Caractéristiques génomiques du genre fongique Mucor et évolution adaptative liée à différents modes et conditions de vie au sein du genre

Annie Lebreton

UNIVERSITE ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

UBO
Université de Bretagne Occidentale

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : Génétique, génomique et bio-informatique

Par
Annie LEBRETON

Caractéristiques génomiques du genre fongique *Mucor* et évolution adaptative liée à différents modes et conditions de vie au sein du genre

Thèse présentée et soutenue à Plouzané, le 20 décembre 2018
Unité de recherche : Laboratoire Universitaire de Biodiversité et d'Écologie Microbienne avec l'aide de la plateforme ABiMS de la Station Biologique de Roscoff

Rapporteurs avant soutenance :

Joëlle DUPONT	Professeure	MNHN-CNRS, Sorbonne Université-EPHE
Philippe SILAR	Professeur	Université Paris Diderot, CNRS

Composition du Jury :

Présidente :	Gwenaëlle LE BLAY	Professeure	Université de Bretagne Occidentale
Examinateurs :	Jeanne ROPARS	Chargée de Recherche	CNRS, Université Paris Sud
	Riccardo BARONCELLI	Chargé de Recherche	Université de Salamanque
Dir. de thèse :	Georges BARBIER	Professeur	Université de Bretagne Occidentale

Invités

Erwan CORRE	Ingénieur de Recherche	CNRS, Sorbonne Université
Laurence MESLET-CLADIERE	Maître de Conférences	Université de Bretagne Occidentale
Jean-Luc JANY	Maître de Conférences	Université de Bretagne Occidentale

« *Entre Ce que je pense, Ce que je veux dire, Ce que je crois dire, Ce que je dis, Ce que vous avez envie d'entendre, Ce que vous entendez, Ce que vous comprenez... il y a dix possibilités qu'on ait des difficultés à communiquer. Mais essayons quand même...* »

Bernard Werber

Remerciements

Tout d'abord, je tiens à remercier les membres du jury, Joëlle Dupont, Philippe Silar, Gwenaëlle Le Blay, Jeanne Ropars et Riccardo Baroncelli, qui ont bien voulu évaluer ce travail et pour le temps consacré à la lecture de mon manuscrit.

Je voulais remercier le Pr. Georges Barbier sans qui cette thèse n'aurait pas pu commencer.

Un grand merci à mes encadrants de thèse, Laurence Meslet-Cladière, Jean-Luc Jany et Erwan Corre qui, chacun à leur façon, par leurs qualités aussi bien humaines que professionnelles, m'ont permis de mener à bien ce projet et développer mon esprit critique. Un merci particulier à Laurence, pour beaucoup de choses mais surtout pour avoir partagé sa "grosse bête noire" jusqu'au bout même si parfois j'aurais pu mériter la fameuse "trottinette à vapeur".

Je remercie également les membres de mon comité de thèse, Emmanuelle Morin et Antoine Branca pour leurs conseils et leur intérêt à l'égard de mon sujet de thèse.

Je remercie le Pr. Emmanuel Coton pour ses conseils sur mon projet de thèse.

Egalement, un grand merci aux membres de ma plateforme d'adoption (ABiMS), passés et présents que j'ai eu le plaisir de rencontrer : Romain, Gabriel, Gildas, Loraine, Xi, Misharl, Julien, Camille, Gwendoline, Olivier, Victor, Ehsan, "chef", Eric, Philippe, Joseph, Jean-Michel, Guillaume, Guita, Delphine, Jérémi, Marjorie et Mark. Une petite pensée pour les blagues du mercredi, les discussions géopolitiques initiées à partir de pistaches et bien d'autres encore.

Un grand merci également aux membres du LUBEM pour m'avoir chaleureusement accueilli à chacun de mes retours au laboratoire. Merci à Stella pour sa disponibilité et son aide lors de mon retour en paillasse, à Marielle pour son aide précieuse sur les aspects administratifs. Un merci particulier à mes co-bureaux de Brest Laura, Marine, Maxence, Guillaume et Fabienne pour leur accueil et leur bonne humeur. Merci également à Amélie, MHT et biens d'autres...

Merci au personnel du Gulf Stream pour avoir significativement augmenté mon moral chaque midi au cours de ces trois dernières années.

Merci Laure pour les chocolats de fin de rédaction.

Merci à tous mes amis, à Léopold notament et à la bande de filles : Solène, Florianne, Pauline et Astrid pour leur soutien au cours de ces trois années.

Merci à mon écureuil/mouton préféré, des petites graines pour une grande amitié.

Á toi mon mari, lumière et éclats de rires dans mes longues soirées de rédaction.

Pardon à ceux que j'aurais pu oublier...

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

1n	Haploïde
2n	Diploïde
ADN	Acide désoxyribonucléique
ADNr	Acide désoxyribonucléique nucléaire ribosomique
ADP	Adenosine DiPhosphate
ARN	Acide ribonucléique
ARNi	Acide ribonucléique interférent
ARNInc	Acide ribonucléique long non codant
ARNm	Acide ribonucléique messager
ARNmi	Micro-acide ribonucléique
ARNr	Acide ribonucléique ribosomique
ARNsn	Petit acide ribonucléique nucléaire
ARNt	Acide ribonucléique de transfert
ATP	Adenosine TriPhosphate
aa	Acide aminé
aw	*Water activity* (activité de l'eau)
BGC	*Biosynthetic Gene Cluster* (cluster de gènes impliqués dans une biosynthèse donnée)
BGI	*Beijing Genome Institute*
BP	*Biological Process* (processus biologique)
CAZyme	*Carbohydrate-active enzyme* (enzyme qui agit sur les glucides)
CBS	*Centrallbureau voor Schimmelcultures* (aujourd'hui *Westerdijk Fungal Biodiversity Institute*)
CNV	*Copy Number Variation* (variation du nombre de copies)
COG	*Clusters of Orthologous Groups* (Cluster de groupes de gènes orthologues)
CotH	*Spore coat protein homologs* (invasines)
DIRS	Dictyostellum intermediate repeat sequence
DMAT	Tryptophan dimethylallyltransferase
dN	Substitution de nucléotide non synonyme
DNA	*Deoxyribonucleic acid* (ADN)
DNAseq	Séquençage de l'ADN
dS	Substitution de nucléotide synonyme
EC number	*Enzyme Commission number* (numéro de la commission des enzymes)
et al.	*et alli* (et autres)
FAS	*Fatty Acid Synthase* (enzyme de synthèse des acides gras)
FISH	*Fluorescence in situ Hybridization* (hybridation in situ en fluorescence)
FLC	Fluoconazole
FPKM	*Fragment Per Kilobase Million*
GC%	au sein d'un acide nucléique, pourcentage en désoxyribonucléotides portant soit une guanine (G) soit une cytosine (C)
GO	*Gene Ontology* (ontologie des gènes, projet bioinformatique)
GSP	*Genome Sequencing Project* (projet de séquençage de génome)

HGT	*Horizontal Gene Transfer* (transfert horizontal de gène)
HMG	*High Mobility Group* (domaine protéique à haute mobilité)
lncRNA	*Long non coding ribonucleic acid* (ARNlnc)
indel	Insertion ou deletion
ITS	*Intergenic Transcribed Spacer* (Région intergénique de l'ADNr nucléaire)
JGI	*Joint Genome Institute*
kb	kilobases (unité de taille d'un acide nucléique correspondant à 1000 nucléotides)
KEGG	*Kyoto Encyclopedia of Genes and Genomes* (encyclopédie des gènes et génomes de Kyoto, projet bioinformatique)
LarD	*Large retrotransposon derivative*
LINE	*Long interspersed repeated sequence* (longs éléments nucléaires intercalés)
LTR	*Long terminal repeat sequence* (séquence terminale longue répétée)
LUBEM	Laboratoire de Biodiversité et d'Ecologie Microbienne
MA	*Mutation Accumulation* (accumulation de mutations)
Mb	Mégabases (unité de taille d'un acide nucléique correspondant à 10000 nucléotides)
MC	*Mucor circinelloides*
ME	*Mucor endophyticus*
MF	*Mucor fuscus*
MGC	*Metabolic Gene Cluster* (cluster de gènes impliqués dans une voie métabolique donnée)
miRNA	*Micro ribonucleic acid*
MiTE	*Miniature interverted-repeat transposable element* (élément transposable miniature à répétition inversée)
ML	*Mucor lanceolatus*
MR	*Mucor racemosus*
mRNA	*Messenger ribonucleic acid*
My	*Million years* (millions d'années)
n	Ploïdie (précédé du facteur livrant le niveau de ploïdie)
NADH	Nicotinamide Adenine Dinucleotide sous forme réduite
NCBI	*National Center for Biotechnological Information*
NGS	*Next Generation Sequencing* (séquençage de nouvelle génération)
nr	*non redundant* (base de données protéique non redondantes du NCBI)
NRPS	*Non ribosomal peptide synthase* (Enzyme de synthèse des peptides non ribosomiques)
OLC	*Overlap Layout Sequencing*
PCR	*Polymerase Chain Reaction* (amplification en chaîne par polymérase)
PDA	*Potato Dextrose Agar* (milieu à base d'infusion de pomme de terre, de dextrose et d'agar)
pH	Potentiel hydrogène
PKS	*Polyketide synthase (polycétide synthase)*
PLE	*Penelope-like element* (éléments transposables de la famille Penelope)
polyA	Succession de nombreux ribonucléotide de type adénosine
rDNA	*Ribosomal deoxyribonucleic acid*
RIP	*Repeat Induced Point mutation* (mutagenèse dirigée contre les éléments répétés)
RNA	*Ribonucleic acid* (ARN)
RNAi	*Ribonucleic acid interferent*
RNAseq	Séquençage de l'ARN
rRNA	*Ribosomal ribonucleic acid*
SBC	Séquençage par ligation
SBS	Séquençage par synthèse
SINE	*Short interspersed repetitive element* (petit élément nucléaire intercalé)
SMS	Séquençage de molécules uniques

SNAC *Small autonomous CACTA* (petit transposon autonomes avec séquence CACTA répétée)
SNM *Single Nucleotide Mutation* (mutation d'un nucléotide)
SNP *Single Nucleotide Polymorphism* (polymorphisme dû à la mutation d'un nucléotide)
snRNA *Small nuclear ribonucleic acid*
TE *Transposable Element* (élément transposable)
TIR *Terminal inverted repeat* (répétition en tandem inversée)
TriM *Terminal repeat retrotransposon in miniature* (retrotransposon miniature à répétition terminale)
tRNA *Transfer ribonucleic acid*
TSD T*arget of Site Duplication* (site cible de duplication)
UBOCC Université de Bretagne Occidentale culture collection
UDP Uridine DiPhosphate
UTR *Untranslated Transcribed Region* (région transcrite non traduite)
WGD *Whole Genome Duplication* (duplication complète du génome)

Glossaire

Clade : en phylogénie, groupe qui comprend une espèce ancestrale et tous ses décendants (aussi appelé monophylétique)

Contig : séquence nucléotidique obtenue par assemblage de reads contigus.

Expansion/contraction de famille de gènes : augmentation/réduction du nombre de gènes au sein d'une même famille de gènes.

Famille de gènes : Ensemble de gènes ayant de grandes ressemblances fonctionnelles et structurelles et supposés issus d'une même copie ancestrale.

Gènes homologues : gènes issus d'un même gène ancestral.

Gènes orthologues : deux gènes homologues présents dans deux espèces différentes et descendant d'un gène unique présent dans le dernier ancêtre commun aux deux espèces.

Gènes BUSCO : groupe de protéines orthologues supposées systématiquement présentes en une seule copie dans la lignée taxonomique considérée

GO terms : dans le cadre du projet *Gene Onthology* destiné à structurer la description des gènes et des produits géniques, vocabulaire contrôlé décrivant des propriétés, descriptions ou concepts.

K-mers : fragment de séquence où K représente le nombre de base du fragment.

Mate pair : protocole de circularisation de l'ADN lors de la préparation d'une banque génomique permettant de séquencer deux fragments d'ADN séparés par une distance fixe.

N50 (génome) : taille du plus petit contig (ou scaffolds) tel que 50% du génome soit contenu dans les contigs (ou scaffolds) de taille N50 et plus.

Noeud (phylogénie) : point de rencontre de trois branches ou segments de branche dans un arbre. En phylogénie, le noeud représente un groupe comprenant les taxons frères en aval de ce noeud. Le noeud terminal aussi apellé feuille correspond au taxon.

Monophylétique : en phylogénie, groupe qui comprend une espèce ancestrale et tous ses décendants (aussi appelé clade)

Orthogroupes : groupe de gènes homologues.

Outgroup : groupe externe en phylogénie

Paired end : type de séquençage qui consiste à séquencer les lectures par paires, ces lectures étant séparées par une distance connue.

Phylogenie : étude des relations de parentés entre différents êtres vivants en vue de comprendre leur évolution.

Pseudogène : gène inactif au sein d'un génome, du fait d'altérations génétiques le rendant non fonctionnel et donc incapable de conduire à l'expression d'une protéine.

Read (lecture) : séquence nucléotidique brute générée suite à un séquençage.

Scaffold : séquences nucléotidique composée de contigs orientés les uns par rapport aux autres mais séparés par des bases inconnues (dites ambigües).

Score Phred : score de qualité associé à chaque base lors d'un séquençage.

Siphonnés : hyphe résultant de divisions nucléaires répétées, sans divisions cellulaires concomitantes.

Synténie : conservation d'un ensemble d'éléments génomiques sur le même locus et dans le même ordre le long d'un fragment d'ADN chromosomique que ce soit dans un même génome (relation de co-localisation) ou entre génomes (relation de correspondance)

Chapitre 1

Introduction générale

Les champignons constituent un groupe d'une extrême variété que ce soit en terme de forme ou de mode de vie. Estimé en 2017 à 3,8 millions d'espèces (Hawksworth and Lücking, 2017), ce groupe compte aussi bien des organismes microscopiques (les cellules de la levure de référence *Saccharomyces cerevisiae* mesurent de 6 à 12 micromètres) que des organismes parmi les plus grands sur terre (un individu du genre *Armillaria* s'étend sur 965 hectares (Sipos et al., 2017)). Dépourvus de chlorophylle et non autotrophes par rapport au carbone, les champignons se nourrissent de constituants organiques préexistants qu'ils acquièrent à partir d'autres organismes morts (saprophytisme) ou vivants (parasitisme et symbiose). Les champignons jouent un rôle primordial dans l'écologie de la planète (de Boer et al., 2005; Frac et al., 2018), que ce soit en recyclant la matière organique morte (Treseder and Lennon, 2015), dans leur implication dans les maladies animales (ex : aspergillose) ou végétales (ex : fusariose) (Vallabhaneni et al., 2016; Duba et al., 2018) ou encore en améliorant l'accès d'autres espèces à des ressources (eau et nutriments) ou des molécules de défense (Wurzbacher et al., 2014; Ashwin et al., 2017; Jayne and Quigley, 2014; El-Komy et al., 2015) (Figure 1.1).

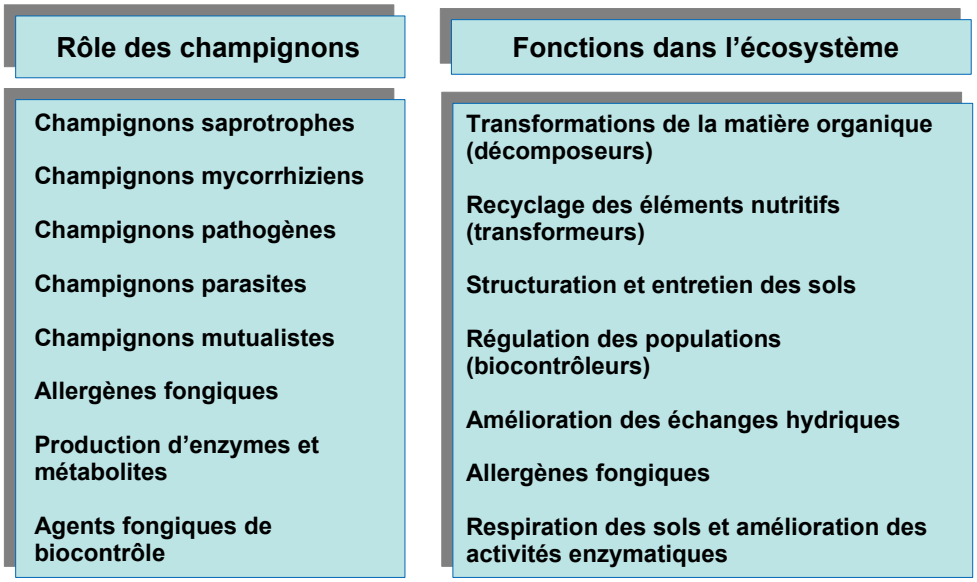


FIGURE 1.1 – Aspects de la biodiversité fongique des sols. Adapté de (Frac et al., 2018).

L'impact des champignons sur l'Homme est ambivalent et concerne de multiples domaines allant de la pharmacologie à l'industrie en passant par l'agriculture et l'agroalimentaire (produits bruts ou transformés). Dans le domaine médical, certaines espèces sont impliquées dans des maladies pouvant être mortelles telles que les Mucormycoses impliquant des espèces du groupe des Mucorales (Roden et al., 2005), tandis que d'autres produisent des antibiotiques utilisés en pharmaceutique, l'un des plus connus étant la pénicilline initialement découverte chez *Penicillium notatum* (Lobanovska and Pilla, 2017). Dans le domaine de l'agroalimentaire, entre 5% et 10% de la production mondiale de nourriture est perdue en raison de détériorations liées à des champignons (Garnier et al., 2017), tandis que quelques autres espèces sont utilisées en consommation directe comme les truffes et les champignons de Paris : en 2005, le marché mondial de l'industrie du champignon était évalué à plus de 45 milliards de dollars (Chang, 2006). Quelques rares autres champignons (ex : *S. cerevisiae*, *Mucor lanceolatus* ou *Penicillium roqueforti*) sont également utilisés pour la production de produits fermentés tels que la bière ou le fromage (Lodolo et al., 2008; Garnier et al., 2017). Les espèces fongiques ne sont pas toujours cantonnées à un mode de vie, leur comportement peut changer selon les conditions environnementales. Ainsi, si les conditions changent, une espèce symbiotique peut se révéler pathogène et une espèce technologique peut devenir contaminante (Porras-Alfaro and Bayman, 2011; Knapp et al., 2018).

Le genre *Mucor*, qui fait l'objet de cette étude, regroupe un grand nombre d'espèces très majoritairement saprophytes capables de coloniser des habitats extrêmement divers (Walther et al., 2013) avec pour point commun une faible tolérance aux faibles activités de l'eau et une stratégie écologique de type rudérale [1], ce qui en fait des espèces pionnières à la croissance et à la sporulation importantes mais peu persistantes (Morin-Sardin et al., 2016).

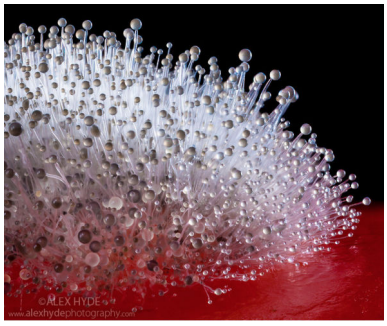


FIGURE 1.2 – *Mucor sp.* sur une tomate. (Source photographie : A. Hide (https ://alexhyde-photo.wordpress.com/2014/01/)

1. Se dit d'une espèce se développant sur des décombres, bords des chemins, friches ou voisinage des habitations. Ces espèces affectionnent les espaces ouverts (à l'inverse de la forêt, qui est un milieu fermé), perturbés ou instables. Ce sont souvent des espèces qui colonisent de nouveaux terrains après un bouleversement ou une modification de l'écosystème local.

Très présents en milieux naturels, les *Mucor spp.* comptent parmi leurs membres aussi bien des espèces ubiquistes que spécialisés. A titre d'exemple, *M. circinelloides* a été identifié en tant que coprophile mais aussi en tant que pathogène animal et végétal (Walther et al., 2013) tandis que *M. endophyticus* n'a jusqu'à présent été retrouvé qu'en tant qu'endophyte de plante ((Zheng and Jiang, 1995; Walther et al., 2013; Morin-Sardin et al., 2017). Cette différence de comportement entre des espèces phylogénétiquement proches fait du genre *Mucor* un cas d'étude pour mieux comprendre les mécanismes pouvant intervenir dans l'adaptation à de multiples milieux ou la spécialisation dans un type d'habitat. Un autre intérêt à l'étude des *Mucor spp.* vient du groupe auquel ils appartiennent : les *Mucoromycota*. Il s'agit de l'un des quatre phylums au sein desquels les champignons sont répartis phylogénétiquement (Spatafora et al., 2016). Avec le groupe des *Zoopagomycota*, il correspond aux champignons ayant divergé le plus tôt dans l'évolution. Ceux ci restent peu étudiés : à titre d'exemple, en 2017, parmi les 1090 espèces fongiques dont les génomes étaient disponibles, 47 correspondaient à une espèce de *Mucoromycota* (Aylward et al., 2017).

Afin de mieux comprendre les caractéristiques des *Mucor* et leur potentielle adaptation à un milieu, des études ont été engagées au Laboratoire Universitaire de Biodiversité et d'Écologie Microbienne (LUBEM, EA3882) avec un intérêt particulier apporté aux espèces retrouvées sur fromage. En 2012, Hermet et al. (2012) ont établi les relations phylogénétiques de 70 souches de *Mucor* (dont 36 issues de milieux fromager) en incluant des analyses morphologiques. Ces travaux ont notamment permis de caractériser une nouvelle espèce, *M. lanceolatus*, pourtant couramment rencontrée et utilisée en affinage de fromage. En 2016, la croissance sur milieu synthétique et fromager d'espèces rencontrées dans des milieux distincts a été étudiée par Morin-Sardin et al. (2016). Cette étude a permis de mettre en évidence une croissance améliorée des espèces utilisées pour l'affinage de fromage sur milieu fromager par rapport à un milieu standard. Ce comportement n'est pas observé pour les espèces contaminantes. *M. endophyticus*, l'espèce endophyte, accuse quant à elle un retard de croissance sur les milieux fromagers. Ces résultats suggèrent que certaines espèces sont plus aptes à se développer sur milieu fromager et interrogent sur une potentielle adaptation de certains *Mucor* à la matrice fromagère. Afin de répondre à cette question, une étude de transcriptomique et génomique comparative a été engagée. Ces travaux réalisés dans le cadre de cette thèse sont présentés ci-après.

Les objectifs de cette thèse étaient d'une part d'améliorer, à partir de donnée génomiques et transcriptomiques, les connaissances concernant le genre *Mucor* et d'autre part chercher s'il existait des éléments indiquant une potentielle adaptation de certains *Mucor* à différents habitats et modes de vie et en particulier à la matrice fromagère. Un intérêt a également été porté aux éléments permettant d'expliquer que des espèces soient retrouvées dans de multiples milieux et que d'autres ne soient retrouvées que dans des habitats spécifiques.

Ce manuscrit est composé de quatre parties : cette introduction générale, une synthèse bibliographique, les analyses sur les transcriptomes, les analyses sur les génomes et une discussion et conclusion générale sont proposées en fin de manuscrit.

Chapitre 2

Synthèse bibliographique

2.1 Les champignons

2.1.1 Qu'est-ce qu'un champignon ?

Le vivant est actuellement classé en trois domaines : les eucaryotes (*Eucarya*), les archées (*Archaea*) et les bactéries (*Bacteria*) (Ciccarelli et al., 2006). C'est parmi les eucaryotes que l'on retrouve les champignons (*Fungi*) (Figure 2.1).

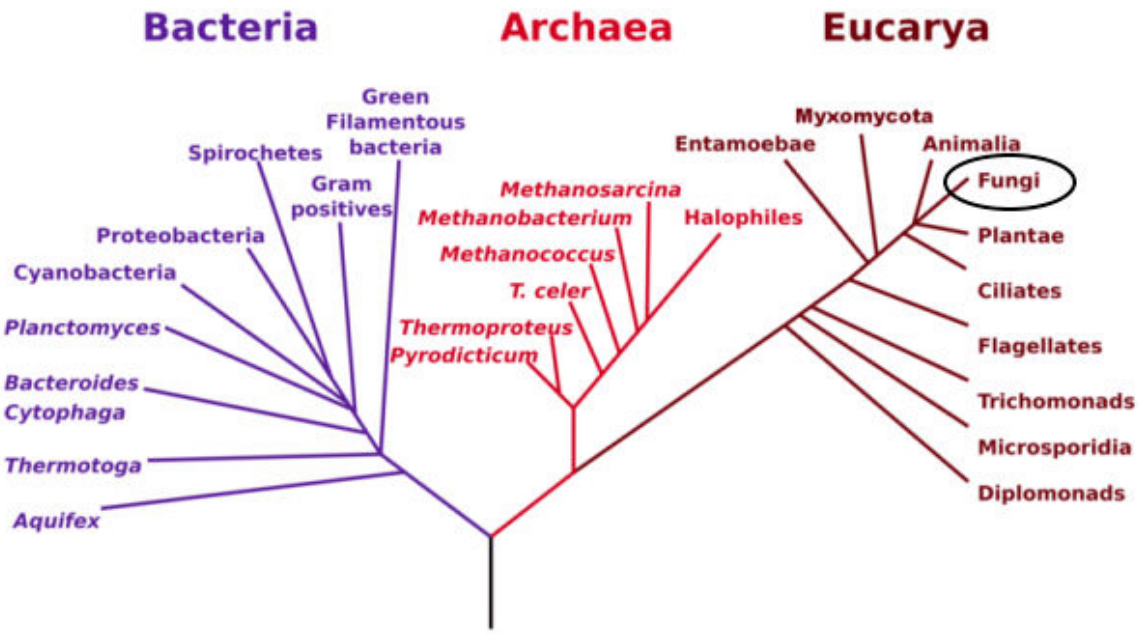


FIGURE 2.1 – Les trois domaines du vivant.

La diversité fongique est extrêmement importante. Au vu des espèces décrites dans la littérature et en se basant sur différentes extrapolations qui prennent en compte l'existence d'habitats inexplorés (Hawksworth and Lücking, 2017), de hotspots de biodiversité non échantillonnés (Scheffers et al., 2012) ou encore de l'existence d'espèces cryptiques (Hawksworth and Rossman, 1997), Hawksworth and Lücking (2017) estiment le nombre d'espèces fongiques à 3,8 millions.

Parmi ces millions d'espèces, 120.000 environ sont bien décrites et se retrouvent dans l'arbre du vivant au niveau des *Fungi*.

2.1.2 Classification des champignons

Cette classification, historiquement basée sur des critères phénotypiques avait permis d'identifier quatre classes (phycomycètes, ascomycètes, basidiomycètes et deutéromycètes ou *fungi imperfecti*) (Ainsworth, 1973). Cependant, l'introduction de critères génotypiques a contribué à une importante évolution de cette classification au cours de ces dernières années.

Encadré 1 : Nomenclature en classification.

La classification des êtres vivants repose sur l'utilisation de rangs taxonomiques : le règne, l'embranchement ou division, la classe, l'ordre, la famille, le genre et l'espèce. Ces rangs peuvent être complétés par des intermédiaires comme le sous-embranchement ou la sous-division, la sous-famille ou la sous-espèce, qui peut elle-même se diviser en variétés. La nomenclature utilisée pour déterminer le nom scientifique des espèces, basée sur les principes énoncés par Carl Von Linné en 1753, est binomiale et fait référence au genre puis à l'espèce. Les différents taxons fongiques présentent une terminologie codifiée où le suffixe définit chaque rang taxonomique dans la classification hiérarchique : -mycota pour la division, -mycotina pour la sous-division, -mycètes pour la classe, -ales pour l'ordre, et -aceae pour la famille.

Le groupe des champignons (ou Eumycètes) est actuellement composé d'un sous-règne, de 7 divisions et de 10 sous-divisions (Hibbett et al., 2007; Spatafora et al., 2016). Le sous-règne des Dikarya comporte deux divisions : les Ascomycota et Basidiomycota. La division des Ascomycota se scinde en trois sous-divisions Pezizomycotina, Saccharomycotina et Taphrinomycotina, tandis que la division des Basidiomycota se décompose en trois sous-divisions Pucciniomycotina, Ustilaginomycotina et Agaricomycotina (Figure 2.2). Les autres divisions sont les Mucoromycota, les Zoopagomycota, les Chytridiomycota, les Blastocladiomycota et les Cryptomycota. La division des Mucoromycota regroupe les trois sous-divisions Mucoromycotina, Mortierellomycotina et Glomeromycotina tandis que la division des Zoopagomycota regroupe les sous-divisions Zoopagomycotina, Kickxellomycotina et Entomophtoromycotina (Spatafora et al., 2016; Morin-Sardin et al., 2017). Cette nouvelle classification a pu être réalisée grâce au séquençage de nombreux génomes fongiques actuellement disponibles (`http://genome.jgi.doe.gov/programs/fungi/index.jsf`, (Spatafora et al., 2016).

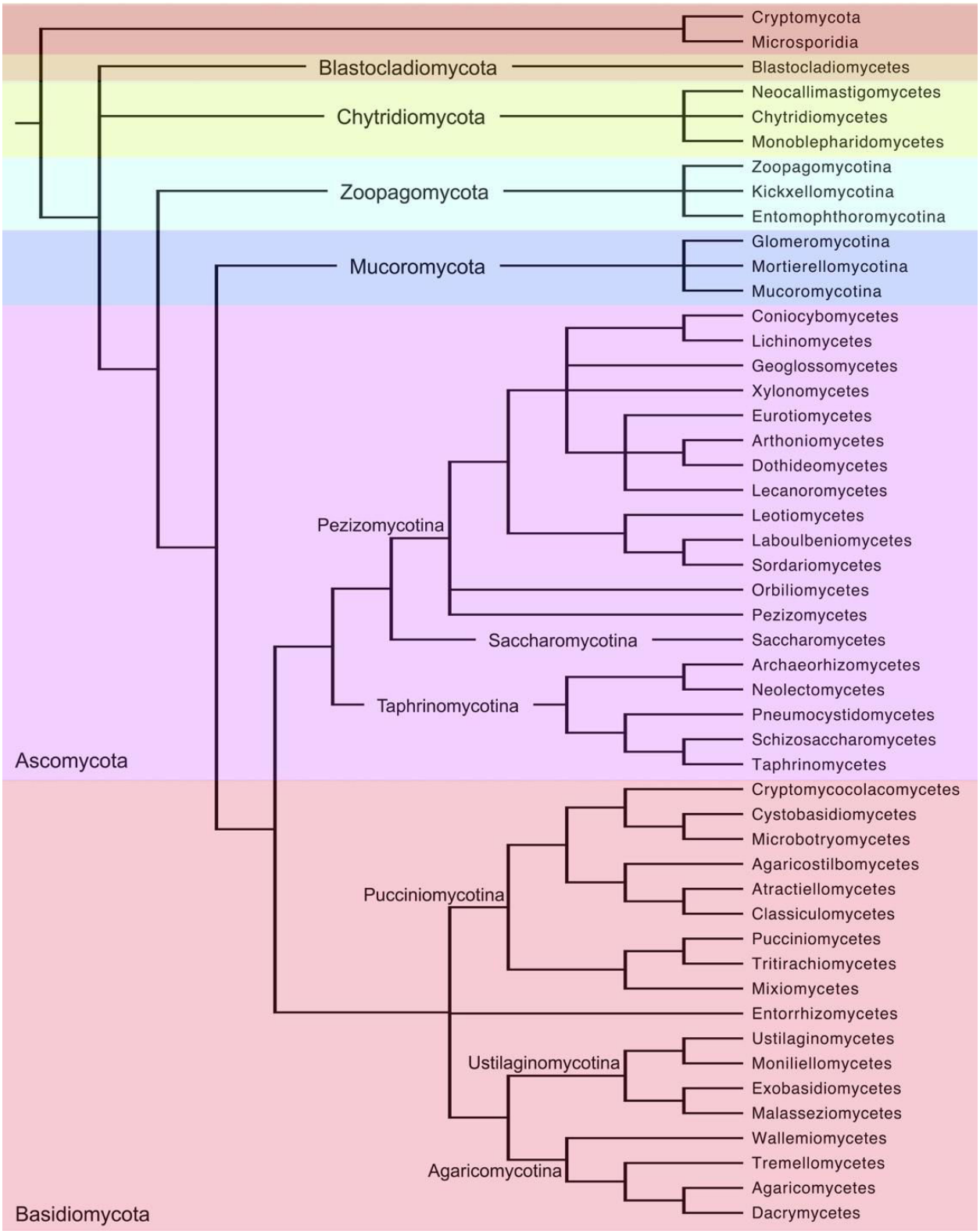


FIGURE 2.2 – Cladogramme présentant les différents groupes fongiques. Les polytomies (section d'arbre phylogénétique ne comprenant pas uniquement de séparations binaires, et donc présentant des branches semblant apparaitre simultanément.) artificielles représentent des régions pour lesquelles la topologie de l'arbre n'est pas résolue. (Source : Spatafora et al. (2017)).

De manière simplifiée, les champignons peuvent être divisés en trois groupes (Spatafora et al., 2017) :

Les champignons à zoospores (Cryptomycota, Blastocladiomycota, *Neocallimastigomycota*, Chytridiomycota et *Monoblépharidomycota*) qui peuvent exister sous forme de thalle filamenteux ou parfois sous forme unicellulaire et qui forment des structures de reproduction appelées zoosporanges. Les zoosporanges peuvent germer en relâchant des cellules avec flagelles et des spores dormantes qui germent après développement d'un tube germinatif.Les *Microsporidia*, sont également souvent classés dans ce groupe bien que d'éventuelles zoospores n'aient pas été observées chez ces endoparasites obligatoires d'animaux.

Les champignons autrefois appelés "zygomycètes" groupe aujourd'hui caduque car non monophylétique et sans caractère synapomorphe. En effet, la zygospore qui est une spore sexuée résultant d'une fusion d'hyphes, n'aurait pas existé uniquement chez ces "zygomycètes" mais aussi chez un ancêtre commun aux phylums Zoopagomycota, Mucoromycota, Ascomycota et Basidiomycota. Les membres de ce phylum caduque ont majoritairement été reversés au niveau des phylums des Zoopagomycota et Mucoromycota. La plupart des espèces de ces phylums possèdent des hyphes siphonnés (c'est à dire non septés hormis au niveau des organes de reproduction et des hyphes en dégénérescence) et forment des zygospores. De manière remarquable, c'est avec l'émergence des "zygomycètes" que le flagelle a disparu lors de la colonisation du milieu terrestre. Chez le phylum Zoopagomycota les spores asexuées peuvent être produites ou non dans des sacs. Les Zoopagomycota comprennent majoritairement des espèces pathogènes et commensales d'animaux et des parasites de champignons et d'amibes et quelques rares endophytes. Chez le phylum Mucoromycota, on retrouve des symbiotes endomycorhiziens (notamment chez les Glomeromycotina) et des saprophytes, parasites et endophytes majoritairement chez les autres sous-phylums.

Les champignons appelés *Dicarya* car à un moment donné de leur stade de vie, leurs hyphes comprennent des espaces cellulaires dicaryotiques (avec deux noyaux génétiquement distincts). Les *Dicarya* comprennent deux phylums : (i) les ascomycètes qui produisent leurs spores sexuées (ascospores) dans des asques (site de la caryogamie et de la méiose). Ils comprennent des espèces aux modes de vie très variés et, (ii) les basidiomycètes qui produisent leurs spores sexuées (basidiospores) au niveau de basides (site de la caryogamie et de la méiose). Les basidiomycètes ont également des modes de

vie très variés (incluant des champignons symbiotiques ectomycorhiziens qu'on retrouve aussi plus rarement chez les ascomycètes).

Longtemps considérés comme des végétaux du fait de la présence d'une paroi cellulaire, de l'absence de mobilité ou encore d'homologies entre leurs cycles de reproduction et celui des algues, ce n'est qu'en 1969 (Whittaker, 1969) que les champignons s'imposent comme un règne distinct dans la classification actuelle du vivant. La définition formelle de ce qu'est un champignon reste cependant un débat actuel puisqu'il n'existe pas de caractère spécifique, qu'il soit génétique, moléculaire ou cellulaire, qui soit retrouvé chez tous les champignons et absent des autres règnes (pour une revue, voir (Richards et al., 2017)). A titre d'exemple, on peut citer (i) l'existence d'une synthèse de lysine via la voie de biosynthèse de l'alpha-aminoadipate qui a aussi été retrouvée chez plusieurs clades eucaryotes et même procaryotes, (ii) la présence d'ergostérol dans les membranes plasmiques retrouvée également chez certains protistes ou encore (iii) la composition spécifique de la paroi cellulaire incluant de la chitine qui n'est pas retrouvée chez tous les champignons suite à une perte secondaire du caractère mais surtout, cette synthèse se retrouve aussi dans de nombreux groupes externes aux champignons.

Même s'il ne se dégage pas de synapomorphies [1] fongiques on retrouve chez la plupart des champignons un certain nombre de caractéristiques structurales et fonctionnelles.

2.1.3 Caractéristiques des champignons

Pour comprendre ce qui caractérise un champignon, il est intéressant de considérer l'étude de Cavalier-Smith (2001) qui a livré sur la base de travaux phylogénétiques un scénario de transition évolutive des protistes vers les champignons. Celui-ci implique la perte du caractère de phagotrophie (acquisition des nutriments sous une forme particulaire ou massive, c'est à dire non dissoute) ainsi que le développement de l'osmotrophie (acquisition de nutriments sous une forme dissoute) en association avec un mode particulier de synthèse de la paroi cellulaire couplé à une croissance polarisée apicale. Richards et al. (2018) paraphrasent cette association croissance radiale/osmotrophie par la périphrase : "je mange comme je croîs" (Figure 2.3).

En complément des éléments caractéristiques des champignons comme les fusions cellulaires asexuées ou sexuées (Encadrés 2 et 3) ou leur mode particulier de dissémination (Encadré 4),

1. Caractère dérivé (ou apomorphique), partagé par deux ou plusieurs taxons.

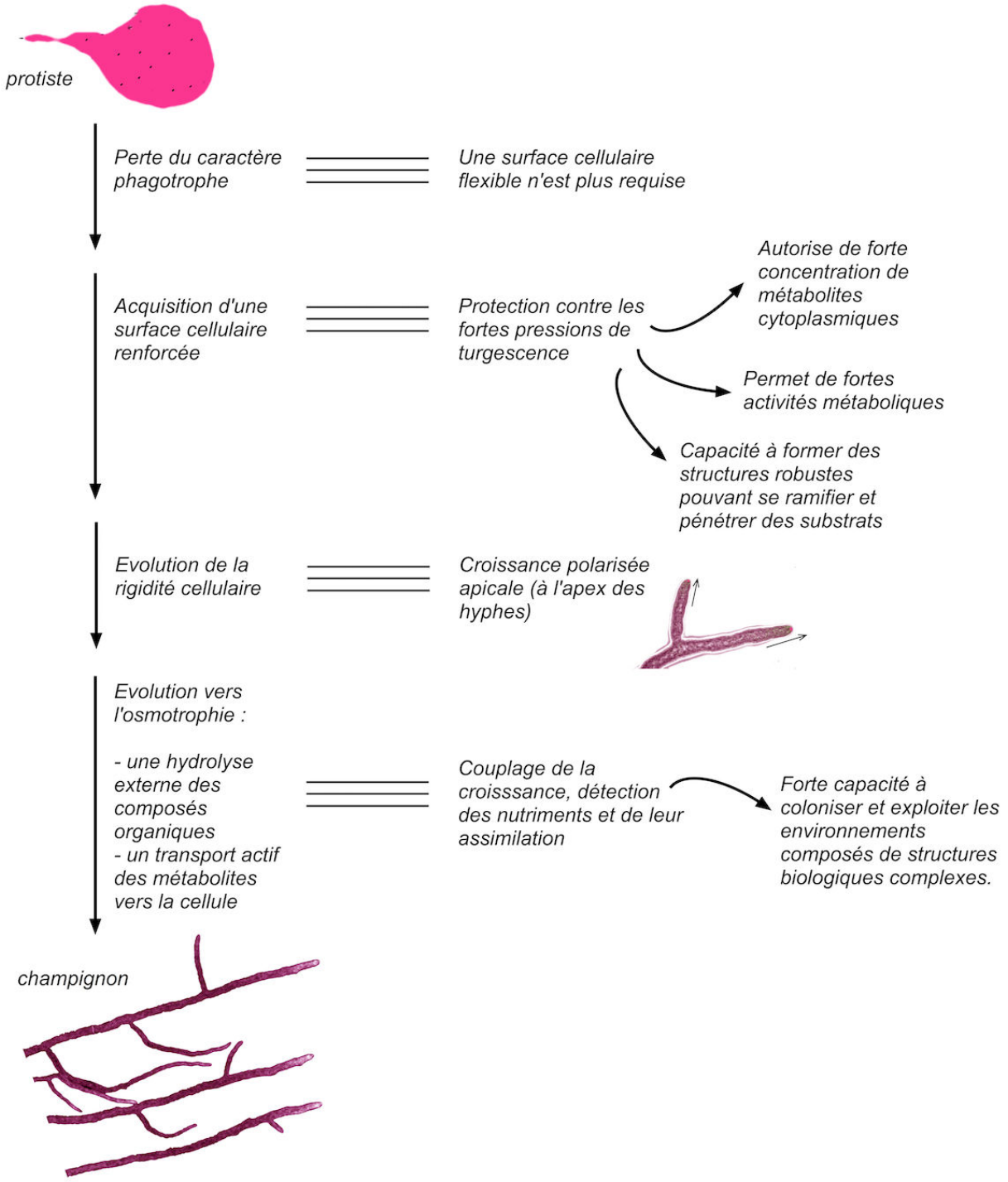


FIGURE 2.3 – Transition évolutive des protistes vers les champignons, selon un scénario proposé par Cavalier-Smith (2001) et la revue bibliographique de Richards et al. (2017).

l'existence de la paroi cellulaire et de la croissance apicale fongique est donc à la base de la structure et du mode de vie fongique.

Encadré 2 : La "reconnaissance du soi" chez les champignons

Le thalle fongique correspond à un entrelacs très dense d'hyphes à croissance apicale qui se ramifient et fusionnent chez de nombreux groupes de champignons (ces fusions n'ont en revanche jamais été observées chez certains groupes comme celui des Mucoromycotina). Cette fusion d'hyphes ou anastomose aboutit à la constitution d'un syncytium, qui permet la distribution du cytosol, des organites ou encore des nutriments dans l'ensemble de l'organisme (à défaut d'avoir recours à un système vascularisé comme chez les animaux ou les végétaux par exemple). Chez les groupes de champignons chez qui les anastomoses végétatives existent, celles-ci peuvent se former très fréquemment au niveau des hyphes d'un même individu pour assurer la formation et le développement du thalle mais aussi moins fréquemment entre individus. Il faut noter pour ce dernier cas que des mécanismes d'incompatibilité végétative pré- et post- fusion restreignent la formation d'hétérocaryons (cellules contenant des noyaux d'origine génétique différente) à des fusions d'hyphes d'entités génétiquement proches. Cette incompatibilité végétative empêche en effet la formation d'hétérocaryons entre individus qui diffèrent génétiquement au niveau de multiples loci (entre 7 et 12 loci appelés *het* ou *vic* chez les ascomycètes et analysés au travers de diverses études). En effet, pour les modèles étudiés, en cas de différence au niveau d'un seul des loci d'incompatibilité végétative, les compartiments hétérocaryotiques sont isolés du reste du mycelium par des septums et un phénomène de mort cellulaire est observé. Cette incompatibilité végétative constituerait notamment un avantage pour limiter la transmission de virus cytoplasmiques. Les mécanismes de fusion et d'incompatibilité végétative ont été étudiés chez des champignons modèles comme *Neurospora* (pour une revue, voir Daskalov et al. (2017)) et impliquent un grand nombre de gènes mais les mécanismes ne sont pas encore élucidés complètement.

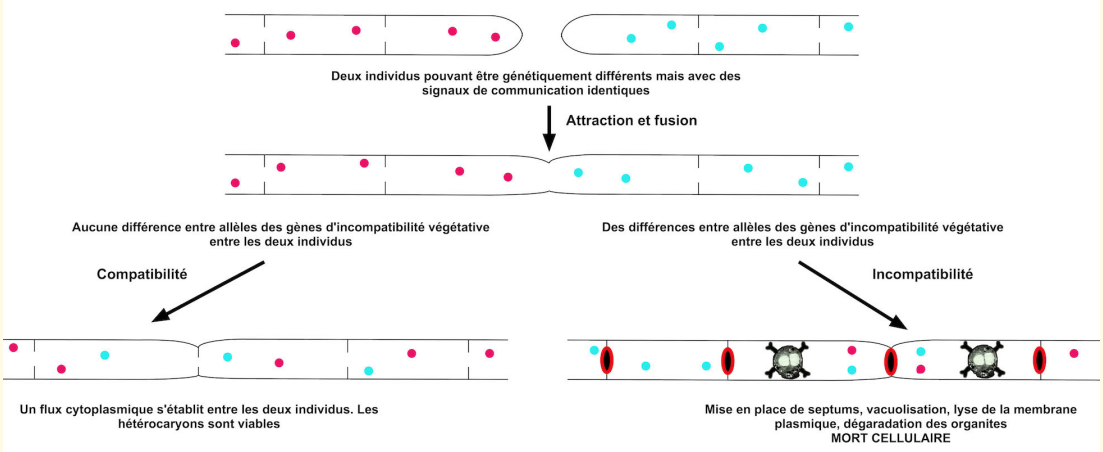


FIGURE 2.4 – Hétérocaryose après anastomose : les différents cas possibles. Les disques bleus et roses représentent les noyaux respectifs des deux individus. Adapté d'après Daskalov et al. (2017).

Encadré 3 : La reproduction sexuée chez les champignons

Une fusion cellulaire se produit lors de la reproduction sexuée. Celle-ci présente une particularité, chez les dicarya, en ce sens que la fusion des cytoplasmes (plasmogamie) conduisant à la formation d'un hétérocaryon n'est pas immédiatement suivie d'une fusion des deux noyaux (caryogamie). Il en résulte la présence d'un stade dicaryotique dans le cycle de vie des ascomycètes (où ce stade est très réduit) et des basidiomycètes. Ce décalage n'existe en revanche pas chez les autres groupes fongiques comme les Mucoromycota. Une méiose, succédant à la caryogamie rétablit dans tous les cas l'haploidie. Chez les champignons, il existe des espèces isogames et anisogames. Chez les anisogames, la fusion des gamètes ou syngamie (qui à l'inverse de la méiose initie une phase diploïde) ne peut se faire qu'entre gamètes différents alors que chez les isogames, il n'y a pas différents types de gamètes. De même, chez les champignons, il existe des espèces hétérothalliques chez qui la fusion ne peut se faire qu'entre deux individus compatibles sexuellement alors que chez les homothalliques la fusion est possible au niveau d'hyphes du même individu (hermaphrodisme). Il faut noter qu'il existe différentes formes d'homothallisme (dont le pseudo-homothallisme). Cette compatibilité sexuelle est déterminée par un gène sexuel (système bipolaire comme chez les Mucoromycota ou les Ascomycota par exemple) ou deux gènes sexuels (système tétrapolaire comme chez les Basidiomycota par exemple). Ces gènes sexuels peuvent coder des régulateurs ou des phéromones et récepteurs de phéromones. Pour le système bipolaire, il existe un gène sexuel possédant le plus souvent deux idiomorphes (deux versions du gène qui ne sont pas appelés allèles car leur homologie ne peut être prouvée) alors que pour le système tétrapolaire, il existe deux gènes sexuels avec une quantité d'idiomorphes variable pouvant aller jusqu'à des centaines dans le cas de certains basidiomycètes (pour une revue, sur la syngamie et la reproduction sexuée voir Billiard et al., 2011). La reproduction sexuée aboutit à la formation de spores sexuées qui permettent la dissémination. Ces dernières ne sont toutefois pas les seules à assurer cette fonction, et chez certains groupes fongiques, la reproduction sexuées étant rare et très dépendante des conditions environnementales, cette fonction est majoritairement assurée par des spores asexuées voire d'autres entités comme des fragments de mycélium ou des agrégats d'hyphes appelés sclérotes.

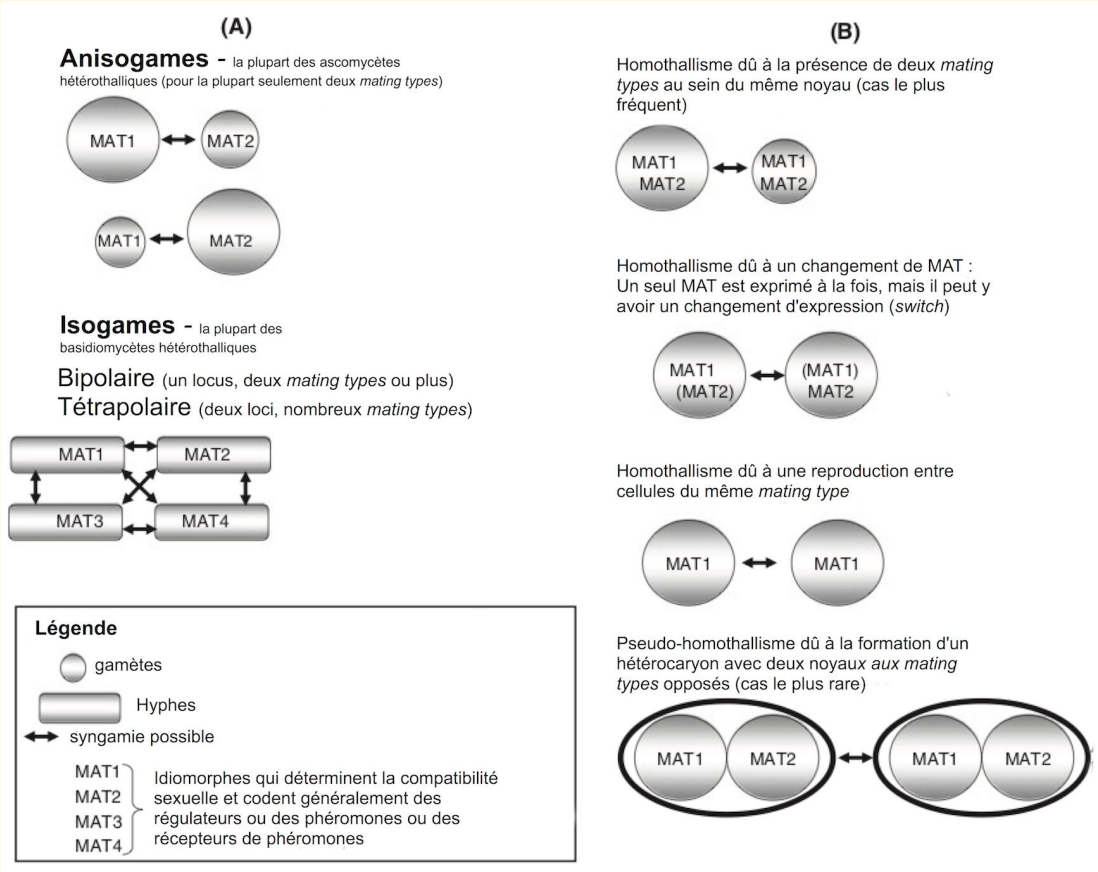


FIGURE 2.5 – Résumé des modes de syngamie chez les champignons. (A) Hétérothallisme ; (B) homothallismes. Adapté de Billiard et al. (2012).

Encadré 4 : La dissémination des champignons

Alors que la colonisation progressive de milieux peut être réalisée par la croissance végétative radiale du thalle fongique, celle-ci n'assure pas à proprement parler une dissémination. Cette dernière peut se faire grâce à des propagules qui peuvent être de types très différents : (i) des spores, qu'elles résultent de la reproduction sexuée (voir ci-avant), ou asexuée. Ces dernières peuvent être produites dans des enveloppes (du type sporange) comme chez les Mucoromycota ou vers l'extérieur comme chez les dicarya (on les appelle alors conidies), par des modes thallique (spores provenant de thalle préexistant et on peut y inclure les chlamydospores) ou blastique (spores issues d'un bourgeonnement à partir d'une structure sporogène). Ces spores peuvent même être flagellées chez les champignons à zoospores. (ii) de fragments de thalle qui peuvent correspondre également à des agrégats d'hyphes comme dans le cas des sclérotes. Ces propagules peuvent être disséminées, en fonction de l'écologie de l'espèce considérée et de l'environnement colonisé, par l'eau, le vent ou encore des vecteurs animaux.

La paroi fongique

Les champignons comprennent généralement une paroi cellulaire (Latgé, 2007), qui contrairement à ce qui est retrouvé chez les végétaux (chez qui elle est composée principalement de cellulose) comprend généralement des microfibrilles de chitine (homopolymère de N-acétylglucosamine lié en β-1,4) associé à des polymères de glucose (β-1,3 glucanes liés de manière covalente à des β-1,6 glucanes). C'est la présence de cette paroi qui permet la rigidité cellulaire associée aux champignons (Figure 2.3). Il faut noter que certaines structures fongiques comme les zoospores de chytridiomycota sont dépourvues de paroi cellulaire (Latgé, 2007) et que certaines espèces comme *Pneumocystis* (Ma et al., 2016) ont perdu secondairement ce caractère. Notons également que chez certains membres des *pseudofungi*, la présence de polymères de N-acétylglucosamine à la structure proche de la chitine a été rapportée (Clay et al., 1991; Mélida et al., 2013).

L'hyphe fongique

La pluricellularité constitue une des transitions majeures dans l'évolution du monde vivant (Love, 2016). Elle a notamment permis l'apparition d'êtres vivants plus grands et plus complexes avec des spécialisations cellulaires et des divisions de tâches entre cellules. Les champignons comme les animaux et les végétaux ont évolué vers une pluricellularité intégrée (et non pas un simple agrégat de cellules individuelles), mais sous une forme particulière : le filament ou hyphe (Nagy et al., 2017). A nouveau ces hyphes ne sont pas spécifiques des champignons car on les retrouve chez d'autres organismes comme par exemple les *pseudofungi*. Il faut aussi indiquer que certains champignons ne forment que des structures unicellulaires : les formes levures. Ces levures ne constituent pas un groupe monophylétique mais plutôt un assemblage polyphylétique de champignon partageant ce caractère morphologique d'unicellularité. Il est également important de noter qu'un grand nombre de champignons (notamment chez les Mucoromycota) décrits morphologiquement comme des levures sont dimorphes, c'est à dire qu'en fonction des conditions environnementales ils peuvent exister soit sous forme unicellulaire, soit sous forme filamenteuse. Cela est particulièrement fréquent chez les espèces pathogènes chez lesquelles la forme levure se repand plus efficacement dans l'hôte, cependant que la phase filamenteuse facilite la phase de transmission et d'adhésion aux tissus (Nemecek et al., 2006).

Cet hyphe à croissance polarisée apicale, également retrouvé chez les oomycètes est une évolution remarquable. La croissance hyphale fait appel à plusieurs phénomènes cellulaires : l'établissement et le maintien de la polarité, le transport de matériels constitutifs des membranes et parois vers l'apex qui est le point de croissance mais aussi l'établissement de fusion des hyphes qui permet au mycelium de se densifier sous forme d'un entrelacs d'hyphes et chez certains champignons (excluant notamment les espèces appartenant aux Mucoromycota) la septation de ces hyphes permettant de définir des espaces cellulaires le plus souvent monocaryotiques ou dicaryotiques. La croissance apicale est assurée principalement grâce à une exocytose polarisée apicale. Cette croissance est initiée grâce à (i) l'établissement d'un site de croissance piloté notamment par le *Spitzenkörper* chez les dicarya (Lopez-Franco and Bracker, 1996) ou le croissant vésiculaire apical chez d'autres champignons, notamment les Mucoromycota (Fisher and Roberson, 2016), (ii) au maintien d'un axe de croissance, jalonné par le cytosquelette associé à diverses protéines motrices, le long duquel migrent depuis l'appareil de Golgi les matériels permettant l'élongation hyphale (membranes sous forme de vésicule et protéines notamment) (Steinberg, 2007). Même dans le cas des hyphes septés, la circulation de ces matériels le long de l'hyphe est rendue possible grâce à l'existence de pores (pouvant être obturés par diverses entités selon les groupes fongiques, comme par exemple les corps de Woronin) entre cellules. Des cascades de kinases et de nombreux régulateurs (voir Steinberg et al. (2017)) sont impliqués dans ce phénomène de croissance apicale.

Les hyphes fongiques agrégés permettent aux champignons de produire des structures plurihyphales complexes, jamais sous forme de vrais tissus mais plutôt sous la forme de plectenchymes tels que les carpophores (fructifications des dicarya portant les structures de production des spores sexuées), les sclérotes (propagules de résistance et de dissémination) ou encore les rhizomorphes (cordons mycéliens).

Le mode de vie fongique

Comme discuté ci-avant, les champignons qui sont des hétérotrophes sont caractérisés par leur osmotrophie couplée à leur mode de croissance hyphale ("je mange comme je croîs"). Au fur et à mesure de son exploration, le champignon peut exploiter la matière organique rencontrée, en la rendant assimilable grâce notamment à l'excrétion d'exoenzymes lytiques (cellulases, laccases, pectinases, estérases, lactases, péroxydases, lipases, protéases etc.). Les nutriments générés sont donc présents à l'extérieur de l'hyphe et le champignon peut donc se trouver en compétition avec

d'autres organismes pour l'import de ces métabolites. Dans cette compétition, le champignon possède deux armes (voir Richards et al. (2017)) : (i) sa croissance permet à la fois aux hyphes individuels d'explorer des zones peu accessibles mais également au thalle de constituer une masse imposante et importante de contact, de lyse mais aussi de transport vers l'intérieur des hyphes, et (ii) les champignons ont une très large gamme de voies de biosynthèses de métabolites secondaires, dont certains vont jouer un rôle non négligeable dans la compétition avec les autres organismes.

Au delà de ce trait commun aux champignons il n'existe pas une seule stratégie de vie fongique puisqu'on retrouve chez les champignons, qui sont des organismes hétérotrophes, des saprobes aussi bien que des mutualistes obligatoires, symbiotes ou parasites. Les saprobes libèrent des nutriments en dégradant la matière organique de divers habitats (sol, bois, végétaux ou animaux morts, fécès ou encore aliments), les symbiotes mycorhiziens grâce à une relation mutualiste avec des végétaux obtiennent du carbone sous forme de glucides issus de la photosynthèse. Les champignons peuvent se développer en tant que commensaux, endophytes de plantes ou se développant sur la peau animale ; ils peuvent être pathogènes, opportunistes ou pathogènes stricts, en envahissant les tissus animaux ou végétaux hôtes après avoir produit des toxines ou des effecteurs pour outrepasser les défenses de leurs hôtes en provoquant des maladies parfois mortelles (voir Stajich (2017)).

2.1.4 Le genre *Mucor*

Le genre *Mucor* est classé parmi la division des Mucoromycota, la sous-division des Mucoromycotina, la famille des *Mucoraceae* et l'ordre des Mucorales. Avec un total de 58 espèces décrites (Walther et al., 2013), le genre *Mucor* est le plus grand genre de l'ordre des Mucorales et de la famille des *Mucoraceae* (Morin-Sardin et al., 2017). Cependant, des phylogénies moléculaires basées sur de multiples loci tels que l'ADNr (ADN ribosomique nucléaire), l'ITS (*internal transcribed spacer*) de l'ADNr , le gène impliqué dans la synthèse du facteur d'élongation alpha (EF-1α) ou des séquences de gènes codant la production d'actine, ont montré que le genre *Mucor* n'était pas monophylétique [2] (Alvarez et al., 2011; de Souza et al., 2012; Voigt and Wostemeyer, 2001; Walther et al., 2013).

A titre d'exemple, lors de l'étude de Walther et al. (2013), incluant près de 400 souches de *Mucor*, les espèces de *Mucor* étaient intégrées dans des groupes phylogénetiques correspondant à

2. Monophylie : Un groupe monophylétique regroupe tous les descendants de l'ancêtre commun.

plus d'une vingtaine de genres tels que les *Chaetocladium*, *Helicostylum*, *Pilaira*, *Pirella*, *Thamnidium* ou *Zygorhynchus* et même des *Mycotypha* and *Choanoephora* qui ne sont pas des membres de la famille des *Mucoraceae* mais correspondent à la famille des *Mycotyphaceae* et *Choanepheraceae*, respectivement.

Cette polyphylétie [3] induit des changements dans la dénomination des espèces. Parmi les espèces les plus citées, *Rhizomucor miehei* apparaît souvent en tant que *Mucor miehei* et *Rhizomucor pusillus* est parfois appellé *Mucor pusillus*. De même,H *Mucor endophyticus* a longtemps été retrouvé sous le nom de *Rhizomucor endophyticus* et *Rhizopus arrhizus* apparaît régulièrement sous le nom de *Mucor rouxii* alors qu'il ne fait pas partie du genre *Mucor* (Hoffmann et al., 2013; Walther et al., 2013). La complexité du classement des espèces de *Mucor* sans apport de données moléculaires vient des caractéristiques morphologiques pouvant être très proches entre les espèces.

2.1.5 Caractéristiques des espèces du genre *Mucor*

Les travaux de Schipper (1967, 1969, 1970, 1976). Zycha et al. (1969) ont participé à la description morphologique des *Mucor*. Les *Mucor* sont caractérisés par une organisation rudimentaire. Ils possèdent un appareil végétatif, le thalle, qui correspond à un faisceau de filaments ou d'hyphes plus ou moins ramifiés qui constitue le mycélium (Figure 2.6).

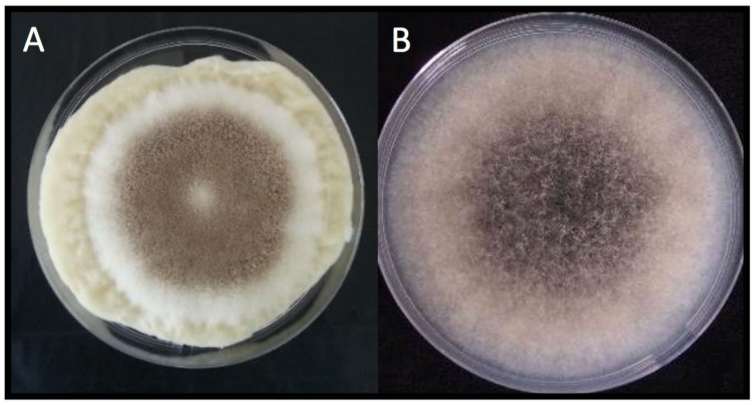


FIGURE 2.6 – Exemples de thalle de *Mucor lanceolatus* sur de la Tomme de Savoie non-affinée (A) et de *M. endophyticus* sur milieu *Potato Dextrose Agar* (B), après 4 jours de croissance à 15 °C (Source photographies : thèse de Morin-Sardin (2016))

3. Un taxon polyphylétique est un taxon défini par une ressemblance qui n'a pas été héritée d'un ancêtre commun. Ce terme désigne des taxons qui n'ont aucune pertinence pour retracer les liens de parenté et donc l'évolution.

Le mycélium, qui constitue la partie végétative du champignon s'accroît rapidement et peut présenter des ramifications pour coloniser le milieu. Les hyphes [4] sont diffus, tubulaires et généralement plus larges que les hyphes des *Dicarya*, avec un diamètre pouvant atteindre 15µm (thèse de Morin-Sardin (2016)). Les hyphes des *Mucor* sont généralement siphonnées, c'est à dire qu'il n'y a pas de cloisons transversales (septums) séparant différentes cellules de l'hyphe, ils contiennent donc plusieurs noyaux non séparés les uns des autres (Figure 2.7). Cette organisation structurale reste peu étudiée et la présence de plusieurs noyaux au sein des hyphes mais aussi des spores représente une particularité importante par rapport aux champignons dits "supérieurs" comme les ascomycètes et les basidiomycètes.

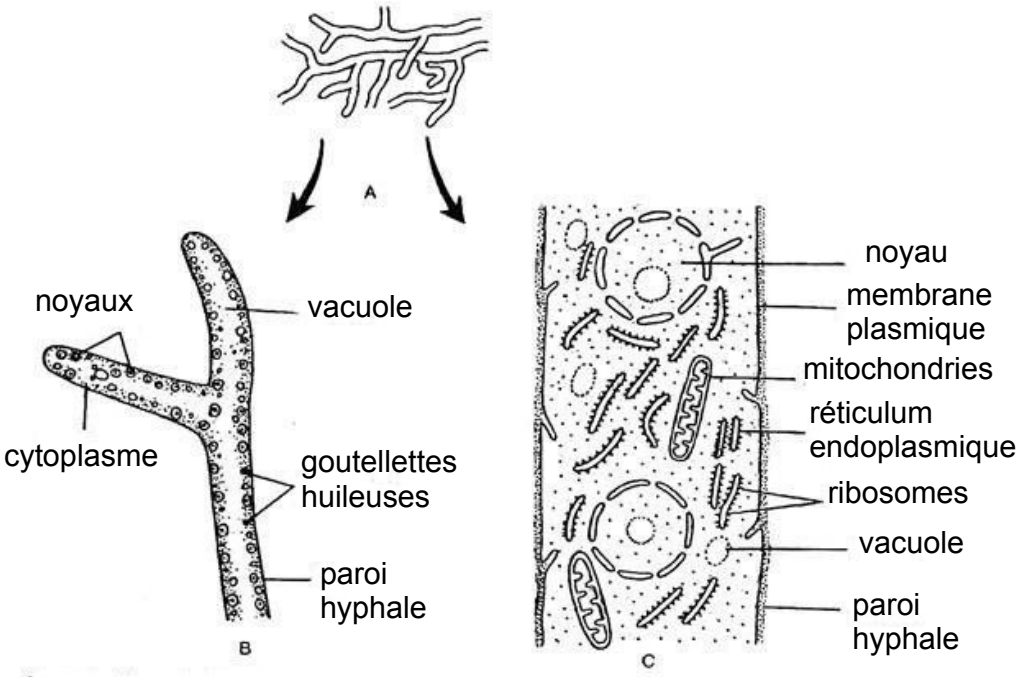


FIGURE 2.7 – Représentation schématique d'éléments mycéliens de *Mucor*. A. mycélium végétatif. B. portion d'hyphe vue sous microscope optique . C. Ultrastructure d'une portion d'hyphe vue sous microscope électronique.

La présence de septums chez les *Mucor* ne se rencontre que pour séparer les organes de reproductions sexuée et asexuée du reste du mycélium (Stajich et al., 2009) ou lors de l'isolement des parties sénescentes pour éviter la propagation de composés toxiques dans tout le mycélium (Carlile, 1995). L'extension mycélienne est générée au niveau de l'apex des hyphes. Les zones en arrière de l'apex peuvent néanmoins émettre des bourgeonnements, néoformations de nouveaux axes à croissance linéaire donnant naissance à des ramifications.

4. Filaments constitutif du mycélium

Le mycélium des *Mucor* est haploïde [5] (et comme nous venons de le préciser multinucléé puisque siphonné), seules les cellules issues de la reproduction sexuée générant les zygospores, sont diploïdes. La formation de la zygospore se fait par fusion (d'où le nom du phylum obsolète zygomycètes dérivé du grec zugos qui signifie fusion) de deux zygophores issus du mycélium d'un même individu (pour les espèces de *Mucor* homothalliques) ou de deux individus compatibles (pour les espèces de *Mucor* hétérothalliques). La compatibilité sexuelle chez les Mucoromycotina reposerait sur l'existence de deux allèles (il s'agit d'un système bipolaire) sexP (allèle +) et sexM (allèle -) qui caractérisent les deux types sexuels. Le gène correspondant à ces deux allèles code un facteur de transcription comportant un domaine HMG (*high mobility group*) comme observé en premier lieu chez *Phycomyces* (Idnurm et al., 2008) puis plus tard chez des *Mucor* : *M. circinelloides*, *M. mucedo* et *M. ambiguus* (Wetzel et al., 2012; Corrochano et al., 2016; Schulz et al., 2017). Le zygotropisme qui permet aux deux zygophores de se diriger l'un vers l'autre puis de fusionner (ils peuvent alors être appelés progamétanges puis gamétanges ou suspenseurs une fois que le zygote entame sa formation) est contrôlée par l'émission de phéromones (Burgeff, 1924; Plempel, 1962) qui ont été identifiées comme étant des acides trisporiques (Austin et al., 1969) dont la synthèse est initiée par le β-carotène. Les acides trisporiques activeraient les facteurs de transcription sexP et sexM (Lee et al., 2014) mais les gènes cibles de ces facteurs de transcription n'ont pas encore été identifiés (Schulz et al., 2016).

La reproduction sexuée dépend donc des conditions environnementales susceptibles d'induire l'expression des gènes dits sexuels, et chez les espèces hétérothalliques qui restent majoritaires, de la compatibilité de deux individus. La production de spores sexuées peut-être très abondante comme par exemple chez *M. endophyticus* (observation personnelle) mais n'est pas toujours présente.

La reproduction asexuée reste généralement toujours abondante chez l'ensemble des espèces de *Mucor*. Les spores asexuées majoritaires, les sporangiospores, ont la particularité d'être produites à l'intérieur de sporanges. Chaque sporange contient un nombre variable de sporangiospores mais on ne retrouve pas au sein du genre *Mucor* des sporanges paucisporés (contenant peu de spores) ou sans columelle vraie (la columelle est une structure en forme de vésicule qui sépare le sporange du sporangiophore et de l'ensemble du mycelium). Ces sporangiospores peuvent être multinuclées (Hammill and Secor, 1983; Tansey et al., 1984). D'autres spores asexuées, les

5. Chaque noyau ne comporte qu'un seul exemplaire de chaque chromosome.

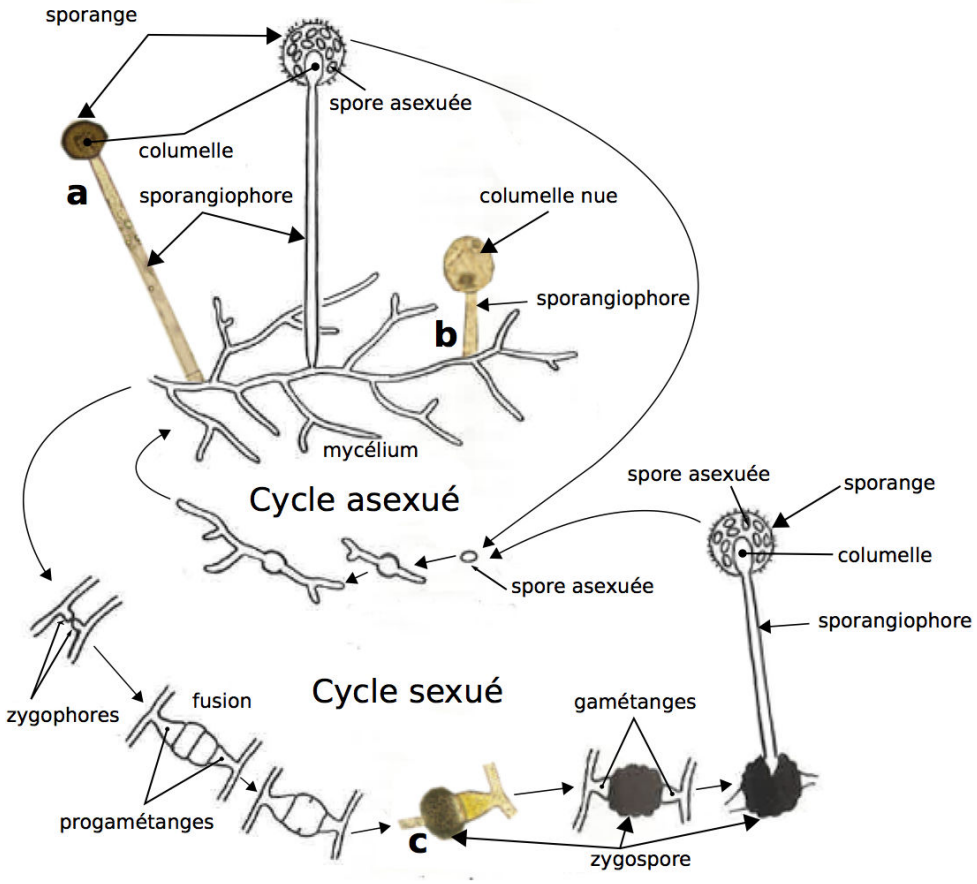


FIGURE 2.8 – Représentation schématique du cycle de vie d'un *Mucor* sp. (d'après une figure de C.T Ingold modifiée, (Ingold and Others, 1978) (Ingold, 1978)) comportant (i) le cycle asexué durant laquelle les spores asexuées sont formées à l'intérieur d'un sporange isolé du reste du mycélium par la columelle et porté par un sporangiophore qui se développe à partir de mycélium non différencié et (ii) le cycle sexué au cours duquel une zygospore (seul élément diploïde du cycle de vie) est formée après fusion entre deux hyphes sexuellement compatibles. (a) sporangiophore et sporange intact contenant les spores asexuées de *M. fuscus* ; (b) sporangiophore portant une columelle nue (la paroi du sporange a éclaté, libérant les spores asexuées) de *M. racemosus* et (c) zygospore de *M. endophyticus*.

chlamydospores, toujours produites à l'intérieur des hyphes (le plus souvent âgées ou sénescentes) chez les *Mucor*, peuvent être présentes en quantité importante chez certaines espèces comme *M. racemosus*.

2.1.6 Modes de vie et habitats des *Mucor*

Les spores de *Mucor* ont une forte capacité d'absorption de l'humidité, d'adhérence à de nombreuses surfaces et de dispersion dans l'air humide. Elles peuvent germer rapidement dans l'environnement si un substrat favorable est trouvé (Et et al., 1972). Ces caractéristiques participent au caractère pionnier des espèces du genre *Mucor*. La stratégie écologique des *Mucor*

consiste en une croissance et sporulation importantes mais une faible persistance dans le milieu (Morin-Sardin et al., 2017). Les conditions de croissance optimale de ces espèces sont identifiées comme suit : température 20 à 25°C, pH de 5-6 et activité de l'eau supérieure à 95% (Dantigny et al., 2005; Morin-Sardin et al., 2016; Panasenko, 1967; Pitt and Hocking, 2009) . Elles présentent une faible tolérance aux faibles activités de l'eau (Morin-Sardin et al., 2016) mais sont capables de coloniser des niches écologiques extrêmement diverses (Walther et al., 2013). Majoritairement saprophytes, ces espèces sont abondantes et fréquentes dans l'environnement (Hoffmann et al., 2013; Voigt et al., 2016; Voigt and Wostemeyer, 2001), certaines pouvant être associées à des habitats ou niches spécifiques. *M. endophyticus* CBS 385-95 dont nous avons étudié le génome dans cette étude a été décrite comme un endophyte du blé (Zheng and Jiang, 1995) tandis que *M. lanceolatus*, également utilisée dans cette étude, n'a été décrite qu'en milieu fromager en tant qu'espèce utilisée pour l'affinage de fromage (Hermet et al., 2012; Walther et al., 2013; Morin-Sardin et al., 2017). Ces deux espèces spécialisées semblent présenter des caractéristiques d'adaptation à leur milieu, en effet Morin-Sardin et al. (2016) ont montré que *M. lanceolatus* et *M. fuscus* (utilisées en affinage de fromage) se développaient mieux sur un milieu fromager que sur un milieu de culture synthétique contrairement à *M. racemosus*, *M. circinelloides*, *M. brunneogriseus*, *M. spinosus* et *M. endophyticus*, de même *M. endophyticus* accusait un retard de croissance sur milieu de culture synthétique par rapport aux six autres espèces de l'étude.

A contrario certaines espèces sont retrouvées dans de multiples milieux comme les saprotrophes *M. circinelloides* et *M. indicus*, également investigués dans le cadre de cette thèse. Ces deux espèces ont notamment la particularité d'être fréquemment associées à l'environnement clinique et constituent des pathogènes opportunistes d'animaux ou de l'homme (Alvarez et al., 2009). Leur succès en tant que pathogènes opportunistes est en partie lié à leur caractère thermophile qui leur donne un avantage compétitif pour se développer au sein des tissus animaux et humains. Robert and Casadevall (2009) lors d'une étude sur la tolérance des champignons à la température indiquaient que la plupart des champignons ne pouvaient pas croître à la température des mammifères, chaque degré de plus dans la fenêtre des 30 à 40°C excluant 6% de plus des 4802 isolats fongiques testés. *M. indicus* par exemple tolère des températures de 40°C (Taj-Aldeen et al., 2017).

2.1.7 Intérêt biotechnologique des *Mucor*

Les caractéristiques associées au mode de vie pionnier de ces espèces, (fort taux de croissance, facilité de développement sur de nombreux milieux, diversité des enzymes produites pour exploiter divers substrats) en font également de bon candidats pour être utilisés dans des applications biotechnologiques. Cela concerne en particulier les espèces ubiquistes capables de croître dans une gamme étendue de pH et température tel que *M. circinelloides* (Zhang et al., 2007; Sautour et al., 2002; Millati et al., 2005; Nahas, 1988; Sajbidor et al., 1988). Parmi les composés d'intérêt pouvant être produits par les *Mucor* on compte des acides organiques (ex : acide lactique et fumarique), de l'éthanol et acides gras (pouvant être utilisé comme biocarburants, éthanol et biodiesel), des acides gras polyinsaturés, des carotenoïdes ou encore des enzymes variées (amylases, cellulases, protéases, lipases). De plus, la biomasse générée peut être utilisée pour l'alimentation des animaux et poissons du fait de leur forte valeur nutritionnelle (Voigt et al., 2016; Karimi and Zamani, 2013; Ferreira et al., 2013)

2.1.8 Métabolisme des *Mucor*

Des études métaboliques ont été conduites sur certains *Mucor* d'intérêt industriel. Les plus étendues portant sur la production de lipides (Tang et al., 2016, 2017) et la production de caroténoïdes (Voigt et al., 2016).

Malgré le potentiel pathogène opportuniste de certains *Mucor*, le seul métabolite secondaire toxique détecté chez les *Mucor* correspond à l'acide neurotoxique 3-nitropropionique, un inhibiteur irréversible de la succinate déhydrogénase menant à des apoptoses anormales (voir Ludolph and Ludolph (1991)) chez *M. circinelloides* (Hollmann et al., 2008). Cependant, une étude de 2014 (Lee et al., 2014) a permis de montrer la présence dans les génomes de *M. circinelloides* (formes *circinelloides* et *lusitanicus*) de gènes impliqués dans la synthèse de métabolites secondaires, gènes que les auteurs rapprochaient d'une potentielle production de toxines. Ces gènes qui correspondraient à des polycétides synthases (PKS), enzymes de synthèses de peptide non ribosomiques (NRPS) et L-triptophane dimethylallyltransférase (DMAT) ont par la suite été identifiés par Voigt et al. (2016) dans les génomes de Mucoromycota publiquement disponibles et par Hermet (2013) dans les génomes de *M. fuscus*, *M. lanceolatus* and *M. racemosus*.

2.1.9 Analyses génomiques retrouvées dans la littérature concernant les *Mucor*

Les génomes de certaines souches de *Mucor spp.* ont été séquencées et sont disponibles publiquement. Le premier génome de *Mucor* disponible fût celui *M. circinelloides* forme *lusitanicus* CBS 277.49 déposé sur le JGI en 2009 (Corrochano et al., 2016), en 2018 dix génomes de *Mucor* étaient accessibles publiquement (Tableau 2.1). Les souches séquencées dont l'origine est connue sont majoritairement originaire d'environnements médicaux.

Taxon	Souche	Source de l'isolat	Référence ou n° d'accession	Annotation disponible
Mucor ambiguus	NBRC 6742	Inconnue	BBKB00000000.1, 2015	oui
Mucor circinelloides	1006PhL	Humain	Findley et al., 2013	oui
Mucor circinelloides	CBS 277.49	Inconnue	Corrochano et al., 2016	oui
Mucor circinelloides	CDC-B8987	Humain	Chibucos et al., 2016	sur demande
Mucor circinelloides	JCM 22480	Inconnue	BCHG00000000.1, 2016	non
Mucor circinelloides	WJ11	Sol	Tang et al., 2015	non
Mucor indicus	CDC-B7402	Humain	Chibucos et al., 2016	sur demande
Mucor irregularis	B50	Humain	AZYI00000000.1, 2014	non
Mucor irregularis	B7584	Inconnue	JNES00000000.1, 2014	non
Mucor velutinous	CDC-B5328	Humain	Chibucos et al., 2016	sur demande

TABLEAU 2.1 – Liste des génomes de *Mucor* accessibles publiquement

Le nombre d'études à l'échelle des génomes de *Mucor* est restreint. A notre connaissance, deux études de génomique comparative se focalisent uniquement sur les *Mucor*. La première, réalisée par Tang et al. (2015), a permis d'identifier les gènes impliqués dans les voies principales du métabolisme du carbone et des lipides et de déterminer quels gènes étaient susceptibles d'être impliqués dans l'importante production de lipides chez une souche de *M. circinelloides*. Ces résultats ont été obtenus grâce à la comparaison de deux souches de *M. circinelloides* l'une à la production de lipides importante, l'autre non. La seconde étude, réalisée par Lopez-Fernandez et al. (2018) a permis, grâce à la comparaison de deux souches de *M. circinelloides*, l'une virulente et l'autre avirulente, d'identifier des éléments pouvant être liés à la pathogénicité de l'espèce. D'autres études plus étendues incluaient également des *Mucor*. C'est le cas des recherches de Corrochano et al. (2016) sur la perception de l'environnement et de la transduction du signal chez les champignons et plus particulièrement des Mucoromycotina (représentés par *Lichtheimia corymbifera*, *Phycomyces blackesleeanus*, *M. circinelloides* et *Rhizopus delmar*). Cette étude a permis d'identifier une duplication complète du génome d'un ancêtre des Mucoromycota ayant permis d'améliorer la perception des signaux environnementaux chez les espèces de cette lignée. L'analyse la plus étendue de génomique comparative s'intéressant aux *Mucor* était celle de Chibucos et al. (2016). Cette étude avait pour objet d'identifier des éléments liés à

la pathogénicité des espèces impliquées dans les *Mucor*mycoses, pour ce faire 27 génomes de l'ordre des Mucorales ont été comparés dont trois *Mucor* [6] (deux souches de *M. circinelloides* [7] et une de *M. indicus*).

Au cours de cette étude deux comparaisons génomiques incluant les *Mucor* ont été réalisées. Ces deux analyses portaient sur la recherche d'expansion/contraction du nombre de gènes de familles spécifique : les chitines déacetylases d'une part et des gènes codant des invasines de type "*spore coat protein homologs*" (*CotH*). Lors de cette étude, a notamment été mise en valeur une corrélation positive entre le nombre de copie de *CotH* et la pathogénicité de l'espèce.

L'augmentation du nombre de copies de *CotH* participe donc probablement à l'adaptation des espèces pathogènes opportunistes à un mode de vie pathogène d'hôtes animaux tout comme la spécialisation de mécanismes d'acquisition du fer (ex : famille de gènes *fet3* (Navarro-Mendoza et al., 2018)) ou encore des mécanismes permettant la tolérance des espèces à des températures supérieures à 37°C (Taj-Aldeen et al., 2017).

La comparaison des génomes d'espèces proches permet de mettre en évidence les différences qui font la diversité des espèces et en cela d'identifier des éléments potentiellement liés à leur adaptation aux conditions environnementales de leur habitat. En effet, la composition et l'organisation des éléments d'un génome est le résultat de mécanismes et processus qui ont permis de diversifier les organismes de la population, les organismes les plus adaptés (fitness élevé) ayant plus de chances de se développer optimalement et d'obtenir une descendance, la population entière devient à terme plus adaptée à l'environnement.

Pour identifier ces éléments et processus ayant contribué à l'adaptation d'un organisme à un milieu, il convient tout d'abord de revenir aux bases : qu'est-ce qu'un génome ? Quels sont les éléments qui le composent ? Quels sont les mécanismes induisant des différences d'un génome à l'autre ? La partie ci-après a pour objet de donner quelques éléments de réponse à ces différentes questions, un accent étant porté sur les éléments associés aux processus d'adaptations.

6. Les deux *M. racemosus* de cette publication ont par la suite été renommés *Rhizopus microsporus* par Gryganskyi et al. (2018).
7. L'une des souches de *M. circinelloides* est indiquée *M. velutinosus* (synonyme de *M. circinelloides* (Walther et al., 2013)) dans la publication.

2.2 Caractéristiques génomiques

Le génome, est l'ensemble du matériel génétique d'un organisme. Celui-ci est codé par des molécules d'acide désoxyribonucléique (ADN). Les chaînes d'ADN, pour pouvoir être contenues dans le noyau de la cellule (cas eucaryote), doivent être compactées (au moins 100 000 fois selon Razin et al. (2007)). En parallèle, certaines régions d'ADN doivent rester accessibles pour les interactions avec des facteurs de transcription et les machineries de transcription/réplication de l'ADN. Pour résoudre ce problème, l'ADN est compacté en chromatine par le biais de multiples étapes de compactions impliquant des ARN (acide ribonucléique) et des protéines (histones et non histones) (Razin et al., 2007; Kalitsis et al., 2017) (Figure 2.9). On distingue deux types de chromatine correspondant à des niveaux différents de compaction : l'euchromatine et l'hétérochromatine. L'euchromatine correspond à une chromatine moins condensée dans laquelle les gènes, plus accessibles, voient leur expression facilitée. L'hétérochromatine correspond à une chromatine plus dense avec un ADN moins facilement accessible. On différencie deux types d'hétérochromatine : l'hétérochromatine constitutive qui correspond à des régions qui ne sont globalement pas exprimées (télomères et centromères par exemple) et l'hétérochromatine facultative qui correspond à des régions qui ne sont pas exprimées mais qui, selon les régulations de la cellule peuvent être décompactées pour former de l'euchromatine (Yadav et al., 2018). Lors de la division cellulaire la compaction de l'ADN augmente encore pour laisser apparaître lors de la métaphase la structure en chromosomes de l'ADN.

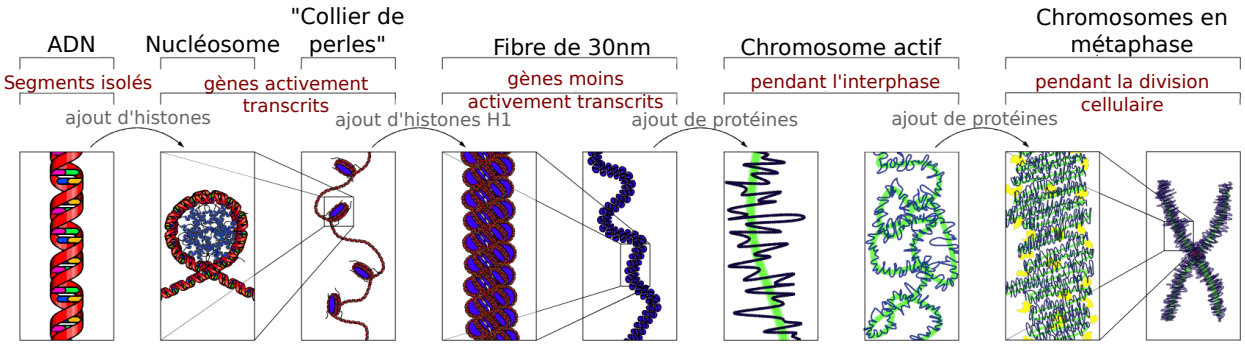


FIGURE 2.9 – Compaction de l'ADN (Source : R. Wheeler).

Chez les eucaryotes, l'ADN est principalement retrouvé dans le noyau (génome nucléaire) mais également dans les mitochondries et d'autres organelles ainsi que dans des plasmides (Smith and Keeling, 2015). Dans une cellule, chaque chromosome peut être présent en un ou plusieurs exemplaires. Lorsque les chromosomes nucléaires sont présents en un seul exemplaire dans le noyau on parle de cellules haploïdes. Lorsque l'ensemble des chromosomes nucléaires

sont répétés, on parle de "ploïdie" précédé d'un préfixe précisant le nombre de répétitions, ainsi, une cellule dont le noyau possède deux copies du lot haploïde sera qualifiée de diploïde. Lorsque le nombre de répétitions n'est pas le même pour les différents chromosomes du lot haploïde on parle d'aneuploïdie (Figure 2.10). Les polyploïdies issues du croisement de deux individus de la même espèce sont qualifiées d'autopolyploïdies, lorsqu'elles sont issues du croisement de deux individus d'espèces différentes on parle d'allopolyploïdie. On note que lorsqu'une cellule possède plusieurs noyaux comme dans les hyphes siphonnés de *Mucor* (Morin-Sardin et al., 2017), le phénotype de la cellule peut être confondu avec celui d'une cellule polyploïde.

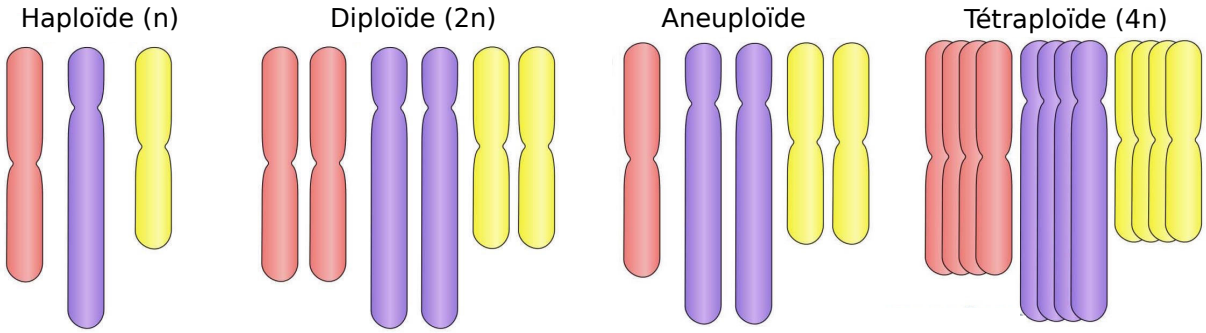


FIGURE 2.10 – Représentation de différents niveaux de ploïdie.

Au sein d'un même organisme ayant recours à la reproduction sexuée, deux phases de vie sont observées l'une diploïde issue d'un processus de syngamie (i.e. fusion de deux cellules haploïdes), l'autre haploïde obtenue a la suite d'un processus de méiose. Comme indiqué plus tôt, chez les *Mucor*, les cellules sont majoritairement haploïdes, seules les cellules issues de la reproduction sexuée générant les zygospores, sont diploïdes (Morin-Sardin et al., 2017).

Au sein d'une même espèce on peut observer des différences importantes entre les génomes des individus. Ces différences peuvent être liées au type sexuel de l'individu, à un niveau de ploïdie différent, à la duplication ou perte de tout ou partie de chromosomes (Zhu et al., 2016; Wertheimer et al., 2016; Scott et al., 2017) (Exemple de *Candida albicans* en Figure 2.11). Ainsi, chez certaines espèces, on peut observer des chromosomes dits "accessoires" qui bien que conservés à l'échelle de la population peuvent être perdus chez un individu (Habig et al., 2017).

A plus petite échelle, les séquences varient également d'un individu à l'autre : polymorphisme nucléotidique (*single nucleotide polymorphism* ou SNPs), insertion et délétion de nucléotides (INDEL), parfois des variations du nombre de copies d'éléments génomiques (gène et élément répétés notamment) (Zhu et al., 2016).

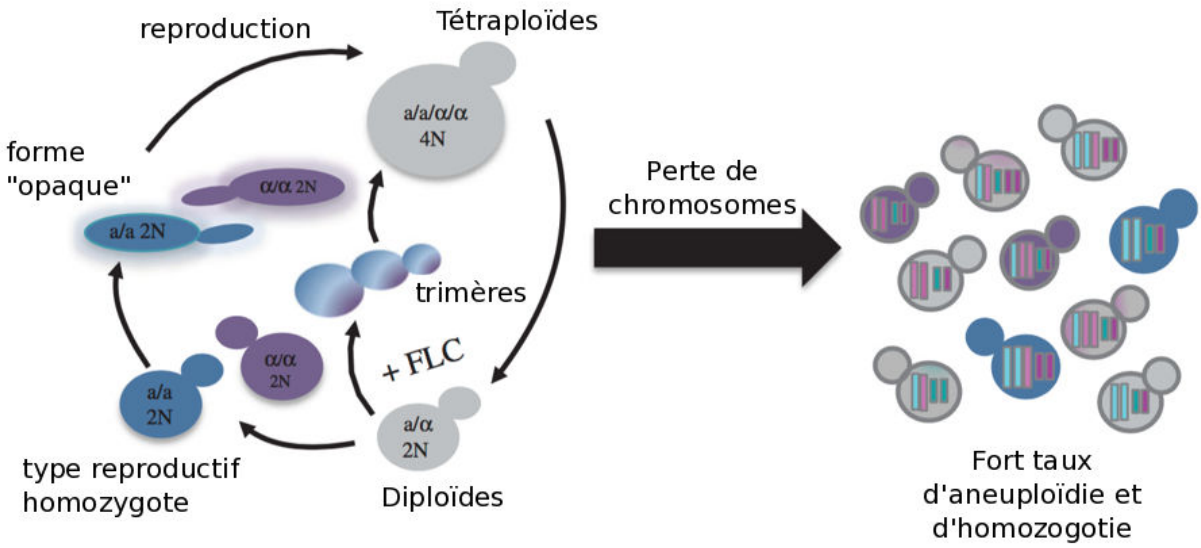


FIGURE 2.11 – Représentation schématique des mécanismes de changements de ploïdie chez *Candida albicans* (source : Wertheimer et al. (2016)). Les individus diploïdes homozygotes au niveau du locus du type sexuel et qui prennent la forme dite "opaque" peuvent se reproduire sexuellement pour former des tétraploïdes. Des tétraploïdes peuvent également être générés, en présence de fluconazole (+FLC), via la formation de trimères. Dans les deux cas, les tétraploïdes peuvent subir des ségrégations incomplètes des chromosomes et/ou des réduction de ploïdies qui aboutissent à une diversité genotypique et phénotypique.

2.2.1 Éléments génomiques

Les génomes sont constitués d'une grande diversité d'éléments aussi bien structuraux que régulateurs. Les plus étudiés sont les gènes codant des protéines mais on retrouve également des gènes non codants (ex : ARN de transfert, ARN ribosomiques, micro ARN, ARN long non codant etc.), des répétitions (ex : microsatellites, éléments transposables), des reliquats d'éléments devenus non fonctionnels (ex : pseudogènes), des régions intergéniques etc. Dans le cadre de ce manuscrit, seuls deux groupes d'éléments génomiques seront abordés : les gènes codant des protéines, et les éléments répétés. En effet, ils représentent les deux classes d'éléments génomiques majoritairement abordés lors des analyses.

Les gènes codant des protéines

La synthèse de protéine à partir d'un gène comprend trois étapes majeures chez les eucaryotes : la transcription, l'épissage et la traduction (Figure 2.12).

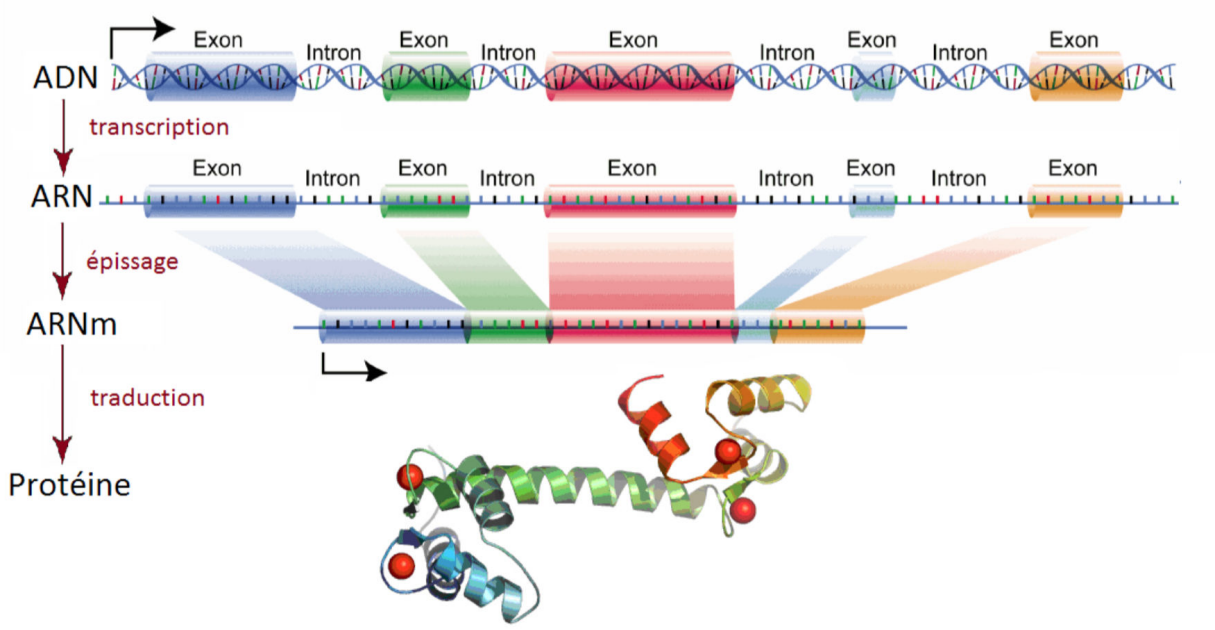


FIGURE 2.12 – Formation d'une protéine à partir de l'ADN (source : Rotival (2011)).

La transcription permet l'encodage de la séquence d'ADN en ARN (acide ribonucléique). L'ADN double brin est séparé en ADN simple brin autour du gène puis transcrit dans le sens 5'-3' par une ARN polymérase pour former un pré-ARN ou ARN primaire. Cet ARN est formé des quatre bases ribonucléiques AUGC (on peut également rencontrer le nucléotide atypique inosine (I) (Morita et al., 2013)). Cet ARN primaire est transformé en ARN messager (ARNm) par épissage. Lors de cette étape, des enzymes retirent des parties non codantes de l'ARN primaire (les introns). Seules les parties codantes (exons) et les régions de début et de fin de gènes (nommées parties 5' non traduite (5'UTR) et 3' non traduite (3'UTR)) sont conservées dans l'ARNm résultant. Chez les champignons, les caractéristiques des exons diffèrent significativement selon les espèces (Galagan et al., 2005). Cependant, Ivashchenko et al. (2009) ont identifiés que la taille moyenne des exons fongiques décroissait avec l'augmentation du nombre d'introns. En moyenne chez les Basidiomycètes, on retrouve 4-5 introns par gènes ; chez les Mucoromycotina 3-4 introns, tandis que chez les Ascomycètes et oomycètes on retrouve 1-2 introns, (Mohanta and Bae, 2015). Les introns fongiques sont généralement courts avec une moyenne de 80 à 150pb chez des Ascomycètes (Loftus et al., 2005). On note que certaines espèces tel que *Cryptococcus*

neoformans possèdent des introns très courts. Chez cette espèce, les introns sont composés de 68pb en moyenne, nombre d'entre eux faisant 35pb. A titre de comparaison chez l'Homme les gènes ont en moyenne 10 introns, les introns faisant en moyenne 6355pb (Piovesan et al., 2016). Lorsqu'un gène est composé de plusieurs exons, des cas d'épissages alternatifs peuvent être observés : certains exons peuvent être retirés, des introns conservés, des bordures introns/exons être altérées (Figure 2.13) (Galagan et al., 2005). Ces variations d'ARNm, qualifié d'isoformes, mènent à la formation de protéines distinctes à partir d'un même gène.

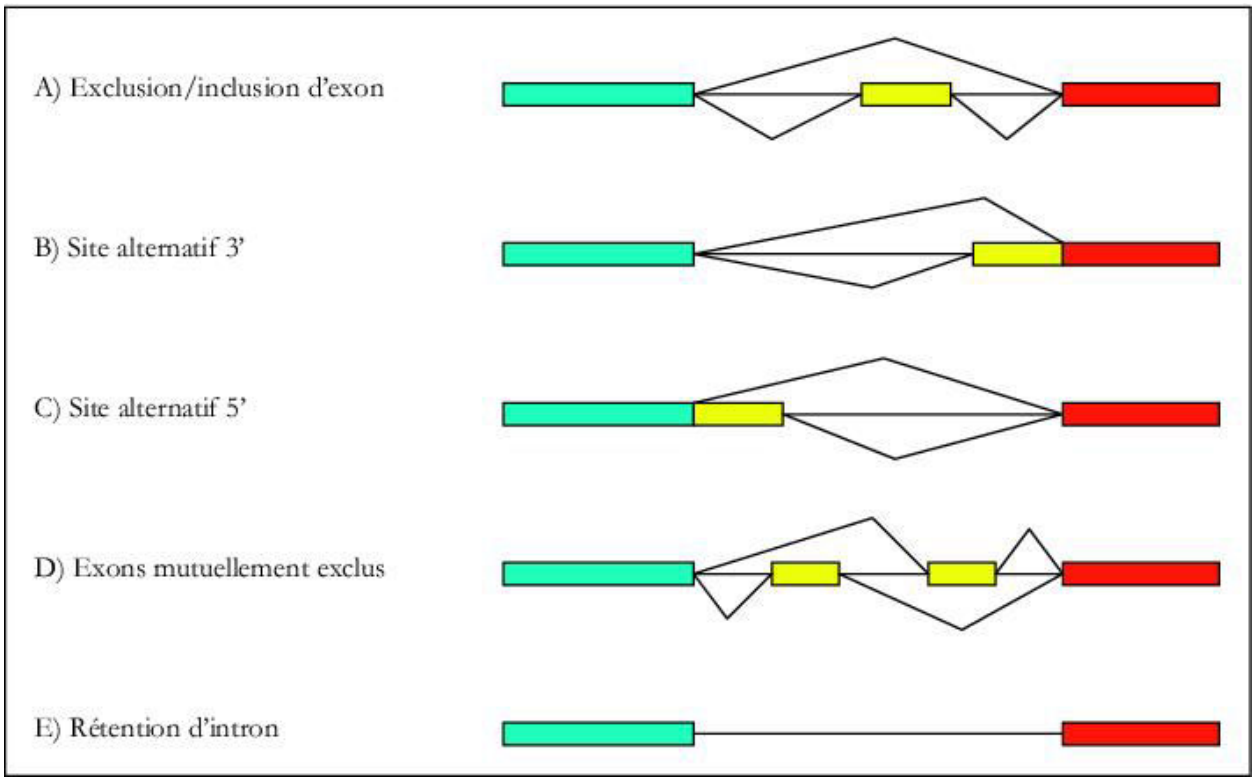


FIGURE 2.13 – Les épissages alternatifs (Source : P. Jeanteur).

D'autres modifications post-transcriptionnelles interviennent également comme l'ajout d'une coiffe au début de l'ARNm et d'une queue polyA (en fin de l'ARNm) qui protègent les ARNm d'une prompte dégradation. Des modifications plus rares existent comme des mécanismes d'édition permettant à la cellule de modifier la séquence de l'ARNm après la transcription. La séquence polypeptidique qui résulte de la traduction de cet ARNm ne correspond donc pas à la séquence exacte du gène correspondant. L'ARNm est par la suite reconnu et traduit par les ribosomes qui assurent la synthèse de la protéine par lecture successive de triplets de bases ribonucléiques (les codons) et assemblage progressif d'une chaîne d'acides aminés. Chaque ribosome fixé à l'ARNm va ainsi synthétiser un exemplaire de la protéine codée par le gène. Là encore des modifications peuvent intervenir sur la protéine synthétisée que ce soit juste après

sa synthèse ou au cours de sa vie dans la cellule (Rotival, 2011). Après leur synthèse, les ARN sont détruits dans la cellule par un groupe d'enzymes de dégradation. Cette dégradation a lieu en continu dans la cellule et prend entre quelques minutes et quelques jours selon l'ARN. La composition de l'ensemble des ARN à un temps donné, aussi appelé transcriptome, peut donc être modifiée rapidement (Rotival, 2011).

Chez les champignons, la densité en gènes codant des protéines est inversement proportionnelle à la proportion de gènes contenant des introns (Ivashchenko et al., 2009). Les gènes peuvent être très rapprochés les uns des autres pouvant générer des chevauchement d'UTR. Un autre trait marquant des champignons vient de l'organisation des gènes impliqués dans la biosynthèse des métabolites secondaires. En effet, chez les Ascomycota et Basidiomycota, ces gènes sont regroupés en cluster métaboliques (*Biosynthetic Gene clusters* ou BGC) (Lind et al., 2017). Ces clusters sont défini comme des gènes physiquement proches codant les enzymes, protéines régulatrices et transporteurs nécessaires à la production, au traitement et à l'export d'un métabolite spécialisé (Medema and Fischbach, 2015) (exemple en Figure 2.15).

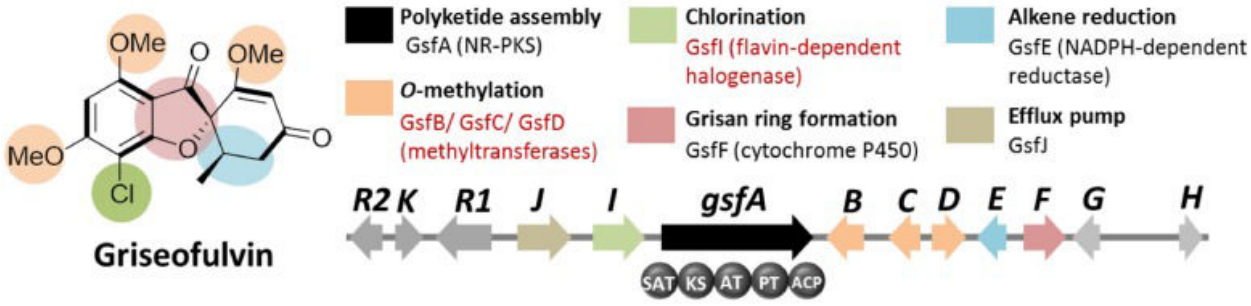


FIGURE 2.14 – Cluster métabolique (BGC) associé à la synthèse de griseofulvine chez *Penicillium aethiopicum*. (source (Cacho et al., 2014)).

Cette caractéristique est parfois étendue aux champignons en général (Wisecaver and Rokas, 2015; Slot, 2017) bien que, jusqu'à présent, aucune publication ne retrace la présence de BCG validés expérimentalement ni chez les Mucoromycota ni chez les Zoopagomycota.

Les éléments répétés

Les éléments répétés sont des séquences d'ADN retrouvées en de nombreuses copies dans un même génome. Certaines, comme les satellites, sont présentes en de longs groupes de motifs similaires au niveau d'un petit nombre de sites tandis que d'autres, en particulier les éléments transposables (TE), sont dispersées sur le génome.

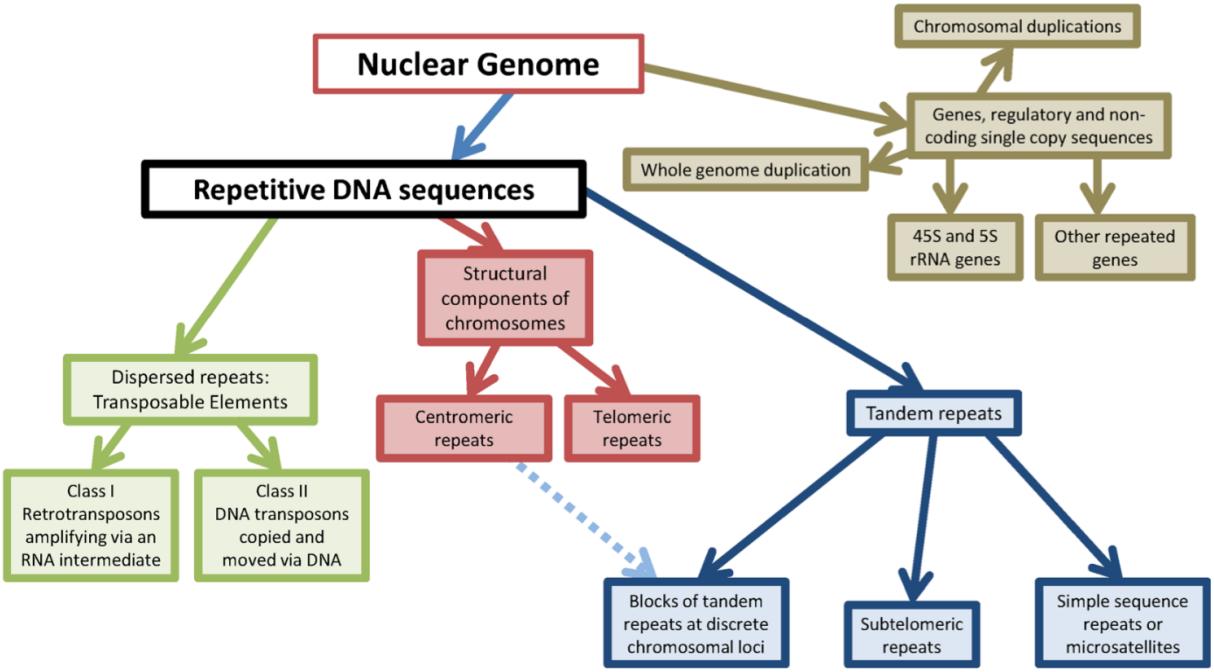


FIGURE 2.15 – Diversité des éléments répétés de génomes eucaryotes (source : Biscotti et al. (2015)). Les éléments répétés peuvent être largement dispersés sur le génome ou situés au niveau de de larges régions chromosomiques. Les éléments répétés en tandem sont généralement localisés en blocs au niveau d'un ou plusieurs sites sur les chromosomes.

Le taux de répétitions présent dans les génomes fongiques a tout d'abord été estimé comme étant faible : en 2005, Galagan et al. (2005) estimaient le taux de répétition "typique" allant de 3% à 10% du génome. Cependant, l'augmentation du nombre de génomes fongiques séquencés a permis de découvrir des espèces possédant un taux de répétition bien plus important (ex environ 74% *Phytophthora infestans*) (Raffaele and Kamoun, 2012; Galagan et al., 2005). La principale source de répétition correspond aux TE. Les TE sont définis comme étant des segments d'ADN discrets capables de se déplacer au sein d'un génome hôte d'une position chromosomique ou plasmidique à une autre et qui n'utilisent pas de machinerie moléculaire spécifique qu'ils codent pour infecter le génome de nouveaux hôtes par transfert horizontal (Piégu et al., 2015). Différents systèmes de classification des familles de TE ont été proposées (pour une revue voir Piégu et al. (2015)), parmi elles la classification de Wicker (Wicker et al., 2007) utilisée dans le cadre de cette

étude est basée sur le mécanisme de transposition, la similarité de séquence et la structure des TE (Figure 2.16). Les TE sont séparés en deux classes principales : les rétrotransposons (classe I) et les transposons à ADN (classe II). Les transposons à ADN se déplacent par un système de "couper coller". Leurs séquences codent une transposase permettant de catalyser l'excision du transposon et son intégration dans une nouvelle région génomique. Les rétrotransposons quant à eux se déplacent par un système de "copier coller". Ils sont la classe la plus abondantes des TEs. Les séquences sont transcrites puis une transcriptase inverse permet d'obtenir à partir de l'ARN la séquence correspondante d'ADN qui sera intégrée dans une nouvelle région génomique. Cinq ordres (LTR, DIRS, PLE, LINE, SINE) sont définis pour la classe I et quatre (TIR, Crypton, Helitron, Maverick) pour la classe II. Chaque ordre est ensuite divisé en une ou plusieurs superfamilles pour un total de 29 superfamilles.

A cette classification est ajoutée la notion d'autonomie des TE. On entend par élément autonome, un élément dont tous les domaines typiquement nécessaires pour sa transposition sont présents. Cela n'implique pas que l'élément soit fonctionnel ni actif (Wicker et al., 2007). On compte quatre catégories principales d'éléments non autonomes : les LArD (*Large retrotransposon derivative*), les MiTE (*Miniature inverted-repeat transposable element*), les SNACs (*small non-autonomous CACTA*) et les TriM (*Terminal repeat retrotransposon in miniature*).

L'identification des TE est complexe du fait de leur grande diversité aussi bien dans un même génome qu'entre génomes différents (Castanera et al., 2016, 2017). Comme pour les gènes, l'identification des TE est dépendante de la qualité et complétude des bases de données (Carr et al., 2012), celles-ci étant biaisées en faveur des espèces modèles, les TE atypiques seront plus difficile à identifier (Steinbiss et al., 2009; Quesneville et al., 2005). De plus, les TE accumulent des mutations au cours du temps, un élément ancien ayant accumulé de nombreuses mutations sera difficile à identifier. Des réarrangements peuvent également avoir lieu : perte d'une région interne, insertion d'un autre élément ou autre modification (Flutre et al., 2011; Castanera et al., 2016). Un TE avec un faible nombre de copies ne pourra pas être identifié par des méthodes de novo (par exemple, lors des premières études sur le génome du riz, 30% des séquences annotées comme des gènes, étaient des éléments transposables ou des fragments d'éléments transposables (Bennetzen et al., 2004). Au contraire, des gènes présents en de très nombreuses copies pourront être identifiés comme TE potentiels (Bennetzen and Park, 2018). Il faut également noter que certains gènes sont issus de la domestication de protéines de TE [8] , c'est par exemple le cas

8. Recrutement de protéines d'élément transposable par l'organisme hôte pour les utiliser comme protéines cellulaires.

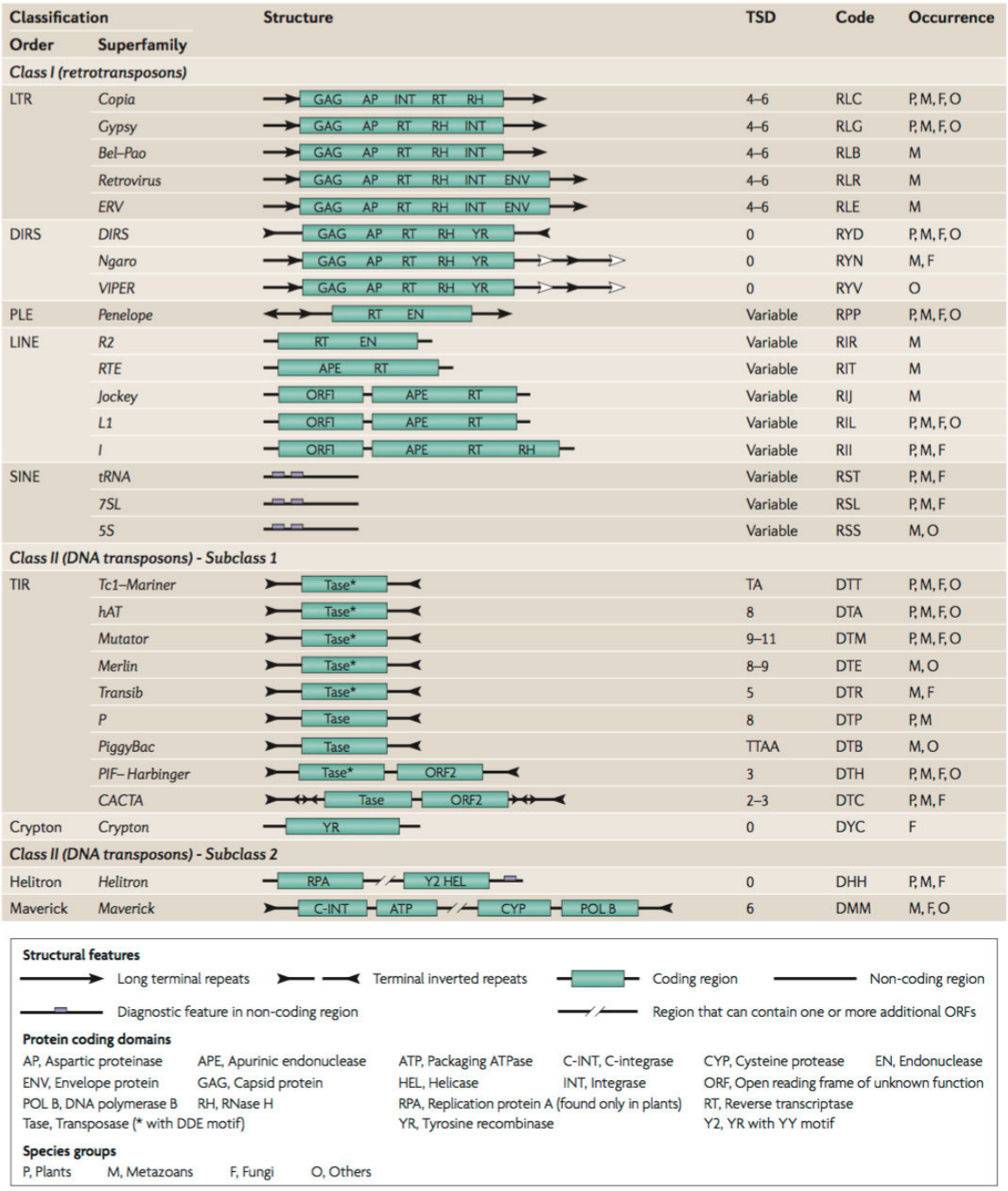


FIGURE 2.16 – Classification des éléments répétés selon Wicker et al. (2007). La classification est hiérarchique et classe les éléments répétés en premier lieu sur la base de la présence/absence d'intermédiaire de transposition sous forme d'ARN.Des sous-classes , des ordres et des familles sont proposées. La taille du site cible de duplication (TSD) qui est généralement spécifique de chaque superfamille peut être utilisé comme un critère de classement. Un code à trois lettres qui décrit chacun des groupes majeurs et ajouté au nom de la famille apparaît dans cette classification, : DIR (séquence répétée intermédiaire de Dyctiostelium), LINE (longs éléments nucléaires intercalés), LTR (séquene terminale longue répétée), PLE (éléments transposables de la famille Penelope), SINE (petit élément nucléaire intercalé)et TIR (répétition en tandem inversée)

des facteurs de transcription impliqués dans l'utilisation et l'homéostasie du fer *Aft1, Rcs1* et *RCbf1* chez *Saccharomyces cerevisiae* (Feschotte and Pritham, 2007)). Du fait de leurs répétitions au sein d'un génome et de leur capacité de se déplacer, les éléments transposables jouent un rôle important dans l'évolution des génomes (Castanera et al., 2016).

2.2.2 Évolution du génome

L'Évolution est communément définie à l'échelle de la population, par les changements de fréquences alléliques au sein de cette population sous l'effet des mutations, de la dérive génétique, des migrations et de la sélection (Zhang et al., 2018). Les empreintes que l'on peut retrouver au niveau d'un génome sont des mutations ponctuelles, divers réarrangements chromosomiques, des gains et des pertes de gènes qui participent à l'évolution du génome. Lorsqu'on définit le génome d'une espèce, on définit donc une entité génomique structurée par ces empreintes qui se retrouvent fixées au sein de la population.

Les mutations ponctuelles

Au cours de la réplication et la réparation de l'ADN, des erreurs peuvent apparaître par rapport à la séquence d'origine. Il peut s'agir de mutations ponctuelles comme des substitutions (SNP), des insertions ou des délétions (INDEL) d'un ou plusieurs nucléotides ou de mutations chromosomiques (Brown, 2002,chapitre 15). Dans le cadre de séquences transcrites puis traduites en protéines, on distingue les substitutions non synonymes (qui change l'acide aminé, dN) des substitutions synonymes (qui ne change pas l'acide aminé, dS). L'impact de ces mutations, qu'il soit négatif, positif ou neutre sur l'organisme, dépend du type de mutation et de l'endroit où la mutation apparaît. L'apparition d'un INDEL peut entraîner des modifications majeures en entraînant un changement de cadre de lecture de la protéine. L'accumulation de mutations peut conduire à la création d'un nouveau domaine fonctionnel ou à l'apparition d'un codon STOP prématuré. Si la mutation touche une promoteur du gène, l'expression du gène peut en être altéré.

Plus la contrainte de conservation de la protéine pour la survie de l'organisme sera importante plus le nombre de substitutions non synonymes sera faible car celles-ci seront éliminées à l'échelle de la population par sélection purificatrice. S'intéresser au ratio dN/dS permet d'avoir une information sur le degré de sélection auquel est soumis l'élément d'intérêt.

Les mutations chromosomiques

Lors de mutations chromosomiques, des fragments de chromosomes peuvent être dupliqués, perdus ou inversés (Figure 2.17). Ces évènements peuvent se dérouler sur un même chromosome ou en impliquer plusieurs. Lorsqu'un fragment de chromosome est déplacé sur un autre chromosome, on parle alors de translocation. Elle est qualifiée de simple si un chromosome perd un fragment au profit de l'autre et de réciproque si deux fragments chromosomiques sont échangés.

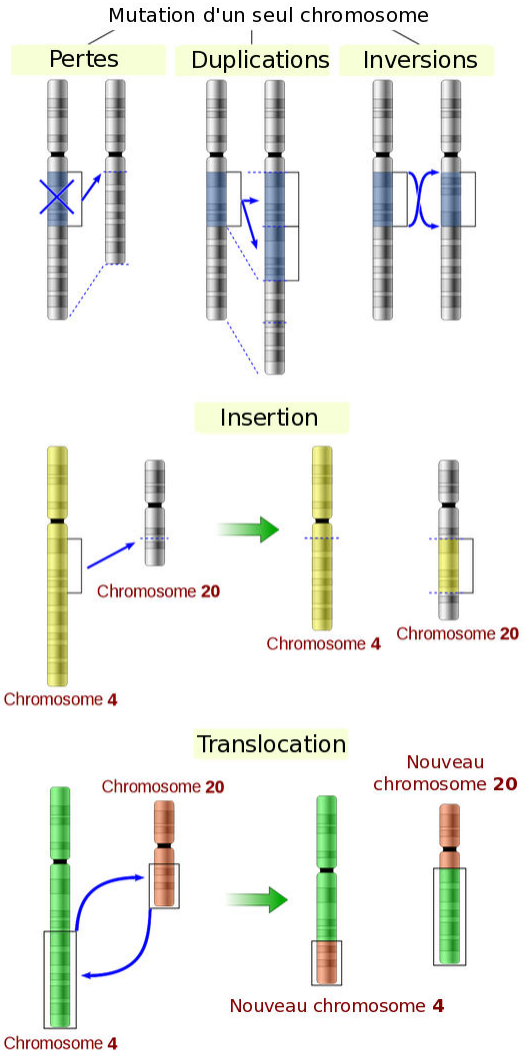


FIGURE 2.17 – Les différents types de mutations chromosomiques

Ces réarrangements chromosomiques sont facilités par la présence de TE. Du fait de leur présence en de multiple copies avec des séquences similaires, ils peuvent provoquer une recombinaison entre deux parties d'un même chromosome ou de chromosomes différents. Modifier la position d'un fragment génomique peut changer la régulation auquel il est soumis pour s'aligner avec la régulation du nouvel environnement génomique dans lequel il se trouve (état de compaction de la chromatine différent, séparation physique d'éventuels stimulateurs). Autre point d'intérêt, lors des réarrangements, des gènes peuvent être coupés en deux et/ou fusionnés avec d'autres gènes adjacents or la fusion de gènes est le principal mécanisme permettant l'acquisition de nouveaux domaines [9] chez les animaux (Marsh and Teichmann, 2010). A l'échelle d'un gène, des réarrangements peuvent avoir lieu. Cela peut être dû à des recombinaisons ou au déplacement de transposons (lors de leur transposition certains transposons peuvent transférer un court fragment d'ADN adjacent). Il existe une corrélation significative entre les bordures d'exons et les bordures de domaines, dupliquer un

9. Les domaines sont des régions protéiques, conservées au cours de l'évolution, ayant généralement des propriétés structurales et fonctionnelles qui leur sont propres (Marsh and Teichmann, 2010).

exon peut donc revenir à dupliquer un domaine fonctionnel, le domaine dupliqué peut alors évoluer créant une nouvelle fonction. A l'échelle du génome, un changement du nombre de chromosomes peut également arriver. On voit ainsi apparaître des cas d'aneuploïdie (1n et 3n).

Les gains et pertes de gènes

L'augmentation du nombre de gènes est majoritairement due à des processus de duplication de gènes présents dans le génome. Les gènes peuvent être dupliqués via (i) la duplication complète du génome (*whole genome duplication*, WGD), (ii) la duplication d'un simple chromosome ou d'un fragment de chromosome et (iii) la duplication d'un gène isolé ou d'un groupe de gènes. La copie du gène obtenue est alors identique au gène initial. Les deux gènes sont qualifiés d'homologues. Les groupes de gènes homologues sont qualifiés de familles de gènes ; ils ont généralement des fonctions similaires. Parmi les homologues, on distingue les orthologues des paralogues. Les orthologues sont des gènes présents dans des espèces différentes qui dérivent d'un même gène présent dans le dernier ancêtre commun des deux espèces. Les paralogues sont des gènes qui dérivent d'un même gène dupliqué au sein d'une même espèce. On distingue les in-paralogues des out-paralogues : les in-paralogues sont des paralogues issus d'une duplication génique qui est apparue après la spéciation tandis que des out-paralogues seront issus d'une duplication génique qui est apparue avant la spéciation (et n'appartiennent donc pas nécessairement à la même espèce) (Figure 2.18).

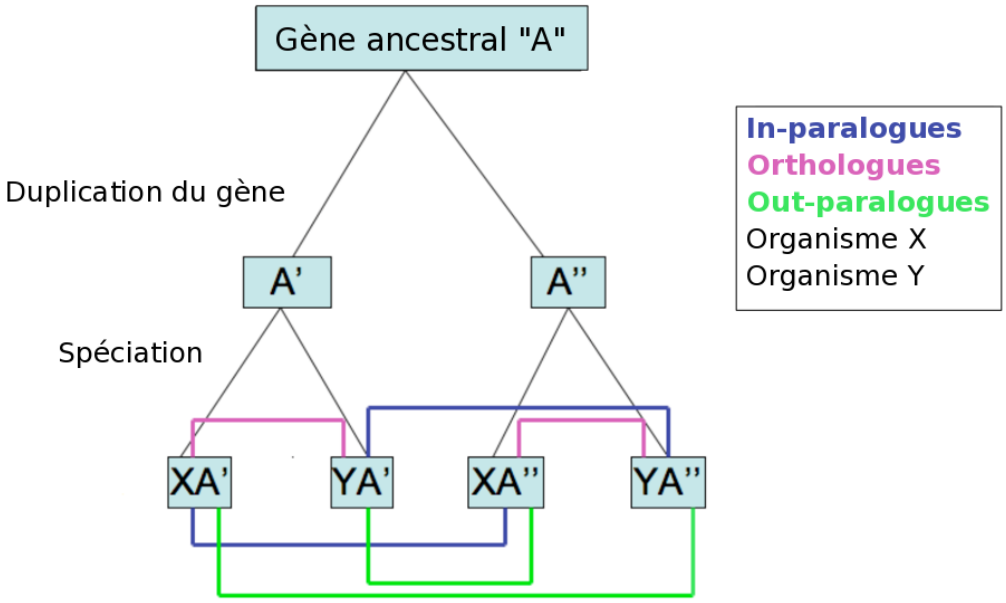


FIGURE 2.18 – Représentation de la relation entre paralogues et orthologues.

Après une duplication, la fonction du gène est assurée par l'une des copies, des modifications peuvent donc apparaître sur l'autre copie sans impacter l'organisme. Les copies peuvent diverger et remplir chacune une partie de la fonction ancestrale (subfonctionnalisation), cela peut être aussi bien une séparation de deux fonctions du gène ancestral qu'une activation régulée par des conditions distinctes. Une des copies peut acquérir une nouvelle fonction (néofonctionnalisation), être perdue suite à des réarrangements chromosomiques (délétion) ou dégénérescence mutationnelle (pseudogénisation) ou être conservée sans modifications majeures. Il existe des cas où plusieurs copies de gènes sont retrouvées identiques, cela s'explique par un mécanisme de conversion génique qui permet de répandre une mutation avantageuse d'une copie dans la famille de gène : on parle alors d'évolution concertée (Brown, 2002,chapitre 15). Des gènes peuvent également être acquis à partir d'autres organismes. Ces transferts dits horizontaux (HGT) peuvent concerner des gènes isolés comme des régions entières. De nouveaux gènes peuvent également être obtenus à partir des éléments déjà présents dans le génome, c'est le cas lors de la domestication de protéines d'éléments transposables ou, de façon exceptionnelle, l'apparition de gènes de novo à partir d'ADN non codant.

Les gènes peuvent être perdus par délétion ou perdre leur fonction initiale via un processus de pseudogénisation. Ces pseudogènes peuvent cependant assurer d'autres fonctions : de nombreux pseudogènes sont transcrits en ARN et certains d'entre eux jouent un rôle essentiel dans la régulation de leur gènes homologues, réduisent la concentration cellulaire de miRNA ou peuvent former des ARN interférents (Tutar, 2012). On distingue deux types de pseudogènes : les pseudogènes non fonctionnels (*unprocessed pseudogenes*) et les rétrogènes (*processed pseudogenes*). Les pseudogènes non fonctionnels sont issus d'une pseudogénisation enclenchée suite à une ou plusieurs mutations non synonymes, INDEL, codons STOP mal placés, changement de cadre de lecture, insertion d'un TE etc. dans un gène existant. Ce gène peut être une copie d'un gène existant (duplicated pseudogenes) ou être le seul de sa famille (unitary pseudogenes). Les rétrogènes correspondent à des pseudogènes qui semblent être des copies ADN d'ARNm ayant subi une maturation normale. Ces copies s'intègreraient au hasard dans le génome donnant des pseudogènes caractérisés par l'absence de promoteurs et d'introns ainsi que par la présence d'une extension d'acide polyadényliques (Raffaele and Kamoun, 2012). Un résumé des mécanismes qui modèlent la structure des génomes fongiques et leur impact sur l'adaptation au milieu est présenté en Figure 2.19.

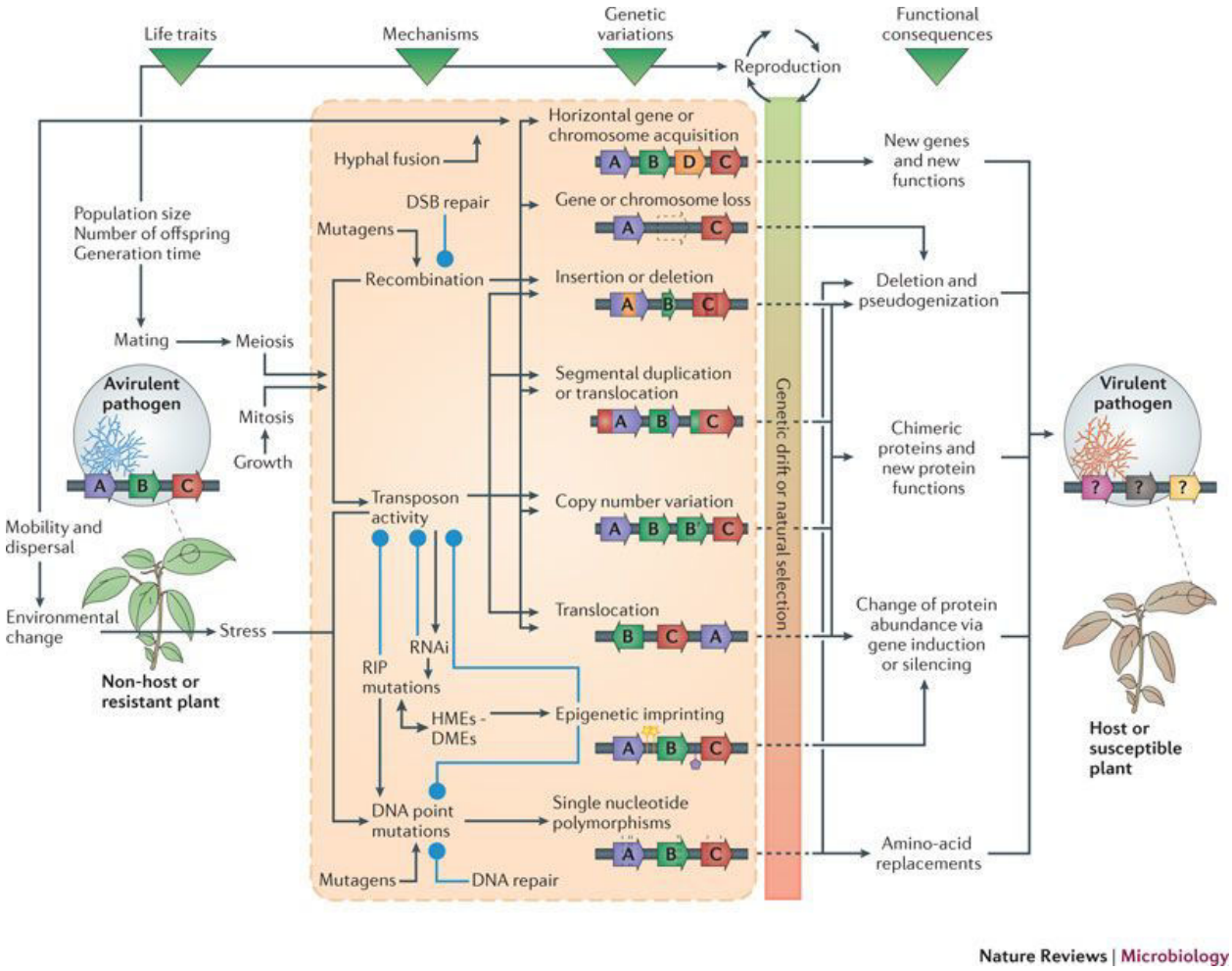


FIGURE 2.19 – Illustration de traits de vie et mécanismes qui modèlent la structure des génomes des champignons filamenteux pathogènes et leurs conséquences génétiques et fonctionnelles(Source Raffaele and Kamoun (2012)). Les relations de cause à effets sont représentées par des flèches noires tandis que les relations d'inhibition sont indiquées en bleu.Certaines relations ont été supprimées par soucis de clarté. La reproduction, la croissance et les stress induits par l'environnement font partie des processus menant aux variations génétiques. Elles sont influencées par des traits de vie tel que la taille de la population, le nombre de descendants, le temps de génération des nouveaux individus et leurs modes de dispersion. Ces processus induisent, plus ou moins directement, une palette de mécanismes qui modèlent la structure et l'expression du génome ce qui inclut également les points de mutations d'ADN, les mécanismes de défense tel que les *repeat induced point mutation* (RIP), les marques épigénétiques, les recombinaisons, les activités des transposons, les transferts horizontaux et transferts de chromosomes. Huit types de variations génétiques majeures sont déclenchés par ces mécanismes : (i) les acquisitions de gènes ou de chromosomes, (ii) les pertes de gènes ou chromosomes, (iii) les petites insertions et délétions, (iv) les remaniements de domaines par duplications de segments ou translocations, (v) les variations du nombre de copies des gènes, (vi) les translocations de gènes, (vii) les marques épigénétiques et (viii) le polymorphisme de nucléotides isolés. Les forces évolutives vont ensuite moduler la fréquence de ces variations dans la population par dérive génétique ou selection naturelle. Les conséquences fonctionnelles peuvent être classées comme : (i) l'acquisition de nouvelles fonctions par l'acquisition de nouveaux gènes, (ii) la perte de gènes et pseudogénisation, (iii) la formation de protéines chimériques menant à de nouvelles fonctions protéiques, (iv) le changement d'abondance protéique lié à des changements d'expressions, (v) le remplacement d'acides aminés dans les séquences protéiques. Dans de nombreux cas, ces variations augmentent l'adaptation du pathogène à son habitat par une augmentation de la virulence. DME, enzymes de méthylation de l'ADN ; DSB, rupture des deux brins d'ADN ; HME, enzymes de méthylation des histones ;RNAi, ARN interférent.

Ces mutations apparaissent et sont conservées à des fréquences différentes. On note par exemple que sur l'apparition d'environ 1000 mutations chez *Saccharomyces cerevisiae*, identifiées grâce au séquençage du génome de 145 lignées accumulant des mutations (Zhu et al., 2014), la majorité correspondent à des mutations ponctuelles mais les cas de courtes délétions et d'aneuploïdies ne sont pas rares (Figure 2.20).

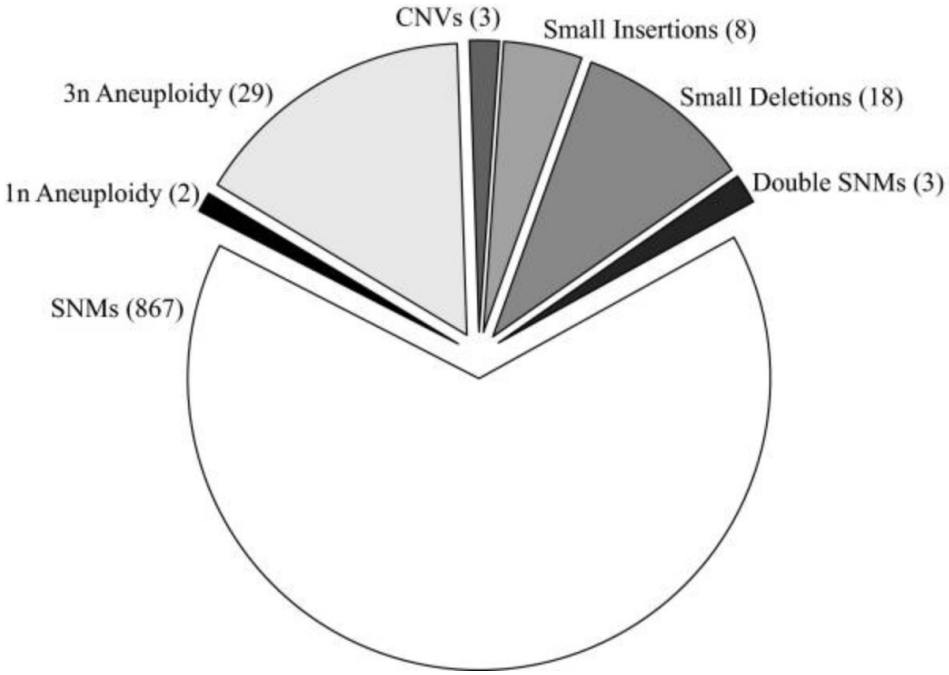


FIGURE 2.20 – Répartition des ±1000 mutations identifiées lors du séquençage de 145 lignées diploïdes accumulant des mutations de la levure *Saccharomyces cerevisiae*. SNMs=Différence de nucléotides isolés (*single nucleotide mutation*) ; CNV= différence du nombre de copies de gènes (*copy number variation*) (Source Zhu et al. (2014)).

2.2.3 Dynamisme des génomes fongiques

Évolution et adaptation (ensemble de modifications provoquées chez un organisme par un changement) sont deux notions distinctes mais étroitement liées. En Biologie Évolutive on utilise le mot adaptation pour parler d'un état nouveau, d'un caractère qui améliore la survie de l'individu et également celle de la population grâce à la reproduction et à la transmission des caractères héréditaires et l'évolution adaptative se définit comme une modification du *pool* génétique d'une population au cours du temps liée aux pressions de sélection exercées par son milieu (Deroy, 2015). C'est notamment cette divergence adaptative qui conduit à la formation d'une diversité d'espèces, aux habitats différents et niches écologiques variées.

Comme décrit ci-avant lors de la définition du mode de vie fongique, on trouve chez les champignons une grande diversité de niches écologiques avec notamment des saprobes aussi bien que des mutualistes obligatoires, symbiotes ou parasites. Il est intéressant de noter que plusieurs études ont montré qu'en fonction de la niche occupée les génomes portaient des empreintes de l'adaptation à cette niche. Les travaux de Kohler et al. (2015) ont montré que chez les champignons symbiotiques ectomycorhiziens les génomes renferment un lot réduit de gènes codant pour des enzymes dégradant les parois cellulaires végétales. Il faut noter que l'action de ces enzymes pourraient entrainer une réponse de défense de la plante (Lamb et al., 1989). De plus ces champignons ectomycorhiziens possèdent un grand nombre de gènes spécifiques et induits lors de la mycorhization (Kohler et al., 2015). Chez les champignons qui dégradent le bois, des familles de gènes impliquées dans la dégradation de la cellulose, hémicellulose et lignine ont été mises en évidence avec une specificité en fonction des substrats dégradés (Ohm et al., 2014). Chez les pathogènes végétaux, de nombreuses expansions de famille de gènes codant des protéines impliquées dans l'interaction avec l'hôte ont été mises en évidence (Raffaele and Kamoun, 2012). Très polymorphes ces familles de gènes livrent donc une plasticité importante et contribuent à l'émergence de nouveaux traits de virulence. Chez les pathogènes, les éléments transposables associés à des gènes de virulence jouent également un rôle important (Möller and Stukenbrock, 2017) tout comme l'existence de chromosomes accessoires (Bertazzoni et al., 2018) qui augmentent la plasticité des génomes et la création de nouveaux facteurs de virulence.

Les génomes portent également des empreintes liées à l'adptation à différents habitats. Par exemple, les champignons aux fructifications hypogées comme les truffes portent au sein de leur génomes des gènes impliquées dans la synthèse de composés volatiles pemettant indirectement leur dissémination par des animaux (Murat et al., 2018), les génomes d'espèces domestiquées se développant sur fromages comportent des régions transférées horizontalement leur conférant un avantage sélectif sur ce milieu fromager (Ropars et al., 2015).

Afin d'avoir une indication sur les processus impliqués dans l'évolution des génomes, ces derniers sont comparés entre eux. Lorsque les espèces sont suffisamment proches, on peut chercher à l'échelle génomique les différences responsables du phénotype. Avec une distribution de 2 Mb (*Encephalitozoon intestinalis*, *Microsporidia* (Pombert et al., 2013)) à 8Gb (*Entomophaga aulicae*, *Zoopagomycota*) et une médiane de 35Mb (sur 325 champignons éloignés phylogénétiquement (Stajich, 2017)) la taille des génomes fongiques est relativement réduite par rapport aux autres eucaryotes (485Mb chez le peuplier *Populus trichocarpa*, ±3 400 Mb chez l'Homme, ±17 000 Mb

chez le blé d'été *Triticum aestivum*). Ceci a grandement facilité leur séquençage ce qui a contribué à faire des champignons un modèle eucaryote. Ainsi, le premier eucaryote dont le génome a été séquencé fut un champignon : la levure ascomycète *Saccharomyces cerevisiae* (Goffeau et al., 1996). Très vite d'autres génomes fongiques sont venus s'ajouter à celui de la levure modèle : des filamenteux, ascomycètes comme *Neurospora crassa* (Galagan et al., 2003) ou basidiomycètes comme *Phanorechaete chrysosporium* (Martinez et al., 2004). En 2005, environ 40 génomes de champignons étaient disponibles (Galagan et al., 2005). Aujourd'hui grâce à des projets tels que le "*1000 fungal genomes*" plus de 1000 génomes fongiques sont publiquement disponibles. L'accessibilité d'obtention des génomes fongiques associée à cette base de données grandissante permet de réaliser des études de génomique comparative.

Malgré une apparente similarité morphologique et physiologique, les champignons peuvent être extrêmement différents à l'échelle des génomes. A titre d'exemple, trois espèces d'*Aspergillus* - *A. nidulans*, *A. fumigatus*, et *A. oryzae*- affichent une distance évolutive comparable à celle entre les humains et les poissons [10] (Galagan et al., 2005). Des changements majeurs tels que la modification de ploïdie et l'aneuploidie sont un moteur majeur d'adaptation de certains champignons pour s'adapter rapidement aux changements environnementaux (Wertheimer et al., 2016). C'est le cas de *Cryptococcus neoformans* normalement retrouvé haploïde (1N) ou diploïde (2N) qui change de ploïdie allant de 4N à > 64N lors de l'infection d'animaux (Todd et al., 2017). En cas de stress lié aux médicaments antifongiques, des cellules aneuploïdes présentant une expansion du nombre de chromosome 1 sont observées (Sionov et al., 2010). Chez *Candida albicans*, le manque de nutriments déstabilise les cellules tétraploïdes ce qui conduit au fil du temps à la perte de chromosomes chez les cellules filles jusqu'à arriver à un état proche de la diploïdie. Des expériences sur *S. cerevisiae* ont montré que la réduction de ploïdie (2N à haploïde) pouvait être réalisée dans un laps de temps relativement court (±50 générations sachant que le temps de doublement d'une population de S. cerevisiae peut être de 2h) (Todd et al., 2017).

La plasticité des génomes fongiques à l'échelle des chromosomes n'est pas limitée au changement de ploïdie. En effet, certaines espèces disposent de chromosomes accessoires. Chez certaines espèces comme les pathogènes de plantes *Fusarium oxysporum*, *Nectria haematococca*, ou encore *Leptosphaeria maculans* ces chromosomes accessoires participent à la virulence et à la spécificité de l'hôte. Chez le pathogène du blé *Zymoseptoria tritici*, 8 des 21 chromosomes sont accessoires (Habig et al., 2017) soit 12% de la taille du génome (pour l'isolat de référence IPO323).

10. Identité des acides aminés de 68% entre chacune des paires d'*Aspergillus sp.*

Ces chromosomes se distinguent des « core » chromosomes par une plus faible densité en gène, une plus grande proportion de gènes uniques et un enrichissement en éléments transposables.

Qu'il s'agisse de chromosomes accessoires ou de core chromosomes, un grand nombre de variations structurales sont visibles. Ces variations en terme d'évolution de structure sont particulièrement frappantes chez de nombreux pathogènes de végétaux pour lesquels le génome est compartimentalisé entre des régions denses en gènes et les régions appauvries en gènes mais enrichies en éléments transposables. Les régions enrichies en répétitions coïncident fréquemment avec les ruptures de synténie [11], confirmant une évolution plus rapide de ces régions par rapport au reste du génome. On y retrouve régulièrement les gènes impliqués dans la virulence et adaptations à l'hôte (Raffaele and Kamoun, 2012).

Les petites inversions, translocations, insertions, délétions et duplications sont courantes (Galagan et al., 2005; Zhu et al., 2014). Les duplications et translocations en particulier sont des réponses communes des levures en cours d'évolution (citepDunham2002, Koszul2004 et contribuent à l'adaptation sur le long terme des génomes fongiques. A titre d'exemple, la comparaison des génomes de *Magnaporthe grisea* et *N. crassa*, des ascomycètes issus d'un ancêtre commun relativement récent (200 million d'années), a permis de mettre évidence une absence de synténie et une identité des acides aminés d'environ 47% (en moyenne) entre protéines homologues (Dean et al., 2005).

Comme ce qui est couramment retrouvé chez les autres eucaryotes, les réarrangements sont plus fréquents près des télomères et sont souvent associés à des éléments répétés (Huynen et al., 2001; Carlton et al., 2002; Coghlan and Wolfe, 2002; Kellis et al., 2003; Galagan et al., 2005; Lephart et al., 2005). Les réarrangements étant plus fréquents en régions subtélomériques, ces régions sont souvent plus riches en gènes évoluant rapidement. Ceux ci permettent d'avoir des indications sur les évolutions récentes de l'organisme et notamment des spécialisations écologiques qui sont apparues (Cuomo and Birren, 2010).

On retrouve également des marques d'évènements de duplication complète du génome (whole genome duplication ou WGD) chez la levure suivis de pertes de gènes massives. Les expansions et les contractions de familles de gènes peuvent permettre une meilleure adaptation au milieu. Cependant une augmentation trop importante du nombre de copies d'un gène peut entraîner l'activation des mécanismes de défense. A titre d'exemple, certaines espèces fongiques disposent d'un mécanisme appelé "*repeat induced point mutation*" (RIP) par lequel les gènes

11. Conservation de l'ordre des éléments génomiques entre deux espèces apparentées.

dupliqués sont fortement mutés durant la méiose. Chez *Neurospora crassa* chez qui ce mécanisme est particulièrement fort, peu de gènes sont dupliqués (Gladyshev, 2017).

Bien que moins fréquents chez les champignons que chez les bactéries, les HGT peuvent avoir un impact très important pour l'adaptation d'une espèce à son milieu. L'acquisition par *Pyrenophora triticirepentis* du gène de *Stagonospora nodorum* permettant la production d'une toxine (ToxA) a permis aux *Pyrenophora* de se développer sur le blé (Friesen et al., 2006). Chez *S. cerevisiae*, les gènes issus de HGT [ab]ont contribué à d'importantes innovations tels que la capacité de synthétiser de la biotine, de pouvoir se développer en conditions anaérobie (Fitzpatrick, 2012) ou encore d'utiliser le sulfate provenant de diverses sources organiques (Hall et al., 2005). Le transfert d'un cluster complet permettant l'assimilation du nitrate d'un Basidiomycète a permis à l'ascomycète *Trichoderma reesei* de s'affranchir d'un style de vie essentiellement mycoparasitaire (Slot and Hibbett, 2007). Il existe également plusieurs travaux montrant l'impact de transfert horizontaux chez les champignons domestiqués et utilisés en agroalimentaire. Par exemple, les génomes de *S. cerevisiae* comprennent plusieurs signatures de transfert horizontaux issus de bactéries ou d'autres levures (Hall et al., 2005; Hall and Dietrich, 2007; Wei et al., 2007). Et dans le cas des souches utilisées pour la vinification, ces régions transférées intègrent plusieurs gènes d'intérêt pour l'élaboration des vins (Novo et al., 2009). Des HGT sont également présents dans les génomes des *Penicillium* et plusieurs espèces technologiques de *Penicillium* possèdent des régions transférées incluant des gènes livrant un avantage sélectif à ces espèces sur fromage (Cheeseman et al., 2014; Ropars et al., 2015). Les transferts horizontaux peuvent donc conduire à un changement de niche écologique ou une meilleure adaptation au milieu dans lesquels ils se trouvent.

Encadré 5 : Exemples d'empreintes génomiques et mécanismes associés à une adaptation.

TABLEAU 2.2 – Exemples d'empreintes génomiques et mécanismes associés permettant une adaptation de champignons au milieu fromager.

Rôle	Empreinte génomiqe	Mécanisme associé	taxons	Référence
Utilisation du substrat	HGT de la région *CheesyTer* contenant 37 gènes dont lactose permease, beta-galactosidase	catabolisme du lactose optimisé entre autres	*Penicillium* spp.	Ropars et al., 2015
Utilisation du substrat	Présence de regions spécifiques (T, X, Z) et duplications	métbolisme du galactose, flocculation optimisés	*Saccharomyces cerevisiae*	Legras et al., 2018
Utilisation du substrat	HGT d'un cluster intégrant des gènes *GAL(*régulateurs GAL4 et GAL80)	métabolisme du galactose optimisé	*S. cerevisiae*	Legras et al., 2018
Utilisation du substrat	Perte de gène	métabolisme du galactose (*GAL1, GAL7* et *GAL10*) optimisé	*S. cerevisiae*	Legras et al., 2018
Utilisation du substrat	Variation du nombre de copies de gènes	Transport de glucides, synthèse de thiamine et métabolisme du galactose optimisés	*S. cerevisiae*	Legras et al., 2018
Tolérance à des basses températures	Capacité	lipolyse et protéolyse augmentées	levures	Corbo et al., 2001
Tolérance à des pH acides	Activité ou meilleure efficacité ou stabilité	Dégradation de l'ATP, transport de protons	levures	Praphailong et Fleet 1997
Compétitivité	HGT de la région *Wallaby* (~500kb)	Régulation de la conidiation et activités antimicrobiennes	*Penicillium*	Ropars et al., 2015
?	Pertes de segments de gènes	?	*S. cerevisiae*	Legras et al., 2018
?	Pseudogeneisation	?	*S. cerevisiae*	Legras et al 2018

2.3 Intérêt et méthodes de séquençage

Les champignons possèdent une grande faculté d'adaptation. Ils peuvent s'adapter à tous types d'environnement. Un exemple parmi les plus frappant est possibilité de trouver deux membres de la même espèce de champignon pouvant vivre sur la terre et dans le monde marin (van de Veerdonk et al., 2017; Redou et al., 2015). L'aspect phénotypique ne renseignant en rien sur la capacité des champignons à croître dans des environnements différents, il est nécessaire de trouver d'autres techniques permettant de mieux comprendre l'adaptation d'une espèce à son environnement.

Un des éléments des êtres vivants qui permet de comprendre les différents modes de vie de ces organismes est l'ARN. Au centre des molécules du vivant, l'ARN permet de savoir quel gène

est exprimé à quel moment et surtout sur quel milieu. En effet, le génome porte l'information génétique mais ce sont les ARNs (codants, non codants, petits, cycliques, ribosomiques, etc.) qui servent de modèles à la synthèse des protéines. Cependant, l'étude de l'ensemble des ARNs (transcriptome) à un instant précis, ne permet d'obtenir qu'une information localisée à un temps et un ensemble de conditions donné. Certains gènes, qui potentiellement pourraient répondre à une adaptation à un environnement ou servir de biomarqueur ne sont de fait pas, ou trop faiblement, exprimés dans certains temps et/ou conditions données. Par exemple, dans les conditions standard de laboratoire, certains clusters de gènes fongiques ne sont pas exprimés car le promoteur est très faiblement ou non activé (Brakhage and Schroeckh, 2011). Nous passons donc à côté d'informations essentielles. C'est pour cela que des études génomiques en parallèle semblent nécessaires. Dans le cadre d'études génomiques d'organismes pour lesquelles les génomes ne sont pas déjà disponibles, les transcriptomes sont d'une importance cruciale pour une bonne prédiction de la structure des gènes. Les traitements analytiques de ces deux types d'études, génomique et transcriptomique, sera donc développées par la suite.

Avant de commencer un projet de séquençage, il convient de réaliser un plan d'expérience. Ce plan doit prendre en compte la question biologique posée, les particularités de l'organisme, les informations déjà disponibles, le type de matériel biologique à disposition, le type de séquençage à réaliser et les contraintes des analyses subséquentes.

2.3.1 Extraction du matériel biologique

Etant le matériel initial, le choix de la méthodologie et la qualité de l'extraction du matériel biologique (ADN ou ARN) aura un impact important sur les analyses subséquentes.

L'ADN

Les points saillant lors de l'extraction de l'ADN pour un séquençage, sont la pureté de l'extrait, la quantité de matériel biologique extraite, et l'intégrité des molécules extraites. Ces points seront détaillés ci-après.

Plus l'ADN sera fragmenté plus la reconstruction (assemblage) du génome sera complexe à réaliser. L'intégrité structurelle de l'ADN est particulièrement importante pour un séquençage impliquant de longues séquences (*mate pair* d'Illumina, PacBio, Oxford Nanopore). L'ADN

initial doit être aussi pur et intègre que possible, la présence de composés tels que les polysaccharides, les protéoglycanes, les protéines ou les métabolites secondaires, extraits avec l'ADN peut conduire à une baisse d'efficacité des séquenceurs et donc une couverture plus faible qu'attendue (Dominguez Del Angel et al., 2018).

Le séquençage de multiples individus augmente la variabilité génétique de l'extraction et conduit à un assemblage plus fragmenté. Si un ensemble d'individus est nécessaire, il conviendra de les choisir tels que leurs patrimoines génétiques soient le plus proches possible.

Les contaminations biologiques peuvent également poser problème. Les contaminations peuvent être introduites lors de l'extraction de l'ADN ou être liées à l'organisme/tissu utilisé pour extraire l'ADN (contaminant/symbiote). Ces contaminations peuvent, dans une certaine mesure, être partiellement écartées a posteriori. Cependant l'effort de séquençage demandé sera réparti entre l'organisme d'intérêt et la contamination. Ainsi, la profondeur de séquençage de l'organisme d'intérêt sera donc plus faible qu'attendue.

Certains tissus sont extrêmement riches en mitochondries et/ou chloroplastes. L'ADN de l'organelle peut alors apparaître en grande concentration par rapport à l'ADN nucléaire ce qui implique, là encore, une plus faible profondeur de séquençage de l'ADN nucléaire.

L'ARN

Les points saillant lors de l'extraction de l'ARN pour un séquençage sont les mêmes que pour l'ADN, l'intégrité structurale étant un peu moins critique du fait de la taille restreinte des transcrits par rapport à la séquence génomique.

Dans le cadre de données d'expression, d'autres éléments entrent en compte pour le séquençage. Le transcriptome est principalement composé de gènes non codants (~95% des transcrits), si seuls les ARNm sont d'intérêt pour l'étude une sélection préliminaire des ARN possédant une queue polyA lors de la création de la librairie permet d'optimiser la proportion de données informatives pour l'étude. On note que des ARNm peuvent être perdus lors de ce type de préparation si la queue polyA des transcrits est dégradée ou dans le cas d'autre types de protection de l'ARN en 3' (queues poly T des ARN de certains organites par exemple).

L'expression des gènes n'est pas uniforme : des gènes sont très faiblement exprimés, les rendant difficile à détecter sans un effort de séquençage important, tandis que d'autres sont très fortement exprimés et représentent une portion importante des ARN en présence. Parmi ces

dernier, on compte notamment les ARN ribosomiques, si un séquençage des ARN totaux est souhaité (en opposition à la sélection d'ARN avec polyA par exemple), il est donc nécessaire d'écarter ces transcrits (déplétion des ARNr) pour éviter que la majeure partie du séquençage soit réalisée sur ces ARNr.

2.3.2 Séquençage

Évolution des coûts de séquençage

Cette étape est devenue de plus en plus accessible au cours du temps du fait de la réduction drastique du temps et des coûts de séquençage (Figure 2.21). A titre d'exemple, entre janvier 2015 et avril 2016 le coût du séquençage d'un génome Humain (~3Gb) a baissé de 4200$ US à 1200$ US (www.genome.gov/sequencingcostsdata). En 2018, ce séquençage pouvait être réalisé en moins d'une journée avec un coût d'environ 1000$ (400 euros au meilleur tarif remisé le 25/09/2018 https ://www.dantelabs.com/).

Dans le cadre d'analyses basées sur des données d'expression, le coût global reste important. En effet, afin de pouvoir exploiter statistiquement les analyses, un minimum de trois réplicats biologiques devront être réalisés (20 vous dira un staticien) pour chaque condition testée. Dans le cas d'une analyse portant sur deux espèces dans deux conditions données, le minimum est donc de douze séquençages.

Les points principaux influençant le prix du séquençage, correspondent au type de séquençage demandé (e.g. un séquençage de *reads* long sera bien plus cher qu'un séquençage de *reads* courts) et à la profondeur de séquençage.

La baisse de coût globale, l'augmentation de la vitesse de séquençage, et pour certaines technologies et l'amélioration de la qualité du séquençage s'expliquent par une évolution extrêmement rapide des technologies.

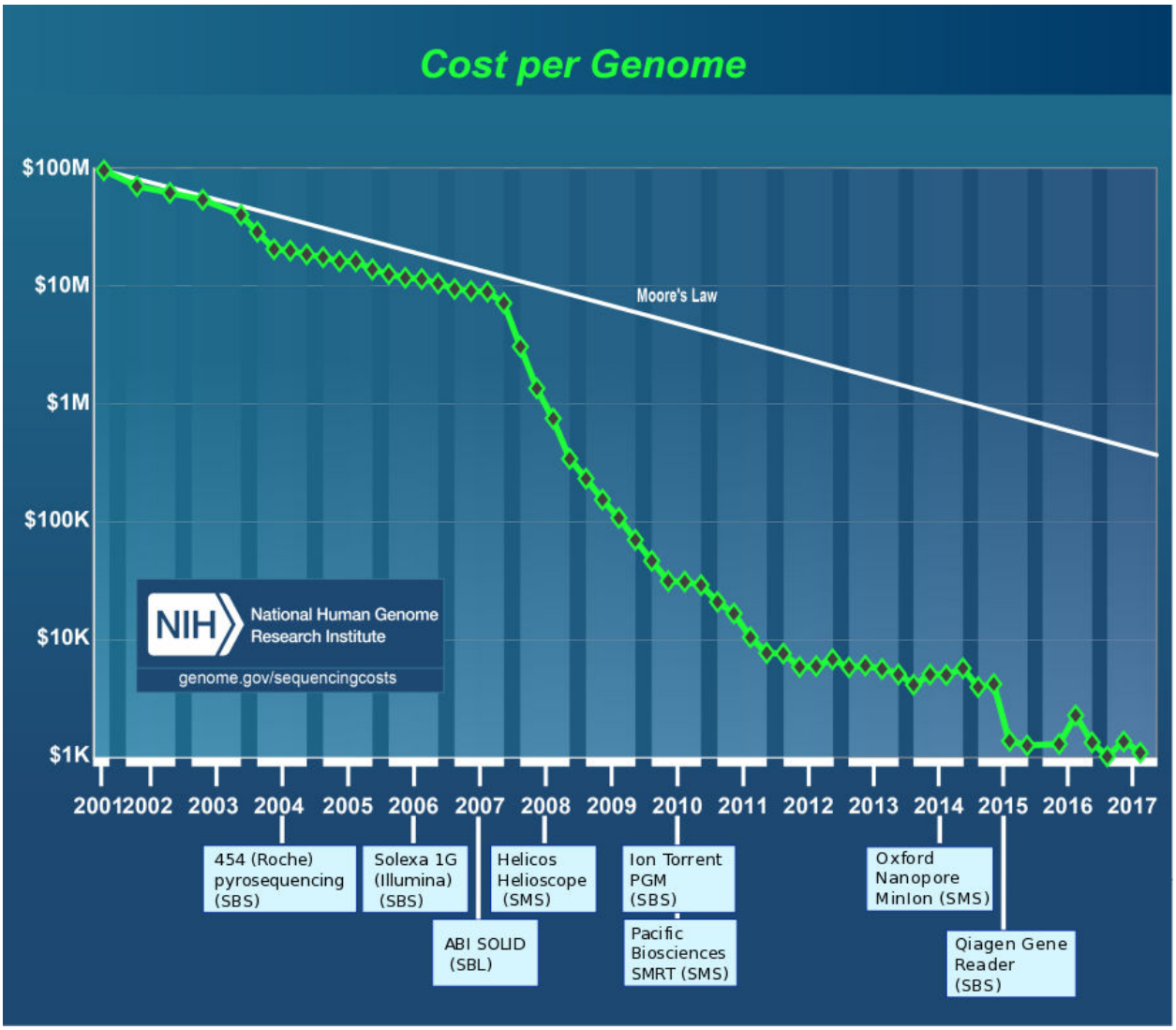


FIGURE 2.21 – Évolution du coût de séquençage du génome humain. Données de septembre 2018 provenant du NHGRI Genome Sequencing Program (GSP) accessible sur : `www.genome.gov/sequencingcostsdata`. Date d'introduction commerciale des différents séquenceurs (NGS) . SBS : séquençage par synthèse. SMS : séquençage de molécules uniques. SBL : séquençage par ligation. Figure adaptée de `www.genome.gov/sequencingcostsdata` et Mardis (2017).

2.3.3 Les technologies de séquençage

Quelle que soit la méthode, le produit de séquençage d'un génome consiste en un ensemble de lectures (*reads*) qui, tel un puzzle géant, doit être reconstruit pour accéder à l'information génomique (représentation schématique de cette problématique en figure 2.22).

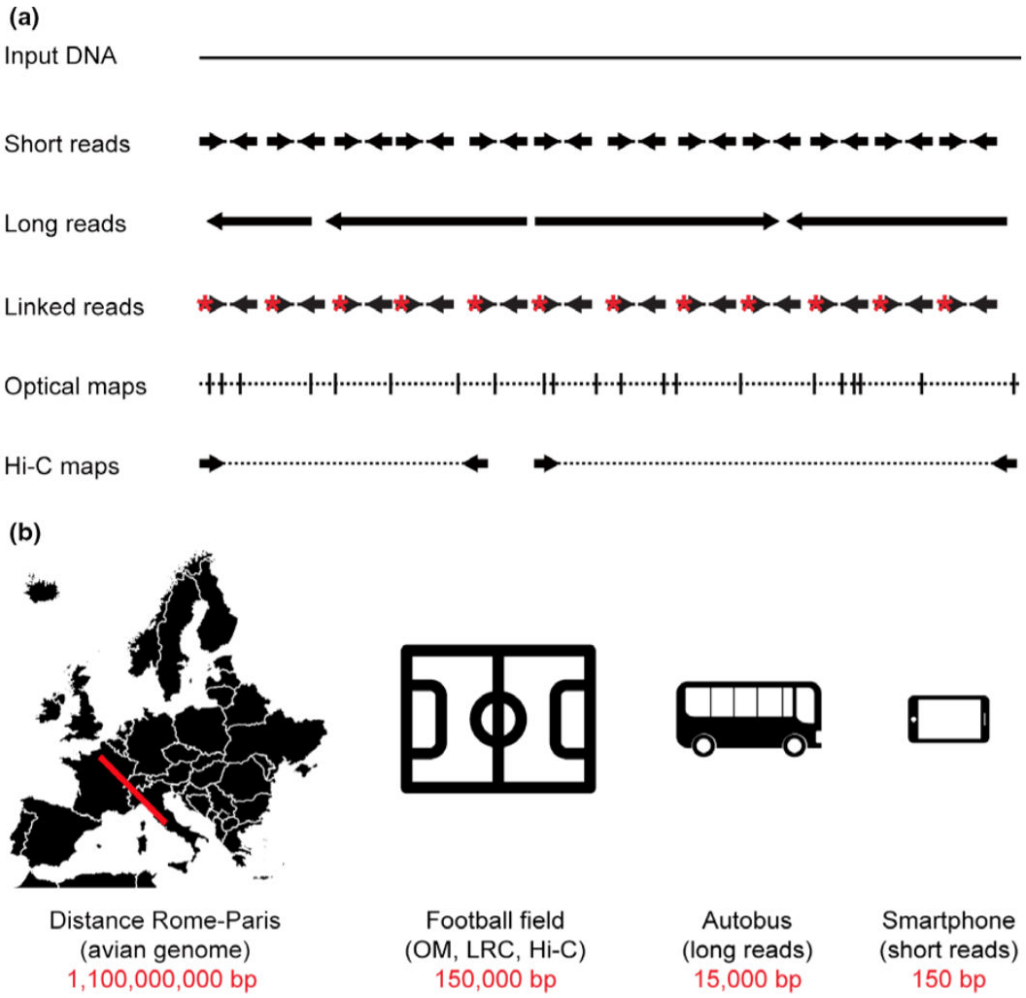


FIGURE 2.22 – Représentation schématique de la problématique d'assemblage. (a) Représentation du résultats des différentes technologies de séquençage disponible sur une séquence d'ADN théorique. Les *reads* courts sont pairés, les *reads* long sont indépendants les uns des autres, les *reads* liées (*linked reads cloud*, LRC) sont des *reads* courts avec un marqueur unique(astérisque rouge) pour chacune des molécules fournies. Les cartes optiques (*Optical maps*, OM) contiennent des distances physiques entre courts motifs de séquences et les cartes de Hi-C (*High Chromosome Contact map*) sont des *reads* courts représentant des interactions obtenues au travers de la conformation de la chromatine. (b) Représentation schématique de la taille des données obtenues à partir de ces différents séquençage (Source : Peona et al. (2018)).

Encadré 6 : Notions de séquençages *mate pair* et *paired end*

Au cours de ce document sera abordé des séquençages qualifiés de "*single end*", "*paired end*" et "*mate pair*". Lors du séquençage de données en *single end*, les fragments d'ADN sont uniquement séquencés à partir d'un côté. Les *reads* obtenus sont donc tous indépendants les uns des autres. Le terme *mate-pair* fait référence à une façon de préparer la banque génomique (protocole de circularisation de l'ADN) (Caboche and Even, 2012). Le but est de séquencer deux fragments d'ADN, c'est-à-dire deux *reads*, séparés par une distance fixe. Pour ce faire, un fragment d'ADN est circularisé, ainsi les deux extrémités deviennent adjacentes et pourront être séquencées (Figure 2.23 gauche). Ce type de banque génomique génère une paire de *reads* éloignés de plusieurs kb (de 2 à 20kb). Le terme *paired end* correspond quant à lui à un type de séquençage qui consiste à séquencer les *reads* par paires, ces *reads* étant séparés par une distance connue (taille du fragment - taille des *reads*) (Figure 2.23 droite). Ce type de séquençage génère une paire de *reads* éloignés de centaines de bp au maximum (< 600bp).

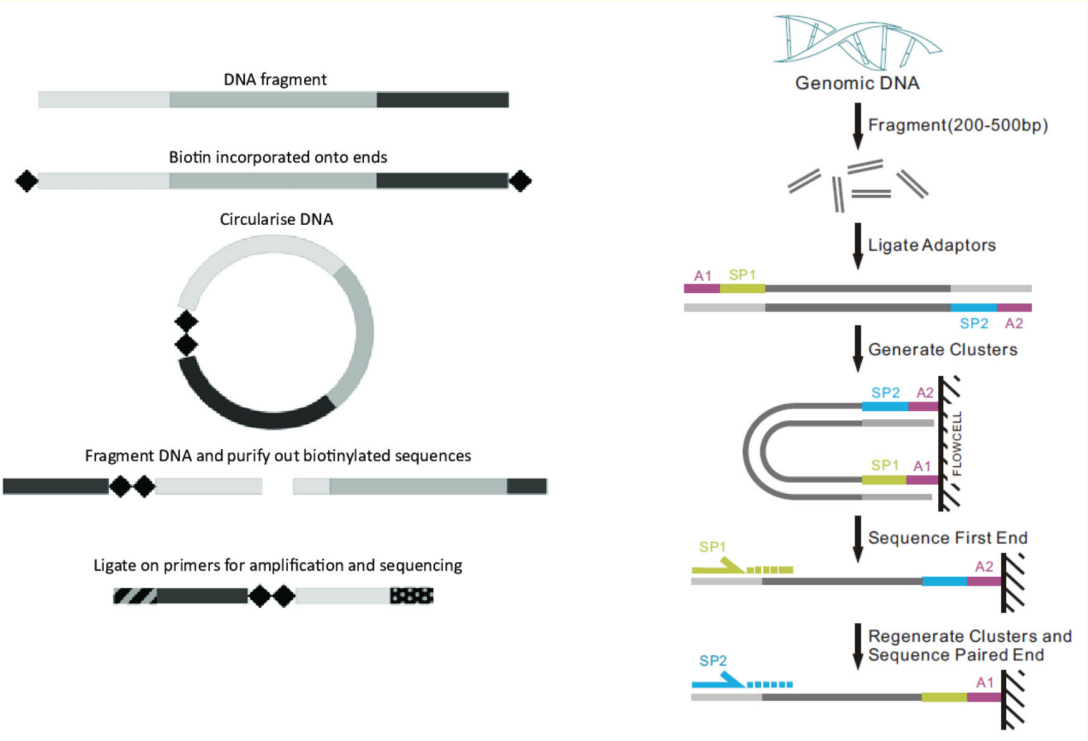


FIGURE 2.23 – Réalisation d'une librairie mate pair à gauche (source (Schlebusch and Illing, 2012)) et séquençage *paired end* à droite avec la technologie d'Illumina (source `illumina.com`). Dans le cadre d'un séquençage dit "*mate pair*" d'Illumina, un séquençage *paired end* est réalisé sur une librairie *mate pair* de fragments d'ADN.

L'orientation des *reads* en *mate pair* et *paired end* est différente et varie selon la technologie de séquencage utilisée. Par exemple avec la technologie d'Illumina, l'orientation est reverse-forward en *mate pair* et forward-reverse en *paired end*.

Lors du séquençage deux stratégies principales coexistent : l'une consiste à séquencer un très grand nombre de petits fragments (c'est le cas des technologies d'Illumina), l'autre privilégie la longueur des séquences au détriment de leur nombre (émergence de technologies dites de troisième générations comme celles de Pacific Biosciences ou d'Oxford Nanopore) (Mardis, 2017) (Figure 2.24). De multiples revues sont disponibles pour expliquer les avantages des différentes techniques comme les articles de Pillai et al. (2017); Heather and Chain (2016) ou encore Mardis (2017)

Company	Read length	Applications	Website
454/Roche	400 bp (single end)	Bacterial and viral genomes, multiplex-PCR products, validation of point mutations, targeted somatic-mutation detection	http://www.454.com/
Illumina	150–300 bp (paired end)	Complex genomes (human, mouse and plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics, noninvasive prenatal testing	http://www.illumina.com/
ABI SOLiD	75 bp (single end) or 50 bp (paired end)	Complex genomes (human, mouse, plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html/
Pacific Biosciences	Up to 40 kb (single end or circular consensus)	Complex genomes (human, mouse and plants), microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection	http://www.pacb.com/
Ion Torrent	200–400 bp (single end)	Multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing.html/
Oxford Nanopore	Variable: depends on library preparation (1D or 2D reads)	Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes	http://nanoporetech.com/
Qiagen GeneReader	107 bp (single end)	Targeted mutation detection, liquid biopsy in cancer	http://www.genereaderngs.com/

FIGURE 2.24 – Comparaison des plateformes NGS disponibles. (Source : Mardis (2017)).

Quelle que soit la stratégie utilisée, des biais associés à la technique transparaissent qu'ils s'agissent d'erreurs de séquençage aléatoires ou récurrentes : les technologies d'Illumina tendent à générer des erreurs provoquées par des répétitions inversées et des séquences GGC (Nakamura et al., 2011; Schirmer et al., 2015). Les technologies de 454/Roche et Oxford Nanopore tendent à générer des insertions ou délétions d'une ou plusieurs bases en rencontrant des séquences homopolymériques (Schirmer et al., 2015; Tyler et al., 2018; Boza et al., 2017).

Dans le cadre d'analyses d'expression, il peut être important de conserver l'information de l'orientation des *reads* lors du séquençage (à titre d'exemple, en cas de mapping sur génome,

ce séquençage permettra de déterminer sur quel brin d'ADN provient le *read*). Ce type de séquençage est qualifié de "*strand specific*".

2.4 Analyses bioinformatiques

2.4.1 Vérification de la qualité

Une fois le séquençage réalisé, il est nécessaire d'en vérifier la qualité. Cette qualité est liée d'une part aux séquences elles mêmes (bases ambigües, score Phred [12], longueur), d'autre part à leur diversité (sur-représentation de K-mers) et à la présence de contaminations potentielles.

L'outil FastQC (Andrews, 2010) est le plus souvent utilisé par la communauté de bioinformatique pour avoir une vue d'ensemble des données et d'en estimer la qualité (d'autres outils existent tels que HTSeq-qa (Anders et al., 2015) ou bien Kraken (Davis et al., 2013) pour remplir des fonctions similaires).

Selon les résultats, une préparation des données peut être nécessaire, qu'il s'agisse de supprimer les bases au score Phred jugé trop faible, des séquences trop courtes, des contaminations liées à la technique (adaptateurs utilisés lors du séquençage) ou d'un problème d'isolement de l'organisme à séquencer. Différents outils sont disponibles pour préparer les données, certains sont spécialisés (SortMeRNA (Kopylova et al., 2012)) et ont pour objectif de séparer les *reads* d'origine ribosomique des autres *reads*. D'autres sont plus génériques et peuvent remplir plusieurs tâches (Trimmomatic (Bolger et al., 2014), cutadapt (Saeidipour and Bakhshi, 2013) permettent de filtrer les adaptateurs, les *reads* au score Phred et taille de séquence trop faible).

2.4.2 Assemblages

Une fois les *reads* obtenus, une étape dite d'assemblage a pour but de reconstituer la séquence d'origine. Ces assemblages peuvent être réalisés avec les données de séquençage seuls (assemblage *de novo*) ou avec l'aide d'un génome de référence (*genome guided*).

12. Un score Phred est assigné à chaque base. Ce score de qualité correspond à -10*log10(probabilité de mauvaise identification de la base). Ainsi, un score Phred de 10 indique une probabilité de 1/10 de mauvaise identification de la base ou d'une probabilité de 90% qu'elle soit correcte, un score Phred de 20 aura probabilité de 1/100 de mauvaise identification de la base ou d'une probabilité de 99% qu ?elle soit correcte.

Assemblages génomiques

Dans le cas d'un génome, l'objectif est généralement de reconstruire les séquences assemblées les plus longues possibles (l'assemblage le moins fragmenté) avec la plus faible proportion d'erreurs d'assemblage. Les séquences assemblées constituées de *reads* chevauchants sont qualifiées de contigs. Des informations peuvent être obtenues sur le positionnement de ces contigs les uns par rapport aux autres sur la séquence d'origine grâce à des séquençages de type *mate pair* par exemple. Ces séquences assemblées et constituées de contigs orientés les uns par rapport aux autres sont qualifiées de scaffolds (Figure 2.25).

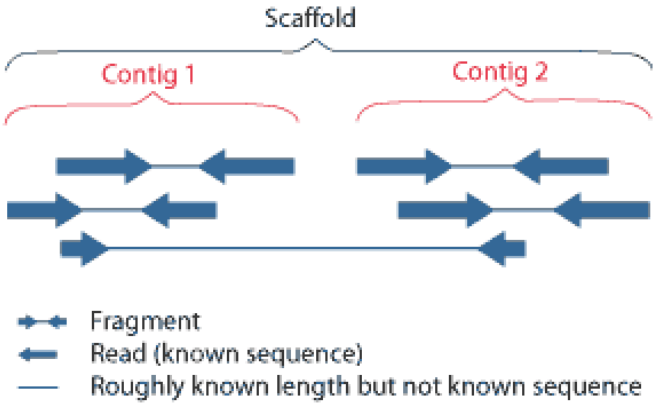


FIGURE 2.25 – Notions de contigs et scaffolds.

Parmi les sources de fragmentation des assemblages, on compte différentes causes qu'elles soient liées à la préparation de l'ADN (présenté plus tôt) ou intrinsèques au génome (Dominguez Del Angel et al., 2018).

En effet, Des taux de répétitions importants dans le génome, poseront des problèmes pour les assembleurs. En effet, comme l'outil ne peut déterminer l'assemblage correct de cette région, il arrête simplement d'étendre le contig à leur bord (Chaisson et al., 2015) (Figure 2.26).

L'assemblage final correspond à une reconstruction d'un génome haploïde. Lorsque l'on séquence un individu hétérozygote, les séquences des allèles homologues peuvent être trop différentes pour être assemblées ensemble. Cela peut mener à une augmentation artificielle du nombre de gènes en présence et une fragmentation accrue du génome (Pryszcz and Gabaldon, 2016). Les mêmes problèmes apparaissent dans les cas de polyploïdies où des séquences très proches risquent d'être fusionnées.

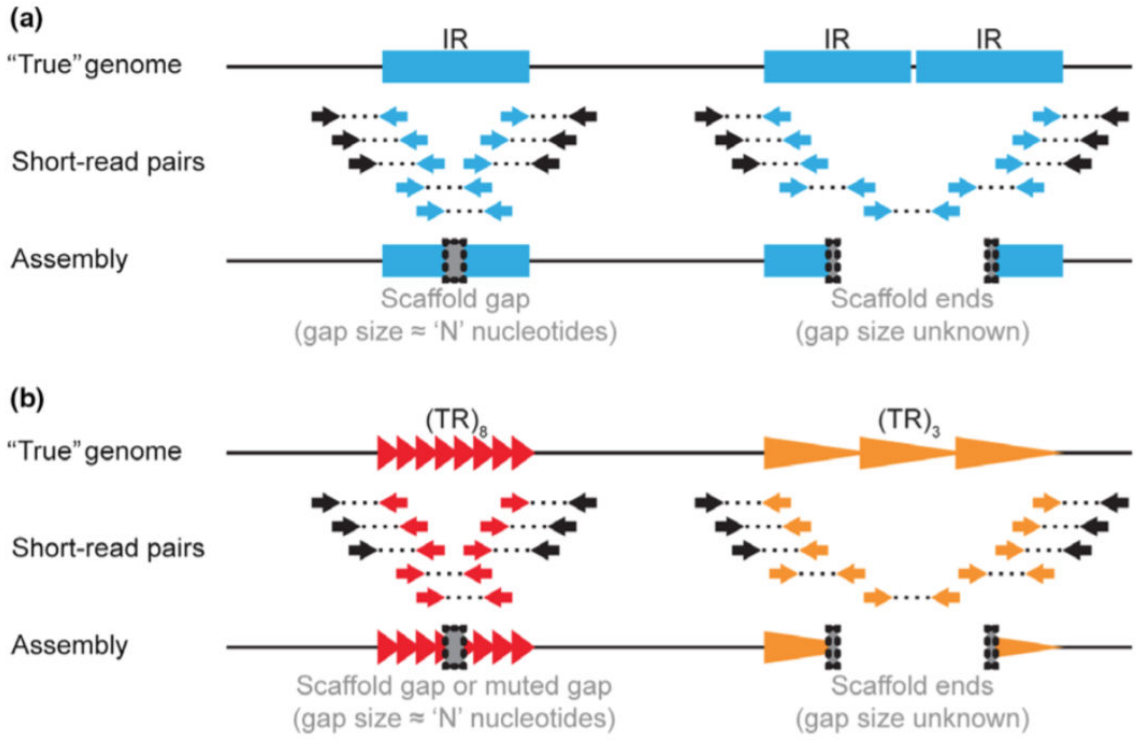


FIGURE 2.26 – Illustration de raisons pour lesquelles les éléments répétés peuvent conduire à des assemblages erronés (source : Peona et al. (2018)). IR : éléments transposables ou virus endogènes. TRs : microsatellites en rouge à gauche ou satellite en orange à droite.

Des taux extrêmement faibles ou hauts en GC dans une région génomique peuvent causer des problèmes lors de l'analyse de séquençages de type Illumina ce qui génère une couverture faible voire absente de ces régions (Chen et al., 2013)

Un assemblage *de novo* nécessite un effort de séquençage conséquent : une profondeur de séquençage adaptée et des fragments longs et de bonne qualité sont nécessaires. Un assemblage avec génome de référence peut être réalisé avec un séquençage moins onéreux mais nécessite de disposer d'un assemblage de référence dont la séquence est proche de celle étudiée. Les régions très différentes de la séquence de référence et dupliquées peuvent poser problème. Des pipelines utilisant les deux méthodes existent (Lischer and Shimizu, 2017).

Pour les assemblages *de novo* de séquences courtes les algorithmes les plus utilisés sont basés sur des graphes de *de Bruijn* (ex : Velvet (Zerbino and Birney, 2008) ou SPADES (Bankevich et al., 2012) mais d'autres algorithmes comme ceux Overlap Layout Consensus (OLC) sont également utilisés (ex : Celera (Cherukuri and Janga, 2016). Les graphes de *de Bruijn* sont basés sur une approche par K-mer : les *reads* sont fragmentés en séquences plus courtes (appelés K-mers, avec "K" représentant ici le nombre de base), l'assembleur réalise des recherches de chevauchement

de ces K-mer par bloc de K-1 pb ce qui permet d'accéder au chevauchement avec le K-mer suivant (Khan et al., 2018) (Figure 2.27).

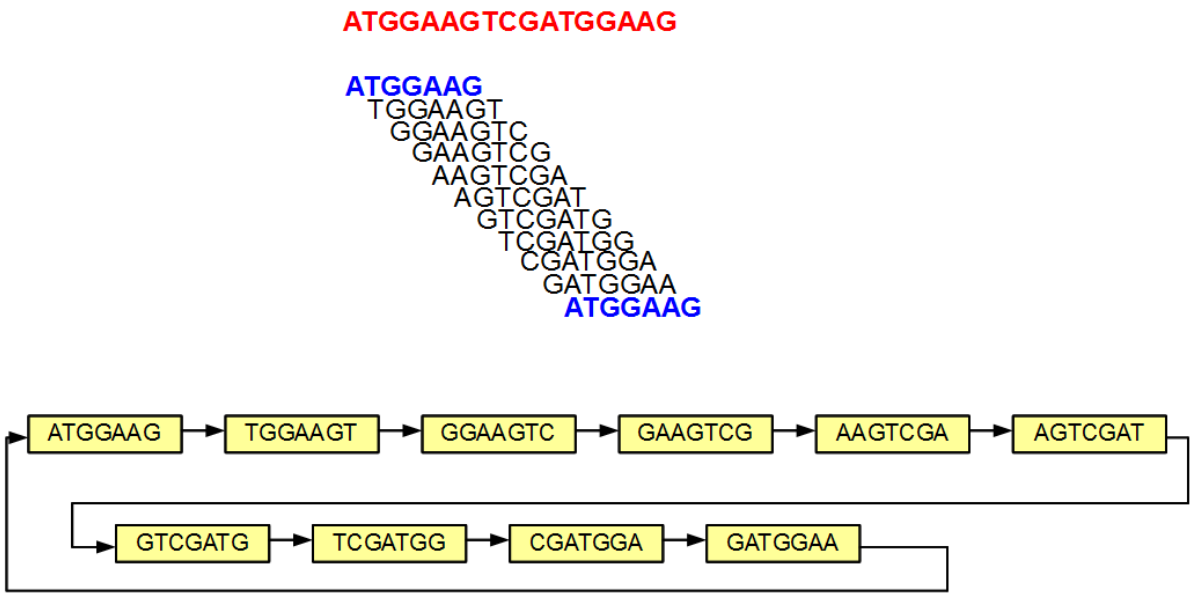


FIGURE 2.27 – Concept du graphe de *de Bruijn*.

De nombreux assembleurs existent, certains sont plus adaptés à certains types de données ; SPADES par exemple fonctionne mieux avec des petits génomes ; mais même à type de données équivalent un assembleur peut donner de bons résultats pour un organisme donné et de mauvais pour un autre (Magoc et al., 2013). L'utilisation de multiples assembleurs et le test de plusieurs paramètres est donc courant dans un projet d'assemblage .

Assemblages transcriptomiques

Dans le cas d'un transcriptome, l'objectif est de reconstruire les transcrits en présence avec la plus faible proportion d'erreurs d'assemblage. Différentes contraintes sont rencontrées : les données de séquençage peuvent ne pas couvrir l'entièreté des ARN (dégradation biologique des ARN avant séquençage, une profondeur de séquençage insuffisante pour les transcrits concernés), des ARN peuvent partager des séquences similaires pouvant générer la fusion artificielle de transcrits lors de leur reconstruction (e.g. domaines fortement conservés dans deux gènes distincts) ou être considérés comme des isoformes d'un même gènes. En effet, le nombre d'isoformes peut être très important : chez les humains par exemple, plus de 95% des gènes montrent des signes d'épissage alternatif (Pan et al., 2008) avec en moyenne, plus de cinq isoformes par gène. Un des exemples les plus frappants est le gène *DSCAM* de la

Drosophile pour lequel environ 30000 isoformes alternatif distincts sont décrits (Celotto and Graveley, 2001). A cela s'ajoute d'autres contraintes techniques comme le chevauchement d'UTR de gènes physiquement proches (fréquent dans les génomes compacts comme ceux de certains champignons) menant dans les assemblage à des fusions artificielles de transcrits. De même, la présence de deux gènes sur la même séquence génomique (l'un sur le brin Watson et l'autre sur le brin Crick) posera problème si le séquençage n'a pas été réalisé en "*strand specific*".

L'assemblage *de novo* des transcriptomes peut être réalisé avec une palette de différents outils, l'un des plus utilisé actuellement étant Trinity (Bankar et al., 2015). Dans le cas de cet assembleur, fonctionnant grâce à un algorithme de graphe *de Bruijn*, des paramètres permettent de tenir compte des différentes contraintes rencontrées (e.g. paramètre "jackard-clip" pour limiter les fusions de transcrits associés à un chevauchement d'UTR). Cet outil regroupe les transcrits au cours leur reconstruction (Kim et al., 2017) en cluster. L'un de ces niveaux de cluster appelé "gène" correspond au regroupement des transcrits (nommés isoformes) prédit comme issu du même gène (au sens biologique du terme).

Lorsqu'un génome de référence est disponible, il est possible de mapper les *reads* sur le génome puis de reconstruire les transcrits à partir ce mapping. Le mapping des *reads* doit être réalisé par un mapper dédié. En effet, il faut tenir compte de la structure des gènes : les données RNAseq étant obtenues à partir d'ARNm et le mapping réalisé sur un génome, il existe donc des *reads* qui sont situés sur les jonctions entre exons, côte à côte sur le mRNA mais pouvant être distants sur le génome à cause de la présence d'introns. Des outils ont été développés afin de tenir compte de cette structure particulière, les « splicing-aware aligner », tels que STAR (Dobin and Gingeras, 2015). Les transcrits sont ensuite reconstruit avec des outils d'assemblage des transcrits, tels que Cufflinks (Trapnell et al., 2010; Roberts et al., 2011) ou StringTie (Pertea et al., 2015, 2016).

L'approche en *genome guided* ne nécessite pas une profondeur de séquençage aussi importante qu'une approche *de novo* mais est très sensible à la qualité du génome de référence (les régions non assemblées du génomes ne pourront pas permettre la prédiction de transcrits, les erreurs d'assemblage du génome engendreront des erreurs d'assemblage des transcrits). L'approche *de novo* est quant à elle particulièrement sensible aux erreurs de séquençage, provoquant la création de transcrits partiels ou de transcrits chimériques (McGettigan, 2013).

Les outils d'assemblages, de génomes aussi bien que de transcriptomes, sont sujets à des développements très fréquents, avec des mises à jour régulières des outils, de nouvelles versions

d'un outil ou encore de nouveaux outils dans le but d'améliorer les performances autant au niveau de la qualité des prédictions que des performances en terme de ressources de calcul.

Déterminer la qualité des assemblages

Ne disposant pas de la séquence réelle, estimer la qualité d'un assemblage est un problème complexe. Une première approche repose sur la taille du génome, le nombre de contigs (ou scaffolds) et la taille du N50 [13]. Ces valeurs informent quant à la continuité de l'assemblage, on notera qu'un assembleur "agressif" peut produire des contigs plus longs au risque de fusionner des régions dans le mauvais ordre ou une mauvaise orientation (Dominguez Del Angel et al., 2018). Le pourcentage de fragments séquencés pouvant être alignés correctement sur le génome permet d'avoir une information sur l'éventuelle portion de génome manquante dans l'assemblage ainsi que sur la quantité d'erreurs qu'il comporte. Lorsqu'un génome de référence est disponible, le nombre d'erreurs trouvées entre les deux assemblages et leur proportion de régions potentiellement manquantes peut être estimé en les comparant (Magoc et al., 2013). Des outils d'évaluation des assemblages tels que Quast (Gurevich et al., 2013) permettent de comparer des assemblages et d'aiguiller l'utilisateur pour déterminer quel est le meilleur assemblage.

Un autre critère d'intérêt correspond à une estimation de la quantité de gènes codants des protéines présentes dans l'assemblage. Le plus utilisé, BUSCO, (Waterhouse et al., 2018) (développé en remplacement de CEGMA `http://www.acgt.me/blog/2015/5/18/goodbye-cegma-hello-busco`) permet de quantifier la présence et l'intégrité des séquences BUSCO [14] (*Benchmarking Universal Single-Copy Orthologs*) correspondant à une lignée taxonomique considérée (bactéries, eucaryotes, protistes, plantes, champignons etc.). En déterminant si les séquences recherchées sont complètes, dupliquées, fragmentées ou manquantes, cet outil permet d'une part de donner une indication sur la proportion du génome portant des gènes codant des protéines manquantes ansi que sur la redondance du génome assemblé au travers des séquences dupliquées (Simão et al., 2015). Le même outil également être utilisé pour les transcriptomes.

Malgré les efforts de séquençage et d'assemblage, une partie significative d'un génome peut rester manquante. A titre d'exemple, parmi les génomes d'oiseaux considérés comme complets, Peona et al. (2018) ont mis en évidence une différence de 7% à 42% (en moyenne 20 +- 9%)

13. Le N50 est la taille du plus petit contig tel que 50% du génome soit contenu dans les contigs de taille N50 et plus.
14. Groupe de protéines orthologues supposées systématiquement présente en une seule copie dans la lignée taxonomique considérée.

entre la taille estimée des génomes et la taille effective des assemblages lorsqu'un séquençage standard avec des séquences courtes était utilisé. Cette différence s'observe également chez les champignons : en 2018, Hess et al. ont estimé la taille des génomes d'*Amanita brunnescens*, *Amanita polypyramis* et *Amanita inopinata* à respectivement 262Mb, 144Mb et 29Mb tandis que la taille des assemblages étaient respectivement de 35Mb, 19Mb et 20Mb. On notera que plus d'une centaine de régions sont toujours inconnues dans le génome humain (Chaisson et al., 2015). Ces régions manquantes ne sont pas équitablement réparties dans le génome. Les régions avec une composition foncièrement différente du reste du génome (ex : très riche en GC) et régions fortement enrichies en ADN répétés (ex : éléments transposables et ADN satellites) sont généralement sous-représentées dans les assemblages (Peona et al., 2018). De même, il a été identifié que les familles de gènes répétées présentes dans les régions subtélomériques étaient souvent mal assemblées voire manquantes des assemblages (Cuomo and Birren, 2010).

2.4.3 Annotations

Une séquence génomique brute n'a que peu d'intérêt pour la plupart des biologistes. L'annotation d'un génome consiste à attacher aux séquences génomiques des informations ayant un sens biologique. De nombreux éléments peuvent être annotés (miRNA, lncRNA, sites de fixation des facteurs de transcriptions etc.), cependant la plupart de l'attention est consacrée à l'annotation des gènes codant des protéines. Cet état de fait est expliqué par Dominguez Del Angel et al. (2018) non pas par un moindre intérêt biologique de ces autres éléments génétiques mais du fait que les approches pour les caractériser sont soit assez directes (par exemple INFERNAL (Nawrocki, 2014) et tRNAscan-SE (Lowe and Chan, 2016) permettent la détection efficace des ARNt sans nécessiter un effort important d'adaptation des paramètres utilisés) ou doivent faire l'objet d'analyses spécialisées (sites de fixation des facteurs de transcriptions, un séquençage spécifique permet de détecter et annoter les miRNA).

Avant de réaliser l'annotation des gènes, il convient de s'intéresser à l'identification des répétitions. Cette étape permet par la suite de décider d'une éventuelle occultation de ces régions vis à vis des outils d'annotations géniques. L'annotation des gènes codant des protéines se divise en deux étapes : l'annotation structurale et l'annotation fonctionnelle. L'annotation structurale permet de positionner les éléments génétiques sur la séquence génomique. L'annotation fonctionnelle permet d'assigner à chacun de ces éléments sa fonction biochimique potentielle. Dans le cadre des transcriptomes, la structure génique est prédite lors de l'assemblage. Ainsi, seule la

partie fonctionnelle, réalisée selon la même méthodologie que pour les annotations génomiques est nécessaire.

L'annotation des éléments répétés

Divers outils ont été développés pour détecter et annoter les éléments transposables (TE) (Flutre et al., 2011; Hoede et al., 2014; El Baidouri et al., 2015; Hoen and Bureau, 2015; Hoen et al., 2015; Zeng et al., 2018), les plus utilisés combinent de nombreux outils pour fonctionner. L'un des plus connus est RepeatMasker qui combine cinq autres outils de recherche (nhmmer, cross_match, ABBlast/WUBlast, RMBlast et Decypher) pour rechercher les éléments répétés des bases de données dédiées Dfam (profile HMM) (Hubley et al., 2016) et Repbase (Bao et al., 2015). Un autre outil largement utilisé est le package REPET (Flutre et al., 2011). Il est composé de deux principaux pipelines : TEdenovo et TEannot. TEdenovo permet de détecter *de novo* et classifier les TE présents dans un génome, il utilise de nombreux outils et méthodes tels que LTRharvest, Recon, Grouper, MCL, Map, Piler, ou encore PASTEC (Hoede et al., 2014). On notera que PASTEC utilise la classification de Wicker (Wicker et al., 2007) pour classer les TE nouvellement identifiés. Une fois les TE identifiés, cette nouvelle base de données peut être utilisée seule et/ou avec d'autres bases de données telle que Repbase par le pipeline TEannot. Ce pipeline permet de tirer parti de multiples outils et méthodes préexistants tels que Blaster, CENSOR, RepeatMasker, TRF, Mreps etc. afin de générer une annotation des éléments répétés.

L'annotation structurale des gènes

Pour réaliser l'annotation structurale des gènes, il existe des méthodes dites intrinsèques (ou *ab-initio*), d'autres extrinsèques et celles qui combinent les deux approches.

Les méthodes *ab-initio* se basent uniquement sur la séquence en elle-même. Les outils associés se basent sur les caractéristiques des séquences géniques propres à chaque groupe d'espèces : biais d'usage du code génétique, fréquence et taille des introns etc. Pour déterminer ces caractéristiques puis établir un modèle de gènes, ces logiciels doivent être entrainés pour l'espèce d'intérêt. Certains logiciels comme GeneMark-ES fungal (Ter-Hovhannisyan et al., 2008) ne nécessitent que la séquence génomique pour réaliser leur entraînement. D'autres, les plus nombreux, se basent sur un jeu de gènes d'entraînement ou des indices fournis par l'utilisateur. C'est le cas notamment d'Augustus (Hoff and Stanke, 2013) qui demande un jeu d'au moins 200 gènes

pour créer un modèle de qualité. Cependant, il est difficile d'obtenir un si grand nombre de gènes avant que l'annotation ne soit déjà à un stade avancé.

Les méthodes extrinsèques font appel à la similarité avec d'autres séquences pour prédire la position des gènes. Ces séquences peuvent provenir de bases de données (tel que la base de protéines non redondante du NCBI, Uniprot ou RefSeq) ou de données expérimentales comme celle d'expression des gènes (RNAseq notamment). Les données associées aux ARN sont d'un grand intérêt car elles fournissent des indications précises sur la structure des gènes (UTR, bordure d'introns, cas de transcrits alternatifs etc.). cependant ces données ne sont pas accessibles pour tous les gènes (les gènes ne sont pas tous exprimés, profondeur de séquençage insuffisante) et certains introns peuvent rester présents à cause d'épissage incomplet. D'un autre côté, utiliser les données protéiques issues des bases de données publiques nécessite de mapper ces protéines sur les six cadres de lectures, la question étant de déterminer à quel seuil considérer qu'un mapping est pertinent : trop stringent et des protéines plus éloignées des bases de données ne seront pas retrouvées, pas assez stringent et des mapping obtenu ne seront pas assez spécifiques ou induiront la prédiction de gènes qui n'existent pas.

Les approches hybrides consistent à utiliser une combinaison de méthodes *in silico* et expérimentales. Des outils tels que MAKER (Cantarel et al., 2008) ou EVidenceModeler (Haas et al., 2008) permettent, à partir des annotations réalisés par des outils de prédictions ab initio et des annotations issues d'informations expérimentales de générer une annotation consensus la plus plausible compte tenu des différentes annotations.

Les annotations fonctionnelles des gènes

Une fois la structure des gènes établie, on peut chercher à identifier la fonction potentielle de ces derniers. Cette recherche est principalement basée sur la structure protéique prédite au travers de la recherche de domaines et des motifs protéiques connus, ainsi que la recherche de correspondances avec des protéines et gènes codants déjà identifiées dans les bases de données.

L'une des méthodes les plus couramment utilisées est la recherche d'une homologie de séquence avec d'autres protéines avec des outils tels que le blastp (Altschul et al., 1990) ou diamond (Mai et al., 2018) contre différentes bases de données. Toutes les bases ne sont pas de qualité égales, certaines telles que SwissProt/UniProtKB (Consortium 2015) sont de haute qualité car manuellement vérifiées ce qui implique un nombre relativement restreint de protéines à disposition(~550 000 dans la version d'octobre 2018 (`www.uniprot.org`)) ainsi qu'une faible

diversité phylogénétique des espèces d'origine de ces séquences. Par contre, d'autres comme la base de données des protéines non redondantes du NCBI sont de moindre qualité mais plus exhaustives (~174 000 000 protéines dans la version d'octobre 2018). Ces deux dernières bases de données sont généralistes, mais selon les objectifs de l'étude, l'utilisation de base de données spécifiques peut se révéler être un complément pertinent (par exemple, une base de données est dédiée aux peptidases (MEROPS) (Baterman et Finn 2018) `https://www.ebi.ac.uk/merops/`) Des informations complémentaires peuvent être obtenues grâce à la prédiction de domaines et la détection de motifs protéiques. Ainsi la prédiction de domaines transmembranaires (avec tmhmm (Krogh et al., 2001) par exemple), ou de motifs associé à un peptide signal (avec SignalP (Nielsen, 2017)), TargetP (Emanuelsson et al., 2000) ou encore SecretomeP (Bendtsen et al., 2004)) permet d'avoir une information sur la localisation cellulaire de la protéine et si elle est potentiellement associée à une membrane. La recherche de domaines fonctionnels connus permet également d'orienter sur la fonction de la protéines. Des bases dédiées telles que PFAM (Finn et al., 2016) ou PROSITE (Sigrist et al., 2013) permettent de recenser un grand nombre de motifs qui peuvent être recherchés sur des clusters de calcul (cas des domaines PFAM avec des outils tels que hmmscan) ou avec des outils en ligne (cas de PROSITE ou de la recherche de domaines conservés associée au NCBI (Marchler-Bauer et al., 2017)). D'autres types d'outils sont également utilisés pour des recherches spécialisées, par exemple dbCan2 (Zhang et al., 2018) permet de rechercher et annoter spécifiquement les Carbohydrate active enzymes (CAZymes). PRIAM (Claudel-Renard et al., 2003) permet de détecter et classer les enzymes selon la classification numérique de "l'Enzyme Commission" (EC numbers). Les métabolites secondaires fongiques peuvent être recherchés grâce à des outils tels que SMURF (Khaldi et al., 2010) ou antiSMASH (Blin et al., 2017) Il est ensuite possible de rapprocher la fonction prédite des gènes aux différents processus biologiques associés, le but étant d'avoir une vue d'ensemble des voies métaboliques et processus biologique en présence et d'identifier les gènes associés à chacun d'entre eux. Différents systèmes existent, comme *Gene Ontology* (GO) (Holliday et al., 2017), KEGG (Kanehisa, 2017) ou encore BioCyc (Paley and Karp, 2017) ou MetaCyc (Caspi et al., 2018).

Les éléments présentés ci-dessus (connaissances sur le genre fongique, éléments génomiques, séquençage et analyses bioinformatiques) ont été utilisés afin de répondre à l'objectif de ce projet, à savoir identifier les caractéristiques génomiques des *Mucor* et des marques d'une potentielle adaptation au milieu. Ces analyses ont été d'une part été réalisées en comparant les transcriptomes de cinq souches de *Mucor* et d'autre part en comparant les génomes de dix souches de *Mucor*. Les analyses bioinformatiques associées à la comparaisons des génomes et transcriptomes seront détaillées dans les parties associées.

Chapitre 3

Approche transcriptomique

3.1 Introduction

Plusieurs centaines d'espèces de *Mucor* ont été répertoriées (Morin-Sardin et al., 2017). Ces espèces sont principalement ubiquistes mais quelques-unes n'ont jusqu'à présent été retrouvées que dans des milieux spécifiques (par exemple, *M. lanceolatus* sur des fromages ou *M. endophyticus* en tant qu'endophyte du blé) (Morin-Sardin et al., 2017). Leur impact sur les activités humaines est ambivalent : certaines sont utilisées en biotechnologie tandis qu'une poignée d'espèces sont des pathogènes opportunistes ((Schwartze et al., 2014; Prakash et al., 2017; Morin-Sardin et al., 2017). Le genre *Mucor*, tout comme les autres genres appartenant aux lignées ayant divergé très tôt dans l'évolution des champignons est bien moins connu que les genres appartenant aux Ascomycètes et Basidiomycètes. En particulier, les données génomiques et transcriptomiques des espèces composant le genre *Mucor* sont restreintes et les quelques génomes disponibles ciblent principalement les espèces pathogènes opportunistes comme *M. circinelloides* (Tang et al., 2015; Corrochano et al., 2016; Lopez-Fernandez et al., 2018) ou *M. indicus* (Chibucos et al., 2016).

Au cours de ce projet a été initié une démarche de génomique comparative intégrant plusieurs espèces séquencées pour l'étude : *M. fuscus*, *M. lanceolatus*, *M. racemosus* et *M. endophyticus*. Ces espèces ont été choisies afin d'étendre le spectre des génomes disponibles en incluant des espèces qui ne sont pas connues comme pathogènes opportunistes. Afin de produire des annotations de qualité, l'adjonction de données transcriptomiques étaient fortement recommandée (Dominguez Del Angel et al., 2018). Dans le cadre de cette étude, les transcriptomes de ces quatre espèces et de la souche référence *M. circinelloides* CBS 277.49 (Corrochano et al., 2016) ont été séquencés. Au début du projet, ne disposant pas encore des données génomiques, ces données transcriptomiques ont été exploitées par elles même pour mieux définir les caractéristiques du

transcriptome du genre *Mucor* et afin de déterminer s'il existait des éléments susceptibles d'indiquer une adaptation de certaines espèces à leur mode de vie et aux niches occupées avec un intérêt plus particulier pour l'environnement fromager (Figure 3.1)

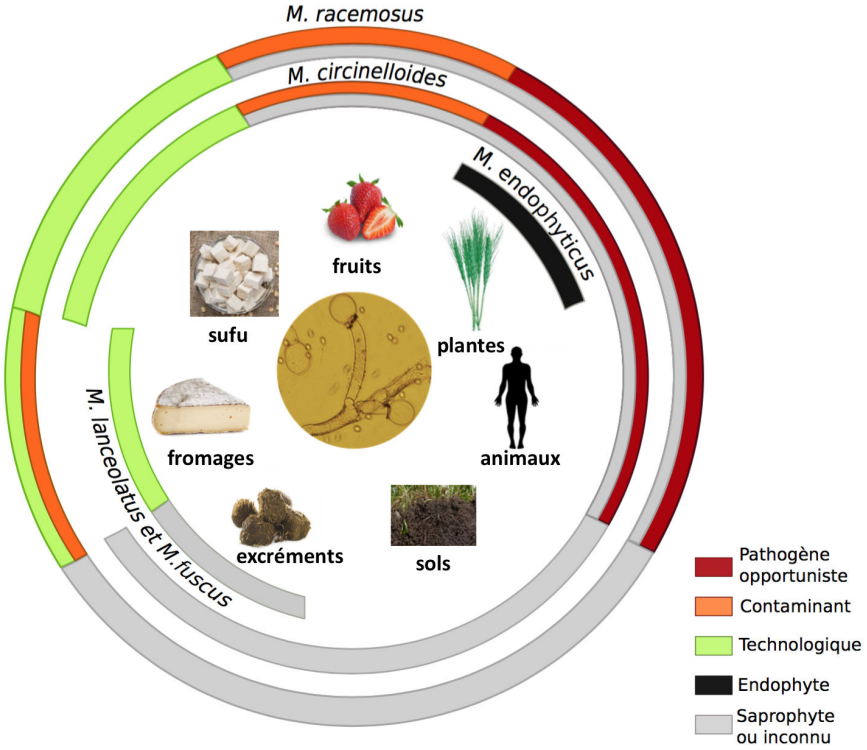


FIGURE 3.1 – Milieu de vie des espèces étudiées dans l'analyse transcriptomique.

Dans un premier temps, les transcriptomes obtenus à partir des cinq souches (*M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153, *M. racemosus* UBOCC-A-109155, *M. endophyticus* CBS 385-95 et *M. circinelloides* CBS 277.49) ont été reconstruits *de novo* et annotés (Figure 3.2). Les premières étapes de l'annotation ont consisté à identifier les ARN ribosomiques et prédire les régions potentiellement codantes des transcrits. Puis, des signatures de domaines connus ont été recherchées sur les protéines prédites, à savoir les domaines répertoriés par la base de données PFAM, les domaines transmembranaires, les domaines indiquant la présence d'un peptide signal. D'autre part, les fonctions potentielles des transcrits ont été identifiées par recherche d'homologies. Ces fonctions potentielles ont été annotées sous la forme de *GO terms* et des annotations complémentaires propres aux enzymes ont été présentées sous la forme d'*EC numbers* . Enfin des familles de gènes spécifiques, impliquées dans l'exploitation du fromage, la virulence, et la synthèse de métabolites secondaires ont été manuellement annotées.

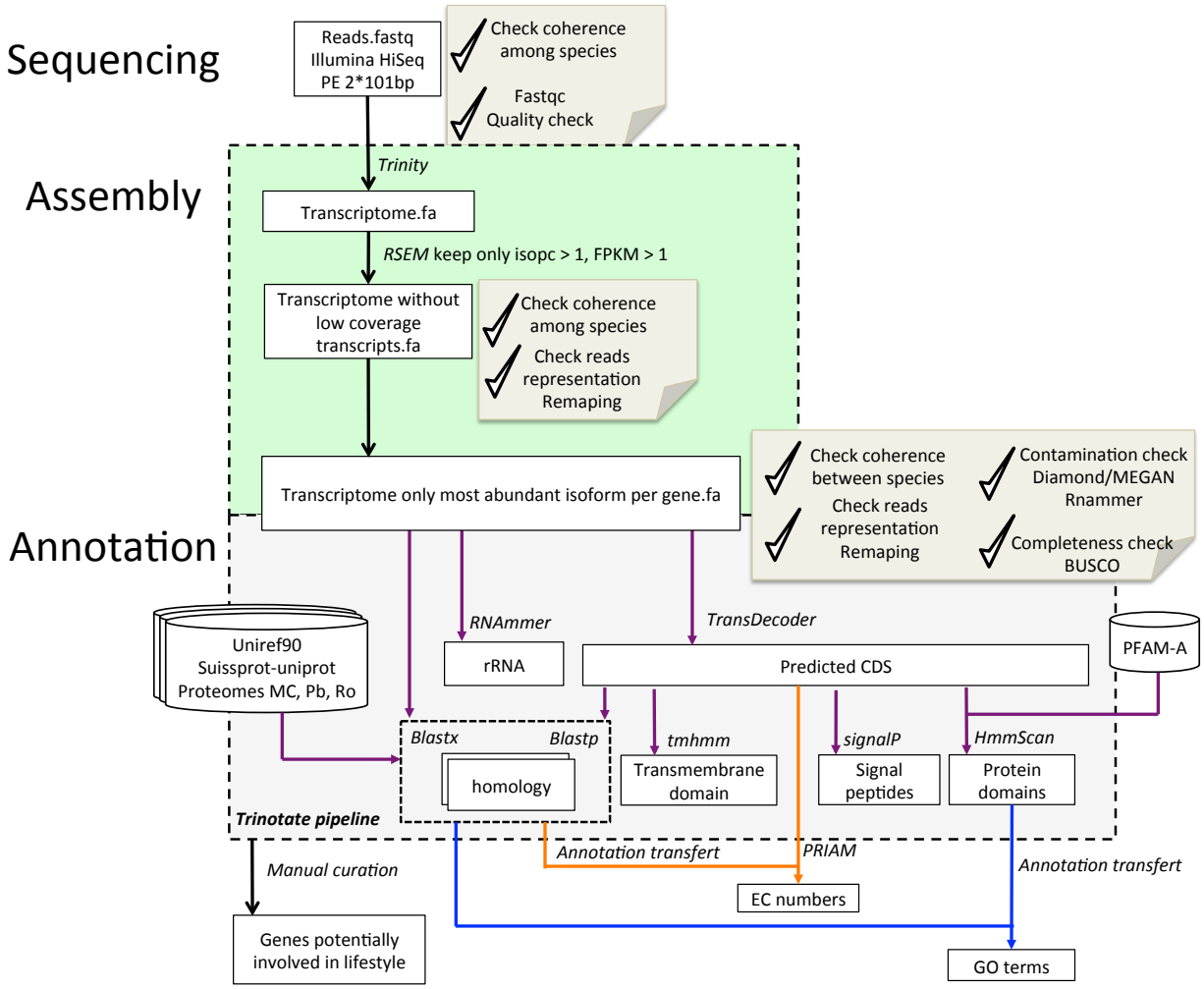


FIGURE 3.2 – Pipeline d'assemblage et annotation des transcriptomes.

Afin de définir les caractéristiques du transcriptome du genre *Mucor* et déterminer s'il existait des éléments indiquant une potentielle adaptation de certaines de ses espèces à leur mode de vie, trois approches ont été utilisées. (i) Les annotations fonctionnelles globales (via les *GO terms*) et plus spécialisées sur les contenus en enzymes (via les *EC numbers*) ont été comparées entre les cinq espèces (la stratégie de travail est représentée en figure 3.3). (ii) À partir des transcrits de chacun des transcriptomes, des protéines ont été prédites. Ces dernières ont été regroupées en familles (appelées orthogroupes) permettant d'accéder aussi bien aux protéines prédites présentes chez les cinq espèces étudiées qu'à celles propres à chacune des espèces. Cette approche a notamment permis de comparer les groupes de protéines prédites partagées par les espèces ayant des modes de vie similaires et d'étudier les protéines prédites propres à chacune des espèces. (iii) Enfin, le contenu en transcrits identifiés dans la littérature comme étant importants pour l'adaptation au milieu de vie des espèces a été comparé entre les différentes

souches.

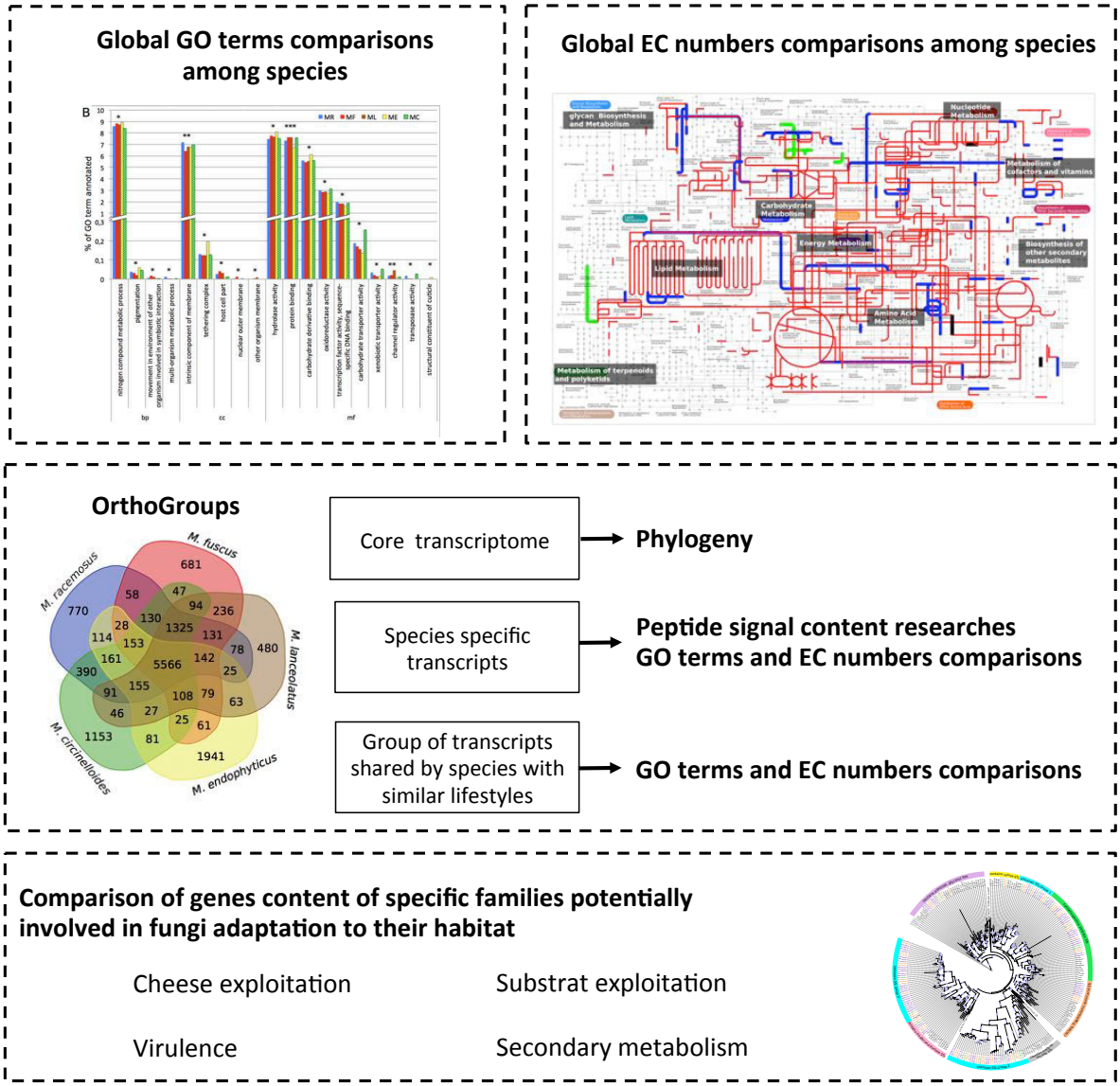


FIGURE 3.3 – Vue d'ensemble de la stratégie de travail adoptée pour la comparaison des annotations réalisées dans le cadre de l'analyse des transcriptomes. (L'analyse de chacune des figures sera détaillée dans la suite du document)

Ces résultats ont été présentés dans le cadre de l'article "*Comparative analysis of five* Mucor *species transcriptomes*" actuellement in press dans Genomics. Des analyses complémentaires à celles de l'article sont présentés à la suite de celui-ci.

3.2 Article : Comparative analysis of five *Mucor* species transcriptomes

Auteurs :

Lebreton Annie[a], Meslet-Cladière Laurence[a], Morin-Sardin Stéphanie[a], Coton Emmanuel[a], Jany Jean-Luc[a], Barbier Georges[a], Corre Erwan[b]

Affiliations :

[a] Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, Université de Brest, Technopôle Brest-Iroise, Plouzané 29280, France

[b] Station biologique de Roscoff, plateforme ABiMS, Sorbonne Université (UPMC), Roscoff 29682, France

Correspondance : corre@sb-roscoff.fr

Article *accepté* dans *Genomics*.

https ://doi.org/10.1016/j.ygeno.2018.09.003

Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Comparative analysis of five *Mucor* species transcriptomes

Lebreton Annie[a], Meslet-Cladière Laurence[a], Morin-Sardin Stéphanie[a], Coton Emmanuel[a], Jany Jean-Luc[a], Barbier Georges[a], Corre Erwan[b,*]

[a] Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, Université de Brest, Technopôle Brest-Iroise, Plouzané 29280, France
[b] Station biologique de Roscoff, plateforme ABiMS, Sorbonne Université (UPMC), Roscoff 29682, France

A B S T R A C T

Mucor species belong to the Mucorales order within the Mucoromycota phylum, an early diverging fungal lineage. Although *Mucor* species are often ubiquitous some species have been reported to specifically occur in certain ecological niches. In this study, similarities and differences of a representative set of *Mucor* species with contrasted lifestyles were investigated at the transcriptome level. Five strains pertaining to five different species were studied, namely *M. fuscus* and *M. lanceolatus*, two species used in cheese production (during ripening), *M. racemosus*, a recurrent cheese spoiler sometimes described as an opportunistic pathogen, *M. circinelloides*, often described as an opportunistic pathogen and *M. endophyticus*, a plant endophyte. A core transcriptome was delimited and a phylogenetic analysis led to an altered phylogenetic placement of *M. endophyticus* compared to previously published topologies. Interestingly, the core transcriptome comprising 5566 orthogroups included genes potentially involved in secondary metabolism. As expected, given the wide taxonomic range investigated, the five transcriptomes also displayed specificities that can be, for some of them, linked to the different lifestyles such as differences in the composition of transcripts identified as virulence factors or carbohydrate transporters.

1. Introduction

Within the filamentous fungi, the *Mucor* genus belongs to the Mucorales order within the Mucoromycota phylum, an early diverging fungal lineage [1]. *Mucor* species are common and often ubiquitous [2,3], their fast growing and high sporulating mycelium, consisting of coenocytic hyphae, are encountered in a large variety of environments, with the exception of low water activities (a_w) substrates. Growth of several *Mucor* species has been documented to be limited to relatively high a_w (> 0.90) [4]. The *Mucor* genus mainly comprises mesophilic species but also some thermotolerant and thermophilic species [2], some of them being animal and human opportunistic pathogens responsible for mucormycoses [5,6] which are increasingly frequent, especially in immunocompromised patients [7]. *Mucor* spp. are mostly saprobes, with some species being described as plant endophytes [8].

Interestingly, several *Mucor* species have an obvious biotechnological interest, for metabolite production (e.g., biofuels) and biotransformations (e.g., terpenoid biotransformations) but also in food production, especially in fermented Asian and African food but also in cheese ripening (e.g. Tommes or Saint-Nectaire in France) (for a review, see [9]). Since *Mucor* strains used for cheese ripening can be considered as technological and have been only described so far in cheese, the question of their potential adaptation to this matrix has been raised [9]. An adaptation hypothesis in cheese technological strains was supported by the results of a recent study that showed that, contrary to other *Mucor* strains tested (*M. racemosus*, *M. circinelloides*, *M. brunneogriseus*, *M. spinosus* and *M. endophyticus)*, *M. lanceolatus* and *M. fuscus* (technological strains) showed higher optimal growth rates (μ_{opt}) on cheese matrices than on Potato Dextrose Agar (PDA) medium [4]. Moreover, lag times of the *M. endophyticus* endophyte strain were strongly extended on cheese related matrices. The apparent adaptation to the cheese environment of *M. lanceolatus* was also confirmed by morphological observations as well as by a higher ratio of over accumulated proteins on Cheese agar versus PDA [2].

A recent large effort to generate genome data concerning the early diverging fungi has helped refine their taxonomy [10] and has shed new light on *Mucor* genome evolution and functions such as sensory perception [11], lipid metabolism [12] or pathogenesis [13]. The lifestyle diversity within the genus *Mucor* offers interesting perspectives to better understand evolutive adaptation to different life modes, e.g., saprobic, pathogenic and even adaptation to anthropogenic conditions. The present study aimed to provide an overview of the common or specific patterns of gene expression of five *Mucor* species with contrasting lifestyles, grown in standard fungal culture.

* Corresponding author.
E-mail address: corre@sb-roscoff.fr (C. Erwan).

https://doi.org/10.1016/j.ygeno.2018.09.003
Received 30 March 2018; Received in revised form 29 August 2018; Accepted 4 September 2018

Table 1
List of *Mucor* strains used in the present study, their origin and reported habitats for the corresponding species according to Walther et al. [38]; Hermet et al. [39]; Zheng and Jiang [8].

Species	Strain	Strain origin	Reported habitat	Reported role
M. racemosus	UBOCC-A-109155	Cheese	Cheese, yogurt, walnuts, sausages, grassland soil, decaying vegetables, human	Food contaminant, technological in cheese production, pathogen
M. fuscus	UBOCC -A-109160	Cheese	Cheese, dung, sediment	Technological in cheese production
M. lanceolatus	UBOCC-A-109153	Cheese	Cheese	Technological in cheese production
M. endophyticus	CBS 385-95	*Triticum aestivum*; leaves;	*Triticum aestivum* endophyte	None
M. circinelloides	CBS 277-49	Unknown	Sufu, corn grain, fungi (basidiomycota), human, forest soil, decaying vegetables	Food contaminant technological in sufu production, pathogen

2. Materials and methods

2.1.1. Mucor strains, culture conditions

Five *Mucor* strains belonging to different species reported to have contrasting lifestyles and habitats, namely *M. fuscus* (MF) and *M. lanceolatus* (ML) (two species used in cheese ripening), *M. racemosus* (MR) (a recurrent cheese spoiler), *M. circinelloides* (MC) (often classified as an opportunistic pathogen) and *M. endophyticus* (ME) (a plant endophyte species) were used in the present study (Table 1). They were obtained from the Université de Bretagne Occidentale Culture Collection, (UBOCC, France; http://www.univ-brest.fr/ubocc) or ordered from the Westerdijk Fungal Biodiversity Institute (The Netherlands, http://www.westerdijkinstitute.nl/). Strains were maintained in the dark at 25 °C on PDA medium (Difco Laboratories, Detroit, Michigan). Spore suspensions of each strain were produced as previously described by Morin-Sardin et al. [4]. Concentrations were adjusted to 10^7 to 10^8 spores·mL^{-1} prior to storage at −80 °C until use. For RNA extraction, each fungal strain was grown on PDA solid medium at 25 °C for 7 days. The whole organism was collected for sequencing.

2.1.2. RNA extraction and sequencing

Total RNA extraction from each strain thallus was performed using the RNeasy plant mini kit (Qiagen, Courtaboeuf, France) following the manufacturer's instructions. The RNA-seq libraries were prepared from total RNA using the Illumina Ribo-Zero rRNA Removal Kit (Epicenter,Madison, WI) and converted to cDNA with the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA) following the manufacturer's instructions. The libraries containing the cDNA from each sample were sequenced on an Illumina HiSeq 2000 system (Illumina) with a sequencing configuration for 100 bp paired-end reads.

2.1.3. RNAseq quality check and assembly

Data quality of the sequencing files was checked using the FastQC software [14]. For each strain, de novo transcriptome was reconstructed using Trinity (release 2014-07-17; [15]) with the '–trimmomatic –normalize_reads –normalize_by_read_set' options. Weakly expressed transcripts (isoform percentage < 1 and fragment per kilobase of transcript per million mapped reads (FPKM) < 1) were removed from the dataset with the RSEM tool [16] included in the Trinity package. For each Trinity-predicted "gene" only the most abundant isoform was conserved in the final dataset.

2.1.4. Transcript annotation

Eukaryotic and prokaryotic ribosomal RNA were screened with Rnammer (v.1.2; [17]). Transcript open reading frames (ORFs) were predicted using Transdecoder (r2014-07-04, available at http://transdecoder.github.io), protein domains were predicted with HMMER (v.3.1b1, [18]) based on the PFAM-A database [19], transmembrane helices were identified with tmhmm (v.2.0, [20]), signal peptide cleavage sites were predicted with SignalP (v.4.1, [21]). All predicted transcripts and predicted proteins were identified using respectively BLASTx and BLASTp [22] against Swissprot-Uniprot (r2017-01-18) and Uniref90 (r201611-02) databases and against the filtered predicted proteomes of *M. circinelloides* CBS 277.49 (v2, [11]), *Rhizopus oryzae* 99-880 [23] and *Phycomyces blackesleeanus* NRRL1555 (v2, [11]) available on the Joint Genome Institute (JGI) database (genome.jgi.doe.gov/). An *E*-value cut-off of 1e-4 was used for all BLASTp and BLASTx analysis, only the best match per database and protein was conserved. Putative functions were assigned to predicted proteins using the Gene Ontology (GO) database [24] and the Enzyme Commission number (EC) classification [25]. Predicted proteins were assigned to GO categories by transferring the annotation of BLAST matches and protein domains. EC numbers were assigned by profile search with PRIAM (r2015-03-04, [26]) and transferred from the predicted proteomes BLAST matches.

2.1.5. Transcriptome quality check

In order to assess the quality of the transcriptomes, statistics such as percentage of realigned reads and N50 were calculated. Transcriptome completeness in terms of gene content was assessed by searching with BUSCO (Benchmarking Universal Single-Copy Orthologs, v2, http://busco.ezlab.org/, [27]) for the presence/absence of the conserved eukaryotic orthologous genes available in the BUSCO software.

2.1.6. Comparative transcriptomics

Significant differences in GO category occurrence among species were identified with an internal script using the Fisher's exact test (pvalue < 0.05, available on github). All EC numbers for each species were mapped onto metabolic networks with iPath2 [28]. When a species specific EC number was found, the transcript annotation was checked manually. All predicted proteomes were compared against each other based on sequence similarity to identify orthologous proteins using the software Orthofinder v.1.1.2 [29] (*E*-value 1e-2, inflation 1.5). Orthologous proteins were clustered according to the reported lifestyle of the producing fungal species. Clusters were then compared: cheese/non-cheese, pathogen/non-pathogen and core transcriptome/non-endophyte (orthogroups composed of proteins of all species except ME). GO categories were assigned to orthogroup by transfer of protein annotation. A single copy of each annotation was kept to avoid annotation redundancy.

2.1.7. Phylogenomic analysis

Single copy orthologs shared by the five studied *Mucor* spp. as well as *Rhizopus oryzae* strain 99-880 and *Phycomyces blackesleeanus* strain NRRL1555 (both later species being considered as outgroups) were kept for phylogenetic reconstruction.

For each of the 1289 obtained clusters, a multiple alignment was inferred using PRANK v.170427 [30], run with default settings. Spuriously aligned regions were excluded with trimAl v1.4.r15 [31] with a 0.2 gap threshold. Subsequent alignments were concatenated in a supermatrix of 727,479 sites. This matrix was used to reconstruct species

tree by maximum likelihood inference and by Bayesian Monte Carlo Markov Chain (MCMC) samples. RAxML PTHREADS v. 8.2.9 [32], a program for Maximum Likelihood based inference, was used with a partitioned WAG + G model, where each data partition represented a single input gene family. A bootstrap analysis with 100 replicates under the same model was performed in RAxML in order to assess branch support of the tree. Alternatively, the PhyloBayes v3.3 MCMC samplers [33] was used with a CAT + GTR model and 3 chains.

2.1.8. Investigation of specific gene families

Transcripts involved in secondary metabolism (*PKS* and *NRPS*), sugar transport, aminoacyl degradation (decarboxylases, transaminases and deaminase), fatty acid degradation (alkaline and acidic lipases) and virulence (*M. circinelloides* CBS 277–49 ferroxidases *fet3* and *ID112092*, *Rhizopus delmar fob*, *FTR1* and *cotH*) were identified in each transcriptome. *PKs* and *NRPs* transcripts were identified using the genes annotated in the *Mucor circinelloides* CBS277.49 genome available at the JGI. Sugar transporters, *cotH* transcripts and *FTR1* transcripts were identified using the corresponding domain profiles: respectively Pfam ID PF00083, PF08757 and PF03239. Alkaline and acidic lipases and sequences of genes involved in virulence were collected from the NCBI database and used to search the *Mucor* predicted proteomes using diamond BLASTp [34]. Matching predicted proteins were then used as queries to search *Mucor* predicted proteomes to identify predicted proteins that might have significant sequence variation compared to the reference sequence. Annotation of each matching sequence was then checked using the conserved domain search tool available on NCBI [35].

A gene tree of putative sugar transporters was reconstructed using similar methodology as presented above. *Aspergillus niger* putative sugar transporters and 61 experimentally characterized fungal sugar transporters presented in Peng et al. [36] were appended to the analysis. Gene tree was displayed using iTOL v3 [37].

3. Availability of supporting data

Sequence data are available in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-6453. Transcriptomes are available upon request. Internal script allowing the analysis of GO terms can be found on github (https://github.com/anlebreton/GO_terms_FisherTest.git).

4. Results

4.1. RNA sequencing, transcriptome assembly and annotation

Sequencing of the five *Mucor* transcriptomes resulted in 25 to 35 million pairs of 101 base paired-end reads (Table 2). Reads were assembled into 16,950 to 21,556 filtered transcripts grouped into 13,655 to 15,554 Trinity “genes”. Except for ML, assemblies correctly represented the read sets. Indeed, the percentage of realigned reads was above 93% for all species except ML which was 70% (73% for the transcriptome including all isoforms). Except for ME, the average transcript length was close to 1200 bp for all species. In addition, > 97% of single-copy orthologs were complete reflecting the high-quality of the assemblies. The assembly of ME was of lower quality as shown by the N50 of 1292 bp, a higher number of predicted genes and the identification of single-copy orthologs against the BUSCO eukaryotic database of which 67% were complete, 20% fragmented and 13% missing. Despite this difference, functional annotation was similar among species: the number of transcripts with protein prediction varied from 9383 to 10,808 and among these predicted proteins 61% to 72% had GO term annotation and 19% to 24% had an EC number assignment (Table 2).

4.2. Functional comparisons

Overall, 1212 different EC numbers were assigned, among them 956 (79%) were shared among all strains (Fig. 1A). We did not identify any strains-specific pathways (data not shown). Even though a significant proportion of EC were lacking for ME (7%) compared to the other strains, all pathways were complete and present in this strain.

Nineteen GO categories were differently represented among strains (level 2) (Fisher exact, p-value < .05). These categories corresponded to primary catabolism and anabolism (e.g. “nitrogen compound metabolic process”), transport (e.g. “carbohydrate transporter activity”) as well as secondary metabolism (e.g. “pigmentation”) (Fig. 1B).

4.3. Ortholog groups

The predicted proteins were clustered in ortholog groups (orthogroups). The core transcriptome of the five *Mucor* species grown on PDA medium comprised 5566 orthogroups, whereas 5017 predicted proteins could not be grouped (singletons) (Fig. 2A).

Based on the obtained single copy orthologs, a phylogenomic tree was reconstructed. This tree was concordant with the previously published results [38] except for the placement of ME which was found basal compared to the other species in the present study (Fig. 2B).

Some orthogroups were composed of multiple predicted proteins associated to the same species. The 21 orthogroups consisting of > 10 predicted proteins associated with one species were investigated. Seven orthogroups corresponded to unknown proteins while 14 could be assigned to a putative function (Fig. 2C). Interestingly, two orthogroups were composed of predicted proteins corresponding to transposons (OG03 and OG12). In the first orthogroup MF and ML had a higher number of predicted proteins by fourfold compared to other species whereas in the second group MF had a higher number of predicted proteins by at least twofold. In addition, the GO category transposase activity was overrepresented in MR and MC compared to the other studied strains. Notably, the number of predicted proteins was found

Table 2
Overview statistics of *Mucor racemosus* (MR), *Mucor fuscus* (MF), *Mucor lanceolatus* (ML), *Mucor endophyticus* (ME) and *Mucor circinelloides* (MC) including transcriptome size, assembly quality and annotation.

species	MR	MF	ML	ME	MC
Filtered reads (2 × 101 bp)	30,203,513	25,511,572	24,835,844	34,852,641	24,642,543
No. genes*	14,035	14,299	14,041	15,554	13,655
No. transcripts	17,368	20,898	21,556	16,950	19,891
Average transcript length	1205	1231	1209	836	1295
Remapped reads (%)	93.08	95.33	70.93	96.94	93.60
Complete BUSCO genes (%)	99,34	97,36	97,36	67,00	97,69
no. of predicted proteins coding genes*	10,808	9963	9383	10,434	10,194
No. predicted proteins with EC numbers assignment	3134	2886	2801	2727	3085
No. predicted proteins with GO term annotation	10,202	9021	8729	11,280	9506

Gene* refers to cluster of transcripts predicted to be transcribed from the same gene by the assembler Trinity.

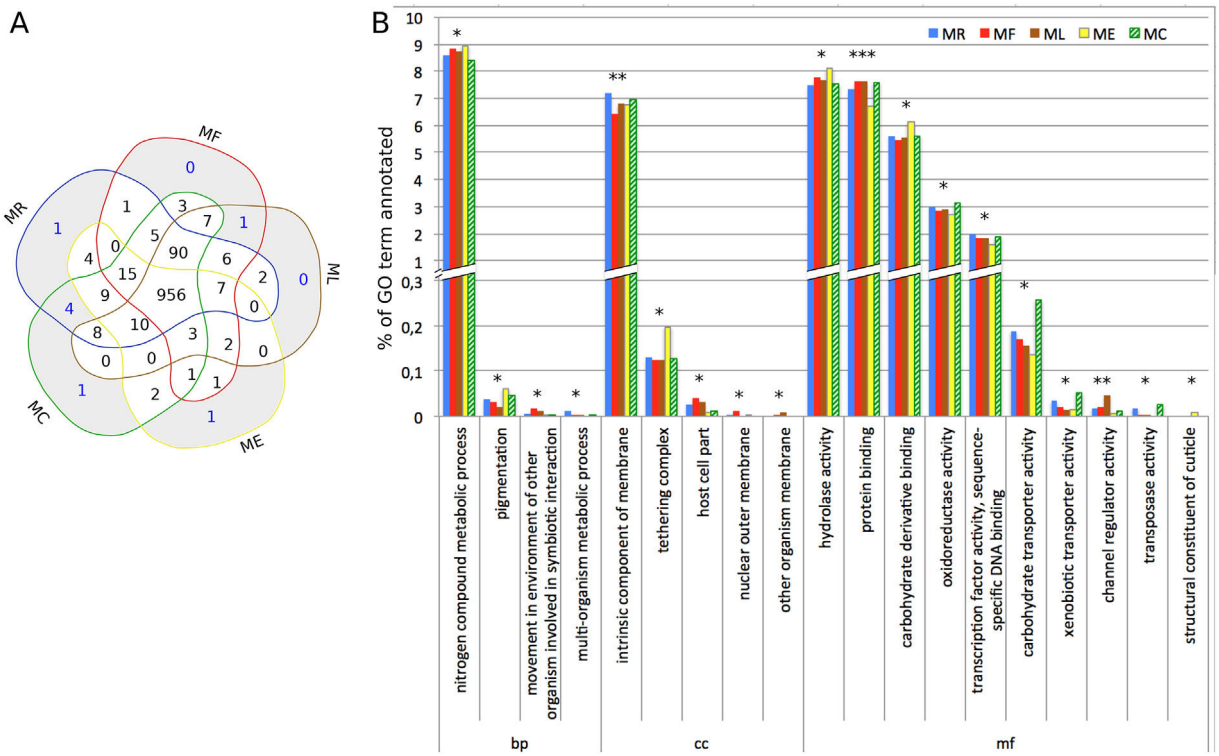


Fig. 1. : Global comparisons of the five transcriptomes functional annotation. (A) The Venn diagram shows the distribution of Enzyme Commission (EC) numbers assigned across the translated products of the five *Mucor* transcriptomes. Shaded sections indicate categories with manual curation of the annotation. (B) Comparison of GO sub-categories (level 2) annotations across the translated products of the five *Mucor* transcriptomes. (Fisher exact, * p value < 0.05, ** p value < 0.005, *** p value < 0.0005). bp: biological process, cc: cellular component, mf: molecular function, MR: *Mucor racemosus*, MF: *M. fuscus*, ML: *M. lanceolatus*, ME: *M. endophyticus*, MC: *M. circinelloides*.

higher by at least twofold in MR and MC than in the other strains in an orthogroup associated with a multidrug resistance gene family. Among the singletons, predicted proteins with a predicted signal peptide were specifically investigated in order to identify putative secreted proteins involved in strain-specific substrate exploitation. However, no predicted proteins with signal peptide were identified in MF while only a few sets could be identified in the species representative strains (e.g. carbohydrate esterase and glycoside hydrolase in MC) (Fig. 2D). At least 48% of the singletons had EC numbers assignation and/or GO term annotation. Although most of the EC numbers were strains-specific, and mapping onto metabolic networks did not reveal any strain-specific pathways.

4.4. Investigation of specific gene families

The expression on PDA medium of (i) different gene families that could be involved in exploitation of the cheese substrate, (ii) virulence factors that could play an important role for *Mucor* opportunistic pathogens (iii) sugar transporters that may vary depending on the fungus lifestyle, and (iv) PKS and NRPS that are important for secondary metabolite production were analyzed in more detail.

MF and ML, two technological strains used for cheese ripening, did not harbor, cultivated on PDA medium, a richer repertoire of transcripts corresponding to genes potentially involved in the exploitation of the cheese substrate such as lactate permease and dehydrogenase and in aroma production through lipolysis and aminoacyl degradation such as lipases, decarboxylases and aminases. Indeed, unexpectedly more of these genes were transcripted in MC and MR.

Among the virulence factors identified in this study, three categories harbor different number of transcripts among strains: spore coat protein homologs (*cotH*), high affinity iron permease *FTR1* and iron transport multicopper ferroxidase *fet3* homologs.

The number of *cotH* transcripts detected was two time higher in MC and MR than in ML and MF transcriptomes (Table 3). A previously described motif involved in the invasin function of *cotH* was identified in three *Rhizopus delmar cotH* genes: *cotH2*, *cotH3* and the *cotH RO3G_15938*. One homolog to *cotH2/cotH3* was found in each *Mucor* transcriptome and one homolog of *RO3G_15938* was found in MC, MR and ME.

MC and MR clearly showed a higher number of transcripts identified as sugar transporters (STs) than the other strains. However, only half of these transcripts were clustered with experimentally characterized fungal transporters (Fig. 3). Three clusters of transcripts were expanded, all three in MR and MC compared to the other *Mucor* strains: Unknown STs group 3, D_galacturonic, quinic acid STs and Glucose, pentose STs.

Among secondary metabolism transcripts, two fragments of non-ribosomal peptides (*NRPs*) were identified in ME and two fragments of *PKs/FAS* were found in MC and MR. In both cases when fragments were concatenated, they formed a complete sequence. If we considered them as one gene, one NRPS and one *PKS/FAS* was found in all transcriptomes.

5. Discussion

The transcriptomes of five *Mucor* strains pertaining to five species within a broad phylogenetic range and reported to have different lifestyles were investigated in order to shed new light on *Mucor* gene expression on a standard medium and get information on specificities that could be linked to specific lifestyles within the *Mucor* genus. Indeed, *Mucor lanceolatus* (ML) and *Mucor fuscus* (MF) that form a clade within the tree reconstructed using 1289 orthogroups composed of single copy orthologs, were encountered in cheeses [39] and were reported to have an optimal growth on cheese [4] whereas *Mucor endophyticus* (ME),

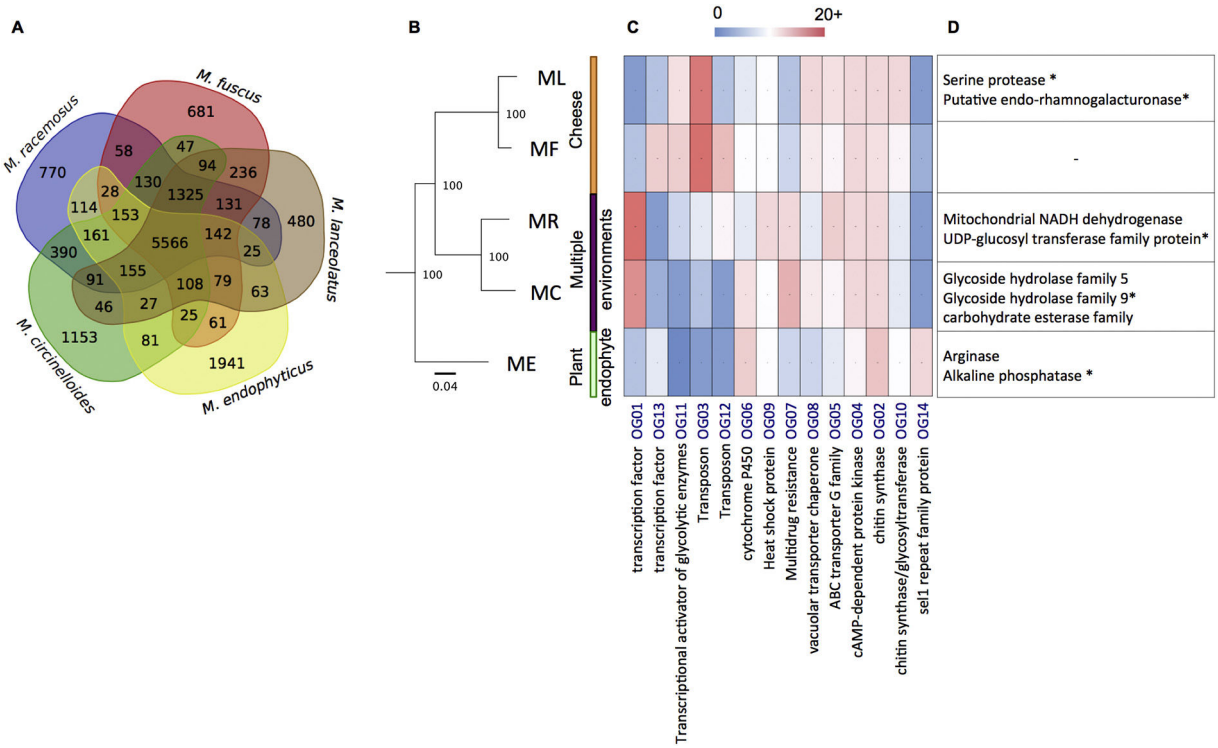


Fig. 2. (A) Distribution of orthogroups among strains. Hereafter, predicted proteins that could not be grouped are referred as singletons. (B) Phylogenetic species tree and reported lifestyle of the five studied *Mucor* species: *M. racemosus* (MR), *M. fuscus* (MF), *M. lanceolatus* (ML), *M. endophyticus* (ME) and *M. circinelloides* (MC). Branch length represents the substitution per site, numbers on node represent the bootstrap support. (C) The heat map represents orthogroups composed of > 10 proteins for at least one strain. For each orthogroup, the number of predicted proteins of each strain is represented by a colour gradient going from blue (no predicted protein) to red (> 20 predicted proteins). Only orthogroups with putative function assignment are displayed. (D) The table lists the function of singletons with signal peptides found in the five transcriptomes. Stars (*) indicate partial predicted proteins. Predicted proteins of unknown function are not shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Number of predicted proteins, potentially involved in adaptation to different lifestyles, across the translated products of the five *Mucor* transcriptomes. *Mucor racemosus* (MR), *Mucor fuscus* (MF), *Mucor lanceolatus* (ML), *Mucor endophyticus* (ME) and *Mucor circinelloides* (MC).

Target	Gene name	References	MC	MR	MF	ML	ME
Cheese exploitation	Lactate permease		2	3	2	2	2
	Lactate dehydrogenase		2	2	1	1	1
	Alkaline lipase	[57–59]	3	1	4	2	0
	Acidic lipase		3	2	1	1	1
	Decarboxylase		21	25	19	22	22
	Transaminase		26	27	25	23	26
	Deaminase		21	20	16	16	21
Virulence factors	*cotH2*, *cotH3*	[13,55]	1	1	1	1	1
	RO3G_15938		1	1	0	0	1
	all *cotH*		16	17	7	8	11
	ID112092 Mucor circinelloides	[60]	1	1	1	1	1
	fob1 (*RO3G_11000*), *fob2* (*RO3G_11000*)	[52,53]	1	1	1	1	1
	FTR1 (*RO3G_03470*)		1	1	1	1	2
	fet3a	[51]	0	0	0	0	1
	fet3b		1	1	1	1	0
	fet3c		1	1	1	1	0
Substrat exploitation	Sugar transporter		47	39	29	25	25
secondary metabolism	*NRPs*	[3]	1	1	1	1	2 fragments
	PKs/FAS		2 fragments	2 fragments	1	1	1

basal to the other species was only described as a wheat leaf endophyte [8]. *Mucor circinelloides* (MC) and *Mucor racemosus* (MR) forming a sister clade to the ML/MF clade are possibly more ubiquitous, MR being known as a recurrent cheese spoiler and MC as an opportunistic pathogen [9]. It is however worth to note that the two latter species include different forms which might display different ecological behaviors.

Some common traits were observed among the five transcriptomes. Global comparison of transcriptome functional annotations in these five species did not reveal noticeable differences among them in terms of enzyme composition nor putative secreted proteins in the medium suggesting that the sets of enzymes mobilized by *Mucor* species to grow on PDA medium were similar among the five species with no lost/altered nor gained pathways despite the different lifestyles and habitats reported for these species. Besides functions related to basal metabolism we noticed interesting traits among the core transcriptome. It is

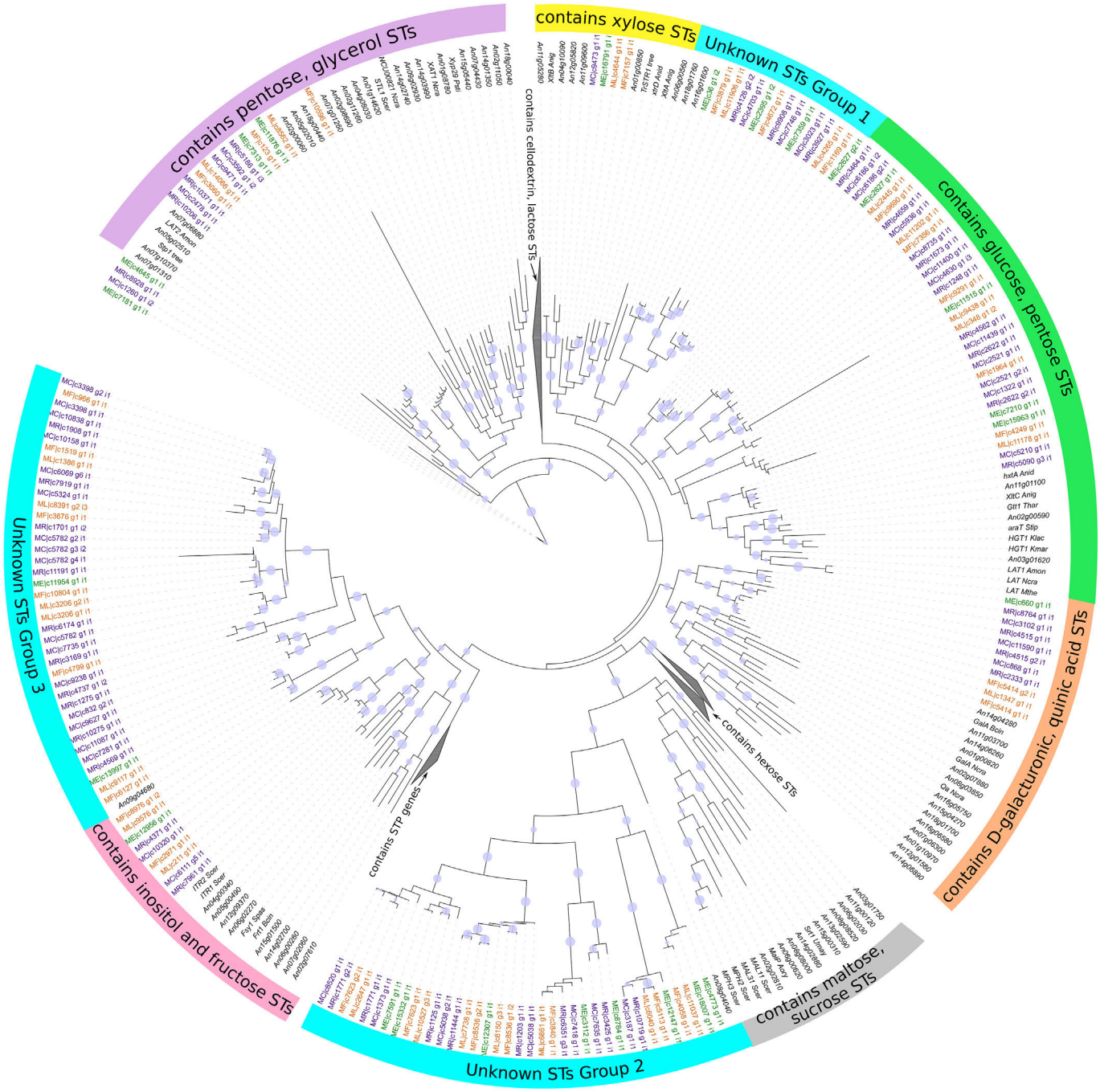


Fig. 3. Phylogenetic classification of putative sugar transporters (STs) in five *Mucor* transcriptomes. The tree includes 165 *Mucor* sugar transporters containing Pfam domain PF00083, 86 *Aspergillus niger* putative sugar transporters containing Pfam domain PF00083 and 61 experimentally characterized fungal transporters. Predicted protein of *Mucor* species found in multiple environments are highlighted with violet font, those found in the endophytic species in green and technological species in orange. Four clades were collapsed as they contain only references genes. Bars around the tree indicates the substrate of experimentally characterized fungal transporters within the cluster. Branches with bootstrap values above 50% are indicated by circles with larger circles representing higher bootstraps. *Mucor* genes begin with the species name tag: MC = *Mucor circinelloides*, MR = *Mucor racemosus,* MF = *Mucor fuscus*, ML = *Mucor lanceolatus* and ME = *Mucor endophyticus*. *An* represent *Aspergillus niger* putative sugar transporter genes. The abbreviation of reference fungal species name is attached to each known transporter gene (*Anid* = *Aspergillus nidulans*, *Anig* = *Aspergillus niger*, *Aory* = *Aspergillus oryzae*, *Amon* = *Ambrosiozyma monospora*, *Bcin* = *Botrytis cinerea*, *Kmar* = *Kluyveromyces marxianus*, *Klac* = *Kluyveromyces lactis*, *Ncra* = *Neurospora crassa*, *Psti* = *Pichia stipitis*, *Scer* = *Saccharomyces cerevisiae*, *Spas* = *Saccharomyces pastorianus*, *Stip* = *Scheffersomyces stipitis*, *Tree* = *Trichoderma reesei*, *Thar* = *Trichoderma harzianum*, umay = *Ustilago maydis*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

noteworthy that repeated elements seemed still active in *Mucor* genomes as shown by the orthogroups OG03 and OG12 that were identified as transposons and expressed in the five *Mucor* species, as well as GO term transposase activity that was differentially represented among species with more expressed genes related to this GO term found in MC which is a thermophilic opportunistic pathogen [38]. It is also worth to note that GO categories related to known factors associated to interactions with other organisms were found: “movement in the environment of other organism involved in symbiotic interaction”, “multi-organism metabolic process”, “host cell part and other organism

membrane". This raises the possibility that bacterial endosymbionts occur in these studied *Mucor* strains especially in the MF and ML strains where the above categories were overrepresented compared to the other strains studied. This has previously been shown in *Rhizopus microsporus*, another Mucorales, which harbors endosymbiotic bacteria [40,41].

Since secondary metabolites can provide a significant advantage for survival in a given ecological niche [42] and might vary depending on the fungus lifestyle, our study investigated transcripts corresponding to *PKS* and *NRPS*. Indeed, *PKs* and non-ribosomal peptides (*NRPs*) are involved in the majority of secondary metabolite biosynthesis (usually 40% and 15% respectively). Despite the different lifestyles reported for the five species studied here, no differences were found among the transcriptomes obtained from cultures on PDA medium. One transcript of *NRPS* was systematically found as well as one transcript corresponding to a gene annotated as a *PKS I* in the *M. circinelloides* genome [11] but which, as noticed by Voigt et al. [3], has the typical structure and domain order of a Fatty Acid Synthase (*FAS*) alpha subunit. *FAS* and *PKS I* share a common evolutionary history [43] and many homologies/similarities, such as the chemistry of catalyzed reactions (a sequence of simple unit condensation resulting in the synthesis of a molecule of higher molecular weight), or their enzymatic activity characteristics (condensation and modification of fatty acids or polyketides). Further studies are needed to determine to which production the *NRPS* could be associated. *Mucor* species are known to produce secondary metabolites such as pigments and terpenoids [3] and has been sometimes suspected to produce harmful toxins [44] but little information has already been obtained concerning production of other secondary metabolites in the *Mucor* genus.

This study did not demonstrate specificities linked to species encountered in cheese and that can be considered as technological species (ML and MF).

The more ubiquitous species MC and MR, which have been isolated from clinical environments [13] especially in the case of MC which is a thermotolerant species [9] involved in mucormycoses [9,45] had transcriptomes with an over-representation of "intrinsic component of membrane", "carbohydrate transporter activity", "xenobiotic transporter activity", "oxidoreductase activity" and "transcription factor activity" on the standard PDA medium. Among carbohydrate transporters, three sugar transporter families were expanded in MR and MC transcriptomes, corresponding to D-galacturonic and quinic acid STs, glucose and pentose STs and an unknown STs group 3. D-Galacturonic acid is the main component of pectin which is an important plant cell-wall polysaccharide. Since pectin is most abundant in the primary cell walls of soft and growing tissues, fruits and vegetables are particularly pectin-rich [46]. Quinic acid can also be extracted from plants sources. Possessing more diverse sugar transporters of this type may be an asset to a plant pathogenic lifestyle. The unknown STs group 3 was rare in ME (the endophyte species) transcriptome while it was expanded in MR and MC. Determining the substrate(s) of this transporter family might contribute to our understanding of how MR and MC are more ubiquist. The GO category xenobiotic transporter activity was over-represented in MR and MC. Furthermore, the orthogroup OG27 associated to a multidrug resistance gene family was also mainly composed of MC and MR transcripts. These genes might contribute to a better resistance to xenobiotics, including drugs, which could facilitate opportunistic infections of these two species which are known to be animal/human pathogens [5,38] and might contribute to the known problem with drug treatments against mucormycosis since pathogenic *Mucor* species are resistant to many classical antifungal products [47–50].

Among the strains studied in this work, the two strains reported as potential opportunistic human pathogens presented expression of all virulence factors checked whereas at least one of them was not detected in the technological and the endophytic transcriptomes.

Indeed, the MC ferroxidases *fet3b* and *fet3c* were absent from ME transcriptome. These genes along with *fet3a,* are overexpressed during infection in a mouse model for mucormycosis [51]. *fet3a* is specifically expressed during yeast growth under anaerobic conditions, whereas *fet3b* and *fet3c* are specifically expressed in mycelium during aerobic growth, *fet3c* being required for virulence during in vivo infections. *FTR1*, another gene involved in iron uptake and linked to virulence showed a different distribution among *Mucor* species. Two *FTR1* transcripts were detected in ME where only one was found in other species, ME display a higher pathogenic susceptibility concerning this aspect. In *R. oryzae* the reduction of *FTR1* copy number by gene disruption reduces the virulence of the fungus in animal models of mucormycosis [52,53]. The most important differences regarding the virulence factor transcripts were observed for spore coat protein homologs (*cotH*). During Human cell invasion by Mucorales, *cotH* genes allow the fungi to bind to glucose-regulated protein 78 (*GRP78*) which act as endothelial cell receptor [54], the *cotH* gene copy number of the species being correlated with its clinical prevalence [13]. Our results concur with this hypothesis as the number of *cotH* transcripts detected was two time higher in the transcriptomes of potentially pathogenic strains than in cheese technological strains. It is worth noting that *cotH* genes does not have the same impact on virulence. Indeed, *Rhizopus delmar cotH3* has higher affinity to *GRP78* than cotH1 leading to a reduce impact of cotH1 in virulence [55]. A motif corresponding to a surface-exposed region against which a therapeutic antibody has been raised was previously proposed [13]. Searching for *cotH* transcripts containing this motif allowed the identification of *cotH2* and *cotH3* known to be important for virulence [55] and the *cotH RO3G_15938* in *Rhizopus delmar* genes. According to this study the duplication of the ancestral gene leading to *cotH2* and *cotH3* happened after the separation of *Rhizopus* and *Mucor* clades. Each of the *Mucor* species used in this study expressed one ortholog of *cotH2/cotH3* gene. However, the ortholog of *cotH RO3G_15938,* that might be an asset for a pathogenic lifestyle, lacked in transcriptomes of MF/ML, the species used for cheese ripening.

6. Conclusion

The transcripts obtained from five different *Mucor* spp. cultivated on PDA allowed us to describe the predicted core proteome of a representative set of *Mucor* species with contrasting lifestyles. This analysis provided insight into *Mucor* characteristics by highlighting the presence of *NRPS* which imply a potential of *Mucor* for the production of secondary metabolites including pigments, siderophores, toxins [56]. It also provided hints as to how *Mucor* may adapt to different lifestyles, for example through expression of a larger set of sugar transporters and a comprehensive array of virulence factors in species that inhabit in multiple environments. On the other hand, species that are associated with cheese did not appear to have over-representation of set of transcripts involved in cheese media usage. Further studies using media mimicking cheese and animal and plant media might highlight more differences in transcripts expression associated to *Mucor* adaptation to a lifestyle.

Acknowledgments

The authors are thankful to Dr. Antoine Hermet for his involvement in the early steps of this project. We also thank Nantes BiogenOuest "Nantes Genomics Core Facility" for technological core facilities. This research was funded by the the Région Bretagne (ARED program) and EQUASA, a technological platform of the Université de Bretagne Occidentale, in the framework of the MUCORSCOPE project. The authors thank the Roscoff Bioinformatic platform: ABiMS (http://abims.sb-roscoff.fr) for providing computational resources. The authors are also grateful to the reviewers for their helpful comments.

References

[1] J.W. Spatafora, Y. Chang, G.L. Benny, K. Lazarus, M.E. Smith, M.L. Berbee, G. Bonito, N. Corradi, I. Grigoriev, A. Gryganskyi, T.Y. James, K. O'Donnell, R.W. Roberson, T.N. Taylor, J. Uehling, R. Vilgalys, M.M. White, J.E. Stajich, A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data, Mycologia 108 (2016) 1028–1046.

[2] S. Morin-Sardin, J.L. Jany, S. Artigaud, V. Pichereau, B. Bernay, E. Coton, S. Madec, Proteomic analysis of the adaptative response of Mucor spp. to cheese environment, J. Proteome 154 (2017) 30–39.

[3] K. Voigt, T. Wolf, K. Ochsenreiter, G. Nagy, K. Kaerger, E. Shelest, T. Papp, D. Hoffmeister (Ed.), Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes, Springer International Publishing, Cham, 2016, pp. 361–385.

[4] S. Morin-Sardin, K. Rigalma, L. Coroller, J.L. Jany, E. Coton, Effect of temperature, pH, and water activity on Mucor spp. growth on synthetic medium, cheese analog and cheese, Food Microbiol. 56 (2016) 69–79.

[5] K. Hoffmann, J. Pawłowska, G. Walther, M. Wrzosek, G.S. de Hoog, G.L. Benny, P.M. Kirk, K. Voigt, The family structure of the *Mucorales*: a synoptic revision based on comprehensive multigene-genealogies, Persoonia 30 (2013) 57–76.

[6] G. Walther, J. Pawlowska, A. Alastruey-Izquierdo, M. Wrzosek, J.L. Rodriguez-Tudela, S. Dolatabadi, A. Chakrabarti, G.S. de Hoog, DNA barcoding in Mucorales: an inventory of biodiversity, Persoonia 30 (2013) 11–47.

[7] L. Mendoza, R. Vilela, K. Voelz, A.S. Ibrahim, K. Voigt, S.C. Lee, Human Fungal Pathogens of Mucorales and Entomophthorales, (2015).

[8] R. Zheng, H. Jiang, Rhizomucor endophyticus sp.nov., an endophytic zygomycetes from higher plants, Mycotaxon 56 (1995) 455–466.

[9] S. Morin-Sardin, P. Nodet, E. Coton, J.-L. Jany, Mucor: a Janus-faced fungal genus with human health impact and industrial applications, Fungal Biol. Rev. 31 (2017) 12–32.

[10] J.W. Spatafora, M.C. Aime, I.V. Grigoriev, F. Martin, J.E. Stajich, M. Blackwell, The Fungal Tree of Life: From Molecular Systematics to Genome-Scale Phylogenies, Microbiology Spectrum, (2017), p. 5.

[11] L.M. Corrochano, A. Kuo, M. Marcet-Houben, S. Polaino, A. Salamov, J.M. Villalobos-Escobedo, J. Grimwood, M.I. Álvarez, J. Avalos, D. Bauer, E.P. Benito, I. Benoit, G. Burger, L.P. Camino, D. Cánovas, E. Cerdá-Olmedo, J.-F. Cheng, A. Domínguez, M. Eliáš, A.P. Eslava, F. Glaser, G. Gutiérrez, J. Heitman, B. Henrissat, E.A. Iturriaga, B.F. Lang, J.L. Lavín, S.C. Lee, W. Li, E. Lindquist, S. López-García, E.M. Luque, A.T. Marcos, J. Martin, K. McCluskey, H.R. Medina, A. Miralles-Durán, A. Miyazaki, E. Muñoz-Torres, J.A. Oguiza, R.A. Ohm, M. Orejas, L. Ortiz-Castellanos, A.G. Pisabarro, J. Rodríguez-Romero, J. Ruiz-Herrera, R. Ruiz-Vázquez, C. Sanz, W. Schackwitz, M. Shahriari, E. Shelest, F. Silva-Franco, D. Soanes, K. Syed, V.G. Tagua, N.J. Talbot, M.R. Thon, H. Tice, R.P. de Vries, A. Wiebenga, J.S. Yadav, E.L. Braun, S.E. Baker, V. Garre, J. Schmutz, B.A. Horwitz, S. Torres-Martínez, A. Idnurm, A. Herrera-Estrella, T. Gabaldón, I.V. Grigoriev, Expansion of signal transduction pathways in fungi by extensive genome duplication, Curr. Biol. 26 (2016) 1577–1584.

[12] X. Tang, H. Chen, Y.Q. Chen, W. Chen, V. Garre, Y. Song, C. Ratledge, Comparison of biochemical activities between high and low lipid-producing strains of *Mucor circinelloides*: an explanation for the high oleaginicity of strain WJ11, PLoS One 10 (2015) e0128396.

[13] M.C. Chibucos, S. Soliman, T. Gebremariam, H. Lee, S. Daugherty, J. Orvis, A.C. Shetty, J. Crabtree, T.H. Hazen, K.A. Etienne, P. Kumari, T.D. O'Connor, D.A. Rasko, S.G. Filler, C.M. Fraser, S.R. Lockhart, C.D. Skory, A.S. Ibrahim, V.M. Bruno, An integrated genomic and transcriptomic survey of mucormycosis-causing fungi, Nat. Commun. 7 (2016) 1–11.

[14] S. Andrews, FASTQC: A Quality Control Tool for High Throughput Sequence Data, (2010).

[15] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (2011) 644–652.

[16] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinform. 12 (2011) 323.

[17] K. Lagesen, P. Hallin, E.A. Rodland, H.H. Staerfeldt, T. Rognes, D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, Nucleic Acids Res. 35 (2007) 3100–3108.

[18] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.

[19] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, S.J. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer, A. Bateman, The Pfam protein families database, Nucleic Acids Res. 36 (2008) D281–D288.

[20] a. B. Krogh, Larsson, G. Von Heijne, E., Sonnhammer predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.

[21] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods 8 (2011) 785.

[22] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[23] L.J. Ma, A.S. Ibrahim, C. Skory, M.G. Grabherr, G. Burger, M. Butler, M. Elias, A. Idnurm, B.F. Lang, T. Sone, A. Abe, S.E. Calvo, L.M. Corrochano, R. Engels, J. Fu, W. Hansberg, J.M. Kim, C.D. Kodira, M.J. Koehrsen, B. Liu, D. Miranda-Saavedra, S. O'Leary, L. Ortiz-Castellanos, R. Poulter, J. Rodriguez-Romero, J. Ruiz-Herrera, Y.Q. Shen, Q. Zeng, J. Galagan, B.W. Birren, C.A. Cuomo, B.L. Wickes, Genomic analysis of the basal lineage fungus Rhizopus oryzae reveals a whole-genome duplication, PLoS Genet 5 (2009) e1000549.

[24] M.A. Harris, Developing an ontology, Methods Mol. Biol. 452 (2008) 111–124.

[25] A.G. McDonald, K.F. Tipton, Fifty-five years of enzyme classification: advances and difficulties, FEBS J. 281 (2014) 583–592.

[26] C. Claudel-Renard, C. Chevalet, T. Faraut, D. Kahn, Enzyme-specific profiles for genome annotation: PRIAM, Nucleic Acids Res. 31 (2003) 6633–6639.

[27] F.A. Simao, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212.

[28] T. Yamada, I. Letunic, S. Okuda, M. Kanehisa, P. Bork, iPath2.0: interactive pathway explorer, Nucleic Acids Res. 39 (2011) W412–W415.

[29] D.M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, Genome Biol. 16 (2015) 157.

[30] A. Loytynoja, Phylogeny-aware alignment with PRANK, Methods Mol. Biol. 1079 (2014) 155–170.

[31] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (2009) 1972–1973.

[32] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[33] N. Lartillot, T. Lepage, S. Blanquart, PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating, Bioinformatics 25 (2009) 2286–2288.

[34] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2014) 59.

[35] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C.J. Lanczycki, S. Lu, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, L.Y. Geer, S.H. Bryant, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures, Nucleic Acids Res. 45 (2017) D200–D203.

[36] M. Peng, M.V. Aguilar-Pontes, R.P. de Vries, M.R. Mäkelä, In silico analysis of putative sugar transporter genes in *Aspergillus niger* using phylogeny and comparative transcriptomics, Front. Microbiol. 9 (2018).

[37] I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees, Nucleic Acids Res. 44 (2016) W242–W245.

[38] G. Walther, J. Pawłowska, A. Alastruey-Izquierdo, M. Wrzosek, J.L. Rodriguez-Tudela, S. Dolatabadi, A. Chakrabarti, G.S. de Hoog, DNA barcoding in < I > Mucorales < /I > : an inventory of biodiversity, Persoonia 30 (2013) 11–47.

[39] A. Hermet, D. Meheust, J. Mounier, G. Barbier, J.L. Jany, Molecular systematics in the genus Mucor with special regards to species encountered in cheese, Fungal Biol. 116 (2012) 692–705.

[40] S.J. Mondo, O.A. Lastovetsky, M.L. Gaspar, N.H. Schwardt, C.C. Barber, R. Riley, H. Sun, I.V. Grigoriev, T.E. Pawlowska, Bacterial endosymbionts influence host sexuality and reveal reproductive genes of early divergent fungi, Nat. Commun. 8 (2017) 1843.

[41] L.P. Partida-Martinez, I. Groth, I. Schmitt, W. Richter, M. Roth, C. Hertweck, Burkholderia rhizoxinica sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus Rhizopus microsporus, Int. J. Syst. Evol. Microbiol. 57 (2007) 2583–2590.

[42] E.M. Fox, B.J. Howlett, Secondary metabolism: regulation and role in fungal biology, Curr. Opin. Microbiol. 11 (2008) 481–487.

[43] H. Jenke-Kodama, A. Sandmann, R. Muller, E. Dittmann, Evolutionary implications of bacterial polyketide synthases, Mol. Biol. Evol. 22 (2005) 2027–2039.

[44] S.C. Lee, R.B. Billmyre, A. Li, S. Carson, S.M. Sykes, E.Y. Huh, P. Mieczkowski, D.C. Ko, C.A. Cuomo, J. Heitman, Analysis of a food-borne fungal pathogen outbreak: virulence and genome of a *Mucor circinelloides* Isolate from yogurt, mBio 5 (2014) e01390–e01391.

[45] S.C. Lee, A. Li, S. Calo, J. Heitman, Calcineurin plays key roles in the dimorphic transition and virulence of the human pathogenic zygomycete *Mucor circinelloides*, PLoS Pathog. 9 (2013) e1003625.

[46] J.P. Benz, R.J. Protzko, J.M. Andrich, S. Bauer, J.E. Dueber, C.R. Somerville, Identification and characterization of a galacturonic acid transporter from *Neurospora crassa* and its application for Saccharomyces cerevisiae fermentation processes, Biotechnol. Biofuels 7 (2014) 20.

[47] R.E. Lewis, O. Lortholary, B. Spellberg, E. Roilides, D.P. Kontoyiannis, T.J. Walsh, How does antifungal pharmacology differ for *Mucormycosis* versus *Aspergillosis*? Clin. Infect. Dis. 54 (2012) S67–S72.

[48] A. Muszewska, J. Pawłowska, P. Krzyściak, Biology, systematics, and clinical manifestations of Zygomycota infections, Eur. J. Clin. Microbiol. Infect. Dis. 33 (2014) 1273–1287.

[49] T.T. Riley, C.A. Muzny, E. Swiatlo, D.P. Legendre, Breaking the mold: a review of *Mucormycosis* and current pharmacological treatment options, Ann. Pharmacother. 50 (2016) 747–757.

[50] A. Skiada, L. Pagano, A. Groll, S. Zimmerli, B. Dupont, K. Lagrou, C. Lass-Florl, E. Bouza, N. Klimko, P. Gaustad, M. Richardson, P. Hamal, M. Akova, J.F. Meis, J.L. Rodriguez-Tudela, E. Roilides, A. Mitrousia-Ziouva, G. Petrikkos, Zygomycosis in Europe: analysis of 230 cases accrued by the registry of the European Confederation of Medical Mycology (ECMM) Working Group on Zygomycosis between 2005 and 2007, Clin. Microbiol. Infect. 17 (2011) 1859–1867.

[51] M.I. Navarro-Mendoza, C. Pérez-Arques, L. Murcia, P. Martínez-García, C. Lax, M. Sanchis, J. Capilla, F.E. Nicolás, V. Garre, Components of a new gene family of ferroxidases involved in virulence are functionally specialized in fungal

dimorphism, Sci. Rep. 8 (2018) 7660.
[52] A.S. Ibrahim, B. Spellberg, T.J. Walsh, D.P. Kontoyiannis, Pathogenesis of Mucormycosis, Clin. Infect. Dis. 54 (2012) S16–S22.
[53] M. Liu, L. Lin, T. Gebremariam, G. Luo, C.D. Skory, S.W. French, T.-F. Chou, J.E. Edwards Jr., A.S. Ibrahim, Fob1 and Fob2 proteins are virulence determinants of rhizopus oryzae via facilitating iron uptake from ferrioxamine, PLoS Pathog. 11 (2015) e1004842.
[54] M. Liu, B. Spellberg, Q.T. Phan, Y. Fu, Y. Fu, A.S. Lee, J.E. Edwards Jr., S.G. Filler, A.S. Ibrahim, The endothelial cell receptor GRP78 is required for mucormycosis pathogenesis in diabetic mice, J. Clin. Invest. 120 (2010) 1914–1924.
[55] T. Gebremariam, M. Liu, G. Luo, V. Bruno, Q.T. Phan, A.J. Waring, J.E. Edwards Jr., S.G. Filler, M.R. Yeaman, A.S. Ibrahim, CotH3 mediates fungal invasion of host cells during mucormycosis, J. Clin. Invest. 124 (2014) 237–250.
[56] P.J. Bhetariya, M. Prajapati, A. Bhaduri, R.S. Mandal, A. Varma, T. Madan, Y. Singh, P.U. Sarma, Phylogenetic and structural analysis of polyketide synthases in *Aspergilli*, Evol. Bioinformatics Online 12 (2016) 109–119.
[57] J. Cerning, J. Gripon, G. Lamberet, J. Lenoir, Les activités biochimiques des Penicillium utilisés en fromagerie, Lait 67 (1987) 3–39.
[58] P.L.H. McSweeney, Chapter 14 - Biochemistry of cheese ripening: introduction and overview, in: P.L.H. McSweeney, P.F. Fox, P.D. Cotter, D.W. Everett (Eds.), Cheese, Fourth Edition, Academic Press, San Diego, 2017, pp. 379–387.
[59] P.L.H. McSweeney, P.F. Fox, F. Ciocia, Chapter 16 - Metabolism of residual lactose and of lactate and citrate, in: P.L.H. McSweeney, P.F. Fox, P.D. Cotter, D.W. Everett (Eds.), Cheese, Fourth Edition, Academic Press, San Diego, 2017, pp. 411–421.
[60] L. López-Fernández, M. Sanchis, P. Navarro-Rodríguez, F.E. Nicolás, F. Silva-Franco, J. Guarro, V. Garre, M.I. Navarro-Mendoza, C. Pérez-Arques, J. Capilla, Understanding Mucor circinelloides pathogenesis by comparative genomics and phenotypical studies, Virulence 9 (2018) 707–720.

Des analyses complémentaires à celles de l'article ont été réalisées et sont présentées dans les parties suivantes :

Partie 3.3 : Un point de méthodologie complémentaire concernant la comparaison des *GO terms* sera abordée. En effet, dans l'article est uniquement indiqué qu'un script interne a été utilisé pour ces comparaisons, cette partie a pour objet d'expliquer les défis liés à la comparaison des *GO terms* et présenter les choix réalisés lors de la conception de ce script.

Partie 3.4 : Dans l'article a été évoqué la possibilité que des bactéries endosymbiotes soient présentes dans les souches de *Mucor* étudiées, or au cours de l'étude des transcriptomes, la vérification de la qualité des données RNAseq de *M. endophyticus* indiquait la présence d'une contamination ou d'un autre type de biais d'échantillonnage. Les analyses présentées en informations complémentaires retracent les recherches liées à ce biais d'échantillonnage.

Partie 3.5 : Lors de la comparaison des transcriptomes, un axe de recherche portait sur la comparaison des enzymes prédites. Dans l'article, cet axe n'a pas été développé, le détail des recherches associées est présenté dans cette section.

Partie 3.6 : De même, les protéines prédites propres à chacune des espèces ont été décrites et des comparaisons des protéines prédites partagées uniquement par les espèces au même mode de vie ont été réalisées. Uniquement évoquées dans l'article ces analyses seront présentées dans cette partie.

3.3 Méthodologie supplémentaire : comparaison des *GO terms*

Comme indiqué dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2), les différences significatives d'occurrences de catégories GO entre espèces ont été identifiées grâce à l'utilisation d'un script basé sur l'application d'un test de Fischer. Le contenu de ce script est explicité ci-après.

Le projet *Gene Ontology* est destiné à structurer la description des gènes et produits géniques. Ceci est réalisé grâce à un vocabulaire contrôlé : des propriétés/descriptions/concepts (*GO terms*) et des relations entre ces concepts (est/régule/etc.). Une représentation de la structure obtenue correspond à un graphe orienté acyclique. Trois principaux domaines sont représentés : *Biological Process*, *Molecular Function* et *Cellular Component*. A gauche dans la figure 3.4 sont présentés les éléments qui relient le GO associé à *apoptotic process* à la racine du graphe *Biological*

Process, ces éléments sont également appelés ancêtre du GO associé à *apoptotic process* ; plus on se rapproche de la racine du graphe, moins les termes sont précis.

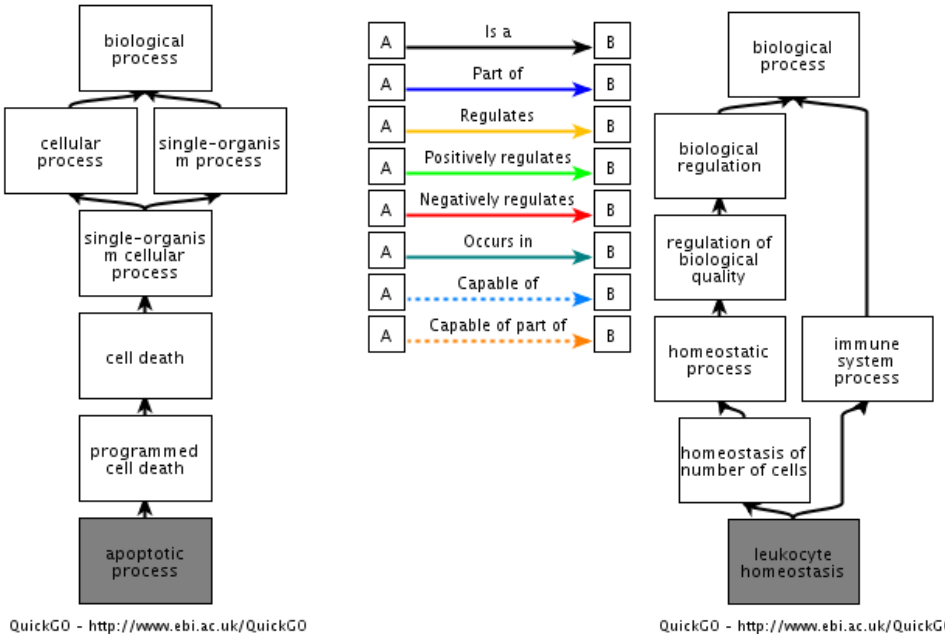


FIGURE 3.4 – Éléments qui relient la racine du graphe *Biological Process* à, à gauche *apoptotic process*, à droite *leukocyte homeostasis*

A l'échelle d'un transcriptome, le nombre de *GO terms* différents était très important (presque 10000 pour *M. racemosus*), il était donc impossible d'avoir une vue d'ensemble pour les cinq espèces en les conservant tous. La précision de chaque *GO term* a donc été diminuée en remplaçant chacun des termes par un de ses ancêtres (ex. *cell death* en lieu et place d'*apoptotic process* (Figure 3.4)). L'ancêtre est choisi à une profondeur fixe dans le graphe en partant de la racine, la profondeur pouvant être déterminée comme étant le nombre de relations nécessaire pour passer d'un GO à un autre. Plusieurs "chemins" de relations sont souvent possibles ce qui fait qu'un GO peut avoir plusieurs ancêtres à une même profondeur (ex : *cellular process* et *single-organism process* sont à 1 relation de *Biological Process* et tous deux sont des ancêtres de *apoptotic process*). De même, à une même profondeur, la précision du terme peut être très différente du point de vue biologique, par exemple à deux relations de *Biological Process*, on peut trouver *leukocyte homeostasis* et *single organism process* (Figure 3.4).

Les outils existants ne permettaient pas de visualiser graphiquement l'occurence de groupes de *GO terms* pour cinq espèces sans apport d'informations de type expression des transcrits ni réplicats pour les analyses statistiques (Tableau 3.1).

Un script personnel a donc été réalisé permettant de retrouver l'ancêtre de chaque *GO term*

Tableau 3.1 – Examples d'outils permettant de traiter les *GO terms* (avril 2017).

Outil	Accès	Type	Commentaire
WEGO	wego.genomics.org.cn	web	Maximum 3 espèces ; base de donnée GO datant de 2009.
DAVID	david.ncifcrf.gov	web	Ne prends pas les *GO terms* tel quel mais des identifiants d'espèces de référence
REVIGO	revigo.irb.hr	web	Ne permet pas de comparer les espèces entre elles
Enigma	bioinformatics.psb.ugent.be	web	S'intéresse d'abord l'expression des transcrits puis incorpore les *GO terms*
topGO	package bioconductor	R package	Nécessite un score associé aux transcrits (suite à une analyse d'expression différentielle par exemple) ; réalise une analyse d'enrichissement
GOstat	package bioconductor	R package	Groupe de fonctions permettant de manipuler les *GO terms* ; base pour d'autres outils
FatiGO	fatigoplus.org babelomics.org	web/ Babelomics environnement	Comparaison de deux listes d'éléments maximum.

à une profondeur donnée puis de compter le nombre de fois où cet ancêtre était retrouvé. Pour ce script, différents choix ont été faits :

— Tous les *GO terms* à la profondeur donnée sont pris en compte (pas de vérification qu'un chemin plus long puisse permettre d'accéder à la racine comme le cas de *leukocyte homeostasis* en Figure 3.4).

— Si deux ancêtres sont à la profondeur demandée pour un même GO term, les deux ancêtres sont comptés.

— Pour un même transcrit, si plusieurs *GO terms* ont le même ancêtre, cet ancêtre, est compté une fois par transcrit.

A l'échelle des catégories principales (*Biological Process*, *Molecular Function* et *Cellular Component*) des différences importantes du nombre de *GO terms* associés à chaque espèce étaient visibles (Figure 3.5), la référence pour chacun des tests ne pouvait donc pas être une moyenne du nombre de *GO terms* obtenu pour les cinq espèces.

Un test de Fisher exact a été choisi pour tester l'hypothèse H0 : pas de différences de proportion entre la distribution des GO chez les espèces à l'échelle d'une catégorie principale (ex. *Molecular Function*) et la sous-catégorie observée qui appartient à cette catégorie principale (ex : *oxydoreductase activity*). L'hypothèse nulle était rejetée si la p-value était inférieure à 0.05. Le test

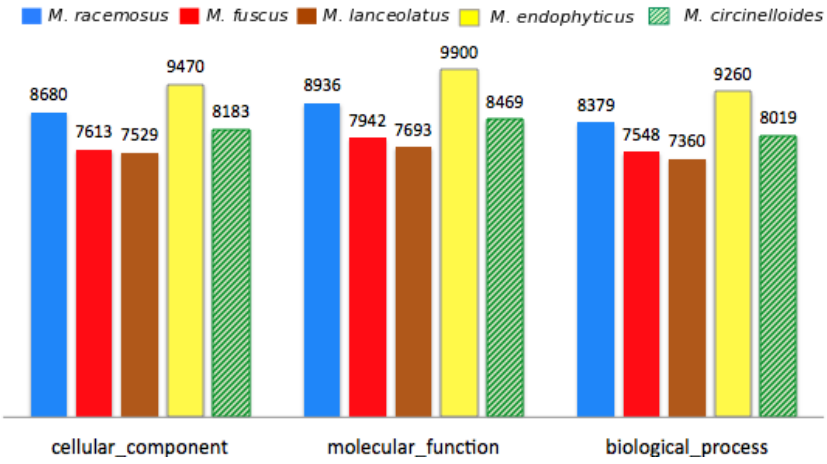


FIGURE 3.5 – Répartition du nombre de *GO terms* des trois principales catégories chez les espèces étudiées.

de Fisher exact a été préféré au Khi2 car ce dernier était inadapté lors de comptages théoriques inférieurs à 5, cas fréquement rencontré lors de la comparaison des annotations GO des protéines propres à chacune des espèces (Tableau 3.2).

TABLEAU 3.2 – Proportion de *GO terms* pour lesquels le Khi2 est applicable. Profondeur 3 relations.

Categorie de *GO terms*	Nombre de *GO terms*	Nombre de *GO terms* pour lesquel le Khi2 était **possible**	Nombre de *GO terms* significativement différents	Nombre de *GO terms* pour lesquel le Khi2 était **impossible**	Différence d'au moins 10 occurrences entre espèces	Différence d'au moins 20 occurrences entre espèces
Biological Process	513	55	4	458	62	15
Cellular Component	343	23	3	320	23	7
Molecular Function	219	22	4	197	38	11
Toute catégories	1075	100	11	975	123	33

Aucune correction associée aux tests multiples n'a été réalisée car le très grand nombre de tests aurait rendu l'obtention des différences significatives impossible du fait de la stringence du test.

Ce script a été utilisé pour les comparaisons de *GO terms* de la publication sur les transcriptomes et a été déposé sur github (`https://github.com/anlebreton/GO_terms_FisherTest.git`) afin d'être accessible à la communauté.

3.4 Recherche de l'origine du biais d'échantillonage des données RNAseq de *M. endophyticus*

3.4.1 Introduction

L'analyse de la qualité des données de séquençage est une étape préliminaire indispensable à toute analyse RNAseq, en particulier la distribution des *reads* en fonction de leurs pourcentage en GC est un critère important de contrôle qualité (Figure 3.6).

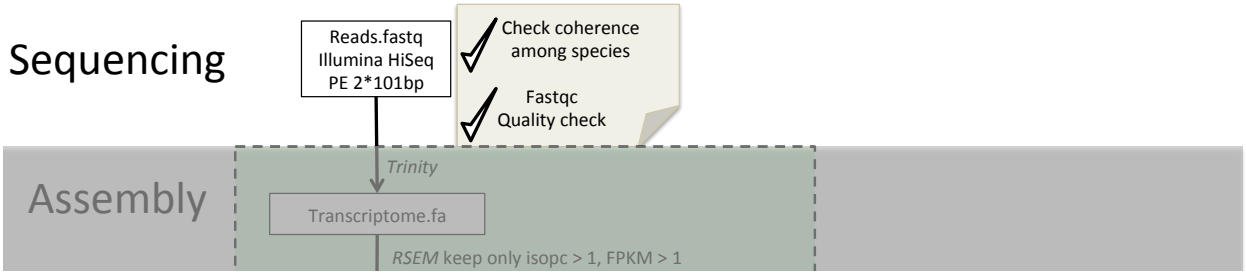


FIGURE 3.6 – Premières étapes du pipeline d'assemblage et annotation des transcriptomes.

Chez *M. endophyticus*, cette distribution s'est révélée trimodale (Figure 3.7) or lorsque l'on s'intéresse au génome ou transcriptome d'une seule espèce, cette distribution est généralement normale (ou bimodale chez les graminées (Clément et al., 2015)). Ce type de profil peut indiquer la présence d'une contamination ou un biais d'échantillonnage (Andrews, 2010).

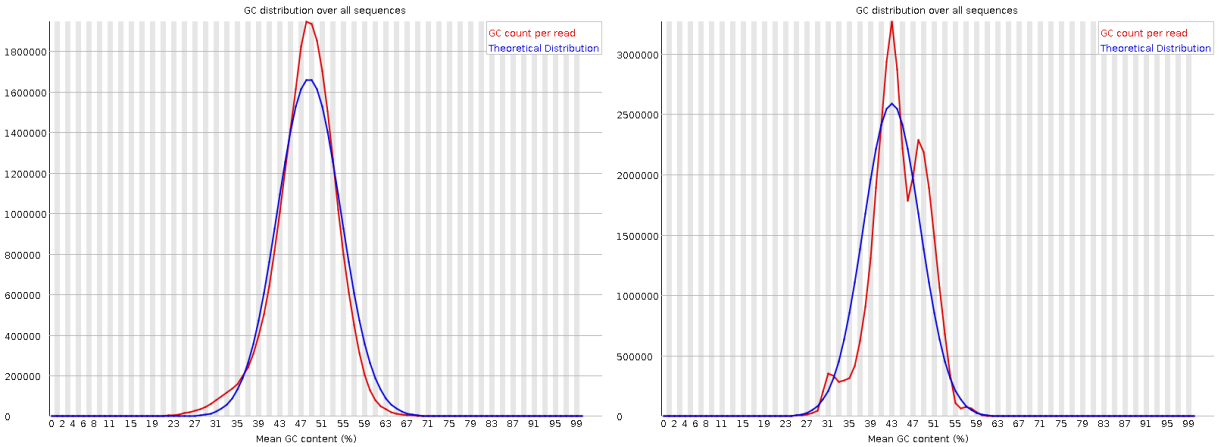


FIGURE 3.7 – Distribution des *reads* en fonction de leurs pourcentage en GC observée à gauche chez *M. circinelloides* et à droite chez *M. endophyticus*. En bleu : distribution théorique. En rouge : distribution observée.

L'origine de cette distribution particulière a été recherchée afin de déterminer si plusieurs types d'organismes avaient été séquencés ou si ce profil était lié à une particularité de *M. endophyticus*.

3.4.2 Matériels et méthodes

Identification des ARNs séquencés

Afin de déterminer si plusieurs types d'organismes avaient été séquencés au sein de la même librairie, les ARN ribosomiques (ARNr) des transcrits reconstruits *de novo* ont été identifiés avec RNAmmer (Lagesen et al., 2007) dans les cinq transcriptomes.

Une identification taxonomique a ensuite été réalisée avec MEGAN v6.6.7 (Huson et al., 2011) sur (i) tous les transcrits reconstruits *de novo*, (ii) tous les *reads* et (iii) sur trois groupes de *reads* de *M. endophyticus* séparés en fonction de leur pourcentage en GC (groupe 1, *reads* contenant moins de 37%GC (pic1 sur la figure 3.7), groupe 2, *reads* contenant entre 37% et 46%GC (pic2 sur la figure 3.7) et groupe 3, *reads* contenant plus de 46%GC (pic3 sur la figure 3.7)) .

La présence d'ARNr bactériens a également été recherchée dans l'ADN de *M. endophyticus* et *M. lanceolatus* par PCR en utilisant des amorces propres aux ARNr 16S (amorce 16Sfwd 5'-ccgaattcgtcgacaacagagtttgatcc-3' , amorce 16Srev 5'-cccgggatccaagcttacggctaccttgt-3'). Les régions amplifiées ont été séquencées puis identifiées par homologie de séquence.

Recherche du déterminisme du profil de distribution des *reads*

Afin d'estimer l'incidence respective de la présence d'ARNr séquencés et de duplications, deux tests ont été réalisés.

En premier lieu, deux groupes de *reads* ont été constitués à partir du jeu de données complet de *M. endophyticus* : l'un constitué des *reads* potentiellement issus des ARNr, identifiés en utilisant ribopicker (Schmieder et al., 2012), l'autre contenant tous les autres *reads*, avec pour but d'estimer l'incidence de la présence des ARNr dans le taux de GC des *reads*.

En second lieu, les *reads* ont été dédupliqués (lorsque plusieurs *reads* possédaient la même séquence, un seul a été conservé) pour chacun des groupes précédemment réalisés afin d'estimer l'importance des duplications dans le profil trimodal obtenu.

3.4.3 Résultats

Des ARNr potentiellement bactériens ont été identifiés dans les cinq transcriptomes, cependant les assignations taxonomiques de ces transcrits différaient d'un transcriptome à l'autre : *Escherichia coli* chez *M. endophyticus*, *Janthinobacterium spp.* chez *M. racemosus*, *M. fuscus* et *M. lanceolatus*, *Lactobacillus spp.* chez *M. lanceolatus* et une bactérie non cultivée (*Uncultured bacterium*) chez *M. circinelloides*.

Identification des ARNs séquencés

Lorsque l'identification est réalisée sur l'ensemble des transcrits, la moitié des transcrits ayant une correspondance avec les bases de données ont été assignés aux *Mucor*, l'autre moitié ont été considérés par l'outil comme étant de faible complexité (composé de nucléotides répétés ou de court motifs nucléotidiques répétés (Toll-Riera et al., 2012) (Figure 3.8)).

Lorsque l'identification est menée sur l'ensemble des *reads*, la majorité des *reads* possédant une assignation taxonomique correspondaient à des *Mucor*, les autres ne se regroupaient pas dans un taxon particulier. Le même résultat a été obtenu lorsque l'étude a été réalisée sur les groupes de *reads* correspondant aux différents pics de la distribution du pourcentage GC des données de *M. endophyticus*.

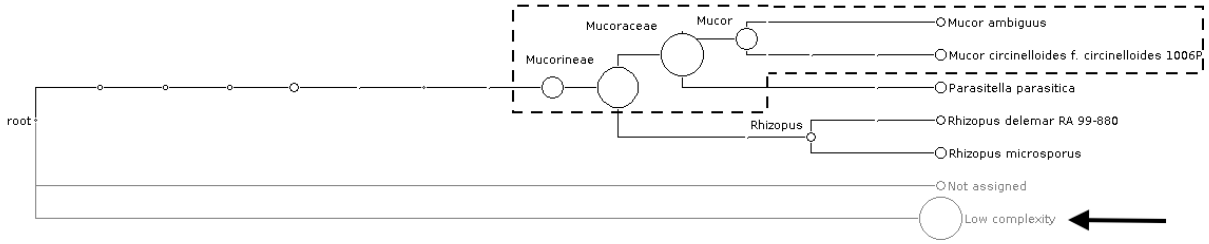


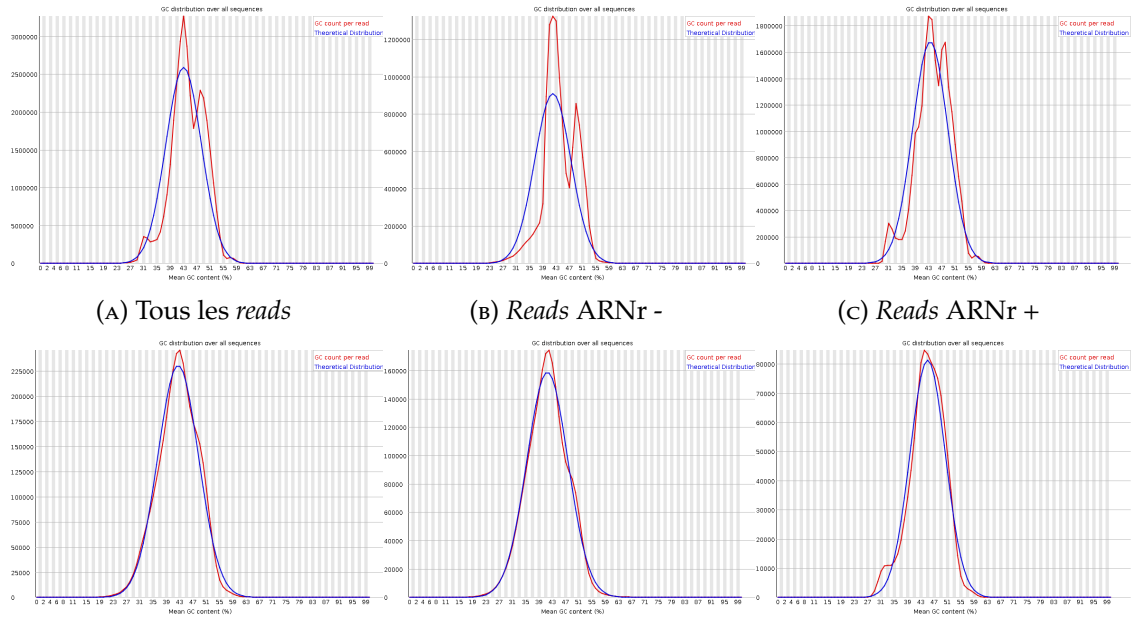
FIGURE 3.8 – Assignation taxonomique des transcrits reconstruits *de novo* à partir des données RNAseq de *M. endophyticus* ; plus le cercle est grand, plus le nombre de transcrits assignés au taxon est important. Les taxons correspondant à des espèces de *Mucor* sont encadrés, le groupe pour lequel les transcrits n'ont pu être assignés à cause d'une trop faible complexité de séquence est fléché.

De potentiels ADN ribosomiques (ADNr) bactériens ont été recherchés chez *M. endophyticus* et *M. lanceolatus* par PCR. Dans les deux cas, un amplicon a été obtenu mais la région amplifiée, plus courte que la taille attendue pour un amplicon d'ADNr 16S, présentait l'identité la plus forte avec les séquences d'ARNr 18S de *Mucor* retrouvées dans les bases de données (amplicon de *M. lanceolatus* 98% d'identité, e-value 0, avec l'ADNr 18S de *M. lanceolatus*, amplicon de *M.*

endophyticus 100% d'identité, e-value 0, avec l'ADNr 18S de *M. endophyticus* et *M. hiemalis*). De façon inattendue, les amorces utilisées ont donc pu s'hybrider sur deux régions de l'ADNr 18S des deux espèces de *Mucor*. Aucun amplicon bactérien n'a pu être obtenu en utilisant la méthode de PCR "classique".

Recherche du déterminisme du profil de distribution des *reads*

Afin d'avoir une indication de l'incidence de la présence des ARNr sur la distribution GC des *reads* de *M. endophyticus*, les *reads* potentiellement issus des ARNr (ARNr +) ont été séparés des autres *reads* (ARNr -). Le groupe ARNr - présentait une distribution bimodale avec un épaulement en lieu et place du premier pic (Figure 3.9B). Le groupe ARNr + présentait quant à lui une distribution trimodale (Figure 3.9C). Les duplications correspondent à l'origine principale de la distribution observée (Figure 3.9D), en effet, lorsque les *reads* sont dédupliqués seul le groupe de *reads* ARNr + conserve deux pics (à 31% et 43%GC, Figure 3.9F), le groupe ARN - ne conserve quant à lui que le pic principal (43%GC) ainsi qu'un épaulement en lieu et place du troisième pic (Figure 3.9E). Ces résultats indiquent que le premier pic observé (à 31%GC) est lié à des *reads* d'ARNr relativement diversifiés, le second pic (à 43%GC) est lié à des *reads* diversifiés d'ARNr ou non, le troisième pic (à 53%GC) est lié à des *reads* dupliqués d'ARNr ou non.



(A) Tous les *reads* (B) *Reads* ARNr - (C) *Reads* ARNr +

(D) *Reads* dédupliqués (E) *Reads* ARNr - dédupliqués (F) *Reads* ARNr + dédupliqués

FIGURE 3.9 – Distribution des *reads* en fonction de leurs pourcentage en GC observés sur les différents groupes de *reads* traités. En rouge : distribution observée. En bleu : distribution théorique. ARNr - : groupe de *reads* appauvris en *reads* d'origine ribosomique. ARNr + : groupe composé de *reads* identifiés comme étant d'origine ribosomique.

Les transcrits non ribosomiques participant à la formation du troisième pic ont été recherchés. Ces transcrits étant fortement exprimés, la répartition des FPKM [1] des transcrits a été observée (Tableau 3.3). La majorité des transcrits avaient un FPKM compris entre 1 et 100, sept transcrits avaient un FPKM de plus de 1 000. Parmi ces sept transcrits deux ont été identifiés comme provenant d'ARNr (FPKM de 200 000 et 130 000), les cinq autres, au FPKM moins élevés (12 000, 2 700 puis en dessous de 1 500), n'avaient pas d'assignations fonctionnelles.

FPKM range	(0,1]	(1,10]	(10,100]	(100,1e+03]	(1e+03,Inf]
number of transcripts	136	15623	1103	81	7

Tableau 3.3 – Nombre de transcrits avec un FPKM donné sur le transcriptome de *M. endophyticus*

Dans un second temps, la répartition de l'expression des transcrits en fonction de leur pourcentage en GC a été recherchée (Figure 3.10). Un faible nombre de transcrits disposaient d'un pourcentage en GC supérieur à 45% parmi eux, 25 présentaient un FPKM supérieur à 50 (dont deux supérieur à 1 000).

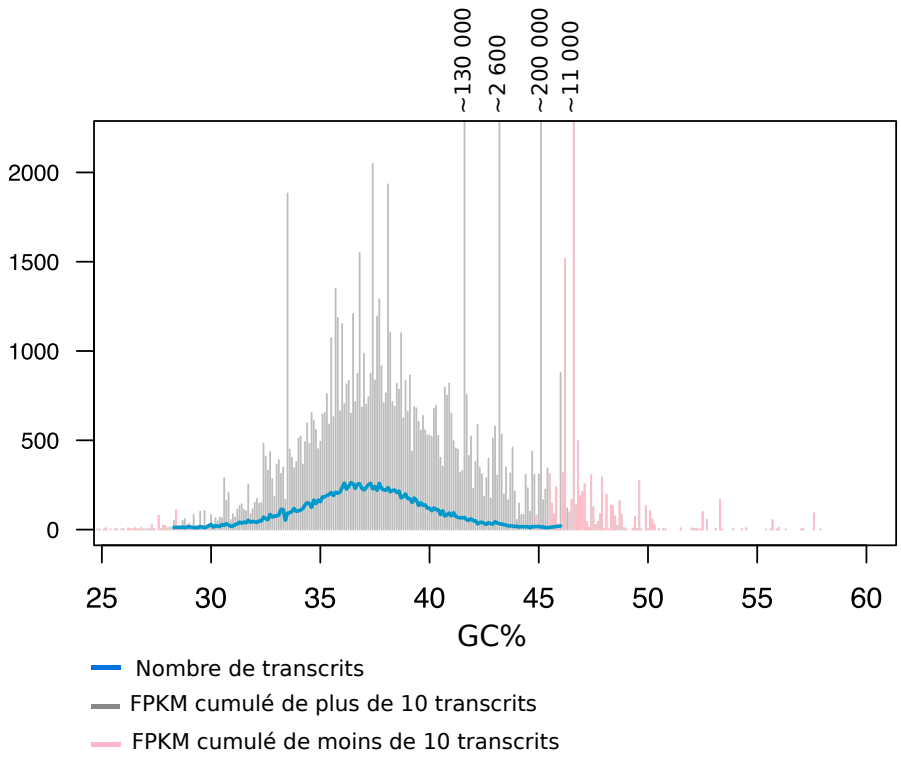


Figure 3.10 – Répartition de l'expression des transcrits en fonction de leur pourcentage en GC.

1. Le FPKM (Fragments Per Kilobase Million) correspond à une valeur normalisée d'expression des transcrits. Plus précisément, il s'agit du nombre de fragments (*reads* dans le cas de séquençage single end, paire de *reads* dans le cas de séquençage paired end) qui s'alignent sur un transcrit avec une normalisation pour la profondeur de séquençage et pour la longueur du transcrit. Plus d'informations sur http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/.

Parmi les 25 transcrits, dix-sept ont été identifiés comme similaires à des ARN ribosomiques. Parmi les huit autres (présentés en Tableau 3.4), cinq transcrits n'ont pas trouvé de correspondance dans les bases de données et ont été uniquement identifiés dans le transcriptome de *M. endophyticus*, les trois autres transcrits ont été retrouvés dans les transcriptomes des cinq espèces. Ils ont été identifiés comme heat shock protein, hypothetical guanine nucletide-binding protein et enolase.

Identifiant du transcrit	FPKM	%GC	Longueur du transcrit (pb)	Protéine prédite	Homologies	EC	Peptide signal	Domaine transmembranaire	PFAM	Espèces présentes dans l'orthogroupe
c2559_g1_i1	11 801	46.6	279	-	-	-	-	-	-	-
c9077_g1_i1	1 437	46.2	795	oui	-	-	oui	-	-	ME
c173_g1_i2	151	47.4	673	oui	-	-	oui	oui	-	ME
c3014_g2_i1	144	46.1	1 981	oui	hypothetical heat shock protein	-	-	-	Hsp70, MreB/Mbl	toutes
c173_g1_i1	141	47.9	733	-	hypothetical	-	-	-	-	-
c2971_g2_i1	82	45.5	1 121	oui	hypothetical guanine nucletide-binding	myosin-heavy-chain kinase	-	-	WD domain; anaphase-promoting; translation initiation factor	toutes
c2852_g2_i1	73	45.5	1 402	oui	enolase	phosphopyruvate hydratase/enolase	-	-	Enolase	toutes
c2673_g1_i1	60	46.1	469	-	-	-	-	-	-	-

Tableau 3.4 – Points remarquables sur les huit transcrits retrouvés dans le transcriptome de *M. endophyticus* à la teneur en GC supérieure à 45%, dont le FPKM est supérieur à 50 et ne provenant pas d'ARN ribosomaux.

Parmi les transcrits non identifiés, des peptides signaux menant vers l'extérieur de la cellule ont été identifiés chez deux transcrits, l'un de ces transcrits possédait également un domaine transmembranaire. Le premier transcrit code donc probablement une protéine secrétée dans le milieu tandis que l'autre code potentiellement une protéine localisée sur la membrane à l'interface entre milieu intracellulaire et milieu extérieur.

3.4.4 Discussion et conclusion

La vérification de la qualité des données de séquençage est importante pour l'interprétation des analyses qui en découlent. Les contaminations biologiques par exemple, si elles ne sont pas écartées du jeu de données, peuvent conduire à une mauvaise interprétation du contenu en gènes de l'espèce d'intérêt.

Les résultats précédents ont montré que la distribution trimodale du pourcentage GC des *reads* de *M. endophyticus* n'était pas liée à une contamination, mais à l'expression d'ARN ribosomiques et d'un nombre restreint de transcrits non ribosomiques fortement exprimés. Parmi ces transcrits fortement exprimés cinq n'ont pas pu être identifiés et n'ont été retrouvés que chez *M. endophyticus*, unique endophyte parmi les espèces sélectionnées. Trois transcrits en particulier pourraient fournir des informations importantes sur *M. endophyticus* : le premier était le transcrit non ribosomal le plus fortement exprimé du transcriptome, cependant, sa faible taille (279pb) pourrait suggérer un artéfact d'assemblage, les deux autres correspondent aux transcrits dont la protéine prédite possède un peptide signal. En effet, les protéines sécrétées jouent un rôle très important dans l'adaptation d'une espèce à un milieu que ce soit pour dégrader les substances présentes dans l'environnement comme des protéases et lipases (Alfaro et al., 2014; Pellegrin et al., 2015) ou pour interagir avec leur compétiteurs ou partenaires qu'ils soient microbiens, végétaux ou animaux (Stergiopoulos and de Wit, 2009; Essig et al., 2014).

L'objectif de cette recherche était d'identifier si des génomes d'organismes contaminants avaient été séquencés avec les génomes de *Mucor* séquencés dans le cadre de ce projet. Tous les transcriptomes de *Mucor* semblaient contenir des ARNr de bactéries. Les échantillons de *M. racemosus*, *M. fuscus* et *M. lanceolatus*, espèces "fromagères", contenaient des ARNr bactériens proches de *Janthinobacterium spp.*, genre bactérien pouvant être retrouvées lors de l'affinage de fromage (Fuka et al., 2013), *M. lanceolatus* contenait un ARNr bactérien supplémentaire proche de *Lactobacillus spp.*, genre de bactérie lui aussi retrouvé lors de l'affinage de fromage (Fuka et al., 2013). Cependant, aucun transcrit n'a été identifié comme provenant de l'une ou l'autre de ces bactéries dans les transcriptomes de ces trois *Mucor* (mêmes résultats pour les deux autres *Mucor*). Une explication possible reposerait sur la faible proportion (en nombre) de ces bactéries vivant en endosymbiose/épibiose avec ces champignons au point que seul les ARNr (très fortements exprimés) aient été détectés. Cette hypothèse est supportée par la présence de *GO terms movement in the environment of other organism involved in symbiotic interaction*, *multi-organism metabolic process*, *host cell part and other organism membrane* (cf article en partie 3.2) et par le fait

qu'une telle endosymbiose a été retrouvée chez un autre membre des Mucorales : *Rhizopus microsporus* avec des bactéries du genre *Burkholderia* (Partida-Martinez et al., 2007; Mondo et al., 2017). Cependant, à ce jour aucune bactérie n'a été identifiée au sein des hyphes des cinq espèces de *Mucor* étudiées dans notre étude. D'autre part les *GO terms* associés à une interaction endosymbiotique pourraient être expliqués par les interactions de ces champignons avec les bactéries présentes dans le milieu extérieur. Par exemple, Zhang et al. (2018) ont montrés que des bactéries "nageaient" dans le liquide de revêtement des hyphes mycélien de champignons, notamment de *Mucor*, pour se disperser dans la matrice fromagère, cette dispersion étant plus aisée sur les hyphes de *Mucor* que ceux d'autres espèces utilisées en affinage tel que *Galactomyces geotrichum*. Le fait de ne pas avoir détecté d'ADNr 16S par méthode de PCR classique ne va pas à l'encontre de cette hypothèse, puisque la quantité d'ADN bactérien dans le cas d'endosymbioses intrahyphales peut s'avérer faible et bien souvent la détection passe par des techniques d'amplificaion par PCR nichée (Desirò et al., 2015) ou par l'utilisation de FISH (Bertaux et al., 2003).

3.5 Comparaison des *EC numbers* à l'échelle des transcriptomes

3.5.1 Introduction

Comme indiqué dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2), des comparaisons des enzymes prédites ont été réalisées. Dans cet article, cet axe n'a pas été développé, le détail des recherches associées est présenté dans cette section.

Les enzymes peuvent être classifiées selon une nomenclature réalisée par l'Enzyme Commission [2] (*Enzyme Commission number* ou EC number). Il s'agit d'une classification numérique des enzymes basée sur la réaction chimique qu'elles catalysent.

Dans les analyses ci-après, à partir des protéines prédites, les enzymes et leurs *EC numbers* ont été recherchées. Les premières analyses présentées ci-après ont été réalisées avant l'obtention des données RNAseq de *M. circinelloides* ce qui explique l'absence de cette espèce dans les analyses présentées.

3.5.2 Matériels et méthodes

L'annotation des *EC numbers* a été détaillée dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2). Brièvement, sur les transcrits obtenus dans le cadre de cette étude ont été identifiés des régions potentiellement codantes. La fonction de ces protéines prédites a ensuite été déterminée. Les enzymes ont été identifiées par transfert d'annotation et par recherche de profils spécifiques en utilisant un outil nommé PRIAM (Claudel-Renard et al. (2003), version de mars 2015). Les fonctions de ces enzymes sont identifiées selon la classification des *EC numbers* .

Ces *EC numbers* ont été mappés sur un réseau métabolique avec iPath2 (Yamada et al., 2011) pour avoir une vision d'ensemble des enzymes en présence et leur rôle dans le réseau métabolique.

2. Cette Commission a été créée lors d'un Congrès International de Biochimie en 1955 afin d'établir des règles de nomenclature des enzymes (`http://www.sbcs.qmul.ac.uk/iubmb/enzyme/history.html`)

3.5.3 Résultats

Comparaison globales des enzymes retrouvées dans les transcriptomes

À partir des transcriptomes, de 14 800 (*M. endophyticus*) à 18 800 (*M. lanceolatus*) protéines ont été prédites. Parmi elles, 9 à 12% sont des enzymes (Tableau 3.5). Près de la moitié des fonctions enzymatiques identifiées (représentées par les *EC numbers*) sont assurées par plusieurs protéines (42% chez *M. endophyticus* à 53% chez *M. lanceolatus*). Certaines protéines prédites ont également plusieurs fonctions enzymatiques identifiées (100 protéines prédites chez *M. endophyticus* à 145 chez *M. lanceolatus*).

Espèce	nbr protéines prédites	nbr d'enzymes identifiées	enzymes constituant le protéome prédit	nbr d'enzymes différentes identifiées	nbr de protéines avec plusieurs fonctions enzymatiques prédites	nbr d'enzymes retrouvées en plusieurs copies	proportion d'enzymes retrouvées en plusieurs copies
M. endophyticus	14 852	1 289	8,6%	494	100	208	42,1%
M. fuscus	17 523	1 926	11,0%	588	135	299	50,9%
M. lanceolatus	18 802	2 116	11,3%	594	145	312	52,5%
M. racemosus	15 131	1 780	11,8%	601	132	278	46,3%

TABLEAU 3.5 – Description des enzymes identifiées à partir des protéines prédites des transcriptomes.

Les fonctions prédites des enzymes (au travers des *EC numbers*) ont été réparties au sein des espèces (Figure 3.11). La majorité des *EC numbers* (473) étaient communs aux quatre espèces étudiées, un grand nombre (93) étaient retrouvés chez les trois souches isolées à partir de milieu fromager et quelques *EC numbers* étaient propres à chacune des espèces. De façon intéressante, trois des six *EC numbers* retrouvés uniquement chez les espèces non technologiques (*M. racemosus* et *M. endophyticus*) étaient impliqués dans la synthèse de pigments.

Lors du mapping des *EC numbers* sur un réseau métabolique, la majorité des voies métaboliques semblaient être représentées par les enzymes retrouvées chez toutes les espèces. Les enzymes retrouvées chez les souches isolées à partir de fromage et absentes de la souche endophyte ne semblaient pas se concentrer sur des voies métaboliques spécifiques (Figure 3.13).

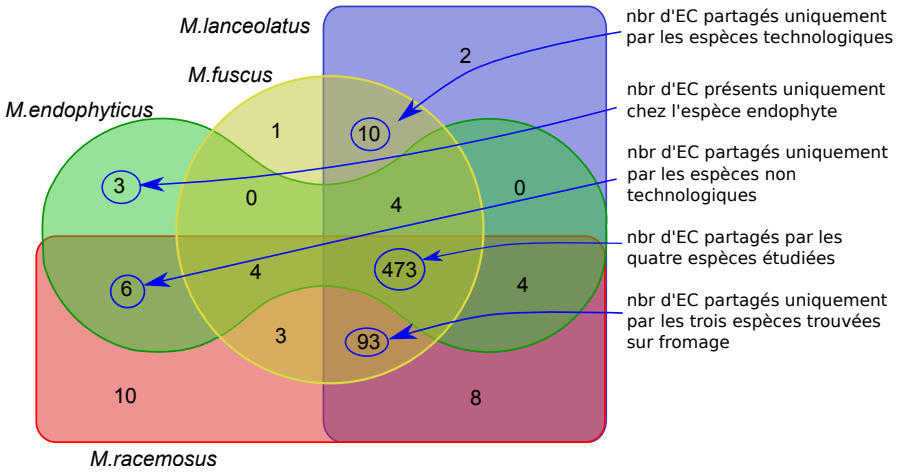


FIGURE 3.11 – Répartition des *EC numbers* au sein des espèces.

Comparaison des enzymes dupliquées au sein des transcriptomes

Le pourcentage d'*EC numbers* dupliqués était plus important chez les espèces technologiques (50,9% et 52,5%) que celles non technologiques (42,1% et 46,3%). La majorité des *EC numbers* dupliqués (166) étaient partagés par toutes les espèces, puis une part importante (61) étaient uniquement partagées par les espèces trouvées sur fromage suivi par les enzymes dupliquées uniquement chez *M. lanceolatus* (44) (Figure 3.12 gauche). Les *EC numbers* dupliqués, ne l'étaient pas un même nombre de fois entre les espèces (Figure 3.12 droite). Les *EC number* dupliqués ont été mappés sur le réseau métabolique présenté en Figure 3.14. L'ajout de *M. circinelloides* à ces analyses a fourni des résultats similaires (résultats non présentés ici).

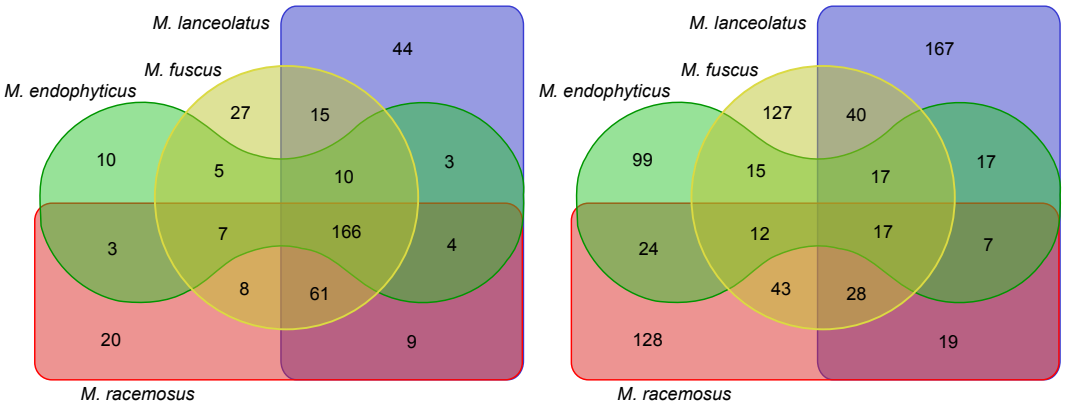


FIGURE 3.12 – Répartition des *EC numbers* dupliqués au sein des espèces ; à gauche *EC numbers* dupliqués ; à droite *EC numbers* dupliqués un même nombre de fois

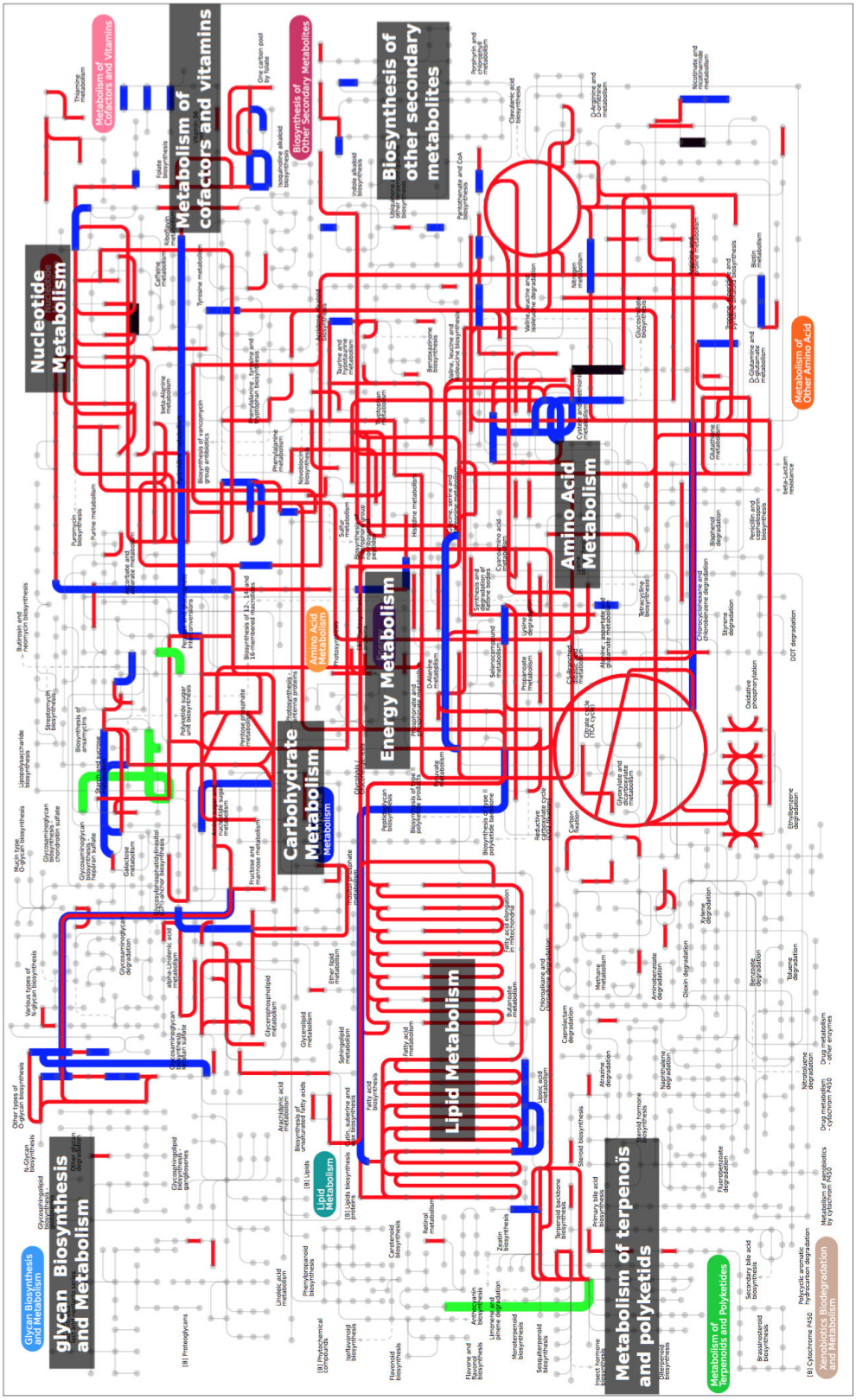


FIGURE 3.13 – Réseau métabolique des *Mucor* ; en rouge EC commun entre toutes les espèces ; en bleu *EC numbers* commun entre les espèces présentes sur fromage ; en vert plus épais *EC numbers* présents uniquement chez les espèces non technologiques (*M. racemosus* et *M. endophyticus*) et en noir *EC numbers* présents uniquement chez les espèces technologiques (*M. lanceolatus* et *M. fuscus*).

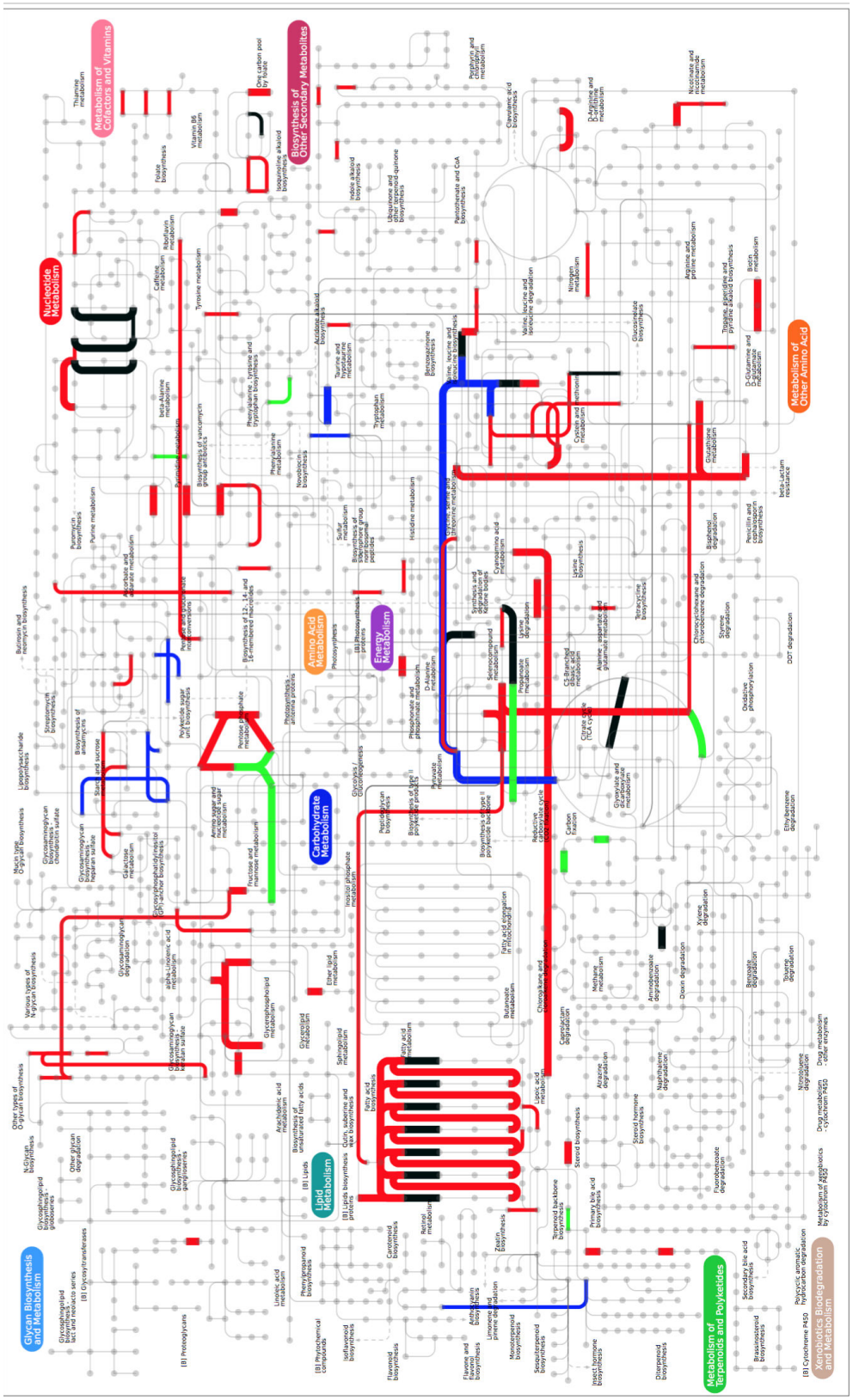


FIGURE 3.14 – Réseau métabolique des *Mucor* ; en rouge *EC number* commun entre les espèces présentes sur fromage ; en vert EC présents uniquement chez *M. endophyticus* ; en bleu EC présents uniquement chez les espèces non technologiques (*M. racemosus* et *M. endophyticus*) et en noir EC présents uniquement chez les espèces technologiques (*M. lanceolatus* et *M. fuscus*). Les lignes fines correspondent à une présence de l'*EC number* uniquement dans le groupe, les lignes épaisses correspondent à un *EC number* dupliqué uniquement dans le groupe.

3.5.4 Discussion

Le pourcentage d'*EC numbers* par rapport au nombre de protéines était cohérent avec les résultats de Wisecaver et al. (2014) qui décrivaient leur proportion par rapport au nombre de protéines à 15.4% chez les Saccharomycotina, 12.6% chez les Pezizomycotina et 8.9% chez les Agaricomycota. Dans cette même étude, les auteurs ont montré qu'en moyenne 88.7% des gènes avec *EC numbers* ont été dupliqués une ou plusieurs fois, ce pourcentage étant plus faible dans les lignées ayant divergé tôt dans l'évolution des espèces fongiques (ex : chez *Encephalitozoon cuniculi* du groupe des microsporidians, 49.0% des gènes avec *EC number* ont été dupliqués). Le pourcentage d'*EC numbers* retrouvés dupliqués dans cette étude était faible par rapport à ces informations mais (i) les *Mucor* font partie des lignées ayant divergé tôt dans l'évolution des champignons, (ii) il s'agit d'un groupe relativement peu étudié, il est donc encore plus complexe d'obtenir des annotations précises et exhaustives car les annotations sont principalement basées sur les connaissances acquises sur d'autres espèces, (iii) les données étudiées sont des transcriptomes tandis que la littérature se base sur des génomes, une partie des enzymes n'est peut être pas exprimée sur milieu synthétique.

Le nombre d'enzymes prédites était plus faible chez *M. endophyticus*, la seule souche endophyte étudiée, cependant cela correspond à son plus faible nombre de protéines prédites par rapport aux autres souches. Lors des analyses suivantes, aucune voie métabolique ne semblait spécifiquement impactée par ce nombre plus restreint d'enzymes prédites. L'espèce endophyte semble donc globalement posséder les mêmes types d'enzymes que les espèces retrouvées sur fromage et ubiquistes. Observer les spécificités des différentes espèces de *Mucor* nécessite donc de s'intéresser à des voies métaboliques spécifiques.

De façon intéressante, trois des six *EC numbers* retrouvés uniquement chez les espèces non technologiques (*M. racemosus* et *M. endophyticus*) étaient impliqués dans la synthèse de pigments, ce qui concorde avec la variation de la couleur des spores de *Mucor* selon les espèces (observation personelle). La synthèse de caroténoïdes a été reportée chez les *Mucor* (Zhang et al., 2016), or la voie de biosynthèse des caroténoïdes possède des étapes communes avec la synthèse de différents types de terpènes (Schmidt-Dannert, 2015). Il serait intéressant par la suite de voir si d'autres éléments de cette voie métabolique diffèrent selon les espèces.

3.6 Détail des analyses sur les protéines prédites propres à chacune des espèces et partagées par les espèces au même mode de vie

3.6.1 Introduction

Comme indiqué dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2), les protéines prédites des cinq transcriptomes ont été regroupées en groupes d'orthologues (Figure 3.15). Cette recherche a permis d'accéder au core transcriptome, aux transcrits partagés uniquement par les espèces au même mode de vie ainsi qu'aux transcrits propres à chacune des espèces. Ces derniers ont été décrits et les annotations fonctionnelles des transcrits partagés uniquement par les espèces au même mode de vie ont été comparées. Ces deux derniers points, qui n'ont pas été détaillés dans l'article, seront présentées ci-après.

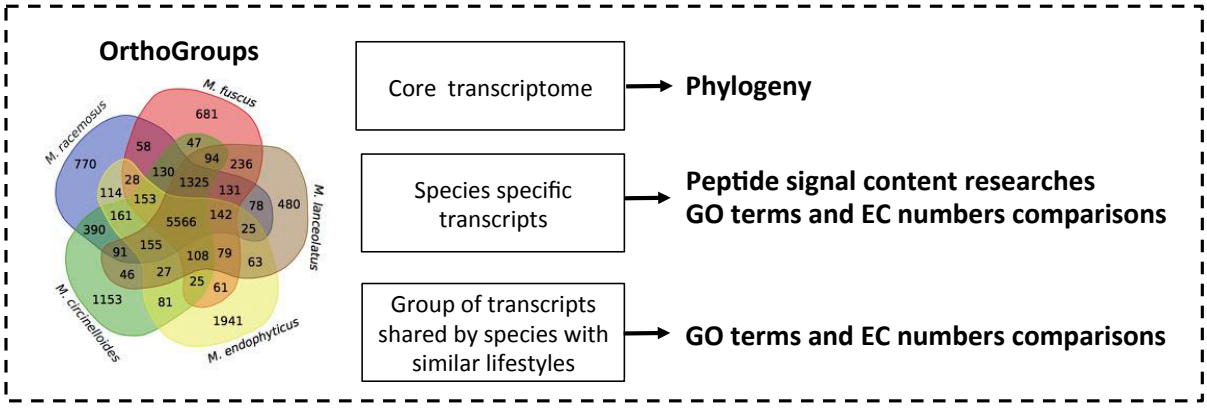


FIGURE 3.15 – Répartition des orthogroupes entre les transcriptomes des espèces.

3.6.2 Matériels et méthodes

L'annotation des transcrits a été détaillée dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2). Les recherches de voies métaboliques dans lesquelles les *EC numbers* étaient impliquées ont été réalisées avec IPath2 (Yamada et al., 2011). La recherche des fonctions des *GO terms* a été réalisée d'une part avec Revigo (Supek et al., 2011) et d'autre part avec le script personnel utilisé dans l'article "*Comparative analysis of five* Mucor *species transcriptomes*" (3.2) et détaillé en partie 3.3.

3.6.3 Résultats

Description générale des protéines prédites propre à chaque espèce

Une première description des protéines prédites propres à chaque espèce (ou protéines isolées) a été réalisée (Tableau 3.6). Ces protéines prédites représentaient 4% (*M. lanceolatus*) à 18% (*M. endophyticus*) des protéomes prédits des espèces. Bien que certaines protéines prédites soient de taille conséquente (plus de 1000aa) la majorité était de plus petite taille que les protéines retrouvées dans le transcriptome complet (N50 d'environ 200aa chez les protéines propres à l'espèce contre 500aa à l'échelle du transcriptome complet).

Espèce	*M. racemosus*	*M. fuscus*	*M. lanceolatus*	*M. endophyticus*	*M. circinelloides*
Nbr de protéines prédites dans tout le transcritome	11 728	11 157	10 674	10 646	11 953
Nbr de protéines prédites propre à l'espèce	771	687	484	1 941	1 155
% de protéines prédites propre à l'espèce	7%	6%	4%	18%	10%
Taille maximale de ces protéines (aa)	1 168	1 309	911	1 083	1 014
N50 de ces protéines (aa)	194	197	187	193	323
Nbr d'EC numbers annotés	183	97	110	637	212
Nbr d'EC numbers différents	108	77	86	279	117
Nbr de protéines possédant une annotation GO	513	328	264	1 433	547
Nbr de protéines avec un peptide signal prédit	38	25	18	60	75
Nbr de protéines sans annotations GO ni EC	257	358	219	502	603
% de protéines sans annotations GO ni EC	33%	52%	45%	26%	52%

TABLEAU 3.6 – Description des protéines prédites propres à chaque espèce identifiées dans les transcriptomes. Le nombre de protéines prédites propres à l'espèce est différent des valeurs retrouvées en Figure 3.15 car certains orthogroupes étaient constitués de plusieurs protéines prédites appartenant à la même espèce.

Comparaison des annotations fonctionnelles de type *GO terms* et *EC numbers*

Les protéines isolées pour lesquelles un EC a été annoté ne représentaient pas de voies métaboliques spécifiques. La comparaison des *EC numbers* identifiés chez les différentes espèces a révélé que ceux-ci étaient majoritairement propres à l'espèce. Les *GO terms* associés aux protéines propres à chacune des espèces étaient similaires. La comparaison des *GO terms* des groupes de protéines propres à chacune des cinq espèces a mené à l'identification de onze *GO terms* possédant une distribution significativement différente de celle attendue (test de Fisher exact, p-value < 0.05 ; Figure 3.16).

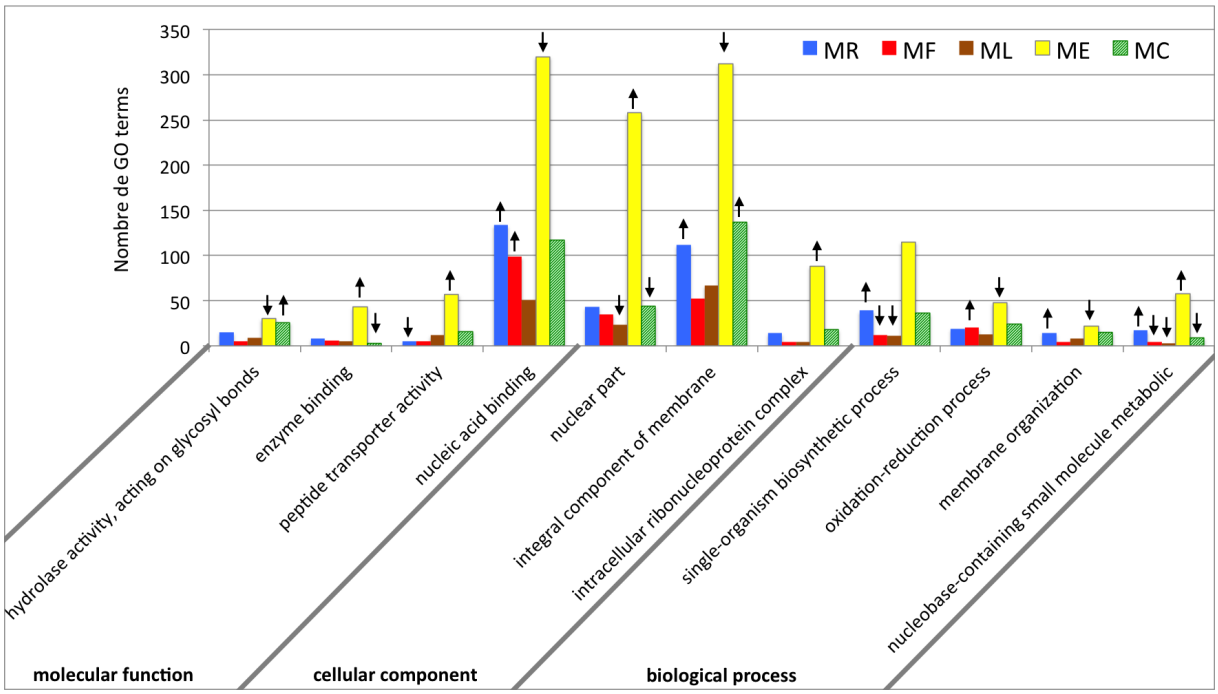


FIGURE 3.16 – *GO terms* avec une distribution entre espèces significativement différente de celle attendue (test de Fisher exact, p-value 0.05). Les *GO terms* ont été préalablement synthétisés à une profondeur 3 relations "is a". MR : *M. racemosus*. MF : *M. fuscus*. ML : *M. lanceolatus*. ME : *M. endophyticus*. MC : *M. circinelloides*. Une flèche pointant vers le haut indique une valeur plus importante qu'attendue, une flèche vers le bas indique une valeur moins importante qu'attendue.

Protéines avec peptide signal

Parmi les protéines prédites propres à chaque espèce, le nombre de protéines disposant de peptides signaux variait selon les espèces (*M. circinelloides* 75, *M. endophyticus* 60, *M. racemosus* 38, *M. fuscus* 25 et *M. lanceolatus* 18). Ces protéines étaient globalement petites et souvent partielles et ne possédaient pas de correspondances avec les bases de données ou étaient annotées *Hypothetical Unknown protein* à l'exception de deux protéines prédites avec peptides signaux identifiées chez *M. lanceolatus* (une serine protéase et une endo-rhamnogalacturonase potentielle), deux chez *M. racemosus* (une NADH dehydrogenase mitochondriale et une UDP-glycosyl transferase), trois chez *M. circinelloides* (deux glycoside hydrolase et une carbohydrate esterase) et deux chez *M. endophyticus* (une arginase et une phosphatase alcaline).

Comparaison des *GO terms* entre groupe de protéines retrouvées chez des espèces au même mode de vie

Les annotations fonctionnelles de type *GO terms* ont été comparées entre les familles de gènes (au travers des orthogroupes) retrouvées chez les espèces présentes en milieu fromager avec celles non retrouvées (ou de façon exceptionnelle) sur milieu fromager (Figure 3.17).

Cette recherche a permis de mettre en évidence 18 *GO terms* représentés significativement différemment entre les deux groupes. Parmi les dix *GO terms* sur-représentés chez les espèces non retrouvées sur milieu fromager, on compte des éléments liés à l'organisation de la cellule comme "*maintenance of location*", des éléments liés à la croissance de l'organisme ("*filamentous growth*" et "*cell cycle process*") et des activités enzymatiques ("*hydrolase activity*", "*ligase activity*", "*deaminase activity*"). Parmi les huit *GO terms* sur-représentés chez les espèces retrouvées sur fromage, on compte des éléments associés à la structure de l'organisme ("*external encapsulating structure*" et "*host cell part*"), des éléments associés à la fixation de composés ("*heterocyclic compound binding*" et "*organic cyclic compound binding*") et les transporteurs de sucre ("*carbohydrate transporter activity*").

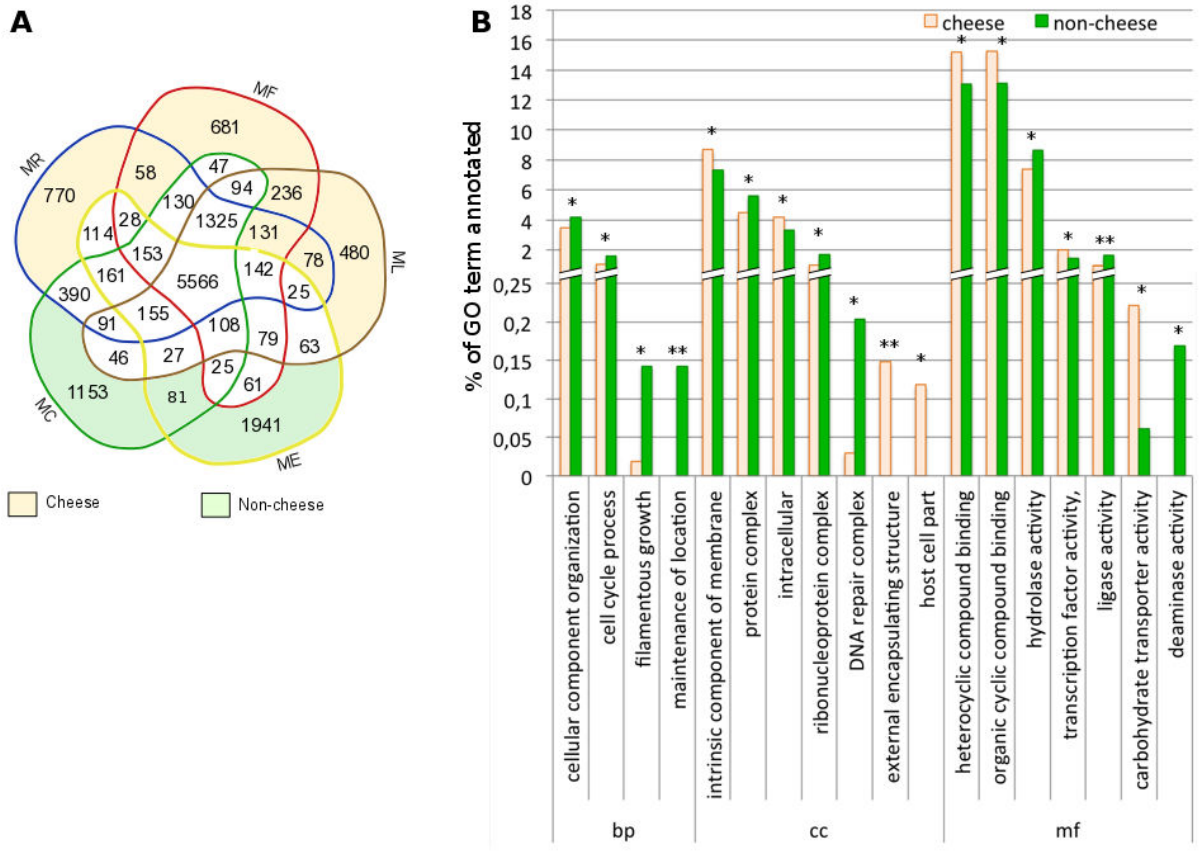


FIGURE 3.17 – A : Répartition des orthogroupes au sein des espèces. B : Comparaison des annotations de type *GO terms* (profondeur deux relations "is a") entre les orthogroupes retrouvés chez les espèces présentes en milieu fromager et celles non retrouvées (ou de façon exceptionnelle) sur milieu fromager. (Fisher exact, * p value < 0.05, ** p value < 0.005, *** p value < 0.0005). bp : *Biological Process*, cc : *Cellular Component*, mf : *Molecular Function*, MR : *Mucor racemosus*, MF : *M. fuscus*, ML : *M. lanceolatus*, ME : *M. endophyticus*, MC : *M. circinelloides*.

La même recherche a été réalisée entre les familles de gènes retrouvées chez les espèces identifiées comme pathogènes opportunistes avec celles non pathogènes (Figure 3.18). Cette recherche a permis de mettre en évidence 23 *GO terms* représentés significativement différemment entre les deux groupes. Parmi les cinq groupes sur-représentés chez les pathogènes, on compte des éléments liés à la position des composés dans le milieu ("*establishment of localization*", "*single organism localization*" et "*maintenance of location*") au catabolisme ("*catabolic process*") et aux composant de la membrane ("*intrinsic component of membrane*"). Parmi les 18 *GO terms* sur-représentés chez les espèces non-pathogènes, on retrouve des *GO terms* impliqués dans le contrôle du nombre de cellules ("*cell proliferation*", "*maintenance of cell number*"), des facteurs de transcription ("*transcription factor activity*") ou encore des activités enzymatiques ("*deaminase activity*", "*protein serine/threonine phosphatase*").

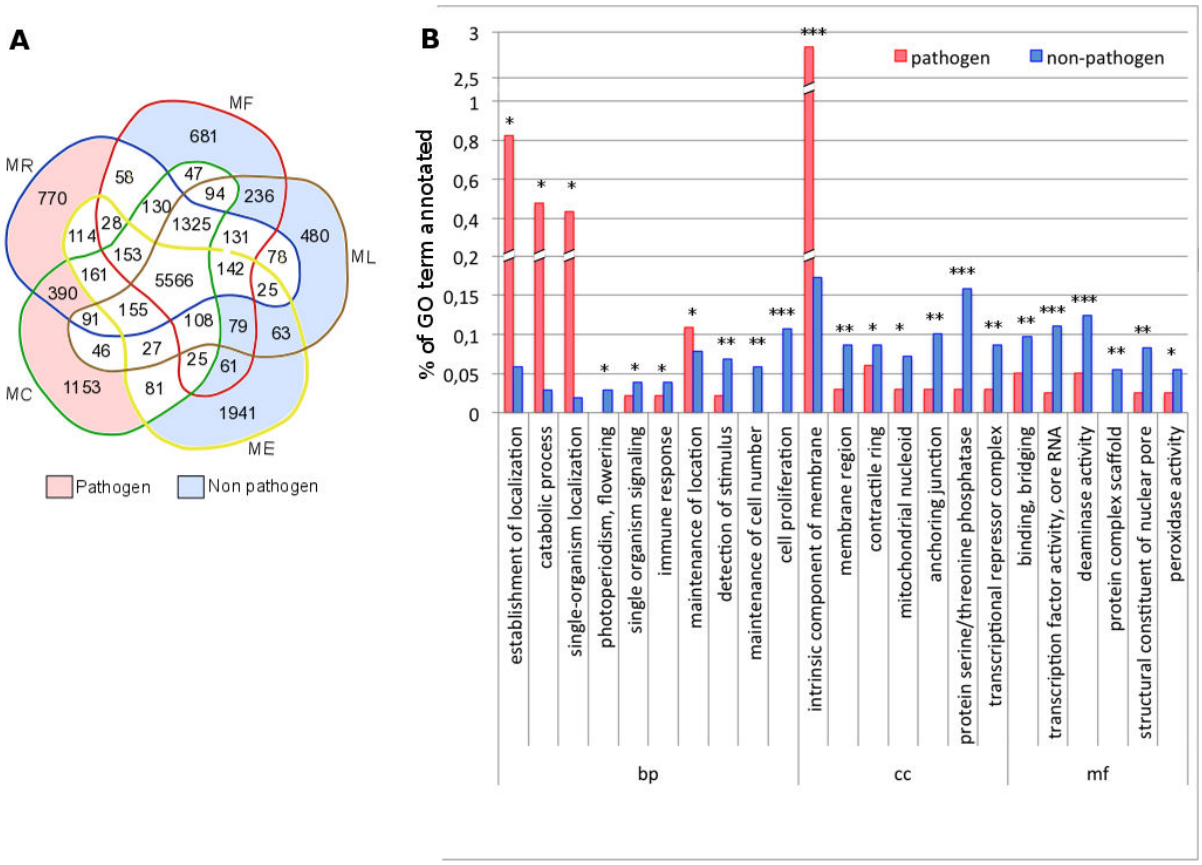


FIGURE 3.18 – A : Répartition des orthogroupes au sein des espèces. B : Comparaison des annotations de type *GO terms* (profondeur deux relations "is a") entre les orthogroupes retrouvés chez les espèces pathogènes opportunistes et celles non pathogènes. (Fisher exact, * p value < 0.05, ** p value < 0.005, *** p value < 0.0005). bp : *Biological Process*, cc : *Cellular Component*, mf : *Molecular Function*, MR : *Mucor racemosus*, MF : *M. fuscus*, ML : *M. lanceolatus*, ME : *M. endophyticus*, MC : *M. circinelloides*.

3.6.4 Discussion

Lorsque l'on s'est intéressé aux protéines prédites propres à chacune des espèces, la comparaison des annotations fonctionnelles n'a pas permis d'identifier de différence associée au mode de vie. Le nombre de protéines prédites avec peptide signal était deux à quatre fois plus important chez les espèces retrouvées sur fromage que chez les espèces non retrouvées sur fromage. Cependant, presque aucune de ces protéines prédites n'a pu être identifiée par homologie de séquence avec les séquences connues.

Les résultats obtenus lors de la comparaison des *GO terms* entre les familles de gènes (au travers des orthogroupes) retrouvées chez les espèces partageant le même milieu, étaient complexes à interpréter, d'une part du fait du manque de spécificité de certaines des catégories (ex : *Cellular Component organization*, *protein complex*) d'autre part car des éléments apparaissaient aussi bien dans les comparaisons sur les espèces fromagères que pathogènes (ex : *intrinsic component of membrane* sur-représenté chez les espèces fromagères et chez les espèces pathogènes).

L'obtention du transcriptome de quatre espèces de *Mucor* aux modes de vie distincts pour lesquelles aucune donnée ni génomique (au début de cette étude) ni transcriptomique n'était disponible a permis d'étendre les connaissances sur le répertoire du transcriptome au niveau du genre. Ces connaissances sont d'autant plus importantes que jusqu'ici les recherches se sont principalement concentrées sur les espèces pathogènes opportunistes (et parmi elles, principalement *M. circinelloides*) (Morin-Sardin et al., 2017) alors que notre étude cible trois espèces non pathogènes. Ces travaux permettent donc de mieux caractériser le genre *Mucor* dans sa diversité et par là même, de révéler des différences entre espèces de *Mucor* tels que les éléments liés à la pathogénicité des rares espèces capables d'infecter les humains.

L'axe principal de recherche de cette analyse portait sur l'identification d'éléments potentiellement impliqués dans l'adaptation au milieu présents chez les cinq *Mucor* spp.. Cette recherche est d'autant plus complexe que les séparations entre milieux de vie ne sont pas nettes : les deux espèces considérées comme pathogènes opportunistes sont ubiquistes (Walther et al., 2013). Il est également possible que les espèces spécialisées puissent être rencontrées dans d'autre milieux mais n'aient jusqu'à présent été détectées que sur un seul type de substrat. *M. endophyticus* est capable de se développer (avec un temps de latence (Morin-Sardin et al., 2016)) sur milieux PDA et fromager, il n'est donc pas un endophyte obligatoire. Et si une espèce comme *M. lanceolatus* n'a été décrite que sur fromage, il n'est pas exclu qu'elle soit beaucoup plus ubiquiste que ne le

suggèrent les études précédentes. Les analyses réalisées ici tendent à montrer que les cinq *Mucor* spp. ne présentent pas de différences majeures, aussi bien en termes de composition d'enzymes que de fonctions des protéines, associées à une adaptation au milieu de vie. Des différences plus marquées pourraient être détectées (i) a des niveaux soit plus précis ; des différences en termes de facteurs de virulence ont été détectées en s'intéressant à des familles de gènes spécifiques ; (ii) parmi les nombreux éléments pour lesquels aucune fonction n'a pu être prédites ; c'est le cas de la protéine de *M. circinelloides* ID112092 pour laquelle aucune fonction n'est prédite mais qui a été identifiée par Lopez-Fernandez et al. (2018) comme étant liée à la virulence. Les transcriptomes sont le reflet de l'expression des génomes dans une condition donnée. Nous avons choisi dans cette étude de les comparer sur un milieu et des conditions de culture standards. La difficulté à trouver des différences entre les espèces peut être liée au fait que toute ces espèces ont été cultivées dans des conditions standards, éloignées de leurs conditions de vie naturelle. Afin d'avoir accès a l'ensemble des gènes potentiellement exprimés, et ce quel que soit les conditions, une approche de génomique comparative a été initiée. Cette étude permettra notamment de vérifier si les éléments non retrouvés dans les transcriptomes des différentes espèces sont réellement absents ou s'ils ne sont pas exprimés dans les conditions utilisées.

Parmi les transcrits présents chez l'ensemble des espèces analysées, des transcrits impliqués dans la synthèse de métabolites secondaires ont été identifiés dans chacune des souches : un transcrit correspondant à une NRPS et un transcrit possédant la structure typique et l'ordre des domaines d'une FAS annotée dans les bases de données en tant que PKS (Voigt et al., 2016). De manière intéressante, l'analyse des annotations fonctionnelles a également mis en évidence des *GO terms* associés à des interactions avec d'autres organismes : "*movement in the environment of other organism involved in symbiotic interaction*", "*multi-organism metabolic process*" ou encore "*host cell part and other organism membrane*". Ces éléments peuvent indiquer la présence de bactéries endosymbiotes/epibiotes chez les souches étudiées, cas ayant déjà été rencontré chez *Rhizopus microsporus*, un autre membre de l'ordre des Mucorales (Partida-Martinez et al., 2007; Mondo et al., 2017).

L'étude des transcriptomes n'a pas permis de mettre en évidence d'éléments associés à une adaptation au milieu fromager chez les espèces utilisées lors de l'affinage de fromage (*M. fuscus* et *M. lanceolatus*). Les espèces ubiquitaires (*M. racemosus* et *M. circinelloides*) présentaient des annotations fonctionnelles de type *GO terms* enrichies dans les catégories "*intrinsic component of*

membrane", "*carbohydrate transporter activity*", "*xenobiotic transporter activity*", "*oxidoreductase activity*" et "*transcription factor activity*" sur milieu PDA. Lorsque les transporteurs de sucres ont été ciblés, trois familles de transporteurs intégrant davantage de membres chez *M. racemosus* et *M. circinelloides* que chez les autres espèces ont été mis en évidence. Il s'agissait de transporteurs d'acides D-galacturonique et quinique, de transporteurs de glucose et pentose et d'une famille de transporteur de sucre non identifiée. L'acide D-galacturonique est le composant principal de la pectine, un polysaccharide largement utilisé dans la paroi des plantes notamment des fruits (Benz et al., 2014), l'expansion de ce type de transporteurs de sucre pouvait donc présenter un avantage pour ces espèces. La famille de transporteurs de sucres non identifiée est très restreinte chez *M. endophyticus* et au contraire très étendue chez les espèces pathogènes opportunistes, déterminer sa fonction plus précisément pourrait donc permettre d'avoir des indices sur l'adaptation des *Mucor* au milieu. D'autre part, la catégorie GO "*xenobiotic transporter*" était sur-représentée chez *M. racemosus* et *M. circinelloides*, de même un orthogroupe associé à une famille de gènes annotée "*multidrug resistance*" était principalement composée de protéines prédites associées à ces deux espèces. Ce type de gènes pourrait faciliter une infection d'organismes par ces deux espèces connues pour être des pathogènes opportunistes. Lorsque des facteurs de virulence rapportés dans la littérature (Ibrahim et al., 2012; Gebremariam et al., 2014; Liu et al., 2015; Chibucos et al., 2016; Lopez-Fernandez et al., 2018; Navarro-Mendoza et al., 2018) ont été recherchés, tous ont été retrouvés chez les deux espèces connues pour être des pathogènes opportunistes tandis qu'au moins l'un d'entre eux n'a pas été détecté chez les espèces technologiques et l'espèce endophyte. En effet, les ferroxidases *fet3b* et *fet3c* étaient absents du transcriptome de *M. endophyticus* or ces gènes sont nécessaires pour la virulence de *M. circinelloides* chez la souris (Navarro-Mendoza et al., 2018), de même le nombre de "*spore coat proteins homologs*" (cotH), impliqués dans la virulence des Mucorales chez les humains, était deux fois plus faible chez les espèces technologiques par rapport aux espèces potentiellement pathogènes.

Chapitre 4

Approche génomique

4.1 Introduction

Le projet de comparaison des génomes a été initié avec le séquençage des génomes de *M. fuscus*, *M. lancelolatus*, *M. racemosus* et *M. endophyticus*. Les objectifs de ce projet sont de : (i) mieux caractériser le pan-génome du genre *Mucor* et (ii) rechercher s'il existe au sein des génomes de *Mucor* des éléments permettant d'expliquer des adaptations à différents habitats et modes de vie. Dans le cadre de ce projet, j'ai contribué à l'extraction d'ADN des isolats de *M. lancelolatus* et *M. endophyticus* et aux assemblages des génomes des quatre espèces. J'ai ensuite réalisé l'annotation structurale et fonctionnelle des quatre génomes (Figure 4.1).

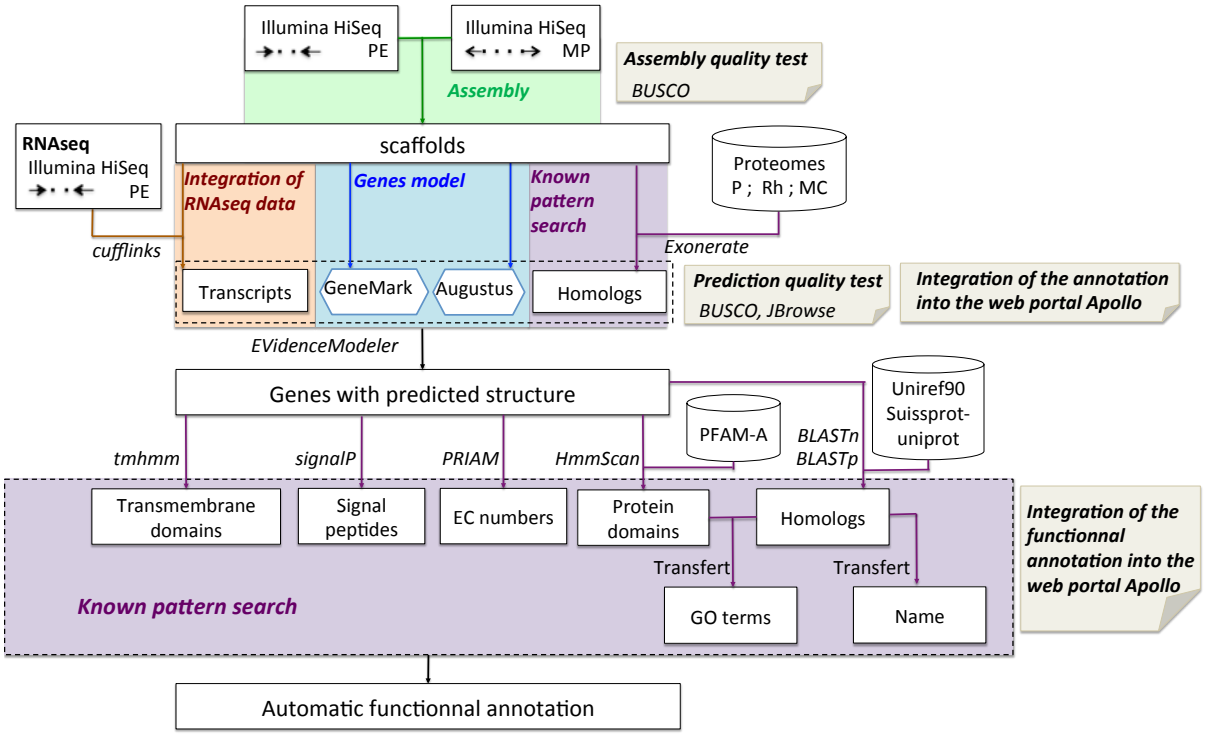


FIGURE 4.1 – Méthodes d'assemblage et d'annotation utilisées pour les génomes de *M. fuscus*, *M. lancelolatus*, *M. racemosus* et *M. endophyticus*, séquencés dans le cadre de ce projet.

Lorsque l'étude a été initiée, six génomes de *Mucor* disposant d'annotations étaient disponibles dans les bases de données publiques (Tableau 4.1). Pour mieux répondre aux objectifs du projet, ces génomes : à savoir *M. ambiguus* NBRC 6742 (considéré comme un *M. circinelloides* par la collection nationale fongique des États-Unis (https://nt.ars-grin.gov)), *M. circinelloides* 1006PhL, *M. circinelloides* CBS 277.49, *M. circinelloides* CDC-B8987, *M. circinelloides* CDC-B5328 (synonyme de *M. velutinosus* (Walther et al., 2013)), et *M. indicus* CDC-B7402 ont été intégrés aux comparaisons des génomes des quatres souches de *Mucor* que le laboratoire possède ainsi que deux isolats CDC-B9645 et CDC-B9738 identifiés comme *M. racemosus* par Chibucos et al. (2016) puis réassignés très récemment à l'espèce *Rhizopus microsporus* par Gryganskyi et al. (2018).

Taxon	Souche	Source de l'isolat	Référence ou n° d'accession	Annotation disponible
Mucor ambiguus	NBRC 6742	Inconnue	BBKB00000000.1, 2015	oui
Mucor circinelloides	1006PhL	Humain	Findley et al., 2013	oui
Mucor circinelloides	CBS 277.49	Inconnue	Corrochano et al., 2016	oui
Mucor circinelloides	CDC-B8987	Humain	Chibucos et al., 2016	sur demande
Mucor circinelloides	JCM 22480	Inconnue	BCHG00000000.1, 2016	non
Mucor circinelloides	WJ11	Sol	Tang et al., 2015	non
Mucor indicus	CDC-B7402	Humain	Chibucos et al., 2016	sur demande
Mucor irregularis	B50	Humain	AZYI00000000.1, 2014	non
Mucor irregularis	B7584	Inconnue	JNES00000000.1, 2014	non
Mucor velutinosus	CDC-B5328	Humain	Chibucos et al., 2016	sur demande

TABLEAU 4.1 – Liste des génomes de *Mucor* accessibles publiquement. *M. velutinosus* est un synonyme de *M. circinelloides* (Walther et al., 2013), *M. ambiguus* est considéré comme un *M. circinelloides* par la collection nationale fongique des Etats Unis (https://nt.ars-grin.gov)

Pour ces génomes complémentaires, une annotation structurale et fonctionnelle était disponible pour les gènes. Cependant aucune information n'était présente concernant les éléments répétés. Le pipeline d'annotation des éléments répétés utilisé pour les quatre souches séquencées dans le cadre de ce projet a donc été appliqué sur ces génomes publics. De même, des annotations spécifiques, à savoir sur les peptidases (MEROPS) et Carbohydrate-active enzymes (CAZymes), ont été réalisées sur ces génomes publiquement accessibles de la même façon que pour les quatre génomes des souches séquencées dans le cadre de ce projet.

Afin de définir les caractéristiques du génome du genre *Mucor* et déterminer s'il existait des éléments indiquant une potentielle adaptation de certaines de ces espèces à leur mode de vie différentes caractéristiques ont été étudiées.

(I) Les caractéristiques structurales du génome du genre *Mucor* ont été étudiées au travers de (i) la recherche de synténie entre les génomes, (ii) la description de la structure des gènes

(nombre et taille des introns par exemple), (iii) l'étude des régions intergéniques de grande taille et (iv) la recherche de clusters métaboliques associés à la synthèse de métabolites secondaires.

(II) L'évolution des familles de gènes et d'éléments transposables au sein du genre a été étudiée. La première étape de ces analyses a consisté à reconstruire la phylogénie des espèces étudiées, cette phylogénie a été calibrée sur une échelle de temps. Par la suite, la proportion et la composition des éléments répétés identifiés sur les génomes ont été comparées à la lumière de cette phylogénie. Les protéines prédites à partir des gènes annotés sur chacun des génomes ont ensuite été regroupées en familles (orthogroupes) ce qui a notamment permis de définir le *core genome* des *Mucor* et les gènes propres à chaque espèce. Des analyses d'expansion et contraction de familles de gènes sont actuellement en cours afin d'identifier s'il existe des familles de gènes réduites ou étendues spécifiquement dans des groupes d'espèces partageant le même habitat. La fonction de ces familles sera par la suite étudiée pour déterminer si ces expansions/contractions participent potentiellement à l'adaptation de chacune de ces espèces à leur l'habitat.

(III) En parallèle de ces études globales, ont été réalisées des comparaisons fonctionnelles plus ciblées. En effet, des gènes identifiés dans la littérature comme étant importants pour l'adaptation au milieu de vie des espèces ont été comparés entre les différentes souches : pour l'instant, les gènes comparés sont associés à la synthèse de métabolites secondaires (PKS/FAS, NRPS et terpènes synthases et cyclases) et les gènes associés au métabolisme du fer.

Comme explicité ci-avant, les génomes des deux isolats CDC-B9645 et CDC-B9738 identifiés comme *M. racemosus* (Chibucos et al., 2016) puis réassignés très récemment à l'espèce *R. microsporus* par Gryganskyi et al. (2018) ont été intégrés à notre étude.

N'ayant obtenu cette dernière information qu'en septembre 2018, notre analyse du core-génome associé au genre *Mucor* intègre actuellement les génomes de ces deux souches alors qu'elles n'appartiennent pas à ce genre. De même, l'analyse détaillée de la composition du core-génome, des gènes propres à chacune des espèces et de ceux partagés entres espèces aux mêmes modes de vie, devra être réalisée sur cette nouvelle base avant soumission de l'article.

L'intégration de ces deux génomes a également posé problème lors de l'analyse des expansions/contractions de familles de gènes. Celle ci a été réalisée afin de voir si l'expansion (ou au contraire la contraction) de certaines familles avait accompagné un processus d'adaptation évolutive à certains habitats (ex. fromage) ou niche (parasitisme, endophytisme). En effet, la présence de ces génomes a conduit à des résultats incohérents (comme expliqué dans la partie

4.6) dont l'origine pourrait résider, d'après nos premières analyses, dans le caractère hybride de ces deux génomes (voir partie 4.7). Cette analyse des expansions/contractions de familles de gènes sera donc réalisée sur un jeu de données dépourvu de ces deux génomes.

Une fois ces analyses terminées et intégrées à l'article "Comparative genomics applied to *Mucor* species with different lifestyles", ce dernier sera soumis pour publication.

Les résultats de ces travaux sont présentés dans l'article en préparation : "Comparative genomics applied to *Mucor* species with different lifestyles".

4.2 Article : Comparative genomics applied to *Mucor* species with different lifestyles

Auteurs :

Annie Lebreton[a], Erwan Corre[b], Jean-Luc Jany[a], Loraine Gueguen[b], Carlos Perez-Arques[c], Misharl Monsoor[b], Antoine Hermet[a], Robert Debuchy[d], Emmanuel Coton[a], Georges Barbier[a], Laurence Meslet-Cladière[a]

Affiliations :

[a] Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, Université de Brest, Technopôle Brest-Iroise, Plouzané 29280, France

[b] Station biologique de Roscoff, plateforme ABiMS, Sorbonne Université (UPMC), Roscoff 29682, France

[c] Department of Genetics and Microbiology. Faculty of Biology. University of Murcia. 30100 Murcia Spain

[d] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, Université Paris-Saclay, CEDEX, 91198 Gif-sur-Yvette, France

Correspondance : laurence.meslet@univ-brest.fr

Article en cours de rédaction.

Comparative genomics applied to *Mucor* species with different lifestyles

Annie Lebreton[1], Erwan Corre[2], Jean-Luc Jany[1], Loraine Gueguen[2], Carlos Perez-Arques[3], Misharl Monsoor[2], Antoine Hermet[1], Robert Debuchy[4], Emmanuel Coton[1], Georges Barbier[1], Laurence Meslet-Cladière [1]

[1]Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, Université de Brest, Technopôle Brest-Iroise, 29280 Plouzané, France.

[2]Station biologique de Roscoff, plateforme ABiMS, CNRS: FR2424, Sorbonne Université (UPMC) - Paris VI, Place Georges Teissier - BP 74 29682 ROSCOFF CEDEX - France.

[3]Department of Genetics and Microbiology. Faculty of Biology. University of Murcia. 30100 Murcia Spain.

[4]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, Université Paris-Saclay, CEDEX, 91198 Gif-sur-Yvette, France

Contributions:

Research idea conception: ECn, GB, LMC and JLJ

DNA Extraction: AL, RD and AH

Analytical approach design: ECe, JLJ and LMC

Apollo instance deployment for manual annotation: LG and MM

Manual annotation: AL, JLJ, CPA and LMC

Data analysis: AL, ECe, JLJ and LMC

Manuscript writing: AL, JLJ and LMC

Manuscript discussion and revision: ECn, GB and ECe

All authors read and approved the final manuscript.

Abstract

Despite the growing number of investigations on fungal early diverging lineages species, this lineage has not been as extensively characterized in comparison to ascomycetes or basidiomycetes. The *Mucor* genus, pertaining to one of this lineage (Mucoromycota phylum, Mucorales order) is not an exception. To this date, a restricted number of *Mucor* annotated genomes is publicly available and mainly correspond to the reference species, *Mucor circinelloides,* and medically relevant species. However, this genus is composed of a large number of ubiquitous species as well as few species that have been reported to specifically occur in certain habitats. In this study, we report four newly sequenced *Mucor* genomes with contrasted lifestyles namely *M. fuscus* and *M. lanceolatus*, two species used in cheese production (during ripening), *M. racemosus*, a recurrent cheese spoiler sometimes described as an opportunistic pathogen and *M. endophyticus*, a plant endophyte. Comparison of these new genomes with those previously available of six *Mucor* isolates and two *Rhizopus* isolates, formerly identified as *M. racemosus,* appended the descriptive data set for *Mucor* genomes, pointed out the complexity of obtaining a robust phylogeny even with multiple genes families, emphasized the unique structure of FAS genes and provided hints on potentially lifestyle-associated genes (e.g. iron metabolism).

Introduction

The *Mucor* genus belongs to the most prominent order of the Mucorales, a phylogenetically ancient group of fungi pertaining to the ‘’early diverging fungi’’ [1]. From the first microscopic observation of a *Mucor* specimen in 1665 up until now, several hundreds of potential *Mucor* species have been reported [2]. *Mucor* species are common and predominantly saprotrophs [3]. These ubiquitous microorganisms may colonize multiple and contrasted environments from dungs or dead plant materials to plant and animal tissues. Members of the *Mucor* genus have an ambivalent impact on human activities. Regarding their negative impact, some *Mucor* species, in particular the thermotolerant species *M. indicus, M. ramosissimus* and *M. circinelloides,* have been shown to be human and animal pathogens responsible for mucormycosis [2]. Mucormycosis has been recently described as the third most common angioinvasive fungal infection and can lead to death [4]. Another negative impact concerns the ability of different species of the genus to spoil raw and transformed foods and feeds [5] On the contrary, some *Mucor* species have an important biotechnological potential thanks to their high growth rates in a large range of temperatures [6], existence of a yeast state in certain *Mucor* spp. [7], and high proteolytic and lipolytic enzymatic activities [8], making them good candidates for biotechnologists. Interestingly, some few species are also used in food manufacturing of Asian fermented food production (such as sufu, ragi, tempeh, furu or mureha) or for French cheese ripening (such as Tomme or Saint-Nectaire) [2].

The increasing number of infections associated with *Mucor* species as well as the biotechnological potential of the genus have led to a large effort to better know these fungi. In this context, Vongsangnak et al. (2018) proposed a metabolic network of the oleaginous strain *M. circicinelloides* WJ11 [9], Corrochano *et al*., (2016) shed new light on *Mucor* sensory perception [10] and multiple genes potentially involved in virulence were investigated and discovered in *Mucor* spp. [11-17]. Following the obtention and annotation of the first *Mucor* genome sequence (*M. circinelloides* CBS 277.49), researches on *Mucor* benefited and will continue benefit from different sequencing projects including the Zygolife

initiative (http://zygolife.org/home/) which aims to provide a better phylogenetical classification to the formerly called Zygomycetes which include the genus *Mucor* (see [1]).

This phylogenetical classification appears to be challenging as stated by inconsistencies among previous works; e.g. *M. indicus* CDC-B7402 placement was modified between the phylogeny of Álvarez et al. and Whalter et al. (2013), *M. endophyticus* CBS 385-95 placement was modified between Whalter et al. (2013) and Lebreton et al. (2018) [18-20]. Moreover, the uncertain taxonomic assignment of some *Mucor* strains used in published studies may lead to confusion. For example, following genomic studies, strain 97-1192 was reassigned from *M. racemosus* to *Rhizopus oryzae* and strain CDC-B9738 (initially *R. microsporus)* was consecutively assigned to *M. racemosus* by Chibucos et al. (2016) and more recently reassigned to *R. microsporus* by Gryganskyi et al. (2018) [11, 21]. As stated by Gryganskyi et al. (2018), the closely related genus *Rhizopus* can not be deciphered with a single or even a handful of gene families [21]. However the range of *Mucor* genome sequences exploited is limited and those available with annotations even scarcer. Furthermore, these genome sequencing projects are mainly limited to *Mucor* species with a biotechnological or pathogenic potential. Indeed, at the time of this study, only six *Mucor* annotated genomes were freely available, five of them corresponding to *M. circinelloides* (or potential synonyms of *M. circinelloides*) strains.

The present study aimed to use comparative genomics to identify potential genomic imprints of adaptation to different environments and lifestyles in the *Mucor* genus. To do so, four genomes, corresponding to *M. fuscus* UBOCC-A-109160 and *M. lanceolatus* UBOCC-A-109153 (used in cheese production, during ripening), *M. racemosus* UBOCC-A-109155 (a cheese spoiler sometimes described as an opportunistic pathogen) and *M. endophyticus* CBS 385-95 (a plant endophyte), were sequenced and compared to those of six publicly available *Mucor* and two *Rhizopus* (formerly identified as *Mucor*) isolates.

Materials and methods

Biological material

The genomes of twelve representative strains were investigated in the present study (Table 1). Four of them were sequenced in the framework of this study while the eight other were publicly available [10, 11, 22]. *M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153, *M. endophyticus* CBS 385-95 (UBOCC-A-113049) and *M. racemosus* UBOCC-A-109155 used for genome sequencing were cultivated in the dark at 25°C on PDA medium (Difco Laboratories, Detroit, Michigan). Spore suspensions of each strain were produced as previously described by Morin-Sardin et al. (2016) [6]. Concentrations were adjusted to 10^7 to 10^8 spores·mL^{-1} prior to storage at -80°C until use. For genomic DNA extraction, the fungal strains was grown on PDA solid medium at 25°C for 7 days.

Table 1: List of isolates used in this study and their reported habitat. Newly sequenced isolates are in red. *The two names are synonyms according to the U.S. National Fungus Collections (https://nt.ars-grin.gov). ** The two names are synonyms according to Walther et al. 2013 [20]. § Initially referenced as *Mucor racemosus* [11] but later assigned to *Rhizopus microsporus* [21].

Taxon	Strain	Strain isolation source	Reported habitat of the species	Reported habitat references	Genome reference or accession
M. endophyticus	CBS 385-95 (UBOCC-A-113049)	*Triticum aestivum*, leaves	Plant endophyte	[23]	This study
M. fuscus	UBOCC-A-109160	Cheese	Cheese	[24]	This study
M. lanceolatus	UBOCC-A-109153	Cheese	Cheese	[24]	This study
M. racemosus	UBOCC-A-109155	Cheese	Cheese, yogurt, walnuts, sausages, grassland soil, decaying vegetables, human	[24] [20] [25]	This study
M. ambiguus (syn. M. circinelloides)*	NBRC 6742	Unknown	Vegetable, tanned sole leather	[26]	BBKB00000000
M. circinelloides f.circinelloides	1006PhL	Skin of a healthy human	Cheese, sufu starter, decaying vegetables, human,	[18] [24] [20]	[28]

M. circinelloides f.lusitanicus	CBS 277.49 (UBOCC-A-108085)	Unknown	soda, air, soil, dung, sediment	[27]		[10]
M. circinelloides	*CDC-B8987*	Human BL line				[11]
*M. circinelloides f. janssenii (syn. M. velutinosus **)*	*CDC-B5328*	Human: nasal				[11]
M. indicus	*CDC-B7402*	Human: unknown	Human, dung, *Dioscorea tuber*, sorghum malt	[20] [29] [30]		[11]
R. microsporus (formerly M. racemosus §)	*CDC-B9645*	Clean room floor	Human, dust, sorghum malt, stored cereals	[20]	[11]	[11]
R. microsporus (formerly M. racemosus §)	*CDC-B9738*	Human abdomen				GCA_000697275.1

Genome sequencing and assembly

Genomic DNA from *M. fuscus* UBOCC-A-109160*, M. racemosus* UBOCC-A-109155*, M. lanceolatus* UBOCC-A-109153 and *M. endophyticus* CBS 385-95 was extracted from fresh mycelium, following the CTAB method proposed by the Joint Genome Institute (Kohler et al., 2011) with an optional step using Qiagen genome-tips (Qiagen). Due to the low efficiency of the CTAB method for *M. lanceolatus* UBOCC-A-109153, genomic DNA of this strain was also extracted following the protocol developed by Cheeseman et al. (2014) with a purification by a cesium chlorid gradient with DAPI [31].

Genomes were sequenced with Illumina technology (San Diego, CA) technology at different sequencing facilities (Table S1). For each of the four species, DNA were paired end sequenced (read length 2x100pb, insert size 500 bp). An additional mate pair sequencing was performed for *M. lanceolatus* UBOCC-A-109153 and *M. endophyticus* CBS 385-95 (read length 2x100pb, insert size 9-12kb). Sequences were quality checked with FastQC [32]. Adaptors were removed, reads were quality trimmed (bases kept had a phred score above 25) and reads shorter than 20pb were dropped with Cutadapt [33]. Mate pair reads of *M.*

lanceolatus UBOCC-A-109153 were mapped with STAR [34] on a preliminary version of the assembly (by providing only *M. lanceolatus* UBOCC-A-109153 paired end data to CLC Genomics Workbench -CLCbio, Seoul, Korea-). Mate pair reads separated by less than 500bp and oriented in forward-reverse were dropped. This new set of reads was used in further *M. lanceolatus* UBOCC-A-109153 assemblies. *M. lanceolatus* UBOCC-A-109153 and *M. endophyticus* CBS 385-95 were assembled using Velvet [35] (option "shortMatePaired", k-mer of 67 for *M. lanceolatus* UBOCC-A-109153 and k-mer of 85 for *M. endophyticus* CBS 385-95), while *M. racemosus* UBOCC-A-109155 and *M. fuscus* UBOCC-A-109160 were assembled with SOAPdenovo [36]. Genome assembly quality was checked with BUSCO v3 [37] using the fungal library and *Rhizopus* Augustus training.

Genome annotation of the four newly sequenced genomes

Genome assembly scaffolds were annotated using combinations of *ab initio* predictors, RNAseq data support and homology researches. As for *ab initio* predictors, Genemark-ES [38], with self-training, and Augustus [39], with *Rhizopus* training available within the toolAugustus, were used. RNAseq transcripts were extracted and sequenced as previously described in Lebreton et al. (2018) and reconstructed using two methods [19]: (i) by mapping RNAseq reads on genome with STAR [40] and reconstructing transcripts with Cufflinks [41], and (ii) by reconstructing transcripts *de novo* with Trinity [42] and mapping the transcripts on the genomes with gmap [43]. Predicted proteins of *M. circinelloides* CBS 277-49 [10], *Rhizopus delemar* RA-99880 [8] and *Phycomyces blakesleeanus* NRRL1555 [10] were searched on genomes with Exonerate [44]. Consensus gene models were generated from all predictions by EVidenceModeler [45].

The obtained gene predictions were functionally annotated as follows: transmembrane domains were predicted with TMHMM [46], peptide signal with SignalP v4 [47] and Pfam domains with HMMER [48] using the PFAM-A database [49]. Sequences homologies were searched using tBLASTx and BLASTp [50], (with an e-value threshold inferior to 10^{-5}, against Swissprot-Uniprot and Uniref90 databases as well as *M. circinelloides* CBS 277-49, *R.*

delemar RA-99880 and *P. blakesleeanus* NRRL*1555* filtered proteins obtained from the JGI platform [51]. EC numbers were predicted using PRIAM [52] and were transferred from homology researches. GO terms were transferred from homology researches. Gene names were assigned with AHRD (Automated Assignment of Human Readable Descriptions) available on Github (https://github.com/groupschoof/AHRD).

Non coding RNA were predicted with tRNAscan-SE [53], RNAmmer [54] and Infernal [55] using the Rfam database [56]. The obtained data were integrated in an instance of the genome viewer Apollo [57] allowing experts to validate gene prediction quality and perform manual curation.

Complementary annotation of the full set of genomes

Transposable elements (TE) were annotated using the REPET pipeline [58] that includes a *de novo* prediction and TE classification [59]. Carbohydrate-active enzymes were searched using dbCAN2 [59] which searches for sequences available in the CAZy database [60] with HMMER (E-Value < $1e^{-15}$, coverage > 0.35), DIAMOND [61] (E-Value < $1e^{-102}$) and Hotpep [62] (Frequency > 6.0, Hits > 2.6). Only annotations predicted by at least two different tools were subsequently considered. Peptidases and their inhibitors available in the MEROPS database [63] were searched using BLASTp (E-Value 10^{-5}).

Syntheny search

Syntheny blocks were searched among the twelve studied strains using Mummer [64] and Mauve [65].

Phylogenetic reconstruction

Predicted proteomes of the twelve studied strains, as well as those of *R. delemar* RA-99880 and *P. blakesleeanus* NRRL1555 (both latter species being considered as outgroups), were compared based on sequence similarity to identify orthologous proteins using the Orthofinder v.2.2.0 software [66] (E-value 10^{-5}, inflation 1.5). The 64 obtained single copy orthologs were selected to reconstruct the phylogeny of the studied species. Multiple alignment was inferred

using PRANK v.170427 [67], run with default settings. Spuriously aligned regions were excluded with TrimAl v1.4.r15 [68] with a 0.2 gap threshold. Based on the alignments, 13 orthogroups were manually discarded due to low percent of identical sites or high number of gaps among orthologs. Subsequent alignments were concatenated in a supermatrix of 23398 sites. This matrix was used to reconstruct species tree by Maximum Likelihood inference and by Bayesian Monte Carlo Markov Chain (MCMC) samples. RAxML PTHREADS v. 8.2.9 [69], a program for Maximum Likelihood based inference, was used with a partitioned LG+G model, in which each data partition represented a single input gene family. A bootstrap analysis with 100 replicates under the same model was performed in RAxML in order to assess tree branch support. Alternatively, the PhyloBayes v3.3 MCMC samplers [70] was used with a CAT+GTR model and 3 chains. The RAxML tree obtained was used to estimate the divergence time between species with the Langley-Fitch method with r8s v1.8 [71] by calibrating against the assessed origins of *P. blakesleeanus* NRRL1555 and *R. delemar* RA-99880 at 468 MY [72].

Evolution of genes families

Based on OrthoFinder results and the obtained ultrametric tree, expansion and contraction of gene families were reconstructed with CAFE v4 [73]. Birth and death parameters were estimated independently using orthologous groups containing less than 75 genes per strain. The analysis was done in two parts: (i) on all strains except *R. microsporus* CDC-B9645 and *R. microsporus* CDC-9738, and (ii) only on *R. microsporus* CDC-B9645 and *R. microsporus* CDC-9738. Rapidly evolving families were predicted by CAFE using the Viterbi algorithm.

Focus on specific genes families

Gene cluster associated with secondary metabolites were searched with FungiSMASH [74] and SMURF [75]. Crucial genes (Polyketide Synthase (PKS), Non Ribosomal Peptide Synthetase (NRPS), Terpene Synthase (TPS), DiMethylAllyTryptophane Synthase

(DMATS)) involved in secondary metabolism and genes potentially involved in adaptation to the environment were searched in each species.

Availability of supporting data

Genomes assembly and annotation are available on the ABiMS platform (http://application.sb-roscoff.fr/project/mucor_project/) and the Mycocosm platform of the JGI (https://genome.jgi.doe.gov/programs/fungi/index.jsf). Noteworthy, functional annotations found on the Mycocosm platform differ from the ones exploited here since the structural annotation was functionally re-annotated by the pipelines associated with the Mycocosm platform.

Results

Genome description

Genome sequences and assembly

The genomes of four strains were sequenced, assembled and annotated in the context of this study. Among them, three corresponded to cheese isolates (*M. fuscus* UBOCC-A-109160 and *M. lanceolatus* UBOCC-A-109153 used for cheese ripening, and *M. racemosus* UBOCC-A-109155 identified as a cheese contaminant [24]), while the fourth one was a wheat endophyte (*M. endophyticus* CBS 385-95, [23]) (Table 1). Their genome features were compared to eight previously sequenced and annotated genomes from six *Mucor* strains and two *Rhizopus* strains formerly identified as *Mucor s*pp. [10, 11, 22]. Among them, four corresponded to isolates collected from human/clinical environments (Table 1).

The number of scaffolds in *M. circinelloides* CBS 277-49 assembly was the lowest with 21 scaffolds. *M. endophyticus* CBS 385-95 was the second less fragmented assembly with 159 scaffolds. The ten other assemblies were composed of 470 to 3888 scaffolds (Table 2). Despite these differences in genome fragmentation, at least 95% of the 290 single copy fungal orthologous genes searched by BUSCO were found complete in all genomes. It is worth noting that respectively 52% and 87% of the searched genes were found duplicated in *R. microsporus* CDC-B9645 and *R. microsporus* CDC-9738 whereas, in the other species, duplicated genes only represented 17 to 26% of the searched genes. The two *Rhizopus* strains also exhibited the highest genome size with 65Mb and 75Mb for *R. microsporus* CDC-B9645 and *R. microsporus* CDC-9738, respectively while the average genome size for the other strains is approximately 39 Mb. Strain CBS 385-95 pertaining to the endophytic species *M. endophyticus* CBS 385-95, had the smallest genome size with 35Mb (Table 2).

Table 2: Genome assembly features and structural annotation of the ten *Mucor* strains and two *Rhizopus* strains. The four species newly sequenced in the context of this work are in red. IG: intergenic regions, TE: transposable elements.

Taxon	*M. endophyticus* CBS 385-95	*M. fuscus* UBOCC-A-109160	*M. lanceolatus* UBOCC-A-109153	*M. racemosus* UBOCC-A-109155	*M. ambiguus* NBRC 6742	*M. circinelloides* 1006PhL	*M. circinelloides* CBS 277.49	*M. circinelloides* CDC-B8987	*M. circinelloides* CDC-B5328	*M. indicus* CDC-B7402	*R. microsporus* CDC-B9645 §	*R. microsporus* CDC-B9738 §
Genome size (Mb)	35,00	40,60	43,42	46,92	40,74	36,35	36,57	36,77	35,61	39,62	64,72	74,89
# Scaffolds > 1000pb	159	3 819	1 531	3 506	1 283	470	21	1 022	1 016	798	3 888	2 676
%GC	34	36	34	32	32	37	42	39	40	36	32	33
%N	2	0	5	0	22	6	0	0	0	0	0	0
Maximum scaffold (kb)	4 527	142	681	162	539	664	6 050	340	424	831	518	3 513
N50 (kb)	1 957	27	142	25	114	141	4 318	92	79	271	38	90
N90 (kb)	575	4	13	7	26	51	1 075	23	21	33	10	14
# Genes	11 799	12 571	10 924	11 604	11 726	12 410	11 936	10 437	9 997	11 703	15 153	21 229
Gene density (gene/Mb)	337	310	252	247	288	341	326	284	281	295	234	283
Ave.gene length (pb)	1 677	1 542	1 677	1 633	1 616	1 491	1 408	1 609	1 608	1 491	1 517	1 543
Ave. exon frequency	4,54	4,33	4,55	4,27	4,05	3,83	3,73	4,09	4,11	3,99	4,00	4,16
Ave. exon length (pb)	313	302	301	335	340	340	311	336	335	316	324	318
% genes with introns	86	82	87	84	81	82	81	83	84	83	81	83
Ave. intron length (pb)	82	76	95	72	77	66	91	76	74	76	74	69
Ave. IG size (pb)	1 293	1 284	2 040	1 791	1 673	1 374	1 691	1 677	1 758	1 760	2 151	1 713
# Prot. predicted	11 437	12 310	10 636	11 269	11 343	12 227	11 709	10 190	9 716	11 390	14 656	20 723
%TE coverage	5,08	14,28	22,8	36,8	6,03	15,2	23,44	22,18	24,43	16,51	38,47	34,46

§ Initially assigned to *Mucor racemosus* [11] but later reassigned to *R. microsporus* [21].

Genome annotation.

The number of predicted genes were in accordance with genome sizes with respectively 15153 and 21229 predicted genes for *R. microsporus* CDC-B9645 and *R. microsporus* CDC-9738, *i.e.* a gene density of 234 and 283 genes/Mb, respectively. In *Mucor* strains, the number of predicted genes fluctuated from 9997 to 12571, *i.e.* a gene density ranging from 234 to 341 genes/Mb (Table2).

Gene characteristics were well conserved among the *Mucor* and *Rhizopus* strains: the average gene length was 1568bp, 83% of genes had predicted introns, genes had an average frequency of 4.1 exons and average intron size was approximately 60bp.

Noteworthy, within all genomes, the intergenic distance was variable: 25% of intergenic regions were shorter than ~300pb whereas the largest intergenic regions exceeded 20kb.

When repeated elements were taken into account for the analysis, regions up to 15kb with neither gene nor repeated elements were still detected (Figure 1).

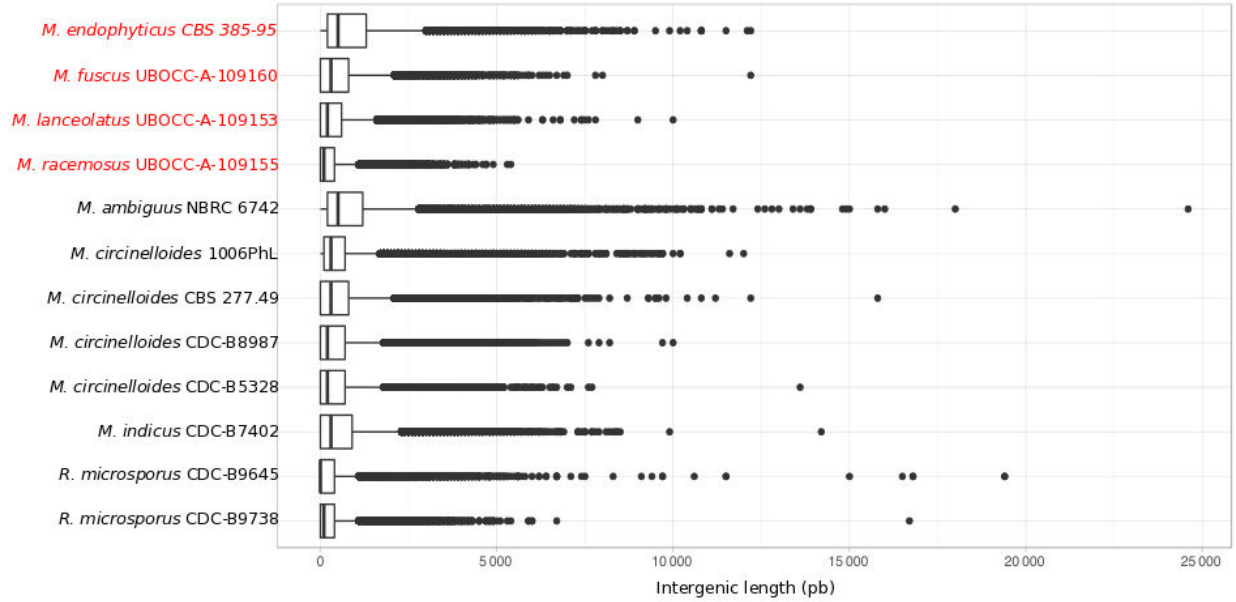


Figure 1: Repartition of the intergenic length within the twelve studied *Mucor* and *Rhizopus* genomes. The four species newly sequenced in the context of this study are in red.

The boxes represent intergenic length arboured by 25% to 75% (sorted by length) of the intergenic regions, the line within the box represent the median length, dots represents each values corresponding to the 10% highest lengths represented.

Genome comparisons

Synteny blocks.

Synteny was investigated in order to better understand the evolution of the *Mucor* genome architecture; however, even for phylogenetically close species, no synteny at the genome scale was observed among the twelve analysed genomes.

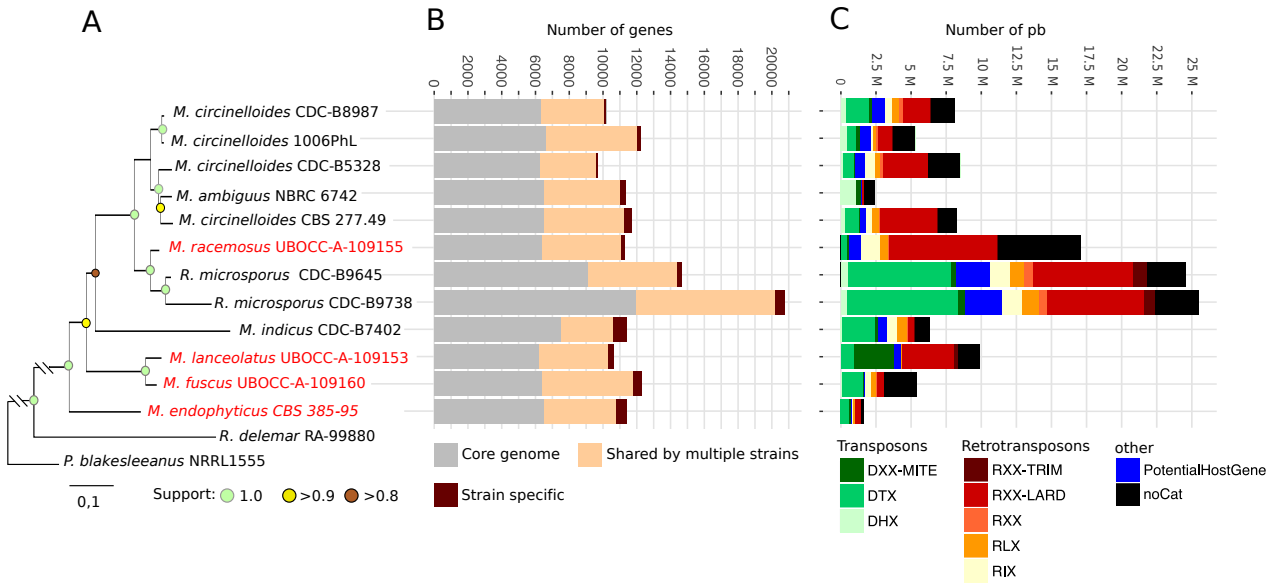
Phylogenomic reconstruction.

The 181601 predicted proteins of the ten *Mucor* strains, two *R. microsporus* strains, *R. delemar* RA-99880 and *P. blakesleeanus* NRRL1555 strains were grouped in 20588 orthogroups. Among them, 4240 orthogroups were composed of predicted proteins belonging to all strains while 64 were composed of single copy orthologs.

Among the 64 single copy orthogroups, 52 were selected to reconstruct the species tree using Maximum Likelihood and Bayesian methods, *R. delemar* RA-99880 and *P. blakesleeanus* NRRL1555 being used as outgroups (Figure 2A). The phylogeny of the present study confirmed the placement of *M. ambiguus* NBRC 6742 within the *M. circinelloides* clade as stated in the U.S. National Fungus Collections (https://nt.ars-grin.gov), it will therefore be considered heareafter as a synonym of *M. circinelloides*. The placement of *M. indicus* CDC-B7402 was altered compared to the topology obtained by Whalter et al. (2013) but concurred with the topology published by Álvarez et al. [18, 20]. Similarly, the placement of *R. microsporus* CDC-B9645 *and R. microsporus* CDC-B9738 was concordant with the topology of Chibucos et al. (2016) which identified these two strains as *M. racemosus* strains, but differed from the one by Gryganskyi et al. (2018) in which the two strains were clustered with *R. microsporus* spp. (and renamed accordingly to this result) [11, 21], thus raising questions concerning their actual position and the genetic bases associated with these incongruences. The other studied strains had concordant phylogenetic placements with previously published studies [11, 18-20].

From the biological point of view, the *M. fuscus* UBOCC-A-109160 and *M. lancelotatus* UBOCC-A-109153 strains, that are close in the tree, pertain to species mainly encountered in cheeses, the basal singleton strain *M. endophyticus* CBS 385-95 belong to a species only described as a wheat leaf endophyte whereas the other *Mucor* strains that can be grouped in the same clade corresponded to species which are described as potential pathogenic species.

Figure 2: Phylogenomic tree and gene and TE elements representation for the studied *Mucor* and *Rhizopus* strains (initially assigned to *M. racemosus*) . A: Phylogenomic tree of the studied strains reconstructed with RAxML based on 52 single copy ortholog families using *R. delmar* RA-99880 and *P. blakesleeanus* NRRL1555 as outgroup (not shown). The branch length corresponds to the number of substitutions per site. The four species sequenced in this study are in red. **B:** Representation of the gene number in the studied strains. **C:** Representation of the TE coverage in each genome depending on the TE category (classification of Wicker et al. 2007). DXX-MITE: unknown non-autonomous transposon, MITE-like. DTX: TIR transposon. DHX: Helitron transposon. RXX-TRIM: unknown non-autonomous retrotransposon, TRIM-like. RXX-LARD: unknown non-autonomous retrotransposon, LARD-like. RXX unknown retrotransposon. RLX: LTR retrotransposon. RIX: LINE retrotransposons. noCat: potential transposable element that could not be identified.



Transposable elements

The studied genomes contained contrasting transposable element (TE) coverages (Figure 2C). The plant endophyte, *M. endophyticus* CBS 385-95, held the lowest TE coverage (5%) whereas the ubiquist *M. racemosus* UBOCC-A-109155 contained the highest *Mucor* TE coverage (37%). Even between close species, TE coverage and composition notably differed, e.g. between *M. lanceolatus* UBOCC-A-109153 and *M. fuscus* UBOCC-A-109160 or between *M. ambiguus* NBRC 6742 (*M. circinelloides* according to the phylogenomic analysis performed in this study) and *M. circinelloides* strains. *M. lanceolatus* UBOCC-A-109153 showed higher TE coverage than *M. fuscus* UBOCC-A-109160 (23% and 14%, respectively) for both transposons and retrotransposons but the predicted elements in the *M. lanceolatus* UBOCC-A-109153 strain were mainly non-autonomous (*i.e.* not including all the domains necessary for their transposition). On the contrary, almost all *M. fuscus* UBOCC-A-109160 predicted TE were autonomous, the main represented category being terminal inverted repeat (TIR) but few retrotransposons, long interspersed nuclear elements (LINE) and long terminal repeat (LTR), were also detected.

Among the *M. circinelloides* clade, four strains displayed a similar pattern in terms of TE composition and, except for *M. circinelloides* 1006PhL, in terms of coverage. The fifth one, *M. ambiguus* NBRC 6742, displayed striking differences from the other genomes. Indeed, the

TE coverage of this strain was reduced by at least two folds compared to the other strains (Figure 2). Furthemore, the main TE category was identified as belonging to the helitrons order, a marginal category in the four other strains of the clade.

Evolution of genes families

When using the whole genome dataset, CAFE-, DupliPhyML- and Notung-based analyses yielded non concordant results (data not shown) with inconsistent placements of expansion/contraction events within the cluster encompassing *R. microsporus* CDC-9738*, R. microsporus* CDC-9645 and *M. racemosus* UBOCC-A-109155. This behaviour was interpreted as a side effect of the putative whole genome duplication or hybridization observed in the genomes of the two *R. microsporus* strains. The phylogenetic placement of these two strains within the *Mucor* genus being also questioned, we decided to remove them from the CAFE analysis. CAFE identified 44 rapidly evolving gene families on the *M. lanceolatus* UBOCC-A-109153/*M.fuscus* UBOCC-A-109160 branch (pertaining to the two species associated to cheese ripening). Among these families, two were associated to secondary metabolism, namely an acyl-CoA synthetase and a cytochrome p450 encoding gene families, both with reduced number of genes in cheese ripening strains (*M. fuscus* UBOCC-A-109160 and *M. lanceolatus* UBOCC-A-109153) compared to other strains. A cysteine hydrolase gene family was also reduced in the two cheese-associated strains. Another family (less conserved) with genes identified as encoding putative transcriptional activators of glycolytic enzyme was expanded in the cheese technological species genomes. Other families were either unknown or similar to TE sequences.

In the endophyte *M. endophyticus* CBS 385-95, at the node separating technological species from pathogenic species, and at the node separating *M. indicus* CDC-B7402 from other species, 49, 4 and 9 gene families were considered as rapidly evolving, respectively. However, these gene families were either of unknown function or similar to TE sequences.

Comme explicité en 4.1 cette partie sera amendée avant la préparation finale de l'article et sa soumission.

Focus on genes involved in secondary metabolism

Since secondary metabolites can provide a significant advantage for survival in a given habitat and might vary depending on the fungus lifestyle, our study investigated genes encoding Polyketide Synthase (PKS), Non Ribosomal Peptide Synthetase (NRPS), Terpene Synthase (TPS) and DiMethylAallyTryptophane Synthase (DMATS). Among genes associated with terpene biosynthesis, some corresponding to squalene cyclases, squalene synthases, bifunctional lycopene cyclase, squalene/phytoene synthases and geranylgeranyl pyrophosphate (*ggpp*) synthases were identified in each strain. The number of genes in each category was well conserved among strains except for *R. microsporus* CDC-B9738 and CDC-B9645 which had a number of gene higher by twofold than in the other strains (Table 3).

Table 3: Number of genes involved in secondary metabolites found in the twelve studied *Mucor* and *Rhizopus* strains. The four species sequenced in this study are in red. For each gene category, maxima are highlighted in orange and minima in blue.

	Taxon	*M. endophyticus* CBS 385-95	*M. fuscus* UBOCC-A-109160	*M. lanceolatus* UBOCC-A-109153	*M. racemosus* UBOCC-A-109155	*M. ambiguus* NBRC 6742	*M. circinelloides* 1006PhL	*M. circinelloides* CBS 277.49	*M. circinelloides* CDC-B8987	*M. circinelloides* CDC-B5328	*M. indicus* CDC-B7402	*R. microsporus* CDC-B9645	*R. microsporus* CDC-B9738
Terpenes	Squalene cyclase	1	1	1	1	1	1	1	1	1	1	2	2
	Squalene synthase	3	3	3	3	3	3	3	3	3	3	4	6
	Lycopene cyclase & squalene/phytoene synthase*	1	1	1	1	1	1	1	1	1	1	2	2
	Geranylgeranyl pyrophosphate synthase	4	5	5	5	3	5	5	5	5	4	6	8
NRPS	NRPS	1	1	1	1	1	1	1	1	1	1	3	1
	NRPS-like 3/4 domains	0	0	0	0	0	0	1	0	0	1	0	1
PKS	PKSI	2	1	1	2	2	2	2	2	2	3	1	2
	3-oxoacyl synthase II (PKS-like)	1	1	1	1	1	1	1	1	1	1	1	2

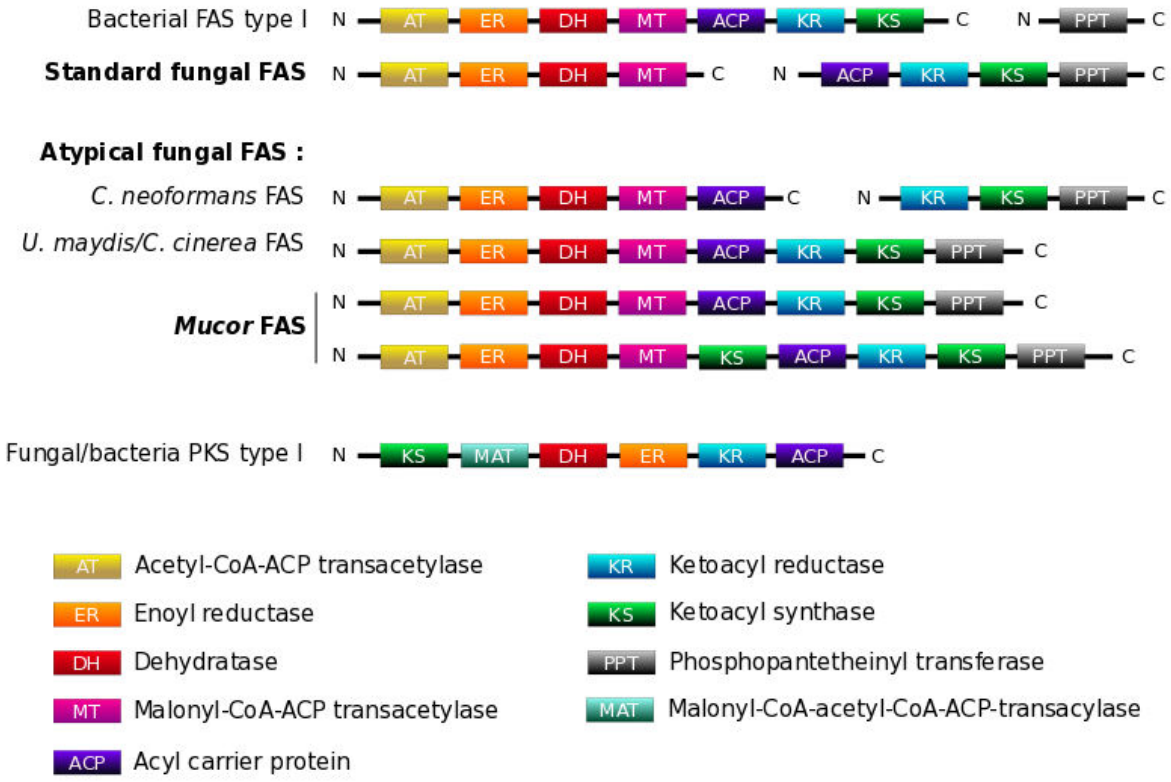
* these genes encode a putative bifunctional enzyme

In all studied strains, a single complete *nrps* gene; *i.e.* having at least one condensation domain, one carrier domain, one phosphopantetheine attachment site and one AMP-binding domain; was detected except in *R. microsporus* CDC-B9645 which had three complete NRPS-encoding genes. Three other genes containing three of the four mandatory NRPS domains were found: one in *R. microsporus* CDC-B9738 and *M. indicus,* both of them lacking the condensation domain, and one in *M. circinelloides* CBS 277.49, lacking the AMP-binding domain.

In all strains, no gene encoding for DMATS were identified.

Genes similar to PKS I encoding genes were detected in all the investigated genomes; three genes were identified in *M. indicus* CDC-B7402, one in *M. fuscus* UBOCC-A-109160 and *M. lanceolatus* UBOCC-A-109153 (both pertaining to cheese ripening species) and in *R. microsporus* CDC-B9645 and two in the other studied strains. The different domains of these genes were close in terms of composition and organisation to typical fungal Fatty Acid Synthases type I (FAS) domains with a major variation. Typical fungal FAS domains are encoded by two distinct genes, whereas in the investigated *Mucor* species all the domains were found within a single gene (Figure 3). Regarding their composition and structure, these genes were identified as putative FAS genes. Among the different *Mucor* putative FAS genes, three held an additional KS domain: one in *M. circinelloides* CDC-B5328*,* one in *M. indicus* CDC-B7402 and one in *M. racemosus* UBOCC-A-109155. Noteworthy, these genes were not expressed on PDA medium [19]. None of the potential secondary metabolism associated genes determined in the different *Mucor* genomes were organized in metabolic clusters.

Figure 3: Domain organization of FAS within the studied *Mucor* genomes in comparison to other reported FAS organizations. Fungal/bacterial PKS type I was presented for purpose of comparison. *C. neoformans*: *Cryptococcus neoformans. U. maydis: Ustilago maydis. C. cinerea: Coprinopsis cinerea.*



Iron uptake

Iron uptake is determinant for virulence but also for development in iron-depleted medium such as cheese. Homologs of genes encoding proteins involved in the different iron uptake mechanisms identified so far in fungi [76] were found in the analyzed *Mucor* genomes (Table 4; Figure 4).

Regarding the siderophore mediated iron uptake, genes playing a role in the carboxylate and hydroxamate siderophore synthesis were searched for. At least one ortholog of the *R. delmar rfs* gene, necessary for the carboxylate siderophore rhizoferrin production [77] was found in

each *Mucor* strain. Other genes that might be involved in this rhizoferrin mediated iron uptake mechanism were identified based on their homology to the bacterial genes of the *Francisella tularensis* rhizoferrin operon [78]. Homologs of the *FslB* and *FslC F. tularensis* genes were detected in each *Mucor* genome, numerous potential *Mucor* genes belonging the major facilitator family matched to *FslD* but *FslA*, *FslE*, *FslF* and the operon regulator *Fuf* could not be detected in *Mucor* genomes. Genes involved in hydroxamate siderophore synthesis could not be detected but predicted orthologous genes corresponding to the *Aspergillus* MirB siderophore permease (group 1) encoding gene and another gene coding for a MirB-like siderophore permease (group 2) were identified (Table 4).

Regarding the reductive iron acquisition iron uptake, homologous sequences of the gene encoding the *FTR1* high-affinity permease and *fet3* ferroxidase genes were detected (*fet3a* was not detected in *M. fuscus UBOCC-A-109160,* and *M. lanceolatus UBOCC-A-109153* and *M. racemosus* UBOCC-A-109155 genomes). Except for the FTR1 encoding gene, no gene involved in heme uptake was identified. The FET4 low affinity iron permease encoding gene was identified in the different *Mucor* genomes. When focusing on the iron uptake regulation, homologs to the *SreA* iron uptake repressor gene were detected but no gene involved in activation of iron acquisition pathways such as *HapX* or *Aft* genes have been identified yet. Two ferritin encoding genes of similar lengths were present in each studied strain (except *R. microsporus* CDC-B9546 and CDC-B9738 which had 4 genes), in each case, the two genes shared 70% similarity and 40% identity.

Table 4: Number of genes involved in iron uptake found in the twelve studied *Mucor* and *Rhizopus* strains. The four species sequenced in this study are in red. For each gene category, maxima are highlighted in orange and minima in blue. Proteins encoded by the different genes and their role in iron uptake mechanisms are presented in Figure 4.

		Taxon	*M. endophyticus* CBS 385-95	*M. fuscus* UBOCC-A-109160	*M. lanceolatus* UBOCC-A-109153	*M. racemosus* UBOCC-A-109155	*M. ambiguus* NBRC 6742	*M. circinelloides* 1006PhL	*M. circinelloides* CBS 277.49	*M. circinelloides* CDC-B8987	*M. circinelloides* CDC-B5328	*M. indicus* CDC-B7402	*R. microsporus* CDC-B9645	*R. microsporus* CDC-B9738
uptake	Siderophore (sid) mediated iron uptake	Sid-transporter: mirB-like group 1	0	0	1	1	1	1	1	1	1	1	1	1
		Sid-transporter: mirB-like group 2	1	1	1	1	1	1	2	1	1	2	2	2
		Sid-transporter: fslB	2	1	2	2	2	2	2	2	2	2	3	5
		Sid-biosynthesis: rfs	1	1	1	1	1	1	1	1	1	2	2	2
		Sid-biosynthesis: fslC	1	1	1	1	1	1	1	1	1	1	2	2
	Reductive Iron Acquisition (RIA)	Ferric reductase: Fre	4	4	3	5	4	4	4	3	4	3	7	7
		High-affinity iron permease: FTR1	2	1	1	1	2	2	2	2	2	2	2	3
		Ferroxidase: fet3a	1	0	0	0	1	1	1	1	1	1	1	1
		Ferroxidase: fet3b	1	1	1	1	1	1	1	1	1	2	1	2
		Ferroxidase: fet3c	1	1	1	1	1	1	1	1	1	2	1	1
		Ferrioxamine binding: Fob	2	2	1	2	2	2	2	2	2	2	2	3
	Heme degradation	Heme oxygenase: HOXG	2	2	2	2	2	2	2	2	2	3	3	3
	Low affinity iron permease	Low affinity iron permease: FET4	1	1	1	1	2	2	2	2	2	3	1	1
Iron regulation		Transcription factor: sreA	2	2	2	2	2	2	2	2	2	1	5	4
Iron utilization		Iron-binding protein frataxin: yfh1	1	1	1	1	1	1	1	1	1	1	1	2
Iron storage		Vacuolar iron importer: ccc1	2	2	2	2	2	2	2	2	2	2	4	3
		Vacuolar iron transporter: smf3	3	3	3	3	3	3	3	3	2	3	3	6
		Ferritin	2	2	2	2	2	2	2	2	2	2	4	4

Interestingly, apart from *R. microsporus* CDC-B9546 and CDC-B9738 genomes, the number of genes involved in iron uptake was always higher in the genome of the opportunistic pathogen *M. indicus* than in the other genomes (Table 4). Indeed, in the *M. indicus CDC-B7402* genome*,* (i) for the siderophore pathways, the *rfs* rhizoferrin biosynthesis gene was duplicated as well as a MirB-like siderophore permease encoding gene; (ii) for the reductive iron acquisition (RIA), the *fet3b* and *fet3c* ferroxidase genes were duplicated; (iii) for heme degradation, a supplementary heme oxygenase gene homolog was detected and, (iv) for direct Fe^{2+} uptake, three orthologs of the FET4 low affinity permease were identified, whereas two genes were found for *M. circinelloides* spp. and one for the others studied strains. Finally, only one copy of the *sreA*-like gene, involved in down regulation of iron acquisition, was found in *M. indicus* CDC-B7402 when at least two copies were found for all other strain.

On the contrary, cheese ripening species exhibited a reduced number of genes involved in iron uptake within their genomes *MirB*-like and *fslB*-like siderophore permease encoding genes were absent from the *M. fuscus* UBOCC-A-109160 genome, *M. lanceolatus* UBOCC-A-109153 lost one copy of the cell surface receptor *fob* gene and both strains/species lacked the *FTR1* high affinity permease and *fet3a* ferroxidase genes. The latter genes were also absent from the cheese contaminant *M. racemosus* UBOCC-A-109155 genome.

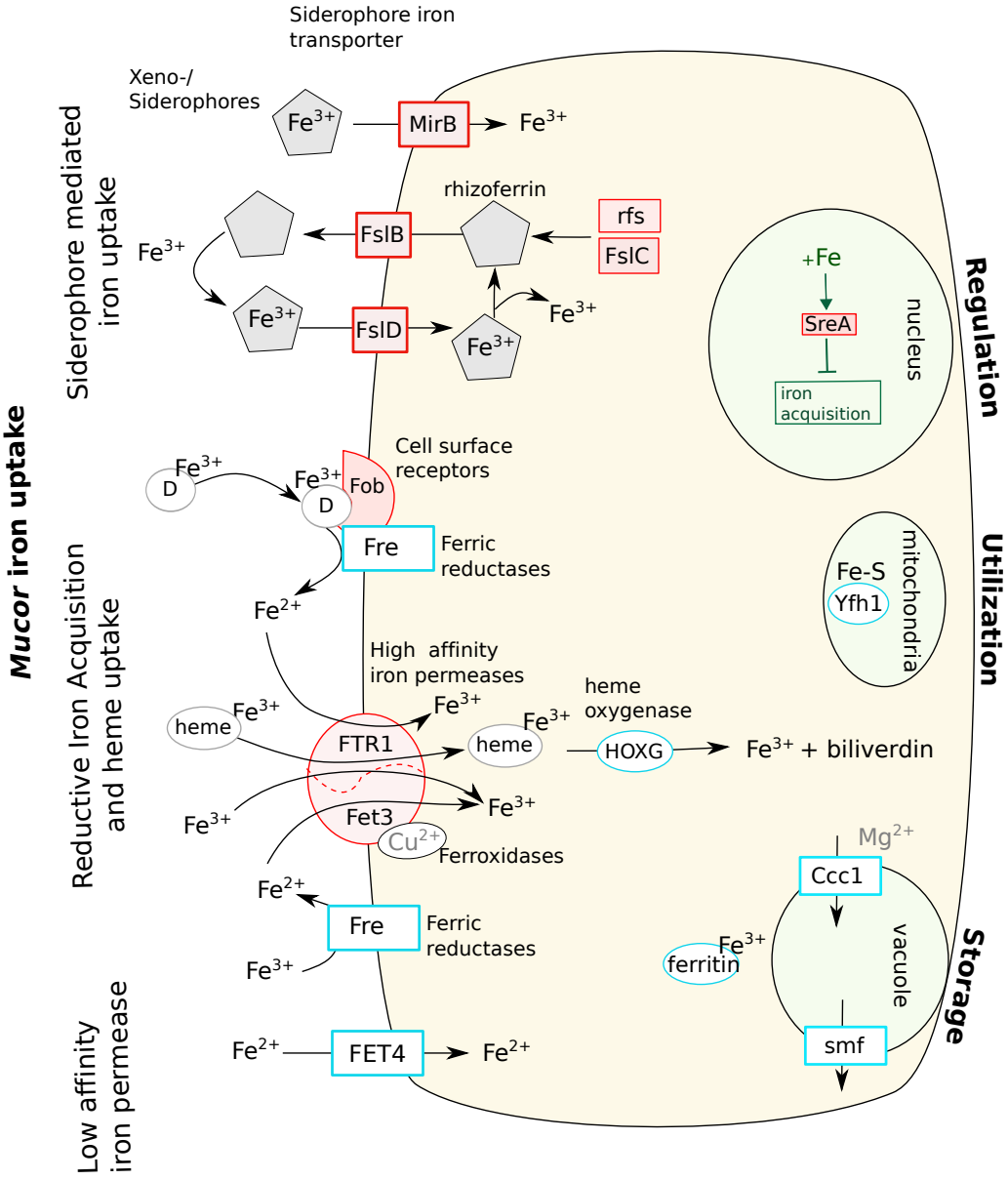


Figure 4: Proposed mechanisms of iron uptake by *Mucor spp*. as well as elements used for iron regulation, utilization and storage. Elements in red were described in literature as important for *Mucor* pathogenicity. D: Deferoxamine.

Discussion

Although benefiting from a growing interest due to their involvement in mucormycosis but also their biotechnological potential, a restricted number of *Mucor* genomes are available to this date, leading to scarce whole genome comparative studies: [11, 13, 79]. These comparative studies mainly focused on human/clinical environments species. Yet, only a handful of *Mucor* species are known to cause human infections [2]. In the context of this study, a comparative analysis of four newly sequenced *Mucor* genomes isolated from non-clinical environments (cheese and plant) with the six publicly available *Mucor* genome sequences provided a better overview of the *Mucor* pan-genome and allowed identifying gene candidates that could contribute to habitat and lifestyle adaptation within this diverse genus. The determined phylogeny, based on a large set of orthologous genes (52), that integrated the four genome sequences obtained in this study was non concordant with previously published phylogenies. The most ambiguous phylogenetic placement concerned the *R. microsporus* CDC-B9738 and CDC-B9645 strains, as they appeared as a sister group of *M. circinelloides* and closely related to *M. racemosus* both in the present study and in Chibucos et al. (2016) (76 analyzed orthologs) while they clustered into the *R. microsporus* clade with *M. circinelloides* species as an outgroup in Gryganskyi et al. study (2018) (192 orthologues analyzed) [11, 21]. Topologies may vary depending on the selected genes and on the reconstruction pipeline. If, as highlighted by Gryganskyi et al. (2018), single-gene phylogenetic reconstructions are of limited value, it is worth noting that approaches based on the analysis of large sets of genes may also generate inconsistent topologies [21]. The contrasted placements of the two *R. microsporus* strains among the different studies may arise from the whole/partial-genome duplication events and/or hybridization observed in their genomes [21]. Indeed, when considering gene families involved in secondary metabolism and iron acquisition (which were investigated in this study for comparative genomics approaches), CDC-B9738 and CDC-B9645 genes were often found duplicated. Phylogenetic

reconstructions performed using the duplicates yielded incongruent topologies with either CDC-B9738 or CDC-B9645 strains clustering with *R. delemar* clade or alternatively with *M. racemosus* clade suggesting a possible hybrid genomic content for these two strains.

Whatever the phylogenetic placement of the strains or their proposed habitat/lifestyle, the current study revealed that the gene features among *Mucor* species (gene number, size, exon length etc.) were globally conserved. However, as shown by the lack of macro and micro synteny, species within this genus experienced extensive genomic rearrangements. These rearrangements can be partially explained by transposable elements (TE) which displayed high degree of diversity within the available genomes and has already been reported to have a major role in fungal genomic diversity and genome evolution [80].

Interestingly, large regions (above >5kb) with neither gene nor TE were systematically observed throughout the *Mucor* genomes. Although lack of synteny among the different genomes did not allow to check whether or not these regions were conserved among species, this raises the question regarding their putative role and evolutive advantage. Among different hypotheses these regions might be involved in three-dimensional folding of the chromatin, plasticity of the genome or might hold unknown functional structures.

The average genome size of the *Mucor* species analyzed in this study (39Mb) is congruent with what is observed at the scale of Mucoromycotina subphylum (38Mb) and also in Ascomycota (37Mb). The gene number (9997 to 12571) is also in concordance with the gene numbers observed in Mucoromycotina and Ascomycota [81]. The core genome detected in this study includes approximately 6000 genes. The analysis of the functions of genes included in the core genome as well as those pertaining to single species or to species sharing similar lifestyles (e.g. cheese ripening species or opportunistic pathogens) will inform on genes possibly involved in lifestyle adaptation in the genus *Mucor* (this task is under progress).

While the presence or absence of metabolic pathways involved in primary metabolism can inform about the feeding ecology of different fungi, examination of the presence or absence

of secondary metabolic pathways can shed light on fungal ecology and adaptation to a particular lifestyle [82]. Thus, in this study, we investigated more specifically genes that are essential to most of the secondary metabolites biosynthesis (PKS, NRPS, DMATS…) but also genes involved in iron uptake which have been shown to be fundamental both for fungal virulence in opportunistic pathogens [83, 84] (a category represented in this study by the *M. indicus* CDC-B7402 or *M. circinelloides* strains) and for cheese colonizing microorganisms (represented in this study by the *M. lanceolatus* UBOCC-A-109153 and *M. fuscus* UBOCC-A-109160) given that cheese is a highly iron-restricted medium [85]. The search for genes involved in secondary metabolite biosynthesis are often initiated by the search for PKS, NRPS or DMATS genes clustered with additional genes involved in the biosynthesis pathway considered. Indeed, in higher fungi, genes involved in secondary metabolism are usually located in metabolic gene clusters (MGCs) (for a review see [82]) MGCs being defined as "close linkage of two or more genes that participate in a common metabolic or developmental pathway" [86]. This characteristic has often been extended to the whole fungal kingdom [82, 87-89] although for some authors, this assertion is premature due to the scarcity of information [3]. Indeed, the few studies depicting MCG in early diverging fungi relied on automatized analyses with no manual checking of the gene participation to a common metabolic or developmental pathway [82, 90]. A group formed by *carRP* and *carB* genes associated with their promoters at a single genome location have been identified in *M. circinelloides* CBS 277.49 [91] and detected in the present study within some of the *Mucor* genomes associated with FAS/PKS gene (*M. fuscus* UBOCC-A-109160 and *M. endophyticus* CBS 385-95) but no typical MGCs involved in secondary metabolism and including a PKS, FAS, NRPS, NRPS-like gene or terpene synthesis related genes were detected within the *Mucor* spp. genomes. Still, secondary metabolites pathways have been characterized in Mucorales (see [3]) and different genes involved in secondary metabolism were identified in the present and previous studies. MGC clustering might represent an adaptation against the accumulation of toxic intermediates within biosynthetic pathways but also might facilitate fungal adaptation by enabling wholesale acquisition (through duplication or HGT) or loss of

entire metabolic pathways (see [82]). The apparent absence or at least scarcity of MGCs within the *Mucor* genus (and possibly at a broader scale within the Mucoromycotina) might appear enigmatic since the genus encompasses species with contrasting lifestyles similar to what exists in Ascomycetous fungi in which MCGs are abundant. This raises the question why the selective advantage conferred by MCGs to higher fungi would not apply to *Mucor* species. The answer to this question might lie in some ecological specificities of the Mucorales which are considered as ruderal species avoiding stress and competition [92, 93] or to structural specificities such as their coenocytic structure [94] ,and may shed new light regarding the evolution of MCGs in eukaryotes. All the *Mucor* species genomes investigated here included a NRPS gene which role has still to be determined as well as different genes encoding for enzymes of the terpene biosynthesis. As discussed in Voigt et al., (2016), no PKS were encountered in the Mucorales genomes they analyzed (n=4) but PKS-like genes (with a typical structure and domain order of FAS have been detected and Lebreton et al. showed these genes were systematically expressed on Potato dextrose medium (PDA) in five of the *Mucor* species studied in this study [3, 19]. Type I fatty acid synthases (FASs) are giant multifunctional proteins allowing fatty acid biosynthesis. Two strikingly different architectures of FAS exists: yeast-type FAS (yFAS) and metazoan FAS (mFAS), the latter one being related to PKS [95]. FASs are considered to be associated to primary metabolism whereas PKSs are associated to secondary metabolism [96]. Within the *Mucor* genomes studied in the present study, one yFAS gene was identified in the cheese-related *M. lanceolatus* UBOCC-A-109153 and *M. fuscus* UBOCC-A-109160 strains, three in the *M. indicus* CDC-B7402 opportunistic pathogen and two in all other strains. Only one yFAS gene was expressed in *M. circinelloides* CBS 277.49, *M. endophyticus* CBS 385-95 and *M. racemosus* UBOCC-A-109155 cultivated on PDA medium [19] while these genomes harbour two yFAS genes, suggesting that the second yFAS gene has lost its functionality or displays a different role since FAS are expected to be expressed on PDA medium to participate to the fatty acid synthesis for membrane synthesis or energy storage [97]. The tandem duplication observed in the *M. endophyticus* CBS 385-95 genome, suggests a specific regulation of the

yFAS genes. Duplication of *pks* gene and its subsequent functional diversification has been shown to increase the adaptive flexibility of fungal species ([98]), duplication of the *Mucor* yFAS encoding genes might play have a similar impact.

Another interesting feature was the atypical structure of the identified yFAS. Indeed, the different fungal FAS domains are usually encoded by two genes (Figure 3) [95, 99]. However, the identified *Mucor* yFAS encoding genes included all the different typical yFAS domains. Noteworthy, *M. circinelloides* CDC-B5328, *M. indicus* CDC-B7402 and *M. racemosus* UBOCC-A-109155 yFAS possessed a supplementary KS domain. To our best knowledge, the occurrence of the different fungal yFAS domains within a single gene was only discovered so far in the basidiomycetes *Ustilago maydis* and *Coprinopsis cinerea* [95, 99]. This result tends to support the hypothesis of a primordial contiguous FAS gene encoding the entire yFAS that have been split in dikarya through evolution.

Among the secondary metabolites, terpenes play a role in natural pigment synthesis such as the carotenoid biosynthesis [3]. Carotenoid synthesis has been described in *M. circinelloides* and in particular in overproducing strains [3, 100, 101]. Although the different species within the mucoromycotina subphylum are expected to produce and accumulate large amount of carotenoids, differences might exist among species [102]. Terpenes play also an important role in flavour production which is an important trait for cheese-ripening fungi that could have been under human-directed selection in species used for cheesemaking [103]. This study indicated that the main genes involved in terpene biosynthesis were conserved among the analysed *Mucor* species without any important differences in terms of gene number at the exception of GGPP genes for which the number is lower in the endophytic species as well as in the *M. ambiguus* NRBC 6742 and the *M. indicus* CDC-B7402 opportunistic pathogen.

Iron is an essential nutrient involved in variety of cellular processes such as respiration, oxidative stress and synthesis of amino acids [104]. Iron uptake has been described as a virulence factor for pathogenic fungal strains [83, 84] and of primary importance in cheese microorganisms since cheese is a highly iron-depleted medium [85]. In fungi, four different iron uptake mechanisms have been described so far: (i) siderophore mediated Fe^{3+} uptake,

(ii) reductive iron assimilation (RIA), (iii) heme uptake and (iv) direct Fe^{2+} uptake [83] (Figure 4). Homologs of genes coding for proteins participating to the four mechanisms were found in the analysed *Mucor* genomes. These results suggest that the different *Mucor* spp. investigated here rely on carboxylate siderophore rhizoferrin as it is the case for *M. circinelloides* CBS 277.49 and *R. delmar* RA-99880 [77]. This type of siderophore is not used by ascomycetous and basidiomycetous fungi, and would thus be a specificity of early diverging fungi in the fungal kingdom. Noteworthy, rhizoferrin encoding gene sequence are only described in bacterial genomes [105, 106] and coding sequences pertaining to the rhizoferrin operon in *F. tularensis* were used here to search for homologs in *Mucor* spp. In *F. tularensis*, genes involved in rhizoferrin synthesis and transport are located in an operon regulated by the *Fur* gene [78]. Based on these results, new candidate genes involved in rhizoferrin synthesis, import and export are proposed in the present study (Figure 4). Although no true evidence of *Mucor* ability to the produce hydroxamate siderophores was detected, some of these siderophores could be used by *Mucor* species as xeno-siderophores as suggested by the presence of *mirB*-like siderophore transporter genes in some of the *Mucor* genomes. It is worth noting that the bacterial siderophore deferoxamine is also used by *Mucor* spp. as xeno-siderophores [107].

Iron uptake is determinant for virulence but also for development in iron-depleted medium such as cheese. Interestingly, the three strains sampled from cheeses, including the two strains associated to cheese ripening *M. fuscus* UBOCC-A-109160 and *M. lanceolatus* UBOCC-A-109153, and the cheese contaminant *M. racemosus* UBOCC-A-109155, presented a reduced number of genes related to iron acquisition compared to the other strains. Indeed, the genome of these strains lack a FTR1 gene copy as well as the *fet3a* gene. It could be hypothesized that the latter genes would have a specific role in *Mucor* pathogenicity since *R. delmar* RA-99880 mutants with FTR1 reduced gene copies or with decreased FTR1 expression had reduced virulence in the deferoxamine-treated mouse model of mucormycosis [108]. The *fet3a* gene appeared as the less important among the *fet3* genes regarding *M. circinelloides* pathogenicity, but inactivation of two *fet3* genes lead to

a reduced virulence [15] and the loss of *fet3a* led to an increased sensibility to the mutations on *fet3b/fet3c* in terms of fungal pathogenicity. Furthermore, one copy of the ferroxamine binding (Fob) cell surface protein gene is absent in the *M. lanceolatus* UBOCC-A-109153 genome, the production of this protein being required for full virulence of *R. delmar* RA-99880 in a deferoxamine-treated mouse model of mucormycosis [12]. *M. fuscus* UBOCC-A-109160 also lacks MirB-like and fslB-like siderophore permease genes, which might reduce its potential to acquire iron.

On the contrary, *M. indicus* CDC-B7402, a species that is considered as the most threatening opportunistic human and animal pathogen amongst the *Mucor* spp. [6] harbours an increased set of genes involved in iron uptake which might be an asset to its opportunistic pathogenic lifestyle.

Acknowledgements

The authors are thankfull to Vincent Bruno for providing the *Mucor* annotation of strains CDC-B8987, CDC-B7402, CDC-B9645, CDC-B5328 and CDC-B9738, and to Zhengqiang Jiang for providing the predicted proteome of *Rhizomucor miehei.* Jonathan Dorival for advices on PKS. Antoine Branca and Emmanuelle Morin for stimulating discussion. Stephen Mondo for the integration of *Mucor* genomes in the mycocosm platform. This research was funded by the the Région Bretagne (ARED program) and EQUASA, a technological platform of the Université de Bretagne Occidentale, in the framework of the MUCORSCOPE project.

LITERATURE CITED

1. Spatafora JW, Aime MC, Grigoriev IV, Martin F, Stajich JE, Blackwell M: **The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies**. *Microbiology spectrum* 2017, **5**(5).
2. Morin-Sardin S, Nodet P, Coton E, Jany J-L: **Mucor: A Janus-faced fungal genus with human health impact and industrial applications**. *Fungal Biology Reviews* 2017, **31**(1):12-32.
3. Voigt K, Wolf T, Ochsenreiter K, Nagy G, Kaerger K, Shelest E, Papp T: **15 Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes**. In. Edited by Hoffmeister D. Cham: Springer International Publishing; 2016: 361-385.
4. Petrikkos G, Skiada A, Lortholary O, Roilides E, Walsh TJ, Kontoyiannis DP: **Epidemiology and Clinical Manifestations of Mucormycosis**. *Clinical Infectious Diseases* 2012, **54**(suppl_1):S23-S34.
5. Pitt JI, Hocking AD: **Fungi and Food Spoilage**: Springer US; 2009.
6. Morin-Sardin S, Rigalma K, Coroller L, Jany JL, Coton E: **Effect of temperature, pH, and water activity on Mucor spp. growth on synthetic medium, cheese analog and cheese**. *Food Microbiol* 2016, **56**:69-79.
7. Orlowski M: **Mucor dimorphism**. 1991, **55**(2):234-258.
8. Ma L-J, Ibrahim AS, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF, Sone T *et al*: **Genomic Analysis of the Basal Lineage Fungus Rhizopus oryzae Reveals a Whole-Genome Duplication**. *PLOS Genetics* 2009, **5**(7):e1000549.
9. Vongsangnak W, Kingkaw A, Yang J, Song Y, Laoteng K: **Dissecting metabolic behavior of lipid over-producing strain of Mucor circinelloides through genome-scale metabolic network and multi-level data integration**. *Gene* 2018, **670**:87-97.
10. Corrochano LM, Kuo A, Marcet-Houben M, Polaino S, Salamov A, Villalobos-Escobedo JM, Grimwood J, Álvarez MI, Avalos J, Bauer D *et al*: **Expansion of signal transduction pathways in fungi by extensive genome duplication**. *Current biology : CB* 2016, **26**(12):1577-1584.
11. Chibucos MC, Soliman S, Gebremariam T, Lee H, Daugherty S, Orvis J, Shetty AC, Crabtree J, Hazen TH, Etienne KA *et al*: **An integrated genomic and transcriptomic survey of mucormycosis-causing fungi**. *Nature Communications* 2016, **7**:1-11.
12. Liu M, Lin L, Gebremariam T, Luo G, Skory CD, French SW, Chou T-F, Edwards JE, Jr., Ibrahim AS: **Fob1 and Fob2 Proteins Are Virulence Determinants of Rhizopus oryzae via Facilitating Iron Uptake from Ferrioxamine**. *PLOS Pathogens* 2015, **11**(5):e1004842.
13. López-Fernández L, Sanchis M, Navarro-Rodríguez P, Nicolás FE, Silva-Franco F, Guarro J, Garre V, Navarro-Mendoza MI, Pérez-Arques C, Capilla J: **Understanding Mucor circinelloides pathogenesis by comparative genomics and phenotypical studies**. *Virulence* 2018, **9**(1):707-720.
14. López-Muñoz A, Nicolás FE, García-Moreno D, Pérez-Oliva AB, Navarro-Mendoza MI, Hernández-Oñate MA, Herrera-Estrella A, Torres-Martínez S, Ruiz-Vázquez RM, Garre V *et al*: **An Adult Zebrafish Model Reveals that Mucormycosis Induces Apoptosis of Infected Macrophages**. *Scientific Reports* 2018, **8**(1):12802.
15. Navarro-Mendoza MI, Pérez-Arques C, Murcia L, Martínez-García P, Lax C, Sanchis M, Capilla J, Nicolás FE, Garre V: **Components of a new gene family of ferroxidases**

involved in virulence are functionally specialized in fungal dimorphism. *Scientific Reports* 2018, **8**(1):7660.

16. Patino-Medina JA, Maldonado-Herrera G, Perez-Arques C, Alejandre-Castaneda V, Reyes-Mares NY, Valle-Maldonado MI, Campos-Garcia J, Ortiz-Alvarado R, Jacome-Galarza IE, Ramirez-Diaz MI *et al*: **Control of morphology and virulence by ADP-ribosylation factors (Arf) in Mucor circinelloides**. *Current Genetics* 2018, **64**(4):853-869.
17. Trieu TA, Navarro-Mendoza MI, Pérez-Arques C, Sanchis M, Capilla J, Navarro-Rodriguez P, Lopez-Fernandez L, Torres-Martínez S, Garre V, Ruiz-Vázquez RM *et al*: **RNAi-Based Functional Genomics Identifies New Virulence Determinants in Mucormycosis**. *PLoS Pathogens* 2017, **13**(1):e1006150.
18. Álvarez E, Cano J, Stchigel AM, Sutton DA, Fothergill AW, Salas V, Rinaldi MG, Guarro J: **Two new species of Mucor from clinical samples**. *Medical Mycology* 2011, **49**(1):62-72.
19. Annie L, Laurence M-C, Stéphanie M-S, Emmanuel C, Jean-Luc J, Georges B, Erwan C: **Comparative analysis of five Mucor species transcriptomes**. *Genomics* 2018.
20. Walther G, Pawłowska J, Alastruey-Izquierdo A, Wrzosek M, Rodriguez-Tudela JL, Dolatabadi S, Chakrabarti A, de Hoog GS: **DNA barcoding in <I>Mucorales</I>: an inventory of biodiversity**. *Persoonia - Molecular Phylogeny and Evolution of Fungi* 2013, **30**(1):11-47.
21. Gryganskyi AP, Golan J, Dolatabadi S, Mondo S, Robb S, Idnurm A, Muszewska A, Steczkiewicz K, Masonjones S, Liao H-L *et al*: **Phylogenetic and Phylogenomic Definition of Rhizopus Species**. *G3: Genes|Genomes|Genetics* 2018, **8**(6):2007-2018.
22. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, Program NCS *et al*: **Human Skin Fungal Diversity**. *Nature* 2013, **498**(7454):367-370.
23. Zheng R, Jiang H: **Rhizomucor endophyticus sp.nov., an endophytic zygomycetes from higher plants**. *Mycotaxon* 1995, **56**:455-466.
24. Hermet A, Meheust D, Mounier J, Barbier G, Jany JL: **Molecular systematics in the genus Mucor with special regards to species encountered in cheese**. *Fungal Biol* 2012, **116**(6):692-705.
25. Ozturkoglu-Budak S, Wiebenga A, Bron PA, de Vries RP: **Protease and lipase activities of fungal and bacterial strains derived from an artisanal raw ewe's milk cheese**. *International Journal of Food Microbiology* 2016, **237**:17-27.
26. Falkiewicz-Dulík M: **6.8 - LEATHER AND LEATHER PRODUCTS**. In: *Handbook of Material Biodegradation, Biodeterioration, and Biostablization (Second Edition).* Edited by Falkiewicz-Dulik M, Janda K, Wypych G: ChemTec Publishing; 2015: 133-256.
27. Joichi Y, Chijimatsu I, Yarita K, Kamei K, Miki M, Onodera M, Harada M, Yokozaki M, Kobayashi M, Ohge H: **Detection of <I>Mucor velutinosus</I> in a Blood Culture After Autologous Peripheral Blood Stem Cell Transplantation : A Pediatric Case Report**. *Medical Mycology Journal* 2014, **55**(2):E43-E48.
28. Lee SC, Billmyre RB, Li A, Carson S, Sykes SM, Huh EY, Mieczkowski P, Ko DC, Cuomo CA, Heitman J: **Analysis of a Food-Borne Fungal Pathogen Outbreak: Virulence and Genome of a Mucor circinelloides Isolate from Yogurt**. *mBio* 2014, **5**(4):e01390-01314.

29. Singh P, Paul S, Shivaprakash MR, Chakrabarti A, Ghosh AK: **Stress response in medically important Mucorales**. 2016, **59**(10):628-635.
30. Taj-Aldeen SJ, Almaslamani M, Theelen B, Boekhout T: **Phylogenetic analysis reveals two genotypes of the emerging fungus Mucor indicus, an opportunistic human pathogen in immunocompromised patients**. *Emerging Microbes & Infections* 2017, **6**(7):e63.
31. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, Abraham A-L, Ceppi M, Conseiller E, Debuchy R *et al*: **Multiple recent horizontal transfers of a large genomic region in cheese making fungi**. *Nature Communications* 2014, **5**:2876.
32. Andrews S: **FASTQC: A quality control tool for high throughput sequence data**. In.; 2010.
33. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *2011* 2011, **17**(1):3 %J EMBnet.journal.
34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
35. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. 2008, **18**(5):821-829.
36. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**:18-18.
37. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics**. *Molecular Biology and Evolution* 2018, **35**(3):543-548.
38. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training**. 2008, **18**(12):1979-1990.
39. Stanke M, Schöffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources**. *BMC Bioinformatics* 2006, **7**(1):62.
40. Dobin A, Gingeras TR: **Mapping RNA-seq Reads with STAR**. *Current protocols in bioinformatics* 2015, **51**:11.14.11-11.14.19.
41. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms**. *Nature biotechnology* 2010, **28**(5):511-515.
42. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data**. *Nature biotechnology* 2011, **29**(7):644-652.
43. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.
44. Slater GSC, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC Bioinformatics* 2005, **6**(1):31.
45. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biology* 2008, **9**(1):R7.

46. Krogh a, Larsson B, von Heijne G, Sonnhammer E: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *Journal of molecular biology* 2001, **305**(3):567-580.
47. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nature Methods* 2011, **8**:785.
48. Eddy SR: **Accelerated Profile HMM Searches**. *PLOS Computational Biology* 2011, **7**(10):e1002195.
49. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al*: **The Pfam protein families database: towards a more sustainable future**. *Nucleic Acids Research* 2016, **44**(D1):D279-D285.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
51. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F *et al*: **MycoCosm portal: gearing up for 1000 fungal genomes**. *Nucleic Acids Research* 2014, **42**(D1):D699-D704.
52. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM**. *Nucleic Acids Res* 2003, **31**(22):6633-6639.
53. Lowe TM, Chan PP: **tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes**. *Nucleic Acids Research* 2016, **44**(Web Server issue):W54-W57.
54. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Res* 2007, **35**(9):3100-3108.
55. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics* 2013, **29**(22):2933-2935.
56. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J *et al*: **Rfam 12.0: updates to the RNA families database**. *Nucleic Acids Research* 2015, **43**(Database issue):D130-D137.
57. Dunn N, Munoz-Torres M, Unni D, Y E, Rasche E, Bretaudeau A, Diesh C, Elsik C, Holmes I, Lewis S: **GMOD/Apollo: Apollo2.0.6(JB#29795a1bbb)**. 2017.
58. Flutre T, Duprat E, Feuillet C, Quesneville H: **Considering Transposable Element Diversification in De Novo Annotation Approaches**. *PLOS ONE* 2011, **6**(1):e16526.
59. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H: **PASTEC: An Automatic Transposable Element Classification Tool**. *PLOS ONE* 2014, **9**(5):e91929.
60. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Research* 2014, **42**(Database issue):D490-D495.
61. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**(1):59-60.
62. Busk PK, Pilgaard B, Lezyk MJ, Meyer AS, Lange L: **Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function**. *BMC Bioinformatics* 2017, **18**(1):214.
63. Rawlings ND, Barrett AJ, Finn R: **Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors**. *Nucleic Acids Research* 2016, **44**(D1):D343-D350.

64. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A: **MUMmer4: A fast and versatile genome alignment system**. *PLOS Computational Biology* 2018, **14**(1):e1005944.
65. Darling AE, Mau B, Perna NT: **progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement**. *PLOS ONE* 2010, **5**(6):e11147.
66. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy**. *Genome Biology* 2015, **16**(1):157-157.
67. Löytynoja A: **Phylogeny-aware alignment with PRANK**. *Methods Mol Biol* 2014, **1079**:155-170.
68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009, **25**(15):1972-1973.
69. Stamatakis A: **Using RAxML to Infer Phylogenies**. 2015, **51**(1):6.14.11-16.14.14.
70. Lartillot N, Rodrigue N, Stubbs D, Richer J: **PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment**. *Systematic Biology* 2013, **62**(4):611-615.
71. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock**. *Bioinformatics* 2003, **19**(2):301-302.
72. Zhou P, Zhang G, Chen S, Jiang Z, Tang Y, Henrissat B, Yan Q, Yang S, Chen C-F, Zhang B *et al*: **Genome sequence and transcriptome analyses of the thermophilic zygomycete fungus Rhizomucor miehei**. *BMC Genomics* 2014, **15**(1):294.
73. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW: **Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3**. *Molecular Biology and Evolution* 2013, **30**(8):1987-1997.
74. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de los Santos Emmanuel LC, Kim HU, Nave M *et al*: **antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification**. *Nucleic Acids Research* 2017, **45**(W1):W36-W41.
75. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND: **SMURF: genomic mapping of fungal secondary metabolite clusters**. *Fungal genetics and biology : FG & B* 2010, **47**(9):736-741.
76. Thieken A, Winkelmann G: **Rhizoferrin: a complexone type siderophore of the Mucorales and entomophthorales (Zygomycetes)**. *FEMS microbiology letters* 1992, **73**(1-2):37-41.
77. Carroll CS, Grieve CL, Murugathasan I, Bennet AJ, Czekster CM, Liu H, Naismith J, Moore MM: **The rhizoferrin biosynthetic gene in the fungal pathogen Rhizopus delemar is a novel member of the NIS gene family**. *The International Journal of Biochemistry & Cell Biology* 2017, **89**:136-146.
78. Ramakrishnan G: **Iron and Virulence in Francisella tularensis**. *Frontiers in Cellular and Infection Microbiology* 2017, **7**:107.
79. Tang X, Chen H, Chen YQ, Chen W, Garre V, Song Y, Ratledge C: **Comparison of Biochemical Activities between High and Low Lipid-Producing Strains of Mucor circinelloides: An Explanation for the High Oleaginicity of Strain WJ11**. *PloS one* 2015, **10**(6):e0128396-e0128396.
80. Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev IV *et al*: **Transposable Elements versus**

the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. *PLOS Genetics* 2016, **12**(6):e1006108.
81. Mohanta TK, Bae H: **The diversity of fungal genome**. *Biological procedures online* 2015, **17**:8-8.
82. Rokas A, Wisecaver JH, Lind AL: **The birth, evolution and death of metabolic gene clusters in fungi**. *Nature Reviews Microbiology* 2018.
83. Bairwa G, Hee Jung W, Kronstad JW: **Iron acquisition in fungal pathogens of humans**. *Metallomics : integrated biometal science* 2017, **9**(3):215-227.
84. Haas H: **Iron - A Key Nexus in the Virulence of Aspergillus fumigatus**. *Frontiers in microbiology* 2012, **3**:28-28.
85. Monnet C, Loux V, Gibrat J-F, Spinnler E, Barbe V, Vacherie B, Gavory F, Gourbeyre E, Siguier P, Chandler MJPO: **The Arthrobacter arilaitensis Re117 genome sequence reveals its genetic adaptation to the surface of cheese**. 2010, **5**(11):e15489.
86. Keller NP, Hohn TM: **Metabolic Pathway Gene Clusters in Filamentous Fungi**. *Fungal Genetics and Biology* 1997, **21**(1):17-29.
87. Brown DW, Lee S-H, Kim L-H, Ryu J-G, Lee S, Seo Y, Kim YH, Busman M, Yun S-H, Proctor RH *et al*: **Identification of a 12-Gene Fusaric Acid Biosynthetic Gene Cluster in Fusarium Species Through Comparative and Functional Genomics**. *Molecular Plant-Microbe Interactions* 2014, **28**(3):319-332.
88. Reynolds HT, Slot JC, Divon HH, Lysøe E, Proctor RH, Brown DW: **Differential Retention of Gene Functions in a Secondary Metabolite Cluster**. *Molecular Biology and Evolution* 2017, **34**(8):2002-2015.
89. Slot JC, Rokas A: **Horizontal Transfer of a Large and Highly Toxic Secondary Metabolic Gene Cluster between Fungi**. *Current Biology* 2011, **21**(2):134-139.
90. Schwartze VU, Winter S, Shelest E, Marcet-Houben M, Horn F, Wehner S, Linde J, Valiante V, Sammeth M, Riege K *et al*: **Gene Expansion Shapes Genome Architecture in the Human Pathogen Lichtheimia corymbifera: An Evolutionary Genomics Analysis in the Ancient Terrestrial Mucorales (Mucoromycotina)**. *PLOS Genetics* 2014, **10**(8):e1004496.
91. Velayos A, Papp T, Aguilar-Elena R, Fuentes-Vicente M, Eslava AP, Iturriaga EA, Alvarez MI: **Expression of the carG gene, encoding geranylgeranyl pyrophosphate synthase, is up-regulated by blue light in Mucor circinelloides**. *Curr Genet* 2003, **43**(2):112-120.
92. Andrews JHJTFCio, ecosystem rit: **Fungal life-history strategies**. 1992, **2**:119-145.
93. Cooke RC, Rayner AD: **Ecology of saprotrophic fungi**: Longman; 1984.
94. Howard DH: **Pathogenic fungi in humans and animals**: CRC Press; 2002.
95. Herbst DA, Townsend CA, Maier T: **The architectures of iterative type I PKS and FAS**. *Natural Product Reports* 2018.
96. Crawford JM, Vagstad AL, Ehrlich KC, Udwary DW, Townsend CA: **Acyl-Carrier Protein–Phosphopantetheinyltransferase Partnerships in Fungal Fatty Acid Synthases**. 2008, **9**(10):1559-1563.
97. Maier T, Leibundgut M, Boehringer D, Ban N: **Structure and function of eukaryotic fatty acid synthases**. *Quarterly Reviews of Biophysics* 2010, **43**(3):373-422.
98. Zeng G, Zhang P, Zhang Q, Zhao H, Li Z, Zhang X, Wang C, Yin W-B, Fang W: **Duplication of a Pks gene cluster and subsequent functional diversification facilitate environmental adaptation in Metarhizium species**. *PLOS Genetics* 2018, **14**(6):e1007472.

99. Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, DeAngelis YM, Reeder NL, Johnstone KR *et al*: **Dandruff-associated Malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens**. 2007, **104**(47):18730-18735.
100. Csernetics Á, Nagy G, Iturriaga EA, Szekeres A, Eslava AP, Vágvölgyi C, Papp TJFg, biology: **Expression of three isoprenoid biosynthesis genes and their effects on the carotenoid production of the zygomycete Mucor circinelloides**. 2011, **48**(7):696-703.
101. Navarro E, Sandmann G, Torres-Martínez SJEm: **Mutants of the carotenoid biosynthetic pathway of Mucor circinelloides**. 1995, **19**(3):186-190.
102. Zhang Y, Navarro E, Cánovas-Márquez JT, Almagro L, Chen H, Chen YQ, Zhang H, Torres-Martínez S, Chen W, Garre V: **A new regulatory mechanism controlling carotenogenesis in the fungus Mucor circinelloides as a target to generate β-carotene over-producing strains by genetic engineering**. *Microbial cell factories* 2016, **15**:99-99.
103. Ropars J, Lo Y-C, Dumas E, Snirc A, Begerow D, Rollnik T, Lacoste S, Dupont J, Giraud T, López-Villavicencio M: **Fertility depression among cheese-making Penicillium roqueforti strains suggests degeneration during domestication**. *Evolution; international journal of organic evolution* 2016, **70**(9):2099-2109.
104. Winkelmann G: **Specificity of iron transport in bacteria and fungi**. In: *Handbook of Microbial Iron Chelates (1991).* CRC Press; 2017: 73-114.
105. Franken ACW, Lechner BE, Werner ER, Haas H, Lokman BC, Ram AFJ, van den Hondel CAMJJ, de Weert S, Punt PJ: **Genome mining and functional genomics for siderophore production in Aspergillus niger**. *Briefings in Functional Genomics* 2014, **13**(6):482-492.
106. Khan A: **Synthesis, nature and utility of universal iron chelator − siderophore: a review**. *Microbiological research* 2017:2017.
107. Szebesczyk A, Olshvang E, Shanzer A, Carver PL, Gumienna-Kontecka E: **Harnessing the power of fungal siderophores for the imaging and treatment of human diseases**. *Coordination Chemistry Reviews* 2016, **327-328**:84-109.
108. Ibrahim AS, Spellberg B, Walsh TJ, Kontoyiannis DP: **Pathogenesis of Mucormycosis**. *Clinical Infectious Diseases* 2012, **54**(suppl_1):S16-S22.
109. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR *et al*: **CDD/SPARCLE: functional classification of proteins via subfamily domain architectures**. *Nucleic Acids Research* 2017, **45**(D1):D200-D203.

Table S1: Extraction and sequencing information corresponding to the four newly sequenced *Mucor*.

Taxon	Extraction protocol	Sequencing type	Sequencing date	Sequencing place	# million filtered reads
M. racemosus UBOCC-A-109155	CATB + phenol-chloroform purifications	Paired-end (Illumina)	2012	BGI	37
M. fuscus UBOCC-A-109160	CATB + phenol-chloroform purifications	Paired-end (Illumina)	2012	BGI	19
M. lanceolatus UBOCC-A-109153	[31]	Paired-end (Illumina)	2013	Biogenouest Génomique	74
	CATB + Qiagen	Mate-pair (Illumina)	2016	Macrogen	11
M. endophyticus CBS 385-95	CATB + Qiagen	Paired-end (Illumina)	2016	Macrogen	60
	CATB + Qiagen	Mate-pair (Illumina)	2016	Macrogen	19

Supplementary materials

Doc S1: Details of materials and methods on specific gene family annotation

Genes involved in iron uptake, storage and regulation referenced in supplementary table (N) were searched by BLASTp (default e-value, coverage above 50%) against the predicted proteome of the twelve strains. Matching sequences were then searched on the predicted proteomes of the twelve strains to gather sequences that could have been too different from the reference sequence. When a specific “PFAM id” was available, sequences with this annotation were added to the pool of potential predicted proteins. Conserved domains of each of these potential predicted proteins were searched using the CD-Search tool ([109]) to verify that the conserved domains concurred between the reference sequence and the potential predicted protein; if not, the sequence was discarded. This set of filtered predicted proteins was then aligned with MUSCLE to create a gene tree with the Geneious (v9) tree

builder, sequences behaving as outgroups of the reference sequences were discarded. When missing, genes were searched by BLAT on the Apollo genome viewer.

Des analyses complémentaires à celles de l'article ont été réalisées et sont présentées dans les parties suivantes :

4.3 Méthodologie supplémentaire : Assemblages des génomes

4.3.1 Introduction

Comme indiqué dans l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*" quatre souches ont été séquencées lors de cette étude : *M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153, *M. racemosus* UBOCC-A-109155 et *M. endophyticus* CBS 385-95 (UBOCC-A-113049). Dans le cadre de cet article, il a été précisé que les séquençages ont été réalisés avec la technologie Illumina sur différentes plateformes (Tableau S1 de l'article), que tous les génomes des souches ont été séquencés en *paired end* et que les génomes de *M. lanceolatus* et de *M. endophyticus* avaient bénéficié d'un séquençage supplémentaire en *mate pair*. Il est également précisé qu'un nettoyage des données issues du séquençage du génome de *M. lanceolatus* a été réalisé. Cependant, les analyses ayant mené à ce choix méthodologique n'ont pas été développées. Elles seront donc explicitées ci-après.

4.3.2 Matériels et méthodes

Données de séquençage initiales

A l'origine du projet de comparaisons génomiques, en 2012, a été réalisé le séquençage des génomes de *M. racemosus* UBOCC-A-109155 et de *M. fuscus* UBOCC-A-109160 en *paired end* par la plateforme du BGI. Les données de séquençage étaient accompagnées de l'assemblage correspondant réalisé avec l'assembleur SOAPdenovo. En 2013 a été séquencé le génome de *M. lanceolatus* UBOCC-A-109153 également en *paired end* mais cette fois ci par la plateforme Biogenouest Génomique de Nantes et sans assemblage associé.

Assemblage des génomes à partir de séquençage en *paired end* seul

Au cours du projet, les assembleurs *CLC Genomics Workbench*, Velvet, SPADES et SOAPdenovo2 ont été testés pour l'assemblage de novo des trois souches. Velvet et SPADES sont des assembleurs basés sur des algorithmes associés au graphe de *de Bruijn* (approche par K-mers voir synthèse bibliographique sur les assembleurs), différentes tailles de K-mer ont donc été testées (de 55pb à 99pb avec un pas de 2pb). Ces deux même assembleurs peuvent être, et ont

été, utilisés en *genome guided* en se basant sur le meilleur assemblage préliminaire *de novo* réalisé par un autre assembleur.

Séquençages complémentaires

En 2016, un séquençage complémentaire du génome de *M. lanceolatus* a été réalisé, cette fois ci, en *mate pair* sur la plateforme de Macrogen. En parallèle, la souche de *M. endophyticus* a été intégrée au projet. Son génome a été séquencé d'une part avec un séquençage en *paired end* et d'autre part avec un séquençage en *mate pair* tous deux sur la plateforme de Macrogen. Les données de séquençage du génome de *M. endophyticus* étaient accompagnées de l'assemblage correspondant réalisé avec l'assembleur Platanus.

Description des données *mate pair* obtenues dans le génome de *M. lanceolatus*

La taille d'insert des données *mate pair* a été estimée d'une part par l'assembleur *CLC Genomics Workbench*, d'autre part par un script utilisant BWA (Li and Durbin, 2009). Les *reads* ont été mappés sur le génome le moins fragmenté de *M. lanceolatus* avec STAR. La distribution de la longueur entre les *reads* pairés a été examinée ainsi que l'orientation des *reads* appartenant à la même paire. A partir de ce mapping, les 90000 paires de *reads* orientées en FR avec un insert inférieur à 500pb ont été supprimées du jeu de données. Par la suite, il sera fait référence à ce nouveau jeu de données sous de nom de "*mate pair filtrés*".

Assemblage des génomes à partir de séquençage en *paired end* et *mate pair*

Les quatre assembleurs présentés plus tôt (SOAPdenovo2, SPADES, *CLC Genomics Workbench* et Velvet) ont été de nouveau utilisés pour assembler les génomes de *M. lanceolatus* et *M. endophyticus*. Ils ont été utilisés selon la même méthodologie aux exceptions suivantes près. (i) Velvet a été utilisé en ajoutant l'option "shortMatePaired". (ii) Pour *CLC Genomics Workbench* des assemblages de *M. lanceolatus* ont été réalisés en précisant différentes tailles d'insert des données *mate pair* : estimation de l'outil, 500pb et 7500pb. Un scaffolding supplémentaire a été réalisé sur l'assemblage le moins fragmenté en utilisant les données RNAseq *via* l'outil L_RNA_scaffolder (Xue et al., 2013)).

4.3.3 Résultats

Séquençage

La description des données issues des différents séquençages est présentée dans l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*".

Assemblage des génomes à partir de séquençage en *paired end* seul

Pour les génomes de *M. fuscus* et de *M. racemosus*, les meilleurs assemblages réalisés étaient d'une qualité équivalente à celle des assemblages fournis par le BGI. Ces assemblages étaient composés de 3 819 scaffolds (*M. fuscus*) et 3506 scaffolds (*M. racemosus*) de plus de 1000pb. Le meilleur assemblage du génome de *M. lanceolatus* obtenu comptait 13 035 contigs de plus de 200pb (6946 de plus de 1000pb). Cet assemblage a été obtenu avec SPADES utilisant un assemblage préliminaire de *CLC Genomics Workbench* comme guide.

Description des données *mate pair* obtenues pour le génome de *M. lanceolatus*

Le script utilisant BWA a permis d'estimer la taille d'insert des données *mate pair* du génome de *M. lanceolatus* à 5310pb +/- 2911pb. L'assembleur *CLC Genomics Workbench* l'estimait quant à lui à 538pb. Lors du mapping des *reads* sur le génome le moins fragmenté de *M. lanceolatus*, 87% d'entre eux ont pu être alignés. Cependant, aussi bien la taille d'insert que l'orientation des *reads* d'une même paire étaient extrêmement variables selon les paires de *reads* considérées (Figure 4.2).

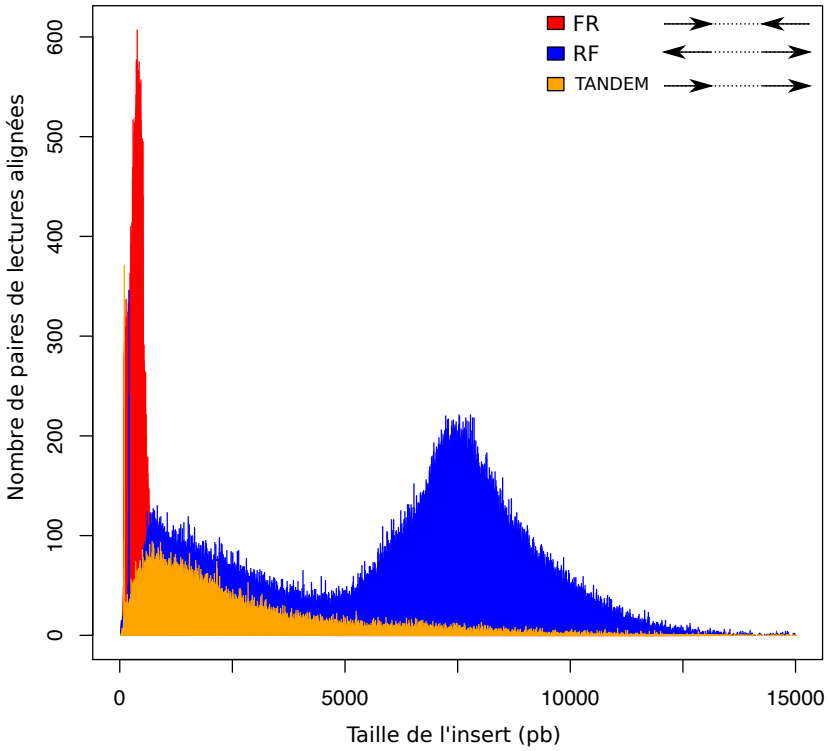


FIGURE 4.2 – Distribution de la taille d'insert des *reads* pairés *mate pair* de *M. lanceolatus* lors de leur mapping sur l'assemblage de novo généré par SPADES en utilisant un assemblage préliminaire de CLC genomic workbench comme guide. FR, RF et TANDEM font référence à l'orientation des *reads* de la paire.

Assemblage du génome de *M. lanceolatus*

À l'exception de l'assembleur Velvet utilisé avec l'option "shortMatePaired", l'utilisation conjointe des séquençages *paired end* et *mate pair* par les assembleurs testés a conduit à des assemblages plus fragmentés que lorsque seules les données *paired end* étaient utilisées (Tableau 4.2).

Assembleur	Séquençage utilisé	Option	# Scaffolds > 200pb	Scaffold max (kb)	Taille (Mb)	N50 (kb)	N80 (kb)
SOAPdenovo	PE	*M. fuscus - BJI*	3 819	142,3	40,6	27,3	9,8
SOAPdenovo	PE	*M. racemosus - BJI*	3 506	161,9	46,9	24,9	13,0
SOAPdenovo2	PE	*par défaut*	21 918	42,0	57,4	6,6	2,4
	PE & MP	*par défaut*	18 168	175,9	81,2	17,5	5,8
SPADES	PE	*genome guided CLC*	13 035	67,5	49,2	7,7	4,1
	PE & MP	*par défaut*	20 504	347,5	51,0	9,2	2,4
	PE & MP filtrées	*genome guided CLC*	20 974	79,0	51,0	9,0	2,6
	PE & MP	*genome guided velvet k67*	14 452	304,5	52,2	28,7	3,8
CLC	PE	*par défaut*	18 090	44,4	45,3	6,3	2,4
	PE & MP	*clc sans estimations*	17 687	47,9	46,4	7,6	2,8
	PE & MP	*clc estimations PE*	18 457	47,9	46,6	7,3	2,5
	PE & MP	*clc estimations MP*	19 903	46,7	51,3	7,6	2,6
	PE & MP filtrées	*clc estimations MP*	18 253	57,9	50,0	7,8	2,9
Velvet	PE & MP	*k67 - shortMatePaired*	7 134	556,3	45,4	116,4	37,6
	PE & MP	*k69 - shortMatePaired*	7 385	937,2	45,6	114,1	30,8
	PE & MP filtrées	*k67 - shortMatePaired*	6 898	537,9	45,2	110,6	33,1
	PE & MP filtrées	*k69 - shortMatePaired*	7 044	802,0	46,1	107,1	35,7
	PE & MP filtrées	*k71 - shortMatePaired*	7 242	1 026,6	46,0	108,0	36,4

TABLEAU 4.2 – Assemblages du génome de *Mucor lanceolatus* réalisés en utilisant les données *mate pair* (MP). Les assemblages des génomes de *M. fuscus* et *M. racemosus* alors à disposition sont grisés. Les meilleurs assemblages pour les données *paired end* (PE) seules d'une part et MP et PE conjointes d'autre part sont surlignés en vert. "MP filtrées" fait référence à la suppression de lectures contaminantes *paired end* présentes dans les données *mate pair*. "*genome guided* CLC" indique l'utilisation de l'assemblage reconstruit avec les données PE seules avec *CLC Genomics Workbench* (options par défaut) comme guide. "clc estimations PE" correspond au test en précisant la taille d'insert des données MP à 500pb (taille d'insert de données PE). "clc estimations MP" correspond au test en précisant la taille d'insert des données MP à 7 500pb. k# fait référence à la taille de K-mer utilisé par l'assembleur, avec # le nombre de bases.

L'assemblage obtenu en utilisant Velvet avec l'option "shortMatePaired" et un K-mer de 67, a été conservé. Le scaffolding supplémentaire a permis de diminuer le nombre de scaffolds de 6 898 à 6 566 (de plus de 200pb). Parmi ces 6 566 scaffolds, seuls 1 531 disposaient d'une taille supérieure à 1000pb. Cet assemblage sera par la suite celui auquel on fera référence en tant qu'assemblage du génome de *M. lanceolatus*.

Assemblage du génome de *M. endophyticus*

L'assemblage le moins fragmenté a été obtenu avec Velvet avec K-mer de 85pb. Celui-ci après une étape de scaffolding était constitué de 159 scaffolds de plus de 1000pb (Tableau 4.3). Cet assemblage sera utilisé par la suite. Cet assemblage présentait une faible taille de génome par rapport aux autres espèces mais ce résultat était cohérent avec les assemblages préliminaires réalisés par les cinq assembleurs.

Assembleur	Option	# Scaffolds > 1 000pb	Scaffold max (kb)	Taille (Mb)	N50 (kb)	N80 (kb)	%N
Platanus - macrogen		406	539	32,5	125	61	0,06
CLC Genomics Workbench	*par défaut*	816	422	34,2	108	50	1,5
SOAPdenovo2	*par défaut*	256	3 164	39,6	648	341	15,35
SPADES	genome guided Platanus	307	1 349	34	497	209	0,23
Velvet	k85	161	4 526	35	1 957	910	2
Velvet	k85 LNRA	159	4 526	35	1 957	910	2

TABLEAU 4.3 – Assemblages du génome de *Mucor endophyticus* testés. L'assemblage utilisant Platanus est celui généré par Macrogen avec le séquençage des données. L'assemblage de SPADES a été généré en utilisant l'assemblage de Platanus comme guide. LNRA : scaffolding supplémentaire avec l'outil L_RNA_scaffolder.

4.3.4 Conclusion et discussion

Au cours de l'assemblage des génomes des quatre souches de *Mucor*, des difficultés ont été rencontrées pour l'assemblage du génome de *M. lanceolatus* UBOCC-A-109153. En effet, sur le séquençage en *paired end* seul, malgré une profondeur de séquençage deux fois plus importante pour le génome de *M. lanceolatus* que pour les génomes de *M. fuscus* et de *M. racemosus*, l'assemblage du génome de *M. lanceolatus* obtenu était deux fois plus fragmenté (~7000 scaffolds) que ceux de *M. racemosus* (~3500 scaffolds) et *M. fuscus* (~3800 scaffolds). La fragmentation d'un assemblage est souvent due à des répétitions dans les génomes (Peona et al., 2018; Dominguez Del Angel et al., 2018). Ce résultat suggère donc que la composition et/ou le taux de répétition est plus important dans le génome de *M. lanceolatus* que dans les génomes de *M. racemosus* et *M. fuscus*, ce qui a été confirmé par la suite (Figure 2C de l'article). Pour améliorer l'assemblage du génome de *M. lanceolatus*, un séquençage en *mate pair* a été réalisé. Cependant, cette librairie *mate pair* était contaminée par des *paired end*. Dans ce type de contamination, les lectures erronées se comportent comme des lectures *paired end* : orientation opposée par rapport aux *mate pair* et taille d'insert bien plus courtes (Sahlin et al., 2016), ce qui correspondait au premier pic (en rouge) de la Figure 4.2. Ces contaminations sont liées à la préparation de la

librairie *mate pair* : durant sa création, une fraction inconnue de fragments qui ne contiennent pas de jonctions permettant la circularisation est séquencée. Les contaminations des données *mate pair* par des *paired end* peuvent conduire à des erreurs d'assemblage et notamment des problèmes liés à l'ordre relatif des contigs les uns par rapport aux autres (Figure 4.4).

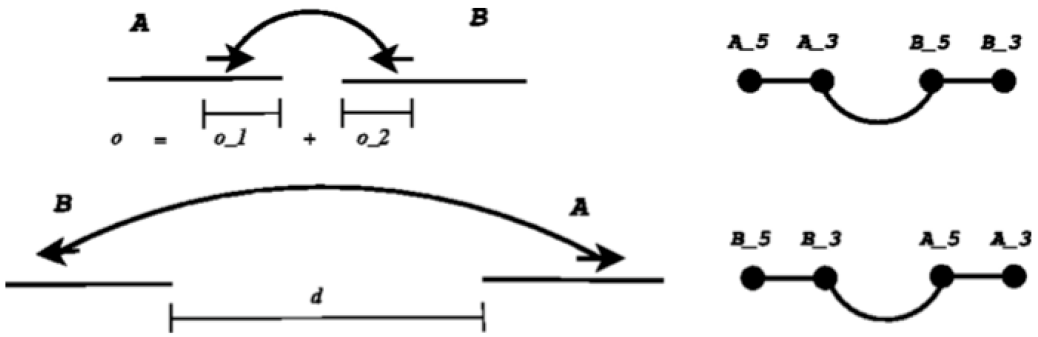


Tableau 4.4 – Impact des contaminations PE dans des données MP. Gauche : deux positionnements possibles des contigs si on ne connait pas l'orientation relative des lectures pairées représentées. A droite : l'assemblage des contigs correspondants. "5" et "3" correspondent au 5′ et 3′ de la séquence. Cette figure est extraite de publication de Sahlin et al. (2016)

Parmi les assembleurs testés (SOAPdenovo2, SPADES, *CLC Genomics Workbench* et Velvet), seul l'assembleur Velvet disposait d'une option permettant de tenir de compte et filtrer les lectures *paired end* potentiellement contaminantes dans une library *mate pair* (option "shortMate-Paired") ce qui explique les meilleurs résultats de Velvet par rapport aux autres assembleurs. Au final, les quatre génomes des souches séquencés, à savoir *M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153, *M. racemosus* UBOCC-A-109155 et *M. endophyticus* UBOCC-A-113049 (CBS 385-95), ont été assemblées en 3819, 1531, 3506 et 159 scaffolds respectivement.

4.4 Mise en évidence de longues régions génomiques dépourvues d'annotations de gènes et éléments répétés

4.4.1 Introduction

Au cours de l'annotation experte, des sections génomiques sans annotations ont été détectées (d'une taille allant jusqu'à 65kb dans le génome de *M. lanceolatus*). La présence de ces régions est d'autant plus surprenante que lorsque l'on observe les régions géniques des génomes de *Mucor* à disposition, les gènes sont rarement séparés par plus de 1000pb. Ces régions blanches présentées dans l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*" (section 4.2), ont fait l'objet de recherches qui n'ont pas été développées dans l'article mais seront présentées ici.

4.4.2 Matériels et méthodes

Recherche de gènes non annotés sur la région de 65kb sans annotations

Les recherches ci-après se basent principalement sur les données produites dans le cadre de l'article. L'absence de modèles de gènes, à savoir les prédictions des prédicteurs *ab initio* de Genemark-ES et Augustus, qui n'auraient pas été retenus par EvidenceModeler (outil permettant d'obtenir une annotation consensus à partir de multiples modèles de gènes, données d'homologies et d'expression) a été vérifiée. La présence de gènes exprimés sur la région a été recherchée d'une part par un mapping RNAseq sur le génome de *M. lanceolatus* avec STAR, d'autre part par le mapping des transcrits reconstruits *de novo* avec Trinity (chapitre 2) sur le génome avec gmap. Des recherches d'éléments disposant d'homologies avec les bases protéiques sur la région d'intérêt a été vérifiée.

Recherche de caractéristiques particulières associées à la séquence

Dans un premier temps, un biais lié à l'assemblage dans la région a été recherché : (i) en identifiant son pourcentage de base ambigües (N), (ii) en comparant le pourcentage de GC de la région à celui du reste du génome, (iii) en cherchant si la couverture en *reads* génomiques était cohérente entre cette région et le reste du génome. Cette dernière vérification a été réalisée par mapping des données de séquençage génomique (DNAseq) de *M. lanceolatus* sur la région génomique sans annotation avec Bowtie. Par la suite, un dot plot de la séquence contre elle même a été réalisé pour identifier la présence d'éventuelles répétitions.

Recherche de correspondances

Une correspondance avec les bases de données a été recherchée par BLASTn et BLASTx entre la base protéique et nucléotidique non redondante du NCBI (nr et nt) et la séquence de la région sans annotations (entière et fragmentée). L'ensemble de la région sans annotation et un de ses fragment (de 8kb) ont été recherchés par BLASTn sur l'ensemble des génomes de *M. fuscus*, *M. lanceolatus*, *M. racemosus*, *M. endophyticus* et *M. circinelloides*.

Estimation du nombre et de la taille de ces régions sans annotations

Des régions sans annotations ont été recherchées chez *R. delemar* et *P. blakesleeanus* sur l'interface d'annotation du JGI (`https://genome.jgi.doe.gov/programs/fungi/index.jsf`). La distribution de la taille des régions entre les annotations (gènes et éléments répétés) a été observée chez les dix *Mucor* et deux *Rhizopus microsporus* (anciennement nommés *M. racemosus*) étudiés.

4.4.3 Résultats

Recherche de gènes non annotés sur la région de 65kb sans annotations

Aucune trace d'expression de gènes sous la forme d'un mapping RNAseq ou de transcrits reconstruits de novo n'a été identifiée sur la région. Les deux prédicteurs de gènes *ab initio* (augustus et genemark-ES), n'ont pas prédit de gènes sur la région. Aucun mapping de protéines recensées dans les bases de données (Swissprot/Uniprot, Uniref90 et protéome prédits de *M. circinelloides*, *R. delemar* et *P. blakesleeanus*) n'a été identifié sur la région.

Recherche de caractéristiques particulières associées à la séquence

La région génomique sans annotations est couverte de façon homogène et similaire à des régions géniques par les *reads* lors du mapping des données DNAseq sur génome. La région ne présente pas de différence en terme de pourcentage GC par rapport au reste du génome bien qu'environ 15% de la séquence soit constituée de bases ambiguës.

Un dot plot de la séquence contre elle même (Figure 4.3) a permis de mettre en évidence des sections largement répétées composées de très petites répétitions (les "carrés" dans le dot plot de la Figure 4.4) et des régions non répétées (début et fin de la séquence de la Figure 4.3, sans correspondance dans la séquence).

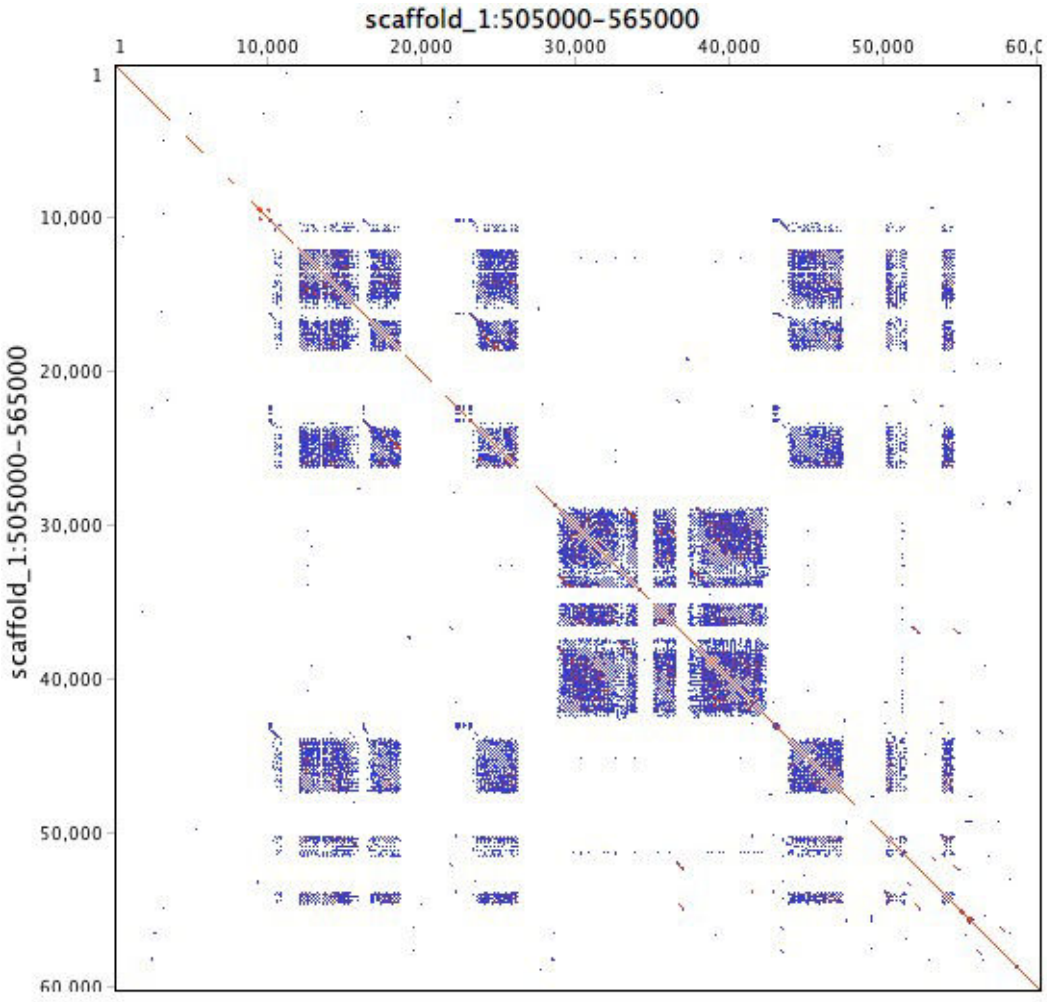


FIGURE 4.3 – Dot plot d'une région du génome de *M. lanceolatus* de 65kb sans annotations (60kb présentés ici) contre elle même. Chaque point représente une correspondance entre les deux séquences comparées. La diagonale centrale correspond à l'alignement de la séquence contre elle même, les sections blanches dans cette diagonale correspondent aux bases ambiguës. Les points en dehors de cette diagonale principale correspondent aux motifs répétés dans la séquence.

Recherche de correspondances

Aucune correspondance n'a été retrouvée sur tout ou une partie de la séquence ni contre nr/nt ni dans le génome d'origine (*M. lanceolatus*) ni dans les génomes des quatre autres *Mucor* alors étudiés (*M. fuscus*, *M. lanceolatus*, *M. racemosus*, *M. endophyticus* et *M. circinelloides*).

Estimation du nombre et de la taille de ces régions sans annotations

La distribution de la taille des régions entre les annotations (gènes et éléments répétés) a été présentée dans l'article (Figure 1, de l'article présenté en 4.2). Des régions sans annotations supérieures à 5kb ont été observées également chez *R. delemar* et *P. blakesleeanus*.

4.4.4 Discussion

De multiples et longues régions sans annotations ont été identifiées au sein des génomes des souches de *Mucor* étudiées. Pour mieux les caractériser une des plus grande région détectée de ce type a été analysée. La région sélectionnée était présente dans le génome de *M. lanceolatus* et s'étendait sur 65kb.

Cette étude a permis de montrer que l'absence d'annotation au niveau de cette longue région n'était pas due à un biais d'assemblage (mapping DNAseq homogène avec les parties géniques, bases ambiguës concentrées sur certaines parties de la séquence), qu'il ne s'agissait probablement pas de l'introduction d'un fragment d'ADN contaminant par l'assembleur (pas de biais du pourcentage GC), que cette absence n'était pas non plus associée à un problème d'annotation (absence de modèles de gènes, de marques d'expression génique et de correspondance avec des protéines des bases de données) et qu'une partie de la séquence était composée de motifs répétés très courts. Ce type de région sans annotations est présent dans les 14 génomes observés (les 10 *Mucor spp.* de l'étude, les deux *R. microsporus*, *R. delemar* et *P. blakesleeanus*), cependant ces régions semblent différentes en terme de composition de séquences (pas de correspondances détectées entre elles ni avec les bases de données). Il reste à déterminer leur rôle et l'avantage évolutif de conserver de telles régions. Parmi les différentes hypothèses, ces régions peuvent avoir un intérêt pour la structuration tridimensionnelle de la chromatine, la plasticité du génome ou elle peuvent encore contenir des structures fonctionnelles inconnues et/ou non détectées par les méthodes utilisées.

4.5 Expansions et contractions des familles de gènes au sein des *Mucor spp.*

4.5.1 Introduction

Afin d'identifier les familles de gènes et/ou fonctions associées à une potentielle adaptation au milieu, des analyses sur les expansions et contractions de famille de gènes ont été engagées. Ces analyses, actuellement en cours, et dont les premiers résultats ont été présentés dans l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*" (section 4.2) ont permis, au travers des analyses préliminaires, de dégager des pistes intéressantes sur l'évolution des *Mucor spp.*. Ces analyses préliminaires font l'objet de cette partie. Ces analyses se décomposent en trois parties : la première a été réalisée à l'échelle du sous-phylum Mucoromycota avec Notung, la

deuxième se focalise sur le genre *Mucor* avec DupliPHYML et la troisième a pour objet de mieux comprendre les événements d'expansions et contractions de familles de gènes au sein du genre *Mucor* et plus particulièrement au sein du clade alors formé par *M. racemosus* et les souches CDC-B9645 et CDC-B9738 (alors identifiées comme *M. racemosus* (Chibucos et al., 2016)) mais ensuite réassignée à l'espèce *R. microsporus* (Gryganskyi et al., 2018)).

4.5.2 Matériels et méthodes

Analyse préliminaire de l'expansion/contraction des familles de gènes à l'échelle du sous-phylum Mucoromycota

En complément des quatres souches de *Mucor* séquencées pour le projet, douze souches d'espèces appartenant aux Mucoromycota dont les génomes étaient disponibles sur le portail du JGI ont été utilisées (Tableau 4.5).

Souche	Sous-division	Référence
Mortierella elongata AG-77 v2.0	Mortierellomycotina	Uehling et al., 2017
Lobosporangium transversale NRRL 3116 v1.0	Mortierellomycotina	Mondo et al., 2017
Rhizophagus irregularis DAOM 181602 v1.0	Glomeromycota	Tisserant et al., 2013
Hesseltinella vesiculosa NRRL3301 v2.0	Mucoromycotina	Mondo et al., 2017
Absidia repens NRRL 1336 v1.0	Mucoromycotina	Mondo et al., 2017
Lichtheimia corymbifera JMRC:FSU:9682	Mucoromycotina	Schwartze et al., 2014
Syncephalastrum racemosum NRRL 2496 v1.0	Mucoromycotina	Mondo et al., 2017
Phycomyces blakesleeanus NRRL1555 v2.0	Mucoromycotina	Corrochano et al., 2016
Saksenaea vasiformis B4078	Mucoromycotina	Chibucos et al., 2016
Rhizopus microsporus ATCC11559 v1.0	Mucoromycotina	Lastovetsky et al., 2016
Rhizopus microsporus var. chinensis CCTCC M201021	Mucoromycotina	Wang et al., 2013
Mucor circinelloides CBS277.49 v2.0	Mucoromycotina	Corrochano et al., 2016
M. endophyticus CBS 385-95 (UBOCC-A-113049)	Mucoromycotina	cette étude
M. fuscus UBOCC-A-109160	Mucoromycotina	cette étude
M. lanceolatus UBOCC-A-109153	Mucoromycotina	cette étude
M. racemosus UBOCC-A-109155	Mucoromycotina	cette étude

Tableau 4.5 – Liste des souches utilisées dans l'analyse d'expansion/contraction de familles de gènes portant sur les Mucoromycota. En rouge sont indiquées les souches dont les génomes ont été séquencés dans le cadre de cette étude.

Les familles de gènes et l'arbre des espèces reconstruit avec OrthoFinder v1.1.8 ont été utilisés par Notung pour prédire les expansions et contractions de familles de gènes. Les *GO terms* des gènes présents dans les familles présentant une expansion ou une contraction chez l'ancêtre le plus proche des espèces technologiques (*M. fuscus* et *M. lanceolatus*) ont été extraits puis représentés de manière synthétique à l'aide de l'outil REViGO.

Analyse des expansions/contractions de familles de gènes à l'échelle du genre *Mucor*

Une deuxième analyse a été réalisée. Dans un premier temps, celle ci portait uniquement sur les quatre génomes des espèces de *Mucor* séquencés dans le cadre de ce projet : *M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153, *M. racemosus* UBOCC-A-109155 et *M. endophyticus* CBS 385-95 ; ainsi que la souche de référence *M. circinelloides* CBS 277.49.

Dans un second temps, cette analyse a été étendue au sept autres génomes disponibles alors identifiés comme appartenant au genre *Mucor* et dont les annotations étaient accessibles, à savoir *M. circinelloides* souches CDC-B8987, 1006PhL, CDC-B5328 et NBRC 6742, *M. indicus* CDC-B7402 et les souches CDC-B9738 et CDC-B9546 (Tableau 2 de l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*" (section 4.2)).

Dans les deux cas, les familles de gènes ont été reconstruites avec OrthoFinder v2.2.0 puis l'arbre phylogénétique a été reconstruit selon la méthodologie présentée dans l'article (section 4.2). Brièvement, l'arbre a été reconstruit d'une part selon une approche par maximum likelihood avec RaxML et d'autre part par une approche bayésienne avec PhyloBayes à partir des orthologues du *core genome* présents en une seule copie dans chacune des espèces. Les recherches d'expansions/contractions de familles de gènes ont été réalisées avec DupliPHYML. Les gènes associés aux sidérophores ont été recherchés par mots clefs (*iron transporters, siderophores*) dans les annotations à disposition. Les expansions et contractions des familles de gènes identifiés comme *drug transporters* ont été recherchées manuellement.

Analyse sur les *Mucor* avec CAFE

Un autre outil a été utilisé (CAFE), l'analyse associée correspond à celle présentée dans l'article "*Comparative genomics applied to* Mucor *species with different lifestyles*" (section 4.2). CAFE nécessite un arbre calibré sur une échelle de temps (arbre ultramétrique) ce qui a été réalisé avec le programme r8S en se basant sur l'arbre généré précédemment avec RAxML et l'estimation de l'origine de *R. delemar* et *P. blakesleeanus* à 468MY (Zhou et al., 2014). Il sera fait référence par la suite en tant que "groupe dupliqué" au groupe contenant les souches CDC-B9645 et CDC-9738 et l'ancêtre direct de ces deux souches.

A la différence de DupliPHYML, CAFE permet (i) de prendre en compte la distance évolutive entre les espèces, (ii) d'assigner une vitesse de gain et perte de gènes différente entre groupes d'espèces spécifiés par l'utilisateur, (iii) de séparer l'estimation de la vitesse de gain de gènes de celle de perte de gènes (par défaut la vitesse moyenne de gain et perte de gènes est considéré

comme identique) et (iv) d'identifier les familles de gènes pour lesquelles une expansion ou contraction est significativement plus importante qu'attendue.

Différentes analyses ont été réalisées avec CAFE : (I) en estimant une même vitesse de gain/perte de gènes pour toutes les espèces, (II) en spécifiant la présence de deux groupes ayant des vitesses de gain/perte de gènes différentes, le groupe dupliqué d'une part et les autres espèces d'autre part, (III) en spécifiant la présence de quatre groupes ayant des vitesses de gain de gènes différentes, (i) la souche CDC-B9645, (ii) la souche CDC-9738, (iii) l'ancêtre des deux souches précédentes et (iv) les autres espèces, (IV) en séparant les analyses entre le groupe dupliqué d'une part et le reste de l'arbre d'autre part, ces deux analyses étant réalisées estimant une même vitesse de gain/perte de gènes pour les espèces considérées. (V) Toutes les analyses précédentes ont été réalisées de nouveau en spécifiant des vitesses différentes entre gains et pertes de gènes.

Afin de vérifier que l'augmentation du nombre de paramètres décrivait mieux les données qu'un seul, un test de vraisemblance réalisé par CAFE avec 100 simulations a été réalisé.

4.5.3 Résultats

Analyse préliminaire de l'expansion/contraction des familles de gènes à l'échelle du sous-phylum Mucoromycota

La majorité des duplications et pertes de gènes étaient prédites sur les noeuds terminaux de l'arbre phylogénétique (au niveau des espèces) (exemple du groupe des *Mucor* en Figure 4.4). Dans le groupe des *Mucor*, *M. fuscus* présentait 1405 duplications de gènes (Figure 4.4 à gauche), ayant eu lieu dans environ 350 familles de gènes (Figure 4.4 droite).

Parmi les prédictions d'expansion et contraction de familles de gènes, 248 pertes de gènes et 29 duplications ont été prédites chez l'ancêtre le plus proche des espèces technologiques (n17 sur la Figure 4.4). Les 29 gènes dupliqués étaient impliqués dans la fixation d'ions métalliques (Zinc notamment) ou correspondaient à des peptidases (endopeptidase aspartique notamment), des acetyltransferases, et des éléments liés aux ribosomes et aux rétrotransposons. Les *GO terms* associés aux 248 gènes perdus ont été synthétisés et sont présentés en figure 4.5. Ces résultats seront développés par la suite dans ce manuscrit.

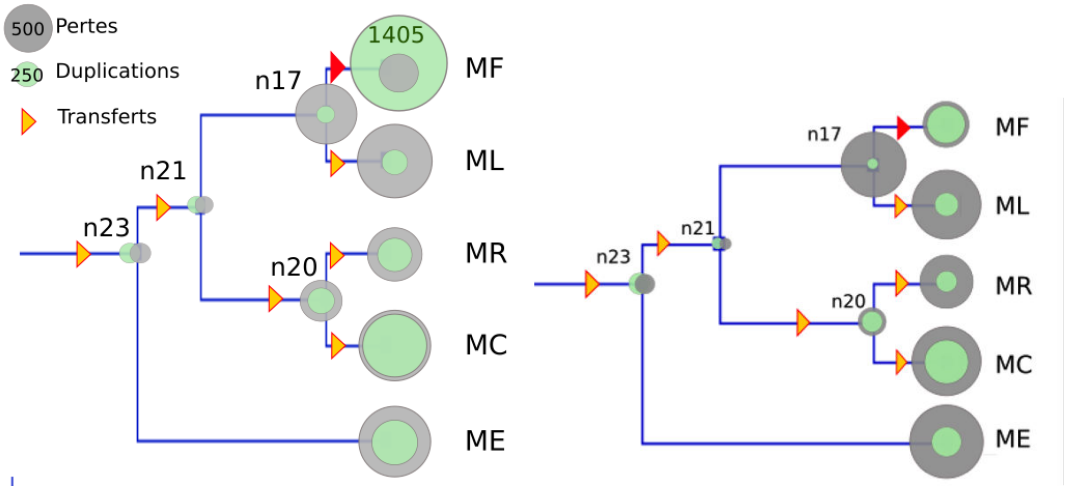


FIGURE 4.4 – Représentation schématique des prédictions des gains et pertes de gènes au cours de la spéciation des cinq *Mucor spp.* étudiées. A gauche, nombre de duplications, pertes et transferts horizontaux prédits pour chacune des espèces et leurs ancêtres. A droite, nombre de familles de gènes ayant subi un événement de duplication. MF : *M. fuscus*. ML : *M. lanceolatus*. MR : *M. racemosus*. MC : *M. circinelloides*. ME : *M. endophyticus*. Un triangle jaune indique la présence de 5 à 10 transferts horizontaux prédits. Un triangle rouge de 20 à 65 transferts horizontaux prédits.

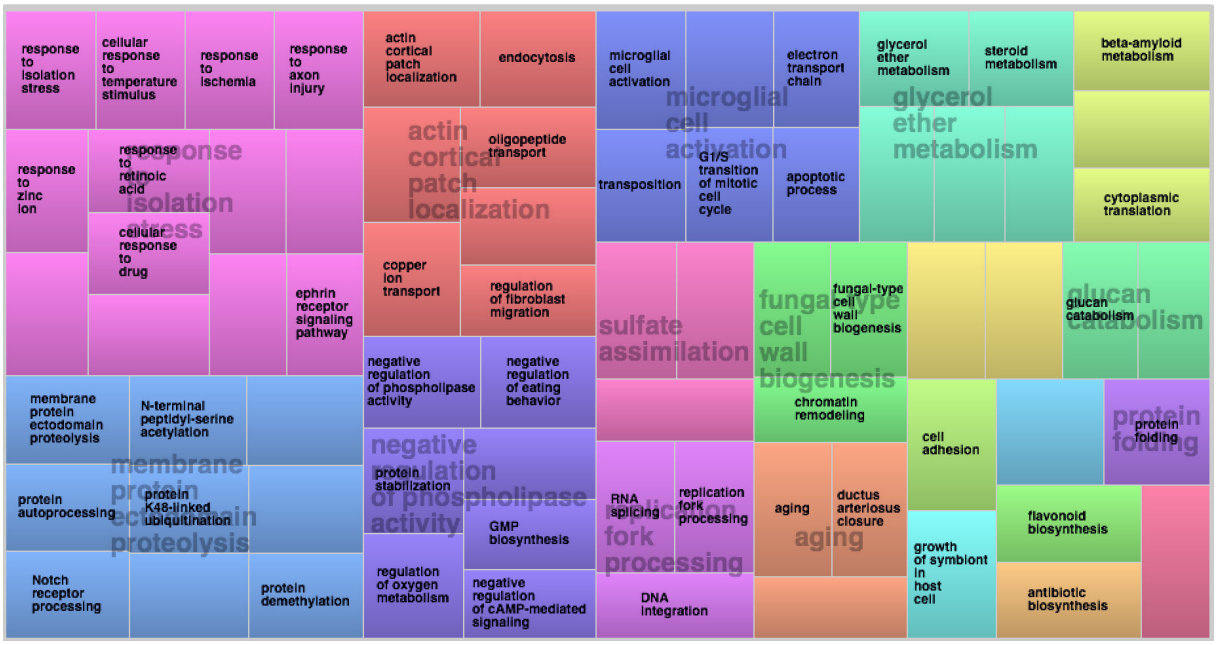


FIGURE 4.5 – Synthèse des *GO terms* de type *Biological Process* annotés sur les 248 gènes perdu chez l'ancêtre le plus proche des espèces technologiques (noeud n17 en figure 4.4).

Analyse des expansions/contractions de familles de gènes à l'échelle du genre *Mucor*

Les résultats associés à cette analyse sont présentés en Figure 4.6. Des différences importantes avec les prédictions de Notung à l'échelle des Mucoromycota ont été observées : dans cette analyse, les gains et pertes de gènes étaient répartis de façon plus uniforme dans l'arbre phylogénétique. Par exemple, là où l'ancêtre le plus proche des espèces technologiques arborait 29 gènes dupliqués dans l'analyse sur les Mucoromycota avec Notung, 648 gènes ont été identifiés comme étant dupliqués dans cette analyse focalisée sur les *Mucor* avec DupliPhyML.

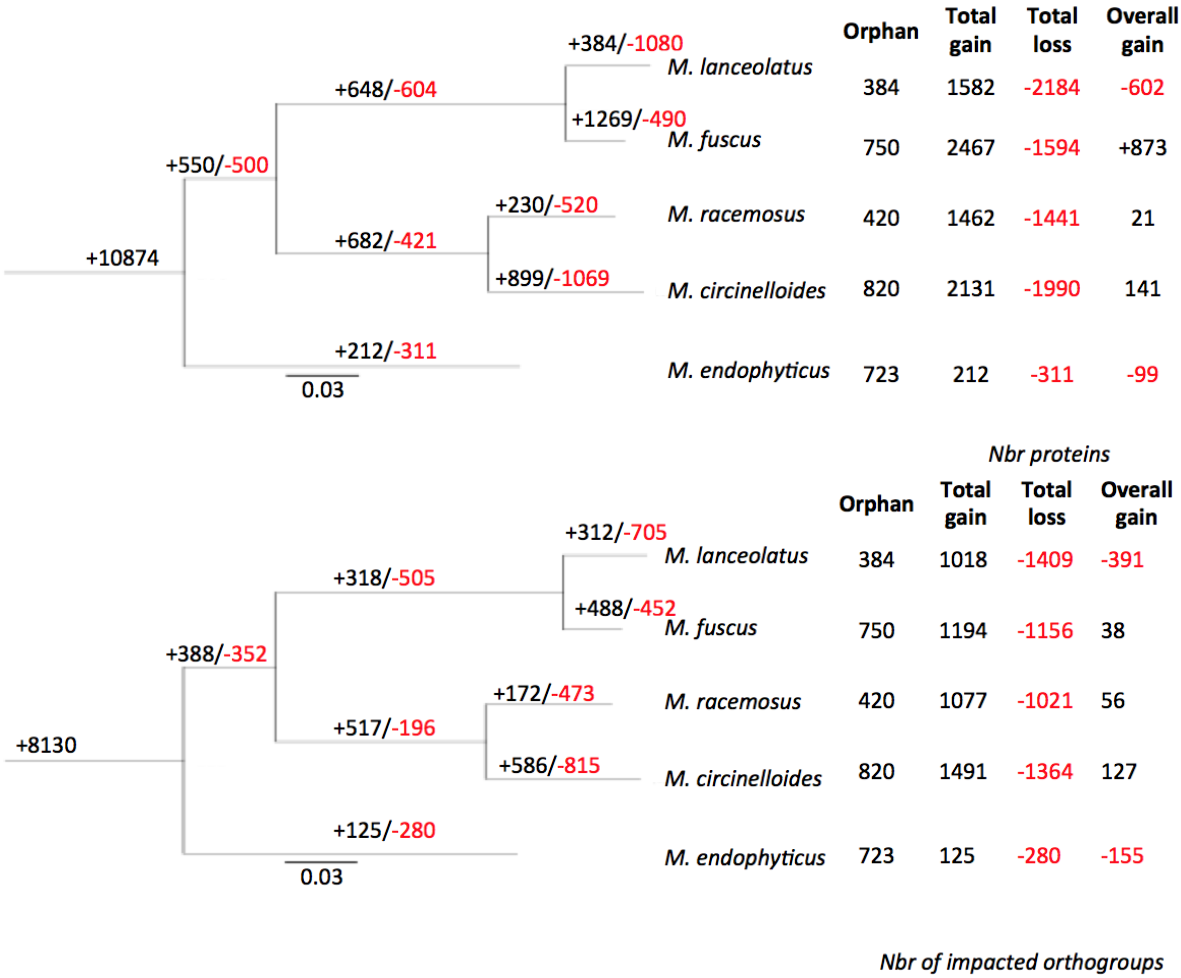


FIGURE 4.6 – Prédiction des expansions et contractions de familles de gènes au cours de l'évolution des *Mucor* au travers de cinq espèces représentatives du genre. En haut estimation du nombre de gènes gagnés et perdus au cours de cette évolution. En bas, le nombre de famille de gènes ayant subit des expansions et contractions au cours de l'évolution de ces cinq *Mucor*.

La recherche de gènes associés aux sidérophores (acquisition du fer) a permis de montrer que ces gènes étaient dupliqués principalement chez les espèces pathogènes opportunistes et en particulier chez *M. circinelloides* (Figure 4.7 gauche). Chez *M. lanceolatus*, on note une perte et

deux acquisitions de gène associés aux sidérophores. De façon plus spécifique, lors d'une autre recherche, une famille de gène liée à un "*drug transporter*" a été perdue chez l'ancêtre des espèces technologiques (Figure 4.7 droite).

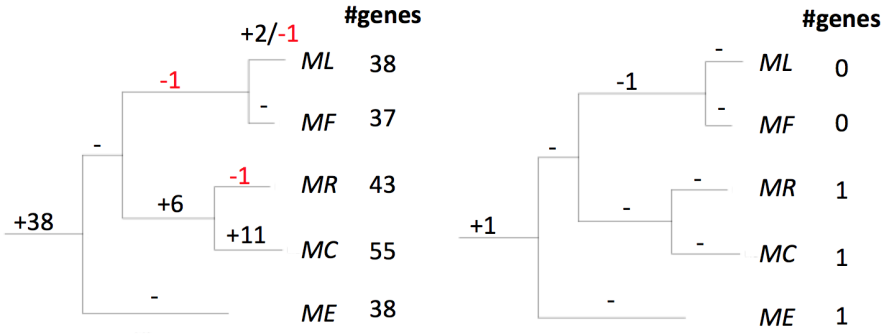


FIGURE 4.7 – Expansions et contractions de familles de gènes A gauche les gènes associés à des sidérophores, à droite famille de "*drug transporter*". ML : *M. lanceolatus*. MF : *M. fuscus*. MR : *M. racemosus*. MC : M.circinelloides. ME : *M. endophyticus*.

Lors de l'extension de l'analyse aux huit autres *Mucor*, les résultats associés aux espèces proches de la racine (*M. endophyticus*, *M. lanceolatus*, *M. fuscus* et leurs ancêtres directs) étaient en accord avec l'analyse restreinte aux cinq espèces. Cependant, des estimations incohérentes étaient obtenues pour le clade associé à *M. racemosus* : *M. racemosus* UBOCC-A-109155, *M. racemosus* B9738 et *M. racemosus* B9645 (ces deux dernières souches ayant été par la suite réassignées à l'espèce *R. microsporus*). En effet l'ancêtre identifié à l'origine de ce clade ainsi que le noeud terminal de la souche B9645 possédaient une estimation d'aucun gain ni perte de gène (Figure 4.8).

Analyse sur les *Mucor* avec CAFE

Le nombre global d'expansions et contractions de familles de gènes prédits par CAFE était bien moindre que lors des analyses précédentes. La plupart des expansions de familles de gènes prédites étaient localisées au niveau des noeuds terminaux (Figure 4.9) tandis que les contractions de familles de gènes étaient répartis de façon plus homogène. Lors de la vérification de la vraisemblance des prédictions, l'utilisation de quatre vitesses de gains/pertes de gènes expliquait mieux la taille et composition des familles de gènes observées au sein des espèces étudiées, puis deux vitesses de gains/pertes de gènes et enfin une vitesse globale.

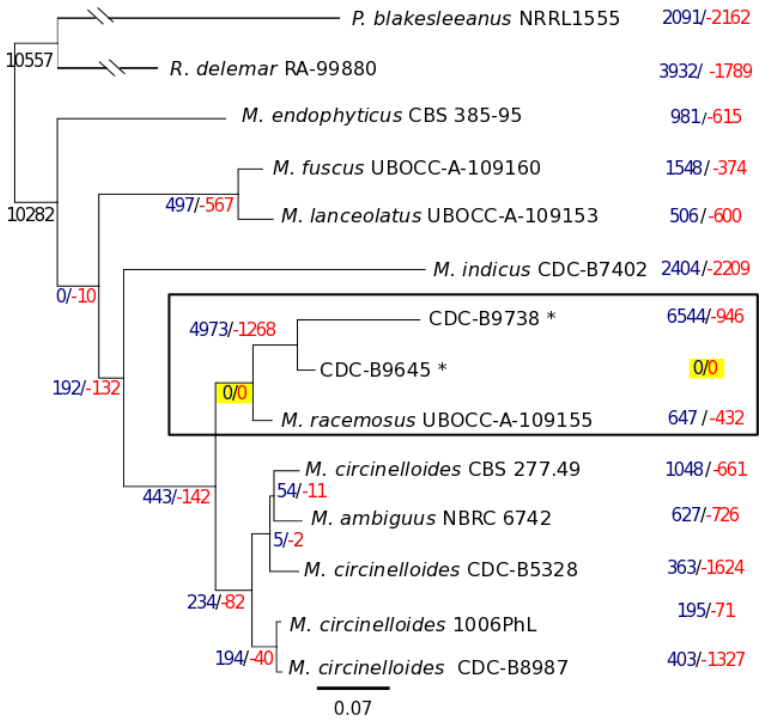


FIGURE 4.8 – Estimation des expansions et contractions de familles de gènes prédits avec DupliPHYML au sein du genre *Mucor*. * Les souches CDC-B9645 et CDC-B9738 étaient lors de cette étude identifiées comme étant des *M. racemosus* (Chibucos et al., 2016) mais ont été par la suite assignées à l'espèce *Rhizopus microsporus* (Gryganskyi et al., 2018).

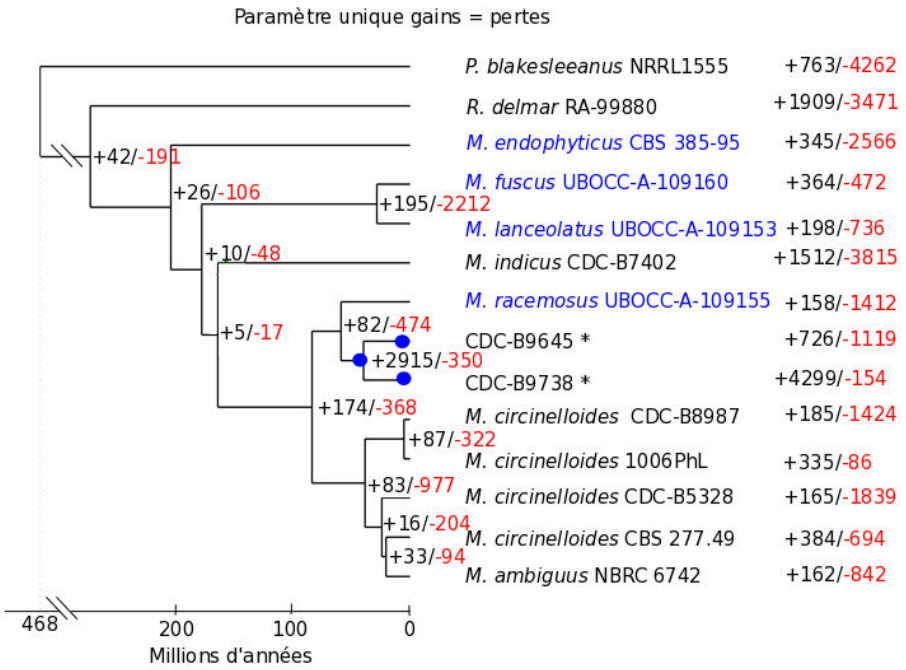


FIGURE 4.9 – Expansions et contractions de familles de gènes prédites par CAFE avec une seule vitesse de gains et pertes de gènes précisée. Les souches séquencées pour cette étude sont présentées en bleu. Des marqueurs bleu sur l'arbre indiquent le groupe dit "dupliqué". * souches identifées par Chibucos et al. (2016) comme étant des *M. racemosus*, et par Gryganskyi et al. (2018) comme étant des *R. microsporus*.

Trois points principaux se dégageaient de l'analyse réalisée : une perte massive de gènes chez l'ancêtre des souches appartenant aux espèces utilisées en affinage des fromages, des variations importantes entre les *M. circinelloides* (le nombre de perte de gènes entre le groupe contenant les souches 1006PhL et CDC-B8987 et le groupe composé des autres *M. circinelloides* était environ trois fois plus important qu'entre le groupe associé aux *M. racemosus* et le groupe des *M. circinelloides*) et la répartition des gains de gènes au sein du groupe dupliqué.

Quelle que soit l'approche, les prédictions de contractions de familles de gènes étaient similaires, cependant des différences notables étaient présentes pour les prédictions d'expansions de familles de gènes au niveau du groupe dupliqué. Globalement, deux profils étaient observés : soit une partie des expansions étaient prédites sur l'ancêtre des deux souches dupliquées et l'autre partie sur le noeud terminal correspondant à la souche CDC-B9738 (tableau 4.6A,B,C), soit les expansions étaient reportées sur les noeuds adjacents (ancêtre des *M. racemosus* par exemple) et des noeuds terminaux (tableau 4.6D,E). Le premier cas était observé lors de la prédiction de vitesses équivalentes entre gains et pertes de gènes, le second cas lorsque les pertes de gènes étaient estimées indépendamment des gains de gènes.

	*une vitesse **globale***	deux vitesses de gains de gènes: **groupe dupliqué / autres souches**	quatre vitesses de gains de gènes : **un par noeud du groupe dupliqué et un pour les autres souches**
Vitesse de gain et perte de gènes estimée **conjointement**	**A** Mr (158) 82 B9645 (726) 2915 B9738 (4299)	**B** Mr (159) 59 B9645 (1020) 2851 B9738 (4578)	**C** Mr (166) 54 B9645 (734) 3034 B9738 (4312)
Vitesse de gain et perte de gènes estimée **séparément**		***D*** Mr (98) 5827 B9645 (0) 0 B9738 (0)	**E** Mr (188) 80 B9645 (3195) 0 B9738 (6837)

Tableau 4.6 – Impact des paramètres de CAFE sur les prédictions des expansions de familles de gènes associées au groupe dupliqué et de l'espèce la plus proche de ce groupe : *M. racemosus* UBOCC-A-109155 (Mr).

4.5.4 Discussion

Les différentes analyses réalisées ont permis d'obtenir des estimations d'expansions et contractions de familles de gènes au sein du genre *Mucor*. Les résultats de la première analyse portant sur les Mucoromycota montraient que les prédictions d'expansions et contractions se concentraient au niveau des noeuds terminaux (espèces) voire sub-terminaux (noeuds vers des espèces proches phylogénétiquement) alors qu'en amont (cas des noeuds n23 et n21 en Figure 4.4), relativement peu de contractions/expansions étaient prédites (cas des noeuds n23 et n21 en Figure 4.4). Ce profil asymétrique inhabituel pourrait s'expliquer biologiquement par une faible contribution évolutive des expansions/contractions lors de l'émergence et de l'évolution du genre *Mucor*. Le moteur évolutif n'aurait alors pas tant reposé sur des processus d'expansion/contraction de familles de gènes que sur, par exemple, des modifications de gènes (via des mutations non synonymes menant à des modifications de l'activité des enzymes codées ou à leur spécificité) ou de leur expression *via* leur régulation. Toutefois, au vu de l'importance des processus de duplication de gènes en tant que force évolutive (Zhang, 2003), cette dernière hypothèse concernant leur faible contribution au cours de l'évolution chez les *Mucor* reste peu probable et incite à considérer que la dissymétrie observée repose sur un biais analytique. Ce biais potentiel nous a conduit par la suite dans un premier temps à focaliser l'étude uniquement sur le genre *Mucor* et dans un second temps à tester un autre outil permettant d'estimer les expansions et contractions de familles de gènes.

Malgré des différences entre les prédictions des outils, différentes caractéristiques étaient retrouvées. Par exemple, au niveau du noeud correspondant à l'ancêtre le plus récent des souches *M. lanceolatus* UBOCC-A-109153 et *M. fuscus* UBOCC-A-109160 (souches appartenant aux espèces utilisées pour l'affinage de fromages), un grand nombre de contractions de familles de gènes a été prédit tandis qu'un nombre restreint d'expansions a été identifié (résultats de Notung (Figure 4.4) et CAFE (Figure 4.9)).

Parmi les expansions de familles de gènes, sont notées des peptidases, gènes codant des protéines pouvant jouer un rôle important dans l'exploitation du substrat "matrice fromagère" et pour la génération d'arômes (Sousa et al., 2001; Ardö, 2006) ainsi que des gènes codant des protéines associées à la fixation d'ions métalliques, déjà identifiés comme permettant un avantage sélectif chez *Glutamicibacter arilaitensis* sur milieu fromager (Monnet et al., 2010). L'accès aux ions métalliques est également critique dans d'autres milieux. L'expansion des gènes liés à l'acquisition du fer est notamment important chez les pathogènes opportunistes (Gerwien et al.,

2018). La recherche des gènes associés codant des sidérophores (étude avec DupliPHYML sur cinq souches) a montré une importante expansion de ces gènes chez les espèces pathogènes opportunistes (*M. racemosus* UBOCC-A-109155 et *M.circinelloides* CBS 277.49) et chez *M. circinelloides* CBS 277.49 en particulier (souche la plus virulente de cette analyse). Une analyse plus poussée sur les génomes des dix souches de *Mucor* à disposition, présentée dans le cadre de l'article présenté plus tôt (section 4.2), a confirmé l'expansion de familles de gènes impliquées dans l'acquisition du fer (sidérophores et acquisition par réduction du fer notamment) chez les espèces pathogènes opportunistes. Cette analyse montre que chez les *Mucor* l'acquisition du fer peut être un élément important pour la pathogénicité de l'espèce. On note également lors de la première recherche des sidérophores que la souche de *M. lanceolatus* UBOCC-A-109153 présentait à la fois expansion et contraction de gènes associés à cette famille ce qui interroge sur une éventuelle spécialisation de ces gènes impliqués dans l'acquisition du fer au sein de la souche *M. lanceolatus* UBOCC-A-109153 appartenant à une espèce jusqu'à présent uniquement retrouvée sur fromage.

Lors des analyses réalisées, un grand nombre de perte de gènes a été prédit sur l'ancêtre des souches *M. lanceolatus* UBOCC-A-109153 et *M. fuscus* UBOCC-A-109160. Parmi les fonctions associées aux pertes de gènes (pertes identifiées avec Notung), on note surtout des gènes associés à la réponse au stress dont la présence pourrait être sélectivement moins avantageuse en milieu fromager que dans l'environnement naturel aux conditions moins stables. De même, des familles de gènes impliquées dans le transport de drogues, et donc permettant potentiellement une résistance à des éléments toxiques, ont été spécifiquement perdues chez ces deux souches retrouvées sur fromage. Il a déjà été suggéré que les familles de gènes liées à la résistance au stress étaient en général davantage sujettes aux expansions/contractions (Wapinski et al., 2007). Hormis ces gènes associés à la réponse au stress, des gènes aux fonctions diverses semblaient être perdus. Dans le cas d'espèces qui pourraient être considérées comme technologiques (Morin-Sardin et al., 2016), ces pertes de gènes pourraient être dues à un relâchement global des contraintes sélectives sur des populations restreintes à l'environnement fromager conduisant d'une part à des pertes de gènes et des contractions de familles de gènes (voir Demuth and Hahn (2009), et d'autre part à une sélection forte par l'environnement fromager ou par l'homme pour un faible nombre de trait d'intérêt.

Un autre point d'intérêt est le nombre important d'expansions et contractions de familles de gènes détectées au sein du groupe des *M. circinelloides*. Ce résultat supporte une éventuelle

scission future de l'espèce *M. circinelloides* en plusieurs espèces distinctes correspondant aux différentes formes de l'espèce actuelle (Pawlowska et al., 2013).

Enfin, l'évolution des familles de gènes au sein du clade alors identifié comme correspondant à l'espèce *M. racemosus* (souches UBOCC-A-109153, CDC-B9645 et CDC-B9738) s'est révélé complexe. Au début de l'étude le nombre de gènes bien plus important chez les deux souches CDC-B9645 et CDC-B9738 suggérait un niveau de ploïdie différent des autres souches. Ces modifications peuvent être dues à la duplication complète ou importante soit du génome de l'ancêtre des deux souches, soit de chacune des souches de façon indépendante ou encore à une hybridation. Cette augmentation massive du nombre de gènes sur deux noeuds terminaux de l'arbre phylogénétiquement proches mais disposant d'un nombre de gènes significativement différents (25% de moins chez la souche CDC-B9645 que chez la souche CDC-B9738) semblait générer un biais analytique impactant également les prédictions associées aux noeuds proches. Les résultats obtenus avec CAFE semblaient pallier ce problème. Environ 4000 gènes étaient identifiés comme propre à la souche CDC-B9738 et jusqu'à 1000 gènes propres à la souche CDC-9645. Ce résultat peut être expliqué par une perte importante de gènes chez les deux souches de façon indépendante après l'hypothétique gain de gènes massif ayant eu lieu chez leur ancêtre commun. Au cours de cette étude le placement phylogénétique des souches CDC-B9645 et CDC-B9738 a été identifié comme proche de *M. racemosus* UBOCC-A-109153 correspondant à l'analyse de Chibucos et al. (2016). Cependant une analyse récente de Gryganskyi et al. (2018) plaçait ces deux souches comme étant proche de *R. microsporus*, ce résultat supporte l'hypothèse d'une hybridation entre deux espèces, l'une proche d'un *M. racemosus* et l'autre proche d'un *R. microsporus*.

4.6 Discussion et conclusion de l'approche génomique présentée

L'obtention du génome de quatre souches appartenant à quatre espèces de *Mucor* pour lesquelles ce type d'information n'a jamais été accessible publiquement et qui sont associées à des habitats et niches écologiques distinctes a permis d'étendre les connaissances sur le genre *Mucor*. Ces connaissances sont d'autant plus importantes que jusqu'à présent seuls les génomes de trois espèces du genre sont disponibles publiquement : *M. irregularis* (deux souches), *M. indicus* (une souche) et *M. circinelloides* (six souches) (Findley et al., 2013; Tang et al., 2016; Chibucos et al., 2016; Corrochano et al., 2016). Parmi ces génomes, seuls six disposent d'annotations accessibles publiquement (ou obtenus sur demande), ces six génomes correspondant à cinq souches de *M. circinelloides* et à la souche de *M. indicus*. Il s'agit de la première étude génomique qui s'intéresse spécifiquement au genre *Mucor*, qui inclut plus de deux espèces du genre et qui intègre des souches venant de milieux non cliniques (Chibucos et al., 2016; Corrochano et al., 2016; Tang et al., 2016; Lopez-Fernandez et al., 2018) permettant par là même d'avoir plus de recul sur le genre *Mucor* dans sa diversité. Au travers de l'étude réalisée, des particularités structurales ont pu être détectées de même que des marques d'une potentielle adaptation à l'habitat et/ou niche écologique des espèces et en particulier concernant la pathogénicité. Ces travaux auront un intérêt, aussi bien dans les domaines de l'agroalimentaire ; en effet, les *Mucor* sont impliqués dans des pertes importantes de produits agroalimentaires bruts et transformés (Garnier et al., 2017) et la production de certains produits fermentés (Walther et al., 2013; Hermet et al., 2012) ; que pour la recherche de médicaments permettant de traiter les Mucormycoses, maladie de plus en plus fréquente et pouvant être mortelle (Ibrahim et al., 2012; Chibucos et al., 2016; Lin et al., 2017; Lopez-Fernandez et al., 2018; Pilmis et al., 2018).

Au cours de cette étude, des caractéristiques communes au genre ont été identifiées. Les génomes de *Mucor* étudiés présentaient une taille comprise entre 35 et 47Mb, ce qui est cohérent avec la taille moyenne des génomes séquencés de Mucoromycotina (38Mb) tout comme leur nombre de gènes compris entre (9997 et 12571) (Mohanta 2015). La structure de ces génomes est extrêmement changeante : aucune synténie n'a été détectée entre génomes proches que ce soit à l'échelle des génomes ou à l'échelle de groupe de gènes spécifiques. Ainsi, même lors de la recherche de cluster de gènes associés à la production de métabolites secondaires, bien que des gènes d'ossature (FAS/NRPS/terpènes synthases) soient détectés aucune synténie entre espèces n'a été observée sur les régions génomiques portant ces gènes. En effet, aucun gène

commun entre les génomes des différentes espèces n'a été retrouvé dans l'environnement génétique/génomique de ces gènes. Or, en plus d'un manque de synténie, ces études des différents génomes ont permis de mettre en évidence l'absence de cluster de gènes impliqués dans la synthèse de métabolites secondaires pourtant couramment retrouvés chez les Ascomycota et les Basidiomycota et dont la présence est souvent étendue aux *Fungi* en général (Lind et al., 2017; Wisecaver and Rokas, 2015). Les différentes études menées sur la potentielle présence de ces clusters chez les *Mucor* n'avaient pas prouvé avec certitude leur absence. Notre étude, sur une douzaine de génomes de *Mucor*, démontre qu'il n'existe pas de cluster génétique liés à la synthèse de métabolites secondaires chez les souches étudiées.

De façon intéressante, lors des analyses d'expansions contraction de familles de gènes, l'ancêtre le plus proche des souches appartenant aux espèces utilisées en affinage de fromage (*M. fuscus* et *M. lanceolatus*) présentait un faible nombre gènes dupliqués ayant des fonctions spécifiques d'intérêt pour l'adaptation à la matrice fromagère : fixation des ions métalliques (le fromage est un milieu dans lequel l'accès au ions métalliques est restreint) et peptidases (enzymes pouvant jouer un rôle important dans l'exploitation du substrat "matrice fromagère" et pour la génération d'arômes (Sousa et al., 2001; Ardö, 2006). D'autre part, un grand nombre de gènes était perdu avec parmi eux des gènes de résistance au stress. Dans un contexte d'espèces technologiques (Morin-Sardin et al., 2016), ces pertes de gènes pourraient être dues à un relâchement global des contraintes sélectives sur des populations restreintes à l'environnement fromager conduisant d'une part à des pertes de gènes et des contractions de familles de gènes (Demuth et Hahn, 2009), et d'autre part à une sélection forte par l'environnement fromager ou par l'homme pour un faible nombre de trait d'intérêt. Parmi ces traits d'intérêts, l'accession au ions métalliques et en particulier au fer est un élément critique, aussi bien pour l'adaptation à la matrice fromagère que pour la pathogénicité des espèces pathogènes opportunistes. Au cours de cette étude, les gènes associés aux mécanismes actuellement répertoriés chez les champignons permettant l'acquisition du fer ainsi que des éléments clefs de la régulation, le stockage et l'utilisation de ce dernier ont été identifiés chez chacune des dix souches de *Mucor*. D'autre part des gènes impliqués dans l'acquisition du fer et expérimentalement identifiés comme importants pour la pathogénicité de *M. circinelloides* ou *R. delemar* ont été retrouvés en plus grand nombre chez les souches au mode de vie de pathogène opportuniste et en quantité restreinte chez les souches utilisées en affinage de fromage.

Parmi les résultats de l'étude, l'un des plus surprenants fut l'incohérence de placement

phylogénétique des souches CDC-B9645 et CDC-B9738 initialement assigné en tant que *M. racemosus* (Chibucos et al., 2016) et par la suite réassignées dans l'étude de Gryganskyi et al. (2018) à l'espèce *R. microsporus*. Les résultats présenté dans cette étude tendent à montrer que ces résultats seraient potentiellement liés à une hybridation d'une espèce proche d'un *M. racemosus* et d'une autre espèce proche d'un*R. microsporus* chez l'ancêtre de ces deux souches.

Initialement intégrées à notre collection de souches en tant que *M. racemosus* les deux souches CDC-B9645 et CDC-B9738 (Chibucos et al., 2016) ont plus tard été réassignées dans l'étude de Gryganskyi et al. (2018) à l'espèce *R. microsporus*. Ces deux souches appartiennent bien à un clade intégrant majoritairement des espèces de *Mucor* bien séparées du clade regroupant différentes espèces de *Rhizopus* (dont *R. microsporus*) dans l'étude de Chibucos et al. (2016) reposant sur l'utilisation de séquences de 76 gènes orthologues pour la reconstruction phylogénetique. En revanche, il apparaît clairement au sein du clade *R. microsporus* dans l'étude plus récente de Gryganskyi et al. (2018) reposant sur l'analyse de 192 gènes orthologues avec une méthodologie différente. Si cela pose le problème (comme évoqué en section 4.2 dans le manuscrit « Comparative genomics applied to *Mucor* species with different lifestyles » en cours de préparation) de l'impact de la méthode de reconstruction phylogénomique utilisée pour établir les relations entres espèces, un autre questionnement est également apparu : Gryganskyi et al. (2018) montre que la taille des génomes de ces deux souches CDC-B9645 et CDC-B9738 est beaucoup plus importante que celle du plus petit génome disponible de *R. microsporus*, ce qui suggère pour ces auteurs l'existence d'une duplication complète de génome (WGD). On sait que cette duplication est un évènement récurrent chez [et]les Mucorales mais il peut également s'agir d'une hybridation avec d'autres espèces de *Rhizopus*. En effet on sait qu'il est possible de réaliser expérimentalement cette hybridation chez les *Rhizopus* (Schipper et al., 1985), les barrières aux interactions sexuelles non spécifiques étant faibles selon Schipper et al. (1985). Dans notre propre reconstruction phylogénomique basée sur l'utilisation de 52 orthologues (section 4.2, « Comparative genomics applied to *Mucor* species with different lifestyles »), les souches CDC-B9645 et CDC-B9738 se placent dans un clade frère d'un singleton *M. racemosus* et éloigné phylogénétiquement de l'outgroup (racine) intégrant une souche de *R. delemar* (seul *Rhizopus* de notre étude) mais lorsque nous avons réalisé des reconstructions distinctes avec des gènes différents (données non présentées), le placement de CDC-B9645 et CDC-B9738 alterne entre le singleton *M. racemosus* et le groupe de *R. delemar*. Ces dernières données suggérerait davantage l'existence d'un génome hybride (voire de manière beaucoup moins probable, car jamais observé, la présence de deux génomes distincts au sein des hyphes siphonnés de ces souches qui seraient

alors dicaryotiques). Une perspective immédiate, qui est en cours de réalisation au laboratoire, est l'identification au sein de chaque famille de gènes de l'espèce dont le gène est le plus proche de chacun des gènes de la famille appartenant aux souches CDC-B9645 et CDC-B9738. Cette identification est réalisée par clustering en utilisant d'une part le groupe d'espèces utilisé dans la publication en préparation auquel a été ajouté la souche *R. microsporus* var. chinensis CCTCC M201021 dont la taille de génome correspond à celle du genre *Mucor* et d'autre part à partir d'un groupe d'espèce composé de *M. racemosus* UBOCC-A-109155, *M. circinelloides* CBS 277.49, *R. delemar* RA 99-880, *R. microsporus* var. chinensis CCTCC M201021 et les souches CDC-B9645 et CDC-B9738. Cette étude pourrait mettre en avant le caractère hybride de ces génomes possédant à la fois les orthologues de *M. racemosus* et de *R. microsporus* expliquant le positionnement alterné de ces deux souches en fonction des études (Chibucos et al., 2016; Gryganskyi et al., 2018) et/ou des gènes choisis ainsi que d'identifier si l'une ou l'autre des espèces se rapproche plus des *Mucor* ou des *Rhizopus*. Des expériences d'hybridation in situ en fluorescence (FISH) utilisant des sondes spécifiques d'un orthologue de *M. racemosus* d'une part et de *R. microsporus* d'autre part, voire des PCR ciblées menées avec des amorces spécifiques de ces orthologues sur des noyaux micro-disséqués seraient une solution pour vérifier expérimentalement l'absence de dicaryose.

Chapitre 5

Conclusions et perspectives

Les objectifs de ce projet étaient d'une part d'améliorer les connaissances concernant le genre *Mucor* et d'autre part de chercher s'il existait des éléments indiquant une potentielle adaptation de certains *Mucor* à différents habitats et modes de vie. Pour atteindre ces objectifs, des comparaisons de transcriptomes et génomes de *Mucor* ont été engagées. En cela, les études réalisées et présentées dans ce manuscrit ont permis l'acquisition des génomes et transcriptomes de quatre souches (*M. endophyticus* CBS 385-95, *M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153 et *M. racemosus* UBOCC-A-109155) appartenant à des espèces pour lesquelles ce type d'information n'était pas disponible. La comparaison de leurs transcriptomes avec la souche de référence *M. circinelloides* CBS 277.49 et de leurs génomes avec les six génomes de *Mucor* alors disponibles a permis de mettre en évidence des points saillants concernant le genre *Mucor* et l'adaptation potentielle des différentes espèces de *Mucor* à leur habitat et mode de vie.

Tout d'abord, les analyses transcriptomiques ont révélé que sur milieu PDA, il y avait environ 18 000 gènes qui étaient exprimés (nombre de transcrits en moyenne). Une des différences retrouvées a été que *M. endophyticus* CBS 385-95 exprimait moins de gènes que les autres *Mucor spp.* de cette étude. De la même façon, la taille des transcrits a révélé une différence entre *M. endophyticus* CBS 385-95 et les autres *Mucor spp.*. En effet, les transcrits de ce dernier sont de plus petite taille (836 pb contre 1200 pb en moyenne) ce qui pourrait être lié à la présence d'un biais d'assemblage.

Les analyses des séquences des génomes ont montré de nombreuses différences en terme de taille. *M. endophyticus* CBS 385-95 possède un génome plus petit que les génomes des autres *Mucor spp.* de cette étude (*M. fuscus* UBOCC-A-109160, *M. lanceolatus* UBOCC-A-109153 et *M. racemosus* UBOCC-A-109155). La taille de son génome est de 35 Mb contre par exemple 46 Mb pour *M. racemosus* UBOCC-A-109155. Une autre caractéristique remarquable est la taille importante des génomes de deux souches initialement assignées à *M. racemosus* (CDC-B9645 et

CDC-B9738) qui ont été réassignées en 2018 à l'espèce *Rhizopus microsporus* (Gryganskyi et al., 2018). Alors que *M. racemosus* UBOCC-A-109155 a un génome d'une taille de 46 Mb, les souches CDC-B9645 et CDC-B9738 ont des génomes avec une taille de 64 Mb et 75 Mb respectivement, ce qui est plus du double de celle du plus petit génome disponible de *R. microsporus* (Gryganskyi et al., 2018). Ce dernier point suggère d'après Gryganskyi et al. (2018) l'existence possible d'une duplication complète de génome (*whole genome duplication*, WGD). Cette hypothèse correspond aux observations en terme de nombre de gènes : les quatre *Mucor* de notre étude possèdent en moyenne 11 724 gènes alors que le génome de CDC-B9738 comprend environ 21 000 gènes détectés.

Parmi les résultats obtenus, ceux concernant la structure des génomes de ces champignons apparaissent également d'intérêt. Les *Mucor spp.* sont des "*early divergent fungi*, et leur structure génomique a été explorée ici afin de vérifier les similitudes et différences avec les génomes mieux connus des Ascomycota et Basidiomycota notamment. Il a été remarqué que les génomes de *Mucor* sont soumis à des modifications importantes et fréquentes comme montré par l'absence de synténie entre les souches étudiées.

Une différence majeure entre les génomes concerne les éléments transposables (TE). Pour les deux souches ayant été isolées sur des matrices fromagères (*M. lanceolatus* UBOCC-A-109153 et *M. fuscus* UBOCC-A-109160), le pourcentage de TE dans le génome est plus important que pour la souche endophyte (entre 15 et 22% par rapport à 5%). La souche *M. lanceolatus* UBOCC-A-109153 semble être passée par une étape d'expansion de TE qu'il s'agisse des transposons ou des retrotransposons après la spéciation qui la sépare de *M. fuscus* UBOCC-A-109160. Ces éléments sont principalement non autonomes. Cette spécialisation à un milieu aurait entrainé un relâchement des pressions de sélection sur de nombreux gènes n'étant pas sous sélection positive, facilitant l'insertion de TE dans le génome à de multiples endroits. L'expansion exponentielle de ces TE aurait par la suite été endiguée par les mécanismes de protection du génome via la mutation spécifique des éléments largement répétés ce qui les aurait rendus non autonomes. *M. fuscus* quant à lui est retrouvé dans d'autres environnements que le milieu fromager. Le génome de ce dernier contient moins d'éléments transposables mais ceux-ci sont autonomes. En effet, ils ont aussi été détectés comme étant actifs puisque de nombreux transcrits ont été retrouvés dans l'étude des transcriptomes. Dans le génome de *M. racemosus* UBOCC-A-109155, près de 37% du génome est représenté par des TE. Parmi eux, certains pourraient avoir un rôle de régulation des gènes. En effet, des études ont montrés que l'insertion de TE à proximité de gènes pouvaient

modifier leur régulation et notamment les rendre inductibles au stress, ce qui pourrait avoir un impact positif sur le caractère ubiquiste de la souche et sa capacité de développement selon plusieurs modes de vies (saprophitisme ou pathogène/contaminant).

Au cours de l'annotation experte des génomes, de grandes sections génomiques (jusqu'à 65kb) sans aucune annotation (ni gènes, ni éléments répétés) ont été détectées. La présence de ces régions est d'autant plus étonnante que lorsque l'on observe des régions géniques, les gènes sont rarement séparés par plus de 1000pb. Ces " régions blanches" de grande taille sont présentes au sein des génomes de l'ensemble des espèces que nous avons étudiées. Le rôle de ces régions et la manière dont elles ont émergé dans le génome restent à l'heure actuelle une énigme. Les hypothèses que nous pourrions formuler seraient que ces sections contribuent à la structure tridimenstionnelle de l'ADN ou encore alors qu'elles comportent des éléments jusqu'à présent inconnus et/ou non identifiés. De la même façon que pour le manque de synténie, le séquençage d'autres souches du genre *Mucor* permettra de savoir si cette structure génomique est conservée chez toutes les espèces.

L'une des dernières particularités structurales a été la découverte de l'absence de clusters de gènes liés à la production de métabolites secondaires. Cette découverte est d'ailleurs en accord avec l'absence de synténie au sein des génomes étudiés. Les clusters (c'est-à-dire des regroupements de gènes codant des protéines impliquées dans une même voie de biosynthèse) sont très bien décrits chez les champignons supérieurs tels que les Ascomycota et les Basidiomycota (Rokas et al., 2018), notamment dans le cadre des synthèses des métabolites secondaires (les *Biosynthetic Gene Clusters* ou BGC). Leurs rôles réels dans l'évolution n'est pas encore clarifié. Les phénomènes de regroupements en clusters de gènes font partie du mécanisme évolutif des champignons et même si leurs rôles exacts dans l'évolution ne sont pas clairement définis, il a été suggéré qu'ils correspondraient à une adaptation contre l'accumulation de composés toxiques et pourraient faciliter l'acquisition ou la perte de voies complètes par expansion/contraction ou HGT (Rokas et al., 2018).

L'une des questions posées au début de cette étude concernait l'existence d'empreintes génomiques expliquant l'adaptation à un habitat ou à un mode de vie : entre autres s'est posé la question de savoir s'il existait des gènes spécifiques d'un mode de vie ou d'une niche écologique. Pour cela il convenait de déterminer ce qu'était le *core genome* et le *pan genome*. Au sein du *core genome*, des gènes codant des enzymes impliquées dans la synthèse de métabolites secondaires ont été identifiés et leur expression a été constatée sur milieu PDA. Des gènes identifiés comme

codant des Polyketides synthases (PKS) (par analogie aux annotations déjà réalisées pour la souche de *M. circinelloides* CBS 385-95 dont le génome était disponible publiquement), des Terpènes synthase (TPS) et des Peptides Non Ribosomiques Synthetase (NRPS) ont été retrouvées dans les différents génomes de *Mucor* étudiés. Cependant, en examinant la structure des gènes codant les PKS putatives, il s'est avéré qu'il ne s'agissait probablement pas de gènes codant des PKS mais de gènes codant des *Fatty acid Synthase* (FAS). Chez certaines souches étudiées, deux gènes codant ces FAS ont été retrouvés. Par contre chez les souches qui ont été isolées sur milieu fromager (*M. lanceolatus* et *M. fuscus*), il n'existe qu'un seul gène. De manière intéressante, les analyses de transcriptomique ont mis en évidence, chez les souches possédant deux gènes codant des FAS, qu'un seul des deux gènes était exprimé sur milieu PDA. Les souches dites technologiques auraient pu perdre un gène codant une FAS sous l'effet d'un relâchement des pressions de sélection positives sur certaines régions génomiques lié à l'adaptation à l'utilisation technologique par l'Homme. La fonction de ces gènes reste à ce jour à déterminer. De plus, dans le génome de *M. racemosus* UBOCC-A-109155, un domaine Ketosynthase (KS) supplémentaire a été trouvé. A quoi sert-il ? Offre-t-il un avantage évolutif pour une adaptation à d'autres milieux ? Et surtout dans quelles conditions ce gène est-il exprimé puisqu'il ne l'est pas sur milieu synthétique dans les conditions de laboratoire ?

Des gènes associés aux mécanismes actuellement répertoriés chez les champignons permettant l'acquisition du fer ainsi que des éléments clefs de la régulation, le stockage et l'utilisation de ce dernier ont été identifiés chez chacune des dix souches de *Mucor*. Cette recherche a permis d'une part d'identifier des candidats associés au transport et à la production de rhizoferrine, un type de sidérophore retrouvé uniquement chez les Mucoromycota et les bactéries. Jusqu'à présent, seul le gène *rfs* impliqué dans la synthèse de ce sidérophore avait été identifié chez les Mucorales (représentés par *R. delemar* (Carroll et al., 2017)) les autres gènes candidats (transport, synthèse) associés au métabolisme de la rhizoferrine sont les premiers identifiés chez les Mucoromycota. D'autre part des gènes impliqués dans l'acquisition du fer et expérimentalement identifiés comme importants pour la pathogénicité de *M. circinelloides* ou *R. delemar* ont été retrouvés en plus grand nombre chez les souches au mode de vie de pathogène opportuniste et en quantité restreinte chez les souches utilisées en affinage de fromage.

Des analyses en cours qui seront réalisées à court terme par mes soins devraient permettre d'amener de nouveaux éléments de réponse concernant la présence d'empreintes génomiques de l'adaptation à un habitat ou à un mode de vie. Ces analyses viendront compléter l'article

en cours de préparation « *Comparative genomics applied to* Mucor *species with different lifestyles* ». Elles concernent en premier lieu la recherche de familles de gènes identifiables et des voies métaboliques KEGG liées appartenant : (i) à l'ensemble des espèces de *Mucor* étudiées, (ii) à chacune des espèces de manière spécifique et (iii) aux espèces au même habitat ou même mode de vie afin de délimiter le répertoire de gènes du genre *Mucor* (*core genome*) et de le comparer à celui d'autres lignées. Il s'agirait aussi d'identifier des familles de gènes qui peuvent participer à la diversification des traits au sein du genre (*species-specific genes*) et celles potentiellement impliquées dans l'adaptation à un habitat ou mode de vie particulier (*lifestyle-specific genes*). Nous réaliserons également des analyses fonctionnelles à partir des résultats des analyses d'expansions et contractions de familles de gènes sur les différents noeuds qui sont présentés dans le chapitre 4.5. Enfin, nous conduirons des analyses complémentaires ciblant des familles de gènes spécifiques et potentiellement en rapport avec les modes de vie d'intérêt. Par exemple, concernant des gènes pouvant jouer un rôle lors du développement sur fromage, nous rechercherons l'existence des gènes impliqués dans une voie métabolique spécifique impliquée dans la synthèse de caroténoïde identifiée chez la souche oléagineuse *M. circinelloides* WJ11 (Vongsangnak et al., 2018) ou encore la présence d'une séquence codant un motif spécifique de lipase chez cette même souche (Komeda et al., 2014). Concernant, des gènes importants pour la pathogénicité, nous rechercherons par exemple la présence de *Single Nucleotide Polymorphism* (SNPs) déjà identifiés chez plusieurs souches de *Mucor* au sein de gènes codant des cibles de l'azole (Lupetti et al., 2002; Caramalho et al., 2017). La résistance à l'azole et ses dérivés, couramment utilisés comme antifongiques, pouvant expliquer la résistance d'opportunistes pathogènes aux traitements. Nous rechercherons également les gènes codant divers facteurs de virulence comme ceux codant les adhésines (Chibucos et al., 2016).

Toujours dans le cadre de la soumission de cet article, il semble important de résoudre la question liée au génome dupliqué ou résultant d'une hybridation des deux souches CDC-B9596 et CDC-B9738 assignées à l'espèce *R. microsporus* (Gryganskyi et al., 2018). Des données préliminaires (incongruences de reconstructions phylogénétiques et placement alterné au sein d'un groupe *Rhizopus* ou au sein d'un groupe *Mucor* en fonction des gènes utilisés) ont davantage suggéré l'existence d'un génome hybride ce qui est actuellement vérifié par des méthodes bioinformatiques de clustering d'orthologues.

En dehors du cadre de la soumission de cet article, nous pouvons définir des objectifs à plus long terme qui pourront être réalisés au Laboratoire de Biodiversité et Ecologie Microbienne

(LUBEM) en partenariat avec la plateforme ABiMS comme une suite logique à ce travail de thèse. La découverte de structures peu communes pour les FAS (*Fatty Acid Synthases*) identifiées dans cette étude, avec l'ensemble des domaines fonctionnels regroupés sur un seul gène mais également pour la FAS identifiée chez *M. racemosus* UBOCC-A-109155, *M. circinelloides* CDC-B5328 et *M. indicus* CDC-B7402 ou la présence inhabituelle d'un domaine KS supplémentaire et une absence d'expression sur milieu synthétique PDA, nous conduisent à placer en perspectives des études visant à élucider le rôle joué par ces protéines. Des outils d'inactivation étant disponibles chez *M. circinelloides* (Garre et al., 2015), il peut être envisagé d'éteindre l'expression des gènes codant les FAS chez cette espèce afin de possiblement détecter l'incidence de cette suppression sur le phénotype (croissance sur différents milieux, dans différentes conditions, détection de métabolites produits). Des expériences de RT-PCR quantitative pourraient être également envisagées en faisant croître *M. racemosus* UBOCC-A-109155 dans des conditions contrastées afin de vérifier si l'expression de la FAS au domaine KS supplémentaire non exprimée sur milieu PDA peut être induite.

Comme spécifié dans la synthèse bibliographique, les transferts horizontaux (HGT) sont des éléments importants à considérer lorsqu'on s'intéresse à l'évolution adaptative. D'abord surtout décrits chez les procaryotes, on sait qu'ils sont également importants chez les eucaryotes et chez les champignons notamment (Fitzpatrick, 2012). Plusieurs travaux ont montré le rôle qu'ils ont joué dans l'adaptation aux différentes niches chez les champignons pathogènes (Gluck-Thaler and Slot, 2015) chez les espèces fongiques technologiques (Hall et al., 2005; Hall and Dietrich, 2007; Wei et al., 2007) et notamment dans le cadre d'un développement sur fromage (Cheeseman et al., 2014). Des analyses bioinformatiques complémentaires devront donc être menées pour identifier d'éventuels transferts horizontaux. Il semble également important de mener des analyses visant à indentifier les gènes sous sélection positive ($dN/dS > 1$) parmi les génomes de *Mucor* à disposition. Elles pourront affiner la détection de gènes critiques pour l'adaptation au milieu des espèces.

Bibliographie

Alfaro, M., Oguiza, J. A., Ramírez, L. and Pisabarro, A. G. (2014). Comparative analysis of secretomes in basidiomycete fungi. Journal of Proteomics *102*, 28–43.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410.

Alvarez, E., Cano, J., Stchigel, A. M., Sutton, D. A., Fothergill, A. W., Salas, V., Rinaldi, M. G. and Guarro, J. (2011). Two new species of Mucor from clinical samples. Medical mycology *49*, 62–72.

Alvarez, E., Sutton, D. A., Cano, J., Fothergill, A. W., Stchigel, A., Rinaldi, M. G. and Guarro, J. (2009). Spectrum of zygomycete species identified in clinically significant specimens in the United States. Journal of clinical microbiology *47*, 1650–1656.

Anders, S., Pyl, P. T. and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. http ://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ardö, Y. (2006). Flavour formation by amino acid catabolism. Biotechnology advances *24*, 238–242.

Ashwin, N. M., Barnabas, L., Ramesh Sundar, A., Malathi, P., Viswanathan, R., Masi, A., Agrawal, G. K. and Rakwal, R. (2017). Comparative secretome analysis of Colletotrichum falcatum identifies a cerato-platanin protein (EPL1) as a potential pathogen-associated molecular pattern (PAMP) inducing systemic resistance in sugarcane. Journal of Proteomics *169*, 2–20.

Austin, D. G., Bu'Lock, J. D. and Winstanley, D. J. (1969). Trisporic acid biosynthesis and carotenogenesis in Blakesleea trispora. The Biochemical journal *113*, 34P.

Aylward, J., Steenkamp, E. T., Dreyer, L. L., Roets, F., Wingfield, B. D. and Wingfield, M. J. (2017). A plant pathology perspective of fungal genome sequencing. IMA Fungus *8*, 1–45.

Bankar, K. G., Todur, V. N., Shukla, R. N. and Vasudevan, M. (2015). Ameliorated de novo transcriptome assembly using Illumina paired end sequence data with Trinity Assembler. Genomics data *5*, 352–359.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A. (2012). SPAdes : a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology *19*, 455–477.

Bao, W., Kojima, K. K. and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA *6*, 11.

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G. and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. Protein engineering, design & selection : PEDS *17*, 349–356.

Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004). Consistent over-estimation of gene number in complex plant genomes. Current Opinion in Plant Biology *7*, 732–736.

Bennetzen, J. L. and Park, M. (2018). Distinguishing friends, foes, and freeloaders in giant genomes. Current Opinion in Genetics and Development *49*, 49–55.

Benz, J. P., Protzko, R. J., Andrich, J. M., Bauer, S., Dueber, J. E. and Somerville, C. R. (2014). Identification and characterization of a galacturonic acid transporter from Neurospora crassa and its application for Saccharomyces cerevisiae fermentation processes. Biotechnology for Biofuels *7*, 20.

Bertaux, J., Schmid, M., Prevost-Boure, N. C., Churin, J. L., Hartmann, A., Garbaye, J. and Frey-Klett, P. (2003). In situ identification of intracellular bacteria related to Paenibacillus spp. in the mycelium of the ectomycorrhizal fungus Laccaria bicolor S238N. Applied and Environmental Microbiology *69*, 4243–4248.

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C. and Hane, J. K. (2018). Accessories Make the Outfit : Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. Molecular Plant-Microbe Interactions *1*, MPMI–06–17–0135.

Billiard, S., López-Villavicencio, M., Hood, M. E. and Giraud, T. (2012). Sex, outcrossing and mating types : unsolved questions in fungi and beyond. Journal of evolutionary biology *25*, 1020–1038.

Biscotti, M. A., Olmo, E. and Heslop-Harrison, J. S. P. (2015). Repetitive DNA in eukaryotic genomes.

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de Los Santos, E. L. C., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T. and Medema, M. H. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic acids research *45*, W36–W41.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic : a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) *30*, 2114–2120.

Boza, V., Brejova, B. and Vinar, T. (2017). DeepNano : Deep recurrent neural networks for base calling in MinION Nanopore reads. PLoS ONE *12*, e0178751.

Brakhage, A. A. and Schroeckh, V. (2011). Fungal secondary metabolites - strategies to activate silent gene clusters. Fungal genetics and biology : FG & B *48*, 15–22.

Brown, T. (2002). Genomes. 2 edition, Garland Science.

Burgeff, H. (1924). Untersuchungen über Sexualität und Parasitismus bei Mucorineen, vol. 1,. G. Fischer.

Caboche, S. U. d. L. and Even, G. G. D. (2012). Paired-end versus mate-pair. Technical report Universite de Lille.

Cacho, R. A., Tang, Y. and Chooi, Y.-H. (2014). Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. Frontiers in microbiology *5*, 774.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A. and Yandell, M. (2008). MAKER : an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research *18*, 188–196.

Caramalho, R., Tyndall, J. D. A., Monk, B. C., Larentis, T., Lass-Flörl, C. and Lackner, M. (2017). Intrinsic short-tailed azole resistance in mucormycetes is due to an evolutionary conserved aminoacid substitution of the lanosterol 14α-demethylase. Scientific Reports *7*, 15898.

Carlile, M. J. (1995). The Success of the Hypha and Mycelium. In The Growing Fungus SE - 1 pp. 3–19. 978-0-412-46600-7.

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., Van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. and Carucci, D. J. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature *419*, 512–519.

Carr, M., Bensasson, D. and Bergman, C. M. (2012). Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. PLoS ONE *7*, e50978.

Carroll, C. S., Grieve, C. L., Murugathasan, I., Bennet, A. J., Czekster, C. M., Liu, H., Naismith, J. and Moore, M. M. (2017). The rhizoferrin biosynthetic gene in the fungal pathogen Rhizopus delemar is a novel member of the NIS gene family. The international journal of biochemistry & cell biology *89*, 136–146.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S., Subhraveti, P. and Karp, P. D. (2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic acids research *46*, D633–D639.

Castanera, R., Borgognone, A., Pisabarro, A. G. and Ramírez, L. (2017). Biology, dynamics, and applications of transposable elements in basidiomycete fungi. Applied Microbiology and Biotechnology *101*, 1337–1350.

Castanera, R., Lopez-Varas, L., Borgognone, A., LaButti, K., Lapidus, A., Schmutz, J., Grimwood, J., Perez, G., Pisabarro, A. G., Grigoriev, I. V., Stajich, J. E. and Ramirez, L. (2016). Transposable Elements versus the Fungal Genome : Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS genetics *12*, e1006108.

Cavalier-Smith, T. (2001). What are fungi ? In Systematics and Evolution pp. 3–37. Springer.

Celotto, A. M. and Graveley, B. R. (2001). Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. Genetics *159*, 599–608.

Chaisson, M. J., Wilson, R. K. and Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. Nature Reviews Genetics *16*, 627–640.

Chang, S.-T. (2006). The World Mushroom Industry : Trends and Technological Development, vol. 8,. begell.

Cheeseman, K., Ropars, J., Renault, P., Dupont, J., Gouzy, J., Branca, A., Abraham, A.-L., Ceppi, M., Conseiller, E., Debuchy, R., Malagnac, F., Goarin, A., Silar, P., Lacoste, S., Sallet, E., Bensimon, A., Giraud, T. and Brygoo, Y. (2014). Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nature communications *5*, 2876.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y. and Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE *8*, e62856.

Cherukuri, Y. and Janga, S. C. (2016). Benchmarking of de novo assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. BMC Genomics *17*, 507.

Chibucos, M. C., Soliman, S., Gebremariam, T., Lee, H., Daugherty, S., Orvis, J., Shetty, A. C., Crabtree, J., Hazen, T. H., Etienne, K. A., Kumari, P., O'Connor, T. D., Rasko, D. A., Filler, S. G., Fraser, C. M., Lockhart, S. R., Skory, C. D., Ibrahim, A. S. and Bruno, V. M. (2016). An integrated genomic and transcriptomic survey of mucormycosis-causing fungi. Nature Communications *7*, 1–11.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science (New York, N.Y.) *311*, 1283–1287.

Claudel-Renard, C., Chevalet, C., Faraut, T. and Kahn, D. (2003). Enzyme-specific profiles for genome annotation : PRIAM. Nucleic Acids Research *31*, 6633–6639.

Clay, R. P., Benhamou, N. and Fuller, M. S. (1991). Ultrastructural detection of polysaccharides in the cell walls of two members of the Hyphocytriales. Mycological research *95*, 1057–1064.

Clément, Y., Fustier, M.-A., Nabholz, B. and Glémin, S. (2015). The Bimodal Distribution of Genic GC Content Is Ancestral to Monocot Species. Genome Biology and Evolution *7*, 336–348.

Coghlan, A. and Wolfe, K. H. (2002). Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Research *12*, 857–867.

Corbo, M. R., Lanciotti, R., Albenzio, M. and Sinigaglia, M. (2001). Occurrence and characterization of yeasts isolated from milks and dairy products of Apulia region. International journal of food microbiology *69*, 147–152.

Corrochano, L. M., Kuo, A., Marcet-Houben, M., Polaino, S., Salamov, A., Villalobos-Escobedo, J. M., Grimwood, J., Álvarez, M. I., Avalos, J., Bauer, D., Benito, E. P., Benoit, I., Burger, G., Camino, L. P., Cánovas, D., Cerdá-Olmedo, E., Cheng, J. F., Domínguez, A., Eliáš, M., Eslava, A. P., Glaser, F., Gutiérrez, G., Heitman, J., Henrissat, B., Iturriaga, E. A., Lang, B. F., Lavín, J. L., Lee, S. C., Li, W., Lindquist, E., López-García, S., Luque, E. M., Marcos, A. T., Martin, J., McCluskey, K., Medina, H. R., Miralles-Durán, A., Miyazaki, A., Muñoz-Torres, E., Oguiza, J. A., Ohm, R. A., Olmedo, M., Orejas, M., Ortiz-Castellanos, L., Pisabarro, A. G., Rodríguez-Romero, J., Ruiz-Herrera, J., Ruiz-Vázquez, R., Sanz, C., Schackwitz, W., Shahriari, M., Shelest, E., Silva-Franco, F., Soanes, D., Syed, K., Tagua, V. G., Talbot, N. J., Thon, M. R., Tice, H., de Vries, R. P., Wiebenga, A., Yadav, J. S., Braun, E. L., Baker, S. E., Garre, V., Schmutz, J., Horwitz, B. A., Torres-Martínez, S., Idnurm, A., Herrera-Estrella, A., Gabaldón, T. and Grigoriev, I. V. (2016). Expansion of Signal Transduction Pathways in Fungi by Extensive Genome Duplication. Current Biology *26*, 1577–1584.

Cuomo, C. A. and Birren, B. W. (2010). The fungal genome initiative and lessons learned from genome sequencing. Methods in enzymology *470*, 833–855.

Dantigny, P., Guilmart, A. and Bensoussan, M. (2005). Basis of predictive mycology. International Journal of Food Microbiology *100*, 187–196.

Daskalov, A., Heller, J., Herzog, S., Fleissner, A. and Glass, N. L. (2017). Molecular mechanisms regulating cell fusion and heterokaryon formation in filamentous fungi. Microbiol Spectrum *5*.

Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. and Enright, A. J. (2013). Kraken : a set of tools for quality control and analysis of high-throughput sequence data. Methods (San Diego, Calif.) *63*, 41–49.

de Boer, W., Folman, L. B., Summerbell, R. C. and Boddy, L. (2005). Living in a fungal world : impact of fungi on soil bacterial niche development. FEMS microbiology reviews *29*, 795–811.

de Souza, J. I., Pires-Zottarelli, C. L. A., Dos Santos, J. F., Costa, J. P. and Harakava, R. (2012). Isomucor (Mucoromycotina) : a new genus from a Cerrado reserve in state of Sao Paulo, Brazil. Mycologia *104*, 232–241.

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. and Birren, B. W. (2005). The genome sequence of the rice blast fungus Magnaporthe grisea. Nature *434*, 980.

Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. Bioessays *31*, 29–39.

Deroy, A. (2015). Evolution et adaptation des champignons saprophytes : les systèmes impliqués dans la dégradation du bois chez Trametes versicolor. PhD thesis,.

Desirò, A., Faccio, A., Kaech, A., Bidartondo, M. I. and Bonfante, P. (2015). Endogone, one of the oldest plant-associated fungi, host unique Mollicutes-related endobacteria. New Phytologist *205*, 1464–1472.

Dobin, A. and Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. Current protocols in bioinformatics *51*, 11.14.1–19.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M. and Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research *7*.

Duba, A., Goriewa-Duba, K. and Wachowska, U. (2018). A review of the interactions between wheat and wheat pathogens : Zymoseptoria tritici, fusarium spp. and parastagonospora nodorum. International Journal of Molecular Sciences *19*.

El Baidouri, M., Kim, K. D., Abernathy, B., Arikit, S., Maumus, F., Panaud, O., Meyers, B. C. and Jackson, S. A. (2015). A new approach for annotation of transposable elements using small RNA mapping. Nucleic acids research *43*, e84.

El-Komy, M. H., Saleh, A. A., Eranthodi, A. and Molan, Y. Y. (2015). Characterization of novel trichoderma asperellum isolates to select effective biocontrol agents against tomato fusarium wilt. Plant Pathology Journal *31*, 50–60.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of molecular biology *300*, 1005–1016.

Essig, A., Hofmann, D., Münch, D., Gayathri, S., Künzler, M., Kallio, P. T., Sahl, H. G., Wider, G., Schneider, T. and Aebi, M. (2014). Copsin, a novel peptide-based fungal antibiotic interfering with the peptidoglycan synthesis. Journal of Biological Chemistry *289*, 34953–34964.

Et, M. B., Centeleghe, J. L. and Milliere, J. B. (1972). Etude d ' un accident en fromagerie de type Â« Camembert Â» causé par des mucorales. Le Lait *52*, 141–148.

Ferreira, J. A., Lennartsson, P. R., Edebo, L. and Taherzadeh, M. J. (2013). Zygomycetes-based biorefinery : Present status and future prospects. Bioresource technology *135*, 523–532.

Feschotte, C. and Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genetics *41*, 331–368.

Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J. A., Schoenfeld, D., Nomicos, E., Park, M., Kong, H. H. and Segre, J. A. (2013). Topographic diversity of fungal and bacterial communities in human skin. Nature *498*, 367–370.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. and Bateman, A. (2016). The Pfam protein families database : towards a more sustainable future. Nucleic acids research *44*, D279–85.

Fisher, K. E. and Roberson, R. W. (2016). Hyphal tip cytoplasmic organization in four zygomycetous fungi. Mycologia *108*, 533–542.

Fitzpatrick, D. A. (2012). Horizontal gene transfer in fungi. FEMS Microbiology Letters *329*, 1–8.

Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS ONE *6*, e16526.

Frac, M., Hannula, S. E., Belka, M. and JÈ©dryczka, M. (2018). Fungal biodiversity and their role in soil health. Frontiers in Microbiology *9*, 707.

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., Rasmussen, J. B., Solomon, P. S., McDonald, B. A. and Oliver, R. P. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. Nature Genetics *38*, 953–956.

Fuka, M. M., Wallisch, S., Engel, M., Welzl, G., Havranek, J. and Schloter, M. (2013). Dynamics of bacterial communities during the ripening process of different Croatian cheese types derived from raw ewe's milk cheeses. PLoS ONE *8*, e80734.

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. C., 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. and Birren, B. (2003). The genome sequence of the filamentous fungus Neurospora crassa. Nature *422*, 859–868.

Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A. and Birren, B. (2005). Genomics of the fungal kingdom : Insights into eukaryotic biology. Genome Research *15*, 1620–1631.

Garnier, L., Valence, F. and Mounier, J. (2017). Diversity and Control of Spoilage Fungi in Dairy Products : An Update. Microorganisms *5*, 42.

Garre, V., Barredo, J. L. and Iturriaga, E. A. (2015). Transformation of Mucor circinelloides f. lusitanicus protoplasts. In Genetic Transformation Systems in Fungi, Volume 1 pp. 49–59. Springer.

Gebremariam, T., Liu, M., Luo, G., Bruno, V., Phan, Q. T., Waring, A. J., Edwards, J. E., Filler, S. G., Yeaman, M. R. and Ibrahim, A. S. (2014). CotH3 mediates fungal invasion of host cells during mucormycosis. Journal of Clinical Investigation *124*, 237–250.

Gerwien, F., Skrahina, V., Kasper, L., Hube, B. and Brunke, S. (2018). Metals in fungal virulence. FEMS microbiology reviews *42*.

Gladyshev, E. (2017). Repeat-Induced Point Mutation and Other Genome Defense Mechanisms in Fungi. Microbiology spectrum *5*.

Gluck-Thaler, E. and Slot, J. C. (2015). Dimensions of Horizontal Gene Transfer in Eukaryotic Microbial Pathogens. PLoS pathogens *11*, e1005156.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 genes. Science (New York, N.Y.) *274*, 546,563–567.

Gryganskyi, A. P., Golan, J., Dolatabadi, S., Mondo, S., Robb, S., Idnurm, A., Muszewska, A., Steczkiewicz, K., Masonjones, S., Liao, H.-L., Gajdeczka, M. T., Anike, F., Vuek, A., Anishchenko, I. M., Voigt, K., de Hoog, G. S., Smith, M. E., Heitman, J., Vilgalys, R. and Stajich, J. E. (2018). Phylogenetic and Phylogenomic Definition of Rhizopus Species. G3 (Bethesda, Md.) *8*, 2007–2018.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013). QUAST : quality assessment tool for genome assemblies. Bioinformatics (Oxford, England) *29*, 1072–1075.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R. and Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology *9*, R7.

Habig, M., Quade, J. and Stukenbrock, E. H. (2017). Forward genetics approach reveals host genotype-dependent importance of accessory chromosomes in the fungal wheat pathogen Zymoseptoria tritici. mBio *8*.

Hall, C., Brachat, S. and Dietrich, F. S. (2005). Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. Eukaryotic Cell *4*, 1102–1115.

Hall, C. and Dietrich, F. S. (2007). The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. Genetics *177*, 2293–2307.

Hammill, T. M. and Secor, D. L. (1983). The Number of Nuclei in Sporangiospores of Mucor mucedo. Mycologia *75*, 648–655.

Hawksworth, D. L. and Lücking, R. (2017). Fungal Diversity Revisited : 2.2 to 3.8 Million Species. Microbiology Spectrum *5*.

Hawksworth, D. L. and Rossman, A. Y. (1997). Where are all the undescribed fungi ? Phytopathology *87*, 888–891.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers : The history of sequencing DNA. Genomics *107*, 1–8.

Hermet, A., Méheust, D., Mounier, J., Barbier, G. and Jany, J. L. (2012). Molecular systematics in the genus Mucor with special regards to species encountered in cheese. Fungal Biology *116*, 692–705.

Hess, J., Skrede, I., Chaib De Mares, M., Hainaut, M., Henrissat, B. and Pringle, A. (2018). Rapid Divergence of Genome Architectures Following the Origin of an Ectomycorrhizal Symbiosis in the Genus Amanita. Molecular Biology and Evolution *1*, msy179–msy179.

Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., Huhndorf, S., James, T., Kirk, P. M., Lücking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P. B., McLaughlin, D. J., Powell, M. J., Redhead, S., Schoch, C. L., Spatafora, J. W., Stalpers, J. A., Vilgalys, R., Aime, M. C., Aptroot, A., Bauer, R., Begerow, D., Benny, G. L., Castlebury, L. A., Crous, P. W., Dai, Y. C., Gams, W., Geiser, D. M., Griffith, G. W., Gueidan, C., Hawksworth, D. L., Hestmark, G., Hosaka, K., Humber, R. A., Hyde, K. D., Ironside, J. E., Kõljalg, U., Kurtzman, C. P., Larsson, K. H., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J. M., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J. D., Roux, C., Ryvarden, L., Sampaio, J. P., Schüßler, A., Sugiyama, J., Thorn, R. G., Tibell, L., Untereiner, W. A., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M. M., Winka, K., Yao, Y. J. and Zhang, N. (2007). A higher-level phylogenetic classification of the Fungi. Mycological Research *111*, 509–547.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. and Quesneville, H. (2014). PASTEC : an automatic transposable element classification tool. PloS one *9*, e91929.

Hoen, D. R. and Bureau, T. E. (2015). Discovery of novel genes derived from transposable elements using integrative genomic analysis. Molecular biology and evolution *32*, 1487–1506.

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D. D., Quesneville, H., Smit, A., Wheeler, T. J., Bureau, T. E. and Blanchette, M. (2015). A call for benchmarking transposable element annotation methods. Mobile DNA *6*, 13.

Hoff, K. J. and Stanke, M. (2013). WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic acids research *41*, W123–8.

Hoffmann, K., Pawlowska, J., Walther, G., Wrzosek, M., de Hoog, G. S., Benny, G. L., Kirk, P. M. and Voigt, K. (2013). The family structure of the Mucorales : a synoptic revision based on comprehensive multigene-genealogies. Persoonia *30*, 57–76.

Holliday, G. L., Davidson, R., Akiva, E. and Babbitt, P. C. (2017). Evaluating Functional Annotations of Enzymes Using the Gene Ontology. Methods in molecular biology (Clifton, N.J.) *1446*, 111–132.

Hollmann, M., Razzazi-Fazeli, E., Grajewski, J., Twaruzek, M., Sulyok, M. and Böhm, J. (2008). Detection of 3-nitropropionic acid and cytotoxicity inMucor circinelloides. Mycotoxin Research *24*, 140–150.

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A. and Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. Nucleic acids research *44*, D81–9.

Huson, D., Mitra, S. and Ruscheweyh, H. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Research *21*, 1552–1560.

Huynen, M. A., Snel, B. and Bork, P. (2001). Inversions and the dynamics of eukaryotic gene order. Trends in genetics : TIG *17*, 304–306.

Ibrahim, A. S., Spellberg, B., Walsh, T. J. and Kontoyiannis, D. P. (2012). Pathogenesis of mucormycosis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America *54 Suppl 1*, S16–22.

Idnurm, A., Walton, F. J., Floyd, A. and Heitman, J. (2008). Identification of the sex genes in an early diverged fungus. Nature *451*, 193–196.

Ingold, C. T. and Others (1978). The biology of Mucor and its allies. Edward Arnold.

Ivashchenko, A. T., Tauasarova, M. I. and Atambayeva, S. A. (2009). Exon-intron structure of genes in complete fungal genomes. Molecular Biology *43*, 24–31.

Jayne, B. and Quigley, M. (2014). Influence of arbuscular mycorrhiza on growth and reproductive response of plants under water deficit : a meta-analysis. Mycorrhiza *24*, 109–119.

Kalitsis, P., Zhang, T., Marshall, K. M., Nielsen, C. F. and Hudson, D. F. (2017). Condensin, master organizer of the genome. Chromosome Research *25*, 61–76.

Kanehisa, M. (2017). Enzyme Annotation and Metabolic Reconstruction Using KEGG. Methods in molecular biology (Clifton, N.J.) *1611*, 135–145.

Karimi, K. and Zamani, A. (2013). Mucor indicus : biology and industrial application perspectives : a review. Biotechnology advances *31*, 466–481.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H. and Fedorova, N. D. (2010). SMURF : Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology : FG & B *47*, 736–741.

Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N. and Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers : Current Challenges and Future Prospective. Evolutionary Bioinformatics *14*, 117693431875865.

Kim, C. S., Winn, M. D., Sachdeva, V. and Jordan, K. E. (2017). K-mer clustering algorithm using a MapReduce framework : application to the parallelization of the Inchworm module of Trinity. BMC bioinformatics *18*, 467.

Knapp, D. G., Németh, J. B., Barry, K., Hainaut, M., Henrissat, B., Johnson, J., Kuo, A., Lim, J. H. P., Lipzen, A., Nolan, M., Ohm, R. A., Tamás, L., Grigoriev, I. V., Spatafora, J. W., Nagy, L. G. and Kovács, G. M. (2018). Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. Scientific Reports *8*, 6321.

Kohler, A., Kuo, A., Nagy, L. G., Morin, E., Barry, K. W., Buscot, F., Canbäck, B., Choi, C., Cichocki, N., Clum, A., Colpaert, J., Copeland, A., Costa, M. D., Doré, J., Floudas, D., Gay, G., Girlanda, M., Henrissat, B., Herrmann, S., Hess, J., Högberg, N., Johansson, T., Khouja, H. R., Labutti, K., Lahrmann, U., Levasseur, A., Lindquist, E. A., Lipzen, A., Marmeisse, R., Martino, E., Murat, C., Ngan, C. Y., Nehls, U., Plett, J. M., Pringle, A., Ohm, R. A., Perotto, S., Peter, M., Riley, R., Rineau, F., Ruytinx, J., Salamov, A., Shah, F., Sun, H., Tarkka, M., Tritt, A., Veneault-Fourrey, C., Zuccaro, A., Tunlid, A., Grigoriev, I. V., Hibbett, D. S. and Martin, F. (2015). Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nature Genetics *47*, 410–415.

Komeda, H., Yamasaki-Yashiki, S., Hoshino, K. and Asano, Y. (2014). Identification and characterization of D-xylulokinase from the D-xylose-fermenting fungus, Mucor circinelloides.

Kopylova, E., Noe, L. and Touzet, H. (2012). SortMeRNA : fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics (Oxford, England) *28*, 3211–3217.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model : application to complete genomes. J Mol Biol *305*, 567–580.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T. and Ussery, D. W. (2007). RNAmmer : Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research *35*, 3100–3108.

Lamb, C. J., Lawton, M. A., Dron, M. and Dixon, R. A. (1989). Signals and transduction mechanisms for activation of plant defenses against microbial attack. Cell *56*, 215–224.

Latgé, J.-P. (2007). The cell wall : a carbohydrate armour for the fungal cell. Molecular microbiology *66*, 279–290.

Lee, S. C., Blake Billmyre, R., Li, A., Carson, S., Sykes, S. M., Huh, E. Y., Mieczkowski, P., Ko, D. C., Cuomo, C. A. and Heitman, J. (2014). Analysis of a food-borne fungal pathogen outbreak : Virulence and genome of a Mucor circinelloides isolate from yogurt. mBio *5*, e01390–14.

Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P. and Dequin, S. (2018). Adaptation of S. cerevisiae to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. Molecular biology and evolution *35*, 1712–1727.

Lephart, P. R., Chibana, H. and Magee, P. T. (2005). Effect of the major repeat sequence on chromosome loss in Candida albicans. Eukaryotic Cell *4*, 733–741.

Lin, E., Moua, T. and Limper, A. H. (2017). Pulmonary mucormycosis : clinical features and outcomes. Infection *45*, 443–448.

Lind, A. L., Wisecaver, J. H., Lameiras, C., Wiemann, P., Palmer, J. M., Keller, N. P., Rodrigues, F., Goldman, G. H. and Rokas, A. (2017). Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLoS Biology *15*, e2003583.

Lischer, H. E. L. and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics *18*, 474.

Liu, M., Lin, L., Gebremariam, T., Luo, G., Skory, C. D., French, S. W., Chou, T. F., Edwards, J. E. and Ibrahim, A. S. (2015). Fob1 and Fob2 Proteins Are Virulence Determinants of Rhizopus oryzae via Facilitating Iron Uptake from Ferrioxamine. PLoS Pathogens *11*, e1004842–e1004842.

Lobanovska, M. and Pilla, G. (2017). Penicillin's Discovery and Antibiotic Resistance : Lessons for the Future ? The Yale journal of biology and medicine *90*, 135–145.

Lodolo, E. J., Kock, J. L. F., Axcell, B. C. and Brooks, M. (2008). The yeast Saccharomyces cerevisiae- the main character in beer brewing. FEMS yeast research *8*, 1018–1036.

Loftus, B. J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I. J., Fraser, J. A., Allen, J. E., Bosdet, I. E., Brent, M. R., Chiu, R., Doering, T. L., Donlin, M. J., D'Souza, C. A., Fox, D. S., Grinberg, V., Fu, J., Fukushima, M., Haas, B. J., Huang, J. C., Janbon, G., Jones, S. J. M., Koo, H. L., Krzywinski, M. I., Kwon-Chung, J. K., Lengeler, K. B., Maiti, R., Marra, M. A., Marra, R. E., Mathewson, C. A., Mitchell, T. G., Pertea, M., Riggs, F. R., Salzberg, S. L., Schein, J. E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C. A., Suh, B. B., Tenney, A., Utterback, T. R., Wickes, B. L., Wortman, J. R., Wye, N. H., Kronstad, J. W., Lodge, J. K., Heitman, J., Davis, R. W., Fraser, C. M. and Hyman, R. W. (2005). The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science (New York, N.Y.) *307*, 1321–1324.

Lopez-Fernandez, L., Sanchis, M., Navarro-Rodriguez, P., Nicolas, F. E., Silva-Franco, F., Guarro, J., Garre, V., Navarro-Mendoza, M. I., Perez-Arques, C. and Capilla, J. (2018). Understanding Mucor circinelloides pathogenesis by comparative genomics and phenotypical studies. Virulence *9*, 707–720.

Lopez-Franco, R. and Bracker, C. E. (1996). Diversity and dynamics of the Spitzenk6rper in growing hyphal tips of higher fungi. Protoplasma *195*, 90–111.

Love, A. C. (2016). Explaining the Origins of Multicellularity : Between Evolutionary Dynamics and Developmental Mechanisms. Multicellularity : Origins and Evolution *1*, 277–295.

Lowe, T. M. and Chan, P. P. (2016). tRNAscan-SE On-line : integrating search and context for analysis of transfer RNA genes. Nucleic acids research *44*, W54–7.

Ludolph, A. C. and Ludolph, A. G. (1991). 3-Nitropropionic acid-abundant xenobiotic excitotoxin linked to putaminal necrosis and tardive dystonia. ANNALS OF \ldots.

Lupetti, A., Danesi, R., Campa, M., Del Tacca, M. and Kelly, S. (2002). Molecular basis of resistance to azole antifungals. Trends in molecular medicine *8*, 76–81.

Ma, L., Chen, Z., Huang, D. W., Kutty, G., Ishihara, M., Wang, H., Abouelleil, A., Bishop, L., Davey, E., Deng, R., Deng, X., Fan, L., Fantoni, G., Fitzgerald, M., Gogineni, E., Goldberg, J. M., Handley, G., Hu, X., Huber, C., Jiao, X., Jones, K., Levin, J. Z., Liu, Y., Macdonald, P., Melnikov, A., Raley, C., Sassi, M., Sherman, B. T., Song, X., Sykes, S., Tran, B., Walsh, L., Xia, Y., Yang, J., Young, S., Zeng, Q., Zheng, X., Stephens, R., Nusbaum, C., Birren, B. W., Azadi, P., Lempicki, R. A., Cuomo, C. A. and Kovacs, J. A. (2016). Genome analysis of three Pneumocystis species reveals adaptation mechanisms to life exclusively in mammalian hosts, vol. 7,.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J. and Salzberg, S. L. (2013). GAGE-B : An evaluation of genome assemblers for bacterial organisms. Bioinformatics *29*, 1718–1725.

Mai, H., Zhang, Y., Li, D., Leung, H. C.-M., Luo, R., Wong, C.-K., Ting, H.-F. and Lam, T.-W. (2018). AC-DIAMOND v1 : accelerating large-scale DNA-protein alignment. Bioinformatics (Oxford, England) *34*, 3744–3746.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y. and Bryant, S. H. (2017). CDD/SPARCLE : functional classification of proteins via subfamily domain architectures. Nucleic acids research *45*, D200–D203.

Mardis, E. R. (2017). DNA sequencing technologies : 2006-2016. Nature Protocols *12*, 213–218.

Marsh, J. A. and Teichmann, S. A. (2010). How do proteins gain new domains ? Genome Biology *11*, 126.

Martinez, D., Larrondo, L. F., Putnam, N., Sollewijn Gelpke, M. D., Huang, K., Chapman, J., Helfenbein, K. G., Ramaiya, P., Detter, J. C., Larimer, F., Coutinho, P. M., Henrissat, B., Berka, R., Cullen, D. and Rokhsar, D. (2004). Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78. Nature Biotechnology *22*, 695–700.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. Current opinion in chemical biology *17*, 4–11.

Medema, M. H. and Fischbach, M. A. (2015). Computational approaches to natural product discovery. Nature Chemical Biology *11*, 639–648.

Mélida, H., Sandoval-Sierra, J. V., Diéguez-Uribeondo, J. and Bulone, V. (2013). Analyses of extracellular carbohydrates in oomycetes unveil the existence of three different cell wall types. Eukaryotic Cell *12*, 194–203.

Millati, R., Edebo, L. and Taherzadeh, M. J. (2005). Performance of Rhizopus, Rhizomucor, and Mucor in ethanol production from glucose, xylose, and wood hydrolyzates. Enzyme and Microbial Technology *36*, 294–300.

Mohanta, T. K. and Bae, H. (2015). The diversity of fungal genome. Biological procedures online *17*, 8.

Möller, M. and Stukenbrock, E. H. (2017). Evolution and genome architecture in fungal plant pathogens. Nature Reviews Microbiology *15*, 756–771.

Mondo, S. J., Lastovetsky, O. A., Gaspar, M. L., Schwardt, N. H., Barber, C. C., Riley, R., Sun, H., Grigoriev, I. V. and Pawlowska, T. E. (2017). Bacterial endosymbionts influence host sexuality and reveal reproductive genes of early divergent fungi. Nature Communications *8*, 1843.

Monnet, C., Loux, V., Gibrat, J.-F., Spinnler, E., Barbe, V., Vacherie, B., Gavory, F., Gourbeyre, E., Siguier, P., Chandler, M., Elleuch, R., Irlinger, F. and Vallaeys, T. (2010). The arthrobacter arilaitensis Re117 genome sequence reveals its genetic adaptation to the surface of cheese. PloS one *5*, e15489.

Morin-Sardin, S. (2016). Etude physiologique et moleculaire de l'adaptation des Mucor a la matrice fromagere. PhD thesis, UBO.

Morin-Sardin, S., Nodet, P., Coton, E. and Jany, J.-L. (2017). Mucor : A Janus-faced fungal genus with human health impact and industrial applications. Fungal Biology Reviews *31*, 12–32.

Morin-Sardin, S., Rigalma, K., Coroller, L., Jany, J. L. and Coton, E. (2016). Effect of temperature, pH, and water activity on Mucor spp. growth on synthetic medium, cheese analog and cheese. Food Microbiology *56*, 69–79.

Morita, Y., Shibutani, T., Nakanishi, N., Nishikura, K., Iwai, S. and Kuraoka, I. (2013). Human endonuclease v is a ribonuclease specific for inosine-containing RNA. Nature Communications *4*, 2273.

Murat, C., Payen, T., Noel, B., Kuo, A., Morin, E., Chen, J., Kohler, A., Krizsán, K., Balestrini, R., Silva, C. D., Montanini, B., Hainaut, M., Levati, E., Barry, K. W., Belfiori, B., Cichocki, N., Clum, A., Dockter, R. B., Fauchery, L., Guy, J., Iotti, M., Tacon, F. L., Lindquist, E. A., Lipzen, A., Malagnac, F., Mello, A., Molinier, V., Miyauchi, S., Poulain, J., Riccioni, C., Rubini, A., Sitrit, Y., Splivallo, R., Traeger, S., Wang, M., Žifčáková, L., Wipf, D., Zambonelli, A., Paolocci, F., Nowrousian, M., Ottonello, S., Baldrian, P., Spatafora, J. W., Henrissat, B., Nagy, L. G., Aury, J.-M., Wincker, P., Grigoriev, I. V., Bonfante, P. and Martin, F. M. (2018). Pezizomycetes genomes reveal the molecular basis of ectomycorrhizal truffle lifestyle. Nature Ecology & Evolution *1*, 1.

Nagy, L., Tóth, R., Kiss, E., Slot, J. and Gácser, A. (2017). Six key traits of fungi : their evolutionary origins and genetic bases. Microbiology spectrum *5*, 1–22.

Nahas, E. (1988). Control of Lipase Production by Rhizopus oligosporus under Various Growth Conditions. Microbiology *134*, 227–233.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N. and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic acids research *39*, e90.

Navarro-Mendoza, M. I., Pérez-Arques, C., Murcia, L., Martínez-García, P., Lax, C., Sanchis, M., Capilla, J., Nicolás, F. E. and Garre, V. (2018). Components of a new gene family of ferroxidases involved in virulence are functionally specialized in fungal dimorphism. Scientific Reports *8*, 7660.

Nawrocki, E. P. (2014). Annotating functional RNAs in genomes using Infernal. Methods in molecular biology (Clifton, N.J.) *1097*, 163–197.

Nemecek, J. C., Wüthrich, M. and Klein, B. S. (2006). Global control of dimorphism and virulence in fungi. Science *312*, 583–588.

Nielsen, H. (2017). Predicting Secretory Proteins with SignalP. Methods in molecular biology (Clifton, N.J.) *1611*, 59–73.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.-L., Wincker, P., Casaregola, S. and Dequin, S. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proceedings of the National Academy of Sciences *106*, 16333–16338.

Ohm, R. A., Riley, R., Salamov, A., Min, B., Choi, I.-G. and Grigoriev, I. V. (2014). Genomics of wood-degrading fungi. Fungal Genetics and Biology *72*, 82–90.

Paley, S. and Karp, P. D. (2017). Update notifications for the BioCyc collection of databases. Database : the journal of biological databases and curation *2017*.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics *40*, 1413–1415.

Panasenko, V. T. (1967). Ecology of microfungi. The Botanical Review *33*, 189–215.

Partida-Martinez, L. P., Groth, I., Schmitt, I., Richter, W., Roth, M. and Hertweck, C. (2007). Burkholderia rhizoxinica sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus Rhizopus microsporous. International Journal of Systematic and Evolutionary Microbiology *57*, 2583–2590.

Pawlowska, J., Walther, G., Wilk, M., de Hoog, S. and Wrzosek, M. (2013). The use of compensatory base change analysis of ITS2 as a tool in the phylogeny of Mucorales, illustrated by the Mucor circinelloides complex. Organisms Diversity & Evolution *13*, 497–502.

Pellegrin, C., Morin, E., Martin, F. M. and Veneault-Fourrey, C. (2015). Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins. Frontiers in Microbiology *6*, 1278.

Peona, V., Weissensteiner, M. H. and Suh, A. (2018). How complete are "complete" genome assemblies ? - An avian perspective.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols *11*, 1650–1667.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T. and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology *33*, 290–295.

Piégu, B., Bire, S., Arensburger, P. and Bigot, Y. (2015). A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. Molecular Phylogenetics and Evolution *86*, 90–109.

Pillai, S., Gopalan, V. and Lam, A. K.-Y. (2017). Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. Critical Reviews in Oncology / Hematology *116*, 58–67.

Pilmis, B., Alanio, A., Lortholary, O. and Lanternier, F. (2018). Recent advances in the understanding and management of mucormycosis. F1000Research *7*.

Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. and Vitale, L. (2016). GeneBase 1.1 : A tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database *2016*.

Pitt, J. I. and Hocking, A. D. (2009). The Ecology of Fungal Food Spoilage. In Fungi and Food Spoilage pp. 3–9. Springer US Boston, MA.

Plempel, M. (1962). Die zygotropische Reaktion bei Mucorineen. 1. Planta, Berlin *55*, 254–258.

Pombert, J. F., Xu, J., Smith, D. R., Heiman, D., Young, S., Cuomo, C. A., Weiss, L. M. and Keeling, P. J. (2013). Complete genome sequences from three genetically distinct strains reveal high intraspecies genetic diversity in the microsporidian Encephalitozoon cuniculi. Eukaryotic Cell *12*, 503–511.

Porras-Alfaro, A. and Bayman, P. (2011). Hidden Fungi, Emergent Properties : Endophytes and Microbiomes. Annual Review of Phytopathology *49*, 291–315.

Prakash, H., Rudramurthy, S. M., Gandham, P. S., Ghosh, A. K., Kumar, M. M., Badapanda, C. and Chakrabarti, A. (2017). Apophysomyces variabilis : Draft genome sequence and comparison of predictive virulence determinants with other medically important Mucorales. BMC Genomics *18*, 736.

Praphailong, W., Van Gestel, M., Fleet, G. H. and Heard, G. M. (1997). Evaluation of the Biolog system for the identification of food and beverage yeasts. Letters in applied microbiology *24*, 455–459.

Pryszcz, L. P. and Gabaldon, T. (2016). Redundans : an assembly pipeline for highly heterozygous genomes. Nucleic acids research *44*, e113.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M. and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. PLoS Computational Biology *1*, 0166–0175.

Raffaele, S. and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens : Why bigger can be better. Nature Reviews Microbiology *10*, 417–430.

Razin, S. V., Iarovaia, O. V., Sjakste, N., Sjakste, T., Bagdoniene, L., Rynditch, A. V., Eivazova, E. R., Lipinski, M. and Vassetzky, Y. S. (2007). Chromatin Domains and Regulation of Transcription. Journal of Molecular Biology *369*, 597–607.

Redou, V., Navarri, M., Meslet-Cladiere, L., Barbier, G. and Burgaud, G. (2015). Species richness and adaptation of marine fungi from deep-subseafloor sediments. Applied and environmental microbiology *81*, 3571–3583.

Richards, T. A., Leonard, G. and Wideman, J. G. (2017). What Defines the "Kingdom" Fungi ? Microbiology spectrum *5*.

Robert, V. A. and Casadevall, A. (2009). Vertebrate endothermy restricts most fungi as potential pathogens. The Journal of infectious diseases *200*, 1623–1626.

Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics (Oxford, England) *27*, 2325–2329.

Roden, M. M., Zaoutis, T. E., Buchanan, W. L., Knudsen, T. A., Sarkisova, T. A., Schaufele, R. L., Sein, M., Sein, T., Chiou, C. C., Chu, J. H., Kontoyiannis, D. P. and Walsh, T. J. (2005). Epidemiology and Outcome of Zygomycosis : A Review of 929 Reported Cases. Clinical Infectious Diseases *41*, 634–653.

Rokas, A., Wisecaver, J. H. and Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. Nature reviews. Microbiology *16*, 731–744.

Ropars, J., Rodríguez De La Vega, R. C., López-Villavicencio, M., Gouzy, J., Sallet, E., Dumas, É., Lacoste, S., Debuchy, R., Dupont, J., Branca, A. and Giraud, T. (2015). Adaptive horizontal gene transfers between multiple cheese-associated fungi. Current Biology *25*, 2562–2569.

Rotival, M. (2011). Approches integrees du genome et du transcriptome dans les maladies complexes humaines. PhD thesis, Universite Paris 11.

Saeidipour, B. and Bakhshi, S. (2013). The relationship between organizational culture and knowledge management,& their simultaneous effects on customer relation management. Advances in Environmental Biology *7*, 2803–2809.

Sahlin, K., Chikhi, R. and Arvestad, L. (2016). Assembly scaffolding with PE-contaminated mate-pair libraries. Bioinformatics *32*, 1925–1932.

Sajbidor, J., Certik, M. and Dobronova, S. (1988). Influence of different carbon sources on growth, lipid content and fatty acid composition in four strains belonging to Mucorales. Biotechnology letters *10*, 347–350.

Sautour, M., Soares Mansur, C., Divies, C., Bensoussan, M. and Dantigny, P. (2002). Comparison of the effects of temperature and water activity on growth rate of food spoilage moulds. Journal of Industrial Microbiology and Biotechnology *28*, 311–315.

Scheffers, B. R., Joppa, L. N., Pimm, S. L. and Laurance, W. F. (2012). What we know and don't know about Earth's missing biodiversity. Trends in ecology & evolution *27*, 501–510.

Schipper, M. A. (1967). Mucor strictus hagem, a psychrophilic fungus, and Mucor falcatus sp.n. Antonie van Leeuwenhoek *33*, 189–195.

Schipper, M. A. (1969). Zygosporic stages in heterothallic Mucor. Antonie van Leeuwenhoek *35*, 189–208.

Schipper, M. A. (1970). Two species of Mucor with oval- and spherical-spored strains. Antonie van Leeuwenhoek *36*, 475–488.

Schipper, M. A. (1976). Induced azygospore formation in Mucor (Rhizomucor) pusillus by Absidia corymbifera. Antonie van Leeuwenhoek *42*, 141–144.

Schipper, M. A. A., Gauger, W. and Van Den Ende, H. (1985). Hybridization of Rhizopus species. Microbiology *131*, 2359–2365.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic acids research *43*, e37.

Schlebusch, S. and Illing, N. (2012). Next generation shotgun sequencing and the challenges of de novo genome assembly .

Schmidt-Dannert, C. (2015). Biosynthesis of terpenoid natural products in fungi. Advances in Biochemical Engineering/Biotechnology *148*, 19–61.

Schmieder, R., Lim, Y. W. and Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. Bioinformatics *28*, 433–435.

Schulz, E., Wetzel, J., Burmester, A., Ellenberger, S., Siegmund, L. and Wöstemeyer, J. (2016). Sex loci of homothallic and heterothallic Mucorales. Endocytobiosis Cell Res. *27*, 39–57.

Schulz, E., Wetzel, J., Burmester, A., Ellenberger, S., Siegmund, L. and Wostemeyer, J. (2017). Sex loci of homothallic and heterothallic Mucorales, vol. 27,.

Schwartze, V. U., Winter, S., Shelest, E., Marcet-Houben, M., Horn, F., Wehner, S., Linde, J., Valiante, V., Sammeth, M., Riege, K., Nowrousian, M., Kaerger, K., Jacobsen, I. D., Marz, M., Brakhage, A. A., Gabaldón, T., Böcker, S. and Voigt, K. (2014). Gene Expansion Shapes Genome Architecture in the Human Pathogen Lichtheimia corymbifera : An Evolutionary Genomics Analysis in the Ancient Terrestrial Mucorales (Mucoromycotina). PLoS Genetics *10*, e1004496.

Scott, A. L., Richmond, P. A., Dowell, R. D. and Selmecki, A. M. (2017). The Influence of Polyploidy on the Evolution of Yeast Grown in a Sub-Optimal Carbon Source. Molecular biology and evolution *34*, 2690–2703.

Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2013). New and continuing developments at PROSITE. Nucleic acids research *41*, D344–7.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M. (2015). BUSCO : Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Sionov, E., Lee, H., Chang, Y. C. and Kwon-Chung, K. J. (2010). Cryptococcus neoformans Overcomes Stress of Azole Drugs by Formation of Disomy in Specific Multiple Chromosomes. PLoS Pathogens *6*, e1000848.

Sipos, G., Prasanna, A. N., Walter, M. C., O'Connor, E., Bálint, B., Krizsán, K., Kiss, B., Hess, J., Varga, T., Slot, J., Riley, R., Bóka, B., Rigling, D., Barry, K., Lee, J., Mihaltcheva, S., Labutti, K., Lipzen, A., Waldron, R., Moloney, N. M., Sperisen, C., Kredics, L., Vágvölgyi, C., Patrignani, A., Fitzpatrick, D., Nagy, I., Doyle, S., Anderson, J. B., Grigoriev, I. V., Güldener, U., Münsterkötter, M. and Nagy, L. G. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nature Ecology and Evolution *1*, 1931–1941.

Slot, J. C. (2017). Fungal Gene Cluster Diversity and Evolution. Advances in Genetics *100*, 309–328.

Slot, J. C. and Hibbett, D. S. (2007). Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi : A phylogenetic study. PLoS ONE *2*, e1097.

Smith, D. R. and Keeling, P. J. (2015). Mitochondrial and plastid genome architecture : Reoccurring themes, but significant differences at the extremes. Proceedings of the National Academy of Sciences of the United States of America *112*, 10177–10184.

Sousa, M. J., Ardö, Y. and McSweeney, P. L. H. (2001). Advances in the study of proteolysis during cheese ripening. International Dairy Journal *11*, 327–345.

Spatafora, J. W., Aime, M. C., Grigoriev, I. V., Martin, F., Stajich, J. E. and Blackwell, M. (2017). The Fungal Tree of Life : from Molecular Systematics to Genome-Scale Phylogenies. Microbiology spectrum *5*.

Spatafora, J. W., Chang, Y., Benny, G. L., Lazarus, K., Smith, M. E., Berbee, M. L., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., James, T. Y., O'Donnell, K., Roberson, R. W., Taylor, T. N., Uehling, J., Vilgalys, R., White, M. M. and Stajich, J. E. (2016). A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia *108*, 1028–1046.

Stajich, J. E. (2017). Fungal Genomes and Insights into the Evolution of the Kingdom. The Fungal Kingdom *5*, 619–633.

Stajich, J. E., Berbee, M. L., Blackwell, M., Hibbett, D. S., James, T. Y., Spatafora, J. W. and Taylor, J. W. (2009). The fungi. Current biology : CB *19*, R840–5.

Steinberg, G. (2007). Hyphal growth : A tale of motors, lipids, and the spitzenkörper. Eukaryotic Cell *6*, 351–360.

Steinberg, G., Peñalva, M. A., Riquelme, M., Wösten, H. A. and Harris, S. D. (2017). Cell Biology of Hyphal Growth. Microbiology Spectrum *5*, 1–34.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research *37*, 7002–7013.

Stergiopoulos, I. and de Wit, P. J. (2009). Fungal Effector Proteins. Annual Review of Phytopathology *47*, 233–263.

Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

Taj-Aldeen, S. J., Almaslamani, M., Theelen, B. and Boekhout, T. (2017). Phylogenetic analysis reveals two genotypes of the emerging fungus Mucor indicus, an opportunistic human pathogen in immunocompromised patients. Emerging microbes & infections *6*, e63.

Tang, X., Chen, H., Gu, Z., Zhang, H., Chen, Y. Q., Song, Y. and Chen, W. (2017). Comparative Proteome Analysis between High Lipid-Producing Strain Mucor circinelloides WJ11 and Low Lipid-Producing Strain CBS 277.49. Journal of Agricultural and Food Chemistry *65*, 5074–5082.

Tang, X., Zan, X., Zhao, L., Chen, H., Chen, Y. Q., Chen, W., Song, Y. and Ratledge, C. (2016). Proteomics analysis of high lipid-producing strain Mucor circinelloides WJ11 : an explanation for the mechanism of lipid accumulation at the proteomic level. Microbial cell factories *15*, 35.

Tang, X., Zhao, L., Chen, H., Chen, Y. Q., Chen, W., Song, Y. and Ratledge, C. (2015). Complete genome sequence of a high lipid-producing strain of mucor circinelloides WJ11 and comparative genome analysis with a low lipid-producing strain CBS 277.49. PLoS ONE *10*, e0137543.

Tansey, M. R., Kamel, S. M. and Shamsai, R. (1984). The number of nuclei in sporangiospores of Rhizomucor species : taxonomic and biological significance. Mycologia *1*, 1089–1094.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome research *18*, 1979–1990.

Todd, R. T., Forche, A. and Selmecki, A. (2017). Ploidy Variation in Fungi : Polyploidy, Aneuploidy, and Genome Evolution. Microbiology spectrum *5*.

Toll-Riera, M., Rado-Trilla, N., Martys, F. and Alba, M. M. (2012). Role of low-complexity sequences in the formation of novel protein coding sequences. Molecular biology and evolution *29*, 883–886.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq

reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511–515.

Treseder, K. K. and Lennon, J. T. (2015). Fungal Traits That Drive Ecosystem Dynamics on Land. Microbiology and Molecular Biology Reviews *79*, 243–262.

Tutar, Y. (2012). Pseudogenes. Comparative and functional genomics *2012*, 424526.

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R. and Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Scientific Reports *8*, 10931.

Vallabhaneni, S., Mody, R. K., Walker, T. and Chiller, T. (2016). The Global Burden of Fungal Diseases. Infectious disease clinics of North America *30*, 1–11.

van de Veerdonk, F. L., Gresnigt, M. S., Romani, L., Netea, M. G. and Latge, J.-P. (2017). Aspergillus fumigatus morphology and dynamic host interactions. Nature reviews. Microbiology *15*, 661–674.

Voigt, K., Wolf, T., Ochsenreiter, K., Nagy, G., Kaerger, K., Shelest, E. and Papp, T. (2016). 15 Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes. In Biochemistry and Molecular Biology, (Hoffmeister, ed.), pp. 361–385. Springer Verlag, Berlin, Heidelberg, New York 3 edition.

Voigt, K. and Wostemeyer, J. (2001). Phylogeny and origin of 82 zygomycetes from all 54 genera of the Mucorales and Mortierellales based on combined analysis of actin and translation elongation factor EF-1alpha genes. Gene *270*, 113–120.

Vongsangnak, W., Kingkaw, A., Yang, J., Song, Y. and Laoteng, K. (2018). Dissecting metabolic behavior of lipid over-producing strain of Mucor circinelloides through genome-scale metabolic network and multi-level data integration. Gene *670*, 87–97.

Walther, G., Pawłowska, J., Alastruey-Izquierdo, A., Wrzosek, M., Rodriguez-Tudela, J. L., Dolatabadi, S., Chakrabarti, A. and de Hoog, G. S. (2013). DNA barcoding in Mucorales : An inventory of biodiversity. Persoonia : Molecular Phylogeny and Evolution of Fungi *30*, 11–47.

Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. Nature *449*, 54.

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V. and Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution *35*, 543–548.

Wei, W., McCusker, J. H., Hyman, R. W., Jones, T., Ning, Y., Cao, Z., Gu, Z., Bruno, D., Miranda, M., Nguyen, M., Wilhelmy, J., Komp, C., Tamse, R., Wang, X., Jia, P., Luedi, P., Oefner, P. J., David, L., Dietrich, F. S., Li, Y., Davis, R. W. and Steinmetz, L. M. (2007). Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proceedings of the National Academy of Sciences *104*, 12825–12830.

Wertheimer, N. B., Stone, N. and Berman, J. (2016). Ploidy dynamics and evolvability in fungi. Philosophical transactions of the Royal Society of London. Series B, Biological sciences *371*.

Wetzel, J., Burmester, A., Kolbe, M. and Wostemeyer, J. (2012). The mating-related loci sexM and sexP of the zygomycetous fungus Mucor mucedo and their transcriptional regulation by trisporoid pheromones. Microbiology (Reading, England) *158*, 1016–1023.

Whittaker, R. H. (1969). New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science (New York, N.Y.) *163*, 150–160.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics *8*, 973–982.

Wisecaver, J. H. and Rokas, A. (2015). Fungal metabolic gene clusters-caravans traveling across genomes and environments. Frontiers in Microbiology *6*, 161.

Wisecaver, J. H., Slot, J. C. and Rokas, A. (2014). The evolution of fungal metabolic pathways. PLoS genetics *10*, e1004816.

Wurzbacher, C., Rösel, S., Rychła, A. and Grossart, H. P. (2014). Importance of saprotrophic freshwater fungi for pollen degradation. PLoS ONE *9*, e94643.

Xue, W., Li, J.-T., Zhu, Y.-P., Hou, G.-Y., Kong, X.-F., Kuang, Y.-Y. and Sun, X.-W. (2013). L_RNA_scaffolder : scaffolding genomes with transcripts. BMC genomics *14*, 604.

Yadav, T., Quivy, J.-P. and Almouzni, G. (2018). Chromatin plasticity : A versatile landscape that underlies cell fate and identity. Science *361*, 1332–1336.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. and Bork, P. (2011). iPath2.0 : interactive pathway explorer. Nucleic Acids Res *39*, W412–5.

Zeng, L., Kortschak, R. D., Raison, J. M., Bertozzi, T. and Adelson, D. L. (2018). Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. PloS one *13*, e0193588.

Zerbino, D. R. and Birney, E. (2008). Velvet : algorithms for de novo short read assembly using de Bruijn graphs. Genome research *18*, 821–829.

Zhang, Y., Adams, I. P. and Ratledge, C. (2007). Malic enzyme : the controlling activity for lipid production ? Overexpression of malic enzyme in Mucor circinelloides leads to a 2.5-fold increase in lipid accumulation. Microbiology (Reading, England) *153*, 2013–2025.

Zhang, Y., Kastman, E. K., Guasto, J. S. and Wolfe, B. E. (2018). Fungal networks shape dynamics of bacterial dispersal and community assembly in cheese rind microbiomes. Nature Communications *9*, 336.

Zhang, Y., Navarro, E., Cánovas-Márquez, J. T., Almagro, L., Chen, H., Chen, Y. Q., Zhang, H., Torres-Martínez, S., Chen, W. and Garre, V. (2016). A new regulatory mechanism controlling carotenogenesis in the fungus Mucor circinelloides as a target to generate β-carotene overproducing strains by genetic engineering. Microbial Cell Factories *15*, 99.

Zheng, R. and Jiang, H. (1995). Rhizomucor endophyticus sp.nov., an endophytic zygomycetes from higher plants. Mycotaxon *56*, 455–466.

Zhou, P., Zhang, G., Chen, S., Jiang, Z., Tang, Y., Henrissat, B., Yan, Q., Yang, S., Chen, C.-F., Zhang, B. and Du, Z. (2014). Genome sequence and transcriptome analyses of the thermophilic zygomycete fungus Rhizomucor miehei. BMC genomics *15*, 294.

Zhu, Y. O., Sherlock, G. and Petrov, D. A. (2016). Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. G3 (Bethesda, Md.) *6*, 2421–2434.

Zhu, Y. O., Siegal, M. L., Hall, D. W. and Petrov, D. A. (2014). Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences of the United States of America *111*, E2310–8.

Zycha, H., Siepmann, R. and Linneman, G. (1969). Mucorales. 355pp., 155 fig. J. Cramer. Keys.(A revision of Zycha.

Annexe A

Posters et formations

Au cours de ma thèse j'ai participé à divers congrès, trois d'entre eux pour lesquels un poster a été réalisé pour l'occasion. D'autres part j'ai été impliqués dans deux processus de formations. Ces valorisations et formations sont présentés ici.

Poster

JOBIM (Journées Ouvertes en Biologie, Informatiques et Mathématiques),

28-30 juin 2016 (Lyon)

Etude comparative des génomes et transcriptomes de *Mucor* spp.

Annie LEBRETON, Laurence MESLET-CLADIÈRE, Jean-Luc JANY, Georges BARBIER and Erwan CORRE

Etude comparative des génomes et transcriptomes de *Mucor* spp.

Lebreton Annie[1], Meslet-Cladière Laurence[1], Jany Jean-Luc[1], Barbier Georges[1], Corre Erwan[2].

1 : Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (LUBEM), Université de Bretagne Occidentale (UBO), ESIAB - Parvis Blaise Pascal - Technopôle Brest-Iroise - 29280 Plouzané - France
2 : Station biologique de Roscoff, plateforme ABiMS, CNRS : FR2424, Université Pierre et Marie Curie (UPMC) - Paris VI, Place Georges Teissier - BP 74 29682 ROSCOFF CEDEX - France

Introduction

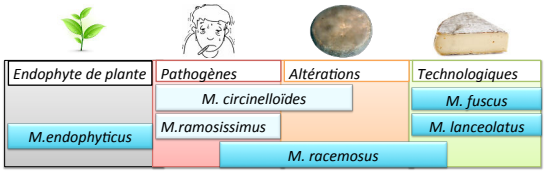
Un champignon « basal » aux niches écologiques et rôles variés

Le genre fongique *Mucor* appartient au phylum des Mucoromycota; l'un des quatre groupes issus de lignées ayant divergé très tôt dans l'évolution des espèces fongiques.

Le genre *Mucor* est un genre ubiquiste. Plusieurs espèces comme *M. indicus, M. ramosissimus* ou *M. circinelloides (MC)* sont des pathogènes animaux et humains alors qu'un petit nombre d'espèces sont retrouvées en industrie agroalimentaire et notamment fromagère : *M. fuscus (MF)* et *M. lanceolatus (ML)* sont des espèces technologiques d'affinage des fromages alors que d'autres comme *M. racemosus (MR)* peuvent être des agents d'altération.

Problématique:
Existe-t-il des spécificités au niveau des génomes de différentes espèces de *Mucor* liées aux modes de vie différents ? Trouve-t-on dans les génomes et leur expression des traces d'adaptation à une niche écologique, des traces de domestication chez les espèces technologiques ?

Espèces étudiées



X.xx Séquençage génomique et transcriptomique réalisés dans le cadre du projet.

X.xx Utilisation de données génomiques et transcriptomiques publiques.

Les données génomiques de *Phycomyces blackesleanus (PB)* et *Rhizopus oryzae (RO)*; espèces non fromagères du phylum Mucoromycota; ont également été utilisées.

Approches

Travail initial

Extraction — Les ADNc sont extraits après un même temps de culture. L'ADN est extrait lorsque la masse fongique est suffisante .

Séquençage — Les données RNAseq sont pairées, en ARN total et brin non spécifique. Pour l'ADN les séquençages sont réalisés en paired-end (PE) et mate pair (MP).

Assemblage — Les méthodes d'assemblage et annotation, génomique et transcriptomique, sont décrites ci-contre. Les statistiques associées aux assemblages et annotations sont comparées.

Annotation

Analyses transcriptomiques

Un « core transcriptome » est recherché ainsi que des fonctions géniques différentiellement représentées entre les espèces.

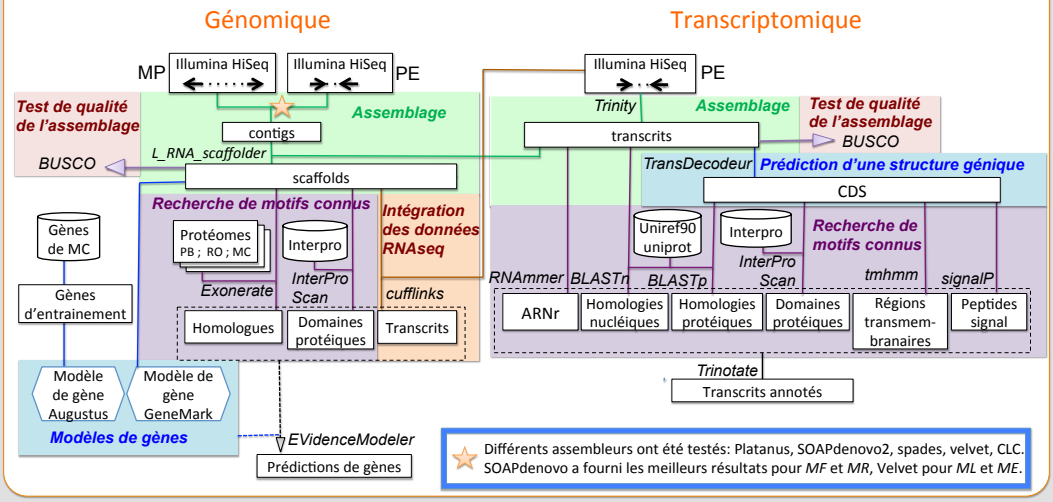
Annotation → OrthoFinder → OrthoGroupes
OrthoFinder → Analyses GO

Analyses génomiques

Les gènes impliqués dans la synthèse de métabolites secondaires sont organisés en clusters chez les champignons dits 'supérieurs'. Nous cherchons à vérifier si c'est le cas chez les *Mucor*.

Assemblage → antiSMASH → Métabolites secondaires
Annotation → SMURF → Clusters potentiels

Assemblage et annotation



Résultats préliminaires

Statistiques associées aux génomes	MF	ML	MR	ME
Taille du génome (Mb)	40,6	43,4	46,9	35,2
Nbr scaffolds > 1000pb	3819	1531	3506	159
N50 (kb)	27	141	24	1957
Taille max des scaffolds (kb)	142	681	161	539
% de gènes BUSCO trouvés	85	83	86	86
Nbr de transcrits cufflinks	15230	15514	12759	13108
Nbr de gènes EVidenceModeler	12310	11060	11269	11491

Statistiques associées aux transcriptomes	MF	ML	MR	ME
Nbr transcrits trinity	20898	21556	17368	22581
% de gènes BUSCO trouvés	86	82	84	55
Nbr de gènes trinity	14035	14299	14041	19865
% de gènes avec GO terms	53	53	59	55
Nbr transcrits avec régions transmembranaires	3128	3516	2904	2344
Nbr transcrits avec peptide signal	1035	1162	919	624

Quatre génomes et transcriptomes ont été assemblés et annotés. La recherche des 1438 gènes BUSCO fongiques montre que ces assemblages, aussi bien génomiques que transcriptomiques, sont de qualité suffisante pour identifier la majorité des gènes.

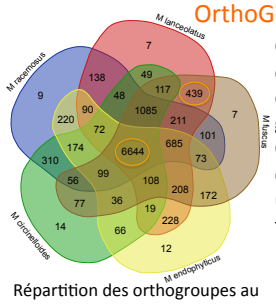
Résultats préliminaires

Métabolites secondaires

SMURF a permis de détecter entre un et trois clusters associés aux métabolites secondaires par espèce.
AntiSMASH a permis de détecter des gènes d'intérêt (NRPS par exemple) à partir desquels des clusters peuvent être recherchés.

Résultats d'antiSMASH					
Espèce	MF	ML	MR	ME	MC
Nbr de gènes trouvés	14	18	13	23	35

OrthoGroupes



Répartition des orthogroupes au sein des espèces

OrthoFinder a permis de mettre en évidence 6644 orthogroupes qui contiennent au moins un gène de chaque espèce.
On note également la présence de 439 orthogroupes constitués uniquement de gènes d'espèces technologiques.

Analyse des termes GO

La comparaison des annotations GO entre les espèces a mis en valeur de nombreuses différences. Par exemple, les annotations 'drug transporter' et 'organelle lumen' sont sous-représentées chez les espèces technologiques par rapport aux espèces pathogènes et endophyte. L'annotation 'protein binding' est quant à elle sur-représentée chez les espèces technologiques par rapport à toutes les autres espèces étudiées.

Conclusion & perspectives

Quatre génomes et transcriptomes de *Mucor* ont été assemblés et annotés. L'annotation des génomes reste cependant à améliorer avant la mise en place d'un consortium qui réalisera l'annotation experte de certaines familles de gènes et voies métaboliques d'intérêt.
D'autre part, cette étude a permis d'estimer la taille et la composition du «*core transcriptome*» et d'identifier un premier groupe de 439 gènes qui serait conservé chez les espèces technologiques fromagères et absentes des autres espèces étudiées. Il reste à déterminer les caractéristiques de ce groupe.
Une analyse des annotations GO a permis d'avoir une première vue d'ensemble de fonctions d'intérêt, il faut désormais réaliser des recherches plus fines pour appréhender l'implication de ces fonctions dans l'adaptation aux niches écologiques.
Des clusters de gènes potentiels ont été détectés, il reste à vérifier si les gènes de ces clusters sont effectivement co-exprimés/co-régulés.

Financements:

Poster

JOBIM (Journées Ouvertes en Biologie, Informatiques et Mathématiques),

3-6 juillet 2017 (Lille)

A transcriptional study of five fungal *Mucor* strains

Annie LEBRETON, Laurence MESLET-CLADIÈRE, Jean-Luc JANY, Georges BARBIER and Erwan CORRE

Transcriptional studies of five fungal *Mucor* species

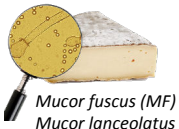
Lebreton Annie[1], Meslet-Cladière Laurence[1], Jany Jean-Luc[1], Barbier Georges[1], Corre Erwan[2]

1 : Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (LUBEM), Université de Bretagne Occidentale (UBO), ESIAB - Parvis Blaise Pascal - Technopôle Brest-Iroise - 29280 Plouzané - France.
2 : Station biologique de Roscoff, plateforme ABiMS, CNRS : FR2424, Université Pierre et Marie Curie (UPMC) - Paris VI, Place Georges Teissier - BP 74 29682 ROSCOFF CEDEX - France.

annie.lebreton@univ-brest.fr

INTRODUCTION

The fungal genus *Mucor*, belonging to the early diverging fungi, includes a large number of ubiquitous species which show diverse lifestyles



Mucor fuscus (MF)
Mucor lanceolatus (ML)
Used for cheese ripening

Mucor racemosus (MR)
Cheese spoiler

Mucor endophyticus (ME)
Plant endophyte

Mucor circinelloides (MC)
Mucor racemosus (MR)
Plant and animal pathogen

Aims:
Whether *Mucor* species different lifestyles are linked to different gene expression ?

STUDIED STRAINS

Five species were grown on PDA for 7 days and total RNA was extracted

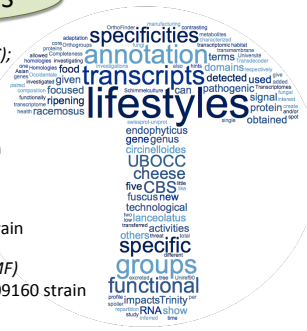
M. circinelloides (MC);
CBS 277-49 strain

M. racemosus (MR)
UBOCC-A-109155 strain

M. lanceolatus (ML)
UBOCC-A-109153 strain

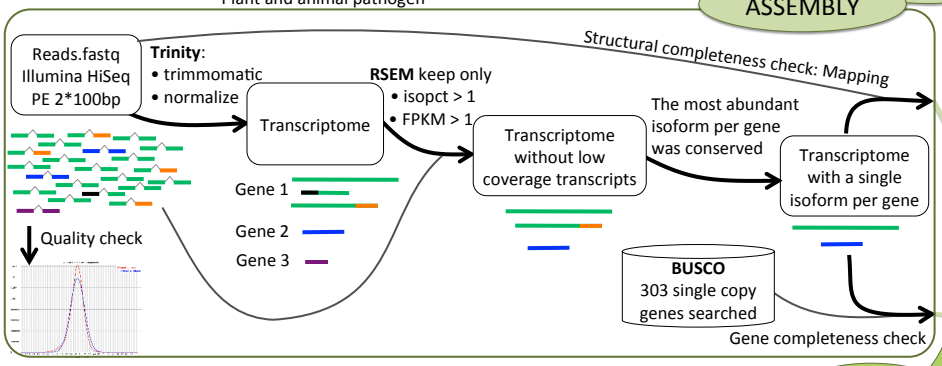
M. fuscus (MF)
UBOCC-A-109160 strain

Mucor endophyticus (ME)
CBS 385-95 strain



RNA was paired-end (PE) sequenced and transcriptomes were assembled

ASSEMBLY



ASSEMBLY RESULTS

Species	MR	MF	ML	ME	MC
Reads					
Million of Filtered Paired-end reads (2×100 bp)	30.2	25.5	24.8	34.9	24.6
Trinity *de novo* Assembly					
Nbr of trinity genes	14 035	14 299	14 041	15 554	13 655
Nbr of transcripts	17 368	20 898	21 556	16 950	19 891
genes with isoforms	16.5%	26.5%	31.1%	6.9%	27.6%
Used transcriptome - only the most abundant transcript per gene					
Transcripts N50 (bp)	1 761	1 831	1 938	1 292	1 964
Nbr of genes > 1 Kb	6 654	6 912	6 541	4 338	7 025
Realigned reads	93.1%	95.3%	70.9%	96.9%	93.6%
Busco eukaryotic lineage					
Completes genes found (duplicated and unique)	99.3%	97.4%	97.4%	67.0%	97.7%
Fragmented genes found	0.0%	0.3%	0.3%	20.1%	0.7%

Transcriptomes were functionnally annotated

ANNOTATION

Trinotate pipeline

RNAmmer	rRNA	*OrthoFinder*	OrthoGroups	Groups of orthologs and paralogs
TransDecoder	Predicted CDS	*PRIAM*	EC numbers	numerical classification scheme for enzymes
Blastx, Blastp (Uniref90, Swissprot-uniprot, Proteomes MC, Pb, Ro)	Homology	*Transfert**	GO terms	concepts/classes used to describe gene function
HmmScan (PFAM-A)	Protein domains	*Transfert**		
tmhmm	Transmembrane domain		Secreted proteins	
signalP	Signal peptides			

** Personal scripts*

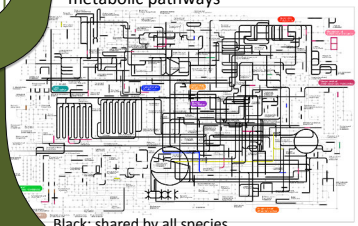
ANNOTATION RESULTS

Species	MR	MF	ML	ME	MC
Nbr of predicted proteins	11 728	11 157	10 674	10 646	11 953
Protein N50 (aa)	509	521	513	422	513
Transcripts with blast hits (E < e-5)	85.8%	76.7%	73.6%	86.2%	85.6%
Transcripts with GO terms transfert	72.3%	62.6%	61.8%	72.3%	69.0%
Transcripts with EC numbers	23.6%	19.8%	20.4%	24.9%	23.7%
Nbr of predicted secreted proteins	476	461	453	336	585

Annotations were compared among species

EC numbers

EC numbers were mapped on metabolic pathways

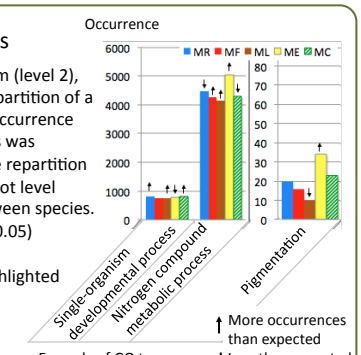


Black: shared by all species
Purple: shared by all species except ME
Yellow: only found for ME
Blue: only found for MR

GO terms

For each GO term (level 2), the observed repartition of a given GO term occurrence between species was compared to the repartition of GO term at root level occurrence between species. (Chi^2, p-value < 0.05)

15 GO terms highlighted



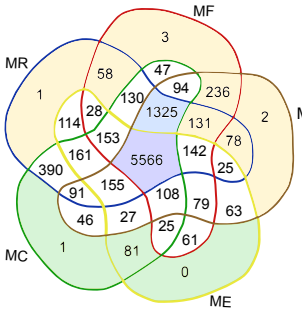
Example of GO terms highlighted in Biological process

And among species specifics orthogroups

Orthogroups

9421 orthogroups were detected

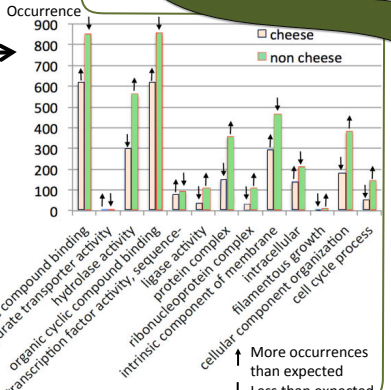
Orthogroups and singletons were clustered according to the lifestyle of the producing fungal species. Cluster were then compared (e.g. cheese/non-cheese, pathogen/non-pathogen).



Orthogroup distribution among species.

Orthogroups composed of proteins predicted on species:
- found in cheese
- rarely found on cheese
- found in all species
- found in all species except ME

GO terms comparison between cheese and non-cheese related groups.



5018 proteins could not be assigned to an orthogroup
- 37% had function annotation
- No function identified for proteins with a signal peptide
- No proteins involved in species specific metabolic pathway

Conclusions

Five *Mucor* transcriptomes were assembled and annotated.
Three functional criteria were examined: Proteins with **signal peptide, GO terms and EC numbers.**
Annotation were compared:

- **Species level**. As expected differences in transcripts content were found for GO terms such as "pigmentation" (ME and ML have darker spores than other species) and "developmental process" (ME grows slower) but differences were also detected for additional GO terms such as "nitrogen compound metabolic process".
- **Groups of orthogroups.** GO terms potentially linked to species lifestyle were found such as "heterocyclic compound binding" and "organic cyclic compound binding" which were over-represented on cheese related species and "cell cycle process" over-represented on non-cheese related groups.

No specific metabolic pathways was found and species/lifestyle specific proteins with peptide signal could not be functionally annotated.

Perspectives

Further comparatives studies will focus on **gene family expansion/contraction**. Another transcriptonnal **study on cheese/plant media** would allow to detect lifestyle specific transcripts unexpressed on standard media.

JOBIM Lille 3-6 juillet 2017 Journées Ouvertes en Biologie, Informatique et Mathématiques 2017

Thanks for funding:

Poster

ECFG14 (14th European Conference of Fungal Genetics),

25-28 février 2018 (Haifa, Israel)

Genomic comparison of five early diverging fungi encounter in different environments and belonging to the *Mucor* genus

Annie LEBRETON, Erwan CORRE, Jean-Luc JANY, Loraine GUEGUEN, Carlos PEREZ-ARQUES, Misharl MONSOOR, Victoriano GARRE, Emmanuel COTON, Georges BARBIER and Laurence MESLET-CLADIERE

Genomic comparison of five early diverging fungi encountered in different environments and belonging to the *Mucor* genus

Annie LEBRETON[1], Erwan CORRE[2], Jean-Luc JANY[1], Loraine GUEGUEN[2], Carlos PEREZ-ARQUES[3], Misharl MONSOOR[2], Victoriano GARRE[3], Emmanuel COTON[1], Georges BARBIER[1] and Laurence MESLET-CLADIÈRE[1]

1 Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (LUBEM) – Université de Bretagne Occidentale (UBO), 970 Avenue du Technopôle, 29280, Plouzané, France
2 Station biologique de Roscoff - plateforme ABiMS - CNRS : FR2424 - Sorbonne Université - Place Georges Teissier, BP 74 29682, Roscoff CEDEX, France
3 Department of Genetics and Microbiology, Faculty of Biology, University of Murcia, 30100 Murcia, Spain

Contact: annie.lebreton@univ-brest.fr

Introduction

The *Mucor genus* includes a large number of ubiquitous species with diverse lifestyles

Mucor fuscus (MF)
Mucor lanceolatus (ML)
Used for cheese ripening

Mucor racemosus (MR)
Cheese spoiler

Mucor endophyticus (ME)
Plant endophyte

Mucor circinelloides (MC)
Mucor racemosus (MR)
Plant and animal pathogens

Aim: Investigating genomic specificities potentially linked to *Mucor* species lifestyles

Genomic DNAs were extracted from 4 species and sequenced

Studied strains

UBO CC UBO Culture Collection

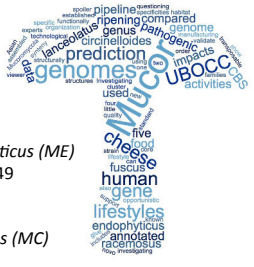
M. racemosus (MR)
UBOCC-A-109155

M. lanceolatus (ML)
UBOCC-A-109153

M. fuscus (MF)
UBOCC-A-109160

Mucor endophyticus (ME)
UBOCC-A-113049
CBS 385-95

M. circinelloides (MC)
CBS 277-49
Reference species



Genomes were assembled and annotated

Assembly & annotation protocols

1. Assembly

Selected tools

Velvet
SOAPdenovo

2. Annotation

- *ab initio* gene predictors: *(Genemark, Augustus)*
- Homology with public annotations
- RNAseq data support
- Repeated elements detection: *(REPET pipeline)*

- Genic structure prediction
- Known pattern search (protein domains, signal peptide prediction etc.)
- Transfert of GO terms and EC numbers
- Name assignation

3. Visualization

abims.sb-roscoff.fr

Apollo : genome browser

4. Expert curation

Manual check of genes involved in:
- RNAi pathway
- lipid metabolism
- DNA replication
- secondary metabolism
-...

Assembly & annotation results

Species	MR	MF	ML	ME	MC*
Millions of filtered PE reads	37	19	74	60	-
Millions of filtered MP reads	-	-	11	19	-
# of scaffolds (>1000pb)	3 506	3 819	1 531	159	21
Genome size (Mb)	46.9	40.6	43.4	35	36.6
GC%	32	36	34	33	42
# genes	11 269	12 310	11 060	11 491	11936
Gene size	1666	1562	1707	1714	1406
Exon size	334	301	300	313	309
Gene density, (genes / Mb scaffold)	240	303	254	328	326

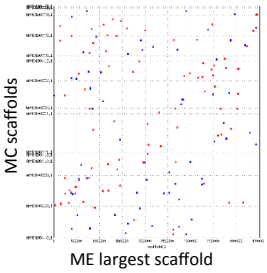
*MC genome assembly and annotation were extracted from publicly available databases [2]

The number of predicted genes was similar among species despite the observed differences among genome sizes.

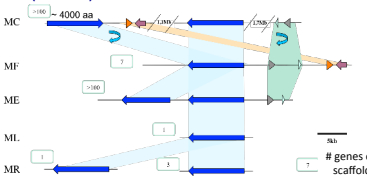
Genome structures were compared

Synteny

Synteny check at genome scale *(Mummer)*



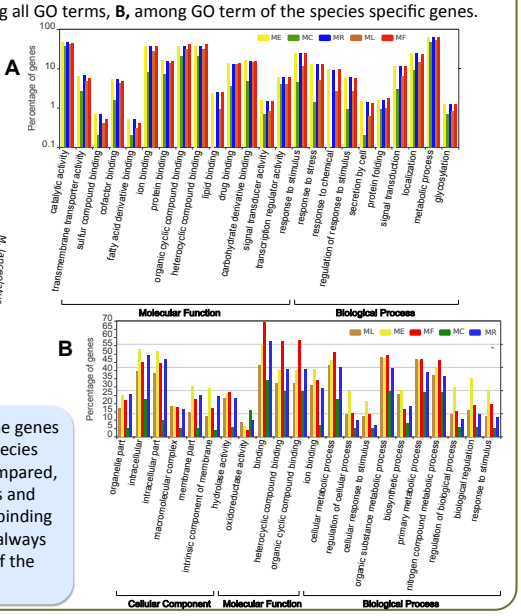
Local synteny check on genes potentially involved in secondary metabolism (blue arrow correspond to FAS/PKS) *(Mauve)*



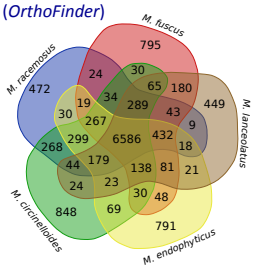
No synteny detected at genome and local scales.
No cluster organization was observed.

Gene annotations were compared among species

Sample of GO terms with significant differences among the five species: **A,** among all GO terms, **B,** among GO term of the species specific genes. *(WEGO)*



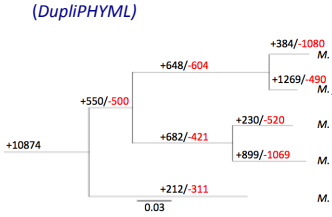
Genes were clustered in ortholog groups *(OrthoFinder)*



A core and pan genomes were delimited. Most of the genes had orthologs in all species and the number of species specific genes was similar among species. When compared, GO categories associated to response to stimulus and regulation were over-represented in ME, whereas binding was over-represented in MF. GO categories were always under-represented in MC , showing the impact of the difference in annotation methodology.

Evolution of gene families

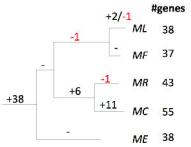
Gene gain and loss prediction *(DupliPHYML)*



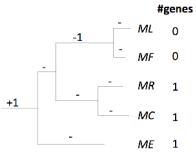
	Orphan	Total gain	Total loss	Overall gain
M. lanceolatus	384	1582	-2184	-602
M. fuscus	750	2467	-1594	+873
M. racemosus	420	1462	-1441	21
M. circinelloides	820	2131	-1990	141
M. endophyticus	723	212	-311	-99

Gene family evolution

Example: 15 orthogroups related to siderophores



Example: a family involved in drug transporter activity



Gains and losses of genes highly differed among species. In particular, MF and ML showed different patterns despite their phylogenetic proximity and their shared ecological niche. A focus on specific gene families provided hints to determine potential adaptation imprints to different lifestyles on *Mucor* genomes: e.g. expansion/contraction of genes involved in iron scavenging. This study is still under progress.

Conclusions

The five analyzed *Mucor* genomes differed structurally with no synteny observed among them. None of them included biosynthetic gene clusters such as it is currently observed for secondary metabolism in higher fungi. A core and a pan genomes were determined. Although analyses of GO term repartition has not yet unraveled genome specificities linked to different lifestyles, an ongoing gene family evolution study yielded some hints for revealing genomic specificities related to *Mucor* diverging lifestyles.

Citations:
[1] Voigt, K., T. Wolf, K. Ochsenreiter, G. Nagy, K. Kaerger, *et al.*, 2016 pp. 361–385 in 15 Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes.
[2] Corrochano, L. M., A. Kuo, M. Marcet-Houben, S. Polaino, A. Salamov, *et al.*, 2016 Expansion of signal transduction path- ways in fungi by extensive genome duplication.

ECFG14 HAIFA 2018
14th European Conference on Fungal Genetics

Thanks for funding:

Formation dispensée

Galaxy Initiation

Formatrices : Loraine GUEGUEN et **Annie LEBRETON**

15 mars 2017 (Roscoff)
23 avril 2018 (Roscoff)

Formation en Bioinformatique
Plateforme ABiMS
2018

Module
Galaxy Initiation

Objectifs

- Savoir analyser ses données sous l'environnement Galaxy.
- Être en mesure de créer ses workflows.

Programme

- Téléchargement des données à traiter.
- Manipulation de fichiers.
- Traitement des données.
- Visualisation des résultats.
- Création de workflows.
- Partage de résultats et de workflows.

Public
Personnel scientifique et technique

Pré-requis
Aucun

Modalités pédagogiques
Cours réalisé en salle TP informatique IGM
Théorie: 20% / Pratique: 80%
Un poste de travail par stagiaire

Durée: 1 journée

Intervenants
Loraine Guéguen / Gildas Le Corguillé / Annie Le Breton

Renseignements formation
mark.hoebeke@sb-roscoff.fr

Tél: 02 98 29 25 68

Pre-inscriptions en ligne

http://abims.sb-roscoff.fr/formation

Formation dispensée

École de Bioinformatique AVESIAN-IFB 2016/2017

Intervention en tant que tutrice pour deux chercheurs

Initiation au traitement des données de génomique obtenues par séquençage à haut débit

Du 20/11/2016 au 25/11/2017 (Roscoff)
Du 12/11/2017 au 17/11/2017 (Roscoff)

photo W. Thomas

5ème Ecole de bioinformatique AVIESAN – IFB 2016

Initiation au traitement des données de génomique obtenues par séquençage à haut débit

20-25 novembre 2016, Station Biologique, Roscoff

Objectifs

Les domaines des sciences du vivant liés à l'analyse du génome ont vu au cours des dernières années une accumulation explosive des données provenant des techniques de séquençage à haut débit. Les progrès accomplis ont considérablement augmenté les possibilités expérimentales dans des domaines tels que la génomique (séquençage de nouveaux génomes, variants génétiques), la transcriptomique (expression génétique, ARNs non codants) et les interactions ADN-protéine (immuno-précipitation de chromatine) et modifications de la chromatine. AVIESAN organise une quatrième école de bioinformatique, dont les objectifs sont d'apporter aux biologistes des notions et une pratique leur permettant d'appréhender le traitement et l'analyse des données de séquençage à haut débit.

Participants

Cette formation est destinée aux biologistes (ingénieurs, doctorants, chercheurs, enseignants-chercheurs, …) confrontés à l'analyse de données NGS, et qui ne disposent pas des compétences bioinformatiques suffisantes.

Contenu

La formation est une initiation à l'utilisation des outils bioinformatiques permettant d'aborder la diversité des applications du NGS. Cette école, qui se veut généraliste, sera organisée en deux groupes thématiques principaux : (1) régulation, transcriptome et épigénome et (2) variations génomiques. Elle couvrira une série de techniques dérivées du séquençage à haut débit : RNA-seq, ChIP-seq, identification et annotation de variants, assemblage de novo de RNA-seq. Le but de l'école est de couvrir plusieurs technologies largement utilisées, plutôt que de se concentrer sur une seule. L'école sera basée sur des ateliers pratiques sous l'environnement convivial Galaxy.

Les participants sélectionnés pourront bénéficier d'un tutorat personnalisé pour discuter de leur plan d'analyse, et effectuer les premières étapes de traitement de leurs propres données ou de celles de leur plateforme.

Attention : Cette formation n'a pas vocation à mener à bien l'analyse complète des données des participants, ni à les rendre indépendants sur l'intégralité d'une analyse bioinformatique.

Modalités d'inscription

Date limite de pré-inscription : 16 mai 2016 (Sélection des participants : mi-juin 2016)

Remplir en ligne la fiche de **pré-inscription** (**http://www.france-bioinformatique.fr/EBA2016/preinscriptions/**). Le nombre de places étant limité à 40, le comité d'organisation sélectionnera les participants d'après les renseignements portés sur cette fiche. Le degré de maturité du projet scientifique impliquant l'analyse de données de séquençage sera un des critères d'évaluation.

Renseignements : AVIESAN - ITMO Génétique, Génomique et Bioinformatique, **ecole-bioinfo@aviesan.fr**

Site Web (matériel de cours, informations complémentaires): **http://www.france-bioinformatique.fr/eba2016**

Frais d'inscription pour les personnels académiques : 500 € (coût déjà couvert pour les personnels rémunérés par l'Inserm) ; pour les industriels : 1750 €. L'hébergement et la restauration sont inclus.

Coordination scientifique : Christophe Caron (INRA, Rennes), Jacques van Helden (AMU, Marseille), Matthias Zytnicki (INRA, Toulouse).

Enseignants/Encadrants : 30 formateurs provenant des organismes et universités suivants: CNRS, INRA, Inserm, AgroParisTech, Institut Curie, Gustave Roussy, ENS, Aix-Marseille Université, UPMC, IRISA Rennes. Avec le soutien de l'Institut Français de Bioinformatique (IFB) et d'AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé).

Plateformes : IFB core (Gif-sur-Yvette), ABiMS (CNRS/UPMC, Roscoff), eBIO (Univ. Paris Sud), Genouest (CNRS/IRISA, Rennes), Genotoul (Toulouse), Institut Gustave Roussy (Villejuif), I2BC (Gif-sur-Yvette), Institut Curie - U900 (Paris), MIAT (INRA Toulouse), Sigenae (Toulouse), TGML/TAGC (Marseille), URGI (INRA Versailles).

Gestion : Christine Lemaitre (AVIESAN, ITMO GGB, Paris), Katy Main (AVIESAN, Paris).

aviesan
alliance nationale
pour les sciences de la vie et de la santé
ITMO Génétique, génomique
et bio-informatique

ifb

INRA
SCIENCE & IMPACT

Instituts thématiques
Inserm
Institut national
de la santé et de la recherche médicale

cnrs
dépasser les frontières

Photo W. Thomas

6ème Ecole de bioinformatique AVIESAN – IFB 2017

Traitement des données de génomique obtenues par séquençage à haut débit

12-17 novembre 2017, Station Biologique, Roscoff

Objectifs

La formation s'adresse à des chercheurs et ingénieurs directement impliqués dans des projets "Next Generation Sequencing" (NGS). Cette édition de l'école s'adresse aux nouveaux enjeux technologiques: elle inclura notamment une ouverture aux technologies lectures longues, qui transforment les approches en matière d'assemblage de génomes et l'identification de transcrits pleine longueur, ainsi que trois ateliers optionnels (lectures courtes : RNA-seq, ChIP-seq, variants), et une introduction à l'intégration des données.

L'école vise à introduire les concepts et à manipuler les outils informatiques qui permettront aux participants d'analyser ensuite leurs propres données de séquençage. Elle sera basée sur une alternance de courtes sessions théoriques et d'ateliers pratiques. Les participants bénéficieront d'un tutorat personnalisé pour discuter de leur plan d'analyse, et effectuer les premières étapes de traitement de leurs propres données ou de celles de leur plateforme. **Attention** : cette formation n'a pas pour vocation de réaliser l'analyse complète des données des participants.

Participants

Cette formation est destinée aux biologistes (ingénieurs, doctorants, chercheurs, enseignants-chercheurs,…) confrontés à l'analyse de données NGS, et qui ne disposent pas des compétences bioinformatiques suffisantes.

Environnement de travail

Nouveau : L'ensemble de la formation reposera sur l'utilisation de commandes en ligne (terminal Linux) pour les analyses bioinformatiques, et du langage R pour les analyses statistiques.

Prérequis

Aucune connaissance préalable de l'environnement Linux ou R n'est requise: la formation débutera par une introduction aux commandes en ligne qui sera progressivement approfondie au fil des sessions thématiques.

Modalités d'inscription

Date limite de pré-inscription : 19 mai 2017 (sélection des participants : mi-juin 2017). Le nombre de places étant limité, le comité d'organisation sélectionnera les participants d'après les renseignements portés sur cette fiche. Le degré de maturité du projet scientifique impliquant l'analyse de données de séquençage sera un des critères d'évaluation.

Renseignements : ecole-bioinfo@aviesan.fr
Informations et **inscriptions** : http://www.france-bioinformatique.fr/eba2017

Frais d'inscription pour les personnels académiques : 500€HT=600€TTC (coût déjà couvert pour les personnels rémunérés par l'Inserm) ; pour les industriels : 1.750€HT=2.100€TTC. L'hébergement et la restauration sont inclus.

Coordination scientifique : Christophe Caron (INRA), Jacques van Helden (AMU), Matthias Zytnicki (INRA).

Enseignants/Encadrants : une vingtaine de formateurs provenant des organismes et universités suivants: CNRS, INRA, Inserm, AgroParisTech, Institut Curie, Institut Pasteur, Gustave Roussy, ENS, Aix-Marseille Université. Avec le soutien de l'Institut Français de Bioinformatique (IFB) et d'AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé).

Plateformes : ABiMS (CNRS/UPMC, Roscoff), BIOGER (INRA Grignon), C3BI (Institut Pasteur, Paris), eBIO (Univ. Paris Sud), IFB core (Gif-sur-Yvette), IGBMC (Strasbourg), Institut Curie - U900 (Paris), Institut Gustave Roussy (Villejuif), I2BC (Gif-sur-Yvette), Genotoul (Toulouse), Genouest (CNRS/IRISA, Rennes), MIAT (INRA Toulouse), Sigenae (INRA Toulouse), TAGC (Marseille).

Gestion : Christine Lemaitre (AVIESAN, ITMO GGB, Paris).

UNIVERSITE ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

UBO
Université de Bretagne Occidentale

Titre : Caractéristiques génomiques du genre fongique *Mucor* et évolution adaptative liée à différents modes et conditions de vie au sein du genre

Mots clés : *Mucor*, Transcriptomique, Génomique comparative, Évolution adaptative

Résumé : Le genre *Mucor* appartient au phylum des *Mucoromycota*, un groupe issu de l'une des lignées ayant divergé très tôt dans l'évolution des espèces fongiques (*early diverging lineages*). Ces groupes restent encore très peu connus par rapport aux Ascomycètes et Basidiomycètes. Le genre *Mucor* est un genre d'espèces saprophytes, avec cependant une certaine diversité au niveau du mode de vie. Il existe en effet au sein du genre, des endophytes de plantes (comme *M. endophyticus*) ou encore des pathogènes opportunistes d'animaux (comme les espèces thermophiles *M. circinelloides* ou *M. indicus*). Le genre est ubiquiste mais il existe des associations à certains habitats qui semblent dénoter une certaine spécialisation. L'objectif de cette thèse était de mieux connaître les potentialités génétiques du genre *Mucor* lui permettant ce mode de vie ubiquiste, son potentiel d'adaptation mais également de mieux comprendre l'existence au sein du genre d'espèces semblant s'être spécialisées en colonisant préférentiellement ou exclusivement certains habitats comme le fromage. Afin d'atteindre cet objectif des études transcriptomiques et génomiques comparées ont été menées dans le cadre de cette thèse , afin de déterminer les principales caractéristiques des génomes de *Mucor* aussi bien structurelles que fonctionnelles, identifier les similitudes au niveau des espèces étudiées et aussi leur spécificités e en fonction des modes de vie/habitats et déterminer s'il existe chez les espèces fréquemment rencontrées dans les fromages (et notamment pour celles considérées comme technologiques) des traces d'adaptation voire de domestication.

Title : Genomic characteristics of the fungal genus *Mucor* and adaptive evolution linked to different modes and conditions of lifestyle within the genus

Keywords: *Mucor*, Transcriptomics, Comparative genomics, Adaptive evolution

Abstract: The genus *Mucor* belongs to the phylum *Mucoromycota*; a group that derived from the lineages that diverged early in the evolution of fungal species (early diverging lineages). These groups have been less well studied and are less well understood in comparison to Ascomycetes and Basidiomycetes. The genus *Mucor* is composed of saprophytic species, but also encompasses species with diverse lifestyles. For example, it includes plant endophytes (such as *M. endophyticus*) or opportunistic animal pathogens (such as the thermophilic species *M. circinelloides* or *M. indicus*). The genus is ubiquitous but there are some associations with specific habitats which seem to indicate specialisation. The aim of this thesis is to better understand the genetic potential of the genus Mucor in particular, to decipher how it maintains this ubiquitous lifestyle, its capacity to adapt to diverses habitats and to better understand the existence within the genus of species that may have undergone specialization allowing them to preferentially or exclusively colonise certain habitats, such as cheese. In order to achieve this, we have performed comparative transcriptomic and genomic studies in order to determine the main structural and functional characteristics of the Mucor genomes, identify similarities among the species studied and also assess whether there exist specific genetic associations with lifestyle/habitat and determine whether the species frequently found in cheese (in particular those species considered as technological) harbour imprints of adaptation or even domestication.