# Étude de l'impact mutationnel d'une perte de méthylation de l'ADN chez Arabidopsis thaliana

Victoire Baillet

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL
Préparée à l'Ecole Normale Supérieure

# Etude de l'impact mutationnel d'une perte de méthylation de l'ADN chez *Arabidopsis thaliana*


Soutenue par
**Victoire BAILLET**
Le 20 décembre 2018


Ecole doctorale n° 577
**Structure et dynamique des systèmes vivants**

Spécialité
**Sciences de la vie et de la santé**

Composition du jury :

| | |
|---|---|
| Pierre CAPY<br>PR, Université Paris Sud | *Président* |
| Laurent DURET<br>DR CNRS, Université Lyon 1 | *Rapporteur* |
| Malika AINOUCHE<br>PR, Université Rennes 1 | *Rapportrice* |
| Camille BERTHELOT<br>CR INSERM, Ecole Normale Supérieure | *Examinatrice* |
| Caroline HARTMANN<br>PR émérite, Université Paris Diderot | *Examinatrice* |
| Vincent COLOT<br>DR CNRS, Ecole Normale Supérieure | *Directeur de thèse* |

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

# Remerciements

Je tiens en premier lieu à remercier les membres de mon jury, Malika Ainouche et Laurent Duret qui ont accepté d'être rapporteurs de ma thèse, ainsi que Camille Berthelot, Pierre Capy et Caroline Hartmann qui se sont joints à eux pour évaluer mes travaux.

Merci à Vincent Colot, mon directeur de thèse, de m'avoir donné l'occasion de rejoindre l'équipe. La liberté que tu m'a laissée m'a donné l'opportunité d'apprendre énormément par moi-même et de prendre mon indépendance, et en cela je t'en remercie. Je me dois par ailleurs de témoigner toute ma gratitude à celles et ceux qui ont su répondre présent au cours des derniers mois. En particulier, merci à Barbara Desprès, Daniel Bouyer et Leandro Quadrana pour leur investissement durant la période d'écriture, et à Pierre Capy et Antoine Triller pour la confiance qu'ils m'ont accordée. J'espère que le résultat final sera à la hauteur des espérances de tous.

Merci aux membres de mon comité de thèse, Olivier Loudet, Clémentine Vitte et Florian Maumus de s'être réunis par deux fois pour évaluer mes travaux. Merci également à Frantz Depaulis de les avoir rejoints et de m'avoir fait bénéficier de ses connaissances et contacts en évolution moléculaire. Merci en particulier à Thomas Bataillon et Charlie Baer pour leurs conseils et suggestions.

Au cours de ma thèse, j'ai eu la chance d'être associée à plusieurs collaborations, pour lesquelles je remercie Vincent de m'avoir offert ces opportunités. Merci à Jin-Hoe Huh (Université de Séoul) de nous avoir fait bénéficier d'outils en cours de développement dans son laboratoire. Merci à Fredy Barneche et Anne-Sophie Fiorucci (IBENS) de m'avoir fait participer au projet Epishade, l'histoire est encore loin d'être achevée et je compte bien me remettre en quête du fameux locus causal du QTL(epi ?) du Chr3 dès que ce manuscrit sera rendu ! Merci également à Maria Manzanares-Dauleux, Mélanie Jubault et Benjamin Liégard (INRA Rennes) de m'avoir initiée à la hernie au travers de la thèse de ce dernier. J'ai appris de vous au moins autant que vous avez appris de moi.

Merci à l'ensemble des membres de l'équipe A2E que j'ai côtoyé ces dernières années. Merci en particulier à : Erwann Caillieux, force tranquille à l'humeur égale et expert *ès* Gateway, pour m'avoir formée lorsque nécessaire, guidée bon nombre de fois, et s'être investi dans l'axe CRISPR-dCas9 de mon projet quand je n'ai plus été en mesure d'y consacrer le temps nécessaire. Mathilde Etcheverry et Benoit Lahouze, prédécesseurs en epiRILeries avec qui j'ai partagé épiphalotypes, rires et frustrations. Amanda Silveira, soutien moral et scientifique des soirées et week-ends au labo. Daniel Bouyer, germain au grand coeur (proportionnalité oblige) et au nombre incalculable de projets tous réalisés à des heures indues au son de sonates sifflotées, ta créativité ne peux que forcer l'admiration. François Roudier, qui a érigé au rang d'art la mauvaise humeur de façade pour mieux cacher son sens de la joute verbable et ses bons conseils. Claudia Chica, pour sa patience indéfectible et ses conseils avisés, qu'il s'agisse de tests statistiques ou de choix de la langue la plus appropriée pour insulter un ordinateur et bien plus encore. Barbara

Desprès, pour nos pauses café-potin et le temps que tu as su me dégager, tout particulièrement ces dernières semaines. Leandro Quadrana, qui a plusieurs fois su me pousser à me faire confiance et à aller de l'avant. Amira Kramdi, partenaire fidèle de la team Readme, pour tous les coups de main ponctuels en bioinfo et plus encore (sans oublier les baklawas. burp). Merci également aux pièces rapportées à l'équipe A2E : Imen Mestiri d'un côté et Ricardo Rodriguez de la Vega de l'autre, pour les multiples discussions scientifiques que nous avons eu et continuerons d'avoir autour d'une sélection d'IPA ; ainsi qu'à leurs F1 respectives qui ont toujours réussi à m'arracher au moins un sourire.

Merci à Fredy Barneche (sans-accents-c'est-basque) et à son équipe qui nous ont rejoints aux lab meetings et dont j'ai colonisé le bureau aux pauses café-goûter, avec ou sans l'excuse des epiRILs. Merci pour les discussions scientifiques et moins scientifiques que nous avons pu avoir. Gérald Zabulon, ton grimoire de protocoles, ton congélateur-caverne d'ali baba et ton sourire qui tous ensemble permettent de venir à bout des clonages les plus récalcitrants (le rhum vieux est peut-être un facteur confondant, mais pour une fois les facteurs confondants on s'en fiche, hips). Anne-Sophie Fiorucci, pilier de la team Epishade, qui aurait deviné jusqu'où nous aurons menées les QTLepi ? Merci également à Gianluca Teano et à Clara Bourbousse pour les bons moments passés ensemble.

J'adresse également toute ma reconnaissance à Pierre Vincens et aux membres de la plateforme informatique de l'IBENS dont j'ai moult fois éprouvé la patience à grands coups d'I/O. Merci pour votre patience, votre professionnalisme et votre gentillesse. Au même étage, merci à Alexandra Louis, Yves Clément et d'autres encore, pour les multiples discussions que nous avons pu avoir et vos conseils avisés. Merci plus spécifiquement à ce dernier d'avoir pris à bras le corps le problème de mon incapacité pathologique au networking et joué à "Have you met Victoire ?" dans les tréfonds du Texas.

Je tiens également à remercier l'ensemble des membres de l'IBENS et notamment de la section "Ecologie et Biologie de l'Evolution" que j'ai pu croiser au cours de ces années pour les échanges scientifiques ou non que nous avons pu avoir. La force de l'IBENS est de réunir en un même bâtiment une multitude de scientifiques aux parcours, nationalités et sensibilités distinctes, un melting pot qui garantit des interactions hautement fructueuses sur tous les plans.

Merci à mes "eigen-mentors", qui ont su être moteurs dans mon orientation personnelle et professionnelle. Ceux qui m'ont donné ma chance, ceux qui m'ont mis le pied à l'étrier, ceux qui m'ont conduite à me dépasser, ceux qui m'ont inculqué l'exigence et la rigueur, ceux qui m'ont permis d'aiguiser ma curiosité scientifique et ceux qui m'ont poussée à la canaliser, ceux qui sont toujours disponibles pour un conseil. J'ai énormément appris de vous aussi bien professionnellement qu'humainement, et sous réserve que ma carrière scientifique dépasse le stade embryonnaire, j'espère transmettre à mon tour ce que vous m'avez transmis. Je vous dois énormément.

Merci également à Donald Knuth et Linus Torvald de m'avoir donné l'opportunité de perdre un temps considérable tout en pouvant dire que oui, je travaille sur ma thèse (et merci Phi-Phong pour le script de system restore). Merci aussi aux thrips des chambres de cultures de m'avoir contrainte de faire évoluer mon sujet de thèse et à plus long terme mes centres d'intérêts scientifiques, moins d'avoir grignoté une par une ou presque mes

transgéniques pleines de sucre toutes droit sorties d'in vitro après avoir annihilé mes populations de cartographie et mes phénotypages, et encore moins d'être allées cacher vos larves dans le terreau pour faire des réplicats techniques de la destruction de mon matériel biologique sur plusieurs générations successives...

Cette partie scientifique des remerciements ne peut s'achever sans mention du nerf de la guerre. Aussi, merci à l'IdEx Paris Sciences et Lettres puis au LabEx MemoLife d'avoir pris en charge mon salaire aux cours de ces quatre années ; ainsi qu'à la Société Française de Génétique d'avoir contribué à m'envoyer présenter mes travaux à SMBE 2017.

Pour la partie moins scientifique à présent :
Merci aux copains. Les anciens et les récents, les originels et les pièces rapportées, les amis et les potes, ceux qui m'ont demandé comment ma thèse avançait et ceux qui s'en sont savamment abstenus, ceux qui découvrent le mot "génome" et ceux qui ont des questions techniques. Parmi eux tous, je dédie cette thèse à Sophie, à qui la vie n'a pas laissé le temps d'achever la sienne. Le courage et la ténacité qui te caractérisent et avec lesquels tu t'es battue ont pour moi valeur d'exemple.

Merci à mes parents, ainsi qu'à ceux qui ne sont plus là pour voir ce travail s'achever. Ca y est, c'est (presque) fini, vous pouvez arrêter de retenir de votre souffle ! Merci de m'avoir laissée partir à plusieurs centaines de kilomètres (et parfois encore plus loin) tout en poursuivant les approvisionnements réguliers en kouign-amann, caramel au beurre salé et surgelés maison, et de ne pas m'avoir (trop) reproché de ne pas faire médecine et donc de ne prendre la suite ni de l'un ni de l'autre. La recherche étant ce qu'elle est, j'ai manifestement eu tort, mais de même que vous m'avez fait confiance toutes ces années vous pouvez être sûrs que je saurai faire ce qu'il faut pour continuer à faire ce que j'aime sans pour autant vivre à vos crochets (ou à ceux de Pôle Emploi) plus que nécessaire.

Pour finir, merci à Baptiste, arrivé dans la dernière ligne droite pour subir mon stress. Merci d'être là quand il le faut, de me rappeler qu'il faut manger et dormir, et accessoirement de t'abstenir de m'assommer avec l'un de tes cailloux quand je parle taux de mutation à minuit passé, peste contre les réformes de l'ESR, me lamente que je n'y arriverai jamais, ou les trois à la suite dans le même quart d'heure. Je présume que c'est surtout pour éviter d'endommager ledit caillou, mais dans le doute merci de t'abstenir encore quelques semaines, j'ai besoin de l'intégralité de mes neurones jusqu'à la soutenance :-) ! !

# Table des matières

# Liste des figures

# Liste des tables

# Liste des abréviations

5hmC 5-hydroxyméthylcytosine

5mC 5-méthylcytosine

ADN Acide Désoxyribonucléique

ARN Acide Ribonucléique

ARNt ARN de transfert

*CMT2 CHROMOMETHYLASE2*

*CMT3 CHROMOMETHYLASE3*

*CNR COLORLESS NON-RIPENING*

Cas9 *CRISPR-associated protein*

cM centiMorgan

CNV *Copy Number Variation*, variations du nombre de copies d'un locus

CO *Crossing-Over*

Col-0 Accession *Columbia* d'*Arabidopsis thaliana*

CRISPR *Clustered Regularly Interspersed Short Palindromic Repeats*

*DDM1 DECREASE IN DNA METHYLATION 1*

*DME DEMETER*

*DML DEMETER-LIKE*

*DNMT1 DNA methyltransferase 1*

*DNMT3a DNA methyltransferase 3a*

*DRM2 DOMAINS REARRANGED METHYLTRANSFERASE 2*

dCas9 *dead Cas9*, protéine Cas dont la fonction catalytique a été abolie

DMP *Differentially Methylated Position*, cytosine différentiellement méthylée

DMR *Differentially Methylated Region*, région différentiellement méthylée

DP *Dipyrimidine site*, site dipyrimidique

DSB *Double Strand Break*, cassure double-brin

*EN/SPM enhancer/supressor of mutator*

EMS Ethylméthylsulfonate

epiRIL *Epigenetic Recombinant Inbred Lines*, lignées recombinantes épigénétiquement différenciées

ET Élément Transposable

EWAS *Epigenome-Wide Association Study*

*FWA FLOWERING WAGENINGEN*

FoSTeS *Fork Stalling and Template Switching*, arrêt des fourches de réplication et changement de matrice

gBM *gene body methylation*, méthylation du corps des gènes

GWAS *Genome-Wide Association Study*

HIF *Heterogeneous Inbred Family*

HMM *Hidden Markov Model*, modèle de Markov caché

IGV *Integrative Genome Viewer*

indel insertion/délétion

*KYP KRYPTONITE*

kb kilobase

Ler Accession *Landsberg* d'*Arabidopsis thaliana*

LTR *Long Terminal Repeat*

*MET1 METHYLTRANSFERASE1*

MA *Mutation Accumulation*

MAGIC *Multi-parent Advanced Generation Inter-Cross*

Mb mégabase

MeDIP *Methylated DNA ImmunoPrecipitation*, immunoprécipitation de l'ADN méthylé

meQTL *Methylation quantitative trait loci*, variant ADN associé à des variations du niveau de méthylation

MMR *Mismatch Repair*

NAHR *Non-Allelic Homologuous Recombination*, Recombinaison Homologue Non-Allélique

Ne Taille efficace d'une population

NER *Nucleotide Excision Repair*

NHEJ *Non-Homologous End Joining*, ligature d'extrémités non homologues

NIL *Nearly-Isogenic Lines*

nt nucléotide

pb paire de bases

PCR *Polymerase Chain Reaction*, réaction de polymérisation en chaîne

PTGS *Post-Transcriptional Gene Silencing*

QTL *Quantitative Trait Locus*

*ROS1 REPRESSOR OF SILENCING 1*

RdDM *RNA-dependant DNA Methylation*

RIL *Recombinant Inbred Lines*

RIP *Repeat-Induced Point mutation*

RNAi *RNA interference*, ARN interférence

RViMR *Regional Variation in Mutation Rates*, variations locales du taux de mutation

SBS *single base substitution*

sgRNA *single guide RNA*

SNP *single nucleotide polymorphism*

SV *Structural Variation*, variation structurale

*TET* *Ten-Eleven Translocation*

TALE *Transcription Activator-Like Effector*

TCR *Transcription-Coupled DNA repair*

TGS *Transcriptional Gene Silencing*

TIR *Terminal Inverted Repeat*

TSD *Target Site Duplication*

UV Ultraviolets

*VIM1* *VARIATION IN METHYLATION1*

VIGS *Viral-Induced Gene Silencing*

WGBS *Whole Genome Bisulphite Sequencing*, séquençage génome-entier après traitement bisulphite

WT *Wild Type*, génotype sauvage

ZFP *Zinc-Finger Protein*, protéine à doigt de zinc

# Préambule

Un constat saisissant qui peut être fait lorsque l'on observe différents individus d'une même espèce est l'importante diversité phénotypique qui les distingue les uns des autres, et ce tout particulièrement dès lors qu'on s'intéresse aux traits dits complexes comme la taille ou le poids. Cette diversité phénotypique provient essentiellement de la variabilité génétique pré-existante au sein de cette population, et qui elle-même résulte des effets conjoints de l'apparition de nouveaux variants (mutation), de leur réassortiment lors de la recombinaison, et enfin des forces évolutives ou démographiques qui influenceront leur fréquence dans la population. Ce pool de variabilité, ou diversité génétique, constitue un paramètre clé de la capacité d'une espèce à répondre à des challenges environnementaux et donc à persister sur le long terme. Dans la mesure où sans source continuelle de nouveaux variants il n'y a aucune perspective d'adaptation, le processus de mutation peut être considéré comme la pierre angulaire du processus d'évolution, et il apparaît dès lors crucial de déterminer le taux et le patron de mutation (c'est-à-dire la fréquence à laquelle apparaissent les différents types de mutations) d'un organisme, ainsi que les différentes forces pouvant l'influencer.

Il est néanmoins une seconde source de variabilité entre individus, dont la compréhension a gagné en importance au cours de ces dernières années, et qui concerne non plus les polymorphismes le long du génome, mais le long de l'épigénome : il s'agira par exemple de variants dans les profils de méthylation de l'ADN. Contrairement aux variants ADN qui une fois créés sont transmis aux générations successives selon les lois de Mendel, ces variants épigénomiques, ou "épivariants", sont fréquemment effacés lors du passage à la génération suivante. Chez les plantes, où cette remise à zéro des profils épigénomiques est limitée, de tels épivariants peuvent cependant persister au travers des générations et dès lors participer à la fraction héritable (c'est-à-dire transmise à la descendance) du déterminisme des traits complexes.

Si ces observations ouvrent des perspectives importantes quant au potentiel évolutif de variations épigénomiques, il est crucial de garder à l'esprit que la diversité épigénomique ne peut pas être pensée indépendamment de la diversité génétique : non seulement les épivariants sont fréquemment associés à des variants ADN sous-jacents (typiquement des éléments transposables), mais les différents états épigénomiques, en influençant la formation de lésions dans l'ADN et/ou leur réparation, peuvent modeler le patron de mutation et donc à terme la diversité génétique. Dans cette interconnexion résident plusieurs écueils :

comment identifier les épivariants contribuant aux traits complexes, et comment isoler leur contribution ? Comment évaluer directement l'impact effectif de la variation épigénétique sur la création *in fine* de variation génétique ?

Durant ma thèse, je me suis intéressée à la méthylation de l'ADN, une modification épigénomique, comme source de variants génétiques et épigénétiques tous deux susceptibles de contribuer à la variabilité phénotypique héritable entre individus, en utilisant comme système modèle la plante *Arabidopsis thaliana*. Dans ce contexte, mes travaux se sont articulés autour de deux axes complémentaires : l'étude de l'effet d'une perte de méthylation de l'ADN sur le patron de mutation de l'arabette, qui constitue la majeure partie de ce manuscrit, mais également la mise en place d'approches expérimentales visant à identifier les épivariants causaux de QTL "épigénétiques" détectés dans une population de lignées épigénétiquement différenciées (epiRIL) établie au laboratoire.

# Organisation du manuscrit

Le manuscrit est organisé comme suit :
La première section (Introduction générale, chapitres 1 à 3) présente un panorama des différents concepts nécessaires à la bonne compréhension des travaux effectués. Elle intègre également une revue grand public dont je suis premier auteur.
La deuxième section (Résultats, chapitres 4 à 7), est divisée en quatre chapitres indépendants décrivant des travaux à des degrés divers d'avancement, tous précédés et suivis d'une mise en contexte et d'une discussion spécifique. Le premier chapitre de résultats (chapitre 4) décrit la caractérisation de la variation nucléotidique présente dans la population epiRIL. Les chapitres 5 et 6 tirent profit des données obtenues afin d'analyser le spectre mutationnel de la population epiRIL sous deux aspects distincts : impact d'une perte de méthylation de l'ADN sur le spectre des mutations ponctuelles (chapitre 5, rédigé sous la forme d'un article en préparation dont je suis premier auteur) et impact mutationnel de la remobilisation des éléments transposables (chapitre 6, auquel est joint un article dont je suis co-auteur). Enfin, le chapitre 7 présente et discute les approches mises en place pour effectuer la caractérisation de QTL "épigénétiques" identifiés au laboratoire.
Le manuscrit se conclut par une discussion des principaux résultats obtenus, de leur importance et des perspectives ouvertes par ce travail de thèse.

Première partie

# Introduction générale

Chapitre 1

# Aux origines de la variation génétique : la mutation

La création de variation génétique étant fortement dépendante du processus mutationnel, il est critique de comprendre ce dernier. Dans cette section, je décrirai les différents types de mutations et les approches employées pour estimer le taux auquel elles apparaissent. Je présenterai également les variations observées dans le taux et le patron de mutation entre les différentes espèces mais également le long du génome, ainsi que l'origine de ces variations.

## 1.1 Nature et cause des mutations

La mutation est un processus aléatoire qui conduit à la création de changements dans le matériel génétique d'un individu. Les variants ainsi créés sont ensuite transmis de façon stable au travers des divisions cellulaires, et dans le cas où l'évènement mutationnel se soit produit dans une cellule de la lignée germinale, il sera ensuite transmis à la génération suivante selon les lois de Mendel.
Le terme de mutation recouvre un large spectre d'évènements, qui vont des modifications ponctuelles de la séquence de l'ADN (substitutions nucléotidiques, courtes insertions et délétions) à des réarrangements chromosomiques (délétions, duplications, inversions, translocations), voire même des altérations dans l'organisation du génome (aneuploïdisation, polyploïdisation). A l'exception de ce dernier cas, les spécificités de chacun de ces types de mutation sont décrites dans les paragraphes suivants.

### 1.1.1 Mutations ponctuelles

#### Substitutions

Ce type de mutation consiste en la substitution d'un nucléotide par un autre, ce des suites d'une erreur soit de la machinerie de réplication de l'ADN entraînant l'insertion d'un nucléotide erroné, soit du système de réparation de l'ADN suite à une altération physique (radiation ionisante, rayons UV) ou chimique (exposition à des agents mutagènes) de la molécule d'ADN. Deux types de substitutions peuvent être distingués : les transitions et

les transversions, qui correspondent respectivement au remplacement d'une purine (adénine A ou guanine G) ou d'une pyrimidine (cytosine C ou thymine T) par une autre base du même type, et à la substitution entre purines et pyrimidines ; avec le constat quasi-universel que les transitions sont beaucoup plus fréquentes que les transversions. Il est à noter que ce biais mutationnel est purement chimique et résulte du phénomène de tautomérisation : chaque base de l'ADN est présente dans le noyau sous des formes alternatives, appelées tautomères, qui résultent d'une perte ou d'un gain d'une liaison hydrogène qui rend possible leur appariement avec une base incorrecte. A titre d'exemple, au lieu de former un dimère C:G, la forme imino de la cytosine peut s'apparier avec une adénine et ainsi conduire à une transition C → T.

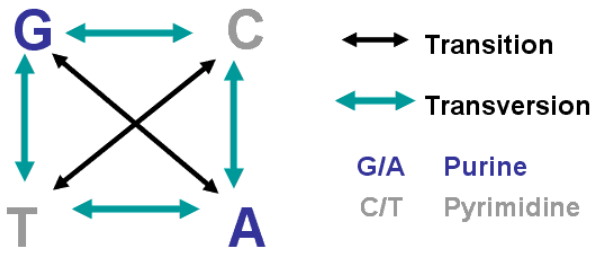


FIGURE 1.1 – Les 6 types de substitutions

Par ailleurs, l'étude des taux relatifs des différentes substitutions (aussi appelé spectre mutationnel ou patron de mutation) a permis de mettre en évidence une autre propriété quasi-universelle du taux de mutation ; à savoir un biais mutationnel vers AT. Cela se traduit notamment par le fait que parmi les deux types possibles de transitions, A:T → G:C (où ":" représente la liaison entre les deux brins) et C:G → T:A, ce dernier type est majoritaire aussi bien chez les procaryotes que chez les eucaryotes (Hershberg et al. 2010 ; Hildebrand et al. 2010 ; Lind et al. 2008 ; Lynch 2010b), à quelques exceptions près (Dillon et al. 2015).
En raison de ce biais, le contenu des génomes devrait tendre vers une réduction du pourcentage de sites G:C (%GC), or ce n'est pas ce qui est observé avec bon nombre d'organismes aux génomes plus riches en GC qu'attendu, notamment parmi les bactéries (Lynch 2010b). Cela signifie que ce biais mutationnel est contrebalancé par divers mécanismes, parmi lesquels on trouve la sélection sur les codons (avec un biais d'usage du code génétique favorisant les codons GC-riches) ou la conversion génique biaisée, qui va favoriser la réparation des mésappariements G:T en G:C plutôt qu'en T:A dans les régions fortement recombinantes (Duret et Galtier 2009 ; Meunier et al. 2004).

### Indels

Le terme d'indel (insertion-délétion) recouvre à la fois les insertions et délétions de nucléotides de taille inférieure à 50 pb. Il s'agit donc d'une catégorie de mutation fortement

hétérogène, ce qui pose une limite conceptuelle dès lors qu'il s'agit d'estimer un taux de mutation pour les indels. Aussi, la majorité de la littérature se concentre sur la formation d'indels au niveau des courtes répétitions en tandem (*short tandem repeats*), également appelées microsatellites. Le mécanisme le plus couramment proposé pour la formation d'indels est celui d'un dérapage de la polymérase (*polymerase slippage* ou *slipped strand mispairing*), illustré FIGURE 1.2. Durant la réplication, la polymérase et le brin en cours de synthèse vont se dissocier de façon transitoire du brin matrice, ce qui peut s'avérer particulièrement critique dans les régions répétitives du génome : la polymérase peut alors se réassocier à une position de même séquence nucléotidique, mais qui sera une ou plusieurs répétitions en amont ou en aval de sa position initiale. Ce mécanisme affecte notamment les régions présentant des répétitions en orientation directe (répétitions en tandem) de un à quatre nucléotides, et va résulter en l'insertion ou la délétion d'une à deux de ces répétitions. En fonction de la longueur de l'élément répété (en l'occurrence, si ce n'est pas un multiple de trois), l'apparition d'un indel dans une séquence codante peut provoquer une altération du cadre du lecture (*frameshift mutation*), avec des conséquences potentiellement plus délétères qu'un simple gain ou perte de codon.

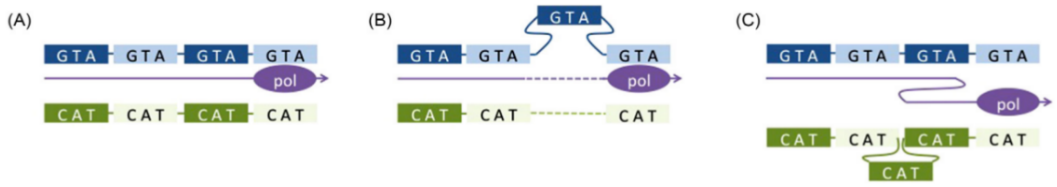


FIGURE 1.2 – Mécanisme de formation d'indels par dérapage de la polymérase
Les séquences en bleu et vert correspondent respectivement au brin matrice et à son complémentaire en cours de synthèse, la ligne violette représente le déplacement de la polymérase. (A) représente la situation normale, (B) et (C) les pertes et gain d'une répétition GTA en fonction du sens de dérapage de la polymérase. Figure extraite de (Sehn 2015).

Des approches expérimentales ont permis de mettre en évidence que de tels indels dérivaient d'un mésalignement initial, lequel peut être résolu soit avant synthèse d'ADN par l'activité proofreading de la polymérase, soit de façon postréplicative au moyen de la machinerie du *mismatch repair* (MMR). Cette voie de réparation joue un rôle prévalent dans la mesure où à la fois l'activité proofreading et la sélectivité (au profit de la processivité) de la polymérase vont décroître à mesure que la polymérase travaille sur un run long d'homopolymères (Garcia-Diaz et al. 2006). L'importance du MMR dans la résolution des indels est illustrée par le fait que que le taux d'indels dans les séquences répétées accroît de près de 10.000 fois lorsque la machinerie MMR est inactivée. Chez l'Homme, cela se traduit notamment par le phénotype dit de "*microsatellite instability*" observé dans les types de cancers présentant une inactivation mutationnelle de la voie MMR.

Récemment, un travail réalisé à l'échelle du génome entier chez l'Homme (Montgomery et al. 2013) a permis de mettre en évidence que près des trois quarts des indels résultaient de ce mécanisme. Pour expliquer le quart restant, et particulièrement les indels situés hors des régions répétitives, plusieurs mécanismes ont été être proposés, parmi lesquels le FoSTeS et le NHEJ qui sont également associés à la formation de variations structurales (voir Section 1.1.2). Par ailleurs, la remobilisation d'éléments transposables est également vecteur d'indels (voir Section 1.1.3).

### 1.1.2 Variations structurales

Le terme de variation structurale (*structural variation*, SV) recouvre toute altération de la séquence de l'ADN d'au moins 1 kb. Quatre types de SV "simples" peuvent être distingués : les délétions et duplications, aussi appelées variations du nombre de copies (*copy number variation*, CNV), les inversions et les translocations. Des réarrangements plus complexes peuvent être obtenus par une combinatoire de plusieurs SV (par exemple, une duplication avec inversion), un phénomène fréquent notamment dans les cellules cancéreuses (Stephens et al. 2011 ; Weckselblatt et al. 2015 ; Willis et al. 2015). Du fait de cette diversité, il est difficile d'estimer précisément un taux de SV : si la survenue de ce genre d'évènement apparaît relativement rare en comparaison du taux de substitutions, les SV peuvent néanmoins être considérés comme un type majeur de mutation eut égard au nombre de nucléotides affectés. Les conséquences des SV peuvent être de grande ampleur : en premier lieu des effets de dose si le SV n'est pas équilibré (c'est à dire, si le nombre de copies a été altéré, à l'exemple des CNV ou des translocations non-réciproques), mais également des effets positionnels (le SV affectant le niveau d'expression des gènes à proximité ou inversement), la formation de gènes chimériques ou des isolements reproductifs du fait d'anomalies d'appariement à la méiose dans le cas des inversions ou des translocations.

#### Mécanismes provoquant la formation de SV

La formation de SV est principalement attribuée à des erreurs lors de la réparation de cassures double-brin dans l'ADN, mais également à des erreurs durant la réplication (Carvalho et al. 2016). Sur la base de l'organisation génomique de la région dans laquelle s'est produit un SV et de la signature laissée au point de cassure (*breakpoint*), il est possible d'inférer quel mécanisme moléculaire en a été à l'origine (Weckselblatt et al. 2015). Les trois mécanismes principaux, décrits ci-dessous, sont illustrés Figure 1.3.

- **Recombinaison homologue non-allélique (NAHR)**
  Les évènements de recombinaison homologue non-allélique (*non allelic homologuous recombination*, NAHR) se produisent au cours de la recombinaison (mito-

tique ou méiotique) et résultent de l'alignement puis de la formation d'un crossing-over entre deux séquences homologues mais non-alléliques, donc des séquences suffisamment identiques pour pouvoir s'hybrider mais situées à des endroits différents du génome. Pour avoir lieu, la NAHR nécessite au moins 300 pb d'homologie stricte entre les deux brins d'ADN mis en jeu, et son efficacité va dépendre de la longueur, de l'identité et de la proximité physique des séquences homologues impliquées (Lupski 2004 ; Stankiewicz et al. 2002). Aussi, ce mécanisme va concerner tout particulièrement certains types d'annotations génomiques, notamment les éléments transposables, les duplications segmentaires et les *low copy repeats*. En fonction de l'orientation des séquences matrices et de la façon dont le crossing-over sera résolu, le NAHR peut générer des duplications, délétions ou réarrangements (décrits FIGURE 1.6) ; mais la majorité des évènements résultent d'un crossing-over entre des répétitions situées sur un même chromosome (NAHR intrachromosomal) ou sur des chromosomes homologues - à l'opposé, il est rare d'observer de tels évènements entre des répétitions situées sur des chromosomes non homologues. Dans le cas où des SV apparaissent être "récurrents", c'est-à-dire que des individus non apparentés vont présenter un SV affectant une même région du génome, et que le site de cassure est localisé au sein d'une région d'homologie, l'origine NAHR va être proposée. Réciproquement, la présence de longues régions d'homologie à proximité d'un intervalle génomique donné est un facteur de risque d'instabilité de cette région (Carvalho et al. 2016).

- **Ligature d'extrémités non-homologues (NHEJ)**
  Les réarrangements associés au NHEJ (*non homologuous end joining*) correspondent à l'opposé à des réarrangements non-récurrents (Carvalho et al. 2016). Le NHEJ est une voie prédominante en phases G1 et M du cycle cellulaire et qui effectue la réparation des cassures double-brin en "recollant" les bouts libres. Ce mécanisme laisse une signature caractéristique au site de cassure, sous la forme d'une jonction nette (*blunt end*) entre les partenaires de réparation, avec présence d'un indel les séparant, la formation de ce dernier résultant de l'édition des brins libres (ajout ou clivage de quelques nucléotides) effectuée préalablement à la religation. Ce mécanisme a longtemps été proposé dès lors qu'aucune région de longue homologie (donc matrice pour une recombinaison) ne pouvait être identifiée au site de cassure.

- **Arrêt des fourches de réplication et changement de matrice (FoSTeS)**
  Contrairement aux deux mécanismes décrits ci-dessus, le FoSTeS fait partie des mécanismes réplicatifs associés à la formation de SV, une catégorie dont l'ampleur effective a gagné en reconnaissance au cours de ces dernières années et qui est elle aussi associée à la formation de réarrangements non-récurrents (Carvalho et al.

2016). Au cours de la réplication, il peut arriver que la fourche de réplication s'arrête (*fork stalling*). Le brin en cours de synthèse va alors pouvoir envahir une fourche tierce, s'hybrider au niveau d'une région de microhomologie (2-5 pb) et poursuivre sa synthèse à partir de cette nouvelle matrice (*template switching*). Dans le cas où la synthèse effectuée à ce nouveau site ne se poursuit que sur quelques nucléotides avant que le brin ne retrouve son site de synthèse initial, le FoSTeS peut créer un indel, qui sera alors caractéristique de part sa signature d'une courte microhomologie. Cette même signature au *breakpoint* d'un SV, là encore en l'absence de régions d'homologie, est spécifique des SV résultat d'un FoSTeS. En fonction de l'origine de l'arrêt de la fourche (collision entre fourches, collision réplication-transcription, réparation d'une cassure, dérapage au niveau de séquences répétitives) et compte-tenu de la proximité physique dans des foci de réplication de régions du génome distinctes mais répliquées selon le même programme temporel (Figure 1.11), le FoSTeS et les autres mécanismes réplicatifs peuvent être associés à la formation de réarrangements complexes, notamment des réarrangements juxtaposant des fragments issus de plusieurs chromosomes.

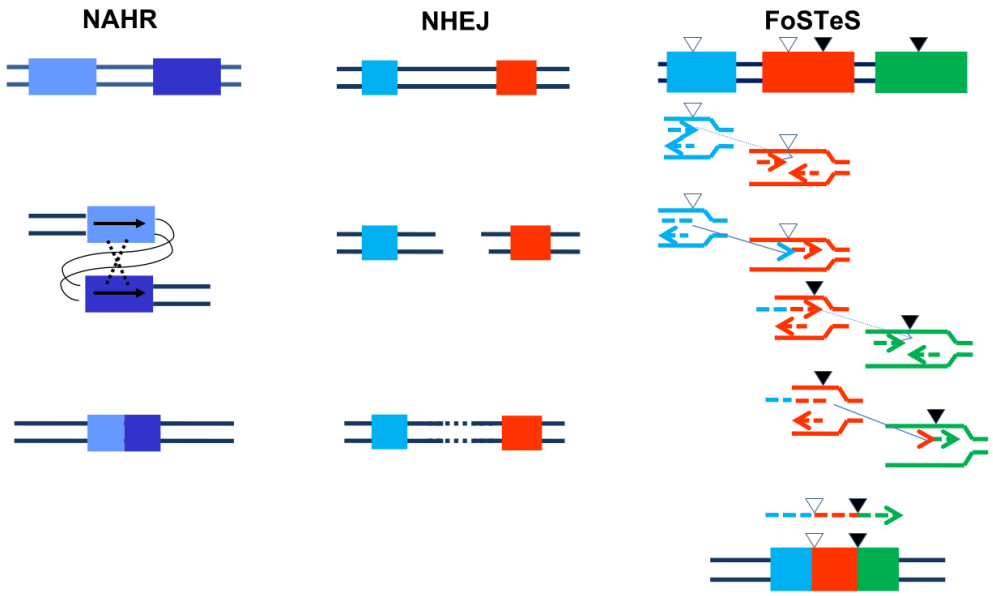


FIGURE 1.3 – Comparaison des trois mécanismes principaux associés à la formation de SV. Les lignes noires représentent les deux brins d'ADN, les couleurs des rectangles indiquent leur niveau d'homologie (nuances de bleu : régions de forte homologie, couleurs distinctes, absence d'homologie).
Panel de gauche : NAHR intra-chromatide. L'évènement de recombinaison va provoquer la délétion ou -réciproquement- la duplication d'une partie des séquences homologues ainsi que de la région comprise entre les deux.
Panel du milieu : NHEJ. Suite à la formation de cassures double-brin entre deux séquences non homologues, la machinerie NHEJ va permettre la religation entre les deux séquences, conduisant à la délétion de la région comprise entre les deux DSB ainsi que la création d'un indel au site de cassure.
Panel de droite : Formation d'une délétion complexe résultant d'un FosTeS associant deux fragments. Les fragments ne présentent pas d'homologie (couleurs distinctes des rectangles), néanmoins des régions de microhomologie (2-5pb) sont présentes entre les différentes régions (triangles blancs et noirs). Les fourches de réplication sont figurées de la même couleur que le fragment correspondant. Dans l'exemple, ce phénomène se produit une deuxième fois, causant la délétion des deux régions flanquées par chaque paire de microhomologies. Extrait de (Gu et al. 2008).

### 1.1.3 Contribution des éléments transposables aux paysages mutationnels

Les éléments transposables (ET) sont des séquences ayant la capacité de se déplacer au sein d'un génome. Sur la base des mécanismes employés pour transposer et de leur structure, deux classes d'ET peuvent être distinguées (FIGURE 1.4) : les ET de classe I, ou rétrotransposons, qui utilisent un intermédiaire ARN lors de leur mobilisation selon un mécanisme de type "copier/coller" notamment gouverné par une transcriptase inverse, et les ET de classe II, ou transposons à ADN, qui se mobilisent sous forme d'ADN selon un mécanisme de type "couper/coller" gouverné par une transposase.

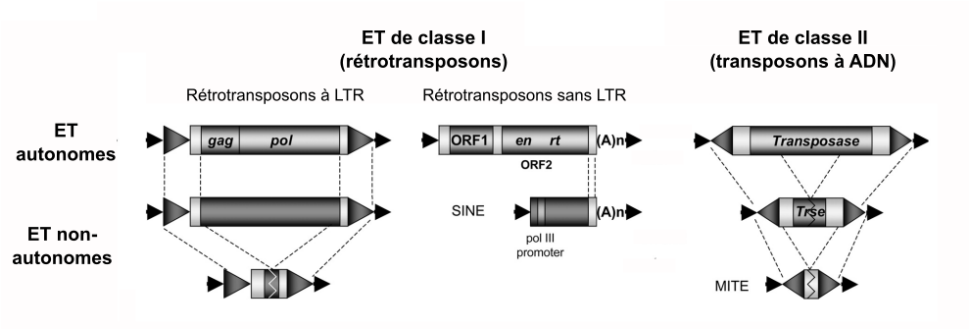


FIGURE 1.4 – Organisation et classification des types majoritaires d'ET.
Les ET sont séparés en deux classes, en fonction de leur intermédiaire de transposition (ARN pour les classe I, ADN pour les classe II). Les éléments de classe I sont en plus divisées en deux groupes sur la base de leur structure et de leur mécanisme de transposition : les rétrotransposons à LTR (b) et ceux dépourvus de LTR (c). Chaque classe comporte des éléments autonomes ainsi que non-autonomes. Les éléments autonomes codent les protéines requises pour la transposition (*gag*, *capsid-like protein* ; *pol*, *reverse transcriptase* ; ORF1, *gag-like protein* ; *en*, *endonuclease* ; *rt*, *reverse transcriptase*), tandis que les éléments non-autonomes en sont dépourvus mais comportent cependant les séquences nécessaires leur permettant d'être mobilisés. Les triangles noirs flanquant chaque ET correspondent aux TSD formés lors de la transposition, les triangles gris correspondent aux répétitions terminales inversées (*terminal inverted repeats*, TIR) et aux longues répétitions terminales (*long terminal repeats*, LTR) aux extrémités des ET de classe II et des rétrotransposons à LTR respectivement. Modifié d'après (Wessler 2006).

Les ET peuvent générer un spectre très large de mutations, allant de changements subtils dans la régulation de l'expression des gènes à d'importants réarrangements chromosomiques, comme illustré FIGURES 1.5 et 1.6. Deux cas de figure peuvent être distingués : les effets directs de la transposition (donc qui résultent de l'insertion ou de l'excision d'un ET), et les effets indirects (donc dûs à la présence d'ET en un grand nombre de copies dispersées dans le génome).

**Effets directs**

Indépendamment de leur mécanisme de transposition, en s'insérant dans une séquence co-

dante ou des régions régulatrices, les ET ont la capacité d'inactiver un gène ou d'en altérer son patron d'expression ; mais également conduire à la création de gènes chimériques. Un exemple en est *SETMAR*, gène chimérique résultant de la fusion entre le domaine transposase d'un ET et un gène codant une histone méthyltransférase (Cordaux et al. 2006).
Les éléments de classe II, qui se mobilisent par excision, contribuent additionnellement au paysage mutationnel au travers de ce mécanisme. En effet, lorsqu'un ET (de classe I ou II) s'insère à un locus donné, les mécanismes de réparation mis en jeu au site d'insertion vont conduire à la formation d'une courte duplication (*target site duplication*, TSD) au point exact de cassure. Lors d'une excision, cette signature restera présente, ainsi qu'un ou plusieurs nucléotides additionnels issus ou de la séquence du transposon ou des sites alentours, en fonction des caractéristiques de la famille de transposon. Cette "*excision footprint*", peut conduire à l'altération du cadre de lecture d'un gène, avec de potentielles conséquences phénotypiques. Cette situation est par exemple bien illustrée par le transposon *Ascot* chez le champignon *Ascobolus immersus* (Colot et al. 1998). L'intégration d'*Ascot* dans le gène *b2* qui détermine la couleur de la spore conduit à des spores incolores, et les diverses *footprints* générées par son excision ont pu être associées à des réversions partielles à totales du phénotype (Figure 1.5, panel du bas, droite). Par ailleurs, les mécanismes de réparation impliqués pour résoudre la cassure double brin provoquée par une excision peuvent également être responsables d'évènements de recombinaison intrachromosomale ou ectopique (Athma et al. 1991 ; Shalev et al. 1997).
Des évènements de transposition complexe, alors qualifiée transposition alternative, peuvent également avoir lieu. Lors d'une transposition "conventionelle", la transposase va reconnaître les répétitions terminales inversées (TIR) à chaque extrémité de l'élément. Mais si la transposase reconnaît deux TIR issues de deux copies différentes qui sont inversées ou en orientation directe l'une par rapport à l'autre, la réaction de transposition elle-même peut conduire à des inversions ou délétions, des formations de gènes chimériques ou encore des cassures de chromosomes (D. Wang et al. 2015 ; J. Zhang et Peterson 2004 ; J. Zhang, F. Zhang et al. 2006).

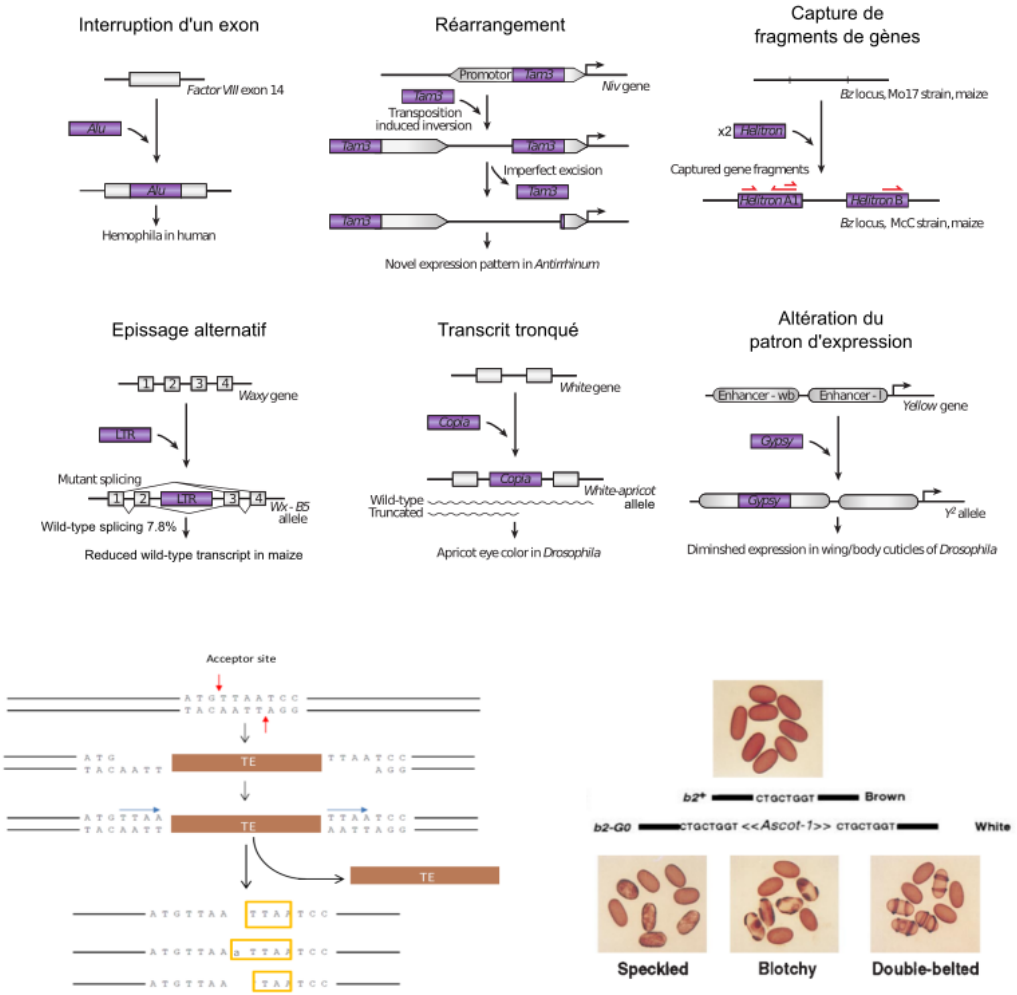


FIGURE 1.5 – Conséquences mutationnelles directes des ET. Exemples non-exhaustifs d'altérations génomiques provoquées par des insertions d'ET. Extrait de (C. Huang et al. 2012).
Dans le cas des ET de classe II, leur excision est associée à la formation d'une *footprint* qui va correspondre ou au TSD, ou au TSD +/- quelques nucléotides (panel du bas, gauche). Illustration des réversions phénotypiques associées à différentes *footprints* formées au cours de l'excision d'*Ascot* du locus *b2*. La présence de l'ET est à l'origine d'une couleur marron des spores. Extrait de (Colot et al. 1998).

### Effets indirects

Du fait de leur présence en un grand nombre d'exemplaires dans les génomes, les ET sont de parfaites matrices pour former des réarrangements : l'abondance de copies de séquences similaires dans un génome peut conduire à des réarrangements chromosomiques importants. Les évènements ainsi induits peuvent conduire à des inversions, des délétions et également des duplications (Gray 2000). Le produit de ces réarrangements va dépendre de la localisation et l'orientation des copies-matrices (FIGURE 1.6) : si les copies sont en sens direct et localisées sur deux chromosomes différents, la recombinaison conduit à une translocation. Si les deux copies sont en orientation inverse, la recombinaison mène à la

formation d'un chromosome dicentrique et d'un chromosome acentrique. Si les deux copies sont localisées sur le même chromosome, la recombinaison entre deux répétitions directes conduit à une duplication et/ou une délétion. Quand les deux copies sont inversées, la recombinaison conduit à l'inversion de la région comprise en ces deux répétitions. De tels réarrangements peuvent donc avoir un impact majeur sur l'intégrité du génome.

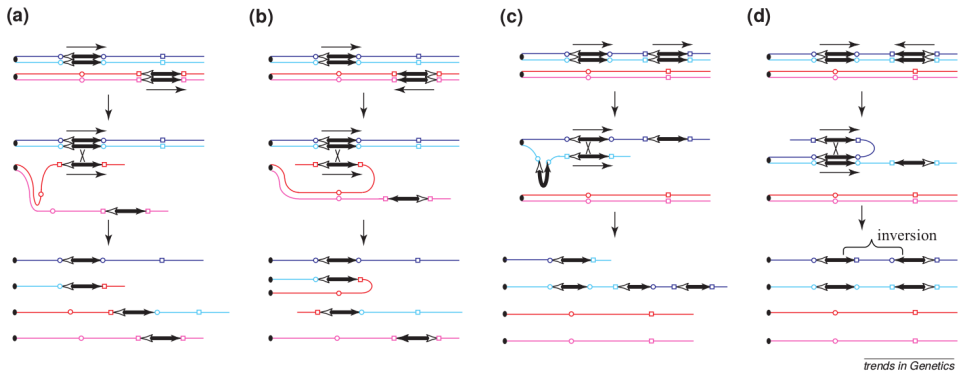


FIGURE 1.6 – Conséquences mutationnelles indirectes des ET : formation de délétions, duplications et inversions au travers d'évènements de recombinaison ectopique. Pour chaque chromosome homologue (nuances de bleu ou nuances de rouge), les deux chromatides sœurs sont figurées, chacune représentée par une ligne et reliées l'une à l'autre par le centromère (rond noir). Les ET sont représentés en noir et en gras, l'orientation de l'élément est figurée par les flèches à chaque extrémité. La recombinaison homologue est figurée par une croix. Dans le cas où les ET sont situés chacun sur un homologue, la recombinaison va conduire à la délétion ou réciproquement à la duplication de la région située entre les deux copies si les ET sont dans la même orientation (a), ou à la formation d'un chromosome dicentriquue et réciproquement d'un fragment acentriquue si les ET sont en orientation inversée (b). Dans le cas où les ET sont situés sur le même chromosome, si les copies sont en orientation directe, la recombinaison va conduire à la délétion d'une part et à la duplication d'autre part de la région située entre les deux copies, avec dans ce dernier cas une augmentation du nombre de copies (c). Si les copies sont en orientation inversée, la recombinaison va être à l'origine d'une inversion (d). Extrait de (Gray 2000).

## 1.2 Comment estimer un taux de mutation ?

Estimer un taux de mutation, c'est-à-dire le nombre d'évènements de mutation, par génome, par génération, est un paramètre crucial en génétique des populations. Quel que soit le type d'altération, les mutations n'en demeurent pas moins des évènements rares et aléatoires, et il est de fait complexe d'obtenir une évaluation directe et non-biaisée du taux et du patron de mutation. Dans cette section, je présente les différentes méthodes couramment employées ainsi que leurs limites.

### 1.2.1 Approches basées sur des systèmes rapporteurs

Historiquement, notre connaissance empirique des taux de mutations spontanées chez les micro-organismes et quelques espèces animales dérive d'estimations basées sur des systèmes rapporteurs (Baer et al. 2007 ; F. A. Kondrashov et al. 2010). Dans ce type de dispositif, l'apparition de mutations est révélée par une altération du phénotype, et il est possible de dériver un taux de mutation par site par division cellulaire en rapportant la fréquence d'apparition de mutants à la longueur de l'espace mutationnel (*mutational target size* ; ici, la longueur en pb du transgène) (Drake et al. 1998). Des stratégies fondées sur le même principe ont été ensuite développées chez le Nématode, sur la base non pas d'un transgène mais d'un locus endogène ; mais également chez l'Homme, en tirant profit de la fréquence d'apparition des maladies génétiques à déterminisme monogénique (A. S. Kondrashov 2003).

Si ces approches élégantes permettent d'évaluer expérimentalement (et donc directement) le taux de mutation, elles présentent également des limites importantes. En effet, dans la mesure où elles se basent sur la mise en évidence d'un phénotype attendu, des mutations sans effet visible (mutation silencieuse) ou conduisant à un phénotype subtil ne seront pas révélées, et donc pas prises en compte dans l'estimation. Par ailleurs, dans la mesure où l'estimation qui est obtenue est spécifique à un gène, il peut être incorrect d'extrapoler ce résultat au génome entier en raison de variations locales du taux de mutation le long du génome (voir Section 1.3.2). De plus, si cette approche nous indique la présence ou l'absence d'une mutation, elle ne nous renseigne nullement quand au type de mutation, donc ne documente en rien le patron de mutation de l'organisme étudié.

### 1.2.2 Approches génome-entier indirectes

Jusqu'à peu, l'étude du patron de mutation d'un organisme se faisait à partir du patron de divergence entre espèces apparentées, soit donc le nombre de différences qui se sont accumulées dans la séquence d'ADN depuis leur séparation (Kumar et al. 2002 ; Nachman et al. 2000). Cette approche se base sur le fait qu'aux sites dits neutres (donc

non soumis à pression de sélection), le degré de divergence entre les deux espèces va être égal au taux de mutation (Kimura 1968). Dès lors, le taux de substitution à des sites neutres peut être converti en un taux de mutation par milliers ou millions d'années. Si cette approche garantit d'avoir un large nombre de mutations et ne requiert aucun séquençage additionnel, elle présente toutefois d'importantes limites (Ségurel et al. 2014). La qualité de l'estimation obtenue va en effet être fortement dépendante de variables additionnelles : l'estimation qui est faite le temps de génération de l'organisme considéré (puisqu'il va s'agir de convertir un taux de mutation par millions d'années en un taux de mutation par génération), celle de la date de séparation des deux espèces (par exemple des données fossiles), mais également la façon dont les sites neutres ont été déterminés comme tels, dans la mesure où il s'agit bien souvent d'un a priori génomique (pseudogènes, sites dégénérés quatre fois) plus qu'une mesure effective de l'absence de signature de sélection à ces sites.

Une stratégie dérivée consiste à utiliser non plus la divergence entre deux espèces, mais les polymorphismes entre individus de la même espèce (Messer 2009 ; Ségurel et al. 2014). Les théories de génétique des populations prédisent en effet que le niveau de polymorphisme dans une population donnée va être fonction de la taille de cette population et du taux de mutation, et il est donc possible d'inférer cette dernière valeur connaissant les deux précédentes. En pratique, résoudre cette équation requiert d'avoir à disposition d'autres estimateurs qui permettent d'expliciter le lien entre niveau de polymorphisme et taux de mutation, mais ceux-ci sont bien souvent fortement dépendants de la neutralité effective des sites échantillonnés. Par ailleurs, de telles approches sont fortement sensibles à la démographie. Ces contraintes peuvent être atténuées en prenant en compte exclusivement les polymorphismes présents à une faible fréquence allélique, lesquels reflètent plus fidèlement le patron de mutation. Cela requiert néanmoins l'utilisation de cohortes de grande taille (plusieurs dizaines de milliers d'individus), ce que les projets de séquençage à large échelle rendent à présent possible, avec en sus dans le cas des organismes modèles la possibilité d'évaluer le taux de mutation dans des conditions autres que celles de laboratoire. En ce sens, cette approche a récemment été appliquée notamment chez la Levure (Y. Zhu et al. 2017) et la Drosophile (Assaf et al. 2017).

Aussi, si de telles approches permettent d'avoir une image sur l'ensemble du génome du taux et du patron de mutation d'un organisme, elle se contentent de l'inférer à partir d'une image partiellement brouillée par la sélection et ne permettent pas d'en avoir une vision directe.

### 1.2.3 Dispositifs d'accumulation de mutations

L'avènement des progrès du séquençage a permis de remettre au goût du jour une stratégie ancienne, initiée par Muller (Muller 1928), celle des lignées dites d'accumulation de mutations (MA lines, FIGURE 1.7). La visée initiale de cette approche est alors d'inférer le taux de mutation affectant des traits quantitatifs et la distribution des effets sur la fitness (P. Keightley 1994 ; Mukai 1964 ; Shaw et al. 2002). Comparativement, ce n'est que récemment que l'on s'est employé à analyser les génomes des MA lines afin d'estimer directement le taux de mutation à l'échelle du génome entier.

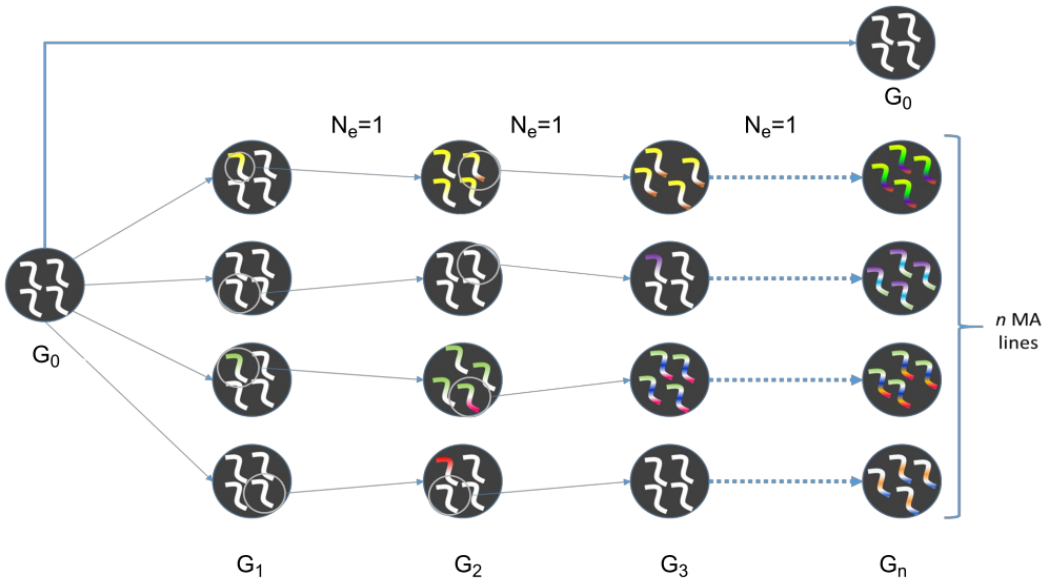


FIGURE 1.7 – Schéma illustrant l'établissement d'une population de lignées MA. G indique le numéro de la génération, Ne le nombre d'individus employé pour propager la population d'une génération à la suivante. Explications dans le texte.

A partir d'un individu "fondateur" rendu aussi homogène génétiquement que possible par des cycles d'autofécondation (ou de croisements frères-sœurs), on dérive des lignées sur un grand nombre de générations, en ne sélectionnant à chaque fois qu'un seul individu pour donner la génération suivante. Cette modalité de propagation par goulots d'étranglements successifs, appelée filiation monograine chez les plantes, garantit que la sélection sera drastiquement réduite au profit de la dérive génétique. Aussi, à l'exception des mutations létales ou quasi-létales, le séquençage des lignées "évoluées" nous donne accès à la totalité des mutations accumulées au travers des générations, permettant dès lors d'estimer le taux ainsi que le patron de mutation. De plus, en comparaison des approches présentées plus haut, la puissance des dispositifs MA lines réside en ce qu'ils permettent également d'évaluer les interactions entre les différentes mutations, leur relation

de dominance-récessivité ainsi que la dépendance environnementale des effets (Halligan et al. 2009).

Néanmoins, compte-tenu du faible nombre de mutations par génération, il est nécessaire d'avoir à disposition un grand nombre de lignées ainsi qu'un grand nombre de générations afin de dériver une estimation robuste du taux de mutation, ce qui peut s'avérer être un écueil chez les organismes à temps de génération long ou lorsqu'on cherche à évaluer le taux de mutation pour des évènements plus rares comme les réarrangements chromosomiques.

Une approche dérivée, favorisée par l'accessibilité toujours meilleure des technologies de séquençage, est celle du séquençage de trios parents-enfants, soit un équivalent aux MA lines pour des organismes non-modèles comme l'Homme. Si cette stratégie est intuitive et permet d'associer de façon non-ambiguë les néo-mutations à l'un ou l'autre des haplotypes parentaux, sa mise en pratique se heurte à plusieurs écueils (Ségurel et al. 2014). En effet, le fond génétique des individus étant fortement hétérogène et les néo-mutations étant attendues à l'état hétérozygote, une telle approche va nécessiter d'avoir une bonne profondeur de séquençage. Par ailleurs, les mutations somatiques et germinales sont alors confondues et il s'avère donc nécessaire de séquencer une génération additionnelle (comme illustré dans (Yang et al. 2015)), un contrôle encore trop rarement effectué en raison de son coût.

## 1.3 Variations interspécifique et intragénomique du taux de mutation

Les technologies de séquençage actuelles nous permettent d'avoir accès au taux de mutation ainsi qu'au spectre mutationnel pour un grand nombre d'espèces, mais également pour différentes souches d'une même espèce ou pour différents types cellulaires. Ce large nombre d'évènements à disposition permet également d'étudier plus finement comment le taux de mutation est distribué le long du génome et quelles forces (épi)génomiques peuvent l'influencer.

### 1.3.1 Variation du taux de mutation entre organismes

Dans la mesure où les indels et variations structurales restent encore complexes à identifier dans les produits de séquençage traditionnels, seul le taux de mutation affectant un seul site (*single-base substitutions*) a pu être estimé de façon robuste chez un grand nombre d'espèces procaryotes comme eucaryotes. Une liste des taux de mutations disponibles à ce jour pour une sélection d'organismes est présentée Table 1.1. Il peut

être observé que le taux de mutation (par site, par génération) varie de près d'un facteur 1000 au travers de l'arbre du vivant, allant de $10^{-11}$ substitutions/site/génération chez les Archées ou la Paramécie à $10^{-8}$ chez l'Homme. Ce dernier taux se traduit par près de 100 SBS distinguant le génome d'un nouveau-né de celui de ses parents (contre un seul SBS par génération chez Arabidopsis par exemple), donc un nombre conséquent de mutations introduites dans les populations humaines à chaque génération, ce qui pose en creux la question de leur fardeau génétique (*genetic burden*). Néanmoins, il s'agit de souligner qu'une fois mis en regard des autres forces évolutives à l'œuvre et notamment de la taille effective des populations (voir plus bas), le taux de mutation dans la lignée germinale des populations humaines ne se révèle plus être excessivement élevé, contre-carrant le qualificatif de "*human exceptionalism*" qui lui a longtemps été accolé (Lynch 2016).

Afin d'expliquer ces disparités inter-espèces, plusieurs hypothèses ont pu être avancées, lesquelles ont trait à la biochimie et à la physiologie des organismes (temps de génération, taux métabolique, réparation de l'ADN) ou se fondent sur des concepts de génétique des populations (barrière de dérive) (Baer et al. 2007 ; Lynch 2011) :

- L'hypothèse du "temps de génération" se fonde sur le principe que la réplication est une source primaire de mutations. A taux de mutation par division cellulaire identique, un organisme qui atteint la maturité reproductive plus tôt transmettra moins de mutations à sa descendance qu'un organisme dont la lignée germinale a effectué des mitoses additionnelles. (Bromham et al. 1996 ; Mooers et al. 1994 ; Ohta 1993) Une illustration de cette hypothèse est le "*male mutation bias*" (Ellegren 2007) : en raison du caractère continu de la spermatogenèse tout au long de la vie d'un individu (environ 20 divisions par an, contre 30 divisions au total pour l'oogenèse), 80% des 100 mutations ponctuelles portées par un nouveau-né seront portées par l'allèle paternel, et similairement l'effet de l'âge du père sur l'apparition de maladies génétiques est un paramètre bien connu (Kong et al. 2012 ; Venn et al. 2014).

- L'hypothèse du "taux métabolique" trouve son origine non pas dans la réplication comme source d'erreurs, mais dans la présence d'agents mutagènes endogènes comme les radicaux libres (*reactive oxygen species*).Selon cette hypothèse, les organismes présentant des taux métaboliques (les proxys en étant la respiration ou la masse corporelle) plus élevés vont produire plus de radicaux libres, d'où une augmentation des dommages à l'ADN provoqué par ce stress oxydatif et donc un taux de mutation plus élevé (Martin et Palumbi 1993 ; Nunn et al. 1998). Cette théorie est notamment soutenue par le fait que les vertébrés à sang chaud présentent des taux métaboliques ainsi que des taux d'évolution plus élevés que ceux à sang froid, et pourrait également expliquer l'origine du taux de mutation bien plus élevé dans

les mitochondries, siège de la respiration, que dans le génome nucléaire.

- L'hypothèse de la "réparation de l'ADN" propose que plus le système de réparation des dommages à l'ADN d'un organisme est efficace, plus son taux de mutation sera faible, dans la mesure où un plus grand nombre de mutations sera corrigé avant le passage à la génération suivante. Ce phénomène est particulièrement bien illustré par la comparaison des taux de mutation entre procaryotes et virus : ces derniers, qui ne disposent pas de polymérases avec une activité proofreading, présentent un taux de mutation très élevé, de l'ordre de $10^{-6}$ à $10^{-8}$ mutations par site par génération pour les virus à ADN et jusqu'à $10^{-3}$ à $10^{-5}$ pour les virus à ARN (Sanjuán et Domingo-Calap 2016 ; Sanjuán, Nebot et al. 2010).

- L'hypothèse de la "barrière de dérive" (Lynch 2011 ; Sung, Ackerman, Miller et al. 2012) se fonde sur un paramètre de génétique des populations, la taille efficace des populations (Ne), laquelle va déterminer le "rapport de force" entre dérive génétique et sélection naturelle quant au devenir des variants ADN apparus par mutation dans une population. Ainsi, dans les populations à Ne élevé, la sélection prend le pas sur la dérive, tandis que cette dernière sera prédominante dans les populations où Ne est faible. Cette théorie confronte le coût de l'investissement dans des mécanismes de réparation de l'ADN plus fidèles (donc le coût de la sélection qui y est associée) avec l'efficacité avec laquelle ces innovations pourront être sélectionnées (donc la force de la dérive dans la population). Dans ce contexte, les organismes présentant un Ne élevé auront un taux de mutation moindre que ceux à Ne faible, ce puisque dans ces dernières populations l'investissement dans des innovations moléculaires qui permettraient de réduire le taux de mutation est contrebalancé par une dérive forte.
  La force de ce cadre théorique est qu'il permet d'unifier les patrons observés au travers de l'arbre du vivant (bactéries, eucaryotes uni- et multi-cellulaires, voir Figure 1.8) en une relation linéaire simple valable pour tous les organismes analysés à ce jour : le taux de mutation dans la fraction du génome sous sélection (approximée par la fraction codant pour des protéines) est inversement proportionnelle à la taille efficace de la population (Lynch, Ackerman et al. 2016 ; Sung, Ackerman, Miller et al. 2012).

TABLE 1.1 – Taux de mutation (substitutions) par site par génération pour une sélection d'organismes, obtenus par approches d'accumulations de mutation. Les étoiles indiquent des estimations basées sur des systèmes rapporteurs.

| | Espèce | Taux de substitution (par site, par génération) | Taille du génome (Mb) | Référence |
|---|---|---|---|---|
| Eucaryotes multicellulaires | *Homo sapiens* (8) | $1{,}35.10^{-8}$ | 3300,00 | (Campbell et al. 2012) |
| | *Pan troglodytes* (11) | $1{,}20.10^{-8}$ | 3524,00 | (Venn et al. 2014) |
| | *Aotus nancymaae* | $8{,}10.10^{-9}$ | 2861,68 | (Thomas et al. 2018) |
| | *Oryza sativa* (10) | $7{,}10.10^{-9}$ | 389,00 | (Yang et al. 2015) |
| | *Arabidopsis thaliana* (2) | $6{,}95.10^{-9}$ | 119,70 | (Ossowski et al. 2010) |
| | *Apis mellifera* (1) | $6{,}80.10^{-9}$ | 262,00 | (Yang et al. 2015) |
| | *Daphnia pulex* (5) | $5{,}69.10^{-9}$ | 250,00 | (Keith et al. 2016) |
| | *Mus musculus* (9) | $5{,}40.10^{-9}$ | 2717,00 | (Uchimura et al. 2015) |
| | *Drosophila melanogaster* (6) | $5{,}17.10^{-9}$ | 168,70 | (Schrider et al. 2013) |
| | *Heliconius melpomene* (7) | $2{,}90.10^{-9}$ | 273,79 | (P. D. Keightley et al. 2015) |
| | *Caenorhabditis elegans* (4) | $1{,}45.10^{-9}$ | 100,30 | (Denver et al. 2012) |
| | *Caenorhabditis briggsae* (3) | $1{,}33.10^{-9}$ | 104,00 | (Denver et al. 2012) |
| Eucaryotes unicellulaires | *Neurospora crassa** (14) | $4{,}10.10^{-9}$ | 38,64 | (Lynch 2010a) |
| | *Plasmodium falciparum** (16) | $2{,}08.10^{-9}$ | 22,85 | (Lynch 2010a) |
| | *Trypanosoma brucei** (19) | $1{,}38.10^{-9}$ | 26,08 | (Lynch 2010a) |
| | *Chlamydomonas reinhardtii* (13) | $3{,}80.10^{-10}$ | 111,10 | (Ness et al. 2015) |
| | *Saccharomyces cerevisiae* (17) | $2{,}63.10^{-10}$ | 12,46 | (Y. O. Zhu et al. 2014) |
| | *Schizosaccharomyces pombe* (18) | $2{,}17.10^{-10}$ | 19,63 | (Behringer et al. 2016 ; Farlow et al. 2015) |
| | *Paramecium tetraurelia* (15) | $1{,}94.10^{-11}$ | 72,09 | (Sung, Tucker et al. 2012) |
| Procaryotes | *Mesoplasma florum* (26) | $9{,}78.10^{-9}$ | 0,79 | (Sung, Ackerman, Miller et al. 2012) |
| | *Helicobacter pylori** (25) | $1{,}90.10^{-9}$ | 1,66 | (Lynch 2010a) |
| | *Staphylococcus epidermidis* (32) | $7{,}40.10^{-10}$ | 2,56 | (Sung, Ackerman, Dillon et al. 2016) |
| | *Mycobacterium smegmatis* (27) | $5{,}27.10^{-10}$ | 6,99 | (Kucukyildirim et al. 2016) |
| | *Deinococcus radiodurans* (23) | $4{,}99.10^{-10}$ | 3,28 | (Long et al. 2015) |
| | *Agrobacterium tumefaciens* (20) | $2{,}92.10^{-10}$ | 5,67 | (Sung, Ackerman, Dillon et al. 2016) |
| | *Escherichia coli* (24) | $2{,}00.10^{-10}$ | 4,64 | (Lee et al. 2012) |
| | *Mycobacterium tuberculosis** (28) | $1{,}95.10^{-10}$ | 4,41 | (Lynch 2010a) |
| | *Thermus thermophilus** (33) | $1{,}38.10^{-10}$ | 2,13 | (Lynch 2010a) |
| | *Vibrio cholerae* (34) | $1{,}15.10^{-10}$ | 3,95 | (Sung, Ackerman, Dillon et al. 2016) |
| | *Pseudomonas aeruginosa* (29) | $7{,}92.10^{-11}$ | 6,53 | (Dettman et al. 2016) |



FIGURE 1.8 – Relations linéaires entre taille du génome, Ne, et taux de mutation pour une sélection de bactéries, eucaryotes unicellulaires et eucaryotes pluricellulaires. Les numéros se réfèrent aux organismes listés TABLE 1.1. Extrait de (Lynch, Ackerman et al. 2016). Figure non reproduite dans la version de diffusion de la thèse.

### 1.3.2 Variation du taux du mutation le long du génome

Le taux de mutation ne varie pas seulement entre organismes mais également le long du génome, un constat qui s'illustre par le fait que les différents types de mutations ne sont pas distribués uniformément le long des chromosomes (Duret 2009 ; Hodgkinson et Eyre-Walker 2011 ; Wolfe et al. 1989).
Ces variations locales dans les taux de mutations, aussi appelées RViMR (*Regional Variation in Mutation Rates*, (Makova et al. 2015)) s'observent à plusieurs échelles, de variations dans le taux de mutation d'un nucléotide à l'autre jusqu'à des patrons de mutation de l'ordre du mégabase. Par ailleurs, comme illustré FIGURE 1.9, elles sont associées à des propriétés intrinsèques de la séquence de l'ADN (contenu local en GC, nucléotides amont et aval, présence de séquences répétées), mais peuvent également dépendre de processus cellulaires (réplication, recombinaison, transcription) ou de modifications chromatiniennes.

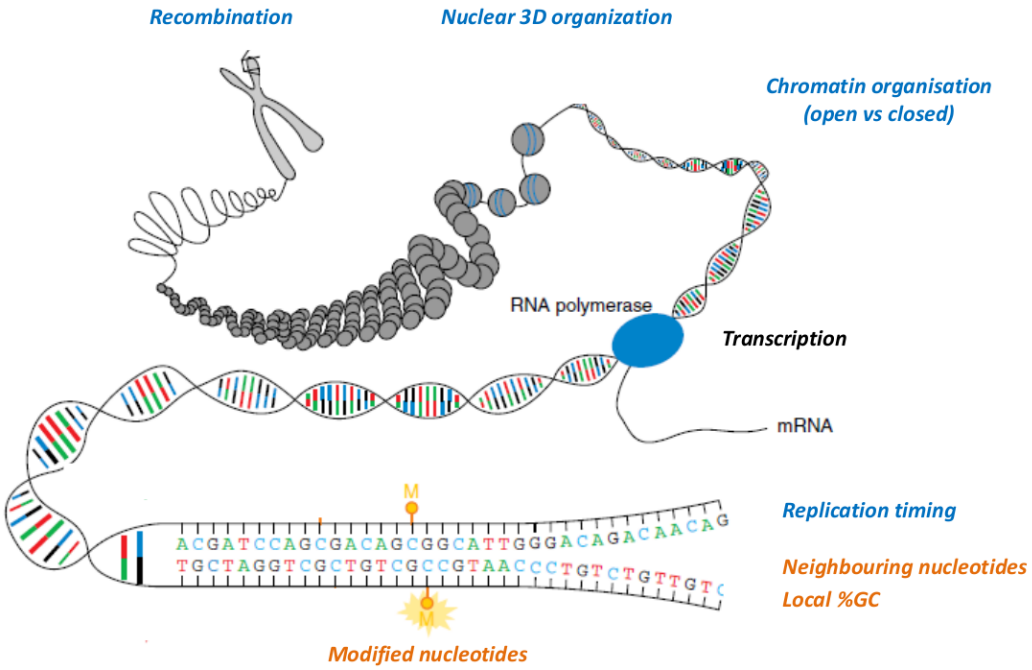


FIGURE 1.9 – Récapitulatif des différents facteurs influençant le taux de mutation le long du génome. Les facteurs associés à des variations locales et à large échelle sont mentionnés en orange et bleu respectivement. Modifié d'après (Acuna-Hidalgo et al. 2016).

Un exemple bien connu de cette combinaison entre propriétés du génome et de l'épigénome est le taux élevé de transitions C → T aux cytosines méthylées, dont le mécanisme est présenté FIGURE 1.10. La méthylation des cytosines, qui sera décrite en plus amples détails au CHAPITRE 2, est une modification chromatinienne qui consiste en l'ajout post-réplicatif d'un groupement méthyle sur le carbone 5 de la cytosine. Ces cytosines ainsi modifiées sont chimiquement plus instables que les C non méthylées, et leur désamination

spontanée produit une thymine, d'où un mésappariement G:T qui sera corrigé ou en G:C, ou en A:T, ce second cas de figure donnant lieu à la formation d'une transition C → T (Bird 1986 ; Ehrlich et al. 1981 ; Holliday et al. 1993).
Dans les génomes globalement méthylés comme le sont ceux des mammifères, cette hypermutabilité (parfois qualifiée de "CpG effect") a été associée à l'érosion à l'échelle de l'évolution du dinucléotide CG, contexte de méthylation préférentiel chez ces organismes (Bird 1980). Une illustration en est le génome humain, où hors des "îlots" CpG le dinucléotide CG n'est présent qu'à 20% de sa fréquence théorique attendue (Lander et al. 2001). En plus de cette mutabilité accrue conférée par l'ajout du groupement méthyle, il a pu être mis en évidence que le contexte nucléotidique local pouvait également influencer le taux de mutation à ces sites, avec un taux de mutation moindre au sein des régions GC-riches, probablement en lien avec une meilleure stabilité de la double-hélice donc une probabilité réduite de désamination (Fryxell et al. 2005). L'accumulation des données génomiques et épigénomiques a permis d'identifier que la 5mC n'était pas la seule base modifiée à être associée à un type de mutation spécifique : ainsi, les 5hmC, produit d'oxydation des 5mC, présentent quant à elles un taux accru de transversions C → G (Supek, Lehner et al. 2014).

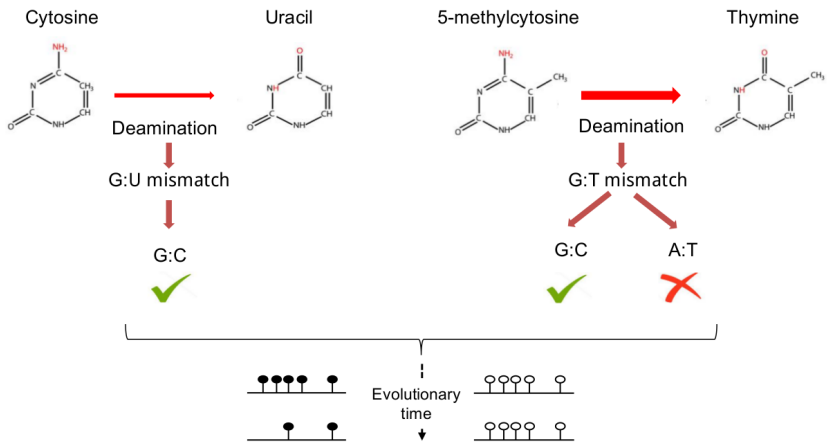


FIGURE 1.10 – Mutabilité accrue des cytosines méthylées. Explications dans le texte. Les ronds noirs et blancs représentent respectivement les cytosines méthylées et non méthylées, la disparition du pictogramme indique la disparition de la cytosine.

En plus du contenu local en GC, dont la contribution effective aux patrons de mutation est difficile à déterminer particulièrement chez les vertébrés en raison de facteurs confondants (organisation du génome en isochores, recombinaison), il peut être observé que les nucléotides proximaux vont également influencer le taux de mutation du nucléotide "focal". L'incertitude réside cependant quant à la fenêtre à employer pour définir de tels effets locaux. Ainsi, s'il est connu depuis longtemps que le taux de mutation va varier d'un triplet à l'autre (Blake et al. 1992), des travaux plus récents ont mis en évidence qu'un

contexte à 7 nucléotides expliquait plus en profondeur les patrons de polymorphisme le long du génome (Aggarwala et al. 2016).
De plus, des effets à plus longue distance ont également été identifiés, sans pour autant que leur origine exacte ne soit déterminée (Hodgkinson, Ladoukakis et al. 2009). En outre, il a été proposé que les régions d'hétérozygotie présentent un taux de mutation plus élevé, une observation qui peut être expliquée par un mauvais appariement à la méiose (Yang et al. 2015).

Du côté des facteurs responsables de variations à plus large échelle, il peut être constaté que la transcription influençait le patron de mutation. Ainsi, au niveau des gènes activement transcrits, la voie du NER (*nucleotide excision repair*) va corriger les dommages à l'ADN, au niveau du seul brin transcrit toutefois. Dès lors, l'intervention de ce mécanisme dit de *transcription-coupled DNA repair* (TCR) va être détectée en raison d'un biais mutationnel brin-spécifique (Mugal, Grünberg et al. 2009). Pour autant, il est une seconde modalité de la transcription qui au contraire va se traduire par un taux de mutation plus élevé. En effet, il faut garder à l'esprit que la transcription impose de désenrouler localement la double hélice d'ADN, d'où un stress de torsion qui peut se traduire par des cassures double-brin ainsi qu'à l'exposition de l'ADN simple brin à des mutagènes. Deux autres modalités associent la transcription à la création de mutations : d'une part au travers de la collision entre fourches de réplication et machinerie de transcription, et d'autre part en raison du regroupement sous forme de foci dans le noyau des gènes transcrits au même moment, qui peuvent conduire à la formation de SV entre segments issus de plusieurs chromosomes. De façon similaire, les gènes répliqués au même moment seront soumis à un stress torsionnel alors qu'ils sont en contact étroit dans le noyau au sein des foci de réplication. Ainsi, comme illustré FIGURE 1.11, la réplication seule ou de part ses interactions avec la transcription peut conduire à la formation de SV (Sima et al. 2014).

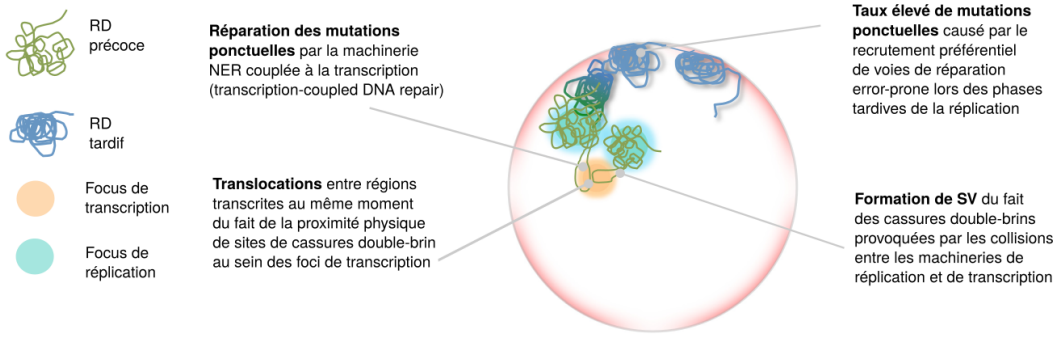


FIGURE 1.11 – Liens entre taux de mutation et réplication. RD : domaine de réplication. Modifié d'après (Sima et al. 2014).

Un dernier facteur jouant un rôle non négligeable quant à la variation du taux de mutation le long du génome est l'état chromatinien (Makova et al. 2015).
La richesse de données génomiques et épigénomiques pour différents types de cancer ont permis à ce jour d'obtenir un panorama précis de l'impact des différentes modifications chromatiniennes sur le patron de mutation. Ainsi, il a pu être décrit que les différentes modifications chromatiniennes constituaient un déterminant majeur des variations locales dans le taux de mutation. Notamment, il a été montré que l'état H3K9me3, caractéristique de l'hétérochromatine compactée, était associé à des taux élevés de mutations ponctuelles, une corrélation qui persiste au travers des types cancéreux étudiés et qui explique à lui seul 40% de la variation du taux de mutation (Schuster-Böckler et al. 2012). Comme mis en évidence peu de temps après, ce taux de mutation plus élevé dans l'hétérochromatine s'explique par une accessibilité moindre de ces régions à la machinerie de réparation de l'ADN : il a pu être montré qu'un rétablissement de la transcription dans les régions hétérochromatiques suffisait à restaurer l'accessibilité aux machineries du DNA repair (Zheng et al. 2014), et plus précisément à la machinerie du MMR, dont l'accès aux différentes fractions du génome sous-tend pour partie la variation dans le taux de mutation le long du génome (Supek et Lehner 2015).
Ainsi, les facteurs associés à la variation des taux de mutation le long du génome sont multiples et souvent interconnectés, rendant la détermination de la contribution des uns et des autres au RViMR particulièrement ardue (Makova et al. 2015).

Chapitre 2

# De la chromatine à l'épigénomique

L'existence d'états chromatiniens différentiels chez les eucaryotes et les propriétés génériques qui y sont associées ont été mentionnées brièvement dans la section précédente du manuscrit. Dans ce chapitre, je décrirai plus en détails l'épigénome d'Arabidopsis, et plus particulièrement une modification chromatinienne spécifique, la méthylation de l'ADN. Je présenterai également l'état des connaissances concernant les modifications ciblées de l'épigénome.

## 2.1 La chromatine, plus qu'une structure d'empaquetage de l'ADN

### 2.1.1 Aspects historiques

Chez les eucaryotes, l'ADN génomique est organisé autour de protéines histones en un complexe nucléoprotéique appelé chromatine, dont la fonction première est de parvenir à un niveau de compaction suffisamment élevé pour faire entrer une molécule d'ADN de l'ordre du mètre dans un noyau de quelques microns de diamètre. Le terme "chromatine" est introduit pour la première fois en 1882 par Flemming en raison de la capacité forte présentée par ces structures du noyau à retenir les colorants (*chromos*, couleur) en microscopie (voir (Paweletz 2001) et (D. Olins et al. 2003) pour revue). L'existence de différentes "formes" de chromatine est ensuite établie par Heitz en 1928 : au moyen d'approches cytologiques, ce dernier a pu mettre en évidence des régions chromosomiques qui restent condensées tout au long du cycle cellulaire, qu'il a qualifié d'hétérochromatine, par opposition à l'euchromatine qui correspond aux régions qui se décondensent et deviennent peu à peu invisibles en microscopie au cours de l'interphase (voir (Passarge 1979) pour revue).

Ces observations ont conduit Heitz à émettre dès 1929 l'hypothèse que l'hétérochromatine, systématiquement maintenue dans un état compacté, puisse correspondre à un état fonctionnellement inactif du génome (voir (Passarge 1979) pour revue). L'hétérochromatine sera par la suite divisée entre hétérochromatine constitutive et facultative, ce afin de distinguer les régions qui sont systématiquement dans un état condensé sur les deux homologues et dans tous les types cellulaires (à l'exemple des centromères) des régions

qui peuvent alterner entre état actif et inactif au cours du développement ou d'un type cellulaire à l'autre.

Si la fonction de la chromatine est donc pressentie dès les années 30, sa structure ne sera résolue que plus tard, au bénéficie des progrès de la microscopie : en microscopie électronique, la chromatine apparaît sous la forme d'une structure répétitive de 11nm de diamètre ayant l'aspect d'un "collier de perles". Ces perles prendront par la suite le nom de nucléosome (A. Olins et al. 1974 ; D. Olins et al. 2003) et leur structure cristallographique sera affinée dans les années 90 (Luger et al. 1997 ; Richmond et al. 1984). Les nucléosomes constituent le premier niveau de compaction de la chromatine, qui va ensuite subir des repliements additionnels jusqu'à atteindre un niveau de compaction ultime sous la forme d'un chromosome métaphasique (FIGURE 2.1).

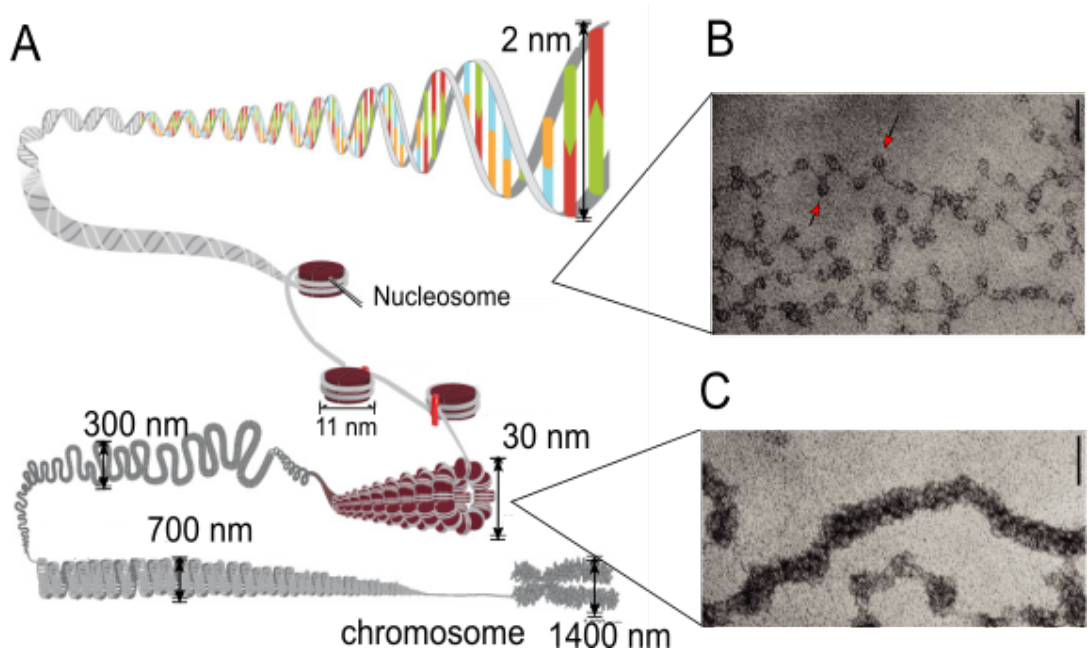


FIGURE 2.1 – Les différents niveaux de compaction de la chromatine. Adapté de (D. Olins et al. 2003).

### 2.1.2 Structure et composition de la chromatine

Le nucléosome, unité de base de la chromatine, correspond à l'enroulement de 147 pb d'une molécule d'ADN (soit environ deux tours) autour d'un octamère d'histones. Cet octamère, illustré FIGURE 2.2, est une structure protéique globulaire comprenant deux exemplaires de chacune des histones H2A, H2B, H3 et H4, sous la forme de deux hétérodimères H2A-H2B et deux hétérodimères H3-H4 (Luger et al. 1997).
Les régions N-terminales (aussi dites queues amino-terminales) des histones sont exposées

à l'extérieur du nucléosome et peuvent être modifiées de manière post-traductionnelle, par exemple au travers de l'ajout d'un ou plusieurs groupements méthyle ou acétyle sur certains résidus (Luger et al. 1997). Ces modifications post-traductionnelles sont covalentes et réversibles, et elles vont ainsi influencer localement le niveau de compaction de l'ADN, tout comme le remplacement des histones canoniques (mentionnés ci-dessus) par des variants spécifiquement associés à des situations particulières (FIGURE 2.2). Quelques exemples en sont H2A.X, qui remplace H2A aux sites de cassure double-brin, ou bien cenH3, variant de H3 spécifique des régions centromériques.

Cette succession de nucléosomes, séparés les uns des autres par un segment d'ADN appelé "linker" de taille variable selon les organismes (entre 20 et 90 pb environ) ne constitue que la première étape de compaction de l'ADN. Par la suite, cette structure va subir une compaction additionnelle pour former la fibre chromatinienne de 30 nm, dans laquelle les nucléosomes s'enroulent en solénoïde ou en zigzag. Ce niveau de compaction modéré est celui observé dans l'euchromatine, et la stabilisation de la structure est favorisée par l'histone H1, dit histone "linker", qui relie le nucléosome aux portions d'ADN linker qui l'entourent (Harshman et al. 2013). Dans les régions hétérochromatiques, cette fibre subit des repliements additionnels dont le détail biophysique est à ce jour encore débattu ; le niveau maximal de compaction étant atteint en métaphase sous la forme des chromosomes tels que visibles en microscopie (FIGURE 2.1).

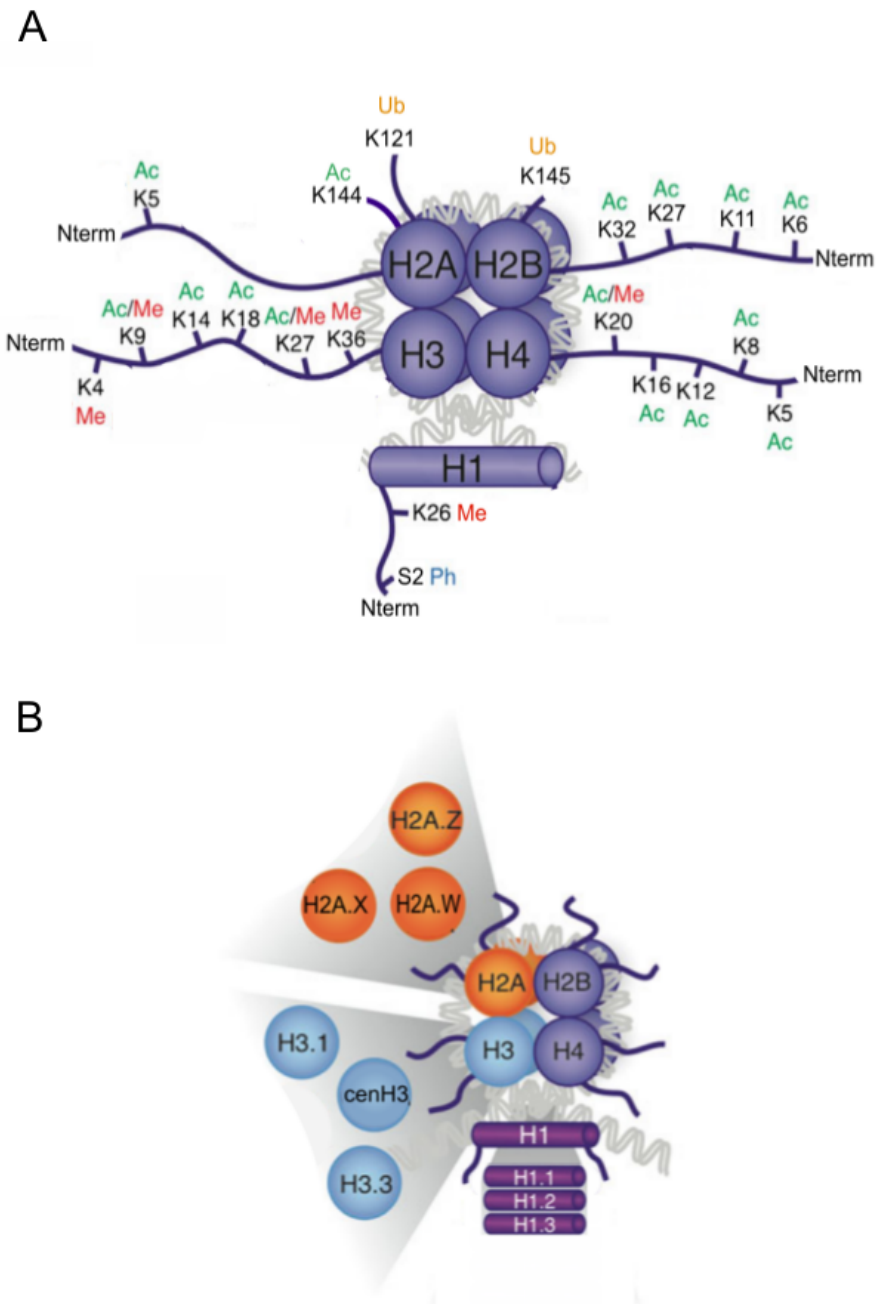


FIGURE 2.2 – Modifications post-traductionnelles et variants d'histones chez Arabidopsis. A. Représentation des principales modifications post-traductionnelles d'histones. Ac, Me, Ph indiquent l'ajout d'un ou plusieurs groupements acétyle, méthyle ou phosphate sur les différents résidus, majoritairement des lysines (K). B. Représentation des différents variants associés à chaque histone canonique.

Il est à noter que l'euchromatine et l'hétérochromatine ne sont pas distribuées au hasard dans le génome. Ainsi, comme illustré FIGURE 2.3, chez *Arabidopsis thaliana*, l'euchromatine correspond aux bras chromosomiques riches en gènes tandis que l'hétérochromatine va être essentiellement localisée au niveau des régions centromériques et péricentromériques riches en séquences répétées. Dans le noyau interphasique, l'hétérochromatine apparaît sous la forme de foci denses au microscope appelés chromocentres, au nombre de 6 à 10, d'où partent des boucles d'euchromatine (Fransz et al. 2002).



FIGURE 2.3 – Organisation du génome et de la chromatine chez *Arabidopsis thaliana*.
A. Représentation schématisée des 5 chromosomes d'Arabidopsis (2n=10), accession Col-0. Les longueurs des chromosomes sont figurées sous chacun d'eux. Les régions péricentromériques sont indiquées en vert, les régions centromériques (constituées de répétitions de 180pb) en gris. Les loci d'ADNr 45S (bleu) sont portés par les chromosomes II et IV, les loci d'ADNr 5S (rouge) par les chromosomes III, IV et V. B. Organisation en microscopie du noyau d'Arabidopsis (gauche) et schéma d'un chromocentre du chr IV (droite). Le marquage DAPI fait apparaître une dizaine de structures d'hétérochromatine fortement condensées, appelées chromocentres (cc). L'euchromatine apparaît en gris diffus, le nucléole n'est pas visible au marquage au DAPI. Echelle : 5µm. Les chromocentres regroupent les structures hétérochromatiques, desquelles sont extrudées les régions d'euchromatine sous forme de boucle (gris foncé) : ainsi, les régions transcriptionnellement actives du 45S et du 5S extrudent respectivement vers le nucléole et le compartiment euchromatique. Extrait de (Benoit et al. 2013). Figure non reproduite dans la version de diffusion de la thèse.

### 2.1.3 Etats chromatiniens et épigénome

La composition de la chromatine en marques chromatiniennes et variants d'histones le long du génome est appelée l'épigénome. Ainsi, si toutes les cellules d'un organisme pluricellulaire possèdent (aux mutations près) la même séquence d'ADN, chaque type cellulaire présente un épigénome qui lui est propre, qui peut varier au cours du cycle cellulaire, mais également en réponse à des signaux internes (développementaux) ou externes (environnementaux) (Roudier, F. Teixeira et al. 2009 ; W. Xie et al. 2013 ; Y. Zhang et al. 2018).

A ce jour, plusieurs centaines de modifications d'histones ont pu être identifiées, d'où un nombre quasi-infini de combinaisons possibles. Néanmoins, il peut être observé que la majorité des modifications vont préférentiellement s'associer en un nombre restreint de combinaisons, permettant dès lors de définir un nombre réduit d'états chromatiniens. Il a ainsi pu être proposé que la complexité de la chromatine puisse être réduite à cinq et six états chromatiniens principaux chez la Drosophile et l'Homme respectivement (Filion et al. 2010 ; Rao et al. 2014).
L'organisation chromatinienne d'Arabidopsis (Figure 2.4) apparaît être similaire : l'analyse combinatoire de 12 marques chromatiniennes a permis la définition de quatre états chromatiniens principaux, chacun d'entre eux se caractérisant par une combinaison spécifique de marques chromatiniennes et étant préférentiellement associé à certaines annotations génomiques (Roudier, Ahmed et al. 2011).
Ainsi :

- L'état CS1 (Chromatin State 1) est principalement associé aux gènes transcriptionnellement actifs. Il se caractérise par un enrichissement en di- et triméthylation de la lysine 4 de l'histone H3 (H3K4me2/3), triméthylation de la lysine 9 de l'histone H3 (H3K9me3), monoubiquitination de l'histone H2B (H2Bub), triméthylation de la lysine 36 de l'histone H3 (H3K36me3) et l'acétylation de la lysine 56 de l'histone H3 (H3K56ac). Un signal de méthylation de l'ADN peut également être détecté sur quelques domaines CS1.
- L'état CS2 est principalement associé aux gènes réprimés ciblés par le complexe PRC2. Il se caractérise par un enrichissement en H3K27me3, marque spécifiquement déposée par les complexes Polycomb.
- L'état CS3 peut être interprété sur le plan cytologique comme les régions qui forment l'hétérochromatine. Il se caractérise par un enrichissement en H3K9me2, H3K27me1/2, H4K20me1 et en ADN méthylé, et correspond aux ET transcriptionnellement inactivés.
- L'état CS4 ne présente aucun enrichissement particulier en une marque chromati-

nienne et semble légèrement associé aux régions intergéniques. Il peut être interprété comme indiquant une absence d'activité transcriptionnelle.

A l'exception de l'état CS3, qui forme de large blocs au niveau des régions péricentromériques et du knob du chromosome 4, les états CS1, 2 et 4 sont associés à des domaines de petite taille (de l'ordre d'une unité transcriptionnelle) qui alternent sans ordre apparent le long des chromosomes (FIGURE 2.4). Il est à noter que cette organisation diffère de celle observée chez les mammifères ou chez la Drosophile.

Cette vision a depuis été enrichie par la prise en compte de marques chromatiniennes additionnelles, notamment des variants d'histones (Sequeira-Mendes et al. 2014 ; C. Wang et al. 2015). Cela a abouti à la définition de 9 à 11 états chromatiniens, lesquels donnent lieu à une caractérisation affinée des gènes transcriptionnellement actifs en permettant de distinguer promoteur, site de démarrage de la transcription (TSS) et corps du gène ; sans pour autant invalider la description originelle à 4 états.

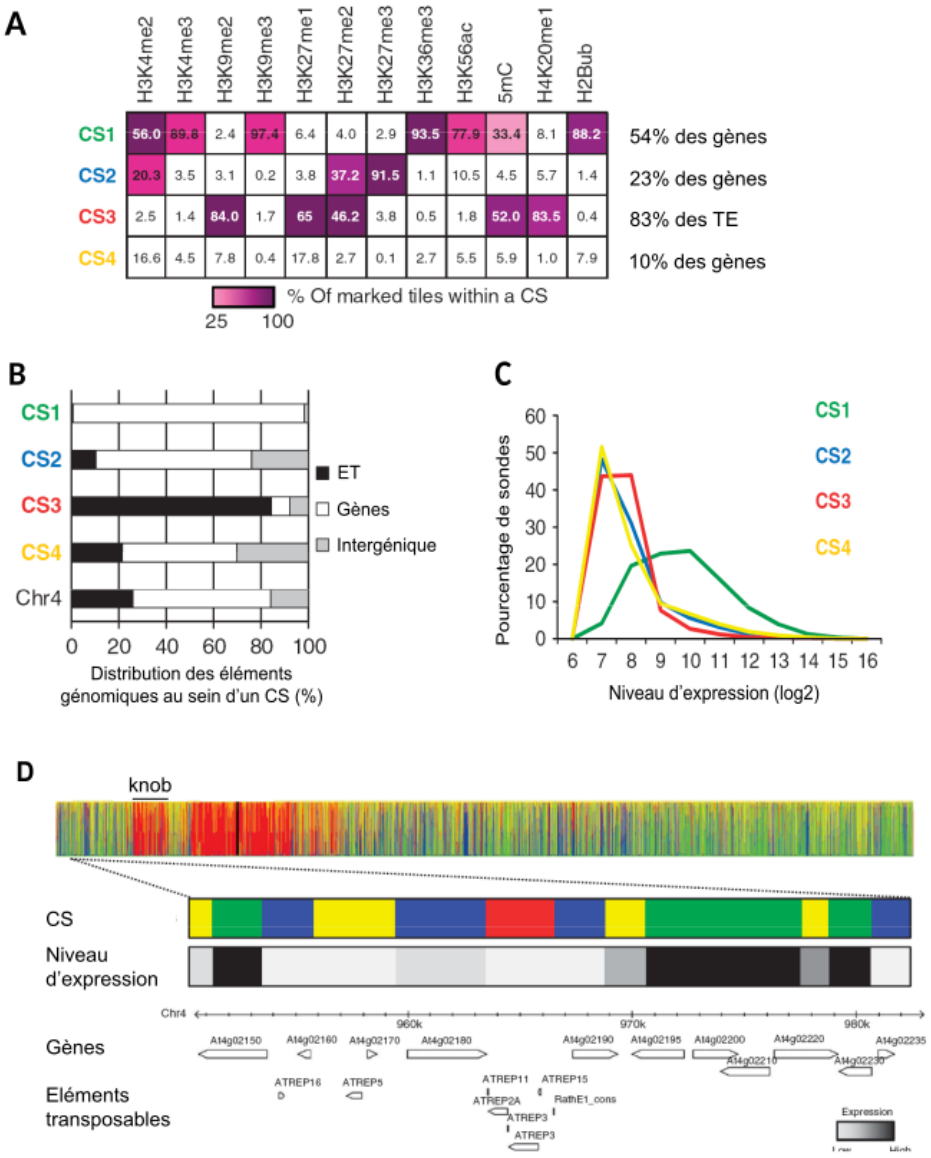


FIGURE 2.4 – Etats chromatiniens chez Arabidopsis : composition et distribution génomique.
D'après (Roudier, Ahmed et al. 2011). A. Composition des quatres états chromatiniens (CS1-4) en les 12 modifications chromatiniennes étudiées par ChIP- ou MeDIP-chip. Les modifications représentatives de chaque CS sont indiquées par un gradient de couleurs. B. Proportions relative de gènes, ET et régions intergéniques dans chaque CS le long du chromosome 4. C. CS et expression des gènes. Le pourcentage de sondes associées à chacun des CS est donné pour chaque niveau d'expression. D. Distribution des quatre CS le long du chromosome 4 et détail d'une région d'euchromatine. La couleur de chaque sonde (panel du haut) illustre son appartenance à l'un ou l'autre des CS.

## 2.2 La méthylation de l'ADN

### 2.2.1 Généralités

La méthylation de l'ADN est l'une des modifications chromatiniennes les mieux décrites à ce jour. Elle concerne majoritairement les cytosines et consiste en l'ajout par des enzymes spécifiques, les méthyltransférases, d'un groupement méthyle sur le carbone n°5 des cytosines, à partir du substrat donneur S-adénosylméthionine.

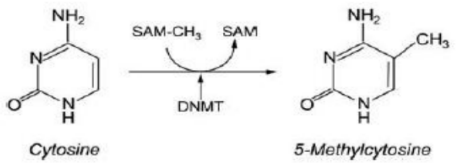


FIGURE 2.5 – Méthylation en C5 de la cytosine.
DNMT, DNA méthyltransférase ; SAM, S-adénosylméthionine

D'importantes différences peuvent être observées entre les patrons de méthylation des plantes et des mammifères (FIGURE 2.7). Tandis que chez les mammifères, la méthylation de l'ADN est répartie sur l'ensemble du génome et concerne presque uniquement les cytosines en contexte CG (Z. Chen et al. 2011 ; Suzuki et al. 2008), chez les plantes, cette modification est plus particulièrement retrouvée au niveau des séquences répétées (Cokus et al. 2008 ; Lister et al. 2008 ; X. Zhang, Yazaki et al. 2006), et les trois contextes de séquence (CG, CHG et CHH, où H = A, T ou C) peuvent être ciblés par la méthylation. Ainsi, le fait que seules 6% des cytosines d'Arabidopsis soient méthylées s'interprète en regard du faible contenu en ET dans son génome (FIGURES 2.6 et 2.7). On peut également constater que les différents contextes sont méthylés dans des proportions distinctes (FIGURE 2.6), qui peuvent par ailleurs varier en fonction des stades développementaux (Bouyer et al. 2017 ; Kawakatsu, Nery et al. 2017 ; Luo et al. 2018 ; H. Zhang et al. 2018).

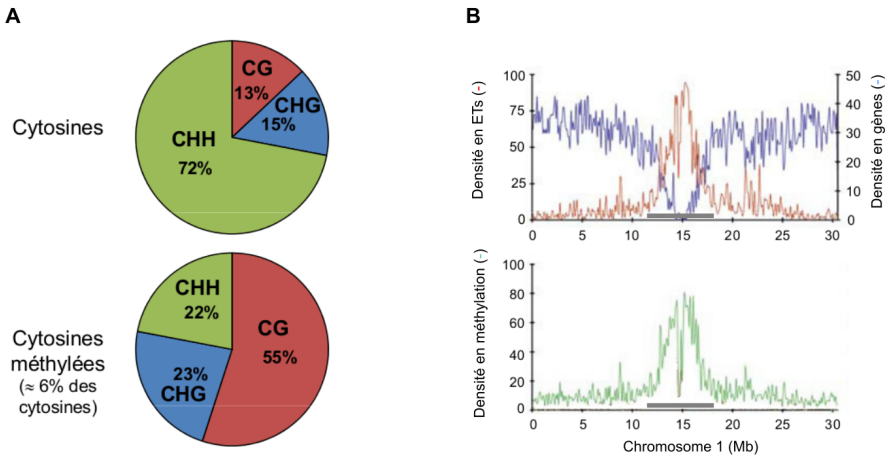


FIGURE 2.6 – Distribution génomique de la méthylation de l'ADN chez Arabidopsis. A. Répartition dans les trois contextes de séquence (CG, CHG, CHH) des cytosines du génome (haut) et de la fraction méthylée (bas) (Cokus et al. 2008 ; Lister et al. 2008). B. Distribution des gènes, des éléments transposables (haut) et de la méthylation de l'ADN (bas) le long du chromosome 1. La région péricentromérique est figurée par un rectangle gris foncé.

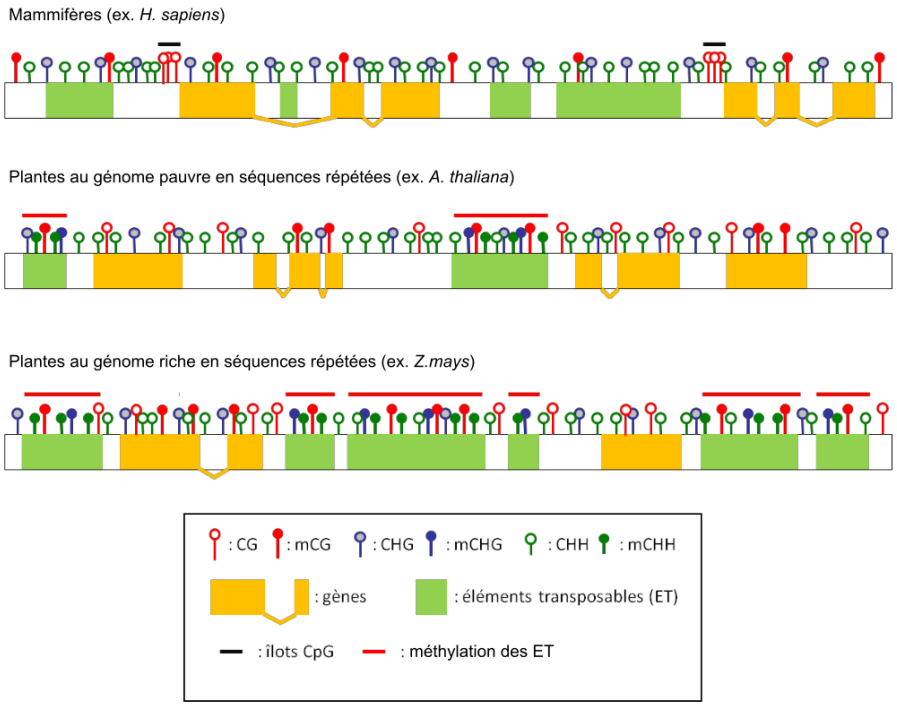


FIGURE 2.7 – Représentation schématique des profils type de méthylation de l'ADN chez l'Homme, Arabidopsis et le Maïs. On notera la méthylation sur la majorité des sites CG à l'exclusion des îlots CpG chez l'Homme, lesquels sont situés dans les promoteurs des gènes et ne sont méthylés que de manière spécifique au cours du développement ; ainsi que les différences dans la densité en ET entre Arabidopsis et le Maïs.

### 2.2.2 Rôles de la méthylation de l'ADN

#### Répression de l'activité des éléments transposables

Comme nous l'avons vu précédemment, les ET sont susceptibles de poser des menaces majeures quant au maintien de l'intégrité des génomes. Une fonction générique de la méthylation de l'ADN, quel que soit l'organisme considéré, a à voir avec le maintien à l'état transcriptionnellement inactif des éléments transposables.
Aussi, chez Arabidopsis, l'hétérochromatine péricentromérique ainsi que les quelques éléments transposables dispersés le long des bras chromosomiques sont fortement méthylés dans les trois contextes. Chez les mutants défectifs pour la méthylation de l'ADN, une dérépression transcriptionnelle des ET est ainsi observée, sans pour autant qu'une transposition extensive n'ait été observée à ce jour, un patron qui peut être expliqué par la présence de mécanismes de régulations post-transcriptionnels. Il est à noter que le patron de dérépression est différent d'un mutant à l'autre, mettant en exergue la complexité à la fois de la machinerie de méthylation et de la régulation de l'expression des ET.
Il est par ailleurs à noter que d'autres organismes ont évolué des moyens de défense contre les ET allant au-delà de la méthylation de l'ADN ou des autres modifications chromatiniennes répressives. Plus particulièrement, le champignon *Neurospora crassa* emploie le RIP (*repeat-induced point mutation*), qui combine mutations C:G $\rightarrow$ T:A et méthylation des quelques cytosines restantes afin d'inactiver les séquences répétées non plus seulement transcriptionnellement, mais également "en dur" par induction de mutations dans la séquence de l'ADN. Si le RIP semble être une stratégie particulièrement efficace dans la mesure où aucun transposon actif n'a été identifié chez cet organisme à ce jour, le prix à payer est que toutes les séquences répétées quelles qu'en soit leur nature vont être ciblées, ce qui constitue un frein à l'évolution du génome (Galagan et al. 2004).

#### Régulation de l'expression génique

La contribution de la méthylation de l'ADN à la régulation de l'expression génique diffère radicalement entre plantes et mammifères.
Chez ces derniers, comme illustré Figure 2.7, des patchs de dinucléotides CG, appelés îlots CpG, sont situés au niveau des séquences promotrices des gènes et la méthylation dynamique de ces structures au cours du développement ou d'une cellule à l'autre va être associée à l'expression ou à la répression transcriptionnelle des gènes. A l'opposé, il n'existe pas d'îlots CpG chez les plantes, et les gènes dont l'expression apparaît régulée par la méthylation se révèlent être accolés à des séquences répétées, à l'exemple de *FWA* (voir Chapitre 3). Aussi, la contribution de la méthylation de l'ADN à l'expression des gènes chez les plantes peut être vue comme un sous-produit de la présence d'ET à proximité, lesquels sont les cibles primaires de la méthylation.
Une conséquence en est que la majorité des mutants de méthylation sont embryoniques

létaux ou présentent des phénotypes drastiques chez les mammifères, mais sont parfaitement viables chez Arabidopsis. Une exception en est le Maïs chez lequel des altérations phénotypiques importantes peuvent être décrites dans les mutants de méthylation. Ces observations peuvent néanmoins être expliquées par l'importante proportion d'éléments transposables dans le génome de cet organisme (FIGURE 2.7), d'où un grand nombre de gènes se retrouvant à proximité d'ET et donc sous le contrôle de leur état de méthylation.

### Méthylation du corps des gènes

Comme mentionné précédemment, un signal de méthylation de l'ADN est détecté ponctuellement dans les régions CS1, qui correspondent aux gènes transcriptionnellement actifs.
Cette méthylation du corps des gènes (*gene body methylation*, gBM) concerne près d'un tiers des gènes chez Arabidopsis et se caractérise par de la méthylation en contexte CG qui va être répartie le long du corps du gène, essentiellement au niveau des exons, tandis que les sites d'intiation et de fin de la transcription vont en être dépourvus (Bewick et R. J. Schmitz 2017). A l'opposé de l'extinction transcriptionnelle retrouvée au niveau des ET, les gènes présentant de la gBM correspondent à des gènes constitutivement exprimés à des niveaux modérés (Bewick et R. J. Schmitz 2017 ; H. Zhang et al. 2018). L'analyse de mutants de méthylation affectés pour le gBM suggère qu'un effet sur le niveau d'expression des gènes est à exclure, et si plusieurs fonctions dont la régulation de l'épissage alternatif et la stabilisation de la transcription (Zilberman, Coleman-Derr et al. 2008 ; Zilberman, Gehring et al. 2007) ont pu proposées, aucune n'a reçu de réel support à ce jour (Zilberman 2017). Dans la mesure où la gBM tend à être associée aux mêmes sets de gènes parmi les différentes plantes vasculaires analysées à ce jour, il semble difficile de nier une possible importance fonctionnelle, mais le fait qu'au moins deux plantes à fleurs (*E. salsugineum* et *C. Planisiliqua*) en soient intégralement dépourvues (Bewick, Ji et al. 2016) suggère *a minima* que son importance biologique au sein des Angiospermes soit espèce-dépendante.

### 2.2.3 Établissement, maintenance et effacement de la méthylation

Au contraire des mécanismes associés au gBM, nous avons à ce jour une vision relativement détaillée bien que non exhaustive des différents acteurs impliqués dans l'établissement initial de la méthylation puis de sa maintenance post-réplicative au niveau des séquences répétées (FIGURE 2.8). En fonction du contexte génomique et chromatinien, la mise en place de la méthylation va être effectuée par des voies distinctes, qui vont également différer par le taux de méthylation qu'elles produisent. Ainsi, les sites CG, CHG et CHH présentent respectivement des taux de méthylation élevés (>80%), intermédiaires (40-60%) et faibles (20%) (Cokus et al. 2008 ; Lister et al. 2008).

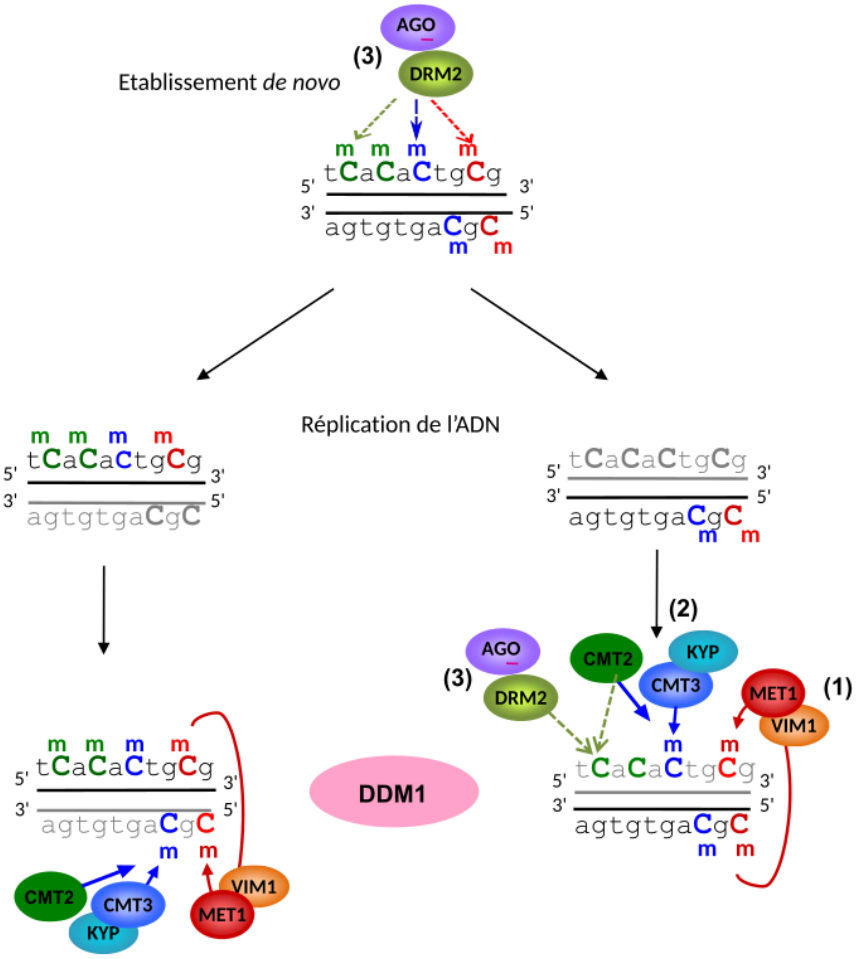


FIGURE 2.8 – Établissement et maintenance de la méthylation de l'ADN au niveau de séquences répétées. (1), (2) et (3) correspondent respectivement à la maintenance dans le contexte CG et CHG (1 et 2, détaillés FIGURE 2.10), et à l'établissement et à la maintenance dans le contexte CHH (3, détaillé FIGURE 2.9).

### Etablissement

L'établissement *de novo* de la méthylation dans les trois contextes est effectuée par la voie dite du RdDM (*RNA-directed DNA methylation*), schématisée FIGURE 2.9 (Matzke et al. 2014). Cette machinerie fait intervenir deux polymérases spécifiques des plantes, PolIV et PolV. Les transcrits générés par PolIV sont pris en charge par la polymérase ARN-dépendante RDR2 qui synthétise le brin complémentaire. Ces longs ARN double-brin sont reconnus par DCL3, qui les clive en petits ARN interférents (siRNA) de 24nt. Chacun des deux brins de ces siRNA est chargé dans un complexe AGO4, qui est ciblé au locus producteur de siRNA ainsi qu'aux autres locus de même séquence par l'appariement entre le siRNA et des transcrits encore mal définis générés à ces locus par PolV. Cette interaction aboutit *in fine* au recrutement de la méthyltransférase dite "*de novo*" DRM2 (DOMAINS REARRANGED METHYLTRANSFERASE2, homologue à Dnmt3 chez les mammifères) et à la méthylation des séquences ciblées.

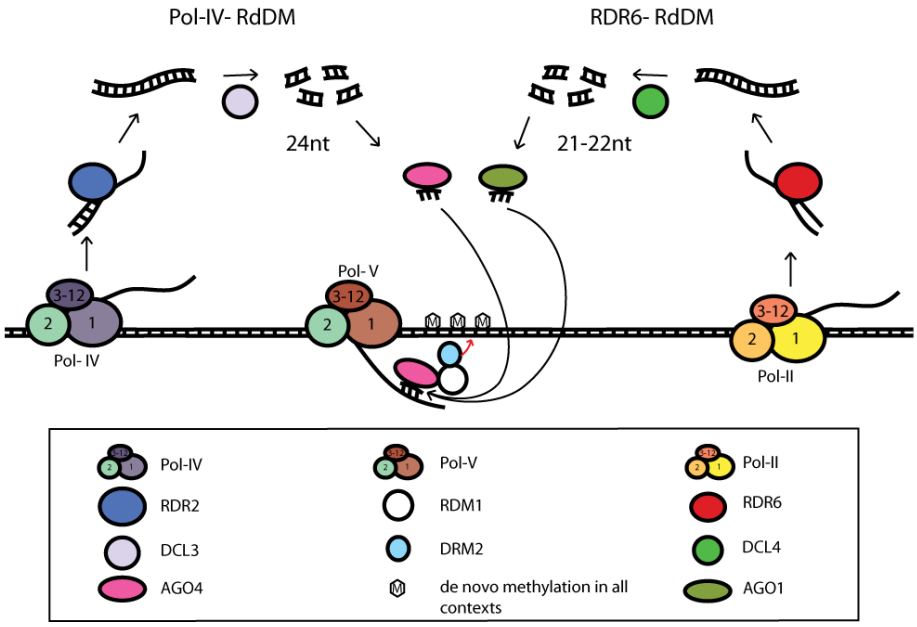


FIGURE 2.9 – Mécanisme du RdDM pour l'établissement de la méthylation *de novo* dans les 3 contextes et la maintenance dans le contexte CHH. Détails dans le texte.

Plusieurs observations suggèrent néanmoins que le recrutement de PolIV et PolV puisse à un locus donné puisse être dépendant d'une méthylation préalable, ce qui a pu remettre en question l'implication de ce RdDM canonique (ou PolIV-RdDM) dans l'acquisition de méthylation *de novo* plutôt que dans sa maintenance. A la place, il semblerait que ce soit une forme alternative du RdDM, dite RDR6-RdDM, qui soit impliquée dans le ciblage initial d'un locus pour y établir une méthylation *de novo*. Si la méthylation est là encore effectuée par DRM2, cette voie implique non plus des siRNA de 24nt mais de 21-22nt, dérivés de transcrits de PolII, dont le complémentaire est synthétisé par RDR6, et dont

le clivage puis le chargement fait intervenir DCL4 et AGO1 respectivement (Cuerda-Gil et al. 2016 ; Fultz, Choudury et al. 2015 ; Fultz et Slotkin 2017 ; Panda et al. 2016).

### Maintenance

Les patrons de méthylation établis *de novo* par le RdDM sont ensuite maintenus au cours des divisions cellulaires au moyen de machineries spécifiques à chaque contexte.
Dans le cas des contextes symétriques, CG et CHG, la méthylation peut être perpétuée par maintenance sur le brin opposé au cours de la réplication de l'ADN : après réplication, les deux molécules d'ADN double brin sont hémiméthylées en CG ou CHG, c'est-à-dire méthylées uniquement sur le brin d'origine (Figure 2.8).
En contexte CG (Figure 2.10, panel de gauche), la méthylation est maintenue par MET1 (METHYLTRANSFERASE1, homologue à Dnmt1 chez les mammifères), et son action est favorisée par le facteur VIM1 (VARIATION IN METHYLATION1) qui s'associe physiquement à l'ADN hémiméthylé (Kankel et al. 2003 ; Woo et al. 2007).
En contexte CHG (Figure 2.10, panel de droite), la méthylation est maintenue par CMT3 (CHROMOMÉTHYLASE3), une MTase spécifique aux plantes (Lindroth et al. 2001)). Il existe par ailleurs une boucle de renforcement entre CMT3 et KYP (KRYPTONITE), une histone méthyltransférase impliquée dans le dépôt de H3K9me2 : CMT3 interagit avec H3K9me2 au travers de son chromodomaine et catalyse la méthylation de la cytosine du triplet CHG, tandis que KYP se lie à cette cytosine méthylée via son domaine SRA et dépose H3K9me2 via son domaine SET (Du, Johnson et al. 2014 ; Du, Zhong et al. 2012).

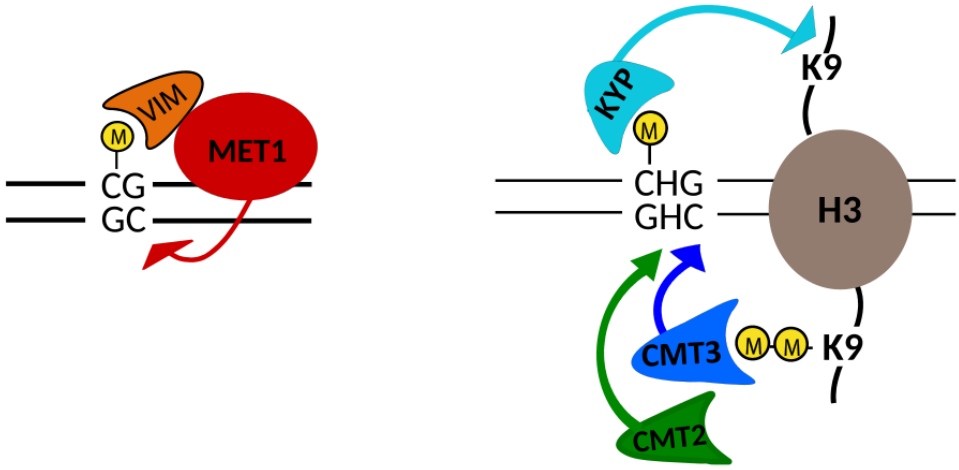


Figure 2.10 – Maintenance de la méthylation en contextes CG, à gauche, et CHG, à droite. Explications dans le texte.

A l'opposé, la maintenance en contexte CHH ne peut que s'effectuer au travers d'un ré-établissement de novo de la méthylation à ces sites du fait de leur asymétrie. Ce ré-établissement se fait alors via la voie du RdDM canonique décrite plus haut (FIGURE 2.9) au niveau des ET courts (moins de 500 pb). Au niveau des ET longs (plus de 4kb), seules les extrémités sont reméthylées par la voie du RdDM, tandis que la partie centrale fait intervenir CMT2, un paralogue de CMT3 qui de façon similaire à cette dernière interagit avec H3K9me2 (FIGURE 2.8). Par ailleurs, il a pu être montré que l'action de CMT2 était également requise pour la méthylation de certains sites CHG (FIGURE 2.10) de façon redondante avec CMT3 (Stroud et al. 2014 ; Zemach et al. 2013).

De plus, la maintenance des patrons de méthylation au niveau des ET (FIGURE 2.8) va nécessiter l'intervention d'autres facteurs et notamment de DDM1, un facteur de remodelage de la chromatine qui va induire une ouverture locale de la chromatine au niveau des régions hétérochromatiques condensées afin d'y permettre l'accès aux différentes ADN méthyltransférases et notamment à CMT2 (Jeddeloh et al. 1999 ; Kakutani, Jeddeloh et Richards 1995 ; Vongs et al. 1993 ; Zemach et al. 2013).

**Effacement**

Faute du maintien ou du ré-établissement de la méthylation à chaque division cellulaire, celle-ci va être inexorablement perdue au cours des réplications. A cette déméthylation dite passive s'ajoute une déméthylation active, qui s'effectue selon deux modalités distinctes entre plantes et mammifères. Chez ces derniers, la 5mC est tout d'abord oxydée en 5hmC par l'enzyme TET avant que la machinerie de réparation ne vienne exciser la base correspondante, tandis que les "déméthylases" des plantes sont des ADN glycosylases qui reconnaissent puis excisent elles-mêmes la 5mC.
Le génome d'Arabidopsis code pour quatre de ces glycosylases, qui peuvent effectuer la déméthylation dans les trois contextes CG, CHG et CHH : DME (DEMETER), la première déméthylase identifiée chez les eucaryotes (Choi et al. 2002), et trois autres enzymes apparentées : DML2, DML3 (DEMETER-LIKE 2 et 3) et ROS1 (REPRESSOR OF SILENCING 1) (Gehring et al. 2009 ; J. Zhu 2009).
Alors que DML2, DML3 et ROS1 sont exprimées dans l'ensemble de la plantes, DME présente une fenêtre d'expression restreinte au cours du développement (expression exclusive dans les cellules associées aux gamètes mâle et femelle), (Hsieh et al. 2009 ; Park et al. 2017 ; Schoft et al. 2011), en lien avec une régulation par la méthylation de l'expression de gènes spécifiques au cours du développement.
En effet, ces différentes déméthylases présentent des spécificités de ciblage, qui vont notamment dépendre de caractéristiques chromatiniennes. Ainsi, la déméthylation effectuée par DME cible majoritairement les ET de petite taille, AT-riches et situés dans les régions euchromatiques (d'où une expression altérée des gènes à proximité à l'exemple de *FWA*,

décrit Chapitre 3) (Frost et al. 2018) ; tandis que ROS1 va être essentiellement associé à une déméthylation des extrémités des ET ciblés par le RdDM, contrecarrant ainsi un possible effet sur l'expression des gènes à proximité (Tang et al. 2016 ; Williams et al. 2015), notamment en raison du mécanisme de *spreading* au travers duquel la méthylation associée à l'ET "dépasse" sur les séquences régulatrices des gènes proximaux (Ahmed et al. 2011).

## 2.3 Modifications ciblées de l'épigénome

Comme nous l'avons décrit au cours de ce chapitre, les différentes modifications chromatiniennes jouent (avec les facteurs de transcription) un rôle clé dans la régulation spatio-temporelle de l'activité du génome.
Pour autant, il demeure complexe d'associer l'état chromatinien d'un locus donné à son expression et au phénotype qui en découle autrement que par des approches indirectes ; à savoir l'établissement de corrélations entre données transcriptomiques (RNAseq) et épigénomiques (WGBS, ChIPseq) pour une diversité de marques et l'utilisation de transgéniques dans lesquels l'expression du gène d'intérêt ou le profil global d'une ou plusieurs marque(s) chromatinienne(s) vont être altérés. De telles approches présentent deux limites importantes : d'une part l'interprétation des phénotypes mutants est complexe en raison des effets confondants potentiels associés à la dérégulation à l'échelle du génome d'une modification chromatinienne, et d'autre part elles ne permettent en aucun cas d'établir un lien de causalité.
Dans ce contexte, la perspective de pouvoir effectuer des modifications ciblées du génome et de l'épigénome ouvre la porte à une meilleure compréhension du lien entre génotype et phénotype.

### 2.3.1 Des protéines à doigts de zinc à CRISPR-dCas9

Depuis la mise en place des premières protéines de fusion destinées à aller effectuer des cassures double-brin à un locus d'intérêt (Kim et al. 1996), les vingt dernières années ont vu se développer d'importantes avancées dans le domaine dit de l'édition du génome et de l'épigénome (Perez-Pinera, Ousterout et al. 2012).
Indépendamment de l'outil employé (Figure 2.11), ces approches se fondent sur la combinaison d'un domaine de reconnaissance de l'ADN (module de liaison), qui se lie à une séquence déterminée par l'expérimentateur, et d'un domaine enzymatique (module effecteur), qui va permettre d'altérer la séquence du génome ou l'état chromatinien au locus d'intérêt.

Les premières protéines de fusion développées utilisent des domaines à doigts de zinc

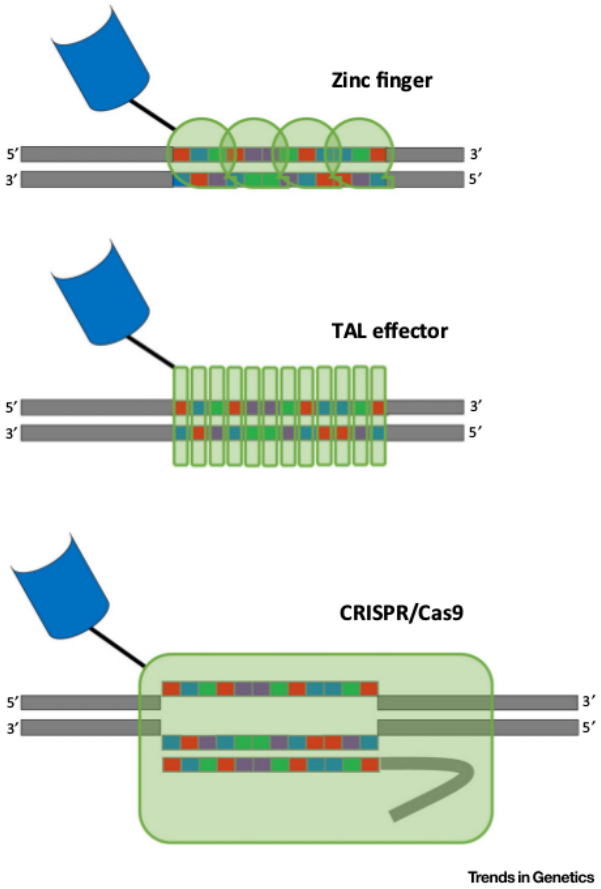


FIGURE 2.11 – ZFP, TALE et CRISPR/Cas9. Les modules effecteur et de liaison à l'ADN sont figurés en bleu et vert respectivement. Les spécificités de chacun de ces systèmes sont détaillés dans le corps du texte. Extrait de (Kungulovski et al. 2015).

(*Zinc Finger, ZF*), un type de domaine protéique responsable de la liaison à l'ADN des facteurs de transcription eucaryotes. Chaque domaine ZF, d'environ 3kDa, reconnaît une combinaison de trois nucléotides, et la combinaison de plusieurs domaines ZF va permettre de cibler une région précise dans l'ADN. Si le faible encombrement stérique engendré par leur petite taille garantit une bonne efficacité, seules les régions du génome pour lesquelles une combinaison de triplets auxquels correspondent un domaine ZF peuvent être ciblées, ce qui restreint la modularité de cette approche, sans compter la difficulté associé au design et au clonage des différents domaines ZF.
Une amélioration substantielle est apportée avec la mise en évidence des TALE (*Transcription Activator-Like Effector*) chez la bactérie *Pseudomonas*. Les TALE sont constituées d'une succession de répétitions de résidus protéiques, chacun d'entre eux (d'une taille d'environ 50kDa) étant spécifique d'un seul nucléotide, ce qui simplifie grandement le code séquence protéique-séquence ADN et permet en théorie de dessiner une TALE pour chaque séquence possible du génome. Néanmoins, comme dans le cas des ZFP (*Zinc-Finger Protein*), le module de liaison et le module effecteur sont partie intégrante d'une même protéine, et changer le module effecteur ou la séquence ciblée va nécessiter de re-

dessiner/recloner une protéine entière, soit donc un travail laborieux. Dans ce contexte, l'arrivée du système CRISPR/Cas représente une avancée majeure.

Les séquences CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*) et les protéines Cas (*CRISPR-associated protein*) constituent les composants du système immunitaire adaptatif des Procaryotes et des Archées, avec des variantes en fonction des espèces. Le système CRISPR/Cas de type II d'où dérivent les technologies d'édition du génome est constitué des acteurs suivants : des crRNA (*CRISPR-RNA*), qui sont transcrits à partir des séquences CRISPR, une protéine Cas9 d'environ 160kDa à activité nucléase, et des tracrRNA (*trans-activating crRNA*) qui relient la Cas9 et le crRNA. Le crRNA a la capacité à s'hybrider à une région d'ADN d'une vingtaine de nucléotides dont il est le complémentaire et qui est précédée d'une séquence PAM (NGG, où N peut être A, T, C ou G), et la liaison du complexe Cas9-crRNA-tracrRA à cette région va conduire à la formation d'une coupure double-brin à un site précis au sein de la région de complémentarité (Figure 2.12).

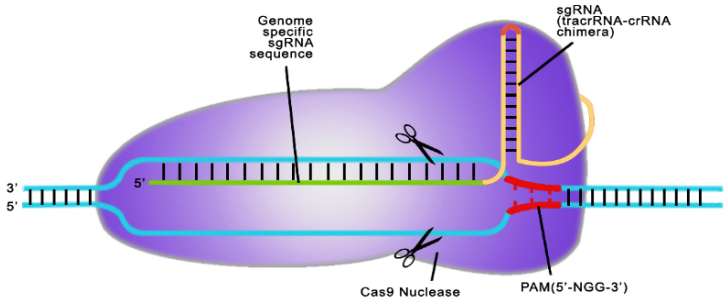


Figure 2.12 – Les différents composants du système CRISPR/Cas9

Ce système présente de nombreux avantages pour l'édition du génome ou de l'épigénome : le module de liaison étant un acide nucléique, son design est bien plus facile que celui d'une protéine et les effets off-target peuvent être prédits plus précisément ; tandis que la séparation des modules effecteur et de liaison permet de cibler aisément plusieurs régions du génome à la fois. Des développements techniques, conduisant pour l'un au remplacement du duplex crRNA-tracrRNA par un guide unique (*single guide RNA*, gRNA, Figure 2.12) et pour l'autre à la création d'une forme catalytiquement inactive de la protéine (dCas9, *dead Cas9*), qui peut alors être liée à différents effecteurs, vont par la suite diversifier et faciliter plus encore l'utilisation du système CRISPR/(d)Cas9 pour l'édition du génome et de l'épigénome.

### 2.3.2 Edition de l'épigénome médiée par dCas9 : considérations techniques

Depuis les premières tentatives de ciblage locus-spécifique d'une activité enzymatique par approche CRISPR/dCas9, les deux composantes de ce système ont bénéficié de développements techniques importants visant à améliorer l'efficacité de la modification recherchée, par exemple une répression transcriptionnelle du locus d'intérêt. Plusieurs stratégies ont été mises en place en ce sens, les trois dernières sont illustrées FIGURES 2.13 et 7.7 :

- La fusion de plusieurs domaines effecteurs, identiques ou de fonctions similaires, à une même protéine dCas9. Il s'agit d'une approche à présent standard dans le cadre de l'emploi d'une fusion dCas9-GFP pour obtenir un meilleur signal en microscopie, par exemple dans le cadre de l'analyse de la localisation nucléaire d'un locus d'intérêt (X. Ma et al. 2015 ; Pradeepa et al. 2016)
- L'emploi de multiples ARN guides ciblant la région d'intérêt. L'objectif est ici d'effectuer un tiling de la région pour amplifier l'effet recherché voire promouvoir le spreading de la modification induite aux sites adjacents (C. Liu et al. 2016 ; Morita et al. 2016 ; Perez-Pinera, Kocak et al. 2013 ; Thakore et al. 2015).
- Le recours au système SunTag (Tanenbaum et al. 2014). Le SunTag correspond à un ensemble de répétitions d'un peptide, dont l'épitope est reconnu par un anticorps minimal (ScFv). En fusionnant le SunTag avec la dCas9 et l'anticorps avec l'effecteur, il va être possible de recruter des copies multiples de l'effecteur, qui peuvent par ailleurs agir plus loin qu'au site de liaison (Morita et al. 2016).
- La modification du scaffold (région non variable du guide, correspondant au tracrRNA initial) des ARN guides, avec ajout d'un domaine de reconnaissance ARN-protéine à l'exemple de MS2. De façon similaire à l'approche SunTag, l'objectif est de permettre le recrutement d'un grand nombre d'effecteurs fusionnés avec le domaine protéique correspondant (Konermann et al. 2015 ; Xu, Tao et al. 2016).

Dans leur ensemble, ces méthodes de multiplexage et de multimérisation, utilisées séparément ou combinées, surpassent grandement les fusions de "première génération" entre dCas9 et un domaine effecteur unique (Garcia-Bloj et al. 2016 ; Konermann et al. 2015) et représentent très probablement le futur de l'édition de l'épigénome.

Un autre paramètre important quant à l'altération de la chromatine à un locus donné est le choix du domaine effecteur : en raison d'une activité catalytique généralement meilleure (Tahiliani et al. 2009) et du plus faible encombrement stérique, le domaine catalytique seul est généralement préféré à la protéine entière.

A ce jour, le système CRISPR/dCas9 a été employé avec succès pour induire et réprimer la transcription à des loci d'intérêt, au travers de fusions employant pour domaines

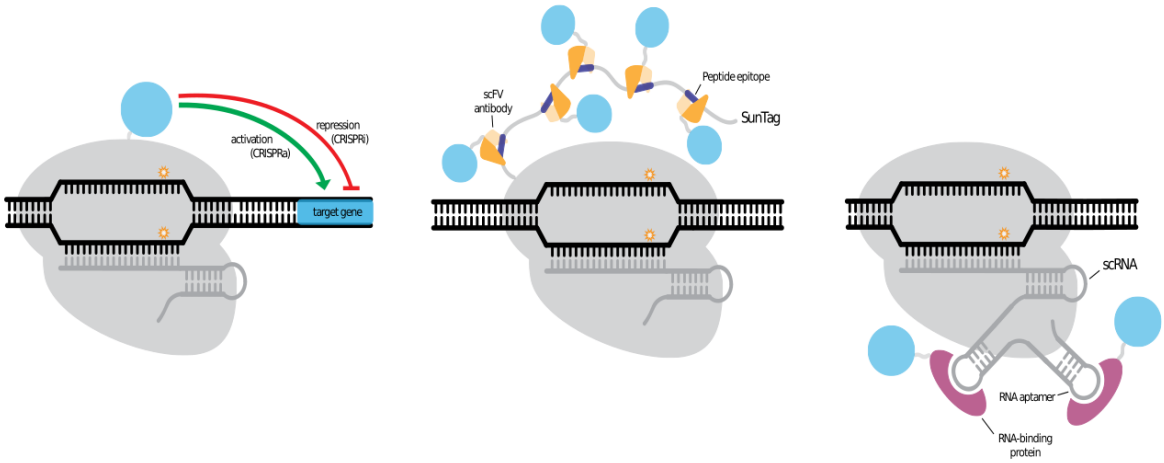


FIGURE 2.13 – Nouvelle génération de fusions dCas9. Les domaines effecteurs sont représentés par des cercles bleus. De gauche à droite : fusion dCas9-effecteur classique, SunTag, scaffold modifié avec un domaine de liaison à une protéine. Modifié d'après (Brocken et al. 2018)

effecteurs des activateurs transcriptionnels viraux (VP64) ou endogènes (p300) ainsi que des domaines catalytiques permettant d'altérer les modifications post-traductionnelles des histones ainsi que le profil de méthylation de l'ADN, cette dernière situation étant présentée dans le détail dans la section qui suit (voir (Adli 2018 ; Brocken et al. 2018) pour revue). Cependant, il peut être constaté que de façon générale, à effecteur identique, CRISPR/Cas va être plus efficace que les TALE pour induire une répression mais moins pour induire une activation, en possible lien avec l'encombrement stérique associé au module de liaison à l'ADN employé (X. Liu et al. 2016).

### 2.3.3 Édition du profil de méthylation de l'ADN

En combinant l'un des systèmes de guidage mentionnés plus haut (ZFP, TALE ou dCas9) avec un effecteur de type DNA déméthylase ou DNA méthyltransférase, il est possible d'aller altérer spécifiquement le patron de méthylation d'un locus donné. La TABLE 2.1 répertorie les différentes approches publiées à ce jour. La majorité des travaux ont été effectués dans des cellules de mammifères en culture, et utilisent comme effecteur le domaine catalytique de la méthyltransférase *de novo* des mammifères Dnmt3a, seul ou en fusion avec Dnmt3L, une protéine de la famille des Dnmt3 dépourvue d'activité catalytique mais qui stimule celle de Dnmt3a (Chédin et al. 2002). En raison de leur taille réduite et de leur organisation simple, des méthyltransférases bactériennes ont également pu être employées.

Pour la déméthylation ciblée, les domaines catalytiques des enzymes de la famille TET, responsables de l'oxydation des 5mC préalable à leur excision par la glycosylase TDG chez les mammifères, ont été privilégiés, et ce même chez Arabidopsis (Gallego-Bartolomé et

TABLE 2.1 – Stratégies d'altération locus-spécifique du patron de méthylation publiées à ce jour

| **Type de changement de méthylation** | **Module de guidage** | **Module effecteur** | **Modèle** | **Réference** |
|---|---|---|---|---|
| Méthylation | GAL4/UAS | DNMT3B-CD | Cellules humaines en culture | (F. Li et al. 2007) |
| | GAL4/UAS | DNMT3A-CD | Cellules humaines en culture | (F. Li et al. 2007) |
| | ZFP | M.EcoHK31I | *E. coli* | (Meister et al. 2010) |
| | ZFP | M.HhaI | *E. coli* | (Chaikind, Kilambi et al. 2012) |
| | ZFP | M.SssI | *E. coli* | (Chaikind et Ostermeier 2014) |
| | ZFP | DNMT3B-CD | Cellules humaines en culture | (F. Li et al. 2007) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (F. Li et al. 2007) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (Cui et al. 2015) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (Kungulovski et al. 2015) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (Nunna et al. 2014) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (Rivenbark et al. 2012) |
| | ZFP | DNMT3A-CD | Cellules humaines en culture | (Stolzenburg et al. 2015) |
| | ZFP | DNMT3A-3Lsc | Cellules humaines en culture | (Siddique et al. 2013) |
| | ZFP | SUVH9 | *A. thaliana* | (Johnson et al. 2014) |
| | TALE | M.SssI | Cellules murines en culture | (Yamazaki et al. 2017) |
| | TALE | DNMT3A-CD | Cellules humaines en culture | (Amabile et al. 2016) |
| | TALE | DNMT3A-CD | Cellules murines en culture | (Lo et al. 2017) |
| | TALE | DNMT3A-3Lsc | Cellules humaines en culture | (Bernstein et al. 2015) |
| | TALE | DNMT3A-3Lsc | Cellules humaines en culture | (Mlambo et al. 2018) |
| | dCas9 | M.SssI | Cellules humaines en culture | (Xiong et al. 2015) |
| | dCas9 | MQ1 | Cellules humaines en culture et embryons de souris | (Lei et al. 2017) |
| | dCas9 | DNMT3A | Cellules murines en culture | (X. Liu et al. 2016) |
| | dCas9 | DNMT3A-CD | Cellules humaines en culture | (Vojta et al. 2016) |
| | dCas9 | DNMT3A-CD | Cellules humaines en culture | (McDonald et al. 2016) |
| | dCas9 | DNMT3A-CD | Cellules humaines en culture | (Amabile et al. 2016) |
| | dCas9 | DNMT3A-CD | Cellules humaines et murines en culture | (Galonska et al. 2018) |
| | dCas9 | DNMT3A-3Lsc | Cellules humaines en culture | (Stepper et al. 2017) |
| | dCas9-SunTag | DNMT3A-CD | Cellules humaines en culture | (Y.-H. Huang et al. 2017) |
| | dCas9-SunTag | DNMT3A-CD | Cellules humaines en culture | (Pflueger et al. 2018) |
| Déméthylation | GAL4/UAS | ROS1-CD | Cellules humaines en culture | (Parrilla-Doblas et al. 2017) |
| | ZFP | TDG | Cellules murines en culture | (Gregory et al. 2013) |
| | ZFP | TET2-CD | Cellules humaines en culture | (H. Chen et al. 2014) |
| | ZFP | TET1-CD | *A. thaliana* | (Gallego-Bartolomé et al. 2018) |
| | TALE | TET1-CD | Cellules humaines en culture | (Maeder et al. 2013) |
| | TALE | TET1-CD | Cellules murines en culture | (Lo et al. 2017) |
| | TALE | TET1-CD | Cellules humaines en culture | (Amabile et al. 2016) |
| | dCas9 | TET1-CD | Cellules humaines en culture | (Amabile et al. 2016) |
| | dCas9 | TET1-CD | Cellules murines en culture | (C. Liu et al. 2016) |
| | dCas9 | TET1-CD | Cellules humaines en culture | (Choudhury et al. 2016) |
| | dCas9 | TET1-CD | Cellules murines en culture | (Okada et al. 2017) |
| | dCas9 | TET3-CD | Cellules humaines et murines en culture | (Xu, X. Tan et al. 2018) |
| | dCas9-SunTag | TET1-CD | Cellules humaines en culture et embryons de souris | (Morita et al. 2016) |
| | dCas9-SunTag | TET1-CD | *A. thaliana* | (Gallego-Bartolomé et al. 2018) |
| | dCas9 /sgRNA-MS2 | TET1-CD | Cellules humaines en culture | (Xu, Tao et al. 2016) |

al. 2018) en dépit des différences mécanistiques dans le processus de déméthylation active. D'autre part, l'utilisation de la déméthylase des plantess ROS1 n'a été tentée qu'une seule fois, dans le cadre d'une expression transitoire au moyen du système UAS/GAL4 (Parrilla-Doblas et al. 2017). Dans leur ensemble, les fusions présentées TABLE 2.1 sont parvenues à altérer le profil de méthylation du locus ciblé ainsi que l'expression du gène à proximité, même si le degré de (dé)méthylation induit ne se reflète pas nécessairement dans le profil d'expression du gène (X. Liu et al. 2016 ; McDonald et al. 2016). Une question également en suspens est celle de la persistance de l'état de méthylation induit après élimination du transgène ; un point qui n'est que peu documenté et qui le cas échéant donne lieu à des résultats variables d'un locus et d'un système à l'autre. Par ailleurs, des effets off-target, en l'occurrence des gains de méthylation ectopiques et une réduction du niveau global de méthylation, ont également été documentés (Gallego-Bartolomé et al. 2018 ; Galonska et al. 2018), pointant la nécessité d'ajouter une condition contrôle dans les approches expérimentales de (dé)méthylation ciblée.

En outre, la description d'une (dé)méthylation ciblée n'a été documentée que deux fois chez les plantes, et exclusivement chez Arabidopsis, et l'une des deux descriptions reprend par ailleurs une fusion déjà publiée en système mammifère (Gallego-Bartolomé et al. 2018 ; Morita et al. 2016). La première démonstration d'une édition du profil de méthylation à un locus donné a été effectuée au moyen d'une fusion entre SUVH9, une protéine de la voie du RdDM, et une ZFP dirigée contre le locus *FWA* (Johnson et al. 2014). Il a pu être montré que cette fusion suffisait à recruter la machinerie du RdDM au locus visé et que cette méthylation ciblée est à l'origine d'une réversion phénotypique (voir CHAPITRE 3 pour la description de *FWA*). Néanmoins, il faut garder à l'esprit que le locus testé était certes dépourvu de méthylation dans le mutant utilisé, mais qu'il est toujours source de siRNA, ce qui constitue une situation très particulière. Aussi, il n'est en rien garanti que ce résultat puisse être reproduit à un locus plus standard. La déméthylation de ce même locus a ensuite été décrite au moyen d'une fusion dCas9-TET1-CD, avec maintien de l'état hypométhylé après ségrégation du transgène, néanmoins ce maintien n'a pas été observé au second locus étudié, pointant une fois encore une susceptibilité variable d'un locus à l'autre quant à l'altération du patron de méthylation (Gallego-Bartolomé et al. 2018).

Par ailleurs, il faut souligner qu'au moins chez les plantes, en raison d'une interconnexion entre le RdDM décrit précédemment et les machineries d'ARN interférence (RNAi), il est possible d'aller altérer le patron de méthylation de façon locus spécifique au moyen des stratégies RNAi classiques que sont l'emploi de transgènes *hairpin* (illustré FIGURE 2.14) et le VIGS (*virus induced gene silencing*).

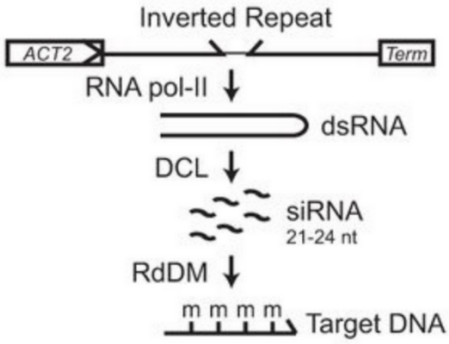


FIGURE 2.14 – Principe de la stratégie *hairpin*. Les "m" indiquent la méthylation induite.

Si la fonction première du RNAi est le PTGS (*post-transcriptional gene silencing*), donc la dégradation du transcrit et non pas le TGS (*transcriptional gene silencing*), donc l'extinction transcriptionnelle, ces deux voies sont reliées sur un plan mécanistique et il a pu être constaté que le ciblage par les outils d'ARN interférence d'un locus pouvait conduire à l'inactivation transcriptionnelle par méthylation de ce dernier (Dalakouras et al. 2009 ; Kanazawa et al. 2011 ; Mette et al. 2000 ; Sijen et al. 2001).
L'une et l'autre de ces approches sont fondées sur l'emploi d'un transgène de séquence identique à celle du locus à cibler, qui va être ou organisée sous forme de répétitions inversées afin de créer un transcrit en "épingle à cheveux" dont la partie double-brin va être employée comme matrice pour former des siRNA et conduire à un PTGS dans le cas de l'approche *hairpin*, ou bien intégrée à un vecteur viral qui va être reconnu par la machinerie de défense de la plante et ciblé par le PTGS dans le cas du VIGS. L'efficacité de ces deux approches a pu être démontrée là encore au locus *FWA* (Bond et al. 2015 ; Kinoshita et al. 2007), avec dans ce second cas le maintien de l'état reméthylé après élimination du transgène.
Cependant, les observations faites à des loci endogènes moins spécifiques que *FWA* donnent des résultats plus contrastés (Cigan et al. 2005 ; Melquist et al. 2003, 2004 ; Yelina et al. 2015) qui dans leur ensemble suggèrent que l'initiation de la méthylation n'est pas systématique, et que plusieurs générations sont requises pour atteindre un niveau de méthylation suffisant pour être perpétué au travers des générations, en lien avec un renforcement du RdDM au cours de la phase reproductive (F. Teixeira et al. 2010).
Au-delà de la susceptibilité individuelle des différents loci ciblés, il apparaît également que le design du transgène soit critique (Dadami et al. 2014), tout comme dans le cas des *hairpin* la capacité d'intégration dans le génome (Sunitha et al. 2012).
Aussi, ces stratégies ne sont à ce jour pas satisfaisantes dans le cadre d'une méthylation ciblée à l'envi.

Chapitre 3

# L'épigénome, un support additionnel de variation héritable

Nous avons pour le moment principalement évoqué les variants ADN, que leur apparition soit ou non influencée par les différents états chromatiniens, comme source ultime de différences génétiques entre individus. En effet, bien que la variation phénotypique héritable soit en premier lieu dépendante des variations dans la séquence de l'ADN entre individus, des travaux récents indiquent que d'autres facteurs, et notamment la méthylation de l'ADN, peuvent également y prendre part. Cela est particulièrement vrai chez les plantes, où des variants dans les profils de méthylation, appelés "épiallèles", peuvent être transmis au travers des générations et contribuer pour partie à la variation phénotypique héritable, c'est à dire la fraction de la variation observée pour des traits complexes (comme la taille de la plante ou la date de floraison) qui est expliquée par les allèles (et épiallèles) dont l'individu a hérité.

Je présenterai l'état de l'art concernant la variation épiallélique chez Arabidopsis, et discuterai les approches expérimentales permettant d'évaluer la contribution de ces épimutations à la variation phénotypique héritable. Cette section est par ailleurs complétée par une revue à destination du grand public portant sur l'épigénétique transgénérationnelle.

## 3.1 Nature et cause des "épimutations"

### 3.1.1 Définition et propriétés

Les épimutations peuvent être définies comme des changements héritables de l'épigénome à un locus donné, en l'absence de modifications dans la séquence de l'ADN (Richards 2006). Deux points importants sont à noter dans cette définition : d'une part elle ne présuppose en rien une altération du profil d'expression des gènes et encore moins une différence phénotypique ; d'autre part elle recouvre aussi bien des altérations portées par des modifications post-traductionnelles d'histones (ou encore des variants d'histones ou des petits ARN) que des différences dans le profil de méthylation de l'ADN. Je me focaliserai ici sur ce second cas de figure, mais il faut noter que des épimutations ont pu être identifiées chez des eucaryotes dépourvus de méthylation de l'ADN, par exemple chez le Nématode *C.elegans* (Greer et al. 2011).

En fonction de la taille de la région affectée, il est possible de distinguer les épimutations affectant une position -en l'occurrence, une cytosine- de celles affectant une région ; on parle alors respectivement de DMP (*differentially methylated position*) ou de DMR (*differentially methylated region*). En intégrant l'information du contexte de méthylation, on peut dès lors distinguer des CG-, CHG- ou CHH- DMP ou DMR, mais également de C-DMR lorsque les trois contextes sont affectés.

Les épimutations présentent par ailleurs une propriété qui leur est très spécifique : au contraire des mutations dans la séquence de l'ADN, qui lorsqu'elles sont créées sont transmises selon les lois de Mendel au travers des divisions cellulaires, les épimutations ont la capacité de réverter en mitose ou en méiose. Cette réversion, particulièrement flagrante lorsque l'épimutation est à effet phénotypique, peut être totale (retour au phénotype initial) ou incomplète (apparition de phénotypes intermédiaires), et sera dans ce dernier cas associée à un niveau de méthylation intermédiaire.

De même que pour les mutations, plusieurs approches permettent d'évaluer le taux d'épimutation, ou plus exactement les taux de formation (*forward*) et de réversion (*backward*) des épimutations : fréquence d'apparition/réversion d'un phénotype (Miura et al. 2009 ; Quadrana et Colot 2016), lignées MA ou équivalents (Becker et al. 2011 ; Graaf et al. 2015 ; R. Schmitz, Y. He et al. 2013 ; R. Schmitz, Schultz, Lewsey et al. 2011) mais aussi polymorphisme de méthylation dans les populations naturelles (Hagmann et al. 2015 ; R. Schmitz, Schultz, Urich et al. 2013). Ces différents travaux ont permis de donner un aperçu de taux et du spectre d'épimutations. Les taux d'épimutation obtenus au moyen d'approches MA lines chez Arabidopsis sont donnés Table 3.1. On constate que si le taux d'apparition de DMR par génération est du même ordre de grandeur que le taux de mutation d'Arabidopsis (environ 1 SBS par génome par génération, Chapitre 1), le taux d'apparition de DMP est près de 1000 fois supérieur, au moins pour les CG-DMP pour lesquels une estimation fiable est disponible.
Il peut également être observé que les CG-DMR et C-DMR sont retrouvées dans les corps des gènes soumis à la gBM et dans les ET respectivement, ces dernières pouvant influencer le niveau d'expression des gènes proximaux. A l'opposé, les CG-DMP sont retrouvés préférentiellement au niveau des gènes éloignés des ET, d'où l'hypothèse que ce taux élevé résulte de l'absence d'une maintenance robuste des patrons de méthylation dans ces régions (Becker et al. 2011 ; R. Schmitz, Schultz, Lewsey et al. 2011).
Les conclusions qui découlent de l'analyse détaillée des taux de formation et de réversion des CG-DMP soutiennent par ailleurs cette hypothèse : le ratio des taux de formation et de réversion diffère d'une annotation génomique à l'autre, avec des taux *forward* équivalents entre gènes et ET mais un taux *reverse* plus élevé dans les gènes que dans les ET, ces derniers présentant une maintenance robuste de la méthylation (Graaf et al. 2015).

TABLE 3.1 – Taux d'épimutation chez Arabidopsis obtenus par analyse de données bisulphite de MA lines (Becker et al. 2011 ; Graaf et al. 2015 ; R. Schmitz, Schultz, Lewsey et al. 2011). Le taux de DMP non-CG n'est pas figuré faute d'une estimation robuste à ce jour, les différences de méthylation dans les contextes CHG et CHH étant statistiquement plus difficiles à identifier. Adapté de (Quadrana et Colot 2016).

| | Taux estimé |
|---|---|
| CG-DMP | 1000/génération |
| CG-DMP (*forward*) | $2{,}56.10^{-4}$/site CG/génération |
| CG-DMP (*reverse*) | $6{,}30.10^{-4}$/site CG/génération |
| C-DMR | 0,34/génération |
| CG-DMR | 0,8-1,3/génération |

Cela conforte par ailleurs l'hypothèse plus générale selon laquelle les épimutations, qu'il s'agisse de DMP ou de DMR, résultent en premier lieu d'une maintenance incorrecte de la méthylation aux loci concernés (Quadrana et Colot 2016 ; R. J. Schmitz et al. 2012).

### 3.1.2 Dépendance au génotype

Une autre observation qui peut être faite, et ce particulièrement au bénéfice des méthylomes d'accessions naturelles, est que certains variants de méthylation sont identifiés exclusivement dans un génotype donné, tandis que d'autres sont pervasifs. En fonction de cette dépendance au génotype, une classification des épiallèles en trois classes (obligatoires, facilités et purs ; FIGURE 3.1) a été proposée (Richards 2006).



FIGURE 3.1 – Classification des épiallèles en fonction de leur dépendance au génotype, d'après (Richards 2006). De gauche à droite, épiallèles obligatoire, facilité et pur. x et y représentent des variants ADN constituant des génotypes distincts, les rectangles blancs et rouges deux épigénotypes. Le basculement d'une forme épiallélique à une autre est illustrée par le passage d'un rectangle blanc à un rouge. Extrait de (Richards 2006). Figure non reproduite dans la version de diffusion de la thèse.

Ainsi, les épiallèles obligatoires sont intégralement dépendants de la présence d'un variant ADN donné, tandis que les épiallèles facilités sont induits initialement par la présence d'un variant ADN, mais peuvent persister en son absence.
Par exemple, dans le cas d'un gain de méthylation dans le promoteur d'un gène du fait

de sa proximité avec un ET, la disparition ou à l'opposé la persistance de ce gain de méthylation après élimination de l'ET par ségrégation traduit respectivement le caractère obligatoire ou facilité de cette épimutation.
A l'opposé, les épiallèles dits purs sont indépendants de toute variation de la séquence de l'ADN, et peuvent donc apparaître dans n'importe quel génotype.
Si cette distinction théorique a le bénéfice de la clarté, déterminer la classe d'un épiallèle donné est en réalité complexe, particulièrement dans les populations naturelles. En effet, le patron de méthylation d'un locus peut être influencé par des meQTL (QTL de méthylation), en *cis* (présence/absence d'un variant ADN à ce locus) mais également en *trans* (polymorphisme dans l'un des composants des voies de méthylation, à l'exemple du variant de *VIM1* dans l'accession Bor-4 d'Arabidopsis (Woo et al. 2007)). Ce critère va prendre toute son importance lorsqu'il va s'agir de démontrer la causalité d'un épiallèle potentiellement causatif pour un phénotype, puisque les épiallèles "obligatoires" ne sont en rien épigénétiques.
En l'espèce, l'analyse des données de variation naturelle dans les profils des méthylation a conduit à la mise en évidence que la majorité des différences de méthylation observées entre individus ont une base génétique (Dubin et al. 2015 ; Kawakatsu, S.-S. C. Huang et al. 2016 ; Quadrana, Bortolini Silveira et al. 2016 ; R. Schmitz, Schultz, Urich et al. 2013).

### 3.1.3 Epimutations naturelles et induites

Compte-tenu des progrès techniques dans le champ de l'épigénomique, il est à présent bien plus facile d'identifier de nouveaux épiallèles, d'où un acroissement récent du nombre d'épimutations décrites. Les Tables 3.2 et 3.3 référencent différentes épimutations, naturelles et induites, à effet phénotypique décrites à ce jour chez les plantes.
Qu'il s'agisse d'épimutations naturelles -survenues spontanément dans des individus de génotype sauvage- ou d'épimutations induites -survenues dans un contexte particulier (mutant de méthylation ou propagation somaclonale) et qui persistent après rétablissement des conditions initiales (en ce sens, elles sont épigénétiques)- ; les mêmes caractéristiques sont observées : la majorité d'entre elles sont associées à la présence de séquences répétées (ET ou répétition inversée), et peuvent être qualifiées de "métastables", dans la mesure où elles alternent entre un état méthylé et non-méthylé.

A l'instar de *QQS* (Tables 3.2), il est possible d'avoir également des épimutations "pures", au sens où elles ne sont pas associées à des séquences répétées et sont apparues dans des accessions non apparentées. Cependant, de tels évènements sont rares en comparaison des épimutations "facilitées" ou "obligatoires" qui constituent la majorité des deux tables.

Parmi les épimutations induites, *FWA* (FIGURE 3.2) est l'une des mieux connues à ce jour. Le gène *FWA* est réprimé dans l'ensemble de la plante et exprimé exclusivement dans l'albumen (tissu nourricier de l'embryon) de la graine, où il est soumis à l'empreinte parentale avec l'expression du seul allèle maternel (Kinoshita et al. 2007). Cette expression spécifique résulte de la déméthylation par DME des répétitions dérivées d'un SINE localisés en 5' du gène.
L'épimutation *fwa* se caractérise par une perte de méthylation dans l'ensemble des tissus de la plante, d'où une expression ectopique de *FWA* qui conduit à un retard drastique de floraison. Si cette épimutation n'a à ce jour jamais été observée dans des populations naturelles probablement en raison d'une contre-sélection de l'allèle hypométhylé, elle apparaît sporadiquement chez des mutants globalement hypométhylés comme *met1* et *ddm1* et une aggravation du phénotype associée à une perte de méthylation de plus en plus importante à ce locus peut être décrite au cours des générations successives d'autofécondation (Kakutani 1997 ; Kakutani, Jeddeloh, Flowers et al. 1996). Cette hypométhylation (et le phénotype qui en découle) est alors maintenue au travers des générations même après restauration de la fonction des gènes mutés, alors qu'une réversion est possible dans les générations précoces (Kakutani 1997).

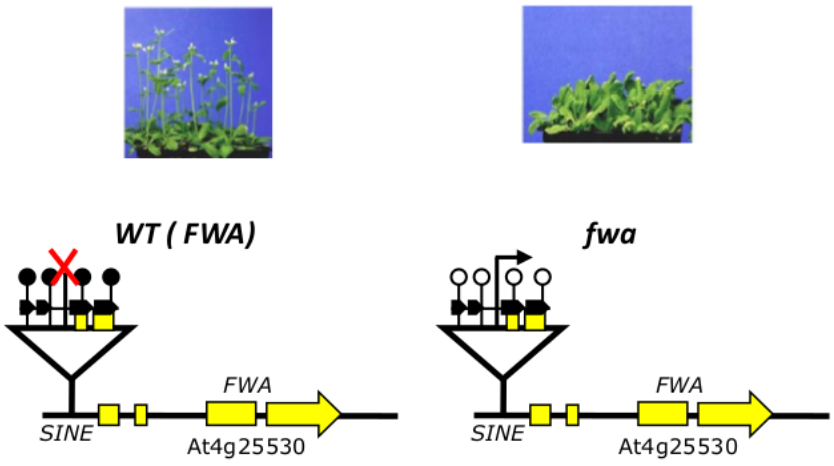


FIGURE 3.2 – *FWA*, un exemple d'épimutation induite. L'état de méthylation des cytosines est figurée par des ronds noirs pleins (méthylé) ou vides (hypométhylé). En contexte sauvage (panel de gauche, épiallèle *FWA*), les répétitions dérivées d'un SINE (triangle) en amont du promoteur du gène sont méthylées dans les tissus végétatifs. L'épimutation *fwa* (panel de droite) se traduit par une perte de méthylation généralisée de ces séquences répétées, laquelle induit l'expression ubiquite du gène et un retard de floraison.

TABLE 3.2 – Epimutations naturelles à effet phénotypique décrites chez les plantes.*cnr* et *lcyc* sont illustrés dans la revue grand public jointe à ce chapitre.

| **Espèce** | **Nom** | **Gène affecté** | **Phénotype** | **Type de changement de méthylation** | **Association à une séquence répétée** | **Stabilité** | **Références** |
|---|---|---|---|---|---|---|---|
| *Solanum lycopersicum* | *VTE* | *VTE3(1)* | Contenu en vitamine E | Gain de méthylation dans la région promotrice | Oui (SINE en 5') | Métastable | (Quadrana, Almeida et al. 2014) |
| | *cnr* | *LeSPL-CNR* | Altération du mûrissement du fruit | Gain de méthylation dans la région promotrice | Oui | Métastable | (Manning et al. 2006) |
| *Oryza sativa* | *Epi-d1* | *DWARF1* | Taille réduite | Gain de méthylation dans la région promotrice | Oui | Métastable | (Miura et al. 2009) |
| | *Epi-rav6* | *RAV6* | Angle de la feuille, taille des graines | Perte de méthylation dans la région promotrice | Oui (MITE en 5') | - | (X. Zhang, Sun et al. 2015) |
| *Linaria vulgaris* | *peloric* | *lcyc* | Altération de la symétrie florale | Gain de méthylation dans la région promotrice | - | Métastable | (Cubas et al. 1999) |
| *Cucumis melo* | - | *CmWIP1* | Altération du sexe des fleurs | Gain de méthylation dans la région promotrice | Oui | Métastable | (Martin, Troadec et al. 2009) |
| *Arabidopsis thaliana* | - | *PAI* | Déficience de biosynthèse du tryptophane | Gain de méthylation pour les quatre gènes de la famille multigénique | Oui (réarrangement aboutissant à la répétition en tandem de deux gènes) | - | (Luff et al. 1999) |
| | - | *QQS* | Biosynthèse de l'amidon | Gain de méthylation dans la région promotrice | Oui (répétitions en tandem dans le promoteur et le 5'UTR) | - | (Silveira et al. 2013) |
| | *NMR19-4* | *PPH* | Sénescense des feuilles | Gain de méthylation dans la région promotrice | Oui | - | (L. He et al. 2018) |
| | - | *AtFOLT1* | Incompatibilité allélique Col-0 - Shadara | Gain de méthylation dans la région promotrice | Oui (réarrangement aboutissant à une répétition en tandem) | Métastable | (Durand et al. 2012) |
| | - | *TAD3* | Incompatibilité allélique Col-0 - Nok-1 | Gain de méthylation dans la région promotrice et dans le corps du gène | Oui | Métastable | (Agorio et al. 2017) |
| | - | *HISN6* | Incompatibilité allélique Col-0 - Cvi | Gain de méthylation dans la région promotrice | - | Métastable | (Blevins et al. 2017) |

TABLE 3.3 – Epimutations induites à effet phénotypique décrites chez les plantes

| **Espèce** | **Nom** | **Gène affecté** | **Induction** | **Phénotype** | **Type de changement de méthylation** | **Association à une séquence répétée** | **Stabilité** | **Références** |
|---|---|---|---|---|---|---|---|---|
| *Arabidopsis thaliana* | *FWA* | *FWA* | *ddm1*, *met1* | Floraison tardive | Hypométhylation des séquences répétées en 5' du gène | Oui (SINEs dégénérés en tandem en 5' du gène) | Métastable | (Kakutani 1997 ; Lippman et al. 2004) (Kinoshita et al. 2007 ; Soppe et al. 2000) |
| | *SUPERMAN* | - | *met1* | Augmentation du nombre d'étamines | Hyperméthylation du gène | - | Métastable | (Jacobsen et Meyerowitz 1997) |
| | *AGAMOUS* | - | *met1* | Perte des étamines et du carpelle | Hyperméthylation du 1er intron du gène | - | Métastable | (Jacobsen, Sakai et al. 2000) |
| | *BONSAI* | - | *ddm1* | nanisme | Hyperméthylation du gène | Oui (LINE en 3') | Métastable | (Saze et al. 2007) |
| | *SDC* | - | *drm1 drm2 cmt3* | Taille réduite, surenroulement des feuilles | Hypométhylation des séquences répétées en 5' du gène | Oui | Métastable | (Henderson et al. 2008) |
| *Elaeis guineensis* | *KARMA* | *MANTLED* | somaclonale | Transformation homéotique et parthénocarpie, fruit impropre à l'utilisation. | Hypométhylation d'un ET situé dans un intron du gène, conduisant à un épissage alternatif et une isoforme tronquée | Oui (insertion d'un LINE dans un intron) | Métastable | (Ong-Abdullah et al. 2015) |

## 3.2 Des épimutations à la variation épigénétique héritable

Le fait que des épimutations puissent être transmises au travers des générations et avoir des effets phénotypiques ouvre la possibilité que de tels variants puissent contribuer à l'héritabilité de traits complexes.

### 3.2.1 Les plantes, un modèle d'épigénétique transgénérationnelle

Les paragraphes précédents se sont restreints aux épimutations identifiées chez les plantes, et la question des mammifères n'a pas été abordée. De telles épimutations ont pu être identifiées chez ces organismes (voir encadré dédié dans l'annexe à ce chapitre), et si elles présentent les mêmes propriétés que celles décrites chez les plantes (association à des séquences répétées et métastabilité), seul un nombre très restreint a pu être décrit. Une explication à cela est qu'au cours de la gamétogénèse puis de l'embryogénèse précoce, les épigénomes des mammifères font l'objet d'une reprogrammation drastique, ce qui limite grandement la possibilité que des épimutations puissent être transmises à la génération suivante.
A l'opposé, il n'y pas chez les plantes de véritable reprogrammation : le niveau de méthylation des sites CHG et surtout des sites CHH va certes varier mais le contexte CG reste non altéré et les ET maintiennent leur niveau de méthylation élevé au cours du passage à la génération suivante, ce qui autorise une transmission de variants de méthylation au travers des générations successives.
En ce sens, au contraire des mammifères, une hérédité épigénétique transgénérationnelle pervasive est tout à fait envisageable chez les plantes (Heard et al. 2014).

Dans ce contexte où une variation phénotypique héritable peut exister en l'absence de différences dans la séquence de l'ADN, l'hypothèse que des changements induits par l'environnement puissent être transmis aux générations suivantes est particulièrement attrayante, avec l'idée sous-jacente qu'un tel système permettrait aux plantes de générer rapidement de la variation adaptative héritable. Néanmoins, l'ensemble des observations faites à ce jour avec des design expérimentaux rigoureux mettent en évidence que si un stress, par exemple salin ou thermique, peut induire effectivement des altérations dans le patron de méthylation de l'ADN, celles-ci ne seront pas transmises au fil des générations et les effets phénotypiques d'"adaptation", s'ils existent, ne persistent pas au-delà de la première génération. Ainsi donc, il s'agit là d'effets parentaux, majoritairement d'effets maternels, et nullement d'effets transgénérationnels (Secco et al. 2015 ; Van Dooren et al. 2018 ; Wibowo et al. 2016). En ce sens, sans que l'ensemble des détails soit connu à ce

jour, plusieurs mécanismes ont été proposés qui permettent d'effectuer un resetting des effets du stress sur l'épigénome (Baubec et al. 2014 ; Crevillén et al. 2014 ; Iwasaki et al. 2014).

### 3.2.2 Apport des populations de lignées epiRIL

Ces observations concernant les effets phénotypiques et la stabilité des épimutations ouvrent tout un champ de recherches concernant la possible contribution de tels variants à la variation héritable pour une diversité de traits complexes.

L'identification des déterminants génétiques des traits complexes se fait au moyen de deux stratégies (Bazakos et al. 2017) : d'une part, l'analyse de liaison, qui emploie des population expérimentales, et d'autre part les études d'association (GWAS, *genome-wide association studies*), qui tirent profit de la variation naturelle entre accessions. Dans un cas comme dans l'autre, il va s'agir d'identifier des QTL, c'est à dire des régions du génome qui contribuent pour partie à la variation héritable d'un caractère quantitatif. La présence d'un QTL est détectée par la mise en évidence d'une corrélation statistiquement significative entre valeur phénotypique et ségrégation d'un ou plusieurs marqueurs génétiques ; la "probabilité" d'observer un QTL dans un intervalle compris entre deux marqueurs est donnée par le LOD score (plus ce score est important, plus la vraisemblance d'avoir un QTL dans cet intervalle est forte).

Dans le cas de l'analyse de liaison, il s'agit de phénotyper pour différents traits des individus issus de populations en ségrégation, les QTL étant détectés par la co-ségrégation génotype-phénotype. Plusieurs types de populations expérimentales existent, l'une des plus fréquentes étant les dispositifs RIL (*recombinant inbred lines*) : à partir du croisement entre deux accessions, des individus F2 vont être propagés par autofécondation sur au moins quatre générations afin d'atteindre un degré élevé d'homozygotie. Chaque RIL va se caractériser par une importante homozygotie et constitue une mosaïque unique des deux génomes parentaux. Si l'analyse de liaison permet la détection de variants rares, une contrainte importante vient du fait que les QTL sont localisés dans des intervalles de grande taille, bien souvent de l'ordre du Mb, ce qui va requérir toute une procédure de cartographie fine pour identifier le variant causal. Afin d'augmenter la résolution (donc le nombre de recombinaisons) ainsi que la diversité allélique en ségrégation, des populations au schéma de croisement plus complexe ont été établies, à l'exemple des populations MAGIC (*Multi-parent Advanced Generation Inter-Cross*) qui dérivent de croisements multiples entre plusieurs accessions avant d'initier les cycles d'autofécondation qui permettront de créer des RIL.

En comparaison à une analyse de liaison, l'approche GWAS offre une bien meilleure résolution : en raison du grand nombre de générations qui distinguent les différentes accessions en comparaison aux différentes RIL d'une population, un grand nombre d'évènements de recombinaison se sont produits, et il est dès lors possible de réduire les intervalles de confiance autour d'un QTL à quelques kb. Néanmoins, un inconvénient notable de ce type d'approche a à voir avec les effets confondants associés au fond génétique des populations analysées : pour peu que des groupes d'individus présentent un fort apparentement au sein d'une population (donc une structuration de la population), un déséquilibre de liaison "parasite" peut apparaître entre le variant causal du QTL et des variants situés ailleurs dans le génome, ce qui va brouiller le signal d'association phénotype-génotype.

Si ces deux approches sont généralement complémentaires pour l'identification des variants ADN qui sous-tendent la variation héritable pour les traits complexes, elles présentent l'une et l'autre un écueil important dès lors que l'on s'intéresse aux épivariants : les polymorphismes dans les profils de méthylation d'une part et dans la séquence de l'ADN d'autre part sont entremêlés, aussi il est complexe de déterminer le niveau de dépendance au génotype des différences de méthylation observées et donc d'évaluer la contribution effective des seules différences de méthylation au phénotype d'intérêt.

La nécessité de s'affranchir des effets confondants de la variation nucléotidique pour ne prendre en compte que la seule variation épigénétique constitue le fondement théorique de l'établissement de populations dites de lignées epiRIL (*epigenetic recombinant inbred lines*), par analogie aux populations RIL (Johannes, Colot et al. 2008). Dans cette optique de minimiser la variation nucléotidique et de maximiser la variation épigénétique, deux populations epiRIL ont été développées en parallèle chez Arabidopsis, à partir des mutants de méthylation *met1* et *ddm1*.

Les epiRIL *met1* dérivent du croisement entre le mutant *met1-3* et un individu sauvage isogénique : après autofécondation de la F1 issue de ce croisement, les individus de génotype sauvage au locus *MET1* sont sélectionnés et propagés par autofécondation sur 8 générations successives (Reinders et al. 2009). Compte-tenu d'une importante mortalité (28%) au cours de l'établissement de la population, celle-ci ne comporte au final que 68 lignées, un taille peu compatible avec des approches de génétique quantitative. Par ailleurs, si une variation phénotypique a pu être décrite pour plusieurs traits dans cette population, l'important nombre de variations non-parentales de méthylation détectées au cours de l'analyse des méthylomes de quelques lignées ainsi que la contribution de nouvelles insertions d'ET aux phénotypes observés ne vont pas dans le sens d'une variation causée par la ségrégation d'épiallèles parentaux (Mirouze et al. 2009 ; Reinders et al. 2009).

Les epiRIL *ddm1* dérivent du croisement entre le mutant *ddm1-2* et un individu sauvage isogénique : après backcross de la F1 avec un individu sauvage issu du même lot de graines que le parent employé au cours du premier croisement, les individus *DDM1/DDM1* sont sélectionnés et propagés par filiation monograine sur 6 générations successives (FIGURE 3.3). Seule une faible mortalité (0.8%) a pu être constatée au cours de la propagation des lignées, et la population finale est constituée de 505 epiRIL.

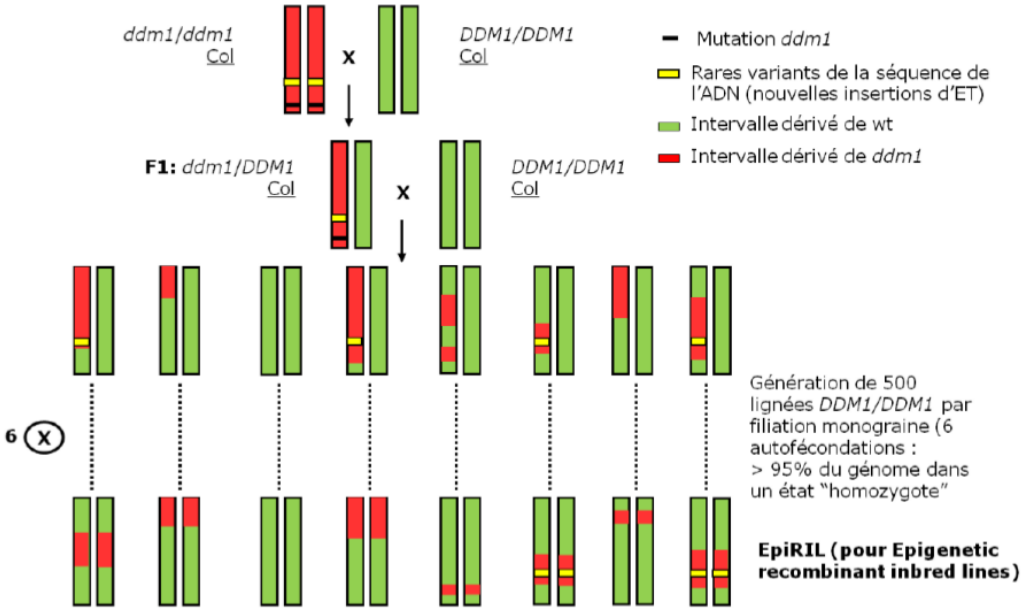


FIGURE 3.3 – Schéma d'obtention de la population epiRIL. Détails dans le texte

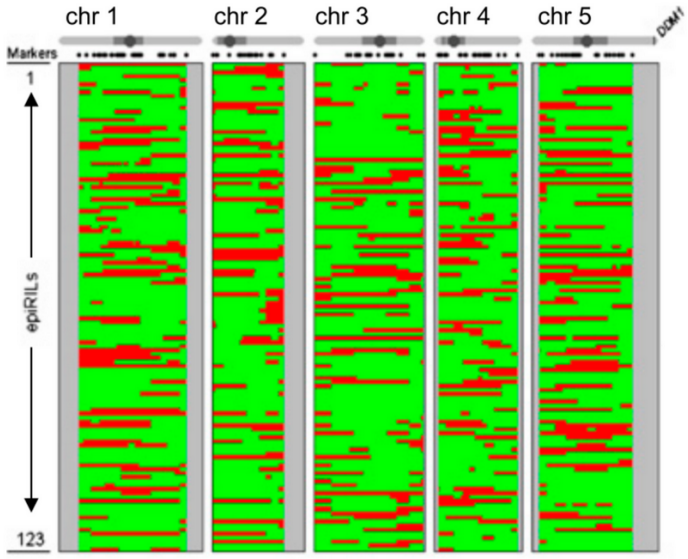


FIGURE 3.4 – Epihaplotypes des 123 epiRIL épigénotypées. Le vert et le rouge désignent respectivement les régions d'origine wt et *ddm1*. Les positions des 126 marqueurs employés pour construire la carte génétique sont représentées par des points noir sous les schémas des chromosomes. Chaque ligne représente l'épihaplotype d'une epiRIL. Extrait de (Colomé-Tatché et al. 2012).

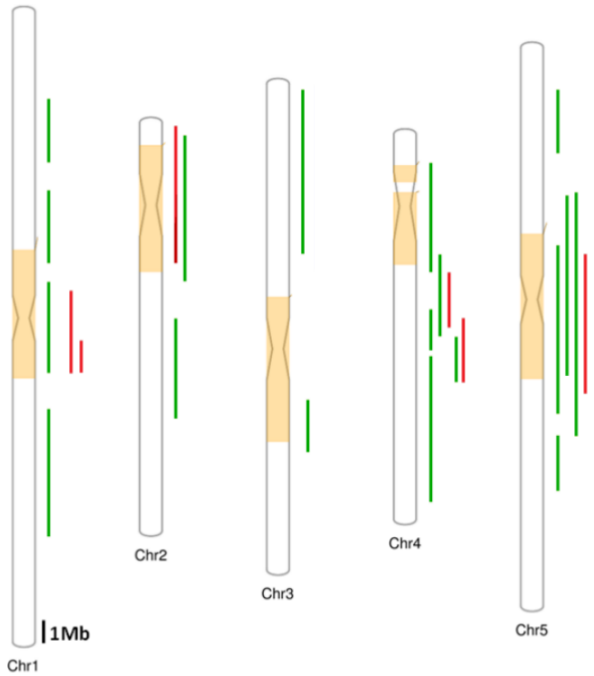


FIGURE 3.5 – Représentation des différents QTL "épigénétiques" publiés à ce jour. Les intervalles de support des 6 et 18 QTLepi décrits dans (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) et (Kooke et al. 2015) sont figurés en rouge et vert respectivement. On notera la taille importante des intervalles, ainsi que la co-localisation de plusieurs QTLepi qui laisse suggèrer des effets pléiotropes.

Il a pu être mis en évidence que cette population présente une variation continue (ce qui est caractéristique d'un trait quantitatif), de faible amplitude mais néanmoins en grande partie héritable pour plusieurs traits complexes, ce qui suggère la ségrégation dans la descendance de variants d'origine parentale (Johannes, Porcher et al. 2009). Il a par la suite été établi que près d'un tiers des 3000 régions différentiellement méthylées (DMR) identifiées entre les 2 parents étaient transmises de façon stable sur au moins 8 générations, le reste étant progressivement reméthylé par la machinerie RdDM (F. K. Teixeira et al. 2009).

Parmi ces 867 DMR parentales stables, 126 ont été retenues pour établir une carte génétique, qui couvre près de 80% du génome (Colomé-Tatché et al. 2012). En suivant les états de méthylation (méthylé vs hypométhylé, origine wt ou *ddm1*) le long des chromosomes, il est possible de déterminer pour chaque epiRIL son "épihaplotype", c'est à dire, quelle région est héritée de quel parent (Figure 3.4).

A partir de cette carte génétique, la première fondée exclusivement sur des marques épigénétiques, il a été possible d'appliquer une approche classique de cartographie de QTL afin de localiser sur le génome des régions sous-tendant la variation phénotypique héritable identifiée dans la population. Dans la continuité d'un article princeps décrivant la cartographie de QTL "épigénétiques" (QTLepi) pour la date de floraison et la longueur de la racine primaire (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014), plus d'une vingtaine de QTLepi ont été identifiés (Figure 3.5.) dans les epiRIL dérivées de *ddm1* pour une diversité de traits (Aller et al. 2018 ; Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014 ; Kooke et al. 2015).
Néanmoins, aucune de ces études n'a à ce jour poursuivi jusqu'à l'identification de l'épiallèle causatif au sein de l'intervalle QTL.

### 3.2.3 Mise en évidence de la causalité des (épi)variants associés à un QTL

Un fois que la présence d'épiallèle potentiel a été détectée au travers d'une recherche de QTL, il va s'agir de mettre en évidence sa causalité. Cela requiert :
- la démonstration que l'effet observé n'est en rien associé à une variation nucléotidique,
- la validation du QTL et sa cartographie fine,
- la complémentation fonctionnelle des variants candidats.

#### Exclusion de la variation nucléotidique

Parmi les 123 epiRIL employées pour la détection des QTLepi, 121 ont été intégralement reséquencées (Gilly et al. 2014 ; Quadrana, Etcheverry et al. 2018) afin de vérifier

que les nouvelles insertions d'ET survenues en raison de l'hypométhylation de ces loci n'étaient pas responsables des associations phénotype-épigénotype détectées.
À prime abord, il semble démesuré de reséquencer des génomes entiers à seule fin de valider qu'une différence de méthylation est responsable d'un phénotype. Néanmoins, seule une telle approche permet d'exclure que le variant causal soit effectivement une différence de méthylation (et non pas un variant ADN), et le cas échéant que celle-ci soit indépendante d'un meQTL en *cis* ou *trans* (variant ADN qui co-ségrège avec l'épivariant), une seconde information qu'il n'est pas possible d'obtenir par le séquençage de la seule région proximale. Cela est d'autant plus critique que près des deux tiers des différences de méthylation identifiées dans la nature dépendent de meQTL et correspondent donc à des épiallèles obligatoires (R. Schmitz, Schultz, Urich et al. 2013).
En raison du caractère laborieux d'une telle analyse, celle-ci n'a été effectuée que dans de rares cas, à l'exemple de (Oey et al. 2015). Néanmoins, les progrès techniques dans le domaine du séquençage ainsi que la baisse des coûts associés font que cette validation du caractère épigénétique des épimutations est appelée à se démocratiser, de même qu'elle s'initie pour l'identification rapide des variants ADN causaux (Michael et al. 2018).
Dans le contexte des QTLepi identifiés dans la population epiRIL, il a pu être mis en évidence que les nouvelles insertions d'ET localisées dans les intervalles de support des différents QTL n'étaient pas responsables de l'effet de ces différents, ce qui rend ces derniers a priori épigénétiques (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014 ; Kooke et al. 2015). Cependant, la contribution des autres types de variants ADN, et notamment des mutations qui se sont accumulées au cours de la production des lots de graines d'où sont extraits les parents wt et *ddm1*, n'a pas été étudiée.

### Restriction de l'intervalle de support du QTL

Comme nous l'avons décrit plus haut et illustré au moyen des QTLepi identifiés à ce jour (Figure 3.5), l'utilisation de populations RIL va permettre d'identifier des QTL pour les traits d'intérêts, mais sans fournir une localisation précise.
Afin d'avoir une idée plus précise des variants potentiellement causatifs pour ce QTL, dits variants candidats, il est possible de "croiser" les localisations des intervalles QTL avec des intervalles obtenus ou dans d'autres populations de cartographie (RIL dérivées du croisement entre d'autres accessions ou populations MAGIC), ou par GWAS. Si cela peut suggérer des candidats solides à tester en priorité par complémentation fonctionnelle, une telle stratégie ne se substitue en rien à la restriction physique de l'intervalle de support au travers d'un protocole de cartographie fine.
Cette approche requiert en premier lieu de mendéliser le QTL d'intérêt, c'est-à-dire, d'aller employer (au besoin, établir) une population de plantes additionnelle dans laquelle ne ségrège que la région à tester, et de vérifier si la ségrégation de cette région va être associée à une ségrégation du phénotype d'intérêt. Cette étape est nécessaire dans la mesure où

la détection d'un QTL est une approche statistique, et il peut être fréquemment observé qu'en fonction du modèle utilisé (*multiple QTL model*, *composite interval mapping* ou régression Haley-Knott) l'intervalle de support du QTL peut varier de façon significative. De façon similaire, deux QTL sur un même bras de chromosome peuvent être réunis en un seul et même intervalle. Cela est particulièrement critique dans les zones où la densité en marqueurs est faible.
Une fois le QTL "validé" par mendélisation, il va s'agir d'aller réduire la taille de la région de support au moyen de lignées dites d'introgression ou de NIL (*Nearly Isogenic Lines*) : au travers de cycles de backcross avec un même individu, il va être possible d'accumuler des évènements de recombinaison additionnels, jusqu'à ce qu'il n'y ait plus qu'une courte région à tester par co-ségrégation génotype-phénotype. Cette stratégie classique est fortement résolutive, mais extrêmement chronophage s'il n'existe pas de population pré-existante. Aussi, une stratégie plus rapide et qui permet à la fois de mendéliser et d'initier la cartographie fine est celle dite des HIF (*Heterogeneous Inbred Families*). De telles familles sont développées en tirant profit de l'hétérozygotie résiduelle présente dans des RIL : à partir d'une RIL hétérozygote dans la région d'intérêt, il va être possible de fixer dans sa descendance l'état homozygote issu de l'un et de l'autre des parents, et donc d'évaluer les phénotypes des plantes possédant l'un ou l'autre des génotypes au locus d'intérêt, dans un fond génétique commun.

### Complémentation fonctionnelle

Enfin, il va s'agir d'aller effectuer la complémentation fonctionnelle des variants candidats. Cette étape requiert d'aller manipuler l'état de méthylation du locus d'intérêt, et de valider l'effet phénotypique de cette réversion. Il s'agit d'un point critique, dans la mesure où à ce jour, aucune des associations profil de méthylation/état d'expression/phénotype décrites chez les plantes n'a fait l'objet de cette validation formelle (Springer et al. 2017). Cette validation finale peut se faire au moyen des stratégies d'édition de l'épigénome décrites dans le chapitre précédent.

# Annexe à l'introduction : revue grand public

---

Cet article de vulgarisation fait partie d'une édition spéciale du périodique Biofutur portant sur l'épigénétique et vise à fournir en des termes accessibles une introduction au concept d'épigénétique transgénérationnelle, en mettant l'accent sur le modèle végétal.

Référence : Baillet, V. et Colot, V. (2014), "La génétique au-delà de la séquence de l'ADN" *Biofutur* 359, p.34-38.

Pour des raisons de droit d'auteur, cette revue ne peut être reproduite dans la version de la thèse disponible en ligne.

Deuxième partie

# Résultats

Chapitre 4

# Caractérisation de la variation nucléotidique présente au sein de la population epiRIL

## 4.1 Introduction

Comme décrit Chapitre 3, le recours à la population epiRIL pour tester la contribution de variants épigénétiques à la variation héritable pour une diversité de traits complexes se fonde sur le principe que les variants en ségrégation dans cette population sont essentiellement sous la forme de DMR (ou épiallèles) et non pas de variations dans la séquence de l'ADN.

Pour autant, les différentes lignées de la population ne peuvent être intégralement dépourvues de variation nucléotidique. En l'espèce, deux catégories de variants ADN sont attendues : d'une part les variants apparus ou déjà présents dans les plantes employées pour établir la population, qui vont ségréger parmi les epiRIL selon les fréquences alléliques attendues d'après le schéma de croisement ; et d'autre part les variants qui se sont accumulés au cours de la propagation des lignées et qui seront propres à chaque lignée. Parmi ces variants, en plus des mutations spontanées (SBS, indels, SV), on s'attend plus particulièrement à trouver de nouvelles insertions d'éléments transposables, compte-tenu du rôle joué par la méthylation de l'ADN dans l'extinction épigénétique de ces séquences. Dans le cadre d'un projet précédent du laboratoire visant à étudier la dynamique de remobilisation des éléments transposables (thèse de Mathilde Etcheverry, (Quadrana, Etcheverry et al. 2018)), 121 lignées epiRIL ont été reséquencées et les nouvelles insertions d'ET ont été recensées.

Mon travail a consisté en la détection des variants ADN non-ET présents parmi les epiRIL, qu'il s'agisse de variants propres à chaque lignée ou en ségrégation dans la population. Ce premier chapitre de résultats se focalise sur l'identification et l'interprétation du patron de ségrégation de ce second type de variants.

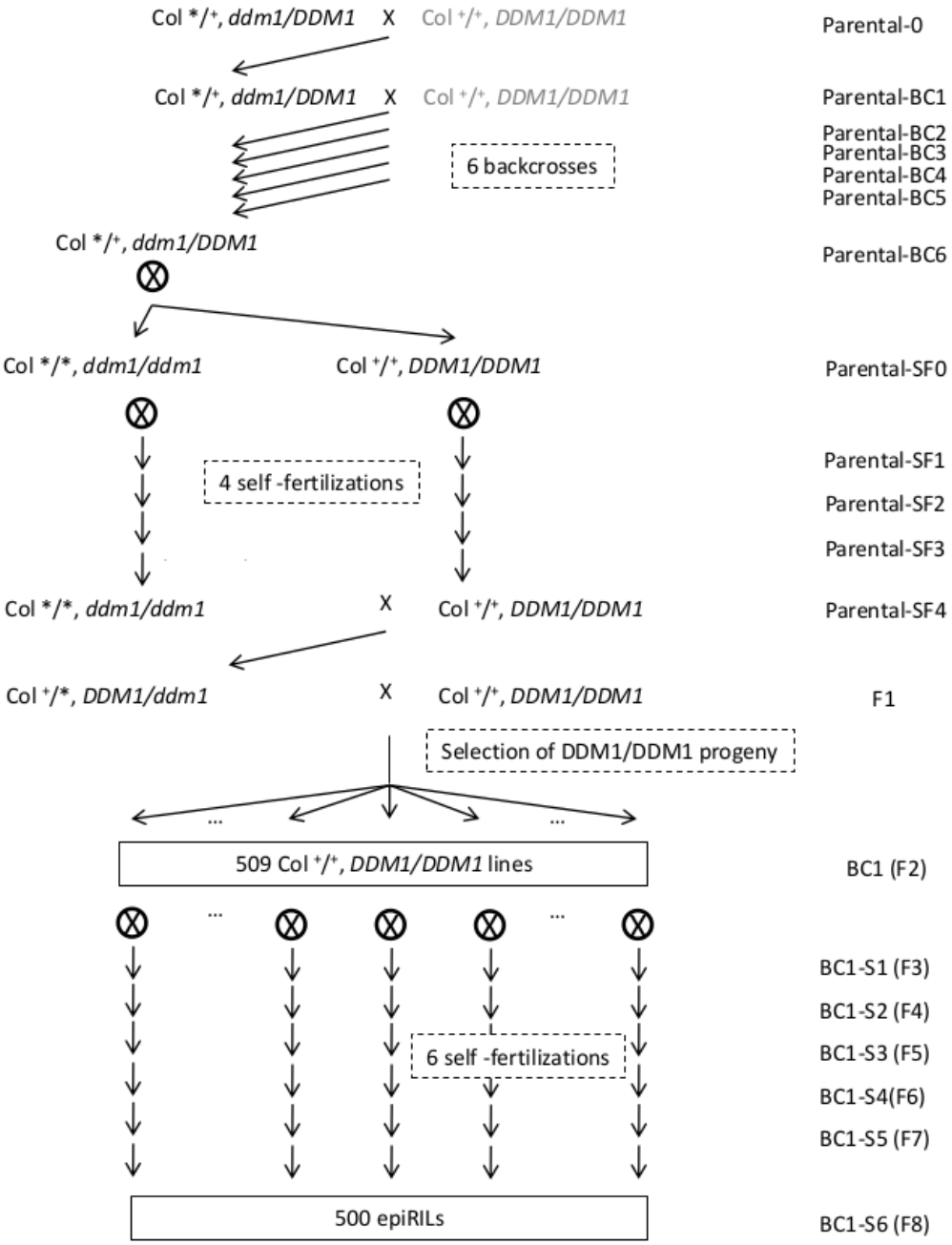


FIGURE 4.1 – Schéma étendu de l'obtention de la population epiRIL. Le mutant *ddm1-2* employé ici a été obtenu par mutagénèse EMS d'une plante Col-0, ce que reflète le *.

## 4.2 Résultats

### 4.2.1 Identité, fréquence allélique et distribution génomique des variants ADN en ségrégation

Afin d'identifier les variants présents dans la population, j'ai effectué la détection conjointe des SBS et indels dans les 121 epiRIL simultanément à l'aide de l'outil GATK HaplotypeCaller. Si une telle approche requiert des ressources computationnelles plus importantes, sa force réside en ce qu'elle favorise un effet de levier qui va permettre le calling de variants à des localisations chromosomiques ponctuellement de moins bonne qualité génomique (couverture verticale plus faible) dans l'une ou l'autre des lignées, réduisant ainsi le nombre de faux-négatifs concernant les variants partagés.
Après applications successive de plusieurs filtres visant à ne conserver que les variants de qualité, j'ai pu identifier 162 variants partagés par au moins deux lignées, parmi lesquels 131 SBS et 31 indels (Table 4.2), ce qui excède grandement les ca. 9 variants attendus au maximum en raison d'un taux de SBS/génome/génération évalué à 1 chez Arabidopsis et des 9 générations séparant les individus WT et *ddm1* à l'origine de la population (Figure 4.1).

Concernant la distribution des fréquences alléliques de ces variants en ségrégation (Figure 4.2), il apparaît que la majorité des variants en ségrégation présentent une fréquence allélique de 50%. Dans la mesure où seuls le F1 et l'individu employé lors du backcross contribuent au matériel génétique des epiRIL à 50% chacun, ce résultat suggère que la majorité des différences nucléotidiques en ségrégation correspondent à des polymorphismes ADN distinguant ces deux individus. Il s'agit d'un résultat inattendu puisque ces deux individus sont censés être apparentés, la plante sauvage utilisée pour le backcross étant issue du même lot de graines que le parent WT initial.

Par ailleurs, cette observation va à l'encontre des prédictions faites d'après le schéma d'établissement de la population epiRIL, puisque la majorité des polymorphismes attendus devraient distinguer l'individu sauvage utilisé lors du croisement initial puis du back-cross de l'individu *ddm1*. Ces variants correspondraient alors ou à des mutations qui sont apparues durant les 4 générations d'autofécondation qui ont permis d'obtenir des individus *ddm1* et WT quasi-isogéniques à partir l'hétérozygote dont ils sont issus, ou à des variants encore à l'état hétérozygote (en dépit des 6 backcross) chez ce même individu et qui vont ségréger dans sa descendance (Figure 4.1) . L'un dans l'autre, il s'agira de variants qui vont distinguer l'individu *ddm1* de l'individu sauvage, et qui dès lors devraient ségréger 75-25 dans la population comme attendu dans le cas d'un croisement initial suivi d'un backcross (75% pour les variants présents chez le sauvage et absent chez *ddm1*, et 25%

pour ceux présents chez *ddm1* et absent chez le WT).

En l'espèce, comme illustré par le spectre des fréquences alléliques, on observe effectivement des variants présents dans 25% de la population et qui peuvent donc correspondre aux variants issus de *ddm1*. A l'opposé, parmi les variants identifiés, aucun ne ségrège à 75%, ce qui signifie qu'il n'existe aucun variant ADN qui soit à la fois commun aux individus wt employés lors du croisement initial et qui les distingue de l'individu *ddm1*. Combiné à la présence de variants ségrégeant à 50%, ce résultat suggère que les deux individus sauvages employés pour le croisement initial et le backcross sont assurément plus divergents qu'attendu.

Le spectre des fréquences alléliques met également en évidence des variants présents dans moins de 25% des epiRIL, et correspondent le plus fréquemment à des variants partagés par 2 à 5 lignées, suggérant qu'ils correspondent à des mutations qui se seraient produites tardivement dans le cycle de vie des individus utilisés lors du backcross (WT ou F1) et qui ne seront alors transmises qu'à une fraction de la descendance. Il n'est cependant pas possible d'exclure que ces variants ségrègent en fait à 25% dans la population, mais ne soient pas détectés comme tels en raison d'un effet d'échantillonnage, puisque moins d'un quart (121 lignées sur 500) de la population epiRIL a été séquencée. Quoi qu'il en soit, en raison de ces faibles fréquences, ces variants (43 au total) ne seront pas considérés dans les analyses à suivre.

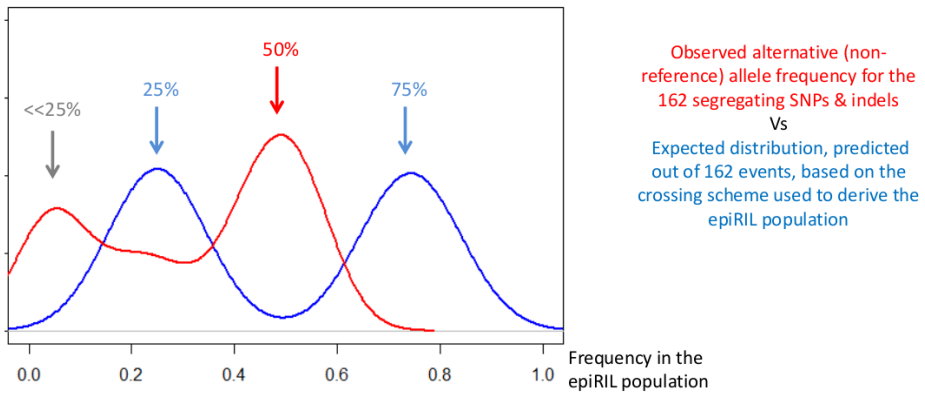


FIGURE 4.2 – Spectre des fréquences alléliques dans les epiRIL. La hauteur des pics pour la distribution attendue (bleu) est donnée pour illustration et ne reflète pas le nombre effectif de variants attendus pour chaque fréquence allélique.

### 4.2.2 Reconstruction des haplotypes parentaux et inférence du pedigree de la population epiRIL

Afin de caractériser plus avant ces résultats inattendus, j'ai mis en relation les différents variants ADN avec leur origine parentale telle qu'inférée sur la base des états de méthylation. En effet, comme décrit dans l'introduction, sur la base de la succession le long des chromosomes des états de méthylation (méthylé versus hypométhylé) aux DMR stables, il est possible de déterminer pour chaque epiRIL son "épihaplotype", c'est-à-dire quelle région a été héritée du parent *ddm1* ou d'un individu sauvage. Néanmoins, comme illustré Figure 3.4, il n'est pas possible de déterminer une origine parentale pour les variants situés aux extrémités des chromosomes, ce qui ne permet pas d'inférer l'origine de 21 variants au total.

En combinant ces épihaplotypes avec les variants présents dans 50% des lignées, on observe que les régions d'épihaplotype WT (en vert) sont partitionnées de la façon suivante : un tiers de ces régions sont associées à des variants ADN également présents dans toutes les régions d'épihaplotype *ddm1*, tandis que les deux autres tiers portent des variants exclusivement propres à ces régions, ce qui va là encore dans le sens de l'emploi lors du backcross d'un individu certes WT au regard de son profil de méthylation, mais distinct de part son génotype des individus WT et *ddm1* ayant donné la F1. On peut en effet noter qu'il y a un non-overlap strict entre ces deux types de variants : aucun d'entre eux n'est partagé ou par les 2 WT, ou par le second WT et l'individu *ddm1*. En ce sens, sur la base de la combinaison des épihaplotypes et des variants ségrégeant à 50%, il est possible de distinguer deux haplotypes, dits haplotype "bleu" (associé au second WT) et "jaune" (partagé par les WT et *ddm1* du croisement initial) (Table 4.1, Figure 4.4).

J'ai ensuite cherché à préciser les génotypes de ces trois individus (*ddm1*, WT1 et WT2) en examinant la co-ségrégation entre les épihaplotypes (WT ou *ddm1*), les haplotypes nouvellement définis ("bleu" ou "jaune"), et les variants ADN présents dans 25% des epiRIL. Parmi ces variants ségrégeant à 25%, trois cas de figure mutuellement exclusifs peuvent être observés : les variants présents exclusivement dans des régions d'origine ddm1, les variants présents exclusivement dans les régions d'épihaplotype WT et d'haplotype "jaune" (associés au premier WT), et les variants présents exclusivement dans les régions d'épihaplotype WT et d'haplotype "bleu" (associés au second WT). Au total, il est possible de classifier les 98 variants (Table 4.2) en cinq catégories dont les spécificités sont synthétisées Table 4.1, et qui sont illustrées Figure 4.4). On peut notamment observer que 9 SBS et 2 indels distinguant les individus *ddm1* et WT1, ce qui cohérent avec le nombre de mutations attendues compte-tenu du nombre de générations séparant ces individus.

TABLE 4.1 – Classification des variants ADN en ségrégation dans les epiRIL

| Code couleur | Fréquence allélique | (Epi)haplotype(s) associé(s) | Origine probable | #SBS | #indels |
|---|---|---|---|---|---|
| Bleu | 50% | Présents exclusivement au sein d'épihaplotypes wt et dans les 2/3 d'entre eux | Présents exclusivement chez le WT employé lors du backcross et à l'état homozygote chez cet individu | 31 | 7 |
| Jaune | 50% | Présents dans 1/3 des épihaplotypes wt et dans tous les épihaplotypes *ddm1* | Présents chez les individus WT et *ddm1* employés lors du croisement initial et à l'état homozygote chez ces individus | 30 | 7 |
| Turquoise | 25% | Présents dans tous les épihaplotypes *ddm1* | Présents à l'état hétérozygote dans le F1 ou dans le *ddm1* employé lors du croisement initial ou à l'état homozygote chez ce dernier | 3 | 1 |
| Orange | 25% | Présents exclusivement dans les 2/3 d'épihaplotypes wt auxquels sont associés des variants "bleus" | Présents exclusivement dans l'individu WT employé lors du backcross et à l'état hétérozygote chez cet individu | 9 | 3 |
| Saumon | 25% | Présents exclusivement au sein des 1/3 d'épihaplotypes wt auxquels sont associés des variants "jaunes" | Présents à l'état hétérozygote dans le F1 ou dans le WT employé lors du croisement initial ou à l'état homozygote chez ce dernier | 6 | 1 |

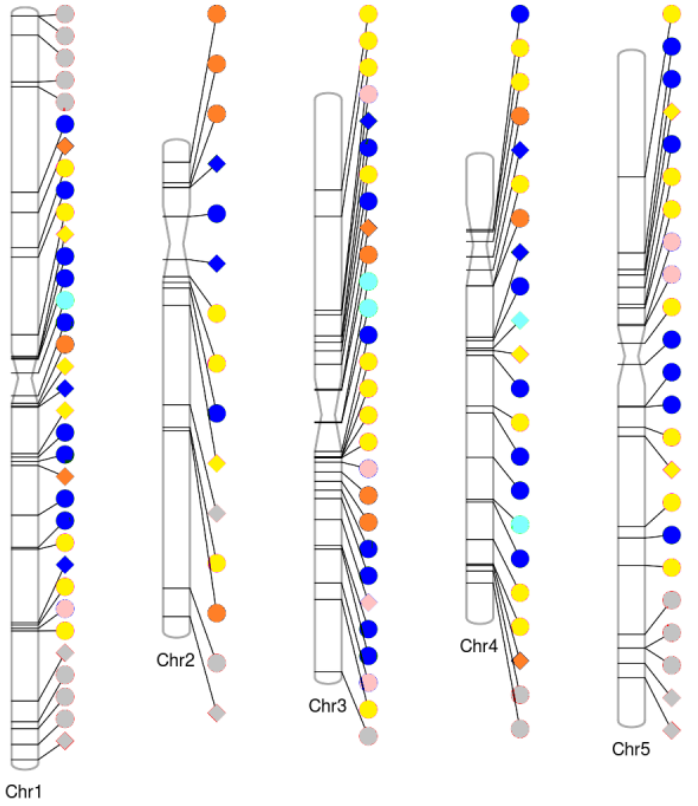


FIGURE 4.3 – Distribution génomique des variants en ségrégation dans les epiRIL. Les SBS et indels sont représentés par des cercles et des losanges respectivement. Le code couleur est tel que décrit TABLE 4.1, les variants grisés correspondent à des variants situés au-delà des DMR stables les plus distales et pour lesquels il n'est pas possible d'inférer avec certitude l'origine parentale.

En considérant le taux de mutation par génome par génération calculé chez Arabidopsis, cela se traduit par 60 générations d'évolution indépendante entre l'individu sauvage employé lors du backcross et les individus WT et *ddm1* employés lors du croisement initial, soit une divergence significative pour des individus considérés comme quasi-isogéniques et

issus du même lot du graines, ce qui laisse supposer une contamination par un individu d'un génotype tiers mais néanmoins dans un fond génétique Col-0. Au moyen de données de séquençage disponibles pour des pools de plantes cousines des individus WT et *ddm1*, j'ai cherché à retrouver si des traces de cet haplotype "bleu" au sein des génomes d'individus apparentés à ceux employés pour établir les générations précoces des epiRIL. Aucun variant associé à cet haplotype ne peut être identifié ; au contraire, l'ensemble de ces individus présentent un haplotype "jaune" ainsi que quelques variants associés à cet haplotype, ce qui tend à confirmer que la présence de l'haplotype "bleu" dans les epiRIL résulte d'un contamination lors du backcross.

Par ailleurs, et confirmant d'autant la divergence entre cet individu et le reste des individus associés à la création de la population, j'ai pu mettre en évidence qu'à cet haplotype "bleu" étaient également associées des néo-insertions du rétrotransposon *ATCOPIA78*, qui avaient précédemment été interprétées comme s'étant produites dans l'individu F1 en raison de la confrontation dans le zygote des épigénomes méthylé (wt) et non-méthylé (*ddm1*).
Ce résultat est également renforcé par le fait que parmi les 7 insertions observées, l'une d'entre elles est présente à l'extrémité sud du chromosome 5, donc en déséquilibre de liaison avec le locus *DDM1*. Dans la mesure où la descendance du backcross a été génotypée pour ne sélectionner que les individus de génotype *DDM1/DDM1*, si la remobilisation s'était produite dans la F1, alors cette insertion devrait être présente dans 50% des lignées et les six autres dans 25%, or les sept insertions ségrègent à 25%.
Par ailleurs, le fait que ces néo-insertions soient présentes dans 25% et non 50% des lignées mettent en évidence le fait qu'elles sont à l'état hétérozygote dans le WT2, ce qui peut suggérer une remobilisation récente mais de faible ampleur. Même si nous savons à présent que*ATCOPIA78* présente dans les accessions naturelles d'Arabidopsis une variation dans le nombre de copies en relation avec un gradient de température (Quadrana, Bortolini Silveira et al. 2016), la remobilisation de cet ET dans l'accession Col-0 n'a été observée que dans des plantes défectives dans la voie du RdDM soumises à un stress chaleur (Cavrak et al. 2014 ; Ito et al. 2011 ; Sanchez et al. 2017). Aussi, j'ai recherché parmi les variants associés à ce WT2 si l'annotation révélait des mutations pouvant se traduire par une remobilisation de *ATCOPIA78*. Il s'avère que parmi ces variants (Table 4.2), aucun ne se situe ou au sein d'une annotation d'*ATCOPIA78*, ou au sein ou à proximité de gènes connus comme étant associés aux différentes voies de méthylation de l'ADN. Une autre hypothèse concernant ce second individu est qu'il puisse dériver d'un mutant T-DNA pour l'un de ces gènes. A cette fin, j'ai extrait les régions génomiques dérivées de cet individu afin de vérifier si des séquences correspondant à des fragments de T-DNA ou à des gènes de résistance pouvaient être identifiées. Aucune trace de telles séquences ne peut être identifiée parmi les différentes epiRIL (données non montrées), En conclusion,

l'identité tout autant que l'origine de ce second WT restent à ce jour indéterminées.

### 4.2.3 Validation du pedigree

Des travaux précédents du laboratoire avaient pu mettre en évidence qu'approximativement 8% des 500 epiRIL étaient issues non pas du backcross mais de l'autofécondation de l'individu F1 (Johannes, Porcher et al. 2009).
Dès lors, si le pedigree proposé est correct, l'on devrait retrouver parmi les epiRIL analysées une proportion équivalente de lignées qui ne présentent aucun variant ADN associé à l'individu employé lors du backcross (néo-insertions de *ATCOPIA78* et variants "bleus" et "oranges").
Il s'avère que 9 epiRIL sur les 121 (soient 7,4% des lignées, ce qui est compatible avec la fraction prédite) présentent ces caractéristiques, suggérant que le pedigree proposé et illustré FIGURE 4.4 est effectivement correct.

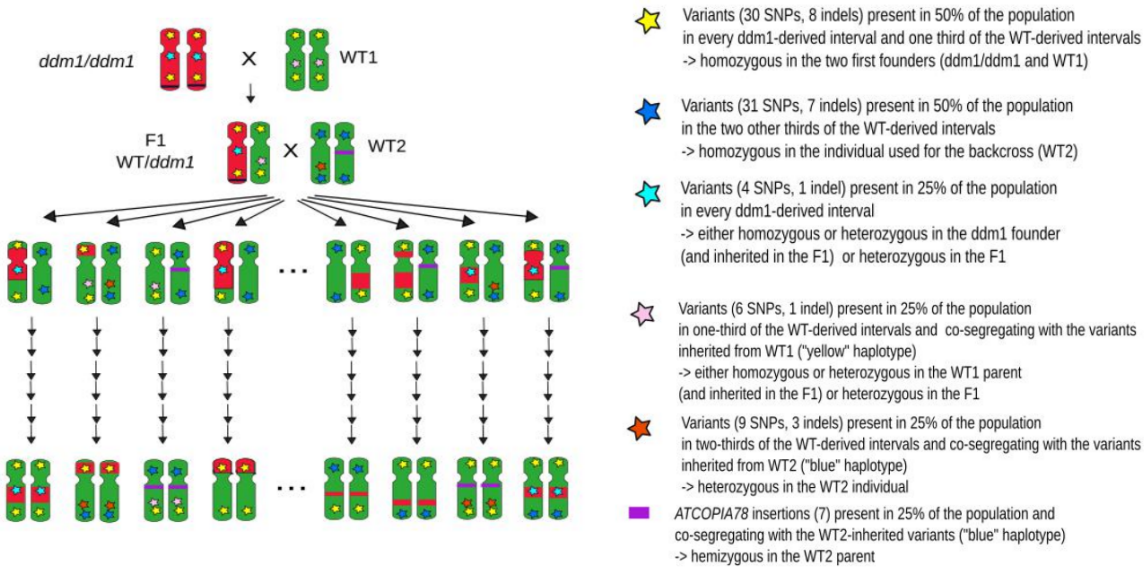


FIGURE 4.4 – Pedigree révisé de la population epiRIL sur la base de l'incorporation des données de variants en ségrégation.

## 4.3 Conclusion et discussion

### 4.3.1 Perspectives d'exploitation des données de variants partagés

Au travers de la caractérisation des variants partagés par plusieurs lignées et de leur patron de ségrégation, mon travail a permis de mettre en évidence que le pedigree effectif de la population epiRIL est plus complexe que décrit initialement (Johannes, Porcher et al. 2009).

Quelle peut être la portée de ces résultats, au regard des travaux précédemment effectués avec cette population (Aller et al. 2018 ; Colomé-Tatché et al. 2012 ; Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014 ; Johannes, Porcher et al. 2009 ; Kooke et al. 2015 ; Latzel et al. 2013 ; Roux et al. 2011 ; Y.-Y. Zhang et al. 2013) ? Alors que deux accessions diffèrent l'une de l'autre par plusieurs milliers de SNP et indels, on n'observe ici qu'un très faible polymorphisme nucléotidique, laissant supposer que celui-ci est sans grande incidence, sauf peut-être si ces quelques variants en ségrégation contribuent à la variation héritable décrite pour différents traits complexes dans la population epiRIL. Compte-tenu du faible nombre de variants caractérisés (98 au total) et exploitables (c'est-à-dire pour lesquels l'absence du variant est non-ambiguë) et de leur distribution le long des chromosomes, il n'apparaît pas possible de construire une carte génétique, cette fois-ci basée sur des marqueurs ADN et qui permettrait de rechercher des QTL portés par variants nucléotidiques. On peut en effet constater, comme illustré Figure 4.3, que ces variants ne sont pas distribués uniformément le long des chromosomes, mais pour la majorité d'entre eux organisés en clusters. Une explication à cette distribution particulière peut être que la majorité de ces clusters sont retrouvés dans les régions péricentromériques, dans lesquelles l'intensité de recombinaison est la plus faible. Aussi, il est envisageable que ces régions n'aient pas été "nettoyées" au cours des backcross successifs, d'où un niveau élevé de polymorphisme.
Néanmoins, concernant les QTLepi détectés à ce jour (Figure 3.5), l'absence de co-ségrégation entre épiphaplotype et haplotype pour la majorité (94/98) des variants rend moins plausible la causalité de variants ADN aux QTLepi, ce qui suggère que les QTLepi détectés à ce jour sont effectivement de nature épigénétique. Ces points seront explorés plus avant dans le Chapitre 7.
Quoi qu'il en soit, il peut être envisagé de tirer profit de ces variants en ségrégation pour développer des marqueurs ADN qui permettraient d'identifier chacune des 121 epiRIL par simple génotypage ; un outil qui peut s'avérer précieux pour qui utilise les epiRIL au laboratoire.

### 4.3.2 Les epiRIL, une population de lignées MA avec des états de méthylation mosaïques

Comme discuté dans l'introduction, les populations de lignées d'accumulation de mutations se caractérisent par i) un fond génétique homogène, ii) des modalités de propagation avec présence de goulot d'étranglement à chaque génération et iii) une propagation en l'absence de pression de sélection.
Compte-tenu de la faible ampleur de la variation ADN en ségrégation et des modalités de création et de propagation des epiRIL, cette population peut être assimilée à une population de lignées MA, laquelle présenterai des épigénomes mosaïques de régions méthylées et hypométhylées.
Ceci nous permet donc d'évaluer de façon directe l'impact de la méthylation de l'ADN sur le taux de mutation (Chapitre 5).

Table 4.2: Identité des variants ADN non-ET en ségrégation

| Type | Classe | Chr | Position | REF | ALT | Annotation | PCR |
|---|---|---|---|---|---|---|---|
| SNP | Bleu | Chr1 | 7241579 | G | A | ET \| AT1TE23395 | . |
| Indel | Orange | Chr1 | 8067517 | AAT | A | Gène \| AT1G22770 | . |
| SNP | Jaune | Chr1 | 9497509 | G | A | Gène \| AT1G27340 | . |
| SNP | Bleu | Chr1 | 9883998 | G | A | Région intergénique | X |
| SNP | Jaune | Chr1 | 13033405 | C | T | Région intergénique | X |
| indel | Jaune | Chr1 | 13908828 | CA | C | Région intergénique | . |
| SNP | Bleu | Chr1 | 13967107 | C | T | ET \| AT1TE45845 | . |
| SNP | Bleu | Chr1 | 14023610 | C | T | ET \| AT1TE46065 | X |
| SNP | Azur | Chr1 | 14590272 | T | C | ET \| AT1TE48035 | . |
| SNP | Bleu | Chr1 | 15509104 | G | A | ET \| AT1TE51070 | X |
| SNP | Orange | Chr1 | 15818726 | A | G | ET \| AT1TE52085 | X |
| indel | Jaune | Chr1 | 15935454 | TGGTGAAG | T | ET \| AT1TE52520 X . | |
| indel | Bleu | Chr1 | 15972511 | AT | A | ET \| AT1TE52650 | . |
| indel | Jaune | Chr1 | 17863273 | TC | T | Région intergénique | . |
| SNP | Bleu | Chr1 | 18002233 | T | C | ET \| AT1TE59745 | X |
| SNP | Bleu | Chr1 | 18173880 | C | T | Gène \| AT1G49120 | . |
| indel | Orange | Chr1 | 18343059 | TA | T | Gène \| AT1G49560 | . |
| SNP | Bleu | Chr1 | 20362809 | G | A | ET \| AT1TE67280 | . |
| SNP | Bleu | Chr1 | 21675141 | G | A | Région intergénique | . |
| SNP | Jaune | Chr1 | 21745570 | C | T | Région intergénique | . |
| indel | Bleu | Chr1 | 24729023 | CA | C | ET \| AT1TE81230 | . |
| SNP | Jaune | Chr1 | 24825541 | G | A | Gène \| AT1G66540 | . |
| SNP | Saumon | Chr1 | 24961964 | T | A | Gène \| AT1G66910 | . |
| SNP | Jaune | Chr1 | 25064916 | G | A | Gène \| AT1G67110 | . |
| SNP | Orange | Chr2 | 1486806 | G | A | ET \| AT2TE06885 | X |
| SNP | Orange | Chr2 | 1672192 | T | A | ET \| AT2TE07745 | . |
| SNP | Bleu | Chr2 | 1685655 | C | CA | ET \| AT2TE07805 | . |
| SNP | Bleu | Chr2 | 2865618 | C | T | Région intergénique | X |
| indel | Bleu | Chr2 | 4618505 | G | GAT | Région intergénique | . |
| SNP | Jaune | Chr2 | 5306035 | C | T | ET \| AT2TE21605 | . |
| SNP | Jaune | Chr2 | 5542775 | C | A | Région intergénique | X |
| SNP | Bleu | Chr2 | 5774917 | G | A | ET \| AT2TE23480 | . |
| indel | Jaune | Chr2 | 6469233 | GT | G | ET \| AT2TE26425 | . |
| SNP | Jaune | Chr2 | 11434437 | A | C | Gène \| AT2G26810 | . |
| SNP | Orange | Chr2 | 11560143 | G | C | ET \| AT2TE50140 | . |
| SNP | Jaune | Chr3 | 3666840 | A | G | Région intergénique | . |
| SNP | Jaune | Chr3 | 4740011 | C | T | Région intergénique | . |
| SNP | Jaune | Chr3 | 8547929 | G | A | Région intergénique | . |
| SNP | Saumon | Chr3 | 8733951 | A | C | Gène \| AT3G24170 | X |
| indel | Bleu | Chr3 | 9589063 | TA | T | Région intergénique | . |
| SNP | Bleu | Chr3 | 9858537 | T | C | ET \| AT3TE41045 | . |
| SNP | Jaune | Chr3 | 10202051 | A | G | ET \| AT3TE42410 | . |
| SNP | Bleu | Chr3 | 10743038 | G | A | ET \| AT3TE44650 | . |
| indel | Orange | Chr3 | 11762942 | CTATGGAAAA | C | ET \| AT3TE48990 | . |

Table 4.2: Identité des variants ADN non-ET en ségrégation

| Type | Classe | Chr | Position | REF | ALT | Annotation | PCR |
|---|---|---|---|---|---|---|---|
| SNP | Orange | Chr3 | 11795091 | C | T | ET \| AT3TE49090 | . |
| SNP | Azur | Chr3 | 13092809 | C | T | ET \| AT3TE53605 | . |
| SNP | Bleu | Chr3 | 14288480 | A | T | Région intergénique | . |
| SNP | Jaune | Chr3 | 14509867 | G | T | ET \| AT3TE59480 | . |
| SNP | Jaune | Chr3 | 14518339 | C | G | ET \| AT3TE59525 | . |
| SNP | Jaune | Chr3 | 14530903 | G | A | ET \| AT3TE59535 | . |
| SNP | Jaune | Chr3 | 14545591 | C | A | ET \| AT3TE59600 | X |
| SNP | Saumon | Chr3 | 14725289 | T | A | Gène \| AT3G42640 | . |
| SNP | Orange | Chr3 | 15129360 | A | C | ET \| AT3TE61585 | . |
| SNP | Orange | Chr3 | 15507720 | G | A | ET \| AT3TE62815 | . |
| SNP | Bleu | Chr3 | 15861145 | C | T | ET \| AT3TE64160 | . |
| SNP | Bleu | Chr3 | 16205156 | C | T | ET \| AT3TE65635 | . |
| indel | Saumon | Chr3 | 17060365 | G | GT | ET \| AT3TE69150 | . |
| SNP | Bleu | Chr3 | 18112512 | G | A | ET \| AT3TE73435 | . |
| SNP | Bleu | Chr3 | 18253358 | C | G | Gène \| AT3G49220 | . |
| SNP | Saumon | Chr3 | 19637526 | A | G | ET \| AT3TE79750 | . |
| SNP | Jaune | Chr3 | 20306786 | C | T | ET \| AT3TE82555 | . |
| SNP | Bleu | Chr4 | 2834559 | G | A | ET \| AT4TE13315 | X |
| SNP | Jaune | Chr4 | 2920544 | G | A | Région intergénique | . |
| SNP | Jaune | Chr4 | 3318778 | C | G | ET \| AT4TE15155 | X |
| SNP | Orange | Chr4 | 3946307 | T | A | ET \| AT4TE17205 | . |
| indel | Bleu | Chr4 | 4474111 | ACCAGAT | A | ET \| AT4TE18865 | . |
| SNP | Jaune | Chr4 | 5115842 | T | A | ET \| AT4TE21220 | . |
| SNP | Orange | Chr4 | 5126890 | A | C | ET \| AT4TE21290 | X |
| indel | Bleu | Chr4 | 7214677 | C | CATTATTTGT | Région intergénique | . |
| SNP | Bleu | Chr4 | 7344157 | C | A | Région intergénique | . |
| indel | Azur | Chr4 | 7652616 | AC | A | ET \| AT4TE33480 | . |
| indel | Jaune | Chr4 | 7750735 | G | GT | Gène \| AT4G13310 | . |
| SNP | Bleu | Chr4 | 7944780 | G | A | Région intergénique | . |
| SNP | Jaune | Chr4 | 10012411 | A | G | Gène \| AT4G18030 | X |
| SNP | Bleu | Chr4 | 10291133 | A | T | Région intergénique | . |
| SNP | Bleu | Chr4 | 12093042 | C | T | ET \| AT4TE56060 | . |
| SNP | Azur | Chr4 | 13798054 | A | T | Gène \| AT4G27640 | . |
| SNP | Bleu | Chr4 | 13909063 | A | C | Région intergénique | . |
| SNP | Jaune | Chr4 | 15427181 | A | G | Gène \| AT4G31890 | . |
| SNP | Jaune | Chr4 | 16428223 | A | G | Région intergénique | . |
| SNP | Jaune | Chr5 | 4896800 | G | A | Gène \| AT5G15110 | . |
| SNP | Bleu | Chr5 | 7980674 | A | T | Région intergénique | . |
| SNP | Bleu | Chr5 | 8654576 | G | C | Région intergénique | . |
| indel | Jaune | Chr5 | 8872537 | TG | T | Région intergénique | . |
| SNP | Bleu | Chr5 | 9381623 | C | A | ET \| AT5TE34035 | X |
| SNP | Jaune | Chr5 | 10075739 | G | A | Région intergénique | X |
| SNP | Jaune | Chr5 | 10214637 | G | A | Région intergénique | X |
| SNP | Saumon | Chr5 | 10874531 | C | G | ET \| AT5TE39495 | . |

Table 4.2: Identité des variants ADN non-ET en ségrégation

| **Type** | **Classe** | **Chr** | **Position** | **REF** | **ALT** | **Annotation** | **PCR** |
|---|---|---|---|---|---|---|---|
| SNP | Saumon | Chr5 | 10943157 | C | T | Gène \| AT5G28920 | . |
| SNP | Jaune | Chr5 | 11645737 | T | C | ET \| AT5TE42015 | . |
| SNP | Bleu | Chr5 | 12506158 | G | A | ET \| AT5TE44310 | . |
| SNP | Bleu | Chr5 | 14215841 | A | T | ET \| AT5TE50805 | . |
| SNP | Bleu | Chr5 | 14240551 | T | A | Gène \| AT5G36180 | . |
| SNP | Jaune | Chr5 | 15070083 | T | C | Gène \| AT5G37860 | X |
| indel | Jaune | Chr5 | 15431687 | G | GT | Région intergénique | . |
| SNP | Jaune | Chr5 | 19102123 | G | A | ET \| AT5TE68695 | . |
| SNP | Bleu | Chr5 | 19562194 | G | A | Gène \| AT5G48250 | . |
| SNP | Jaune | Chr5 | 20669976 | G | A | ET \| AT5TE74470 | . |

Chapitre 5

# Impact d'une perte de méthylation de l'ADN sur le spectre des mutations ponctuelles

## 5.1 Introduction

Comme décrit dans le Chapitre 1, la méthylation de l'ADN est intrinsèquement mutagène, compte-tenu du taux plus élevé de désamination des 5mC que des cytosines non méthylées, et de la formation d'une base ADN (T) au cours de la désamination des 5mC. Les mésappariements qui en résultent sont dès lors plus difficiles à identifier et à corriger, d'où une forte incidence de transitions C $\rightarrow$ T aux cytosines méthylées.
Si ce mécanisme est bien connu, l'impact mutationnel de la 5mC est dans les faits évalué le plus fréquemment au travers d'approches indirectes (corrélations entre données de polymorphisme ou de divergence et profils de méthylation) ; et à ce jour aucune description directe et à l'échelle du génome entier n'a été effectuée chez quelque organisme que ce soit. Afin d'effectuer une telle analyse, il s'agirait de se placer dans un système expérimental dans lequel le profil de méthylation peut être altéré ; puis d'en caractériser le patron de mutation.
Contrairement à l'emploi d'agents chimiques pouvant altérer le patron de méthylation (5-azacytidine), le recours à des mutants de méthylation permet de modifier de façon robuste et reproductible les profils de méthylation. Si comme décrit au Chapitre 2 les mutants de méthylation chez les mammifères présentent une létalité embryonnaire qui ne permet pas une description du patron de mutation ; chez les plantes et plus spécifiquement chez Arabidopsis de tels mutants sont viables et fertiles. Par ailleurs, comme présenté Chapitre 3, des états de méthylation différentiels peuvent être transmis au travers des générations chez ces organismes, ce qui autorise l'étude de l'impact mutationnel de la 5mC dans un fond génétique sauvage, évitant ainsi de potentiels effets confondants.

Dans le Chapitre 4, j'ai pointé le fait que les modalités de propagation de la population epiRIL (absence de sélection, filiation monograine) ainsi que son très faible niveau de polymorphisme rendent cette population assimilable à une population de lignées MA,

avec un particularité notable ; les méthylomes mosaïques de régions méthylées et hypométhylées de chaque lignée.
Dans ce contexte, j'ai entrepris d'employer la population epiRIL pour analyser l'impact mutationnel de pertes héritables de méthylation. Comme mentionné dans l'introduction, cette hypométhylation peut conduire à la réactivation transcriptionnelle et dans certains cas à la remobilisation d'éléments transposables, un cas de figure particulier qui est traité dans le Chapitre 6.

## 5.2 Manuscrit en préparation

Les pages suivantes présentent l'ébauche d'un article en préparation décrivant le spectre des mutations ponctuelles dans la population epiRIL et en particulier la réduction du taux de transitions C $\rightarrow$ T dans cette population.
Dans cet article, je i) mets en évidence la similarité entre epiRIL et MA lines, ii) caractérise les spectres des mutations ponctuelles dans ces deux populations et iii) propose un modèle explicatif pour la réduction du taux de transitions C $\rightarrow$ T dans les epiRIL.
Des analyses complémentaires qui permettront de préciser les résultats obtenus sont en cours et seront discutées dans la section suivante de ce chapitre.

# Towards a direct, genome-wide assessment of the mutational impact of cytosine methylation in Arabidopsis thaliana

Victoire Baillet, Leandro Quadrana and Vincent Colot
Institut de Biologie de l'Ecole Normale Superieure, CNRS UMR 8197 - INSERM U1024 - ENS, PSL University, France
Correspondence to: colot@biologie.ens.fr

**Abstract**

DNA methylation is an epigenetic modification that is pivotal in ensuring proper genome function and integrity, notably through the silencing of transposable elements (TEs). However, as spontaneous deamination of 5-methylcytosine (5mC), which can lead to C→T transitions, is more frequent than that of unmethylated C, DNA methylation is also inherently mutagenic, yet a direct, genome-wide assessment of its mutational impact is still lacking. Here, we make use of a population of Arabidopsis mutation accumulation lines with mosaic methylation profiles to investigate the effect of heritables losses of DNA methylation on the spectrum of spontaneous mutations. We show in this population a strong reduction in the rate of C→T, which we interpret as the combination of the lower number of mCs at-risk for deamination and the higher DNA repair over heterochromatic regions, presumably due to the re-establishment of transcription at the loci. Altogether, these results propose a in-depth model of the mutational forces influencing the evolution of TE sequences.

## Introduction

In both plants and mammals, DNA methylation plays a pivotal role in ensuring proper genome function and integrity, notably through the epigenetic silencing of transposable elements (TEs). Importantly, the presence of 5-methylcytosine (5mC) leaves considerable genomic footprints, as exemplified by the depletion in CpG dinucleotides in mammalian genomes which are typically methylated at CG sites except in so-called CpG islands [3].
The higher mutability of 5mC compared to its unmethylated counterpart can be explained by the combination of two factors. First, if C and 5mC are both susceptible to spontaneous deamination, 5mC is more liable to deamination than C, which promotes a higher mutation rate of 5mC compared to C Second, deamination of C converts C into uracil (U), which is not found naturally in the DNA sequence. As a consequence, U can be easily detected and removed by uracil glycosylase with the subsequent replacement by the correct base. On the other hand, deamination of 5mC converts 5mC into T, which is one of the building blocks of DNA. The mechanism to detect the deaminated 5mC as a mutated base is therefore less efficient, which consequently reduces its rate of repair [17, 9]
Despite this well-characterized effect of DNA methylation, we still lack a comprehensive view of its impact on the whole spectrum of mutations in any given organism. Studies to date have been limited in scope [4] or mainly relied on indirect approaches, such as comparisons between methylomes and inter- or intra-species DNA polymorphisms [39, 21, 26, 13, 28], while one would ideally need to alter DNA methylation patterns in a controlled manner and investigate the mutational consequences genome-wide. This strategy is however difficult to implement, particularly in mammals where a loss of DNA methylation leads to embryonic lethality.
In contrast, the flowering plant Arabidopsis thaliana combines multiple advantages to perform such a study: most mutants affecting DNA methylation are viable and fertile, perturbation of DNA methylation patterns can be induced reproducibly by genetic means and a loss of DNA methylation can in part be stably inherited even in the absence of the inducing mutation, therefore allowing to investigate the genomic consequences of a loss of DNA methylation in a wild-type background, therefore avoiding the confonding effects associated with further
Here, we make use of a population of so-called epigenetic recombinant inbred lines (epiRILs) to investigate the impact of DNA methylation on the rate and spectrum of spontaneous mutation. The epiRILs derive from a cross between the *ddm1* mutant, which displays an extensive loss of DNA methylation over TE loci, and a near-isogenic wild-type individual [18].Following a back-cross, lines selected as wild-type at the *DDM1* locus were advanced by single-seed descent for six generations under minimal selection, a design which is similar to the one used to establish populations of so-called Mutation Accumulation lines, an experimental strategy that has been employed to study spontaneous mutation in a large number of organisms under the following rationale: at each generation, the use of small effective population size (here equal to one) minimizes natural selection and maximises genetic drift, therefore allowing the accumulation of all but the most deleterious mutations [14, 1, 22].
Such a MA lines population has been established in

Arabidopsis thaliana: starting from a single, highly inbred founder individual, Shaw and colleagues established a set of 120 lines, out of which 10 were then sequenced following 30 generations of independent evolution [34, 27]. Direct comparison between the epiRILs and the MA lines allows us to infer how much do the heritable losses of DNA methylation in the epiRILs did affect the spectrum of spontaneous mutations in only a few generations.

## Results

### The epiRILs are *bona fide* MA lines

By design, DNA sequence variation in the epiRILs is expected to occur at two levels : first, the variants that are shared by at least two lines and must have occurred during the initial stages of the population setup and second, those that are private to a single epiRIL and therefore must have occurred during the propagation of the lines. As the epiRILs derive from inbred founders, the amount of shared DNA sequence variation is expected to be limited, with the exception of transposable element remobilisation events that could have occurred in the *ddm1* mutant or in the F1 due to the reduced DNA methylation. 122 epiRILs that underwent six generations of single-seed descend and for which methylome information is available have been resequenced and both TE remobilisation events and other types of DNA sequence mutations (point mutations, indels, structural variations) have been profiled [12, 31], the former being described extensively elsewhere [31].
In order to tease apart standing DNA sequence variation from mutations that accumulated during the propagation of the lines, we performed a population-level joint discovery of small-scale variants (single-base substitutions and short insertions or deletions, here after referred to as SBS and indels, respectively) (see **Methods** section). This analysis led us to identify 131 SBS and 31 indels that segregate among the epiRILs and that could unequivocally be associated to a parental origin (**Figure S2**, **Table S2**)).
This level of standing variation is orders of magnitude lower than that of a classical RIL design deriving from a cross between two accessions, in which the founders differ by tens of thousands of DNA polymorphisms including large-scale rearrangements in some instances [41]. It was recently shown that a high level of heterozygosity was associated with a high mutation rate in the form of a feed-forward loop [40]. Here, given the low level of divergence between the founding individuals, this parameter is expected to be of little influence on the pattern of mutation accumulation in the epiRILs and these can therefore be approximated to an regular MA line population (**Figure1**). Before pursuing further on characterizing the rate and spectrum of spontaneous mutations in the epiRIL population, we investigated whether the different DNA repair pathways were affected in the *ddm1* mutant as well as in the epiRILs.
Indeed, *DDM1* encodes a chromatin remodeler from the SWI/SNF family, which are involved in DNA repair in a wide range of organisms, and previous work could show a higher sensitivity of *ddm1* mutant to gamma irradiation and UV-C damage [33], presumably in line with an impaired nuclear organization and chromatin structure [29].
We looked for the differential expression of 190 genes previously curated and associated with the main DNA repair pathways in the RNAseq data available for five epiRILs and the *ddm1* mutant and did not find any of these genes to be misregulated in these individuals. In addition, none of the segregating DNA sequence variants (**Table S2**) were found to overlap with these DNA repair-associated genes, prompting us to conclude that the epiRILs do not display any major constitutive alteration of the main DNA repair pathways.

The census of all shared DNA sequence variants allows us to confidently identify these that are private to a single line, and are relevant to the study of the mutation rate. Using only the 113 F9 epiRILs for which we have sequence data of sufficient quality to estimate mutation rates (**Figure S1**, **Table S1**), we identify 388 single-base substitutions and 185 indels (**Supplementary Table 2**). We found no excess of synonymous SBS ($\chi$-squared test, $P$=0.348), which indicates a lack of selection during the propagation of the lines, reinforcing the notion that the epiRILs can truly be assimilated to a MA population (**Figure1**).
In order to draw a comparison with the MA lines population, we used the same pipeline to reanalyze publicly available sequence data for 10 G30 MA lines, including the 5 lines (MA29, MA39, MA49, MA59, and MA69) that had been used to provide the first direct estimate of the Arabidops mutation spectrum [27]. The 259 SBS and 74 indels we identified among the 10 lines all include the 99 and 13 previously reported in [27] (**Supplementary Table 3**).

### Patterns of small-scale mutation in the epiRILs and the MA lines

Building on the identification of private, small-scale (less than 50bp) mutation events that occurred in the MA lines and the epiRILs, we analyzed the rate and pattern of point mutation and indels in both populations. Strikingly, the ratio of point mutations to indels, equals to 2.05 in the epiRILs, deviates from estimates obtained in other mutation accumulation experiments in Arabidopsis (3.5 and 3.7 in the Col-0 MA lines and a parent-progeny sequencing Col-0/Ler design [40], respectively). Further investigation of this pattern identifies a bias of the epiRIL indel spectrum towards 3- to 5bp long insertions specifically, which correspond for most of them (82/96) to the excision footprints that were left by the extensive remobilisation of one DNA transposon in the epiRILs, described elswhere [31]. Point

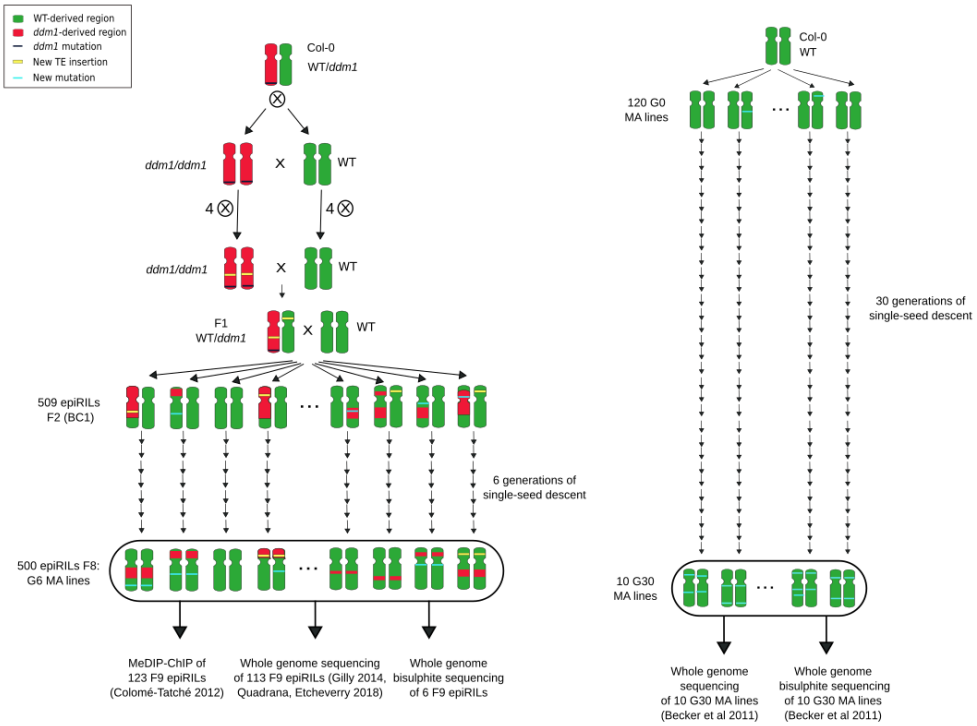


Figure 1: Comparison between the epiRIL (left panel) [18] and the MA line population (right panel) [34]. Note the similarities in the design (propagation by single-seed descent) as well as the differences between the two populations in terms of number of lines and number of generations.

mutations that have accumulated in the epiRILs are comparatively more abundant in TE-rich, pericentromeric regions than in gene-rich chromosome arms (**Figure S4**). Furthermore, GC-rich triplets have a tendency to mutate the most and transitions are more frequent than transversions, again suggesting that the global pattern of point mutations resembles that of Arabidopsis mutation accumulation experiments [27, 40] (**Figure S4**)

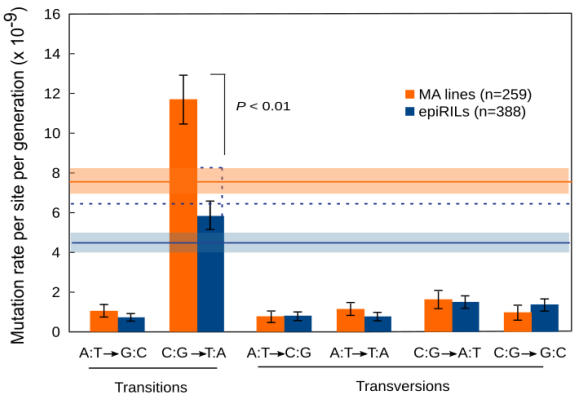


Figure 2: Single-base substitution spectra in the epiRILs (blue) and the MA lines (orange). Horizontal lines of the same colour depict the global mutation rate, coloured shadow the SEM. Dashed lines indicate the rates upon correction for maximal possible underestimation given the distinct number of generations between the two settings [16].

However, the comparison of the mutation spectra of the MA lines and the epiRILs reveal a strong and significant decrease specifically in the rate of C:G $\rightarrow$ T:A transitions in the latter ($P < 0.001$, Fisher's exact test) which in turn triggers a reduction of the genome-wide rate of single-base substitution (4.48 +/- 0.23 x 10-9 base substitution per site per generation in the epiRILs, versus 7,63 +/- 0.49 x 10-9 in the MA lines, which is in line with previous estimates [27, 40]). Because the MA lines and the epiRILs differ in the number of generations of propagation of the lines and because we only consider homozygous mutations, the rate of mutation we find in the epiRILs may be an underestimation of the actual one. In order to account for this possible confounding effect, we corrected the rates of C $\rightarrow$ T transitions as well as the genome-wide one for the maximal possible understimation. Still, we find a significant decrease in both rates in the epiRILs compared to the MA lines.

## Dissecting the basis of the distinct rate of C $\rightarrow$ T in the two settings

Because the epiRILs display reduced methylation levels exclusively in *ddm1*-derived regions, we investigated further the rates of C $\rightarrow$ T transitions between the MA lines and the wt- and *ddm1*-derived

intervals of the epiRIL population. We find a reduction in this rate between wt- and *ddm1*-regions, in line with the methylation status between the two, but also between wt-derived intervals of the epiRILs and the MA lines (**Figure 3A**). In order to find out the origin of the latter difference, we determined the trinucleotide context of the different mutated positions in the two settings. More precisely, we searched for a differential enrichment of so-called dipyrimidine sites (C or T, immediatly preceded or followed by C or T) among the mutated sites. Indeed, the initial description of the Arabidopsis mutation spectrum using MA lines associated the strong bias towards C $\rightarrow$ T to the combined effects of spontaneous deamination of methylated cytosines and UV-B induced mutagenesis at dipyrimidine sites, which both give rise to [27].
Strikingly, we found the mutated positions to be enriched in such DP sites in the MA lines but not in the epiRILs (**Figure 3B**). Because UV-B-induced mutagenesis actually corresponds to daylight intensity which can vary from one region of the world to the next, we hypothetized that these differences may be reflect the distinct growth conditions of the two populations.

## Rate of methylation-induced C $\rightarrow$ T and dampening effect of transcription

In order to avoiding confounding effects associated with the aforementioned high rate of C $\rightarrow$ T transitions at dipyrimidine sites in the MA lines, we then focused exclusively on differences between wt- and *ddm1*- derived intervals.
We first aimed to estimate a rate of C $\rightarrow$ T transitions in line with the methylation status of the cytosine in both type of intervals. Because the wt- and *ddm1*-derived regions were determined using the MeDIP-chip technique [6] for methylome profiling [5] and therefore do not provide single-cytosine resolution, we made use of the whole-genome bisulphite sequencing datasets available for 5 epiRILs to establish a so-called composite methylome. Indeed, because the methylation status of individual cytosines were shown to vary readily between samples and generations, with additionnal confounding effects associated with the difficulty to produce a robust methylation call in non-CG contexts [32, 2], such an approach allows us to dampen sample-specific individual variation.
This composite methylome, illustrated **Figure 4** allows us to define the methylation status of single cytosines across the two-thirds of the genome, with a relative fraction of pericentromeric regions and chromosome arms comparable to the one of the full-length genome (**Figure 4A and B**).

Albeit the fact that methylated cytosines are associated with a higher rate of C $\rightarrow$ T transitions than their unmethylated counterparts is well established, the actual extent of this difference has been proposed to range from 10 to 40-fold. [7, 36, 8]. Here, we estimate a rate of C $\rightarrow$ T transitions at methylated cytosines in wt-derived regions to be 10 times higher than that of their unmethylated counterparts in both CG and CHG contexts, which is at odds with a previous estimate (12) derived from the pattern of CpG depletion in vertebrate genomes [36]. Conversely, this difference appears milder in the CHH context, a pattern that may result from a lower statistical power to detect methylated sites in this context and/or from the lower methylation level at these sites. In comparison, the rate of C $\rightarrow$ T transitions is reduced in all three methylation contexts in *ddm1*-derived intervals compared to wt, a pattern that can not be attributed to the simple lower number of methylated cytosines in these regions. Alternatively, this would suggest either that the methylation level of individual cytosines is lower in these regions, which would translate into a lower probability of mutation, or that the methylated sites in ddm1-derived regions were subjected to additionnal features than would trigger a lower rate of damage and/or a higher rate of repair (**Figure 4C**). As a support for the latter hypothesis, the global mutation rate over TEs is also lower in *ddm1*-derived regions compared to wt-ones, a pattern that is not observed in genes which reside in euchromatic regions and are not affected by the *ddm1*-induced hypomethylation (**Figure 4D**)
Because TEs located in pericentromeric intervals of *ddm1* origin are found to be transcriptionally reactivated [31] and because the re-establishment of transcription and concomitant transcription-coupled DNA repair to heterochromatin suffices to trigger a reduction in the rate of mutation [42], we propose that this reduced rate of C $\rightarrow$ T transitions is the consequence of both a lower rate of mutation (due to the lower number of mCs) and a higher of repair (due to an ectopic transcription-coupled DNA repair).

## Conclusion

In this work, we aimed to provide a clear assessement of the methylation-induced rate of mutations over transposable element sequences. Such TE annotations have been previously found to undergo a high rate of mutation, which in no small part is due to the mutagenic effect of the deamination of methylated cytosines [24, 25, 19, 11].
However, because mutation rates are known to vary across the genome due to the combination of several factors ranging from local GC content to chromatin compaction as well as replication timing [23, 15, 35], knowing the exact rate at which these sequences mutate is crucial to provide a better assessment of their rate of evolution.
Here, using a population of mutation accumulation lines segregating heritable losses of DNA methylation over TE loci, we suggest that the reduction in the rate of C $\rightarrow$ T transitions is the combination of the lower number of methylated Cs and of the re-establishment of transcription at these loci typically

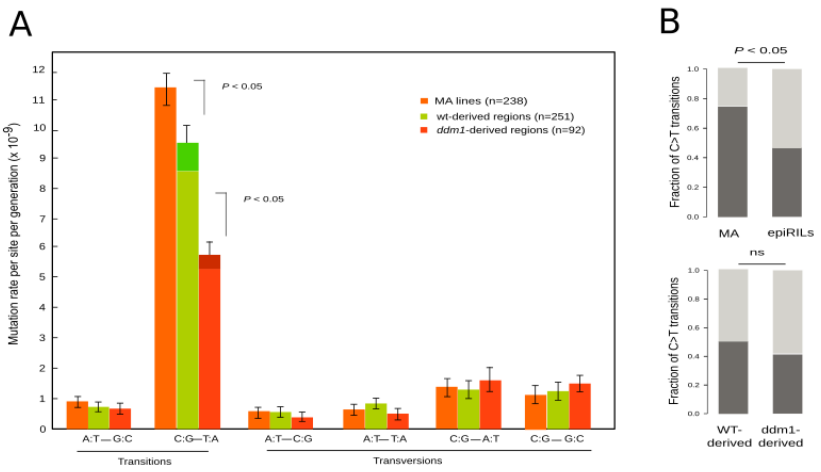


Figure 3: Higher mutation rates at cytosines in a dipyrimidine context explain differences in the rates of C → T transitions between the MA lines and the wt-derived intervals of the epiRILs. (A) SBS spectrum of the MA lines (orange) and the epiRIL in wt-derived (green) and *ddm1*-derived (red) regions. Dark green and dark red rectangles indicate the correction for maximal possible underestimation of the rate due to differences in the number of generations. (B) Proportion of C → T transitions at dipyrimidine (dark grey) and non-dipyrimide (light grey) sites, p-value from Fisher's Exact test.

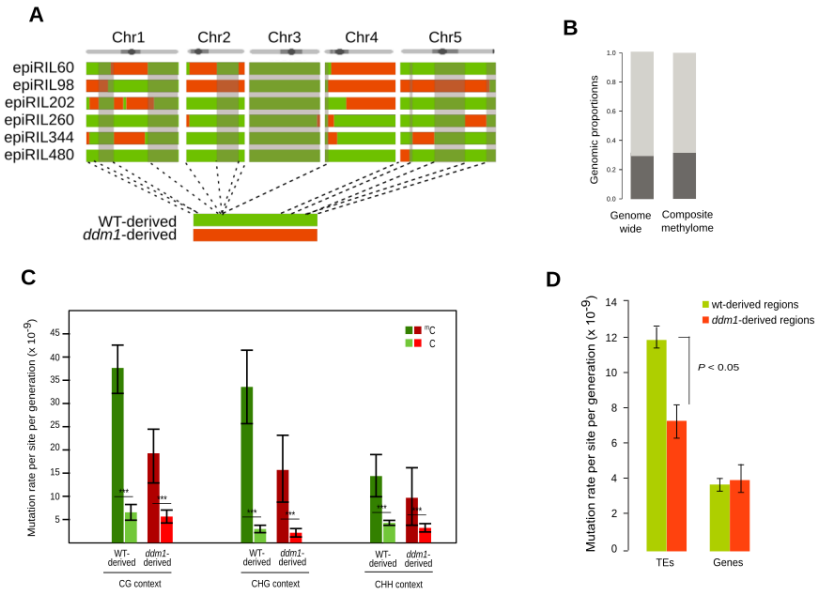


Figure 4: Rate of C → T transitions in the epiRILs and influences thereof. (A,B) Composite methylome constitution and proportion of pericentromeric regions (dark gray) and chromosome arms (light grey) in comparison to the complete genome. (C) Rate of mutation at methylated and unmethylated cytosines in the three methylation contexts in wt- and *ddm1*-derived regions of the epiRIL. Error bars denote SEM. Stars denote statistically significant differences from Fisher's exact test. (D) Rate of mutation in the TE and gene fraction of the genome in wt and *ddm1*-derived intervals. P denotes statistical significance of Fisher's exact test, error bars denote SEM.

found in compacted heterochromatin.
Further work will be needed to incorporate the effects of the global level of chromatin decompaction over the whole pericentromeric regions (rather than the sole re-establishment of transcription over TE loci) as well as the effect of the local GC content, which is generally low over TE sequences [20, 24] and is known to influence the liability to spontaneous deamination at a local scale [10, 26].

## Materials and Methods

**Dataset origin**
Mate-pair WGS data for 121 F9 epiRIL previously produced in our group are available in GEO under SRP XXXXXX [31, 12]. RNAseq and WGBS data for 5 and 6 epiRILs respectively are deposited under the same SRP. MA line WGS and WGBS data described in [2, 27] were retrieved from GEO accession XXXXXX. Wt- and *ddm1*-derived regions are defined based on previously published work [5].

**Reads mapping and variant calling**
After quality check, trimmed reads were mapped to the Arabidopsis Col-0 TAIR10 reference genome using BWA-MEM v 0.7.10, which was shown to perform well on mate-pair sequencing data with 76nt-long reads. Duplicate reads were then removed using Picard MarkDuplicates. The joint identification of small-scale variants (single-base substitutions and indels) was perfomed using GATK HaplotypeCaller in –joint-genotyping mode on the epiRIL (parents included) and MA lines cohorts separately. To ensure an accure calling of point mutations, raw variant calls were filtered to keep only the variants which quality scores were equal or higher than the ones of known true mutations in the samples (the *ddm1-2* mutation, PCR-validated SBS that segregate in the epiRIL population and SBS identified in [27]). Variant calls falling in regions of the Arabidopsis Col-0 genome that display recurrent coverage abnormalities [30] were filtered out. All variant calls passed these criteria were subjected to visual inspection using Integrative Genome Viewer [37].

**Mutation rate calculation**
Mutation rate per generation $mg$ or per site $ms$ was estimated as $m = n/g$ or $n/s$ respectively, where $n$ stands for the number of homozygous mutation per line, $s$the number of sites at risk for mutation and $g$ the total elapse of generations. Because of a finite number of generations, the calculated mutation rate is either an overestimation or an under estimation of the actual, depending on whether heterozygous mutations or only homozygous ones were taken into account. In order to be able to compare mutation rates between the MA lines and epiRIL settings, we applied the correction for maximal possible underestimation as devised in [16].

**Calling of methylated cytosines in the composite methylome**
WGBS datasets were analyzed as described in [38]. Only 5 epiRILs out of the 6 for which WGBS information are available were used ; the last one was filtered out due to a high content of *ddm1*-derived regions which may have driven down the methylation level of individuals sites.


### Acknowledgements

We thank members of the Colot lab for discussions as well as Anne Britt for the identity of DNA repair associated genes. This work was supported by the European Union Seventh Framework Programme Network of Excellence EpiGeneSys (Award 257082, to VC), the Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, ANR-11-IDEX-0001-02 PSL* Research University and ANR-12-ADAP-0020-01 (to VC). VB was supported by a PhD studentship from IdEx Paris Sciences et Lettres Research University (ANR-10-IDEX-0001-02 PSL) and LabEx MemoLife (ANR-10-LABX-54 MEMO LIFE).


### Author contributions

VB and VC conceived the project. VB performed all bioinformatic analyses except the differential expression analysis over TE sequences, which was performed by LQ. VB interpreted the data and wrote the paper with significant help from VC.

### Supplementary information

**Figure S1**: Fraction of the genome with a 5X,10X and 15X coverage for each of the 121 sequenced epiRILs
**Figure S2**: Crossing scheme of the epiRILs population illustrating the parental origins for each type of segregating DNA sequence variant
**Figure S3**: Pattern of DNA repair genes in the epiRILs
**Figure S4**: Pattern of point mutations in the epiRIL and in the MA lines

**Table S1**: EpiRILs used for the detection of shared variants vs. for the mutation rates analyses
**Table S2**: Identity of the segregating DNA polymorphisms and haplotypes of the 121 epiRILs
**Table S3**: Identity of the SBS and indels that accumulated during the propagation of the epiRILs.
**Table S4**: Identity of the SBS and indels that accumulated during the propagation of the MA lines.

## References


[1] Baer, Charles F, Miyamoto, Michael M, and Denver, Dee R. "Mutation rate variation in

multicellular eukaryotes: causes and consequences." *Nature Reviews Genetics* 8 (Aug. 2007), pp. 619–631. DOI: 10.1038/nrg2158.

[2] Becker, Claude et al. "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome." *Nature* 480 (Sept. 2011), pp. 245–249. DOI: 10.1038/nature10555.

[3] Bird, AP. "DNA methylation and the frequency of CpG in animal DNA." *Nucleic Acids Res.* 8 (1980), pp. 1499–1504.

[4] Chen, RZ et al. "DNA hypomethylation leads to elevated mutation rates." *Nature* 395 (1998), pp. 89–93.

[5] Colome Tatche, Maria et al. "Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation." *Proc Natl Acad Sci U S A* 109 (Oct. 2012), pp. 16240–16245. DOI: 10.1073/pnas.1212955109.

[6] Cortijo, S et al. "Genome-wide analysis of DNA methylation in Arabidopsis using MeDIP-chip." *Methods Mol Biol* 1112 (2014), pp. 125–149.

[7] Coulondre, C et al. "Molecular basis of base substitution hotspots in Escherichia coli." *Nature* 274 (1978), pp. 775–780.

[8] Duncan, BK and Miller, JH. "Mutagenic deamination of cytosine residues in DNA". *Nature* 287 (1980), pp. 560–561.

[9] Ehrlich, M and Wang, RY. "5-Methylcytosine in eukaryotic DNA." *Science* 212 (1981), pp. 1350–1357.

[10] Fryxell, KJ and Moon, WJ. "CpG mutation rates in the human genome are highly dependent on local GC content." *Mol Biol Evol.* 22 (2005), pp. 650–658.

[11] Galagan, James E. and Selker, Eric U. "RIP: the evolutionary cost of genome defense". *Trends in Genetics* 20 (2004), pp. 417–423. DOI: 10.1016/j.tig.2004.07.007.

[12] Gilly, Arthur et al. "TE-Tracker: systematic identification of transposition events through whole-genome resequencing." *BMC Bioinformatics* 15 (Nov. 2014), p. 377. DOI: 10.1186/s12859-014-0377-z.

[13] Glastad, KM et al. "Effects of DNA Methylation and Chromatin State on Rates of Molecular Evolution in Insects." *G3 (Bethesda)* 6 (2015), pp. 357–363.

[14] Halligan, DL and Keightley, PD. "Spontaneous Mutation Accumulation Studies in Evolutionary Genetics". *Annu. Rev. Ecol. Evol. Syst.* 40 (2009), pp. 151–172. DOI: 10.1146/annurev.ecolsys.39.110707.173437.

[15] Hodgkinson, A. and Eyre-Walker, A. "Variation in the mutation rate across mammalian genomes". *Nature Reviews Genetics* 12 (Oct. 2011), pp. 756–766. DOI: 10.1038/nrg3098.

[16] Hoffman, PD et al. "Rapid accumulation of mutations during seed-to-seed propagation of mismatch-repair-defective Arabidopsis." *Genes Dev* 18 (2004), pp. 2676–2685.

[17] Holliday, R and Grigg, GW. "DNA methylation and mutation." *Mutat Res* 285 (1993), pp. 61–67.

[18] Johannes, Frank et al. "Assessing the impact of transgenerational epigenetic variation on complex traits." *PLoS Genetics* 5 (June 2009), e1000530. DOI: 10.1371/journal.pgen.1000530.

[19] Kricker, MC, Drake, JW, and Radman, M. "Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes." *Proc Natl Acad Sci U S A* 89 (1992), pp. 1075–1079.

[20] Lerat, E., Capy, P., and Biemont, C. "Codon usage by transposable elements and their host genes in five species." *J. Mol. Evol.* 54 (2002), pp. 625–637.

[21] Li, J et al. "Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome." *PLoS Genet.* 8 (2012), e1002692.

[22] Lynch, Michael et al. "Genetic drift, selection and the evolution of the mutation rate." *Nature Reviews Genetics* 17 (Oct. 2016), pp. 704–714. DOI: 10.1038/nrg.2016.104.

[23] Makova, Kateryna D and Hardison, Ross C. "The effects of chromatin organization on variation in mutation rates in the genome." *Nature Reviews Genetics* 16 (Apr. 2015), pp. 213–223. DOI: 10.1038/nrg3890.

[24] Maumus, Florian and Quesneville, Hadi. "Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana." *Nature Communications* 5 (June 2014), p. 4104. DOI: 10.1038/ncomms5104.

[25] Meunier, Julien et al. "Homology-dependent methylation in primate repetitive DNA." *Proc Natl Acad Sci U S A* 102 (Apr. 2005), pp. 5471–5476. DOI: 10.1073/pnas.0408986102.

[26] Mugal, Carina F and Ellegren, Hans. "Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content." *Genome Biology* 12 (June 2011), R58. DOI: 10.1186/gb-2011-12-6-r58.

[27] Ossowski, Stephan et al. "The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana." *Science* 327 (Jan. 2010), pp. 92–94. DOI: 10.1126/science.1180677.

[28] Panchin, Alexander Y., Makeev, Vsevolod J., and Medvedeva, Yulia A. "Preservation of methylated CpG dinucleotides in human CpG islands". *Biology Direct* 11 (2016), p. 11.

[29] Probst, AV et al. "Two means of transcriptional reactivation within heterochromatin." *Plant J.* 33 (2003), pp. 743–749.

[30] Quadrana, Leandro et al. "The Arabidopsis thaliana mobilome and its impact at the species level." *eLife* 5 (June 2016). DOI: 10.7554/eLife.15716.

[31] Quadrana, Leandro et al. "Transposon accumulation lines uncover histone H2A.Z-driven integration bias towards environmentally responsive genes". *bioRxiv* (2018). DOI: 10.1101/447870.

[32] Schmitz, R.J. et al. "Transgenerational epigenetic instability is a source of novel methylation variants". *Science* 334 (Oct. 2011), pp. 369–373. DOI: 10.1126/science.1212959.

[33] Shaked, H, Avivi-Ragolsky, N, and Levy, AA. "Involvement of the Arabidopsis SWI2/SNF2 chromatin remodeling gene family in DNA damage response and recombination." *Genetics.* 173 (2006), pp. 985–994.

[34] Shaw, RG, Byers, DL, and E, Darmo. "Spontaneous mutational effects on reproductive traits of Arabidopsis thaliana." *Genetics* 155 (2000), pp. 369–378.

[35] Sima, J. and Gilbert, D.M. "Complex correlations: replication timing and mutational landscapes during cancer and genome evolution." *Curr Opin Genet Dev.* 25 (Apr. 2014), pp. 93–100. DOI: 10.1016/j.gde.2013.11.022..

[36] Sved, J and Bird, A. "The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model." *Proc Natl Acad Sci* 87 (1990), pp. 4692–4696.

[37] Thorvaldsdottir, Helga, Robinson, James T., and Mesirov, Jill P. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". *Briefings in Bioinformatics* 14 (2013), pp. 178–192. DOI: 10.1093/bib/bbs017.

[38] Van Dooren, Tom et al. "Mild drought induces phenotypic and DNA methylation plasticity but no transgenerational effects in Arabidopsis". *bioRxiv* (2018). DOI: 10.1101/370320.

[39] Xia, Junfeng, Han, Leng, and Zhao, Zhongming. "Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome." *BMC Genomics* 13 Suppl 8 (2012), S7. DOI: 10.1186/1471-2164-13-S8-S7.

[40] Yang, Sihai et al. "Parent-progeny sequencing indicates higher mutation rates in heterozygotes." *Nature* 523 (July 2015), pp. 463–467. DOI: 10.1038/nature14649.

[41] Zapata, Luis et al. "Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms." *Proc Natl Acad Sci U S A* 113 (July 2016), E4052–E4060. DOI: 10.1073/pnas.1607532113.

[42] Zheng, CL et al. "Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes." *Cell Rep.* 9 (Jan. 2014), pp. 1228–1234. DOI: 10.1534/g3.114.015545.

## 5.3 Conclusion et discussion

Dans ce chapitre, j'ai pu mettre en évidence une réduction spécifique du taux de transitions C → T dans les epiRIL, en lien avec les pertes héritables de méthylation dans cette population. J'ai également pu proposer que cette réduction soit due à la combinaison du plus faible nombre de cytosines méthylées mais également du rétablissement de la transcription (et donc de l'intervention du transcription-coupled DNA repair) au niveau des ET, qui sont les cibles primaires de la machinerie de méthylation et sont hypométhylés dans les epiRIL. Je décris par ailleurs qu'une partie des différences dans le taux de transitions C → T entre epiRIL et MA lines a à voir avec un enrichissement en sites dipyrimidiques des positions mutées dans cette dernière population, ce qui peut être imputé à des conditions distinctes de croissance des lignées entre ces deux populations.

Les résultats présentés dans les pages précédentes sont novateurs, notamment puisqu'ils décrivent la première évaluation directe à l'échelle du génome entier de l'impact mutationnel de la méthylation des cytosines. Cependant, plusieurs points peuvent être discutés et méritent pour certains des analyses complémentaires. En premier lieu, il faut mentionner qu'en raison de la ségrégation des régions dérivées de wt et de *ddm1* au cours de la propagation des lignées ainsi que de la reméthylation au fil des générations de près des deux tiers des DMR parentales (F. K. Teixeira et al. 2009), les travaux présentés ici ne peuvent fournir une estimation exacte du taux de transitions C → T associé à la méthylation des cytosines. Une estimation plus précise aurait pu être obtenue en effectuant une analyse équivalente à celle des trios parents-enfants, à savoir en séquençant un grand nombre d'epiRIL en génération F2 et non pas F8. Dans ce cadre, les états de méthylation différentiels ne seraient pas encore fixés mais à l'état "épihétérozygote", ce qui pourrait également amener des effets confondants si de même que dans les régions hétérozygotes l'épihétérozygotie module le taux de mutation localement. Une explication à cela pourrait être un état chromatinien particulier dans ce contexte, qui influencerait l'accessibilité à l'ADN.
Un premier point qu'il sera nécessaire d'éclaircir est le suivant : les différences dans le taux de mutation au niveau des régions péricentromériques dérivées de *ddm1* sont-elles exclusivement provoquées par le *transcription coupled DNA repair* au niveau des ET, ou un état chromatinien plus ouvert de l'ensemble de ces régions y contribue-t-il également ? Ceci pourrait être précisé par l'incorporation des données ATACseq en cours d'obtention au laboratoire, qui permettent de capturer le degré d'ouverture de la chromatine.
Un facteur confondant additionnel a à voir avec le fait que de façon générale, les ET présentent un contenu en GC plus faible que le reste du génome (Lerat et al. 2002), avec des variations entre familles d'ET et entre annotations, en raison de leur "âge" dans ce dernier cas (Maumus et al. 2014). Comme décrit Chapitre 1, un contenu localement plus

élevé en GC réduit le taux de transitions C → T aux cytosines méthylées. Des analyses sont actuellement en cours afin de tester l'association entre ces deux paramètres et donc de préciser l'impact effectif de la 5mC sur le taux de mutation dans ces différents types d'annotations.

Il peut également être mentionné que le séquençage bisulphite de 121 epiRIL, actuellement conduit au laboratoire, permettra de préciser le taux de transitions C → T dans les différents contextes de méthylation mais également dans les différentes fractions du génome (régions péricentromériques vs bras dérivés du wt ou de *ddm1*).

Quoi qu'il en soit, ces différentes analyses seront limitées dans leur résolution en raison du faible nombre total de mutations dans les epiRIL tout comme dans les MA lines.

Chapitre 6

# Impact de la remobilisation des ET sur le spectre mutationnel de la population epiRIL

## 6.1 Introduction

Comme nous l'avons vu dans l'introduction, le ET sont des contributeurs importants aux paysages mutationnels, et ce sur plusieurs plans : création de cassures double-brin dans l'ADN lors de l'intégration à un nouveau site ou lors de l'excision, formation d'indels associés à cette excision, et réarrangements chromosomiques résultant notamment d'évènements de recombinaison ectopique entre copies dispersées dans le génome.
Les observations effectuées à ce jour, desquelles résultent ces conclusions, dérivent cependant pour la majorité d'entre elles d'approches indirectes ou locus-spécifiques. Quelques exemples en sont (Wicker et al. 2016), qui se fonde sur le polymorphisme à proximité d'annotations d'ET pour évaluer l'impact mutationnel du mécanisme d'insertion ou encore (Colot et al. 1998), qui effectue une description extensive des différentes *footprints* d'excision d'*Ascot* du locus *b2*. Aussi, il est pertinent d'aller étudier à l'échelle du génome entier quelles vont être les conséquences mutationnelles, autres que les seules nouvelles insertions d'ET, de la remobilisation soudaine d'une à plusieurs familles d'éléments transposables.
Il a pu être mis en évidence au laboratoire qu'en raison des pertes de méthylation provoquées par la mutation *ddm1* au niveau des séquences répétées, les epiRIL présentaient une remobilisation extensive (environ 1800 nouvelles insertions) d'une dizaine de familles d'ET. Par ailleurs, la majorité des ces néo-insertions sont propres à chaque epiRIL, ce qui indique que cette remobilisation a lieu pour l'essentiel au cours de la propagation des lignées et non pas dans le mutant *ddm1* ou la F1 (Figure 6.1).
Dans ce contexte, j'ai entrepris d'analyser les conséquences mutationnelles de cette activité extensive des ET. Ce chapitre récapitule les résultats obtenus et intègre également le manuscrit (Section 6.2.3) décrivant la remobilisation des ET dans les epiRIL dont je suis co-auteur (Quadrana, Etcheverry et al. 2018).

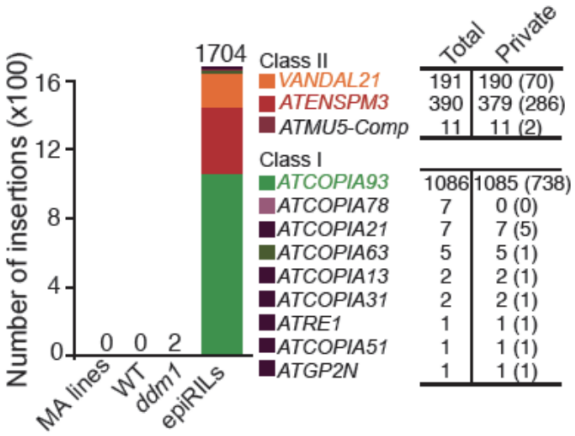


FIGURE 6.1 – Remobilisation des ET dans les epiRIL, en comparaison des MA lines et des individus wt et *ddm1* employés pour établir la population. La colonne "*private*" recense les évènements de remobilisation présents dans une seule epiRIL, donc s'étant produits au cours de la propagation des lignées. Les nombres entre parenthèses indiquent les néo-insertions détectées à l'état hétérozygote. 3 familles d'ET constituent l'essentiel des néo-insertions : l'ET de classe I *ATCOPIA93* et les ET de classe II *VANDAL21* et *ATENSPM3*. Comme décrit dans le CHAPITRE 4, *ATCOPIA78* n'est pas remobilisé dans les epiRIL : les 7 insertions détectées sont toutes partagées. Extrait de (Quadrana, Etcheverry et al. 2018).

## 6.2 Résultats

### 6.2.1 Le spectre des indels de la population epiRIL est modelé par la remobilisation d'*ATENSPM3*

Dans la continuité de mon travail portant sur les spectres mutationnels de la population epiRIL et des MA lines, décrit dans le chapitre précédent, j'ai pu faire l'observation que le spectre des indels de la population epiRIL, illustré FIGURE 6.2, présentait en comparaison de celui des MA lines un excédent d'insertions de 3 à 5 nucléotides de long.

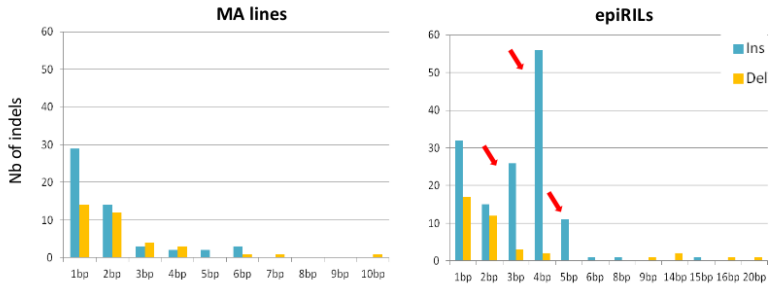


FIGURE 6.2 – Spectre des indels dans les epiRIL (droite) et les MA lines (gauche). Les flèches indiquent le biais vers des insertions de 3 à 5pb dans les epiRIL.

Comme nous l'avons vu dans l'introduction, l'excision des ET de classe II peut conduire

à la formation de courtes insertions d'une longueur comparable à celle du TSD formé au cours de l'insertion. Comme illustré FIGURE 6.3, il peut être envisagé que des indels accumulés au cours de la propagation des lignées correspondent à de telles *footprints* d'excision. Le cas échéant, ces *footprints* putatives doivent présenter plusieurs propriétés : une taille compatible avec celle du TSD, une localisation génomique similaire aux sites préférentiels d'insertion de l'ET, et une séquence correspondant à la répétition directe de celle du TSD. De plus, elles devraient être retrouvées exclusivement dans les epiRIL dans lesquels l'ET en question est mobile, et pas dans les autres.

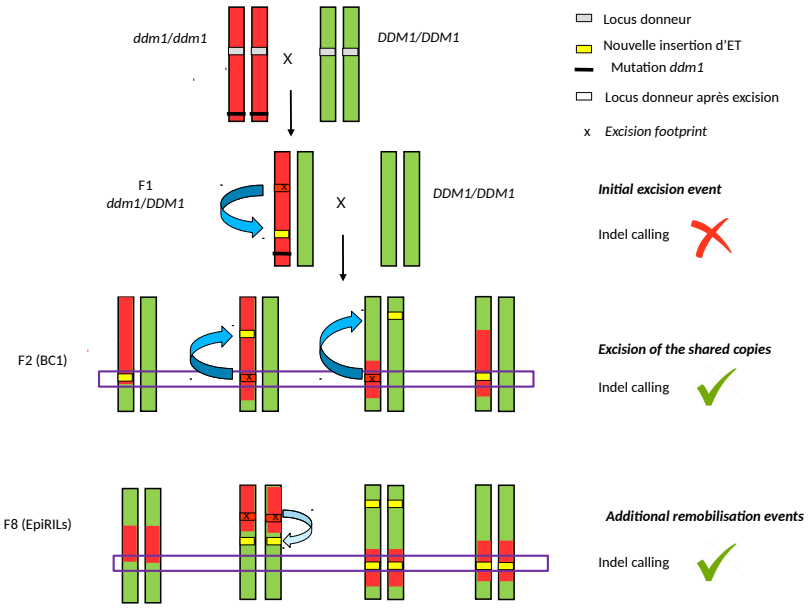


FIGURE 6.3 – Origine des *footprints* d'excision détectées par recherche d'indels. Les flèches bleues indiquent la remobilisation de l'ET.

La longueur du TSD formée au cours d'une insertion est spécifique de chaque famille d'ET. Les deux familles d'ET de classe II les plus mobiles dans les epiRIL, *VANDAL21* et *ATENSPM3*, forment des TSD de 8 et 3 pb respectivement, ce qui suggère que l'excès d'insertions de 3 à 5 pb reflète les excisions de ce dernier transposon.

Afin de valider cette hypothèse, j'ai en premier lieu vérifié si aux loci correspondants à des insertions d'*ATENSPM3* partagées entre plusieurs epiRIL (FIGURE 6.3) des indels de taille similaire pouvaient être identifiés, lesquels traduiraient alors l'excision epiRIL-spécifique de cette néo-insertion. Comme illustré FIGURE 6.4, de tels évènements peuvent être observés, avec une insertion de longueur compatible avec les *footprints* putatives qui se reflètent dans le spectre des indels.


FIGURE 6.4 – Copie d'écran IGV illustrant une *footprint* d'excision au niveau d'une insertion d'*ATENSPM3* partagée par plusieurs epiRIL. Le TSD est figuré en vert (flèche mettant en évidence le pic de 3pb de long dans le profil de couverture verticale dans l'epiRIL où l'insertion est encore présente -panel du haut-, nucléotides en vert dans la séquence restantes au site d'excision dans la deuxième epiRIL illustrée ici -panel du bas-), la *footprint* d'excision en violet (barres violettes dans le track IGV de l'epiRIL 466 dans laquelle l'ET s'est excisé, séquence en violet). On notera le gain d'un nucléotide dans la *footprint* d'excision (4pb), en comparaison des 3 pb de long du TSD.

Pour autant, toute insertion de 3 à 5 pb détectée n'est pas nécessairement une *footprint* d'excision. Afin d'identifier parmi les 96 indels de cette taille les *footprints* d'excision "vraies", j'ai vérifié si leur séquence correspondait à la répétition directe du TSD à proximité, avec insertion d'au maximum 2 nucléotides supplémentaires. Pour les 82 indels vérifiant cette condition, j'ai également analysé si leur distribution au sein des différents états chromatiniens (présentés CHAPITRE 2) et des différentes epiRIL était cohérente avec celles des nouvelles insertions d'*ATENSPM3*. Comme illustré FIGURE 6.5, les *footprints* putatives laissées par la remobilisation d'*ATENSPM3* présentent tout comme les néo-insertions un enrichissement dans l'état CS2, qui se traduit par la présence de H3K27me3, marque déposée par les complexes Polycomb.
Par ailleurs (FIGURE 6.6), le nombre de *footprints* putatives est positivement corrélé au

nombre de néo-insertions d'*ATENSPM3* dans les différentes epiRIL.

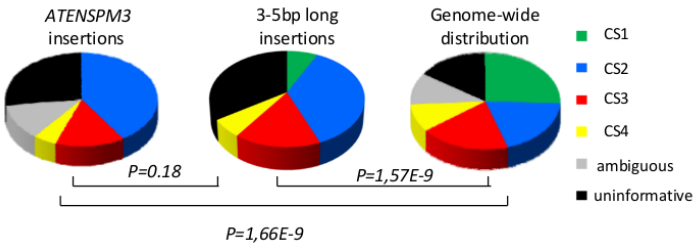


FIGURE 6.5 – Répartition dans les différents états chromatiniens des insertions d'*ATENSPM3* et des *footprints* d'excision putatives. Les états CS 1, 2, 3 et 4 sont définis d'après (Roudier, Ahmed et al. 2011), les états "ambiguous" et "uninformative" correspondent respectivement aux régions du génomes pour lesquelles des marques correspondant à plusieurs états distincts ont été détectées et pour lesquelles aucun enrichissement en une quelconque marque que ce soit est détecté. Les *P*-values sont issues d'un test du $\chi^2$.

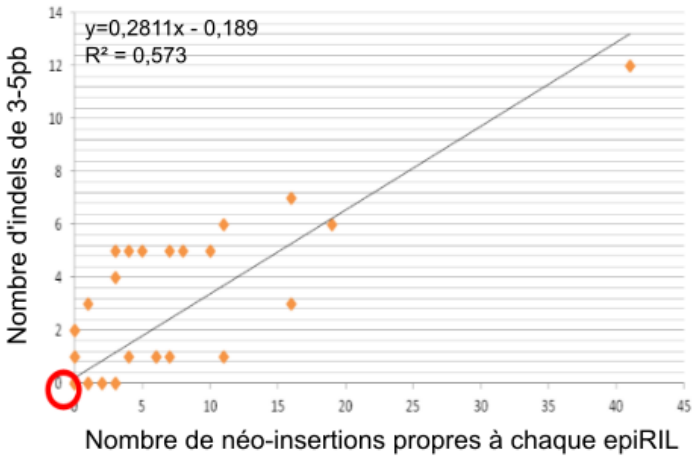


FIGURE 6.6 – Corrélation entre néo-insertions d'*ATENSPM3* et *footprints* d'excision putatives. Seules les néo-insertions propres à chaque lignée et présentes à l'état homozygote ont été considérées. Le cercle rouge indique les 41 epiRIL pour lesquelles ni insertion d'*ATENSPM3*, ni *footprint* putative d'excision n'est mise en évidence.

Au total, 82 indels correspondant à des *footprints* d'excision peuvent être identifiés, soient autant d'évènements additionnels de remobilisation d'*ATENSPM3* dans les epiRIL. Dans leur ensemble, ces données illustrent l'impact de la remobilisation extensive d'un ET de classe II sur le spectre des indels d'une population et proposent un proxy pour l'identification des cicatrices d'excision laissées par ce type d'ET. Les limites d'une telle approche sont discutées plus bas.

### 6.2.2 La remobilisation extensive des ET ne se traduit pas par d'importants réarrangements chromosomiques dans les epiRIL

En raison de l'importante remobilisation des ET dans les epiRIL, des réarrangements chromosomiques peuvent être attendus. Aussi, j'ai employé une combinaison d'approches basées sur la couverture du génome (*read-depth*), l'orientation des paires de reads (*read-pairs*) et la présence de reads recouvrant un breakpoint (*split-reads*) afin de documenter le patron de SV (autres que les nouvelles insertions d'ET) dans les epiRIL en comparaison des MA lines dans lequelles aucune remobilisation d'ET n'est mise en évidence (Figure 6.1).

La localisation chromosomique de ces différents SV est illustrée Figure 6.7.
Sur un ensemble de 92 epiRIL pour lesquelles la qualité du séquençage permet une détection de SV, un total de 15 CNV peut être identifié, parmi lesquels 8 délétions de 2 à 14 kb et 7 duplications de 5 à 238 kb ; contre une duplication de 200 kb et 3 délétions de 700 pb à 12 kb dans les MA lines. Ni translocation, ni inversion ne sont détectées dans aucune des deux populations. On notera par ailleurs que les délétions du locus donneur pour les ET de classe II remobilisés dans les epiRIL ne sont pas comptabilisées parmi les CNV.

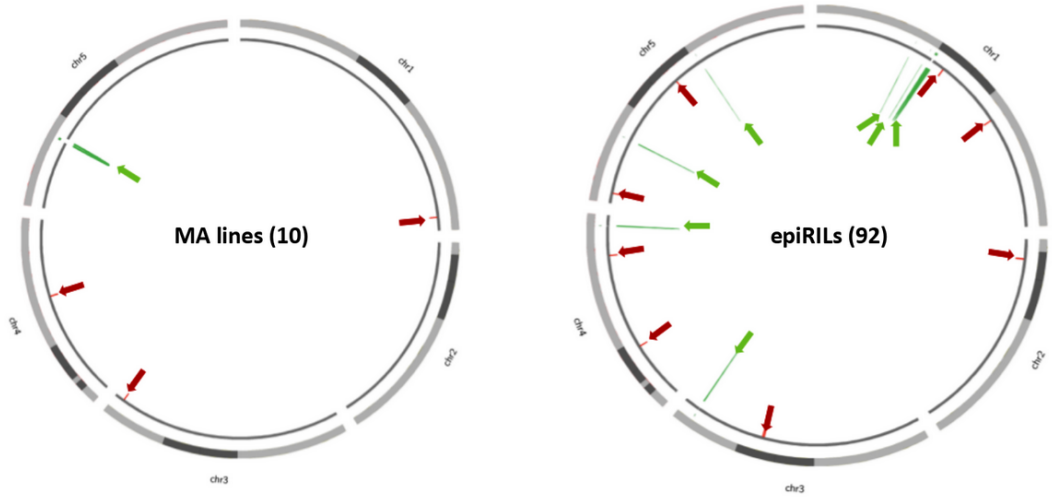


Figure 6.7 – Localisation des délétions et duplications le long des chromosomes dans les epiRIL et les MA lines. Les duplications sont figurées en vert et les délétions en rouge, les flèches de même couleur indiquent leur position. Les régions péricentromériques sont représentées en gris foncé, les bras en gris clair. Le nombre de lignées est donné entre parenthèses.

En raison du faible nombre total d'évènements, il n'est pas possible d'extraire ni de

dériver un taux de SV par génération, ni de tirer des conclusions fortes du patron de CNV dans ces deux populations. Néanmoins, quelques observations peuvent être faites.
En premier lieu, les duplications détectées correspondent systématiquement à des duplications en tandem, et pour les 5 duplications (1/1 parmi les MA lines, 4/7 parmi les epiRIL) pour lesquelles il est possible de définir précisément le site de cassure ; toutes sont situées au sein d'une séquence codante et présentent une courte région de microhomologie au breakpoint en l'absence de région d'homologie plus longue à proximité, suggérant une origine réplicative. Un exemple est illustré FIGURE 6.8. Il peut être souligné que la même observation a été faite dans le cadre d'une analyse détaillée des breakpoints de CNV chez l'Homme (Newman et al. 2015), mais à notre connaissance aucune description n'a été faite chez les plantes et en particulier chez Arabidopsis.

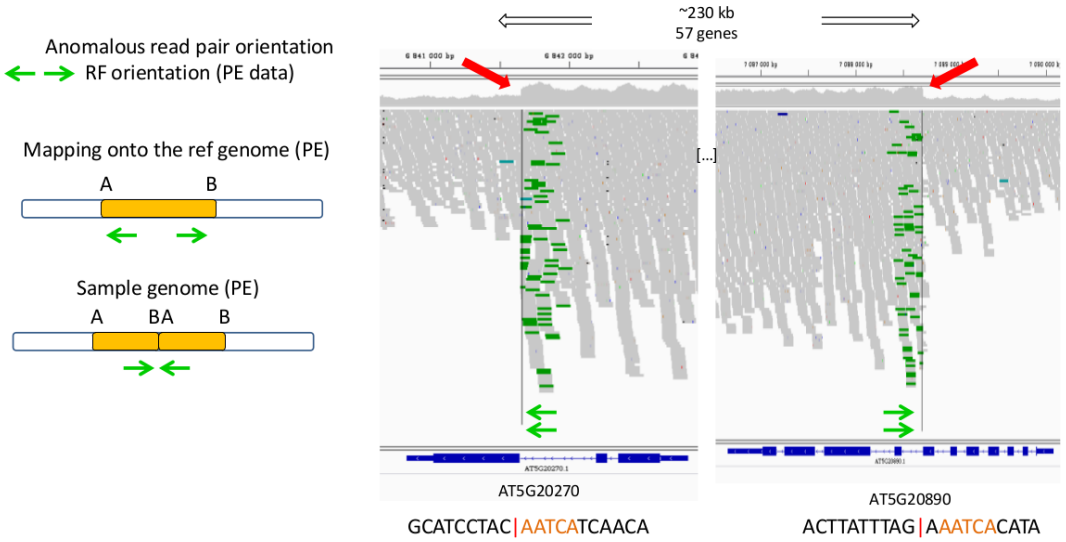


FIGURE 6.8 – Duplication en tandem représentative des 5 pouvant être analysées en détail au sein des MA lines et des epiRIL. Le panel de gauche décrit l'interprétation des orientations des paires de reads ayant conduit à la conclusion selon laquelle il s'agit d'une duplication en tandem (PE : *paired-end*). La copie d'écran IGV illustre la duplication de 200 kb identifiée dans la MA line 49. Les flèches rouges indiquent le gain de la couverture verticale dans ces régions, la séquence située en dessous représente la séquence retrouvée au breakpoint. La microhomologie est figurée en orange.

Concernant les délétions, on observe de façon surprenante qu'elles sont systématiquement associées des annotations d'ET (pleine longueur ou reliques) à la fois dans les MA lines et les epiRIL, avec dans cette dernière population l'observation additionnelle que les 4/7 délétions observées dans les régions péricentromériques se sont produites dans des lignées où ces régions sont dérivées de *ddm1* et donc globalement hypométhylées. A l'opposé, les 3 délétions identifiées dans les MA lines sont situées dans les bras des chromosomes.
Avec toutes les précautions qu'il s'agit de prendre en raison du faible nombre d'évène-

ments, il est tentant de rappeler qu'un niveau réduit de méthylation de l'ADN au niveau des séquences répétées a pu être associé à la formation de réarrangements chez les mammifères (Carbone et al. 2009 ; R. Chen et al. 1998 ; J. Li et al. 2012) tandis qu'un effet inhibiteur de la méthylation sur la recombinaison méiotique a été démontré chez Arabidopsis (Yelina et al. 2015) et *Ascobolus* (Maloisel et al. 1998).
En effet, en raison de la longueur des régions d'homologie et de la distance les séparant, il est probable que les différentes délétions observées aient été produites par NAHR, donc par recombinaison.
Dès lors, il serait attrayant de pouvoir conclure que la présence de délétions dans les régions péricentromériques exclusivement lorsque celles-ci sont dérivées de *ddm1* soit la conséquence d'une levée de l'inhibition médiée par la méthylation de l'ADN sur la recombinaison.

Du plus, alors que la NAHR peut aussi bien produire des délétions que des duplications, tous les CNV associés à des ET aussi bien dans les MA lines que dans les epiRIL ont été résolus en délétions et aucun en duplication, une observation qui, comme discuté plus bas, peut être mise en regard de travaux précédents ayant mis en évidence une tendance chez Arabidopsis à une réduction du génome au travers d'un biais vers la délétion des ET (Devos et al. 2002).

### 6.2.3 Manuscrit associé

Le manuscrit joint dans les pages qui suivent présente les résultats de l'analyse extensive de la remobilisation des ET dans les epiRIL, réalisée par Leandro Quadrana et Mathilde Etcheverry (Quadrana, Etcheverry et al. 2018). Mes travaux concernant l'impact mutationnel des ET se basent sur l'identification préalable des néo-insertions, décrites dans cet article.
Ma contribution à ce manuscrit réside dans la mise en évidence des évènements d'excision d'*ATENSPM3* qui y sont exploités.

# Transposon accumulation lines uncover histone H2A.Z-driven integration bias towards environmentally responsive genes

Leandro Quadrana[1,6,*], Mathilde Etcheverry[1,6], Arthur Gilly[2,4], Erwann Caillieux[1], Mohammed-Amin Madoui[3], Julie Guy[2], Amanda Bortolini Silveira[1,5], Stefan Engelen[2], Victoire Baillet[1], Patrick Wincker[3], Jean-Marc Aury[2] and Vincent Colot[1,*]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Ecole Normale Supérieure, PSL Research University, Paris, France. [2]Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, F-91057 Evry, France. [3]Génomique Métabolique, Genoscope, Institut de biologie François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France. [4]Present address: Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [5]Present address: Translational Research Department, Institut Curie, PSL Research University, Paris, France. [6]These authors contributed equally

*Correspondence: quadrana@biologie.ens.fr (L.Q.); colot@biologie.ens.fr (V.C)

**Inherited transposition events are important drivers of genome evolution but because transposable element (TE) mobilization is usually rare, its impact on the creation of genetic variation remains poorly characterized. Here, we used a population of *A. thaliana* epigenetic recombinant inbred lines (epiRILs) to characterize >8000 de novo insertions produced by several TEs families also active in nature. Integration was strongly biased towards genes, with evident deleterious effects. Biases were TE family-specific and associated with distinct chromatin features. Notably, we demonstrate that the histone variant H2A.Z guides the preferential integration of *Ty1/Copia* LTR-retrotransposons within environmentally responsive genes and that this guiding function is evolutionary conserved. Finally, we uncover an important role for epigenetic silencing in exacerbating or alleviating the effects of TE insertions on target genes. These findings establish chromatin as a major determinant of the spectrum and functional impact of TE-generated mutations, with important implications for adaptation and evolution.**

## INTRODUCTION

Transposable elements (TEs) are sequences that self-propagate and accumulate to various levels in the genome of eukaryotic species. TEs fall into two broad classes, depending on their mechanism of transposition: DNA transposons, which move through a "cut and paste" mechanism and Long Terminal Repeat (LTR) as well as non-LTR retrotransposons, which move through an RNA intermediate that is reverse transcribed. The content of the different classes and subclasses of TEs varies markedly across species. Thus, mammalian genomes are laden with non-LTR retrotransposons whereas plant genomes are populated by LTR retrotransposons and DNA transposons mainly (Lisch 2013; Chuong, Elde, and Feschotte 2016). However, most of these sequences are relics of once active TEs and although TEs are arguably among the main drivers of the evolution of genome size, organization and

function (Chuong et al., 2016), we still know little about their contribution to the generation of somatic or heritable mutations in extant genomes (Huang, Burns, and Boeke 2012; Sultana et al. 2017). This situation stems in large part from the fact that transposition is typically rare in nature, notably because of the strong epigenetic silencing mechanisms, such as DNA methylation in plants and mammals, that target TEs (Slotkin and Martienssen 2007).

With the advent of population genomics, it is now possible to exploit natural variation in the number and distribution of recent TE insertions to gain information about the set of TEs likely mobile in extant genomes. However, because of the deleterious effects typically associated with TE mobilization, natural selection prevents the proper assessment of the complete range of the mutations generated by transposition (Huang, Burns, and Boeke 2012; Sultana et al. 2017; Quadrana et al. 2016).

Here, we have exploited a population of *A. thaliana* epigenetic Recombinant Inbred Lines (epiRILs; Johannes et al., 2009) that are operationally similar to mutation accumulation (MA) lines (Ossowski et al. 2010; Zhu et al. 2014; Denver et al. 2009; Keightley et al. 2009), but in which transposition was kick-started for several TEs, to obtain a first comprehensive assessment of the actual rate, spectrum and genome-wide distribution of heritable mutations generated by TEs. The epiRILs were derived from a cross between a wild type and a near isogenic parent that contains thousands of transcriptionally active TE sequences because of compromised DNA methylation (Johannes et al., 2009; Figure 1A). Based on whole genome sequencing for more than 100 epiRILs and the characterization of >8000 *de novo* heritable insertions, we show that two DNA transposons and one LTR-retrotransposon target non-overlapping sets of genes, which are characterized by distinct chromatin features. We further demonstrate that the preferential targeting of environmentally responsive genes by LTR-retrotransposons of the *Ty1/Copia* superfamily is guided by the histone variant H2A.Z. We also provide evidence that this guiding function is ancestral and we propose that the contrasted fate of *Ty1/Copia* in plants and animal genomes results from the differential incorporation of this histone variant in environmentally responsive and developmentally regulated genes in these two groups of organisms, respectively. Finally, we showed that epigenetic silencing of new insertions soon after their occurrence can either alleviate or exacerbate their effects on gene transcription. Our findings illustrate the role of chromatin as a major determinant of the spectrum and impact of TE-generated mutations, with important implications for adaptation as well as the evolutionary success of TEs and the organisms in which they reside.

## RESULTS

### TE insertions accumulate in the epiRILs

We produced high quality whole genome sequencing data for 107 epiRILs at generation F8 as well as of close relatives of the wild type and *ddm1* parental lines (Figure 1A) using Illumina mate-pair libraries built from size-selected genomic fragments of ~5.5 kb. Mate-pair reads were mapped to the reference genome sequence (accession Col-0) and non-reference (i.e. *de novo*) insertions were detected based on mapping discordance between mate-pair reads and using also split reads (Gilly et al., 2014; Quadrana et al., 2016; Figure S1A). Thanks to the large physical distance between mate-pair reads, the complete sequence of the *de novo* insertions was also determined, thus enabling the identification of the exact or most likely donor TE in each case. Consistent with the low frequency of TE mobilization in nature (Quadrana et al. 2016) there were no *de novo* TE insertions in the two wild type siblings sequenced nor in five *A. thaliana* MA lines (Ossowski et al. 2010). In contrast, two non-reference TE insertions were detected in the

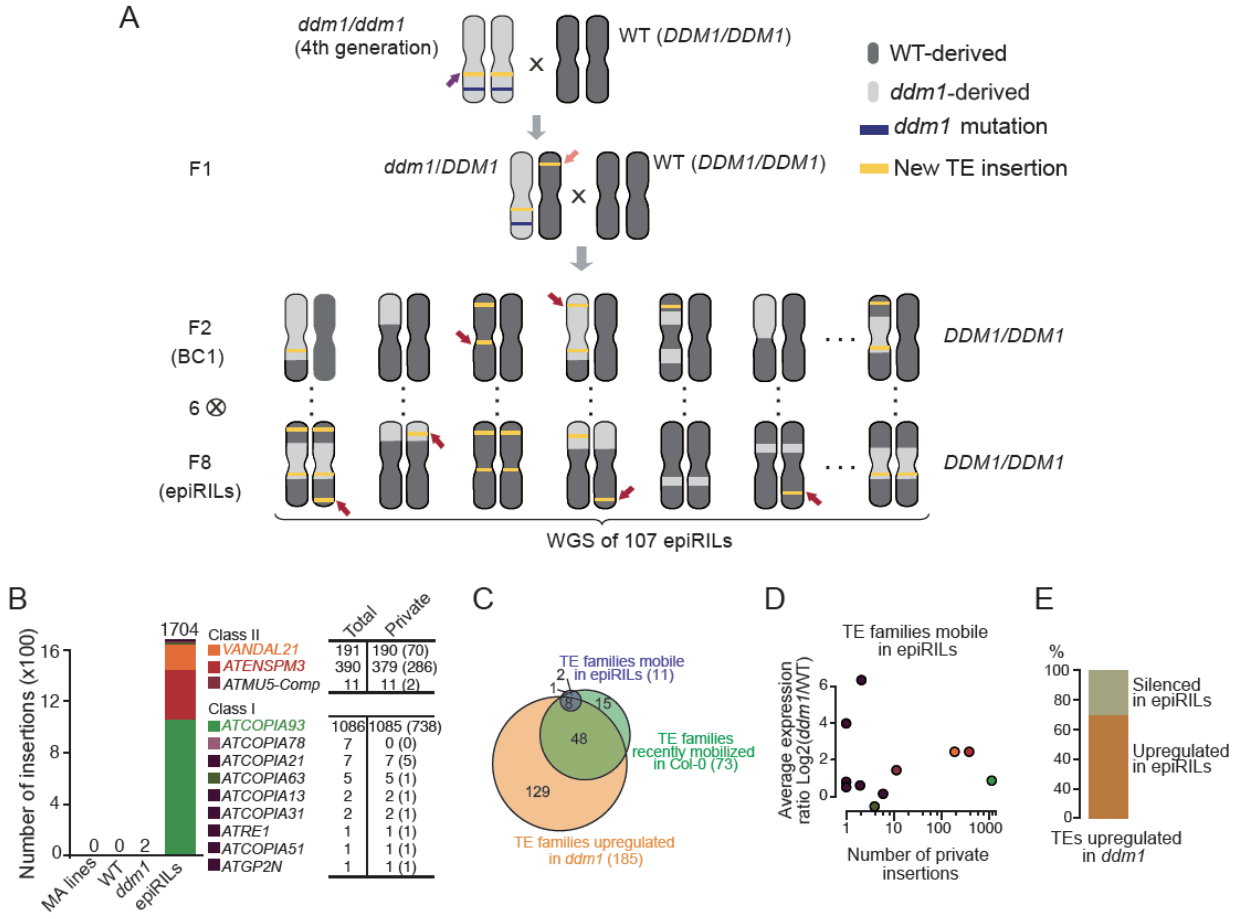


**Figure 1. TE insertions accumulate in the epiRILs.**

**A**. Crossing scheme used to generate the epiRIL population. **B**. Number and identity of insertions accumulated in wild-type, *ddm1* and the epiRILs. Number of heterozygous insertions are indicated within brackets. **C**. Overlap between TE families transcriptionally reactivated in *ddm1,* those potentially mobile in Col-0 and those that transposed in at least one epiRIL. **D**. Relation between number of private TE insertions and average expression ratio (*ddm1*/wild-type) for the 11 TE families that transposed in the epiRILs. **E**. Proportion of reference TE sequences upregulated in *ddm1* and either silenced or stably upregulated in the epiRILs.

sequenced *ddm1* individual and many more in the epiRILs. Specifically, 95% of the 107 epiRILs harbored between 1 and 97 heritable *de novo* insertions and only 8 had none (Figure S1B). Almost all (98.7%) of these insertions were private and therefore did not occur in the *ddm1* parent but rather either in the inflorescences of the F1 individual used for the backcrossing or during the propagation of the epiRILs, starting from the F2 (Figure 1A). In fact, 1107 out of the 1670 private insertions detected in total were heterozygous (Figure 1B, S1B), which is strong evidence that transposition was ongoing in the epiRILs. Given that they were propagated by repeated selfing and single seed descent, the epiRILs could therefore be defined as TE accumulation (TEA) lines. In other words, notwithstanding the possibility of excision in the case of DNA transposons (see below), TE insertions that occurred during the propagation of the epiRILs and that were potentially heritable should be lost or fixed in a neutral fashion through random segregation, except for those disrupting essential genes.

Overall, we detected private insertions for two DNA transposon and eight retrotransposon families as well as for three composite DNA transposons (Figure 1B). These composite elements were uncovered thanks to our mate-pair

sequencing approach and are made up for two of them of a protein-coding gene of unknown function flanked by two distinct truncated copies of *ATENSPM3*. The third composite DNA transposon is more complex and contains several truncated TEs, the largest of which is related to the *ATMU5* family (Figure S1C). The number of transposition events varied considerably among the eleven TE families, with the two DNA transposons *VANDAL21* and *ATENSPM3* (together with the two composite *ATENSPM3* elements) and the LTR-containing retrotransposon *ATCOPIA93* contributing 11.2%, 22.5% and 64.4% of the *de novo* insertions, respectively (Figure 1B).

We showed previously that 73 of the 326 TE families annotated in the *A. thaliana* genome transposed in the recent past in the reference accession Col-0 (Quadrana et al. 2016), which was used to derive the epiRIL population. Consistent with this finding the 10 annotated TE families with private insertions in the epiRILs are also mobile in nature and belong to the Col-0 mobilome. Nonetheless, the epiRIL mobilome is only a small subset of that of Col-0, which indicates that the widespread loss of DNA methylation induced by *ddm1* does not translate in an equally widespread remobilization of TE families. To determine if differences in the level of *ddm1*-induced transcriptional reactivation between TE families could explain at least in part the limited mobilome of the epiRILs, we performed RNA-seq on whole seedlings for five epiRILs as well as one sibling of the two parents. Between 731 and 1056 (considering unique or multiple mapping reads, respectively) TEs, mostly full-length, were significantly overexpressed at least 2-fold ($P$<0.05) in *ddm1* compared to wild type, consistent with previous RNA seq data sets (Figure S2). Moreover, >75% of the 73 TE families that compose the Col-0 mobilome were upregulated in *ddm1*, including eight of the 11 TE families that are part of the epiRIL mobilome (Figure 1C). However, levels of upregulation in *ddm1* did not correlate with the number of private insertions in the epiRILs (Figure 1D). Finally, 72% of the *ddm1*-upregulated TEs for which expression could be ascertained unambiguously in the epiRILs remained transcriptionally active when derived from the *ddm1* parent (Figure 1E). Thus, levels of transcription in seedlings did not reflect transposition activity. Whether this also holds true in the cells, starting from the zygote, that will ultimately pass on their DNA to the next generation remains to be determined, as these cells are where transposition must occur for insertions to be inherited.

**Invasion dynamics differ between TE families**

The large number of private insertions for *VANDAL21*, *ATENSPM3* and *ATCOPIA93* enabled us to investigate in depth for each of these three TE families their mode and tempo of mobilization across generations. Mate-pair reads were first used to determine the sequence of the private insertions. In the case of *VANDAL21* and *ATCOPIA93,* almost all (99%) were identical to a single one of the 11 and four full-length copies present in the reference genome, respectively. This single copy is the same that was reported previously to be mobile in *ddm1* or another DNA methylation mutant background (Fu et al. 2013; Mirouze et al. 2009). In contrast, the identity of the *ATENSPM3* private insertions was more diverse, with 58% matching a single one of the nine full-length copies and the remaining private insertions matching either one of the two composite copies also present in the reference genome (Figure 2A). Each of these two composite *ATENSPM3* copies contains a single gene of unknown function with no similarity to any known transposase gene, indicating that they were likely mobilized in *trans*, presumably using the transposase

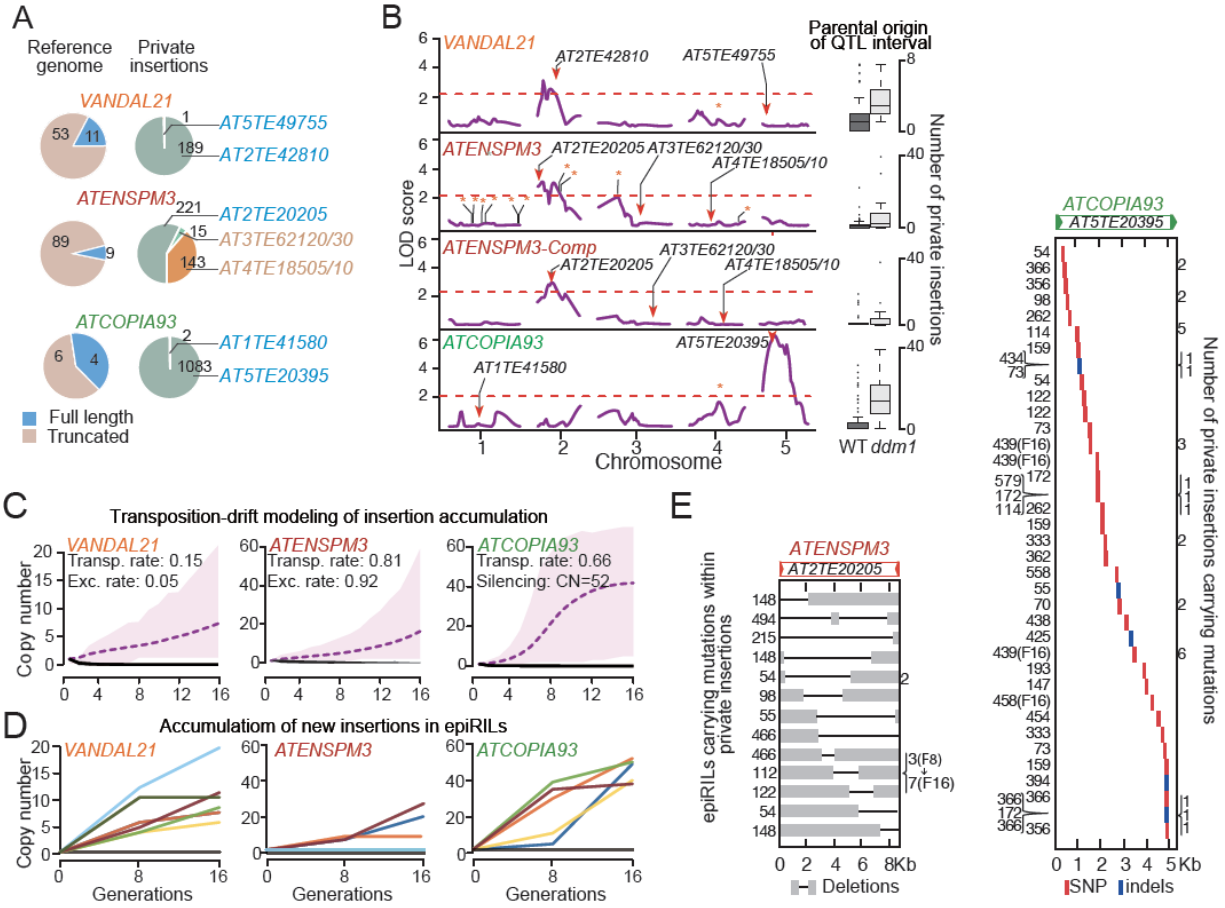


**Figure 2. Invasion dynamics differ between TE families**

**A.** Identification of donor copies for each transposition event. The number of truncated and full-length copies annotated in the *A. thaliana* reference genome is also shown. **B**. QTL mapping results for the mobilization of *VANDAL21, ATENSPM3,* the two composite *ATENSPM3* and *ATCOPIA93*. TEs annotated in the reference *A. thaliana* genome as well as donor and shared insertions are indicated by bars, arrowheads and stars, respectively. Box-plots on the right indicate for each TE family the numbers of insertions in the epiRILs in relation to the parental origin of the relevant QTL interval. **C.** Predicted dynamics of insertion accumulation for the three TE families based on the best fitted transposition-drift model. Estimates of the transposition and excision rates as well as of the number of TE insertions required for triggering concerted epigenetic silencing are indicated. **D.** Number of *VANDAL21, ATENSPM3,* and *ATCOPIA93* insertions observed at generation F8 and F16 in ten epiRILs. **E**. Mutations detected within private insertions of *ATENSPM3* and *ATCOPIA93*. Carrier epiRILs as well as the number of private insertions with the mutation are indicated in each case on the left and right side of each panel, respectively.

encoded by the mobile full-length *ATENSPM3* reference copy. Thus, for each of these three TE families, a single *ddm1*-derived reference element is at the origin of most transposition events that accumulated in the epiRILs.

To confirm this conclusion and explore in more detail the genetic architecture underlying the large variation in the number of private insertions accumulated in the epiRILs, we performed a quantitative trait locus (QTL) analysis for each of the three TE families. We took advantage of the fact that the epiRILs are virtually isogenic but segregate hundreds of parental differentially methylated regions (DMRs) that can be used as *bona fide* genetic markers (Colomé-Taché et al, 2012; Cortijo et al., 2014) to identify loci whose epigenetic state is associated with transposition activity. Using the number of private insertions as the quantitative trait, this approach revealed a single QTL interval for each TE family, which included the full-length reference copy identified in the previous step (Figure 2B). Furthermore, the same QTL interval was obtained whether all *ATENSPM3* private insertions were considered or only those that match the two composite *ATENSPM3* present in the reference genome, thus demonstrating that these

were mobilized in t*rans* by the single full-length, *ddm1*-derived reference copy of *ATENSPM3*.

To determine if these observations were compatible with a transposition-drift scenario, we modeled TE mobilization for each of the three TE families based on three key parameters. These are the rate of transposition as well as of excision in the case of the DNA transposons, as this is an integral part of their mechanism of transposition, and copy-number dependent inhibition of transposition, as *ATCOPIA93* was shown to be epigenetically silenced when reaching around 40 copies (Marí-Ordóñez et al. 2013). For each TE family we selected the model that produced the best fit to the observed number and ratio of homozygous vs heterozygous private insertions at generation F8 and ran each model for another eight generations (F16; Figure S3A). The selected models predicted that in addition to the original reference TE donor, all new insertions are equally able to transpose, at a rate of 0.15, 0.81 and 0.66 new copy/per donor/per generation for *VANDAL21*, *ATENSPM3* and *ATCOPIA93,* respectively (Figure 2C and S3B). The best-fit models also predicted low and high rates of excision for *VANDAL21* and *ATENSPM3*, as was observed in maize for TEs belonging to the same two superfamilies *Mu* and *En/Spm*, respectively (Doseff, Martienssen, and Sundaresan 1990; Masson et al. 1987). In agreement with these predictions, the mobile reference copy of *VANDAL21* was almost always retained in the epiRILs, unlike that of *ATENSPM3*, which was systematically missing from the corresponding *ddm1*-derived chromosome interval (Figure S3C). Furthermore, indels compatible with *ATENSPM3* excision footprints were consistently detected across the genome and their number correlated positively with that of *ATENSPM3* private insertions (Figure S3D). Finally, our modeling predicted that transposition of *ATCOPIA93* should stop when reaching 52 copies (Figure 2C and S3B). To evaluate further the predictive power of our models, we sequenced ten epiRILs that were propagated for another 8 generations (F16). Consistent with numerical simulations (Figure 2C), *VANDAL21* and *ATENSPM3* continued to transpose in most of the lines with initial transposition activity, with no evidence of copy-number-dependent transposition inhibition. In contrast, although *ATCOPIA93* continued to transpose beyond F8, none of the F16 lines had accumulated more than 50 copies (Figure 2D). Furthermore, almost all copies were in the homozygous state at F16, consistent with the concerted epigenetic silencing expected at this copy number (Marí-Ordóñez et al. 2013).

Last, we took advantage of the fact that mobilized TEs tend to mutate during the transposition process to obtain direct evidence that in addition to the initial *ddm1*-derived donor copy, new insertions were also mobilized in the epiRILs. While all new insertions were identical to the mobile reference element for *VANDAL21*, approximately 5% of *ATENSMP3* and *ATCOPIA93* new insertions identified at F8 and/or F16 contained large internal deletions and point mutations or small indels, respectively (Figure 2E). Some mutations were carried by more than one new insertion within an epiRIL, but rarely between epiRILs. Notably, for one of the 10 epiRILs also sequenced at F16, we observed an increase in the number of *ATENSPM3* insertions containing the same mutation at that later generation (Figure 2E). These findings provided therefore further evidence that newly inserted copies were also mobilized in the epiRILs.

**TEs exhibit strong and diverse chromatin-associated insertion biases towards genes**

As in nature (Quadrana et al. 2016), private insertions for *VANDAL21, ATENSPM3* and *COPIA93* were distributed evenly across the five chromosomes in the epiRILs, overall (Figure 3A). However, because the genomes of the epiRILs are epigenetic mosaics, each being composed on average of 25% *ddm1*- and 75% wild-type-derived

segments (Colomé-Tatché et al, 2012), we searched for potential differences in insertion frequency associated with parental origin. For all three TE families, the percentage of insertions in *ddm1*-derived intervals was slightly higher than expected by chance at the whole genome level but much higher (35-55%) when considering the pericentromeric regions only (Figure 3A and S4A), which lose their heterochromatic features in *ddm1* as well as in subsequent generations (Soppe et al. 2002; Lippman et al. 2004; Colome-Tatche et al. 2012). These findings reinforced the notion that euchromatin is the preferred substrate for the integration of *VANDAL21, ATENSPM3* and *COPIA93*.

Given that the methylome as well as the transcriptome of wild-type-derived intervals in the epiRILs are with few exceptions identical to their wild type parental counterparts (Colome-Tatche et al., 2012; Ito et al., 2015; Figure S2), we then used the private insertions located within these intervals to obtain information on integration preferences. Insertions were strongly overrepresented within or adjacent to genes for the three TE families (Figure 3B). However, while the fraction of essential genes (Lloyd et al. 2015) that were targeted was that expected by chance for *VANDAL21*, it was much lower for *ATENSPM3* and *ATCOPIA93*, suggesting that their integration, unlike that of *VANDAL21*, strongly affects the expression of target genes. Consistent with this interpretation, the fraction of essential genes targeted by *ATCOPIA93* was further reduced when only considering insertions in the homozygous state (Figure S4B). Unexpectedly though, we found an opposite pattern for *ATENSPM3* insertions within essential genes, which were less frequent in the heterozygous than the homozygous state. Given the high rate of excision associated with *ATENSPM3* mobilization, this last result suggested stronger deleterious effects after excision of *ATENSPM3*. Consistently, none of the 98 homozygous excision footprints detected in the epiRILs were located within essential genes (Figure S4B).

Using the nine main chromatin states defined in *A. thaliana* based on epigenomic maps obtained from diverse organs and tissues (Sequeira-Mendes et al. 2014), we found that *VANDAL21* mainly targeted promoters and 5' UTRs of genes characterized as active (chromatin states 1&2; Figure 3C). In contrast, *ATENSPM3* and *ATCOPIA93* preferentially inserted within or close to genes that are typically marked by H3K27me3 and slightly enriched for the variant histone H2A.Z (chromatin state 5) and, in the case of *ATCOPIA93,* also in or adjacent to genes that were only enriched for that histone variant (chromatin state 6; Figure 3C).

Meta-analysis of insertion sites confirmed these findings and revealed that *VANDAL21* integration within wild-type-derived intervals tended to coincide with peaks of DNase I hypersensitivity (Figure 3D) at the transcriptional start site (TSS) of genes (Figure S4C). Furthermore, insertions were typically in the same orientation as the target genes (Figure S4C). These results support previous observations reported for *VANDAL21* in *ddm1* mutant plants (Fu et al., 2013) as well as in nature (Quadrana et al. 2016). In contrast, insertion sites were enriched for H3K27me3 in the case of *ATENSPM3* and for nucleosomal DNA as well as H2A.Z in the case of *ATCOPIA93* (Figure S4D). However, *ATCOPIA93* did not insert preferentially within well-positioned nucleosomes (Figure S4E; Lyons and Zilberman, 2017) nor within so-called "+1 nucleosomes", which tend to incorporate H2A.Z (Henikoff and Smith 2015), but rather within genes (Figure S4F) that are enriched in *H2A.Z* throughout their body (Figure S4G). Such pattern of H2A.Z deposition is specific to plants and was shown previously to concern mainly responsive genes (Coleman-Derr and Zilberman 2012; Zahraeifard et al. 2018). In agreement, this class of genes was strongly overrepresented among *ATCOPIA93* targets (Figure 3E).

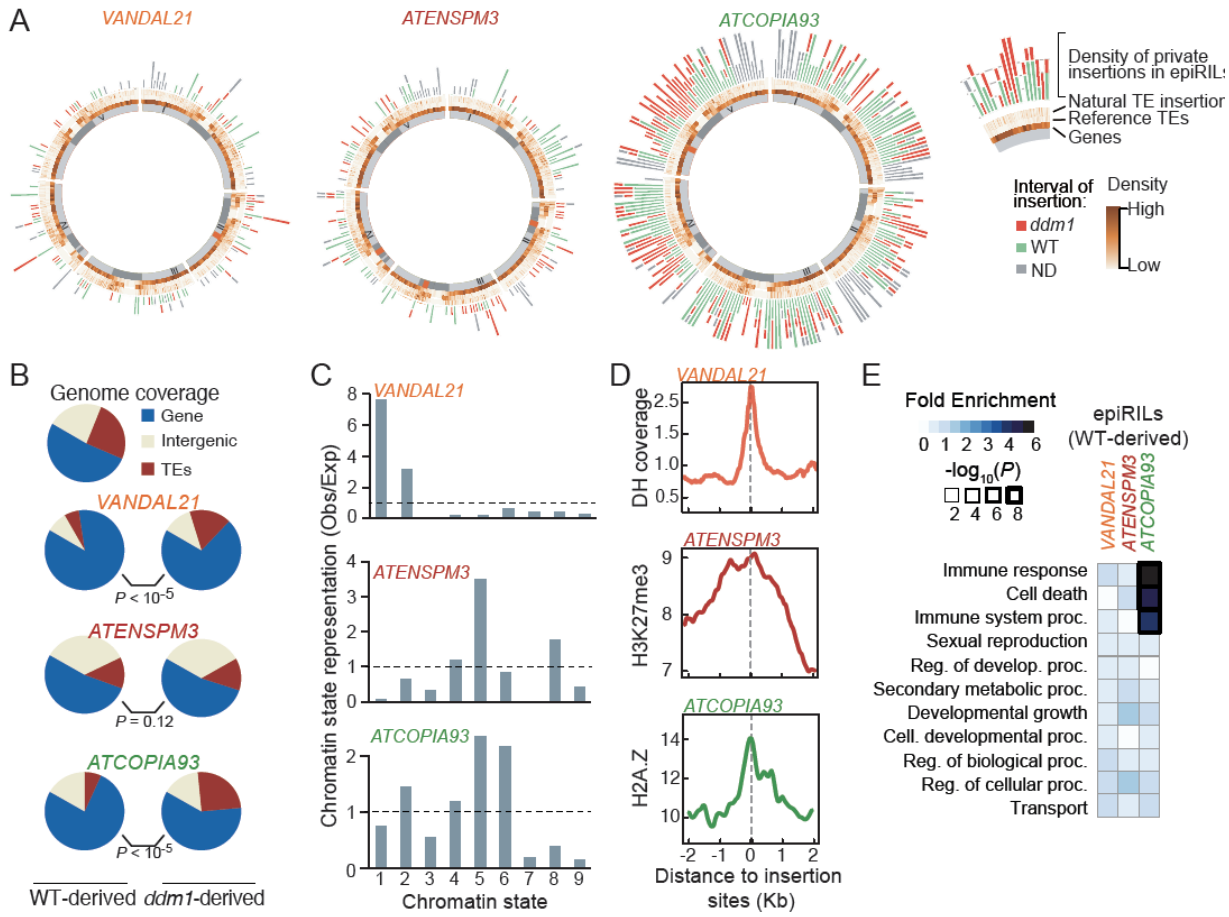


**Figure 3. TEs exhibit strong and diverse chromatin-associated insertion biases towards genes.**

**A.** Circos representation of private TE insertions detected for *VANDAL21*, *ATENSPM3* and *ATCOPIA93*. The exterior circle represents the density of private insertions within wild-type- and *ddm1*-derived intervals. Density of non-reference (natural) TE insertions, reference TEs as well as genes are represented inwards in this order. **B.** Fraction of TE insertions in wild-type- and *ddm1*-derived intervals in genes, TEs and intergenic regions. Statistical significance for each comparison was obtained using the Chi-square test. **C.** Observed/expected ratio (Obs/Exp) of insertion sites within wild-type-derived regions in relation to the nine chromatin states defined in *A. thaliana*. Error bars represent the 95% confidence interval obtained by 1000 boots-traps. **D.** Coverage of DNAse hypersensitivity (DH; top panel), H3K7me3 (middle panel) and H2A.Z (bottom panel) around private insertion located within wild-type-derived regions for *VANDAL21*, *ATENSPM3* and *ATCOPIA93*, respectively. **E.** GO term analysis of genes with private TE insertions in the epiRILs.

### H2A.Z directs the integration of *ATCOPIA93*

The *Ty1/copia* superfamily of LTR-retrotransposons is one of the main contributors of genome size inflation seen in many plant species (Huang, Burns, and Boeke 2012). We therefore explored further the function of chromatin in the integration of *ATCOPIA93* and first noted that the proportion of insertions within TE sequences was much higher in *ddm1*-derived than in wild-type-derived intervals (Figure 3B). This finding is entirely consistent with the fact that TE sequences tend to acquire H2A.Z when hypomethylated (Zilberman et al. 2008). Indeed, H2A.Z enrichment levels measured in the hypomethylated background *met1* that was used for this previous analysis were on average much higher, in *ddm1*-derived intervals, for TE sequences targeted by *ATCOPIA93* than for those that were not (Figure S4H).

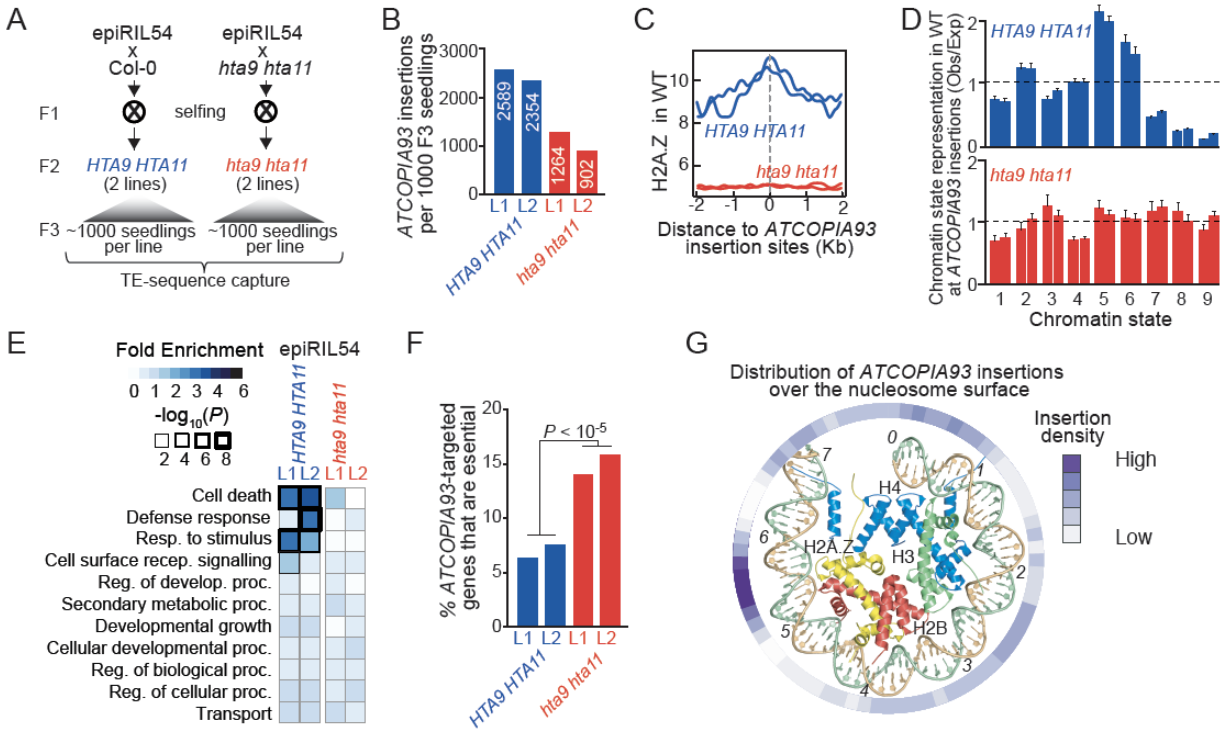


**Figure 4. H2A.Z directs the integration of *ATCOPIA93*.**

**A**. Experimental strategy for determining the role of H2A.Z in the integration of *ATCOPIA93*. epiRIL54 was crossed with wild type and the double or double mutant *hta9 hat11*. Two wild type (*HTA9 HTA11* L1 and L2) and two double mutant (*hta9 hta11* L1 and L2*)* F2 plants were selected and 1000 F3 seedlings were collected in each case to perform TE-sequence capture. **B**. Number of new insertions detected in 1000 *HTA9 HTA11* and *hta9 hta11* F3 seedlings. **C**. Meta analysis of levels of H2A.Z around *ATCOPIA93* insertion sites. **D**. Observed/expected ratio (O/E) of insertion sites in relation to the nine chromatin states defined in *A. thaliana*. Error bars represent the 95% confidence interval obtained by 1000 boots-traps. **E**. GO term analysis of genes containing new *ATCOPIA93* insertions in *HTA9 HTA11* or *hta9 hta11*. **F.** Fraction of essential genes among those targeted by *ATCOPIA93* in *HTA9 HTA11* or *hta9 hta11* F3 seedlings. Statistically significant differences were calculated using the chi-square test. **G**. Density of *ATCOPIA93* insertions over the surface of mouse H2A.Z containing nucleosomes (PDB 1F66). Only 73bp of DNA and associated proteins are viewed down the superhelical dyad and each number (0-7) represents one DNA double helix turn, starting from the central base pair. The exterior circle shows the density of new *ATCOPIA93* insertions per base pair.

To demonstrate the involvement of H2A.Z in the integration of *ATCOPIA93,* we crossed one epiRIL (epiRIL54, F8) containing 23 active (mainly heterozygous) *ATCOPIA93* copies to a wild type or an *hta9 hta11* double mutant parent, which lacks most H2A.Z (March-Díaz et al. 2008). Two F1 individuals were selfed in each case and two homozygous wild type as well as two homozygous double mutant F2 lines were selected (Figure 4A) to produce F3 progeny for TE-sequence capture (Quadrana et al. 2016). DNA was extracted from approximately 1000 F3 seedlings from each of the four F2 lines. Over 2000 new *ATCOPIA93* insertions were detected in each of the two wild type F3 progenies and less than twice that number in the two *hta9 hta11* F3 progenies (Figure 4B). In addition, while the strong *ATCOPIA93* integration preferences observed in the epiRILs was confirmed in wild type F3 seedlings, they were totally abolished in the *hta9 hta11* lines (Figure 4C, 4D, 4E). Moreover, the proportion of essential genes targeted by *ATCOPIA93* almost tripled in the double mutant (Figure 4F). These results demonstrated that H2A.Z acts at two levels, to promote *ATCOPIA93* retrotransposition and to guide integration within environmentally responsive genes.

To investigate at the nucleosome-scale the guiding function of H2A.Z, we crossed the list of the *ATCOPIA93* insertion sites detected in the wild type F3 progeny with the list of well-positioned nucleosomes previously produced for *A. thaliana* (Lyons and Zilberman 2017). A major peak of integration was observed ~55 bp away from the nucleosome

dyad (Figure 4G). This position is a main point of contact between DNA and H2A or H2A.Z and it is also where these two histones differ by several amino acids (Suto et al. 2000; Zlatanova and Thakar 2008). These findings further support a direct role of H2A.Z in guiding *ATCOPIA93* integration.

**H2A.Z-directed integration is evolutionarily conserved**

Unlike *Ty3/Gypsy* LTR-retrotransposons, those of the *Ty1/Copia* superfamily tend to insert within euchromatin (Sultana et al. 2017). To determine if the guiding function of histone H2A.Z is evolutionary conserved, we first focused on *ATCOPIA78,* which is distantly related to *ATCOPIA93* (Figure 5A). There were no private *ATCOPIA78* insertions in the epiRILs, but previous work showed that *ATCOPIA78* can be transcriptionally reactivated by heat stress and mobilized if stressed plants are defective in RNA-directed DNA methylation, such as in the *nrpd1* mutant background (Ito et al., 2011). We therefore assessed the mobilization of *ATCOPIA78* in pools of *nrpd1* seedlings that were subjected or not to heat stress and subsequently grown under normal conditions to produce seeds. One thousand F1 seedlings from each pool were grown under normal conditions and used to perform TE-sequence capture (Figure 5B). A total of 279 *ATCOPIA78* insertions were recovered in the progeny of heat-stressed plants, compared to only two in the progeny of non-stressed plants (Figure 5C). Insertion preferences were similar to those observed with *ATCOPIA93* (Figure 5D, E, F).

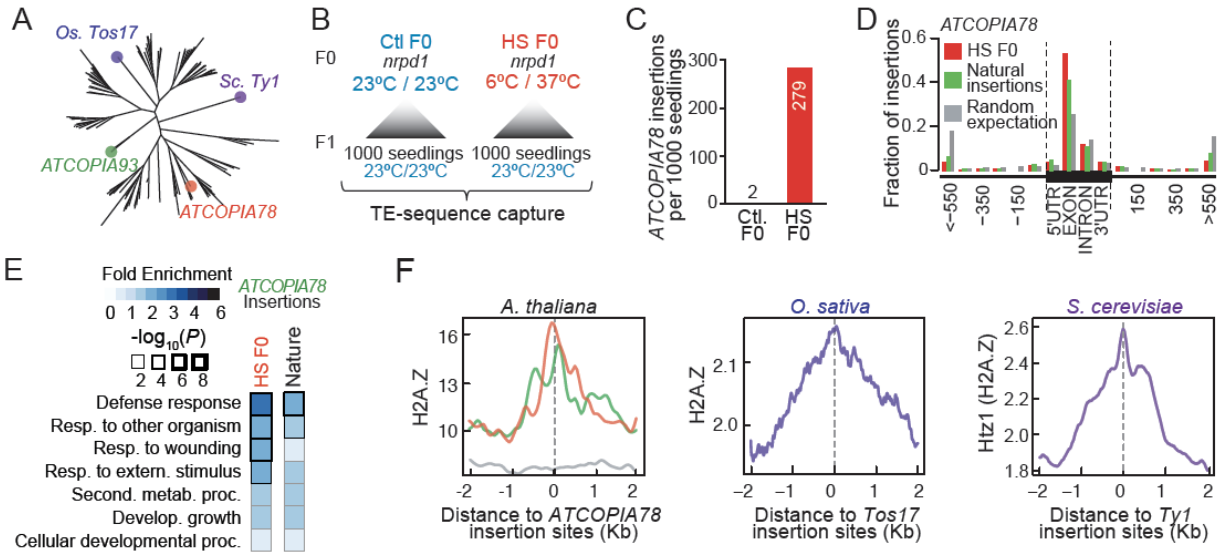


**Figure 5. H2A.Z-directed integration of *Ty1/Copia* retrotransposons is evolutionarily conserved.**

**A**. Phylogenetic analysis of *Ty1/Copia* LTR-retrotransposons from *A. thaliana* as well as *Tos17* and *Ty1* from rice and budding yeast, respectively. **B**. Experimental strategy for studying transposition landscape in *A. thaliana* of the heat-responsive *ATCOPIA78* LTR-retrotransposon. 1000 *nrpd1* F1 seedlings derived from plants grown under control conditions (Ctl F0) or subjected to heat-stress (HS F0) were subjected to TE-sequence capture. **C**. Number of new insertions detected in Ctl F0 or HS F0. **D**. Metagene analysis showing the distribution of new *ATCOPIA78* insertions detected in heat-stressed *nrpd1* mutant plants (HS F0) or in natural population of *A. thaliana* (Natural insertions). UTR, untranslated transcribed region. **E**. GO term analysis of genes containing new *ATCOPIA78* insertions in heat-stressed *nrpd1* mutant plants (HS F0) or in natural population of *A. thaliana* (Natural insertions). **F**. Metanalysis of *A. thaliana*, rice (*O. sativa*) and yeast (*S. cerevisiae*) H2A.Z levels around *ATCOPIA78*, *Tos17* and *Ty1* insertion sites, respectively. For *A. thaliana*, experimental and natural insertions are depicted in green and red, respectively.

Unlike for *ATCOPIA93*, numerous recent *ATCOPIA78* insertions were found in nature (Quadrana et al. 2016). Although patterns of integration were similar in the experimental and natural settings (Figure 5D, E and F), the

purifying effect of selection was already evident in nature. Specifically, the fraction of natural private insertions within exons was much reduced compared to that found experimentally (Figure 5D). Similarly, none of the 147 recent natural insertions examined were within essential genes, compared to 2.8% of the *ATCOPIA78* insertions detected experimentally (Figure S5). These findings establish a key role for H2A.Z in directing the integration of *COPIA* retrotransposons in nature and further highlight the strong deleterious effects they can cause.

Last, we investigated the integration patterns of *Ty1/Copia* retrotransposons in relation to H2A.Z in a distant plant (rice; Miyao et al., 2003) and in widely divergent species (*S. cerevisiae;* Mularoni et al., 2012). In both cases, experimentally induced insertions were located at sites enriched for H2A.Z (Figure 5F), thus indicating that the guiding role of H2A.Z has been evolutionarily conserved since the last common ancestor of plants and fungi.

**Intronic *ATCOPIA* insertions create epigenetically-inducible alternative transcripts**

To assess the functional impact of the strong, chromatin-based insertion biases towards genes observed in the epiRILs, we measured gene expression in six lines by RNA-seq. None of the homozygous *VANDAL21* and *ATENSPM3* insertions had detectable effects on the expression of target genes (Figure 6A), consistent with findings in natural accessions (Quadrana et al. 2016). In marked contrast, six of the 16 distinct homozygous *ATCOPIA93* insertions located within genes affected negatively their expression (Figure 6A). We also performed RNA-seq on one of the six epiRILs taken at a later generation (F16, epiRIL394). At this generation, there were approximately 40 *ATCOPIA93* insertions (counting homozygous insertions as two) per epiRIL and all were silenced and presumably methylated (Figure 6B).

One *ATCOPIA93* insertion present at F8 and F16 was located in the 5'UTR of gene *AT1G33560,* which encodes an NBS-LRR disease resistance protein, and had a much stronger dampening effect on gene expression at F16 than at F8 (Figure 6C). This result resembles that obtained for another gene in another epiRIL (Marí-Ordóñez et al. 2013). Moreover, reduction in *AT1G33560* expression correlated with down-regulation of *PR-1* (Figure 6C), a gene that is implicated in the response of plants to pathogens and whose expression depends on that of *AT1G33560* (Collier et al., 2011). These findings indicate that epigenetic silencing of newly inserted *ATCOPIA93* insertions can reinforce gene repression with potential phenotypic consequences.

Two *ATCOPIA93*-containing genes, *AT4G17190* and *AT5G20280*, which are implicated respectively in defense against aphids and in nectar secretion (Lin et al. 2014; Bhatia et al. 2015), also showed reduced expression, but only at generation F8. In both cases, *ATCOPIA93* was located within an intron and either in the same or the opposite orientation relative to gene transcription and was associated with transcript truncation at F8, but not at F16 (Figure 6C), when the insertions are epigenetically silent. Examination of DNA methylation, copy number and expression of *ATCOPIA93* as well as expression of the two genes across several generations between F8 and F16 revealed that DNA methylation was established at F12 and coincided with the silencing of all *ATCOPIA93* copies (Figure 6B) as well as the proper splicing of the insertion-containing intron (Figure 6D). These results demonstrate that the deleterious effects of intronic *ATCOPIA93* insertions can be alleviated once they become epigenetically silent.

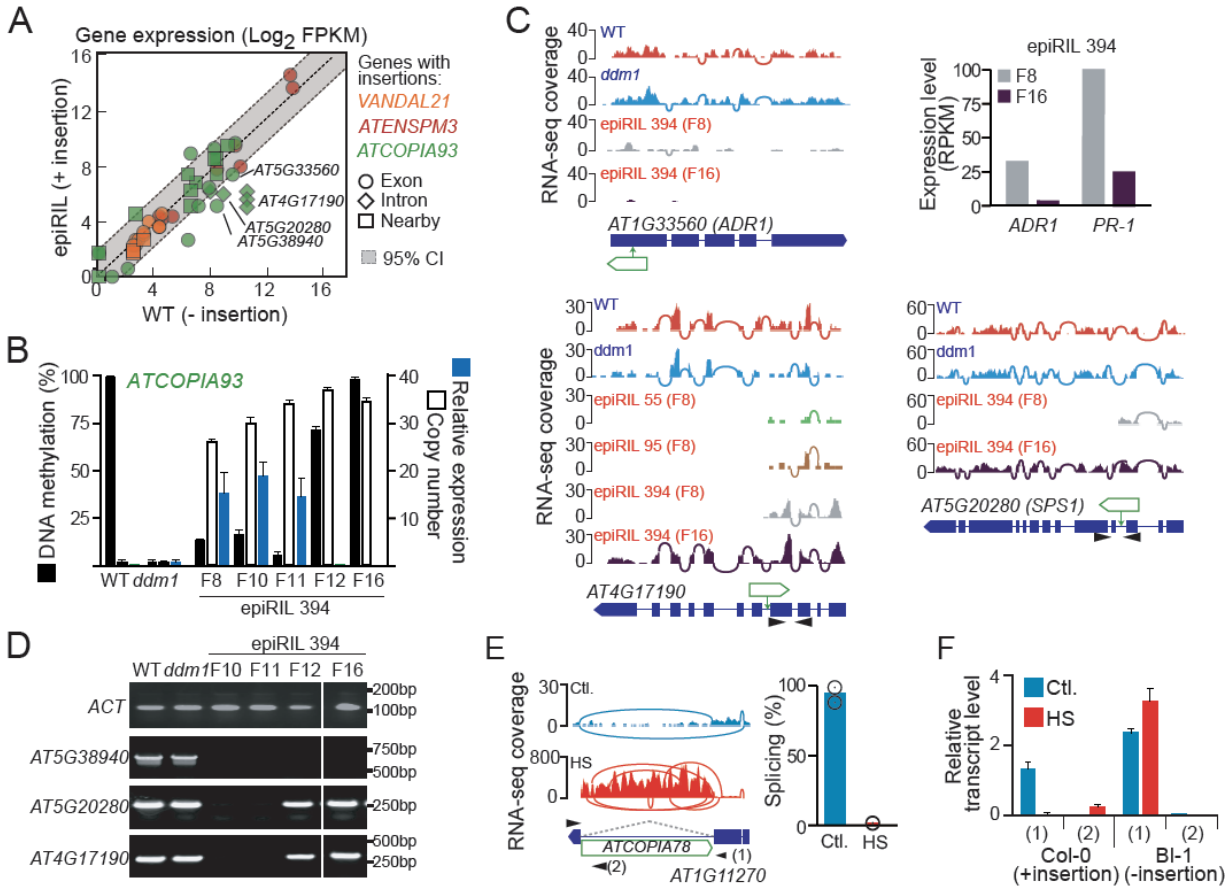


**Figure 6. Intronic *Ty1/COPIA* insertions create epigenetically-inducible alternative transcripts.**

**A.** Expression ratios between epiRILs and wild type plants for genes harboring homozygous *VANDAL21*, *ATENSPM3* and *ATCOPIA93* insertions. Genes with exonic, intronic or nearby insertions are indicated by circles, boxes or diamonds, respectively. Expected expression ratios and 95% confidence intervals were obtained by sampling 1000 random set of 55 genes and calculating their expression ratio. **B**. q-PCR analyses of DNA methylation, copy number and expression of *ATCOPIA93* in wild type, *ddm1* and epiRIL394, taken at F8 and more advanced generations. Data are mean ± s.d. (n = 2 independent biological experiments). **C**. Genome browser view of RNA-seq coverage over selected genes containing new *ATCOPIA93* insertions. Exon-exon junctions detected by split-reads are represented by arcs that connect exons. Samples containing or lacking TE insertions are highlighted in red and blue, respectively. Expression levels for the gene *ADR1 and PR-1* are shown on the upper right corner. **D**. RT–PCR analysis of genes containing new *ATCOPIA93* insertions in epiRIL394. Primers are indicated by arrow-heads in **B**. **E**. Genome browser view of RNA seq data for a gene containing an *ATCOPIA78* insertion in Col-0 in plants grown under control conditions (Ctl) or subjected to heat stress (HS). Percentage of splicing for the TE-containing intron (indicated by dashed lines) is shown on the right panel. Data are mean (n = 2 independent biological experiments). **F**. qRT-PCR analyses of *AT1G11270* expression in response to heat stress (HS) in accessions containing (Col-0) or lacking (Bl-1) the intronic *ATCOPIA78* insertion**.** Primers are indicated in **E** by arrow-heads. Data are mean ± s.d. (n = 2 independent biological experiments).

We last investigated a natural intronic *ATCOPIA78* insertion that is present in the reference Col-0 genome but absent from other accessions (Figure S6). Analysis of publicly available RNA-seq data (Pietzenuk et al. 2016) indicated that the insertion is epigenetically reactivated by heat stress and that reactivation is associated with the production of a truncated transcript from the target gene (Figure 6E). To demonstrate causality, we compared the splicing level of the second intron of the target gene *(AT1G11270)* in Col-0 plants subjected or not to heat stress as well as in plants of the Bl-1 accession, which does not contain any *ATCOPIA78* insertion within this gene. While the second intron was spliced in heat-stressed and control Bl-1 plants, this was not the case for Col-0 plants, which showed splicing only under non-stress conditions (Figure 6F). Given that *ATCOPIA78* is among the most active TE families in nature and that its activity correlates with several geo-climatic variables (Quadrana et al. 2016), it is likely that numerous

natural *ATCOPIA78*-containing alleles of genes are similarly endowed with the possibility to modulate their expression in response to environmental stress. Whether such alleles confer selective advantages remains to be determined.

## Discussion

### TEA lines: when mobile TEs are caught in the act

Whole genome sequencing of MA lines has been extremely valuable in determining the rate, spectrum and genome-wide distribution of spontaneous point and other small-size mutations in several model organisms (Ossowski et al. 2010; Zhu et al. 2014; Denver et al. 2009; Keightley et al. 2009). Results of these experiments indicated that small-size mutations occur almost randomly throughout the genome, although local modulations were observed, notably in relation to certain DNA sequences, recombination, transcription activity, and chromatin marks, including DNA methylation (Makova and Hardison 2015). However, MA lines did not provide much information about the heritable mutational landscape generated by TE mobilization, mainly because transposition is typically rare and also because non-reference TE insertions are difficult to detect using standard short-read sequencing strategies.

Based on a unique experimental system of *A. thaliana* epiRILs, in which transposition activity was kick-started but that otherwise resembled MA lines, we documented >8000 new TE insertions. These were produced mainly by three TEs, which belong to two DNA transposon (*VANDAL21* and *ATENSPM3*) and one LTR-retrotransposon (*ATCOPIA93*) families that are among the most active in nature (Quadrana et al. 2016). Furthermore, our modeling indicated that the insertion mutations produced by each of these three TEs in the epiRILs likely accumulated following a transposition-genetic drift scenario. Thus, the epiRILs can be defined operationally as TEA lines and as such they provide a first comprehensive assessment of the actual rate and genome-wide distribution of heritable mutations generated by TE mobilization in any species.

Using the epiRILs and additional experimental populations, we showed that *VANDAL21*, *ATENSPM3, ATCOPIA93* as well as another member (*ATCOPIA78*) of the *Ty1/Copia* superfamily of LTR-retrotransposons preferentially integrate within or close to three distinct sets of genes, each characterized by specific chromatin features. Moreover, we obtained strong evidence that *ATENSPM3* and the two *ATCOPIA* families generate highly deleterious mutations that are immediately purged by natural selection.

Collectively, our findings provide a first comprehensive experimental demonstration that mutations generated by TEs have radically distinct properties than the spontaneous small size mutations documented in MA lines. First, because of strong TE-specific integration preferences linked to chromatin, TE-caused mutations are distributed non-uniformly across the genome, with the implication also that their repertoire may vary substantially for any given TE as a result of changes in chromatin states, such as the loss of heterochromatin. Second, because many TEs target chromatin states that are associated with genes, their insertion, as well as their excision in the case of DNA transposons, tend to have drastic effects on gene expression. This is however not an obligate outcome, as exemplified by the lack of any discernible consequences of *VANDAL21* integration on gene transcription. Third, the functional impact of TE insertions can vary in a reversible manner over very short times, through their epigenetic silencing, which can be influenced by the environment, as in the case for *ATCOPIA78*. Last, TE insertions accumulate discontinuously, as a result of episodic, TE-specific reactivation, and at rates that may differ widely between TEs and environments.

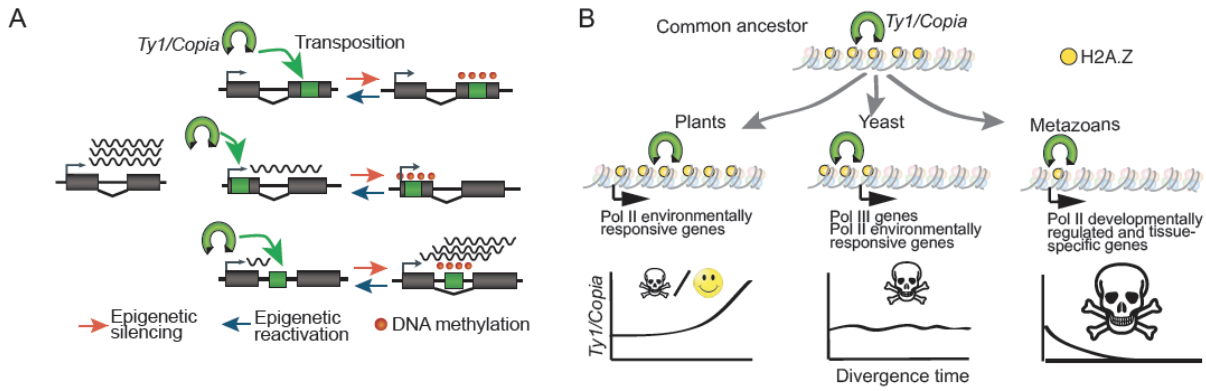


**Figure 7. A model to depict the functional consequences of new TE insertions and the evolutionary fate of *Ty1/Copia* retrotransposons in different organisms.**

**A.** *De novo* retrotransposon insertions within genes impact their expression in multiple ways. While retrotransposition within internal exons systematically disrupt gene expression, insertions in 5' regions or introns have an immediate dampening effect, which can be either aggravated or mitigated following epigenetic silencing of the inserted TE, respectively. Note that DNA methylation is depicted as spreading from the *COPIA* insertion into adjacent sequences, which remains to be determined experimentally. **B.** H2A.Z-guided integration of *Ty1/Copia* is ancestral. Functional diversification of H2A.Z between kingdoms determined the preferential integration of *Ty1/Copia* towards fast evolving genes in plants and yeast and towards developmentally regulated genes in animals. This differentiation may explain the high and low invasions success of *Ty1/Copia* retrotransposons in plants and animals, respectively.

### Retrotransposition-driven allelic heterogeneity associated with adaptive traits

Although WGS has revealed that TEs are powerful engine of genome as well as organism evolution (Chuong, Elde, and Feschotte 2016), we still lack a clear understanding of the impact of their mobilization within any given species. Our finding that LTR-retrotransposons belonging to the *Ty1/Copia* superfamily preferentially integrate within environmentally responsive genes provide valuable information in this respect. As we have also shown previously that mobilization of diverse *ATCOPIA* elements is associated with climate (Quadrana et al. 2016), it is tempting to speculate that this superfamily of TEs facilitates adaptation to changing or local environments. Consistent with this hypothesis, disease resistance genes are characterized by a high load of *COPIA* insertions and extensive allelic heterogeneity in nature (Quadrana et al, 2016; Kawakatsu et al., 2016). Thus, recurrent retrotransposition within these genes may play a critical role in the rapid evolution and expansion of the innate immune system in plants. We have demonstrated however that most *ATCOPIA* insertions have severe deleterious effects, athough some of these effects may be mitigated or fully erased once the insertions are epigenetically silenced. This mitigation can be environmentally dependent, thus endowing TE-containing alleles with unique properties in fluctuating environments (Figure 7A).

### H2A.Z and the different fate of *Ty1/Copia* retrotransposons in plants, yeast and animals

TEs need to keep moving in order to prevent their demise by the accumulation of inactivating mutations. However, because uncontrolled mobilization compromises host survival and that of TEs, different mechanisms have evolved that limit TE activity, notably epigenetic silencing, or that target integration to reduce the mutational load they generate (Sultana et al. 2017). For example, the LTR-retrotransposon *Ty5* from yeast interacts with the heterochromatic factor

silent information regulator 4 (Sir4), which directs *Ty5* integration within gene-poor regions (Sultana et al. 2017). Here, we have uncovered another targeting mechanism based on the histone variant H2A.Z, which is the among the most conserved histone H2As (Henikoff and Smith 2015). H2A.Z guides the integration of *Ty1/Copia* elements not only in *A. thaliana* but presumably also in rice and yeast. Furthermore, as in plants, the *S. cerevisiae Ty1* retrotransposon integrates preferentially within the arrays of H2A.Z-containing nucleosomes that are located upstream of RNA polymerase III (Pol III)-transcribed genes (Figure 5F; Albert et al., 2007; Baller et al., 2012), aided by the Pol III subunit AC40 (Bridier-Nahmias et al. 2015). Disruption of the AC40-integrase interaction leads to a redistribution of *Ty1* insertions towards subtelomeres (Bridier-Nahmias et al. 2015), which are enriched in environmentally responsive genes (Brown, Murray, and Verstrepen 2010). As in plants, these genes are enriched in H2A.Z-containing nucleosomes (Meneghini, Wu, and Madhani 2003; Sadeghi et al. 2011; Albert et al. 2007). These findings suggest therefore that AC40 serves to further restrict the integration of *Ty1* to a subclass of genes containing arrays of H2A.Z, which otherwise would integrate preferentially within or around responsive genes.

In animals, H2A.Z is essential and mainly found at developmentally regulated genes, where it occupies nucleosomes that flank the promoter (Henikoff and Smith 2015). Given the conserved role of H2A.Z in guiding the integration of *Ty1/Copia* retrotransposons, their mobilization should therefore have catastrophic effects in animals, which may explain why this superfamily of retrotransposons is virtually absent from animal genomes (Huang, Burns, and Boeke 2012). In other words, functional diversification of H2A.Z between plants and animals may have directed the opposite fate of *Ty1/Copia* retrotransposons in these two kingdoms (Figure 7B). This in turn opens up the possibility that the evolutionary fate of other TEs could also be explained by similar chromatin-directed integration biases.

**Acknowledgments**

We thank members of the Colot group and especially Pierre Baduel for discussions and critical reading of the manuscript. We thank the Genomics and Informatics facilities at IBENS for their help. Support was from the Agence National de la Recherche (ANR-09-BLAN-0237 to V.C. and P.W.), the Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, ANR-11-IDEX-0001-02 PSL* Research University (to V.C.), the European Union Seventh Framework Programme Network of Excellence EpiGeneSys (Award 257082, to V.C.) and the Centre National de la Recherche Scientifique (MOMENTUM program, to L.Q.). M.E. was supported by a PhD studentship from the French Ministry of Research and by a postdoctoral fellowship from MEMOLIFE. Initial support for L.Q. was provided by MEMOLIFE and EpiGeneSys. V.B. was supported by a PhD studentship from PSL Research University and from MEMOLIFE.

**Author contributions**

VC, PW and JMA conceived the project, with substantial additional input from LQ. ME and EC performed DNA and RNA extraction; JG, SE and JMA produced the WGS data; AG, ME, JMA and LQ performed the detection of TE insertions in the epiRILs; VB identified excision footprints in the epiRILs. LQ performed all of the other bioinformatic analyses, as well as tthe TE sequence capture and RT-QPCR experiments. LQ and AS performed the heat-stress experiments. LQ and VC interpreted the results. LQ and VC wrote the manuscript.

**Declaration of interests**

The authors declare no competing financial interest.

**MATERIALS AND METHODS**

***Experimental model and subject details***

The following *A. thaliana* plants were used: wild type Col-0 and Bl-1 accessions. The Col-0 *ddm1-2* mutant and the epiRILs

population (Johannes et al. 2009), the Col-0 *nrpd1-3* mutant (H. Ito et al. 2011) and the Col-0 *hta9-1 hta11-2* double mutant (March-Díaz et al. 2008). Unless stated otherwise, all plants were grown in long-days (16h:8h light:dark) at 23°C.

***Genomic DNA sequencing using mate-pairs***

Genomic sequencing was performed as described before (Gilly et al, 2014). Briefly, DNA was extracted from 10-20 seedlings grown under long day conditions, using DNeasy Qiagen kits. About 10 μg of genomic DNA was sonicated to a 4-6 kb size range using the E210 Covaris instrument (Covaris, Inc., USA). Libraries were prepared following Illumina's protocol (Illumina Mate Pair library kit), starting with size-selected (approximately 5kb) fragments, which were end-repaired, biotin labeled and circularized. Linear DNA was eliminated by digestion and circularized DNA was fragmented to 300-700 bp using the E210 Covaris. Biotinylated DNA junctions were purified using streptavidin, end-repaired and 3′-adenylated in order to ligate Illumina adapters. Junction fragments were PCR-amplified using Illumina adapter-specific primers and amplified fragments within the 350-650 bp size range were selected for sequencing. Each library was sequenced using 100 base-length read chemistry in a paired-end flow cell on the Illumina GAIIx (2 lanes) or HiSeq2000 (1 lane) (Illumina, USA).

***Mapping and detection of TE insertions using WGS***

Reads were mapped with BWA v.0.6.1 using the parameters -R 10000 -l 35 -O 11, and the parameters n 10000 N 10000 -s for sampe, onto the TAIR10 reference sequence. Reads hanging over chromosome ends were removed using picard CleanSam and duplicate pairs were removed using picard MarkDuplicates. TE insertions were detected by implementing TE-Tracker software (available at http://www.genoscope.cns.fr/TE-Tracker) exactly as described before (Gilly *et al.*, 2014). WGS did not produce sufficient coverage (<10X) for 16 of the 123 epiRILs analyzed at generation F8, and these 16 epiRILs were not considered further (Table S1). TE-Tracker is a computational method that we have previously developed for accurately detecting both the identity and destination of newly mobilized TEs in genomes re-sequenced using mate-pair libraries (Gilly et al, 2014). Importantly, TE-Tracker does not rely on prior annotation, yet is able to integrate it, making the results easily interpretable. Briefly, TE-Tracker uses paired reads mapping information to identify discordant reads that mapped partially over TE-sequences to detect the position of TE insertions. Insertion site positions were refined at the single nucleotide resolution by exploiting sequence information contained in split-reads. To this end, we implemented the software SPLITREADER (available at https://github.com/LeanQ/SPLITREADER; Quadrana et al., 2016). Homozygous and heterozygous insertions were defined based on the normalized number of reads supporting each insertion event (Table S2). In addition, this approach enabled us also to identify insertions that were likely present in only one of the 10-20 seedlings used to extract DNA and which reflect transposition during the reproductive phase of the parent. These insertions were also called heterozygous, as this was likely the case and to reflect their very recent ancestry. Conversely, our approach was designed to exclude poorly supported insertions, which could reflect either mapping artifacts or rare somatic events. Finally, visual inspection was carried out for a random sample of over 200 insertion events and their homozygous or heterozygous status was confirmed in each case.

***TE-sequence capture***

TE sequence capture was performed on exactly 1000 seedlings in all cases except for the F3 progeny of *hta9 hta11* line 2, where only 477 seedlings were recovered (see main text and Figure 4A and 5B for details of the plant materials used). Seedlings were grown under control (long-day) conditions and genomic DNA was extracted using the CTAB method (Murray and Thompson 1980). In order to assess the sensitivity of TE-sequence capture, we added 1ng of genomic DNA extracted from epiRILs 394 (generation F16) to 1ug of genomic DNA extracted from the 477 F3 seedling of *hta9 hta11* line 2 prior to library construction (i.e. 1:1000 dilution of the spiked-in genomic DNA). Libraries were prepared using 1μg of DNA and TruSeq paired-end kit (Illumina) following manufacturer instructions. Libraries were then amplified through 7 cycles of ligation-mediated PCR using the KAPA HiFi Hot Start Ready Mix and primers AATGATACGGCGACCACCGAGA and CAAGCAGAAGACGGCATACGAG at a final concentration of 2$\mu$M. 1μg of multiplexed libraries were then subjected to TE-sequence capture exactly as previously reported (Quadrana et al. 2016). Enrichment for captured TE sequences was confirmed by qPCR and estimated to be higher than 1000 fold. Pair-end sequencing was performed using one lane of Illumina NextSeq500 and 75bp reads. About 42 million pairs were sequenced per library and mapped to the TAIR10 reference genome using Bowtie2 v2.3.2 (Langmead and Salzberg 2012) with the arguments --mp 13 --rdg 8,5 --rfg 8,5 --very-sensitive. An improved version of SPLITREADER (available at https://github.com/LeanQ/SPLITREADER) was used to detect new TE insertions. Briefly, split-reads as well as discordant reads mapping partially on the reference sequence of *ATCOPIA93* and *ATCOPIA78* (obtained from RepBase update) were identified, soft clipped and remapped to the TAIR10 reference genome using Bowtie2 (Langmead and Salzberg 2012). Putative insertions supported by at least one split- and/or discordant-reads at each side of the insertion sites were retained. Insertions spanning centromeric repeats or coordinates spanning the corresponding donor TE sequence were excluded. In addition, putative TE insertions detected in more than one library were excluded to retain only sample-specific TE insertions (Table S3). More than 80%

of new TE insertions present in epRIL394 F16 were detected, confirming that the sensitivity of our TE-sequence capture and computational approach is higher than 1:1000. In addition, no non-reference insertions were detected in 1000 F1 seedlings of wild type Col-0, highlighting the specificity of our approach.

***Detection of Tos17 non reference insertions in rice genomes***

A total of 16,784 *Tos17* non-reference flanking sequences (Miyao et al. 2003) were retrieved from GeneBank and mapped to the reference rice genome using Minimap2 v2.11-r797 (H. Li 2018), which enabled us to identify 14,258 insertion points with high confidence (Table S4).

***Detection of mutations within transposed copies***

Discordant mate-pair reads mapping within a 6kb interval either upstream or downstream of each insertion site were extracted and re-mapped using Bowtie2 (Langmead and Salzberg 2012) over a library constructed with the specific donor TE sequence only. Sequence variants were detected using samtools mpileup V1.2.1 and only variants with a quality of at least 30 were kept. Long deletions were detected as regions without coverage and breakpoints were reconstructed by local assembly using Velvet V1.2.09 (Zerbino and Birney 2008).

***Analysis of global and local enrichment of new TE insertions***

To assess if new TE insertions are enriched in pericentromeric regions, their number within these regions was compared with that expected from a random distribution. The expected distribution was calculated by randomizing $10^4$ times the position of new TE insertions across the chromosomes (genomic regions showing coverage deviation, the inner pericentromeres, or coordinates spanning the corresponding donor TE sequence were excluded). This set of random positions was used as a control for all subsequent analyses. Insertion distribution over wild-type- and *ddm1*-derived regions was obtained by counting the number of new TE insertions within intervals delimited by at least two consecutive stable hypermethylated or hypomethylated regions, respectively (Cortijo et al. 2014). Overrepresentation over genes and neighboring sequences was performed using a meta-gene analysis. Briefly, protein coding gene features were extracted from the TAIR10 annotation and coordinates of non-reference TE insertions with TSDs were crossed with the set of genic features according to the following stepwise hierarchy: 5' UTR > 3' UTR> exon > intron > intergenic regions. For insertions that do not overlap protein-coding genes, the distance to the closest gene was calculated and reported as negative or positive distance according to the gene orientation. Overrepresentation over chromatin states was performed by comparing the number of new TE insertions and randomly generated TE insertions located within each chromatin domain (Sequeira-Mendes et al. 2014). Density of ATCOPIA93 insertions were obtained by calculating the distance between insertion sites and the middle point of the nearest well positioned nucleosome mapped in Col-0 (Table S5; Lyons and Zilberman, 2017).Gene ontology (GO) analyses were performed using AGRIGO (http://bioinfo.cau.edu.cn/agriGO/) and as input the ID of genes that contain a TE insertion within the limits of their annotation.

***Analysis of chromatin features at insertion sites***

4kb regions centered around insertion sites were defined and used to extract normalized coverage of DnaseI hypersensitivity (Sullivan et al. 2014), Mnase accessibility (G. Li et al. 2014), H3K27me3 (C. Li et al. 2015), H2A.Z enrichment level (Coleman-Derr and Zilberman 2012) and well-positioned nucleosomes (Lyons and Zilberman 2017). The same approach was used for the analysis of H2A.Z enrichment in rice (Zahraeifard et al. 2018) and of htz1 from yeast (Gu et al. 2015). Average normalized coverage was then calculated for each bp and plotted using the *smooth.spline* function in R.

***epiQTL mapping of transposition activity***

Using TE copy number as phenotype and a total of 126 parental differentially methylated regions (DMRs) that segregate in a Mendelian fashion in the epiRILs (i.e. stable DMRs) as physical markers (Cortijo et al. 2014), we implemented the multiple QTL model (mqmsacn) from the R/qtl package. Genome-wide significance was determined empirically for each trait using 1000 permutations of the data. LOD significance thresholds were chosen to correspond to a genome-wide false positive rate of 5%.

***Transposition-drift modeling of insertion accumulation***

In the absence of selection, TE invasion is mostly determined by the rate of transposition and rate of fixation of insertions by random segregation (i.e. drift). Thus, we constructed individual-based transposition-drift models, all starting with an initial number of copies all equally active (with a rate $K$ of new copies per generation) and with each new copy being also equally active. New copies arise in the heterozygous state and can be inherited following a Poisson distribution according to Mendelian segregation. The model also considers TE elimination by excision, which occurs at rate $E$ per transposition event. Additionally, concerted silencing of all active copies may occur when copy number reach the threshold $I$. Simulations were run 1000 times using a wide

space of parameter values ($K$={0,0.1,…,1}, $E$={0,0.2,…,1} and $l$={20,21,...,60}. Distribution of homozygous and heterozygous insertions between simulated and observed data at F8 was evaluated using a two-dimensional goodness-of-fit test.

***Transcriptome analysis***

RNA from wild type, *ddm1*, epiRIL55 (F8), epiRIL95 (F8), epiRIL260 (F8), epiRIL439 (F8), epiRIL454 (F8) epiRIL394 (F8) and epiRIL394 (F16) was isolated using Rneasy Plant Minikit (Qiagen) according to the supplier's instructions. Contaminating DNA was removed using RQ1 DNase (Promega). One μg of total RNA was processed using TruSeq Stranded Total RNA kit (Illumina) according to the supplier's instructions. About 20M 76nt-long single-end reads were obtained per sample on the Illumina HiSeq2000. Expression level was calculated by mapping reads using STAR v2.5.3a (Dobin et al. 2013) on the *A. thaliana* reference genome (TAIR10) with the following arguments --outFilterMultimapNmax 50 --outFilterMatchNmin 30 --alignSJoverhangMin 3 --alignIntronMax 10000. Duplicated pairs were removed using picard MarkDuplicates. Counts were normalized and annotations were declared differentially expressed between samples using DESeq2 (Love, Huber, and Anders 2014). When specified, uniquely mapped reads (mapping quality > 10) were selected with samtools (H. Li et al. 2009). RNAseq data from heat-stressed plants were obtained from (Pietzenuk et al. 2016) and analyzed as described above. Splicing of intronic TE insertions was calculated as described previously (Teixeira et al. 2017). Briefly, the number of split-reads (SR) and non-split-reads (NSR, which should be fully and uniquely contained within the interval surrounding the same exon–intron junction) mapping to an exon-intron junction was extracted and the ratio SR/(SR+NSR)*100 was then calculated.

***Quantification of expression level, copy number and DNA methylation***

RNA was extracted using the RNeasy plant mini kit (Qiagen) from Col-0 or Bl-1 plants grown under normal conditions (10 days old seedlings grown in liquid medium) or subjected to heat shock treatment (H. Ito et al. 2011). RT-qPCR was performed as described previously (Quadrana et al. 2016). Primer sequences are provided in Table S6. RT-qPCR results (two biological replicates) are indicated relative to those obtained for a gene (*AT5G13440*) that shows invariant expression under multiple conditions. Copy number and DNA methylation of *ATCOPIA93* was performed as described before (Marí-Ordóñez et al. 2013).

***Data and software availability***

Sequencing data has been deposited in the European Nucleotide Archive (ENA) under project XXXXX.

## REFERENCES

Albert, Istvan, Travis N. Mavrich, Lynn P. Tomsho, Ji Qi, Sara J. Zanton, Stephan C. Schuster, and B. Franklin Pugh. 2007. "Translational and Rotational Settings of H2A.Z Nucleosomes across the Saccharomyces Cerevisiae Genome." *Nature* 446 (7135): 572–76. https://doi.org/10.1038/nature05632.

Baller, Joshua A., Jiquan Gao, Radostina Stamenova, M. Joan Curcio, and Daniel F. Voytas. 2012. "A Nucleosomal Surface Defines an Integration Hotspot for the Saccharomyces Cerevisiae Ty1 Retrotransposon." *Genome Research* 22 (4): 704–13. https://doi.org/10.1101/gr.129585.111.

Bhatia, Varnika, Jaya Maisnam, Ajay Jain, Krishan Kumar Sharma, and Ramcharan Bhattacharya. 2015. "Aphid-Repellent Pheromone E-β-Farnesene Is Generated in Transgenic Arabidopsis Thaliana over-Expressing Farnesyl Diphosphate Synthase2." *Annals of Botany* 115 (4): 581–91. https://doi.org/10.1093/aob/mcu250.

Bridier-Nahmias, Antoine, Aurélie Tchalikian-Cosson, Joshua A. Baller, Rachid Menouni, Hélène Fayol, Amando Flores, Ali Saïb, Michel Werner, Daniel F. Voytas, and Pascale Lesage. 2015. "An RNA Polymerase III Subunit Determines Sites of Retrotransposon Integration." *Science (New York, N.Y.)* 348 (6234): 585–88. https://doi.org/10.1126/science.1259114.

Brown, Chris A., Andrew W. Murray, and Kevin J. Verstrepen. 2010. "Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts." *Current Biology* 20 (10). Elsevier Ltd: 895–903. https://doi.org/10.1016/j.cub.2010.04.027.

Chuong, Edward B, Nels C Elde, and Cédric Feschotte. 2016. "Regulatory Activities of Transposable Elements: From Conflicts to Benefits." *Nature Reviews. Genetics* 18 (2): 71–86. https://doi.org/10.1038/nrg.2016.139.

Coleman-Derr, Devin, and Daniel Zilberman. 2012. "Deposition of Histone Variant H2A.Z within Gene Bodies Regulates Responsive Genes." *PLoS Genetics* 8 (10): e1002988. https://doi.org/10.1371/journal.pgen.1002988.

Collier, Sarah M., Louis-Philippe Hamel, and Peter Moffett. 2011. "Cell Death Mediated by the N-Terminal Domains of a Unique and Highly Conserved Class of NB-LRR Protein." *Molecular Plant-Microbe Interactions* 24 (8): 918–31. https://doi.org/10.1094/MPMI-03-11-0050.

Colome-Tatche, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze, A. Sarazin, M. Etcheverry, et al. 2012. "Features of the Arabidopsis Recombination Landscape Resulting from the Combined Loss of Sequence Variation and DNA Methylation." *Proceedings of the National Academy of Sciences* 109 (40): 16240–45. https://doi.org/10.1073/pnas.1212955109.

Colomé-Tatché, Maria, Sandra Cortijo, René Wardenaar, Lionel Morgado, Benoit Lahouze, Alexis Sarazin, Mathilde Etcheverry, et al. 2012. "Features of the Arabidopsis Recombination Landscape Resulting from the Combined Loss of Sequence Variation and DNA Methylation." *Proceedings of the National Academy of Sciences* 109 (40): 16240–45. https://doi.org/10.1073/pnas.1212955109.

Cortijo, Sandra, Rene Wardenaar, Maria Colome, Arthur Gilly, Mathilde Etcheverry, Karine Labadie, and Fréderic Hospital. 2014. "Mapping the Epigenetic Basis of Complex Traits." *Science (New York, N.Y.)* 343 (6175): 1145–48. https://doi.org/10.1038/nrg2719.

Denver, D. R., P. C. Dolan, L. J. Wilhelm, W. Sung, J. I. Lucas-Lledo, D. K. Howe, S. C. Lewis, et al. 2009. “A Genome-Wide View of Caenorhabditis Elegans Base-Substitution Mutation Processes.” *Proceedings of the National Academy of Sciences* 106 (38): 16310–14. https://doi.org/10.1073/pnas.0904895106.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1). Oxford University Press: 15–21.

Doseff, A, R Martienssen, and V Sundaresan. 1990. “Somatic Excision of the Mu1 Transposableelement of Maize.” *Nucleic Acids Research Acids Res* 19 (3): 579–84.

Fu, Yu, Akira Kawabe, Mathilde Etcheverry, Tasuku Ito, Atsushi Toyoda, Asao Fujiyama, Vincent Colot, Yoshiaki Tarutani, and Tetsuji Kakutani. 2013. “Mobilization of a Plant Transposon by Expression of the Transposon-Encoded Anti-Silencing Factor.” *EMBO Journal* 32 (17): 2407–17. https://doi.org/10.1038/emboj.2013.169.

Gilly, Arthur, Mathilde Etcheverry, Mohammed-Amin Amin Madoui, Julie Guy, Leandro Quadrana, Adriana Alberti, Antoine Martin, et al. 2014. “TE-Tracker: Systematic Identification of Transposition Events through Whole-Genome Resequencing.” *BMC Bioinformatics* 15 (1). BioMed Central Ltd: 377. https://doi.org/10.1186/s12859-014-0377-z.

Gu, Muxin, Yanin Naiyachit, Thomas J. Wood, and Catherine B. Millar. 2015. “H2A.Z Marks Antisense Promoters and Has Positive Effects on Antisense Transcript Levels in Budding Yeast.” *BMC Genomics* 16 (1): 1–11. https://doi.org/10.1186/s12864-015-1247-4.

Henikoff, Steven, and Mitchell Mitchell Smith. 2015. “Histone Variants and Epigenetics.” *Cold Spring Harbor Perspectives in Biology* 7 (1). https://doi.org/10.1101/cshperspect.a019364.

Huang, Cheng Ran Lisa, Kathleen H. Burns, and Jef D. Boeke. 2012. “Active Transposition in Genomes.” *Annual Review of Genetics* 46 (1): 651–75. https://doi.org/10.1146/annurev-genet-110711-155616.

Ito, Hidetaka, Hervé Herve Gaubert, Etienne Bucher, Marie Mirouze, Isabelle Vaillant, Jerzy Paszkowski, Hervé Herve Gaubert, et al. 2011. “An SiRNA Pathway Prevents Transgenerational Retrotransposition in Plants Subjected to Stress.” *Nature* 472 (7341). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 115–19. https://doi.org/10.1038/nature09861.

Ito, Tasuku, Yoshiaki Tarutani, Taiko Kim To, Mohamed Kassam, Evelyne Duvernois-Berthet, Sandra Cortijo, Kazuya Takashima, et al. 2015. “Genome-Wide Negative Feedback Drives Transgenerational DNA Methylation Dynamics in Arabidopsis.” *PLoS Genetics* 11 (4): e1005154. https://doi.org/10.1371/journal.pgen.1005154.

Johannes, Frank, Emmanuelle Porcher, Felipe K Teixeira, Vera Saliba-colombani, Juliette Albuisson, Fabiana Heredia, Vincent Colot, et al. 2009. “Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits.” *Plos Genetics* 5 (6): e1000530. https://doi.org/10.1371/journal.pgen.1000530.

Kawakatsu, Taiji, Shao shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J J. Schmitz, Mark A A. Urich, Rosa Castanon, et al. 2016. “Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions.” *Cell* 166 (2): 492–506. https://doi.org/10.1016/j.cell.2016.06.044.

Keightley, Peter D, Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L Blaxter. 2009. “Analysis of the Genome Sequences of Three Drosophila Melanogaster Spontaneous Mutation Accumulation Lines,” 1195–1201. https://doi.org/10.1101/gr.091231.109.more.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Li, Chenlong, Chen Chen, Lei Gao, Songguang Yang, Vi Nguyen, Xuejiang Shi, Katherine Siminovitch, et al. 2015. “The Arabidopsis SWI2/SNF2 Chromatin Remodeler BRAHMA Regulates Polycomb Function during Vegetative Development and Directly Activates the Flowering Repressor Gene SVP.” *PLoS Genetics* 11 (1): 1–25. https://doi.org/10.1371/journal.pgen.1004944.

Li, Guang, Shujing Liu, Jiawei Wang, Jianfeng He, Hai Huang, Yijing Zhang, and Lin Xu. 2014. “ISWI Proteins Participate in the Genome-Wide Nucleosome Distribution in Arabidopsis.” *Plant Journal* 78 (4): 706–14. https://doi.org/10.1111/tpj.12499.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16). Oxford University Press: 2078–79.

Lin, I Winnie, Davide Sosso, Li-Qing Chen, Klaus Gase, Sang-Gyu Kim, Danny Kessler, Peter M Klinkenberg, et al. 2014. “Nectar Secretion Requires Sucrose Phosphate Synthases and the Sugar Transporter SWEET9.” *Nature* 508 (March). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 546. http://dx.doi.org/10.1038/nature13082.

Lippman, Zachary, Anne-Valérie Gendrel, Michael Black, Matthew W Vaughn, Neilay Dedhia, W Richard McCombie, Kimberly Lavine, et al. 2004. “Role of Transposable Elements in Heterochromatin and Epigenetic Control.” *Nature* 430 (6998): 471–76. https://doi.org/10.1038/nature02651.

Lisch, Damon. 2013. “How Important Are Transposons for Plant Evolution?” *Nature Reviews Genetics* 14 (1): 49–61. https://doi.org/10.1038/nrg3374.

Lloyd, John P., Alexander E. Seddon, Gaurav D. Moghe, Matthew C. Simenc, and Shin-Han Shiu. 2015. “Characteristics of Plant Essential Genes Allow for Within- and between-Species Prediction of Lethal Mutant Phenotypes.” *The Plant Cell* 27 (8): 2133–47. https://doi.org/10.1105/tpc.15.00051.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12). BioMed Central: 550.

Lyons, David B., and Daniel Zilberman. 2017. “DDM1 and Lsh Remodelers Allow Methylation of DNA Wrapped in Nucleosomes.” *ELife* 6: 1–20. https://doi.org/10.7554/eLife.30674.

Makova, Kateryna D., and Ross C. Hardison. 2015. “The Effects of Chromatin Organization on Variation in Mutation Rates in the Genome.” *Nature Reviews Genetics* 16 (4). Nature Publishing Group: 213–23. https://doi.org/10.1038/nrg3890.

March-Díaz, Rosana, Mario García-Domínguez, Jorge Lozano-Juste, José León, Francisco J. Florencio, and José C. Reyes. 2008. “Histone H2A.Z and Homologues of Components of the SWR1 Complex Are Required to Control Immunity in Arabidopsis.” *Plant Journal* 53 (3): 475–87. https://doi.org/10.1111/j.1365-313X.2007.03361.x.

Marí-Ordóñez, Arturo, Antonin Marchais, Mathilde Etcheverry, Antoine Martin, Vincent Colot, and Olivier Voinnet. 2013. “Reconstructing de Novo Silencing of an Active Plant Retrotransposon.” *Nature Genetics* 45 (9): 1029–39. https://doi.org/10.1038/ng.2703.

Masson, Patrick, Richard Surosky, Jeffrey A Kingsbury, and Nina V Fedoroff. 1987. “Genetic and Molecular Analysis of the Spm-Dependent a-M2 Alleles of the Maize a Locus.” *Genetics* 177: 117–37.

Meneghini, Marc D., Michelle Wu, and Hiten D. Madhani. 2003. “Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin.” *Cell* 112 (5): 725–36. https://doi.org/10.1016/S0092-8674(03)00123-5.

Mirouze, Marie, Jon Reinders, Etienne Bucher, Taisuke Nishimura, Korbinian Schneeberger, Stephan Ossowski, Jun Cao, Detlef Weigel, Jerzy Paszkowski, and Olivier Mathieu. 2009. “Selective Epigenetic Control of Retrotransposition in Arabidopsis.” *Nature* 461 (7262). Nature Publishing Group: 427–30. https://doi.org/10.1038/nature08328.

Miyao, Akio, Katsuyuki Tanaka, Kazumasa Murata, Hiromichi Sawaki, Shin Takeda, Kiyomi Abe, Yoriko Shinozuka, Katsura Onosato, and Hirohiko Hirochika. 2003. “Target Site Specificity of the Tos17 Retrotransposon Shows a Preference for Insertion Withn Genes and against Insertion in Retrotransposon-Rich Regions of the Genome.” *The Plant Cell* 15 (August): 1771–80. https://doi.org/10.1105/tpc.012559.ements.

Mularoni, Loris, Yulian Zhou, Tyson Bowen, Sunil Gangadharan, Sarah J. Wheelan, and Jef D. Boeke. 2012. “Retrotransposon Ty1 Integration Targets Specifically Positioned Asymmetric Nucleosomal DNA Segments in TRNA Hotspots.” *Genome Research* 22 (4): 693–703. https://doi.org/10.1101/gr.129460.111.

Murray, G, and William F Thompson. 1980. “Rapid Isolation of High Molecular Weight Plant DNA.” *Nucleic Acids Research* 8 (19). Oxford University Press: 4321–26.

Ossowski, Stephan, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M Clark, Ruth G Shaw, Detlef Weigel, and Michael Lynch. 2010. “The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana.” *Science* 327 (5961): 92–94. https://doi.org/10.1126/science.1180677.

Pietzenuk, Björn, Catarine Markus, Hervï¿½ Gaubert, Navratan Bagwan, Aldo Merotto, Etienne Bucher, and Ales Pecinka. 2016. “Recurrent Evolution of Heat-Responsiveness in Brassicaceae COPIA Elements.” *Genome Biology* 17 (1). Genome Biology: 1–15. https://doi.org/10.1186/s13059-016-1072-3.

Quadrana, Leandro, Amanda Bortolini Silveira, George F. Mayhew, Chantal LeBlanc, Robert A. Martienssen, Jeffrey A. Jeddeloh, and Vincent Colot. 2016. “The Arabidopsis Thaliana Mobilome and Its Impact at the Species Level.” *ELife* 5 (JUN2016): 1–25. https://doi.org/10.7554/eLife.15716.

Sadeghi, Laia, Carolina Bonilla, Annelie Strålfors, Karl Ekwall, and J. Peter Svensson. 2011. “Podbat: A Novel Genomic Tool Reveals Swr1-Independent H2A.Z Incorporation at Gene Coding Sequences through Epigenetic Meta-Analysis.” *PLoS Computational Biology* 7 (8). https://doi.org/10.1371/journal.pcbi.1002163.

Sequeira-Mendes, J., I. Araguez, R. Peiro, R. Mendez-Giraldez, X. Zhang, S. E. Jacobsen, U. Bastolla, and C. Gutierrez. 2014. “The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States.” *The Plant Cell* 26 (6): 2351–66. https://doi.org/10.1105/tpc.114.124578.

Slotkin, R Keith, and Robert Martienssen. 2007. “Transposable Elements and the Epigenetic Regulation of the Genome.” *Nature Reviews. Genetics* 8 (4): 272–85. https://doi.org/10.1038/nrg2072.

Soppe, Wim J.J., Zuzana Jasencakova, Andreas Houben, Tetsuji Kakutani, Armin Meister, Michael S. Huang, Steven E. Jacobsen, Ingo Schubert, and Paul F. Fransz. 2002. “DNA Methylation Controls Histone H3 Lysine 9 Methylation and Heterochromatin Assembly in Arabidopsis.” *EMBO Journal* 21 (23): 6549–59. https://doi.org/10.1093/emboj/cdf657.

Sullivan, Alessandra M., Andrej A. Arsovski, Janne Lempe, Kerry L. Bubb, Matthew T. Weirauch, Peter J. Sabo, Richard Sandstrom, et al. 2014. “Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. Thaliana.” *Cell Reports* 8 (6): 2015–30. https://doi.org/10.1016/j.celrep.2014.08.019.

Sultana, Tania, Alessia Zamborlini, Gael Cristofari, and Pascale Lesage. 2017. “Integration Site Selection by Retroviruses and Transposable Elements in Eukaryotes.” *Nature Reviews. Genetics* 18 (5). Nature Publishing Group: 292–308. https://doi.org/10.1038/nrg.2017.7.

Suto, Robert K., Michael J. Clarkson, David J. Tremethick, and Karolin Luger. 2000. “Crystal Structure of a Nucleosome Core Particle Containing the Variant Histone H2A.Z.” *Nature Structural Biology* 7 (12): 1121–24. https://doi.org/10.1038/81971.

Teixeira, Felipe Karam, Martyna Okuniewska, Colin D. Malone, Rémi Xavier Coux, Donald C. Rio, and Ruth Lehmann. 2017. “PiRNA-Mediated Regulation of Transposon Alternative Splicing in the Soma and Germ Line.” *Nature* 552 (7684). Nature Publishing Group: 268–72. https://doi.org/10.1038/nature25018.

Zahraeifard, Sara, Maryam Foroozani, Aliasghar Sepehri, Dong-Ha Oh, Guannan Wang, Venkata Mangu, Bin Chen, Niranjan Baisakh, Maheshi Dassanayake, and Aaron P Smith. 2018. “Rice H2A.Z Negatively Regulates Genes Responsive to Nutrient Starvation but Promotes Expression of Key Housekeeping Genes.” *J Exp Bot*, no. June. https://doi.org/10.1093/jxb/ery244.

Zerbino, Daniel R., and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhu, Y. O., M. L. Siegal, D. W. Hall, and D. A. Petrov. 2014. “Precise Estimates of Mutation Rate and Spectrum in Yeast.” *Proceedings of the National Academy of Sciences* 111 (22): E2310–18. https://doi.org/10.1073/pnas.1323011111.

Zilberman, Daniel, Devin Coleman-Derr, Tracy Ballinger, and Steven Henikoff. 2008. “Histone H2A.Z and DNA Methylation Are Mutually Antagonistic Chromatin Marks.” *Nature* 456 (7218): 125–29. https://doi.org/10.1038/nature07324.

Zlatanova, Jordanka, and Amit Thakar. 2008. “H2A.Z: View from the Top.” *Structure* 16 (2): 166–79. https://doi.org/10.1016/j.str.2007.12.008.

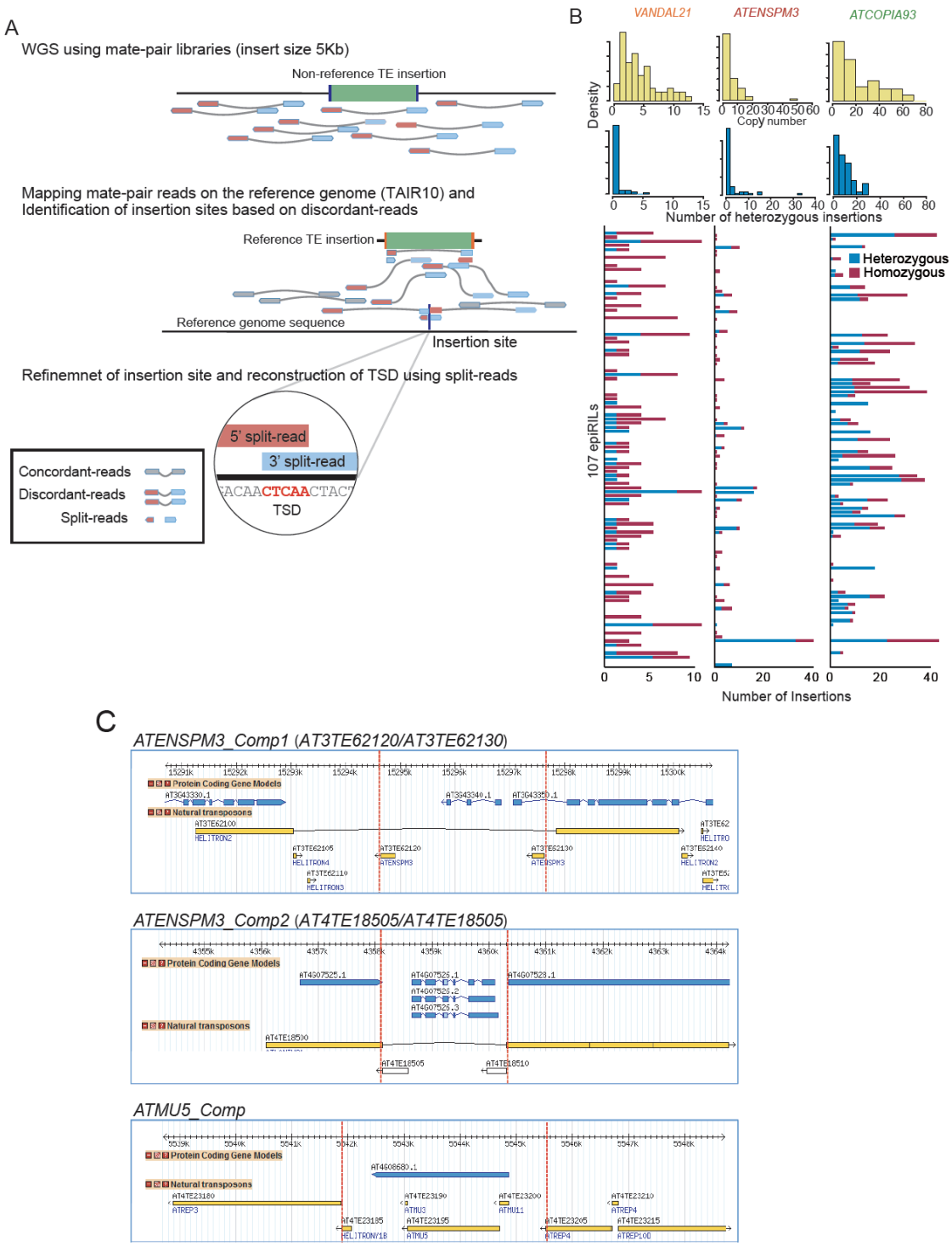


**Figure S1. Identification of new TE insertions in epiRILs.**
**A**. Schematic representation of the bioinformatic method used to detect new TE insertions. Sequenced reads around a newly inserted TE-copy (top half) produce discordant read mappings when aligned with the reference sequence (bottom half). Dashed light-red and light-blue arrows represent the mate-pairs reads linking the left and right extremities of the insertion breakpoint with the donor TE sequence. **B.** The top two rows of panels indicate the distribution of the number of private insertions (top: all; bottom: heterozygous only) among epiRILs for each TE family. The number of homozygous and heterozygous private insertions in each epiRIL (ordered by name) is indicated below for each of the three TE families. **C**. Genome browser view of the three composite mobile TEs (delineated by the dotted red lines) identified in the present study.

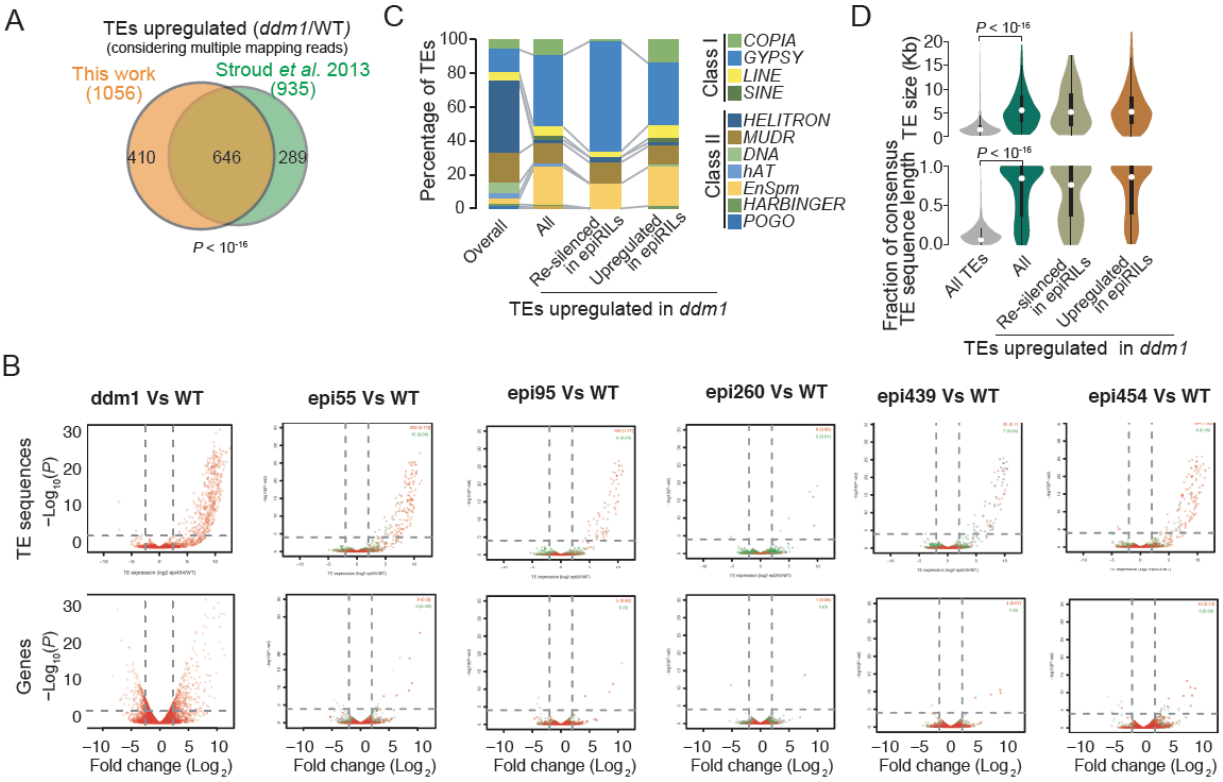


**Figure S2. Sustained transcriptional activation of TEs does not associate systematically with transposition.**
**A.** Comparison of the number of TE sequences upregulated in siblings of the *ddm1* parent used to generate the epiRILs population with a that obtained previously (Stroud et al, 2013). Statistical significance of the overlap was obtained using the Chi-square test. **B.** Differential expression analysis of TEs and genes between wild type and *ddm1*as well as five epiRILs. Annotations located in wild-type- or *ddm1*-detived regions are indicated by green and red dots, respectively. **C.** Identity of the TE sequences upregulated in *ddm1* and either re-silenced or stably upregulated in the five epiRILs. **D**. Distribution of sequence lengths and fraction of TE consensus length of TEs upregulated in *ddm1* and either re-silenced or stably upregulated in the five epiRILs.

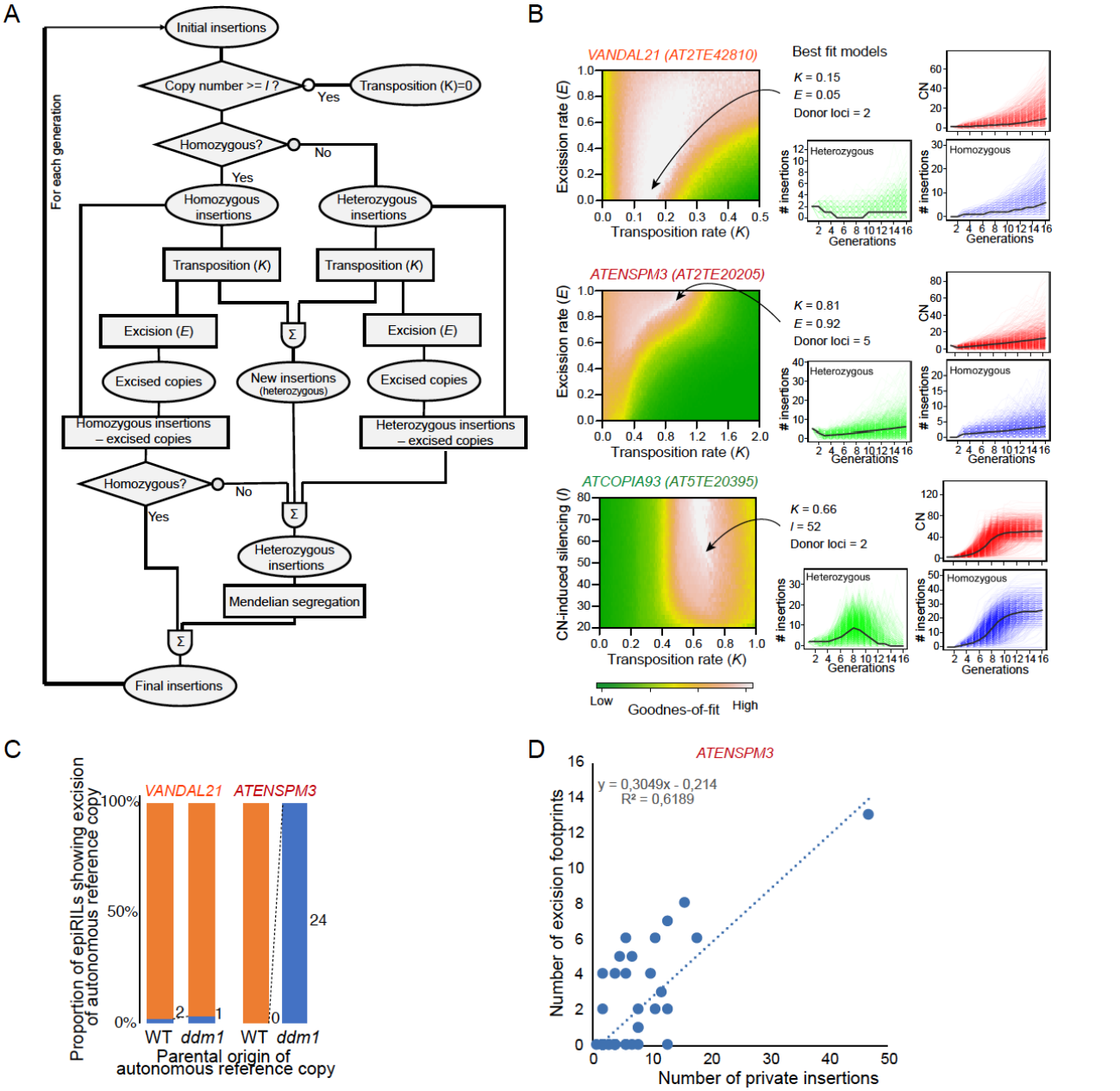


**Figure S3. Modeling of insertion accumulation dynamics.**
**A**. Schematic representation of the transposition-drift model developed to reconstruct transposition dynamics. **B**. Goodness-of-fit between simulated and observed data. Patterns of insertion accumulation produced by the best fitted models. **C**. Proportion and number of epiRILs showing excision of the autonomous reference copy for *VANDAL21* and *ATENSPM3* in relation to their parental origin. **D**. Correlation between the number of private *ATENSPM3* insertions and excision footprints detected in the epiRILs.

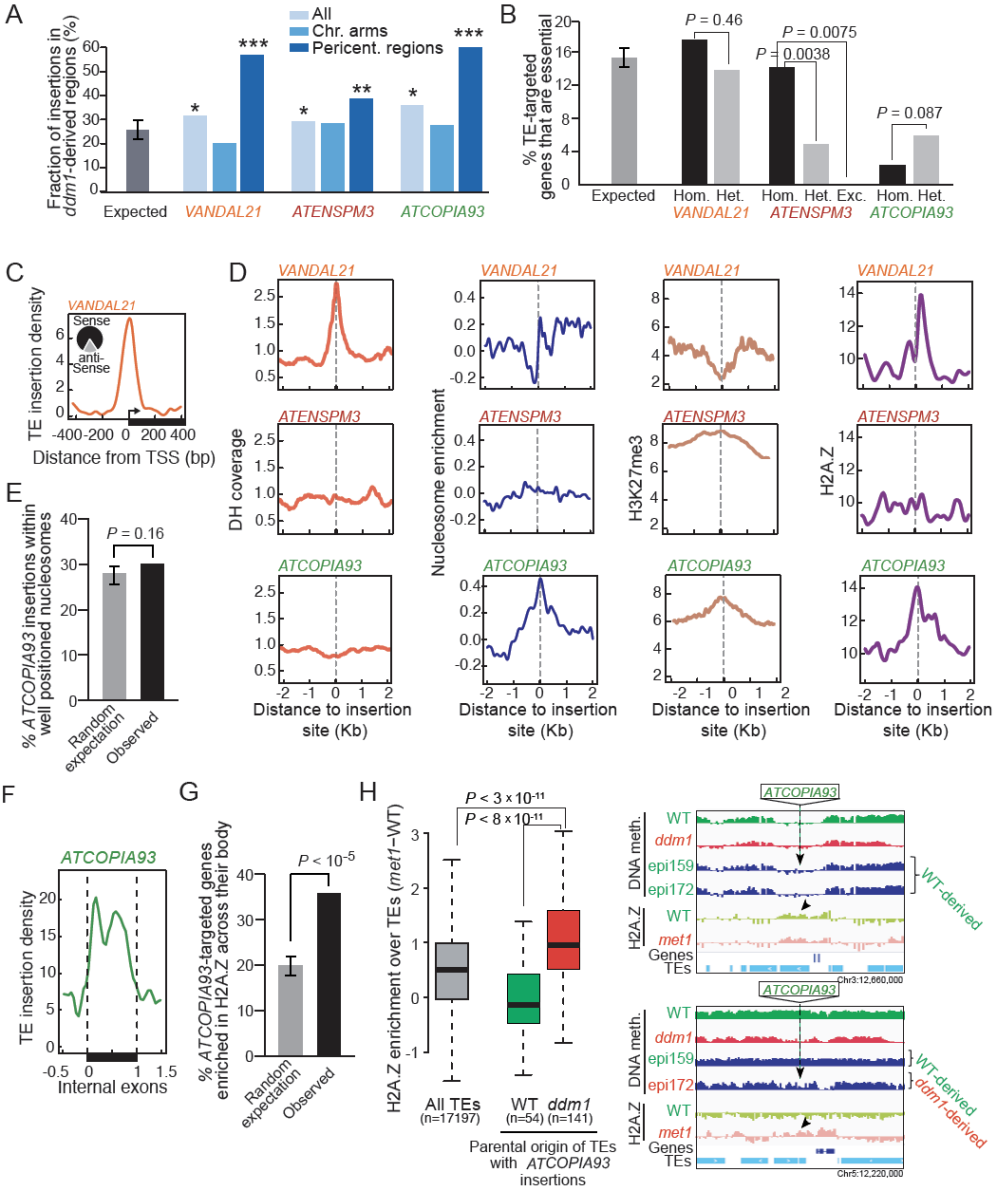


**Figure S4. *ddm1*-derived pericentromeric regions are preferentially target by TE insertions.**
**A.** Proportion of private TE insertions in *ddm1*-derived intervals in relation to their chromosomal position. Statistically significant differences compared to the expected values are indicated (Chi-square test,* P <0.05, ** P < 0.001, *** P<0.0001). **B.** Fraction of essential genes among those targeted by *VANDAL21*, *ATENSPM3* or *ATCOPIA93* in the epiRILs. Fraction of essential genes among those containing short indels compatible with *ATENSPM3* excision footprints is also indicated. Statistical significance for each comparison was obtained using the Chi-square test. **C.** Density of *VANDAL21* insertions around transcriptional start sites (TSS). The fraction of insertions that are in the same (sense) or opposite (antisense) orientation relative to the targeted gene are indicated. **D**. Meta-analysis of of DNAse hypersensitivity (DH) levels as well as nucleosome, H3K7me3 and H2A.Z densities around insertion sites for *VANDAL21*, *ATENSPM3* and *ATCOPIA93*. **E**. Proportion of *ATCOPIA93* private insertions within well-positioned nucleosomes. Expected distribution was obtained by randomizing 1000 times insertion site positions across the genome and performing a randomization test. Errors bars represent the 95% confidence interval. **F**. Density of *ATCOPIA93* insertions within internal exons of protein coding genes. **G**. Fraction of genes containing *ATCOPIA93* private insertions that are also enriched in H2A.Z across their body. Expected distribution was obtained by randomizing 1000 times insertion site positions across the genome and performing a randomization test. Errors bars represent the 95% confidence interval. **H**. Relative level of H2A.Z in *met1* compared to wild type (Zilberman et al 2008) over all TEs (grey box) and in the subset of TEs that contain *ATCOPIA93* insertions in the epiRILs (green box: TEs located in wild-type-derived regions; red box: TEs located in *ddm1*-derived regions (in green and red, respectively). Genome browser views of DNA methylation and H2A.Z over TEs containing *ATCOPIA93* insertions within wild-type- and *ddm1*-derived intervals (top and bottom panel, respectively) are depicted on the right. Insertion sites are indicated by an arrow and samples showing hypomethylation across the region are highlighted in red. Statistical significance for each comparison were obtained by Mann–Whitney test.

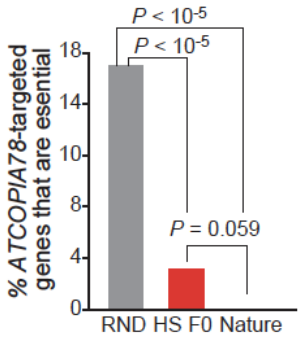


**Figure S5. *ATCOPIA78* avoids integration within essential genes.**
**A**. Fraction of essential genes among those targeted by *ATCOPIA78* in response to heat stress or in nature. Statistical significance for each comparison was obtained using the Chi-square test.

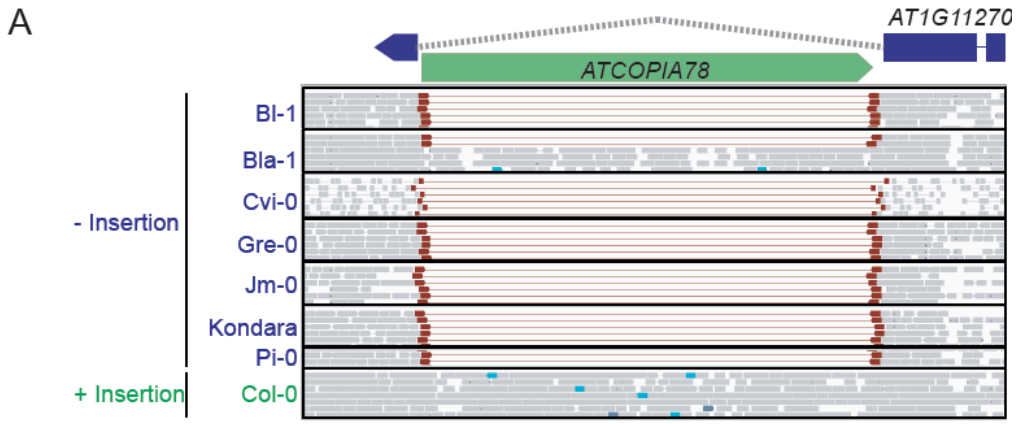


**Figure S6. Col-0 allele of *AT1G11270* contains a recent *ATCOPIA78* insertion in the second intron.**
**A.** Genome browser view of sequence reads produced by WGS over *AT1G11270* for Col-0, which contains an *ATCOPIA78* insertion within the second intron, and seven other accessions that lack this insertion.

## 6.3 Conclusion et discussion

Dans ce chapitre, j'ai présenté et exploité ce qui fait de la population epiRIL un système unique pour l'étude directe et à grande échelle de la contribution des ET aux paysages mutationnels.

Je mets en évidence que le spectre des indels reflète le patron de remobilisation du transposon à ADN *ATENSPM3*, qui est réactivé de façon importante dans les epiRIL avec 665 nouvelles insertions détectées (379 homozygotes, 286 hémizygotes), auxquelles il faut ajouter les 82 évènements d'excision que je détecte sous la forme d'insertions de 3 à 5 pb, correspondant à une répétition en orientation directe du TSD et présentant les mêmes préférences d'insertions que le transposon (état chromatinien CS2, enrichi en H3K27me3). Cette détection à l'échelle du génome des évènements d'excision d'un élément transposable n'avait pas été réalisée à ce jour. Il est à noter qu'un deuxième ET de classe II, *VANDAL21*, est également activement remobilisé dans la population epiRIL, néanmoins aucune trace d'excision ne peut être détectée. Cela est à mettre en regard du faible taux d'excision de cet ET notamment dans la lignée germinale (Fu et al. 2013 ; Quadrana, Etcheverry et al. 2018).

Peut-on imaginer employer une détection des évènements d'excision comme celle présentée ici pour préciser la dynamique des ET de classe II au sein des populations naturelles d'Arabidopsis, dans desquelles la mobilisation de plusieurs de ces ET (dont *ATENSPM3*) a pu être mise en évidence (Quadrana, Bortolini Silveira et al. 2016) ?
Cela semble peu probable en raison de l'important niveau de polymorphisme distinguant ces accessions (Zapata et al. 2016) qui va grandement complexifier la détection des indels, notamment du fait de mésalignements fréquents (Y. Jiang et al. 2015). Néanmoins, compte-tenu de la préférence pour l'état chromatinien CS2 mise en évidence pour cet ET, il pourrait être pertinent d'analyser plus spécifiquement ces régions dans les différentes accessions. Cela présuppose néanmoins et que cette préférence d'insertion soit conservée parmi les différentes accessions, et qu'il en soit de même pour le marquage H3K27me des gènes. Dans la mesure où les préférences d'insertions sont conservées entre espèces distantes pour au moins certaines familles d'ET (ET de type COPIA chez *A. thaliana*, *O. sativa* et *S. cerevisiae* - décrit dans (Quadrana, Etcheverry et al. 2018)), et que de façon similaire le marquage H3K27me3 est partiellement conservé parmi les Brassicacées (*A.thaliana*, *A.alpina* et *A.lyrata* - décrit dans (Chica et al. 2017)), une telle analyse semble envisageable.

Au contraire du spectre des indels qui reflète la remobilisation d'*ATENSPM3*, aucun évènement de réarrangement chromosomique associé à la remobilisation des ET ne peut être identifié dans les epiRIL. Plusieurs hypothèses peuvent être proposées pour expliquer

cette observation.

Premièrement, il est possible de tels évènements aient lieu non pas dans les cellules qui contribueront à la génération suivante mais dans les tissus somatiques et qu'ils ne soient donc pas détectables ici, ou encore qu'ils aient pu se produire et conduire à un arrêt précoce de la méiose, un cas de figure qui conduirait à une létalité et ne peut une fois encore pas être documenté ici.
Une deuxième hypothèse aurait à voir avec le nombre et la localisation chromosomique des néo-copies, entre elles ainsi que par rapport aux copies existantes. Au contraire des copies résidentes (les annotations d'ET du génome) qui sont localisées dans les régions péricentromériques, la majorité des néo-insertions est localisée dans les bras, or comme nous l'avons vu dans l'introduction une NAHR requiert que les séquences matrices de la recombinaison ne soient pas trop distantes l'une de l'autre. De façon plus générale, la faible proportion en ET du génome d'Arabidopsis en comparaison de celui du riz ou du maïs pourrait expliquer pourquoi de tels réarrangements sont plus fréquemment décrits dans ces derniers organismes (J. Ma et al. 2006 ; B. Tan et al. 2017) mais peu voir pas du tout chez Arabidopsis. Comme illustré Figure 6.1, près de la moitié (1107/1670) des néo-insertions d'ET détectées dans les epiRIL sont à l'état hémizygote, ce qui indique qu'en génération F8, la remobilisation est toujours en cours et un nombre accru de nouvelles insertions est attendu dans les générations suivantes. Il serait intéressant de suivre si dans les générations avancées des epiRIL, la plus forte densité en ET peut conduire à l'observation de réarrangements. Dans ce contexte, j'ai pu identifier dans une epiRIL en génération F16 un réarrangement putatif entre deux nouvelles copies d'*ATCOPIA93*. Ces deux néo-insertions, l'une détectée dans la F8 et l'autre apparue entre la F8 et la F16, sont situées à approximativement 1Mb de distance et en orientation inverse l'une par rapport à l'autre, et l'orientation des paires de reads dans cette région laisse suspecter une inversion de la région située entre ces deux copies (données non montrées), dont une validation expérimentale serait nécessaire.
La troisième hypothèse a trait à la nature des ET remobilisés dans les epiRIL. Si les 10 familles d'ET remobilisées dans les epiRIL le sont également dans les populations naturelles de l'accession Col-0, elles ne représentent qu'une fraction de la diversité des familles d'ET remobilisées parmi les accessions naturelles d'Arabidopsis (Quadrana, Bortolini Silveira et al. 2016), et cette même remarque peut être faite en comparant les familles d'ET actives chez Arabidopsis et chez d'autres organismes. Il est dès lors envisageable que les ET remobilisés ici ne fassent pas partie de ceux qui, comme *Ac/Ds* chez le Maïs ou l'élément P chez la Drosophile, créent des réarrangements au cours de la transposition. Cette dernière hypothèse apparaît d'autant plus envisageable que la transposition alternative peut créer une diversité de réarrangements, et qu'à ce jour la quasi-totalité des évènements de réarrangements impliquant des ET correspondent à ce type d'évènement (Gray 2000).

Si les analyses décrites dans ce chapitre visaient en premier lieu à détecter des réarrangements associés aux ET, elles permettent également d'identifier des CNV spontanés. Une analyse similaire avait été conduite par le passé et avait conclu à un taux extrêmement élevé de CNV dans l'accession Col-0 d'Arabidopsis (DeBolt 2010). Ici, en combinant les observations faites dans les MA lines et les epiRIL, qui dans leur ensemble représentent un peu de plus de 800 générations d'évolution indépendante (92 epiRIL sur 6 générations plus 10 MA lines sur 30 générations), je n'identifie pas plus de 15 CNV. Par ailleurs, la réanalyse des données disponibles pour une population de type MA lines soumise à un stress salin (C. Jiang et al. 2014) met également en évidence un nombre restreint de CNV, suggérant que même dans des conditions de stress ce type de mutation ne se produit pas à un taux accru chez Col-0 (données non montrées).
Ces conclusions opposées peuvent être expliquées en partie par des différences dans les méthodes de détection : séquençage et analyse CNV/read-pair dans les travaux présentés ici contre hybridation comparative du génome sur puce à ADN (*comparative genome hybridization*) dans (DeBolt 2010), cette dernière étant peu résolutive et "bruitée" en raisons de variations dans le signal d'hybridation. Par ailleurs, (DeBolt 2010) identifie des CNV récurrents, apparus dans plusieurs lignées apparentées, ce qui pourrait suggérer qu'il s'agisse à la place d'un CNV en ségrégation dans la lignée initiale, une validation qui n'est pas décrite.
Parmi les CNV détectés et dans les MA lines et dans les epiRIL, la majorité d'entre eux correspondent à des duplications en tandem situées au sein d'une annotation génique et présentant un signature de microhomologie indicative d'un mécanisme réplicatif ; soient les mêmes spécificités que décrites chez l'Homme. Ces similitudes entre epiRIL et MA lines sont par ailleurs très cohérentes compte-tenu l'absence de fonction de *ddm1* au niveau des gènes.
Dans ces deux populations, toutes les délétions sont associées à des annotations d'ET et résultent probablement d'une NAHR. Il est surprenant d'observer qu'à l'opposé, aucune duplication correspondant ne peut être décrite ici. Ce résultat peut être interprété dans le contexte des observations précédentes effectuées chez Arabidopsis et qui suggèrent une tendance à la réduction du génome chez cet organisme par réduction du contenu en ET (Devos et al. 2002 ; Vu et al. 2017). Une observation générique concernant les génomes des plantes est qu'ils présentent une taille importante, en lien avec un contenu élevé en ET. Au moyen d'épisodes de transposition soudains et extensifs (*burst*), les ET peuvent conduire à un accroissement rapide de la taille du génome, cependant les mécanismes pouvant permettant un retour en arrière sont moins bien décrits, d'où l'hypothèse dite du "*one-way ticket to genome obesity*" (Bennetzen et al. 1997). Les résultats présentés ici suggèrent, dans la continuité de descriptions précédentes, qu'*Arabidopsis thaliana* tout comme *Arabidopsis lyrata* ait une tendance à la réduction de la taille de son génome par

formation de délétions associées aux ET (Hu et al. 2011).

Par ailleurs, dans la suite des travaux décrits dans (Wicker et al. 2016), il serait pertinent d'évaluer si les évènements d'excision ou de néo-insertion d'ET se traduisent par un taux de mutation accru à ces loci en raison du recrutement de polymérases *error-prone* pour résoudre les cassures double-brins provoquées par de tels évènements. Ces analyses sont actuellement en cours et permettront de mieux documenter les conséquences mutationnelles d'une mobilisation des ET.

Chapitre 7

# Vers la caractérisation de QTL épigénétiques

## 7.1 Introduction

Ce dernier chapitre de résultats porte sur un axe distinct de mes travaux de thèse. Mon projet visait initialement à la caractérisation de QTL épigénétiques préalablement mis en évidence au laboratoire, au travers d'une combinaison d'approches de cartographie fine et de validation de loci candidats par complémentation fonctionnelle. Ce projet nécessitait à la fois la création de populations de plantes supplémentaires ainsi que la mise en place d'outils moléculaires fondés sur le système CRISPR/Cas pour aller altérer le profil de méthylation de l'ADN à des loci spécifiques.

Ces travaux n'ont pas pu être menés à leur terme dans le temps imparti à ma thèse en raison d'aléas techniques (invasions récurrentes d'insectes dans les enceintes de culture ayant conduit à la destruction d'une grande partie du matériel biologique développé), mais l'approche CRISPR-Cas est poursuivie par Erwann Caillieux, ingénieur au laboratoire.

Dans ce chapitre, je présente et discute l'impact des variants ADN en ségrégation dans les epiRILs sur les QTLepi détectés, la cartographie fine de l'un de ces QTL ainsi que la stratégie expérimentale mise en place pour effectuer à terme la complémentation fonctionnelle des épiallèles candidats. La caractérisation des variants en ségrégation m'a par ailleurs amenée à être impliquée dans plusieurs projets collaboratifs avec des équipes recherchant des QTLepi au sein de la population epiRIL ; l'un des manuscrits qui en résulte est joint à ce chapitre.

## 7.2 Résultats

### 7.2.1 Contribution des variants ADN en ségrégation à la variation héritable détectée au sein des epiRIL

Comme décrit Chapitre 4, le fait que la majorité des variants ADN en ségrégation dans les epiRIL ne co-ségrège pas avec les épihaplotypes suggère que ceux-ci ne soient pas à l'origine des QTLepi détectés.
En effet, comme nous l'avons vu dans le Chapitre 3 de l'introduction, la détection des QTLepi s'est faite sur la base des profils de méthylation. Aussi, seuls les variants qui sont associés à l'un ou à l'autre des épihaplotypes sont susceptibles d'être à l'origine des QTLepi détectés. Dans la mesure où l'épihaplotype WT est partitionné entre variants dérivés du WT2 d'une part et variants dérivés du WT employé lors du croisement initial d'autre part, il n'existe aucun variant ADN qui co-ségrège exclusivement avec cet épihaplotype. A l'opposé, seuls les variants (au nombre de 4) hérités de l'individu *ddm1* et qui co-ségrègent donc avec l'épihaplotype hypométhylé sont susceptibles d'être causatifs pour les QTLepi.

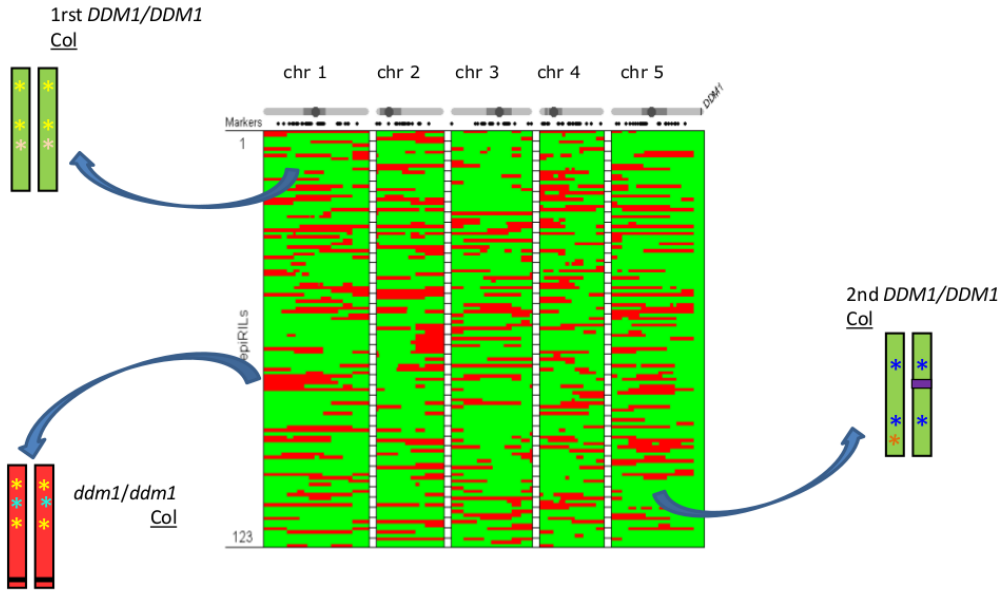


Figure 7.1 – Illustration du patron de ségrégation haplotype-épihaplotype. L'épihaplotype WT est partitionné en deux haplotypes, l'un d'entre eux étant commun avec l'épihaplotype *ddm1*

Comme décrit dans le Chapitre 3, des analyses précédentes menées au laboratoire en collaboration avec l'équipe de Frank Johannes avaient visé à déterminer si les nouvelles insertions d'ET partagées par plusieurs lignées et localisées au sein des intervalles de support des 6 QTLepi détectés (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) expliquaient une fraction significative de l'effet de ces différents QTL.

Pour ma part, je me suis concentrée sur les quatre mutations ponctuelles que j'ai pu identifier comme co-ségrégeant systématiquement avec l'épihaplotype *ddm1*. Il s'avère que parmi ces quatre variants, un seul est localisé au sein de l'intervalle de support de l'un des 6 QTLepi identifiés dans (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) et celui-ci correspond à une mutation au sein d'une annotation d'ET (Table 4.2. En conséquence, il semble peu probable qu'il soit à l'origine de la variation héritable détectée pour le trait. Ce résultat complète les analyses réalisées par l'équipe de F. Johannes pour les insertions d'ET partagées (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) et suggère donc que ces 6 QTLepi sont effectivement causés par des épiallèles.

Néanmoins, puisque ces variants ADN représentent une seconde source de variants parentaux en ségrégation au sein des epiRIL (la première étant les DMRs stables), ils sont susceptibles de contribuer à la variation héritable identifiée pour les différents traits analysés à ce jour. Sans qu'ils se substituent aux QTL "épigénétiques" (portés par des polymorphismes de méthylation) détectés, nous aurions alors des QTL ADN "classiques" (portés par des polymorphismes nucléotidiques) que je me suis employée à rechercher.
En raison du faible nombre de variants ADN distinguant le WT2 des individus *ddm1* et wt, il n'est pas possible de construire une carte génétique qui permettrait de détecter de tels QTLadn. J'ai donc testé pour chaque variant ADN la corrélation entre génotype à ce marqueur et et valeur phénotypique dans chacune des epiRIL. Il faut noter qu'il s'agit là d'une analyse de première intention, au sens où elle permet de détecter la présence d'un QTL mais pas d'y associer un intervalle de support.

Les profils de LOD score pour ces deux traits sont donnés Figure 7.2A. Pour les deux traits, 3 variants ADN atteignent ou dépassent le seuil de significativité. L'un d'entre eux, mis en évidence par un rond azur, est commun aux deux profils et est au sein des intervalles QTLepi décrits sur le chromosome 1 pour l'un et l'autre des traits considérés. Dans la mesure où ce variant est hérité du parent *ddm1*, cette association avec le phénotype est attendue puisqu'il n'est pas possible de le distinguer des épihaplotypes.
A l'opposé, on constate pour la date de floraison que les variants qui dépassent le seuil de significativité correspondent à des variants qui distinguent le WT2 des individus WT et *ddm1* à l'origine de la population. Cela correspond à un QTL "ADN" et non pas épigénétique, porté non pas par des différences de méthylation mais des polymorphismes distinguant le WT2 et les individus fondateurs, et qui ne recoupe pas le QTLepi détecté pour le même trait sur ce chromosome. Concernant l'effet associé à l'un ou l'autre de ces génotypes (et non pas épigénotypes), on constate que la présence de l'un ou l'autre des haplotypes -indépendamment du profil de méthylation- est associé à une différence d'un peu plus de 2 jours dans la date de floraison, soit du même ordre de grandeur que celui observé pour les 3 QTLepi décrits pour la date de floraison dans les epiRILs (Figure

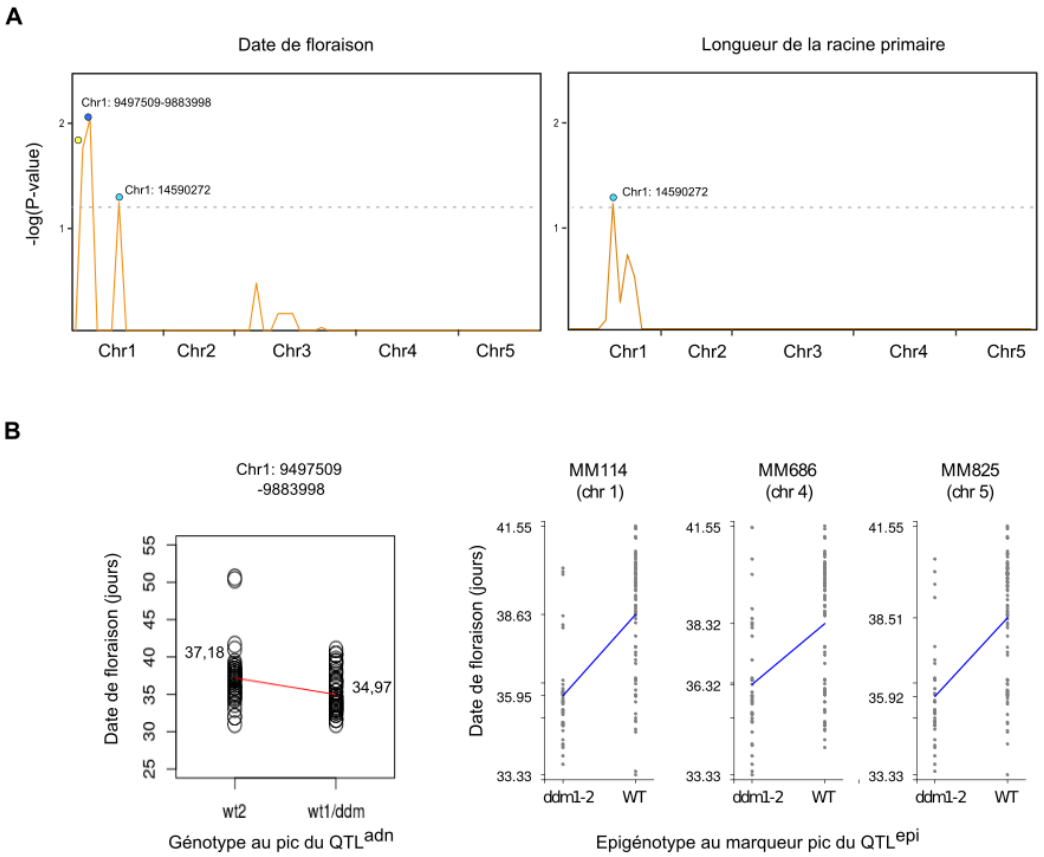


FIGURE 7.2 – Résultats de la recherche de QTLadn pour la date de floraison et la longueur de la racine primaire par analyse simple marqueur.
A. Profils de LOD score. La couleur des ronds indique l'origine parentale du variant, les coordonnées chromosomiques leur identité. B. Amplitude et direction des effets pour un QTLadn affectant la date de floraison dans les epiRIL (gauche) et comparaison avec les QTLepi détectés pour ce trait (droite, panel tiré de (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014)).

7.2B).

Aucun des gènes situés à proximité de ces deux variants n'est un candidat évident pouvant expliquer une variation dans la date de floraison, une observation qui est peu surprenante compte-tenu de la faible amplitude des effets [1]. Il faut par ailleurs garder à l'esprit que la date de floraison est un trait fortement susceptible à des stimuli multiples, en l'occurrence dans le cas de QTL à effet mineur comme celui-ci (ou même les 3 QTLepi identifiés pour ce trait), il est probable que les loci affectés aient des rôles plus "pléiotropes" et touchent plus généralement à la vigueur de la plante, dont la date de floraison est un read-out parmi d'autres (biomasse, taille de la plante).

1. Dans le contexte de la date de floraison, une différence de 2 jours est considérée comme un QTL à faible effets. Des QTL à fort effet correspondent à des loci expliquant une variation de l'ordre de la dizaine de jours, comme il en existe entre les différentes accessions d'Arabidopsis. Cependant, un faible effet peut être fortement héritable comme c'est le cas dans les epiRIL, et inversement.

Cette observation met en évidence la nécessité de ré-évaluer la fraction de l'héritabilité pour la date de floraison expliquée par les QTL détectés pour ce trait. L'héritabilité -au sens large- s'évalue sur des bases phénotypiques (la variance inter-lignée, (Johannes, Porcher et al. 2009 ; Lynch et Walsh 1998)), et il a été estimé que les 3 QTLepi contribuaient au total à environ 85% de l'héritabilité pour la date de floraison (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014). Dans ce contexte, il est tentant de spéculer que les 15% restant puissent être expliqués au moins en partie par ce QTLadn. Dans tous les cas, en l'absence de QTLadn détecté pour la longueur de la racine primaire, cette hypothèse ne peut être invoquée pour expliquer les 40% d'héritabilité manquante après prise en compte des 3 QTLepi pour ce trait.

Une deuxième approche peut être envisagée afin d'évaluer l'impact potentiel des variants ADN en ségrégation sur les QTLepi identifiés au sein des epiRIL. De façon similaire aux analyses effectuées dans (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) et (Kooke et al. 2015) pour évaluer l'impact des ET présents dans les intervalles QTLepi, il s'agit de recalculer l'effet de chaque QTL en substituant le marqueur au pic par le variant ADN à tester, et ainsi d'interroger si le variant ADN explique plus que le marqueur de méthylation.
Une telle analyse a été effectué dans le cadre d'une collaboration avec l'équipe de Maria Manzanares (manuscrit Liégard et al, Section 7.2.4). Parmi les 20 QTLepi identifiés dans la population epiRIL pour la résistance au pathogène responsable de la hernie des Brassicacées, après prise en compte des ET et autres variants ADN en ségrégation, nous avons pu montrer que si 16 QTL sont à proprement causés par des variants de méthylation, pour quatre d'entre eux, un variant ADN hérité du WT2 explique une fraction plus importante de l'effet que le marqueur de méthylation au pic. Deux explications peuvent être formulées pour expliquer ce résultat : ou bien le variant ADN en question est causatif pour le QTL, une hypothèse qui n'est pas privilégiée en raison de la localisation génomique du variant, ou il est en déséquilibre de liaison avec le vrai polymorphisme causatif, qu'il s'agisse ou non d'un épivariant.

Dans leur ensemble, ces observations suggèrent qu'en dépit de la faible amplitude de variation nucléotidique en ségrégation au sein de la population epiRIL, il est nécessaire de la garder à l'esprit lorsque l'on s'intéresse à des traits complexes.

### 7.2.2 Mise en place de populations de cartographie fine pour un QTLepi

Comme décrit Chapitre 3, démontrer la causalité d'un épiallèle associé à un trait complexe va requérir la même stratégie que pour un variant ADN classique : il s'agit en premier lieu d'isoler la région d'intérêt (mendélisation), de valider la co-ségrégation du phénotype et de l'épigénotype dans cette nouvelle population, puis dans un second temps d'effectuer la cartographie fine à proprement parler, c'est-à-dire réduire la taille de la région incluant le QTL au moyen de cycles successifs de recherche de recombinants et phénotypage.

Cette étape de mendélisation est une étape cruciale : en effet, la détection primaire d'un QTL, telle qu'effectuée dans les epiRIL, va résulter en des intervalles de confiance statistiques de l'ordre du Mb. Cependant, ces intervalles ne sont en rien des intervalles "physiques", et de la même façon que le variant causatif n'est pas nécessairement situé au pic du QTL (loin s'en faut), il peut aussi bien être localisé au-delà des limites de l'intervalle. Il va donc s'agir de mettre en place des populations de plantes dans lesquelles ne ségrègent que l'intervalle QTL à tester afin de confirmer la co-ségrégation du phénotype et de l'épigénotype.

Parmi les 6 QTL affectant la longueur de la racine primaire et la date de floraison décrits dans (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014), je me suis concentrée sur le QTL RLch4 (*root length-chromosome 4*, Figure 7.3). Bien que ce QTL présente le plus faible effet, ce choix a été motivé par les deux raisons suivantes. D'une part, les ressources disponibles au laboratoire ne permettent pas d'effectuer un phénotypage de la date de floraison à grande échelle et de façon répétée, ce qui restreint les possibilités aux QTL de longueur de racine, un phénotype observable en 11 jours après semis des graines (contre environ un mois pour la date de floraison, comme illustré Figure 7.2 et aisément réplicable. D'autre part, parmi les trois QTL affectant la longueur de la racine illustrés Figure 7.3, seul RLch4 est situé hors des régions péricentromériques. Dans ces régions, l'intensité de recombinaison est plus faible que dans les bras (Colomé-Tatché et al. 2012), ce qui limite donc la possibilité d'obtenir des recombinants. De plus, seul un faible nombre de DMR sont présentes dans cet intervalle, ce qui réduit d'autant la liste des épimutations causales.

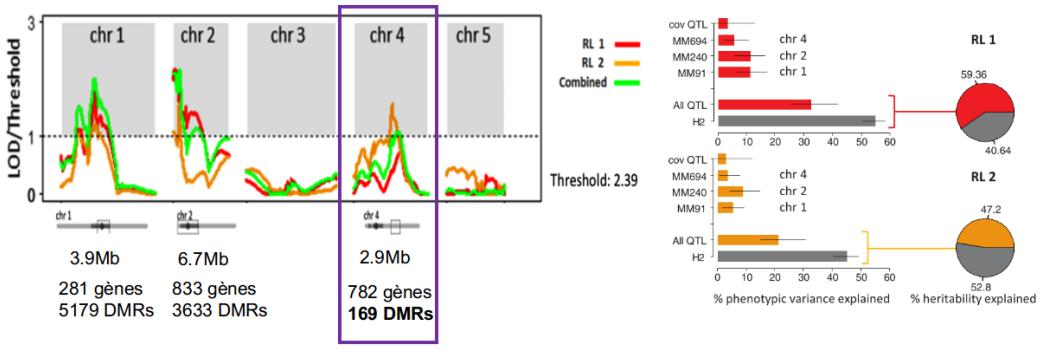


FIGURE 7.3 – QTLepi pour la racine primaire. Les profils de LOD score pour les deux mesures de racine primaire effectuées (RL1 et RL2) ainsi que les valeurs d'héritabilité et effets des différents QTL détectés sont figurés en rouge et orange respectivement. Pour les 3 QTL affectant la longueur de la racine, le positionnement par rapport aux régions péricentromériques, la longueur de l'intervalle de support et le nombre de gènes et de DMR dans cet intervalle est donné sous le profil de LOD. Le QTL RLch4 est encadré en violet. Modifié d'après (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014).

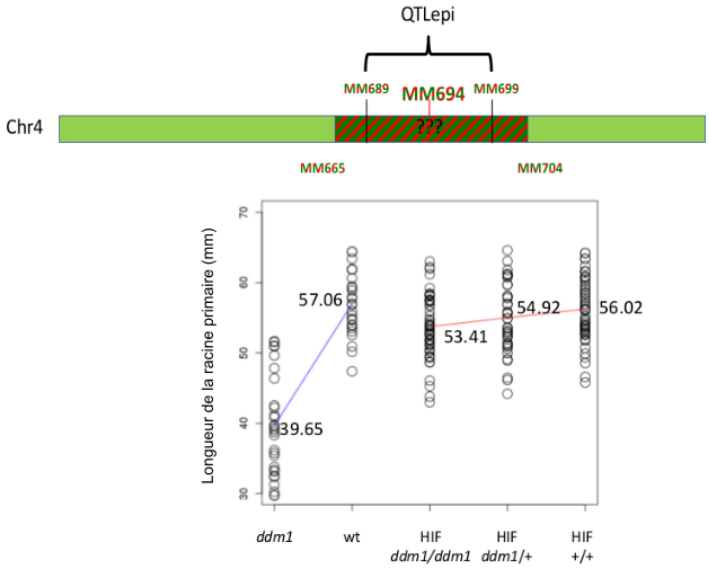


FIGURE 7.4 – Validation de l'HIF établie à partir de l'epiRIL172, dans laquelle ségrège l'intervalle incluant le QTL RLch4. Panel du haut : Illustration du chromosome 4 de l'HIF. L'intervalle en ségrégation, délimité par MM665 et MM704, est figuré par des rayures rouges et vertes. Panel du bas : longueur de la racine primaire dans les lignées issues de l'HIF et dans lesquelles l'intervalle à tester est dérivé de *ddm1*, dérivé du wt ou épihétérozygote ; ainsi que chez le mutant *ddm1* et un individu sauvage.

Afin d'établir des populations ségrégeant exclusivement l'intervalle de support du QTL RLch4, une approche rapide consiste en l'établissement d'HIF (voir Chapitre 3). Dans le cas des epiRIL, cette "épihétérozygotie" peut être détectée sur la base d'une succession de sondes présentant des niveaux de méthylation intermédiaire (sondes I). Une telle lignée n'ayant pas pu être identifiée pour la région d'intérêt, j'ai entrepris de créer des "épihétérozygotes" pour la région incluant l'intervalle de support du QTL RLch4 par croisement entre un individu wt et des epiRILs sélectionnées selon les critères suivants : état dérivée de *ddm1* pour la région RLch4, état dérivé de wt pour les intervalles de support des autres QTL de longueur de racine (afin que leur co-ségrégation avec RLch4 ne fausse pas les phénotypages), cohérence génotype-épigénotype (racine courte), faible contenu en régions dérivées de *ddm1* sur l'ensemble du génome (afin de réduire le nombre de régions en ségrégation dans la descendance du croisement).

Parmi les différentes epiHIF établies, l'une d'entre elle a pu être validée (Figure 7.4). Depuis l'obtention de ce résultat, le matériel biologique correspondant a été détruit et il sera donc nécessaire de re-créer cette ressource afin de poursuivre la cartographie fine de cet intervalle. Cependant, cette validation est d'ores et déjà un résultat important, puisque j'ai pu isoler la région incluant le QTL et confirmer ses effets. Des approches complémentaires à la cartographie fine, et qui pourront s'y substituer partiellement pour identifier le variant causal de ce QTL une fois qu'une HIF correspondante aura été réétablie, sont proposées dans la discussion de ce chapitre.

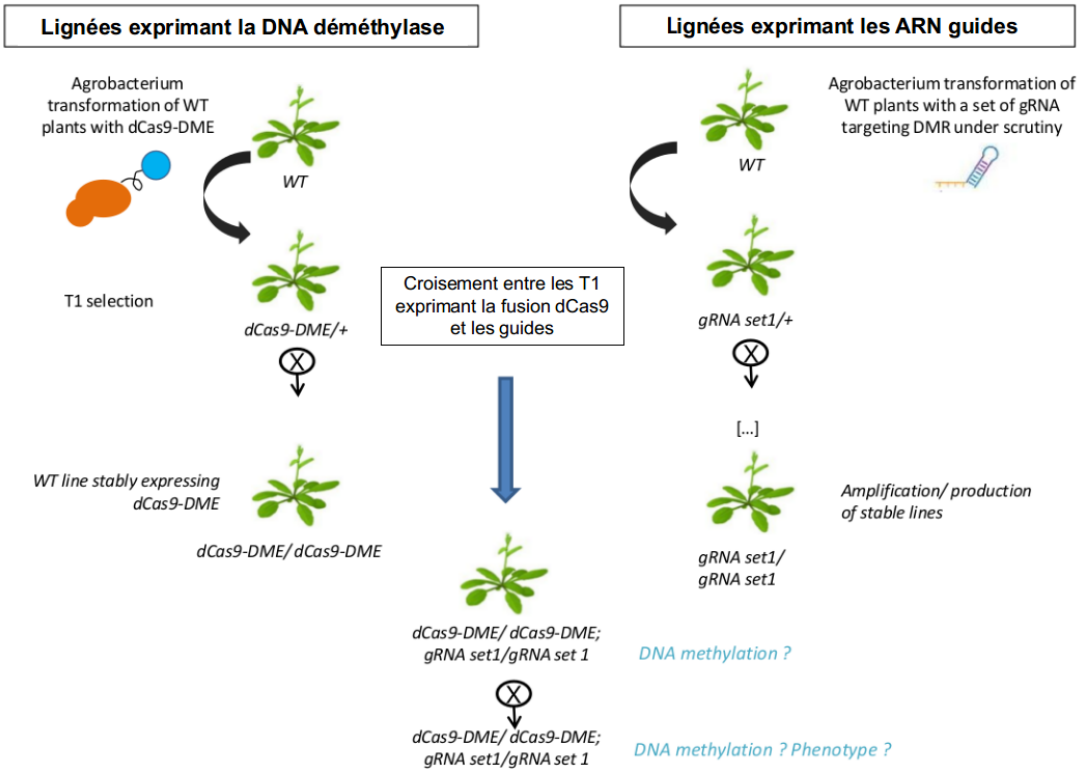


FIGURE 7.5 – Stratégie expérimentale de déméthylation ciblée. Détails dans le texte.

### 7.2.3 Vers une complémentation fonctionnelle des épiallèles par édition de l'épigénome

Comme présenté Chapitres 2 et 3, la démonstration de la causalité d'un épiallèle requiert sa complémention fonctionnelle, laquelle peut à présent s'effectuer au travers de l'édition de l'épigénome.

Dans ce contexte, j'ai mis en place au laboratoire les outils de biologie moléculaire visant à effectuer une modification ciblée de l'état de méthylation d'un locus candidat.

L'approche que nous avons choisie consiste à aller induire la déméthylation, plutôt que la reméthylation, du locus d'intérêt. En effet, une telle approche nous permet de travailler dans un fond génétique sauvage -tandis que rechercher une reméthylation imposerai d'être dans un fond génétique déméthylé au moins au locus ciblé-, donc de s'affranchir des contraintes associées à l'emploi de mutants de méthylation, et notamment dans le cas du mutant *ddm1* l'apparition d'altérations phénotypiques au fil des générations d'autofécondation successives, ce qui s'avère rédhibitoire dans la mesure où plusieurs cycles d'autofécondation sont nécessaires pour valider un transgénique.

Plutôt qu'utiliser un vecteur bifonctionnel, c'est-à-dire qui porte à la fois les ARN guides et la protéine effectrice, nous avons choisi de développer un système à deux composantes, illustré Figure 7.5. Cette approche a été privilégiée pour les raisons suivantes : i) elle rend possible la sélection des "meilleures" lignées effectrices au regard du niveau d'expression du transgène, ii) elle permet de faire ségréger l'effecteur (la Cas9-déméthylase) des guides et donc de tester la transmission de l'état déméthylé en l'absence de l'inducteur et iii) une fois mise en place, elle constituera un système hautement versatile au sens où la déméthylation d'un locus donné pourra être réalisée par simple croisement avec toute plante portant le set de guides ciblant la région choisie.

Les deux composantes, effecteur et guides, sont décrites dans les paragraphes qui suivent.

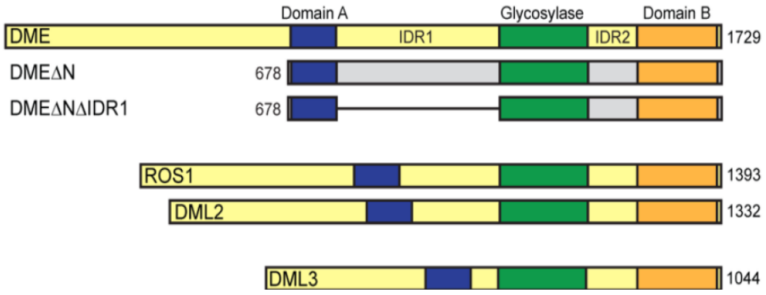


FIGURE 7.6 – Représentation schématique des différentes DNA déméthylases d'Arabidopsis. DMEΔNΔIRD1, caractérisée dans (Mok et al. 2010), est la DNA déméthyase "minimale" employée ici. Les domaines A et B sont conservés parmi les paralogues de DME, ne présentent aucune homologie à une séquence protéique connue et sont nécessaires à la fonction de DME. IRD : *interdomain region*, région sans structure définie reliant deux domaines. Dans DMEΔNΔIRD1, IRD1 a été remplacé par un dodécapeptide synthétique à fonction exclusive de linker. Extrait de (Brooks et al. 2015)

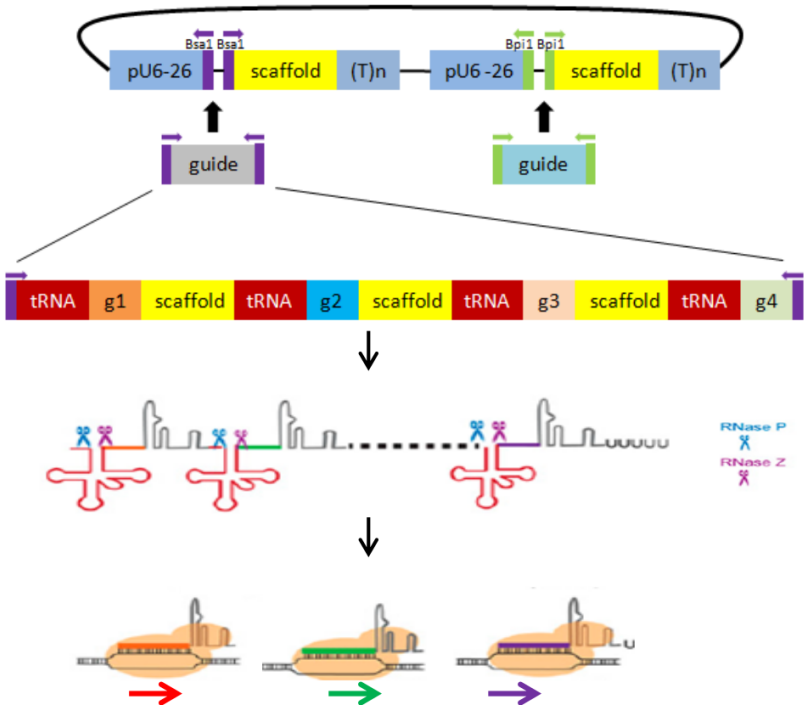


FIGURE 7.7 – Multiplexage d'ARN guides au moyen d'un polycistron. Modifié d'après (K. Xie et al. 2015).

### Fusion dCas9 effectrice

Comme décrit Chapitre 2, l'utilisation du seul domaine catalytique de la protéine effectrice d'intérêt se révèle être plus efficace que la protéine entière, probablement en raison d'un encombrement stérique moindre. Dans ce contexte, notre choix s'est porté sur une protéine DME "minimale" précédemment caractérisée (Mok et al. 2010). Dans le cadre d'une collaboration avec l'équipe de Jin-Hoe Huh (Université de Séoul), nous avons bénéficié d'une fusion dCas9-DMEΔNΔIRD1, dont le domaine fonctionnel est illustré Figure 7.6.

### Dispositif d'expression des ARN guides

Concernant les guides, j'ai adapté au laboratoire une stratégie de multiplexage des guides initialement mise en place chez le riz (K. Xie et al. 2015). Afin de maximiser l'efficacité de ciblage par une activité enzymatique, il s'agit de réaliser un *tilling* du locus d'intérêt, à raison d'environ un guide par section de 100 pb (Thakore et al. 2015). Si dans le cas de cellules en cultures il est possible de transfecter un grand nombre de guides, chez Arabidopsis le moyen le plus facile d'obtenir des lignées transgéniques est de transformer les plantes en plongeant les boutons floraux dans une solution d'Agrobactéries possédant le vecteur portant le transgène d'intérêt. En raison de la faible efficacité de transformation, il est peu envisageable de co-transformer une plante à l'aide d'une solution équimolaire contenant plusieurs vecteurs portant les transgènes d'intérêt. Dans ce contexte, il est nécessaire de trouver un moyen de mettre un maximum de guides sur un même vecteur de transformation stable. Ici, la stratégie de multiplexage employée (illustrée Figure 7.7) combine les différents guides en un transcrit polycistronique tRNA-gRNA (PTG), dans lequel chaque ARN guide est séparé du suivant par une séquence codant pour un ARN de transfert. Cette méthode tire profit des RNAses P et Z endogènes, responsables du clivage du pré-tRNA : ainsi, le transcrit primaire issu du polycistron va être clivé au niveau des pré-tRNA, ce qui libère les différents guides. Dans la mesure où les guides sont tous issus du même transcrit, leur niveau d'expression est identique, d'où une incertitude de moins concernant l'efficacité de ciblage. Par ailleurs, il est à noter que les unités tRNA présentent des séquences de reconnaissance de TFIIA et TFIIB, deux co-facteurs de PolIII, la polymérase responsable de l'expression des tRNA mais également des guides. Il avait pu décrit dans la littérature que l'efficacité des promoteurs employés pour l'expression des guides était variable d'une séquence à l'autre ; aussi, la présence des "boîtes" TFIIA et TFIIB favorise la bonne expression de la construction, qui a été insérée au sein d'un vecteur de clonage de guides "classique".

**Test de l'approche de déméthylation ciblée**

A terme, cette stratégie vise à aller modifier l'état de méthylation quel que soit le locus d'intérêt. Pour l'heure, l'objectif est de tester cette stratégie en allant cibler le locus *FWA* (introduit Chapitre 2). Ce locus présente deux intérêts : d'une part *FWA* est une cible native de DME dans l'albumen, et d'autre part le phénotype attendu (illustré Figure 3.2) est facile à mettre en évidence. Par ailleurs, les différentes constructions ont été introduites également dans un fond génétique *nrpd1a*, qui est défectif pour le RdDM et dans lequel les répétitions en amont du locus ne sont plus ciblées par des siRNA. Il n'y a alors pas de compétition entre la machinerie de déméthylation amenée à ce locus par la construction d'une part, et le RdDM d'autre part. Les expériences sont toujours en cours et des résultats sont attendus prochainement.

### 7.2.4 Manuscrit associé

Ce manuscrit dont je suis deuxième auteur résulte d'une collaboration avec Mélanie Jubault, Benjamin Liégard et Maria Manzanares-Dauleux (INRA Rennes) et a récemment été accepté à *New Phytologist*.

Research

New Phytologist

# Quantitative resistance to clubroot infection mediated by transgenerational epigenetic variation in Arabidopsis

**Benjamin Liégard[1], Victoire Baillet[2], Mathilde Etcheverry[2], Evens Joseph[1], Christine Lariagon[1], Jocelyne Lemoine[1], Aurélie Evrard[1], Vincent Colot[2], Antoine Gravot[1], Maria J. Manzanares-Dauleux[1] and Mélanie Jubault[1]**

[1]IGEPP, INRA, AGROCAMPUS OUEST, Université de Rennes, F-35000, Rennes, France; [2]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), F-75005, Paris, France

Author for correspondence:
*Mélanie Jubault*
*Tel: +33 2 23 48 56 37*
*Email:*
*melanie.jubault@agrocampus-ouest.fr*

Received: *25 June 2018*
Accepted: *26 October 2018*

*New Phytologist* (2019) **222:** 468–479
**doi**: 10.1111/nph.15579

**Key words:** *Arabidopsis*, clubroot, epigenetics, EpiRIL, methylome, *Plasmodiophora brassicae*, quantitative resistance, transgenerational.

## Summary

- Quantitative disease resistance, often influenced by environmental factors, is thought to be the result of DNA sequence variants segregating at multiple loci. However, heritable differences in DNA methylation, so-called transgenerational epigenetic variants, also could contribute to quantitative traits. Here, we tested this possibility using the well-characterized quantitative resistance of Arabidopsis to clubroot, a *Brassica* major disease caused by *Plasmodiophora brassicae*.
- For that, we used the epigenetic recombinant inbred lines (epiRIL) derived from the cross *ddm1-2* × Col-0, which show extensive epigenetic variation but limited DNA sequence variation. Quantitative loci under epigenetic control ($QTL^{epi}$) mapping was carried out on 123 epiRIL infected with *P. brassicae* and using various disease-related traits.
- EpiRIL displayed a wide range of continuous phenotypic responses. Twenty $QTL^{epi}$ were detected across the five chromosomes, with a *bona fide* epigenetic origin for 16 of them. The effect of five $QTL^{epi}$ was dependent on temperature conditions. Six $QTL^{epi}$ co-localized with previously identified clubroot resistance genes and QTL in Arabidopsis.
- Co-localization of clubroot resistance $QTL^{epi}$ with previously detected DNA-based QTL reveals a complex model in which a combination of allelic and epiallelic variations interacts with the environment to lead to variation in clubroot quantitative resistance.

## Introduction

Clubroot caused by the protist *Plasmodiophora brassicae* is a major disease of *Brassicaceae* including the three most economically important *Brassica* species, *B. napus*, *B. rapa* and *B. oleracea* (Dixon, 2009), and the model plant *Arabidopsis thaliana* (Koch *et al.*, 1991). Infection with *P. brassicae* leads to tumorous club formation on roots resulting from cell hyperplasia and hypertrophy (Ingram & Tommerup, 1972). Cropping practices and crop protection products have limited success in controlling clubroot (Dixon, 2009). Currently, one of the most effective ways to limit the impact of this disease is to use resistant varieties (Diederichsen *et al.*, 2009). To date, both qualitative and quantitative trait loci (QTL) for clubroot resistance have been identified in different *Brassicaceae* species (Manzanares-Dauleux *et al.*, 2000a; Rocherieux *et al.*, 2004; Alix *et al.*, 2007; Jubault *et al.*, 2008; Piao *et al.*, 2009; Lee *et al.*, 2016). Current approaches to generating resistant varieties rely mainly on a few loci controlling qualitative resistance, with the inevitable outcome of rapid adaptation of the pathogen populations (Diederichsen *et al.*, 2009). In this context, diversification and access to other sources of clubroot resistance variability is becoming necessary.

Numerous studies (Dowen *et al.*, 2012; Luna *et al.*, 2012; Zhang *et al.*, 2013; Liu *et al.*, 2015; Aoun *et al.*, 2017; Zheng *et al.*, 2017) have reported that plant responses to abiotic (temperature, drought) and biotic stresses could be associated with epigenetic variation in addition to nucleotidic variation. For instance, Dowen *et al.* (2012) and Yu *et al.* (2013) have shown that Arabidopsis mutants altered in the maintenance of DNA methylation in the CG, CHG and CHH contexts, showed strong resistance to *Pseudomonas syringae* pv *tomato* strain DC3000. The role of epigenetics in the expression of adaptive plant traits thus suggests that epigenetic variability could be used for generating stress-tolerant or resistant plants. Furthermore, the occurrence of natural DNA methylation variants (epialleles) in plants (Becker *et al.*, 2011; Schmitz *et al.*, 2011) and their implication in evolution (Weigel & Colot, 2012) suggest that epialleles could be considered as a source of variability in plant breeding. In this 'epigenetic breeding' approach (Gallusci *et al.*, 2017), two conditions are needed: transgenerational inheritance of epialleles and a clear connection between epigenotype and observed phenotype. Previous studies demonstrated that epialleles could be stably transmitted across at least eight generations (Johannes *et al.*, 2009; Teixeira *et al.*, 2009), and that such heritable differences in DNA methylation could

be associated with heritable phenotypic variation for several complex traits (Johannes *et al.*, 2009; Reinders *et al.*, 2009). However, linking heritable phenotypic variation to epigenetic variation remains challenging because of the difficulty in teasing apart its effects to that of DNA sequence variation in natural settings (Johannes *et al.*, 2008; Quadrana & Colot, 2016). This problem can, however, be greatly alleviated in Arabidopsis by using experimental populations of so-called epigenetic recombinant inbred lines (epiRIL), which show extensive epigenetic variation but limited DNA sequence variation (Johannes *et al.*, 2009; Reinders *et al.*, 2009). One such population (Johannes *et al.*, 2009) was indeed used to build a genetic map based solely on heritable differences in DNA methylation (differentially methylated regions, DMR) (Colomé-Tatché *et al.*, 2012) and to identify epigenetic QTL ($QTL^{epi}$) for several complex traits (Cortijo *et al.*, 2014; Kooke *et al.*, 2015; Aller *et al.*, 2018).

In the present study, we used this same epiRIL population to determine the impact of heritable differences in DNA methylation on the response of Arabidopsis to clubroot. We first showed that *ddm1* mutants were less susceptible to clubroot than the wild-type Col-0 and that the assessed subset of 123 epiRIL displayed a wide range of continuous phenotypic responses to clubroot. Twenty $QTL^{epi}$ were detected across the five chromosomes, with a *bona fide* epigenetic origin for 16 of them. We have thus demonstrated that heritable differences in DNA methylation also could contribute to quantitative resistance to clubroot. Six $QTL^{epi}$ co-localized with previously identified clubroot resistance genes and QTL in Arabidopsis, revealing that quantitative resistance to clubroot in natural accessions could be controlled by both nucleotidic and epigenetic variations.

## Materials and Methods

### Plant material

**Plant stocks** Mutant plants altered in genes encoding chromatin modifiers were ordered from the NASC. All plants were obtained from T-DNA insertion in the Columbia (Col-0) genetic background and showed a homozygous insertion in the plant genome. The mutant lines used were *drm2-2* (SALK_150863), *hda 15* (SALK_004027C), *atxr5* (SALK_130607C), *hac1* (SALK_082118C), *srt2* (SALK_149295C) and *ddm1* (SALK_000590C). T-DNA insertions and homozygosity were confirmed by PCR using the set of appropriate primers designed with the T-DNA Primer Design interface (http://signal.salk.edu/tdnaprimers.2.html, Supporting Information Table S1). The epigenetic recombinant inbred lines (epiRIL) population is that derived from a cross between the *Arabidopsis thaliana* ecotype Col-0 and a Col-0 line carrying the *ddm1-2* mutant allele (Johannes *et al.*, 2009). Note that the *ddm1-2* allele was obtained by EMS mutagenesis, not T-DNA transformation (Vongs *et al.*, 1993). EpiRIL population seeds were obtained from the Versailles Arabidopsis Stock Center (http://publiclines.versailles.inra.fr/) at the Institute Jean-Pierre Bourgin.

### Clubroot pathogen

All clubroot tests were performed with the *Plasmodiophora brassicae* eH isolate described by Fahling *et al.* (2003). Isolate eH belongs to the P1 pathotype according to the classification using the differential host set of Some *et al.* (1996). One millilitre of resting spore suspension ($10^7$ spores $ml^{-1}$) prepared according to Manzanares-Dauleux *et al.* (2000b) was used for pathogen inoculation 10 d after germination (stage 1.04; Boyes *et al.*, 2001). The inoculum was applied to the bottom of the stem base of each seedling.

### Growth conditions

The Arabidopsis accession Col-0 and the six mutant lines (Col-0 background) were evaluated in a randomized complete block design (with three blocks) against the eH *P. brassicae* isolate. For $QTL^{epi}$ detection, the 123 epiRIL that have been epigenotyped previously (Colomé-Tatché *et al.*, 2012), together with the two parental lines, were phenotyped in four biological replicates, split in two growth rooms, in a randomized complete block design (with two blocks per replicate and six plants per epigenotype per block). In growth room-2, four temperature sensors per block were placed at the height of plants. For each pathological test, seed germination was synchronized by placing seeds on wet blotting paper in Petri dishes for 2 d at 4°C. Seeds were sown individually in pots (4 cm diameter) containing a sterilized mix (by autoclaving) composed of 2/3 compost and 1/3 vermiculite. Per block, six plants per genotype (T-DNA mutants, Col-0, *ddm1-2* and epiRIL) were grown in controlled conditions of 16 h light (110 µmol $m^{-2}$ $s^{-1}$) at 21°C and 8 h dark at 18°C.

### Phenotyping

Phenotyping was performed 3 wk after inoculation (21 d post-inoculation (dpi)). Plants were thoroughly rinsed with water and photographed. Infected roots were removed and frozen in liquid nitrogen. For the epiRIL and their parents, four disease-related traits were assessed: longest leaf length (Lfi), disease index (DI), root biomass (Rbi) and pathogen DNA quantity (Pb), whereas for mutants, only DI was evaluated. DI was calculated according to Manzanares-Dauleux *et al.* (2000b) and pathogen DNA quantification was determined by quantitative polymerase chain reaction (qPCR). These phenotypic traits were chosen due to their relevance for the study of the quantitative response to eH isolate on Arabidopsis (Jubault *et al.*, 2008; Gravot *et al.*, 2011; Lemarié *et al.*, 2015a,b). The longest leaf length (Lfni) and the root biomass (Rbni) were also measured at 28 d post sowing on noninfected plants cultivated in growth room-2 in a randomized complete block design with 24 plants per (epi)genotype. Two new variables were then obtained by calculating the difference between these 'control' values with those obtained on infected plants; these new variables were termed ΔLf and ΔRb.

### Pathogen DNA quantification

Total genomic DNA (plant and pathogen) was extracted from lyophilized infected roots with the NucleoSpin Plant II kit according to manufacturer's instructions. After the evaluation of DNA concentration by Nanodrop ND-1000, 5–10 ng of total genomic DNA extract and specific primers were used to quantify by qPCR pathogen the DNA quantity in each epiRIL and the parental lines. The pathogen-specific primers designed on the 18S rRNA of *P. brassicae* used were PbF-K1 5′ TTGGG TAATTTGCGCGCCTG 3′; PbR-K1 5′ CAGCGGCAGGTCA TTCAACA 3′. qPCR reactions were carried out in a thermocycler (LightCycler 480 II Roche), with Syber green (LightCycler® 480 SYBR Green I Master). The qPCR conditions used were: 50 cycles of denaturation at 95°C for 15 s, annealing/extension at 63°C for 30 s and 72°C for 30 s. Absolute quantification of pathogen DNA was carried out using a standard curve obtained based on a series of known amounts of pathogen DNA. The average of the pathogen quantity present in each epiRIL (ratio of pathogen DNA quantity to total DNA used for qPCR) was used in further statistical analyses and QTL$^{epi}$ detection.

### Data analysis

Two generalized linear models (glm) were used to determine the effects of epigenome, temperature and epigenome × temperature interaction with R/GLM2 (Marschner, 2011) in R v.3.2.2 (R Development Core Team, 2015). For each model, the family distribution option of the glm function was adapted according to the data distribution. The first glm (glm1) described in Eqn 1 was used to estimate the epigenotype effect of each epiRIL across biological replicates in each growth room:

$$Y_{ijk} = \mu + G_i + R_j + B_k(R_j) + e_{ijk} \quad \text{Eqn 1}$$

where $\mu$, mean general effect; $G_i$, differential effect between epigenotypes; $R_j$, differential effect between replicates; $B_k$ $(R_j)$, interaction between blocks and replicates; and $e_{ij}$, residual variance. Based on this model, broad sense heritability was estimated using the following equation:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + (\frac{\sigma^2_e}{n})} \quad \text{Eqn 1.1}$$

($\sigma^2_G$, estimated epigenetic variance; $\sigma^2_e$, estimated environmental variance; and $n$, number of replicates per line).

The second generalized linear model (glm2) described in Eqn 2 was used to estimate the epigenotype × temperature effect in growth room-2 for each epiRIL across the biological replicates:

$$Y_{ijkl} = \mu + G_i + T_j + R_k + B_l(R_k) + GT_{ij} + TR_{jk} + TB(R)_{jlk} + e_{ijkl} \quad \text{Eqn 2}$$

where $\mu$, mean general effect; $G_i$, differential effect between epigenotypes; $T_j$, differential temperature conditions; $R_k$, differential effect between replicates; $B_l$ $(R_k)$, interaction between blocks and replicates; $GT_{ij}$, interaction between epigenotype and temperature condition; $TR_{jk}$, interaction between replicate and temperature condition; $TB(R)_{jlk}$, interaction between block and temperature condition; and $e_{ijkl}$, residual variance). Based on this model, broad sense heritability was estimated using the following equation:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{GT} + (\frac{\sigma^2_e}{n})} \quad \text{Eqn 2.1}$$

where $\sigma^2_G$, estimated epigenetic variance; $\sigma^2_{GT}$, estimated epigenetic × temperature variance; $\sigma^2_e$, estimated environmental variance; and $n$, number of replicates per line.

For all traits, least square means on each effect according to the two models described above were estimated with the function lsmeans of the R/LSMEANS package (Lenth, 2016) in R v.3.2.2 (R Core Team, 2015). The LSMEANS function also was used to extract the epigenotype effect $(G)$ and the interaction epigenotype × temperature $(GT)$ of each trait according to the generalized model used. Differences in longest length leaf $(\Delta Lf)$ and root biomass $(\Delta Rb)$ between infected and control plants were calculated from $G$.

### QTL$^{epi}$ detection

The G and GT of each trait were treated with the package R/QTL (Broman *et al.*, 2003) in R v.3.2.2 (R Development Core Team, 2015). The package SNOW in R (Tierney *et al.*, 2015) allowing the use of processor cores as cluster was used to reduce permutation calculation time. Simple Interval Mapping (SIM) was first carried out to identify potential QTL$^{epi}$ with the Haley–Knott (hk) method (Broman & Sen, 2009), using a step size of 2 cM and a window size of 10 cM. One thousand permutations with the hk method were carried out in order to determine SIM threshold levels for each condition and trait analysed. The significance level of threshold was fixed at $\alpha = 0.05$. In order to integrate the possibility of the presence of multiple QTL$^{epi}$ on the same chromosome, a manual multiple QTL mapping (MQM) approach (Broman & Sen, 2009) was used based on the results of SIM analysis. For this, the stepwise function was used in order to select the QTL$^{epi}$ (forward and backward system, option 'additive.only = FALSE') based on the preliminary putative QTL$^{epi}$ identified by SIM. For each trait, a minimum of two potential QTL$^{epi}$ was used in the stepwise function even if only one potential QTL$^{epi}$ was detected in SIM. Logarithm of odds (LOD) thresholds were calculated using 1000 permutations with the function scantwo with a significance level of $\alpha = 0.05$. Once the QTL$^{epi}$ selected, the model was fitted (fitqtl function), in order to calculate the LOD scores and the percentage of variation explained by each and all QTL$^{epi}$ $(R^2)$. The confidence interval of each QTL$^{epi}$ was calculated with the lodint function with a LOD drop of one parameter. The epiallele effect was evaluated with the function effectplot. Putative interactions among the QTL$^{epi}$ incorporated in the model were tested with the function addint according to Broman & Sen (2009).

### DNA sequence variation

Whole genome sequence data are available for the 123 epigenotyped epiRIL (Gilly *et al.*, 2014). The joint identification of small-scale variants (single-base substitutions and indels < 100 bp) was performed using GATK HaplotypeCaller on the whole-genome sequencing data available for 122 epiRIL. Raw variant calls were then filtered following GATK Best Practice suggestions and additional scripts. All variants were visually inspected using IGV and a subset of the detected single nucleotide polymorphism (SNP) was validated by PCR. For the analysis of transposition events (TE), TE-Tracker (Gilly *et al.*, 2014) was used to identify nonreference TE insertions in 102 epiRIL of sufficient genomic coverage. Raw calls were then filtered using in-house scripts. All of the detected new TE insertions were visually inspected using IGV and a fraction of them were validated by PCR. Shared small-scale and TE insertion variants are defined as present in at least 25% of the population. In order to evaluate the impact of shared sequence variants on the heritable variation observed in the epiRIL in response to clubroot, we compared various QTL models as described by Kooke *et al.* (2015). Three different models were tested: (1) using all QTL peak epigenetic markers (MM); (2) using each shared DNA sequence variant (SNP, indel, TE insertion) located in the confidence interval instead of the epigenetic marker at the peak; and (3) using both peak epigenetic markers and the shared DNA sequence variants included in the confidence interval. QTL detected using the three models were compared. If the DNA-based markers had a more significant effect than the peak QTL epigenetic markers, then the model fitted was considered to be better or identical to the model with the peak epigenetic markers.

## Results

### *ddm1* mutants have reduced susceptibility to *P. brassicae*

In order to determine if epigenetic variations could be associated with variations in response to clubroot infection, we assessed the response to infection of several Arabidopsis mutants affected in genes encoding chromatin modifiers. Infection of Col-0 and six T-DNA insertion mutants in genes encoding chromatin modifiers (*atxr5, ddm1, drm2, hac1, hdc15* and *srt2*) with the *P. brassicae* eH isolate led to a range of phenotypic responses. A statistically significant DI effect (ANOVA: $F = 4.99$, $P$-value = 0.005) (Table S2a) was identified 21 dpi suggesting that epigenetics was involved in plant response to infection by *P. brassicae*. A Dunnett's post hoc test (Table S2) revealed a significant difference between *ddm1* and Col-0 ($t = -4.437$, $P$-value = 0.003) for the DI trait, *ddm1* showing reduced symptoms compared to Col-0 (Fig. 1).

### Heritable differences in DNA methylation are associated with differential susceptibility to *P. brassicae*

**EpiRIL response to *P. brassicae* is quantitative** In order to identify epialleles involved in clubroot response, QTL$^{epi}$ detection was carried out on the subset of 123 lines of the epiRIL population used previously in Colomé-Tatché *et al.* (2012). In total, each genotype was assessed against *P. brassicae* isolate eH in four biological replicates, split in two growth rooms, each biological replicate being composed of two blocks. Distribution of the four disease-related traits assessed on the 123 epiRIL showed continuous distribution suggesting polygenic control of these traits (Fig. 2). However, significant differences were observed for all traits between the biological replicates (Mann–Whitney $U$-test) set up in the two growth rooms (DI: $W = 3685.5$, $P$-value < 2.2e-16; Lfi: $W = 233\,530$, $P$-value < 2.2e-16; Pb: $W = 4826$, $P$-value < 2.2e-16), suggesting an influence of the growth conditions on the epiRIL response to clubroot (Table S3). Indeed, higher levels of disease symptoms were observed on the Col-0 parent line (growth room-1: DI = 53.25 ± 2.36; growth room-2: DI = 90.75 ± 9.64) and on the epiRIL population (growth room-1: DI = 51.17 ± 11.42; growth room-2: DI = 86.27 ± 10.74) growing in growth room-2 compared to growth room-1. Consequently, we decided to analyse data from the two growth rooms independently. Analysis of biological replicates grown in each growth room showed a significant epigenotype effect (glm1, Eqn 1) for nearly all phenotypic traits measured ($P$-value ranged from 0.02 to < 2.2e-16 in growth room-1, and from 0.35 to < 2.2e-16 in growth room-2; Table S4). Broad-sense heritability ($H^2$) was estimated for each trait using Eqn 1.1 and ranged from 0.46 to 0.76 in the growth room-1 and from 0.44 to 0.65 in the growth room-2 depending on the trait studied (Table 1).

**QTL analysis of data obtained in growth room-1** Phenotypic data measured on the two biological replicates set up in growth room-1 were used in glm1 (Eqn 1) to extract the epigenotype effect (G) with the lsmean function. The epigenotype effect G identified for each trait was then used for the QTL$^{epi}$ analysis. In total, five QTL$^{epi}$ were detected (Fig. 3). Two QTL$^{epi}$ were identified for Pb on chromosomes 1 and 4 (Pb1$^{epi}$-At1, Pb1$^{epi}$-At4) explaining 14.64% and 9.65% of the phenotypic variability, respectively. Three QTL$^{epi}$ were detected for Lfi on chromosomes 1, 3 and 5 (Lfi1$^{epi}$-At1, Lfi1$^{epi}$-At3, Lfi1$^{epi}$-At5) explaining 15.49%, 15.01% and 8.11% of phenotypic variation, respectively (Table 2). The variance explained by each fitted QTL model was of 43.82% and 19.59% for Pb and Lfi, respectively. Surprisingly, no QTL$^{epi}$ was identified for DI despite a significant effect of epigenotype on this trait. Confidence intervals of the QTL$^{epi}$ detected ranged from 6.23 to 36 cM (Table 2). No epistatic interaction was found between QTL$^{epi}$ for either trait. For the three QTL$^{epi}$ detected for Lfi (markers nearest of the peak LOD score: MM52, MM427 and MM728), wild-type (WT) epialleles were associated with an increase in the trait values. The *ddm1*-derived epiallele was associated with an increase in pathogen quantity at QTL$^{epi}$ on chromosome 1 (peak marker: MM123) whereas it was associated with a decrease in the Pb value on chromosome 4 (peak marker MM550) (Fig. S1; Table 2).

**QTL analysis of data from growth room-2** As above, phenotypic data measured on the two biological replicates in growth


**Fig. 1** Clubroot symptoms shown by Col-0 wild-type and the *ddm1* mutant (Col-0 background). Arabidopsis Col-0 and mutants were inoculated 10 d after germination and phenotyped 3 wk post-inoculation. (a) Col-0 ecotype; (b) *ddm1* mutant. The *ddm1* mutant showed a significant decrease in disease symptoms (Dunett's test, $P < 0.02$) compared to Col-0. Plant individuals are representative of standard observations made in our experimental conditions. Bars, 1 cm.

room-2 were used in glm1 to extract the epigenotype effect with the lsmean function. Once again, the epigenotype effect G identified for each trait was then used for the QTL$^{epi}$ analysis. In total, seven QTL$^{epi}$ were detected (Fig. 3). One QTL$^{epi}$ was detected on chromosome 4 for Pb (Pb2$^{epi}$-At4) and two QTL$^{epi}$ on chromosomes 1 and 3 were detected for Lfi (Lfi2$^{epi}$-At1, Lfi2$^{epi}$-At3). QTL mapping also revealed four QTL$^{epi}$ controlling Rbi, which was measured only in this growth room: one on chromosome 1, two on chromosome 2 and one on chromosome 4 (Rbi$^{epi}$-At1, Rbi$^{epi}$-At2a, Rbi$^{epi}$-At2b, Rbi$^{epi}$-At4). Again, no QTL$^{epi}$ was identified for DI despite a significant effect of the epigenotype. The variance explained by each fitted QTL model was of 10.86-%, 26.16% and 44.59% for Pb, Lfi and Rbi, respectively. The QTL$^{epi}$ identified for Pb on chromosome 4 explained 10.86% of the variability. The two QTL$^{epi}$ detected on chromosomes 1 and 3 for Lfi explained (respectively) 15.65% and 12.06% of the phenotypic variation (Table 2). Confidence intervals of the detected QTL$^{epi}$ ranged from 7.62 to 54.33 cM (Table 2). No epistatic interaction was found between QTL$^{epi}$ for Lfi and Rbi traits. The additive allele effect of the two QTL$^{epi}$ detected for Lfi (peak markers: MM91 and MM515) and of three of the four QTL$^{epi}$ detected for Rbi (peak markers: MM147, MM383, MM686) was in the same direction: WT-derived epialleles were associated with an increase in trait value. For the Pb QTL (peak marker: MM693) and one of the four QTL$^{epi}$ of Rbi (peak marker: MM385), the *ddm1*-derived epiallele was associated with a decrease in trait value (Fig. S1; Table 2).

For Lfi, QTL$^{epi}$ detected on chromosomes 1 and 3 in growth room-1 (Lfi1$^{epi}$-At1, Lfi1$^{epi}$-At3) co-localized with the QTL$^{epi}$ detected on chromosomes 1 and 3 in growth room-2 (Lfi2$^{epi}$-At1, Lfi2$^{epi}$-At3). The QTL$^{epi}$ on chromosome 5 was detected only in growth room-1. The Pb QTL$^{epi}$ detected in the two growth rooms were different. These results indicated that the QTL detection was possibly dependent on the growth room conditions (Fig. 3; Table 2).

As *ddm1* mutants displayed smaller roots and leaf lengths than WT in control conditions (i.e. noninoculated; Kakutani *et al.*, 1996; Cortijo *et al.*, 2014; Table S3), a similar experiment was carried out also in growth room-2 without clubroot infection to determine the impact of developmental alteration due to the mutation on clubroot symptoms. The Lfni and Rbni data assessed on the 123 epiRIL in control condition showed continuous distribution, suggesting polygenic control of these traits (Fig. 2). Data analysis showed a significant epigenotype effect (glm1) on both phenotypic traits (Lfni: $\chi^2 = 15\,771.8$, $P$-value < 2.2e-16; Rbni: $\chi^2 = 33.09$, $P$-value = 3.48e-10). In total, five QTL$^{epi}$ were detected (Fig. 3): three QTL$^{epi}$ were identified for Lfni on chromosomes 1, 3 and 5 (Lfni$^{epi}$-At1, Lfni$^{epi}$-At3 and Lfni$^{epi}$-At5) explaining 19.66%, 12.93% and 10.49% of the phenotypic variability, respectively. Two QTL$^{epi}$ were detected for Rbni on chromosomes 1 and 4 (Rbni$^{epi}$-At1, Rbni$^{epi}$-At4) explaining 10.44% and 11.51% of phenotypic variation, respectively (Table 2). The additive allele effect of all the QTL$^{epi}$ detected for Lfni

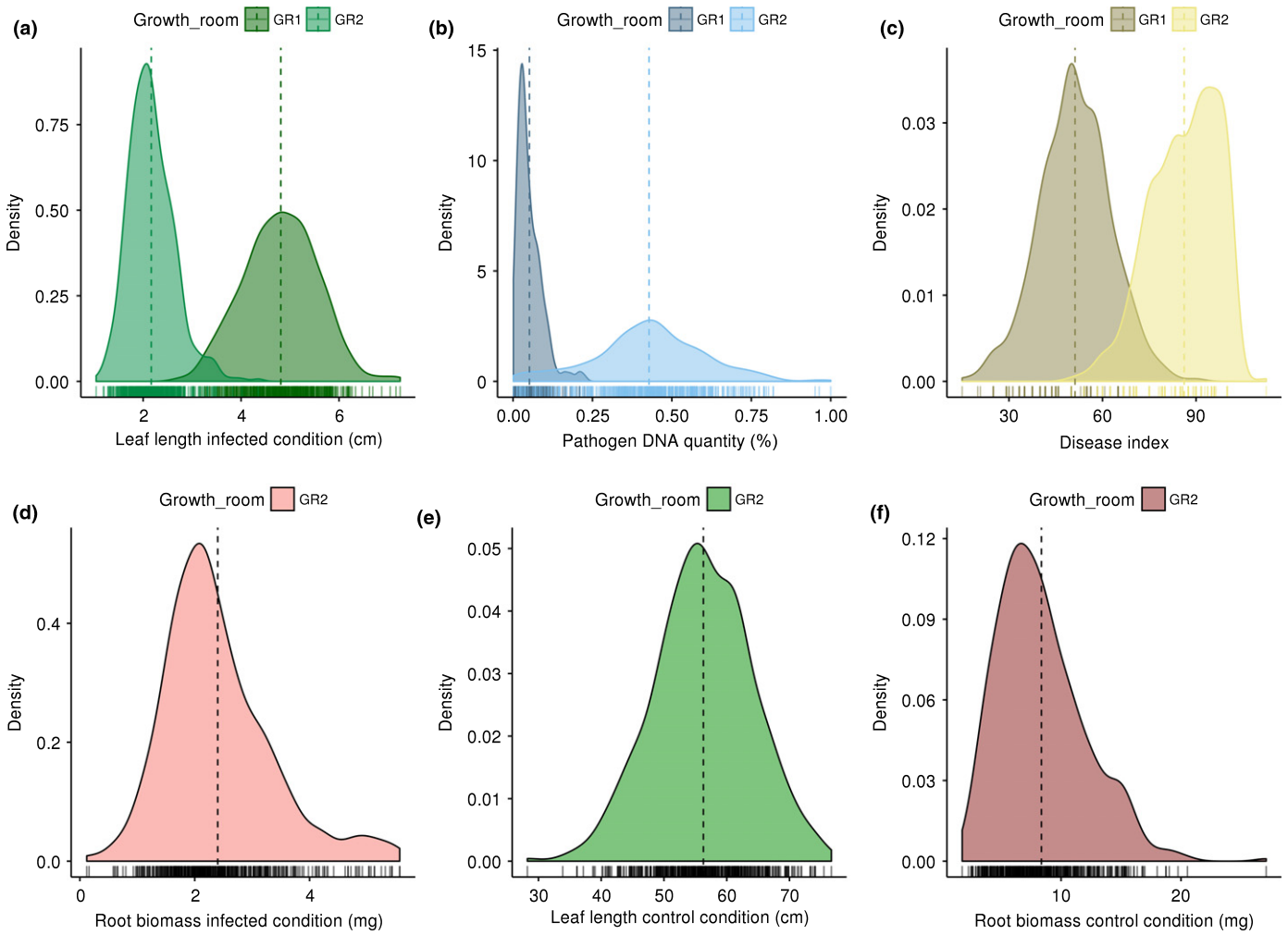


**Fig. 2** Distribution of phenotypic data measured in the two growth rooms used for pathological tests. For each trait studied, data are coloured to show in which growth chamber the Arabidopsis plants were grown. Vertical coloured dashed lines indicate means of traits in each growth room. Data distribution is shown for leaf length (a), pathogen DNA quantity (b), disease index (c) and root biomass (d) in infected condition, and for leaf length (e) and root biomass (f) in noninfected condition. GR1 corresponds to growth room-1. GR2 corresponds to growth room-2.

(peak markers: MM126, MM515 and MM713) and Rbni (peak markers: MM126 and MM691) was in the same direction: WT-derived epialleles were associated with an increase in trait value (Fig. S1). Three QTL$^{epi}$ were detected with ΔLf on chromosomes 1, 3 and 5 (ΔLf$^{epi}$-At1, ΔLf$^{epi}$-At3 and ΔLf$^{epi}$-At5) explaining 10.44%, 10.11% and 8.38% of the phenotypic variability, respectively (Fig. 3). No QTL$^{epi}$ was identified for ΔRb. The additive allele effect of all the QTL$^{epi}$ detected for ΔLf (peak markers: MM10, MM515 and MM854) was also in the same direction but in this case the *ddm1-2* epialleles were associated with the decrease of the value (Fig. S1).

## Temperature affects the plant response to *P. brassicae* in the epiRIL population

In order to explain the differences observed between the two growth rooms, we paid specific attention to the temperature conditions, as this parameter was shown to be critical for the development of clubroot disease (Siemens *et al.*, 2002; Sharma *et al.*, 2011). Although similar values of global mean temperature were noted in both growth rooms (growth room-1 = 19.95°C, growth room-2 = 20.06°C), significant differences ($F$ = 2.67, $P$-value = 0.002) in global temperature variances were registered between the two rooms (Fig. S2a). We then analysed mean and variance temperature values for each photoperiod (day and night). Day and night mean temperatures were similar in both chambers (growth room-1: day temperature = 20.98°C, night temperature = 17.83°C; growth room-2: day temperature = 20.85°C, night temperature = 18.51°C) and were very close to the required values (day temperature = 21°C, night temperature = 18°C). Conversely, the temperature variance in the two growth rooms and for both periods was significantly different (Fisher permutation test; day period: $F$ = 9.05, $P$-value = 0.002; night period: $F$ = 3.67, $P$-value = 0.002). The temperature range in growth room-1 was 10.83°C for the day period and 4.98°C for the night period; the temperature range in growth room-2 was 3.51°C for the day period and 1.88°C for the night period (Fig. S2b,c).

**Effect of temperature variation on plant response to clubroot** Based on these observations, a more detailed analysis was carried out in growth room-2 to evaluate the impact of the temperature variability on epiRIL response to clubroot. For this, we used data from temperature sensors placed at the height of the plants in growth room-2. Significant differences (Kruskal–Wallis test) in the median temperatures were observed ($\chi^2 = 4115.3$, df = 15, *P*-value < 2.2e-16) according to the location of the sensors. Pairwise analysis (Dunnett's test) of the differences between temperatures measured by the sensors showed significant temperature variations at several positions in the growth room (Table S5) and a temperature gradient from 21.4 to 24.1°C (Table S6).

A significant temperature effect (glm2) was shown for two of the phenotypic traits measured (*P*-value ranged from 0.54 to < 2.2e-16; Table S4). Moreover, a significant interaction between temperature and epigenotype effects (GT interaction) on the phenotypic traits measured was identified for nearly all traits (*P*-value ranged from 0.44 to < 2.2e-16; Table S4). Heritability ($H^2$) in the case of temperature and epigenotype interaction was estimated for each trait using the formula (2.1). Heritability ranged from 0.50 to 0.67 according to the dataset used (Table 1).

**Table 1** Heritability estimates for each trait in the Arabidopsis epigenetic recombinant inbred lines population.

| Trait | Heritability model | $H^2$ Growth room-1 | $H^2$ Growth room-2 |
|---|---|---|---|
| DI | glm1 | 0.55 | 0.50 |
| Pb | glm1 | 0.46 | 0.44 |
| Lfi | glm1 | 0.76 | 0.65 |
| Rbi | glm1 | NA | 0.62 |
| DI | glm2 | NA | 0.50 |
| Pb | glm2 | NA | 0.50 |
| Lfi | glm2 | NA | 0.67 |
| Rbi | glm2 | NA | 0.50 |

$H^2$ is the broad sense heritability calculated with the formula 1: $H^2 = \sigma^2_G/[\sigma^2_G + (\sigma^2_e/n)]$ and 2: $H^2 = \sigma^2_G/[\sigma^2_G + \sigma^2_{GT} + (\sigma^2_e/n)]$ including the variance of the temperature × epigenotype interaction. $\sigma^2_G$, the estimated epigenetic variance; $\sigma^2_{GT}$, the estimated temperature × epigenetic interaction variance; $\sigma^2_e$, the estimated environmental variance; *n*, the number of replicates per line. DI, disease index; Pb, pathogen DNA quantity; Lfi, leaf length in infected condition; Rbi, root biomass in infected condition. NA, not available.

**Detection of epigenotype × temperature QTL$^{epi}$** In order to determine whether temperature variation could influence plant response to clubroot in the epiRIL population, we associated each epiRIL with the mean temperature data measured with the temperature sensor placed in the block where the epiRIL was grown. These data were used in glm2 (Eqn 2) to extract the GT interaction using the lsmean function. The GT interaction values identified for each trait were then used for the QTL$^{epi}$ analysis. Two QTL$^{epi}$GT (epigenotype × temperature QTL$^{epi}$) were detected on the chromosomes 1 and 3 for Lfi (Lfi$^{epi}$GT-At1, Lfi$^{epi}$GT-At3), and one QTL$^{epi}$GT on chromosome 3 was detected for Rbi (Rbi$^{epi}$GT-At3). No QTL$^{epi}$ was detected for Pb at a 5% significance level using the SIM by stepwise approach but two QTL$^{epi}$ were detected on the chromosome 4 using a 10% significance level (Pb$^{epi}$GT-At4a, Pb$^{epi}$GT-At4b). Again, no QTL$^{epi}$ could be detected for DI (Fig. 3). The variance explained by the fitted QTL model was 19.34% for Lfi, 9.06% for Rbi and 15.89% for Pb. The QTL$^{epi}$ found for Lfi on chromosomes 1 and 3 explained 10.67% and 7.81% of the variability. The QTL$^{epi}$ found for Rbi explained 9.06% of phenotypic variation. The two

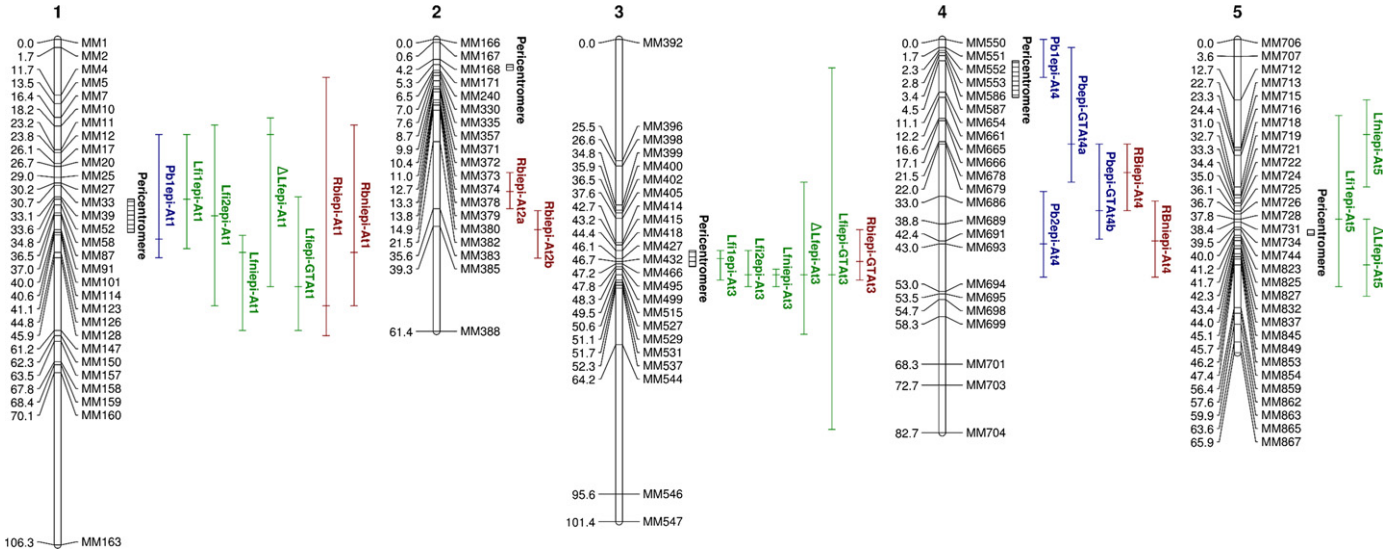


**Fig. 3** Epigenetic quantitative trait loci (QTL) identified in the epigenetic recombinant inbred lines (epiRIL) population in response to *Plasmodiophora brassicae* infection. Numbers above chromosomes indicate *Arabidopsis* chromosomes. MM indicates markers. For each marker, position in cM is indicated. The vertical bar length is equal to the one- logarithm of odds (LOD) likelihood confidence interval. The dash in the bar indicates the peak position. QTL name indicates the trait associated and chromosomal localization.

**Table 2** Summary of epigenetic quantitative trait loci (QTL$^{epi}$) detected in the Arabidopsis epigenetic recombinant inbred lines (epiRIL) population by multiple QTL mapping (MQM).

| Trait | Condition | Test location | QTL | Chr | Pos | Lod | Closest peak marker | Confidence interval (cM) | | $R^2$ (%) | $R^2$ for all QTL by trait | Favourable allele |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G-Rbi | Infected | Growth room-2 | Rbi$^{epi}$-At1 | 1 | 56.00 | 3.04 | MM147 | 8.00 | 62.33 | 6.67 | 44.59 | WT |
| | | Growth room-2 | Rbi$^{epi}$-At2a | 2 | 32.00 | 7.48 | MM383 | 28.00 | 35.62 | 17.91 | | WT |
| | | Growth room-2 | Rbi$^{epi}$-At2b | 2 | 40.00 | 6.18 | MM385 | 36.00 | 46.00 | 14.42 | | *ddm1-2* |
| | | Growth room-2 | Rbi$^{epi}$-At4 | 4 | 28.00 | 4.79 | MM686 | 22.02 | 36.00 | 10.88 | | WT |
| G-Lfi | Infected | Growth room-1 | Lfi1$^{epi}$-At1 | 1 | 33.60 | 6.51 | MM 52 | 20.00 | 44.00 | 15.50 | 43.82 | WT |
| | | Growth room-1 | Lfi1$^{epi}$-At3 | 3 | 46.10 | 6.32 | MM427 | 44.36 | 50.59 | 15.01 | | WT |
| | | Growth room-1 | Lfi1$^{epi}$-At5 | 5 | 37.80 | 3.60 | MM728 | 16.00 | 52.00 | 8.11 | | WT |
| | | Growth room-2 | Lfi2$^{epi}$-At1 | 1 | 37.05 | 5.14 | MM91 | 18.00 | 56.00 | 15.65 | 26.16 | WT |
| | | Growth room-2 | Lfi2$^{epi}$-At3 | 3 | 49.45 | 4.04 | MM515 | 44.37 | 52.00 | 12.06 | | WT |
| G-Pb | Infected | Growth room-1 | Pb1$^{epi}$-At1 | 1 | 42.00 | 4.36 | MM123 | 20.00 | 45.93 | 14.64 | 19.59 | WT |
| | | Growth room-1 | Pb1$^{epi}$-At4 | 4 | 0.00 | 2.95 | MM550 | 0.00 | 8.00 | 9.65 | | *ddm1-2* |
| | | Growth room-2 | Pb2$^{epi}$-At4 | 4 | 42.99 | 3.05 | MM693 | 32.00 | 50.00 | 10.86 | 10.86 | *ddm1-2* |
| GT-Lfi | Infected | Growth room-2 | Lfi$^{epi}$GT-At1 | 1 | 52.00 | 3.32 | MM128 | 33.07 | 61.20 | 10.67 | 19.34 | WT |
| | | Growth room-2 | Lfi$^{epi}$GT-At3 | 3 | 49.45 | 2.47 | MM547 | 6.00 | 82.00 | 7.81 | | WT |
| GT-Rbi | Infected | Growth room-2 | Rbi$^{epi}$GT-At3 | 3 | 46.65 | 2.52 | MM432 | 40.00 | 50.59 | 9.06 | 9.06 | WT |
| GT-Pb | Infected | Growth room-2 | Pb$^{epi}$GT-At4a | 4 | 22.00 | 2.38 | MM679 | 1.73 | 30.00 | 8.04 | 15.89 | *ddm1-2* |
| | | Growth room-2 | Pb$^{epi}$GT-At4b | 4 | 36.00 | 4.50 | MM689 | 22.02 | 42.00 | 15.88 | | *ddm1-2* |
| G-Lfni | Noninfected | Growth room-2 | Lfni$^{epi}$-At1 | 1 | 44.80 | 8.32 | MM126 | 41.15 | 61.20 | 19.66 | 46.20 | WT |
| | | Growth room-2 | Lfni$^{epi}$-At3 | 3 | 49.46 | 5.75 | MM515 | 48.32 | 52.00 | 12.93 | | WT |
| | | Growth room-2 | Lfni$^{epi}$-At5 | 5 | 20.00 | 4.76 | MM713 | 12.73 | 31.00 | 10.49 | | WT |
| G-Rbni | Noninfected | Growth room-2 | Rbni$^{epi}$-At1 | 1 | 44.80 | 3.39 | MM126 | 18.00 | 56.00 | 10.44 | 22.93 | WT |
| | | Growth room-2 | Rbni$^{epi}$-At4 | 4 | 42.43 | 3.72 | MM691 | 34.00 | 50.00 | 11.51 | | WT |
| G-ΔLf | Noninfected – Infected | Growth room-2 | ΔLf$^{epi}$-At1 | 1 | 20.00 | 3.79 | MM10 | 16.45 | 52.00 | 10.44 | 31.55 | *ddm1-2* |
| | | Growth room-2 | ΔLf$^{epi}$-At3 | 3 | 49.46 | 3.68 | MM515 | 30.00 | 62.00 | 10.11 | | *ddm1-2* |
| | | Growth room-2 | ΔLf$^{epi}$-At5 | 5 | 47.36 | 3.08 | MM854 | 37.80 | 54.00 | 8.38 | | *ddm1-2* |

Confidence intervals are in cM. Peak positions are indicated in cM and with the marker nearest to the logarithm of odds (LOD) score peak. $R^2$, phenotypic variation explained by the QTL$^{epi}$. Chr, chromosome. For Lfi and Rbi, favourable alleles are associated with an increase in the value. For Pb, favourable alleles are associated with a decrease in the value. G-Rbi, G-Lfi and G-Pb represent epigenetic QTL for each trait in infected condition. G-Lfni and G-Rbni represent epigenetic QTL for each trait in control condition. G-ΔLf represents epigenetic QTL obtained by difference between leaf length in infected condition and leaf length in control condition. GT-Rbi, GT-Lfi and GT-Pb represent epigenetic × temperature QTL for each trait.WT, wild-type; *ddm1-2*, mutant allele.DI, disease index; Pb, pathogen DNA quantity; Lfi, leaf length in infected condition; Rbi, root biomass in infected condition; Lfni and Rbni, respectively, leaf length and root biomass in control (noninfected) condition; ΔLf, change in leaf length.

QTL$^{epi}$ found for Pb on chromosome 4 explained 8.04% and 15.88% of the variation (Table 2). Confidence intervals of the detected QTL$^{epi}$ ranged from 10.59 to 76 cM (Table 2). No epistatic interaction was found between QTL$^{epi}$ for Lfi and Pb traits. For all QTL$^{epi}$ detected, the WT epialleles were associated with an increase in the values (Fig. S1; Table 2). Comparison of the QTL$^{epi}$ and QTL$^{epi}$GT (taking into account the interaction with temperature) showed that all QTL$^{epi}$GT co-localized totally or partially with QTL$^{epi}$ detected in at least one growth room (Fig. 3). Indeed, the comparison of the confidence intervals of QTL$^{epi}$ detected using the leaf length trait showed that the QTL$^{epi}$ Lfi1$^{epi}$-At1 and Lfi2$^{epi}$-At1 overlapped with the QTL$^{epi}$GT Lfi$^{epi}$GT-At1. Likewise, Lfi1$^{epi}$-At3 and Lfi2$^{epi}$-At3 confidence intervals co-localized with the confidence interval of Lfi$^{epi}$GT-At3 (Fig. 3; Table 2). For the QTL$^{epi}$ detected using the quantification of the pathogen DNA (Pb) two overlaps were identified on chromosome 4, one between Pb1$^{epi}$-At4 and Pb$^{epi}$GT-At4a in the region extending from the short arm to the pericentromeric region and another between Pb2$^{epi}$-At4, Pb$^{epi}$GT-At4a and Pb$^{epi}$GT-At4b on the long arm of chromosome 4 (Fig. 3; Table 2).

## Impact of DNA sequence variation within QTL$^{epi}$ confidence intervals

Although the epiRIL population was designed to minimize as much as possible DNA sequence variation, the presence of a small number of segregating nucleotidic variants cannot be avoided, notably as a result of the known effect of loss of DNA methylation on TE activity. In order to investigate the potential contribution of parental DNA sequence variants to the differential susceptibility to *P. brassicae* associated with the 20 QTL$^{epi}$, we identified using whole genome sequencing data all sequence variants shared by more 25% of the epiRIL. In total, 63 small-scale and 11 TE insertion variants were located within the 20 QTL$^{epi}$, respectively. Eleven shared insertions of TE were detected in the QTL$^{epi}$ confidence intervals, with 18 QTL$^{epi}$ showing at least one insertion and two QTL$^{epi}$ showing no insertion. All QTL$^{epi}$ detected included at least one small-scale sequence polymorphism in their confidence interval (Dataset S1).

Effects of the shared TEs and sequence variants were tested as described in Kooke *et al.* (2015). For 16 of 20 QTL$^{epi}$, the effect of the epigenetic marker at the QTL peak was more significant

than the effect of the TE or the small-scale sequence variant (Dataset S2). However, for Lfi$^{epi}$GT-At1, Lfi$^{epi}$GT-At3, Lfi1$^{epi}$-At5 and ΔLf$^{epi}$-At3, the significant effect observed for a SNP or a TE was greater than for the epigenetic marker at the QTL peak (Dataset S2). In this case, we considered that these four QTL$^{epi}$ were actually caused by DNA sequence variants. However, a linkage disequilibrium between the significant genetic markers and a causal epiallele could also be possible.

## Discussion

The aim of the present study was to investigate the role of epigenetic modifications in the Arabidopsis response to *Plasmodiophora brassicae*. To determine whether epigenetic regulation does play a role in clubroot quantitative resistance, a reverse genetics approach, using six T-DNA insertional mutants (Col-0 background) in genes involved in epigenetic pathways, was first carried out. For the six mutants tested, the disease index (DI) was only significantly reduced in the *ddm1* mutant compared to Col-0. This finding suggests that *ddm1* confers decreased susceptibility of Arabidopsis to *P. brassicae*. This observation is in agreement with previous results obtained by Kellenberger *et al.* (2016) and Sharma *et al.* (2017) in the *Brassica* genus, which linked plant responses to biotic and abiotic stress to epigenetic regulations. Moreover, the involvement of *ddm1* in the Arabidopsis–*P. brassicae* interaction further supports the hypothesis that DNA methylation plays a role in plant pathogen infections (Dowen *et al.*, 2012; López Sánchez *et al.*, 2016; Hewezi *et al.*, 2017). To decipher the epigenetic architecture of the clubroot resistance in Arabidopsis, we then carried out a epigenetic QTL (QTL$^{epi}$) detection experiment using the epigenetic recombinant inbred lines (epiRIL) population. Four disease-related traits (DI, pathogen DNA quantity (Pb), leaf length (Lfi) and root biomass (Rbi)), mostly used previously in Arabidopsis (Jubault *et al.*, 2008; Gravot *et al.*, 2011) to evaluate plant response to clubroot, were monitored. These traits were chosen in order to characterize disease development (disease index and pathogen DNA quantity) and the consequences of *P. brassicae* infection on plant development (Rbi and Lfi).

### Epigenetic QTL control Arabidopsis clubroot resistance

From our experiments using four biological replicates, we have shown a significant epigenetic effect on the response to *P. brassicae* infection. Thus, heritable plant responses induced by DNA methylation appear to be involved in the plant response to *P. brassicae*. The moderate to high heritability values (from 0.33 to 0.76) observed were similar to those described in control and abiotic stress conditions for this population (Johannes *et al.*, 2009; Colomé-Tatché *et al.*, 2012; Kooke *et al.*, 2015). Moreover, these heritability values also were similar to those described by Jubault *et al.* (2008) on a RIL population infected with *P. brassicae*. The moderate heritability levels calculated for the traits mean that this epigenetic variability could be considered for use in breeding. As the plant response to *P. brassicae* varied depending on the growth room used, we carried out the QTL$^{epi}$ detection experiments using data from each growth room. In total, sixteen additive QTL$^{epi}$ grouped in six genomic regions were identified distributed throughout four *A. thaliana* chromosomes. Among the 16 QTL$^{epi}$, five QTL$^{epi}$ were involved in pathogen multiplication (Pb) and 11 were involved in plant (foliar and root) development in response to *P. brassicae* infection. The identification of three QTL$^{epi}$ for differences in longest length leaf (ΔLf) highlighted the modulation by *P. brassicae* infection of the foliar development variation already present in the epiRIL population (illustrated in Fig. S3). Clubroot resistance response is therefore composed of factors reducing the impact of the clubroot infection on foliar development as well as factors reducing pathogen development. For four of the 20 QTL$^{epi}$ initially detected (Lfi$^{epi}$GT-At1, Lfi$^{epi}$GT-At3, Lfi$^{epi}$-At5 and ΔLf$^{epi}$-At3), analysis of shared transposition events (TE) and sequence variants included in their confidence intervals showed that the effect of DNA-based markers was greater than the effect of epigenetic markers, suggesting that these QTL are not *bona fide* epigenetic. Among the four phenotypic traits evaluated, no QTL$^{epi}$ was detected for DI despite a significant epigenotype effect in the two growth rooms ($P$ = 1.57e-06 and 1.65e-04 for growth room-1 and -2, respectively). Results of the reverse genetics approach showed that *ddm1* was less susceptible to *P. brassicae* compared to Col-0 suggesting that *ddm1* epialleles confer a reduction in symptoms. In this context, the absence of QTL$^{epi}$ detected for the DI trait may be explained by a sampling effect, because only a subset (123 of 505 lines) of the epiRIL population was phenotyped in this study, and/or a low proportion of *ddm1* epi-haplotypes in the population subset (27%), which may have been insufficient to detect small QTL effects (Holland, 2007). Most of the QTL$^{epi}$ detected in this study co-localized with the pericentromeric regions. This may be explained by a more stable loss of methylation in those regions due to the loss of methylation maintenance on repeats and TE in *ddm1* (Kooke *et al.*, 2015). Concerning the effect of the epialleles, the majority of the wild-type (WT) epialleles were associated with an increase in the morphological trait values (length of leaves and root biomass). However, the *ddm1*-derived epiallele led to a decrease in the amount of pathogen DNA for two of the three QTL$^{epi}$ detected on chromosomes 1 and 4. This finding highlighted the positive effect of this epiallele on plant resistance, in agreement with the results obtained in the mutant test of this study and by Dowen *et al.* (2012) who showed a modest increase in *ddm1* resistance to *P. syringae*.

### Stability and pleiotropy of clubroot epigenetic QTL

Several QTL detected were stable across growth rooms. Indeed, despite the temperature variations between the two growth rooms, two overlapping leaf length QTL$^{epi}$ (growth room-1: Lfi1$^{epi}$-At1 and Lfi1$^{epi}$-At3; growth room-2: Lfi2$^{epi}$-At1 and Lfi2$^{epi}$-At3) were detected in each growth room. Furthermore, for Pb, the QTL$^{epi}$ Pb1$^{epi}$-At4, detected in the growth room-1 overlapped with the QTL$^{epi}$ Pb$^{epi}$GT-At4a detected in the growth room-2 (Fig. 3; Table 2). The co-localization on chromosome 1 of four QTL$^{epi}$ controlling three different traits (Lfi, Pb and Rbi) and on chromosome 4 of three QTL$^{epi}$ controlling two traits (Pb and Rbi) may

suggest the presence of pleiotropic QTL$^{epi}$ (Fig. 3; Table 2). However, analysis of the correlation (Spearman rho correlation) between traits showed only a moderate correlation ($\rho = 0.52$, $P < 2.2e\text{-}16$) between Lfi and Rbi. Fine mapping is necessary to overcome the bias imposed by the large size of the QTL$^{epi}$ confidence intervals and make further conclusions on the possibility of a pleiotropic gene or an effect due to linked genes.

### Temperature modulates Arabidopsis clubroot responses

Taking into consideration the temperature variations in growth room-2, we tested the hypothesis that temperature plays a potential role in epiRIL and Col-0 clubroot symptom variations. These observations are in agreement with the variations in clubroot severity observed on *Brassica rapa* subsp. *chinensis* and *B. napus* according to the temperature used for growing the plants (Sharma *et al.*, 2011; Gossen *et al.*, 2014). Siemens *et al.* (2002) had already evoked a possible link between temperature, clubroot response and epigenetics when studying the Arabidopsis mutant *tu8* (mutant in the *LIKE HETEROCHROMATIN PROTEIN 1 LHP1*) which presented different levels of response to clubroot depending on the temperature conditions. Similar environmental effects on the modulation of QTL controlling clubroot response also have been shown with nitrogen supply variations in *B. napus* (Laperche *et al.*, 2017; Aigu *et al.*, 2018) and with flooding in Arabidopsis (Gravot *et al.*, 2016). Here, our analyses suggested that the temperature effect was partly triggered by the interaction with the plant epigenome for the traits Rbi and Pb. The identification of temperature-dependent QTL$^{epi}$ controlling pathogen quantity (Pb$^{epi}$GT-At4a and Pb$^{epi}$GT-At4b) suggests the presence of QTL$^{epi}$ involved in the control of the pathogen development under temperature dependence. Similar observation concerning the detection of QTL temperature dependance was carried out by Aoun *et al.* (2017) using the model Arabidopsis–*Ralstonia solanacearum*. In their study, Aoun *et al.* (2017) showed that an increase of the temperature of 3°C during the interaction between Arabidopsis and *R. solanacearum* led to an increase of the sensitivity of most accessions and the loss of detection of one major QTL associated with the resistance. The influence of the temperature on the epiRIL response to *P. brassicae* could be explained by the increase of the environmental sensitivity triggered by the DNA hypomethylation suggested by Kooke *et al.* (2015).

### Clubroot resistance, a sophisticated system of regulation involving genetics and epigenetics

Interestingly, the comparison of the clubroot genetic QTL identified previously (Jubault *et al.*, 2008) in Arabidopsis and the clubroot epigenetic QTL identified in this study highlighted overlapping of some confidence intervals. Indeed, the confidence intervals of six QTL$^{epi}$ (Rbi$^{epi}$-At1, ΔLf$^{epi}$- At1and Pb1$^{epi}$-At1, Lfi1$^{epi}$-At1, Lfi2$^{epi}$-At1, ΔLf$^{epi}$-At1) overlapped with two QTL (Pb-At1 and Pb-At4) found in the Bur-0 × Col-0 RIL population by linkage analysis (Jubault *et al.*, 2008) and the major gene *RPB1* described by Arbeiter *et al.* (2002), respectively. These co-localizations suggest that quantitative resistance to clubroot is modulated by a system involving both nucleotidic and epigenetic variations. These results illustrate the fact that in classical populations used for QTL detection, dissociation between causal genetic and epigenetic variations, both in linkage disequilibrium with markers, is nearly impossible (Schmitz *et al.*, 2013a,b). In addition, our findings are consistent with the suggestion that during plant–pathogen interactions, plant transgenerational changes in genome structure and in DNA methylation patterns are possible (Boyko & Kovalchuk, 2011). The identification of two QTL$^{epi}$ (Rbi$^{epi}$-At2 and Pb1$^{epi}$-At4) that did not show any overlap with previously reported QTL for these traits suggests that at these loci only epigenetic variation may be responsible for plant response variation. However, an absence of co-localization could also be due to the absence of nucleotidic and/or epigenetic variations at these loci in the previously studied populations but does not exclude their existence in other genotypes.

This first study on the role of epialleles in the Arabidopsis–*P. brassicae* interaction brings to light the possibility of a complex model of quantitative resistance where alleles and epialleles act in concert. Furthermore, our study has shown that the temperature variations could influence epiRIL response to *P. brassicae*. In order to confirm whether epialleles are involved in plant response to clubroot infection, the QTL$^{epi}$ confidence intervals must be reduced to find causal epialleles, notably through a fine-mapping approach. The assessment of epiRIL in pathological tests in contrasting controlled temperature conditions is also needed to validate the possible role of temperature in modulating the epigenetic plant response to clubroot infection.

## Acknowledgements

We acknowledge our IGEPP Colleagues for their technical support and the Biological Resource Centers BrACySol and Versailles Arabidopsis Stock Center for providing Brassica and epiRIL seeds, respectively. This research was supported by AGROCAMPUS OUEST, INRA and Université de Rennes. Benjamin Liégard was a PhD student co-funded by INRA BAP department and Brittany Region.

## Author contributions

BL, EJ, AE, MJ, AG, JL and CL carried out the experiments and collected the data; BL, JL and CL carried out the molecular biology work; BL and VB carried out the genetic analyses; BL and AE carried out the bioinformatics analyses; VB, ME and VC provided DNA sequence information about the epiRIL used in this study; BL, VB, MJM-D and MJ wrote the article, assisted by AG and VC for manuscript editing; and BL, MJM-D and MJ designed and coordinated the study.

## ORCID

Victoire Baillet http://orcid.org/0000-0002-1081-2921
Vincent Colot http://orcid.org/0000-0002-6382-1610
Antoine Gravot http://orcid.org/0000-0001-9125-1494
Mélanie Jubault http://orcid.org/0000-0001-9925-8578
Benjamin Liégard http://orcid.org/0000-0001-8986-761X

Maria J. Manzanares-Dauleux http://orcid.org/0000-0002-6452-0393

## References


**Aigu Y, Laperche A, Mendes J, Lariagon C, Guichard S, Gravot A, Manzanares-Dauleux MJ. 2018.** Nitrogen supply exerts a major/minor switch between two QTLs controlling *Plasmodiophora brassicae* spore content in rapeseed. *Plant Pathology* **67**: 1574–1581.

**Alix K, Lariagon C, Delourme R, Manzanares-Dauleux MJ. 2007.** Exploiting natural genetic diversity and mutant resources of *Arabidopsis thaliana* to study the *A. thaliana–Plasmodiophora brassicae* interaction. *Plant Breed* **126**: 218–221.

**Aller EST, Jagd LM, Kliebenstein DJ, Burow M. 2018.** Comparison of the relative potential for epigenetic and genetic variation to contribute to trait stability. *G3* **8**: 1733–1746.

**Aoun N, Tauleigne L, Lonjon F, Deslandes L, Vailleau F, Roux F, Berthomé R. 2017.** Quantitative disease resistance under elevated temperature: genetic basis of new resistance mechanisms to *Ralstonia solanacearum*. *Frontiers in Plant Science* **8**: 1387.

**Arbeiter A, Fahling M, Graf H, Sacristán MD, Siemens J. 2002.** Resistance of *Arabidopsis thaliana* to the obligate biotrophic parasite *Plasmodiophora brassicae*. *Plant Protection Science - Prague* **38**: 519–522.

**Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011.** Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**: 245–249.

**Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Görlach J. 2001.** Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* **13**: 1499–1510.

**Boyko A, Kovalchuk I. 2011.** Genetic and epigenetic effects of plant–pathogen interactions: an evolutionary perspective. *Molecular Plant* **4**: 1014–1023.

**Broman KW, Sen Ś. 2009.** Fit and exploration of multiple-QTL models. In: *A guide to QTL mapping with R/qtl. Statistics for biology and health*. New York, NY, USA: Springer, 241–282.

**Broman KW, Wu H, Sen Ś, Churchill GA. 2003.** R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–890.

**Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E *et al.* 2012.** Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proceedings of the National Academy of Sciences, USA* **109**: 16240–16245.

**Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury J-M, Wincker P *et al.* 2014.** Mapping the epigenetic basis of complex traits. *Science* **343**: 1145–1148.

**Diederichsen E, Frauen M, Linders EGA, Hatakeyama K, Hirai M. 2009.** Status and perspectives of clubroot resistance breeding in Crucifer crops. *Journal of Plant Growth Regulation* **28**: 265–281.

**Dixon GR. 2009.** The occurrence and economic impact of *Plasmodiophora brassicae* and clubroot disease. *Journal of Plant Growth Regulation* **28**: 194–202.

**Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR. 2012.** Widespread dynamic DNA methylation in response to biotic stress. *Proceedings of the National Academy of Sciences, USA* **109**: E2183–E2191.

**Fahling M, Graf H, Siemens J. 2003.** Pathotype separation of *Plasmodiophora brassicae* by the host plant. *Journal of Phytopathology* **151**: 425–430.

**Gallusci P, Dai Z, Génard M, Gauffretau A, Leblanc-Fournier N, Richard-Molard C, Vile D, Brunel-Muguet S. 2017.** Epigenetics for plant improvement: current knowledge and modeling avenues. *Trends in Plant Science* **22**: 610–623.

**Gilly A, Etcheverry M, Madoui M-A, Guy J, Quadrana L, Alberti A, Martin A, Heitkam T, Engelen S, Labadie K *et al.* 2014.** TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics* **15**: 377.

**Gossen BD, Deora A, Peng G, Hwang S-F, McDonald MR. 2014.** Effect of environmental parameters on clubroot development and the risk of pathogen spread. *Canadian Journal of Plant Pathology* **36**: 37–48.

**Gravot A, Grillet L, Wagner G, Jubault M, Lariagon C, Baron C, Deleu C, Delourme R, Bouchereau A, Manzanares-Dauleux MJ. 2011.** Genetic and physiological analysis of the relationship between partial resistance to clubroot and tolerance to trehalose in *Arabidopsis thaliana*. *New Phytologist* **191**: 1083–1094.

**Gravot A, Richard G, Lime T, Lemarié S, Jubault M, Lariagon C, Lemoine J, Vicente J, Robert-Seilaniantz A, Holdsworth MJ *et al.* 2016.** Hypoxia response in *Arabidopsis* roots infected by *Plasmodiophora brassicae* supports the development of clubroot. *BMC Plant Biology* **16**: 251.

**Hewezi T, Lane T, Piya S, Rambani A, Rice JH, Staton M. 2017.** Cyst nematode parasitism induces dynamic changes in the root epigenome. *Plant Physiology* **174**: 405–420.

**Holland J. 2007.** Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* **10**: 156–161.

**Ingram DS, Tommerup IC. 1972.** The life history of *Plasmodiophora brassicae Woron. Proceedings of the Royal Society of London. Series B: Biological Sciences* **180**: 103–112.

**Johannes F, Colot V, Jansen RC. 2008.** Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews. Genetics* **9**: 883–890.

**Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P *et al.* 2009.** Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics* **5**: e1000530.

**Jubault M, Lariagon C, Simon M, Delourme R, Manzanares-Dauleux MJ. 2008.** Identification of quantitative trait loci controlling partial clubroot resistance in new mapping populations of *Arabidopsis thaliana*. *TAG. Theoretical and Applied Genetics* **117**: 191–202.

**Kakutani T, Jeddeloh JA, Flowers SK, Munakata K, Richards EJ. 1996.** Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proceedings of the National Academy of Sciences, USA* **93**: 12406–12411.

**Kellenberger RT, Schlüter PM, Schiestl FP. 2016.** Herbivore-induced DNA demethylation changes floral signalling and attractiveness to pollinators in *Brassica rapa*. *PLoS ONE* **11**: e0166646.

**Koch E, Cox R, Williams PH. 1991.** Infection of *Arabidopsis thaliana* by *Plasmodiophora brassicae*. *Journal of Phytopathology* **132**: 99–104.

**Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJB. 2015.** Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant Cell* **27**: 337–348.

**Laperche A, Aigu Y, Jubault M, Ollier M, Guichard S, Glory P, Strelkov SE, Gravot A, Manzanares-Dauleux MJ. 2017.** Clubroot resistance QTL are modulated by nitrogen input in *Brassica napus*. *TAG. Theoretical and Applied Genetics* **130**: 669–684.

**Lee J, Izzah NK, Choi B-S, Joh HJ, Lee S-C, Perumal S, Seo J, Ahn K, Jo EJ, Choi GJ *et al.* 2016.** Genotyping-by-sequencing map permits identification of clubroot resistance QTLs and revision of the reference genome assembly in cabbage (*Brassica oleracea* L.). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* **23**: 29–41.

**Lemarié S, Robert-Seilaniantz A, Lariagon C, Lemoine J, Marnet N, Jubault M, Manzanares-Dauleux MJ, Gravot A. 2015a.** Both the jasmonic acid and the salicylic acid pathways contribute to resistance to the biotrophic clubroot agent *Plasmodiophora brassicae* in Arabidopsis. *Plant & Cell Physiology* **56**: 2158–2168.

**Lemarié S, Robert-Seilaniantz A, Lariagon C, Lemoine J, Marnet N, Levrel A, Jubault M, Manzanares-Dauleux M, Gravot A. 2015b.** Camalexin contributes to the partial resistance of *Arabidopsis thaliana* to the biotrophic soilborne protist *Plasmodiophora brassicae*. *Frontiers in Plant Science* **6**: 539.

**Lenth RV. 2016.** Least-squares means: the R Package lsmeans. *Journal of Statistical Software* **69**: 1–33.

**Liu J, Feng L, Li J, He Z. 2015.** Genetic and epigenetic control of plant heat responses. *Frontiers in Plant Science* **6**: 267.

**López Sánchez A, Stassen JHM, Furci L, Smith LM, Ton J. 2016.** The role of DNA (de)methylation in immune responsiveness of *Arabidopsis*. *The Plant Journal: for Cell and Molecular Biology* **88**: 361–374.

**Luna E, Bruce TJA, Roberts MR, Flors V, Ton J. 2012.** Next-generation systemic acquired resistance. *Plant Physiology* **158**: 844–853.

**Manzanares-Dauleux MJ, Delourme R, Baron F, Thomas G. 2000a.** Mapping of one major gene and of QTLs involved in resistance to clubroot in *Brassica napus*. *TAG. Theoretical and Applied Genetics* **101**: 885–891.

**Manzanares-Dauleux MJ, Divaret I, Baron F, Thomas G. 2000b.** Evaluation of French *Brassica oleracea* landraces for resistance to *Plasmodiophora brassicae*. *Euphytica* **113**: 211–218.

**Marschner IC. 2011.** glm2: Fitting generalized linear models with convergence problems. *The R Journal* **3**: 12–15.

**Piao Z, Ramchiary N, Lim YP. 2009.** Genetics of clubroot resistance in *Brassica* species. *Journal of Plant Growth Regulation* **28**: 252–264.

**Quadrana L, Colot V. 2016.** Plant transgenerational epigenetics. *Annual Review of Genetics* **50**: 467–491.

**R Development Core Team. 2015.** *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [WWW document] URL https://www.R-project.org/.

**Reinders J, Wulff BBH, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J. 2009.** Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes & Development* **23**: 939–950.

**Rocherieux J, Glory P, Giboulot A, Boury S, Barbeyron G, Thomas G, Manzanares-Dauleux MJ. 2004.** Isolate-specific and broad-spectrum QTLs are involved in the control of clubroot in *Brassica oleracea*. *TAG. Theoretical and applied genetics* **108**: 1555–1563.

**Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G *et al.* 2013a.** Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research* **23**: 1663–1674.

**Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR. 2011.** Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**: 369–373.

**Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ *et al.* 2013b.** Patterns of population epigenomic diversity. *Nature* **495**: 193–198.

**Sharma K, Gossen BD, McDonald MR. 2011.** Effect of temperature on cortical infection by *Plasmodiophora brassicae* and clubroot severity. *Phytopathology* **101**: 1424–1432.

**Sharma R, Vishal P, Kaul S, Dhar MK. 2017.** Epiallelic changes in known stress-responsive genes under extreme drought conditions in *Brassica juncea* (L.) Czern. *Plant Cell Reports* **36**: 203–217.

**Siemens J, Nagel M, Ludwig-Muller J, Sacristan MD. 2002.** The interaction of *Plasmodiophora brassicae* and *Arabidopsis thaliana*: parameters for disease quantification and screening of mutant lines. *Journal of Phytopathology* **150**: 592–605.

**Some A, Manzanares MJ, Laurens F, Baron F, Thomas G, Rouxel F. 1996.** Variation for virulence on *Brassica napus* L. amongst *Plasmodiophora brassicae* collections from France and derived single-spore isolates. *Plant Pathology* **45**: 432–439.

**Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF *et al.* 2009.** A role for RNAi in the selective correction of DNA methylation defects. *Science* **323**: 1600–1604.

**Tierney L, Rossini AJ, Li N. 2009.** Snow: a parallel computing framework for the R system. *International Journal of Parallel Programming* **37**: 78–90.

**Vongs A, Kakutani T, Martienssen RA, Richards EJ. 1993.** *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**: 1926–1928.

**Weigel D, Colot V. 2012.** Epialleles in plant evolution. *Genome Biology* **13**: 249.

**Yu A, Lepère G, Jay F, Wang J, Bapaume L, Wang Y, Abraham A-L, Penterman J, Fischer RL, Voinnet O *et al.* 2013.** Dynamics and biological relevance of DNA demethylation in *Arabidopsis* antibacterial defense. *Proceedings of the National Academy of Sciences, USA* **110**: 2389–2394.

**Zhang Y-Y, Fischer M, Colot V, Bossdorf O. 2013.** Epigenetic variation creates potential for evolution of plant phenotypic plasticity. *New Phytologist* **197**: 314–322.

**Zheng X, Chen L, Xia H, Wei H, Lou Q, Li M, Li T, Luo L. 2017.** Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Scientific Reports* **7**: 39843.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Dataset S1** Location of segregating DNA sequence variants (SNP, indel, TE, SV) among the clubroot QTL intervals.

**Dataset S2** Statistic analysis of TE and sequence variant effect compared to methylated markers on QTL$^{epi}$ detection for each trait study.

**Fig. S1** Epiallele effects at the closest QTL$^{epi}$ peak marker.

**Fig. S2** Boxplot of temperature data recorded in each growth room.

**Fig. S3** Comparison of the epiallele effects at the closest QTL$^{epi}$ peak marker between Lfi and Lfni in growth chamber-2.

**Table S1** List of primer sets used to confirm homozygosity of the T-DNA insertion in mutants.

**Table S2** Disease index for Col-0 and mutants after infection by *P. brassicae*, and Dunnett's post hoc test results.

**Table S3** Phenotypic responses of epiRIL and their parent lines to infection by *P. brassicae*.

**Table S4** Analysis of epigenotype, temperature and interaction between temperature and epigenotype effect for each trait in infected condition.

**Table S5** Dunnett's test comparison of temperatures recorded by each temperature sensor in growth room-2.

**Table S6** Median temperature monitored by each temperature sensor in growth room-2.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

## 7.3 Conclusion et discussion

### 7.3.1 QTL épigénétiques et prise en compte des variants ADN en ségrégation

Dans la mesure où les modèles statistiques de détection de QTL existant à ce jour ne permettent pas d'incorporer un pedigree du type de celui mis en évidence pour la population epiRIL, nous sommes ici face à un problème complexe.
En effet, pour la détection de QTL, on considère soit un schéma classique à deux origines parentales (que les origines soient déterminées sur la base de différences dans les profils de méthylation ou dans la séquence du génome), comme effectué dans (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014 ; Kooke et al. 2015) ; soit un schéma plus complexe à l'image de celui employé dans les populations MAGIC en raison des multiples origines parentales ; mais à notre connaissance il n'existe de pas de modèle permettant de considérer pour la détection des QTL ainsi que l'estimation de l'héritabilité et des effets un schéma dans lequel la ségrégation des origines parentales suit deux plans de croisement différents (backcross classique pour les origines inférées sur la base des profils de méthylation, croisement entre deux individus non apparentés pour les origines inférées sur la base des polymorphismes nucléotidiques).
Aussi, il est difficile d'aller au-delà des deux approches décrites ici, et le développement d'un méthode appropriée serait requise. Quoi qu'il en soit, il est important de garder à l'esprit que les QTLepi préalablement détectés ne sont en rien invalidés par la présence de variants ADN, et compte-tenu de la non-coségrégation pour la majorité d'entre eux il est même envisageable de conclure que ceux-ci sont à proprement parler portés par des épiallèles.

### 7.3.2 Stratégie d'identification de l'épimutation causative : limites de la cartographie fine et autres approches envisageables

Quant bien même la cartographie fine du QTL situé sur le chromosome 4 et influençant la longueur de la racine primaire n'ait pas été menée à terme, l'emploi de cette approche peut être discuté.
A la différence des populations RIL classiques qui dérivent du croisement entre deux accessions, le polymorphisme de méthylation est limitant dans les epiRIL, ce qui va limiter la résolution de la cartographie fine. En effet, si la carte génétique de la population epiRIL est fondée sur 123 marqueurs, ce qui est du même ordre de grandeur que les cartes génétiques établies dans des RIL ; il n'y a en fait que 867 DMR pouvant faire office de marqueurs

dans cette population (Colomé-Tatché et al. 2012), contre plusieurs dizaines de milliers de polymorphisme ADN dans les RIL. Une limite additionnelle est que comme ces marqueurs génétiques sont fondés sur des différences stables de méthylation, ils co-localisent avec des ET ou d'autres séquences répétées qui ne sont pas répartis uniformément dans le génome mais au contraire concentrés dans les régions péricentromériques. Ainsi donc, la densité en marqueurs est plus faible dans les bras chromosomiques que dans les régions péricentromériques, or c'est justement dans les bras que l'intensité de recombinaison est la plus forte. Aussi, cette distribution spécifique des marqueurs dans les epiRIL rend plus complexe l'obtention de recombinants à partir desquels poursuivre une cartographie fine.

On peut cependant noter qu'en assouplissant les critères de définition d'une DMR (par exemple, en imposant une différence de méthylation moins importante), il est possible d'augmenter le nombre de marqueurs potentiels et le séquençage bisulphite génome-entier actuellement mené au laboratoire pour 121 epiRIL devrait permettre d'identifier à la résolution de la cytosine des différences de méthylation additionnelles qui pourront également servir de marqueurs. Néanmoins, l'on reste plusieurs ordres de grandeurs en deçà du niveau de polymorphisme dans un dispositif RIL. Si cela s'avère problématique quant à la résolution qui peut être atteinte par cartographie fine, le corrolaire en est que puisque le polymorphisme est plus restreint dans les epiRIL, alors le variant causal sous-tendant le QTLepi est à rechercher parmi un nombre plus réduit de candidats.

Dans le cas présent, dans la mesure où il est probable que des QTLepi correspondent à des eQTL (ou QTL d'expression), on peut s'attendre à ce que l'épivariant causatif puisse être détecté via l'altération du patron d'expression de gènes à proximité, et pourrait donc être identifié par RNAseq. Néanmoins, à la différence des QTLepi impliqués dans la réponse à un stimulus environnemental biotique (réponse à un pathogène, Liégard et al, Section 7.2.4) ou abiotique (évitement de l'ombre, collaboration avec l'équipe Barneche, Fiorucci et al, in prep), les traits que nous étudions ici, la longueur de la racine et la date de floraison, résultent de l'intégration d'un grand nombre de voies à différentes étapes du développement. Aussi, il n'est pas garanti qu'un RNAseq entre plantes à racines courtes vs longues récoltées à l'issue du phénotypage révèle les loci dérégulés par des altérations du profil de méthylation. A l'opposé, une stratégie qui pourrait être envisagée pour réduire le nombre de DMR candidates serait de réaliser un séquençage bisulphite génome-entier sur des pools de plantes à racines courtes versus racines longues issues de la descendance de l'HIF validée. Dans le cadre de cette approche, les régions différentiellement méthylées entre plantules à racine courte et à racine longue sont autant d'épiallèles candidats.

Un commentaire général quant à l'identification par approches gène-candidat de variants (ou épivariants) potentiellement causaux, est qu'il est complexe d'établir une liste de loci prioritaires, a fortiori -et comme c'est le cas ici- lorsqu'il ne s'agit pas d'un QTL à

fort effet et pour lequel il n'y a aucune suspicion de dérégulation des gènes majeurs de la voie. Au cours de mon stage de Master 2 au sein du laboratoire, j'ai proposé des approches gènes-candidats fondées sur des relations de co-expression entre gène dans l'intervalle QTL et gènes associés à l'un ou l'autre des traits, sans pour autant parvenir à identifier de candidat évident. Une approche combinée peut être celle d'un EWAS (*Epigenome-Wide Association Study*), à présent que les génomes, méthylomes et données de phénotypage sont disponibles pour un grand nombre d'accession d'Arabidopsis (Consortium 2016 ; Kawakatsu, S.-S. C. Huang et al. 2016 ; Seren et al. 2017), sous réserve bien évidemment que les déterminants pour la date de floraison ou la longueur de la racine soient les mêmes à la fois entre populations naturelles et en laboratoire, ce qui constitue une limite théorique à l'extrapolation directe des résultats. Néanmoins, au cours de mon stage, j'ai pu proposer comme candidat pour la date de floraison le gène *AT4G15530*, indépendamment identifié dans le cadre d'un EWAS pour ce même trait (données non publiées du laboratoire, en collaboration avec Frank Johannes) et par ailleurs différentiellement exprimé entre sauvage et *ddm1*. Ce gène apparaît donc être un candidat solide pour le QTLepi influençant la date de floraison détecté sur le chromosome 4 (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) et une tentative de déméthylation ciblée de la DMR qui lui est associée sera effectuée dès que l'approche dCas9-DME aura été validée.

### 7.3.3 Avancées dans le domaine de l'édition de l'épigénome

Comme décrit dans l'introduction, l'état des connaissances dans le domaine de l'édition de l'état de méthylation d'un locus donné a grandement évolué au cours du temps imparti à ma thèse. Aussi, l'approche employée peut être discutée au regard des travaux publiés à ce jour. Il semble avéré que l'emploi d'un système de multiplexage d'ARN guides soit une stratégie pertinente chez les plantes, puisque des approches similaires ont été publiées dans l'intervalle, aussi bien pour du ciblage d'activité enzymatique que pour l'inactivation de plusieurs gènes (Čermák et al. 2017 ; Lowder et al. 2015 ; X. Ma et al. 2015 ; Qi et al. 2016 ; Vazquez-Vilar et al. 2016 ; W. Wang et al. 2016 ; Z. Wang et al. 2018 ; Yan et al. 2016). Du côté de la protéine effectrice, des travaux récents ont réussi à effectuer une déméthylation locus-spécifique chez Arabidopsis, ce en employant une fusion avec le domaine catalytique de TET, multiplexée au moyen d'un dispositif SunTag (Gallego-Bartolomé et al. 2018). Dans ce contexte, on peut se demander quelle sera l'efficacité de déméthylation de la fusion directe dCas9-DME que nous employons ici, néanmoins le système à deux composantes que nous avons développé nous permettra éventuellement de substituer cette fusion directe par une version de type SunTag. Dans tous les cas, l'emploi de DME pour une déméthylation n'a pas été publiée à ce jour, en ce sens nos travaux en cours demeurent originaux.

Troisième partie

# Discussion générale et perspectives

# Conclusion générale et perspectives

## Bilan des résultats obtenus et discussion

La majeure partie des travaux présentés dans ce manuscrit ont visé à évaluer l'impact d'une perte de méthylation de l'ADN sur le taux et le spectre des mutations spontanées chez la plante modèle *Arabidopsis thaliana*.

Au moyen d'une population de lignées assimilables à des lignées d'accumulation de mutation (MA lines) mais aux épigénomes mosaïques de régions méthylées et hypométhylées, j'ai pu mettre en évidence une réduction spécifique du taux de transition C → T en lien avec à la fois un nombre réduit de cytosines méthylées dans cette population mais également avec une réactivation transcriptionnelle des ET, ce qui suggère la contribution additionnelle des voies de réparation de l'ADN couplée à la transcription (*transcription-coupled DNA repair*).
Si ces travaux proposent une évaluation directe de l'impact mutationnel de la 5mC qui complète l'état de l'art, ils nécessitent à ce jour plusieurs analyses complémentaires comme discuté dans la section dédiée. En effet, afin d'obtenir une image plus précise des facteurs additionnels susceptibles d'influencer cette réduction du taux de C → T, il s'agira d'incorporer à la fois le contenu local en GC ainsi que l'état d'ouverture de la chromatine dans les régions péricentromériques dérivées de *ddm1*. La première de ces analyses est en cours, la seconde pourra être effectuée dans les prochaines semaines au profit des données de compaction de la chromatine par ATACseq prochainement disponibles au laboratoire. De plus, l'obtention prochaine des méthylomes à la résolution de la cytosine unique (séquençage bisulphite, WGBS) pour 121 epiRIL permettra de préciser l'estimation du taux de transition C → T induit par la méthylation dans les différents contextes (CG, CHG, CHH) et compartiment génomiques (bras et régions péricentromériques, hérités de WT ou *ddm1*), avec en plus la possibilité de lier le niveau de méthylation au taux de mutation ; une analyse qui n'était pas permise par les méthylomes MeDIP-chip (Colomé-Tatché et al. 2012 ; Cortijo, Wardenaar, Colomé-Tatché, Johannes et al. 2014).

Par ailleurs, mon analyse de l'impact mutationnel de la remobilisation extensive des ET dans les epiRIL met en évidence une absence de réarrangements chromosomiques associés aux ET remobilisés, avec comme discuté Chapitre 6 plusieurs hypothèses envisageables dont l'une ayant trait à la nature des ET mobilisés dans cette population. Il peut en outre être envisagé que comme mis en évidence chez la levure où les répétitions *Ty* sont associées à un nombre d'évènements de NAHR bien plus faible qu'attendu (Sa-

saki et al. 2010), la machinerie de recombinaison d'Arabidopsis ait développé un moyen similaire de limiter les réarrangements entre séquences homologues dispersées.
J'identifie également à la fois dans les epiRIL et les MA lines une tendance à la formation de duplications en tandem par un mécanisme réplicatif qui pourrait être le FoSTeS ; ainsi qu'à la formation de délétions mais pas de duplications par NAHR au niveau d'annotations associées à des ET, une seconde observation qui fait écho à la tendance à la réduction du génome par élimination des ET proposée chez Arabidopsis. Ces deux observations nécessiteraient cependant un nombre d'évènements plus élevé pour pouvoir conclure. De façon similaire, il semblerait que de tels évènements de délétions associés à des annotations d'ET soient plus fréquents dans les régions péricentromériques dérivées de *ddm1*, suggérant comme proposé chez les mammifères l'hypothèse attrayante d'une levée de l'inhibition de la recombinaison médiée par la méthylation de l'ADN dans ces régions. Là encore il faudrait un nombre d'évènements plus élevé pour espérer tirer une telle conclusion.

Dans leur ensemble, mes travaux de thèse fournissent des informations additionnelles quant au patron des mutations spontanées chez Arabidopsis.
En raison des spécificités de la population epiRIL présentées dans ce manuscrit, une ligne de recherche complémentaire peut être envisagée. Puisque les intervalles dans lesquels se sont produits les crossing-over peuvent être identifiés sur la base des différences stables de méthylation mais également des variants ADN en ségrégation ; il serait intéressant d'étudier la corrélation entre taux de mutation et intensité de recombinaison dans cette population fortement homogène génétiquement. Une telle analyse, actuellement en cours, permettra d'évaluer l'impact mutationnel de cette force génomique dans le contexte d'une méiose ColxCol, complémentant ainsi des travaux précédents ayant conduit à la mise en évidence d'un effet mutationnel de la recombinaison au cours d'une méiose ColxLer (Yang et al. 2015), donc fortement polymorphe (Zapata et al. 2016) et peu représentative d'une méiose standard.

Au cours de ma thèse, j'ai également effectué une caractérisation détaillée de la variation nucléotidique en ségrégation dans la population epiRIL. Ce travail m'a permis de confirmer que les différents QTLepi préalablement identifiés au laboratoire pour la date de floraison et la longueur de la racine (Cortijo, Wardenaar, Colomé-Tatché, Gilly et al. 2014) étaient effectivement épigénétiques. Dans l'optique de poursuivre leur caractérisation, j'ai entrepris d'établir des populations de cartographie, qui m'ont conduite à valider l'un d'entre eux. J'ai par ailleurs contribué à mettre en place une approche expérimentale d'altération du patron de méthylation locus-spécifique, qui permettra à terme d'effectuer la complémentation fonctionnelle des variants de méthylation causatifs des QTLepi.
Néanmoins, comme illustré dans le cadre du manuscrit joint à ce chapitre, et en dépit

de la complexité inattendue du pedigree de la population epiRIL que j'ai mise en évidence, certains de ces variants ADN en ségrégation contribuent potentiellement aux QTL "épigénétiques" détectés sur la base de différences de méthylation. Par ailleurs, j'ai pu mettre en évidence un QTLadn influençant la date de floraison dans la population epiRIL, qui s'ajoute aux 3 QTLepi identifiés pour ce trait. Aussi, la population epiRIL pourrait constituer un système idéal dans lequel évaluer la contribution relative des variants ADN et épialléliques à la variation héritable pour les traits complexes.

# Perspectives

Mes travaux de thèse, résumés dans la section précédente, ont eut trait à l'analyse du patron de mutation d'Arabidopsis mais ont également touché du doigt des concepts et méthodes de génétique quantitative. Aussi, des perspectives à long terme de ce travail de thèse peuvent être envisagées au travers de ces deux axes.

## Patrons de mutation et évolution du génome

Il a pu être proposé que la méthylation de l'ADN joue un rôle significatif dans l'évolution du contenu en GC des génomes (Mugal, Arndt et al. 2015). Cette observation peut-elle cependant être extrapolée aux plantes, et plus spécifiquement à Arabidopsis ? Dans la mesure où la méthylation est presque exclusivement restreinte aux ET qui ne constituent qu'une faible fraction du génome d'Arabidopsis, ce n'est assurément pas ce taux de mutation lié à la méthylation qui va altérer à l'échelle de l'évolution le contenu en GC de cet organisme. Néanmoins, il serait pertinent d'évaluer plus avant l'effet de cette hypermutabilité des cytosines méthylées dans d'autres génomes plus riches en ET à l'exemple du Maïs.

Par contre, si l'on considère la seule fraction du génome correspondant aux annotations d'ET, les conclusions sont toutes autres.
Il a pu être mis en évidence que les ET présentaient un mode d'accumulation des mutations en deux stades successifs ; l'un médié majoritairement par la désamination des 5mC, puis l'autre au taux de mutation "standard" une fois qu'il ne reste plus de cytosines méthylées à ces séquences (Maumus et al. 2014). Ainsi, les observations faites dans le cadre de ce travail pourraient être exploitées pour mieux caractériser le patron d'évolution de ces séquences, notamment en permettant le calcul d'une "demi-vie" des ET, c'est-à-dire, le moment où le taux de mutation de ces séquences va présenter une inflexion avec le passage d'un taux "rapide" (désamination des 5mC) à un taux "lent" (accumulation de mutations au taux attendu dans les régions hétérochromatiques).
A l'opposé, il faut noter qu'il ne semble pas adapté d'employer les travaux présentés ici

pour estimer les conséquences mutationnelles de la méthylation du corps des gènes, puisqu'il s'agit d'un contexte chromatinien très distinct.

De plus, une observation générale est que l'ensemble des travaux présentés ici ont été effectués dans l'accession Col-0 d'Arabidopsis. Des travaux menés dans plusieurs organismes et notamment chez le Nématode ont pu mettre en évidence que le taux et le spectre de mutation pouvaient varier entre souches de la même espèce. Aussi, il serait intéressant d'analyser le patron de mutation d'autres accessions d'Arabidopsis, par exemple au travers de l'établissement de MA lines. Cela permettrait de fournir une image plus globale des grandes tendances mutationnelles chez cet organisme, et tout particulièrement de valider si cette tendance à la réduction du génome par délétion des ET y est pervasive.

## Mutations, épimutations et architecture génétique des traits complexes

Pouvoir quantifier la fraction de la variation héritable qui peut être attribuée aux nouvelles mutations est une question centrale en génétique quantitative. Dans le contexte des epiRIL, mi-RIL, mi-MA lines, elle prend une dimension particulière.

En effet, comme précisé plus haut, nous sommes ici dans un système expérimental dans lequel il est possible d'évaluer à la fois la contribution des variants de méthylation ainsi que des variants ADN à la variation héritable dans les traits complexes.
Il pourrait être pertinent d'aller au-delà, et de chercher à estimer dans quelle mesure les mutations qui se sont accumulées lors de la propagation des lignées influencent les différents traits, et particulièrement les traits d'histoire de vie (biomasse, date de floraison, nombre de graines), puisque ce sont ce type de traits qui vont influencer la fitness et donc être "pertinents" sur le plan de l'évolution.
En détail, il s'agirait d'aller analyser la distribution des effets sur la fitness (*distribution of fitness effects*, DFE) des différentes mutations. Une limite toutefois a à voir avec le fait que les mutations présentes dans les epiRIL sont pour leur majorité des nouvelles insertions d'ET, aussi en comparaison des mutations à large effets qu'ils ont pu générer, il sera complexe d'évaluer l'impact des seules mutations ponctuelles. Cependant, on peut envisager retourner cet argument et viser au contraire à évaluer l'impact sur la fitness des nouvelles insertions d'ET, ce qui compléterait les observations décrites dans (Quadrana, Etcheverry et al. 2018) concernant les préférences d'insertions des différentes familles d'ET.

Quatrième partie

# Matériels et méthodes

# Matériels et méthodes

## Jeux de données utilisés

### Données epiRIL

#### Séquencage du génome des epiRIL

Le génome de 121 epiRIL (parmi les 123 lignées pour lesquelles le méthylome est disponible), d'un individu *ddm1* et d'un wt a été séquencé tel que décrit dans (Gilly et al. 2014) et est reproduit dans le manuscrit (Quadrana, Etcheverry et al. 2018).
Brièvement, pour chaque epiRIL, les ADN ont été extraits à partir d'un lot d'une dizaine de plantules F9. Ainsi, les mutations ponctuelles ou les néo-insertions d'ET identifiées dans cette génération correspondent les unes et les autres à des évènements apparues dans la plante F8.
Le séquençage correspond à un séquençage Illumina en paired-end de banques mate-pair, une approche qui avait été privilégiée en raison de l'objectif initial de ce reséquençage ; à savoir la détection de nouvelles insertions d'ET. Le principe de l'une et l'autre des ces approches est illustré Figure 7.8.

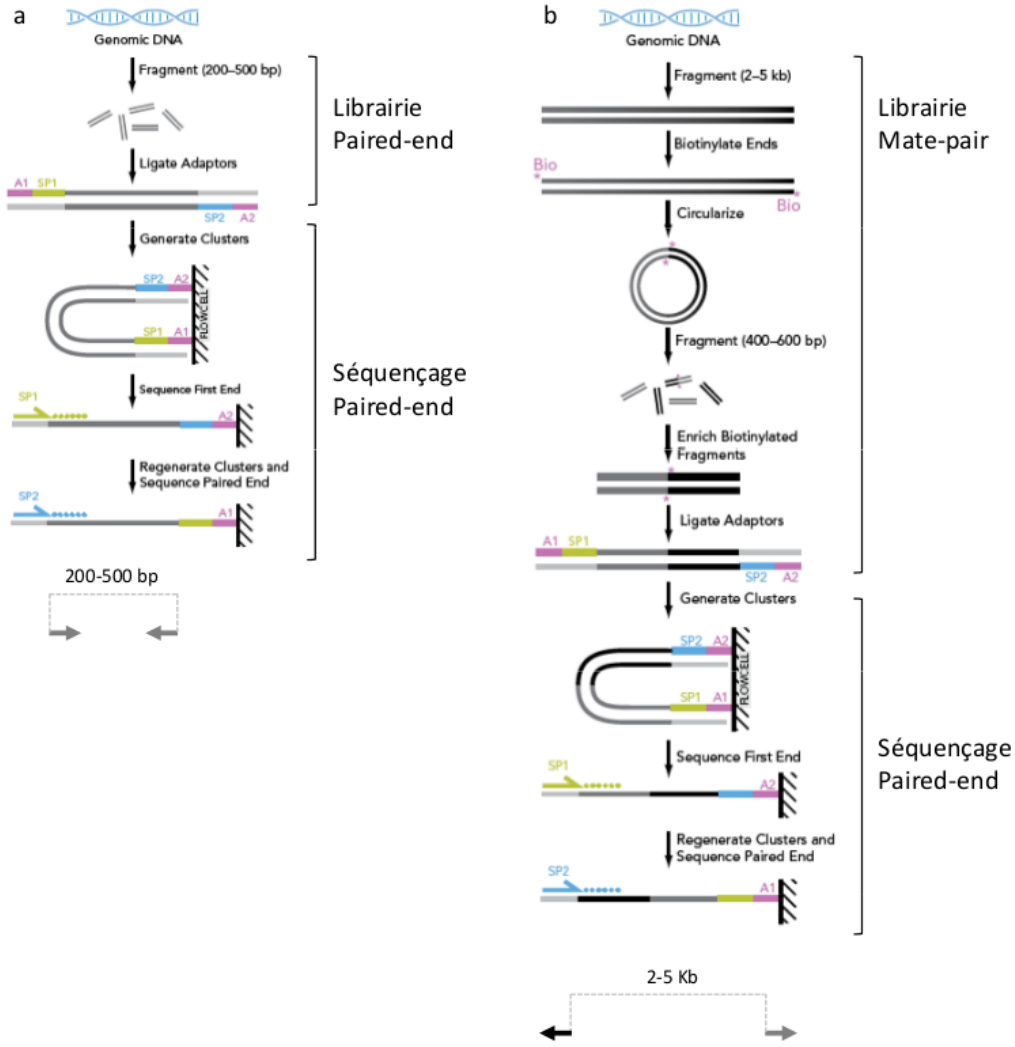


FIGURE 7.8 – Différences entre technologies *mate-pair* et *paired-end*. Principe du séquençage *paired-end* de banques *paired-end*, à gauche, et *mate-pair*, à droite.

### Méthylome MeDIP-chip

L'obtention des méthylomes de 123 epiRIL F8 et d'individus WT et *ddm1* par MeDIP-chip est décrite dans (Colomé-Tatché et al. 2012) et le protocole publié dans (Cortijo, Wardenaar, Colomé-Tatché, Johannes et al. 2014). En bref, dans cette méthode, l'ADN simple brin méthylé est immunoprécipité à l'aide d'un anticorps dirigé contre les groupements méthyles, puis la fraction immunoprécipitée et l'ADN total sont marqués différentiellement et hybridés sur une puce Nimblegen tiling array contenant 711320 sondes choisies dans des fenêtres consécutives de 165 nt le long du génome d'*Arabidopsis thaliana*. L'état de méthylation pour chaque sonde est ensuite inféré au moyen d'un modèle de Markov caché à trois états (*three-state Hidden Markov Model*, HMM) : méthylé (M), non-méthylé (U) ou méthylation intermédiaire (I).

#### Méthylome WGBS

Les WGBS pour 5 epiRIL ont été obtenus et analysés comme décrits dans (Quadrana, Etcheverry et al. 2018).

#### RNAseq

Les RNAseq pour 5 epiRIL ont été obtenus et analysés comme décrits dans (Quadrana, Etcheverry et al. 2018).

### Données MA lines

Les données de séquençage du génome et du méthylome de 10 MA lines ont été obtenues comme décrit dans la section Matériels et Méthodes de l'article en préparation Chapitre 5.

## Identification des variations nucléotidiques

#### Mutations ponctuelles

L'identification des mutations ponctuelles dans les epiRIL et les MA lines a été effectuée comme décrit dans la section Matériels et Méthodes de l'article en préparation Chapitre 5. Brièvement, l'outil HaplotypeCaller de la suite GATK a été utilisé dans le mode *–joint-genotyping* (illustré Figure 7.9), qui offre un cadre computationnel adapté à l'identification de variants au sein de cohortes, qu'ils soient propres à un individu ou en ségrégation dans la population considérée. Cette identification s'effectue en deux temps. La première étape consiste à identifier dans chaque lignée indépendamment les variants probables : pour chaque lignée, l'algorithme recherche les régions (*ActiveRegions*) présentant une état distinct de celui dans le génome de référence fourni, effectue un réassemblage par graphe de De Brujin de ces régions afin d'en extraire un haplotype probable, puis évalue la concordance entre la séquence de chaque read et l'haplotype inféré. L'emploi d'une approche par graphe pour réassembler *de novo* permet notamment une meilleure résolution des variants ponctuels "complexes", par exemple un indel avec des SBS autour. Il en résulte, pour chaque lignée, une matrice (fichier gVCF, *genomic VCF*) récapitulant l'ensemble des positions du génome ainsi que, pour celles qui présentent un état variant par rapport au génome de référence, le génotype à ce locus et le score de vraisemblance qui lui est associé.
Une fois que ces matrices ont été produites pour chaque lignée, un génotypage (*Joint-Genotyping*) est effectué à l'échelle de la cohorte afin de déterminer pour chaque variant identifié son génotype (homozygote référence, homozygote variant ou hétérozygote) dans chaque lignée ainsi que la vraisemblance de ce génotype. Ce génotypage conjoint permet

notamment de ré-évaluer à la hausse le score de vraisemblance d'un variant dans une lignée présentant localement une qualité de séquençage et/ou de mapping limitante, ce sur la base des variants identifiés par d'autres lignées partageant le même haplotype.
A l'issue de cette deuxième étape, l'ensemble des variants identifiés dans chaque lignée et les scores qui y sont associés sont combinés dans un fichier VCF, qui pourra être annoté et filtré sur différents critères afin d'identifier les mutations "vraies". A cette fin, les variants dépassant les scores de qualités minimaux, tels qu'établis sur la base des scores atteints par les mutations validées expérimentalement ainsi que de la mutation *ddm1* ont été sélectionnés. Chacune de ces mutations probables, SBS ou indel, a ensuite été inspectée visuellement à l'aide d'IGV (Thorvaldsdóttir et al. 2013).

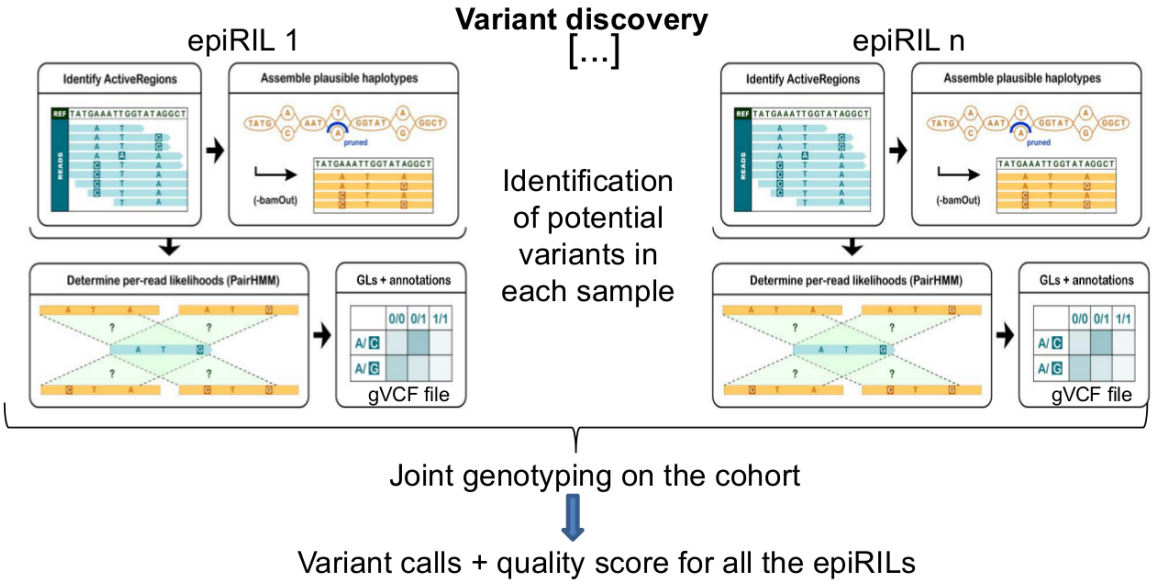


FIGURE 7.9 – Détection de variations ponctuelles avec GATK HaplotypeCaller. Explications dans le texte.

### Variations structurales hors ET

L'identification des SV autres que nouvelles insertions d'ET a été effectuée en combinant les outils Hydra (Lindberg et al. 2015), Delly (Rausch et al. 2012) et Control-FREEC (Boeva et al. 2012), qui emploient respectivement des approches *read-pair*, *split-read* (détection des duplications, délétions, translocations et inversions) et *read depth* (détection des duplications et délétions). La combinaison de différentes approches, dont le principe et la complémentarité pour détecter les principaux types d'évènements sont illustrées FIGURE 7.10, permet de réduire le nombre de faux-positifs et d'obtenir des résultats plus robustes.

L'identification des évènements de translocations parmi les epiRIL n'a pas été achevée à ce jour en raison d'un *runtime* trop important. La majorité des outils existants pour la détection de SV sont en effet optimisés pour des données de séquençage *paired-end*, et

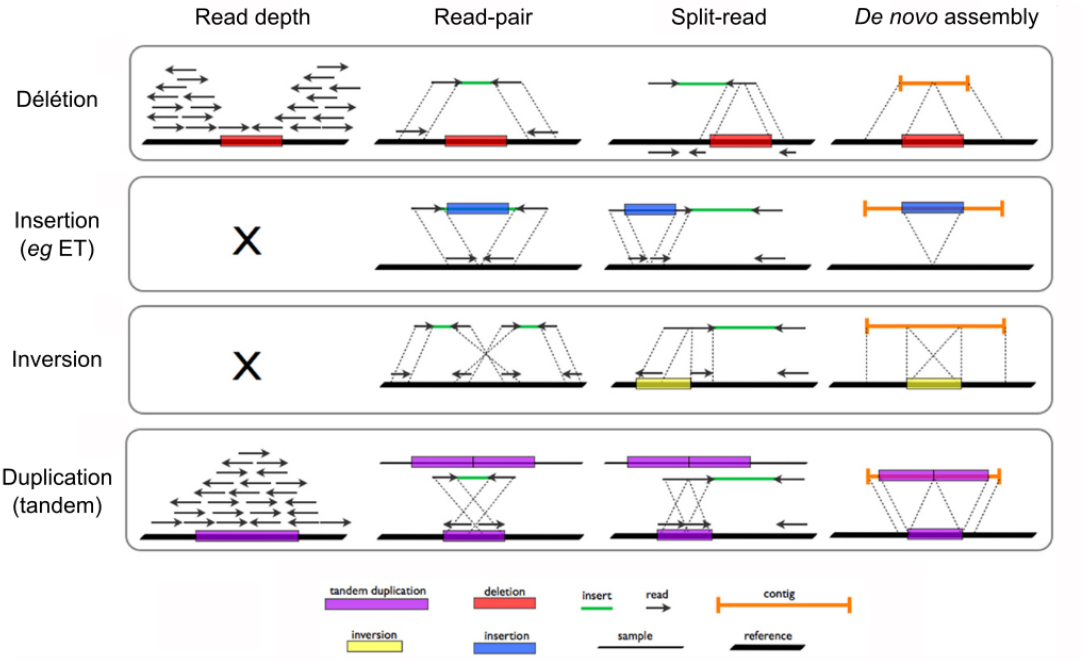


FIGURE 7.10 – Principe des approches de détection de SV. Pour chaque approche (*read depth*, *read-pair*, *split-read*, assemblage *de novo*) sont figurées les modalités de détection des principaux types de SV (délétions, duplications, insertion -par exemple, d'un ET- et inversion). La détection des translocations n'est pas figurée en ce que ce type d'évènement s'interprète comme une délétion couplée à une insertion ailleurs dans le génome, avec ou non inversion par rapport à l'orientation initiale du locus transloqué. Les croix indiquent l'impossibilité d'employer l'approche *read depth* pour la détection de SV autres que les délétions et duplications. L'approche par assemblage *de novo* à l'échelle du génome entier est illustrée à titre informatif mais demeure difficile à mettre en œuvre pour la détection de SV dans les jeux de données short reads qui constituent la majorité des données NGS à ce jour. L'orientation des reads (sens des flèches) correspond à des données de séquençage *paired-end*. Modifié d'après (Tattini et al. 2015).

non pas *mate-pair*, or c'est cette dernière stratégie qui a été employée pour le séquençage des epiRIL (cf supra). D'autres outils censés mieux prendre en compte les spécificités des données *mate-pair* (Gillet-Markowska et al. 2015 ; Iakovishina et al. 2016) ont également été considérés, sans résultat à ce jour sur les données epiRIL. La simulation des signatures de microhomologie aux breakpoints des duplications en tandem a été effectuée au moyen d'un script Perl adapté de (Newman et al. 2015).

### Nouvelles insertions d'ET

La détection des néo-insertions d'ET dans les epiRIL est décrite dans (Quadrana, Etcheverry et al. 2018). Elle combine les outils SPLITREADER (Quadrana, Bortolini Silveira et al. 2016) et TE-tracker (Gilly et al. 2014).

# Conditions de culture des plantes

### Culture *in vitro*

Les graines sont désinfectées dans une solution eau stérile - javel 5%- Tween 20 0,2% pendant 10 minutes, puis lavées à l'éthanol 70% pendant 5 minutes. Après 4 rinçages à l'eau stérile, les graines sont mises à stratifier à 4 °C à l'obscurité pendant 96h. Les graines sont ensuite semées sur milieu MS solide :
- MS standard (MS 0.5X, MES 0.5g/L, sucrose 1%, agar 0.9%, pH 5.7) supplémenté en agent de sélection (Basta, Hygro ou Kana) dans des boîtes de Petri rondes puis fermées au parafilm et mises à pousser en enceinte de culture Percival dans le cas des sélections de plantes transgéniques.
- MS enrichi en sucrose (MS 0.5X, MES 0.5g/L, sucrose 1.5 %, agar 0.8%, pH 5.7) dans des boîtes de Petri carrées (12cm x 12 cm) puis fermées au parafilm et mises à pousser à la verticale en chambre de culture rotative dans le cas des phénotypage de la longueur de la racine primaire (voir section dédiée).

### Culture en terre

La culture en terre s'effectue après traitement du terreau avec un larvicide (Trigard) et passage des graines à -20°C pendant 48h. Les plantes sont cultivées en chambre de culture en conditions de jour long (16h de lumière à 23°C - 8h d'obscurité à 19°C).

# Phénotypage de la longueur de la racine primaire

La validation des "epiHIF" pour le QTL RLch4 requiert de coupler (épi)génotypage et phénotypage.
Le phénotypage s'effectue en chambre de culture rotative en conditions de jour lond. Après semis de 10 graines par boîte à intervalle régulier à 2 cm du rebord, les boîtes sont disposées à la verticale, dos à dos deux par deux, chaque paire de boîte étant séparée de la suivante par une paire de boîtes de Petri vides. Afin de réduire les effets environnementaux préalablement mis en évidence dans ce dispositif expérimental (thèse de Sandra Cortijo), le rang le plus externe du plateau rotatif comporte exclusivement des boîtes de Petri vides disposées selon la même organisation. La mesure des racines s'effectue après 10 jours de culture : les boîtes sont alors ouvertes, égouttées et scannées à 300dpi puis la longueur de la racine primaire est mesurée à l'aide du logiciel ImageJ (outil *segmented lines*). Afin de limiter les biais, les plantes présentant un délai de germination (pas de germination ou germination retardée à J+3), dont la racine a poussé dans l'agar, ainsi que toutes les plantes issues d'une boîte présentant une contamination bactériennes ou fongiques sont exclues des analyses et ne sont pas mesurées. Pour chaque lignée "epiHIF" ainsi que pour

les lignées contrôles (WT, *ddm1*, epiRIL à l'origine de l'HIF), la répartition des différents (épi)génotypes dans les différentes boîtes a été randomisée tout comme la répartition des boîtes dans la chambre de culture. Dans le cadre de la recherche d'une HIF robuste sur laquelle poursuivre la cartographie fine, le test de la ségrégation phénotype-épigénotype a été répété 4 à 6 fois.

## Epigénotypage McrBC-qPCR

La validation du QTL RLch4 ainsi que la recherche de recombinants dans l'intervalle de confiance de ce QTL a nécessité la détermination de l'état de méthylation au niveau de la DMR employée comme marqueur. Cet "épigénotypage" est réalisé par McrBC-qPCR. L'extraction d'ADN génomique des plantes est effectuée à partir de jeunes feuilles à l'aide d'un protocole d'extraction au CTAB. Après quantification Qubit, 100ng d'ADN sont ensuite digérés par l'endonucléase McrBC, qui clive l'ADN double-brin méthylé.
La réaction est effectuée dans les conditions suivantes : 100ng ADN, 0,2uL d'enzyme McrBC (10 U/uL), 5uL de tampon NEB2, 0.5uL GTP (100 mM), 0.5uL BSA (100X), H2O pour un volume final de réaction de 50uL. La réaction est effectuée sur la nuit à 37C. Pour chaque échantillon, une réaction témoin (sans enzyme) est réalisée en parallèle. Une qPCR ciblant le locus est ensuite effectuée sur une quantité égale d'ADN digéré et non-digéré.
Réaction qPCR (10uL volume final) : 5uL Light Cycler 480 SYBR Green I Master mix 2X (Roche 04887352001), 1uL H2O, 1uL ADN concentré à 2ng/uL, et 3uL de mix du couple d'amorces à 1uM pour chaque amorce.
Appareil qPCR : Light Cycler 480 (Roche). Programme de la réaction : dénaturation 10 min à 95°C et 45 cycles de PCR [dénaturation 10sec à 95°C suivi d'hybridation des amorces et élongation en une seule étape de 40sec à 60°C]. Une augmentation lente de latempérature (60 à 95°C) à la fin du programme permet de déterminer la température de fusion des produits amplifiés.
L'état de méthylation d'une séquence est inféré en comparant les valeurs de Ct obtenus dans les conditions "digéré" et "non digéré" : si la séquence d'intérêt n'est pas méthylée, elle ne sera pas clivée lors de la digestion McrBC et donc amplifiée de façon identique dans les deux conditions. Le $\Delta$Ct, calculé comme la différence entre Ct(ADNdigéré) et Ct(ADNnon-digéré), sera donc nul. A l'opposé, une méthylation de la séquence cible se traduira par une amplification plus tardive dans l'échantillon digéré, et donc un $\Delta$Ct supérieur à zéro. Le pourcentage de méthylation d'une séquence parmi le pool de molécules d'ADN, donné en pourcentage de molécules perdues après digestion, est donné par la formule suivante : % de méthylation = [1 - ($2^{(-\Delta ct)}$)] x 100
La qPCR est effectuée en parallèle sur deux régions contrôles, jamais (At5G13440, contrôle négatif) et systématiquement méthylée (région 3' de At5G36220, contrôle positif). Le

contrôle positif permet de constater la bonne digestion, le négatif, que le ΔCt observé pour l'échantillon reflète bien le niveau de méthylation de la séquence et non pas une différence dans les concentrations en ADN des deux conditions (ΔCt attendu =0). Cette technique permet par ailleurs l'identification d'individus "épi-hétérozygotes", chez lesquels une seule copie de la séquence est méthylée, et qui présentent donc 50% de méthylation de la séquence : en effet, il est à noter que lorsqu'on détecte 50% de méthylation pour une séquence avec cette méthode, cela ne signifie pas que 50% de la séquence est méthylée. Seul un traitement au bisulphite (qui convertit les C non méthylées en U), suivi d'un séquençage, permet de donner ce type d'information.
Au sein des DMR, les amorces sont dessinées puis sélectionnées selon les critères suivants : Tm de de 60°C, présence de sites CG dans la seed (afin de maximiser l'efficacité de détection d'une coupure), efficacité d'au moins 85%. Les amorces sont listées en annexe 1.

# Bibliographie

# Bibliographie


Acuna-Hidalgo, R., Veltman, J. et Hoischen, A. (2016), « New insights into the generation and role of de novo mutations in health and disease. », *Genome Biology* 17, p. 241.

Adli, M. (2018), « The CRISPR tool kit for genome editing and beyond », *Nature Communications* 9, p. 1911.

Aggarwala, V. et Voight, B. (2016), « An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. », *Nat Genet* 48, p. 349–355.

Agorio, A. et al. (2017), « An Arabidopsis Natural Epiallele Maintained by a Feed-Forward Silencing Loop between Histone and DNA », *PLOS Genetics* 13, p. 1–23.

Ahmed, I., Sarazin, A., Bowler, C., Colot, V. et Quesneville, H. (2011), « Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. », *Nucleic Acids Research* 39, p. 6919–6931.

Aller, E., Jagd, L., Kliebenstein, D. et Burow, M. (2018), « Comparison of the Relative Potential for Epigenetic and Genetic Variation To Contribute to Trait Stability. », *G3 (Bethesda)* 8, p. 1733–1746.

Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L. et Lombardo, A. (2016), « Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. », *Cell* 167, 219–232.e14.

Assaf, Z., Tilk, S., Park, J., Siegal, M. et Petrov, D. (2017), « Deep sequencing of natural and experimental populations of Drosophila melanogaster reveals biases in the spectrum of new mutations. », *Genome Res.* 27, p. 1988–2000.

Athma, P. et Peterson, T. (1991), « Ac induces homologous recombination at the maize P locus. », *Genetics* 128, p. 163–173.

Baer, C. F., Miyamoto, M. M. et Denver, D. R. (2007), « Mutation rate variation in multicellular eukaryotes : causes and consequences. », *Nature Reviews Genetics* 8, p. 619–631.

Baubec, T., Finke, A., Mittelsten Scheid, O. et Pecinka, A. (2014), « Meristem-specific expression of epigenetic regulators safeguards transposon silencing in Arabidopsis. », *EMBO Rep.* 15, p. 446–452.

Bazakos, C., Hanemian, M., Trontin, C., Jiménez-Gómez, J. et Loudet, O. (2017), « New Strategies and Tools in Quantitative Genetics : How to Go from the Phenotype to the Genotype. », *Annu Rev Plant Biol* 68, p. 435–455.

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K. et Weigel, D. (2011), « Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. », *Nature* 480, p. 245–249.

Behringer, M. G. et Hall, D. W. (2016), « Genome-Wide Estimates of Mutation Rates and Spectrum in Schizosaccharomyces pombe Indicate CpG Sites are Highly Mutagenic Despite the Absence of DNA Methylation », *G3 : Genes, Genomes, Genetics* 6, p. 149–160.

Bennetzen, J. L. et Kellogg, E. A. (1997), « Do Plants Have a One-Way Ticket to Genomic Obesity ? », *The Plant cell* 9, p. 1509–1514.

Benoit, M., Layat, E., Tourmente, S. et Probst, A. (2013), « Heterochromatin dynamics during developmental transitions in Arabidopsis - a focus on ribosomal DNA loci. », *Gene.* 526, p. 39–45.

Bernstein, D., Le Lay, J., Ruano, E. et Kaestner, K. (2015), « TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. », *J Clin Invest.* 125, p. 1998–2006.

Bewick, A. J., Ji, L. et al. (2016), « On the origin and evolutionary consequences of gene body DNA methylation. », *Proc Natl Acad Sci U S A* 113, p. 9111–9116.

Bewick, A. J. et Schmitz, R. J. (2017), « Gene body DNA methylation in plants. », *Current Opinion in Plant Biology* 36, p. 103–110.

Bird, A. (1980), « DNA methylation and the frequency of CpG in animal DNA. », *Nucleic Acids Res.* 8, p. 1499–1504.

— (1986), « CpG-rich islands and the function of DNA methylation. », *Nature* 321, p. 209–213.

Blake, R., Hess, S. et Nicholson-Tuell, J. (1992), « The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. », *J Mol Evol.* 34, p. 189–200.

Blevins, T., Wang, J., Pflieger, D., Pontvianne, F. et Pikaard, C. S. (2017), « Hybrid incompatibility caused by an epiallele », *Proc Natl Acad Sci U S A* 114, p. 3702–3707.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O. et Barillot, E. (2012), « Control-FREEC : a tool for assessing copy number and allelic content using next-generation sequencing data. », *Bioinformatics.* 28, p. 423–425.

Bond, D. et Baulcombe, D. (2015), « Epigenetic transitions leading to heritable, RNA-mediated de novo silencing in Arabidopsis thaliana. », *Proc Natl Acad Sci U S A.* 112, p. 917–922.

Bouyer, D., Kramdi, A., Kassam, M., Heese, M., Schnittger, A., Roudier, F. et Colot, V. (2017), « DNA methylation dynamics during early plant life. », *Genome Biology* 18, p. 179.

Brocken, D., Tark-Dame, M. et Dame, R. (2018), « dCas9 : A Versatile Tool for Epigenome Editing. », *Curr Issues Mol Biol.* 26, p. 15–32.

Bromham, L., Rambaut, A. et Harvey, P. (1996), « Determinants of rate variation in mammalian DNA sequence evolution. », *J Mol Evol.* 43, p. 610–621.

Brooks, S. C., Fischer, R. L., Huh, J. H. et Eichman, B. F. (2015), « 5-Methylcytosine Recognition by Arabidopsis thaliana DNA Glycosylases DEMETER and DML3 », *Biochemistry* 53, p. 2525–2532.

Campbell, C. D. et al. (2012), « Estimating the human mutation rate using autozygosity in a founder population. », *Nature Genetics* 44, p. 1277–1281.

Carbone, L. et al. (2009), « Evolutionary Breakpoints in the Gibbon Suggest Association between Cytosine Methylation and Karyotype Evolution », *PLOS Genetics* 5, p. 1–10.

Carvalho, C. et Lupski, J. (2016), « Mechanisms underlying structural variant formation in genomic disorders. », *Nature Reviews Genetics* 17, p. 224–238.

Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M. et Mittelsten Scheid, O. (2014), « How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation », *PLOS Genetics* 10, p. 1–12.

Čermák, T. et al. (2017), « A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants. », *The Plant Cell* 29, p. 1196–1217.

Chaikind, B., Kilambi, K., Gray, J. et M., O. (2012), « Targeted DNA methylation using an artificially bisected M.HhaI fused to zinc fingers. », *PLoS One.* 7, e44852.

Chaikind, B. et Ostermeier, M. (2014), « Directed evolution of improved zinc finger methyltransferases. », *PLoS One.* 9, e96931.

Chédin, F., Lieber, M. R. et Hsieh, C.-L. (2002), « The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a », *Proceedings of the National Academy of Sciences* 99, p. 16916–16921.

Chen, H., Kazemier, H., Groote, M. de, Ruiters, M., Xu, G. et MG., R. (2014), « Induced DNA demethylation by targeting Ten-Eleven Translocation 2 to the human ICAM-1 promoter. », *Nucleic Acids Res.* 42, p. 1563–1574.

Chen, R., Pettersson, U., Beard, C., Jackson-Grusby, L. et Jaenisch, R. (1998), « DNA hypomethylation leads to elevated mutation rates. », *Nature* 395, p. 89–93.

Chen, Z. et Riggs, A. (2011), « DNA methylation and demethylation in mammals. », *J Biol Chem.* 286, p. 18347–18353.

Chica, C., Louis, A., Roest Crollius, H., Colot, V. et Roudier, F. (2017), « Comparative epigenomics in the Brassicaceae reveals two evolutionarily conserved modes of PRC2-mediated gene regulation », *Genome Biology* 18, p. 207.

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J., Goldberg, R., Jacobsen, S. et Fischer, R. (2002), « DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis », *Cell* 110, p. 33–42.

Choudhury, S., Cui, Y., Lubecka, K., Stefanska, B. et Irudayaraj, J. (2016), « CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. », *Oncotarget.* 7, p. 46545–46556.

Cigan, A., Unger-Wallace, E. et Haug-Collet, K. (2005), « Transcriptional gene silencing as a tool for uncovering gene function in maize. », *Plant J.* 43, p. 929–940.

Cokus, S., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C., Pradhan, S., Nelson, S., Pellegrini, M. et Jacobsen, S. (2008), « Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. », *Nature* 452, p. 215–219.

Colomé-Tatché, M. et al. (2012), « Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. », *Proc Natl Acad Sci U S A* 109, p. 16240–16245.

Colot, V., Haedens, V. et Rossignol, J. (1998), « Extensive, nonrandom diversity of excision footprints generated by Ds-like transposon Ascot-1 suggests new parallels with V(D)J recombination. », *Mol Cell Biol.* 18, p. 4337–4346.

Consortium, 1. G. (2016), « 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. », *Cell* 166, p. 481–491.

Cordaux, R., Udit, S., Batzer, M. et Feschotte, C. (2006), « Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. », *Proc Natl Acad Sci U S A.* 103, p. 8101–8106.

Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Johannes, F. et Colot, V. (2014), « Genome-wide analysis of DNA methylation in Arabidopsis using MeDIP-chip. », *Methods Mol Biol* 1112, p. 125–149.

Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A. et al. (2014), « Mapping the epigenetic basis of complex traits. », *Science* 343, p. 1145–1148.

Crevillén, P., Yang, H., Cui, X., Greeff, C., Trick, M., Qiu, Q., Cao, X. et Dean, C. (2014), « Epigenetic reprogramming that prevents transgenerational inheritance of the vernalized state. », *Nature* 515, p. 587–590.

Cubas, P., Vincent, C. et Coen, E. (1999), « An epigenetic mutation responsible for natural variation in floral symmetry. », *Nature* 401, p. 157–161.

Cuerda-Gil, D. et Slotkin, R. K. (2016), « Non-canonical RNA-directed DNA methylation. », *Nature Plants* 2, p. 16163.

Cui, C., Gan, Y., Gu, L., Wilson, J., Liu, Z., Zhang, B. et Deng, D. (2015), « P16-specific DNA methylation by engineered zinc finger methyltransferase inactivates gene transcription and promotes cancer metastasis. », *Genome Biology* 16, p. 252.

Dadami, E., Dalakouras, A., Zwiebel, M., Krczal, G. et Wassenegger, M. (2014), « An endogene-resembling transgene is resistant to DNA methylation and systemic silencing. », *RNA Biol.* 11, p. 934–941.

Dalakouras, A., Moser, M., Zwiebel, M., Krczal, G., Hell, R. et Wassenegger, M. (2009), « A hairpin RNA construct residing in an intron efficiently triggered RNA-directed DNA methylation in tobacco. », *Plant J.* 60, p. 840–851.

DeBolt, S. (2010), « Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. », *Genome Biol Evol.* 2, p. 441–453.

Denver, D. R., Wilhelm, L. J., Howe, D. K., Gafner, K., Dolan, P. C. et Baer, C. F. (2012), « Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. », *Genome Biology and Evolution* 4, p. 513–522.

Dettman, J., Sztepanacz, J. et Kassen, R. (2016), « The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. », *BMC Genomics.* P. 17–27.

Devos, K., Brown, J. et JL., B. (2002), « Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. », *Genome Res.* 12, p. 1075–1079.

Dillon, M. M., Sung, W., Lynch, M. et Cooper, V. S. (2015), « The Rate and Molecular Spectrum of Spontaneous Mutations in the GC-Rich Multichromosome Genome of Burkholderia cenocepacia », *Genetics* 200, p. 935–946.

Drake, J., Charlesworth, B., Charlesworth, D. et Crow, J. (1998), « Rates of spontaneous mutation. », *Genetics* 148, p. 1667–1686.

Du, J., Johnson, L., Groth, M., Feng, S., Hale, C., Li, S., Vashisht, A., Wohlschlegel, J., Patel, D. et Jacobsen, S. (2014), « Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. », *Mol Cell.* 55, p. 495–504.

Du, J., Zhong, X. et al. (2012), « Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. », *Cell.* 151, p. 167–180.

Dubin, M. et al. (2015), « DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. », *eLife* 4, e05255.

Durand, S., Bouché, N., Perez-Strand, E., Loudet, O. et Camilleri, C. (2012), « Rapid establishment of genetic incompatibility through natural epigenetic variation. », *Curr Biol* 22, p. 326–331.

Duret, L. (2009), « Mutation patterns in the human genome : more variable than expected. », *PLoS Biology* 7, e1000028.

Duret, L. et Galtier, N. (2009), « Biased gene conversion and the evolution of mammalian genomic landscapes. », *Annual Review of Genomics and Human Genetics* 10, p. 285–311.

Ehrlich, M. et Wang, R. (1981), « 5-Methylcytosine in eukaryotic DNA. », *Science* 212, p. 1350–1357.

Ellegren, H. (2007), « Characteristics, causes and evolutionary consequences of male-biased mutation », *Proceedings of the Royal Society of London B : Biological Sciences* 274, p. 1–10.

Farlow, A., Long, H., Arnoux, S., Sung, W., Doak, T. G., Nordborg, M. et Lynch, M. (2015), « The Spontaneous Mutation Rate in the Fission Yeast Schizosaccharomyces pombe. », *Genetics* 201, p. 737–744.

Filion, G. et al. (2010), « Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. », *Cell* 143, p. 212–224.

Fransz, P., De Jong, J., Lysak, M., Castiglione, M. et I., S. (2002), « Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate », *Proc Natl Acad Sci U S A* 99, p. 14584–14589.

Frost, J. et al. (2018), « FACT complex is required for DNA demethylation at heterochromatin during reproduction in Arabidopsis. », *Proc Natl Acad Sci U S A.* 115, E4720–E4729.

Fryxell, K. et Moon, W. (2005), « CpG mutation rates in the human genome are highly dependent on local GC content. », *Mol Biol Evol.* 22, p. 650–658.

Fu, Y., Kawabe, A., Etcheverry, M., Ito, T., Toyoda, A., Fujiyama, A., Colot, V., Tarutani, Y. et Kakutani, T. (2013), « Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. », *The EMBO Journal* 32, p. 2407–2417.

Fultz, D., Choudury, S. G. et Slotkin, R. K. (2015), « Silencing of active transposable elements in plants. », *Current Opinion in Plant Biology* 27, p. 67–76.

Fultz, D. et Slotkin, R. K. (2017), « Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing. », *The Plant Cell* 29, p. 360–376.

Galagan, J. E. et Selker, E. U. (2004), « RIP : the evolutionary cost of genome defense », *Trends in Genetics* 20, p. 417–423.

Gallego-Bartolomé, J., Gardiner, J., Liu, W., Papikian, A., Ghoshal, B., Kuo, H., Zhao, J., Segal, D. et Jacobsen, S. (2018), « Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain. », *Proc Natl Acad Sci U S A.* 115, E2125–E2134.

Galonska, C. et al. (2018), « Genome-wide tracking of dCas9-methyltransferase footprints. », *Nat Commun.* 9, p. 597.

Garcia-Bloj, B. et al. (2016), « Waking up dormant tumor suppressor genes with zinc fingers, TALEs and the CRISPR/dCas9 system », *Oncotarget* 7, p. 60535–60554.

Garcia-Diaz, M. et Kunkel, T. (2006), « Mechanism of a genetic glissando : structural biology of indel mutations. », *Trends Biochem Sci.* 31, p. 206–214.

Gehring, M., Reik, W. et Henikoff, S. (2009), « DNA demethylation by DNA repair », *Trends in Genetics* 25, p. 82–90.

Gillet-Markowska, A., Richard, H., Fischer, G. et Lafontaine, I. (2015), « Ulysses : accurate detection of low-frequency structural variations in large insert-size sequencing libraries. », *Bioinformatics* 31, p. 801–808.

Gilly, A. et al. (2014), « TE-Tracker : systematic identification of transposition events through whole-genome resequencing. », *BMC Bioinformatics* 15, p. 377.

Graaf, A. van der, Wardenaar, R., Neumann, D. A., Taudt, A., Shaw, R. G., Jansen, R. C., Schmitz, R. J., Colomé-Tatché, M. et Johannes, F. (2015), « Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. », *Proc Natl Acad Sci U S A* 112, p. 6676–6681.

Gray, Y. (2000), « It takes two transposons to tango : transposable-element-mediated chromosomal rearrangements. », *Trends in Genetics* 16, p. 461–468.

Greer, E., Maures, T., Ucar, D., Hauswirth, A., Mancini, E., Lim, J., Benayoun, B., Shi, Y. et Brunet, A. (2011), « Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. », *Nature* 479, p. 365–371.

Gregory, D., Zhang, Y., Kobzik, L. et Fedulov, A. (2013), « Specific transcriptional enhancement of inducible nitric oxide synthase by targeted promoter demethylation. », *Epigenetics.* 8, p. 1205–1212.

Gu, W., Zhang, F. et Lupski, J. (2008), « Mechanisms for human genomic rearrangements. », *Pathogenetics.* 1, p. 4.

Hagmann, J. et al. (2015), « Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. », *PLoS Genetics* 11, e1004920.

Halligan, D. et Keightley, P. (2009), « Spontaneous Mutation Accumulation Studies in Evolutionary Genetics », *Annu. Rev. Ecol. Evol. Syst.* 40, p. 151–172.

Harshman, S., Young, N., Parthun, M. et Freitas, M. (2013), « H1 histones : current perspectives and challenges. », *Nucleic Acids Res* 41, p. 9593–9609.

He, L. et al. (2018), « A naturally occurring epiallele associates with leaf senescence and local climate adaptation in Arabidopsis accessions. », *Nat Commun.* 9, p. 460.

Heard, E. et Martienssen, R. (2014), « Transgenerational epigenetic inheritance : myths and mechanisms. », *Cell* 157, p. 95–109.

Henderson, I. et Jacobsen, S. (2008), « Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. », *Genes Dev.* 22, p. 1597–1606.

Hershberg, R. et Petrov, D. (2010), « Evidence That Mutation Is Universally Biased towards AT in Bacteria », *PLoS Genet* 6, e1001115.

Hildebrand, F., Meyer, A. et Eyre-Walker, A. (2010), « Evidence of selection upon genomic GC-content in bacteria. », *PLoS Genet.* 6, e1001107.

Hodgkinson, A. et Eyre-Walker, A. (2011), « Variation in the mutation rate across mammalian genomes », *Nature Reviews Genetics* 12, p. 756–766.

Hodgkinson, A., Ladoukakis, E. et Eyre-Walker, A. (2009), « Cryptic variation in the human mutation rate. », *PLoS Biology* 7, e1000027.

Holliday, R. et Grigg, G. (1993), « DNA methylation and mutation. », *Mutat Res* 285, p. 61–67.

Hsieh, T., Ibarra, C., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. et Zilberman, D. (2009), « Genome-wide demethylation of Arabidopsis endosperm. », *Science* 324, p. 1451–1454.

Hu, T. T. et al. (2011), « The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. », *Nature Genetics* 43, p. 476–481.

Huang, C., Burns, K. et Boeke, J. (2012), « Active transposition in genomes. », *Annu Rev Genet* 46, p. 651–675.

Huang, Y.-H., Su, J., Lei, Y., Brunetti, L., Gundry, M. C., Zhang, X., Jeong, M., Li, W. et Goodell, M. A. (2017), « DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A », *Genome Biology* 18, p. 176.

Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M. et Boeva, V. (2016), « SV-Bay : structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. », *Bioinformatics* 32, p. 984–992.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. et Paszkowski, J. (2011), « An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress », *Nature* 472, p. 115–119.

Iwasaki, M. et Paszkowski, J. (2014), « Identification of genes preventing transgenerational transmission of stress-induced epigenetic states. », *Proc Natl Acad Sci U S A.* 111, p. 8547–8552.

Jacobsen, S. et Meyerowitz, E. (1997), « Hypermethylated SUPERMAN epigenetic alleles in Arabidopsis », *Science* 277, p. 1100–1103.

Jacobsen, S., Sakai, H., Finnegan, E., Cao, X. et Meyerowitz, E. (2000), « Ectopic hypermethylation of flowerspecific genes in Arabidopsis », *Curr Biol* 10, p. 179–186.

Jeddeloh, J., Stokes, T. et Richards, E. (1999), « Maintenance of genomic methylation requires a SWI2/SNF2-like protein. », *Nature Genetics* 22, p. 94–97.

Jiang, C., Mithani, A., Belfield, E. J., Mott, R., Hurst, L. D. et Harberd, N. P. (2014), « Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. », *Genome Research* 24, p. 1821–1829.

Jiang, Y., Turinsky, A. L. et Brudno, M. (2015), « The missing indels : an estimate of indel variation in a human genome and analysis of factors that impede detection. », *Nucl. Ac. Res.* 43, p. 7217–7228.

Johannes, F., Colot, V. et Jansen, R. C. (2008), « Epigenome dynamics : a quantitative genetics perspective. », *Nature Reviews Genetics* 9, p. 883–890.

Johannes, F., Porcher, E. et al. (2009), « Assessing the impact of transgenerational epigenetic variation on complex traits. », *PLoS Genetics* 5, e1000530.

Johnson, L. et al. (2014), « SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. », *Nature* 507, p. 124–128.

Kakutani, T. (1997), « Genetic characterization of late-flowering traits induced by DNA hypomethylation mutation in Arabidopsis thaliana », *Plant J* 12, p. 1447–1451.

Kakutani, T., Jeddeloh, J., Flowers, S., Munakata, K. et Richards, E. (1996), « Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. », *Proc Natl Acad Sci U S A* 93, p. 12406–12411.

Kakutani, T., Jeddeloh, J. et Richards, E. (1995), « Characterization of an Arabidopsis thaliana DNA hypomethylation mutant. », *Nucleic Acids Res* 23, p. 130–137.

Kanazawa, A., Inaba, J., Shimura, H., Otagaki, S., Tsukahara, S., Matsuzawa, A., Kim, B., Goto, K. et Masuta, C. (2011), « Virus-mediated efficient induction of epigenetic modifications of endogenous genes with phenotypic changes in plants. », *Plant J.* 65, p. 156–168.

Kankel, M., Ramsey, D., Stokes, T., Flowers, S., Haag, J., Jeddeloh, J., Riddle, N., Verbsky, M. et Richards, E. (2003), « Arabidopsis MET1 cytosine methyltransferase mutants. », *Genetics* 163, p. 1109–1122.

Kawakatsu, T., Nery, J., Castanon, R. et Ecker, J. (2017), « Dynamic DNA methylation reconfiguration during seed development and germination. », *Genome Biology* 18, p. 171.

Kawakatsu, T., Huang, S.-S. C. et al. (2016), « Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. », *Cell* 166, p. 492–505.

Keightley, P. (1994), « The distribution of mutation effects on viability in Drosophila melanogaster. », *Genetics* 138, p. 1315–1322.

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., Davey, J. W. et Jiggins, C. D. (2015), « Estimation of the spontaneous mutation rate in Heliconius melpomene. », *Molecular Biology and Evolution* 32, p. 239–243.

Keith, N. et al. (2016), « High mutational rates of large-scale duplication and deletion in Daphnia pulex », *Genome Res.* 26, p. 60–69.

Kim, Y. G., Cha, J. et Chandrasegaran, S. (1996), « Hybrid restriction enzymes : zinc finger fusions to Fok I cleavage domain. », *Proc Natl Acad Sci U S A.* 93, p. 1156–1160.

Kimura, M. (1968), « Evolutionary rate at the molecular level. », *Nature* 217, p. 624–626.

Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W., Koornneef, M. et Kakutani, T. (2007), « Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. », *Plant J* 49, p. 38–45.

Kondrashov, A. S. (2003), « Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. », *Human Mutat* 21, p. 12–27.

Kondrashov, F. A. et Kondrashov, A. S. (2010), « Measurements of spontaneous rates of mutations in the recent past and the near future », *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 365, p. 1169–1176.

Konermann, S. et al. (2015), « Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. », *Nature* 517, p. 583–588.

Kong, A. et al. (2012), « Rate of de novo mutations and the importance of father's age to disease risk. », *Nature* 488, p. 471–475.

Kooke, R., Johannes, F., Wardenaar, R., Becker, F., Etcheverry, M., Colot, V., Vreugdenhil, D. et Keurentjes, J. J. B. (2015), « Epigenetic basis of morphological variation and phenotypic plasticity in Arabidopsis thaliana. », *The Plant Cell* 27, p. 337–348.

Kucukyildirim, S., Long, H., Sung, W., Miller, S., Doak, T. et M., L. (2016), « The Rate and Spectrum of Spontaneous Mutations in Mycobacterium smegmatis, a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair Pathway. », *G3 (Bethesda)* 6, p. 2157–2163.

Kumar, S. et Subramanian, S. (2002), « Mutation rates in mammalian genomes », *Proc Natl Acad Sci U S A* 99, p. 803–808.

Kungulovski, G., Nunna, S., Thomas, M., Zanger, U., Reinhardt, R. et Jeltsch, A. (2015), « Targeted epigenome editing of an endogenous locus with chromatin modifiers is not stably maintained. », *Epigenetics Chromatin.* 8, p. 12.

Lander, E. S., Linton, L., Birren, B. et Consortium, T. H. G. (2001), « Initial sequencing and analysis of the human genome. », *Nature* 409, p. 860–921.

Latzel, V., Allan, E., Bortolini Silveira, A., Colot, V., Fischer, M. et Bossdorf, O. (2013), « Epigenetic diversity increases the productivity and stability of plant populations. », *Nature Communications* 4, p. 2875.

Lee, H., Popodi, E., Tang, H. et Foster, P. (2012), « Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. », *Proc Natl Acad Sci U S A.* 109, E2774–83.

Lei, Y., Zhang, X., Su, J., Jeong, M., Gundry, M., Huang, Y., Zhou, Y., Li, W. et Goodell, M. (2017), « Targeted DNA methylation in vivo using an engineered dCas9-MQ1 fusion protein. », *Nat Commun.* 8, p. 16026.

Lerat, E., Capy, P. et Biemont, C. (2002), « Codon usage by transposable elements and their host genes in five species. », *J. Mol. Evol.* 54, p. 625–637.

Li, F., Papworth, M., Minczuk, M., Rohde, C., Zhang, Y., Ragozin, S. et Jeltsch, A. (2007), « Chimeric DNA methyltransferases target DNA methylation to specific DNA sequences and repress expression of target genes. », *Nucleic Acids Res.* 35, p. 100–112.

Li, J. et al. (2012), « Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. », *PLoS Genet.* 8, e1002692.

Lind, P. A. et Andersson, D. I. (2008), « Whole-genome mutational biases in bacteria. », *Proc. Natl. Acad. Sci. USA* 105, p. 17878–17883.

Lindberg, M., Hall, I. et Quinlan, A. (2015), « Population-based structural variation discovery with Hydra-Multi. », *Bioinformatics.* 31, p. 1286–1289.

Lindroth, A., Cao, X., Jackson, J., Zilberman, D., McCallum, C., Henikoff, S. et Jacobsen, S. (2001), « Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation », *Science* 292, p. 2077–2080.

Lippman, Z. et al. (2004), « Role of transposable elements in heterochromatin and epigenetic control. », *Nature* 430, p. 471–476.

Lister, R., O'Malley, R., Tonti-Filippini, J., Gregory, B., Berry, C., Millar, A. et Ecker, J. (2008), « Highly integrated single-base resolution maps of the epigenome in Arabidopsis. », *Cell* 133, p. 523–536.

Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M. et Weigel, D. (2016), « Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. », *Genome Research* 26, p. 1057–1068.

Liu, X., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R. et Jaenisch, R. (2016), « Editing DNA Methylation in the Mammalian Genome. », *Cell* 167, 233–247.e17.

Lo, C., Choudhury, S., Irudayaraj, J. et Zhou, F. (2017), « Epigenetic Editing of Ascl1 Gene in Neural Stem Cells by Optogenetics. », *Sci Rep.* 7, p. 42047.

Long, H., Kucukyildirim, S., Sung, W., Williams, E., Lee, H., Ackerman, M., Doak, T., Tang, H. et Lynch, M. (2015), « Background Mutational Features of the Radiation-Resistant Bacterium Deinococcus radiodurans. », *Mol Biol Evol.* 32, p. 2383–2392.

Lowder, L., Zhang, D., Baltes, N., Paul, J., Tang, X., Zheng, X., Voytas, D., Hsieh, T., Zhang, Y. et Qi, Y. (2015), « A CRISPR/Cas9 Toolbox for Multiplexed Plant Genome Editing and Transcriptional Regulation », *Plant Physiol.* 169, p. 971–985.

Luff, B., Pawlowski, L. et J., B. (1999), « An epigenetic mutation responsible for natural variation in floral symmetry », *Mol Cell* 3, p. 505–511.

Luger, K., Mader, A., Richmond, R., Sargent, D. et Richmond, T. (1997), « Crystal structure of the nucleosome core particle at 2.8 A resolution », *Nature* 389, p. 251–260.

Luo, C., Hajkova, P. et JR., E. (2018), « Dynamic DNA methylation : In the right place at the right time », *Science* 361, p. 1336–1340.

Lupski, J. (2004), « Hotspots of homologous recombination in the human genome : not all homologous sequences are equal. », *Genome Biology* 5, p. 242.

Lynch, M. (2010a), « Evolution of the mutation rate. », *Trends in Genetics* 26, p. 345–352.

Lynch, M. (2011), « The lower bound to the evolution of mutation rates. », *Genome Biol Evol.* 3, p. 1107–1118.

Lynch, M. et Walsh, J. (1998), *in* : *Genetics and Analysis of Quantitative Traits.* Sunderland, MA : Sinauer Associates, p. 980.

Lynch, M. (2010b), « Rate, molecular spectrum, and consequences of human mutation », *Proc. Natl. Acad. Sci. USA* 107, p. 961–968.

— (2016), « Mutation and Human Exceptionalism : Our Future Genetic Load », *Genetics* 202, p. 869–875.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K. et Foster, P. L. (2016), « Genetic drift, selection and the evolution of the mutation rate. », *Nature Reviews Genetics* 17, p. 704–714.

Ma, J. et Bennetzen, J. (2006), « Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. », *Proc Natl Acad Sci U S A.* 103, p. 383–388.

Ma, X. et al. (2015), « A Robust CRISPR/Cas9 System for Convenient, High-Efficiency Multiplex Genome Editing in Monocot and Dicot Plants. », *Mol Plant.* 8, p. 1274–1284.

Maeder, M. et al. (2013), « Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. », *Nat Biotechnol.* 31, p. 1137–1142.

Makova, K. D. et Hardison, R. C. (2015), « The effects of chromatin organization on variation in mutation rates in the genome. », *Nature Reviews Genetics* 16, p. 213–223.

Maloisel, L. et Rossignol, J. (1998), « Suppression of crossing-over by DNA methylation in Ascobolus. », *Genes Dev.* 12, p. 1381–1389.

Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A., King, G., Giovannoni, J. et Seymour, G. (2006), « A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. », *Nat Genet.* 38, p. 948–952.

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., Pitrat, M., Dogimont, C. et Bendahmane, A. (2009), « A transposon-induced epigenetic change leads to sex determination in melon », *Nature* 461, p. 1135–1138.

Martin, A. et Palumbi, S. (1993), « Body size, metabolic rate, generation time, and the molecular clock. », *Proc Natl Acad Sci U S A.* 90, p. 4087–4091.

Matzke, M. et Mosher, R. (2014), « RNA-directed DNA methylation : an epigenetic pathway of increasing complexity », *Nature Review Genetics* 15, p. 394–408.

Maumus, F. et Quesneville, H. (2014), « Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. », *Nature Communications* 5, p. 4104.

McDonald, J., Celik, H., Rois, L., Fishberger, G., Fowler, T., Rees, R., Kramer, A., Martens, A., Edwards, J. et Challen, G. (2016), « Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation. », *Biol Open.* 5, p. 866–874.

Meister, G., S, C. et Ostermeier, M. (2010), « Heterodimeric DNA methyltransferases as a platform for creating designer zinc finger methyltransferases for targeted DNA methylation in cells. », *Nucleic Acids Res.* 38, p. 1749–1759.

Melquist, S. et Bender, J. (2003), « Transcription from an upstream promoter controls methylation signaling from an inverted repeat of endogenous genes in Arabidopsis. », *Genes Dev.* 17, p. 2036–2047.

— (2004), « An internal rearrangement in an Arabidopsis inverted repeat locus impairs DNA methylation triggered by the locus. », *Genetics* 166, p. 437–448.

Messer, P. (2009), « Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. », *Genetics* 182, p. 1219–1232.

Mette, M., Aufsatz, W., Winden, J. van der, Matzke, M. et Matzke, A. (2000), « Transcriptional silencing and promoter methylation triggered by double-stranded RNA. », *EMBO J.* 19, p. 5194–5201.

Meunier, J. et Duret, L. (2004), « Recombination drives the evolution of GC-content in the human genome. », *Molecular Biology and Evolution* 21, p. 984–990.

Michael, T., Jupe, F., Bemm, F., Motley, S., Sandoval, J., Lanz, C., Loudet, O., Weigel, D. et JR., E. (2018), « High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. », *Nat Commun.* 9, p. 541.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J. et Mathieu, O. (2009), « Selective epigenetic control of retrotransposition in Arabidopsis. », *Nature* 461, p. 427–430.

Miura, K., Agetsuma, M., Kitano, H., Yoshimura, A., Matsuoka, M., Jacobsen, S. et Ashikari, M. (2009), « A metastable DWARF1 epigenetic mutant affecting plant stature in rice. », *Proc Natl Acad Sci U S A.* 106, p. 11218–11223.

Mlambo, T., Nitsch, S., Hildenbeutel, M., Romito, M., Müller, M., Bossen, C., Diederichs, S., Cornu, T., Cathomen, T. et Mussolino, C. (2018), « Designer epigenome modifiers enable robust and sustained gene silencing in clinically relevant human cells. », *Nucleic Acids Res.* 46, p. 4456–4468.

Mok, Y. G., Uzawa, R., Lee, J., Weiner, G. M., Eichman, B. F., Fischer, R. L. et Huh, J. H. (2010), « Domain structure of the DEMETER 5-methylcytosine DNA glycosylase », *Proc Natl Acad Sci U S A of the United States of America* 107, p. 19225–19230.

Montgomery, S. B. et al. (2013), « The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. », *Genome Research* 23, p. 749–761.

Mooers, A. et Harvey, P. (1994), « Metabolic rate, generation time, and the rate of molecular evolution in birds. », *Mol Phylogenet Evol.* 3, p. 344–350.

Morita, S. et al. (2016), « Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. », *Nat Biotechnol.* 34, p. 1060–1065.

Mugal, C., Grünberg, H.-H. von et Peifer, M. (2009), « Transcription-induced mutational strand bias and its effect on substitution rates in human genes. », *Molecular Biology and Evolution* 26, p. 131–142.

Mugal, C., Arndt, P., Holm, L. et Ellegren, H. (2015), « Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. », *G3 (Bethesda)* 5, p. 441–447.

Mukai, T. (1964), « The Genetic Structure of Natural Populations of Drosophila Melanogaster. I. Spontaneous Mutation Rate of Polygenes Controlling Viability », *Genetics* 50, p. 1–19.

Muller, H. (1928), « The Measurement of Gene Mutation Rate in Drosophila, Its High Variability, and Its Dependence upon Temperature. », *Genetics* 13, p. 279–357.

Nachman, M. et Crowell, S. (2000), « Estimate of the mutation rate per nucleotide in humans. », *Genetics* 156, p. 297–304.

Ness, R. W., Morgan, A. D., Vasanthakrishnan, R. B., Colegrave, N. et Keightley, P. D. (2015), « Extensive de novo mutation rate variation between individuals and across the genome of Chlamydomonas reinhardtii. », *Genome Research* 25, p. 1739–1749.

Newman, S., Hermetz, K., Weckselblatt, B. et Rudd, M. (2015), « Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. », *Am J Hum Genet.* 96, p. 208–220.

Nunn, G. et Stanley, S. (1998), « Body size effects and rates of cytochrome b evolution in tube-nosed seabirds. », *Mol Biol Evol.* 15, p. 1360–1371.

Nunna, S., Reinhardt, R., Ragozin, S. et Jeltsch, A. (2014), « Targeted methylation of the epithelial cell adhesion molecule (EpCAM) promoter to silence its expression in ovarian cancer cells. », *PLoS One.* 9, e87703.

Oey, H., Isbel, L., Hickey, P., Ebaid, B. et Whitelaw, E. (2015), « Genetic and epigenetic variation among inbred mouse littermates : identification of inter-individual differentially methylated regions. », *Epigenetics Chromatin.* 8, p. 54.

Ohta, T. (1993), « An examination of the generation-time effect on molecular evolution. », *Proc Natl Acad Sci U S A.* 90, p. 10676–10680.

Okada, M., Kanamori, M., Someya, K., Nakatsukasa, H. et Yoshimura, A. (2017), « Stabilization of Foxp3 expression by CRISPR-dCas9-based epigenome editing in mouse primary T cells. », *Epigenetics Chromatin* 10, p. 24.

Olins, A. et Olins, D. (1974), « Spheroid chromatin units (v bodies) », *Science* 183, p. 330–332.

Olins, D. et Olins, A. (2003), « Chromatin history : our view from the bridge. », *Nat Rev Mol Cell Biol.* 4, p. 809–814.

Ong-Abdullah, M. et al. (2015), « Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. », *Nature* 525, p. 533–537.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D. et Lynch, M. (2010), « The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. », *Science* 327, p. 92–94.

Panda, K., Ji, L., Neumann, D. A., Daron, J., Schmitz, R. J. et Slotkin, R. K. (2016), « Full-length autonomous transposable elements are preferentially targeted by expression - dependent forms of RNA - directed DNA methylation. », *Genome Biology* 17, p. 170.

Park, J. et al. (2017), « Control of DEMETER DNA demethylase gene transcription in male and female gamete companion cells in Arabidopsis thaliana. », *Proc Natl Acad Sci U S A.* 114, p. 2078–2083.

Parrilla-Doblas, J., Ariza, R. et Roldán-Arjona, T. (2017), « Targeted DNA demethylation in human cells by fusion of a plant 5-methylcytosine DNA glycosylase to a sequence-specific DNA binding domain. », *Epigenetics* 12, p. 296–303.

Passarge, E. (1979), « Emil Heitz and the concept of heterochromatin : longitudinal chromosome differentiation was recognized fifty years ago. », *Am J Hum Genet.* 31, p. 106–115.

Paweletz, N. (2001), « Walther Flemming : pioneer of mitosis research. », *Nat Rev Mol Cell Biol.* 2, p. 72–75.

Perez-Pinera, P., Kocak, D. et al. (2013), « RNA-guided gene activation by CRISPR-Cas9-based transcription factors », *Nat Methods.* 10, p. 973–976.

Perez-Pinera, P., Ousterout, D. et Gersbach, C. (2012), « Advances in targeted genome editing. », *Curr Opin Chem Biol* 16, p. 268–277.

Pflueger, C., Tan, D., Swain, T., Nguyen, T., Pflueger, J., Nefzger, C., Polo, J., Ford, E. et Lister, R. (2018), « A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. », *Genome Research* 28, p. 1193–1206.

Pradeepa, M., Grimes, G., Kumar, Y., Olley, G., Taylor, G., Schneider, R. et Bickmore, W. (2016), « Histone H3 globular domain acetylation identifies a new class of enhancers. », *Nat Genet.* 48, p. 681–686.

Qi, W., Zhu, T., Tian, Z., Li, C., Zhang, W. et Song, R. (2016), « High-efficiency CRISPR/Cas9 multiplex gene editing using the glycine tRNA-processing system-based strategy in maize. », *BMC Biotechnol.* 16, p. 58.

Quadrana, L., Almeida, J. et al. (2014), « Natural occurring epialleles determine vitamin E accumulation in tomato fruits. », *Nature Communications* 5, p. 3027.

Quadrana, L., Bortolini Silveira, A., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A. et Colot, V. (2016), « The Arabidopsis thaliana mobilome and its impact at the species level. », *eLife* 5.

Quadrana, L. et Colot, V. (2016), « Plant Transgenerational Epigenetics. », *Annual Review of Genetics* 50, p. 467–491.

Quadrana, L., Etcheverry, M. et al. (2018), « Transposon accumulation lines uncover histone H2A.Z-driven integration bias towards environmentally responsive genes », *bioRxiv*.

Rao, S. et al. (2014), « A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping », *Cell* 159, p. 1665–1680.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A., Benes, V. et Korbel, J. (2012), « DELLY : structural variant discovery by integrated paired-end and split-read analysis. », *Bioinformatics*. 28, p. i333–i339.

Reinders, J., Wulff, B., Mirouze, M., Mari-Ordonez, A., Dapp, M., Rozhon, W., Bucher, E., Theiler, G. et Paszkowski, J. (2009), « Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. », *Genes Dev* 23, p. 939–950.

Richards, E. (2006), « Inherited epigenetic variation : revisiting soft inheritance », *Nat. Rev. Genet.* 7, p. 395–401.

Richmond, T., Finch, J., Rushton, B., Rhodes, D. et Klug, A. (1984), « Structure of the nucleosome core particle at 7 A resolution. », *Nature* 311, p. 532–537.

Rivenbark, A., Stolzenburg, S., Beltran, A., Yuan, X., Rots, M., Strahl, B. et Blancafort, P. (2012), « Epigenetic reprogramming of cancer cells via targeted DNA methylation. », *Epigenetics* 7, p. 350–360.

Roudier, F., Teixeira, F. et Colot, V. (2009), « Chromatin indexing in Arabidopsis : an epigenomic tale of tails and more. », *Trends Genet* 25, p. 511–517.

Roudier, F., Ahmed, I. et al. (2011), « Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. », *The EMBO Journal* 30, p. 1928–1938.

Roux, F., Colomé-Tatché, M., Edelist, C., Wardenaar, R., Guerche, P., Hospital, F., Colot, V., Jansen, R. C. et Johannes, F. (2011), « Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. », *Genetics* 188, p. 1015–1017.

Sanchez, D., Gaubert, H., Drost, H., Zabet, N. et Paszkowski, J. (2017), « High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. », *Nat Commun.* 8, p. 1283.

Sanjuán, R. et Domingo-Calap, P. (2016), « Mechanisms of viral mutation », *Cell Mol Life Sci.* 23, p. 4433–4448.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. et Belshaw, R. (2010), « Viral Mutation Rates », *Journal of Virology* 84, p. 9733–9748.

Sasaki, M., Lange, J. et Keeney, S. (2010), « Genome destabilization by homologous recombination in the germ line », *Nat Rev Mol Cell Biol* 11, p. 182–195.

Saze, H. et Kakutani, T. (2007), « Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1 », *EMBO J* 26, p. 3641–3652.

Schmitz, R. J. et Ecker, J. R. (2012), « Epigenetic and epigenomic variation in Arabidopsis thaliana. », *Trends in Plant Science* 17, p. 149–154.

Schmitz, R., He, Y. et al. (2013), « Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population », *Genome Res* 23, p. 1663–1674.

Schmitz, R., Schultz, M., Lewsey, M., O'Malley, R., Urich, M., Libiger, O., Schork, N. et Ecker, J. (2011), « Transgenerational epigenetic instability is a source of novel methylation variants », *Science* 334, p. 369–373.

Schmitz, R., Schultz, M., Urich, M. et al. (2013), « Patterns of population epigenomic diversity », *Nature* 495, p. 193–198.

Schoft, V. et al. (2011), « Function of the DEMETER DNA glycosylase in the Arabidopsis thaliana male gametophyte. », *Proc Natl Acad Sci U S A.* 108, p. 8042–8047.

Schrider, D., Houle, D., Lynch, M. et Hahn, M. (2013), « Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. », *Genetics* 194, p. 937–954.

Schuster-Böckler, B. et Lehner, B. (2012), « Chromatin organization is a major influence on regional mutation rates in human cancer cells. », *Nature* 488, p. 504–507.

Secco, D., Wang, C., Shou, H., Schultz, M., Chiarenza, S., Nussaume, L., Ecker, J., Whelan, J. et Lister, R. (2015), « Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. », *eLife* 4.

Ségurel, L., Wyman, M. J. et Przeworski, M. (2014), « Determinants of mutation rate variation in the human germline. », *Annual Review of Genomics and Human Genetics* 15, p. 47–70.

Sehn, J. K. (2015), « Chapter 9 - Insertions and Deletions (Indels) », *in* : *Clinical Genomics*, sous la dir. de S. Kulkarni et J. Pfeifer, Boston : Academic Press, p. 129–150.

Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S., Bastolla, U. et Gutierrez, C. (2014), « The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States. », *The Plant Cell* 26, p. 2351–2366.

Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K. et Korte, A. (2017), « AraPheno : a public database for Arabidopsis thaliana phenotypes. », *Nucleic Acids Research* 45, p. D1054–D1059.

Shalev, G. et Levy, A. (1997), « The maize transposable element Ac induces recombination between the donor site and an homologous ectopic sequence. », *Genetics* 147, p. 1143–1151.

Shaw, F., Geyer, C. et Shaw, R. (2002), « A comprehensive model of mutations affecting fitness and inferences for Arabidopsis thaliana. », *Evolution* 56, p. 453–463.

Siddique, A., Nunna, S., Rajavelu, A., Zhang, Y., Jurkowska, R., Reinhardt, R., Rots, M., Ragozin, S., Jurkowski, T. et Jeltsch, A. (2013), « Targeted methylation and gene silencing of VEGF-A in human cells by using a designed Dnmt3a-Dnmt3L single-chain fusion protein with increased DNA methylation activity. », *J Mol Biol.* 425, p. 479–491.

Sijen, T., Vijn, I., Rebocho, A., Blokland, R. van, Roelofs, D., Mol, J. et Kooter, J. (2001), « Transcriptional and posttranscriptional gene silencing are mechanistically related. », *Curr Biol.* 11, p. 436–440.

Silveira, A. B., Trontin, C., Cortijo, S., Barau, J., Del Bem, L. E. V., Loudet, O., Colot, V. et Vincentz, M. (2013), « Extensive natural epigenetic variation at a de novo originated gene. », *PLoS Genetics* 9, e1003437.

Sima, J. et Gilbert, D. (2014), « Complex correlations : replication timing and mutational landscapes during cancer and genome evolution. », *Curr Opin Genet Dev.* 25, p. 93–100.

Soppe, W., Jacobsen, S., AlonsoBlanco, C., Jackson, J., Kakutani, T., Koornneef, M. et Peeters, A. (2000), « The late flowering phenotype of fwa mutants is caused by gainoffunction epigenetic alleles of a homeodomain gene. », *Mol Cell* 6, p. 791–802.

Springer, N. M. et Schmitz, R. J. (2017), « Exploiting induced and natural epigenetic variation for crop improvement. », *Nature Reviews Genetics.*

Stankiewicz, P. et Lupski, J. (2002), « Genome architecture, rearrangements and genomic disorders. », *Trends in Genetics* 18, p. 74–82.

Stephens, P. et al. (2011), « Massive genomic rearrangement acquired in a single catastrophic event during cancer development. », *Cell* 144, p. 27–40.

Stepper, P., Kungulovski, G., Jurkowska, R., Chandra, T., Krueger, F., Reinhardt, R., Reik, W., Jeltsch, A. et Jurkowski, T. (2017), « Efficient targeted DNA methylation with chimeric dCas9-Dnmt3a-Dnmt3L methyltransferase. », *Nucleic Acids Res.* 45, p. 1703–1713.

Stolzenburg, S., Beltran, A., Swift-Scanlan, T., Rivenbark, A., Rashwan, R. et Blancafort, P. (2015), « Stable oncogenic silencing in vivo by programmable and targeted de novo DNA methylation in breast cancer. », *Oncogene.* 34, p. 5427–5435.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D. et Jacobsen, S. (2014), « Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis », *Nature Structural and Molecular Biology* 21, p. 64–72.

Sung, W., Ackerman, M. S., Dillon, M. M., Platt, T. G., Fuqua, C., Cooper, V. S. et Lynch, M. (2016), « Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. », *G3 (Bethesda, Md.)* 6, p. 2583–2591.

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. et Lynch, M. (2012), « Drift-barrier hypothesis and mutation-rate evolution. », *Proc Natl Acad Sci U S A* 109, p. 18488–18492.

Sung, W., Tucker, A. E., Doak, T. G., Choi, E., Thomas, W. K. et Lynch, M. (2012), « Extraordinary genome stability in the ciliate Paramecium tetraurelia », *Proc Natl Acad Sci U S A.* 109, p. 19339–19344.

Sunitha, S., Shivaprasad, P. V., Sujata, K. et Veluthambi, K. (2012), « High Frequency of T-DNA Deletions in Transgenic Plants Transformed with Intron-Containing Hairpin RNA Genes », *Plant Molecular Biology Reporter* 30, p. 158–167.

Supek, F. et Lehner, B. (2015), « Differential DNA mismatch repair underlies mutation rate variation across the human genome. », *Nature* 521, p. 81–84.

Supek, F., Lehner, B., Hajkova, P. et Warnecke, T. (2014), « Hydroxymethylated cytosines are associated with elevated C to G transversion rates. », *PLoS Genet.* 10, e1004585.

Suzuki, M. et Bird, A. (2008), « DNA methylation landscapes : provocative insights from epigenomics. », *Nat Rev Genet* 9, p. 465–476.

Tahiliani, M. et al. (2009), « Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. », *Science.* 324, p. 930–935.

Tan, B., Guan, J., Ding, S., Wu, S., Saunders, J., Koch, K. et McCarty, D. (2017), « Structure and Origin of the White Cap Locus and Its Role in Evolution of Grain Color in Maize », *Genetics* 206, p. 135–150.

Tanenbaum, M., Gilbert, L., Qi, L., Weissman, J. et Vale, R. (2014), « A protein-tagging system for signal amplification in gene expression and fluorescence imaging. », *Cell* 159, p. 635–646.

Tang, K., Lang, Z., Zhang, H. et Zhu, J. (2016), « The DNA demethylase ROS1 targets genomic regions with distinct chromatin modifications. », *Nat Plants.* 2, p. 16169.

Tattini, L., D'Aurizio, R. et Magi, A. (2015), « Detection of Genomic Structural Variants from Next-Generation Sequencing Data. », *Frontiers in Bioengineering and Biotechnology* 3, p. 92.

Teixeira, F. K. et Colot, V. (2009), « Gene body DNA methylation in plants : a means to an end or an end to a means ? », *The EMBO Journal* 28, p. 997–998.

Teixeira, F. et Colot, V. (2010), « Repeat elements and the Arabidopsis DNA methylation landscape. », *Heredity (Edinb)* 105, p. 14–23.

Thakore, P., D'Ippolito, A., Song, L., Safi, A., Shivakumar, N., Kabadi, A., Reddy, T., Crawford, G. et Gersbach, C. (2015), « Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. », *Nat Methods.* 12, p. 1143–1149.

Thomas, G. W. et al. (2018), « Reproductive longevity predicts mutation rates in primates », *BioRxiv*.

Thorvaldsdóttir, H., Robinson, J. T. et Mesirov, J. P. (2013), « Integrative Genomics Viewer (IGV) : high-performance genomics data visualization and exploration », *Briefings in Bioinformatics* 14, p. 178–192.

Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J. et Yagi, T. (2015), « Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. », *Genome Res.* 25, p. 1125–1134.

Van Dooren, T., Silveira, A., Gilbaut, E., Jimenez-Gomez, J. M., Martin, A., Bach, L., Tisne, S., Quadrana, L., Loudet, O. et Colot, V. (2018), « Mild drought induces phenotypic and DNA methylation plasticity but no transgenerational effects in Arabidopsis », *bioRxiv*.

Vazquez-Vilar, M., Bernabé-Orts, J., Fernandez-Del-Carmen, A., Ziarsolo, P., Blanca, J., Granell, A. et D, O. (2016), « A modular toolbox for gRNA-Cas9 genome engineering in plants based on the GoldenBraid standard. », *Plant Methods.* 12, p. 10.

Venn, O., Turner, I., Mathieson, I., Groot, N. de, Bontrop, R. et McVean, G. (2014), « Strong male bias drives germline mutation in chimpanzees. », *Science* 344, p. 1272–1275.

Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M. et Zoldoš, V. (2016), « Repurposing the CRISPR-Cas9 system for targeted DNA methylation. », *Nucleic Acids Res.* 44, p. 5615–5628.

Vongs, A., Kakutani, T., Martienssen, R. et Richards, E. (1993), « Arabidopsis thaliana DNA methylation mutants », *Science* 260, p. 1926–1928.

Vu, G., Cao, H., Reiss, B. et Schubert, I. (2017), « Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. », *New Phytol.* 214, p. 1712–1721.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C. et Weigel, D. (2015), « Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. », *Genome Research* 25, p. 246–256.

Wang, D., Yu, C., Zuo, T., Zhang, J., Weber, D. et Peterson, T. (2015), « Alternative Transposition Generates New Chimeric Genes and Segmental Duplications at the Maize p1 Locus », *Genetics* 201, p. 925–935.

Wang, W., Akhunova, A., Chao, S. et Akhunov, E. (2016), « Optimizing multiplex CRISPR/Cas9-based genome editing for wheat », *bioRxiv*.

Wang, Z., Wang, S., Li, D., Zhang, Q., Li, L., Zhong, C., Liu, Y. et Huang, H. (2018), « Optimized paired-sgRNA/Cas9 cloning and expression cassette triggers high-efficiency multiplex genome editing in kiwifruit. », *Plant Biotechnol J* 16, p. 1424–1433.

Weckselblatt, B. et Rudd, M. (2015), « Human Structural Variation : Mechanisms of Chromosome Rearrangements. », *Trends in Genetics* 31, p. 587–599.

Wessler, S. R. (2006), « Transposable elements and the evolution of eukaryotic genomes », *Proc Natl Acad Sci U S A* 103, p. 17600–17601.

Wibowo, A. et al. (2016), « Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. », *eLife* 5.

Wicker, T. et al. (2016), « DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. », *Nature Communications* 7, p. 12790.

Williams, B., Pignatta, D., Henikoff, S. et Gehring, M. (2015), « Methylation-sensitive expression of a DNA demethylase gene serves as an epigenetic rheostat », *PLoS Genet.* 11, e1005142.

Willis, N. A., Rass, E. et Scully, R. (2015), « Deciphering the Code of the Cancer Genome : Mechanisms of Chromosome Rearrangement », *Trends in Cancer* 1, p. 217–230.

Wolfe, K., Sharp, P. et Li, W. (1989), « Mutation rates differ among regions of the mammalian genome. », *Nature* 337, p. 283–285.

Woo, H., Pontes, O., Pikaard, C. et Richards, E. (2007), « VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization », *Genes Dev* 21, p. 267–277.

Xie, K., Minkenberg, B. et Yang, Y. (2015), « Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. », *Proc Natl Acad Sci U S A.* 112, p. 3570–3575.

Xie, W. et al. (2013), « Epigenomic analysis of multilineage differentiation of human embryonic stem cells. », *Cell* 153, p. 1134–1148.

Xiong, T., Meister, G., Workman, R., Kato, N., Spellberg, M., Turker, F., Timp, W., Ostermeier, M. et Novina, C. (2015), « Targeted DNA methylation in human cells using engineered dCas9-methyltransferases. », *Sci Rep.* 7, p. 6732.

Xu, X., Tan, X. et al. (2018), « High-fidelity CRISPR/Cas9-based gene-specific hydroxymethylation rescues gene expression and attenuates renal fibrosis. », *Nat Commun.* 9, p. 3509.

Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G. et Hu, R. (2016), « A CRISPR-based approach for targeted DNA demethylation. », *Cell Discov.* 2, p. 16009.

Yamazaki, T. et al. (2017), « Targeted DNA methylation in pericentromeres with genome editing-based artificial DNA methyltransferase. », *PLoS One* 12, e0177764.

Yan, W., Chen, D. et Kaufmann, K. (2016), « Efficient multiplex mutagenesis by RNA-guided Cas9 and its use in the characterization of regulatory elements in the AGAMOUS gene. », *Plant Methods.* 12, p. 23.

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D. et Tian, D. (2015), « Parent-progeny sequencing indicates higher mutation rates in heterozygotes. », *Nature* 523, p. 463–467.

Yelina, N. E., Lambing, C., Hardcastle, T. J., Zhao, X., Santos, B. et Henderson, I. R. (2015), « DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. », *Genes & development* 29, p. 2183–2202.

Zapata, L. et al. (2016), « Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. », *Proc Natl Acad Sci U S A* 113, E4052–E4060.

Zemach, A., Kim, M. Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S. L. et Zilberman, D. (2013), « The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. », *Cell* 153, p. 193–205.

Zhang, H., Lang, Z. et Zhu, J. (2018), « Dynamics and function of DNA methylation in plants », *Nat Rev Mol Cell Biol.* 19, p. 489–506.

Zhang, J. et Peterson, T. (2004), « Transposition of reversed Ac element ends generates chromosome rearrangements in maize. », *Genetics* 167, p. 1929–1937.

Zhang, J., Zhang, F. et Peterson, T. (2006), « Transposition of reversed Ac element ends generates novel chimeric genes in maize », *PLoS Genet.* 2, e164.

Zhang, X., Sun, J., Cao, X. et Song, X. (2015), « Epigenetic Mutation of RAV6 Affects Leaf Angle and Seed Size in Rice. », *Plant Physiol.* 169, p. 2118–2128.

Zhang, X., Yazaki, J. et al. (2006), « Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. », *Cell* 126, p. 1189–1201.

Zhang, Y.-Y., Fischer, M., Colot, V. et Bossdorf, O. (2013), « Epigenetic variation creates potential for evolution of plant phenotypic plasticity. », *New Phytologist* 197, p. 314–322.

Zhang, Y. et al. (2018), « Dynamic epigenomic landscapes during early lineage specification in mouse embryos. », *Nat Genet.* 50, p. 96–105.

Zheng, C. et al. (2014), « Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. », *Cell Rep.* 9, p. 1228–1234.

Zhu, J. (2009), « Active DNA demethylation mediated by DNA glycosylases. », *Annu Rev Genet* 43, p. 143–166.

Zhu, Y., Sherlock, G. et D.A., P. (2017), « Extremely Rare Polymorphisms in Saccharomyces cerevisiae Allow Inference of the Mutational Spectrum. », *PLoS Genetics* 13, e1006455.

Zhu, Y. O., Siegal, M. L., Hall, D. W. et Petrov, D. A. (2014), « Precise estimates of mutation rate and spectrum in yeast. », *Proc Natl Acad Sci U S A* 111, E2310–E2318.

Zilberman, D., Coleman-Derr, D., Ballinger, T. et Henikoff, S. (2008), « Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. », *Nature* 456, p. 125–129.

Zilberman, D., Gehring, M., Tran, R., Ballinger, T. et Henikoff, S. (2007), « Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. », *Nat Genet* 39, p. 61–69.

Zilberman, D. (2017), « An evolutionary case for functional gene body methylation in plants and animals. », *Genome Biology* 18, p. 87.

## RÉSUMÉ

Chez les plantes et les mammifères, la méthylation de l'ADN est une modification chromatinienne qui joue un rôle pivot dans le maintien de l'intégrité des génomes, notamment au travers de l'extinction épigénétique des éléments transposables (ET). Cependant, dans la mesure où la désamination spontanée des cytosines méthylées, qui peut conduire à des transitions C>T, est plus fréquente que celle des cytosines non méthylées, la méthylation est également intrinsèquement mutagène. Cette mutabilité accrue est de fait très certainement à l'origine de la déplétion en dinucléotides CpG observée dans les génomes de mammifères, naturellement méthylés à ces sites sauf au sein des «îlots CpG». A l'exception de cet effet bien connu, aucune étude à ce jour n'a exploré directement et de façon exhaustive l'impact de la méthylation sur le spectre des mutations spontanées.
Dans ce travail, je tire profit d'une population de lignées epiRIL (epigenetic recombinant inbred lines) établie chez la plante Arabidopsis pour évaluer à l'échelle du génome l'impact de la méthylation de l'ADN sur le paysage mutationnel. Les epiRILs dérivent du croisement entre deux parents quasi- isogéniques, l'un sauvage et l'autre porteur d'une mutation conduisant à une réduction de 70% de la méthylation du génome, et il a pu être mis en évidence que des différences parentales de méthylation pouvaient être héritées de façon stable pour >1000 régions le long du génome.
Au moyen de données de séquençage disponibles pour >100 epiRIL, j'ai effectué la caractérisation exhaustive des variants ADN (autres qu'ET) uniques à chaque lignée mais également en ségrégation parmi les epiRIL, ce qui constitue à terme une ressource pour les différentes équipes qui utilisent cette population. En analysant le patron de variants uniques, j'ai mis en évidence une réduction spécifique du taux de transitions C>T en lien avec l'hypométhylation stable dans les epiRIL. J'ai aussi pu décrire que si la remobilisation extensive des ET dans cette population a modelé le spectre des insertions et délétions ponctuelles, elle ne se traduit pas pour autant par des réarrangements récurrents. Je présente également les développements méthodologiques mis en place afin d'effectuer la caractérisation de QTL (quantitative trait loci) "épigénétiques" préalablement identifiés dans la population.



## MOTS CLÉS

Méthylation de l'ADN, taux de mutation, *Arabidopsis thaliana*, population epiRIL, QTL épigénétique



## ABSTRACT

In both plants and mammals, DNA methylation plays a pivotal role in ensuring proper genome function and integrity, notably through the epigenetic silencing of transposable elements (TEs). However, as spontaneous deamination of 5-methylcytosine (5mC), which can lead to C>T transitions, is more frequent than that of unmethylated C, DNA methylation is also inherently mutagenic. This higher mutability of 5mC has indeed been proposed to explain the depletion in CpG dinucleotides in mammalian genomes, which are typically methylated at these sites except in so-called CpG islands. Despite this well-characterized effect of DNA methylation, we still lack a comprehensive view of its impact on the whole mutation spectrum in any given organism.
Here, I take advantage of a population of so-called epigenetic Recombinant Inbred Lines (epiRILs) established in the flowering plant *Arabidopsis thaliana* to investigate the impact of DNA methylation on the spectrum of spontaneous mutations genome wide. The epiRIL population derives from a cross between a wild-type individual and a near-isogenic mutation deficient in DNA methylation, and it could be shown that parental differences in DNA methylation are stably inherited for at least 8 generations over >1000 regions across the genome.
Building on whole-genome sequencing data available for >100 epiRILs, I performed a thorough characterization of non-TE DNA sequence variants that are either private to one line or segregating in the population, therefore establishing a resource for research groups that make use of the epiRIL population. Based on the pattern of private variants, I show a specific reduction in the rate of C>T transitions in the epiRILs, in line with the heritable hypomethylation in this population. I also describe that the extensive TE remobilisation at play among the epiRILs shapes the spectrum of short insertions and deletions yet does not translate into recurrent large-scale mutation events. On another note, I also present methodological developments aimed towards the identification of causal (epi)variants underlying so-called "epigenetic QTL" (quantitative trait loci) previously described in the epiRIL population.



## KEYWORDS

DNA methylation, mutation rate, *Arabidopsis thaliana,* epiRIL population, epigenetic QTL