of the amplification product on the input sequence (Table 2).

Sequence files and directory

FONZIE allows traceability of each step of the pipeline by generating a directory and associated files. FONZIE displays a DAT file created by TRF which contains the information of all tandem repeats detected on the input sequence. A text file shows the different consensus sequences of each marker, its position on the input sequence as well as period size, copy number, consensus size, percent matches, percent indels, the score and the status of each marker.

A directory named "MARKERS_SEQUENCES" is created at the end of the first screening step which contains all the markers found by FONZIE, in Fasta format. Another directory named "AMPLIF_PRODUCTS_SEQUENCES", contains the Fasta files for the sequence of each amplification product. A last directory named "MARKERS_RECAP" contains, for each MS, a text file summarizing the marker sequence, the marker sequence with flanking regions, the amplification product sequence, the amplification product status, the positions of the amplification product on the query sequence, and the output of Primer3 indicating the characteristics of the primers and their location in relation to the MS (Figure 3).

Results and discussion

Performance

FONZIE was firstly assessed on the whole *L. maculans* genome. The 45 Mbs were analysed in 23 minutes (Table 1). FONZIE found 2606 minisatellites, with 2405 amplification products occurring as unique copies, 146 in multiple copies. Six amplification products had no BLAST results, and 49 markers were identified for which no primers could be designed. 92% of the markers identified by FONZIE, following exclusion of repeat-rich AT-rich isochores thus corresponded to putatively unique sequences in the genome and could be further used for analysis of polymorphism in isolates and genetic mapping and/or population genetics studies.

In a second step, FONZIE was used to mine 6 other fungal and Oomycete genomes. These genomes, sized from 34 Mb to 228 Mb (Table 1), range from compact, repeat poor genomes (*Aspergillus niger*, *Stagonospora nodorum*, *Pyrenophora tritici repentis* and *Botrytis cinerea*) to repeat rich genomes of the basidiomycete *Laccaria bicolor* [29] and the Oomycete *P. infestans* [28]. MS were found using the same parameters as for *L. maculans*, but without defining user-specified genomic regions to be excluded from the analysis. In all cases, numerous MS could be identified and extracted in a very limited time interval, i.e. between 2 minutes to cover the *A. niger* whole genome to 3 h21 min to scan the *P. infestans* genome sized 228 Mb over 4921 contigs (Table 1). This analysis suggests some genomes are extremely rich in MS (e.g., the *B. cinerea* genome; Table 1). It mainly indicates that an average of 70-83% single-copy markers are identified whenever the genomes investigated are poor in repeats or when repeat-rich regions are excluded from the analysis, as done here *for L. maculans* (Table 1). In contrast, only 23-30% of markers were found to be single-copy markers in the genomes of the repeat-rich fungi and Oomycete *L. bicolor* and *P. infestans* if analysing the whole genomes with FONZIE, without user-defined exclusion of repeat-rich sequences.

Biological validation

For validation of the proposed approach, primer pairs for 517 putative single-copy MS markers (i.e. with core motif ≥6), scattered along 17 *L. maculans* supercontigs (with AT-rich isochores as excluded regions), were PCR-assayed on a set of 9 *L. maculans* natural isolates. More than 90% of the primer pairs (475) amplified a PCR product, with size compliant with the expected MS size. This yield is very similar to that of the currently available tools such as STAMP, yielding 83.33 to 91.67% of successful amplifications, depending on the species [30]. Among these PCR-amplified sequences, 44.8% (213 primer pairs) showed resolvable size polymorphism between at least 2 of the 9 tested *L. maculans* isolates. Finally, 24.8% of the 517 MS (i.e. more than 55% of the polymorphic MS markers) were found to be polymorphic between isolates 'a.2' and 'H5', the two parental isolates of the *L. maculans* genetic map [13]. Progeny analysis confirmed that all these markers segregated as single loci, i.e. with a 50:50 ratio, as expected for this haploid organism.

Table 2 Example of the FONZIE final result table, run on Supercontig 16 of the *Leptosphaeria maculans* genome.

MARKER_ID	MOTIF[a]	REPETITION[b]	MARKER_STATUS_ON_GENOMELEPTOV2[c]	LEFT_PRIMER[d]	RIGHT_PRIMER[e]	START_AMPLIF_PRODUCT_ON_SUPERCONTIG_16[f]	END_AMPLIF_PRODUCT_ON_SUPERCONTIG_16[g]	STATUS_AMPLIF_PRODUCT_ON_GENOMELEPTOV2[h]
min_supercontig_16_10	ATAAAAGTAAA CTACTACTTTA	2.0	MULTIPLE_COPIES	GCATAAAGCTAAT CTTCTCTACCCC	GTATAAACTGCCC TTGTGTATACCT	100841	101019	MULTIPLE_COPIES
min_supercontig_16_11	GGATCATCAAGGA	17.3	UNIQUE_COPY	CGTTTTGGCTTT GTTGTTGA	ACTATGAGCCAG GTGAACCG	111896	112241	UNIQUE_COPY
min_supercontig_16_12	CGCTCTCTCTCTC TCTCTTTCTCTCT	4.3	MULTIPLE_COPIES	CGCCAACAAGA CTACCCATC	GAAGCGGTGG CAGTTTTTAG	112524	112812	UNIQUE_COPY
min_supercontig_16_13	CCATGT	5.8	UNIQUE_COPY	ACCTCCCGAGG AAAAGTGAC	CTTGTGTGGT CTGGTTGCAG	134392	134595	UNIQUE_COPY
min_supercontig_16_14	GAGAGAGAG AGAGAGAGA	7.4	MULTIPLE_COPIES	TGACTCGGCGTC TACCCTAC	AGCCAGCCA GCCAGTACTAA	136186	136390	UNIQUE_COPY
min_supercontig_16_15	AAGCAGAAGGC TATTGAGTCGCCAGA GACAAGTCCACAGTCC	2.1	UNIQUE_COPY	AAGTGGCTGGAC CTAGCAGA	ACATCGGCGA CACGTTTAGT	142179	142347	UNIQUE_COPY
min_supercontig_16_16	GTGTGG	11.2	MULTIPLE_COPIES	TGTGGATGATAG GATGGGGT	GTGACAAGCA CATGATTCGC	156524	156707	UNIQUE_COPY

Only a few markers generated by FONZIE are displayed in the table.

[a] consensus sequence of the core motif of the minisatellite (MS)

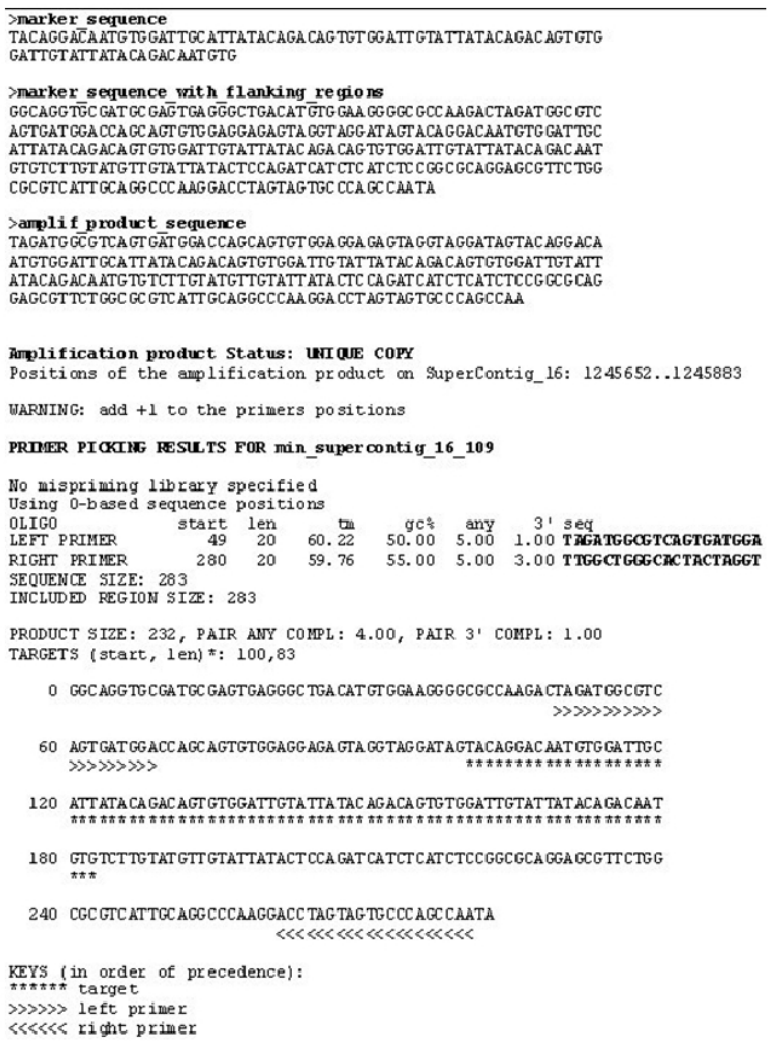
[b] number of repeats of the core motif

[c] results of the first BLAST step on the BLAST database: UNIQUE_COPY, the unique sequence matching the MS (query) is the query sequence; MULTIPLE_COPIES, more than one sequence of the BLAST database match the query sequence and the best hit is obtained for the query sequence (E-value cut-off e-10)

[d] and [e], sequences of the left and right primers, respectively, generated by Primer3 to amplify the minisatellite locus and flanking sequences

[f] and [g], location (in base pairs) of the primers along Super-Contig 16 sequence.

[h], results of the second BLAST step, where the amplification product is blasted on the BLAST database: UNIQUE_COPY, the unique sequence matching the PCR product (query) is the query sequence; MULTIPLE_COPIES, more than one sequence of the BLAST database match the query sequence and the best hit is obtained for the query sequence (E value cut-off e-40).



```
>marker_sequence
TACAGGACAATGTGGATTGCATTATACAGACAGTGTGGATTGTATTATACAGACAGTGTG
GATTGTATTATACAGACAATGTG

>marker_sequence_with_flanking_regions
GGCAGGTGCGATGCGAGTGAGGGCTGACATGTGGAAGGGGCGCCAAGACTAGATGGCGTC
AGTGATGGACCAGCAGTGTGGAGGAGAGTAGGTAGGATAGTACAGGACAATGTGGATTGC
ATTATACAGACAGTGTGGATTGTATTATACAGACAGTGTGGATTGTATTATACAGACAAT
GTGTCTTGTATGTTGTATTATACTCCAGATCATCTCATCTCCGGCGCAGGAGCGTTCTGG
CGCGTCATTGCAGGCCCAAGGACCTAGTAGTGCCCAGCCAATA

>amplif_product_sequence
TAGATGGCGTCAGTGATGGACCAGCAGTGTGGAGGAGAGTAGGTAGGATAGTACAGGACA
ATGTGGATTGCATTATACAGACAGTGTGGATTGTATTATACAGACAGTGTGGATTGTATT
ATACAGACAATGTGTCTTGTATGTTGTATTATACTCCAGATCATCTCATCTCCGGCGCAG
GAGCGTTCTGGCGCGTCATTGCAGGCCCAAGGACCTAGTAGTGCCCAGCCAA


Amplification product Status: UNIQUE COPY
Positions of the amplification product on SuperContig_16: 1245652..1245883

WARNING: add +1 to the primers positions

PRIMER PICKING RESULTS FOR min_supercontig_16_109

No mispriming library specified
Using 0-based sequence positions
OLIGO            start  len      tm      gc%   any    3' seq
LEFT PRIMER         49   20   60.22    50.00  5.00  1.00 TAGATGGCGTCAGTGATGGA
RIGHT PRIMER       280   20   59.76    55.00  5.00  3.00 TTGGCTGGGCACTACTAGGT
SEQUENCE SIZE: 283
INCLUDED REGION SIZE: 283

PRODUCT SIZE: 232, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00
TARGETS (start, len)*: 100,83

    0 GGCAGGTGCGATGCGAGTGAGGGCTGACATGTGGAAGGGGCGCCAAGACTAGATGGCGTC
                                                    >>>>>>>>>>>

   60 AGTGATGGACCAGCAGTGTGGAGGAGAGTAGGTAGGATAGTACAGGACAATGTGGATTGC
      >>>>>>>>>                                  **********

  120 ATTATACAGACAGTGTGGATTGTATTATACAGACAGTGTGGATTGTATTATACAGACAAT
      ************************************************************

  180 GTGTCTTGTATGTTGTATTATACTCCAGATCATCTCATCTCCGGCGCAGGAGCGTTCTGG
      ***

  240 CGCGTCATTGCAGGCCCAAGGACCTAGTAGTGCCCAGCCAATA
                          <<<<<<<<<<<<<<<<<<<

KEYS (in order of precedence):
****** target
>>>>>> left primer
<<<<<< right primer
```

Figure 3 Example of one file from "MARKERS_RECAP" directory generated by FONZIE. From top to bottom: the marker sequence, the marker sequence with flanking regions, the amplification product sequence, and the primer3 output

In previous experiments, MS markers were searched for in *L. maculans* by using the TRF application alone [6], with a yield of only 13% of polymorphic minisatellites out of 30 primer pairs tested, even though a wider range of isolates (21 isolates instead of 9 in the present study) was used for the polymorphism screen. Only 10% of the primer pairs from this preliminary study were found to be polymorphic between isolates 'a.2' and 'H5', compared to 24.8% in the present work. These results not only validate the efficiency of the pipeline but also confirm the potential of MS as genetic markers in *L. maculans*, since it can be extrapolated both from Table 1 and these results that more than 530 MS could be mapped to the reference *L. maculans* map.

Conclusion

FONZIE is a user-friendly application, optimised for MS marker identification and specific PCR primer design. It brings together external software like Tandem Repeat Finder, Primer3 and BLAST, to perform them in one run for PCR primer pair design. FONZIE is time-saving, reduces errors that might be introduced by analysing sequence by hand and is easy to use. Users only need to upload their sequences, a reference database, an optional file containing excluded genomic regions, and select appropriate

parameters. Moreover, FONZIE can cover whole genomes in one run. This tool is generic enough to be used on the DNA of any organism and is particularly well adapted to whole genome sequence from which specific regions (AT-isochores for examples) need to be excluded.

The main innovation of the pipeline is to integrate two successive steps of BLAST to check the specificity and uniqueness of the MS loci identified by TRF. The usefulness of the pipeline has been demonstrated here for the fungus *L. maculans*, since 100% of the polymorphic markers identified were single locus, and the yield of identification of polymorphic markers has been more than doubled compared to manual search for MS without BLAST steps.

Availability and requirements

- Project name: FONZIE
- Operating system: UNIX/LINUX, Windows XP
- Programming language: Python
- Licence: GNU GPL, free for academic and non-academic users
- Any restrictions to use by non-academic: none

Acknowledgements

This work and PB were funded by the French agency ANR ("Agence Nationale de la Recherche") contract ANR-07-GPLA-015 ("AVirLep") under the framework of the Génoplante 2010 programme. JG was funded by the ANR project "Fungeffector" (ANR-06-BLAN-0399) and by the INRA SPE (Santé des Plantes et Environnement) department. The authors thank Siân Deller (INRA BIOGER) for providing useful comments on the manuscript.

Authors' contributions

PB and JG designed the application, JG programmed the application and supervised the implementation of the computational part of this work, PB and JG tested and evaluated the application, TR contributed to validation on fungal genomes and edited the manuscript, PB and MHB performed the biological validation of the markers on *L. maculans*, PB wrote the manuscript, MHB initiated and coordinated the work and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 1 March 2010 Accepted: 29 November 2010
Published: 29 November 2010

References

1. Debrauwere H, Gendrel CG, Lechat S, Dutreix M: **Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites.** *Biochimie* 1997, **79**:577-586.
2. Dharma Prasad M, Muthulakshmi M, Madhu M, Archak Sunil, Mita K, Nagaraju J: **Survey and analysis of microsatellites in the silkworm, *Bombyx mori*: frequency, distribution, mutations, marker potential and their conservation in heterologous species.** *Genetics* 2005, **169**:197-214.
3. Harr B, Kauer M, Schlötterer C: **Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2002, **99**:12949-12954.
4. Ramel C: **Mini- and microsatellites.** *Environ Health Perspect* 1997, **105(Suppl 4)**:781-789.
5. La Rota M, Kantety RV, Yu JK, Sorrells ME: **Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley.** *BMC Genomics* 2005, **6**:23.
6. Eckert M, Gout L, Rouxel T, Blaise F, Jedryczka M, Fitt B, Balesdent MH: **Identification and characterization of polymorphic minisatellites in the phytopathogenic ascomycete *Leptosphaeria maculans*.** *Curr Genet* 2005, **47**:37-48.
7. Vergnaud G, Denoeud F: **Minisatellites: mutability and genome architecture.** *Genome Res* 2000, **10**:899-907.
8. Pouwels D, Simons G: **Fingerprinting microorganisms.** *Food Technol* 2003, **57**:36-40.
9. Parker PG, Snow AA, Schug MD, Booton GC, Fuerst PA: **What molecules can tell us about populations: choosing and using a molecular marker.** *Ecology* 1998, **79**:361-382.
10. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R: **Variable number of tandem repeat (VNTR) markers for human gene mapping.** *Science* 1987, **235**:1616-1622.
11. Arnold J: **Editorial.** *Fungal Genet Biol* 1997, **21**:254-257.
12. Rouxel T, Balesdent MH: **The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era.** *Mol Plant Pathol* 2005, **6**:225-241.
13. Kuhn ML, Gout L, Howlett BJ, Melayah D, Meyer M, Balesdent MH, Rouxel T: **Genetic linkage maps and genomic organization in *Leptosphaeria maculans*.** *Eur J Plant Pathol* 2006, **114**:17-31.
14. Pongam P, Osborn TC, Williams PH: **Genetic analysis and identification of amplified fragment length polymorphism markers linked to the alm1 avirulence gene of *Leptosphaeria maculans*.** *Phytopathology* 1998, **88**:1068-1072.
15. Cozijnsen AJ, Popa KM, Purwantara A, Rolls BD, Howlett BJ: **Genome analysis of the plant pathogenic ascomycete *Leptosphaeria maculans*: mapping mating type and host specificity loci.** *Mol Plant Pathol* 2000, **1**:293-302.
16. Howlett BJ, Idnurm A, Pedras MS: ***Leptosphaeria maculans*, the causal agent of blackleg disease of Brassicas.** *Fungal Genet Biol* 2001, **33**:1-14.
17. Faircloth BC: **Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design.** *Mol Ecol Res* 2008, **8**:92-94.
18. Fukuoka H, Nunome T, Minamiyama Y, Kono I, Namiki N, Kojima A: **Read2Marker: a data processing tool for microsatellite marker development from a large data set.** *Biotechniques* 2005, **39**:472-476.
19. Dereeper A, Argout X, Billot C, Rami JF, Ruiz M: **SAT, a flexible and optimized Web application for SSR marker development.** *BMC Bioinformatics* 2007, **8**:465.
20. Castelo AT, Martins W, Gao GR: **TROLL-Tandem Repeat Occurrence Locator.** *Bioinformatics* 2002, **18**:634-636.
21. Staden R, Beal KF, Bonfield JK: **The STADEN Package.** *Methods Mol Biol* 2000, **132**:115-30.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
23. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
24. Rozen S, Skaletsky H: **Primer3 on the www for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
25. Bernardi G: **The vertebrate genome: isochores and evolution.** *Mol Biol Evol* 1993, **10**:186-204.
26. Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R: **Isochore chromosome maps of eukaryotic genomes.** *Gene* 2001, **276**:47-56.
27. Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, Balesdent MH, Rouxel T: **Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*.** *Mol Microbiol* 2006, **60**:67-80.
28. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T, Bozkurt TO, Ah-Fong AMV, Alvarado L, Anderson VL, Armstrong MR, Avrova A, Baxter L, Beynon J, Boevink PC, Bollmann0 SR, Bos JIB, Bulone V, Cai G, Cakir S, Carrington JC, Chawner M, Conti L, Costanzo S, Ewan R, Fahlgren N, Fischbach MA, Fugelstad J, Gilroy EM, Gnerre S, Green PJ, Grenville-Briggs LJ, Griffith J, Grünwald NJ, Horn K, Horner NR, Hu CH, Huitema E, Jeong DH, Jones AME, Jones JDG, Jones RW, Karlsson EK, Kunjeti SG, Lamour K, Liu Z, Ma LJ, MacLean D, Chibucos MC, McDonald H, McWalters J, Meijer HJG, Morgan W, Morris PF, Munro CA, O'Neill K, Ospina-Giraldo M, Pinzón A, Pritchard L, Ramsahoye B, Ren Q, Restrepo S, Roy S, Sadanandom A, Savidor A, Schornack S, Schwartz DC, Schumann UD, Schwessinger B, Seyer L, Sharpe T, Silvar C, Song J, Studholme DJ, Sykes S, Thines M, van de Vondervoort PJI, Phuntumart V, Wawra S, Weide R, Win J, Young C, Zhou S, Fry W, Meyers BC, Van West P, Ristaino J, Govers F,

Birch PRJ, Whisson SC, Judelson HS, Nusbaum C: **Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*.** *Nature* 2009, **461**:393-398.

29. Martin F, Aerts A, Ahrén D, Brun A, Danchin EG, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buée M, Brokstein P, Canbäck B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbé J, Lin YC, Legué V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Oudot-Le Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kües U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouzé P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV: **The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis.** *Nature* 2008, **452**:88-92.

30. Kraemer L, Beszteri B, Gabler-Schwarz S, Held C, Leese F, Mayer C, Pohlmann K, Frickenhaus S: **STAMP: Extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design.** *BMC Bioinformatics* 2009, **10**:41.

doi:10.1186/1756-0500-3-322
Cite this article as: Bally *et al.*: FONZIE: An optimized pipeline for minisatellite marker discovery and primer design from large sequence data sets. *BMC Research Notes* 2010 **3**:322.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Article 2. Incidence of genome structure, DNA asymmetry and cell physiology on T-DNA integration in chromosomes of the phytopathogenic fungus *Leptosphaeria maculans*.

S Bourras, M Meyer, J Grandaubert, N Lapalu, I Fudal, J Linglin, B Ollivier, F Blaise, MH Balesdent & T Rouxel.

Publié le 1 août 2012 dans *G3 (Bethesda)* **2:**891.

Incidence of Genome Structure, DNA Asymmetry, and Cell Physiology on T-DNA Integration in Chromosomes of the Phytopathogenic Fungus *Leptosphaeria maculans*

Salim Bourras,* Michel Meyer,* Jonathan Grandaubert,* Nicolas Lapalu,† Isabelle Fudal,* Juliette Linglin,* Benedicte Ollivier,* Françoise Blaise,* Marie-Hélène Balesdent,* and Thierry Rouxel*,[1]

*Institut National de la Recherche Agronomique (INRA), Research Unit 1290 BIOGER, F-78850 Thiverval-Grignon, France, and †INRA, Research Unit 1290 BIOGER, F-78026 Versailles Cedex, France

ABSTRACT The ever-increasing generation of sequence data is accompanied by unsatisfactory functional annotation, and complex genomes, such as those of plants and filamentous fungi, show a large number of genes with no predicted or known function. For functional annotation of unknown or hypothetical genes, the production of collections of mutants using *Agrobacterium tumefaciens*–mediated transformation (ATMT) associated with genotyping and phenotyping has gained wide acceptance. ATMT is also widely used to identify pathogenicity determinants in pathogenic fungi. A systematic analysis of T-DNA borders was performed in an ATMT-mutagenized collection of the phytopathogenic fungus *Leptosphaeria maculans* to evaluate the features of T-DNA integration in its particular transposable element-rich compartmentalized genome. A total of 318 T-DNA tags were recovered and analyzed for biases in chromosome and genic compartments, existence of CG/AT skews at the insertion site, and occurrence of microhomologies between the T-DNA left border (LB) and the target sequence. Functional annotation of targeted genes was done using the Gene Ontology annotation. The T-DNA integration mainly targeted gene-rich, transcriptionally active regions, and it favored biological processes consistent with the physiological status of a germinating spore. T-DNA integration was strongly biased toward regulatory regions, and mainly promoters. Consistent with the T-DNA intranuclear-targeting model, the density of T-DNA insertion correlated with CG skew near the transcription initiation site. The existence of microhomologies between promoter sequences and the T-DNA LB flanking sequence was also consistent with T-DNA integration to host DNA mediated by homologous recombination based on the microhomology-mediated end-joining pathway.

KEYWORDS
T-DNA
mutagenesis
Leptosphaeria maculans
genome structure
Gene Ontology

The first eukaryotic (and fungal genome) to be sequenced was that of the budding yeast *Saccharomyces cerevisiae* (Goffeau *et al.* 1996). Since then, an ever-expanding number of fungal genomes has been made available, and the genome sequence of more than 300 isolates from more than 150 fungal species is currently available or in progress (http://cfgp.riceblast.snu.ac.kr/main.php; http://fungalgenomes.org/wiki/Fungal_Genome_Links), with prospects for more fungal genome sequencing, such as the 1000 fungal genome initiative (Grigoriev *et al.* 2011). Whereas high-throughput approaches, such as transcriptomics, proteomics, and comparative genomics between related species, have proved useful in eukaryotic genome annotation to predict the correct gene structure, functional annotation lags behind, and complex genomes, such as those of plants and filamentous fungi, show a large number of genes with no predicted or known function [*e.g.*, Arabidopsis Genome Initiative (2000)]. The dramatic increase in whole-genome

Copyright © 2012 Bourras *et al.*
doi: 10.1534/g3.112.002048
Manuscript received January 21, 2012; accepted for publication June 7, 2012
This is an open-access article distributed under the terms of the Creative Commons Attribution Unported License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
[1]Corresponding author: INRA, UR 1290 BIOGER, Avenue Lucien Brétignières, BP 01, F-78850 Thiverval-Grignon, France. E-mail: rouxel@versailles.inra.fr

sequencing is thus accompanied by a dramatic difficulty to reach the full biological value of the sequenced genomes with accurate identification of the protein-coding genes, as well as the nature of the functional protein products. In yeast and in some model plants, such as *Arabidopsis thaliana*, this was partly achieved with the involvement of a wide community, which promoted the development of strain/line collections in which virtually every protein-coding gene in the genome was modified, for example, by deleting, tagging with green fluorescent protein (GFP), or engineering for overexpression (Jones *et al.* 2008; Huh *et al.* 2003; Winzeler *et al.* 1999; Alonso *et al.* 2003). Even with this wide involvement of the research community, *ca.* 1000 of 5796 (17%) of protein-encoding genes in yeast and *ca.* one third of *A. thaliana* proteins still lack a functional annotation to date (Pena-Castillo and Hughes 2007; Kourmpetis *et al.* 2011).

Although the production of large collections of mutants with disrupted or inactivated genes associated with genotyping and phenotyping has gained wide acceptance for functional annotation of unknown or hypothetical genes, this has only been developed in a few tractable model plant or fungal species, mainly using *Agrobacterium tumefaciens*–mediated transformation (ATMT) (Alonso *et al.* 2003; Michielse *et al.* 2005; Krishnan *et al.* 2009; Thole *et al.* 2010). Furthermore, the whole-genome investigation for T-DNA tag distribution and the biases linked with integration conditioning the possibility to reach saturation mutagenesis has only been investigated for very few model plant species [*i.e.*, *A. thaliana*, rice, and *Brachypodium distachyon* (Alonso *et al.* 2003; Krishnan *et al.* 2009; Thole *et al.* 2010)], and only one phytopathogenic filamentous fungus, *Magnaporthe oryzae* (Choi *et al.* 2007; Meng *et al.* 2007).

Filamentous fungi, including *M. oryzae*, were first believed to have compact genomes with very few repeated elements and repeat-rich genomic regions. Filamentous fungi are amenable to ATMT (Michielse *et al.* 2005). Furthermore, biases linked with T-DNA integration in the genome of *M. oryzae* were indicated to be lower than in plants and the T-DNA integration was suggested to be "more canonical" than in plants (Choi *et al.* 2007). However, the sequencing of numerous fungal species indicates an extreme diversity of genomic complexity, genome size, and genomic landscapes, ranging from those fungi with compact genomes to fungi where massive transposable element (TE) invasion took place, eventually resulting in genome sizes larger than that of *A. thaliana* (*e.g.*, Spanu *et al.* 2010). Filamentous fungi with complex genomes also are characterized by compartmentalized, "two-speed" genomes in which specific compartments of the genome, usually TE-rich, are also enriched in genes involved in niche adaptation, such as pathogenicity effectors in phytopathogenic fungi. Examples of this encompass dispensable 'B' chromosomes of *Fusarium* (Ma *et al.* 2010), TE-rich regions of the powdery mildew fungi (Spanu *et al.* 2010), and AT-rich isochores comprising one third of the genome of *Leptosphaeria maculans* (Rouxel *et al.* 2011). By comparison with *M. oryzae*, there is only preliminary information on how T-DNA integration will happen in such genomes and how it will be useable for saturation mutagenesis of genes involved in niche adaptation. For example, in the case of the fungal pathogen of oilseed rape, *L. maculans*, analysis of 135 T-DNA integration events at a time when the genome sequence was not available indicated T-DNA preferentially integrated as a single copy in gene-rich regions of the fungal genome, but not in AT isochores (Blaise *et al.* 2007). The low frequency of T-DNA tags corresponding to known or putative protein-coding genes (19.3%) also suggested a probable bias toward intergenic and/or regulatory regions (Blaise *et al.* 2007). However, as underlined by some authors (Meng *et al.* 2007), the absence of genome sequence for *L. maculans* limited the analyses that could be performed (*i.e.*, on favored targets for T-DNA integration), and conclusions about the possible bias toward promoter regions drawn by Blaise *et al.* (2007) could not be substantiated.

The objective of this article was to further evaluate the suitability of ATMT for random saturation mutagenesis in the compartmentalized fungal genome of *L. maculans* and to further establish the mechanism of T-DNA integration in fungal genomes, taking advantage of an increased collection of T-DNA–mutagenized isolates and availability of the *L. maculans* genome sequence. One of the main questions to be addressed regarded the accessibility of AT isochores of the genome and thus of genes involved in pathogenicity by the T-DNA. A total of 400 T-DNA tags were generated, and their pattern of integration in the genome was investigated in terms of chromosomal biases, distribution within chromosomes, distribution within protein-coding regions, and targeted motifs. In addition, a Gene Ontology (GO) annotation was done and compared with that of the whole genome to identify possible insertion biases due to the physiological stage of the germinating conidia at the time of ATMT process.

MATERIALS AND METHODS

Transformation of *L. maculans* germinating conidia

All *L. maculans* transformants were issued from the *A. tumefaciens*–mediated transformation (ATMT) of the reference isolate v23.1.3 (Attard *et al.* 2002) sequenced by Genoscope (Rouxel *et al.* 2011). Generation of the collection was described by Blaise *et al.* (2007) and increased for this study to *ca.* 5000 T-DNA–tagged lines. Briefly, ATMT was performed on germinating conidia using the *A. tumefaciens* strain C58pGV2260 harboring the binary vector pBBH. The vector contains a hygromycin B resistance gene (*hph*) under the control of the *Aspergillus nidulans* gpdA promoter (Blaise *et al.* 2007).

Definition of genome compartment for insertion of T-DNA tags

L. maculans chromosomes were first compartmentalized based on nucleic acids composition in AT-rich and GC-equilibrated isochores as described in Rouxel *et al.* (2011).

Following the automated annotation of the genome (Rouxel *et al.* 2011) and regardless of the isochore structure of the genome, we generated four gene-centered collections of sequences: (i) coding regions, defined as sequences from start to stop codons, and further subdivided to introns and exons; (ii) gene-promoter regions; (iii) terminator regions; and (iv) intergenic regions, defined as the remaining genomic sequences. Regulatory regions, and mainly promoters, are often ill-defined in fungi. For this reason, when analyzing T-DNA tag position relative to genes in the *M. oryzae* genome, Meng *et al.* (2007) defined three sets of 500 bp, 750 bp, and 1000 bp for 5′ promoter regions and downstream regulatory regions. However, the increase in size of regulatory regions did not drastically change the features associated with the T-DNA–favored target (Meng *et al.* 2007). In addition, the genome of *L. maculans* is more compact than that of *M. oryzae* in GC isochores [*e.g.*, median size of intergenic regions in the case of head-to-tail ORFs is 670 bp (Rouxel *et al.* 2011)]. We thus only investigated here one range of size for promoters and terminators: 500 bp upstream of gene start codons or downstream of gene stop codons, respectively. Collections of gene-promoter, terminator, and intergenic regions were extracted using a Python script, departing from gene coordinates.

Recovery of T-DNA–flanking sequences and analysis of T-DNA–targeted genes

T-DNA–flanking sequences were recovered from genomic DNA by thermal asymmetric interlaced-PCR (tail-PCR) and PCR-walking techniques as described in Liu *et al.* (1995) and Balzergue *et al.* (2001), respectively. Sequencing was performed on PCR products using a Beckman Coulter CEQ 8000 automated sequencer (Beckman Coulter, Fullerton, CA, USA) according to the manufacturer's instructions. All sequences were cured manually and aligned to *L. maculans* genome sequence using BLASTn with a cutoff e-value of 1e−10. The latter step was automated using a homemade script in Python. The position of an insertion site was defined as the position of the first aligned nucleotide to a flanking sequence. All extracted positions were mapped and plotted on the *L. maculans*–assembled genome using homemade scripts in Python and R. Based on mapping of T-DNA insertion sites, genes with a T-DNA tag in their promoter, terminator, or coding region were extracted, mapped, and analyzed for size, compositional, and structural features. The latter step was automated using homemade scripts in Python.

Functional annotation using GO

GO annotations of *L. maculans*–predicted genes were done with Blast2GO (Götz *et al.* 2008) as described in Rouxel *et al.* (2011). The NCBI "NR" database (October 16, 2009, release version) was queried with all predicted genes using BLAST algorithm version 2.2.21 on the URGI high-throughput computing cluster (128 Intel Xeon E5450). All genes were mapped according to GO, GeneInfo, Gene2accession, and PIR data, and then analyzed with Blast2GO, which applies GO annotations from BLAST search results. This process takes into account sequence similarity and the evidence code (EC) associated with GO annotations. Finally, GO annotations were enriched using Annex and Interproscan data. In this work, we chose to use the "biological process" vocabulary for functional annotation and comparison between T-DNA–tagged genes and all genes of the genome, because this GO vocabulary was found to better fit fungal behavior when described from a physiological or phenotypical point of view. In addition, it is the vocabulary for which the highest annotation number was obtained in yeast (Christie *et al.* 2009).

Statistical analyses

Biases were assessed by calculating the standardized residuals between observed and expected values as follows: SR = (Observed – Expected) / √Expected. SR calculation allows the detection of outlying observations [*i.e.*, those that appear to deviate from other members of the sample in which they occur (Grubbs 1969)]. In general, SR $>$ 0 means the observed value is greater than expected, and by contrast, SR $<$ 0 means the observed value is smaller than expected. To test whether the outlying observations deviate significantly from what is expected, the SR distribution following a normal distribution was estimated using the Kolmogorov-Smirnov test embedded in XLSTAT statistical analysis software version 2009.6.02 (with default parameters). Therefore, when the hypothesis of normal distribution was not rejected, SR exceeding the absolute value of 1.96 was considered a bias (*i.e.*, significantly deviant from the rest of the data).

The Monte Carlo test on contingency tables was used as an alternative to assess biases of T-DNA tags mapping. This nonparametric test based on simulations assesses the independence between rows and columns. Then, when coupled with Fisher's exact test, it determines whether the difference between the observed and the theoretical values is significant. All calculations were performed using the appropriate XLSTAT function with default parameters.

The linear regression option of XLSTAT was used to model the relationships between data sets. A graphical output comprising the regression line and the 95% confidence intervals area was generated using the embedded function of the software.

RESULTS

Generation of the repertoire of T-DNA–flanking sequences

A subset of the collection of 5000 transformants of *L. maculans* obtained by ATMT was selected for sequencing the T-DNA insertion borders. Four-hundred sequences were obtained. Of these, 40 T-DNA–flanking sequences were generated by PCR walking (Balzergue *et al.* 2001) and 360 by tail-PCR (Liu *et al.* 1995). BLASTn searches against the *L. maculans* genome indicated 33 sequences (8.25%) had no BLAST hit (sequences too short for the BLASTn algorithm and sequences corresponding to the bacterial vector). The remaining 367 sequences were filtered for ambiguous BLAST hits (poor homology below the cutoff e-value of 1e−10), resulting in the final repertoire of 318 flanking sequences corresponding to single-locus T-DNA integration events in unique transformants. Of these, 217 sequences were obtained by sequencing the left border (LB) of the T-DNA insertion, and 101 by sequencing its right border (RB).

Compartmentalization of the genome and T-DNA integrations

The *L. maculans* genome is compartmentalized into two distinct genomic landscapes: GC isochores (summing up 64% of the genome and containing 95% of the genes) and AT isochores (summing up 36% of the genome and 5% of the genes, but mainly consisting of mosaics of inactivated and truncated TEs) (Rouxel *et al.* 2011). The

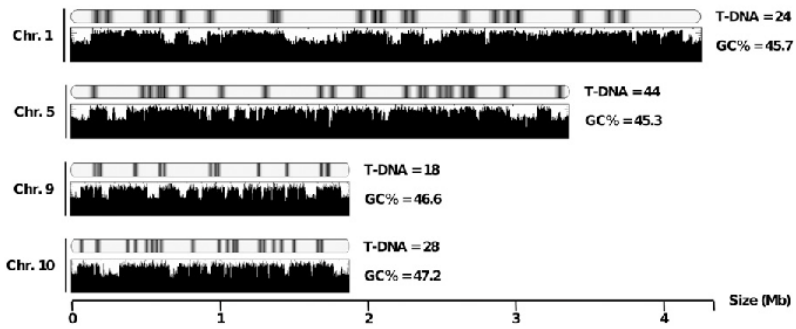


Figure 1 A schematic representation of occurrence of T-DNA insertion events along four *L. maculans* chromosomes. For each chromosome, the upper plot shows the location of the T-DNA integration events, and the lower plot schematizes variations in GC content along the chromosome, defining AT-rich and GC-equilibrated isochores. The average GC percentage of the chromosome is indicated.

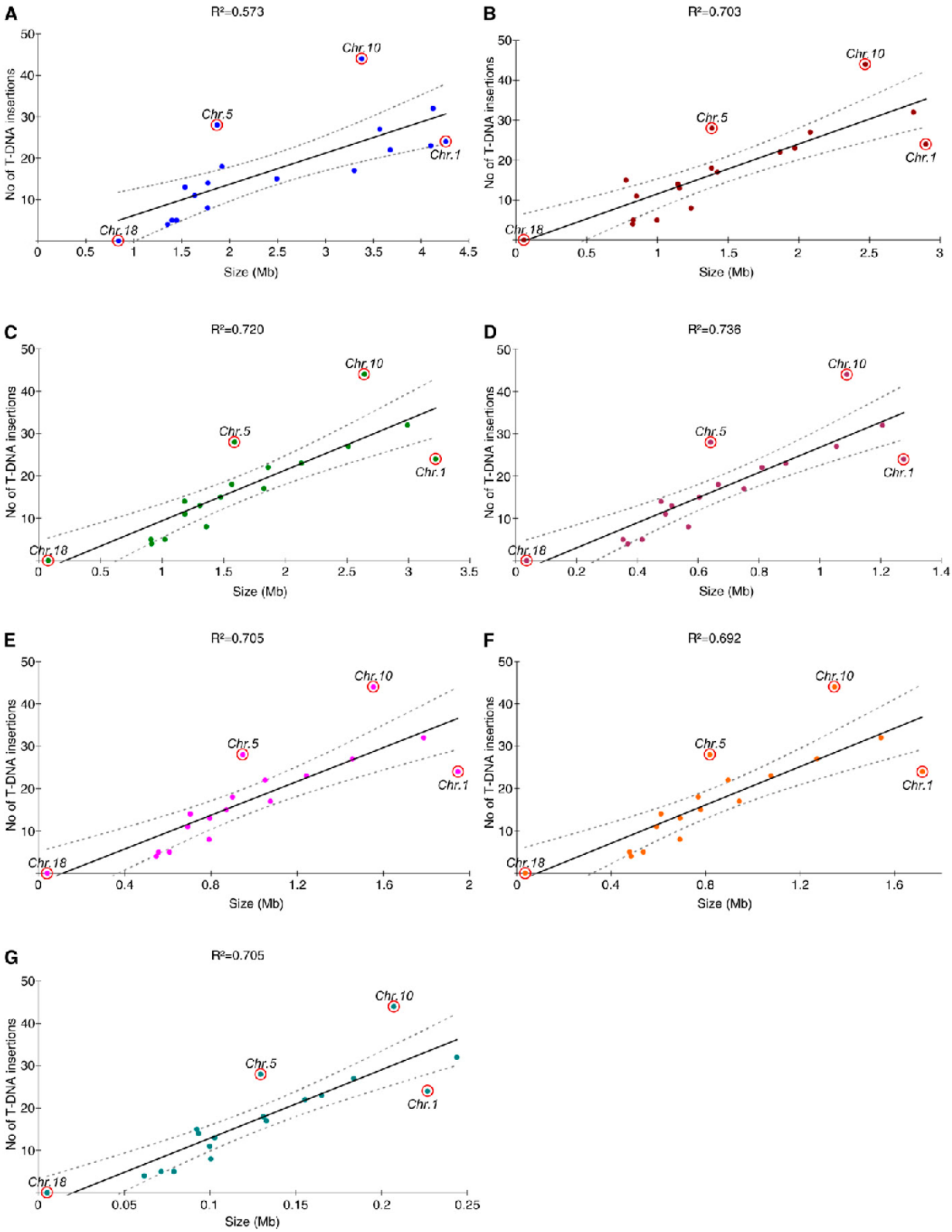

A
R²=0.573
No of T-DNA insertions
Size (Mb)
Chr.10
Chr.5
Chr.1
Chr.18
B
R²=0.703
C
R²=0.720
D
R²=0.736
E
R²=0.705
F
R²=0.692
G
R²=0.705

Table 1 Distribution of T-DNA insertion events within *L. maculans* genomic regions

Genomic Regions			T-DNA Insertion Events		
Type	Size (Mb)	% Genome	Observed	Expected[a]	SR[b]
Regulatory[c]	11.8	26	200	83	12.92
5′ promoting[c]	5.9	13	122	41	12.64
3′ terminating[c]	5.9	13	78	41	5.77
Coding[c]	17.6	39	119	123	−0.37
Exons	15.3	34	86	107	−2.04
Introns	2.3	5	33	16	4.24
Shared[c]	—	—	41	—	
Intergenic[c]	15.7	35	40	110	−6.67

[a] Expected number of T-DNA integration events (T-IE) [= (T-IE genomic density) × (genomic region size)]. Values were approximated to the nearest integer.
[b] Standardized residues. We considered a normal distribution of SRs because we cannot reject the null hypothesis as revealed by the Kolmogorov-Smirnov test (*P*-value = 0.976, $\alpha = 0.05$).
[c] Regulatory regions, defined as the sum of promoting and terminating regions of the 12,469 predicted genes of *L. maculans*; Gene-promoting regions, 500 bp upstream of the start codon; Gene-terminating regions, 500 bp downstream of the stop codon; Gene-coding regions, from start to stop codons, including introns; Shared, common regulatory regions shared by two head-to-tail nearby genes; Intergenic, genomic regions corresponding to none of the previous criteria. Note that overlaps between compartments may occur, leading to a total number of sequences higher than 318.

T-DNA insertions were graphically coincident to GC isochores in an almost systematic fashion (Figure 1), and 96.9% of T-DNA tags were mapped to GC isochores *vs.* only 3.1% that were mapped to AT isochores. AT isochores are further subdivided into three categories: telomeres (representing *ca.* 2% of the genome assembly); large-sized isochores (13–325 kb, representing *ca.* 31% of the genome assembly) corresponding to complex mosaics of TEs; and mid-sized isochores (1–13 kb, summing up *ca.* 3% of the genome assembly) generally corresponding to the insertion of a single TE within a GC isochore (Rouxel *et al.* 2011). The above-mentioned depletion of T-DNA integration in AT isochores was mainly due to a very low amount of integrations in large AT isochores with only two cases observed (0.6%), whereas 1.6% of the tags were found in telomeres that may contain active genes, including numerous copies of a RecQ telomere-linked helicase (Rouxel *et al.* 2011). No T-DNA tag was coincident with multiple loci in the genome, even in AT-rich compartments and telomeres, due to sequence divergence generated by RIP acting on repeated copies of TEs. Actually, even when mapping to a given TE family, the tag sequence, when unambiguous, was always sufficient to derive a single locus due to these sequence polymorphisms.

In addition to these two distinct compartments, the rDNA array summing up 1.7% of the genome assembly (Rouxel *et al.* 2011) was also underrepresented with no T-DNA tags targeting it.

Chromosomal features and T-DNA integrations

The number of T-DNA insertions per chromosome was then compared with seven chromosomal features (Figure 2), and the distribution of T-DNA insertions was plotted against each of these features. Globally, the number of T-DNA integrations was linearly correlated with all investigated features, but it better correlated with size of the GC isochores within chromosomes ($R^2 = 0.703$) than with total size of the chromosome ($R^2 = 0.573$). The favored insertion sites were transcriptional regions ($R^2 = 0.720$), mainly regulatory regions ($R^2 = 0.736$) and introns ($R^2 = 0.705$) (Figure 2). Most chromosomes, except chromosomes 5, 10, and to a lesser extent, 1, showed such a linear correlation between the number of T-DNA integrations and chromosomal features (see below).

Favored T-DNA insertion events in genic regions

Noticing that the chromosomal distribution of T-DNA tags was correlated to the size of gene-regulatory and gene-coding regions within chromosomes, we studied to what extent compartmentalization features are involved in profiling whole-genome T-DNA insertion occurrence. In contrast to what is described in *M. oryzae* (Meng *et al.* 2007; Choi *et al.* 2007), targeting of ORF was not significantly different in *L. maculans* to what would be expected under the hypothesis of random integration in the genome (Table 1). Biases assessment using the SR method showed that T-DNA insertions were less common than expected in intergenic regions (SR = −6.67) and exons (SR = −2.04) and more common in gene regulatory regions (SR = 12.92) and gene introns (SR = 4.24) (Table 1). Biases in favor of regulatory regions were corroborated by the Monte-Carlo test. However, no significant bias was observed in intergenic regions and introns according to this analysis (data not shown).

Promoter features favoring T-DNA targeting

Because promoters are the main genomic regions in which T-DNA integration occurred, we analyzed further promoter regions to investigate the involvement of host-DNA asymmetry and T-DNA–host-DNA shared microhomologies to favor the T-DNA targeting. Previous studies noticed an increased CG skew around transcription start site in *A. thaliana* and other eukaryotes (Tatarinova *et al.* 2003).

Figure 2 Correlation between the number of T-DNA integrations and chromosomal features. The features investigated for each chromosome were (A) chromosome size; (B) total size of the GC isochores; (C) total size of the transcriptional regions [defined as the sum of regulatory sequences (promoter + terminator) and gene-coding sequences (exons + introns)]; (D) total size of the regulatory regions (defined as the sum of promoter and terminator sequences); (E) total size of gene-coding regions (defined as the sum of exonic and intronic sequences); (F) total size of the exonic sequences; and (G) total size of the intronic sequences. Regression curves and the 95% confidence intervals are plotted in continuous and discontinuous lines, respectively.

The targeted promoter sequences were thus analyzed for DNA asymmetry by calculating CG and AT skews. Positive CG and AT skew values indicate an overabundance of C and A residues, respectively, whereas negative CG and AT skew values indicate an overabundance of G and T residues, respectively. Sequences 500 bp upstream the transcription initiation start codon of 122 T-DNA–targeted genes harboring a T-DNA tag in their promoter region were first extracted, and then CG and AT skews were calculated, plotted, and compared with the density of T-DNA integration events in the same intervals (Figure 3A). The T-DNA tag density and CG skew increased gradually when getting closer from the start codon, to reach their maximum value at positions −113 and −50 respectively, and then decreased dramatically (Figure 3A), whereas CG and AT skews plotted differently but shared the same peak position at −50. To corroborate the functional meaning of CG skew peak in promoters targeted by T-DNA, we compared it with the CG skew profile of whole *L. maculans* promoters following the extraction of sequences 500 bp upstream of the transcription initiation start codon for all 12,469 *L. maculans*–predicted genes. Comparison of AT/CG skews between promoters of both collections showed that, in both cases, CG skews reached their peak value at position −50. By contrast, AT skew peak in whole-genome promoters profile plotted at position −25, closer to gene start (Figure 3B).

To assess to what extent DNA asymmetry impacted T-DNA integration, the occurrence of CG and AT skews contexts at T-DNA insertion sites were also analyzed in gene terminator, gene coding and intergenic regions. Sequences 200 bp upstream and downstream the 318 insertion sites were extracted and split out into three groups: (i) gene terminator regions (78 sequences), (ii) gene coding regions (119 sequences) and (iii) intergenic regions (40 sequences). For comparison purposes, the 122 sequences corresponding to gene promoter regions were added.

Skew graphics showed that T-DNA insertions occurred preferentially in increased CG skew context, in all genomic compartments (Figure 4A) and also in a weak AT skewed context for promoter, terminator and gene coding regions. In contrast, AT skew was increased at the insertion sites in intergenic regions (Figure 4B).

Microhomologies between the T-DNA left border and T-DNA preinsertion sites

T-DNA integration to host DNA is mediated by two major mechanisms: nonhomologous recombination (NHR) and homologous recombination (HR) via the T-DNA LB [for review, see Tzfira *et al.* (2004) and Citovsky *et al.* (2007)]. The former is in fact a HR-like mechanism relying on the microhomology-mediated end-joining (MMEJ) pathway and its property to use 5–25 bp microhomologous

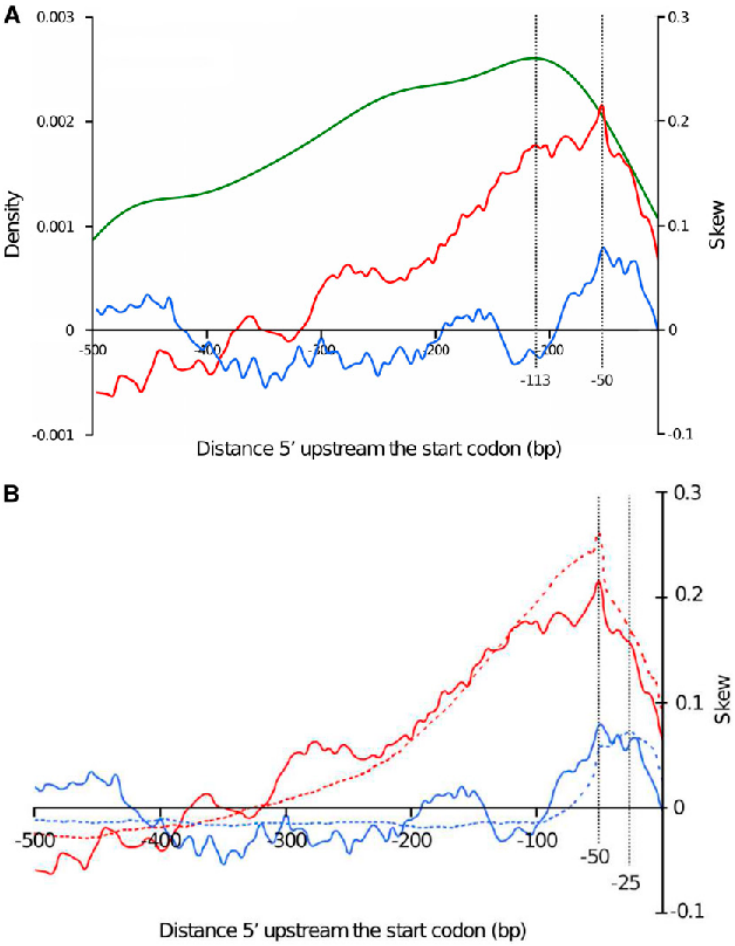


Figure 3 The link between CG skew and AT skew in gene promoter regions and favored T-DNA integration events. A. Density of T-DNA insertions in promoter regions (green curve), CG skew (red curve) and AT skew (blue curve) variations along T-DNA-targeted gene promoter regions, as a function of location from the ATG. B. Comparison of CG skew (red curve) and AT skew (blue curve) variations between promoter regions of T-DNA–targeted genes (plain lines) and promoters of all *L. maculans* predicted genes (dotted lines).

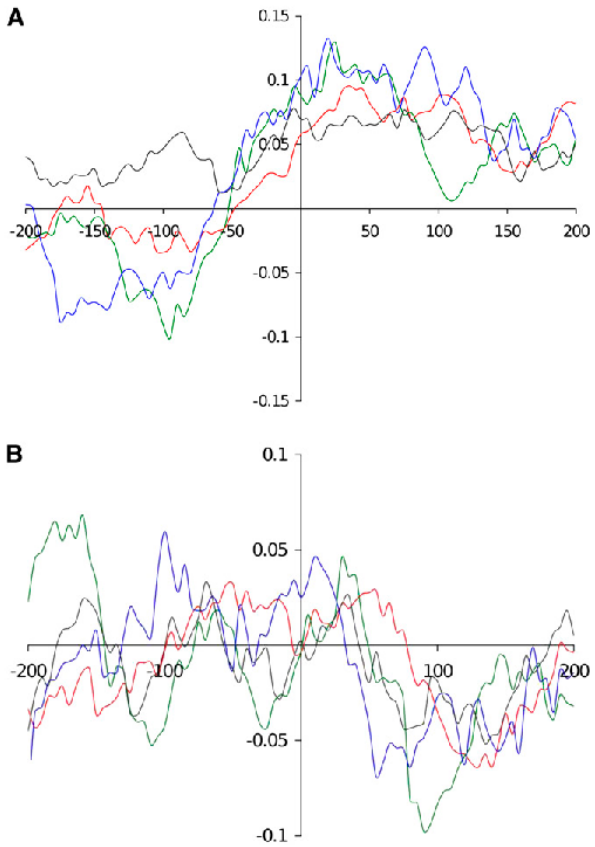


Figure 4 Analysis of CG (A) and AT skews (B) at T-DNA insertion sites in four targeted compartments of the genome. Sequences 200 bp upstream and downstream of the integration sites were extracted and CG/AT skews were calculated. The sequences were then grouped according to four compartments of the genome: promoter (red curves), terminator (green curves), intergenic (blue curves), and protein coding (black curves) regions.

sequences to anneal and join free single-stranded DNA ends [for review, see McVey and Lee (2008)]. We thus investigated whether microhomologies between the 25-bp T-DNA LB and host DNA could be found at the insertion site.

Sequences 25-bp backward of the insertion sites were extracted and chosen so that (i) they correspond to sequences upstream of a junction between T-DNA LB and host DNA; (ii) they were exempt from potential filler DNA at T-DNA–host-DNA junction; and (iii) they were exempt from gaps and undetermined nucleotides (N) in the current version of genome assembly (Rouxel *et al.* 2011). A total of 160 25-bp sequences was thus obtained. We also divided the T-DNA LB into 5-bp-long successive sequences that we named "microhomology motifs" and aligned the 21 resulting motifs to the 160 selected preinsertion sites (Figure 5). Nineteen putative microhomology motifs were found in 69 locations, distributed unequally among 41 preinsertion sites (25.6%) (Figure 5). No single motif was common to all sequences, but 3 microhomology motifs (14.2%) TTGGC (Figure 5, alignments 22–27), ATATA (Figure 5, alignments 45–53), and TATAT (Figure 5, alignments 54–62), were found in 32.4% of the locations, suggesting the presence of homology islands. In addition, TATA-containing motifs were the most represented (32.4%) (Figure 5, alignments 42–65).

Assessment of T-DNA LB sequence affinity with common core promoter elements of eukaryotic genes (TATA box with consensus TATA(T/A)A(A/T) (Breathnach and Chambon 1981; Burley 1996), CAT box with core consensus CCAAT (Bucher 1990), and initiator (Inr) with consensus PPAN(T/A)PP (P, pyrimidine; N, any nucleotide) (Javahery *et al.* 1994) showed that the 25-bp-long T-DNA LB harbored islands of homologies with both TATA box (positions −5 to −10) and Inr (positions −20 to −25) elements. In addition, when we extended the alignment to 15 upstream supplementary bases, an additional homology with CAT box was found (Figure 6). We lastly calculated the frequency of occurrence of each base of the T-DNA LB sequence in microhomology motifs and plotted the values along the T-DNA LB sequence. As shown in Figure 7, T-DNA–host-DNA base identity increased approaching the LB free end. Corroborating the previous observation, the TATA island, but not Inr island, can be postulated to be frequently represented in T-DNA–host-DNA shared microhomologies due to the high ratio of identical bases at this location (Figure 7).

GO annotation of T-DNA–targeted genes

The whole-genome mapping of the 318 T-DNA insertions showed that 279 of these were in gene-coding or regulatory regions, whereas the other T-DNA tags were located in intergenic regions, including AT-rich, gene-poor isochores. A functional profile of the collection of T-DNA–targeted genes was performed by coupling the GO annotation of the "biological process" vocabulary with an assessment of representation bias by calculating the SRs between observed and expected number of annotations per functional category. The proportion of genes coding for hypothetical or predicted proteins of unknown function in the T-DNA–targeted gene collection (73.1%) was comparable to that observed in the whole genome (71.8%). Most of the functional categories represented in the "biological process" vocabulary (15 of 22) were represented in genes tagged by T-DNA (Table 2). The SR values for "pigmentation" (SR = 5.94), "growth" (SR = 2.09), and "cell wall organization or biogenesis" (SR = 3.21) revealed an overrepresentation in the T-DNA–targeted gene collection compared with all predicted genes of the genome, whereas the "signaling" functional category was underrepresented (SR = −2.03) (Table 2). Similar biases were identified using the Monte Carlo method (data not shown).

Functional significance of chromosome bias in T-DNA insertions?

T-DNA insertion events were mapped onto the *L. maculans* genome and plotted along its 18 chromosomes to investigate distribution biases. The T-DNA insertion density varied from 0 insertion event/Mb (chromosome 18) to 14.7 insertion events/Mb (chromosome 11) (Table 3). In most of the cases, the number of tags per chromosome was compliant with a random integration of the T-DNA. However, T-DNA insertion events were found to be statistically more common than expected into chromosomes 5 (SR = 4.08) and 10 (SR = 4.16) (Table 3 and Figure 2), whereas they were less common than expected into chromosome 18 (SR = −2.45) (Table 3). Chromosome 1, whose number of T-DNA integrations was consistent with chromosome size (Figure 2), showed a number of tags markedly lower than the mean confidence intervals for other criteria, such as size of coding or regulatory regions (Figure 2).

Chromosome 18 is very rich in TE and poor in genes (Rouxel *et al.* 2011), and the lack of T-DNA integration in this chromosome is consistent with the above-mentioned preferred integration in gene-

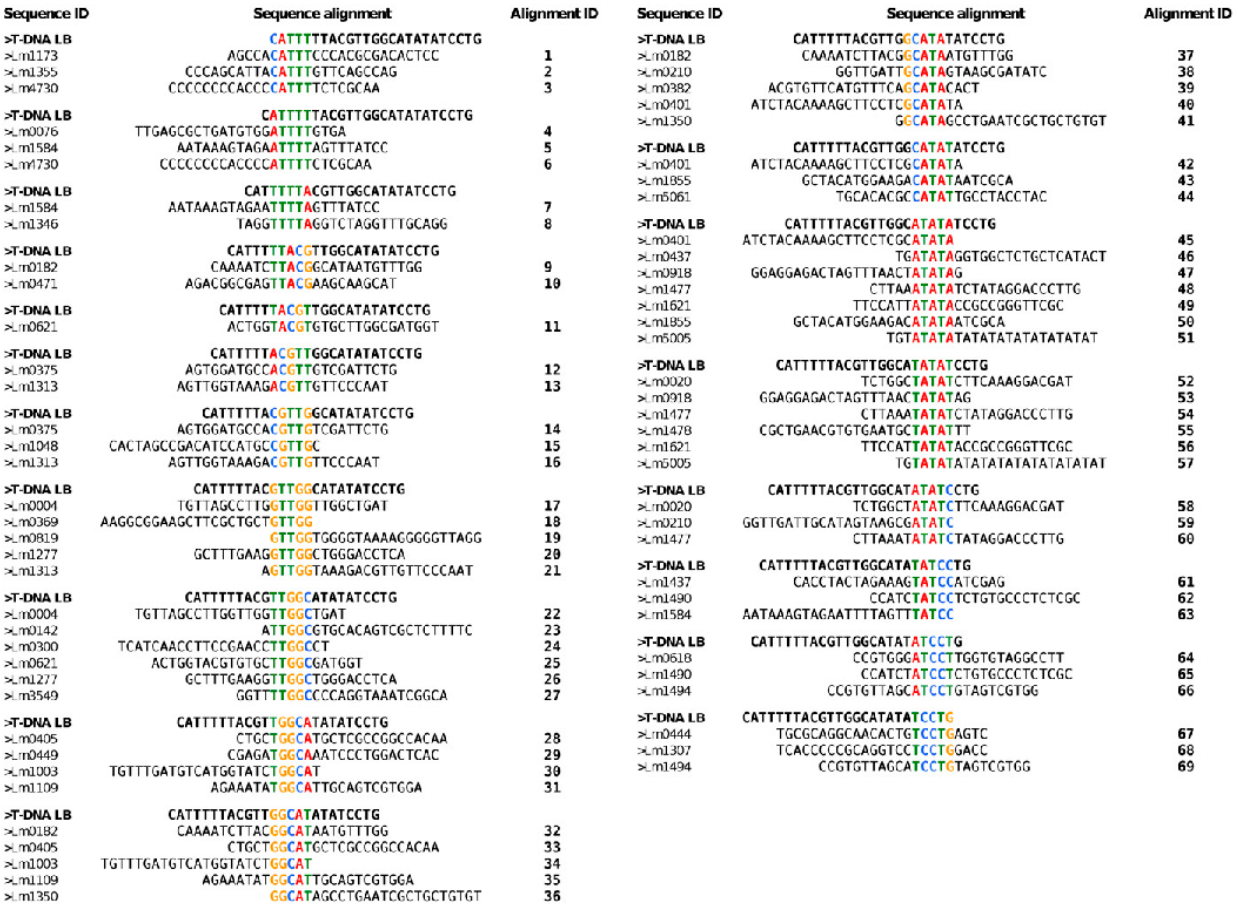


Figure 5 The search for microhomology between the host-DNA and T-DNA left border. One hundred and sixty 25-bp preinsertion sites were investigated for occurrence of 5-bp-long consecutive motifs corresponding to identical motifs in the T-DNA left border. The 41 sequences of preinsertion sites that show identity with consecutive, 1-bp sliding window, and 5-bp-long motifs are displayed.

rich genomics regions. GO annotation indicated that chromosome 5 was significantly enriched in the "pigmentation" (SR = 2.93) and "carbon utilization" (SR = 2.14) functional categories, whereas chromosome 10 showed no significant enrichment in any functional categories (Table 4). Of these, only the "pigmentation" functional category was found to be overrepresented in the collection of T-DNA–targeted genes (Table 2). These data suggest that functional bias is unrelated to genome distribution of T-DNA insertions. Biases assessment using the Monte Carlo method led to the same results (data not shown).

DISCUSSION

Although many fungal species are amenable to ATMT, the mechanisms of T-DNA integration in the fungal genomes are largely unknown compared with what is known in plants, and it is still a matter of debate to know whether T-DNA integration will be random enough to allow a systematic targeting of all genes in the genome for functional identification. In phytopathogenic filamentous fungi, numerous pathogenicity mutants were generated by ATMT, but a systematic analysis of T-DNA integration in the genomes has only been performed in *M. oryzae* (Choi *et al.* 2007; Meng *et al.* 2007) and in *L. maculans* prior to obtainment of the whole-genome sequence (Blaise *et al.* 2007). Here we exploited the *L. maculans* genome sequence to investigate how "canonical" T-DNA integration patterns are in a fungal genome with such contrasted genomic landscapes compared with what is known in *M. oryzae*. This comparison, however, has to be taken with care, as the two articles on *M. oryzae* show some experimental differences with ours. In Choi *et al.* (2007), a very large number of 1246 transformants were investigated, but more than 1100 were chosen so that they harbor phenotypic defects, thus suggesting a bias toward T-DNA integration within coding sequences, in detriment

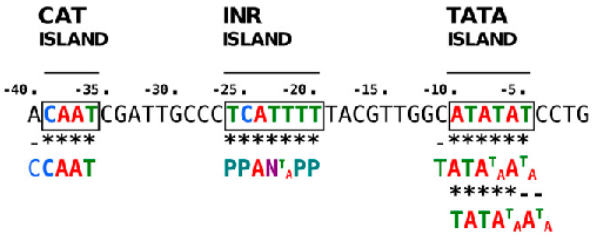


Figure 6 Occurrence of sequence microhomologies to eukaryotic core promoter elements (TATA box, CAT box, and Initiator) in the T-DNA LB and 15 upstream supplementary bases.

to lines in which noncoding regions were targeted. In this sense, Meng *et al.* (2007), who reported on characterization of a much lower number of 175 T-DNA integrations into random T-DNA tagged lines, was less biased for a systematic analysis of T-DNA patterns in filamentous fungi. In addition, the GO annotation of targeted proteins was used to have a better insight into T-DNA integration mechanisms in *L. maculans*, which has not been done in fungi, except budding yeast, to date (Christie *et al.* 2009).

The ultimate goal of ATMT mutagenesis in plants or fungi is to reach saturation mutagenesis in order to eventually reach a functional annotation of the numerous unknown or hypothetical genes in the genomes. For example, in *L. maculans*, only 43% of the predicted proteins in the genome have strongly supported functional annotation, 45% are similar to hypothetical proteins for which no functional annotation is available, and 12% are predicted proteins with no annotation whatsoever (Rouxel *et al.* 2011; J. Grandaubert, unpublished data). In addition, in the case of phytopathogens, the initial objective of the T-DNA insertional mutagenesis strategy is the generation of mutants showing pathogenicity defects, as well as the discovery of novel genes and novel functions involved in pathogenesis. For these objectives, ATMT has to target mostly genic compartments of the genomes and show limited biases in targeted genes or genomic regions. The first advantage of ATMT for this objective is the common single-copy integration of the T-DNA in genomes, and mainly in fungal genomes, including that of *L. maculans* (Michielse *et al.* 2005; Blaise *et al.* 2007). The second point to be stressed in *L. maculans* is the high percentage of recovery of flanking sequences with matches in the fungal genome (around 80%) as was also observed for *M. oryzae* (Meng *et al.* 2007), whereas in plants, the frequencies usually amount to 60–65% [*e.g.* Thole *et al.* (2010)]. As shown for *M. oryzae*, or *Arabidopsis* and other model plant species, the T-DNA integration is shown here to be nonrandom. First T-DNA integration was much rarer than expected in TEs, as is generally the case in plants (Thole *et al.* 2010; Zhang *et al.* 2007), although this bias was not found when analyzing random transformants of *M. oryzae* (Meng *et al.* 2007). One possible explanation for this discrepancy would lie in the fact that all

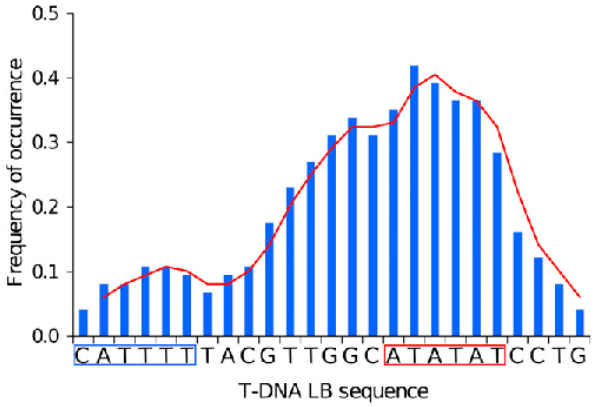


Figure 7 Analysis of microhomology at T-DNA preinsertion sites. Frequency of occurrence of single bases identical to those of the 25-bp T-DNA left border in the genome preinsertion sites were analyzed. The T-DNA LB sequence is illustrated, and homologs of the TATA box and Inr in the LB sequence are boxed.

Table 2 Gene Ontology annotation of T-DNA–targeted genes using the "biological process" vocabulary

	Whole Genome	T-DNA–targeted Genes		
	Annot.[a]	Obs. Annot.[b]	Exp. Annot.[c]	SR[d]
Pigmentation	**1**	**1**	**0.03**	**5.94**
Immune system process	1	0	0.03	−0.16
Cell proliferation	3	0	0.08	−0.28
Death	4	0	0.11	−0.33
Locomotion	4	0	0.11	−0.33
Biological adhesion	6	0	0.16	−0.40
Growth	**6**	**1**	**0.16**	**2.09**
Nitrogen utilization	9	0	0.24	−0.49
Reproduction	14	1	0.38	1.02
Carbon utilization	15	1	0.40	0.94
Multi-organism process	43	1	1.15	−0.14
Cell wall organization or biogenesis	**49**	**5**	**1.31**	**3.21**
Signaling	**153**	**0**	**4.11**	**−2.03**
Cellular component organization	177	5	4.75	0.11
Multicellular organismal process	228	6	6.12	−0.05
Cellular component biogenesis	232	8	6.23	0.71
Developmental process	257	8	6.90	0.42
Response to stimulus	262	4	7.03	−1.14
Biological regulation	469	12	12.59	−0.17
Localization	706	21	18.95	0.47
Cellular process	2489	62	66.79	−0.59
Metabolic process	3070	84	82.39	0.18
Total	8198	220	220	—

[a] Number of annotations generated per category for the 12,469 *L. maculans* predicted genes.
[b] Observed number of annotations generated by the GO analysis.
[c] Expected number of annotations [= (∑annot.) × P(functional category)]. Where (∑annot.) is the sum of all generated annotations, and P (functional category) is whole genome probability of the considered functional category. Values were approximated to two decimals.
[d] Standardized residuals. We considered a normal distribution of SRs because we cannot reject the null hypothesis as revealed by the Kolmogorov-Smirnov test (*P*-value = 0.391, $\alpha = 0.05$). Biased categories are indicated in bold.

Table 3 Distribution of T-DNA insertion events along the *L. maculans* chromosomes

Chromosomes						T-DNA Insertion Events			
No.	SC[a]	Size (Mb)	GC %	Gene Content[b]	GC Size[c] (Mb)	ρ[d] (T-IE/Mb)	Observed	Expected[e]	SR[f]
1	0	4.3	45.7	1276	2.9	5.6	24	30	−1.06
2	2+19	4.1	44.1	1206	2.8	7.8	32	29	0.58
3	6+29+11	3.7	39.7	810	1.9	6.0	22	26	−0.74
4	8+10	3.6	43.2	1055	2.1	7.6	27	25	0.41
5	1	**3.4**	**45.3**	**1089**	**2.5**	**13.0**	**44**	**24**	**4.19**
6	12+15+32	4.1	41.6	888	2.0	5.6	23	29	−1.06
7	20+21+23	3.3	44.7	751	1.4	5.2	17	23	−1.27
8	3+31	2.5	43.3	604	0.8	6.0	15	17	−0.58
9	4	1.9	46.6	666	1.4	9.4	18	13	1.25
10	5	**1.9**	**47.2**	**641**	**1.4**	**15.0**	**28**	**13**	**4.12**
11	9	1.8	45.1	478	1.1	7.9	14	12	0.45
12	7	1.8	46.0	568	1.2	4.5	8	12	−1.25
13	13	1.6	43.9	493	0.9	6.7	11	11	−0.13
14	14	1.5	47.4	513	1.2	8.5	13	11	0.69
15	17	1.4	43.7	416	1.0	3.5	5	10	−1.61
16	16	1.4	44.3	353	0.8	3.6	5	10	−1.53
17	18	1.4	44.7	369	0.8	3.0	4	9	−1.77
18	22	**0.8**	**35.3**	**36**	**0.1**	**0.0**	**0**	**6**	**-2.42**
Un.[g]	—	0.7	—	—	—	—	8	—	—
Genome	—	45.1	44.1	12469	—	7.0	318		

[a] Supercontigs reassembled to make up whole chromosomes.
[b] Number of predicted genes per chromosome.
[c] Total size of GC isochores per chromosome.
[d] T-DNA insertion event (T-IE) density [= (number of T-IEs per chromosome / chromosome size)].
[e] Based on density of T-DNA insertion events in the whole genome (7 T-IE/Mb), the expected number of T-IEs per chromosome was calculated as [(chromosome size) × (T-IE whole-genome density)]. Values were approximated to the nearest integer.
[f] Standardized residuals. We considered a normal distribution of SRs because we cannot reject the null hypothesis as revealed by the Kolmogorov-Smirnov test (*P*-value = 0.475, $\alpha = 0.05$). Chromosomes showing a significant bias in number of T-DNA insertion events are indicated in bold.
[g] Unassembled genomic sequences (summing up approximately to 0.7 Mb).

TEs in the genome of *L. maculans* are strongly degenerated and inactivated (Rouxel *et al.* 2011) and that, as discussed below, T-DNA integration favors transcriptionally active regions of the genome. Also, as is the case for *M. oryzae*, T-DNA tags were not recovered from other large arrays of repeats, such as the rDNA array, or from the mitochondrial genome (Meng *et al.* 2007), whereas the tagging of the rDNA array by T-DNA is overrepresented in some plant species, such as *B. distachyon* (Thole *et al.* 2010). Similarly to what was observed in *M. oryzae*, a marginal chromosomal bias showed some favored or disfavored chromosomes for T-DNA integration in *L. maculans*. As was noticed for *M. oryzae*, the biological significance of this fact remains obscure because no functional specificity was associated with these chromosomes. This bias may only be due to the limited number of tags analyzed in randomly tagged fungal isolates as it does not seem to occur in plants where the nonrandom integrations are observed within a chromosome rather than between chromosomes (Thole *et al.* 2010). More importantly, one main feature of T-DNA integration in the genome [*i.e.*, the favored targeting of 5′ 500-bp regions of genes assumed to be promoters] is a widely shared trait for plants and fungi (Alonso *et al.* 2003; Meng *et al.* 2007; Choi *et al.* 2007; Thole *et al.* 2010). By comparison with *M. oryzae*, the bias toward promoter regions was even more marked in the genome of *L. maculans*, consistent with the common recovery of pathogenic mutants for which the altered pathogenicity was due to T-DNA integration in promoters of genes (Elliott and Howlett 2006; Remy *et al.* 2008a,b, 2009). Lastly, the favored targeting of promoters is consistent with the presence of microhomology motifs (see below) involved in the homologous recombination with the T-DNA border.

When compiled, the observed T-DNA integration biases seem to share at least one common denominator: T-DNA integration takes place in transcriptionally active regions. In a cell, transcriptional activity should be considered the first step in the translation of genomic information into physiological state. Hence, starting from this postulate, it is logical to suppose that cellular activity affects T-DNA–favored insertion sites and that targeted genes should reflect, to a certain extent, the physiological state of the transformed cell. In this study, T-DNA insertion loci were recovered from *L. maculans* transformants obtained by ATMT for which germinating conidia (incubated for 48 hr) were used (Blaise *et al.* 2007). Conidial germination is commonly described as a three-step mechanism (D'enfert 1997; Osherov and May 2001): (i) activation, during which appropriate amounts of water and low-molecular-weight nutrients trigger conidial cell activation for germination; (ii) isotropic growth, during which the conidial cell undergoes morphological changes, uptakes water, and increases its physiological activity, which leads to an increase in size and mass; and (iii) polarized growth, during which a germ tube emerges from the conidial cell and develops, which requires *de novo* synthesis of wall materials. Germination is an asynchronic phenomenon that may differ from one conidia to another. ATMT is thus performed on conidial populations at four physiological stages: (i) ungerminated conidia, (ii) conidia at germination activation, (iii) conidia at isotropic growth, and (iv) conidia at polarized growth. The overrepresentation of "pigmentation," "growth," and "cell wall organization or biogenesis" functional categories in genes targeted by the T-DNA would be consistent with the hypothesis that targeted genes reflect the physiological state of the germinating conidia. First, even though the conidia of *L. maculans* are hyaline under the microscope,

Table 4 Gene Ontology annotation of "biological process" for chromosomes 5 and 10

Whole Genome		Chromosome 5			Chromosome 10		
	Annot.[a]	Obs. Annot.[b]	Exp. Annot.[c]	SR[d]	Obs. Annot.[b]	Exp. Annot.[c]	SR[d]
Pigmentation	**1**	**1**	**0.10**	**2.93**	**0**	**0.06**	**-0.24**
Immune system process	1	0	0.10	−0.31	0	0.06	−0.24
Cell proliferation	3	0	0.29	−0.54	0	0.17	−0.41
Death	4	1	0.38	1.00	0	0.22	−0.47
Locomotion	4	1	0.38	1.00	0	0.22	−0.47
Biological adhesion	6	1	0.57	0.56	0	0.33	−0.58
Growth	6	1	0.57	0.56	1	0.33	1.15
Nitrogen utilization	9	1	0.86	0.15	0	0.50	−0.71
Reproduction	14	2	1.34	0.57	2	0.78	1.38
Carbon utilization	**15**	**4**	**1.43**	**2.14**	**1**	**0.83**	**0.18**
Multi-organism process	43	4	4.11	−0.05	0	2.39	−1.55
Cell wall organization or biogenesis	49	5	4.68	0.15	2	2.73	−0.44
Signaling	153	14	14.61	−0.16	7	8.51	−0.52
Cellular component organization	177	19	16.91	0.51	11	9.85	0.37
Multicellular organismal process	228	19	21.78	−0.59	11	12.68	−0.47
Cellular component biogenesis	232	31	22.16	1.88	17	12.90	1.14
Developmental process	257	21	24.55	−0.72	12	14.30	−0.61
Response to stimulus	262	16	25.02	−1.80	22	14.57	1.95
Biological regulation	469	47	44.79	0.33	31	26.09	0.96
Localization	706	62	67.43	−0.66	33	39.27	−1.00
Cellular process	2489	240	237.73	0.15	149	138.45	0.90
Metabolic process	3070	293	293.22	−0.01	157	170.76	−1.05
Total	8198	783	783	—	456	456	_

[a] Number of annotations generated per category for the 12,469 genes predicted in *L. maculans*.
[b] Observed number of annotations generated for all predicted genes on chromosome 5 and 10.
[c] Expected number of annotations [= (∑annot.) x P(functional category)]. Where (∑annot.) is the sum of all generated annotations, and P(functional category) is whole genome probability of the considered functional category. Values were approximated to two decimals.
[d] Standardized residuals. We considered a normal distribution of SRs because we cannot reject the null hypothesis as revealed by the Kolmogorov-Smirnov test (Chromosome 5: *P*-value = 0.453, α = 0.05. Chromosome 10: *P*-value = 0.188, α = 0.05). Biased categories are indicated in bold.

deposition of melanin and other pigments is generally associated with spore production in fungi in which they seem to function in the protection of microbes against environmental stress such as UV light and heat (Will *et al.* 1987; Brags *et al.* 2006; Rangel *et al.* 2006), consistent with the overrepresentation of the "pigmentation" category. Second, "growth" fits the isotropic growth step of conidial germination. Third, "cell wall organization or biogenesis" fits the polarized growth step of conidial germination.

T-DNA intranuclear targeting is assumed to result from a long evolution of *Agrobacterium* species' transfection mechanisms to fit host cellular machinery. Starting from this postulate, we analyzed this phenomenon mainly from the host point of view. Consequently, we considered T-DNA insertion biases not only as resulting from T-DNA characteristics but also largely depending on the following: (i) host genome characteristics (in *L. maculans* GC isochores/gene-rich *vs.* AT isochores/gene-poor compartments); (ii) gene expression at both transcriptional (machinery) and functional (cell physiological state) levels; and (iii) DNA features characterized by heterogeneity, unequal sensitivity to DNA damages, and organization in a gene-dependent fashion.

Locus biases, which are due to a chromatin-targeting process that guides T-DNA from its entry into the nucleus to its anchorage to host chromatin, were investigated. Previous studies have shown that T-DNA nuclear import is mediated by two bacterial virulence (Vir) proteins, VirD2 and VirE2, which directly associates with T-DNA to form the transport (T) complex [for review, see Tzfira *et al.* (2000) and Zupan *et al.* (2000)]. In addition to T-DNA encapsidation, Vir proteins act as interfaces with host machinery: VirD2 is phosphorylated *in vivo* by CAK2M, a cyclin-dependent kinase-activating kinase, and is tightly associated with TATA box-binding protein (TBP) (Bakó *et al.* 2003), and VirE2 binds VIP1 (VirE2-interacting protein 1), a bZIP transcription factor capable of binding core histones (Li *et al.* 2005; Loyter *et al.* 2005). VIP1, CAK2M, and TBP profile T-DNA insertion loci distribution within the host genome, and their proper functions and properties generate locus biases: CAK2M phosphorylates the C-terminal domain of the RNA polymerase II (RNA Pol II) largest subunit (Bakó *et al.* 2003), which serves as a TBP-binding platform (Yuryev *et al.* 1996); TBP binds the TATA box core promoter element, whose recognition nucleates the assembly of transcription preinitiation complex [for review, see Smale and Kadonaga (2003)]; and VIP1 precisely binds a DNA hexamer motif found in the promoters of various stress-responsive genes and plays a role in immunity signaling by stimulating stress-dependent gene expression, at least in plants (Djamei *et al.* 2007; Pitzschke *et al.* 2009). However, because VIP1 shows no significant homology to known animal or fungal proteins, it could be plant specific. Nevertheless, because T-complex anchorage to host chromatin seems to be a key step for further T-DNA integration, it is consistent to consider the existence of animal and fungal VIP1-like proteins interacting with both T-complex and host chromatin. The T-DNA insertion pattern in *L. maculans* corroborates this tight relation between T-DNA and gene transcription machinery, because T-DNA insertions were predominant in the following: (i) GC isochores, which are gene-rich islands frequently targeted by gene transcription machinery and therefore more likely to be in a relaxed, opened state, rather than AT isochores, which are TE-rich, gene-poor regions assimilated to heterochromatin, therefore condensed and closed (Rouxel *et al.* 2011); (ii) gene-regulatory regions, which border zones between a histone-containing region

capable of anchoring the T-complex and a histone-less region that is the gene-expressing DNA; and (iii) promoter region, in which additional binding opportunities (CAK2M, TBP) increase the probability of T-complex anchorage and strengthen it.

T-complex anchorage to host chromatin is not synonymous with T-DNA insertion. For the latter to occur, DNA damage is mandatory, because the T-DNA integration process abuses HR and NHR pathways, two host-DNA double-strand break (DSB) repair machineries [for review, see Tzfira *et al.* (2004) and Citovsky *et al.* (2007)]. Consequently, as additional factors that affect occurrence of T-DNA integration events, we must consider both DSBs hotspots and DSB repair efficiency. In eukaryotic cells, DSBs are common events resulting from both environmental and endogenous factors. DSBs are also created by converting single-strand lesions (Natarajan 1993) and retrotransposon activity (Gasior *et al.* 2006), and they occur preferentially in opened chromatin (Berchowitz *et al.* 2009) and transcriptionally active promoters, telomeres, and centromeres (Wu and Lichten 1994; Baudat and Nicolas 1997; Blitzblau *et al.* 2007; Buhler *et al.* 2007). However, not all occurring DSBs are repaired with the same efficiency. In fact, telomeric regions and packed heterochromatin are deficient in repair of DSBs (Ricchetti *et al.* 2003). Together, these studies highlight biases of DSB occurrence and repair that correlate with T-DNA mapping biases in *L. maculans*. Actually, T-DNA integration favored gene-rich GC isochores and not AT isochores that exhibit heterochromatin characteristics in which chromatin is packed and DSB repair is likely to be deficient, and T-DNA integration events overmapped to gene promoter regions where DSBs are assumed to occur frequently.

To corroborate mapping biases in T-DNA integration events, we analyzed T-DNA LB, T-DNA preinsertion and insertion site sequences, and T-DNA–targeted genes for particular compositional, structural, and functional signatures, and we showed that T-DNA LB shares microhomologies with preinsertion sites, suggesting that T-DNA integration may occur at least by HR in *L. maculans*. The same was observed by Meng *et al.* (2007) and by Choi *et al.* (2007) in the *M. oryzae* genome, but the authors did not reach conclusive evidence regarding the targeted motifs. Hence, T-DNA LB sequence affinity with host DNA may affect T-DNA integration event distribution. Our results highlighted that T-DNA LB harbored microhomologies with CAAT box, Inr, and TATA box of eukaryotic promoters. Also, TATA-containing micromology motifs were frequently shared between T-DNA LB and its target sequence. These observations correlate with frequent mapping of T-DNA insertions to gene promoter regions, in the sense that sequence affinity is mandatory for DNA end joining by MMEJ [for review, see McVey and Lee (2008)].

DNA asymmetry was observed in both prokaryotic and eukaryotic genomes. It is a consequence of many mechanisms, among which gene expression is one of the better studied, and in which DNA asymmetry is seen as signatures indicating functional signals and DNA modifications (Touchon and Rocha 2008). DNA asymmetry is revealed by the CG skew and AT skew. Previous studies noticed that the CG skew is stronger than the AT skew, at least in eubacteria (Francino and Ochman 1997; Frank and Lobry 1999; Tillier and Collins 2000) and that skew curves are associated with replication origin (Lobry, 1996a,b; Blattner *et al.* 1997; Kunst *et al.* 1997; Grigoriev 1998) and transcription-coupled and splicing-coupled mutations (Touchon *et al.* 2004). In particular, CG skew peak is associated with gene expression level, at least in plants (Alexandrov *et al.* 2006), and transcription initiation starts in eukaryotes, including fungi (Tatarinova *et al.* 2003; Fujimori *et al.* 2005; Alexandrov *et al.* 2006). Altogether, these studies highlight a correlation between DNA asymmetry and cellular activity–driven DNA manipulations and modifications in general (replication, gene expression, mutations) and a tight association between an increased CG skew and gene transcription in particular. T-DNA targeted to CG asymmetric DNA is thus consistent with frequent insertions in transcriptionally active regions and gene promoter sequences.

CONCLUSIONS

Using a fungal genome showing contrasted genomic landscapes, our data substantiate the advantages of ATMT to reach a functional annotation of genes, but they cast doubts on whether this strategy will be able to target species-specific genes involved in pathogenicity that reside in specific AT-rich compartments of the fungal genome. The main particularities substantiating these points are (i) the common single-copy integration of the T-DNA; (ii) the high frequency of integration within protein-coding genes (even if the introns are favored targets), amounting to one third of the integration events in *L. maculans*; and (iii) the common occurrence in promoters favoring the access to genes whose complete inactivation would be too detrimental for the fungus. In contrast, under the hypothesis that genes specifically involved in plant pathogenicity are hosted in specific compartments of the genome, we notice that T-DNA targeting to AT isochores is very low compared with the percentage of these landscapes in the genome, but it is nevertheless consistent with the amount of genes hosted in AT isochores. More importantly, our work suggests the importance of the physiology of the fungus at the time of ATMT and the favored targeting of transcriptionally active regions of the genome. With most of the genes involved in pathogenicity, such as those encoding effector proteins, repressed during the vegetative growth of the fungus and overexpressed at the onset of plant infection (Rouxel *et al.* 2011), these are unlikely to be targeted by the T-DNA.

ACKNOWLEDGMENT

S.B. was funded by a grant from the French Ministry of Research.

LITERATURE CITED

Alexandrov, N. N., M. E. Troukhan, V. V. Brover, T. Tatarinova, R. B. Flavell *et al.*, 2006 Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant Mol. Biol. 60: 69–85.

Alonso, J. M., A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen *et al.*, 2003 Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301: 653–657.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Attard, A., L. Gout, M. Gourgues, M. L. Kuhn, J. Schmit *et al.*, 2002 Analysis of molecular markers genetically linked to the *Leptosphaeria maculans* avirulence gene *AvrLm1* in field populations indicates a highly conserved event leading to virulence on *Rlm1* genotypes. Mol. Plant-Microbe Interact. 15: 672–682.

Bakó, L., M. Umeda, A. F. Tiburcio, J. Schell, and C. Koncz, 2003 The VirD2 pilot protein of *Agrobacterium*-transferred DNA interacts with the TATA box-binding protein and a nuclear protein kinase in plants. Proc. Natl. Acad. Sci. USA 100: 10108–10113.

Balzergue, S., B. Dubreucq, S. Chauvin, I. Le-Clainche, F. Le Boulaire *et al.*, 2001 Improved PCR-walking for large-scale isolation of plant T-DNA borders. Biotechniques 30: 496–504.

Baudat, F., and A. Nicolas, 1997 Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA 94: 5213–5218.

Berchowitz, L. E., S. E. Hanlon, J. D. Lieb, and G. P. Copenhaver, 2009 A positive but complex association between meiotic double-strand break hotspots and open chromatin in *Saccharomyces cerevisiae*. Genome Res. 19: 2245–2257.

Blaise, F., E. Remy, M. Meyer, L. Zhou, J. P. Narcy *et al.*, 2007 A critical assessment of *Agrobacterium tumefaciens*-mediated transformation as a tool for pathogenicity gene discovery in the phytopathogenic fungus *Leptosphaeria maculans*. Fungal Genet. Biol. 44: 123–138.

Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1462.

Blitzblau, H. G., G. W. Bell, J. Rodriguez, S. P. Bell, and A. Hochwagen, 2007 Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. Curr. Biol. 17: 2003–2012.

Brags, G. U. L., D. E. Rangel, S. D. Flint, A. J. Anderson, and D. W. Roberts, 2006 Conidial pigmentation is important to tolerance against solar-simulated radiation in the entomopathogenic fungus *Metarhizium anisopliae*. Photochem. Photobiol. 82: 418–422.

Breathnach, R., and P. Chambon, 1981 Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. 50: 349–383.

Bucher, P., 1990 Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212: 563–578.

Buhler, C., V. Borde, and M. Lichten, 2007 Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in *Saccharomyces cerevisiae*. PLoS Biol. 5: e324.

Burley, S. K., 1996 The TATA box binding protein. Curr. Opin. Struct. Biol. 6: 69–75.

Choi, J., J. Park, J. Jeon, M. H. Chi, J. Goh *et al.*, 2007 Genome-wide analysis of T-DNA integration into the chromosomes of *Magnaporthe oryzae*. Mol. Microbiol. 66: 371–382.

Christie, K. R., E. L. Hong, and J. M. Cherry, 2009 Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. Trends Microbiol. 17: 286–294.

Citovsky, V., S. V. Kozlovsky, B. Lacroix, A. Zaltsman, M. Dafny-Yelin *et al.*, 2007 Biological systems of the host cell involved in *Agrobacterium* infection. Cell. Microbiol. 9: 9–20.

D'Enfert, C., 1997 Fungal spore germination: insights from the molecular genetics of *Aspergillus nidulans* and *Neurospora crassa*. Fungal Genet. Biol. 21: 163–172.

Djamei, A., A. Pitzschke, H. Nakagami, I. Rajh, and H. Hirt, 2007 Trojan horse strategy in *Agrobacterium* transformation: abusing MAPK defense signaling. Science 318: 453–456.

Elliott, C. E., and B. J. Howlett, 2006 Overexpression of a 3-ketoacyl-CoA thiolase in *Leptosphaeria maculans* causes reduced pathogenicity on *Brassica napus*. Mol. Plant-Microbe Interact. 19: 588–596.

Francino, M. P., and H. Ochman, 1997 Strand asymmetries in DNA evolution. Trends Genet. 13: 240–245.

Frank, A. C., and J. R. Lobry, 1999 Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65–77.

Fujimori, S., T. Washio, and M. Tomita, 2005 GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 6: 26.

Gasior, S. L., T. P. Wakeman, B. Xu, and P. L. Deininger, 2006 The human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol. 357: 1383–1393.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science 274: 546–567.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the blast2GO suite. Nucleic Acids Res. 36: 3420–3435.

Grigoriev, A., 1998 Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26: 2286–2290.

Grigoriev, I. V., D. Cullen, S. B. Goodwin, D. Hibbett, T. W. Jeffries, *et al.*, 2011 Fueling the future with fungal genomics. Mycology 2: 192–209.

Grubbs, F. E., 1969 Procedures for detecting outlying observations in samples. Technometrics 11: 1–21.

Huh, W. K., J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson *et al.*, 2003 Global analysis of protein localization in budding yeast. Nature 425: 686–691.

Javahery, R., A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale, 1994 DNA sequence requirements for transcriptional initiator activity in mammalian cells. Mol. Cell. Biol. 14: 116–127.

Jones, G. M., J. Stalker, S. Humphray, A. West, T. Cox *et al.*, 2008 A systematic library for comprehensive overexpression screens in *Saccharomyces cerevisiae*. Nat. Methods 5: 239–241.

Kourmpetis, Y. A., A. D. van Dijk, R. C. van Ham, and C. J. ter Braak, 2011 Genome-wide computational function prediction of *Arabidopsis* proteins by integration of multiple data sources. Plant Physiol. 155: 271–281.

Krishnan, A., E. Guiderdoni, G. An, Y. I. Hsing, C. D. Han *et al.*, 2009 Mutant resources in rice for functional genomics of the grasses. Plant Physiol. 149: 165–170.

Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni *et al.*, 1997 The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.

Li, J., A. Krichevsky, M. Vaidya, T. Tzfira, and V. Citovsky, 2005 Uncoupling of the functions of the *Arabidopsis* VIP1 protein in transient and stable plant genetic transformation by *Agrobacterium*. Proc. Natl. Acad. Sci. USA 102: 5733–5738.

Liu, Y. G., N. Mitsukawa, T. Oosumi, and R. F. Whittier, 1995 Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J. 8: 457–463.

Lobry, J. R., 1996a Origin of replication of *Mycoplasma genitalium*. Science 272: 745–746.

Lobry, J. R., 1996b Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13: 660–665.

Loyter, A., J. Rosenbluh, N. Zakai, J. Li, S. V. Kozlovsky *et al.*, 2005 The plant VirE2 interacting protein 1. A molecular link between the *Agrobacterium* T-complex and the host cell chromatin? Plant Physiol. 138: 1318–1321.

Ma, L. J., H. C. van der Does, K. A. Borkovich, J. J. Coleman, M. J. Daboussi *et al.*, 2010 Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464: 367–373.

McVey, M., and S. E. Lee, 2008 MMEJ repair of double-strand breaks director's cut: deleted sequences and alternative endings. Trends Genet. 24: 529–538.

Meng, Y., G. Patel, M. Heist, M. F. Betts, S. L. Tucker *et al.*, 2007 A systematic analysis of T-DNA insertion events in *Magnaporthe oryzae*. Fungal Genet. Biol. 44: 1050–1064.

Michielse, C. B., P. J. Hooykaas, C. van den Hondel, and A. F. Ram, 2005 *Agrobacterium*-mediated transformation as a tool for functional genomics in fungi. Curr. Genet. 48: 1–17.

Natarajan, A. T., 1993 Mechanisms for induction of mutations and chromosome alterations. Environ. Health Perspect. 101: 225–229.

Osherov, N., and G. S. May, 2001 The molecular mechanisms of conidial germination. FEMS Microbiol. Lett. 199: 153–160.

Pena-Castillo, L., and T. R. Hughes, 2007 Why are there still over 1000 uncharacterized yeast genes? Genetics 176: 7–14.

Pitzschke, A., A. Djamei, M. Teige, and H. Hirt, 2009 VIP1 response elements mediate mitogen-activated protein kinase 3-induced stress gene expression. Proc. Natl. Acad. Sci. USA 106: 18414–18419.

Rangel, D. E., M. J. Butler, J. Torabinejad, A. J. Anderson, G. U. Braga *et al.*, 2006 Mutants and isolates of *Metarhizium anisopliae* are diverse in their relationships between conidial pigmentation and stress tolerance. J. Invertebr. Pathol. 93: 170–182.

Remy, E., M. Meyer, F. Blaise, M. Chabirand, N. Wolff *et al.*, 2008a The *Lmpma1* gene of *Leposphaeria maculans* encodes a plasma membrane H^+-ATPase isoform essential for pathogenicity towards oilseed rape. Fungal Genet. Biol. 45: 1122–1134.

Remy, E., M. Meyer, F. Blaise, U. K. Simon, D. Kuhn *et al.*, 2008b The *Lmgpi15* gene, encoding a component of the glycosylphosphatidylinositol anchor biosynthesis pathway, is required for morphogenesis and pathogenicity in *Leptosphaeria maculans*. New Phytol. 179: 1105–1120.

Remy, E., M. Meyer, F. Blaise, U. K. Simon, D. Kuhn *et al.*, 2009 A key enzyme of the Leloir pathway is involved in pathogenicity of *Leptosphaeria maculans* towards oilseed rape. Mol. Plant-Microbe Interact. 22: 725–736.

Ricchetti, M., B. Dujon, and C. Fairhead, 2003 Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast. J. Mol. Biol. 328: 847–862.
Rouxel, T., J. Grandaubert, J. K. Hane, C. Hoede, A. P. van de Wouw *et al.*, 2011 Diversification of effectors within compartments of the *Leptosphaeria maculans* genome affected by RIP mutations. Nat. Commun. 2: 202.
Smale, S. T., and J. T. Kadonaga, 2003 The RNA polymerase II core promoter. Annu. Rev. Biochem. 72: 449–479.
Spanu, P. D., J. C. Abbott, J. Amselem, T. A. Burgis, D. M. Soanes *et al.*, 2010 Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science 330: 1543–1546.
Tatarinova, T., V. Brover, M. Troukhan, and N. Alexandrov, 2003 Skew in CG content near the transcription start site in *Arabidopsis thaliana*. Bioinformatics 19: i313–i314.
Thole, V., B. Worland, J. Wright, M. W. Bevan, and P. Vain, 2010 Distribution and characterization of more than 1000 T-DNA tags in the genome of *Brachypodium distachyon* community standard line Bd21. Plant Biotechnol. J. 8: 734–747.
Tillier, E. R., and R. A. Collins, 2000 The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. 50: 249–257.
Touchon, M., and E. P. C. Rocha, 2008 A gentle guide to the analysis of strand asymmetry in genome sequences. Biochimie 90: 648–659.
Touchon, M., A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, 2004 Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res. 32: 4969–4978.
Tzfira, T., J. Li, B. Lacroix, and V. Citovsky, 2000 Nucleic acid transport in plant-microbe interactions: the molecules that walk through the walls. Annu. Rev. Microbiol. 54: 187–219.
Tzfira, T., Y. Rhee, M.-H. Chen, and V. Citovsky, 2004 *Agrobacterium* T-DNA integration: molecules and models. Trends Genet. 20: 375–383.
Will, O. H., D. Dixon, A. Birney, and P. L. Thomas, 1987 Effects of far UV and visible light on germination of wild type and albino teliospores of *Ustilago nuda*. Can. J. Plant Pathol. 9: 225–229.
Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson *et al.*, 1999 Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285: 901–906.
Wu, T. C., and M. Lichten, 1994 Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science 263: 515–518.
Yuryev, A., M. Patturajan, Y. Litingtung, R. V. Joshi, C. Gentile *et al.*, 1996 The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. Proc. Natl. Acad. Sci. USA 93: 6975–6980.
Zhang, J., D. Gu, Y. X. Chang, C. J. You, X. W. Li *et al.*, 2007 Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library. Plant J. 49: 947–959.
Zupan, J., T. R. Muth, O. Draper, and P. C. Zambryski, 2000 The transfer of DNA from *Agrobacterium tumefaciens* into plants: a feast of fundamental insights. Plant J. 23: 11–28.

Communicating editor: B. J. Andrews

Article 3. The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*.

MH Balesdent, I Fudal, B Ollivier, P Bally, J Grandaubert, F Eber, AM Chèvre, M Leflon & T Rouxel.

Publié le 13 février 2013 dans *New Phytologist* **198:**887-898.

New Phytologist

Research

The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*

Marie-Hélène Balesdent[1], Isabelle Fudal[1], Bénédicte Ollivier[1], Pascal Bally[1], Jonathan Grandaubert[1], Frédérique Eber[2], Anne-Marie Chèvre[2], Martine Leflon[3] and Thierry Rouxel[1]

[1]INRA, UR1290 BIOGER, Avenue Lucien Brétignières, BP 01, F-78850 Thiverval-Grignon, France; [2]INRA, UMR1349 IGEPP, BP35327, F-35653 Le Rheu Cedex, France; [3]CETIOM, Avenue Lucien Brétignières, BP 01, F-78850 Thiverval-Grignon, France

Author for correspondence:
Marie-Hélène Balesdent
Tel: +33 1 30 81 45 73
Email: mhb@versailles.inra.fr

Received: *10 October 2012*
Accepted: *10 January 2013*

New Phytologist (2013)
doi: 10.1111/nph.12178

Key words: avirulence gene, *Brassica napus*, *Brassica rapa*, dispensable or B chromosomes, effector, fitness, *Leptosphaeria maculans*.

Summary

- Phytopathogenic fungi frequently contain dispensable chromosomes, some of which contribute to host range or pathogenicity. In *Leptosphaeria maculans*, the stem canker agent of oilseed rape (*Brassica napus*), the minichromosome was previously suggested to be dispensable, without evidence for any role in pathogenicity.
- Using genetic and genomic approaches, we investigated the inheritance and molecular determinant of an *L. maculans*–*Brassica rapa* incompatible interaction.
- Single gene control of the resistance was found, while all markers located on the *L. maculans* minichromosome, absent in the virulent parental isolate, co-segregated with the avirulent phenotype. Only one candidate avirulence gene was identified on the minichromosome, validated by complementation experiments and termed *AvrLm11*. The minichromosome was frequently lost following meiosis, but the frequency of isolates lacking it remained stable in field populations sampled at a 10-yr time interval, despite a yearly sexual stage in the *L. maculans* life cycle.
- This work led to the cloning of a new 'lost in the middle of nowhere' avirulence gene of *L. maculans*, interacting with a *B. rapa* resistance gene termed *Rlm11* and introgressed into *B. napus*. It demonstrated the dispensability of the *L. maculans* minichromosome and suggested that its loss generates a fitness deficit.

Introduction

In addition to the normal complement of essential chromosomes, many animal, plant, fungal, and oomycete species contain B chromosomes (also known as supernumerary or dispensable chromosomes) which are inherited in a non-Mendelian manner (Jones, 1995; Covert, 1998). These chromosomes are not essential for life and are lacking in some individuals. Most plant B chromosomes are enriched in transposable elements. As a consequence, they are mainly or entirely heterochromatic, and largely noncoding. Persistence in a species is thus dependent on either higher transmission rates (Muñoz-Pajares *et al.*, 2011) or the presence of the few coding sequences that present a selective advantage, at least under some growth conditions. According to Covert (1998), 'supernumerary chromosomes that confer an adaptive advantage in certain habitats, such as the ability to cause disease on a specific host, may be referred to as 'conditionally dispensable' chromosomes (CDC) in order to reflect their importance in some, but not all, growth conditions.'

In fungi, B chromosomes or CDCs have been described in species belonging to the classes Sordariomycetes and Dothideomycetes, including the plant pathogens *Magnaporthe oryzae* (Chuma *et al.*, 2003, 2011), *Fusarium oxysporum* (Ma *et al.*, 2010), *Nectria haematococca* (Coleman *et al.*, 2009), *Alternaria alternata* (Hatta *et al.*, 2002), *Alternaria arborescens* (Hu *et al.*, 2012), *Cochliobolus heterostrophus* (Tzeng *et al.*, 1992) and *Mycosphaerella graminicola* (Wittenberg *et al.*, 2009) (for additional examples: Covert, 1998). CDCs are not required for saprophytic growth, but their contribution to other stages of the fungal life cycle seems to be variable from one species to another. The most commonly reported selective advantage lies in host range delineation: the presence in the CDC of genes encoding host-selective toxins (HSTs) or effectors allows the pathogen to infect new plant species (Mehrabi *et al.*, 2011) while, in some cases, lack of the CDC results in isolates with a saprophytic behaviour (Johnson *et al.*, 2001). For instance, in *A. alternata* pathotypes, CDCs carry clusters of genes encoding secondary metabolites acting as HSTs responsible for pathogenicity on distinct host plants, some of them being duplicated in the essential genome (Mehrabi *et al.*, 2011). In *N. haematococca*, at least three B chromosomes have been identified, one of which is dispensable for *in vitro* growth and carries genes involved in pathogenicity to pea (*Pisum sativum*), such as genes involved in phytoalexin detoxification (Mehrabi *et al.*, 2011). An *F. oxysporum* CDC was shown to harbour the *SIX* (Secreted In Xylem) genes, encoding small secreted proteins (SSPs) acting as effectors and virulence/

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

avirulence factors. In this species the genes harboured by the CDC were postulated to be the main determinants for adaptation to tomato (*Solanum lycopersicum*) (Ma *et al.*, 2010). While their loss has been associated with reduced fertility, *M. oryzae* CDCs are believed to be nonessential for growth and pathogenicity, probably because of the occurrence of duplicated copies of the avirulence genes in the core genome (Chuma *et al.*, 2003, 2011). In the fungal species with the greatest number of dispensable chromosomes reported to date, *M. graminicola*, dispensable chromosomes harbour, among others, genes with inactivated paralogues in the essential genome, while all pathogenicity genes currently described are harboured by the core genome (Wittenberg *et al.*, 2009; Goodwin *et al.*, 2011).

In *Leptosphaeria maculans*, the causal agent of stem canker of oilseed rape (*Brassica napus*), the smallest chromosome, termed the minichromosome (MC), was shown to be inherited in a non-Mendelian manner following *in vitro* crosses and was lacking in the electrokaryotypes of some isolates. Dispensability of its coding sequences, rather than translocation to larger chromosomes, could, however, not be demonstrated, preventing it from being definitively classified as a CDC (Leclair *et al.*, 1996). Whole-genome sequencing of *L. maculans* showed that the MC is extremely enriched in transposable elements (TEs) and contains only a few predicted genes (Rouxel *et al.*, 2011).

In *L. maculans*, genes encoding effectors involved in gene-for-gene interactions (avirulence genes) are characterized by their genome location in large, TE-rich chromosomal landscapes of the genome termed AT-isochores (Gout *et al.*, 2006b; Fudal *et al.*, 2007; Parlange *et al.*, 2009). In addition, most of the putative effector genes (i.e. encoding SSPs) showing increased expression in the first stage of plant infection are located in such genomic landscapes (Rouxel *et al.*, 2011). These features suggested an improved cloning strategy to rapidly identify new candidate avirulence (*Avr*) genes following phenotyping, by combining genetic mapping of the gene and selection of candidate effector genes based on their genome location and induced expression *in planta*.

In this study, we identified a new resistance (R) source in *Brassica rapa*, a plant species closely related to *B. napus*, and investigated the genetic control of the interaction in the plant and the pathogen. The new resistance was shown to be under monogenic, dominant genetic control, while an apparently more complex genetic control was found in the fungus. Cloning of the avirulence determinants showed that the avirulent phenotype was nevertheless caused by a single gene, *AvrLm11*, resident on the *L. maculans* MC. The dispensability and frequent loss of the MC *in vitro* were then demonstrated, while its maintenance in natural populations suggested its involvement in fungal fitness. The *L. maculans* MC thus can be considered as a CDC that may, like those of other fungal phytopathogens, host genes involved in host range definition.

Materials and Methods

Leptosphaeria maculans isolates

Thirty-one isolates from the IBCN and IMASCORE collections of isolates (Eckert *et al.*, 2005), previously phenotyped for the nine *Avr* alleles (*AvrLm1–AvrLm9*; Balesdent *et al.*, 2005) and originating from Europe, Canada, Australia and New Zealand, along with isolate v11.1.1 (Rouxel *et al.*, 2003), were used in the preliminary screening experiments leading to the identification of a new source of resistance in *Brassica rapa* L. Among these, isolate IBCN14, belonging to race Av5-6 (i.e. combining the avirulence alleles at the *AvrLm5* and *AvrLm6* loci only; Balesdent *et al.*, 2005), was found to be virulent on the resistant *B. rapa* line. *AvrLm11* segregation analysis and mapping were carried out following an *in vitro* cross (cross 38) between two isolates of opposite mating-types (Mat), the sequenced isolate v23.1.3 (Av1-4-5-6-7-8, Mat1-2) and IBCN14 (Av5-6, Mat1-1), from which 84 progeny isolates were recovered and termed v38.*x.y*, with *x* referring to the pseudothecia from which the isolate originates and *y* to the isolate number.

Progeny of three available crosses were used to analyse the segregation and loss of the MC: crosses 11, 28 and 37 (Table 1). One isolate per pair of twin genotypes from tetrad Z (Z1, MC^+; Z2, MC^-; Z4, MC^-, Z6, MC^+; Leclair *et al.*, 1996) were also used. Three hundred and ten single-conidia isolates from a wide-scale sampling performed on naturally infected oilseed rape leaves of a single cultivar in 15 locations in France in 2000–2001 (Balesdent *et al.*, 2006) were randomly selected within each site and screened for the occurrence of the MC (Table 2). Ten years later (2010–2011), a comparable wide-scale sampling was performed in seven of these sites using the same protocol and trap cultivar, giving rise to a collection of 497 isolates (Table 2). Procedures for isolate culture and maintenance, sporulation and *in vitro* crosses were as previously established (Gall *et al.*, 1994; Ansan-Melayah *et al.*, 1995). Conidia suspensions were recovered

Table 1 List and characteristics of *Leptosphaeria maculans in vitro* crosses analysed in this study

Cross no.	Parental isolate 1 (Mat 1-1)	Parental isolate 2 (Mat 1-2)	No. of progeny analysed	Minichromosome (MC) segregation
11[a]	a.2 (MC^+)[b]	H5 (MC^+)	85	82 (MC^+):3 (MC^-) (3.5%)
28	v27.10.1 (MC^+)	v23.1.3 (MC^+)	28	27 (MC^+):1 (MC^-) (3.6%)
37	v36.2.13 (MC^+)	v23.1.3 (MC^+)	179	167 (MC^+):12 (MC^-) (6.7%)
38	IBCN14 (MC^-)	v23.1.3 (MC^+)	84	70 (MC^+):14 (MC^-) (16.7%)

[a]Cross 11 is as described by Ansan-Melayah *et al.* (1995); crosses 28 and 37 are BC3 and BC5 described by Huang *et al.* (2006); cross 38, this study.
[b]MC^+, evidence (obtained using pulsed field gel electrophoresis (PFGE) or amplification of Super-Contig 22 specific markers) for the presence of the minichromosome; MC^-, no PFGE MC band or no amplification of any markers specific for Super-Contig 22.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

Table 2 List and characteristics of French *Leptosphaeria maculans* populations characterized for the occurrence of *AvrLm11* and minichromosome-specific markers

		2000–2001 samplings[b]		2010–2011 samplings[c]	
Location	Region[a]	No. of isolates analysed	No. (%) of isolates without *AvrLm11* and MC markers	No. of isolates analysed	No. (%) of isolates without *AvrLm11* and MC markers
Chartrainvilliers	C	20	1	nd[d]	
Chateauroux	C	18	0	nd	
Lutz	C	20	1	nd	
Oucques	C	20	0	80	4
Toury	C	16	1	80	1
Martincourt	E	20	0	nd	
Mons en Pevèle	N	20	1	nd	
Prémesques	N	19	1	80	2
Grignon	NC	35	2	80	3
Verneuil	NC	19	1	80	1
Versailles	NC	24	2	nd	
La Bénate	W	20	1	nd	
Le Rheu	W	22	3	32	0
Les Alleuds	W	19	0	nd	
Mondonville	S	20	0	65	1
Total		312	14 (4.5)	497	12 (2.4)

[a]C, central region; E, eastern region; NC, north central region; W, western region; S, southern region; for more information on the sites, see Balesdent *et al.* (2006).
[b]Isolates were randomly subsampled from the Balesdent *et al.* (2006) samplings.
[c]In 2010–2011, *L. maculans* populations were sampled in seven selected sites, on the same host plant genotype Drakkar and with the same procedure as for the 2000–2001 sampling.
[d]nd, not done.

from 12-d-old V8-juice agar culture in sterile water. Undiluted spore suspensions (for DNA extraction) or suspensions adjusted to 10^7 conidia ml^{-1} (for inoculation tests) were stored at −20°C until used.

Inoculation tests

Isolates were phenotyped for their virulence towards hosts carrying specific resistance genes following a cotyledon-inoculation test (Balesdent *et al.*, 2001; Supporting Information Methods S1). Spore suspensions were inoculated on 10–12 plants of the *B. napus* lines Westar (no *R* gene), 01-23-2-1 (*Rlm7*), Jet Neuf or Pixel (*Rlm4*), Columbus (*Rlm1*, *Rlm3*), 00-156-2-1 (*Rlm8*) (Balesdent *et al.*, 2002) and 02-159-4-1 (*B. rapa* resistant line; this work). IBCN14 was also inoculated on Surpass400 (*Rlm1*, *RlmS*; van de Wouw *et al.*, 2008) and a *B. napus* line carrying *Rlm10* (Eber *et al.*, 2011). Symptoms were scored 12–21 d after inoculation using a 1 (avirulent) to 6 (virulent) rating scale (Methods S1). Isolates were classified as avirulent or virulent for a given locus whenever the mean disease rating on the corresponding differential line was ≤ 3 or > 3, respectively (Balesdent *et al.*, 2001).

Plant multiplication and crossing

Brassica napus is a natural amphidiploid species (AC genome, $2n = 38$) derived from hybridization between *B. rapa* (A genome, $2n = 20$) and *Brassica oleracea* (C genome, $2n = 18$; U, 1935). One plant (plant 2323) from the *B. rapa* (AA) resistant accession 02-159-4-1 was crossed with a susceptible *B. rapa* doubled haploid line, Z1 (kindly provided by Agriculture and Agri-Food Canada, Saskatoon, Canada), and with two *B. napus* (AACC) winter oilseed rape varieties, Darmor and Eurol (provided by the Genetic Resource Center, BrACySol, UMR IGEPP, Ploudaniel, France). More than 130 seeds were produced from AA F1 resistant hybrids crossed to Z1 (backcross BC1). Simultaneously, in the progeny of AAC resistant F1 hybrids crossed to the recurrent *B. napus* varieties (BC1), >110 plants were analysed and resistant plants with $2n = 38$ were selected for two additional backcrosses (BC2 and BC3). Selfing progenies of two BC3 plants per origin were screened to produce homozygous plants.

Plant cytogenetic analyses

Flow cytometry was used at the seedling stage to assess the plant chromosome number of each F1 hybrid as described by Leflon *et al.* (2006). For the establishment of meiotic behaviour, metaphase I (MI) was analysed for 20 to 30 pollen mother cells of young floral buds. They were fixed in Carnoy's solution (alcohol : chloroform : acetic acid, 6 : 3 : 1) for 24 h at room temperature and stored in 50% ethanol at 4°C. Anthers were squashed and stained in a drop of 1% acetocarmine solution.

DNA and RNA extractions

Fungal DNA was extracted from 1 ml of undiluted conidial suspension, centrifuged at 6000 ***g*** for 10 min. Spores were ground with carbide beads using a Retsch MM380 grinder and DNA was extracted using the BioRobot3000 and the DNeasy 96 Plant

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

Kit (Qiagen S.A., Courtaboeuf, France) according to the manufacturer's recommendations. RNA from *B. napus* cotyledons inoculated with *L. maculans* was extracted 7 d post inoculation as described by Fudal *et al.* (2007).

Genetic mapping

The FONZIE pipeline (Bally *et al.*, 2010) was used to identify single-copy minisatellite (MS) markers and to design primers, departing from the whole-genome sequence (WGS) of isolate v23.1.3. Polymorphic MS markers were amplified with standard procedures (Gout *et al.*, 2006a) and separated in 2% agarose gels. The size of MS bands in progeny isolates was compared with those of the parental isolates used as controls. Linkage analyses among MS or *Avr* loci was performed using the MAPMAKER/EXP 3.0 software (available online; Whitehead/MIT Center for Genome Research, Cambridge, MA, USA) with an LOD (logarithm of odds) score of 4.0 and a maximum recombination frequency of 20 cM.

Gene annotation

Annotation of untranscribed regions (UTRs), transcriptional start and stop sites and intron positions was performed following PCR amplification and sequencing of the 3′ and 5′ ends of cDNA using the Creator SMART cDNA Library Construction Kit (Clontech, Palo Alto, CA, USA) according to the manufacturer's recommendations and using *AvrLm11*-5UTRU, *AvrLm11*-5UTRnestU, *AvrLm11*-3UTRU and *AvrLm11*-3UTRnestU as specific primers (Table S1).

Cloning and transformation

The binary vector pPZPNat1, which carries the nourseothricin acetyltransferase 1 (*nat1*) gene (conferring resistance to the antibiotic nourseothricin) under control of the *Aspergillus nidulans* indole-3-glycerolphosphate synthase (*trpC*) promoter as a selectable marker, was used to introduce the candidate *AvrLm11* gene into *L. maculans* isolate IBCN14 using *Agrobacterium tumefaciens*-mediated transformation (ATMT). *AvrLm11* was amplified from genomic DNA of v23.1.3 using primers AvrLm11U and AvrLm11-XhoL (which introduces a *Xho*I restriction site; Table S1). The 1485-bp fragment obtained was digested by *Nhe*I and *Xho*I and ligated into pPZPNat1 previously digested by *Spe*I and *Xho*I. The construct, termed pPZPNat1-AvrLm11, was cloned into *Escherichia coli*, re-extracted and sequenced. Finally, the construct was introduced into *A. tumefaciens* strain C58 by electroporation at 1.5 kV, 200 ohms and 25 μF and used for transformation of the virulent isolate IBCN14. Transformation of *L. maculans* was performed as described by Gout *et al.* (2006b). Fungal transformants were selected on 50 μg ml^{-1} nourseothricin (Werner BioAgents, Jena, Germany).

Gene expression

Expression of genes located in super-contig (SC) 22 was analysed using whole-genome oligoarray data (Rouxel *et al.*, 2011) corresponding to hybridization using four different conditions (mycelium grown for 7 d in Fries liquid medium, and oilseed rape infected cotyledons 3, 7 and 14 d post inoculation). Background correction, normalization and calculation of average expression levels were performed as described by Rouxel *et al.* (2011). In summary, gene models with an expression higher than three times the median of random probe intensities in at least two of three biological replicates were considered as transcribed. Transcripts showing a 1.5-fold change in transcript level between *in vitro* and *in planta* conditions with a P-value < 0.05 were considered to be significantly differentially expressed during infection compared with mycelial growth.

Multiplex PCR assay

A multiplex PCR assay was used to detect the presence of *AvrLm11* in all field or progeny isolates. This assay aimed to amplify the mating type gene (Cozijnsen & Howlett, 2003) and *AvrLm11*. Using this assay, *AvrLm11* was amplified along with one of the two mating-type alleles, used as internal controls of PCR amplification for isolates for which no *AvrLm11* amplification could be obtained. PCR amplifications were performed in a total volume of 15 μl containing 0.2 μM of each deoxynucleotide triphosphate, 0.67 μM of each of the five primers MATU, MATL1, MATL2, AvrLm11_U2 and AvrLm11_L (Table S1), 0.6 U of *Taq* DNA polymerase (Qbiogen, Illkirch, France), 1.5 μl of the $MgCl_2$-containing 10X reaction buffer supplied with the enzyme, and 1 μl of genomic DNA (between 10 and 30 ng of DNA). PCR amplifications were performed in an Eppendorf Mastercycler EP Gradient thermocycler (Eppendorf, Le Pecq, France), with 30 cycles of 94°C for 30 s, 59°C for 30 s, and 72°C for 60 s, with a final extension at 72°C for 5 min.

Data analyses

Segregations of phenotypes or genotypes were compared with expected segregation ratios using χ^2 tests. Nonparametric tests were used to compare frequencies of genes between different populations of *L. maculans* (Wilcoxon tests for paired observations or Mann–Whitney tests for unpaired observations) or to compare genomic features between different genomic compartments (Kruskal–Wallis test and Steel–Dwass–Critchlow–Fligner multiple comparisons). Exact binomial tests were performed to test the enrichment of the MC in SSPs. All statistical analyses were performed using XLSTAT v2010.5.01 (Addinsoft, Paris, France) or R scripts.

Results

Selection of a resistant *B. rapa* line

In the course of screening a germplasm collection for resistance to European *L. maculans* isolates (Rouxel *et al.*, 2003), one plant out of 10 from the *B. rapa oleifera* accession IPK cr 1564/96 was found to be resistant to four isolates representative of the main European *L. maculans* race, race Av5-6-7-8 (v11.1.1, UK1, PT1

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

and Raw4; Balesdent *et al.*, 2005; Stachowiak *et al.*, 2006). The plant was selfed and the resulting selfing (99-159-1-1) was tested for resistance to v11.1.1, UK1 and 27 isolates from the IBCN collection. The line was fully resistant to 23 isolates, fully susceptible to two isolates including IBCN14, and displayed a heterogeneous resistance (i.e. plants being either resistant or susceptible) to four isolates. Isolates showing either virulent or heterogeneous interaction phenotypes were all characterized by a virulence allele at the *AvrLm7* locus, suggesting that *Rlm7* could have been heterozygous in the selected *B. rapa* plant, as previously described in another *B. rapa* accession (Leflon *et al.*, 2007), in addition to a new R source. Two additional selfings were performed starting from plants selected in line 99-159-1-1 for their susceptibility to both IBCN14 and v38.2.5 (an *AvrLm7* progeny isolate from IBCN14), to ensure the production of a *B. rapa* line (02-159-4-1) with a new fixed resistance but lacking *Rlm7*.

Genetic control of resistance and introgression into *B. napus*

One plant (2323) from line 02-159-4-1 was both selfed and crossed with the susceptible *B. rapa* line Z1. In the selfing (11 plants) and the 2323 × Z1 F1 (23 plants), all plants were susceptible to IBCN14 (Av5-6) and v38.2.5 (Av5-6-7), and resistant to the reference isolate v23.1.2 (Av5-6-7-8), consistent with the hypothesis of a dominant and homozygous resistance gene in plant 2323. One F1 plant was back-crossed to the susceptible *B. rapa* line. In the resulting BC1, all plants were susceptible to IBCN14, while a 66 : 69 resistant : susceptible (R : S) segregation to v23.1.2 was observed, consistent with the expected 50 : 50 R : S segregation under the hypothesis of a single dominant, homozygous resistance gene in plant 2323.

The same plant (2323) was crossed with two *B. napus* varieties, Darmor and Eurol. More than 100 plants of the AAC F1 hybrid progeny crossed with their recurrent parent were analysed for resistance, giving an R : S segregation ratio close to 50 : 50, confirming the presence of a major resistance gene (Tables S2, S3). The chromosome number of 47 and 34 resistant plants from Darmor and Eurol BC1, respectively, was assessed by flow cytometry (Fig. S1) and revealed a segregation of C chromosomes as already described by Leflon *et al.* (2006). Three plants per cross with $2n = 38$ were selected. Their meiotic behaviour was highly unstable, with only 0–25% of cells showing 19 bivalents as expected for *B. napus* (Tables S2, S3). BC1, BC2 and BC3 progeny frequently showed an R : S distortion of segregation which can be explained by the meiotic instability, illustrated by plants with up to $2n = 39$ (Table S2). These data were confirmed by the analysis of the selfing progeny of the four selected BC3 plants (Table S2); the three most stable plants gave rise to 94R : 34S plants, fitting the expected 75R : 25S segregation for one major gene ($P = 0.683$, $\chi^2 < 3.84$), whereas the selfing progeny of the BC3 plant showing only 50% of pollen mother cells with 19 bivalents revealed a distortion (15R : 30S).

Genetic control of avirulence

The virulent isolate IBCN14 (Av5-6, Mat1-1) was crossed with the avirulent isolate v23.1.3 (Av1-4-5-6-7-8, Mat1-2). Apart from avirulence on line 02-159-4-1, three *Avr* genes were polymorphic in this cross, allowing us to test allelism between the avirulent phenotype on line 02-159-4-1 and the three independent *Avr* genes *AvrLm1*, *AvrLm4-7*, and *AvrLm8*. Considering these three genes, the eight possible phenotypic classes were recovered in the progeny, with proportions consistent with those expected for three independent loci ($P_{[\chi 2]} = 0.333$; Table 3). The 14 : 70 segregation ratio of virulence to avirulence towards line 02-159-4-1 was more consistent with that of three independent *Avr* genes ($P = 0.248$) than two ($P = 0.06$) or one gene ($P < 0.001$). Notably, virulent and avirulent isolates were found in almost all previously defined phenotypic classes (Table 3), suggesting that, whatever the number of genes involved, they all are distinct from *AvrLm4-7*, *AvrLm1* or *AvrLm8*. In addition, as *AvrLm2*, *AvrLm3*, *AvrLm5*, *AvrLm6* and *AvrLm9* are not segregating in cross 38, and IBCN14 was not virulent on Surpass400, harbouring *RlmS*, or on *B. napus* lines harbouring *Rlm10* (data not shown), the interaction between v23.1.3 and line 02-159-4-1 involved at least one new avirulence gene.

The avirulence locus is located on the minichromosome

Among 468 MS markers showing size polymorphism in a range of *L. maculans* isolates (Bally *et al.*, 2010), 235 were polymorphic

Table 3 Segregation of interaction phenotypes in *Leptosphaeria maculans* cross 38 (IBCN14 × v23.1.3)

Avr Locus	Parental isolates IBCN14	v23.1.3	Phenotypic classes in the progeny[a]															
AvrLm4-7	V[b]	A	A		A		A		A		V		V		V		V	
AvrLm1	V	A	A		A		V		V		A		A		V		V	
AvrLm8	V	A	A		V		A		V		A		V		A		V	
Number of isolates in each class			13		11		11		16		11		9		4		9	
Virulence on line 02-159-4-1	V	A	V	A	V	A	V	A	V	A	V	A	V	A	V	A	V	A
Number of isolates in each class			1	12	0	11	4	7	3	13	1	10	3	6	0	4	2	7

[a]The eight phenotypic classes as defined by the combination of the three independent avirulence (*Avr*) genes that are polymorphic in cross 38, *AvrLm4-7*, *AvrLm1* and *AvrLm8*.
[b]V, the isolate is virulent, or A, the isolate is avirulent, following inoculation on lines with the corresponding resistance gene (*Brassica napus* Jet Neuf or Pixel (*Rlm4*), Columbus (*Rlm1*, *Rlm3*), *B. rapa* 00-156-2-1 (*Rlm8*) (Balesdent *et al.*, 2002), 02-159-4-1 (*B. rapa* resistant line; this work)).

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

between v23.1.3 and IBCN14. Of these, 227 showed size polymorphism between the alleles amplified in v23.1.3 and IBCN14, whereas seven showed a presence/absence polymorphism, with the locus amplified in v23.1.3 but not in IBCN14. One of these had been generated from the sequence of SC01 and the six others from that of SC22, previously shown to correspond to the *L. maculans* MC (Rouxel *et al.*, 2011) ('Min' markers; Fig. 1d). In the progeny of cross 38, the presence/absence polymorphic marker from SC01 segregated independently from avirulence to line 02-159-4-1. By contrast, all polymorphic markers from SC22 co-segregated with the interaction phenotype on line 02-159-4-1; the 14 virulent isolates did not amplify the markers whereas those amplifying the markers were all avirulent. Consistent with the avirulence segregation ratio, all SC22 MS markers segregated with a very strong segregation distortion. From these data we hypothesized that the *Avr* locus (loci) matching the line 02-159-4-1 *R* gene was (were) located on SC22 and that the whole SC22 was missing in the virulent parental isolate IBCN14. In this context, it was not possible to further reduce the genetic interval around the avirulence locus.

Segregation distortion of MC markers could be explained either by the presence of a duplicated copy of the MC in isolate v23.1.3, although the 14 : 70 segregation of the markers is only poorly supported by statistics ($P = 0.06$), or by lower viability or germination rate of MC^- ascospores. Notably, *in vitro* cross 37 ($MC^+ \times MC^+$ isolates) and cross 38 ($MC^+ \times MC^-$ isolates; Table 1) were established at the same time under the same conditions but gave rise to progeny with clear differences in ascospore germination rate. The germination rate of ascospores discharged from all pseudothecia from cross 37 was > 90%, whereas only 67.3 ± 5% of ascospores ejected from cross 38 pseudothecia germinated.

Analysis of SC22 identified a unique avirulence gene candidate

SC22 (731 443 bp) encompasses nine AT-rich isochores (AT1–AT9; Fig. 1b,c) covering 92.5% of its length (Rouxel *et al.*, 2011) and eight GC-equilibrated isochores representing 54 791 bp only. SC22 contains 36 predicted genes, one of them being located within an AT-rich isochore (Fig. 1). Compared with gene models of the whole *L. maculans* genome, a low percentage (six out of 36) of translated SC22 gene models had BLAST hits to nonredundant protein databases, but most of them (30 genes, including all genes with BLAST hits) were validated by transcriptomic and/or expressed sequence tag (EST) data (Tables 4, S4, S5). Six genes were predicted to encode secreted proteins, among which four were small (< 300 AA) proteins (SSPs). Only one of them, Lema_uP119060.1, combined all the characteristics of previously described *L. maculans* avirulence genes (Gout *et al.*, 2006b; Fudal *et al.*, 2007; Parlange *et al.*, 2009) and therefore was selected as an avirulence candidate gene; it is located within an AT-rich isochore, is over-expressed 7 d post inoculation (Fig. 1e) and encodes a 95 AA, cystein-rich SSP (Fig. 2) with no homology in the databases. Lema_uP119060.1 was further annotated by RACE PCR. One intron of 71 nt, a 5′-UTR of 52 bp and a 3′-UTR of 210 bp were identified (Fig. 2).

AvrLm11, a 'classical' *L. maculans* avirulence gene

To test the ability of Lema_uP119060.1 to confer avirulence towards the 02-159-4-1 *B. rapa* resistant line, a complementation assay was carried out. A fragment of 1453 bp corresponding to the complete open reading frame of Lema_uP119060.1,

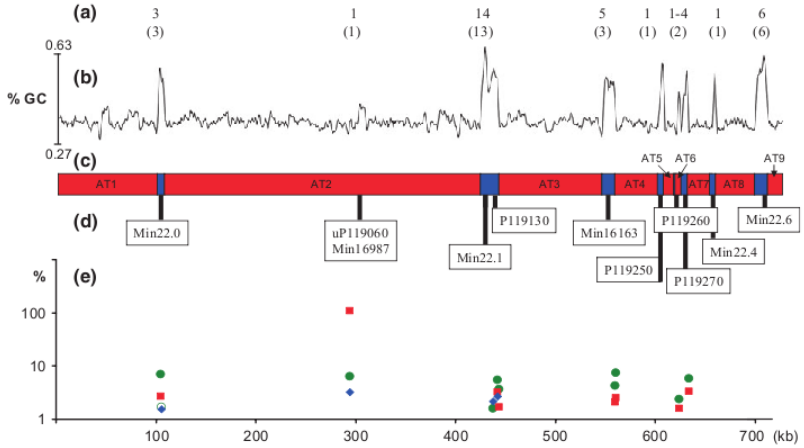


Fig. 1 Genome structure of *Leptosphaeria maculans* Super-Contig 22, gene predictions and their transcriptomic profile. (a) Number of predicted genes along SC22; numbers in brackets indicate the number of genes with at least one transcriptomic support; (b) G+C content variation along SC22; (c) schematic representation of the nine AT-rich (red boxes) and the eight GC-equilibrated (blue boxes) isochores along SC22; (d) location of molecular markers generated along SC22; Minx, minisatellite markers, Px, markers amplifying portions of predicted genes. (e) *In planta* relative expression of predicted genes compared with *in vitro* conditions. Only those genes showing significant over- (closed symbols) or under- (open symbols) expression during oilseed rape infection are shown. Values are the logarithm of the ratio of mean normalized expression values at (green circles) 3 d post infection (dpi), (red squares) 7 dpi or (blue diamonds) 14 dpi over mean normalized expression under *in vitro* conditions, calculated from means of two technical and three biological repeats.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

Table 4 Features of the dispensable chromosome (Super-Contig 22) compared with the whole genome or with AT-rich isochores of *Leptosphaeria maculans*

	Super-Contig 22	Whole v23.1.3 genome	AT-rich isochores[a]
Size (Mb)	0.73	45.12	15.42
Number of predicted genes	36	12469	148
Gene density (nb of genes per kb)	0.049	0.27	0.0096
Core genome size (corresponding to GC-isochores; Mb)	0.055	29.7	NA[b]
Gene density in the core genome (nb of genes per kb)	0.65	0.42	NA
Predicted proteins with BLAST hits (%)[c]	16.6	84.8	24.3
Genes with EST, transcriptomic or proteomic support (%)	82.86	84.8	56.75
Average predicted gene size (bp)	1128	1323	1084
Average predicted protein size (aa)	328	418.4	156
Mean TpA/ApT index of predicted genes	1.08	1.04	1.45
Genes with a TpA/ApT index > 1.5 (%)	13.8	6.9	37
Number (%) of genes encoding predicted Small Secreted Proteins (SSP)	4 (11.1)	651 (5.2)	57 (38.5)

[a]Excluding the small transition zones (mean 859 bp) between AT-rich and GC-equilibrated regions (Rouxel *et al.*, 2011).
[b]NA, not applicable.
[c]Percentage of predicted proteins with a BLAST hit to proteins in the 'Non Redundant' (NR) database, September 2012. Cut-off value $P = e\text{-}10$.

a promoter region of 418 bp and a terminator region of 358 bp (Fig. 2) was cloned into the binary vector pPZPNat1 and introduced into the virulent isolate IBCN14 by ATMT. As a control, the SC22 gene Lema_P119130.1 (Fig. 1), also encoding an SSP, was similarly introduced into IBCN14. Overall, 18 and 14 independent transformants were isolated for Lema_uP119060.1 and Lema_P119130.1, respectively. Only transformation of pPZPNat1-Lema_uP119060.1 into IBCN14 resulted in an avirulent phenotype following inoculation on line 02-159-4-1 for all 18 transformants, whereas their virulence towards Westar (no *R* gene) or 02-23-3-1 (*Rlm7*) was unchanged (Figs 3, S2). Therefore, Lema_uP119060.1 is the gene responsible for the avirulent phenotype on line 02-159-4-1. Based on (1) the single, major gene control of the resistance in the *B. rapa* 02-159-4-1, (2) the control of the corresponding avirulent phenotype by a single avirulence gene and (3) the independence of this interaction with the previously defined *AvrLm1-AvrLm10* and *AvrLmS* Avr genes, Lema_uP119060 was termed *AvrLm11* and the corresponding resistance, *Rlm11*. Further inoculation of a *B. napus* differential set with IBCN14 complemented with *AvrLm11* confirmed that *Rlm11* is distinct from known *R* genes to *L. maculans* (Table S6).

Occurrence of *AvrLm11* and of the MC in progeny of crosses and in field populations

In cross 11, for which a detailed genetic map has been built (Kuhn *et al.*, 2006; Bally *et al.*, 2010), *AvrLm11* was present in both parental isolates and three SC22 markers (Min22.0, Min22.4 and Min17884; Fig. 1d) showed size polymorphism between parental isolates. Three out of 85 (3.5%) cross 11 progeny isolates lacked these three markers, along with *AvrLm11*. Two isolates displayed the two alleles for the three SC22 markers but only one allele (i.e. one or the other parental allele) for the 237 MS markers located on other SCs and mapped in cross 11. These data indicated that mating has led either to the loss or to the duplication of the whole MC with a very high frequency (5.9%). In two additional *in vitro* crosses, the multiplex PCR assay (Fig. 4) revealed the loss of *AvrLm11* in 3.5–6.7% of their progeny (Table 1). Taking into account these three crosses, *AvrLm11*, along with all markers from SC22, were lost with a mean frequency of 4.8% (Table 1; data not shown). Field isolates sampled in France in 2000/2001 and 2010/2011 were multiplex-PCR assayed for *AvrLm11* presence/absence. Depending on the year and the site, *AvrLm11* was absent in 0–13.6% of each population. Overall, 3.2% of French isolates lacked *AvrLm11* (Table 2). No statistically significant evolution of *AvrLm11* frequency could be found in this 10-yr time interval (Wilcoxon test for paired observations (for the seven sites sampled in both decades), $P = 0.248$; Mann–Whitney test (if considering all sites sampled), $P = 0.206$). To confirm the complete dispensability of the MC, additional PCR markers that amplified predicted genes were generated for each of the eight GC-isochores of SC22 ('P*x*' markers; Fig. 1d, Table S1). Similarly to IBCN14, all field isolates lacking *AvrLm11* in both collections also lacked all SC22 PCR markers. Finally, the presence of all SC22 markers, including *AvrLm11*, was analysed in the single-ascospore isolates of the complete tetrad 'Z', in which the loss of the MC band was first identified by pulsed field gel electrophoresis (PFGE) (Leclair *et al.*, 1996). *AvrLm11* and SC22 markers were amplified in the two parental isolates and in the four isolates of the tetrad showing the MC band, whereas they were not amplified in the four isolates lacking the MC band (data not shown).

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

```
gctagctaattaactacttaaactaggacaaatatcttattgttcgtaggataagtaggt     60
attacctaagcttacctttttagtacgcttagctttcttatatgcttacctctcttatac    120
tagtaatcttacctatcttatagtaataagcttacctctcttacaatagtaagcttacct    180
ctcttatatgcttatctctcttgtaatagtaagcttacctctcttataagcttacctctc    240
ttgtaagggtaagcttacctacctagtaagcttacctacctagtaagcttacctacctag    300
taagcttacctacctagtaagcttacctacctagtaagcttacctagtaagcttac        360
ctacctagtaagtatctagtaatagctatcttagtctagtaagcttacctatttacccta    420
tttgcataagattacatatatctccctatcttatcccttccctttccttgatagtattctc    480
tacattcacatcccactctactagttacccccgcttcatc                        520
M  R  F  L  L  P  I  F  S  A  T  L  A  F  A  I  N  E  A  F      20
ATGCGTTTCTTGCTTCCTATATTTTCTGCTACTCTCGCCTTCGCCATAAACGAAGCCTTC    580
E  G  R  P  C  C  Q  K  W  I  E  Q  C  K  L  E  G  A  V  G      40
GAAGGCAGGCCCTGCTGCCAAAAATGGATAGAGCAGTGCAAACTGGAAGGGGCAGTAGGC    640
C  M  L  A  D  Y  N  Y  C  V  V  P  S  S  T  C  Q  E  Q  C      60
TGCATGTTGGCCGACTACAACTACTGCGTAGTGCCATCCTCTACATGTCAAGAACAATGC    700
N  G  F  F  G  R  N  G  R  S  T  H  G  R  D  T  E  Y  V  Y      80
AATGGCTTCTTTGGTCGTAACGGAAGGAGTACACACGGCCGAGACACGGAATATGTGTAC    760
Q  I  H  N                                                      84
CAGATCCATAAgtaggtctgaattttacccggacacgaacttcatctaattttggctcta    820
                              M  Y  T  A  N  E  K  D  P  T  C    95
gtagtatgaatactgaatgcatctgtagCATGTATACCGCGAATGAGAAAGATCCAACTT    880
  *
GCTAA                                                           885
gccaattggatttttcactacatagcaaacctgcagactatgcagattctagctaaagtg    945
aatagtgtgctgtggcacagaggacacttaataaacgaggcttacttagaaatagacagg   1005
ttgaaggtatagggggggggaagcattataggatagaacggcaccttatcacaacttaact   1065
taacttgttaataaaatttcttctgcttaatctatccctcaagttaagttgtttactatg   1125
ttatcttccattaagtctaaactctccccctttcccctttctttaaataggatactctctat  1185
attagtttcatagataatataacttaattacataacgtagacttttacatactagaagat   1245
acctaagctataaccattaattaaaaaggttctaaccttcttatatagtcttttacagtt   1305
atattgtaactagaaattaaacttgcttgtttaattaatcctgttatagatggaataatt   1365
acatcctatttaacgtagaaagctctccatgcatctaagagtgcttctatagatagtatt   1425
agtataactaaggaacgcatatagaggg                                   1453
```

Fig. 2 Nucleotidic sequence of region surrounding *AvrLm11* (accession number CBX90019.1), used for complementation, and amino acid sequence of the corresponding predicted protein. 5′ and 3′ untranslated regions of the gene are underlined. The intron is indicated as italicized lowercase (77 nt). The predicted signal peptide (15 aa) is indicated by dark grey shading and the eight cysteine residues are indicated by grey shading. The asterisk indicates the stop codon.

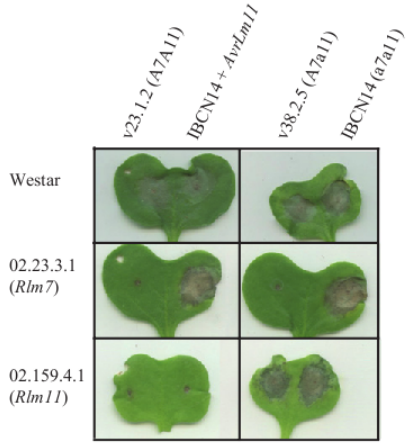


Fig. 3 Differential interaction phenotypes of *Leptosphaeria maculans* isolates on two resistance sources, *Rlm7* and *Rlm11*. Cotyledons of Westar (no R gene), 02-23-3-1 (*Rlm7*) and 02-159-4-1 (*Rlm11*) were inoculated with the wild-type isolates v23.1.2 (Av5-6-7-8-11), IBCN14 (Av5-6) and v38.2.5 (Av4-5-6-7) and with isolate IBCN14 complemented with the *AvrLm11* candidate gene. Photographs were taken 15 d post inoculation and show that the introduction of *AvrLm11* in IBCN14 is sufficient to trigger the *Rlm11*-mediated resistance, whereas the virulent phenotype on the *Rlm7* line is unaltered.

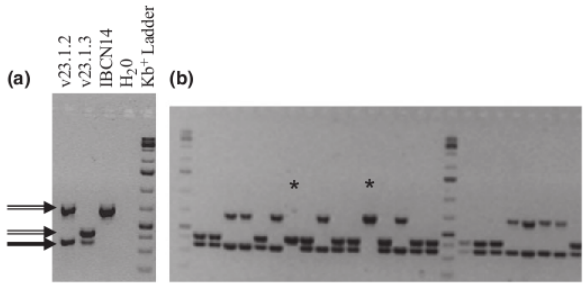


Fig. 4 Multiplex PCR identification of the presence of *AvrLm11* in *Leptosphaeria maculans* isolates. (a) Multiplex assay for the reference isolates v23.1.2 (Mat1-1, *AvrLm11*), v23.1.3 (Mat1-2, *AvrLm11*), IBCN14 (Mat1-1, *avrLm11*) and the water control. (b) An example of the use of the multiplex assay for wide-scale screening of the presence of *AvrLm11* in field isolates. The arrow indicates the PCR product for *AvrLm11*, the double arrows the PCR products for mating-type alleles, and the asterisks field isolates missing *AvrLm11*.

RIP signatures in SC22 genes

Repeat-induced point mutation (RIP) is a pre-meiotic mechanism of inactivation of duplicated sequences in some ascomycetes (Galagan & Selker, 2004). RIP has been shown to be active in *L. maculans* and to contribute to rapid evolution of effector genes, leading to either their diversification or their inactivation (Idnurm & Howlett, 2003; Fudal *et al.*, 2009; Rouxel *et al.*, 2011; Daverdin *et al.*, 2012). As duplication of the MC has been

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

observed to occur in some progeny isolates (2.3% in cross 11; see also Leclair *et al.*, 1996), it was questioned whether this duplication could lead to higher RIP signatures on MC genes compared with other chromosomes. The RIP index TpA/ApT was calculated for all *L. maculans* genes located in GC-isochores, in AT-isochores or in SC22. The RIP index of SC22 genes was intermediate between, but not significantly different from, those of GC-isochores and AT-isochores ($P = 0.837$ and $P = 0.218$, respectively; nonparametric Steel–Dwass–Critchlow–Fligner multiple comparisons). However, five (13.8%) SC22 predicted genes, including *AvrLm11*, had a TpA/ApT > 1.5, which represents twice the proportion observed for the whole *L. maculans* genome (6.9%) but half that for genes within AT-isochores (37%; Table 4).

Discussion

In the 1990s, PFGE of *L. maculans* karyoptypes demonstrated both major chromosome-size polymorphism, generated through meiosis, and the apparent loss of the smallest chromosomal band (Plummer & Howlett, 1993, 1995; Leclair *et al.*, 1996). Leclair *et al.* (1996) failed to identify single-copy sequences specific for the MC, probably because of the large proportion of repeats it harbours, therefore preventing a formal Southern blot demonstration that the absence of the smallest PFGE band really corresponded to loss of genetic material rather than translocation to larger chromosomes. Dispensability of the MC therefore remained questioned until this study. WGS of *L. maculans* confirmed that the MC indeed was mainly made up of repeated elements, in a much higher proportion than in the rest of the genome, with a low GC content (35.2% GC for SC22 compared with 44.1% for the WGS; Rouxel *et al.*, 2011). The search for MS markers targeted to GC-isochores enabled us to generate here single-copy markers along the entire MC. The analysis of their occurrence in isolates from a tetrad previously described as lacking the MC PFGE band (Leclair *et al.*, 1996) demonstrated the complete loss of this chromosome, including all GC-isochores and genes. Therefore, the smallest chromosome of *L. maculans* has to be considered as a B chromosome or a CDC.

The *L. maculans* CDC shares some characteristics with previously described fungal CDCs. Its size, ranging between 650 and 950 kb depending on the isolate (Leclair *et al.*, 1996), is close to those of the supernumerary chromosomes 15 (0.75 Mp) and 17 (0.53 Mp) of *N. haematococca* (Coleman *et al.*, 2009), the 1.05 Mb *A. alternata* CDC (Hatta *et al.*, 2002), the 0.41–0.77 Mb size range of the eight *M. graminicola* CDCs (Goodwin *et al.*, 2011) and the 1 Mb CDC of *A. arborescens* (Hu *et al.*, 2012). Similarly to CDCs of *N. haematococca*, *F. oxysporum*, or *M. graminicola*, a higher repeat content along with a lower GC content was also observed in the *L. maculans* CDC. However, a few striking features of the *L. maculans* CDC are noteworthy. First, its predicted gene content is far lower than those reported in other species, with a gene density of < 0.05 genes per kb, compared with *c.* 0.09–0.18 genes per kb for the three *N. haematococca* CDCs (Coleman *et al.*, 2009), 0.14 genes per kb in *M. graminicola* (Goodwin *et al.*, 2011) and 0.21 genes per kb for the *A. arborescens* CDC (Hu *et al.*, 2012). Secondly, a predicted function could be attributed to none of the CDC predicted proteins, and a very low percentage of them had BLAST hits. For example, 160 out of 209 genes (76.5%) of the *A. arborescens* CDCs had BLAST hits compared with 16.6% for *L. maculans* CDC genes. In this respect, the *L. maculans* CDC more closely resembles those of *M. graminicola*, for which 10% of the genes in the dispensome had BLAST hits compared with 59% for the core chromosome genes. However, the *M. graminicola* CDCs contained a lower proportion of secreted proteins or other pathogenicity genes than the core chromosomes, while in *L. maculans* the MC contained a proportion of SSP-encoding genes twice as high as, but not significantly different (binomial exact test, $P = 0.115$) from, that of the whole genome (Table 4).

The genome of *L. maculans* is structured into alternate regions with homogenous GC content abruptly changing from AT-rich regions (or isochores) to GC-equilibrated regions, and this organization is observed for all chromosomes (Rouxel *et al.*, 2011). Many characteristics of the *L. maculans* CDC are intermediate between those of AT-rich and GC-equilibrated isochores, such as gene density, gene size, RIP indices and effector gene content (Table 4). However, BLAST hits of predicted CDC proteins were much lower than predicted for the rest of the genome, including AT-rich regions, while the percentage of CDC genes with transcriptomic support was similar to that found for the entire *L. maculans* genome. These data highlight the specificity of the gene content of the CDC compared with other chromosomes, with still unknown functions or functions that may be specific for its interaction with its host plants or for its life cycle.

Our work shows that meiosis could generate either the loss or the duplication of the *L. maculans* CDC at an appreciable frequency (*c.* 5%). In *L. maculans*, inactivation of duplicated sequences by RIP is frequent and RIP signatures can be found even in single-copy genes embedded within repeated regions (Rouxel *et al.*, 2011). Duplication of the CDC in some isolates at a nonnegligible frequency, as found here or in previous reports (Leclair *et al.*, 1996), could lead to RIP mutations in the genes it harbours. Activity of RIP to lower the GC content of the dispensome has been described in *M. graminicola* (Goodwin *et al.*, 2011) and RIP following duplication may have led to a diversification of the genes harboured by the MC in *L. maculans*. However, genes located in GC-isochores of SC22 display only slightly higher RIP indexes than the genes located in GC-isochores from other chromosomes, maybe because higher rates of RIP mutations would have led to inactivation of important proteins and have been counter-selected.

The foreign origin and horizontal transfer of CDCs has been suggested or demonstrated in a few fungal species, leading to a change in host range or in pathogenic vs saprophytic status (Mehrabi *et al.*, 2011). The low BLAST hits and lack of predicted function of proteins encoded by the *L. maculans* CDC could be an argument in favour of such a hypothesis; however, the current very low number of BLAST hits of CDC proteins to prokaryotic or eukaryotic sequences prevents any attempt to test this hypothesis and trace back the phylogenetic origin of the *L. maculans*

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

CDC. In addition, because the CDC structure resembles the unique isochore structure of the core chromosomes of *L. maculans,* it is more plausible that its CDC evolved from the core chromosomes.

A few recent reports describe the occurrence of avirulence genes in fungal CDCs. Here we show that the *L. maculans* CDC also shelters a high proportion of genes encoding SSPs, one of which, *AvrLm11,* encodes an avirulence effector. *AvrLm11* is the first cloned *L. maculans Avr* gene corresponding to a resistance source identified in *B. rapa*. *AvrLm11* has all the features of *L. maculans Avr* genes: it is a 'middle of nowhere' gene (Gout *et al.*, 2006b) located within a large (321-kb) AT-rich region, it is expressed during infection, it is over-expressed 7 d after inoculation and it encodes a predicted cystein-rich SSP. Its location on a CDC makes the study of its evolution pattern particularly interesting. We have shown how frequently the loss of the CDC, including *AvrLm11*, may happen during sexual reproduction. Sexual mating occurs each year during the *L. maculans* infectious cycle, and ascospores produced following meiosis are in most countries regarded as the primary source of inoculum. The first leaf lesions observed in autumn in Europe correspond to ascospore infection sites, with each leaf lesion corresponding to a distinct haplotype generated by sexual recombination (Gout *et al.*, 2006a). The frequency of the loss of the CDC following meiosis in controlled conditions (4.8%) was comparable to that observed in samples collected in autumn from leaf lesions (3.2%). Paradoxically, we did not observe any increase in the frequency of isolates having lost the CDC in a 10-yr time interval. These data suggest that maintenance of the *L. maculans* CDC confers a selective advantage for its life cycle completion because, without any fitness cost, annual random matings combined with a 'mutation' rate (loss of the MC) of 0.048 would have led, even in the absence of the *Rlm11* selection pressure in the field, to a shift from 4.5% (2000/2001 sampling) to *c.* 35% of isolates lacking the CDC 10 yr later (2010/2011) (Methods S1, Fig. S3). Leclair *et al.* (1996) tested the host range and virulence of a few *L. maculans* isolates having lost or not the MC and did not find any fitness cost linked to its loss. However, only the early stages of the interaction were investigated, using cotyledon inoculations and qualitative rating scales (Leclair *et al.*, 1996). Genes located on the CDC could be important in the life cycle of the fungus at a later stage, less accessible to controlled condition tests but eventually revealed by field population analyses. Whether this fitness cost is linked to the loss of *AvrLm11* itself, as an effector gene, or of other CDC genes with currently unknown function, is still unresolved. In the present study, cross 38 involving isolate IBCN14, lacking the CDC, gave rise to abundant ascospore discharge, but only 70% of them germinated. In addition, a strong segregation distortion was found in its progeny for the transmission of the CDC, compared with the 1 : 1 expected segregation ratio. It can thus also be hypothesized that the loss of the CDC reduces viability of the progeny and counterbalances the high frequency of MC loss.

All these data therefore question the potential durability of the *AvrLm11*-matching resistance gene, *Rlm11*, and whether the ease by which the CDC can be lost will be counterbalanced by the apparent strong fitness cost resulting from its loss. Molecular events responsible for virulence have been investigated for three *L. maculans Avr* genes located in the core chromosomes (Gout *et al.*, 2007; Fudal *et al.*, 2009; Daverdin *et al.*, 2012). In most cases the resistance was overcome by either gene inactivation or complete deletion of the gene and its AT-rich environment. In contrast to the findings of the present study, complete deletion only encompassed one coding sequence regardless of the size of the deletion, while in the case of *AvrLm11* the whole CDC including all its GC-isochores and genes was lost. The present work has led to the introduction of *Rlm11* from *B. rapa* to oilseed rape genotypes in a context where avirulent isolates represent > 95% of the French *L. maculans* population. This material is now available not only for plant breeding but also to assess in field conditions the durability of an *R* gene corresponding to an *Avr* gene located on a CDC, in comparison with *Avr* genes present in the core chromosomes. The survey of virulent isolates towards *Rlm11* in experimental fields will be facilitated by the multiplex PCR assay to detect *AvrLm11* and the CDC single-copy molecular markers designed here.

Acknowledgements

This work was funded by the French agency Agence Nationale de la Recherche (ANR) contract ANR-07-GPLA-015 ('AVirLep') under the framework of the Génoplante 2010 programme. The 2010–2011 large-scale sampling was funded by the CTPS project 'Evolep' coordinated by Xavier Pinochet (CETIOM). The authors wish to thank Laurent Coudard, Sabrina Frouillou, Julien Carpezat, Juliette Linglin, Bertrand Auclair and Martin Willigsecker for technical assistance.

References

Ansan-Melayah D, Balesdent MH, Buée M, Rouxel T. 1995. Genetic characterization of *AvrLm1*, the first avirulence gene of *Leptosphaeria maculans*. *Phytopathology* **85**: 1525–1529.

Balesdent MH, Attard A, Ansan-Melayah D, Delourme R, Renard M, Rouxel T. 2001. Genetic control and host range of avirulence toward *Brassica napus* cultivars Quinta and Jet Neuf in *Leptosphaeria maculans*. *Phytopathology* **91**: 70–76.

Balesdent MH, Attard A, Kuhn ML, Rouxel T. 2002. New avirulence genes in the phytopathogenic fungus *Leptosphaeria maculans*. *Phytopathology* **92**: 1122–1133.

Balesdent MH, Barbetti MJ, Li H, Sivasithamparam K, Gout L, Rouxel T. 2005. Analysis of *Leptosphaeria maculans* race structure in a worldwide collection of isolates. *Phytopathology* **95**: 1061–1071.

Balesdent MH, Louvard K, Pinochet X, Rouxel T. 2006. A large scale survey of races of *Leptosphaeria maculans* occurring on oilseed rape in France. *European Journal of Plant Pathology* **114**: 53–65.

Bally P, Grandaubert J, Rouxel T, Balesdent MH. 2010. FONZIE: an optimized pipeline for minisatellite marker discovery and primer design departing from large sequence data sets. *BMC Research Notes* **3**: 322.

Chuma I, Isobe C, Hotta Y, Ibaragi K, Futamata N, Kusaba M, Yoshida K, Terauchi R, Fujita Y, Nakayashiki H *et al.* 2011. Multiple translocation of the AVR-Pita effector gene among chromosomes of the rice blast fungus *Magnaporthe oryzae* and related species. *PLoS Pathogens* **7**: e1002147.

Chuma I, Tosa Y, Taga M, Nakayashiki H, Mayama S. 2003. Meiotic behavior of a supernumerary chromosome in *Magnaporthe oryzae*. *Current Genetics* **43**: 191–198.

Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S *et al.* 2009. The genome

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PLoS Genetics* **5**: e1000618.

Covert SF. 1998. Supernumerary chromosomes in filamentous fungi. *Current Genetics* **33**: 311–319.

Cozijnsen AJ, Howlett BJ. 2003. Characterisation of the mating-type locus of the plant pathogenic ascomycete *Leptosphaeria maculans*. *Current Genetics* **43**: 351–357.

Daverdin G, Rouxel T, Gout L, Aubertot JN, Fudal I, Meyer M, Parlange F, Carpezat J, Balesdent MH. 2012. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PLoS Pathogens* **8**: e1003020.

Eber F, Lourgant K, Brun H, Lode M, Huteau V, Coriton O, Alix K, Balesdent MH, Chèvre AM. 2011. Analysis of *Brassica nigra* chromosomes allows identification of a new effective *Leptosphaeria maculans* resistance gene introgressed in *Brassica napus*. 13th International rapeseed congress, Prague 5–9 June 2011.

Eckert M, Gout L, Rouxel T, Blaise F, Jedryczka M, Fitt BDL, Balesdent MH. 2005. Identification and characterization of polymorphic minisatellites in the phytopathogenic ascomycete *Leptosphaeria maculans*. *Current Genetics* **47**: 34–48.

Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, Balesdent MH, Rouxel T. 2009. Repeat-Induced Point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. *Molecular Plant–Microbe Interaction* **22**: 932–941.

Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, Eckert MR, Cattolico L, Bernard-Samain S, Balesdent MH, Rouxel T. 2007. Heterochromatin-like regions as ecological niches for avirulence genes in *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. *Molecular Plant–Microbe Interaction* **20**: 459–470.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. *Trends in Genetics* **9**: 417–423.

Gall C, Balesdent MH, Robin P, Rouxel T. 1994. Tetrad analysis of acid phosphatase, soluble protein patterns, and mating type in *Leptosphaeria maculans*. *Phytopathology* **84**: 1299–1305.

Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, Van der Lee TAJ, Grimwood J, Aerts A *et al.* 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity and stealth pathogenesis. *PLoS Genetics* **7**: e1002070.

Gout L, Eckert M, Rouxel T, Balesdent MH. 2006a. Genetic variability and distribution of mating type alleles in field populations of *Leptosphaeria maculans* from France. *Applied and Environmental Microbiology* **72**: 185–191.

Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, Balesdent MH, Rouxel T. 2006b. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Molecular Microbiology* **60**: 67–80.

Gout L, Kuhn ML, Vincenot L, Bernard-Samain S, Cattolico L, Barbetti M, Moreno-Rico O, Balesdent MH, Rouxel T. 2007. Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. *Environmental Microbiology* **9**: 2978–2992.

Hatta R, Ito K, Hosaki Y, Tanaka T, Tanaka A, Yamamoto M, Akimitsu K, Tsuge T. 2002. A conditionally dispensable chromosome controls host-specific pathogenicity in the fungal plant pathogen *Alternaria alternata*. *Genetics* **161**: 59–70.

Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T. 2012. Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. *BMC Genomics* **13**: 17.

Huang YJ, Li ZQ, Evans N, Rouxel T, Fitt BDL, Balesdent MH. 2006. Fitness cost associated with loss of the *AvrLm4* avirulence function in *Leptosphaeria maculans* (phoma stem canker of oilseed rape). *European Journal of Plant Pathology* **114**: 77–89.

Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genetics and Biology* **39**: 31–37.

Johnson LJ, Johnson RD, Akamatsu H, Salamiah A, Otani H, Kohmoto K, Kodama M. 2001. Spontaneous loss of a conditionally dispensable chromosome from the *Alternaria alternata* apple pathotype leads to loss of toxin production and pathogenicity. *Current Genetics* **40**: 65–72.

Jones RN. 1995. Tansley Review No. 85. B chromosomes in plants. *New Phytologist* **131**: 411–434.

Kuhn ML, Gout L, Howlett BJ, Melayah D, Meyer M, Balesdent MH, Rouxel T. 2006. Genetic linkage maps and genomic organization in *Leptosphaeria maculans*. *European Journal of Plant Pathology* **114**: 17–31.

Leclair S, Ansan-Melayah D, Rouxel T, Balesdent MH. 1996. Meiotic behaviour of the minichromosome in the phytopathogenic ascomycete *Leptosphaeria maculans*. *Current Genetics* **30**: 541–548.

Leflon M, Brun H, Eber F, Delourme R, Lucas MO, Vallée P, Ermel M, Balesdent MH, Chèvre AM. 2007. Detection, introgression and localization of genes conferring specific resistance to *Leptosphaeria maculans* from *Brassica rapa* into *B. napus*. *Theoretical and Applied Genetics* **115**: 897–906.

Leflon M, Eber F, Letanneur JC, Chelysheva L, Coriton O, Huteau V, Ryder C, Barker G, Jenczewski E, Chèvre AM. 2006. Pairing and recombination at meiosis of *Brassica rapa* (AA) × *Brassica napus* (AACC) hybrids. *Theoretical and Applied Genetics* **113**: 1467–1480.

Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B *et al.* 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **464**: 367–373.

Mehrabi R, Bahkali AH, Abd-Elsalam KA, Moslem M, Ben M'Barek S, Gohari AM, Jashni MK, Stergiopoulos I, Kema GHJ, de Wit PJGM. 2011. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. *FEMS Microbiology Reviews* **35**: 542–554.

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M, Cabrero J, Camacho JP, Perfectti F. 2011. A single, recent origin of the accessory B chromosome of the grasshopper *Eyprepocnemis plorans*. *Genetics* **187**: 853–63.

Parlange P, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, Blaise F, Grezes-Besset B, Rouxel T. 2009. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. *Molecular Microbiology* **71**: 851–863.

Plummer KM, Howlett BJ. 1993. Major chromosomal length polymorphisms are evident after meiosis in the phytopathogenic fungus *Leptosphaeria maculans*. *Current Genetics* **24**: 107–113.

Plummer KM, Howlett BJ. 1995. Inheritance of chromosomal length polymorphisms in the ascomycete *Leptosphaeria maculans*. *Molecular and General Genetics* **247**: 416–422.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S *et al.* 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by RIP mutations. *Nature Communications* **2**: 202.

Rouxel T, Willner E, Coudard L, Balesdent MH. 2003. Screening and identification of resistance to *Leptosphaeria maculans* in *Brassica napus* accessions. *Euphytica* **133**: 219–231.

Stachowiak A, Olechnowicz J, Jedryczka M, Rouxel T, Balesdent MH, Happstadius I, Gladders P, Latunde-Dada A, Evans N. 2006. Frequency of avirulence alleles in field populations of *Leptosphaeria maculans* in Europe. *European Journal of Plant Pathology* **114**: 67–75.

Tzeng T, Lyngholm LK, Ford CF, Bronson CR. 1992. A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus*. *Genetics* **130**: 81–96.

U N. 1935. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany* **7**: 389–448.

Wittenberg AHJ, van der Lee TAJ, Ben M'Barek S, Goodwin SB, Kilian A, Visser RGF, Kema GHJ, Schouten HJ. 2009. Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. *PLoS ONE* **4**: e5863.

van de Wouw AP, Marcroft SJ, Barbetti MJ, Li H, Salisbury PA, Gout L, Rouxel T, Howlett BJ, Balesdent MH. 2008. A dual control of avirulence in *Leptosphaeria maculans* towards a *Brassica napus* variety with 'sylvestris-derived'

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

resistance, suggests two resistance genes are operating in the plant. *Plant Pathology* **58**: 305–313.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Percentage of plants with different chromosome numbers assessed by flow cytometry in the progeny of *Brassica rapa* × *Brassica napus* (AAC) hybrids crossed to their recurrent parents Darmor or Eurol.

Fig S2 Interaction phenotypes of wild-type and transformed *Leptosphaeria maculans* isolates with *AvrLm11* candidate genes.

Fig. S3 Evolution of the frequency of the dispensable chromosome in field populations of *Leptosphaeria maculans* under different fitness cost hypotheses.

Table S1 PCR primers used in this study

Table S2 Analysis of the progeny of crosses between a resistant *Brassica rapa* plant ($2n = 20$) and the susceptible *Brassica napus* ($2n = 38$) variety Darmor

Table S3 Analysis of the progeny of crosses between a resistant *Brassica rapa* plant ($2n = 20$) and the susceptible *Brassica napus* ($2n = 38$) variety Eurol

Table S4 List and characteristics of SC22 predicted genes

Table S5 List and characteristics of SC22 predicted proteins

Table S6 Interaction phenotypes of wild-type or transformed *Leptosphaeria maculans* isolates on a *Brassica* differential set

Methods S1 Inoculation tests and modelling the loss of the dispensable chromosome in *Leptosphaeria maculans* populations.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

Article 4. Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*.

JL Soyer, M El Ghalid, N Glaser, B Ollivier, J Linglin, J Grandaubert, MH Balesdent, LR Connolly, M Freitag, T Rouxel & I Fudal.

Soumis le 30 janvier 2013 dans *PloS Pathogens* (en révision)

Plant pathogens secrete an arsenal of small secreted proteins (SSPs) acting as effectors that modulate host immunity to facilitate infection. SSP-encoding genes are often located in particular genomic environments and show waves of concerted expression at diverse stages of plant infection. To date, little is known about the regulation of their expression. The genome of the Ascomycete *Leptosphaeria maculans* comprises alternating gene-rich GC-isochores and gene-poor AT-isochores. The AT-isochores harbor mosaics of transposable elements, encompassing one-third of the genome, and are enriched in putative effector genes that present similar expression patterns, namely no expression or low-level expression during axenic cultures compared to strong induction of expression during primary infection of oilseed rape (*Brassica napus*). Here, we investigated the involvement of one specific histone modification, histone H3 lysine 9 methylation (H3K9me3), in epigenetic regulation of concerted effector gene expression in *L. maculans*. For this purpose, we silenced the expression of two key players in heterochromatin assembly and maintenance, *HP1* and *DIM-5* by RNAi. By using HP1-GFP as a heterochromatin marker, we observed that almost no chromatin condensation is visible in strains in which *LmDIM5* was silenced by RNAi. By whole genome oligoarrays we observed overexpression of 369 or 390 genes, respectively, in the silenced-*LmHP1* and -*LmDIM5* transformants during growth in axenic culture, clearly favouring expression of SSP-encoding genes within AT-isochores. The ectopic integration of four effector genes in GC-isochores led to their overexpression during growth in axenic culture. These data strongly suggest that epigenetic control, mediated by HP1 and DIM-5, represses the expression of at least some effector genes located in AT-isochores during growth in axenic culture. Our hypothesis is that changes of lifestyle and a switch toward pathogenesis lift chromatin-mediated repression, rendering promoters of effector genes accessible to specific transcription factors.

Article 5. Deciphering the regulome of the causal agent of the stem canker agent of oilseed rape, *Leptosphaeria maculans* 'brassicae' to select putative regulators of pathogenicity.

JL Soyer, J Grandaubert, J Linglin, B Ollivier, A Hamiot, RG Lowe, T Rouxel & I Fudal.

En préparation

L. maculans 'brassicae' is an hemibiotrophic fungus, pathogen of crucifers that presents a very complicated life cycle during which it alternates different nutritional strategies. The switch between its different life styles during the primary infection of oilseed rape (*Brassica napus*), notably from the asymptomatic phase that last about 12 days to its first necrotrophic stage reflects its huge adaptability and underlines vast reprogramming of gene expression along with complex regulatory network to allow the switches between its different behaviors. In a study aiming at identifying major pathogenicity regulator of *Leptosphaeria maculans* 'brassicae', we identified its repertoire of transcription factors and processed oligoarray data to select putative regulators to set up their functional analysis with a particular focus on their involvement in the pathogenicity of *L. maculans* 'brassicae'. This fungus is part of a complex species whose members present a different pathogenicity towards oilseed rape. We investigated the repertoire of transcription factors within the *Leptosphaeria* species complex to point out specific *L. maculans* 'brassicae' regulators or those that are found in all the sequenced members of the *L. maculans*-*L. biglobosa* species complex. Several TFs have been identified as possibly required for pathogenicity of *L. maculans* 'brassicae' toward *B. napus* that will be further characterized. This study is a valuable insight to decipher the regulatory pathways of gene expression that underline the infection of oilseed rape by *L. maculans* 'brassicae'.

ANNEXE 2 : BILAN D'ACTIVITÉ

Présentation orale

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?
J Grandaubert, MH Balesdent, M Links, H Borhan, T Rouxel.
26th Fungal Genetics Conference, Dothideomycetes Workshop, March 15-20 2011, Pacific Grove USA.

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?
J Grandaubert, MH Balesdent, M Links, H Borhan, T Rouxel.
Séminaire d'unité, 25 mars 2011, BIOGER.

Comparative and evolutionary genomics within the *L. maculans – L. biglobosa* species complex.
Comité de Thèse 1. Invités : EH Stukenbrock (MPI, Marburg) et E Lerat (CNRS, Lyon), 1 avril 2011, BIOGER.

Génomique évolutive de *Leptosphaeria*.
M2 Module Pathogénie-Symbiose/Défense des plantes, 16 novembre 2011, BIOGER.

Quel rôle jouent les éléments transposables dans la genèse et la diversification des effecteurs chez les espèces du complexe *Leptosphaeria* ?
J Grandaubert, MH Balesdent, M Links, H Borhan, T Rouxel.
9ème Rencontres de Phytopathologie-Mycologie de la SFP, 16 au 20 janvier 2012, Aussois – France.

Comparative and evolutionary genomics within the *L. maculans – L. biglobosa* species complex.
Comité de Thèse 2. Invités : EH Stukenbrock (MPI, Marburg), E Lerat (CNRS, Lyon) et A Hua-Van (CNRS, Gif), 22 octobre 2012, BIOGER.

***Leptosphaeria maculans* 'brassicae' : « Transposable Elements changed my life, I feel different now.**
J Grandaubert, CL Schoch, H Borhan, BJ Howlett, T Rouxel.
27th Fungal Genetics Conference, Phylogenomics session, March 12-17 2013, Pacific Grove USA.

***Leptosphaeria maculans* 'brassicae' : « Transposable Elements changed my life, I feel different now.**
J Grandaubert, CL Schoch, H Borhan, BJ Howlett, T Rouxel.
June 18 2013, Max Planck Institute Marburg, Germany.

Présentation orale (co-auteur)

The genome of *Leptosphaeria maculans* : when structure meets function.
T Rouxel, MH Balesdent, J Amselem, H Quesneville, J Grandaubert, V Dominguez, I Fudal, F Parlange, G Daverdin, L Gout.
EMBL-EBI Industry workshop: Integration of Genomic Information Related to Crop Diseases and Pests, 20 - 21 October 2008, Hinxton, UK.

A systematic analysis of T-DNA insertion pattern in the genome of *Leptosphaeria maculans*.
S Bourras, M Meyer, E Mendes-Pereira, J Grandaubert, MH Balesdent, T Rouxel.
25th Fungal Genetics Conference, March 17-22 2009, Pacific Grove USA.

AT-rich isochores as ecological niches for effectors in the genome of *Leptosphaeria maculans*.
I Fudal, J Grandaubert, A Dilmaghani, N Glaser, P Bally, P Wincker, A Couloux, BJ Howlett, B Profotova, MH Balesdent, T Rouxel.
25th Fungal Genetics Conference, March 17-22 2009, Pacific Grove USA.

AT-rich isochores as ecological niches for effectors in the genome of *Leptosphaeria maculans*.
I Fudal, J Grandaubert, A Dilmaghani, N Glaser, P Wincker, A Couloux, BJ Howlett, MH Balesdent, T Rouxel.
IS-MPMI 2009 XIV International Congress on Molecular Plant-Microbe Interactions, July 19-23,2009, Quebec City, Canada.

***Leptosphaeria maculans* AVRs and SSPs.**
I Fudal, J Grandaubert, A Dilmaghani, N Glaser, P Bally, P Wincker, A Couloux, BJ Howlett, MH Balesdent, T Rouxel.
22nd New Phytologist Symposium, 13–16 September 2009, INRA Versailles Research Centre, Paris, France.

Effectors identification and characterisation in the patchwork genome of *Leptosphaeria maculans*.
I Fudal, J Grandaubert, A Dilmaghani, J Linglin, B Ollivier, B Profotova, N Glaser, P Bally, P Wincker, A Couloux, JK Hane, B Tyler, RP Oliver, BJ Howlett, MH Balesdent, T Rouxel.
Dothideomycete comparative genomics jamboree, March 29, 2010, Noordwijkerhout, Holland.

Présentation Poster (voir Annexe 3)

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?
J Grandaubert, MH Balesdent, M Links, H Borhan, T Rouxel.
26th Fungal Genetics Conference, March 15-19 2011, Pacific Grove USA.

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?
J Grandaubert, MH Balesdent, M Links, H Borhan, T Rouxel.
Comparative Genomics of Eukaryotic Microorganisms, October 15-20 2011, San Feliu de Guixols, Spain.

FungIsochores - Isochores and effectors : genome reshaping and the birth of highly pathogenic species in fungal phytopathogens.
J Grandaubert, MH Balesdent, I Fudal, J Amselem, J Kreplak, N Lapalu, B Le Cam, C Lemaire, T Guillemette, T Rouxel.
2012 Plant Genomics seminar, 3-5 avril 2012, Pont-Royal en Provence, France.

***Leptosphaeria maculans* 'brassicae' : « Transposable Elements changed my life, I feel different now.**
J Grandaubert, CL Schoch, H Borhan, BJ Howlett, T Rouxel.
27th Fungal Genetics Conference, March 12-17 2013, Pacific Grove USA.

Présentation Poster (co-auteur)

The invaded genome of the Dothideomycete *Leptosphaeria maculans*.
T Rouxel, V Dominguez, S Torriani, J Hane, BJ Howlett, S Goodwin, RP Oliver, I Fudal, M Meyer, ML Kuhn, J Grandaubert, B McDonald, J Amselem, H Quesneville, P Wincker, A Couloux, MH Balesdent.
9th European Conference on Fungal Genetics, 5th - 8th April 2008, Edinburgh, Scotland.

What Large-Scale T-DNA Insertional Mutagenesis Tells Us About Pathogenicity? A Functional Genomics Analysis In The Dothideomycete *Leptosphaeria maculans*.
S Bourras, M Meyer, J Grandaubert, N Lapalu, I Fudal, J Linglin, B Ollivier, F Blaise, A Simon, J Amselem, MH Balesdent, T Rouxel.
ECFG10 - 10th European Conference on Fungal Genetics, March 29-April 1st 2010, Noordwijkerhout, Holland.

The genome sequence of *Leptosphaeria biglobosa* 'canadensis', a pathogen of oilseed Brassicas.
AP Van de Wouw, MH Balesdent, J Grandaubert, T Rouxel, BJ Howlett
26th Fungal Genetics Conference, March 15-19 2011, Pacific Grove USA.

Epigenetic control of effector genes in *Leptosphaeria maculans*.
JL Soyer, M El Ghalid, N Glaser, B Ollivier, J Linglin, J Grandaubert, MH Balesdent, T Rouxel, I Fudal.
27th Fungal Genetics Conference, March 12-17 2013, Pacific Grove USA.

Carbohydrate binding proteins of two *Leptosphaeria* pathogens of *Brassica napus*.
RGT Lowe, B Clark, AP Van de Wouw, A Cassin, J Grandaubert, T rouxel, BJ Howlett.
27th Fungal Genetics Conference, March 12-17 2013, Pacific Grove USA.

ANNEXE 3 : POSTERS

Poster 1. Présenté lors du 26th Fungal Genetics Conference à Asilomar (USA) le 17 mars 2011.

A Chicken or Egg Dilemna: are Transposable Elements Drivers of Effector Birth and Diversification in *Leptosphaeria* Species?

J Grandaubert[1], MH Balesdent[1], H Borhan[2] and T Rouxel[1]

[1] INRA-Bioger, Grignon, France ; [2] AAF Saskatoon, Canada

Leptosphaeria maculans and *L. biglobosa* are part of a species complex of fungal pathogens of crucifers. The genomes of two *L. maculans* 'brassicae' (Lmb) isolates (45.12 Mb, assembled into 76 scaffolds and 44.16 Mb assembled into 986 scaffolds, respectively) have an unusual bipartite structure – alternating distinct GC-equilibrated and AT-rich blocks of homogenous nucleotide composition. The AT-rich blocks comprise one third of the genome and contain effector genes and families of transposable elements (TEs), postulated to have recently invaded the genome, both of which are affected by Repeat Induced Point mutation. In silico comparison of the Lmb genomes with that of *L. maculans* 'lepidii' (31.53 Mb, assembled into 123 scaffolds) and *L. biglobosa* 'thlaspii' (32.10 Mb, assembled into 237 scaffolds), shows these species have a much more compact genome with a very low amount of TEs (<1%). In addition some recently expanded TE families are specific of *L. maculans* isolates. Compared to the Lmb genomes, less than 14% of the effector genes and 33% of other genes in AT-blocks are present in the two other genomes, suggesting TEs were key players in gene innovation and that the genome environment promoted rapid sequence diversification and selection of genes involved in pathogenicity.

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?

Jonathan Grandaubert[1], Marie-Hélène Balesdent[1], Hossein Borhan[2], Matthew Links[2], Thierry Rouxel[1]

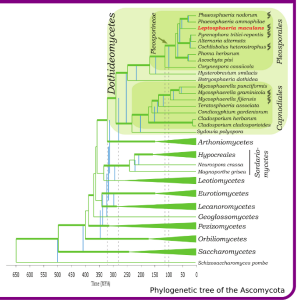
[1]INRA-BIOGER, Thiverval-Grignon, France. [2]AAF Saskatoon, Canada.

1 *Leptosphaeria maculans* 'brassicae'

Leptosphaeria maculans is an Ascomycete, class Dothideomycetes, closely related to three already-sequenced fungi: *Phaeosphaeria nodorum, Pyrenophora tritici-repentis* and *Cochliobolus heterostrophus.*

L. maculans and *L. biglobosa* are part of a species complex of fungal pathogens of crucifers and are collectively responsible for the stem canker disease (also termed « blackleg ») of oilseed rape (*Brassica napus*).

The sequencing of *L. maculans* genome highlights the molecular and evolutionary trends which allow the fungus to adapt rapidly to novel host-derived constraints.



Phylogenetic tree of the Ascomycota

2 *L. maculans* 'brassicae' v23.1.3 genome features

Features of genomes of *L. maculans* and other related Dothideomycetes

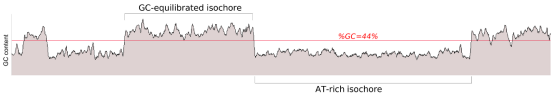
	L. maculans	*P. (Stagonospora) nodorum*	*P. tritici-repentis*	*C. heterostrophus*	*A. brassicicola*	*M. graminicola*
No. of chromosomes	17–18	19	11	15–16	9–11	21
Genome size (Mb)	45.1	36.6	37.8	34.9	30.3	39.7
No of contigs	1,743	496	703	400	4,039	21
No of SuperContigs (SCs)	76	107	47	89	838	21
SC N50 (Mb)	1.8	1.1	1.9	1.3	2.4	NA
Gaps (%)	2.5	0.4	1.7	1.1	5.4	0.01
No. of predicted genes	12,469	10,762	12,141	9,633	10,688	10,952
Average gene length (bp)	1,323	1,326	1,618	1,836	1,523	1,600
GC content (%)	44.1	50.3	50.4	52–54	50.5	55.0
Repeat content (%)	34.2	7.1	16.0	7.0	9.0	18.0
'Core' genome size (Mb)	29.7	34.5	31.7	32.5	27.6	32.6
Gene density/core genome (no. of gene per 10 kb)	4.2	3.1	3.8	3.0	3.9	3.4

Lower GC content
Larger genome size
Higher repeat content

➡ **Invasion of the genome by Transposable Elements (TEs)**

3 A bipartite-structured genome

Alternance of large GC-equilibrated and AT-rich regions all along the chromosomes



AT-isochores represent 36% of the genome, are mainly composed of mosaics of truncated and RIPped TEs and only comprise 5% of the predicted genes (620)

The action of RIP* on TEs is responsible for the low overall GC content

* RIP (Repeat Induced Point mutation) is a premeotic repeat-inactivation mechanism specific to fungi mutating the dinucleotide CpA into TpA and the dinucleotide TpG into TpA.

4 AT-isochores as niches for effectors

Comparative features of SSP-encoding genes in diverse genome environments

	All predicted genes	SSPs in GC-equilibrated regions	SSPs in borders of AT-rich regions	SSPs within AT-rich regions
No.	12,469	529 (4.2%)	65 (0.5%)	57 (0.5%)
BLAST hits (%)	71.3	48.4	15.4	8.8
GC content (%)	54.1	54.6	51.1	48.2
TpA/ApT (RIP index)	1.04	1.20	1.19	1.44
TpA/ApT >1.5 (%)	6.9	16.4	20.0	38.6
EST, transcriptomic or proteomic support (%)	84.8	77.1	56.9	60.0
No. of genes present on the NimbleGen array, and with transcriptomic support	10,524	396	35	33
Genes overexpressed *in planta* 7 dpi (%)	9.9	19.1	13.9	72.7
Genes overexpressed *in planta* 14 dpi (%)	11.0	15.4	22.2	24.2
Average protein size (amino acid)	418.4	167.7	111.6	98.6
% Cysteines in the predicted protein	1.7	2.9	3.8	4.5

Features of known effectors of *L. maculans*

- Small Secreted Proteins (SSPs)
- Lack of homologues
- Low *in vitro* EST support
- Over-expression of the gene during infection
- High cysteine content

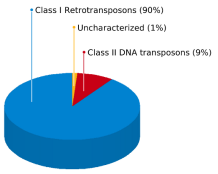
Fit these features
Do not fit these features

AT-isochores are enriched in genes encoding putative effectors (~20%)

SSP-encoding genes within AT-isochores show a biased RIP index:
➡ Overrun of the RIP machinery from TEs to the adjacent single-copy genes

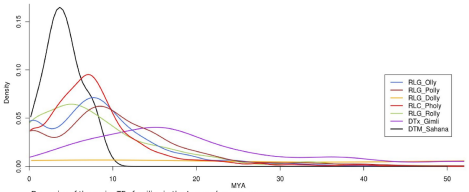
5 History of genome invasion by Transposable Elements

30% of the genome is made of TEs

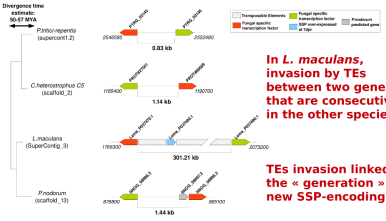


TEs repartition in *L. maculans* genome

Dating divergence times indicate a « recent » TEs invasion (4-20 MYA) posterior to the separation of *L. maculans* from other species (50-57 MYA)



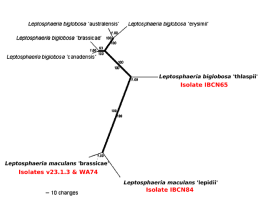
Dynamics of the major TEs families in the *L. maculans* genome



In *L. maculans*, invasion by TEs between two genes that are consecutives in the other species

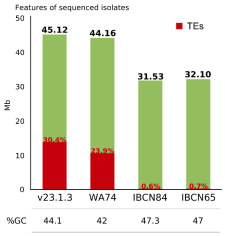
TEs invasion linked with the « generation » of a new SSP-encoding gene

6 Comparative genomics within the species complex *L.maculans – L. biglobosa*



Phylogenetic tree of the species complex

Features of sequenced isolates



Presence of the SSP-encoding genes of v23.1.3 in the sequenced isolates

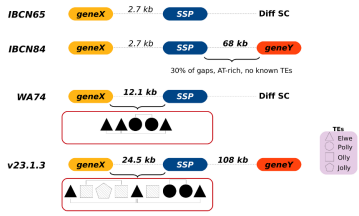
	% SSPs from GC-equilibrated regions	% SSPs from borders of AT-rich regions	% SSPs from within AT-rich regions
WA74	84.3	89.2	65
IBCN84	59.4	**43.1**	**14**
IBCN65	41.4	**20**	**15.8**

In *L. maculans* 'lepidii' and *L. biglobosa* 'thlaspii', 18-30% of the SSP-encoding genes from AT-isochores of v23.1.3 are present

whereas 41-56% of the other genes from the same compartments of v23.1.3 are present

SSP-encoding genes from GC-isochores of v23.1.3 are also present at 41-60%

Microsynteny of a conserved SSP from within the AT-isochores



Ancestral TE mosaics in v23.1.3 and WA74 (black shapes) between *geneX* and *SSP* with new inserted TEs in v23.1.3 (hatched shapes)

L. maculans 'lepidii' and L. biglobosa 'thlaspii':
- **smaller genome,**
- **lower proportion of known TEs**
- **higher GC content**
- **no isochore structure**

7 Our current evolutionary scenario

1/ Massive and recent invasion by a few families of TEs

2/ Waves of overlapping transposition with probable duplication of genes

3/ Duplicated copies of TEs and genes within TE-rich regions underwent the RIP mechanism

4/ Inactivation of TEs and over-diversification of the genes

5/ Selection pressure to maintain effectors genes in AT-isochores and the structure of these latter

? Were TEs originally targeted to pre-existing effector-rich genome regions ?
Origins of the effectors (maybe Large Genomic Transfert events) ?
Regulation of effector genes expression in AT-isochores ?

References

Rouxel, T. *et al.* Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat. Commun.* 2:202 doi:10.1038/ncomms1189 (2011).

Voigt, K. *et al.* Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and β-tubulin sequences. *Mol. Phylogenet. Evol.* 37(2):541-57 (2005).

Contacts

Jonathan Grandaubert: jonathan.grandaubert@versailles.inra.fr
Thierry Rouxel: rouxel@versailles.inra.fr

INRA-BIOGER, Campus AgroParisTech
Avenue Lucien Brétignières
78850 Thiverval-Grignon, FRANCE

Poster 2. Présenté lors du congrès Comparative Genomics of Eukaryotic Microorganisms à San Feliu de Guixols (Espagne) le 18 octobre 2011.

Are Transposable Elements Drivers of Effector Birth and Diversification in *Leptosphaeria* Species?

J Grandaubert[1], MH Balesdent[1], H Borhan[2] and T Rouxel[1]

[1] INRA-Bioger, Grignon, France ; [2] AAF Saskatoon, Canada

Leptosphaeria maculans and *L. biglobosa* are part of a species complex of fungal pathogens of crucifers. The genomes of two *L. maculans* 'brassicae' (Lmb) isolates (45.12 Mb, assembled into 76 scaffolds and 44.16 Mb assembled into 986 scaffolds, respectively) have an unusual bipartite structure – alternating distinct GC-equilibrated and AT-rich blocks of homogenous nucleotide composition. The AT-rich blocks comprise one third of the genome and contain effector genes and families of transposable elements (TEs), postulated to have recently invaded the genome, both of which are affected by Repeat Induced Point mutation. In silico comparison of the Lmb genomes with that of *L. maculans* 'lepidii' (31.53 Mb, assembled into 123 scaffolds) and *L. biglobosa* 'thlaspii' (32.10 Mb, assembled into 237 scaffolds), shows these species have a much more compact genome with a very low amount of TEs (<1%). In addition some recently expanded TE families are specific of *L. maculans* isolates. Compared to the Lmb genomes, less than 14% of the effector genes and 33% of other genes in AT-blocks are present in the two other genomes, suggesting TEs were key players in gene innovation and that the genome environment promoted rapid sequence diversification and selection of genes involved in pathogenicity.

Are transposable elements drivers of effector birth and diversification in *Leptosphaeria* species ?

Jonathan Grandaubert[1], Marie-Hélène Balesdent[1], Hossein Borhan[2], Matthew Links[2], Thierry Rouxel[1]

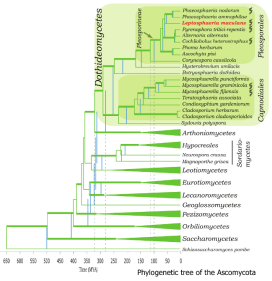
[1]INRA-BIOGER, Thiverval-Grignon, France. [2]AAF Saskatoon, Canada.

1 *Leptosphaeria maculans* 'brassicae'

Leptosphaeria maculans is an Ascomycete, class Dothideomycetes, closely related to three already-sequenced fungi: *Phaeosphaeria nodorum*, *Pyrenophora tritici-repentis* and *Cochliobolus heterostrophus*.

L. maculans and *L. biglobosa* are members of a species complex of fungal pathogens of crucifers of which some are collectively responsible for the stem canker disease (also termed « blackleg ») of oilseed rape (*Brassica napus*).

The sequencing of *L. maculans* genome highlights the molecular and evolutionary trends which allow the fungus to adapt rapidly to novel host-derived constraints.



Phylogenetic tree of the Ascomycota

	L. maculans	*P. (Stagonospora) nodorum*	*P. tritici-repentis*	*C. heterostrophus*	*A. brassicicola*	*M. graminicola*
No. of chromosomes	17-18	19	11	15-16	9-11	21
Genome size (Mb)	45.1	36.6	37.8	34.9	30.3	39.7
No of contigs	1,743	496	703	400	4,039	21
No of SuperContigs (SCs)	76	107	47	89	838	21
SC N50 (Mb)	1.8	1.1	1.9	1.3	2.4	NA
Gaps (%)	2.5	0.4	1.7	1.1	5.4	0.01
No. of predicted genes	12,469	10,762	12,141	9,633	10,688	10,952
Average gene length (bp)	1,323	1,326	1,618	1,836	1,523	1,600
GC content (%)	44.1	50.3	50.4	52-54	50.5	55.0
Repeat content (%)	34.2	7.1	16.0	7.0	9.0	18.0
'Core' genome size (Mb)	29.7	34.5	31.7	32.5	27.6	32.6
Gene density/core genome (no. of gene per 10 kb)	4.2	3.1	3.8	3.0	3.9	3.4

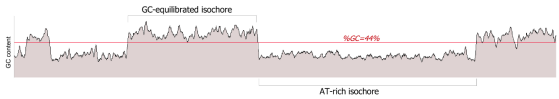
Genome features of *L. maculans* 'brassicae' JN3 and other related Dothideomycetes

Lower GC content
Larger genome size
Higher repeat content

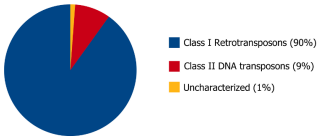
➡ **Invasion of the genome by Transposable Elements (TEs)**

2 A bipartite-structured genome

Alternance of large GC-equilibrated and AT-rich regions all along the chromosomes



AT-isochores represent 36% of the genome and are mainly composed of mosaics of truncated and RIPped* TEs.



TEs repartition in *L. maculans* 'brassicae' genome

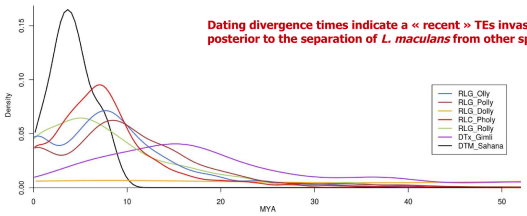
AT-isochores only comprise 5% of the predicted genes and are enriched in genes encoding putative effectors* (~20% vs. 4% in GC-isochores).

Effector-encoding genes within AT-isochores show a biased RIP index:
➡ **Overrun of the RIP machinery from TEs to the adjacent single-copy genes**

* RIP (Repeat Induced Point mutation) is a premeotic repeat-inactivation mechanism specific to fungi mutating the dinucleotide CpA into TpA and the dinucleotide TpG into TpA.

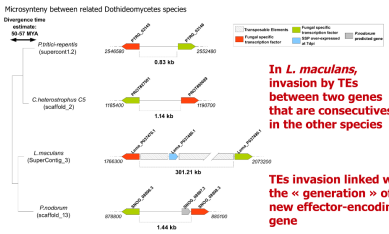
* Effectors are proteins playing key role in pathogenicity. They are usually small secreted proteins (SSP), cysteine rich and their genes are overexpressed during the primary infection.

3 History of *L. maculans* 'brassicae' genome invasion by Transposable Elements



Dating divergence times indicate a « recent » TEs invasion (4-20 MYA) posterior to the separation of *L. maculans* from other species (50-57 MYA)

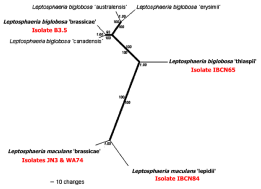
Dynamics of the major TEs families in the *L. maculans* genome



In *L. maculans*, invasion by TEs between two genes that are consecutives in the other species

TEs invasion linked with the « generation » of a new effector-encoding gene

4 Comparative genomics within the *L. maculans* – *L. biglobosa* species complex



Phylogenetic tree of the species complex

4 new isolates of the complex have been sequenced using 454 Roche and Illumina.

The sequences were assembled with Newbler.

Gene models were predicted using a combination of Fgenesh and Genemark.

TEs were identified with REPET.

	JN3 *L. maculans* 'brassicae'	WA74 *L. maculans* 'brassicae'	IBCN84 *L. maculans* 'lepidii'	IBCN65 *L. biglobosa* 'thlaspii'	B3.5 *L. biglobosa* 'brassicae'
Genome size (Mb)	45.12	44.16	31.53	32.1	31.79
No of scaffold	76	986	123	237	606
No of contig	1743	3765	2802	3506	2533
SC N50 (Mb)	1769.6	263.0	1356.3	715.1	779.1
Gaps (%)	2.5	9.6	7.1	8.7	7.4
GC content (%)	44.1	42.0	47.3	46.9	47.6
TEs (%)	30.4	26.8	1.9	3	3.5
No of gene models	12469	10624	11272	11691	11390

Genome features of the sequenced isolates of the complex

***L. maculans* 'lepidii', *L. biglobosa* 'thlaspii'/'brassicae' have :**
- smaller genome
- lower proportion of TEs
- higher GC content
- no isochore structure

***L. maculans* 'lepidii' and *L. biglobosa* 'thlaspii' are not pathogenic to oilseed rape.**

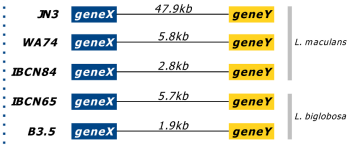
In the new sequenced isolates, only 13-30% of the effector-encoding genes present in AT-isochores of JN3 are found.

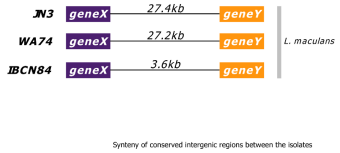
Whereas 44-61% of the other genes from the same compartments in JN3 are found.

Effector-encoding genes from GC-isochores of JN3 are also present at 52-72%

	JN3 *L. maculans* 'brassicae'	WA74 *L. maculans* 'brassicae'	IBCN84 *L. maculans* 'lepidii'	IBCN65 *L. biglobosa* 'thlaspii'	B3.5 *L. biglobosa* 'brassicae'
No of putative effectors	577	632	737	676	665
Mean size (aa)	155.95	152.32	144.84	152.91	153.82
Mean Cysteine proportion (%)	3.1	2.69	2.75	2.79	2.6
Ratio Cysteine in effectors/all predictions	1.82	1.89	1.88	1.89	1.79
Effectors in AT-isochores of JN3 (%)	-	30.33	17.21	18.03	13.11
Others in AT-isochores of JN3 (%)	-	61.45	50.4	44.38	44.98
Effectors in GC-isochores of JN3 (%)	-	72.97	63.14	54.06	52.17
Others in GC-isochores of JN3 (%)	-	76.32	72.19	67.37	66.94

Effectors features and conservation of the genes of the reference isolate JN3 in the new sequenced isolates





Synteny of conserved intergenic regions between the isolates

Synteny analysis of intergenic regions between the isolates will help us to define more precisely the date of TE invasion.

?

Could this TE invasion be interpreted as a speciation (or subspeciation) event ?

What's the link between Transposable Elements, effectors and pathogenicity ?

What's the origin of the isolate-specific effectors ?

References

Rouxel, T., Grandaubert, J. *et al.* Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat. Commun.* 2:202 (2011).

Voigt, K. *et al.* Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and β-tubulin sequences. *Mol. Phylogenet. Evol.* 37(2)(2005).

Contacts

Jonathan Grandaubert: jonathan.grandaubert@versailles.inra.fr
Thierry Rouxel: rouxel@versailles.inra.fr

INRA-BIOGER, Campus AgroParisTech
Avenue Lucien Brétignières
78850 Thiverval-Grignon, FRANCE

Poster 3. Présenté par Thierry Rouxel lors du 2012 Plant Genomics Seminar à Pont-Royal en Provence (France) le 3-5 avril 2012.

FungIsochores

Isochores and effectors : genome reshaping and the birth of highly pathogenic species in fungal phytopathogens

J. Grandaubert[1], M.H. Balesdent[1], I. Fudal[1], J. Amselem[2], J. Kreplak[2], N. Lapalu[2], B. Le Cam[3], C. Lemaire[3], T. Guillemette[3], T. Rouxel[1]

[1] INRA-BIOGER, Thiverval-Grignon ; [2] INRA-URGI, Versailles ; [3] IRHS, Angers (INRA-ACO-Univ.Angers)

Background and Objectives

FungIsochores develops a comparative and evolutionary genomics approach to assess the role of fungal genome reshaping following massive transposable element (TE) invasion on generation of novel species better adapted to new hosts or with increased fitness on a given host plant.

The original fungal model for this study is a pathogen of oilseed rape, *Leptosphaeria maculans* 'brassicae' (Figure 1), whose genome sequence analysis strongly suggested the following events in the course of evolution :

i. massive invasion of the genome by TEs linked with a probable incidence on acquisition of novel effector-encoding genes
ii. TE degeneracy by repeat-induced point mutations (RIP) generating a compartmentalised genome into isochores (Figure 3)
iii. diversification of effector-encoding genes following mild RIP mutation

A second plant pathogenic fungus, *Venturia inaequalis*, the agent of apple scab (Figure 2) is also concerned since preliminary sequence data indicated that its genome is structured into isochores that were postulated to specifically host effector-encoding genes.

This project aims at sequencing and analysing the genomes of three members of the *L. maculans – L. biglobosa* species complex, choosen because they show a divergent adaptation towards oilseed rape and for which preliminary data indicated a low level of invasion by TEs, and at contributing to the sequencing of *V. inaequalis* to validate that its genome is structured into isochores.

Figure 1. Oilseed rape stem canker caused by *Leptosphaeria maculans* 'brassicae'.

Figure 2. Apple scab caused by *Venturia inaequalis*.

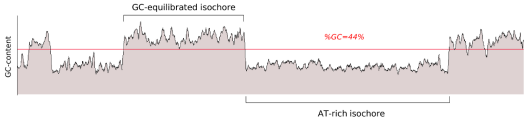


Figure 3. The isochores-structured genome of *Leptopshaeria maculans* 'brassicae'.

Results

- Three members of the *Leptosphaeria* species complex, *L. maculans* 'lepidii', *L. biglobosa* 'thlaspii' and *L. biglobosa* 'brassicae', have been sequenced, assembled and annotated. All isolates have a compact genome sized 31-32 Mb and show a low TE content, only 2-3.5%, compared to 30% in the *L. maculans* 'brassicae' genome (Table 1).

  Comparative genomics between the *Leptosphaeria* genomes indicates a high conservation of chromosomal synteny. This is mainly the case between *L. maculans* 'lepidii' and *L. maculans* 'brassicae' (Figure 4) for which gene order and content are extremely conserved whereas sequence divergence between orthologues is important.

  Very few families of TEs are common between the different species and most of TE invasion took place after the separation between *L. maculans* and *L. biglobosa*, and was not accompanied by massive chromosomal rearrangements. This invasion may have favoured reproductive isolation and recent speciation between the weakly pathogenic *L. maculans* 'lepidii' and the highly pathogenic *L. maculans* 'brassicae'.

  Comparative genomics between the *Leptosphaeria* genomes indicate a highly divergent content in effector-encoding genes.

  While a comparable number of predicted effector-encoding genes is present in the different isolates (ranging from 650 to 740), only 20% are common to all isolates and up to 40% of these genes are isolate- or species-specific. This number is even higher when analysing effector genes hosted in TE-rich genomic landscapes, and TE invasion is shown to be accompanied by « generation » of novel effector genes in a few cases.

	L. maculans 'brassicae'	*L. maculans* 'lepidii'	*L. biglobosa* 'thlaspii'	*L. biglobosa* 'brassicae'
Genome size (Mb)	**45.12**	**31.53**	**32.1**	**31.79**
No of scaffold	76	123	237	606
SC N50 (Mb)	1769.6	1356.3	715.1	779.1
GC content (%)	44.1	47.3	46.9	47.6
TEs (%)	**30.4**	**1.9**	**3**	**3.5**
No of gene models	12543	11272	11691	11390
No of putative effectors	**651**	**737**	**676**	**665**

Table 1. Genome statistics of the sequenced members of the *Leptosphaeria* species complex.

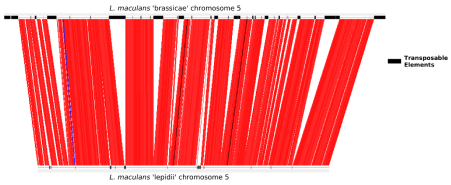


Figure 4. Chromosomal synteny between *L. maculans* 'brassicae' and *L. maculans* 'lepidii'.

- The genome sequence of *V. inaequalis* substantiate our initial postulate that it is structured into contrasted isochores (Figure 5) as the genome of *L. maculans* 'brassicae' is. The genome assembly covers 73.2 Mb, which is much higher than most currently known fungal genomes (45 Mb for *L. maculans* 'brassicae'). But it is still very fragmented and large TE-rich regions are poorly assembled or completely missing.

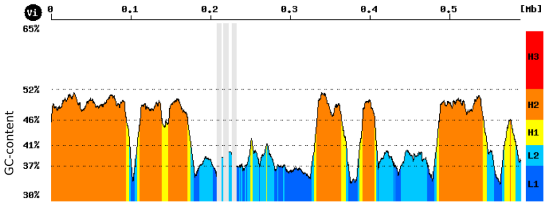


Figure 5. The genome of *V. inaequalis* is structured into isochores (exemple of scaffold00008).

- Automated annotation and setting up of the genome browser has been done for *V. inaequalis*, and a comparative genomics browser (Figure 6) is currently being setup for all *Leptosphaeria* isolates.

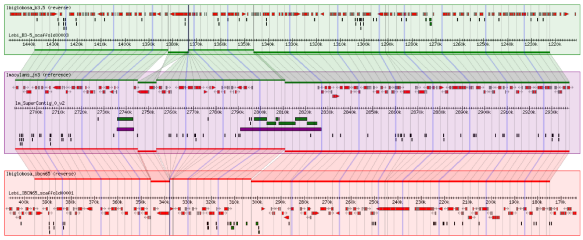


Figure 6. The synteny browser dedicated to sequenced isolates of the *Leptosphaeria* species complex.

Perspectives

Improvment of *V. inaequalis* genome assembly of TE-rich regions for a better visualisation of genes hosted within this genomic environment and analysis of TE families, TE nesting and dating of transposition events.

Evaluation of the incidence of RIP on TE degeneracy and diversification of effector-encoding genes.

Within the *Leptosphaeria* species complex, generation of accurate phylogenies and dating of speciation times (coll. C.L. Schoch, NCBI).

Generation of an extensive repertoire of effectors in all species/isolates and elucidation of their origin and expansion/diversification mechanisms.

Validation of genomes annotation by transcriptomic analysis (microarray).

References

Rouxel, T., Grandaubert, J. *et al.* Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat. Commun.* 2:202 (2011).

Contacts

Jonathan Grandaubert: jonathan.grandaubert@versailles.inra.fr
Thierry Rouxel: rouxel@versailles.inra.fr

INRA-BIOGER, Campus AgroParisTech
Avenue Lucien Brétignières
78850 Thiverval-Grignon, FRANCE

AGENCE NATIONALE DE LA RECHERCHE
ANR
FungIsochores
ANR-09-GENM-028
01.01.2010 - 01.07.2013

Poster 4. Présenté lors du 27th Fungal Genetics Conference à Asilomar (USA) le 14 mars 2013.

Leptosphaeria maculans 'brassicae' : « Transposable Elements changed my life, I feel different now »

J Grandaubert[1], CL Schoch[2], H Borhan[3] , BJ Howlett[4] and T Rouxel[1]

[1] INRA-Bioger, Grignon, France ; [2] NCBI-NIH, Bethesda, USA ;[3] AAF Saskatoon, Canada ; [4] University of Melbourne, Australia

The Dothideomycetes phytopathogens *Leptosphaeria maculans* and *Leptosphaeria biglobosa* form a complex of 8 species and putative subspecies suggested to have diverged “recently”. In 2007, the sequencing of an isolate of *Leptosphaeria maculans* 'brassicae' (Lmb) provided the first reference genome for this fungus. The 45-Mb genome has an unusual bipartite structure, alternating large GC-equilibrated and AT-rich regions. These AT-rich regions comprise one third of the genome and are mainly composed of mosaics of truncated Transposable Elements (TEs) postulated to have “invaded” the genome 5-10 MYA; they also comprise 5% of the predicted genes of which 20% encode putative effectors. In these regions, both genes and TEs are affected by Repeat Induced Point mutation (RIP). To investigate when and how genome expansion took place in the evolutionary series, and the consequences it had on fungal adaptability and pathogenicity, the genomes of five members of the species complex showing contrasted host range and infection abilities were sequenced. *In silico* comparison of the reference genome with that of 30-32-Mb genome of *L. maculans* 'lepidii' (Lml), *L. biglobosa* 'brassicae', *L. biglobosa* 'thlaspii' and *L. biglobosa* 'canadensis', showed these species have a much more compact genome with a very low amount of TEs (<4%). The TE annotation allowed us to identify 121 TE families, all RIP-affected including the expected presence of lineage-specific TEs. Unexpectedly, two of the most expanded TE families in Lmb have been present in the Dothideomycete lineage for 100 million years. This questions how these families, while they have been anciently RIPped, managed to expand recently in Lmb. Interestingly, the comparison between the TE-rich genome of Lmb and the TE-poor genome of Lml, estimated to have diverged 5.5 MYA, indicated a nearly perfect synteny at the chomosomal level, suggesting low incidence of TE expansion on genome reorganisation. The gene annotation produced a similar gene number in each genome (~11000), but compared to the reference genome, less than 20% of the effector genes and 50% of other genes in AT-rich regions are present in the other genomes, suggesting that TEs were key players in gene innovation and that the genome environment promoted rapid sequence diversification and selection of genes involved in pathogenicity.

Leptosphaeria maculans 'brassicae': « Transposable Elements changed my life, I feel different now ! »

280

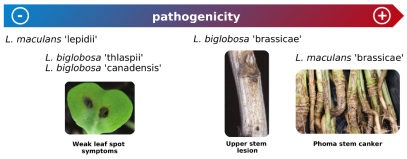
Jonathan Grandaubert[1], Conrad L. Schoch[2], Hossein Borhan[3], Barbara J. Howlett[4] & Thierry Rouxel[1]

[1] INRA-BIOGER, Thiverval-Grignon, France ; [2] NCBI-NIH, Bethesda, USA ; [3] AAF, Saskatoon, Canada ; [4] University of Melbourne, Australia

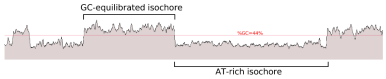
1 The *Leptosphaeria* species complex

Leptosphaeria maculans spp. and *Leptosphaeria biglobosa* spp. form a species complex of fungal pathogens of crucifers. The *Leptosphaeria* species complex belongs to the phylum Ascomycota, class Dothideomycetes which encompasses numerous plant pathogens causing serious crop losses such as species in the genera *Cochliobolus*, *Phaeosphaeria and Pyrenophora* (see Phylogeny).

L. maculans 'brassicae' shows the best adaptation towards oilseed rape (*Brassica napus*) among all members of the species complex.



In 2007, the whole genome sequencing of *L. maculans* 'brassicae' (Lmb), highlighted a peculiar isochore-like structure of the genome.



AT-isochores represent 36% of Lmb genome.
They are mainly composed of RIPped Transposable Elements (TEs) suggested to have expanded in the genome 4 to 20 MYA.

They comprise 5% of the predicted genes of which 20% encode putative effectors.
Cloned effectors in Lmb (AVR genes) are located in AT-isochores.

Genes in AT-isochores are also affected by RIP.

In 2012, the genomes of five members of the complex were sequenced using NGS technologies.

	L. maculans 'brassicae' v23.1.3	*L. maculans* 'lepidii' WA74	*L. biglobosa* 'brassicae' IBCN84	*L. biglobosa* 'canadensis' B3.5	*L. biglobosa* 'thlaspii' J154	*L. biglobosa* 'thlaspii' IBCN65
Genome size (Mb)	45.1	44.2	31.5	31.8	30.2	32.1
Scaffold number	76	986	123	606	6748	237
Scaffold N50 (kb)	1770	263	1356	779	245	715
Gaps (%)	2.5	9.6	7.1	7.4	0.1	8.7
Repeats (%)	35.5	27.5	4.0	4.4	3.9	5.1
TEs (%)	32.5	25.8	2.7	3.2	2.9	4.0
GC-content (%)	45.2	46.5	50.9	51.4	51.1	51.4
Predicted gene number	12543	10624	11272	11390	11068	11691

Smaller genome size for Lml and *L. biglobosa* spp. due to a lower proportion of TEs compared to Lmb.

➡ No isochore structure and a higher GC-content.

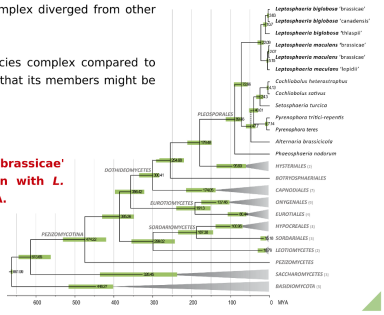
Expansion of TEs occured only in *L. maculans* 'brassicae'

2 Phylogeny and divergence time

The *Leptosphaeria* species complex diverged from other Pleosporales 73 MYA.

Divergence times within the species complex compared to those of other genera suggested that its members might be different species.

TEs expanded in *L. maculans* 'brassicae' genome, after the separation with *L. maculans* 'lepidii' (Lml), 5 MYA.

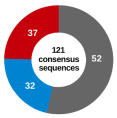


3 TE annotation in *Leptosphaeria*

Identification with the REPET pipeline + manual annotation:

Class I Retrotransposons (LTR, LINE) Clusterised in 24 families

Class II DNA transposons (TIR, MITE) Clusterised in 21 families



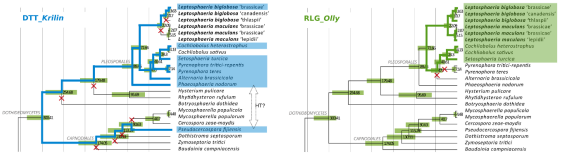
Non-categorized repeats

	L. maculans 'brassicae' v23.1.3	*L. maculans* 'lepidii' WA74	*L. biglobosa* 'brassicae' IBCN84	*L. biglobosa* 'canadensis' B3.5	*L. biglobosa* 'thlaspii' J154	*L. biglobosa* 'thlaspii' IBCN65
TE content (%)	32.5	25.8	2.7	3.2	2.9	4.0
% Retrotransposon	83.3	80.8	50.1	49.2	48.7	39.5
% DNA transposon	8.8	7.5	24.3	31.6	14.5	11.6
% NoCat	7.9	11.7	25.6	19.2	36.8	48.9

Large expansion of retrotransposons in *L. maculans* 'brassicae' genome.

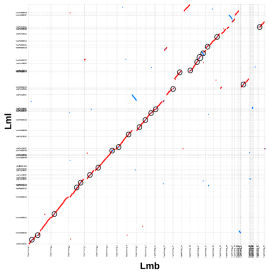
4 Evolutionary dynamics of TEs

66% of the TE families identified in the *Leptosphaeria* clade are specific to this clade. Two families are found in the Capnodiales, but their presence is unclear and might be due to Horizontal Transfer (*e.g.* DTT_*Krilin* below). The other families are majoritarily found within the Pleosporales (90 MYA) and some of them correspond to largely expanded TE families in Lmb (*e.g.* RLG_*Olly* below).

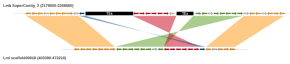


How did families subjected to RIP for million years manage to expand recently in Lmb ?

5 TE incidence on genome structure



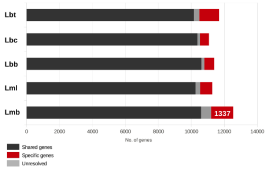
There is a high level of synteny when the TE-cleared genomes of Lmb and Lml are aligned, with only 30 inversions (≥1 kb) (example below).



70% of the inversions have TEs (mainly Lmb-specific) in their vicinity within a 100 bp range.

These inversions are the first step towards mesosynteny.

6 Species-specific genes and TEs in Lmb



Species-specific genes might be linked with the different pathogenic abilities of the members of the complex.

25% of genes in AT-isochores are specific to Lmb (*vs.* 10% in GC-isochores)

In AT-isochores, 41% of effector genes are specific to Lmb (*vs.* 22% for the other genes)

What is the origin of specific effector genes within AT-isochores in Lmb ?

?

- What did trigger the TE expansion in Lmb (or prevent it in the others species) ?
- Is TE expansion responsible for the speciation event between Lmb and Lml ?
- How did TEs present in the Dothideomycete lineage for a long time (*i.e.* more subjected to inactivation) manage to expand recently in Lmb ?
- Are the specific effector genes in AT-isochores the products of HGT or duplication followed by diversification ?

Contacts

Jonathan Grandaubert: jonathan.grandaubert@versailles.inra.fr
Thierry Rouxel: rouxel@versailles.inra.fr

INRA-BIOGER, Campus AgroParisTech
Avenue Lucien Brétignières
78850 Thiverval-Grignon, FRANCE

References

Rouxel, T., Grandaubert, J. *et al.* Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat. Commun.* 2:202 (2011).

Thanks to the Fungal Conference Financial Aid for the financial support.

RÉFÉRENCES

Agrios GN. 2005. *Plant pathology*. 5th ed. Burlington, MA, San Diego, CA: Elsevier Academic Press.

Ambrozová K, Mandáková T, Bures P, Neumann P, Leitch IJ, Koblízková A, Macas J, Lysak MA. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria lilies*. *Ann Bot* **107:** 255-268.

Ansan-Melayah D, Balesdent MH, Buée M, Rouxel T. 1995. Genetic characterization of *AvrLm1*, the first avirulence gene of *Leptosphaeria maculans*. *Phytopathology* **85:** 1525-1529.

Balesdent MH, Gall C, Robin P, Rouxel T. 1992. Intraspecific variation in soluble mycelial protein and esterase patterns of *Leptosphaeria maculans* French isolates. *Mycol Res* **96:** 677-684.

Balesdent MH, Fudal I, Ollivier B, Bally P, Grandaubert J, Eber F, Chèvre AM, Leflon M, Rouxel T. 2013. The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytol* **198:** 887-898.

Bhattacharyya MK, Smith AM, Ellis TH, Hedley C, Martin C. 1990. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell* **60:** 115-122.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Curr Opin Genet Dev* **15:** 621-627.

Biémont C. 2010. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **186:** 1085-1093.

Biémont C, Vieira C, Hoogland C, Cizeron G, Loevenbruck C, Arnault C, Carante JP. 1997. Maintenance of transposabel copy element number in natural populations of *Drosophila melanogaster* and *D. simulans*. *Genetica* **100:** 161-166.

Biémont C & Vieira C. 2006. Junk DNA as an evolutionary force. *Nature* **443:** 521-524.

Birren B, Fink G, Lander E. 2002. Fungal genome initiative: white paper developed by the fungal research community. Cambridge, MA: Whitehead Institute Center for Genome Research.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al*. 2004. Aligning multiple genomic sequences with the threaded blocks*et al*igner. *Genome Res* **14:** 708-715.

Botstein D, Chervitz SA, Cherry JM. 1997. Yeast as a model organism. *Science* **277:** 1259-1260.

Cannon PF. 1997. Strategies for rapid assessment of fungal diversity. *Biodivers Conserv* **6:** 669-680.

Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otillar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N *et al.* 2013. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genet* **9:** e1003233.

Cuomo CA, Güldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M *et al*. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317:** 1400-1402.

Daverdin G, Rouxel T, Gout L, Aubertot JN, Fudal I, Meyer M, Parlange F, Carpezat J, Balesdent MH. 2012. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PloS Pathog* **8:** e1003020.

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H *et al*. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434:** 980-986.

Delourme R, Chevre AM, Brun H, Rouxel T, Balesdent MH, *et al*. 2006. Major gene and polygenic resistance to *Leptosphaeria maculans* in oilseed rape (*Brassica napus*). *Eur Plant Pathol* **114:** 41-52.

Delprat A, Negre B, Puig M, Ruiz A. 2009. The transposon *Galileo* generates natural chromosomal inversions in *Drososphila* by ectopic recombination. *PloS One* **4:** e7883

Desprez-Loustau ML, Robin C, Buée M, Courtecuisse R, Garbaye J, Suffert F, Sache I, Rizzo DM. 2007. The fungal dimension of biological invasions. *Trends Ecol Evol* **22:** 472-480.

Dilmaghani A, Balesdent MH, Didier JP, Wu C, Davey J, Barbetti MJ, Li H, Moreno-Rico O, Phillips D, Despeghel JP *et al*. 2009. The *Leptosphaeria maculans – Leptosphaeria biglobosa* species complex in the American continent. *Plant Pathology* **58:** 1044-1058.

Dilmaghani A, Gladieux P, Gout L, Giraud T, Brunner PC, Stachowiak A, Balesdent MH, Rouxel T. 2012. Migration patterns and changes in population biology associated with the worldwide spread of the oilseed rape pathogen *Leptosphaeria maculans*. *Mol Ecol* **21:** 2519-2533.

Drosophila 12 genomes consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203-218.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10:** 285-311.

Elliott CE & Howlett BJ. 2006. Overexpression of a 3-ketoacyl-CoA thiolase in *Leptosphaeria maculans* causes reduced pathogenicity on *Brassica napus*. *Mol Plant-Microbe Interact* **19:** 588-596.

Elliott CE, Gardiner DM, Thomas G, Cozijnsen A, Van De Wouw A, Howlett BJ. 2007. Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. *Mol Plant Pathol* **8:** 791-802.

Elliott CE, Callahan DL, Schwenk D, Nett M, Hoffmeister D, Howlett BJ. 2013. A gene cluster responsible for biosynthesis of phomenoic acid in the plant pathogenic fungus, *Leptosphaeria maculans*. *Fungal Genet Biol* **53:** 50-58.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9:** 397-405.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet* **5:** 103-107.

Fitt BDL, Brun H, Barbetti MJ, Rimmer SR. 2006. Worldwide importance of phoma stem canker (*Leptosphaeria maculans* and *L. biglobosa*) on oilseed rape (*Brassica napus*). *Eur J Plant Pathol* **114:** 3-15.

Flor AH. 1955. Host-parasite interactions in flax rust-its genetics and other implications. *Phytopathology* **45:** 680-685.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3:** 722-732.

Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, Eckert MR, Cattolico L, Bernard-Samain S, Balesdent MH, Rouxel T. 2007. Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. *Mol Plant-Microbe Interact* **20:** 459-470.

Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, Balesdent MH, Rouxel T. 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. *Mol Plant Microbe Interact* **22:** 932-941.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422:** 859-868.

Galagan JE & Selker EU. 2004. RIP: the evolutionary cost of genome defense. *Trends Genet* **20:** 417-423.

Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B. 2005a. Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res* **15:** 1620-1631.

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Baştürkmen M, Spevak CC *et al*. 2005b. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438:** 1105-1115.

Gardiner DM, Cozijnsen AJ, Wilson LM, Pedras MS, Howlett BJ. 2004. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans*. *Mol Microbiol* **53:** 1307-1318.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al.* 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498-511.

Glawe DA. 2008. The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. *Annu Rev Phytopathol* **46:** 27-51.

Gentles AJ, Wakefield MJ, Kohany O, Gu W, Batzer MA, Pollock DD, Jurka J. 2007. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res* **17:** 992-1004.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*. 1996. Life with 6000 genes. *Science* **274:** 563-567.

Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, Balesdent MH, Rouxel T. 2006. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Mol Microbiol* **60:** 97-80.

Gregory TR. 2013. Animal Genome Size Database. http://www.genomesize.com.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest Angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size. *Plant Biology* **8:** 770-777.

Haas BJ, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461:** 393-398.

Hammond KE & Lewis BG. 1987. Variation in stem infections caused by aggressive and nonaggressive isolates of *Leptosphaeria maculans* on *Brassica napus* var *oleifera*. *Plant Pathology* **36:** 53-65.

Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE *et al*. 2007. Dothideomycete–plant Interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum. Plant Cell* **19:** 3347-3368.

Hawksworth DL. 2012a. Global species numbers of fungi: are tropical studies and molecular approaches contributing to a more robust estimate? *Biodiversity and Conservation* **21:** 2425–2433.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L *et al.* 2010. The genome of the western clawed frog *Xenopus tropicalis*. *Science* **328:** 633-636.

Hock B. 2001. The Mycota, vol 9. Fungal associations [B. Hock (ed)]. Springer, Berlin Heidelberg New York. XVII, 250 pp.

Horbach R, Navarro-Quesada AR, Knogge W, Deising HB. 2011. When and how to kill a plant cell: infection strategies of plant pathogenic fungi. *J Plant Physiol* **168:** 51-62.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L *et al.* 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496:** 498-503.

Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. 2011. The struggle for life of the genome's selfish architects. *Biol Direct* **6:** 19.

I.B.G.S. Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491:** 711-716.

I.C.G.S. Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695-716.

Idnurm A & Howlett BJ. 2001. Pathogenicity genes of phytopathogenic fungi. *Mol Plant Pathol* **2:** 241-255.

Idnurm A & Howlett BJ. 2002. Isocitrate lyase is essential for pathogenicity of the fungus *Leptosphaeria maculans* to canola (*Brassica napus*). *Eukaryot Cell* **1:** 719-724.

Idnurm A, Warnecke DC, Heinz E, Howlett BJ. 2003a. Characterization of neutral trahalase and UDP-glucose:sterol glucosyltransferase genes from the plant pathogenic fungus *Leptosphaeria maculans*. *Physiol Mol Plant Pathol* **62:** 305-313.

Idnurm A, Taylor JL, Pedras MSC, Howlett BJ. 2003b. Small scale functional genomics of the blackleg fungus, *Leptosphaeria maculans*: analysis of a 38 kb region. *Australas Plant Pathol* **32:** 511-519.

Initiative, T.A.G. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Initiative, T.I.B. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463:** 763-768.

Johnson RD & Lewis BG. 1994. Variation in host-range, systemic infection and epidemiology of *Leptosphaeria maculans*. *Plant Pathology* **43:** 269-277.

Kamoun S. 2007. Groovy times: filamentous pathogen effectors revealed. *Curr Opin Plant Biol* **10:** 358-365.

Kämper J, Kahmann R, Bölker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O *et al.* 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444:** 97-101.

Kapitonov VV & Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PloS Biol* **3:** e181

Kapitonov VV & Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet* **23:** 521-529.

Kirk P, Cannon PF, Minter DW, Stalpers JA. 2008. Ainsworth & Bisby's Dictionary of the Fungi. 10th ed. CAB International, Wallingford, UK.

Kloppholz S, Kuhn H, Requena N. 2011. A secreted fungal effector of *Glomus intraradices* promotes symbiotic biotrophy. *Curr Biol* **21:** 1204-1209.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* **304:** 982.

Kubicek CP & Druzhinina IS. 2007. Environmental and Microbial Relationships. Heidelberg, Springer.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lewis ZA, Honda S, Khlafallah TK, Jeffress JK, Freitag M, Mohn F, Schübeler D, Selker EU. 2008. Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. *Genome Res* **19:** 427-437.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K *et al*. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20:** 265-272.

Maside X, Bartolomé C, Assimacopoulos S, Charlesworth B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: in situ hybridization vs. Southern blotting data. *Genetical Research* **78:** 121-136.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* **36:** 344-355.

McGee DC & Petrie GA. 1978. Variability of *Leptosphaeria maculans* in relation to blackleg of oilseed rape. *Phytopathology* **68:** 47-52.

Mendes-Pereira E, Balesdent MH, Brun H, Rouxel T. 2003. Molecular phylogeny of the *Leptosphaeria maculans-L. biglobosa* species complex. *Mycol Res* **107:**1287-1304.

Mendgen K & Hahn M. 2002. Plant infection and the establishment of fungal biotrophy. *Trends Plant Sci* **7:** 352-356.

Miller WJ, Hagemann S, Reiter E, Pinsker W. 1992. P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. *Proc Natl Acad Sci USA* **89:** 4018-4022.

Misumi O, Matsuzaki M, Nozaki H, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Yoshida Y, Kuroiwa H, Kuroiwa T. 2005. *Cyanidioschyzon merolae* genome. A tool for facilitating comparable studies on organelle biogenesis in photosyntethic eucaryotes. *Plant Physiol* **137:** 567-585.

Nekrutenko A & Li WH. 2001. Transposable elements are found in a large number oh human protein-coding genes. *Trends Genet* **17:** 619-621.

Ohno S. 1972. So much "junk" DNA in our genome. InH.H. Smith (Ed.), Proceedings of the 23rd Brookhaven Symposium on Biology, session "Evolution of Genetic Systems" (p. 366-370). Gordon & Breach, New York.

Oliver KR & Greene WK. 2012. Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. *Ecol Evol* **2:** 2912-2933.

Orgel LE & Crick FHC. 1980. Selfish DNA: the ultimate parasite. *Nature* **284:** 604-607.

Parlange F, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, Blaise F, Grezes-Besset B, Rouxel T. 2009. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents Rlm4-mediated recognition through a single amino acid change. *Mol Microbiol* **71:** 851-863.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* **4:** 675-682.

Pöggeler S & Wöstemeyer J. 2011. Evolution of Fungi and Fungal-Like Organisms. The Mycota, Vol. 14 Pöggeler, Stefanie; Wöstemeyer,Johannes (Eds.) 1st Edition., 2011, XIX, 345 p.

Porta-Puglia A & Vannacci G. 2011. Fungal plant diseases in Europe and in the Mediterranean basin. *Encyclopedia of Life Support Systems*.

Pritham EJ. 2009. Transposable elements and factors influencing their success in eucaryotes. *Journal of Heredity* **100:** 648-655.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci* **65:** 2571-2580.

Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: towards new species. *Gene* **454:** 1-7.

Rémy E, Meyer M, Blaise F, Chabirand M, Wolff N, Balesdent MH, Rouxel T. 2008a. The *Lmpma1* gene of *Leptosphaeria maculans* encodes a plasma membrane H+-ATPase isoform essential for pathogenicity towards oilseed rape. *Fungal Genet Biol* **45:** 1122-1134.

Rémy E, Meyer M, Blaise F, Simon UK, Kuhn D, Chabirand M, Riquelme M, Balesdent MH, Rouxel T. 2008b. The *Lmgpi15* gene, encoding a component of the glycosylphosphatidylinositol anchor biosynthesis pathway, is required for morphogenesis and pathogenicity in *Leptosphaeria maculans*. *New Phytol* **179:** 1105-1120.

Ridout CJ. 2009. Profiles in pathogenesis and mutualism: powdery mildews. In: Deising HB, editor. Plant Relationships V. Berlin, Heidelberg: Springer-Verlag. p.51–68.

Rouxel T, Mendes-Pereira E, Brun H, Balesdent MH. 2004. Species complex of fungal phytopathogens: the *Leptosphaeria maculans-L. biglobosa* case study. In: Plant Genome: Diversity and Evolution. Vol. 2: Cryptogams. AK Sharma & A Sharma (Eds.), Science Publishers, Inc., Enfield, USA, pp.33-75.

Rouxel T & Balesdent MH. 2005. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Mol Plant Pathol* **6:** 225-241.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S *et al*. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point Mutations. *Nat Commun* **2:** 202.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al.* 2009. The B73 maize genome: complexity, diversity and dynamics. *Science* **326:** 1112-1115.

Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C *et al*. 2009b. A class-wide phylogenetic assessment of Dothideomycetes. *Stud Mycol* **64:**1-15S10.

Sexton AC & Howlett BJ. 2000. Characterization of a cyanide hydratase gene in the phytopathogenic fungus *Leptosphaeria maculans*. *Mol Gen Genet* **263:** 463-470.

Sexton AC, Paulsen M, Woestemeyer J, Howlett BJ. 2000. Cloning, characterization and chromosomal location of three genes encoding host-cell-wall degrading enzymes in *Leptosphaeria maculans*, a fungal pathogen of *Brassica*. spp. *Gene* **248:** 89-97.

Shoemaker RA & Brun H. 2001. The telemorph of the weakly aggressive segregate of *Leptosphaeria maculans*. *Can J Bot* **79:** 412-419.

Sinzelle L, Izsvak Z, Ivics Z. 2009. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* **66:** 1073-1093.

Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ *et al*. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330:**1543-1546.

Spanu PD. 2012. The genomics of obligate (and nonobligate) biotrophs. *Annu Rev Phytopathol* **50:** 91-109.

Staples RC. 2000. Research on the rust fungi during the twentieth century. *Annu Rev Phytopathol* **38:** 49–69.

Stergiopoulos I & de Wit PJGM. 2009. Fungal effector proteins. *Annu Rev Phytopathol* **47:** 233–263.

Soanes DM, Richards TA, Talbot NJ. 2007. Insights from sequencing fungal and oomycete genomes: what can we learn about plant disease and the evolution of pathogenicity? *Plant Cell* **19:** 3318-3326.

Soanes DM, Alam I, Cornell M, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Oliver SG, Talbot NJ. 2008. Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PloS One* **3:** e2300.

Tyler BM & Rouxel T. 2013. Effectors of fungi and oomycetes: their virulence and avirulence functions and translocation from pathogen to host. In: Molecular Plant Immunity, Guido Sessa (Ed.) John Wiley & Sons, Inc. pp. 123-167.

Van de Lagemaat LN, Landry JR, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19:** 530-536.

Van de Wouw AP, Pettolino FA, Howlett BJ, Elliott CE. 2009. Mutations to *LmIFRD* affect cell wall integrity, development and pathogenicity of the ascomycete *Leptosphaeria maculans*. *Fung Genet Biol* **46:** 695–706.

Van Kan JA. 2006. Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends Plant Sci* **11:** 247-253.

Vincenot L, Balesdent MH, Li H, Barbetti MJ, Sivasithamparam K, Gout L, Rouxel T. 2008. Occurrence of a new subclade of *Leptosphaeria biglobosa* in Western Australia. *Phytopathology* **98:** 321-329.

Voegele RT, Hahn M, Mendgen K. 2009. The uredinales: cytology, biochemistry, and molecular biology. In: Deising HB, editor. Plant Relationships V. 2nd ed. Berlin, Heidelberg: Springer-Verlag. p. 69-98.

Voigt K, Cozijnsen AJ, Kroymann J, Pöggeler S, Howlett BJ. 2005. Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and β-tubulin sequences. *Mol Phylogenet Evol* **37:** 541-557.

West JS, Kharbanda PD, Barbetti MJ, Fitt BDL. 2001. Epidemiology and management of *Leptosphaeria maculans* (phoma stem canker) on oilseed rape in Australia, Canada and Europe. *Plant Pathology* **50:** 10-27.

West JS, Fitt BDL, Leech PK, Biddulph JE, Huang YJ, Balesdent MH. 2002. Effects of timing of *Leptosphaeria maculans* ascospore release and fungicide regime on phoma leaf spot and phoma stem canker development on winter oilseed rape (*Brassica napus*) in southern England. *Plant Pathology* **51:** 454-463.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O *et al.* 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Williams RH & Fitt BDL. 1999. Differentiating A and B groups of *Leptosphaeria maculans*, causal agent of stem canker (blackleg) of oilseed rape. *Plant Pathology* **48:** 161-175.

Wilson LM, Idnurm A, Howlett BJ. 2002. Characterization of a gene (*sp1*) encoding a secreted protein from *Leptosphaeria maculans*, the blackleg pathogen of *Brassica napus*. *Mol Plant Pathol* **3:** 487-493.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S *et al*. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415:** 871-880.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ *et al.* 2004. The genome of *Cryptosporidium hominis*. *Nature* **431:** 1107-1112.

Xu JR, Peng YL, Dickman MB, Sharon A. 2006. The dawn of fungal pathogen genomics. *Annu Rev Phytopathol* **44:** 337-366.

Zander M, Patel DA, Van de Wouw A, Lai K, Lorenc MT, Campbell E, Hayward A, Edwards D, Raman H, Batley J. 2013. Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Funct Integr Genomics* **13:** 295-308.

Zhao H & Bourque G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res* **19:** 934-942.

Résumé : *Leptosphaeria maculans* ‘brassicae’ (Lmb) est un champignon filamenteux de la classe des *Dothideomycètes* faisant partie du complexe d’espèces *Leptosphaeria maculans-Leptosphaeria biglobosa* composé d’agents pathogènes des crucifères. Lmb est particulièrement adapté au colza (*Brassica napus*) et provoque la maladie qui lui est la plus dommageable : la nécrose du collet. Dans le but de mieux comprendre et contrôler cette maladie, l’équipe d’accueil a initié un projet de génomique visant à identifier de façon systématique les gènes impliqués dans le pouvoir pathogène. Les premières données génomiques montraient deux aspects très importants et potentiellement spécifiques de Lmb : (i) tous les gènes d'avirulence caractérisés expérimentalement étaient localisés dans de grandes régions riches en bases AT et composées d'éléments transposables (ET), (ii) ces régions riches en AT préfiguraient une structure génomique particulière, qui, si elle se généralisait à l'ensemble du génome, aurait été totalement inédite chez un micro-organisme eucaryote. La première partie de cette thèse présente la description du génome de Lmb en se focalisant sur sa structure en isochores, résultant d’une invasion du génome par des ET qui ont ensuite été inactivés par un mécanisme de défense spécifique aux champignons ascomycètes, le RIP (*Repeat-Induced Point mutation*). Puis, l’impact potentiel de cette structure sur la diversification et l’évolution des protéines jouant un rôle clé lors de l’interaction agent pathogène-plante a été évalué, mettant ainsi en avant l’existence d’un génome à « deux vitesses ». Afin de mieux comprendre le rôle potentiel joué par les ET au niveau des capacités d’adaptation de Lmb au colza, une étude de génomique comparative et évolutive de cinq membres du complexe d’espèces a été réalisée. Ce travail montre que Lmb est la seule espèce du complexe dont le génome a été envahi par les ET, et que ces derniers sont impliqués dans (i) des réarrangements intrachromosomiques potentiellement liés à la spéciation entre Lmb et l’espèce la plus proche, (ii) la présence de gènes espèce-spécifiques et (iii) des déplacements dans des régions génomiques très dynamiques de gènes codant des effecteurs. Les travaux constituant cette thèse participent à la généralisation du concept selon lequel un lien fort existe chez les champignons filamenteux phytopathogènes entre ET et gènes impliqués dans la pathogenèse ou l’adaptation à l’hôte.

Abstract: *Leptosphaeria maculans* ‘brassicae’ (Lmb) is a filamentous ascomycete from class *Dothideomycetes*. It belongs to the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex which comprises pathogens of crucifers. Lmb is specifically adapted to oilseed rape (*Brassica napus*) and is responsible for the most damaging disease of this crop: “stem canker”. In order to better understand and control the disease, the host team initiated a genomic project aiming at systematically identify genes involved in pathogenicity, analyse genome plasticity and evaluate their incidence on adaptability to host. Preliminary genome data firstly showed that all characterized avirulence genes were localized in large AT-rich regions, mainly composed of Transposable Elements (TEs). In addition, these AT-rich regions were the first hints that the Lmb genome may present a very unusual structure compared to other microorganisms. The first part of this thesis describes the Lmb genome with a special focus on its isochore structure, which is the result of a massive TE invasion of the genome followed by an inactivation of TEs by an ascomycete-specific defense mechanism called RIP (*Repeat-Induced Point mutation*). The potential impacts of this genome structure on diversification and evolution of proteins involved in the plant-pathogen interaction were assessed and highlighted the existence of a “two speed” genome. To better understand how TEs are involved in adaptation of Lmb towards oilseed rape, a comparative and evolutionary genomic analysis of five members of the species complex was conducted. This study shows that Lmb is the only species of the complex with genome invaded by TEs at such an extent, and that TEs are involved in (i) intrachromosomal rearrangements putatively related to the speciation event between Lmb and its closest relative species, (ii) the presence of species-specific genes, (iii) translocations of effector genes into highly dynamic genomic regions. Our data contribute to the generalization of the “two speed” genome concept in filamentous phytopathogens postulating that highly plastic regions of the genome are enriched in genes involved in niche adaptation and that a strong link exists between TEs and genes involved in pathogenesis or host adaptation.