Sommaire

Liste des abbréviations

Ds : Dissociator

Ac : Activator

ET : Elément Transposable

ARN : Acide ribonucléique

ADN : Acide désoxyribonucléique

ADNc : Acide désoxyribonuclétique complémentaire

LTR : Long Terminal Repeat

TIR : Terminal Inverted Repeat

TSD : Target Site Duplication

ORF : Open Reading Frame

GAG : gène codant une polyprotéine de capside

POL : gène codant une polyprotéine impliquée dans la rétrotransposition

RT : transcriptase inverse

LARD : Large Retrotransposon Derivative

TRIM : Terminal Repeat retrotransposition In Miniature

LINE : Long Interspersed Nuclear Element

SINE : Short Interspersed Nuclear Element

EN : Endonucléase

Pol : Polymérase

Poly(A) : polyadénylé

MITE : Miniature Inverted repeat Transposable Element

Rep : domaine intiateur de réplication

Hel : domaine hélicase

INT : Intégrase

EVE : Endogenous Viral Element

ERV : Endogenous Retrovirus

NOR : Nucleolar Organizer Region

rDNA : AND ribosomal

SSR : Simple Sequence Repeat

siRNA : small interfering RNA

TASR : Transposon Annotation using Small RNAs

Mb : Méga bases

Gb : Giga bases

R : Résistance

ARNm : ARN messager

nm : nanomètre

RdDM : RNA-directed DNA methylation

locus FWA : locus du gène Flowering Wagenigen responsible de la date de floraison de la plante

BLAST : Basic Local Alignment Tool, outil d'alignement

C: Cytosine, base de l'ADN et de l'ARN

G: Guanine, base de l'ADN et de l'ARN

H : tous les nucléotides excepté la guanine

DMR : région différentiellement méthylée (Differentially Methylated Region)

RIN : ripening inhibitor

RP : Repeat Poor, région du génome pauvre en répétitions

INT : Intermediate, région intermédiaire du génome

RR : Repeat Rich, région du génome riche en répétitions

UTR : région non transcrite (Untranslated Region)

CDS : séquence d'ADN codante (Coding DNA Sequence)

H0 : hypothèse nulle du test statistique que l'on cherche à vérifier

GO : Gene Ontology

DDM1 : Dicrease in DNA Methylation 1, une protéine remodeleuse de la chromatine

MELT : Mobile Element Locator Tool

ddm1a et ***ddm1b*** : mutants a et b du gène codant la protéine DDM1

VCF : Virtual Card File, format de fichier de données

PCR : amplification en chaîne par polymérase (Polymerase Chain Reaction)

Liste des figures

Liste des tableaux

Introduction

1.1 Les éléments transposables (ET)

1.1.1 Définition, découverte et généralité

Au cours du XXe siècle, la cytogénéticienne américaine Barbara McClintock commença l'étude systématique de la mosaïque des patrons de couleurs des semences de maïs et de l'instabilité de son héritage. Au cours de cette étude elle identifia deux nouveaux locus génétiques dominants interagissant qu'elle nomma *Dissociator* (*Ds*) et *Activator* (*Ac*), et qui avaient des effets variés sur les gènes voisins. En observant les changements dans le patron de coloration des grains de maïs pendant plusieurs générations de croisements contrôlés, elle fit la découverte que ces éléments *Dissociator* et *Activator* pouvaient transposer, c'est-à-dire changer de position sur le chromosome, montrant ainsi le caractère mobile du génome. Elle a également pu observer que la transposition de *Ds* du chromosome 9, sous le contrôle de *Ac*, était accompagné par la brisure du chromosome. Celle-ci permet la synthèse du pigment de coloration dans les cellules, mais la taille de la zone colorée va elle dépendre du stade de développement au moment de la dissociation, ce qui va alors causer la mosaïque de couleurs. De par ces observations, elle a ainsi pu développer une théorie : ces éléments mobiles réguleraient les gènes en inhibant ou en modulant leur action ; les rendant ainsi responsables de l'activation, ou non, de caractéristiques physiques au cours des générations. Elle émit finalement l'hypothèse que ce type de régulation génique pouvait expliquer comment des organismes multicellulaires complexes, composés de cellules aux génomes identiques, pouvaient avoir des cellules aux fonctions différentes.

Les théories émises par McClintock furent assez mal reçues par la communauté scientifique de l'époque, mais après une redécouverte de ses observations, elles furent admises comme des découvertes majeures du fonctionnement du génome et sont aujourd'hui reconnues. Aujourd'hui, on définit ces éléments transposables (ET) comme étant des séquences génétiques répétées, dispersées dans le génome, et qui ont la capacité de se multiplier et de se déplacer au sein des génomes par des mécanismes de transposition. Depuis leur découverte par Barbara McClintock à la fin des années 1940 (McClintock, B. 1950), de nombreux autres éléments transposables ont été identifiés chez quasiment tous les organismes, des bactéries

aux eucaryotes.

Ces éléments qui ne semblent pas contenir de gènes importants pour leur organisme hôte, disposent juste de l'information leur permettant de se répliquer et / ou se déplacer dans les génomes. Ainsi, ils ont longtemps été considérés comme des éléments « égoïstes » et « parasites » uniquement capables de perturber les gènes. Le séquençage de génomes de nombreuses espèces a par la suite montré que les éléments transposables sont un des composants majeurs des génomes eucaryotes (Vieira, C. *et al*. 2012), et notamment chez les plantes. Au fil des découvertes, le dogme d'un génome constitué d'une succession linéaire de gènes parfaitement stable, a été remplacé par la vue d'un génome comme un réseau complexe impliquant génétique, épigénétique, et interactions cellulaires, dans lequel les ET et autres éléments structuraux et fonctionnels sont impliqués.

1.1.2 Classification des éléments transposables et mécanismes de transposition

Bien que les ET semblent être des séquences de types variés, il a rapidement paru important d'être capable de les classer, c'est-à-dire de les regrouper selon leur structure, la similarité de leur séquence nucléotidique ou protéique, ou encore selon leur mode de réplication (Wicker, T. *et al*., 2007).

1.1.2.1 Classe d'élément transposable: le plus haut niveau de classification

De manière générale, il existe trois niveaux de classification des éléments transposables. Le niveau le plus haut est celui de la classe : les éléments peuvent être de la classe I, aussi appelés rétrotransposons, qui utilisent un mécanisme dit de « copier / coller » qui nécessite un intermédiaire d'acide ribonucléique (ARN), et les éléments de type II, ou transposons à acide désoxyribonucléique (ADN), qui ne nécessitent pas d'intermédiaire ARN pour leur mobilisation.

Les ET de classe I sont donc caractérisés par un mode de transposition dit de « copier / coller » (Figure 1). Ce type de transposition engendre systématiquement une augmentation du nombre de copies, pouvant alors expliquer que ce type d'éléments est prédominant dans la plupart des génomes eucaryotes. La transposition de type « copier / coller » implique la transcription d'un acide ribonucléique (ARN) à partir de l'élément par l'ARN Polymérase II

ou III, ARN qui est ensuite convertit en acide désoxyribonucléique complémentaire (ADNc) par réverse transcription. Finalement, cet ADNc sera intégré à une nouvelle position du génome par une intégrase. On identifie principalement ce type d'éléments, par la présence, dans leur séquence, du code d'une proteine « réverse transcriptase », à l'exception de certains qui sont uniquement identifiables par la présence de motifs ADN spécifiques puisqu'ils dérivent des promoteurs de l'ARN Polymérase III.

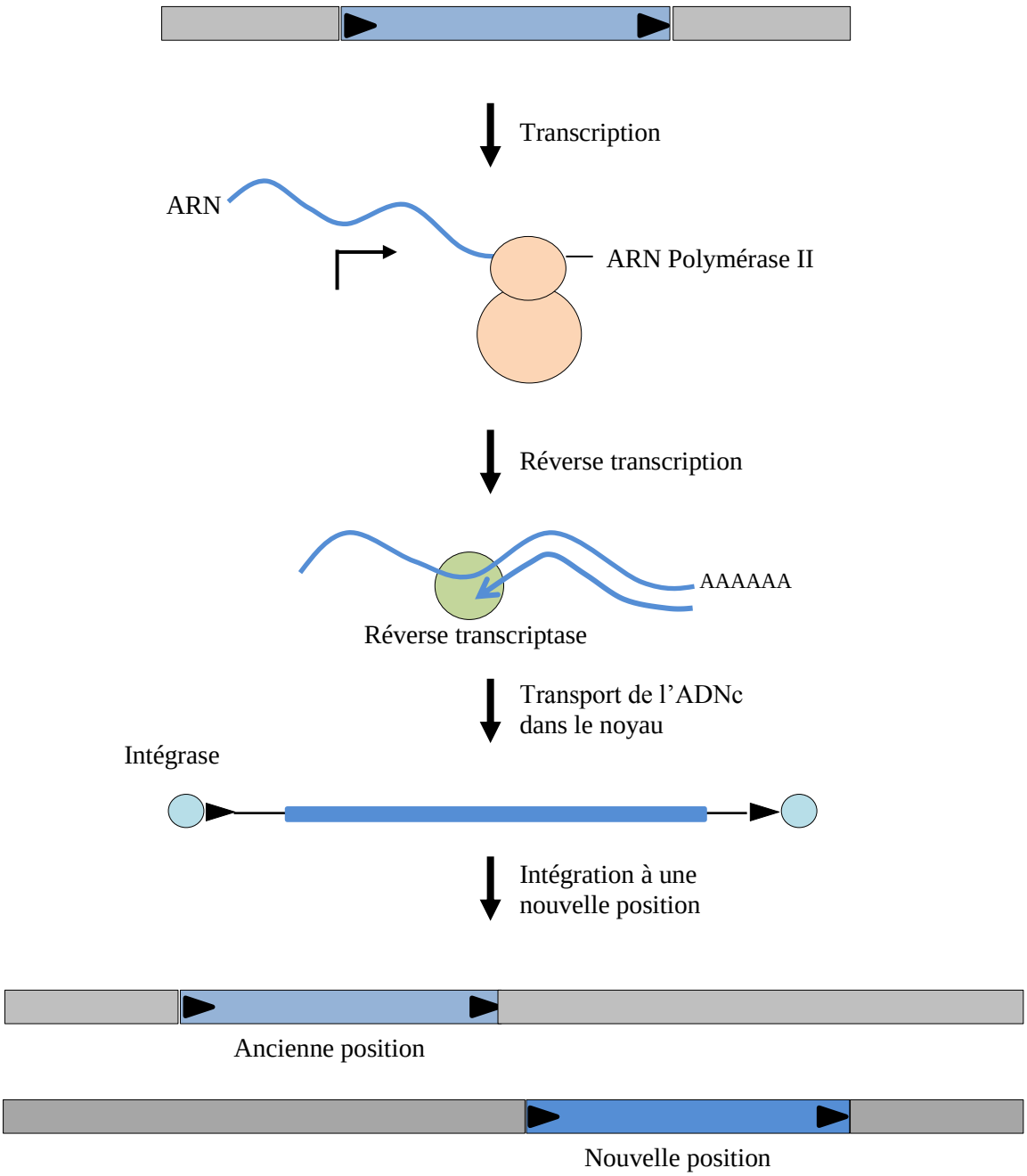


Figure 1 : Mécanisme général de la transposition des éléments de type I. Les grandes répétitions terminales (LTR, Long Terminal Repeat) sont représentées par les flèches pleines noires.

Adapté d'après Lisch, D. *et al*., 2012 et Levin, H . L. *et al*., 2011

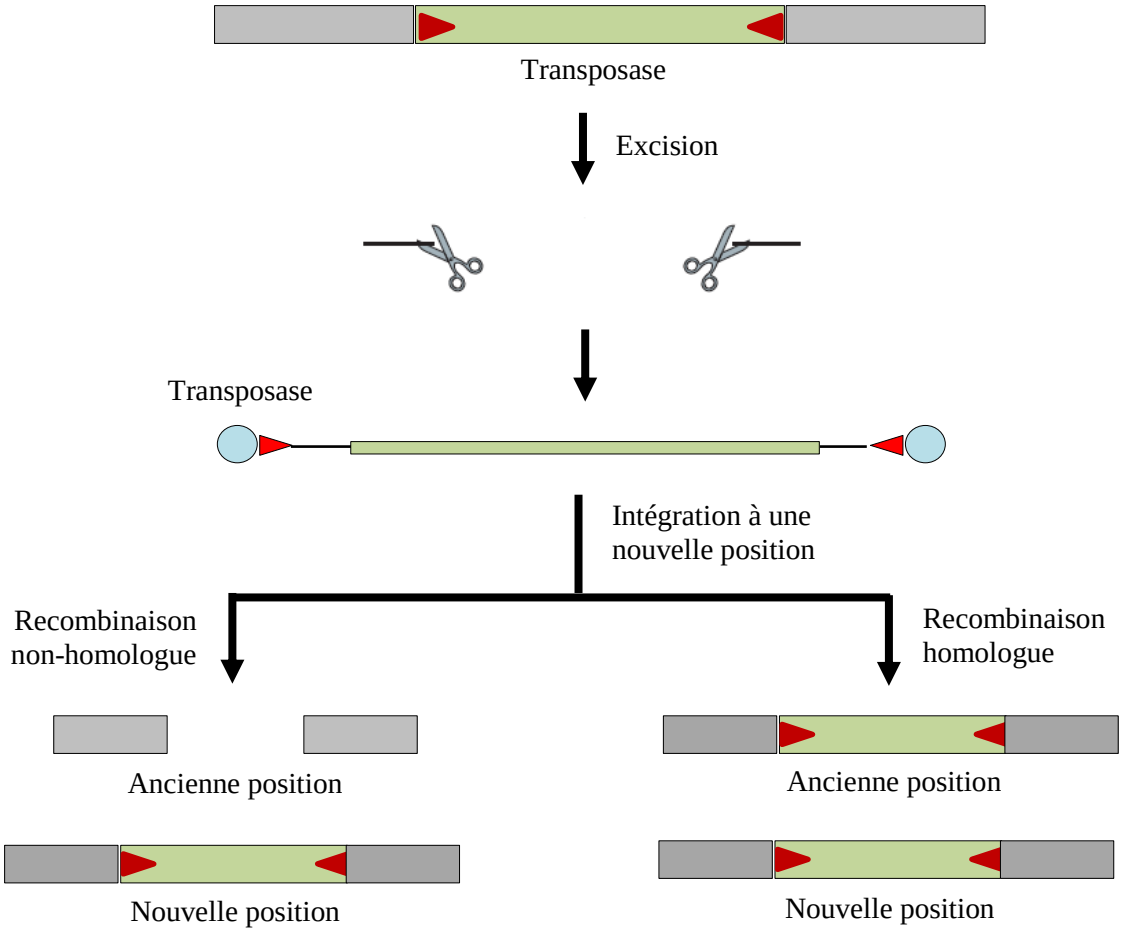


Figure 2 : Mécanisme général de la transposition des transposons à ADN de la sous-classe 1. L'élément transposable est représenté par le rectangle vert, et ses courtes séquences répétées inversées (TIR, Terminal Inverted Repeat) sont représentées par les triangles rouges. Adapté d'après Lisch, D. *et al*., 2012

Les ET de classe II, quant à eux, sont également appelés transposons à ADN et transposent principalement par un mécanisme dit de « couper / coller » **(**Figure 2**).** Contrairement au mode de transposition « copier / coller » des rétroéléments, le mode « couper / coller » n'induit pas toujours une augmentation du nombre de copies. En effet, l'élément est d'abord physiquement excisé du chromosome et est ensuite réintégré à un nouvel endroit dans le génome grâce à la transposase codée par certains ET. La réplication de ce type d'élément est terminée lorsque la cassure double brin de l'ADN est réparée. Si la réparation du site donneur se fait par recombinaison illégitime ou non-homologue (« *non homologous end joining* »), c'est-à-dire qu'elle conduit à la réunion de deux fragments d'ADN ne présentant pas ou très peu d'homologie de séquence, alors la copie au site donneur est perdue. Si, au contraire, la réparation se fait par recombinaison homologue, c'est-à-dire entre deux séquences identiques

situées sur deux molécules d'ADN, ou distantes l'une de l'autre sur la même molécule, alors la copie du site donneur est conservée, l'insertion s'étant produite à un site où la réplication n'avait pas encore eu lieu.

Les mécanismes décrits ci-dessus sont des mécanismes généraux et certains éléments particuliers, notamment au sein de la classe II, ne suivent pas ce mode classique de transposition et reste encore mal connu.

1.1.2.2 Superfamilles d'ET : une marque de la diversité

En fonction des propriétés particulières que présentent les éléments, au sein de chacune de nos deux classes d'ET, ils vont pouvoir être regroupés en différentes superfamilles et familles d'éléments. Mais on pourra également distinguer nos ET en les classant comme des éléments autonomes, c'est-à-dire qu'ils codent l'ensemble des protéines nécessaires à leur transposition, ou non-autonomes qui eux nécessitent la machinerie des éléments autonomes intacts en *trans* pour transposer. La relation entre les copies autonomes et non-autonomes est semblable à du parasitisme : les copies autonomes (les « hôtes ») sont capables de survivre et de se répliquer seuls, alors que les copies non-autonomes (les « parasites ») n'en sont pas capables sans les copies autonomes dont elles vont affecter la survie en s'amplifiant. Les éléments autonomes étant eux-mêmes considérés comme des parasites des génomes, on peut alors considérer que les éléments non-autonomes sont des hyperparasites (Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P. & Hua-Van, A., 2016).

1.1.2.2.1 La diversité de la classe I

La classe des rétroélements peut donc être divisée en deux catégories : les éléments de type I à grandes répétitions terminales (Long Terminal Repeat, LTR), et les éléments de type I non LTR, qui diffèrent également par leur mode de transposition. Cependant, une caractéristique commune à l'ensemble des éléments de la classe I est qu'ils sont tous encadrés par des sites cibles de la duplication (TSD, Target Site Duplication).

Tout d'abord, le groupe des rétroéléments à LTR forme une classe d'éléments très proches des rétrovirus puisqu'ils font partie de la nouvelle famille des *Belpaoviridae* (Krupovic, M. *et al.*, 2018), qui vont se retrouver intégrés aux génomes au cours de leur cycle de réplication ou

par recombinaison illégitime. Ce groupe peut à son tour être séparé en plusieurs superfamilles, comme celles de Gypsy, Copia, DIRS… qui présentent le plus souvent la caractéristique de porter des terminaisons répétées terminales ou LTR, pour Long Terminal Repeat. Ces LTR, situées en orientation directe de part et d'autre de la séquence de l'élément et dont la taille peut varier de quelques centaines à plus d'un millier de paires de bases, portent les régions promotrices de ces ET. Ces superfamilles qui composent le groupe des rétroéléments à LTR forment tout de même deux groupes distincts : les rétroéléments à LTR autonomes et les rétroéléments à LTR non-autonomes (Figure 3a). Les éléments autonomes sont caractérisés par la présence d'au moins deux cadres de lecture ouverte (Open Reading Frame, ORF) caractéristiques : l'une correspondant au gène GAG qui code une polyprotéine de capside, l'autre au gène POL codant une polyprotéine impliquée dans les diverses étapes de la rétrotransposition en regroupant des activités enzymatiques de protéase, transcriptase inverse (RT), RNAse H et intégrase. Les rétroéléments à LTR non-autonomes, eux, dérivent d'éléments autonomes dont la séquence codante est partiellement ou totalement absente, ils regroupent, entre autre, les LARDs (LArge Retrotransposon Derivative), et les TRIMs (Terminal Repeat retrotransposon In Miniature) (Witte, C.-P., Le, Q. H., Bureau, T. & Kumar, A., 2001) (Gao, D., Li, Y., Kim, K. D., Abernathy, B. & Jackson, S. A, 2016). Parmi l'ensemble de ces superfamilles des éléments de type I à LTR, les plus fréquents dans les génomes de plantes sont finalement les éléments autonomes de type Gypsy et Copia.

Les rétroéléments non LTR sont quant à eux également divisés en deux catégories d'éléments (Figure 3b) : les éléments autonomes, ou LINE (Long Interspersed Nuclear Element), et les éléments non autonomes, ou SINE (Short Interspersed Nuclear Element), tous deux présents dans les génomes de nombreuses plantes. Les LINE mesurent environ 6-7kb et sont constitués de l'ORF1, codant une protéine proche de la protéine GAG, et portent également une endonucléase (EN) et une reverse transcriptase (RT) (Feschotte, C. *et al*., 2002). Les SINE, quant à eux, sont des séquences répétées de taille et de structure variables, allant de quelques centaines de paires de bases à plusieurs kilobases. A l'inverse des LINE, ils ne codent pas les séquences nécessaires à leur transposition mais présentent des promoteurs de la Polymérase III (Pol III). Ces deux types d'éléments se terminent par une séquence poly-adénylée (poly(A)), issue de la transcription inverse de l'extrémité poly(A) du transcrit, ou encore microsatellite

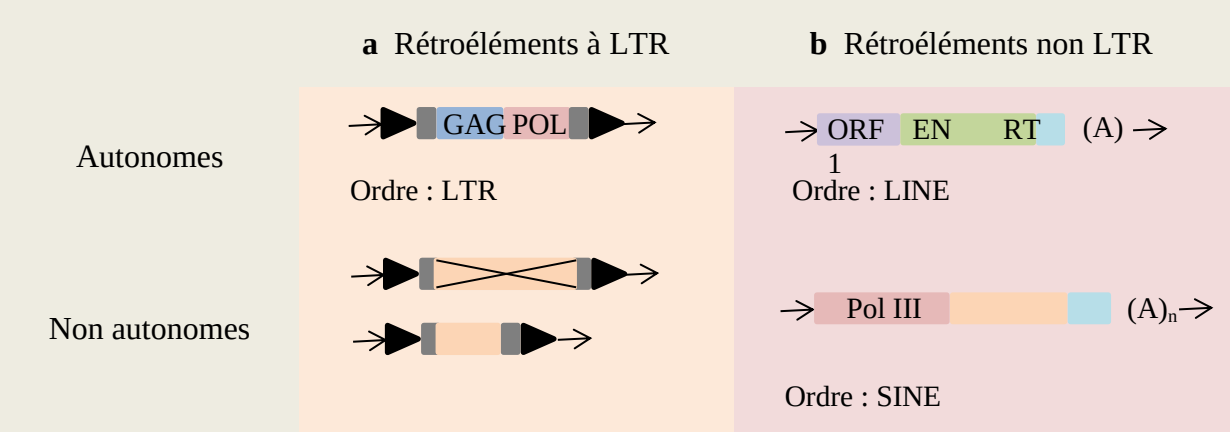


Figure 3 : Classification des éléments transposables de type I, ou rétrotransposons. Tous ces éléments sont encadrés par des sites cibles de la duplication (TSD, Target Site Duplication, petites flèches). a. Classification des éléments de type I à LTR. Le gène GAG code une polyprotéine de capside et le gène POL code une polyprotéine impliquée dans la rétrotransposition de ces éléments qui sont également porteurs de grandes répétitions terminales (LTR, Long Terminal Repeat, triangles noirs). b. Classification des éléments de type I non – LTR. L'ORF1 code une protéine chaperonne d'acide nucléique, l'ORF « EN RT » code à la fois une endonucléase et une réverse transcriptase. Pol III contient des promoteurs de la polymérase III. $(A)_n$ correspond à une queue poly-adénylée.
Adapté d'après Feschotte, C. *et al*., 2002 et Slotkin, R. K. *et al*., 2007)

1.1.2.2.2 La diversité de la classe II

Les éléments de la classe des transposons à ADN, sont divisés en deux sous-classes d'éléments distincts à la fois par leur structure et par leur mode de transposition. La sous-classe 1 (Figure 4a), regroupe à la fois des éléments autonomes et non autonomes. Les éléments autonomes présentent de courtes séquences répétées inversées (Terminal Inverted Repeat, TIR) à leurs extrémités, ainsi que des sites de duplication terminaux (Terminal Site Duplication, TSD) et au moins une ORF correspondant au gène codant une transposase. Cette transposase assure l'excision et l'intégration du transposon grâce à ses activités d'endonucléase et de ligase, cette dernière activité lui permettant de fusionner les extrémités du TE avec la séquence receveuse. Les éléments non autonomes possèdent également des TIR mais le gène codant la transposase est totalement absent ou défectueux, ce qui est, par exemple, le cas des MITE (Miniature Inverted repeat Transposable Elements) (Feschotte, C., Zhang, X. & Wessler, S. R., 2002) (Ye, C., Ji, G. & Liang, C., 2016).

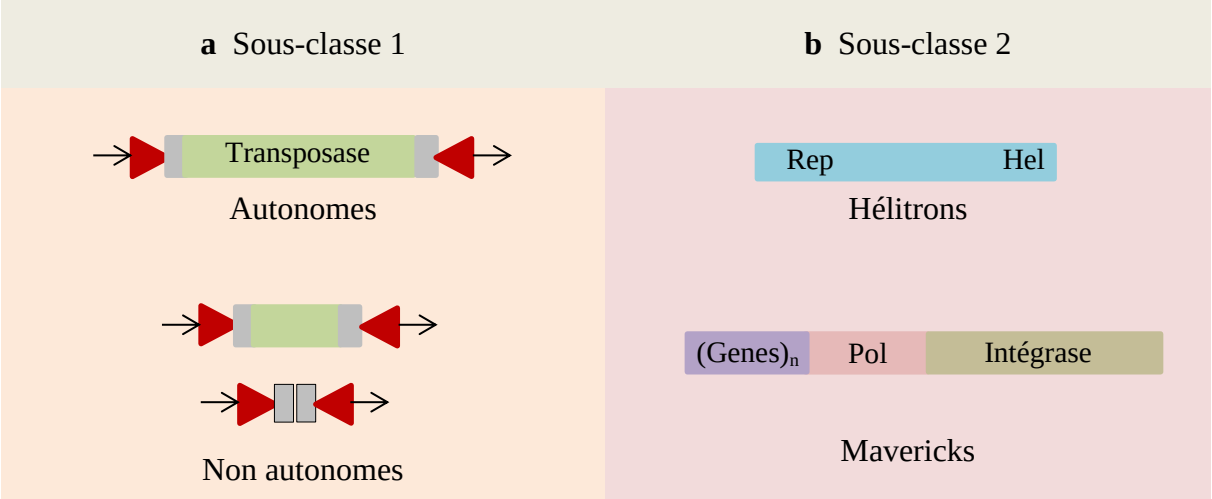


Figure 4 : Classification des éléments transposables de type II, ou transposons à ADN. a. Eléments de la sous-classe 1 des éléments de type II. Ces éléments sont tous porteurs de sites cibles de duplication (TSD, Target Site Duplication, triangles rouges) et de courtes séquences répétées inversées (TIR, Terminal Inverted Repeat, petite flèches). Les éléments autonomes sont de plus porteurs d'une transposase. b. Eléments de la sous-classe 2 des éléments de type II. Cette sous-classe regroupe les Hélitrons, porteurs d'un domaine initiateur de réplication (Rep) et d'un domaine hélicase (Hel), et les éléments Mavericks, pouvant porter jusqu'à 11 gènes dont un codant pour une polymérase (Pol) et un codant pour une intégrase.
Adapté d'après Feschotte, C. *et al*., 2002 et Slotkin, R. K. *et al*., 2007

La sous-classe 2 (Figure 4b) quant à elle est représentée par deux groupes d'éléments : les Hélitrons et les Mavericks. Les Hélitrons (Kapitonov, V. V. & Jurka, J., 2001) sont un groupe étendu et diversifié assez peu fréquents dans l'ensemble des génomes. Bien que classés parmi les éléments de type II, ceux-ci ne présentent pas de structures terminales particulières (absence de LTR et de TIR) et ne génèrent pas non plus de TSD. Ils contiennent cependant deux domaines enzymatiques caractéristiques, présents sur une même protéine RepHel : un domaine Rep, initiateur de réplication, et un domaine Hel, codant une hélicase. Leur mode de transpositipon, encore mal connu, se ferait par un mécanisme de cercle roulant (rolling circle) (Xiong, W., Dooner, H. K. & Du, C., 2016) (Mourier, T., 2016) similaire à celui trouvé chez divers phages et plasmides procaryotes (Figure 5). Les éléments Mavericks, eux (ou Polintrons), sont des transposons eucaryotes découverts plus récemment (Kapitonov, V. V. & Jurka, J., 2006) (Wicker, T. *et al*., 2007) (Feschotte, C. & Pritham, E. J., 2007), mais encore jamais identifiés chez les plantes. Ces éléments d'une longueur de 10 à 20 kb possèdent de longues séquences TIR et peuvent contenir jusqu'à 11 gènes dont parfois une ADN polymérase de type virale (ADN polymérase B), et une intégrase INT (de type *c-int*) proche

de celles de certains éléments de type I mais sans être porteurs de la RT.

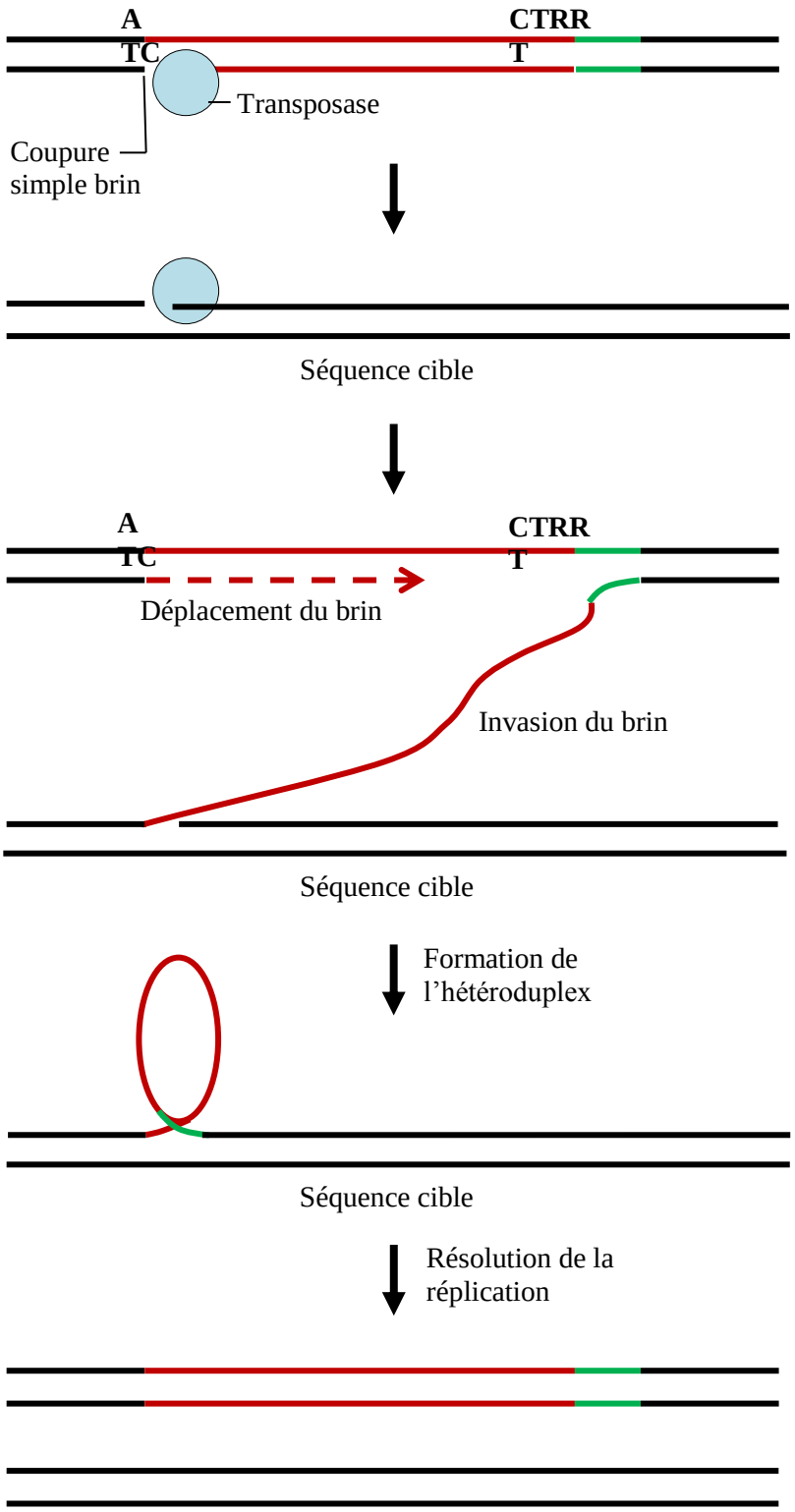


Figure 5 : Mécanisme de transposition par "rolling circle" des Hélitrons. Le R fait référence aux purines (A et G), le fragment rouge correspond à la séquence de l'hélitron et le fragment vert est composé de gènes et de fragments de gènes.

Adapté d'après Lisch, D. *et al*., 2012

Les différentes superfamilles d'éléments transposables peuvent à leur tour être séparées en plusieurs familles. Mais contrairement aux superfamilles, dont les critères de classification sont bien déterminés, il est difficile de définir des règles pour la formation des familles, qui reposent généralement sur un rôle biologique. Cependant, la connaissance précise des critères de reconnaissance et de classification des ET permet aux outils de les identifier dans les nombreux nouveaux génomes séquencés.

1.1.3 Notion de repeatome et stratégies d'annotation

Une partie de chaque génome est composée de ce que l'on appelle le *repeatome* (Maumus, F. & Quesneville, H., 2014), englobant à la fois les éléments transposables et bien d'autres séquences. Compte tenu de l'importance de ces séquences, et particulièrement des ET, il est important de savoir les identifier, différentes méthodes dites d'annotation (Lerat, E., 2010) ont donc été développées dans ce but.

1.1.3.1 Qu'est-ce que le repeatome ?

Le *repeatome* peut être vu comme un ensemble de séquences répétées qui composent le génome. Classiquement, un *repeatome* eucaryote peut contenir un nombre très variable de composants qui peuvent être des éléments transposables (ET), des virus et rétrovirus endogènes (EVEs et ERVs), des séquences non classifiées, des séquences satellites, des *Nucleolar organizer region* (NOR) comprenant notamment de l'ADN ribosomal (rDNA), ou bien des familles de gènes en expansion. Les ET restent toutefois le composant majeur de ce *repeatome* du fait de leur taux élevé de duplication.

Un ensemble d'outils de détection est nécessaire pour identifier, au sein du *repeatome*, à la fois les ET et d'autres séquences répétées. Il peut alors être question de détecter ce que l'on appelle des virus endogènes (Holmes, E. C., 2011) (Doolittle, R. F., Johnson, M. S. & McClure, M. A., 1989) (Hayward, A., 2017), qui sont en fait des séquences ADN dérivées d'un virus et qui sont devenues constitutives d'un organisme non viral. Il peut s'agir d'un génome viral entier (provirus), mais plus généralement de fragments de génomes viraux. Il sera également possible de détecter des séquences identifiées comme fortement répétées dans le génome mais jusqu'alors dépourvue de classification. Et enfin, les outils seront en mesure d'identifier les séquences satellites (Garrido-Ramos, M., 2017), pouvant faire faire jusqu'à

plusieurs millions de paires de bases, qui sont des séquences constituées d'un grand nombre de répétitions d'un même motif appelé unité de répétition. De ce fait, elles présentent également un biais de composition par rapport au reste du génome, ce qui facilite leur détection.

Au sein du *repeatome*, il sera également possible de trouver des composants plus essentiels au bon fonctionnement du génome, tel que les séquences d'ADN ribosomique (rDNA) (Hillis, D. M. & Dixon, M. T., 1991). Ces rDNA correspondent à des séquences d'ADN qui seront transcrites en ARN ribosomique, qui, chez les organismes eucaryotes, prend la forme d'un opéron formé de la répétition de plusieurs domaines. Les ribosomes, quant à eux, sont des complexes de protéines et d'ARN ribosomique au sein desquels se déroule la traduction. Les rDNA font partie des *Nucleolat irganizer region* (NOR) (McStay, B., 2016), des régions chromosomiques spécifiques qui sont associées à un nucléole (Stępiński, D., 2014) après la division du noyau, et qui font également partie des séquences du *repeatome*. Finalement, au sein de ce *repeatome*, on pourra également détecter les familles de gènes en expansion, qui sont généralement des familles multigéniques, c'est-à-dire un ensemble de gènes qui sont issus d'un même gène ancestral ayant subi duplication, mutation et transposition pour donner un ensemble de gènes, au sein d'un génome, présentant des homologies de séquences et donc des fonctions proches.

1.1.3.2 Utilisation de librairies pour l'annotation des éléments transposables

La première méthode d'annotation des éléments transposables, et probablement la plus ancienne qui ne soit pas manuelle, est l'utilisation d'une base ou d'une librairie de séquences connues, comme par exemple la RepBase (Jurka, J., 1998) (Jurka, J., 2000), pour la recherche des ET dans les nouveaux génomes.

La RepBase est donc une base de données de séquences d'éléments répétés, provenant de différentes espèces eucaryotes, et qui a été développée depuis 1990 sous la direction de Jerzy Jurka. La plupart des éléments constituant cette base sont des séquences consensus de grandes familles ou de sous familles de séquences répétées, mais elle contient également des familles plus petites présentes sous forme de séquences exemples uniques. Cette base de données qu'est la RepBase peut alors être utilisée comme collection de références pour annoter et / ou

masquer les éléments répétés dans les génomes à l'aide d'outils d'alignement de séquence comme BLAST (Altschu P, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J., 1990).

1.1.3.3 Les outils dédiés à l'annotation par alignement

A partir des banques de données, certains outils ont été développés pour l'annotation des éléments transposables de nouveaux génomes par des méthodes d'alignement. Parmi ces outils, on trouve entre autres RepeatMasker, CENSOR ou encore Blaster.

Tout d'abord, RepeatMasker (Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker at http://repeatmasker.org) est un outil qui sert principalement pour le masquage des séquences répétées. En effet, le programme va parcourir la séquence d'ADN fournie pour détecter les répétitions et les séquences de faible complexité qu'il va alors comparer avec, par exemple, la RepBase. L'outil fournira ainsi à l'utilisateur au moins trois fichiers : un contenant la séquence du génome soumis dans lequel tous les éléments trouvés sont remplacés par des N ou des X, une table avec l'annotation des séquences marquées (séquence, coordonnées…) et un tableau résumant le contenu en répétitions de la séquence soumise.

Ensuite, il possible d'utiliser CENSOR (Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J., 2006), qui réalisera le même travail de comparaison d'une séquence à une collection de répétitions de référence (comme la RepBase) afin de détecter et masquer les portions homologues à l'aide d'un symbole de masquage.

Enfin, Blaster est une suite de programmes, développée en C++, permettant la recherche et l'annotation d'éléments transposables. Pour cela, il compare deux jeux de séquences : une banque de données de références et une banque de données de soumission à l'aide de BLAST, puis défragmente les résultats à l'aide d'un algorithme de programmation dynamique.

Afin de rendre la détection plus complète et plus efficace, ces différents outils peuvent être intégrés et combinés au sein de pipelines dédiés à la recherche des éléments répétés et notamment à leur annotation.

1.1.3.4 Les méthodes de détection ab initio

Une autre méthodologie d'annotation couramment utilisée est celle des approches de détection *ab initio*, reposant sur les informations d'alignements d'un BLAST « tout-contre-tout ». Dans ce cas, les outils vont généralement chercher à construire des consensus à l'aide d'alignements de séquences et ensuite annoter le génome étudié à partir de cette banque de consensus. Pour ce type d'approche, les principaux outils utilisés sont : RECON, REPET et RepeatModeler.

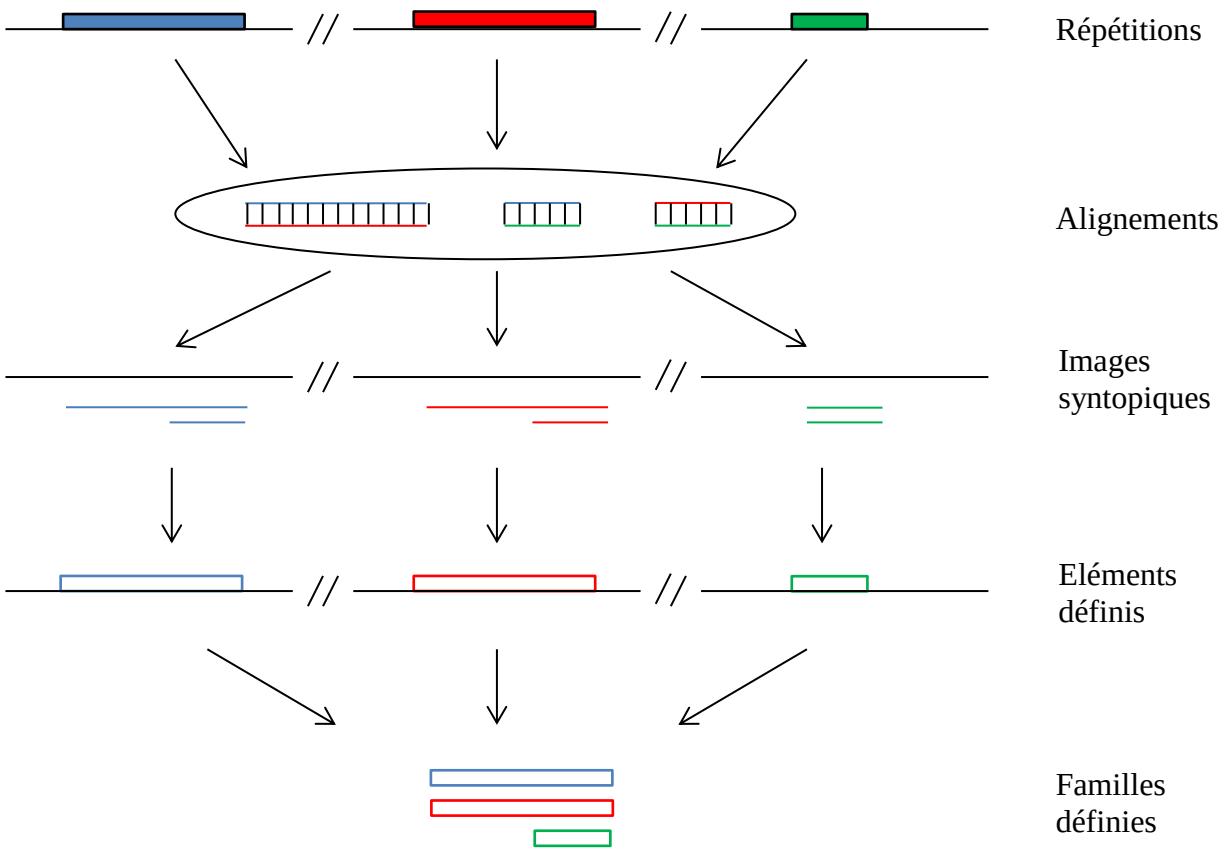


Figure 6 : Schéma de la méthodologie d'anotation *de novo* de l'outil RECON. La séquence génomique de départ contient une famille de répétitions avec trois copies : deux complètes (bleue et rouge) et une partielle (verte). Lors de la comparaison par paire, ces éléments encore inconnus vont donner trois alignements. Les fragments alignés (images) sont regroupés en fonction de leur région génomique et ceux venant d'un même élément (images syntopiques) peuvent être regroupés en fonction de leurs recouvrements. On définit alors les éléments sur la base des jeux syntopiques et on les regroupe en une famille car ils sont similaires les uns aux autres.

D'après Bao, Z., 2002

La première option, RECON (Bao, Z., 2002) est un outil d'identification et de classification *de novo* de séquences répétées qui utilise les informations d'alignements d'un BLAST « tout-contre-tout » pour définir les liens entre les copies individuelles afin de les regrouper par familles d'éléments. Ainsi, si l'on prend une séquence génomique contenant trois copies d'un élément jusqu'alors non identifié (Figure 6), deux copies complètes et une copie partielle, ces éléments fournissent trois alignements lors d'une comparaison par paire. Les fragments alignés, sont ensuite triés en fonction de leur région génomique et ceux venant du même élément peuvent être groupés en raison de leur chevauchement et former ainsi une famille sur la base de leur similarité.

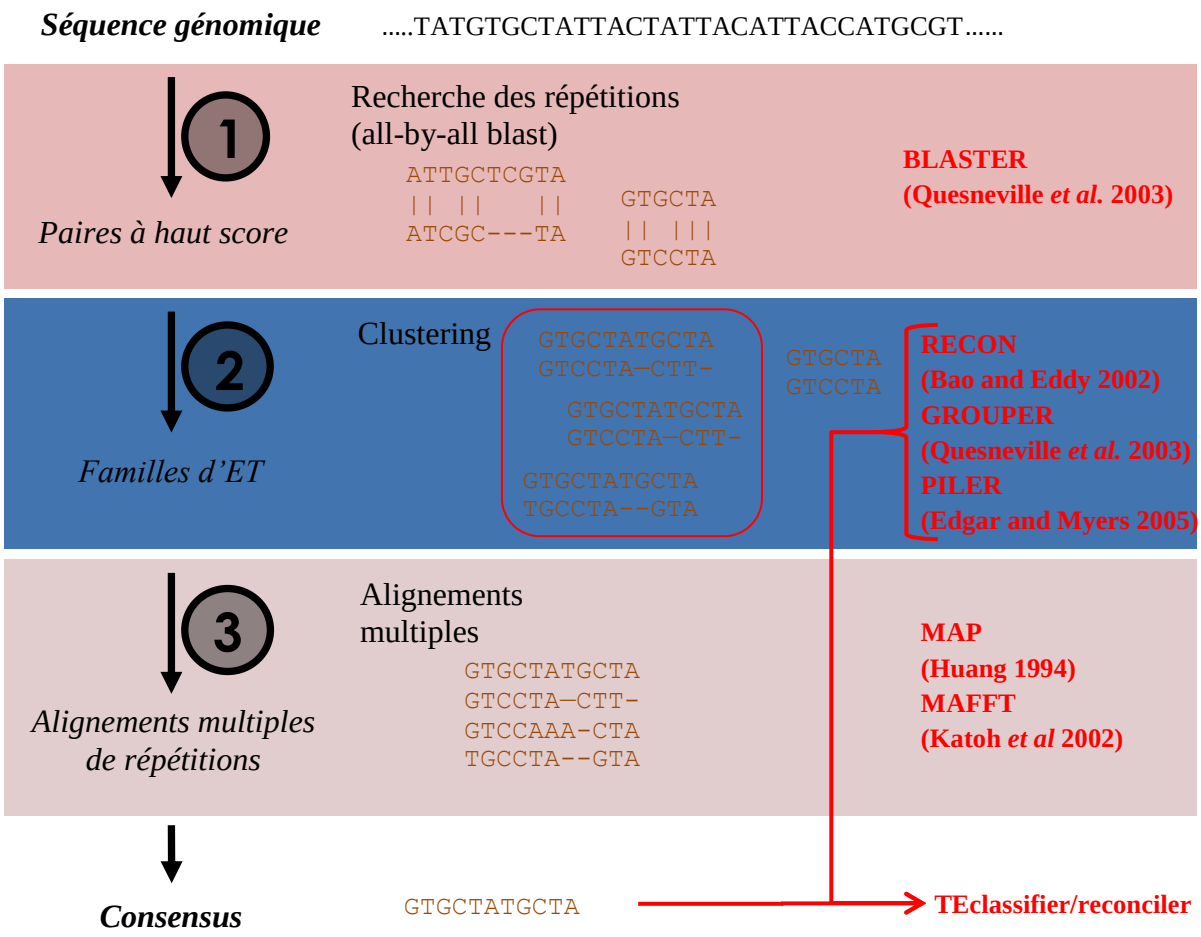


Figure 7 : Protocole de détection d'éléments répétés et de construction de séquences consensus par le pipeline TEdenovo de l'outil REPET. Les séquences répétées sont recherchées par un alignement *all-by-all* puis regrouper selon leur proximité afin de construire un consensus pour une famille d'ET donnée à l'aide d'un alignement multiple.
D'après https://urgi.versailles.inra.fr/Tools/REPET

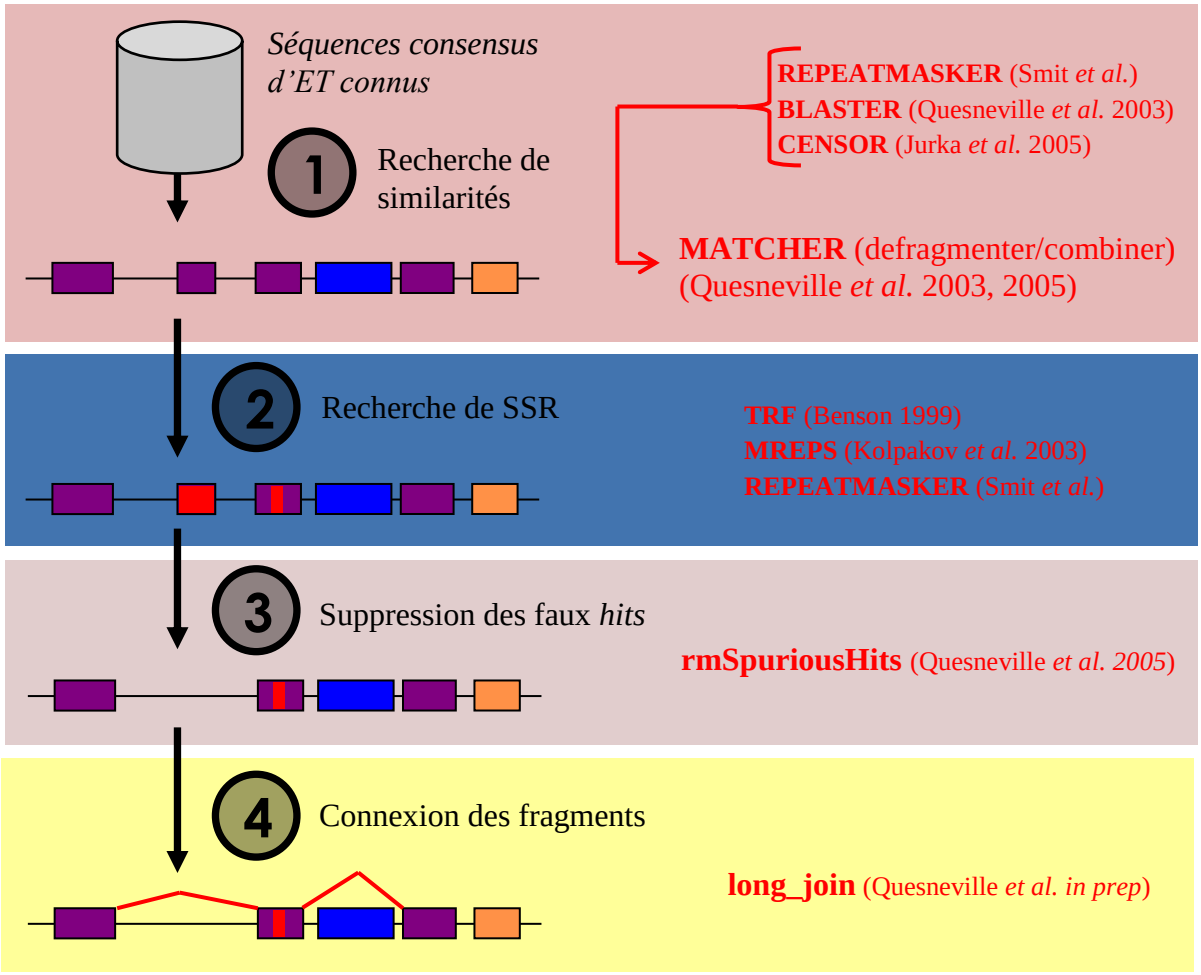


Figure 8 : Protocole d'annotation des éléments répétés par le pipeline TEannot de l'outil REPET. Les consensus fabriqués avec l'aide de TEdenovo sont utilisés pour annoter les éléments répétés du génome. Les résultats sont également nettoyés (détecter des SSR, procédure pour joindre les fragments annotés appartenant à un même élément) avant d'être regroupés dans un même fichier de sortie.
D'après https://urgi.versailles.inra.fr/Tools/REPET

Le second outil nommé REPET (Quesneville, H. *et al.*, 2005) (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011) repose sur l'identification des éléments répétés par l'alignement sur elles-mêmes de séquences génomiques. Cet outil est composé de deux pipelines : TEdenovo, qui réalise la détection *de novo* des éléments répétés afin de former des consensus, et TEannot, qui annote les éléments contenus dans le génome à partir des consensus construits.

Pour commencer, le pipeline TEdenovo (Figure 7) se déroule en quatre grandes étapes : un alignement *all-by-all* afin de rechercher les répétitions dans le génome, un *clustering*, c'est-à-dire un regroupement des répétitions pour former des *clusters* correspondants chacun à une famille d'ET en utilisant plusieurs outils dédiés, et en particulier RECON et Grouper (Quesneville, H., Nouaud, D. & Anxolabéhère, D., 2003), puis au sein de chaque groupe un alignement multiple des séquences est réalisé et enfin un consensus est construit, classifié et la redondance des séquences est éliminée. Toutes ces étapes vont permettre de former une librairie d'ET. Puis, le pipeline TEannot (Figure 8) intervient pour annoter les éléments du génome à l'aide de la librairie de consensus. Pour cela, il s'agit tout d'abord de réaliser une recherche par similarité entre les consensus et la séquence du génome, puis une recherche des séquences satellites. Une fois ces étapes de détection réalisées, les fragments détectés par erreur, ou faux positifs, vont être éliminés et les fragments appartenant à un même élément connectés entre eux.

Finalement Repeat Modeler (Smit, A.F.A., Hubley R.., 2008-2010) est un outil d'identification *de novo* de familles de répétitions qui s'appuie sur les résultats de deux programmes : RECON et RepeatScout. Ceux-ci utilisent des méthodes complémentaires pour identifier les éléments répétés et les liens formant les familles de répétitions à partir d'une séquence donnée. Il permet donc d'automatiser les exécutions de RECON et RepeatScout à partir d'une base de données génomique et utilise le résultat pour construire, affiner et classifier les consensus d'éléments répétés.

1.1.3.5 L'annotation par des approches k-mers

Une dernière méthodologie d'annotation peut être également utilisée afin de détecter les éléments répétés dans les génomes : les approches par k-mers. De nombreux outils emploient désormais cette approche et peuvent travailler à la fois sur des données directement issues de séquençage (reads) ou des séquences assemblées.

Il est donc possible d'identifier, d'une part, certains outils qui vont utiliser les lectures (reads) de séquençage haut débit afin d'identifier les séquences répétées. On trouve notamment l'outil Repeat Explorer (Novák, P., Neumann, P. & Macas, J., 2010), qui effectue une analyse de *clustering* basée sur le graphique des similarités entre lectures. Mais de tels outils ne sont pas les plus répandus et surtout, ils ne sont pas les plus utilisés en raison généralement de la

difficulté à travailler sur les reads pour identifier des ET.

Pour pallier à ce problème, il existe, d'autre part, des outils basés sur l'utilisation de séquences assemblées qui peuvent employer diverses approches pour la détection et l'annotation des éléments transposables. C'est le cas, tout d'abord, des outils utilisant la localisation des petits ARN interférents (small interfering RNA, siRNA), ces siRNA faisant partie de la voie de silencing des ET, pour localiser la position des ET, comme le fait TASR (Transposon Annotation using Small RNAs) (El Baidouri, M. *et al.*, 2015). D'autres outils, comme P-clouds (Gu, W., Castoe, T. A., Hedges, D. J., Batzer, M. A. & Pollock, D. D., 2008), vont identifier les structures des répétitions dans les génomes eucaryotes en utilisant les nuages de k-mers, ou bien d'autres, comme Tallymer (Kurtz, S., Narechania, A., Stein, J. C. & Ware, D., 2008), se basent sur les comptages de ces k-mers pour identifier les potentiels ET. Enfin, il est possible de trouver des outils comme Repeatscout (Price, A. L., Jones, N. C. & Pevzner, P. A., 2005) qui vont rechercher les k-mers puis ensuite étendre la portion identifiée comme potentiel ET à partir de l'alignement de séquences. Dans tous les cas, les consensus obtenus à partir d'un jeu de k-mer vont dépendre des variables de scores d'alignement choisis.

1.1.4 Impacts structuraux

Les éléments transposables sont donc des composants essentiels des génomes qui vont pouvoir influencer à la fois le contenu, mais également la structure de ceux-ci en induisant divers remaniements.

1.1.4.1 Modifications de la composition, de la taille et de la structure des génomes

Lorsque l'on regarde les génomes de plantes, l'élément le plus frappant est peut être leur variation de tailles. En effet, on peut observer de petits génomes allant de 60 Mb pour *Genlisea aurea* à des génomes énormes de plus de 150 Gb pour *Paris japonica*.

Cependant, cette variation de taille n'a pas pour origine principale le niveau de ploïdie du génome ou le nombre de gènes comme on pourrait s'y attendre, mais bien le contenu en éléments transposables (Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A., 2016). En effet, bien qu'ils soient relativement peu fréquents dans les petits génomes, représentant

seulement 5% du génome de la levure *Saccharomyces cerevisiae* (Kin, J. M., Vanguri S., Boeke, J. D., Gabriel, A. & Voytas, D. F., 1998) ou 15% dans le génome de la mouche *Drosophila melanogaster* (Drosophila 12 Genomes Consortium, 2007), ils occupent une grande partie de l'ADN répété des grand génomes, allant de 45% dans le génome de l'Homme (International Human Genome Sequencing Consortium, 2001), à près de 85% dans celui du maïs *Zea mays* (Schnable, P. S. *et al*. 2009) (Figure 9). Mais l'impact des éléments transposables sur la taille des génomes est très aléatoire et imprévisible puisqu'ils ne sont pas soumis aux mêmes contraintes que le reste des composants de celui-ci. Ces changements de taille liés aux répétitions ont alors fréquemment lieu dans le cas où une famille d'ET subit un *burst*, c'est-à-dire une amplification très rapide du nombre de copies d'une famille d'éléments donnée (Bennetzen, J. L. *et al*. 2014) (Fedoroff, N. V. *et al*. 2012). Finalement, malgré l'impact que peut avoir le contenu en répétitions sur la taille du génome, il semble que les plus grandes variations de taille soient principalement liées à des phénomènes d'activation de la transposition, de sélection, ou à des taux de suppression de fragments d'ADN différents, dans le cas où l'organisme change d'environnement ou de climat de manière importante.

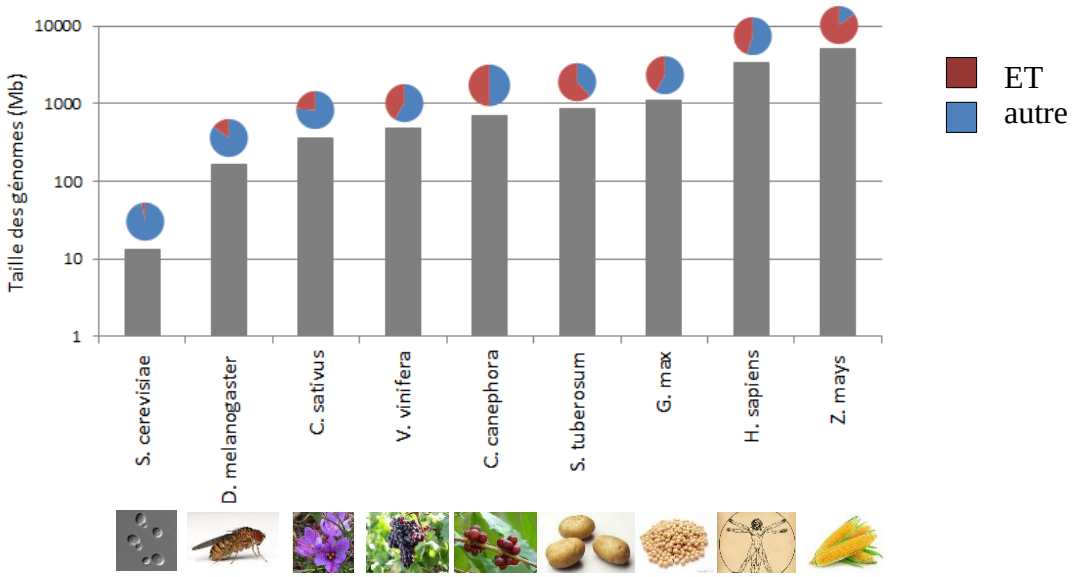


Figure 9 : Taille du génome et proportion d'ET de différentes espèces eucaryotes.
D'après Vieira, C. *et al*. 2012 et Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A., 2016.

Les éléments transposables ne sont cependant pas présents dans les génomes uniquement pour en augmenter la taille. Ils semblent également à l'origine de certaines structures particulières de la chromatine, comme les centromères (Plohl, M., Meštrović, N. & Mravinac, B., 2014) et péricentromères. En effet, l'une des caractéristiques de base des centromères est leur caractère

fortement répété ayant pour principale origine des séquences répétées en tandem ayant elle-même pour origine des éléments transposables qui se sont accumulés dans ces régions. Mais c'est également le cas pour les péricentromères qui sont des régions fortement répétées de l'hétérochromatine qui ont un rôle central dans la stabilité génétique et notamment la délimitation des centromères et la migration correcte des chromatides lors de la division cellulaire.

1.1.4.2 Autres impacts structuraux : réarrangements chromosomiques, recombinaison et duplication de gènes

Un autre impact structurel majeur de l'insertion d'ET sur les génomes est de pouvoir conduire à de grands changements dans l'architecture chromosomique à la fois en embarquant des gènes ou des fragments de gènes lors de leur recombinaison, mais également en induisant des recombinaisons chromosomiques (Feschotte, C. & Pritham, E. J., 2007).

Les réarrangements chromosomiques, pouvant impliquer un certain nombre de mécanismes affectant la structure de l'ADN, participent grandement à l'évolution des génomes. On comprend alors facilement que les ET, par leur présence et leur activité, peuvent faciliter ces réarrangements comme on peut l'observer au sein du génome du maïs où la transposition aberrante d'un élément Ac peut conduire à des délétions, inversions, translocations ou d'autres réarrangements (Yu, C., Zhang, J. & Peterson, T., 2011).

Ce type de mouvements, notamment favorisé par les ET, peut permettre l'insertion et la mobilisation de gènes ou fragments de gènes dans de nouveaux contextes chromosomiques, altérant parfois leur régulation et pouvant alors conduire à l'apparition de nouveaux traits phénotypiques. C'est par exemple le cas pour la tomate, qui, par l'expression tissu-spécifique du gène *IQD12* suite à sa rétrotransposition, va arborer un profil allongé et non rond (Xiao, H., Jiang, N., Schaffner, E. J. & van der Knaap, E., 2008). Ces réarrangements peuvent également conduire à la formation « d'îlots », c'est-à-dire un ensemble de séquences regroupant à la fois des gènes et des ET, qui ont parfois leur propre système de régulation. Ce phénomène est notamment observé dans le cas des gènes de résistance aux maladies, comme les gènes de résistance (R) qui sont fréquemment sujets aux duplications (Cerbin, S. & Jiang, N., 2018) et aux transpositions (Freeling, M. *et al*. 2008).

1.1.5 Impacts sur l'expression des gènes

En plus de participer aux variations de taille et d'organisation des génomes, les ET sont également impliqués dans la régulation de certains gènes et donc dans l'évolution, non seulement de la structure des génomes, mais également de leur physiologie et de la morphologie des organismes vivants (Lisch, D., 2012) (Rebollo, R., Romanish, M. T. & Mager, D. L., 2012) (Hirsch, C. D. & Springer, N. M., 2017).

1.1.5.1 Les mécanismes de régulation de la transcription

La régulation de l'expression des gènes regroupe l'ensemble des mécanismes de régulation permettant de passer d'une information génétique, contenue dans une séquence ADN, à un ARN ou une protéine fonctionnelle (Campbell, N. A. & Reece, J. B., 2004). Cette régulation peut donc s'exercer à n'importe quelle étape permettant l'expression d'un gène, c'est-à-dire : le déroulement de l'ADN, la transcription, la maturation et la traduction des ARNm. Toutes ces étapes ne sont pas toujours nécessaires, mais elles peuvent être la cible des régulations pour activer ou désactiver, augmenter ou diminuer l'expression d'un gène.

1.1.5.1.1 Les modifications de la chromatine

Dans le noyau des cellules eucaryotes, l'ADN se trouve sous une forme compacte et organisée, nommée chromatine. Celle-ci est constituée de l'association de l'ADN, d'ARN et de protéines, principalement des histones. Afin de former cette chromatine, des segments de 146 paires de bases d'ADN s'enroulent autour d'un assemblage de 8 molécules d'histones, appelé nucléosome. Les nucléosomes s'enchaînent alors sur l'ADN pour constituer une structure en collier de perles, structure que l'on observe peu dans les cellules car d'autres histones s'additionnent afin de former les structures plus compactes que l'on nomme la fibre de 10 nm puis la fibre de 30 nm, qui constituent l'unité de base de la chromatine (Figure 10) (Kornberg, R. D., 1977) (McGhee, J. D. & Felsenfeld, G., 1980). On distingue alors deux formes de chromatine (Horn, P. J., 2002) (Németh, A. & Längst, G., 2004) : l'euchromatine qui est la forme la moins condensée où l'ADN est accessible afin de permettre l'expression des gènes, et l'hétérochromatine, forme dense de la chromatine qui ne permet pas l'accès à

l'ADN. L'activité biologique de ces deux formes de chromatine est liée à des marques épigénétiques (Pikaard, C. S. & Mittelsten Scheid, O., 2014) inscrites à la fois dans l'ADN, sous forme de méthylation des bases à certains sites, et dans les histones qui peuvent porter également des modifications au sein des nucléosomes. Ces marques, jouent un rôle important dans de nombreux processus biologiques (Vanyushin, B. F. & Ashapkin, V. V., 2011).

Chez les eucaryotes, la modification de l'ADN la plus connue et la plus fréquente est donc la méthylation. Cette modification consiste en l'addition d'un groupement méthyle sur certaines cytosines de l'ADN à l'issue de la réplication (He, X.-J., Chen, T. & Zhu, J.-K.., 2011) (Pikaard, C. S. & Mittelsten Scheid, O., 2014). Cette méthylation, lorsque l'on compare des gènes identiques au sein des différents types cellulaires, est majoritairement observée dans le cas où les gènes ne sont pas exprimés (Tirado-Magallanes, R., Rebbani, K., Lim, R., Pradhan, S. & Benoukraf, T., 2017). En effet, la déméthylation, c'est-à-dire l'élimination des groupements méthyle, a pour effet d'activer des gènes auparavant inactifs. Chez les plantes, la méthylation a été observée dans les contextes nucléotidiques : CG, CHG, CHH (C = cytosine, G = guanine, et H représentant tous les nucléotides excepté le G) et se trouve transmise au nouveau brin d'ADN lors de la réplication, bien qu'elle puisse également toucher asymétriquement l'ADN et être produite *de novo* grâce à la voie de méthylation de l'ADN dirigée par les ARN (RNA-directed DNA methylation, RdDM) qui fait intervenir les siRNA et longs ARN non-codants pour cibler les séquences à méthyler (Niederhuth, C. E. & Schmitz, R. J., 2017) (He, X.-J., Chen, T. & Zhu, J.-K., 2011).

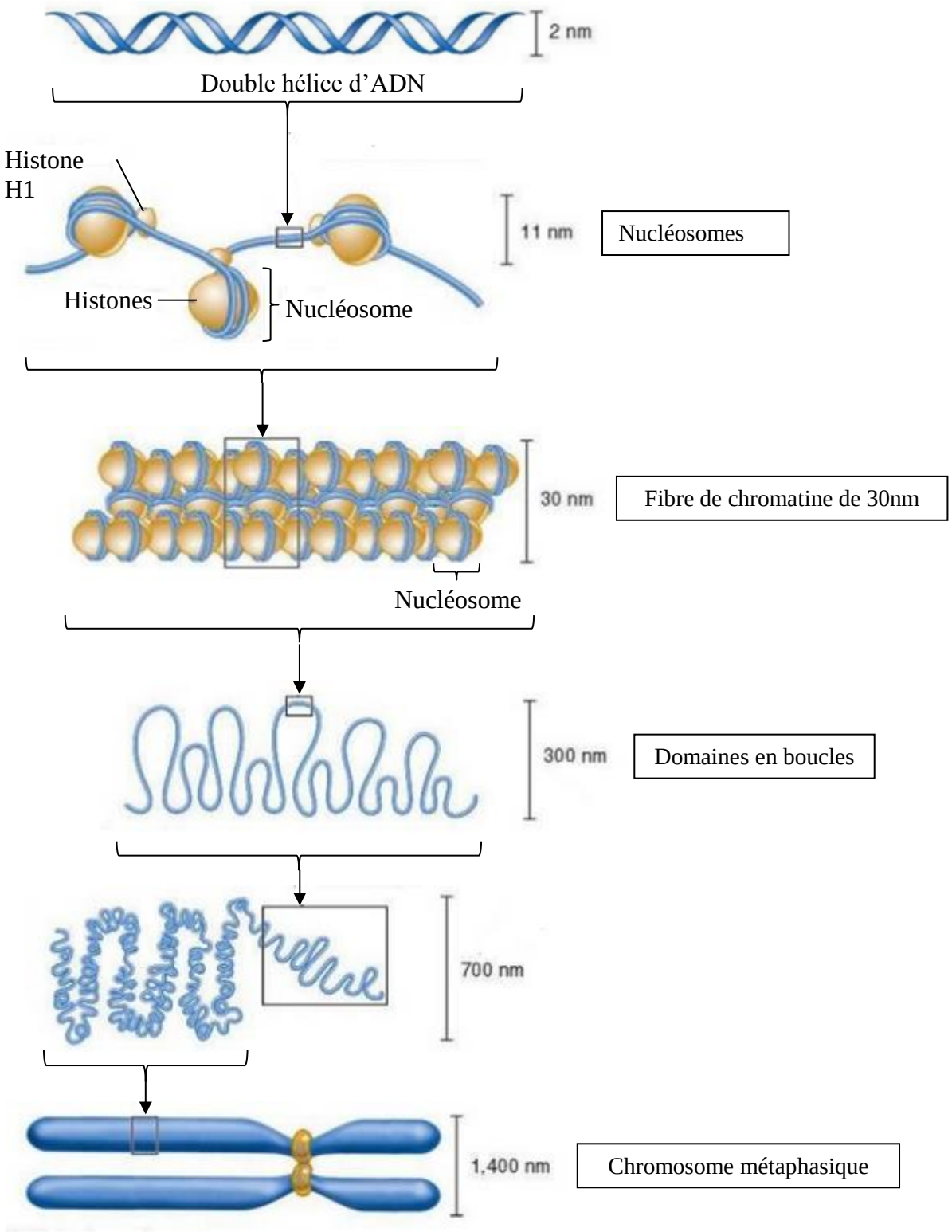


Figure 10 : Les différents niveaux de compaction de la chromatine. Les différents schémas montrent les différents niveaux de compaction de l'ADN selon le modèle actuel de repliement de l'ADN permettant d'obtenir un haut niveau de compaction dans les chromosomes métaphasiques.

Adapté d'après Campbell, N. A. & Reece, J. B., 2004 (Figure 19.1 page 285)

Cependant, ces processus de régulation ne servent pas uniquement au contrôle des gènes. En effet, sachant que les ET peuvent transposer à n'importe quel endroit du génome, et notamment à proximité des gènes ou de leurs éléments de contrôle, on comprend facilement l'importance de contrôler leur transposition ainsi que leur expression afin qu'ils n'interfèrent pas dans le processus d'expression des gènes et donc le bon fonctionnement de la cellule. Pour répondre à cette nécessité, les organismes eucaryotes utilisent différents processus, et justement la voie épigénétique, incluant entre autres comme nous venons de le voir, la modification des queues d'histones, la méthylation de l'ADN ou encore les altérations de la condensation de la chromatine (Slotkin, R. K. & Martienssen, R., 2007). La voie de contrôle la plus connue pour ces éléments reste alors la méthylation de l'ADN. Dans ce cas, les cytosines des ET sont ciblées par les mécanismes de méthylation et notamment grâce à la voie impliquant les siRNA. Bien que l'on observe généralement une corrélation négative entre la présence d'ET méthylés et l'expression des gènes à proximité (Hollister, J. D. & Gaut, B. S., 2009), ceci étant probablement lié aux modifications de conformation de la chromatine, qui se condense en présence de méthylations, on observe parfois des ET actifs et hypométhylés dans ces mêmes régions, pouvant alors eux-mêmes participer à la régulation de l'expression génique.

Finalement, bien que la méthylation soit l'élément central des mécanismes de régulation des génomes, d'autres systèmes existent. C'est par exemple le cas des modifications des histones, incluant l'acétylation, la phosphorylation ou encore la sumoylation. Mais seule la méthylation touche directement les ET du génome.

1.1.5.1.2 Les protéines et séquences de régulation de l'initiation de la transcription

Afin de réguler plus finement l'expression de ses gènes, la cellule utilise des protéines que l'on nomme facteurs de transcription, qui vont interagir avec des séquences d'ADN spécifiques à la régulation de certains gènes (Moore, J. W., Loake, G. J. & Spoel, S. H., 2011). Le passage d'une séquence d'ADN à une molécule d'ARNm (Figure 11), et plus particulièrement à une ARN prémessager, est appelé transcription. Au cours de cette étape, différentes protéines, nommées facteurs de transcription, vont se regrouper et se fixer sur le promoteur du gène, situé à l'extrémité en amont du gène, et former le complexe d'initiation de la transcription. Au sein de ce complexe d'initiation, l'ARN polymérase prend en charge la

transcription du gène, c'est-à-dire une synthèse d'ARN à partir d'un ADN. D'autres étapes, excision des introns, ajout d'une coiffe 5' et d'une queue poly(A), regroupées sous le terme de maturation de l'ARN, viennent ensuite pour transformer cet ARN prémessager en ARNm.
Mais, dans l'ADN, la séquence du gène est souvent accompagnée d'un grand nombre d'éléments de contrôle non codants (Bilas, R., Szafran, K., Hnatuszko-Konba, K. & Kononowicz, A. K., 2016) qui vont permettre, en liant les facteurs de transcription, de réguler l'expression d'un gène donné. En effet, les interactions entre les facteurs de transcription, l'ARN polymérase et le promoteur ne permettent qu'un taux d'initiation de la transcription peu élevé et donc la synthèse d'un faible nombre d'ARN. Ce sont donc les éléments de contrôle proximaux (situés à proximité du promoteur du gène) et distaux (situés à des distances importantes du promoteur, ou bien en aval du gène ou encore à l'intérieur d'un intron), également appelés amplificateurs qui vont augmenter le débit de transcription eucaryote en se liant à des facteurs de transcription supplémentaires afin de rendre les promoteurs beaucoup plus efficaces. Ces amplificateurs, appelés activateurs, ou *enhancer*, s'ils stimulent la transcription du gène, et répresseurs, ou *silencer* s'ils la diminuent, peuvent interagir avec le complexe d'initiation assemblé sur le promoteur grâce à la courbure de l'ADN.

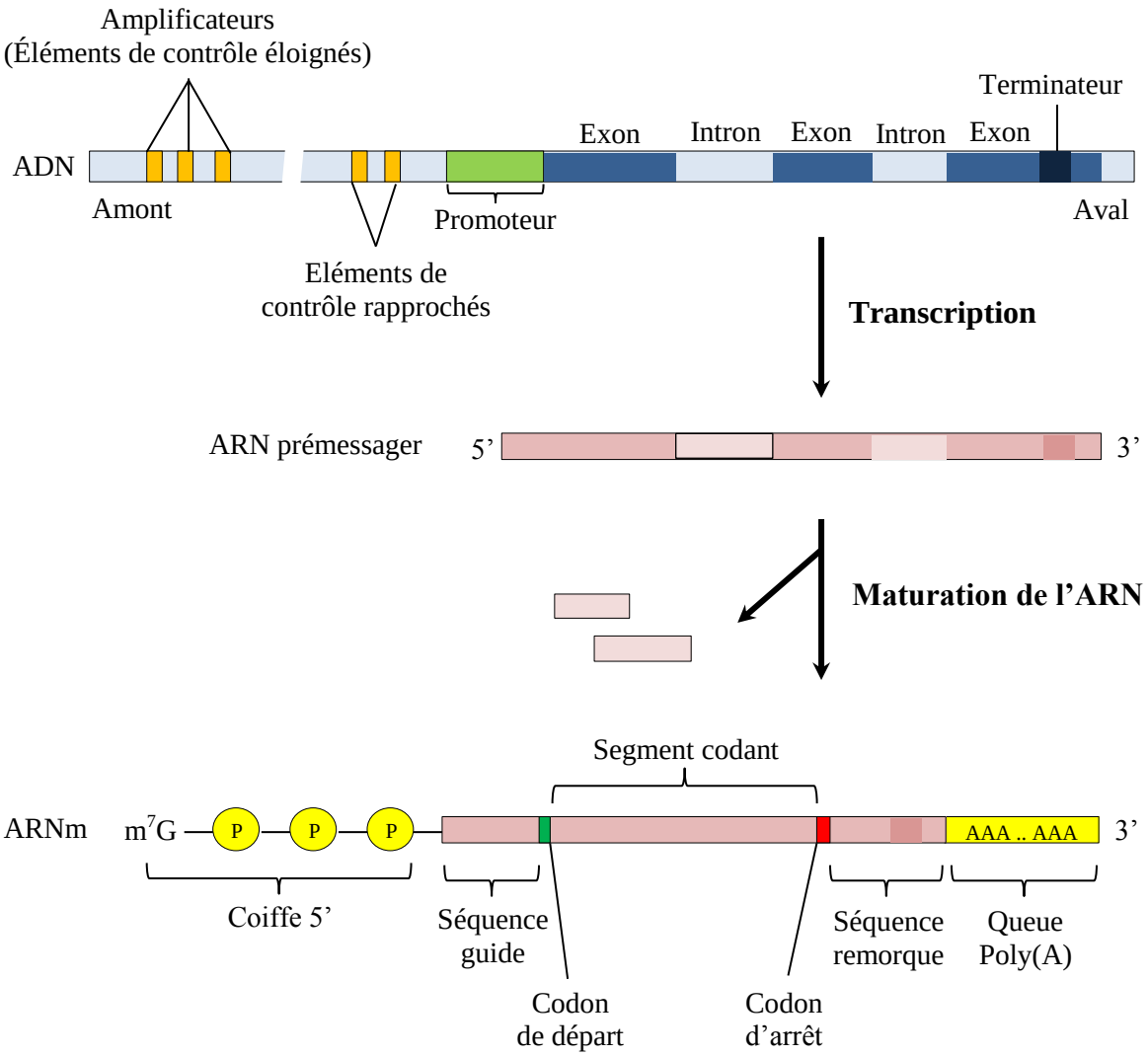


Figure 11 : Structure d'un gène eucaryote, de son ARN prémessager et de son ARN messager. Lors de la maturation, les facteurs de transcription ajoutent une coiffe à l'extrémité 5' du transcrit ainsi qu'une queue poly(A) à l'extrémité 3', et se chargent également de l'excision des introns et de l'épissage des exons.

Adapté d'après Campbell, N. A. & Reece, J. B., 2004 (Figure 19.8 page 394)

1.1.5.2 L'inactivation de gènes par les éléments transposables

La conséquence la plus connue et la mieux caractérisée de l'insertion d'un ET, dans ou à proximité d'un gène, est probablement l'inactivation de ce gène. Cette inactivation du gène génère des mutations nulles, parfois à des taux élevés, mais il ne faut pas oublier que la plupart des mutations nulles observables dans un génome, ne sont pas dues à l'insertion d'un

ET. Cette possibilité de générer des mutations nulles est depuis longtemps exploitée et pourrait notamment avoir eu un rôle dans la domestication des plantes.

L'un des meilleurs exemples permettant d'illustrer cette inactivation de gène se trouve probablement au sein des différentes variétés de raisin (Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H., 2004). En effet, l'insertion d'un rétroélément à LTR, absent dans le génome de la variété *Cabernet*, résulte en une perte de fonction du gène *Vvmby1A* qui conduit, chez la variété *Chardonnay*, à une perte de la coloration du fruit (Figure 12).

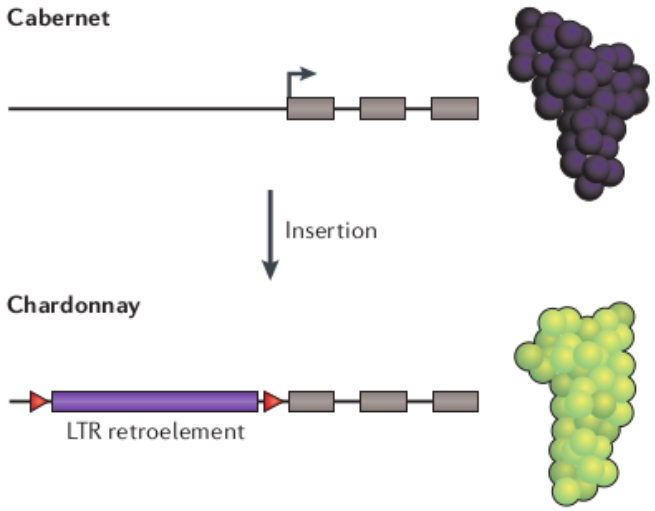


Figure 12 : Effet de l'insertion d'un élément transposable dans le génome du raisin. L'insertion d'un rétroélément à LTR entraîne la perte de coloration du fruit observé chez la variété *Chardonnay*.
Adapté d'après Lisch, D. *et al*., 2012

1.1.5.3 Reprogrammation de l'expression des gènes

Un autre impact possible de l'insertion d'un ET est la modification de la régulation des gènes qui peut s'effectuer de différentes manières.

Tout d'abord, l'insertion d'un élément répété peut avoir un impact sur les fonctions régulatrices positives ou négatives des gènes. Cet impact se produit généralement par l'élimination de ces fonctions de régulations, notamment en supprimant l'activité des *enhancers* et *repressors*. Un exemple connu d'une telle insertion peut être observé chez le maïs (Greene, B., Wako', R. & Hake, S., 1994), au niveau du phénotype des feuilles qui présentent des différences de lobes. Ce trait est directement lié à l'insertion, dans le premier

intron du gène *knotted1*, d'un élément *Mutator* dans une séquence non-codante conservée (CNS).

Ensuite, il est possible que la transposition d'un ET à une nouvelle position dans le génome, introduise une nouvelle information. Cette nouvelle information peut avoir différents impacts sur les gènes situés à proximité. Dans le cas où la séquence de l'élément inséré contient un module de régulation ayant des propriétés fonctionnelles similaires à celles d'un module déjà présent, alors l'organisme peut tolérer une duplication de ces propriétés car cela n'entraîne pas d'effet négatif pour la plante. A long terme, si l'un des modules est perdu, il s'agit environ une fois sur deux, de celui d'origine. Si de tels modules s'intègrent simultanément devant différents gènes participant à un même processus biologique au sein de la plante, cela peut participer à la création de nouveaux réseaux de gènes co-régulés (Bennetzen, J. L. & Wang, H., 2014) (Feschotte, C., 2008).

Une autre altération possible de l'expression des gènes, notamment si l'insertion de l'ET a lieu dans leur extrémité 5' de ceux-ci peut se traduire par un changement dans l'expression du gène qui sera alors exprimé de manière ectopique. On trouve l'exemple de ce type d'expression chez le maïs en comparant les espèces *B. peru* et *B. bolivia* (Dooner, H. K., Robbins, T. P. & Jorgensen, R. A., 1001) (Selinger, D. A. & Chandler, V. L., 1999) (Selinger, D. A. , 2001). En effet, une insertion de 2,5kb dans le premier exon du gène *b1* de l'espèce *B. peru* entraîne une coloration des grains sur l'épi (Figure 13a). Une seconde insertion, d'un élément de plus de 7kb, dans la première, conduit, elle, à l'apparition de la variété *B. bolivia* dont le phénotype de coloration des grains est atténué. Cependant, l'une des particularités des ET est leur capacité à répondre à une grande variété de signaux. Cette aptitude peut alors conduire à l'apparition de traits sélectifs, ce qui peut être observé en comparant les différentes variétés d'orange (Figure 13b) (Butelli, E. *et al.*, 2012). Ainsi, l'orange à chair rouge, de type *Tarocco* découle d'une orange *Navalina* à chair orange ayant subi l'insertion d'un rétrotransposon, de même que la variété *Jingxian*. Ces deux insertions indépendantes, confèrent des propriétés similaires aux deux variétés qui vont exprimer de façon tissue spécifique le gène *Ruby* en réponse au froid. Une recombinaison des LTRs du transposon de *Tarocco* a ensuite entraîné l'apparition d'une variété au phénotype aggravé, l'orange *Maro(I)*. Cette sensibilité des ET à différents stimuli, et notamment aux stress, à la fois biotiques et abiotiques – incluant le sel, la sécheresse, le froid et la chaleur, mais également l'infection par les bactéries et les virus – peut donc, dans le cas où l'insertion a lieu en upstream des gènes de l'organisme hôte, conférer à ces gènes une capacité de réponse au stress (Ivashuta, S. *et al.*,

2002) (Le, T.-N. *et al.*, 2014) (Makarevitch, L. *et al.*, 2015).

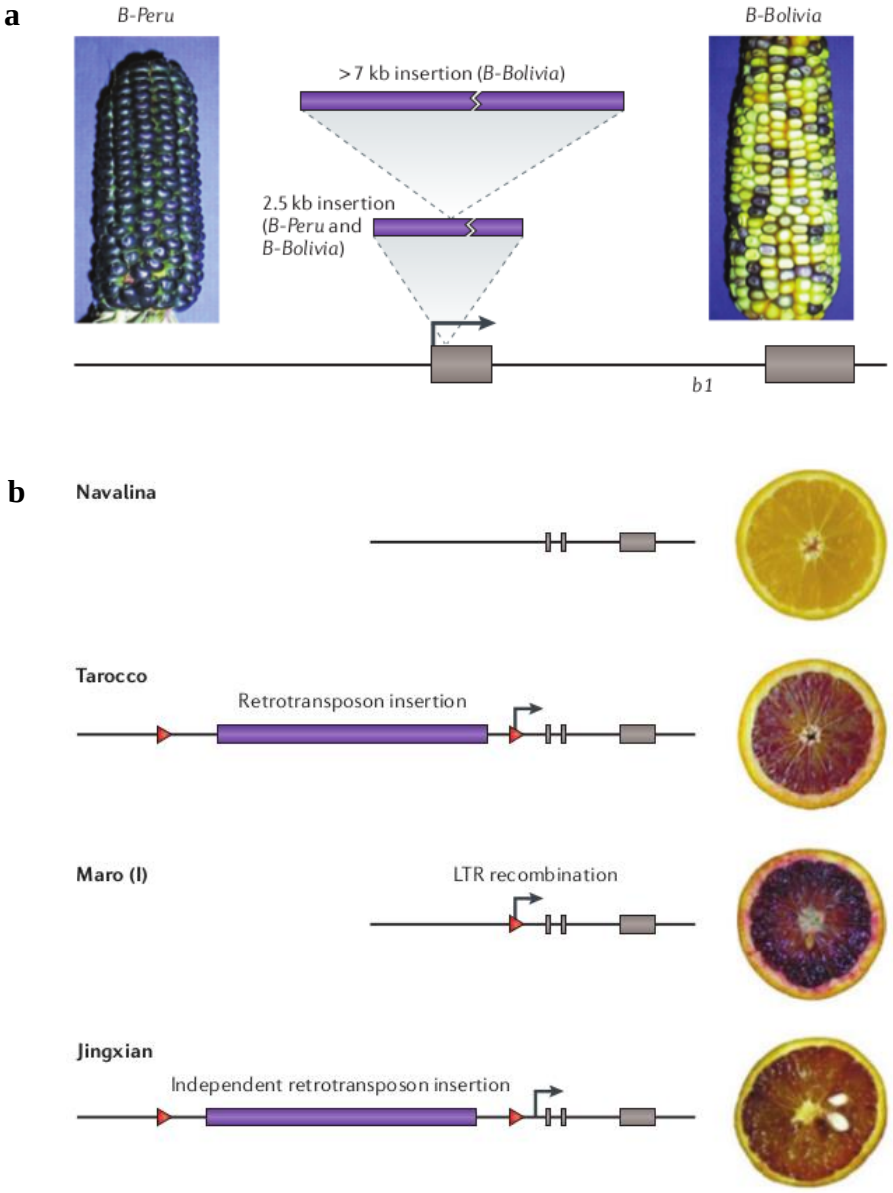


Figure 13 : Effets tissus spécifiques et en réponse au froid de l'insertion d'éléments transposables. a. Effet de deux insertions consécutives sur la coloration des grains de maïs. b. Les insertions indépendantes d'un rétroélément à LTR dans les variétés *Tarocco* et *Jingxian* en upstream du gène *Ruby* entraîne une coloration de l'orange en réponse au froid. La recombinaison de l'élément dans la variété *Maro* entraîne elle une aggravation du phénotype.

Adapté d'après Lisch, D. *et al.*, 2012

1.1.5.4 Exaptation de séquences codantes et mise en place de réseaux de régulation

En évolution, l'exaptation d'une séquence est la sélection d'une adaptation génétique opportuniste, permettant de favoriser les caractères utiles à une nouvelle fonction. Cette exaptation peut permettre à certaines séquences codées par les ET de se maintenir dans les génomes. En effet, en plus d'informations régulatrices, les ET autonomes codent des protéines nécessaires à leur transcription. Ces séquences protéiques peuvent alors être sélectionnées positivement pour favoriser de nouvelles fonctions (Lin, R. *et al*., 2007).

Partant de ce postulat, on peut supposer que si une même séquence d'ET est retenue à une même position chromosomique au sein des génomes de plusieurs espèces plus ou moins proches, alors cette séquence peut potentiellement avoir un impact sur le fonctionnement de son génome hôte. Des fragments d'ET peuvent également être combinés avec des gènes de l'hôte, formant alors de nouveaux gènes chimériques, susceptibles d'être eux aussi maintenus par l'hôte (Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H. & Spillane, C., 2011).

Sachant que certaines séquences d'ET peuvent être sélectionnées et maintenues dans les génomes, on peut s'interroger sur le cas de l'expression co-régulée de différents gènes ayant une fonction apparentée ou participant à un même processus biologique. En effet, dans chaque génome, il existe des dizaines de milliers de gènes. On ne dénombre pourtant que quelques centaines de facteurs de régulation et les éléments de contrôle de l'expression de ces gènes ne comptent que peu de séquences nucléotidiques entièrement différentes. Compte tenu du peu de diversité dans les éléments de régulation, on peut supposer que la présence d'un élément de contrôle donné est moins importante pour le contrôle de l'expression génique que la présence combinée de plusieurs de ces éléments de contrôle. On sait, chez les procaryotes, que les gènes co-régulés sont organisés en opérons (Osbourn, A. E. & Field, B., 2009), c'est-à-dire que les gènes occupent des positions adjacentes sur la molécule d'ADN et qu'ils partagent les mêmes éléments de régulation transcriptionnelle, incluant un même promoteur. De tels opérons n'ont cependant que très rarement été observés dans les organismes eucaryotes. Il a cependant été mis en évidence, pour ces organismes, que les gènes co-régulés partagent les mêmes environnements de contrôle (Michalak, P., 2008), qui ont probablement pour origine des duplications et disséminations à différentes positions du génome d'un même élément de contrôle, cet élément pouvant alors être une séquence d'ET.

1.1.5.5 Modifications épigénétiques

Comme vu précédemment, il arrive parfois que l'insertion d'un ET apporte un avantage sélectif à l'hôte. Cependant, la plupart de ces insertions n'auront aucun impact sur le génome (et seront donc neutre), ou auront un effet négatif. C'est pour cela que les cellules mettent en place d'importants mécanismes pour contrôler les ET. Ce contrôle est principalement réalisé par l'intervention de petits ARN, la méthylation de l'ADN, mais également différentes modifications d'histones, l'ensemble étant regroupé comme processus de *silencing* épigénétique (Pikaard, C. S. & Mittelsten Scheid, O., 2014).

Il est alors logique de penser que ces modifications, visant à réguler les ET, vont également impacter la régulation de l'expression des gènes. On observe néanmoins que la plupart des gènes fonctionnent aussi bien, quel que soit le génome étudié, ce qui nous laisse penser que les organismes eucaryotes sont capables de garder l'équilibre de régulation nécessaire pour garder les ET inactifs et les gènes actifs. Cette observation répond à la définition de la sélection, c'est-à-dire que les variations favorables sont préservées alors que celles qui sont préjudiciables sont rejetées, éteintes. Ainsi, l'organisme ayant besoin de ses gènes pour fonctionner normalement va les maintenir actifs, alors que la majorité des copies d'ET seront inactivées dans la mesure où elles n'apportent pas d'intérêt à l'hôte.

Bien que cette situation soit celle observée le plus fréquemment, il existe des exceptions au sein desquelles on constate que la régulation épigénétique d'un ET peut avoir un effet important sur l'expression des gènes situés à proximité. C'est par exemple le cas chez *Arabidopsis thaliana*, pour le locus FWA (*Flowering Wagenigen*) (Kinoshita, Y. *et al*., 2006) (Fujimoto, R., Kinoshita, Y., Kawabe, A., Kinoshita, T. & Takashima, K., 2008). En effet, la transcription de ce gène débute dans un ET de type SINE qui est porteur d'une courte duplication en tandem. Les petits ARN et la méthylation de l'ADN régulent négativement l'expression de ce gène dans le tissu végétatifs, mais une mutation altérant les petits ARN ou la méthylation, permet l'expression ectopique de FWA ce qui entraîne un phénotype de floraison tardive chez cette plante.

1.1.5.6 Le devenir des insertions

Selon la localisation et l'impact de l'insertion sur le génome hôte, le devenir de l'élément à la

fois sur le court et sur le long terme va varier. Ainsi, le devenir d'une insertion, qu'elle soit délétère, neutre ou adaptative, va être différent (Figure 14).

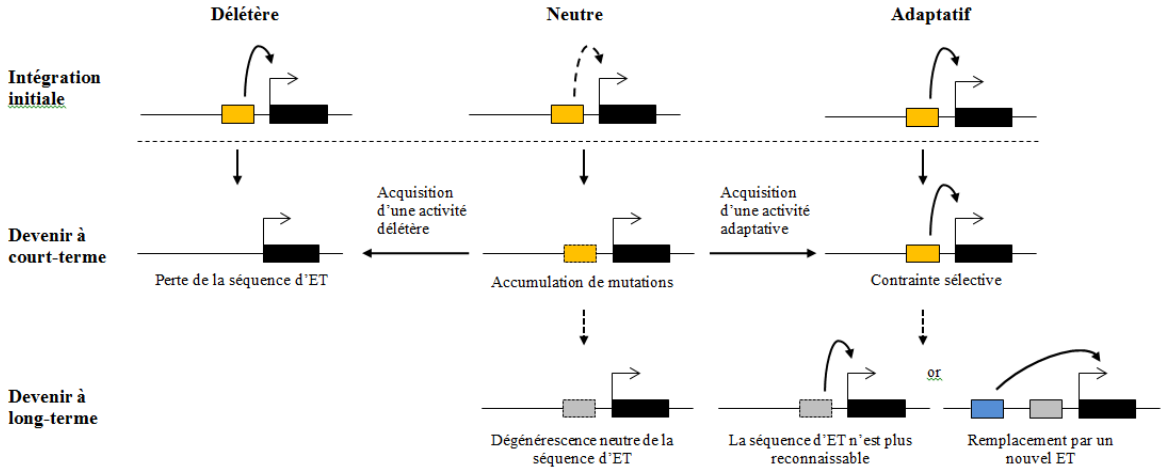


Figure 14 : Schéma du devenir des insertions d'éléments transposables en fonction de leur impact sur les gènes situés à proximité. Les rectangles jaunes, bleu et gris représentent des ET alors que les rectangles noirs représentent les gènes.
D'après Chuong, E. B., 2016

Dans le cas où l'insertion de l'élément est délétère, et a donc un effet négatif sur le fonctionnement de l'hôte, l'organisme va l'éliminer, notamment en entraînant une dégénérescence de la séquence par mutations ou directement par excision de celle-ci. Cependant, l'impact de la sélection (positive ou négative) sur un gène dépend également du taux de recombinaison de son environnement. La sélection sera alors plus efficace quand le taux de recombinaison sera élevé.

Si l'élément a, en revanche, un effet adaptatif sur l'hôte et lui confère un avantage sélectif, alors une contrainte sélective est appliquée sur sa séquence. Cette contrainte va permettre de maintenir la fonction acquise tout en laissant le reste de la séquence, non essentielle, dégénérer. A terme, la séquence de l'ET n'est plus reconnaissable (seul le court fragment d'intérêt n'est pas muté), ou est remplacé par un nouvel ET apportant cette même fonction.

Pour les insertions neutres, qui n'entraînent aucun impact sur le génome hôte, leurs séquences vont accumuler des mutations. Cela peut entraîner trois situations distinctes : la séquence va acquérir une activité délétère et alors être éliminée, ou la dégénérescence neutre de la séquence de l'ET va se poursuivre, ou enfin, la séquence de l'élément va acquérir une activité

adaptative qui sera ensuite sujette à une contrainte sélective. Ces insertions neutres vont donc potentiellement s'accumuler dans le génome et vont à terme, constituer la matière noire des génomes (Maumus, F. & Quesneville, H., 2014).

1.1.5.7 Eléments transposables et théories de l'adaptation et de la domestication

Les éléments transposables peuvent donc avoir de nombreux impacts sur les génomes de plantes et ils peuvent alors, en fonction de ceux-ci, évoluer différemment au cours du temps. Mais l'un des aspects les plus remarquables de ces séquences et leur capacité à permettre au génome hôte une grande adaptabilité à différents changements.

La notion d'adaptation est abordée lorsqu'un organisme vivant subit une modification afin de rester fonctionnel dans de nouvelles conditions. Dans le domaine de la biologie, l'adaptation regroupe l'ensemble des modifications héréditaires découlant d'un changement d'environnement (modification du climat, nouveau prédateur…) et qui sont apparues en quelques milliers à plusieurs centaines de millions d'années. De tels processus peuvent alors être le résultat de mouvements des ET de l'organisme, qui vont servir de régulateurs aux gènes, comme cela peut par exemple être observé dans le génome du maïs. En effet, de nombreux gènes de réponse au stress de cet organisme sont précédés d'un élément transposable qui va être activé en cas de nécessité et permettre la réponse de l'organisme au stress environnemental auquel il est soumis (Makarevitch , I. *et al*., 2015). Ces adaptations se font de façon naturelle et relèvent donc de ce que l'on nomme la *sélection naturelle*.

Dans d'autres cas, les plantes ont subi une modification de leur patrimoine génétique suite à l'intervention de l'Homme qui souhaitait que celle-ci réponde à ses besoins. On parlera alors de domestication et cela comprend à la fois l'isolement d'une population ayant un trait phénotypique souhaité, le changement du génome ou encore la création d'une nouvelle espèce. On va pouvoir observer dans de telles domestications, la sélection d'éléments transposables pour favoriser un trait phénotypique dans différentes espèces, principalement celles d'intérêt agronomique comme par exemple chez le riz, où les espèces cultivées présentent une répartition des éléments transposables et une utilisation de leur séquence très différente de celle des souches sauvages (Li, X. *et al*., 2017). Cette domestication se retrouve aussi dans de nombreuses autres espèces comme l'orange sanguine, où un élément contrôle l'expression du

gène *ruby* responsable de la coloration, ou chez le maïs où une répétition engendre la dominance apicale de l'épi (Chuong, E. B., 2016).

1.2 Les *Solanaceae*, une ressource variée pour l'étude des plantes

1.2.1 La famille des Solanaceae

Les *Solanaceae* sont une famille de plantes qui regroupent 147 genres et environ 3 000 espèces. Ce groupe est le troisième taxon de plantes le plus important économiquement et en termes de produits agricoles. En effet, il regroupe un certain nombre de légumes fruitiers bien connus comme les tomates, les pommes de terre, les aubergines ou les poivrons, des plantes ornementales (pétunias), des plantes à feuilles comestibles (*Solanum aethiopicum, S. macrocarpon*) et des plantes médicinales (Datura, Capsicum).

	Nom latin	Nom commun	Génome
	*Solanum lycopersicum*	Tomate ronde Heinz ou M82	950 Mb
	*Solanum pimpinellifolium*	Tomate groseille	715 Mb
	*Solanum pennellii*	Tomate sauvage	980 Mb
	*Solanum tuberosum*	Pomme de terre	840 Mb
	*Solanum melongana*	Aubergine	850 Mb
	*Capsicum annuum*	Piments et poivrons	950 Mb
	*Nicotiana tabacum*	Tabac	4,5 Gb
	*Nicotiana benthamiana*	Tabac indigène d'Australie	3,5 Gb
	*Petunia axillaris*	Pétunia ancêtre de *P. hybrida*	1,3 Gb
	*Petunia inflata*	Pétunia ancêtre de *P. hybrida*	1,3 Gb

Tableau 1 : Tableau des espèces de Solanaceae séquencées.
Source : *https://solgenomics.net*

Cette famille présente donc une grande variabilité, à la fois en termes d'intérêt agricole et

économique, mais aussi au niveau du phénotype ou encore à l'échelle des génomes (Tableau 1) puisqu'il regroupe des espèces à fruits et à fleurs, sauvages ou cultivées, avec des génomes allant d'environ 700 Mb pour *Solanum pimpinellifolium* à 4,5 Gb pour *Nicotiana tabacum*. Cette diversité est due à la fois à des mécanismes naturels d'évolution, guidés par l'adaptation des espèces à leur environnement et notamment au climat, mais également à une sélection humaine intensive. Ainsi, le niveau d'évolution et d'adaptation des *Solanaceae*, couplé cependant à un haut niveau de conservation de l'organisation des génomes (Tanksley, S. D., Bernatzky, R., Lapitan, N. L. & Prince, J. P., 1988) (Doganlar, S., Frary, A., Daunay, M.-C., Mester, R. N. & Tanksley, S. D., 2002) (Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N., & Tanksley, S. D., 2002), font de cette famille un modèle pour explorer les bases de la diversité phénotypique et de l'adaptation à différents milieux.

Finalement, certaines espèces de *Solanaceae* sont reconnues comme étant elles-mêmes des systèmes modèles importants pour la biologie. On trouve notamment parmi elles : la tomate *Solanum lycopersicum* pour l'étude de la maturation des fruits et la défense des plantes, le tabac pour la défense de la plante, et les pétunias pour la biologie des pigments anthocyanes.

1.2.2 *Solanum lycopersicum* comme organisme modèle

1.2.2.1 Généralités

Parmi les différentes espèces de *Solanaceae* citées précédemment, il a été choisi d'étudier l'espèce *Solanum lycopersicum*, appelée plus généralement tomate commune, qui est définie comme l'une des plantes modèles pour de nombreux domaines.

Tout d'abord, *Solanum lycopersicum* est une plante herbacée originaire du Nord-Ouest de l'Amérique du Sud, largement cultivée pour son fruit qui est l'un des plus populaires dans le monde. Cette espèce de *Solanum* regroupe différentes variétés botaniques, dont la « tomate-cerise », qui peuvent être cultivées soit en plein champs, soit sous abri, dans pratiquement toutes les régions du monde (Figure 15). Ces facilités de culture et ces différentes possibilités de consommation ont entraîné le développement d'une importante industrie de transformation, puisqu'elle sert notamment pour la production de concentrés, de sauces comme le "ketchup", mais aussi pour des jus et des conserves. Mais bien qu'elle serve dans de nombreux produits transformés, elle peut également être consommée crue ou cuite, en tant que légume-fruit et est,

par ce fait, devenue un ingrédient incontournable de la gastronomie dans de nombreux pays.

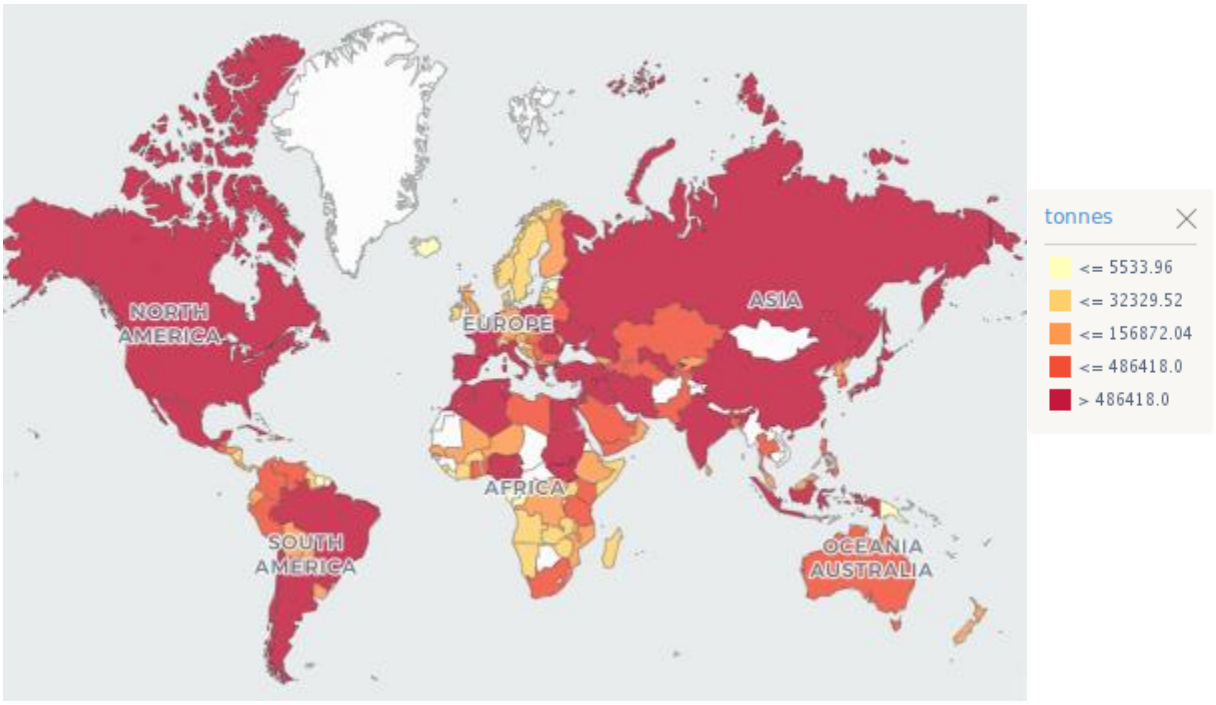


Figure 15 : Carte mondiale des quantités de production de tomates par pays en 2016.
D'après *www.fao.org/faostat/fr/#data/QC/visualize*

On constate, de ce fait, que la tomate *Solanum lycopersicum* représente un enjeu, à la fois agronomique et économique, ce qui participe à faire d'elle un objet de recherches scientifiques privilégié. On la considère notamment comme une plante modèle en génétique, mais également pour l'analyse du développement des fruits ou leurs processus de maturation (Chen, Y.-R. *et al.*, 2013), la tolérance à différents stress, ou encore l'impact des maladies des plantes.

De nombreux moyens ont donc été mobilisés afin de mettre à disposition des chercheurs des ressources importantes et variées. Ainsi, la première séquence complète du génome diploïde de la variété Heinz 1706 a été disponible en 2012 (The Tomato Genome Consortium, 2012) puis elle a été améliorée au fur et à mesure des avancées technologiques. D'autres données, confortant son positionnement comme organisme modèle ont également été rendues publiques : de nombreux marqueurs génétiques et re-séquençages, des collections de mutants, les nombreuses possibilités de croisements, des données épigénétiques (incluant le méthylome), ou encore certains réseaux de régulation de gènes comme celui de la maturation du fruit.

1.2.2.2 Le génome de Solanum lycopersicum

Notre modèle d'étude, *Solanum lycopersicum*, possède un génome diploïde, composé de 12 chromosomes, désormais assez bien caractérisé (Causse, M. *et al*., 2016).

Différentes versions de ce génome sont disponibles (https://solgenomics.net/organism/Solanum_lycopersicum/genome) avec chacune leurs spécificités et de plus en plus complètes grâce à la progression des technologies. Cependant, certaines caractéristiques ne changent pas en fonction des versions du génome. Ainsi, ce génome est, comme dit précédemment, composé de 12 chromosomes qui portent un total d'environ 35 000 gènes. L'importance du nombre de gènes par rapport à la taille du génome peut toutefois être surprenante. Cette caractéristique semble pouvoir être expliquée par l'histoire évolutive du génome qui a subi deux triplications (paléo-hexaploïdie) (The Tomato Genome Consortium, 2012) au cours de son histoire. Lors de tels évènements, l'ensemble du contenu génétique va être copié et une partie seulement sera éliminée, ce qui peut conduire à une redondance de la séquence, notamment des gènes.

Finalement, les analyses faites jusqu'à présent sur ce génome ont également montré qu'il contient un fort pourcentage en éléments répétés, ce qui en fait pour nous un très bon modèle d'étude.

1.2.3 Les ET de Solanum lycopersicum

Le génome de *Solanum lycopersicum*, est donc sujet à de nombreuses études, et les éléments transposables qu'il porte ont déjà été annotés à l'aide de différents outils. Ces premières analyses, ont permis de montrer que les ET sont un composant majeur de ce génome (Mehra, M., Gangwar, I. & Shankar, R., 2015) et que ceux-ci ont grandement participé à son évolution, tant sur le plan génomique, que sur le plan phénotypique (Xiao, H., Jiang, N., Schaffner, E. & Stockinger, E. J., 2008) (Seibt, K. M., Wenke, T., Muders, K., Truberg, B. & Schmidt, T., 2016).

Afin de mener à bien les analyses de cette thèse, il a été choisi de travailler sur la version Heinz 1706 v2.4 puis v2.5 (dès sa mise à disposition). Le passage à la version v2.5 de ce génome a été fait pour permettre, si nécessaire, l'utilisation des nombreuses ressources

disponibles pour cette version (reséquençages, variants…).

Afin de travailler sur ces génomes et de répondre aux questions soulevées dans ce projet, ils ont été annotés par Florian Maumus et moi-même à l'aide de l'outil REPET (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011), développé au sein de l'unité de recherche d'accueil. Par ce biais, le génome de la tomate s'est révélé être composé de 68% d'éléments répétés. Cela représente environ 535Mb de séquence constituée d'ET, regroupant l'ensemble des familles et superfamilles connues.

Cette nouvelle annotation permettra, au cours de la thèse, une étude approfondie, notamment de la composition en familles et superfamilles, mais aussi de leur répartition dans le génome et de leur impact au sein de celui-ci.

1.3 Objectifs de la thèse

Le travail de cette thèse porte sur les éléments répétés du génome de la tomate *S. lycopersicum*, et plus spécifiquement sur leur rôle au sein de ce génome. Deux grands objectifs ont donc été fixés. Le premier est de comprendre l'impact que peuvent avoir ces ET sur la régulation de l'expression des gènes, notamment lors du processus de maturation du fruit. Pour répondre à cette première question, le génome de la tomate est à notre disposition, de même que son annotation en gènes, mais également des données d'expression et de méthylation lors de différents stades de la maturation du fruit. Le second objectif est d'identifier des copies d'ET pouvant avoir été sélectionnées au cours de l'évolution pour leur impact sur les gènes situés à proximité. Pour se faire, nous repartirons de l'annotation des répétitions qui aura été faite sur le génome de la tomate et nous baserons sur les connaissances actuelles concernant les ET, mais aussi les gènes et la régulation de leur expression, afin de mettre au point une méthode de détection *in silico*.

Chapitre 1 : Eléments transposables et processus de maturation de la tomate *Solanum lycopersicum*

2.1 Introduction

L'insertion des éléments transposables dans les génomes va donc pouvoir impacter l'expression des gènes adjacents. Cette altération de l'expression génique peut être observée soit au niveau quantitatif, auquel cas la quantité de protéines produite sera plus ou moins élevée que nécessaire au bon fonctionnement de l'hôte, soit au niveau qualitatif, ce qui implique une modification des propriétés physico-chimique de la protéine produite et peut entraîner une perte de l'activité. Les effets visibles de ces modifications peuvent donc être de différentes natures. D'une part, il peut s'agir d'une modification de l'expression des gènes, notamment en entraînant des changements des séquences régulatrices telles les promoteurs ou les *enhancers* (Rebollo, R., Romanish, M. T. & Mager, D. L., 2012). D'autre part, il peut y avoir création d'un réseau de régulation (Feschotte, C., 2008), si un même élément s'insère devant plusieurs gènes participant à une même voie métabolique et apporte une séquence de régulation. Enfin, cela peut prendre la forme de modifications épigénétiques pouvant impacter l'expression des gènes (Slotkin, R. K. & Martienssen, R., 2007), les éléments transposables étant fréquemment sujets à la méthylation qui entraîne généralement la compaction de la chromatine et peut donc inhiber l'expression des gènes proches. Les impacts des insertions des éléments transposables sont alors souvent délétères pour l'organisme qui aura tendance à les éliminer et à mettre en place de systèmes pour les contrôler. Mais il arrive également que ces insertions aient un impact adaptatif et participent à l'apparition de nouveaux traits chez l'hôte. Ces nouvelles caractéristiques peuvent notamment avoir un intérêt agro-économique et les plantes qui en sont porteuses peuvent alors être sélectionnées par l'Homme pour être cultivées.

Bien que désormais on connaisse un certain nombre de changements adaptatifs ayant pour origine les éléments transposables (couleur de l'orange ou du raisin, forme de la tomate…), leur impact reste à démontrer de façon claire dans la plupart des espèces, mais également pour différents processus et tissus au sein d'une même espèce. Or, l'un des processus les plus importants en agronomie est celui de la maturation du fruit. Il serait donc intéressant de connaître l'implication, ou l'impact des éléments transposables dans celui-ci.

Pour étudier ce processus de maturation, la tomate *Solanum lycopersicum* est un modèle privilégié. En effet, comme nous l'avons vu précédemment, il s'agit de l'un des légumes/fruits les plus cultivés et consommés au monde et il a déjà été montré chez cet organisme que les ET y participaient, à la modification de la forme du fruit (Xiao, H., Jiang, N., Schaffner, E. J. & van der Knaap, E., 2008), mais également à sa qualité (Quadrana, L. *et al.*, 2014). Récemment, une étude s'est intéressée aux changements intervenant dans le méthylome de la tomate au cours de la maturation du fruit (Chen, Y.-R. *et al.*, 2013), mettant ainsi à disposition de nouvelles données sur ce génome. Cette étude a alors permis d'identifier de nombreuses régions présentant des variations de niveaux de méthylation, appelées régions différentiellement méthylées (*Differentially methylated regions*, DMRs), au cours de ce processus. Ces DMRs se trouvent être très fréquemment retrouvées à proximité des gènes dont l'expression change au cours de la maturation du fruit, et plus particulièrement dans leurs régions promotrices. Mais on les retrouve également à proximité des MADS-box des sites de liaison RIN (*ripening inhibitor*), ces sites RIN étant des éléments essentiels dans divers processus du développement de la plante (Ng, M. & Yanofsky, M. F., 2001) (Martel, C., Vrebalov, J., Tafelmeyer, P. & Giovannoni, J. J., 2011).

Compte tenu de tous les effets connus des éléments transposables chez la tomate, la question s'est alors posée de savoir quels autres impacts pourraient-ils avoir, notamment dans le processus de maturation du fruit. Il s'agit donc ici de caractériser la contribution potentielle des éléments transposables aux changements épigénétiques et à la régulation transcriptionnelle des gènes durant la maturation de la tomate. Pour cela, une ré-annotation des éléments répétés dans le génome de *S. lycopersicum* a été effectuée et des analyses cherchant l'existence d'un lien entre les ET, les gènes et les DMRs ont été menées. Cette ré-annotation a été réalisée en combinant plusieurs outils : REPET (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011) (Quesneville, H. *et al.*, 2005), RepeatScout / RepeatMasker (Price, A. L., Jones, N. C. & Pevzner, P. A., 2005) (Smit, A.F.A., Hubley, R. & Green, P., RepeatMasker at http://repeatmasker.org) et Tallymer (Kurtz, S., Narechania, A, Stein, J. C. & Ware, D., 2008). En effet, il a été montré que l'utilisation de différents outils, chacun ayant ses spécificités de détection, était complémentaire (Maumus, F. & Quesneville, H., 2014) pour tendre vers une annotation exhaustives de l'ensemble des séquences répétées dans un génome. A partir de celles-ci, il a ainsi été possible de déterminer que le génome de la tomate, en se basant sur son contenu en gènes et en ET, pouvait être découpé en trois grandes régions : *Repeat Rich* (RR), *Intermediate* (INT) et *Repeat Poor* (RP), chacune de ces régions ayant alors des propriétés et un contenu spécifiques. La présence des répétitions, aux environs

des gènes, pourrait influencer leur expression, bien que l'effet ne soit pas toujours précisément caractérisé, et que l'impact de la méthylation sur cette même expression n'était pas clairement défini. Finalement, en comparant les niveaux d'expression des gènes à différents stades de maturation du fruit, il a été possible d'observer que cette expression pouvait être influencée par des éléments transposables, porteurs d'un grand nombre de DMRs.

2.2 L'analyse approfondie du *repeatome* de la tomate montre un impact potentiel des éléments transposables sur la maturation du fruit

Cette partie présente la méthodologie et les résultats obtenus, concernant la contribution potentielle des ET aux changements épigénétiques et à la régulation transcriptionnelle des gènes au cours de la maturation de la tomate, sous forme d'une publication scientifique parue dans le journal *BMC Genomics* en aout 2016.

Jouffroy et al. BMC Genomics (2016) 17:624
DOI 10.1186/s12864-016-2980-z

BMC Genomics

RESEARCH ARTICLE Open Access

Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening

Ophélie Jouffroy[1], Surya Saha[2], Lukas Mueller[2,3], Hadi Quesneville[1] and Florian Maumus[1*]

Abstract

Background: Plant genomes are populated by different types of repetitive elements including transposable elements (TEs) and simple sequence repeats (SSRs) that can have a strong impact on genome size and dynamic as well as on the regulation of gene transcription. At least two-thirds of the tomato genome is composed of repeats. While their bulk impact on genome organization has been recently revealed by whole genome assembly, their influence on tomato biology and phenotype remains largely unaddressed. More specifically, the effects and roles of DNA repeats on the maturation of fleshy fruits, which is a complex process of key agro-economic interest, still needs to be investigated comprehensively and tomato is arguably an excellent model for such study.

Results: We have performed a comprehensive annotation of the tomato repeatome to explore its potential impact on tomato genome composition and gene transcription. Our results show that the tomato genome can be fractioned into three compartments with different gene and repeat density, each compartment presenting contrasting repeat and gene composition, repeat-gene associations and different gene transcriptional levels. In the context of fruit ripening, we found that repeats are present in the majority of differentially methylated regions (DMRs) and thousands of repeat-associated DMRs are found in gene proximity including hundreds that are differentially regulated. Furthermore, we found that repeats are also present in the proximity of binding sites of the key ripening protein RIN. We also observed that some repeat families are present at unexpected high frequency in the proximity of genes that are differentially expressed during tomato ripening.

Conclusion: Altogether, our study emphasizes the fractionation as defined by repeat content in the tomato genome and enables to further characterize the specificities of each genomic compartment. Additionally, our results present strong associations between differentially regulated genes, differentially methylated regions and repeats, suggesting a potential adaptive function of repeats in tomato ripening. Our work therefore provides significant perspectives for the understanding of the impact of repeats on the maturation of fleshy fruits.

Keywords: Fruit ripening, DNA methylation, Transposable elements, Tomato

Background

The majority of plant genomes contain a large fraction of repetitive DNA, collectively referred to as the repeatome of a species [1]. The major types of repetitive elements in plant genomes comprise transposable elements (TEs), simple sequence repeats (SSRs), and ribosomal DNA. The de novo detection of repeated sequences also commonly reveals the significant contribution of repeated features that remain unclassified. Because of their relatively high duplication rate, TEs, which mediate their own transposition, are generally the most abundant sequences in plant repeatomes. While TE insertions can be deleterious by disrupting genes, mounting evidences demonstrate that some TE copies can also impact the transcriptional regulation of nearby genes and can thereby generate adaptive traits and phenotypes of agro-economical interest [2].

TEs can impact the transcription profile of proximate genes by a variety of means, at the structural and quantitative levels. For instance, TE sequences can distribute new

* Correspondence: florian.maumus@versailles.inra.fr
[1]URGI, INRA, Université Paris-Saclay, 78026 Versailles, France
Full list of author information is available at the end of the article

regulatory regions such as promoters [2]. Being repetitive across genomes, TEs can also build regulatory networks that influence the expression of several genes in a coordinated fashion [3]. TEs can also provide alternative transcription start sites and other transcript isoforms [4]. Furthermore, TE sequences are commonly modified by the cells by the addition of methyl groups on cytosine residues in a process called DNA methylation, that causes local genome compaction and prevent TE transcription [5]. This epigenetic regulation can occasionally act in cis or spread into neighboring genes and affect their expression [6, 7].

Nevertheless, the potential impact of TEs and other repetitive elements remain to be addressed in a comprehensive manner in most plant species at different developmental stages and in a variety of tissues. Of marked interest in agronomy, the role of repeated sequences in the ripening of fleshy fruits remains to be investigated in a comprehensive manner. Tomato, S. lycopersicum, is the most cultivated fleshy fruit/vegetable worldwide with a global production around 160 million tonnes each year (http://faostat3.fao.org). The genome of the inbred tomato cultivar 'Heinz 1706' was sequenced and assembled in 2012, and TEs were found to make a large contribution to the nearly complete assembly of this ~900 megabases (Mb) genome. Previous reports have demonstrated that TEs do play roles in the determination of fruit morphology and quality [8, 9]. In addition, a recent study has investigated the changes of the tomato methylome at single-base resolution and has identified thousands of regions that present dynamic methylation patterns, mostly hypomethylation, during fruit ripening [10]. These differentially methylated regions (DMRs) were found to associate with differentially expressed genes in maturing tomatoes and with binding sites of the RIN (ripening inhibitor) MADS-box transcription factor which is a key regulator of ripening [11].

Here, we have investigated the global impact of repetitive elements on the tomato genome and their potential role in the orchestration of tomato ripening. By studying the composition of the genome in terms of genes and transposable elements contents, we show that it could be divided into three types of regions, each showing specific properties in genes and repeat content. We also found that globally the presence of repeated sequences near genes could slightly influence their expression, but that their methylation could instead have an impact that is still poorly defined. Finally, a comparison of the different stages of maturation reveals that the expression of genes in this process may be partly regulated by TEs and differentially methylated regions (DMRs).

Results

Comprehensive annotation of the Heinz 1706 repeatome

The initial annotation of repetitive elements in the tomato genome has relied on the identification of representative sequences based on the presence of structural features [10, 12] and more recently on the use of a de novo repeat identification tool [13]. We have recently shown the advantage of combining different approaches in order to improve repeatome annotation in genomes [1]. Here, in order to generate a comprehensive annotation of the tomato repeatome, we have employed a combination of similarity- and k-mer-based methods with the REPET and RepeatScout programs, respectively for the de novo construction of libraries of consensus sequences representative of repetitive elements. Alignment-based annotation of the tomato assembly using these libraries yields 68 % (532 Mb) coverage and 72 % coverage of the non-gapped assembly (Fig. 1). For comparison, this annotation covers 96 % of the initial repeat annotation (reciprocally 82 %) and 95 % of the specific annotation of MITE elements established previously [10, 12]. It also covers 96 % of a recently published de novo repeat annotation [13] (reciprocally 84 %). In addition, we have employed a strict mapping of frequent k-mers in the genome assembly in order to detect short repeats that would have been missed by alignment-based strategies. This approach identified 292 Mb of perfect repeats, including 22 Mb that were not detected above. In total, our combined annotation covers 75 % of the ungapped tomato genome assembly.

In line with previous findings [12], we found that LTR-retrotransposons are the most abundant TEs in the Heinz 1706 assembly with Gypsy- and Copia-type elements representing 45 and 14 % of the repeat annotation, respectively (Additional file 1: Figure S1, Additional file 2: Table S1). In addition, Class 2 elements (also called DNA transposons), including both autonomous and non-autonomous elements, were found to contribute 5 % of the repeat annotation. Furthermore, environmental viruses, which can happen to integrate into plant DNA and being vertically transmitted over generations in the form of endogenous viral elements (EVEs), can also represent significant constituents of their nuclear genomes [14, 15]. In tomato, we found that EVEs, including members of the recently described Florendovirus [15] and Mitovirus [16], contribute over 4 Mb of the Heinz 1706 assembly. Finally, SSRs and unclassified repeats were observed to make a substantial contribution to the repeatome annotation (Additional file 1: Figure S1, Additional file 2: Table S1). All the annotations generated here are available at the Sol Genomics Network.

Determining three genomic compartments with contrasted repeatome composition

Plotting the density of repeats and genes along the chromosomes confirms that distal chromosome regions are gene-rich while the remainder of the genome is densely

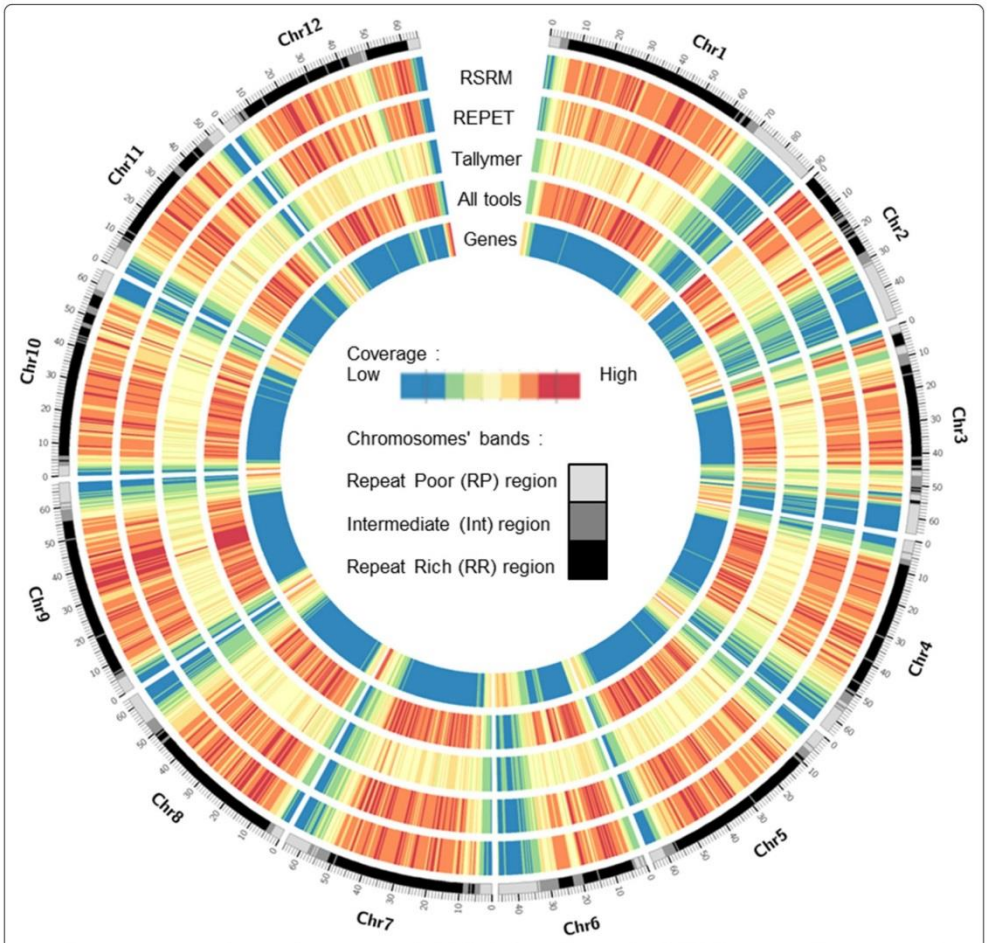


Fig. 1 The tomato repeatome. Repeats and genes coverage by 500 kb windows with an overlap of 50 kb. Repeats annotation was conducted with three different tools: RepeatScout+RepeatMasker (RSRM), REPET, and the k-mer strict program Tallymer. Three types of genomic regions are visible in the chromosome bands and chromosome name and scale (in Mb) are indicated on the outer rim

populated by repetitive elements (Fig. 1). Clustering of the repeat and CDS densities in 500 kilobases (kb) windows allowed empirically determining three kinds of genomic regions (Additional file 3: Figure S2; Additional file 2: Table S2): repeat-rich (485 Mb), repeat-poor (161 Mb), and a third, intermediate category (113 Mb), hereafter referred to as RR, RP and INT compartments, respectively. In line with the repeatome distribution along the chromosomes, INT regions are most of the time positioned in transition zones between RR regions, which are found in pericentromeres, and RP regions, located at the chromosome tips.

The high abundance of repetitive elements, and especially TEs, in the Heinz 1706 genome to a large extent determines the composition of intergenic DNA. We therefore addressed whether the distribution of different types of TEs is homogeneous along the tomato chromosomes. Remarkably, applying a local enrichment statistical analysis [17], we observed a differential distribution of the main types of TEs between the three compartments established above (Fig. 2; Additional file 4: Figure

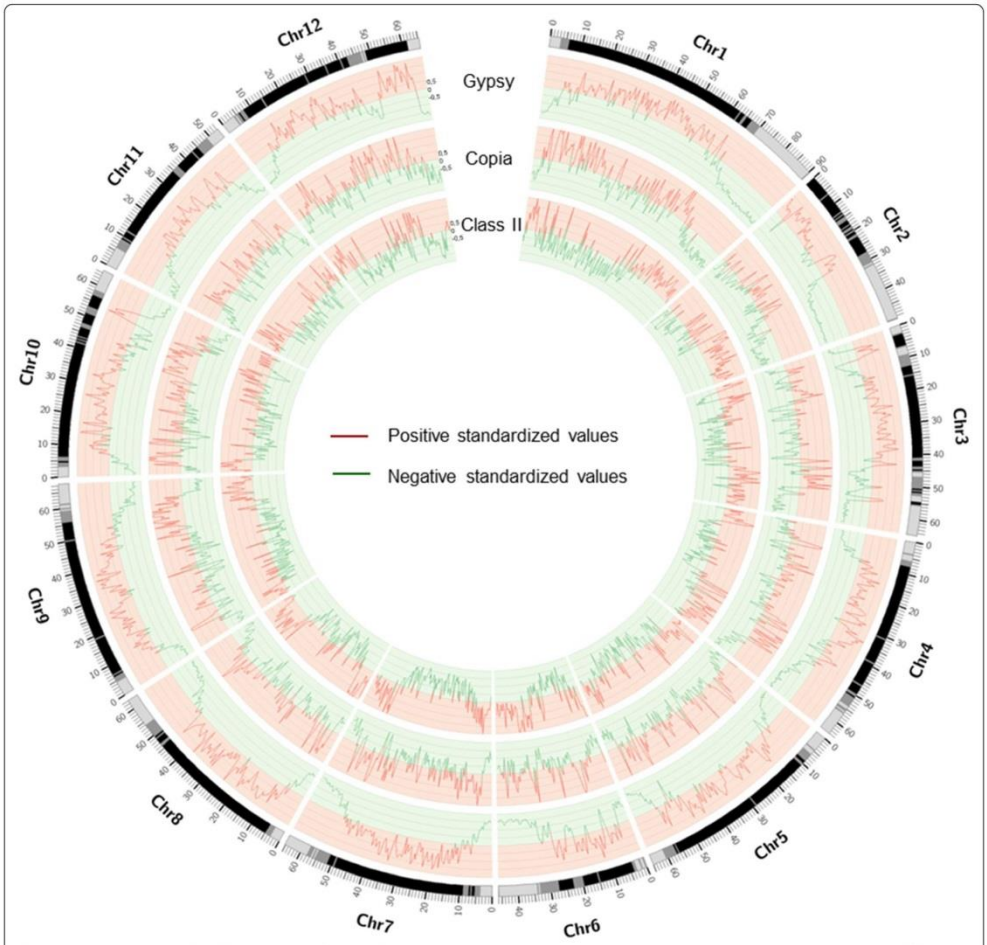


Fig. 2 Genome coverage for different repeats families. This graphic represents the standardized values of coverage of the three major TE families in tomato genome, Gyspsy, Copia and Class II elements. The coverage is calculated by window of 500 kb with an overlap of 50 kb and standardized values are determined based on these calculations. Positive values (in red) reflect an enrichment while negative values (in green) reflect depletion. Chromosome name and scale (in Mb) are indicated on the outer rim. The interval between two lines on the graph corresponds to a value of 0.5

S3). For instance, Copia-type LTR retrotransposons (LTR-RT) are enriched in INT regions and depleted in RP regions. In contrast, Gypsy-type LTR-RTs are predominant in the RR space and under-represented in the RP and INT regions. In addition, Class II elements (autonomous and non-autonomous DNA transposons) are enriched in the RP and INT regions and depleted in the RR compartment. This contrasted distribution of the tomato repeatome in the three genomic compartments may have significant impact on DNA composition. In fact, we observed that the G + C content in the repeatome (REPET + RepeatScout) of each compartment is also different by being relatively higher in RR than in RP, INT showing an intermediate value (Additional file 5: Figure S4A). Correlatively, we found that among the main TEs in tomato, Gypsy-type elements have the highest G + C content followed by Copia-type elements then DNA transposons (Additional file 5: Figure S4B). A

similar bias in repeatome composition has been described in *Arabidopsis thaliana* [18] although the causes seem to be different in the two genomes.

Contrasted gene density between genomic compartments

As predictable, we found that gene density is much lower in RR and INT regions (15 and 57 genes per Mb, respectively) as compared to RP regions (122 genes per Mb). Nevertheless, RR and INT regions contain an important number of predicted genes (7554 and 6466, respectively, i.e. approx. 40 % of the total gene count). Yet the majority of the predicted genes (19,781) are located in RP regions.

Protein-coding TEs are commonly confounded by gene prediction programs, leading to gene models that actually correspond to TEs. We therefore took advantage of this new genome annotation to address the potential contribution of TE-genes to the set of predicted genes in each genomic compartment. We first established a set of confidently classified TEs comprising repetitive elements with similarity to known TEs and/or with TE domains and no other conflicting evidence using the TE classifier PASTEC [19]. The comparison of their positions with those of predicted CDS allowed the identification of 2246 putative TE-genes for which over half of the CDS fraction is covered by high confidence TEs (Additional file 2: Table S3). While most TE-genes are located in RR and INT compartments, these areas still contain a high number of non-TE genes that are distributed in a highly repeated environment in contrast to those that are positioned within the RP space (Additional file 6: Figure S5A). Comparing the expression in leaves using available data [10] of TE-genes and non-TE-genes in different genomic regions showed a significant difference between these two groups. Indeed, while the expression of TE-genes remains at a zero mean, or practically, that of non-TE-genes is contained in a large range of values regardless of the genomic compartment (Additional file 6: Figure S5B). This observation suggests possible TE contamination in the predicted gene set so we decided to exclude putative TE-genes of further analyzes.

Correlation between repeat density and gene expression

The genes that are located in the genomic compartments that we defined on the basis of repeat density appear to be positioned within strikingly different environments in terms of genome dynamics and composition at the chromosome scale. We further examined whether these contrasted landscapes correlate with distinct properties of the gene sets from each compartment including distance to repeats, evolutionary origin, and expression levels.

The different TE densities in the three genomic compartments suggest the respective gene sets may be in different proximity to repeats. For each compartment, we measured the distance from genes to the closest upstream or downstream genomic repeats. As expected, we found that on average, genes from the repeat-rich and intermediate compartments have closer upstream and downstream repeats as compared to those located in the repeat-poor regions (medians: RR = 420, INT = 442, RP = 640; Mann-Withney U test (MWU) [RR vs RP] P value < 0.0001; MWU [INT vs RP] P value < 0.0001, for upstream and downstream repeats) (Additional file 7: Figure S6). Surprisingly, gene sets from the repeat-rich and intermediate compartments show rather similar distances to proximal repeats. More interestingly, we also observe for each compartment that upstream repeats are closer to genes than downstream repeats, especially in the RP compartment.

Because proximal genomic repeats can impact gene expression by a variety of means, we next addressed whether overall, genes within the different genomic compartments may present distinguished expression levels in tomato leaves. We found that, overall, genes from the RR, INT and RP compartments show strikingly different expression levels. Remarkably, the predicted genes from the RR set show very low median RPKM value and relatively few highly expressed genes (median-RR = 0,00; 1567 (21 %) with RPKM > =5) as compared to those located in the INT and RP regions (Fig. 3). Overall, we observe a strong negative correlation between repeat-density and the transcriptional levels of predicted genes (Mann Witney U test P value < 0.001 between each compartment). The observation that genes with closer repeat proximity show lower expression levels could reflect different biological and evolutionary phenomena. For instance, gene expression levels may be impacted by the presence of neighboring, especially upstream repeat-associated heterochromatin that may hamper the access of factors that initiate transcription. Another non-exclusive explanation could be that the different genome compartments enclose gene sets which expression is differentially regulated overall; for instance if constitutively and/or highly expressed genes are over-represented in repeat-poor regions or if stress-responsive genes are mostly present in RR.

Correlation between repeat proximity and gene expression

To investigate the overall differences in gene expression levels observed between genomic compartments, we first addressed whether repeat proximity has a predominant effect on gene transcription levels. Indeed, assuming that repeated elements are generally methylated in the genome [10] and are thus predominantly associated to regions of condensed chromatin, one could expect a negative impact of repeats on the transcription of nearby

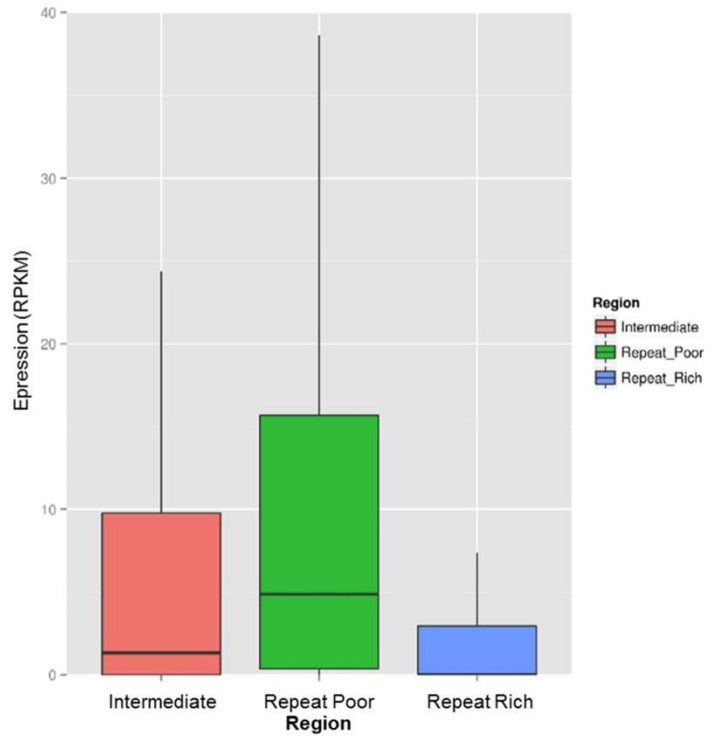


Fig. 3 Global expression in leaf in the three genomic regions. In RR region, median value is equal to 0 and only 1567 (out of 7529) genes have a RPKM value higher than five, while in RP region, median value is equal to 4,87 and 9796 (out of 19780) have a RPKM value higher than five, and in INT region median value is equal to 1,34 with 2299 (out of 6486) genes with a RPKM value higher than five

genes and hence a negative correlation between repeat distance and gene expression levels. However, we could not detect such an overall negative correlation when comparing the expression levels of genes with upstream repetitive element in different 100 bp bins of a 1 kb window (data not shown). We then reason that such a signal could be biased by other gene-proximal repeated elements that may also affect gene expression. Indeed, repeated elements may be positioned upstream, downstream and within introns of the same gene, and all these configurations may influence gene expression by a variety of means [2]. We therefore examined gene expression levels following the presence of repeated elements exclusively in one of the above mentioned configuration. In leaves, we found that the genes without upstream repeats in 1 kb present lower transcript levels than those with an upstream repeat in RR and RP (MWU of RPKM values [No repeat vs 1 kb upstream], in leaves: P value = 0.005 in RR, P value = 0.279 in INT and P value = 0.003 in RP) (Fig. 4). Also, the genes with intronic repeats show robust statistical support of highest median transcript values in the RP and INT compartment (MWU of RPKM [Intronic vs No repeat] in leaves: P value = 0.088 in RR, P value = 0.008 in INT and P value < 0.001 in RP).

To gain more insights on the overall local effects of repeats on gene expression, we specifically analyzed the impact of the DNA methylation of repeats, which can alter the expression of nearby genes. Interestingly, we found that, in leaves, repeat methylation is associated with decreased gene transcript levels when located in an intron or upstream of the gene in RR and INT (MWU of RPKM [methylated vs non methylated repeat]: 0.05 > P value > 0.001) (Additional file 8: Figure S7). Similarly, in INT and RP, an association between gene expression levels and the presence of methylated repeats is visible if they are located downstream, upstream or in an intron of the gene (MWU of RPKM [methylated vs non

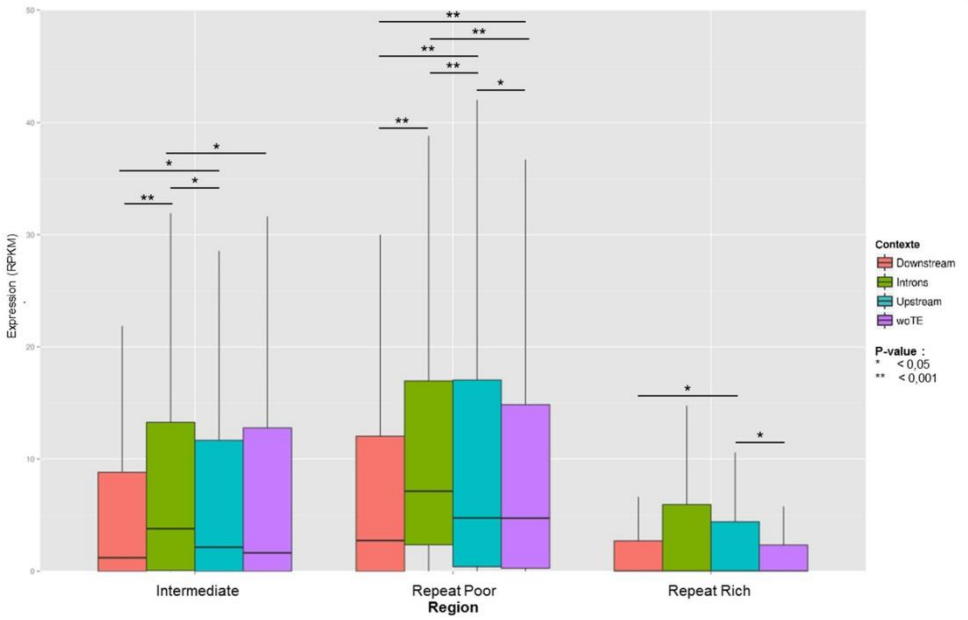


Fig. 4 Gene expression in leaf considering repeat location and genomic region. Considering four different types of associations between genes and repeats, we observe the difference of expression between these categories. Statistical analyzes (Mann – Whitney U tests) were performed to validate the observations and these results are represented by the *P*-value on the graphic

methylated repeat]: $0.05 > P$ value > 0.001). However, we observed that methylated or unmethylated proximal repeats can be associated with globally higher or lower gene expression levels during the different stages of maturation in different genomic regions (data not shown).

Different gene categories between genomic compartments

We also addressed whether the different genomic compartments may contain genes with a biased composition of predicted function. For example, we found that the gene ontologies GO:0005515 (protein binding), GO:0047714 (galactolipase activity) and GO:0009611 (response to wounding) are differentially represented in the RP-vs-RR, RP-vs-INT and INT-vs-RR compartments, respectively (chi-square P value < 0.001 with Bonferroni correction) (Fig. 5; Table 1). In addition, genes without functional category are more frequent in RP regions than in RR and INT. A number of stress genes have been characterized and predicted in tomato (see [20]), which expression profiles are typically expected to be stress-induced. We found that most of these genes (2056) are in RP, while INT contains 604 of these genes and RR only 371. While this distribution corresponds to the expectation if randomly distributed in RP and INT, the number of stress genes in RR is lower than expected (chi-square P value = 0.84321 (degree of freedom X-square = 0.04) in RP vs others, P value = 0.193 (X-square = 1.70) in INT vs others and P value < 0.001 (X-square = 168.25) in RR vs others).

We also investigated whether the genes positioned within the different compartments may show different evolutionary origins. In this scope we used three sets of genes that derive from phylogenetic reconstructions of the gene sets from dozens of plant genomes including three members of Asterids (*Mimulus guttatus*, *Solanum lycopersicum* and *Solanum tuberosum*) [21]. The first set comprises the genes that were present in the Asterid ancestor (ANC2), the second set includes the genes existing in the Solanum ancestor (ANC1), and the third set encompasses genes that are specific to tomato (NEW). Again, we observe significant differences between genomic compartments (Additional file 9: Figure S8). Interestingly, we found that the ANC2 gene set is enriched in the RP regions (chi-square P value < 0.001). Instead the Solanum-derived gene set and the tomato-specific genes are enriched in the RR regions (chi-square P value $<$ 0.001 for both groups). Furthermore, the intermediate

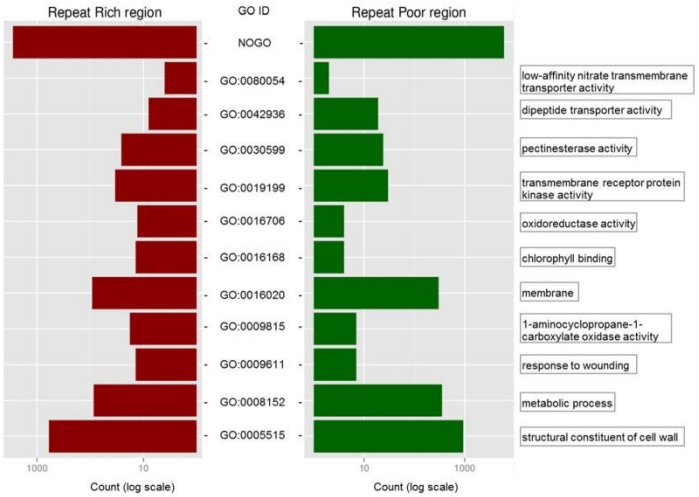


Fig. 5 GO terms distribution between genomic regions. Counting of each GO term in the RR, INT and RP regions give us the possibility to compare their composition in terms of predicted gene functions

compartment is depleted for new genes and enriched in genes of Asterid origin (chi-square P value < 0.05 for both groups).

Several repeat families are associated with differentially regulated genes during ripening

Tomato fruit ripening is accompanied by successive changes in the regulation of thousands of genes as determined by comparison of transcriptomes from pericarps at four different stages of tomato maturation (17 d.p.a = Days Post Anthesis, 39 d.p.a, 42 d.p.a and 52 d.p.a) [10]. Because repeats can provide regulatory elements and can be involved in the establishment of gene regulation networks (see [3]), we have investigated whether copies of any of the tomato repeat consensus would be positioned nearby ripening-modulated genes more frequently than expected if repeat copies were distributed randomly among stable, up- and down-regulated genes (see [22]). Indeed, such distributions could reflect the specific retention and perhaps the selection and function of specific repeats in the proximity of the genes that are differentially regulated during ripening. Interestingly, we observed that 11 and 13 repeat families are present at high frequency compared to expectation nearby genes that are up and down regulated during ripening, respectively (Fig. 6, P values for chi-square with Bonferroni correction in Additional file 2: Table S4). These families comprise unclassified elements and SSRs as well as several putative class 2 (including autonomous and non-autonomous elements) and class 1 TEs (1 LINE and 1 SINE). Remarkably, several elements appear to be enriched both at up- and down-regulated genes, sometimes at different extent and timing during ripening.

Repeats sustain the dynamic methylome during fruit ripening

Tomato ripening also goes together with thousands of changes in DNA methylation levels along the chromosomes that probably have a widespread impact on the regulation of gene expression [10]. Because repeats are generally methylated, we explored the potential contribution of repeat-associated methylation changes on gene transcriptional regulation through tomato ripening. In a dynamic perspective, we examined the presence of Differentially Methylated Regions (DMRs) and their potential link to TEs and genes using methylome data from four different stages of tomato maturation (17 d.p.a = Days Post Anthesis, 39 d.p.a, 42 d.p.a and 52 d.p.a) [10].

We first investigated the associations between DMRs and TEs. Among the 52,095 DMRs present in the Heinz 1706 genome, 72.29 % of them associate with genomic repeats (18.65 % of the DMRs overlapping repeats, 54.57 % are included in a repeat and 0.80 % have a repeat included in their sequence), which is a greater proportion than if DMRs were randomly distributed in the genome (chi-square P value < 0.001 and x-squared = 129.55). The genomic distribution of these associations shows that the majority (72.34 %) are in the RR region, while RP contains 14.23 % of the associations and INT only 13.43 %. Statistical analysis of this distribution

Table 1 GO terms repartition in the three compartments of the genome. Exemple of deviation of repartition of fifteen GO terms between the three genomic regions in the tomato genome. Column '*P*-value' indicates results from the statistical analyzes (chi-square test) while column 'counting' gives the number of a GO term in a specific region

GO	*P*-value			Counting		
	RR vs INT	RR vs RP	INT vs RP	RR	INT	RP
GO:0005199	0	0	1	15	1	3
GO:0005515	0,076	0	0,003	596	644	935
GO:0008152	0,03	0	0	86	139	354
GO:0009611	0	0,003	0,099	14	0	7
GO:0009815	0	0	0,278	18	1	7
GO:0016020	0,002	0	0,385	92	165	303
GO:0016168	0,002	0	1	14	2	4
GO:0016706	0,042	0	0,43	13	5	4
GO:0019199	0	0,001	0,424	34	13	30
GO:0030599	0,572	0,008	0	26	37	24
GO:0042936	0,001	0,844	0	8	34	19
GO:0080054	0,003	0,192	0	4	23	2
NOGO	0,009	0,001	0,761	2834	3570	6037

shows that the observed values do not match those expected if repeat-associated DMRs were randomly distributed in the genome (chi-square P value < 0.001 in all three cases), RR being enriched in repeat-DMR associations while INT and RP are depleted. Looking at the repeats involved in the associations with DMRs, it appears that not all of them are involved in similar proportions (Additional file 10: Figure S9). Copia-type TEs are indeed more frequently associated with DMRs than expected (chi-square P value < 0.001), whereas DNA, non-autonomous DNA, Gypsy, and Line TEs as well as SSRs are observed in a number of associations lower than expected (chi-square P value < 0.001 for each family).

We have next explored the associations between repeat-associated DMRs and genes (upstream, downstream, and intronic). A total of 5021 associations could be found, the majority of them being in the RP region (54.39 %), the rest being distributed between INT and RR regions (22.19 % and 23.42 %, respectively). This distribution does not match the expected one (chi-square for P value < 0.001 for each comparison), the surrounding of RP and INT genes being enriched in repeat-associated DMRs in contrast with RR genes which are so depleted. Therefore, the expression of genes in medium and low repeat density regions is more likely to be influenced by repeat-associated DMRs than those located in RR. It was also observed that 42.12 % of the associations are made with an upstream repeat, 32.49 % have a downstream repeat and 25.39 % of the genes have a repeat-associated DMR in their intron. Among these associations, we found that some repeat families were unevenly represented. For instance, SSRs and Line-type elements are present in these associations more frequently than expected while Gypsy-type elements are less (chi-square P value < 0.001).

Because of the high potential impact of repeat-associated DMRs on gene expression, we were interested in the genes in this configuration. By analyzing the

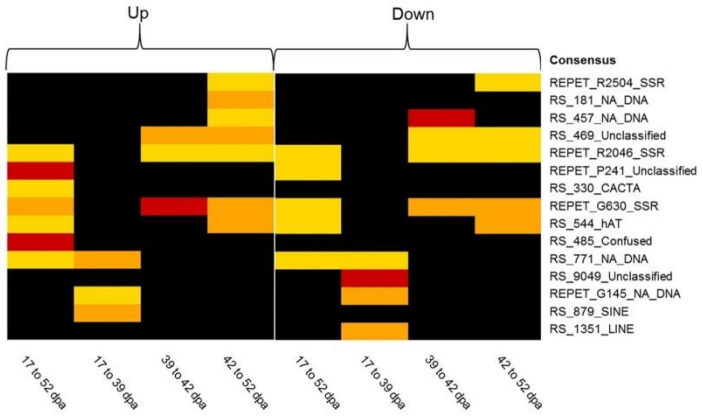


Fig. 6 Enriched repeat families nearby differentially expressed genes. Fold enrichment of associations between consensus and differentially expressed genes during maturation. 'Up' table shows the fold enrichment of consensus within 1 kb of up-regulated genes, while 'Down' table represents the fold enrichment of consensus within 1 kb of down-regulated genes. A yellow box represents a 1-2.5 fold enrichment, orange box: 2.5-6 fold enrichment and red box: 6-9 fold enrichment. NA_DNA stands for « non-autonomous DNA elements »

expression of these genes in each region, we found significant differences (Fig. 7). Interestingly, we found that, in RR and RP regions, the expression of genes with repeat-associated DMRs upstream of their sequence or in one of their introns is higher than if the gene is isolated (MWU *P* value woTE vs Introns and woTE vs Upstream < 0.001). In INT regions, higher gene expression levels are observed only with upstream repeat-associated DMRs (MWU *P* value woTE vs Upstream < 0.001). In order to deepen this study, a more accurate analysis of associations involving differentially expressed genes during tomato ripening was performed. A total of 1773 differentially expressed genes associated with repeat-associated DMRs were found between stages 17 d.p.a and 52 d.p.a (754 upstream, 503 downstream, and 516 in introns). Comparing these observations to the proportions of associations from a random distribution of genes in the genome, we observed that the number of differentially expressed genes associated with repeat-associated DMRs is significantly lower than expected by chance (chi-square *P* value < 0.05). Investigating the contribution of different repeats, we noticed that DNA TEs and SSRs are more frequently associated with genes differentially expressed during ripening, while in contrast, Gypsy elements are less (chi-square *P* value < 0.001 with a Bonferroni correction).

RIN is a MADS-box transcription factor which directly regulates fruit ripening genes and RIN binding sites are typically adjacent to DMRs and hypomethylated during ripening, suggesting that neighboring methylation to a large extent determines RIN access to its binding sites [10]. We therefore explored more specifically the level of association between repeat-associated DMRs and the positions of RIN binding sites. First, we found that 67 RIN binding sites (1.63 %) overlap with DMRs, all of these DMRs being associated to repeats. DMRs are typically adjacent to RIN binding sites [10]. In total, we found that 512 RIN binding sites (12.4 %) associate with DMRs (including adjacent and overlapping), which is a greater fraction than observed in a random distribution of DMRs (chi-square *P* value < 0.001 and x-squared = 49.48). Furthermore, we found that 177 RIN binding sites (4.30 %) overlap or are adjacent to repeat-associated DMRs, which is a lower number than determined from a random distribution of DMRs (chi-square *P* value < 0.001 and x-squared = 215.98). Gypsy and SSR repeat families are involved in those associations in a

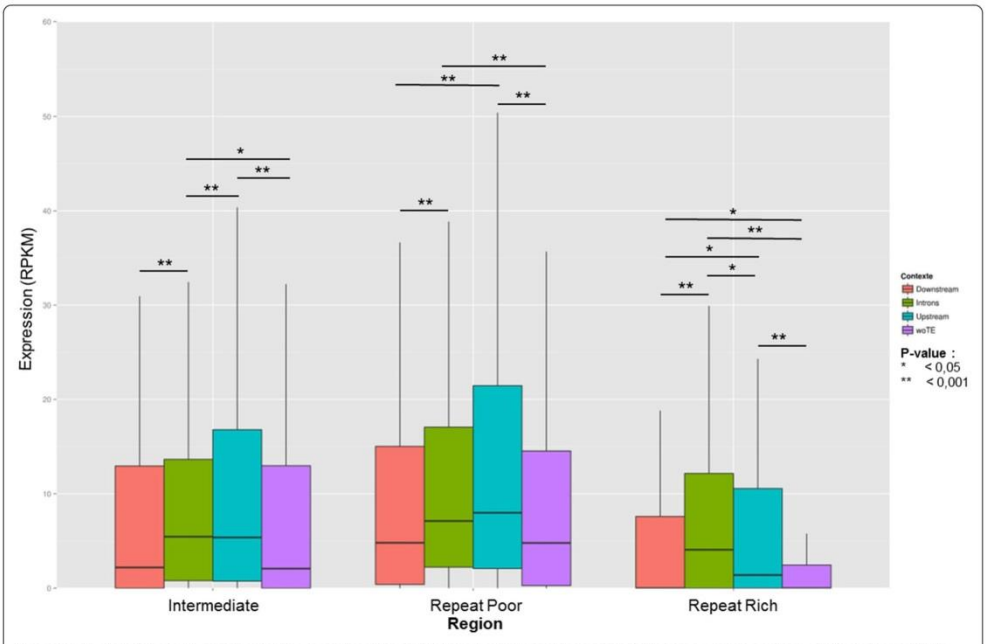


Fig. 7 Expression in leaf of genes related to repeats-supported DMRs. Four types of associations between genes and repeats including a DMR were defined. A study of the expression of these genes was then carried out to assess the impact of the presence of methylated repeats on that expression. Mann–Whitney U tests were performed to validate the observations

different way than expected, Gypsy elements being less present (chi-square P value < 0.001 with a Bonferroni correction, 37 observed and 67 expected) and SSR elements being more present (chi-square P value < 0.001 with a Bonferroni correction, 31 observed and eight expected). Among these, 70 are associated with differentially expressed genes including 14 (Additional file 2: Table S5) of the 292 genes predicted as potential RIN targets [10].

Discussion

The role of repeats in the regulation of gene expression [3] and many biological processes [2] has long been studied and is clearly demonstrated in some organisms. In this study we sought to understand their impact in the tomato genome and specifically the maturation process. Our various analyses, combining comprehensive repeat annotation, gene expression and DNA methylation data, help highlighting many aspects of the impact of repeats on tomato genome and biology.

Accordingly, we were able to demonstrate a compartmentalization of the genome as determined by its gene and repeat content. This separation into three major types of regions shows again the importance of repeats in shaping plant genomes. Indeed, through this division, we have shown that the content of each region is specific; both in terms of repeat content, some families are mainly located in one region or another, but also in terms of gene content, evolutionary origin and predicted function. Taken together, these observations may suggest a difference of dynamic of insertion and deletion, but also a differential control of transposable elements in the various 'territories' of the tomato genome. This aspect provides a new vision of the complexity of the tomato genome, especially in the case where a gene is found to be associated with different repeats. However, although the genes appear frequently associated to repeats, the presence of the latter doesn't have a clear effect on the levels of gene expression overall, and a negative correlation between the distance of repeats upstream and transcripts levels could not be established. Yet, it has been possible to observe variations of the expression taking into consideration the methylation status of the element. The *S. lycopersicum* genome has undergone limited repeat accumulation, the vast majority being ancient [20]. Therefore most ancestral repeat copies that have been retained in the tomato genome are likely to have almost neutral or adaptive effects while the most deleterious copies have likely been purified during evolution.

Thereafter, although a direct link between repeats and gene expression could not be determined, the analysis comparing different fruit maturation stages suggests a link between repeats and the presence of DMRs close to genes. This subset of DMRs could play a critical role in the regulation of gene expression, the repeats being the support of methylation and demethylation thereby influencing genome compaction at and nearby genes.. Most tomato ripening DMRs are linked to repeats and thousands of genes are adjacent to repeat-associated DMRs. Furthermore, 1773 genes that are differentially expressed during ripening are adjacent to repeat-associated DMRs, thereby suggesting a high potential impact of repeats (including both TEs and SSRs) on gene regulation and forecasting a certain degree of adaptiveness in gene-proximal repeat copies.

Conclusion

Repeats seem to be a major element of the structure and organization of the tomato genome and are the significantly associated with methylation, some of these repeats being associated with activation of maturation genes in this organism [9]. Although the impact of specific TE copies on tomato biology is beyond the scope of this paper, further analysis by combining comparative genomics and transcriptomics will allow more targeted analysis of the contribution of repeats in tomato biology. Refining our results could also be considered by exploiting alternative transcripts and detection of other recurrent repeat-derived motifs near genes.

Methods

Repeat annotation

We used the TEdenovo pipeline [23] from the REPET package v2.2 with default parameters (with a minimum of five sequences per group) on the contigs of size > 100 kb in the SL2.40 assembly (representing approx. 340 Mb, gaps excluded) which generated a library of 818 consensus sequences. We used RepeatScout (version 1.0.5) [24] on the contigs of size < 100 kb with default parameters which generated a second library comprising 9085 consensus sequences. The whole SL2.40 assembly was annotated using the TEannot pipeline [25] from the REPET package v2.2 with the TEdenovo library as input. Blaster sensitivity was set to "3" and threshold scores were calculated for each consensus as the 99th percentile value of scores obtained against a simulated genome consisting of the reversed (not complemented) assembly (REPET annotation). The entire SL2.40 assembly was also annotated using RepeatMasker [26] with the RepeatScout library (RepeatScout annotation). Regions identified by the RepeatScout approach that span at least 50 bp and that are not included in REPET annotations (54 Mb) were combined to the latter. Tallymer was launched with a k-mer size = 16 and a minimum of 10 occurrences in the SL2.40 assembly. Fragments of at least 30 bp that do not overlap the REPET and RepeatScout annotations were kept. Consensus sequences from

TEdenovo and RepeatScout were classified using the REPET dedicated utility released as the tool PASTEC [19]. Consensus classified as potential host genes because they contain host gene pfam domains (version 26.0) were excluded from this study. For the detection of TE genes, only the REPET annotations corresponding to TEs classified with high confidence (detection of non-conflicting evidences) were selected and compared to the CDS fraction of predicted genes. Genes which CDS are covered >50 % by selected TEs were categorized as putative TE-genes and not considered for further analysis.

Repeat profile of the tomato genome

Genome coverage in genes (CDS) and transposable elements was calculated by 500 kb window with an overlap of 50 kb. Genes and transposable elements (REPET + RepeatScout + Tallymer) annotations and a karyotype were also needed.

The resulting files were formatted to be visualized graphically through the Circos software [27]. A clustering step of the different coverage windows by the k-mean method using R (*kmeans*), based on the gene content and repeat content in each window has allowed distinguishing three types of genomic regions. Finally, each TE and each gene was located on the genome. The repeats or genes that straddle two regions have been excluded.

G + C content analysis

G + C content analysis was carried out at different scales in the genome: an analysis on the repeatome by genomic region and an analysis by repeat families in all regions. In each case, the sequence in FASTA format and the calculation of the G + C content was then performed by the command *infoseq*.

GO term enrichment analysis

Each of the tomato gene annotation was associated with its GO term. These associations between genes and GO terms were then separated according to the genomic region to which they belong.

The expected values (noted EXP) are calculated by relating the number of genes throughout the genome with a particular function (noted OBS), the number of genes in the region of interest (noted REG) and the total number of genes in the genome (noted TOT):

$$\mathrm{EXP} = \frac{\mathrm{OBS} \times \mathrm{REG}}{\mathrm{TOT}}$$

Then calculating the frequency for each GO term in each region is made and observed numbers are compared to the expected values calculated through chi-square tests for the same GO term between the three compartments.

Evolutionary origin of genes

To investigate the evolutionary origin of genes in the three genomic compartments of the genome, we used three sets of genes that derive from phylogenetic reconstructions [21]. The first set comprises the genes that were present in the Asterid ancestor (named ANC2), the second set includes the genes existing in the Solanum ancestor (named ANC1), and the third set encompasses genes that are specific to tomato (Heinz 1706) (named NEW). Data were obtained by courtesy of Alexandra Louis (Ecole Normale Supérieure, Paris, France).

Expected values (noted EXP) are calculated by linking the number of genes in the overall genome having a particular evolutionary origin (noted OBS), the number of genes in the region of interest (noted REG) and the total number of genes in the genome (noted TOT):

$$\mathrm{EXP} = \frac{\mathrm{OBS} \times \mathrm{REG}}{\mathrm{TOT}}$$

The distribution of these different groups in the three types of genomic regions was then observed as a histogram and statistical analysis checking the fit between the observed and the expected distribution was made with R through such chi-square tests.

Distribution of stress genes

We studied the distribution of a list of known stress genes in tomato (from [20]) within the three genomic regions RP, INT and RR. Each gene has then been assigned its belonging region based on its genomic coordinates, and counting has subsequently been achieved. To assess whether the distribution of these genes correspond to the expected distribution in each compartment based on their gene content, calculation of theoretical numbers and statistical analysis of chi-square type were performed.

The expected values (noted EXP) is calculated by linking the number of genes identified as stress response genes (noted OBS), the number of genes in the region of interest (noted REG) and the total number of genes in the genome (noted TOT):

$$\mathrm{EXP} = \frac{\mathrm{OBS} \times \mathrm{REG}}{\mathrm{TOT}}$$

Detection of repeat methylations

Methylations location data into the genome were available for three different methylation contexts: CG, CHH, CWG. Using a manual script, for each repeat involved in an association, we verified that it was methylated or not

and what was the context of this methylation. For this, and for each repeat, each position of it has been sought in the methylation file to obtain the information necessary for the analysis.

Associations (repeat / gene, DMR / TE and DMR / RIN)

Several types of associations between genes and transposable elements were defined for the analysis of expression: an upstream association (TE witchin 1 kb upstream of the gene), a downstream association (TE within 1 kb downstream of the gene) and an association for genes with a TE overlapping at least one of their introns.

To detect each of these associations, the BEDtools [28] *closest* tool has been used with different options as required: *–io* to ignore overlapping, *–iu* to ignore upstream associations, and *-id* to ignore downstream associations. The associations of interest are selected depending on the distance, between the gene and the nearest repetition, provided by the software in the last column of the result file. In the case of DMR / TE or DMR / RIN associations, only overlaps between the two types of sequences were selected with R after using *bedtools closest*.

To analyze the expression of genes associated with repeats, only genes with a single repeat close to their sequence, i.e. less than 1 kb, have been preserved. Once associations were identified, Mann Witney statistical tests with continuity correction (*wilcox.test()* in R) were performed to see if expression differences are observable according to the location (upstream, downstream, or in a intron) of repeat near gene.

The analysis of enrichment of copies annotated by each consensus of repetitive elements nearby differentially regulated genes was performed according to the methodology used by Makarevitch on maize (see [22]). For each repetitive element present at least once upstream of a gene, we counted the number of s copies observed upstream of genes that are up-regulated, down-regulated and stable. From this table, we calculate the theoretical values of this distribution for each consensus. For this, a calculation involving the number of observations of a consensus for each transcriptional status (up, down, stable) (REGUL_CSS), the total number observed in the genome for the same consensus (EFF_CSS) and the total number for all the consensus for the type of regulation studied is performed (REGUL_TOTAL):

$$\mathrm{THEO} = \frac{\mathrm{REGUL_C SS} \times \mathrm{EFF_C SS}}{\mathrm{REGUL_{TOTAL}}}$$

Statistical analysis checking the fit between the observed and the expected distribution was made with R through such chi-square tests. Finally, a filtering of the results is performed to retain as valid, the consensus associated with at least 10 genes expressed, with an enrichment of the observed value at least twice the expected value and a *p*-value less than 0,001.

For dynamic genome analysis, concerning genes associated with repeat-associated DMRs, all the associations found were preserved, a gene can then be associated with several repeat–associated DMRs. Mann Witney statistical tests with continuity correction (*wilcox.test()* in R) were then carried out to examine differences in gene expression between different classes in pairs.

Graphs and statistical analyzes

The circular graphs in this article were created with the tool *Circos* [27], all other graphics have meanwhile been established under *R v3.0.2* with *ggplot2* library.

Statistical analyzes were also performed with R v3.0.2 using the commands *chisqtest()*, to test the suitability of a data series for a family of probability laws or testing the independence between two random variables, and *wilcox.test()*, that tests the hypothesis that the distribution of data is the same in both groups defined. A Bonferroni correction was applied to analyze that require multiple comparisons.

JBrowse

Coordinates for the repeats identified in this paper using REPET, RepeatScout, Tallymer were converted to SL2.50 reference space from SL2.40 using a script (https://github.com/solgenomics/Bio-GenomeUpdate/blob/master/update_coordinates_gff.pl) and chromosome accessioned genome path files for SL2.40 and SL2.50 assemblies (ftp://ftp.solgenomics.net/tomato_genome/wgs/assembly/). Please note that 15 repeats identified by REPET, RepeatScout and Tallymer in SL2.40 were not ported over to SL2.50 since they straddled scaffolds in SL2.40 that were rearranged in SL2.50. All the repeats mapped to SL2.50 are available for analysis in the JBrowse genome browser at https://solgenomics.net/jbrowse.

RIN binding site analysis

A RIN binding sites analysis was performed using the binding peaks provided in data from Zhong et al. [10], in the Table S8, column “RIN binding peak”, and which have been extended to plus or minus 10 bp. First, we determined which RIN sites are associated with DMRs with the command line *bedtools closest -d -a RINs -b DMRs* from the BEDtools software [28]. RIN binding sites and DMRs involved in overlap were then identified and recovered in R, in bed files, column 19 indicates the distance between the two entities, so we selected distance equal to 0.

For DMRs involved in these associations, we then searched for those associated to repeats, using *bedtools closest -d -a DMRs -b TEs* command line again. The results of interest, i.e. the overlap between DMRS and repeats, were again identified and recovered in R (column 19 of the table is equal to 0).

A second analysis of these RIN binding sites was conducted in the same procedure, but this time, by extending the area of interest around the binding peak at plus or minus 500 bp because DMRs are most often next to RIN binding sites and non-overlapping.

Additional files

Additional file 1: Figure S1. Repeats composition of the genome of *S. lycopersicum*. The different families of repeats have been defined according to the Wicker's classification. The percentage of coverage of each family is calculated relative to the total coverage of the genome by repeats. (TIF 321 kb)

Additional file 2 Table S1. Genome coverage of the different repeat families. For each repeat family, a coverage calculation has been performed and the percentage of coverage of each family was calculated.by relating the cover of every family with respect to the overall repeat coverage of the genome. **Table S2.** Compartmentalization of the tomato genome in three major regions. Coordinates in gff3 format of the three regions of the genome based on the repeats and genes coverage calculation into 500 kb windows with overlap of 50 kb.**Table S3.** Names of identified TE-genes. List of ID and name of each gene identified as a TE-gene. **Table S4.** Statistical results of the analysis of transposable elements consensus by the method of Makarevitch and al. **Table S5.** Functional annotation of differentially expressed genes candidate as RIN targets. (XLSX 89 kb)

Additional file 3: Figure S2. Three categories of genomic regions. K-mean clustering results considering CDS and TE percentage of coverage of each window. We choose to defined three types of regions based on that result. (TIF 61 kb)

Additional file 4: Figure S3. Deviation of genome coverage by the three major repeat families. This boxplot shows the deviation of the genome coverage to the mean value of the three main repeat families of the tomato genome, Gypsy, Copia and Class II elements (bringing together the elements DNA and DNAna). The coverage is calculated by window of 500 kb with an overlap of 50 kb and standardized values are determined based on these calculations. Positive values reflect an enrichment while negative values reflect depletion of that type of repeat. Chi-square *P* values < 0.001 for each family versus others except for Copia in RR. (TIF 97 kb)

Additional file 5: Figure S4. GC content of repeats. (**A**) For each genomic compartment, the percentage of GC in repeats have been calculated. (**B**) Percentage of GC content in the four main families of repeats in tomato genome. (TIF 116 kb)

Additional file 6: Figure S5. TE-genes identification. (**A**) Proportion of putative TE-genes and true genes genes in each genomic region of the genome. (**B**) Comparison of the expression of genes and TE-genes in each genomic region. (TIF 48 kb)

Additional file 7: Figure S6. Distance between repeats and genes varies depending on genomic region. After determining for each gene the nearest repeat upstream and downstream to their sequence, we compare these distances between the three genomic regions. (TIF 101 kb)

Additional file 8: Figure S7. Gene expression levels depending on DNA methylation. Gene expression depending on the location of the repetition and status methylated or not. The four graphs correspond to the three main methylation contexts (CHH, CG, CHG) and all methylation contexts without distinction (All contexts). Mann Whitney statistical analyzes were conducted to test the differences observed and the results are shown in the the « All contexts » with a different symbol depending on the value of the *P*-value. (TIF 275 kb)

Additional file 9: Figure S8. The age of the genes specific to each genomic region. Counting genes considering their phylogenetic origin and comparing that repartition to that expected give us an information about gene age repartition in the three compartments. Statistical analyzes (chi-square tests) were conducted to validate the observations and are represented by the *P*-value on this graphic. (TIF 130 kb)

Additional file 10: Figure S9. Distribution of associations between repeats and DMRs in the different repeat families. After determining the associations between repeats and DMRs, counting of each family has been achieved and the percentage was defined by relating this count to the total number of defined associations. (TIF 146 kb)

Abbreviations

d.p.a., days post anthesis; DMRs, differentially methylated regions; EVEs, endogenous viral elements; SSRs, simple sequence repeats; TEs, transposable elements

Acknowledgements

We are very grateful to Mathilde Causse (INRA, France) for commenting a previous version of this manuscript. We also thank Alexandra Louis (ENS, France) for providing phylogenetic information regarding the ancestry of the tomato genes. Finally, we acknowledge the doctoral school ABIES from AgroParisTech for funding Ophélie Jouffroy.

Funding

Not applicable.

Availability of data and materials

TE annotations are available on Sol Genomics Network (https://solgenomics.net/jbrowse), and genes and DMR annotations can be found on Tomato Epigenome Database (http://ted.bti.cornell.edu/cgi-bin/gb2/gbrowse/tomato/). Gene Ontology annotation comes from Gene Ontology Consortium (http://geneontology.org). Gene expressions are provided Pr. James J. Giovanonni and Dr. Zhangjun Fei. The three sets of genes that derived from phylogenetic reconstruction were obtained by courtesy of Alexandra Louis (Ecole Normale Supérieure, Paris, France). Finally, TE-genes annotation and chromosomic bands are included within the article and its additional files.

Authors' contributions

FM designed the experiments and coordinated the project with contribution from HQ and OJ and performed genome annotation. OJ performed all the analyses and prepared the manuscript for publication. FM and OJ drafted the manuscript with contribution from HQ, SS, and LM. HQ advised on the statistical analysis. SS performed the conversion of genome annotation from SL2.40 to SL2.50. SS and LM made the SL2.50 genome annotation accessible through the ftp and jbrowse of the Sol Genomics Network. Finally, all authors have read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]URGI, INRA, Université Paris-Saclay, 78026 Versailles, France. [2]Boyce Thompson Institute, Ithaca, NY 14853, USA. [3]Department of Plant Breeding, Cornell University, Ithaca, NY 14853, USA.

Received: 15 April 2016 Accepted: 28 July 2016
Published online: 12 August 2016

References

1. Maumus F, Quesneville H. Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS One. 2014;9(4), e94101.
2. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49–61.
3. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397–405.
4. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.
5. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science. 2012;338(6108):758–67.
6. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res. 2011;39(16):6919–31.
7. Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature. 2015; 525(7570):533–7.
8. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008;319(5869):1527–30.
9. Quadrana L, Almeida J, Asis R, Duffy T, Dominguez PG, Bermudez L, Conti G, da Silva JV C, Peralta IE, Colot V, et al. Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat Commun. 2014;5:3027.
10. Zhong S, Fei Z, Chen YR, Zheng Y, Huang M, Vrebalov J, McQuinn R, Gapper N, Liu B, Xiang J, et al. Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat Biotechnol. 2013;31(2):154–9.
11. Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Schuch W, Giovannoni J. A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. Science. 2002;296(5566):343–6.
12. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635–41.
13. Mehra M, Gangwar I, Shankar R. A deluge of complex repeats: the Solanum genome. PLoS One. 2015;10(8), e0133962.
14. Jakowitsch J, Mette MF, van Der Winden J, Matzke MA, Matzke AJ. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc Natl Acad Sci U S A. 1999;96(23):13241–6.
15. Geering AD, Maumus F, Copetti D, Choisne N, Zwickl DJ, Zytnicki M, McTaggart AR, Scalabrin S, Vezzulli S, Wing RA, et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat Commun. 2014;5:5269.
16. Bruenn JA, Warner BE, Yerramsetty P. Widespread mitovirus sequences in plant genomes. PeerJ. 2015;3, e876.
17. Daron J, Glover N, Pingault L, Theil S, Jamilloux V, Paux E, Barbe V, Mangenot S, Alberti A, Wincker P, et al. Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol. 2014;15(12):546.
18. Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat Commun. 2014;5:4104.
19. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. PASTEC: an automatic transposable element classification tool. PLoS One. 2014;9(5), e91929.
20. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Rensen I, Lichtenstein G. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet. 2014;46(9):1034–8.
21. Louis A, Muffato M, Roest Crollius H. Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 2013;41(Database issue):D700–5.
22. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015;11(1), e1004915.
23. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011; 6(1), e16526.
24. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351–8.
25. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1(2):166–75.
26. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org 1996–2010.
27. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

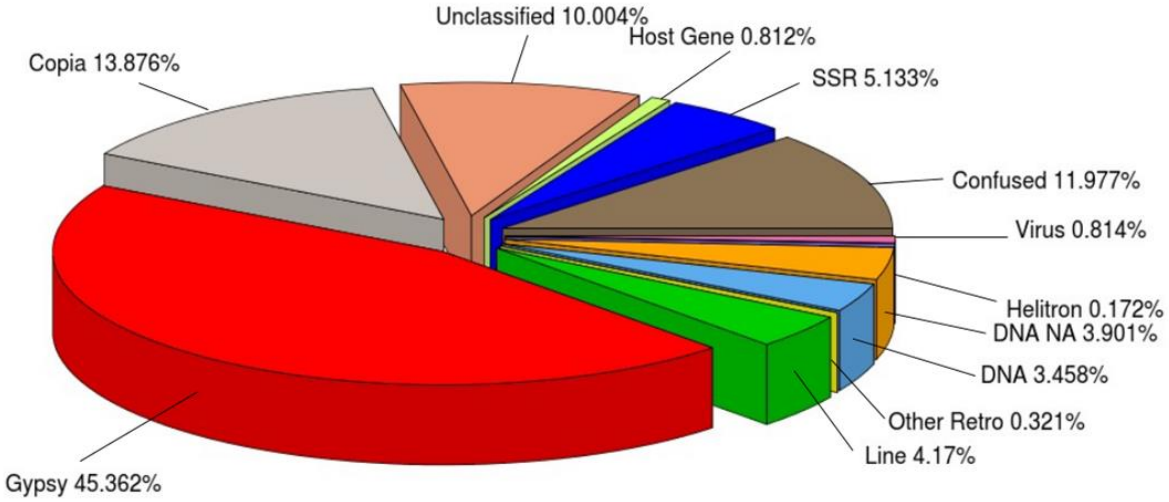


Figure S1. Repeats composition of the genome of S. lycopersicum. The different families of repeats have been defined according to the Wicker's classification. The percentage of coverage of each family is calculated relative to the total coverage of the genome by repeats

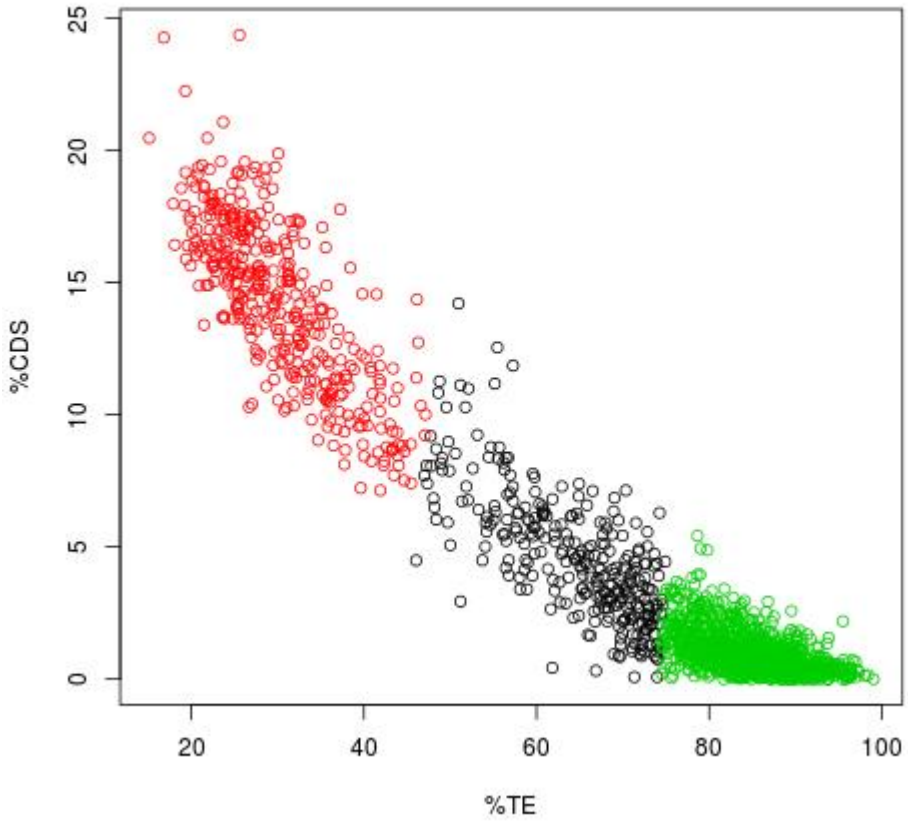


Figure S2. Three categories of genomic regions. K-mean clustering results considering CDS and TE percentage of coverage of each window. We choose to defined three types of regions based on that result.

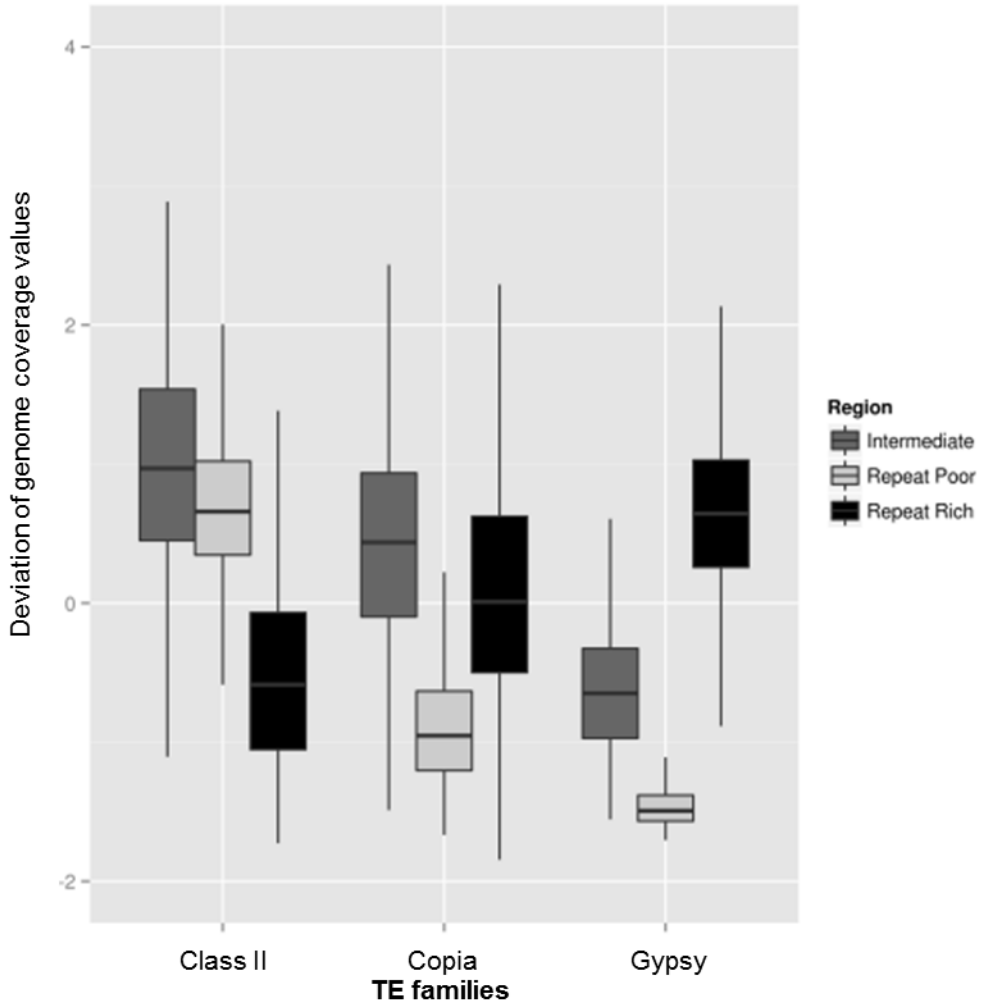


Figure S3. Deviation of genome coverage by the three major repeat families. This boxplot shows the deviation of the genome coverage to the mean value of the three main repeat families of the tomato genome, Gypsy, Copia and Class II elements (bringing together the elements DNA and DNAna). The coverage is calculated by window of 500 kb with an overlap of 50 kb and standardized values are determined based on these calculations. Positive values reflect an enrichment while negative values reflect depletion of that type of repeat. Chi-square P values < 0.001 for each family versus others except for Copia in RR.

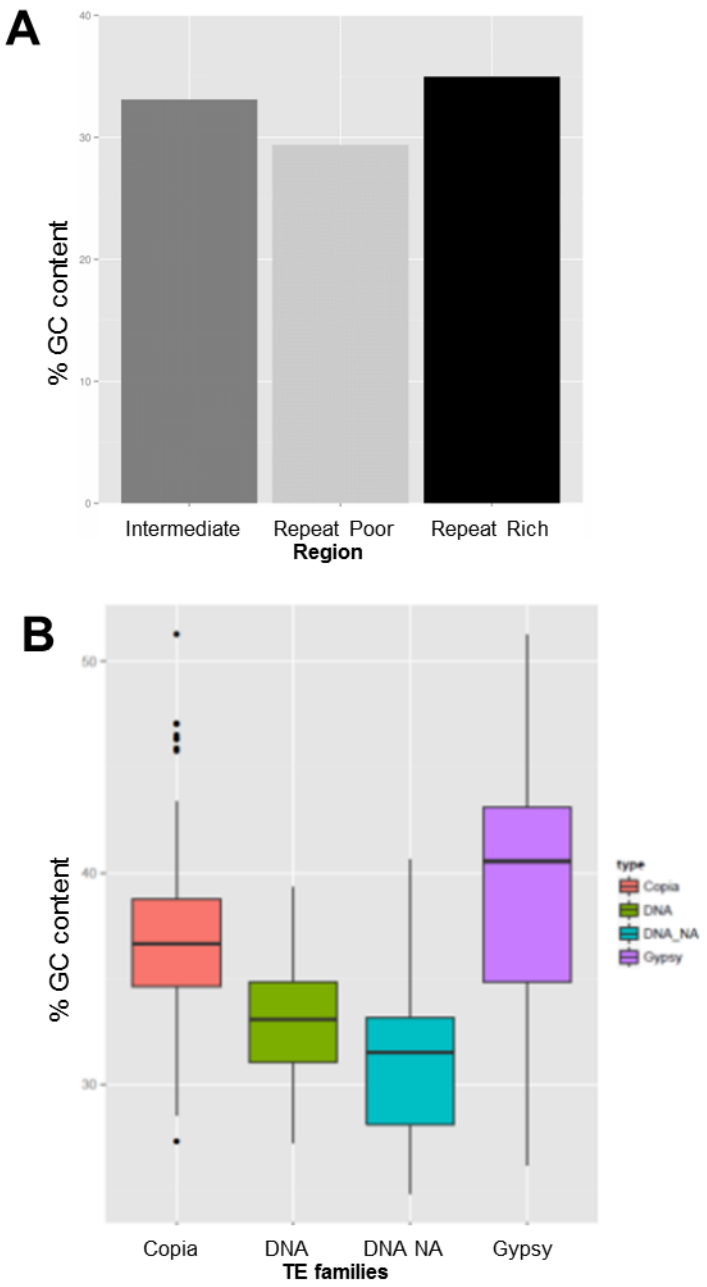


Figure S4. GC content of repeats. (**A**) For each genomic compartment, the percentage of GC in repeats have been calculated. (**B**) Percentage of GC content in the four main families of repeats in tomato genome.

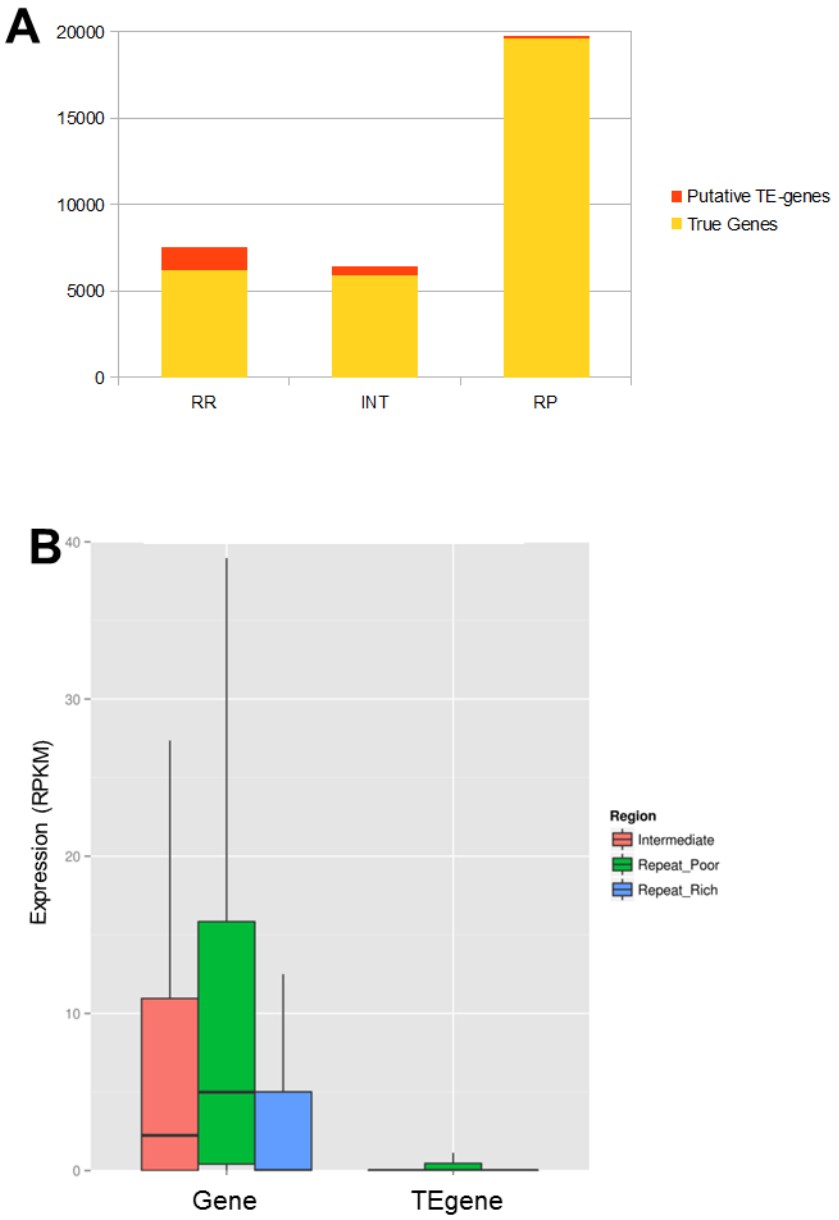


Figure S5. TE-genes identification. (**A**) Proportion of TE-genes and true genes genes in each genomic region of the genome. (**B**) Comparison of the expression of genes and TE-genes in each genomic region.

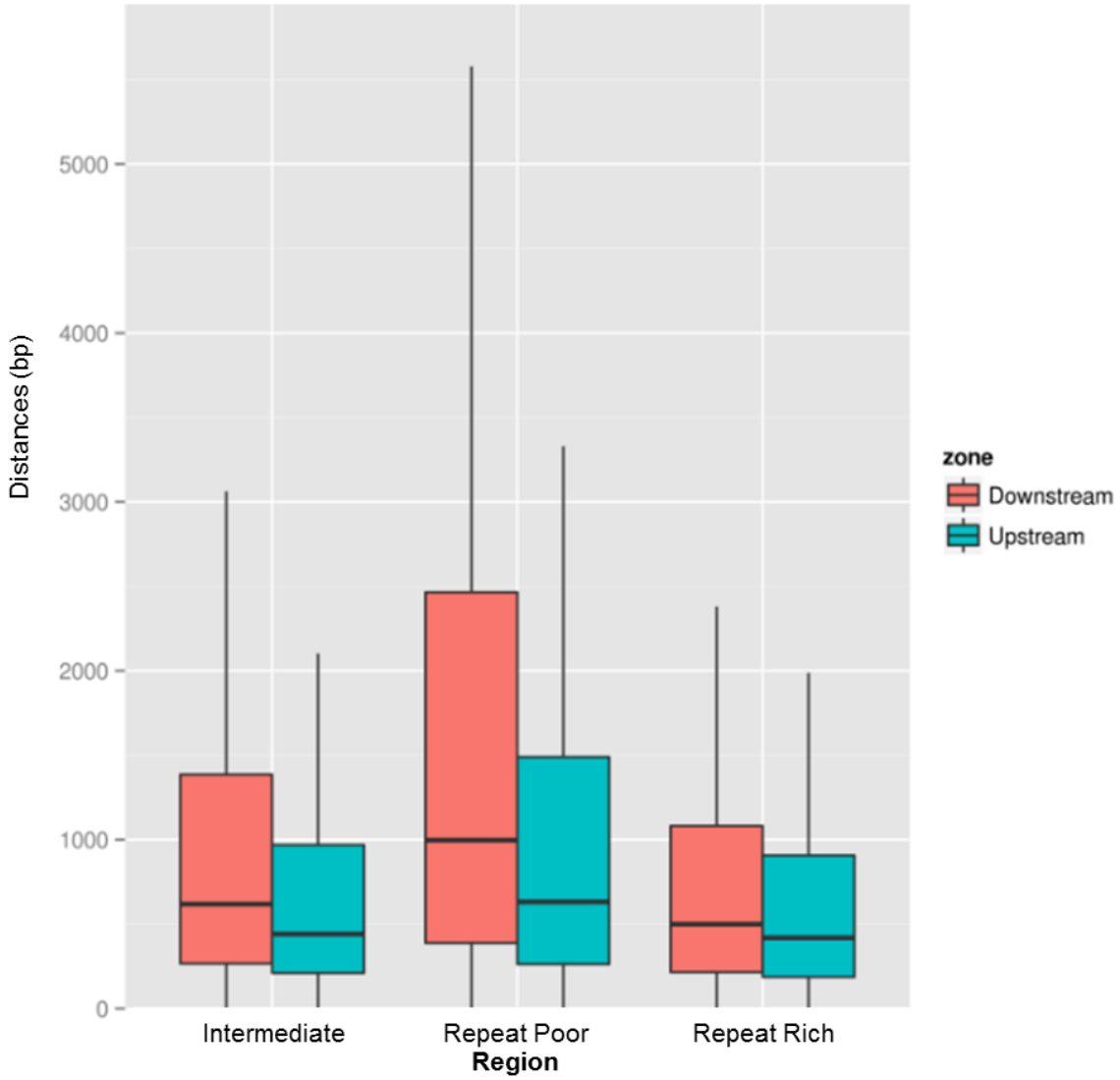


Figure S6. Distance between repeats and genes varies depending on genomic region. After determining for each gene the nearest repeat upstream and downstream to their sequence, we compare these distances between the three genomic regions.

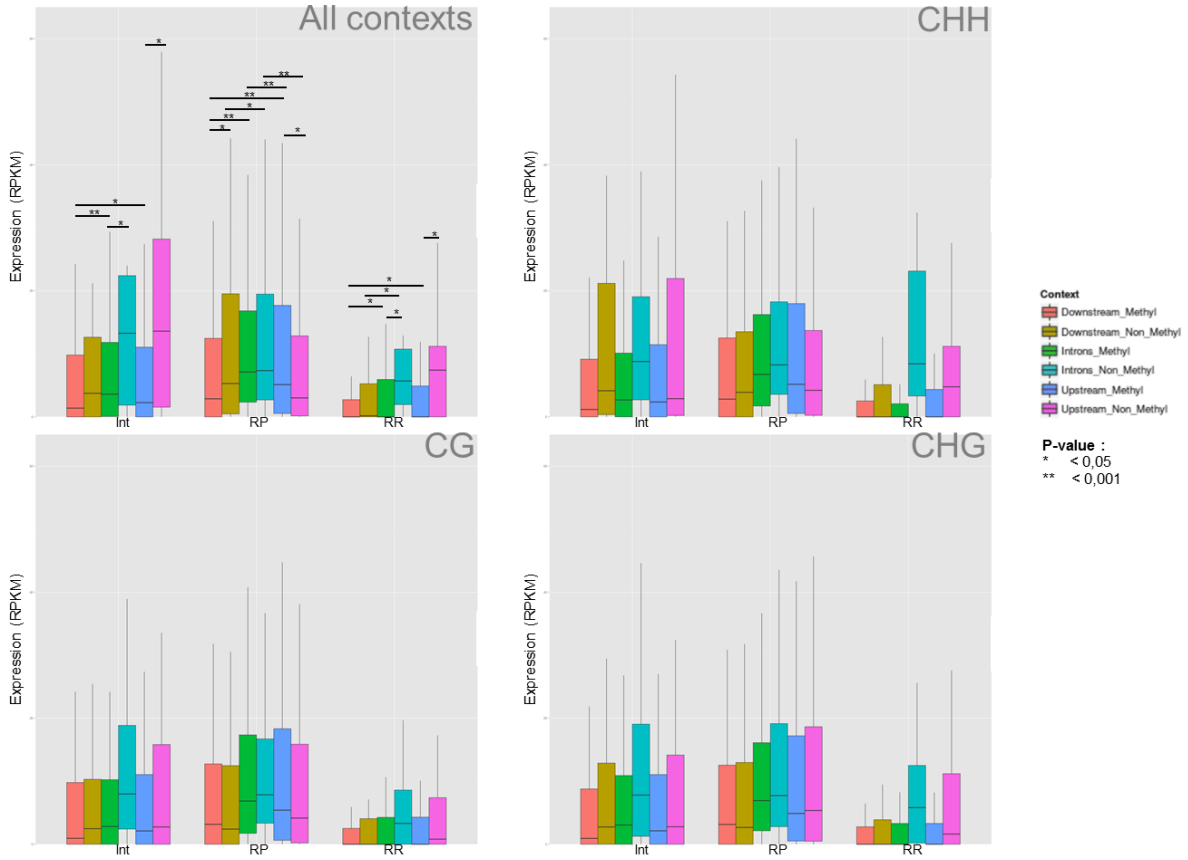


Figure S7. Gene expression levels depending on DNA methylation. Gene expression depending on the location of the repetition and status methylated or not. The four graphs correspond to the three main methylation contexts (CHH, CG, CHG) and all methylation contexts without distinction (All contexts). Mann Whitney statistical analyzes were conducted to test the differences observed and the results are shown in the the « All contexts » with a different symbol depending on the value of the P-value.

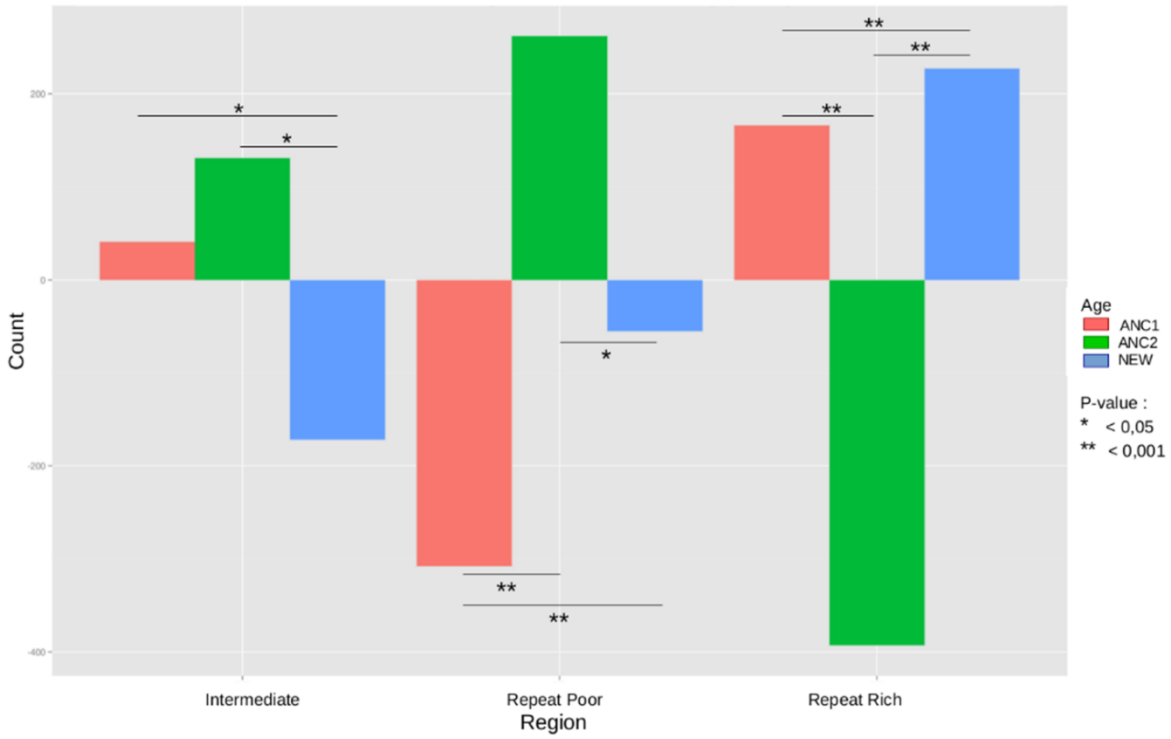


Figure S8. The age of the genes specific to each genomic region. Counting genes considering their phylogenetic origin and comparing that repartition to that expected give us an information about gene age repartition in the three compartments. Statistical analyzes (chi-square tests) were conducted to validate the observations and are represented by the P-value on this graphic.

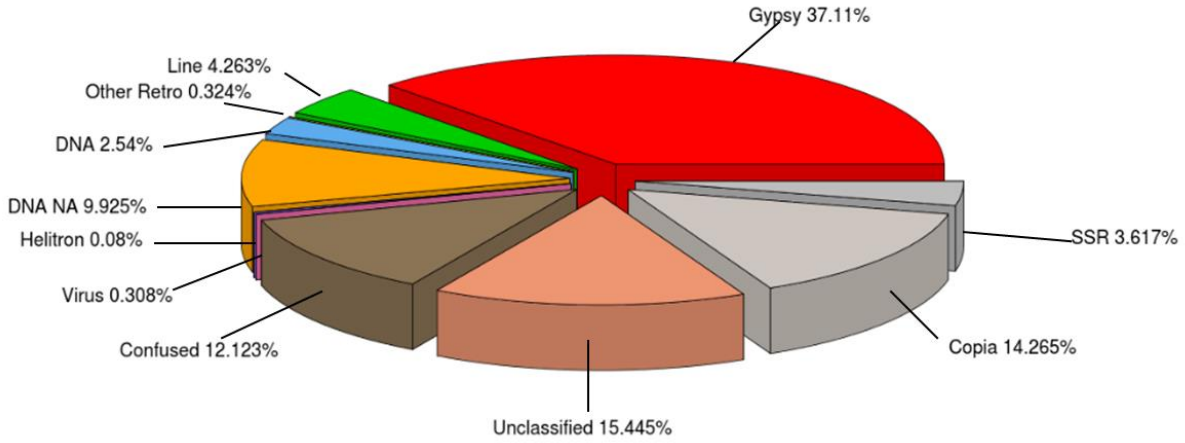


Figure S9. Distribution of associations between repeats and DMRs in the different repeat families. After determining the associations between repeats and DMRs, counting of each family has been achieved and the percentage was defined by relating this count to the total number of defined associations.

2.3 Discussion et conclusion

Le rôle des éléments transposables dans certains processus et chez certains organismes est désormais bien connu. Le but de ce projet était alors de savoir si ces éléments transposables participaient aux modifications épigénétiques et à la régulation de l'expression des gènes dans le processus de maturation chez la tomate. En effet, chez cet organisme, les répétitions sont déjà connues pour avoir un rôle important dans la forme du fruit (ronde ou allongée) (Xiao, H., Jiang, N., Schaffner, E. J. & van der Knaap, E., 2008) et dans la qualité de celui-ci (Quadrana, L. *et al.*, 2014). Pour répondre à ces questions, une ré-annotation complète des éléments répétés, couvrant plus de 70% du génome de la variété Heinz 1706 a été effectuée à l'aide d'outils complémentaires. Des données de méthylome et d'expression des gènes étaient également disponibles. Une série d'analyses, visant à déceler l'impact global des ET sur ce génome, et plus particulièrement le lien pouvant exister entre les ET, les DMRs et les gènes, a alors été mise en place.

Grâce à l'analyse du contenu en gènes et en ET du génome de *S. lycopersicum*, il a été possible de proposer une compartimentation de celui-ci en trois grandes régions génomiques. L'une d'elles est ainsi riche en répétitions et possède une densité en gènes plus faibles, il s'agit de la région *Repeat Rich* (RR). Une autre est riche en gènes et dispose d'une densité en ET plus faible, c'est la région *Repeat Poor* (RP). Enfin, la région *Intermediate* (Int) a une densité en gènes et en ET intermédiaire à celle des deux autres régions. Compte tenu de cette hétérogénéité de la densité en répétitions le long des chromosomes, les analyses ont été menées par type de région afin de ne pas masquer leurs spécificités. Il a ainsi été possible de constater que ces régions se différenciaient non seulement par leur contenu en gènes et en répétitions, mais également par le taux en G+C des répétitions qu'elles contiennent, et par la dynamique des différentes super-familles d'ET et ainsi l'environnement qu'ils constituent pour les gènes. Ainsi, la région RR offre un environnement riche en répétitions au sein duquel les gènes, dont la plupart sont apparus récemment dans la phylogénie des plantes, sont relativement proches des ET. Au contraire, les gènes de la région RP, enrichis en orthologues présents chez l'ancêtre commun aux trois espèces étudiées (*Mimulus guttatus*, *Solanum lycopersicum* et *Solanum tuberosum*), se trouvent plus éloignés des ET. Ces mêmes compartiments présentent une composition biaisée en termes de fonctions de gènes prédites et la répartition des éléments transposables n'y est pas uniforme, la région RR contenant principalement des éléments de type Gypsy et Copia alors que la région RP contient quant à elle essentiellement des éléments de Classe II. Ces nombreuses différences observées entre

nos trois régions suggèrent qu'elles sont soumises à des dynamiques d'insertion et de délétion différentes et que le contrôle des éléments répétés insérés y varie également. Il serait alors intéressant de déterminer si certains gènes des différents compartiments, surtout ceux de la région RR, constituent des néogènes issus de l'acrétion de copies d'ET et si certaines copies d'ET sont maintenues dans le génome au cours de l'évolution pour leur impact sur certains gènes.

Une étude cherchant à déterminer, dans chaque compartiment du génome, s'il existe un lien entre l'expression des gènes, la présence d'ET et la méthylation de ceux-ci, a également été menée. En effet, on suppose que l'impact des ET au sein de chaque compartiment va différer compte tenu des variations de l'environnement des gènes dans chacun d'eux. Cependant, seule une corrélation positive entre la localisation des ET à moins de 1kb en *upstream* ou dans un intron du gène, et le niveau d'expression de ce gène, a pu être observée. L'impact de la méthylation de l'ET sur cette régulation transcriptionnelle n'est également pas clairement visible. L'ensemble de ces résultats ne permet donc pas de prouver l'impact direct des éléments transposables dans le génome de la tomate mais ils suggèrent néanmoins que la présence de répétitions à proximité des gènes peut participer à la régulation de leur expression. Cette idée est soutenue par le fait que la majorité des ET de ce génome sont anciens et que la plupart d'entre eux vont donc avoir un effet neutre ou adaptatif, les copies délétères ayant déjà été éliminées durant l'évolution.

Dans le but d'affiner les résultats obtenus sur l'étude du lien entre les ET et l'expression des gènes, une analyse plus spécifique, ciblant les gènes différentiellement exprimés au cours de la maturation du fruit, a été menée en comparant différents stades de maturation de la tomate. Cette comparaison a montré qu'un nombre important de gènes différentiellement exprimés pendant la maturation se trouvent à proximité d'une DMR portée par un élément transposable. Cette observation suggère que les éléments transposables sont un support privilégié des méthylations et déméthylations, ayant dès lors le potentiel de fournir des contextes chromatiniens contrôlés de façon dynamique au cours de la maturation du fruit, et ce notamment, dans le voisinage des gènes différentiellement exprimés. En effet, si la présence d'une copie d'ET en amont d'un gène est souvent considérée comme potentiellement délétère, certaines deviennent en quelque sorte « domestiquées », c'est-à-dire que leur niveau de méthylation est régulé et agit comme un rhéostat sur l'expression de gènes flanquants.

Une partie des résultats obtenus chez *Solanum lycopersicum* a déjà été observée chez d'autres

espèces. Ainsi, il avait été montré chez *Arabidopsis* que le génome de cet organisme pouvait être divisé en deux principales régions (Ahmed, I., Sarazin, A., Bowler, C., Colot, V. & Quesneville, H., 2011) : une région riche en gènes et pauvre en ET et une région pauvre en gènes et riche en ET, ce qui se rapproche de la compartimentation proposée chez la tomate en trois : la région RR (*Repeat Rich*) située principalement dans les régions péricentromériques, une région RP (*Repeat Poor*) correspondant aux parties distales des chromosomes, et ajoutant une région dite intermédiaire, INT, aux deux autres, servant principalement de « zone tampon » entre celles-ci. Toujours chez *Arabidopsis*, la plupart des ET ont pu être observés avec un statut méthylé dans les trois contextes possibles (CG, CHG, CHH) et l'effet de cette méthylation sur les gènes adjacents semble principalement délétère au niveau de l'initiation de la transcription (Ahmed, I., Sarazin, A., Bowler, C., Colot, V. & Quesneville, H., 2011), là où chez la tomate, l'effet de la méthylation n'a pu être montrée de façon précise. Chez l'Homme et la souris, il a également été montré que certains ET exprimés se trouvaient plus abondamment qu'attendu à proximité des gènes transcrits (Faulkner, G. J. *et al.*, 2009), résultat que l'on retrouve chez la tomate où certaines familles d'éléments transposables sont plus fréquemment présentes à proximité des gènes différentiellement exprimés au cours de la maturation du fruit. Finalement, chez différents hôtes, les ET sont présentés comme des facteurs adaptatifs permettant notamment la réponse au stress (Chénais, B., Caruso, A., Hiard, S. & Casse, N., 2012) (Stapley, J., Santure, A. W. & Dennis, S. R., 2015) (Negi, P., Rai, A. N. & Suprasanna, P., 2016), et les variants de présence / absence de ces ET sont fréquemment associés à des extrêmes locaux d'expression génique et de méthylation, démontrant l'altération de l'expression des gènes voisins et impliquant ces variants comme des sources d'une grande diversité génétique, pouvant conduire à des variations épigénétiques et phénotypiques (Stuart, T. *et al.*, 2016), appuyant ainsi la participation des ET à la régulation transcriptionnelle des gènes adjacents.

Pour terminer, bien que l'impact direct de copies spécifiques sur les gènes adjacents ne soit pas l'objet de cette étude globale, elle montre clairement que les éléments transposables jouent un rôle important chez *S. lycopersicum*, non seulement pour la forme du fruit et sa qualité , mais ils semblent désormais également être le support de la régulation de certains gènes de la maturation du fruit. Des analyses complémentaires, permettant d'obtenir des copies candidates pour leur impact sur les gènes situés à proximité, pourraient permettre d'affiner nos résultats, notamment en testant l'impact de la suppression de ces candidats *in vivo*.

Chapitre 2 : Les éléments transposables : de potentiels éléments de régulation de l'expression des gènes au cours de l'évolution

3.1 Introduction

Chaque organisme vivant dispose d'un patrimoine génétique unique. Afin d'en assurer l'intégrité et le bon fonctionnement, l'expression des gènes doit être constamment contrôlée et finement régulée (Cooper GM). Afin de répondre à cette nécessité biologique, différents mécanismes existent au sein de ces organismes. La majorité de ces mécanismes impliquent l'existence de motifs, ou séquences, de régulation (Harbison, C. T. *et al.*, 2004) (Mariño-Ramírez, L., Tharakaraman, K., Spouge, J. L. & Landsman, D., 2009), qui servent notamment de sites de fixation pour certains des facteurs de régulation de l'expression des gènes. Ces séquences de régulation sont présentes à proximité des gènes qu'ils régulent, mais également à une plus grande distance, auquel cas il est plus difficile de les détecter et de définir à quels gènes ils sont associés.

Ces séquences régulatrices sont maintenues dans les génomes au cours de l'évolution afin de permettre le maintien du potentiel de régulation des différents gènes. Ils permettent de moduler l'expression des gènes selon le contexte dans lequel se trouve l'organisme. Mais, les processus de régulation et les séquences qui y contribuent peuvent être amenés à évoluer en fonction des contraintes de sélection subies par le génome. Ce type d'adaptation des génomes survient fréquemment lors d'un changement d'environnement. Le devenir des modifications génomiques subies peut alors prendre trois directions (Chuong, E. B., 2016) : si la séquence n'a plus d'effet sur le génome alors elle subira un processus de dégénérescence neutre avec des mutations aléatoires, si la séquences a un effet délétère, elle sera éliminée rapidement, et enfin, si la séquence apporte un avantage sélectif, elle se trouvera sous pression de sélection afin d'être conservée au plus près de sa séquence d'origine.

Chez certains organismes, il a été montré que les éléments transposables participent aux processus de régulation de l'expression des gènes de différentes manières. Ils peuvent ainsi apporter de nouvelles régions régulatrices telles les enhancers ou un nouveau promoteur (Rebollo, R., Romanish, M. T. & Mager, D. L., 2012). Les éléments de régulation d'un gène apporté par un élément transposable, s'il n'est pas éliminé par le génome, peut substituer et s'ajouter aux éléments de régulation déjà existants. Chez la tomate S. lycopersicum, les

éléments répétés sont connus pour leur impact sur la forme et la qualité du fruit (Xiao, H., Jiang, N., Schaffner, E. & Stockinger, E. J., 2008), mais ils semblent aussi potentiellement avoir un rôle important dans les processus de maturation du fruit en étant porteurs d'une partie de la dynamique chromatinienne régulant les gènes différentiellement exprimés au cours de ce processus. Dans ce contexte, il est alors possible que certaines copies des répétitions du génome de S. lycopersicum se trouvent dans des positions particulières à proximité des gènes et aient été sélectionnées au cours de l'évolution pour leur impact sur la régulation des gènes.

Afin d'avoir la possibilité de rechercher de telles séquences dans le génome, différents a priori, reposant sur les connaissances actuelles, ont été posés. Tout d'abord, les régions d'intérêt à étudier ont été spécifiées : les régions régulatrices des gènes peuvent se situer à proximité ou à des distances importantes de ceux-ci, mais pour notre étude, seules les régions proches des gènes seront étudiées. Ensuite, le second postulat, est celui de la sélection positive des copies présentant un avantage sélectif pour l'organisme (Chuong, E. B., Elde, N. C. & Feschotte, C., 2016) (González, J., Lenkov, K., Lipatov, M., Macpherson, J. M. & Petrov, D. A., 2008), c'est-à-dire la fixation de modifications génétiques apportant un nouveau phénotype avantageux, et créant alors une nouvelle population. En effet, il est supposé que si une copie d'un élément transposable apporte un motif de régulation qui confère un avantage à l'organisme qui en est hôte, alors elle sera sous contrainte sélective afin d'être conservée au cours de l'évolution. Finalement, certains gènes, participant à une même voie métabolique ou un même processus biologique, sont organisés en réseaux co-régulés. Il est alors possible de supposer que les séquences à l'origine d'une telle co-régulation peuvent découler de l'insertion d'un même élément transposable en différents endroits du génome (Feschotte, C., 2008).

La question est alors de savoir si certaines des copies répondant à nos a priori ont été sélectionnées au cours de l'évolution, et si de telles insertions peuvent être détectées par une approche méthodologique in silico afin d'être étudiées de manière plus approfondie. Pour tenter de répondre à cette problématique, une série d'analyses successives a été mise en place et réalisée sur l'ensemble des familles d'éléments transposables afin de détecter celles dont la répartition et l'orientation en amont des gènes semblent non aléatoire. Parmi les familles sélectionnées, certaines ont également la particularité d'être associées à une fonction de gènes spécifique pouvant indiquer un rôle dans la mise en place de réseaux biologiques.

3.2 Résultats

3.2.1 Constitution du jeu de données

Le génome de S. lycopersicum est composé d'un grand nombre d'éléments transposables. Après avoir montré que ces ET pouvaient avoir un impact global sur la régulation de l'expression des gènes lors du processus de maturation de la tomate (Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. & Maumus, F., 2016), une étude plus précise du lien pouvant exister entre les gènes et ces ET a été réalisée. Pour ce faire, un jeu de données constitué de paires d'éléments répétés et de gènes a été constitué à partir des positions connues de chacune de ces entités.

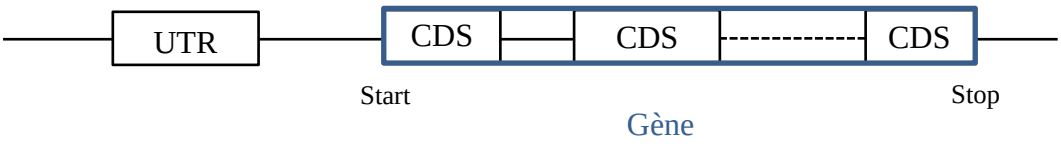


Figure 16 : Schéma du processus pour formater la liste de gènes. L'UTR se trouve exclu de la séquence formatée du gène (cadre bleu) puisque celui-ci débute à la première base de la première CDS du gène et se termine à la dernière base de la dernière CDS du gène.

Afin de constituer ce jeu de données, une liste contenant la position des 34 725 gènes connus du génome de la tomate était disponible. Pour pouvoir exploiter cette liste, différentes étapes permettant de formater et de filtrer les données ont été réalisées. La première étape de formatage des données a été de définir un gène comme limité à sa séquence codante afin de ne pas inclure dans la séquence de ce gène les éléments régulateurs qui peuvent lui être associés. C'est-à-dire que pour chaque gène, la première base faisant partie de la première séquence codante (CDS) qui le constitue est devenue la coordonnée de départ de ce gène, son start, et la dernière base de la dernière CDS de ce gène est devenue sa coordonnée de fin, son stop (Figure 16). Après cette étape, qui ne modifie pas le nombre de gènes inclus dans la liste, les TE-gènes, qui sont des gènes dont au moins la moitié de la séquence codante est chevauchée par un élément transposable et peuvent donc être des erreurs d'annotation, ont été éliminés, réduisant alors à la liste de gènes utilisables à 32 480. Finalement, pour permettre une bonne interprétation des liens qui seront obtenus entre ET et gènes, les gènes partageant un upstream commun et se situant à une distance inférieure à un seuil défini ont été exclus de l'analyse.

Pour définir ce seuil, la distance intergénique médiane, qui correspond à la médiane des distances entre deux gènes successifs, a été évaluée. Elle a alors été calculée de manière globale, sur l'intégralité des gènes du génome, mais également au sein des différentes régions de ce génome. Ce calcul a permis de valider que dans la région Repeat Poor, les gènes sont plus proches les uns des autres (distance médiane de 3 532,5 bases) que dans la région Repeat Rich (distance médiane de 22 787 bases) comme attendu. Cependant, en regardant les résultats obtenus en calculant une distance intergénique moyenne dans les différentes régions (RP : 3 321 bases, INT : 2 948 bases, RR : 2 538 bases), il est possible de remarquer que, bien que les gènes semblent éloignés dans le compartiment RR comme l'indique la médiane de la distance intergénique, la moyenne de ces mêmes distances révèle qu'en réalité les gènes tendent à se regrouper et ainsi former des ilôts de gènes au sein des régions riches en répétitions.

Souhaitant que les résultats futurs soient les moins biaisés possible par ce rapprochement des gènes dans certaines régions, nos analyses reposeront sur une distance médiane. Ainsi, dans l'ensemble du génome, c'est-à-dire sans tenir compte des différents compartiments, la distance intergénique médiane est de 4 524 bases, ce qui permet d'approximer la distance intergénique de la tomate à environ 5kb. Ainsi, pour filtrer la liste de gènes, tous ceux partageant un upstream commun et se trouvant à une distance inférieure à 5kb, ont été éliminés pour la suite de l'étude. Après ce nouveau filtre, seuls 25 621 gènes restent exploitables sur les 32 480 de départ. En parallèle, une liste de 636 643 annotations d'éléments répétés (Figure 17a), excluant les *Potential Host Gene*, est disponible.

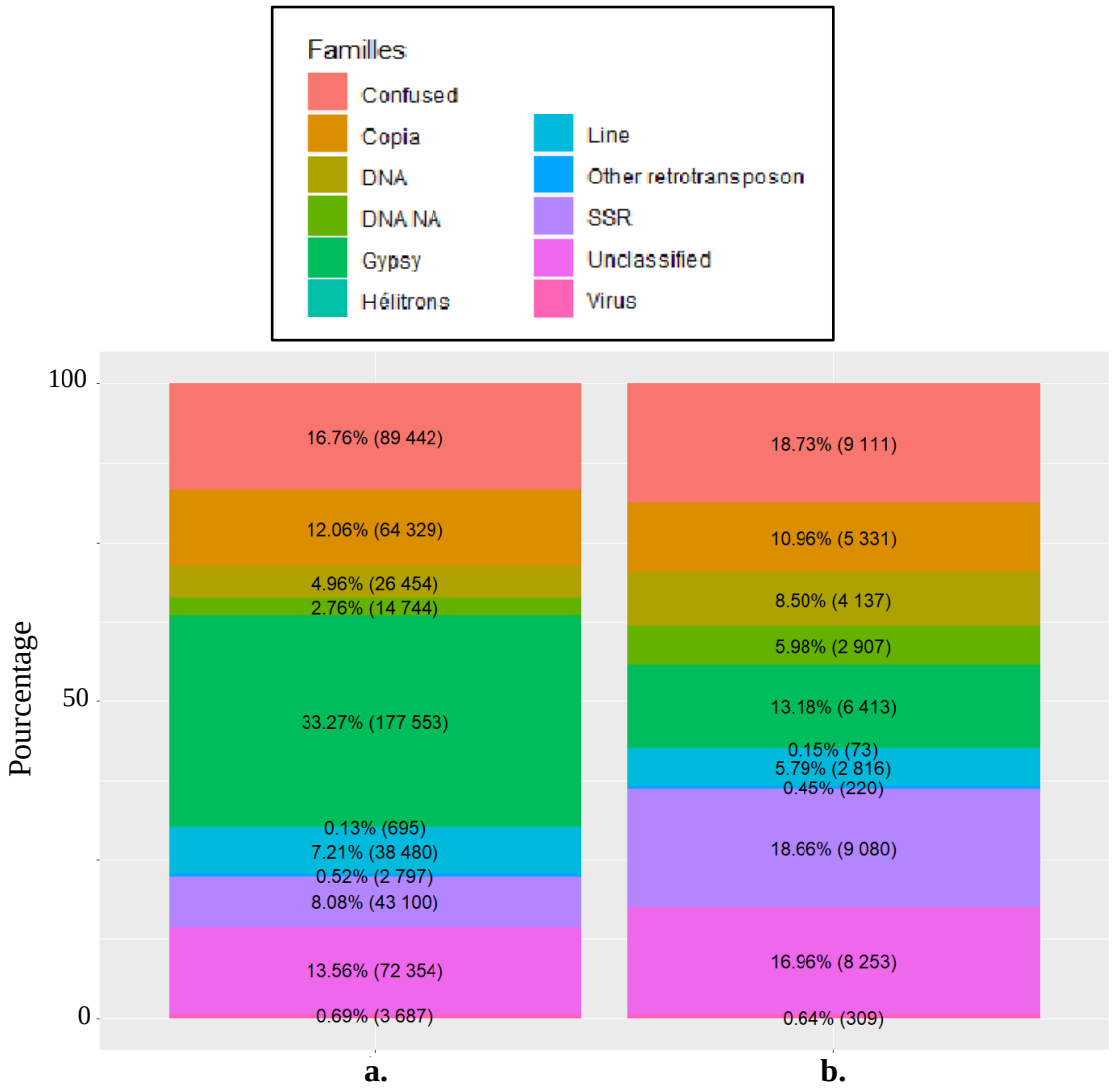


Figure 17 : Répartition des éléments transposables dans les différentes familles. a. Composition du génome entier en différentes familles d'éléments transposables. b. Composition en différentes familles des éléments transposables impliqués dans les tuples avec les gènes.

Une fois l'ensemble de ces étapes préliminaires réalisées, permettant d'obtenir des sets de données nettoyés, une distance d'étude maximale de 2,5kb, soit la moitié de la distance intergénique médiane globale, a été définie comme région upstream de chaque gène. Les paires correspondantes à la combinaison unique d'un gène et d'un élément transposable, ont alors pu être formés. Pour cela, à chaque élément transposable est associé son gène le plus proche de manière à ce que l'ET se trouve toujours dans la région upstream du gène. De cette

manière, un total de 536 734 paires est trouvé avec des distances allant de 0 base (chevauchement entre l'ET et le gène) à 2 502 944 bases. Cependant, les paires présentant un chevauchement entre la répétition et le gène sont éliminés car pour ces cas, il peut être difficile d'établir quelle partie de l'ET pourrait influencer l'expression du gène. De même, seules les paires dont la distance calculée par l'outil est inférieure ou égale à 2,5kb ont été étudiés puisque, pour ce travail, seules les régions proches des gènes veulent être étudiées. Au final, 49 055 paires forment le jeu de données qui a ensuite été étudié, impliquant un total de 20 717 gènes et 49 055 ET ; les SSR (Simple Sequance Repeat, séquences répétées en tandem), et les microsatellites (séquences constituées d'unités répétées de 1 à 4 nucléotides), étant les séquences les plus abondantes suivies par les éléments Gypsy et Copia (Figure 17b). Ce résultat indique qu'environ 81% (20 717 / 25 621) des gènes ont une répétition dans leur région upstream à moins de 2,5kb de leur séquence codante. Or, cette région contenant fréquemment les éléments régulateurs des gènes, une étude de l'impact des ET comme éléments régulateurs potentiels de ces gènes a donc été réalisée.

3.2.2 Une organisation particulière des ET par rapport aux gènes

Afin de savoir si les éléments transposables détectés à proximité des gènes se trouvent ou non là de manière aléatoire en raison de leur grande fréquence dans le génome, une étude de la répartition de ces ET a été menée. Ainsi, la première analyse a eu pour but de déterminer si les répétitions étaient plus ou moins fréquentes qu'attendu à proximité des gènes, c'est-à-dire est-ce que la répartition observée des ET est celle que l'on attendrait sous l'hypothèse d'une répartition aléatoire dans le génome. Pour répondre à cette question, des effectifs théoriques à proximité des gènes ont été calculés et comparés aux effectifs réels par un test statistique.

De manière générale, en étudiant l'intégralité des éléments répétés, les ET ont été trouvés plus fréquemment positionnés à proximité des gènes qu'attendu sous l'hypothèse d'une répartition aléatoire (chi-square P value < 0.001 et x-squared = 639,4884). En réalisant la même étude à l'échelle de chaque famille d'ET, seuls les endovirus, les Hélitrons et la catégorie Other retrotransposons ne passent pas l'analyse statistique (Tableau 2) et correspondent donc à une répartition attendue sous l'hypothèse d'une répartition aléatoire. Cependant, ces trois types d'éléments répétés sont assez peu fréquents dans le génome de la tomate par rapport aux autres : de l'ordre de quelques centaines à quelques milliers de copies selon la famille.

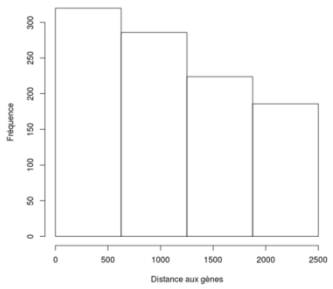
Famille		*Upstream*	Non *upstream*	P-value	X-squarred
Confused	**Réel**	9 111	64 101	< 2,2 e-16	978,76
	Théorique	8 154	81 288		
Copia	**Réel**	5 331	58 998	2,589 e-13	53,4974
	Théorique	5 865	58 464		
DNA	**Réel**	4 137	22 317	< 2,2 e-16	1357,443
	Théorique	2 412	24 042		
DNA NA	**Réel**	2 907	11 837	< 2,2 e-16	1999,996
	Théorique	1 344	13 400		
Gypsy	**Réel**	6 413	171 140	< 2,2 e-16	6493,731
	Théorique	16 187	161 366		
Hélitrons	**Réel**	73	622	0,1864	1,7455
	Théorique	63	632		
Line	**Réel**	2 816	35 664	< 2,2 e-16	150, 1991
	Théorique	3 508	34 972		
Other retrotransposon	**Réel**	220	2 577	0,0215	5,2858
	Théorique	255	2 542		
SSR	**Réel**	9 080	34 020	< 2,2 e-16	7430,426
	Théorique	3 929	39 171		
Unclassified	**Réel**	8 253	64 101	< 2,2 e-16	458,0135
	Théorique	6 596	65 758		
Virus	**Réel**	309	3 378	0,1223	2,3872
	Théorique	336	3 351		

Tableau 2 : Résultats de l'analyse statistique de la répartition des différentes familles d'éléments transposables en *upstream* ou non des gènes.

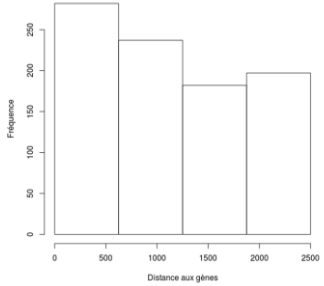
Un biais de répartition des ET dans le génome peut donc être observé, ceux-ci semblant privilégier une présence dans les régions upstream à proximité des gènes. Deux hypothèses sur ce biais peuvent alors être posées. La première est que ce biais est lié à un niveau plus faible de compaction de la chromatine dans ces régions pour permettre l'expression des gènes. L'insertion plus fréquente des ET dans ces régions est donc uniquement liée à ces différences de compaction. La seconde est que les ET s'insèrent de manière relativement homogène dans le génome mais seules certaines insertions sont sélectionnées par le génome pour l'impact

qu'elles ont sur les gènes situés à proximité, et elles sont donc les seules à pouvoir être détectées, les autres étant éliminées. Les deux hypothèses semblent également pouvoir coexister. Cependant, dans le cas où certaines insertions sont sélectionnées, d'autres caractéristiques indiquant une participation des ET à la régulation transcriptionnelle des gènes devraient être observables.

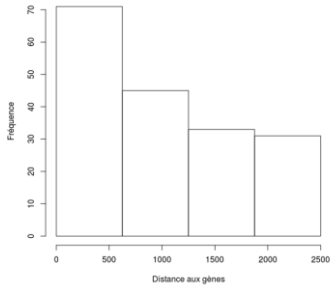
Dans cette même région de 2,5kb en upstream des gènes, un biais de la répartition des séquences des copies d'ET pourrait en effet être observé, les ET ainsi placés agiraient alors potentiellement comme régulateurs à une distance spécifique du gène. Pour vérifier cette possibilité, une analyse de la distribution des ET dans la région upstream de 2,5kb des gènes est réalisée en séparant en 4 fenêtres de 625 bases cette région : 0 à 625 bases, 625 à 1250 bases, 1250 à 1875 bases et 1875 à 2500 bases, la borne inférieure de chaque fenêtre étant non incluse contrairement à la borne supérieure. Pour chacune des fenêtres, un comptage des effectifs réels est réalisé et un calcul des effectifs attendus sous l'hypothèse d'une répartition aléatoire est effectué. Finalement, une analyse statistique comparant la répartition observée à celle attendue est réalisée pour toutes les familles disposant d'au moins 20 copies dans les régions upstream des gènes comme défini ci-dessus. Un nombre de 818 familles de séquences répétées était disponible au début de cette analyse. Une fois le filtre imposant un nombre minimum de 20 copies appliqué, il n'en reste plus que 475 (58,07%). Au final, sur ces 475 familles testées, 31 (6,53%) ont une répartition qui ne correspond statistiquement pas à celle attendue sous l'hypothèse d'une répartition aléatoire. Ces 31 familles ont cependant des profils de distribution en upstream des gènes assez différents (Figure 18), indiquant que leur impact potentiel n'est pas limité à une seule position spécifique mais que celle-ci est éventuellement propre à chaque famille d'ET



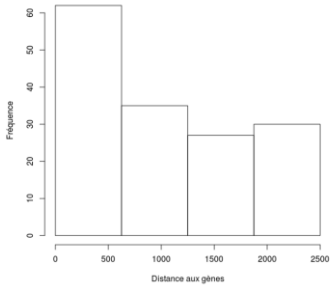
DHX-incomp-chim_Slyco_light-B-R4878-Map5



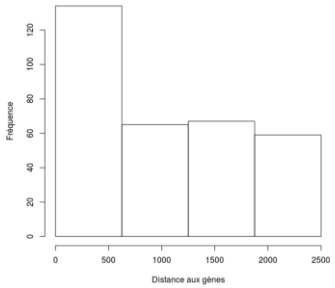
DTX-incomp-chim_Slyco_light-B-R6087-Map5



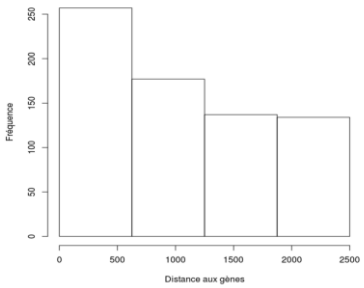
DTX-incomp_Slyco_light-B-G1226-Map7_reversed



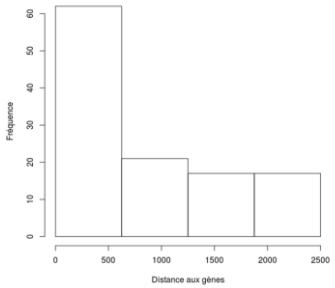
DTX-incomp_Slyco_light-B-R1099-Map20



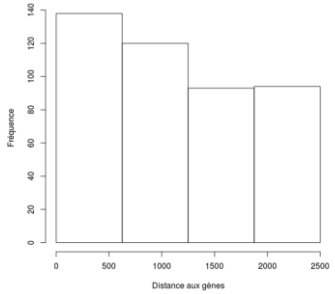
DTX-incomp_Slyco_light-B-R319-Map20_reversed



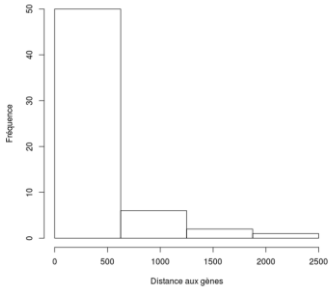
DTX-incomp_Slyco_light-B-R4501-Map6



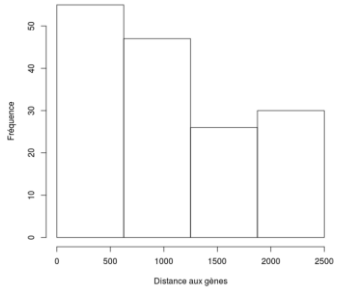
DXX_Slyco_light-B-R1024-Map10



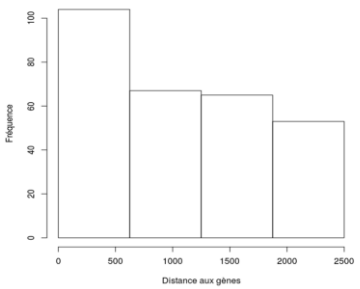
DXX_Slyco_light-B-R2136-Map5_reversed



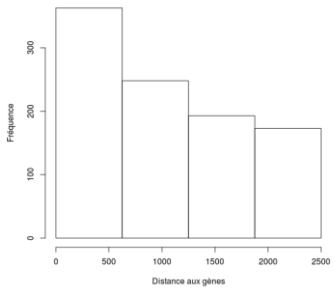
DXX_Slyco_light-B-R732-Map8



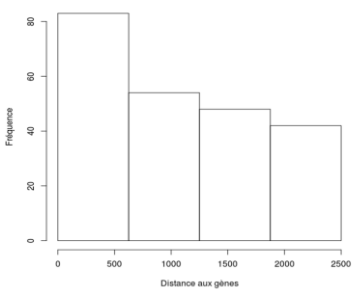
noCat_Slyco_light-B-R1426-Map5



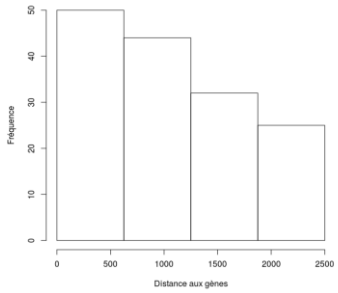
noCat_Slyco_light-B-R2058-Map10



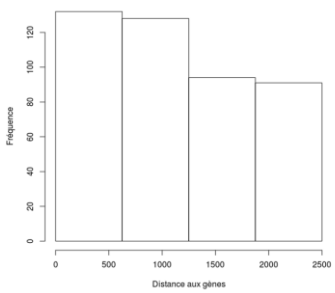
noCat_Slyco_light-B-R2684-Map20



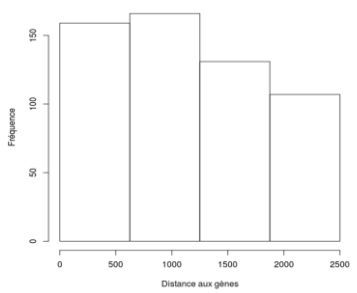
noCat_Slyco_light-B-R5779-Map13



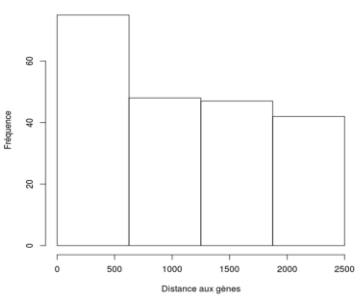
noCat_Slyco_light-B-R5820-Map11



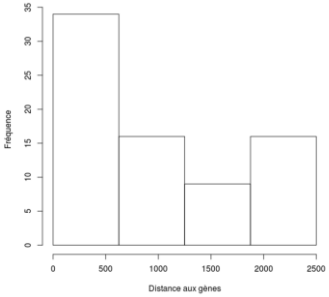
RIX-incomp-chim_Slyco_light-B-R3888-Map7



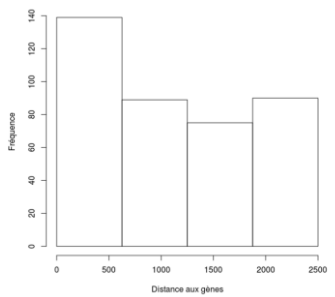
RIX-incomp_Slyco_light-B-R4254-Map8



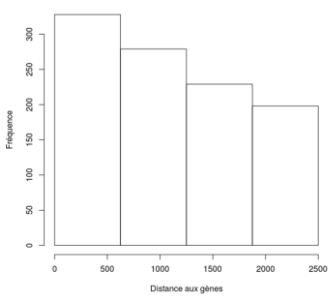
RLX-incomp_Slyco_light-B-R2672-Map5_reversed



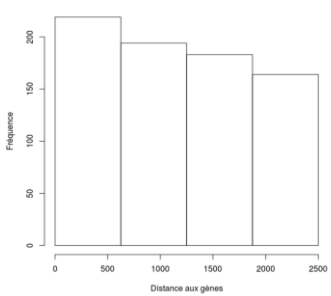
RLX-incomp_Slyco_light-B-R3040-Map10



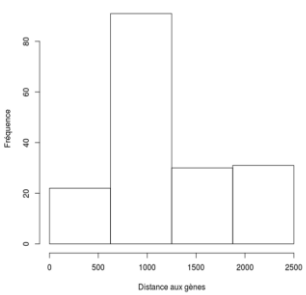
RLX-incomp_Slyco_light-B-R5043-Map13_reversed



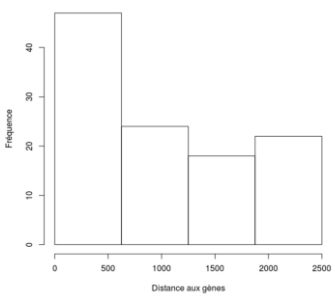
RLX-incomp_Slyco_light-B-R5783-Map5



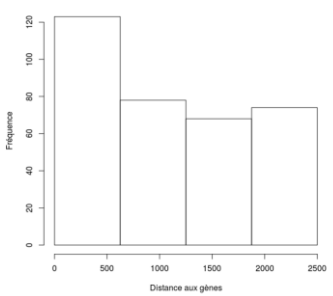
RXX_Slyco_light-B-R1157-Map8



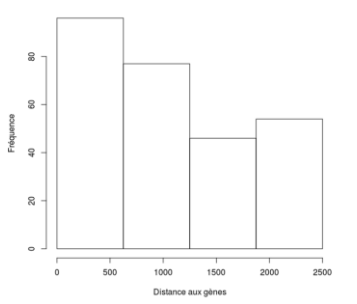
RXX-TRIM_Slyco_light-B-R131-Map20



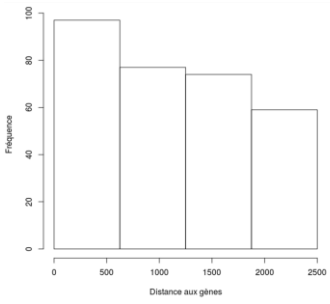
SSR_Slyco_light-B-G988-Map5



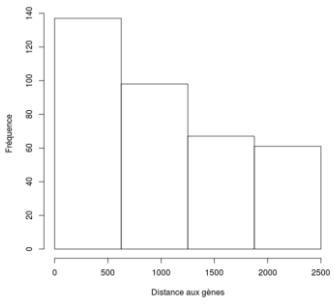
SSR_Slyco_light-B-P170.263-Map11



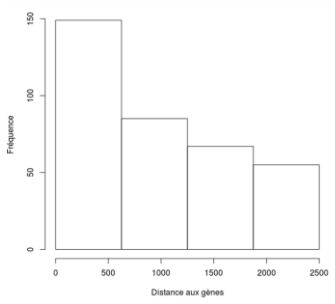
SSR_Slyco_light-B-P35.334-Map5



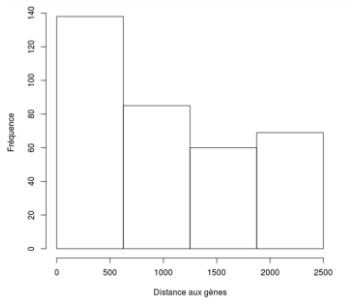
SSR_Slyco_light-B-P52.217-Map20



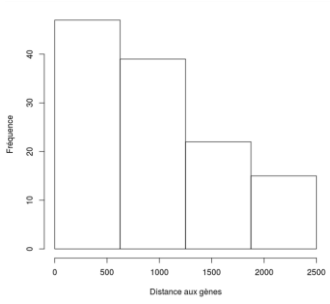
SSR_Slyco_light-B-R1044-Map15



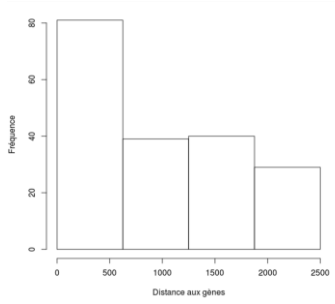
SSR_Slyco_light-B-R2046-Map17



SSR_Slyco_light-B-R2504-Map8



SSR_Slyco_light-B-R630-Map20



SSR_Slyco_light-B-R688-Map20

Figure 18 : Profils de distribution des différentes familles d'éléments transposables en upstream des gènes à une distance maximale de 2,5 kb par fenêtre de 625 bases. Les profils des familles d'éléments transposables présentés ici sont ceux des 31 familles dont le profil ne correspond pas à celui attendu sous l'hypothèse d'une répartition aléatoire d'après les analyses statistiques de type Chi2 réalisées.

.

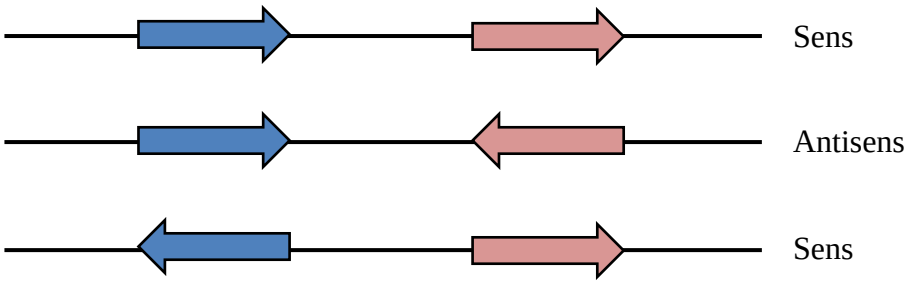


Figure 19 : Schéma des différentes orientations possibles entre les gènes et les éléments transposables. Ce schéma est celui qui permet de définir si les éléments transposables (en bleu) se trouvent en sens ou en antisens par rapport aux gènes (en rouge) afin de réaliser une analyse statistique permettant d'estimer si leur répartition dans ces deux positions correspond à celle sous l'hypothèse d'une répartition aléatoire.

Une autre observation pourrait encore indiquer que certaines familles d'ET ont un impact potentiel sur l'expression des gènes à proximité : l'orientation de la séquence des répétitions par rapport à celle des gènes. C'est-à-dire, est ce que certaines familles d'ET sont retenus dans un sens privilégié en upstream des gènes, leur séquence étant alors soit dans la même orientation, soit en orientation opposée à celle du gène (Figure 19). Pour chaque famille d'ET disposant de plus de 10 copies en upstream, une analyse comparant le nombre réel de copies en sens et en anti-sens par rapport à un attendu, c'est-à-dire autant dans chaque orientation, est réalisé à l'aide d'un test. Sur l'intégralité des familles testées (645), seules 6 (0,93%) ont une orientation préférentielle par rapport au gène. Cependant, bien que ce paramètre révèle que l'orientation des ET pourrait être importante pour leur impact sur les gènes, celui-ci n'est pas exclusif car certains ET présentent des motifs répétés inversés dans leur séquence qui pourraient avoir un impact sur les gènes à proximité et ils ne sont pas retenus lors de cette analyse. Il existe néanmoins des outils afin de détecter de tels motifs qui pourraient donc être utilisés. L'outil einverted (Rice, P. Site : emboss.open-bio.org/rel/dev/apps/einverted.html) a alors servi pour la recherche de tels motifs dans l'ensemble des familles de répétitions. Pour qu'un ET soit considéré comme porteur d'un motif répété inversé, il a été choisi que la séquence de ce motif fasse un minimum de 20 bases. Sur les 818 familles testées au total, 366 (~45%) présentent dans leur séquence un motif répété inversé. Toutefois, sur les 475 consensus comptant plus de 20 copies, c'est-à-dire le sous jeu de données qui est évalué par chaque analyse, 259 (~54,5%) sont concernés par ce type de motif. Les résultats ainsi obtenus,

très nombreux, suggèrent que de nombreuses séquences sont porteuses de motifs inversés, et que pour pouvoir être exploités, les critères permettant de les filtrer devraient être affinés. Les critères de sélection lors de cette analyse étant majoritairement empiriques, les résultats obtenus n'ont pas été définis comme un facteur de sélection des familles considérées comme présentant potentiellement un intérêt.

L'ensemble des analyses effectuées jusqu'à présent révèle donc que 30 consensus ont une distribution particulière en upstream des gènes, que 5 ont une orientation préférentielle dans cette région et qu'un seul présente à la fois une distribution et une orientation particulière (Tableau 3). Il serait alors intéressant de savoir si les gènes associés à ces différentes familles ont des fonctions spécifiques dans le génome de la tomate, c'est-à-dire, s'ils ont une même fonction ou participent à un même processus biologique.

Nom du consensus	Nombre d'associations	P-value : distribution en upstream	P-value : orientation en upstream
DHX-incomp-chim_Slyco_light-B-R4878-Map5	1016	2.7692652e-15	1
DMX-incomp_Slyco_light-B-R733-Map6	68	1	0.00192952125
DTX-incomp-chim_Slyco_light-B-R525-Map18	63	0.2719505625	0.0014906393
DTX-incomp-chim_Slyco_light-B-R6087-Map5	898	4.88907525e-08	1
DTX-incomp_Slyco_light-B-G1226-Map7_reversed	180	4.81509875e-05	1
DTX-incomp_Slyco_light-B-R1099-Map20	154	0.00060786415	1
DTX-incomp_Slyco_light-B-R319-Map20_reversed	325	1.83027e-10	1
DTX-incomp_Slyco_light-B-R4501-Map6	705	8.52493425e-16	1
DXX_Slyco_light-B-R1024-Map10	117	5.312153e-10	8.463892e-08
DXX_Slyco_light-B-R2136-Map5_reversed	445	0.0350676635	1
DXX_Slyco_light-B-R732-Map8	59	7.84727075e-26	1
noCat_Slyco_light-B-R1426-Map5	158	0.0190657495	1
noCat_Slyco_light-B-R2007-Map7	77	1	5.99839025e-15
noCat_Slyco_light-B-R2058-Map10	289	0.000458670925	1
noCat_Slyco_light-B-R2461-Map5	76	1	0.002122483825
noCat_Slyco_light-B-R2684-Map20	977	3.800860225e-28	0.909018425
noCat_Slyco_light-B-R5779-Map13	227	0.000497758675	1
noCat_Slyco_light-B-R5820-Map11	151	0.04405511475	1
RIX-incomp-chim_Slyco_light-B-R3888-Map7	445	0.00552334275	1
RIX-incomp_Slyco_light-B-R4254-Map8	563	0.000594794525	1

RLX-incomp_Slyco_light-B-R2672-Map5_reversed	212	0.0180036305	1
RLX-incomp_Slyco_light-B-R3040-Map10	75	0.0238422355	1
RLX-incomp_Slyco_light-B-R5043-Map13_reversed	393	5.8814595e-07	1
RLX-incomp_Slyco_light-B-R5783-Map5	1034	1.055888425e-12	1
RXX_Slyco_light-B-R1157-Map8	760	0.0049134095	1
RXX-TRIM_Slyco_light-B-R131-Map20	174	0.0107419825	0.4042531675
RYX-incomp-chim_Slyco_light-B-R755-Map20	129	0.061994625	0.00337834535
SSR_Slyco_light-B-G988-Map5	111	0.001405879825	1
SSR_Slyco_light-B-P170.263-Map11	343	0.00011459242	1
SSR_Slyco_light-B-P35.334-Map5	273	3.089536325e-05	1
SSR_Slyco_light-B-P52.217-Map20	307	0.003769726825	1
SSR_Slyco_light-B-R1044-Map15	363	2.317468475e-08	1
SSR_Slyco_light-B-R2046-Map17	356	3.45504075e-16	1
SSR_Slyco_light-B-R2504-Map8	352	3.265236925e-10	1
SSR_Slyco_light-B-R630-Map20	123	0.001844383675	1
SSR_Slyco_light-B-R688-Map20	189	8.481733e-07	1

Tableau 3 : Tableau bilan des résultats des différentes analyses pour les 36 consensus pouvant avoir été sélectionnés au cours de l'évolution pour leur impact sur les gènes. Ce tableau regroupe pour chacun des 36 consensus d'intérêt le nombre de tuples dans lequel il est impliqué, ainsi que les résultats de l'analyse statistique de la distribution des copies de ces consensus en *upstream* des gènes et enfin les résultats de l'analyse statistique de l'orientation de ces copies par rapport aux gènes.

3.2.3 Des fonctions de gènes particulières associées aux ET

D'après les résultats obtenus jusqu'à présent, certaines familles de répétitions se trouvent donc dans des régions particulières des gènes et pourraient avoir un impact sur leur expression. Les gènes sont connus pour être fréquemment organisés en réseaux co-régulés (Chen, D., Yan, W., Fu, L.-Y. & Kaufmann, K., 2018), c'est-à-dire que ceux répondant à une même fonction ou un même processus biologique peuvent être exprimés à partir d'une même séquence promotrice. Chez la tomate, peu de ces réseaux sont cependant connus.

Une vision d'ensemble des fonctions de gènes connus chez la tomate était souhaitée afin de savoir, par la suite, si certaines fonctions sont plus fréquemment associées à la présence d'éléments répétés que ce qui serait attendu sous l'hypothèse d'une répartition aléatoire de ces gènes à proximité des ET. Pour avoir cette vision d'ensemble, la liste des fonctions de gènes et plus spécifiquement leur « Biological process » a été consulté sur le site de référence geneontology.org. D'après ce site, seul environ 33% des gènes de S. lycopersicum ont une fonction connue dans les processus métaboliques ou cellulaires (Figure 20).

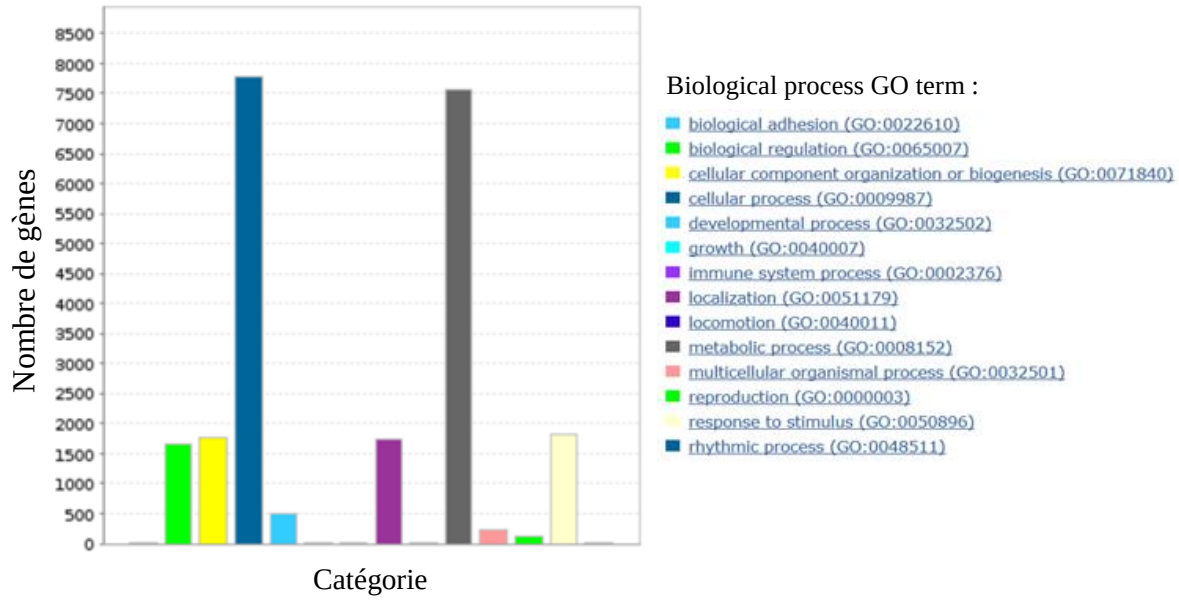


Figure 20 : Proportions en différentes fonctions des gènes de la tomate *S. lycopersicum* d'après le site *geneontology.org*. Cet histogramme présente la proportion de gène dans chaque grande catégorie de *GO term* pour les fonctions de type *Biological process*.

Nom du consensus	*GO term* enrichis
DHX-incomp-chim_Slyco_light-B-R4878-Map5	Defense response Unclassified
DMX-incomp_Slyco_light-B-R733-Map6	Pas de résultat significatif
DTX-incomp-chim_Slyco_light-B-R525-Map18	Pas de résultat significatif
DTX-incomp-chim_Slyco_light-B-R6087-Map5	Chemical homeostasis Response to chemical Regulation of cellular process Metabolic process Unclassified
DTX-incomp_Slyco_light-B-G1226-Map7_reversed	Pas de résultat significatif
DTX-incomp_Slyco_light-B-R1099-Map20	Pas de résultat significatif
DTX-incomp_Slyco_light-B-R319-Map20_reversed	Brassinosteroid homeostasis (lipid homeostasis) Brassinosteroid biosynthetic process (organic cyclic compound biosynthetic prosses) Sterol metabolic process
DTX-incomp_Slyco_light-B-R4501-Map6	Pas de résultat significatif
DXX_Slyco_light-B-R1024-Map10	Pas de résultat significatif
DXX_Slyco_light-B-R2136-Map5_reversed	Pas de résultat significatif
DXX_Slyco_light-B-R732-Map8	Pas de résultat significatif
noCat_Slyco_light-B-R1426-Map5	Pas de résultat significatif
noCat_Slyco_light-B-R2007-Map7	Pas de résultat significatif

noCat_Slyco_light-B-R2058-Map10	Pas de résultat significatif
noCat_Slyco_light-B-R2461-Map5	Cellular metabolic process Primary metabolic process Organic substance metabolic process Unclassified
noCat_Slyco_light-B-R2684-Map20	Pas de résultat significatif
noCat_Slyco_light-B-R5779-Map13	Pas de résultat significatif
noCat_Slyco_light-B-R5820-Map11	Pas de résultat significatif
RIX-incomp-chim_Slyco_light-B-R3888-Map7	Pas de résultat significatif
RIX-incomp_Slyco_light-B-R4254-Map8	Proteasome-mediated ubiquitin-dependent protein catabolic process Protein ubiquitination
RLX-incomp_Slyco_light-B-R2672-Map5_reversed	Pas de résultat significatif
RLX-incomp_Slyco_light-B-R3040-Map10	Pas de résultat significatif
RLX-incomp_Slyco_light-B-R5043-Map13_reversed	Pas de résultat significatif
RLX-incomp_Slyco_light-B-R5783-Map5	Pas de résultat significatif
RXX_Slyco_light-B-R1157-Map8	Transcription, DNA-templated Unclassified
RXX-TRIM_Slyco_light-B-R131-Map20	Pas de résultat significatif
RYX-incomp-chim_Slyco_light-B-R755-Map20	Metabolic process Cellular process Unclassified

SSR_Slyco_light-B-G988-Map5	Pas de résultat significatif
SSR_Slyco_light-B-P170.263-Map11	Unclassified
SSR_Slyco_light-B-P35.334-Map5	Small molecule metabolic process
SSR_Slyco_light-B-P52.217-Map20	Pas de résultat significatif
SSR_Slyco_light-B-R1044-Map15	Pas de résultat significatif
SSR_Slyco_light-B-R2046-Map17	Pas de résultat significatif
SSR_Slyco_light-B-R2504-Map8	Pas de résultat significatif
SSR_Slyco_light-B-R630-Map20	Pas de résultat significatif
SSR_Slyco_light-B-R688-Map20	Pas de résultat significatif

Tableau 4 : Résultats de l'analyse d'enrichissement des *GO term* associant à chaque famille d'éléments transposables les fonctions de gènes qui présentent un enrichissement. L'intitulé « *GO term enrichis* » signifie que la fonction de gène citée est plus fréquente qu'attendue dans cet ensemble par rapport aux proportions trouvées dans l'ensemble des gènes. La catégorie *Unclassified* correspond à des gènes dont la fonction n'était pas encore connue le jour de l'analyse.

Une analyse d'enrichissement des GO term a ensuite été réalisée pour chacune des 36 familles d'éléments répétés qui se trouvent à proximité des gènes dans des configurations particulières suggérant un impact sur la régulation transcriptionnelle de ceux-ci. Pour cela, la liste des noms des gènes associés à une famille spécifique d'ET est soumise à un test d'enrichissement. Parmi les 36 familles, 8 ont un enrichissement pour une ou plusieurs fonctions de gènes spécifiques (Tableau 4). On remarque par exemple que le consensus DHX-incomp-chim_Slyco_light-B-R4878-Map5, un élément de la sous-classe 2 des transposons à ADN, est associé à des gènes pour lesquels l'analyse des GO term révèle un enrichissement dans des fonctions de défense et de réponse au stress de la plante. Ces fonctions sont connues pour être très importantes et pour évoluer rapidement au sein des génomes ce qui rend ce résultat d'autant plus intéressant. Une étude approfondie des séquences d'une telle famille pourrait alors nous en apprendre d'avantage sur une potentielle sélection.

3.2.4 Comparaison des copies proches et éloignées des gènes

Certaines familles d'éléments transposables semblent alors dans des positions particulières par rapport aux gènes. Afin de les étudier de manière plus approfondie, dans le but de savoir si certaines copies pourraient avoir été sélectionnées au cours de l'évolution pour leur impact sur les gènes, un alignement de leurs séquences a été réalisé par famille. Cet alignement a pour objectif de comparer les profils de délétion des séquences de copies proches de gènes à celles des copies qui en sont éloignées afin d'identifier, s'il y en a, de potentiels motifs spécifiques aux copies proches des gènes.

Ce travail a donc été effectué pour les 36 familles d'intérêt. La première difficulté rencontrée a été de définir le nombre de séquences limites à utiliser pour l'alignement, l'outil ne permettant pas un alignement de toutes les séquences au complet. Après avoir testé des jeux de données de tailles différentes, il a été choisis d'utiliser 100 séquences proches des gènes, 100 qui en sont éloignées et d'obligatoirement disposer des 5 séquences de copies les plus longues, constituant un jeu de 200 à 205 séquences. Ce choix a été fait car permettant une bonne représentativité du jeu de données et car il permet à l'alignement d'aboutir et d'être réalisé dans un délai compatible avec notre étude. Chacun des alignements obtenus a été visualisé manuellement sous JalView. Cependant, lors de cette visualisation, aucune différence au niveau séquence, n'a pu être observée entre les copies proches des gènes et celles qui en sont éloignées. D'après ce résultats, il semble donc que ces séquences ne possèdent pas de motifs spécifiques en fonction de leur distance au gène, bien que des

délétions soient fréquemment observées dans les séquences de ces ET (Figure 21). Ce résultat pourrait alors suggérer que les portions inutiles, ou délétères des répétitions sont éliminées lorsque celles-ci se trouve à proximité d'un gène sans pour autant que les copies qui en sont éloignées ne soient modifiées si leur effet sur le génome est neutre.

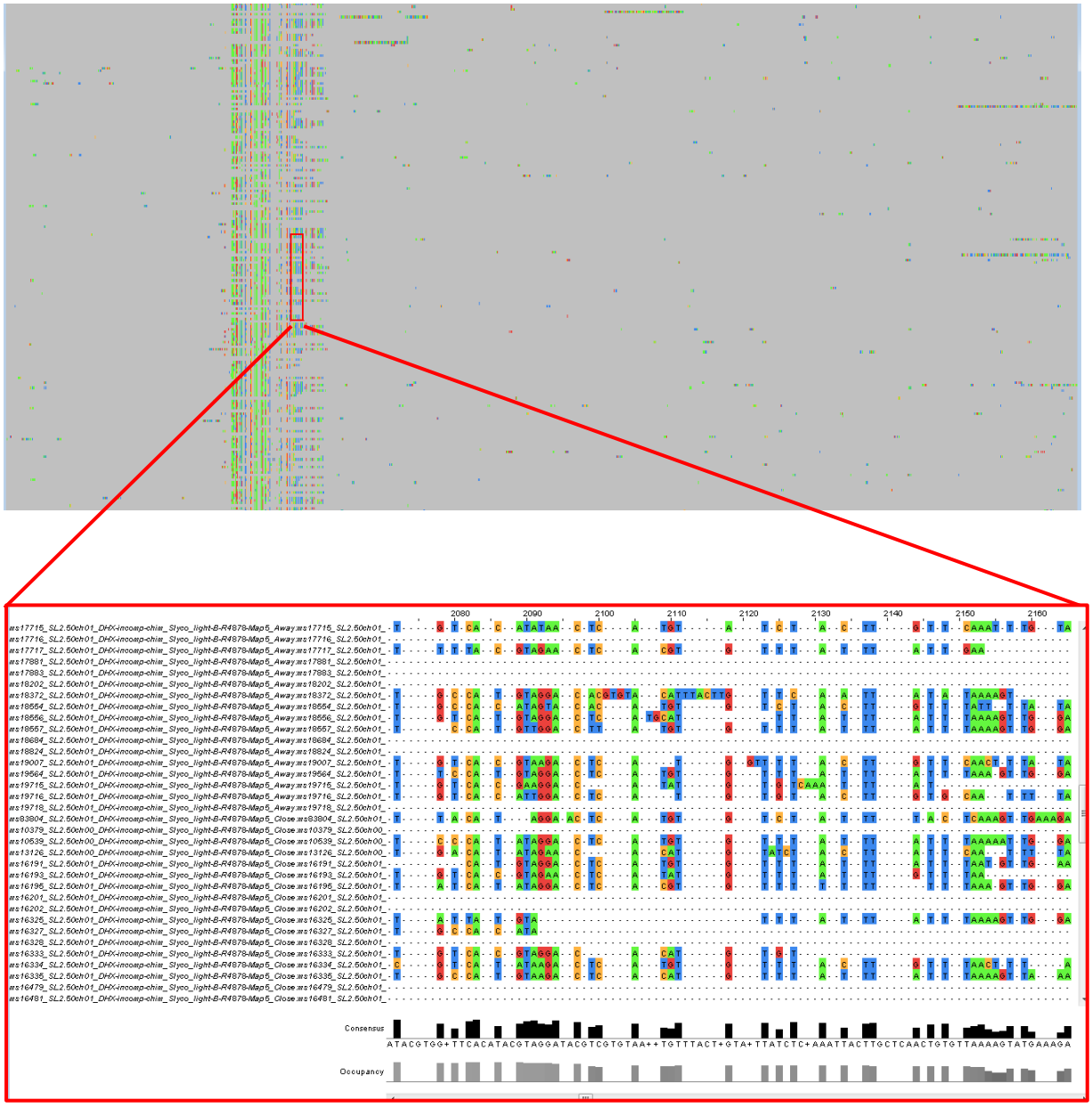


Figure 21 : Profil de séquences de la famille d'éléments issus du consensus DHX-incomp-chim_Slyco_light-B-R4878-Map5 aligné par refalign et visualisé sous JalView.

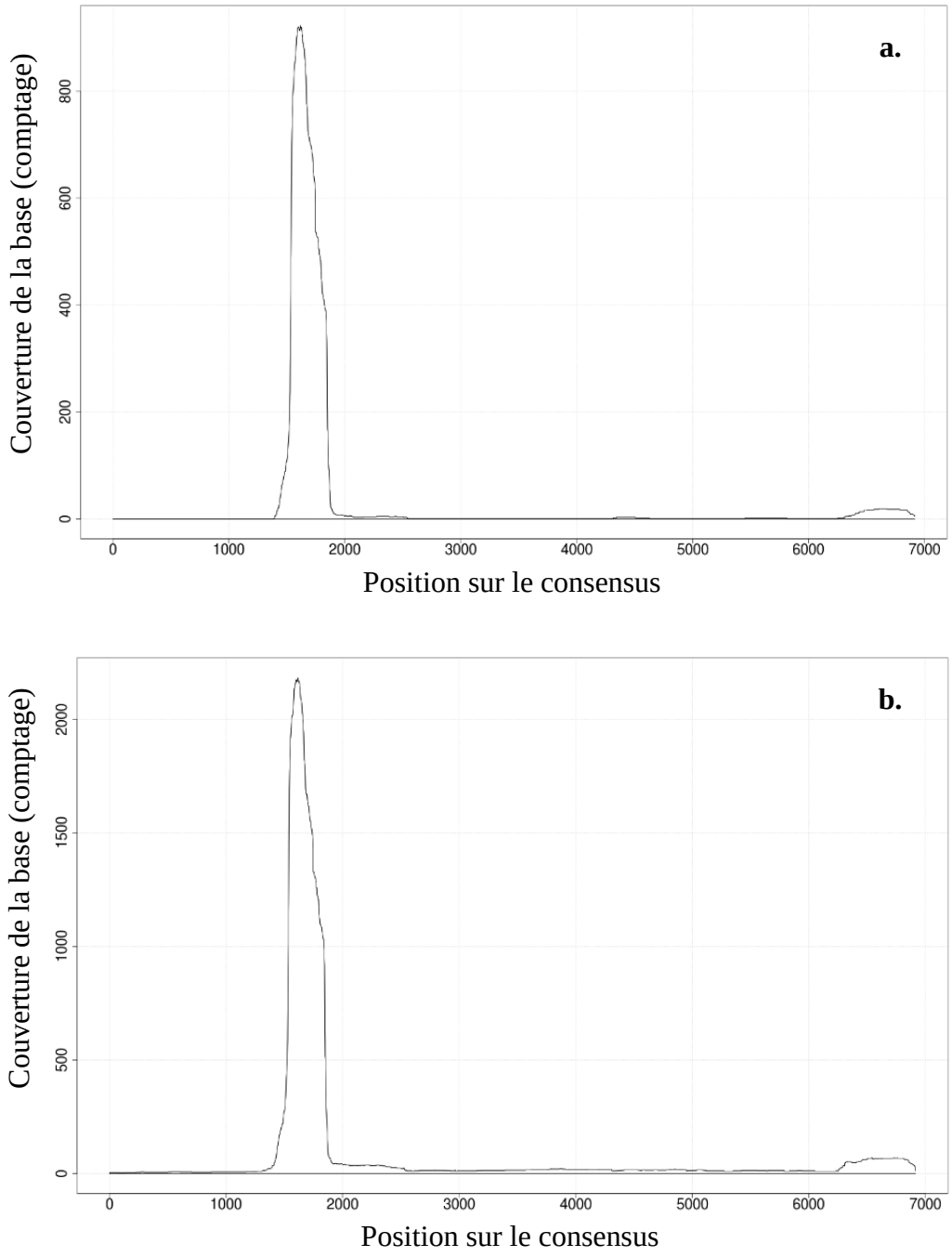


Figure 22 : Graphique issu du PloCoverage des copies du consensus DHX-incomp-chim_Slyco_light-B-R4878-Map5. a. Visualisation graphique de la couverture en bases du consensus par les copies proches des gènes. b. Visualisation graphique de la couverture en bases du consensus par les copies éloignées des gènes.

Cependant, l'alignement réalisé de cette façon ne prend en compte qu'une partie des copies, une analyse complémentaire, utilisant l'intégralité des copies, quel que soit leur nombre, a alors été effectuée en réalisant un graphique appelé PlotCoverage, indiquant la couverture, en nombre de bases, de chaque position d'une séquence, pour analyser la couverture en bases du

consensus de chaque famille. Cette nouvelle approche n'a cependant de nouveau pas permis d'observer de motifs spécifiques aux copies proches de gènes qui pourraient avoir été sélectionnés au cours de l'évolution (Figure 22). La question s'est alors posée de comprendre si ce résultats était réel ou s'il était possible qu'il soit erroné, en raison des données, ou des méthodes utilisées. En analysant les séquences étudiées lors de ces alignements, il a déjà été possible de constater qu'il pouvait s'agir de séquences dont le consensus qui a permis leur détection serait mal construit. En effet, les éléments porteurs des mentions « incomp » ou « noCat » sont souvent difficiles à classifier correctement en raison du manque d'information présent dans leur séquence, de même que pour les éléments classés « RXX » ou « DXX ». Si tel est le cas, l'absence de motifs spécifiques pourrait découler de cette mauvaise construction. Mais il est également possible que les outils utilisés pour l'alignement, bien qu'à l'origine de la construction du consensus, ne parviennent pas à reproduire correctement le premier résultat obtenu, révélant de nouveau une erreur potentielle dans la construction du consensus. Enfin il est tout à fait possible qu'aucun motif particulier ne soit présent dans les copies de ces familles d'ET, une autre raison étant alors à l'origine de leur positionnement particulier dans le génome.

Pour finir, les résultats obtenus par ces analyses d'alignements et de PlotCoverage sont difficiles à interpréter et peuvent soit avoir un sens réel, soit découler de différentes erreurs liées aux données, à la méthode ou à un ou plusieurs des outils utilisés.

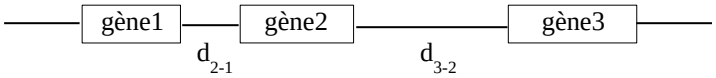
3.3 Matériel et méthodes

a.

BLOC 1 : Identifier les éléments transposables dans la région upstream des gènes

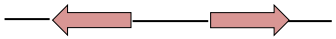
Formatage gènes :

Distance intergéni



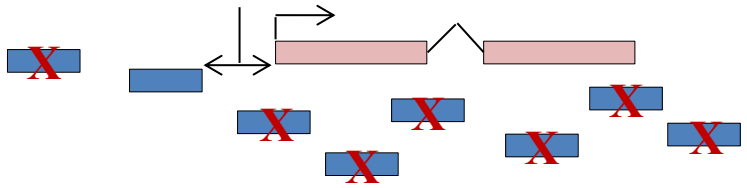
$d_{étude}$ = distance intergénique médiane / 2

Sélection des gènes :



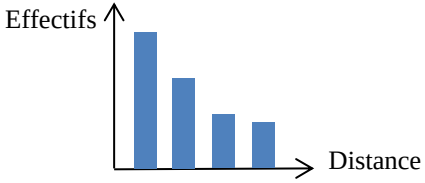
$d < d_{étude} \times 2$

Recherche et sélection des associations entre gènes et éléments transposables :



BLOC 2 : Regarder la répartition des effectifs en upstream des gènes pour chaque famille d'élément transposable et comparer la répartition observée à une répartition attendue sous l'hypothèse d'une distribution aléatoire. ***b.***

Effectifs en upstream des gènes :



Effectifs théoriques en upstream des gènes :

Test statistique :
Chi2 d'homogénéité (obs vs théo) avec correction FDR ou Bonferroni
H0 : la répartition observée correspond à la répartition

Regarder s'il y a présence de motifs inversés dans les séquences

Regarder l'orientation des ET par rapport aux gènes

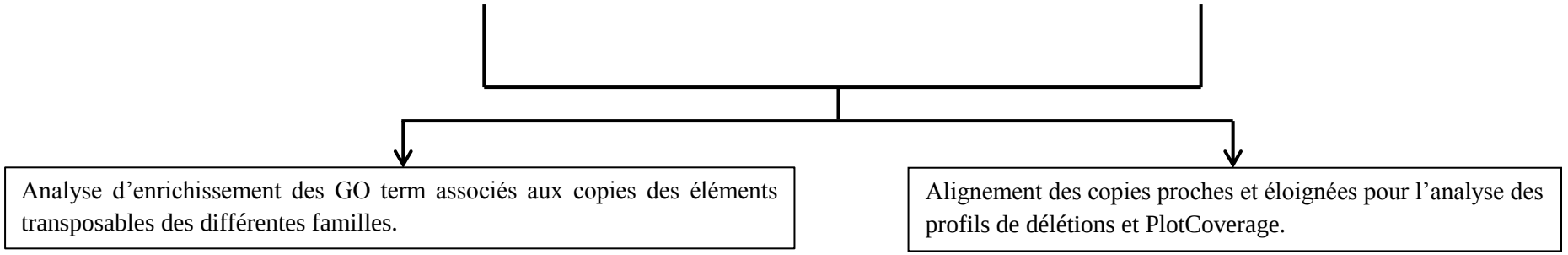


Figure 23 : Schéma de la méthodologie d'analyse *in silico* mise en place pour la détection de séquences d'éléments transposables potentiellement sélectionnés au cours de l'évolution. a. Le bloc 1 regroupe les étapes de formatage des données et de formations des tuples entre gènes et éléments transposables. b. Le bloc 2 regroupe l'ensemble des étapes permettant l'analyse de la répartition des éléments transposables en *upstream* des gènes.

3.3.1 Annotation de l'assemblage SL2.50 du génome de S. lycopersicum

Le pipeline TEdenovo du package REPET v2.5 a été utilisé, avec des paramètres par défaut, mais un minimum de séquences par groupe monté de 3 à 5, sur les contigs d'une taille supérieure à 80kb de l'assemblage SL2.50 du génome de Solanum lycopersicum (soit environ 390Mb, gaps exclus) afin de générer une librairie de 818 séquences consensus. L'intégralité de l'assemblage SL2.50 a alors été annoté en utilisant le pipeline TEannot du package REPET v2.5, avec les paramètres par défaut, en utilisant la librairie de consensus issue de TEdenovo (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011) (Quesneville, H. *et el.*, 2005) (Hoede, C. *et al.*, 2014). Un total de 636 643 séquences d'éléments transposables ont alors été annoté. Les consensus de cette librairie ont été classifiés en utilisant l'outil REPET dédié appelé PASTEC. Les consensus classifiés en tant que potential host gene ont été exclus de cette étude puisqu'ils peuvent contenir des domaines pfam des gènes de l'hôte.

3.3.2 Paires gènes / éléments transposables

Afin de définir les paires de gènes et éléments transposables, les différents fichiers à disposition ont été filtrés et formatés (Figure 23a). Tout d'abord, les TE-gènes, c'est-à-dire les gènes dont au moins 50% des CDS sont recouvertes par un ET, ont été éliminés de la liste, permettant ainsi de filtrer 2 247 gènes. L'ensemble des coordonnées des gènes restant de S. lycopersicum a ensuite été formaté afin de définir chaque gène comme limité à sa séquence codante, éliminant alors les régions UTR. Pour cela, les coordonnées de début de la première CDS, et de fin de la dernière CDS de chaque gène ont été récupérées à l'aide d'un script dédié (Annexe 1), développé en langage python, et servent désormais de coordonnées de début et de fin de gène.

Une fois le fichier de gènes formaté, la distance intergénique médiane et moyenne sur l'ensemble des gènes et au sein de chaque région génomique (RR, INT, RP) a été calculée dans le but, notamment, de définir la distance limite entre un gène et un élément transposable pour en faire un tuple étudié. Pour cela, après une étape pour ordonner le fichier de coordonnées des gènes, par chromosome et par position sur le chromosome, un script spécifique (Annexe 2), développé en python, a été utilisé, permettant pour chaque fichier donné en entrée du script de fournir le résultat du calcul d'une distance intergénique médiane ou moyenne. Une fois cette distance d'étude définie, tous les gènes en sens opposés

partageant une région upstream commune et étant située à une distance inférieure à celle établie précédemment sont éliminés de notre liste (Figure 23a). Cette sélection a pour but de définir les tuples dont les interactions potentielles, entre gène et élément transposable, sont le plus clair possible, afin que celles-ci puissent éventuellement être testées biologiquement.

L'ensemble des conditions d'utilisation des gènes étant désormais respectées, les paires gène / élément transposable vont pouvoir être formés. On va alors chercher à associer à chaque élément transposable le gène qui lui est le plus proche, de manière à ce que l'ET se trouve dans la région upstream du gène, sans le chevaucher, à une distance maximale égale à la moitié de la distance intergénique médiane (Figure 23a). Un gène peut alors être couplé à plusieurs éléments transposables. Pour trouver ces tuples, l'outil Bedtools v2.17.0 a été utilisé avec la commande bedtools closest -D b -id -d -a Annotation_ET.gff3 -b Annotation_Genes.gff3 > Associations_TEs_Genes.bed a été lancée pour détecter toutes les paires possibles avant d'être triées sous R.

3.3.3 Calcul de la répartition des effectifs en upstream des gènes et test de conformité

Une analyse de la répartition des différentes familles d'éléments transposables en upstream des gènes est ensuite réalisée. La distribution des copies de chaque famille d'élément transposable en upstream des gènes est alors calculée par fenêtres de distances de 625 bases, et ce jusqu'à une distance maximale de 2 500 bases. Un comptage des copies dans chaque plage de distance est ensuite réalisé à l'aide d'un script python dédié (Annexe 3) en se basant sur les distances entre gène et répétition calculées par Bedtools closest. Ce groupe de valeurs définira pour chaque famille d'éléments, les effectifs réels de la distribution.

Des effectifs théoriques de répartition ont ensuite été calculés afin de comparer les effectifs réels à une distribution attendue sous l'hypothèse d'une répartition aléatoire. Pour cela, l'effectif total par famille est simplement divisé par quatre afin que les valeurs soient les même dans les différentes classes de distances définies.

Les effectifs théoriques et réels sont finalement comparés à l'aide d'un test de conformité de type Chi2, sous le logiciel R v3.0.2 (en local) ou v3.2.2 (sur le cluster de caclul), pour mesurer si les effectifs observés s'écartent de la distribution attendue : si les copies de la

famille étudiée sont distribuées dans le génome de manière aléatoire (Figure 23b). Pour cette étude de leur répartition, seules les familles ayant un effectif total d'au moins 20 copies ont été conservées. En effet, le test statistique réalisé nécessite que chaque classe théorique contienne un effectif minimal de 5 copies.

3.3.4 Analyse de l'orientation des éléments transposables par rapport aux gènes

L'orientation en sens ou en antisens des éléments transposables par rapport aux gènes a alors été étudiée (Figure 23). On a donc réalisé un comptage des effectifs réels, pour chaque famille de répétition dont l'effectif est supérieur à 10 afin de pouvoir réaliser des analyses statistiques de type Chi2, qui nécessitent que chaque classe théorique ai un effectif minimum de 5, des copies en sens et en antisens. Les effectifs théoriques sous l'hypothèse d'une absence d'orientation préférentielle, ont ensuite été calculés en divisant l'effectif total de chaque famille par deux. Un test de conformité de type Chi2 , sous le logiciel R v3.0.2 (en local) ou v3.2.2 (sur le cluster de calcul), a ensuite été réalisé entre les effectifs réels et théoriques pour définir quelles familles s'éloignent d'un modèle aléatoire et semblent donc avoir une orientation préférentielle par rapport aux gènes.

3.3.5 Analyse d'enrichissement des GO term

Une analyse d'enrichissement des GO term des gènes associés à chaque famille d'éléments transposables est réalisée. Cette étude a pour objectif de tester si un ou plusieurs groupes de répétitions sont associés à certaines fonctions géniques de façon préférentielle. Pour réaliser cette analyse, les listes de gènes associés aux copies des différentes familles sont soumises une à une sur le site geneontology.org (Ashburner, M. *et al.*, 2000) (The Gene Ontology Consortium, 2017) (Mi, H., *et al*., 2017) pour réaliser une analyse d'enrichissement (Enrichment analysis) des processus biologiques en prenant comme liste de gènes de référence celle de Solanum lycopersicum. Une analyse globale des gènes ayant un ET à moins de la moitié de la distance intergénique dans leur région upstream est également réalisée pour comparer les fonctions associées à ces gènes, à la répartition des fonctions chez l'ensemble des gènes de la tomate. Ces analyses d'enrichissement de GO terms n'ont pas été effectuées localement sur nos machines car aucun outil dédié, applicable chez la tomate, n'a été identifié. L'utilisation des données de GO terms pour créer notre propre outil d'analyse n'a également pas été une option retenue car la complexité et le volume de ces données auraient nécessité un

temps de développement important. Le choix a donc été fait d'utiliser l'outil en ligne d'analyse d'enrichissement développé par Panther, sur le site du Gene Ontology Consortitum, geneontology.org, qui exploite déjà les données de façon optimisée.

3.3.6 Comparaison des séquences de copies proches et éloignées des gènes

L'étude des séquences composant les familles d'intérêt a été réalisée en alignant les séquences des copies d'une même famille à la séquence de son consensus à l'aide de l'outil refalign (package Repet) et capable d'aligner plusieurs séquences à une même référence et de gérer les gaps importants qui peuvent ainsi exister. Cette analyse doit permettre de comparer les séquences des copies proches de gènes à celles des copies qui en sont éloignées. Une sélection de 100 copies de chaque catégorie, ainsi que des 5 copies les plus longues a été mise en place en amont de l'alignement, élevant le nombre de copies à aligner de 200 à 205 au total. Cet alignement est lancé grâce à la commande for dossier in Light/*; do refalign $dossier/Consensus.fa $dossier/*_Select.fa; done qui lance de manière automatique l'outil d'alignement sur chaque fichier correspondant chacun à une famille d'ET d'intérêt et la commande for dossier in Light/*; do refalign2fasta.py -i $dossier/*.aligner; done qui permet d'obtenir le fichier d'alignement multiple. Cependant, pour certains fichiers, l'alignement ne se déroule pas correctement et une étape permettant d'obtenir la séquence complémentaire de chaque copie est nécessaire à l'aide de la commande revcomp css_Select.fa > css_Select_revcomp.fa avant de lancer refalign. Les alignements obtenus sont ensuite visualisés sous JalView (Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J., 2009).

Les alignements ainsi réalisés n'utilisent donc pas systématiquement l'intégralité des copies de la famille de répétitions étudiées. Un PlotCoverage, disponible parmi les outils fournis par REPET, avec l'ensemble de séquences a alors été effectué en complément pour permettre une vision globale. Cette étape est réalisée grâce à la commande plotCoverage.py -i .gff3 -f gff3 -q .fa –o lancée sur chaque groupe de copies de façon indépendante.

3.4 Discussion et conclusion

Le génome de la tomate est un génome riche en éléments transposables dont l'étude a déjà montré l'impact tant sur la structure du génome, que sur l'expression des gènes ou encore sur

le processus de maturation de la tomate. Sachant cela, la question s'est posée de savoir si certaines copies avaient été sélectionnées au cours de l'évolution pour leur impact sur les gènes et s'il était possible de les détecter par des analyses in silico. Pour tenter de répondre à cette question, un protocole d'analyse a été développé, en ayant pour objectif d'étudier le positionnement des différentes familles de répétitions par rapport aux gènes, leur orientation, mais également les fonctions de gènes proches desquels ils sont présents, a été développé.

Les analyses ainsi réalisées ont alors montré qu'un certain nombre de familles d'ET se localisaient de façon plus importantes qu'attendu à proximité des gènes, mais également qu'ils pouvaient présenter des profils de distribution particuliers en amont de ces derniers. Il existe deux hypothèses pouvant expliquer cette localisation des répétitions à une distance précise. La première suppose que les ET s'insèrent dans cette région upstream de manière quasi exclusive car elle se trouve être moins condensée que le reste du génome afin de permettre l'expression des gènes. La seconde, considère que les ET s'insèrent aléatoirement dans l'ensemble de ce génome mais qu'ils ne sont soumis à une pression de sélection que s'ils ont un impact positif pour les gènes situés à leur proximité. Les travaux réalisés par la suite n'ont pas permis de trancher entre ces deux hypothèses mais il est possible que les deux mécanismes coexistent au sein de S. lycopersicum. Il a en revanche été possible de révéler que, outre des profils de distribution particuliers, certaines familles de répétitions ont également une orientation spécifique par rapport aux gènes. Ce biais d'orientation n'est cependant pas forcément révélateur d'une sélection au cours de l'évolution car, certains ET porteurs de séquences répétées inversées, ne seront plus sélectionnés pour leur orientation dans le génome mais car ils présentent ce motif dans les deux orientations. Ces éléments particuliers mis à part, 36 familles se sont révélées pouvoir être d'intérêt puisque présentant des caractéristiques définies pouvant suggérer une sélection de séquence au cours de l'évolution.

Pour ces 36 familles de répétitions que l'on suppose avoir été soumises à des mécanismes de sélection, une étude des fonctions de gènes associées à chacun des groupes a été réalisée. Celle-ci a montré qu'une famille d'éléments donnée pouvait être associée à une fonction particulière. On a par exemple pu identifier que la famille de répétitions provenant du consensus DHX-incomp-chim_Slyco_light-B-R4878-Map5, étant un élément appartenant à la sous-classe 2 de la classe II, était plus fréquemment associée qu'attendu, à des gènes dont les fonctions participent aux mécanismes de défense et de réponse au stress chez la tomate. Malheureusement, chez la tomate, encore peu de réseaux de gènes sont connus. Autrement,

une étude complémentaire, employant la démarche inverse serait à envisager. C'est-à-dire que l'on partira d'une liste de gènes co-régulés impliqués dans un même réseau ou un même processus biologique, une détection des répétitions se trouvant en upstream de ces gènes sera alors menée, et enfin une analyse d'enrichissement permettant d'identifier si certaines familles sont plus présentes qu'attendues dans cette région sera réalisée. Peut-être sera-t-il alors possible de retrouver des familles déjà identifiées dans l'analyse dite Bottom-up et d'en découvrir de nouvelles grâce à celle-ci, appelée Top-down.

Pour terminer ce travail, une étude des profils de séquences par famille d'ET d'intérêt a été menée. Celle-ci avait pour but de permettre de comparer les séquences des copies d'ET proches des gènes à celles qui en sont éloignées. Un simple alignement grâce à l'outil refalign, ne prenant pas l'intégralité des copies en entrée mais un jeu de maximum 205 copies, n'a cependant pas permis de mettre en évidence une différence de conservation de séquence notable entre ces deux catégories. L'analyse complémentaire réalisant un PlotCoverage pour l'ensemble des séquences de chaque famille n'a pas non plus fourni de résultat significatif. Cependant, compte tenu de la nature des ET étudiés ici, majoritairement les éléments portant la mention « incomp », « chim » et « noCat », il est possible que la construction des consensus et donc de ces familles ait été erronées, ou bien, que les outils utilisés ne soient pas adaptés pour ce travail, et ne permettent donc pas de retrouver les profils que l'on pourrait attendre, à savoir un fragment de séquence spécifique à l'une des deux catégories.

L'ensemble des résultats obtenus tout au long de cette étude suggèrent donc qu'il est possible de détecter, par des approches strictement in silico, des familles de répétitions avec des caractéristiques particulières, indiquant que les éléments ciblés pourraient avoir été sélectionnés au cours de l'évolution pour leur impact sur les gènes. Néanmoins, l'analyse des séquences de telles familles n'a pas révélé de motif particulier qui pourrait permettre un effet sur les gènes, seules des délétions ayant pu être observées lors de la visualisation. Pour compléter ce travail, une recherche de séquences promotrices connues dans les familles d'ET d'intérêt serait également à envisager pour tenter de valider la possibilité d'une sélection des copies pour leur impact sur les gènes situés à proximité mais que l'on aurait pu identifier lors des alignements. Par la suite, si certaines copies parvenaient à être spécifiquement identifiées comme des cibles de la sélection, des tests in vitro pourraient être menés afin de valider ou invalider l'impact de ces séquences sur l'expression des gènes, notamment en les inactivant.

Conclusion et discussion

Le travail réalisé dans cette thèse avait pour objectif d'étudier le rôle des éléments transposables dans le génome de la tomate *S. lycopersicum*. Pour réaliser cette étude, deux objectifs intermédiaires ont été définis. Le premier était d'analyser l'impact potentiel que pouvaient avoir les ET sur la régulation de l'expression des gènes, et ce notamment dans le cadre du processus de maturation du fruit. Le second était alors d'évaluer la possibilité que certaines copies de répétitions soient sélectionnées au cours de l'évolution pour leur impact sur les gènes situés à proximité et de déterminer s'il était possible de détecter de telles copies par des approches *in silico*.

L'étape de ré-annotation du génome de *S. lycopersicum* a révélé que ce génome est riche en éléments transposables, et plus particulièrement en éléments de type Gypsy et Copia (éléments de type I). Une analyse approfondie de cette ré-annotation a alors permis de définir trois compartiments au sein de ce génome : un compartiment riche en ET et plutôt pauvre en gènes appelé *Repeat Rich* (RR), un compartiment riche en gène et appauvrit en ET appelé *Repeat Poor* (RP) et un compartiment au contenu en gènes et en ET intermédiaire appelé *Intermediate* (INT). L'étude de ces trois compartiments a montré qu'ils présentaient non seulement une densité en gènes et en répétitions particulière, mais aussi que chacune des trois régions avait un contenu spécifique en ET, c'est-à-dire que certaines familles étaient plus abondantes dans une région ou une autre, et enfin que les gènes localisés dans ces différentes régions avaient des origines évolutives différentes. En effet, il a été montré que la région RP présentait essentiellement des gènes « anciens », ce qui signifie qu'ils sont communs à d'autres espèces proches de la tomate (la pomme de terre et la plante *Mimilus gutatus*), alors que la région RR contient principalement des gènes apparus récemment chez la tomate. Cette différence pourrait être expliquée par l'abondance des ET et leur activité dans la région RR, qui pourraient participer à la création de nouveaux gènes. La compartimentation observée au sein du génome de la tomate correspond globalement à ce que l'on pourrait attendre dans la plupart des génomes de plantes ayant des chromosomes monocentriques : la majorité des régions riches en répétitions se trouvent dans les régions péricentromériques ou à l'extrémité des chromosomes dans les régions télomériques. Mais outre cette compartimentation du génome et les propriétés propres à chaque compartiment qui ont été définies, une analyse du lien entre les répétitions, la méthylation de l'ADN et l'expression des gènes a été menée. Cette analyse a montré qu'un grand nombre des régions différentiellement méthylées (DMRs) au cours de la maturation du fruit étaient co-localisées avec des répétitions dans ce génome.

L'association entre éléments répétés et méthylation de l'ADN n'est cependant pas surprenant compte tenu du fait que la méthylation est le principal moyen de contrôle de ces ET dans les génomes. Ce qui l'est d'avantage est l'aspect dynamique de ces méthylations des ET et le lien qu'ils pourraient avoir avec le processus de maturation du fruit. Il a par la suite été montré que de nombreux gènes avaient dans leur région *upstream*, un ET qui est fréquemment porteur d'une DMR. Parmi ces gènes, un certain nombre se trouve être différentiellement exprimé au cours de la maturation du fruit.

D'après les résultats obtenus dans cette première partie, et bien qu'un impact direct des ET sur les gènes n'ai pas été trouvé, cela suggère que les ET participent à la régulation de l'expression des gènes au cours de la maturation du fruit chez *S. lycopersicum* en étant porteurs d'un grand nombre des DMRs détectées dans le cadre de ce processus. Les résultats ainsi obtenus ont déjà été en grande partie observée chez *Arabibopsis thaliana* qui présente également une compartimentation de son génome en fonction de la densité en répétitions, celles-ci étant également porteuses de nombreuses méthylations. Il serait alors intéressant d'affiner nos résultats chez la tomate afin de détecter des copies candidates spécifiques pour étudier leur impact sur les gènes situés à proximité. Ce travail pourrait alors être approfondi par des approches expérimentales in vivo.

Pour tenter de répondre à cette attente, la seconde partie de cette thèse avait donc pour objectif d'identifier des copies d'éléments répétés pouvant avoir été sélectionnées au cours de l'évolution pour leur impact sur les gènes situés à proximité. Cet objectif découlait d'*a priori* : les copies d'éléments répétés présentant un avantage sélectif vont être conservées au cours du temps, les régions régulatrices des gènes les plus aisées à étudier se trouvent en *upstream* de leurs séquences à une courte distance et enfin certains gènes sont organisés en réseaux co-régulés qui pourraient présenter dans leur région *upstream* une même répétition qui pourrait servir de séquence de régulation.

Une méthodologie *in silico* a donc été mise en place, regroupant un ensemble d'analyses, afin d'identifier des éléments d'intérêt pouvant avoir été sélectionnés au cours de l'évolution. Cette approche regroupe plusieurs grandes analyses : l'étude de la distribution des copies en *upstream* des gènes, l'étude de l'orientation en sens ou en antisens de ces copies par rapport aux gènes, la recherche de motifs répétés dans les séquences des répétitions, une étude de l'enrichissement de certaines fonctions de gènes associées aux ET, et enfin l'étude des profils de délétion en comparant l'intégrité des copies proches de gènes à celles qui en sont éloignées.

Grâce à cette méthodologie, il a été possible de montrer que certaines familles d'éléments répétés se trouvaient dans des configurations particulières par rapport aux gènes. D'une part, certains éléments sont plus fréquemment localisés à proximité des gènes dans le génome que l'on pourrait s'y attendre s'ils suivaient une distribution aléatoire, mais aussi dans la région *upstream* des gènes eux-mêmes, on trouve des profils de distribution particuliers. A la place d'une distribution aléatoire, impliquant un même nombre de répétitions quelle que soit la distance aux gènes, on trouve un nombre important de copies d'une famille donnée d'éléments répétés à une distance spécifique des gènes. D'autre part, certains éléments ont une orientation spécifique, en sens ou en antisens, par rapport aux gènes dont ils proches. Cependant, ce résultat concernant l'orientation peut ne pas être concluant car certains éléments sont porteurs de motifs inversés dans leur séquence, comme l'a montré l'analyse par le logiciel *einverted*, impliquant que leur effet pourrait avoir lieu quelle que soit leur orientation. Ces premières étapes ont permis une sélection de 36 consensus : 30 ayant un profil de distribution particulier, 5 présentant une orientation particulière de leurs copies par rapport aux gènes, et 1 regroupant ces deux caractéristiques. Les gènes associés à ces 36 familles de répétitions ont alors été analysés afin de détecter un enrichissement en certaines fonctions, ce qui a été le cas pour 8 de ces familles. On trouve parmi ces fonctions, celles de la défense de l'organisme et de la réponse au stress, ou encore des fonctions dans le recrutement du protéasome par l'ubiquitination. Cet enrichissement de certaines fonctions associées à des éléments répétés dans des configurations particulières suggère alors qu'ils pourraient avoir un impact sur ces gènes, impliqués dans des processus particuliers. Il a alors été supposé que, si ces répétitions avaient un impact sur les gènes situés à proximité, ils devraient présenter, dans leurs séquences, des particularités permettant, par exemple, la régulation de l'expression de ces gènes. Une étude des séquences des copies proches des répétitions par rapport à celles qui en sont éloignées a alors été menée afin d'en étudier les profils de délétions. Malheureusement, aucune des deux approches utilisées pour réaliser cette recherche n'a permis d'identifier de profil particulier à l'une des deux catégories de copies, bien que les copies proches des gènes semblent, dans l'ensemble, subir d'avantage de délétions dans leur séquence que les copies éloignées, dont on peut supposer un impact neutre dans le génome.

Les résultats obtenus dans cette seconde partie ne prouvent pas la sélection de certaines copies mais suggèrent que certaines familles de répétitions pourraient avoir un impact sur les gènes et avoir été sélectionnées au cours de l'évolution compte tenu de leurs propriétés particulières à proximité des gènes. D'autres analyses sont alors à envisager pour compléter cette étude. La

première serait de rechercher, dans les séquences des copies des familles d'intérêt, et plus particulièrement dans les copies proches de gènes, des motifs de régulation de la transcription connus. Trouver, au sein de ces séquences de tels motifs, tendrait à prouver leur impact sur les gènes et leur sélection au cours de l'évolution. Par ailleurs, une approche dite *top-down* peut être envisagée pour atteindre le même objectif. Cette approche consiste à partir d'une liste de gènes connus pour être impliqués dans un même réseau ou un même processus afin de chercher, dans leur région *upstream*, si certaines familles de répétitions s'y retrouvent plus fréquemment que ce que l'on pourrait attendre sous l'hypothèse d'une répartition aléatoire. Malheureusement, chez la tomate, peu de réseaux de gènes co-régulés sont encore connus en dehors de celui de la maturation du fruit, et cette approche a donc dû être mise de côté dans cette étude. Finalement, il serait aussi envisageable de passer par une approche plus large, recherchant des *k-mers* fréquents à proximité des gènes, un *k-mer* étant une courte séquence de quelques nucléotides (souvent 10 à 15), afin de faire potentiellement un lien avec des séquences de régulations connues et de voir si ces *k-mers* sont portés par nos éléments répétés. Toutes ces approches peuvent être vues comme complémentaires afin d'identifier un panel de copies potentiellement importantes pour le génome, qu'il faudrait ensuite valider par des analyse *in vivo* les inactivant ou les supprimant.

L'ensemble de ce travail de thèse a finalement permis de révéler chez la tomate, le rôle apparemment essentiel des éléments répétés pour la régulation de l'expression des gènes, notamment au cours de la maturation du fruit en étant porteurs de nombreuses DMRs à proximité de gènes dédiés à ce processus, mais également la possibilité que certaines copies, retrouvées dans des configurations très particulières, aient été sélectionnées au cours de l'évolution. Ces travaux novateurs, uniquement basés sur des approches *in silico*, semblent permettre d'identifier des copies candidates pouvant avoir été sélectionnées au cours de l'évolution. La plus grande difficulté dans ce type d'analyses est alors de définir des critères de sélection suffisamment précis et basés sur des connaissances, pour sélectionner un petit jeu de copies candidates qui nécessitent par la suite une validation *in vitro*. Pour finir, il a été envisagé au cours de cette thèse, que si certaines copies candidates parvenaient à être identifiées, il serait intéressant de comparer notre génome de référence à celui de plusieurs variétés de tomates obtenus par re-séquençages, mais également à des espèces proches afin d'identifier si l'insertion candidate est nouvelle chez la référence Heinz, commune à différents génomes de la tomate, ou si elle plus ancienne et commune à d'autres espèces de Solanacées. Cette analyse consisterait alors en l'étude des variants de présence / absence de l'insertion dans différents génomes de manière intra ou inter-espèces. Différents outils ont déjà été

évalués dans ce sens en réalisant chez la variété M82 de la tomate, une recherche de nouvelles insertions d'ET suite à leur réactivation dans deux mutants

Bibliographie

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

Altschu, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Molecular Biology* **215**, 403-410 (1990).

Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

Bao, Z. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research* **12**, 1269–1276 (2002).

Bennetzen, J. L. & Wang, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology* **65**, 505–530 (2014).

Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).

Biłas, R., Szafran, K., Hnatuszko-Konka, K. & Kononowicz, A. K. Cis-regulatory elements used to control gene expression in plants. *Plant Cell, Tissue and Organ Culture (PCTOC)* **127**, 269–287 (2016).

Brzeski, J. & Jerzmanowski, A. Deficient in DNA Methylation 1 (DDM1) Defines a Novel Family of Chromatin-remodeling Factors. *Journal of Biological Chemistry* **278**, 823–828 (2003).

Butelli, E. *et al.* Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. *The Plant Cell* **24**, 1242–1255 (2012).

Campbell, N. A. & Reece, J. B. Biologie. 2E éd. *De Boeck*, 391-398 (2004).

Causse, M. *et al*. The tomato genome, *Springer-Verlag Berlin Heidelberg*, DOI 10.1007/978-3-662-53389-5_11 (2016).

Cerbin, S. & Jiang, N. Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development* **49**, 63–69 (2018).

Chen, D., Yan, W., Fu, L.-Y. & Kaufmann, K. Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nature Communications* **9**, (2018).

Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).

Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**, 71–86 (2017).

Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. Regulation of Transcription in Eukaryotes. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9904/ ; Bookshelf ID: NBK9904.

Dalakouras, A. & Wassenegger, M. Revisiting RNA-directed DNA methylation. *RNA Biology* **10**, 453–455 (2013).

Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N. & Tanksley, S. D. A Comparative Genetic Linkage Map of Eggplant (Solanum melongena) and Its Implications for Genome Evolution in the Solanaceae. *Genetics* **161**, 1697-1711 (2002).

Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N. & Tanksley, S. D. Conservation of Gene Function in the Solanaceae as Revealed by Comparative Mapping o.f Domestication Traits in Eggplant. Genetics 161, 1718-1726 (2002)

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H. & Spillane, C. Evolutionary origins of

Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology* **11**, (2011).

Doolittle, R. F., Johnson, M. S. & McClure, M. A. Origins and Evolutionary Relationships of Retroviruses. *The Quarterly Review of Biology* **64**, 1–30 (1989).

Dooner, H. K., Robbins, T. P. & Jorgensen, R. A. Genetic and Developmental Control of Anthocyanin Biosynthesis. *Annu. Rev. Genet.* **25**, 173-199 (1991).

Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–218 (2007).

Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).

El Baidouri, M. *et al.* A new approach for annotation of transposable elements using small RNA mapping. *Nucleic Acids Research* **43**, e84–e84 (2015).

Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* **41**, 563–571 (2009).

Fedoroff, N. V. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**, 758–767 (2012).

Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**, 397–405 (2008).

Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* **3**, 329–341 (2002).

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

Feschotte, C., Zhang, X. & Wessler, S. R. Miniature Inverted-repeat Transposable Elements (MITEs) and their Relationship with Established DNA Transposons. in Craig, N. L. (ed) Mobile DNA II, Chapter 50, pp. 1147–58 (2002).

Fiston-Lavier, A.-S., Carrigan, M., Petrov, D. A. & González, J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Research* **39**, e36–e36 (2011).

Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* **6**, e16526 (2011).

Freeling, M. *et al.* Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Research* **18**, 1924–1937 (2008).

Gao, D., Li, Y., Kim, K. D., Abernathy, B. & Jackson, S. A. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biology* **17**, (2016).

Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Research* **27**, 1916–1929 (2017).

Garrido-Ramos, M. Satellite DNA: An Evolving Topic. *Genes* **8**, 230 (2017).

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M. & Petrov, D. A. High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biology* **6**, e251 (2008).

Greene, B., Wako', R. & Hake, S. Mutator Insertions in an Intron of the Maize knottedl Gene Result in Dominant Suppressible Mutations. *Genetics*, 138, 1275-1285 (1994).

Gu, W., Castoe, T. A., Hedges, D. J., Batzer, M. A. & Pollock, D. D. Identification of repeat structure in large genomes using repeat probability clouds. *Analytical Biochemistry* **380**, 77–83 (2008).

Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature Reviews Molecular Cell Biology* **12**, 483–492 (2011).

Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).

Hayward, A. Origin of the retroviruses: when, where, and how? *Current Opinion in Virology* **25**, 23–27 (2017).

He, X.-J., Chen, T. & Zhu, J.-K. Regulation and function of DNA methylation in plants and animals. *Cell Research* **21**, 442–465 (2011).

Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics* **16**, (2015).

Hillis, D. M. & Dixon, M. T. Ribosomal DNA: Molecular Evolution and Phylogenetic Inference. *The Quarterly Review of Biology* **66**, 411–453 (1991).

Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, 157–165 (2017).

Hoede, C. *et al.* PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* **9**, e91929 (2014).

Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research* **19**, 1419–1428 (2009).

Holmes, E. C. The Evolution of Endogenous Viral Elements. *Cell Host & Microbe* **10**, 368–377 (2011).

Horn, P. J. MOLECULAR BIOLOGY: Chromatin Higher Order Folding--Wrapping up Transcription. *Science* 297, 1824–1827 (2002).

Huang, X. On global sequence alignment. Comput Appl Biosci 10, 227-235 (1994).

Inagaki, S. & Kakutani, T. What Triggers Differential DNA Methylation of Genes and TEs:

Contribution of Body Methylation? *Cold Spring Harbor Symposia on Quantitative Biology* **77**, 155–160 (2012).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

Ivashuta, S. *et al.* Genotype-dependent transcriptional activation of novel repetitive elements during cold acclimation of alfalfa (Medicago sativa). *The Plant Journal* **31**, 615–627 (2002).

Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. & Maumus, F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics* **17**, (2016).

Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **16**, 418–420 (2000).

Jurka, J. Repeats in genomic DNA : mining and meaning. *Current Opinion in Structural Biology* 8, 333-337 (1998).

Kapitonov, V. V. & Jurka, J. Rolling-Circle Transposons in Eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8714–8719 (2001).

Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences* **103**, 4540–4545 (2006).

Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete *Saccharomyces cerevisiae* Genome Sequence. *Genome Research* **8**, 464–478 (1998).

Kinoshita, Y. *et al.* Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats: FWA gene silencing in A. thaliana. *The Plant Journal* **49**, 38–45 (2006).

Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-Induced Mutations in Grape Skin Color. *Science, New Series* **304**, 982 (2004).

Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7 (2006).

Kolpakov, R. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**, 3672–3678 (2003).

Kornberg, R. D. Structure of chromatin. *Annual review of biochemistry* **46,** 931–954 (1977).

Krupovic, M. *et al. Ortervirales* : New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *Journal of Virology* **92**, (2018).

Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).

Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11**, 204–220 (2010).

Le, T.-N. *et al.* DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. *Genome Biology* **15**, (2014).

Lerat, E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**, 520–533 (2010).

Li, X. *et al.* Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. *BMC Genomics* **18**, (2017).

Lin, R. *et al.* Transposase-Derived Transcription Factors Regulate Light Signaling in Arabidopsis. *Science* **318**, 1302–1305 (2007).

Lisch, D. How important are transposons for plant evolution? *Nature Reviews Genetics* **14**,

49–61 (2013).

Lyons, D. B. & Zilberman, D. DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. *eLife* **6**, (2017).

Makarevitch, I. *et al.* Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genetics* **11**, e1004915 (2015).

Mariño-Ramírez, L., Tharakaraman, K., Spouge, J. L. & Landsman, D. Promoter Analysis: Gene Regulatory Motif Identification with A-GLAM. in *Bioinformatics for DNA Sequence Analysis* (ed. Posada, D.) **537**, 263–276 (Humana Press, 2009).

Martel, C., Vrebalov, J., Tafelmeyer, P. & Giovannoni, J. J. The Tomato MADS-Box Transcription Factor RIPENING INHIBITOR Interacts with Promoters Involved in Numerous Ripening Processes in a COLORLESS NONRIPENING-Dependent Manner. *PLANT PHYSIOLOGY* **157**, 1568–1579 (2011).

Maumus, F. & Quesneville, H. Deep Investigation of Arabidopsis thaliana Junk DNA Reveals a Continuum between Repetitive Elements and Genomic Dark Matter. *PLoS ONE* **9**, e94101 (2014).

McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**, 344–355 (1950).

McGhee, J. D. & Felsenfeld, G. Nucleosome structure. *Annual review of biochemistry* **49,** 1115–1156 (1980).

McStay, B. Nucleolar organizer regions: genomic 'dark matter' requiring illumination. *Genes & Development* **30**, 1598–1610 (2016).

Mehra, M., Gangwar, I. & Shankar, R. A Deluge of Complex Repeats: The Solanum Genome. *PLOS ONE* **10**, e0133962 (2015).

Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, D183–

D189 (2017).

Michalak, P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**, 243–248 (2008).

Moore, J. W., Loake, G. J. & Spoel, S. H. Transcription Dynamics in Plant Immunity. *The Plant Cell* **23**, 2809–2820 (2011).

Mourier, T. Potential movement of transposable elements through DNA circularization. *Current Genetics* **62**, 697–700 (2016).

Negi, P., Rai, A. N. & Suprasanna, P. Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. *Frontiers in Plant Science* **7**, (2016).

Nemeth, A. Chromatin higher order structure: Opening up chromatin for transcription. *Briefings in Functional Genomics and Proteomics* **2**, 334–343 (2004).

Ng, M. & Yanofsky, M. F. Function and evolution of the plant MADS-box gene family. *Nature Reviews Genetics* **2**, 186–195 (2001).

Niederhuth, C. E. & Schmitz, R. J. Putting DNA methylation in context: from genomes to gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, 149–156 (2017).

Novák, P., Neumann, P. & Macas, J. M. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. 12 (2010).

Osbourn, A. E. & Field, B. Operons. *Cellular and Molecular Life Sciences* 66, 3755–3775 (2009).

Pikaard, C. S. & Mittelsten Scheid, O. Epigenetic Regulation in Plants. *Cold Spring Harbor Perspectives in Biology* **6**, a019315–a019315 (2014).

Plohl, M., Meštrović, N. & Mravinac, B. Centromere identity from the DNA point of view.

Chromosoma **123**, 313–325 (2014).

Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).

Quadrana, L. *et al.* Natural occurring epialleles determine vitamin E accumulation in tomato fruits. *Nature Communications* 5, (2014).

Quesneville, H. *et al.* Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Computational Biology* **1**, e22 (2005).

Quesneville, H., Nouaud, D. & Anxolabéhère, D. Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. *Journal of Molecular Evolution* **57,** S50–S59 (2003).

Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annual Review of Genetics* **46**, 21–42 (2012).

Rice, P. EMBOSS: The European Molecular Biology Open Software Suite. 2, TIG, volume 16, No. 6 (2000).

Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P. & Hua-Van, A. Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences* **113**, 14763–14768 (2016).

Schnable, P. S. *et al.* The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**, 1112–1115 (2009).

Seibt, K. M., Wenke, T., Muders, K., Truberg, B. & Schmidt, T. Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. *The Plant Journal* **86**, 268–285 (2016).

Selinger, D. A. B-Bolivia, an Allele of the Maize b1 Gene with Variable Expression, Contains a High Copy Retrotransposon-Related Sequence Immediately Upstream. *PLANT*

PHYSIOLOGY **125**, 1363–1379 (2001).

Selinger, D. A. & Chandler, V. L. Major recent and independent changes in levels and patterns of expression have occurred at the b gene, a regulatory locus in maize. *Proceedings of the National Academy of Sciences* **96**, 15007–15012 (1999).

Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* **8**, 272–285 (2007).

Smit, A.F.A., Hubley R. RepeatModeler Open-1.0. RepeatModeler website. Available: http://www.repeatmasker.org/RepeatModeler.html (2008-2010).

Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker open-3.0 at http://repeatmasker.org. Institute for Systems Biology (1996-2004).

Stapley, J., Santure, A. W. & Dennis, S. R. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular Ecology* **24**, 2241–2252 (2015).

Stępiński, D. Functional ultrastructure of the plant nucleolus. *Protoplasma* **251**, 1285–1306 (2014).

Stuart, T. *et al.* Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**, (2016).

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A. & Brown, J. W. Alternative splicing in plants – coming of age. *Trends in Plant Science* 17, 616–623 (2012).

Tanksley, S. D., Bernatzky, R., Lapitan, N. L. & Prince, J. P. Conservation of gene repertoire but not gene order in pepper and tomato. *Proceedings of the National Academy of Sciences* **85**, 6419–6423 (1988).

Tirado-Magallanes, R., Rebbani, K., Lim, R., Pradhan, S. & Benoukraf, T. Whole genome DNA methylation: beyond genes silencing. *Oncotarget* **8**, (2017).

The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**, D331–D338 (2017).

The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).

Vanyushin, B. F. & Ashapkin, V. V. DNA methylation in higher plants: Past, present and future. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1809**, 360–368 (2011).

Vieira, C. *et al.* A comparative analysis of the amounts and dynamics of transposable elements in natural populations of Drosophila melanogaster and Drosophila simulans. *Journal of Environmental Radioactivity* **113**, 83–86 (2012).

Wassenegger, M., Heimes, S., Riedel, L. & Sänger, H. L. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**, 567–576 (1994).

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biology* **17**, (2016).

Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973–982 (2007).

Witte, C.-P., Le, Q. H., Bureau, T. & Kumar, A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences* **98**, 13778–13783 (2001).

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & van der Knaap, E. A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. *Science* **319**, 1527–1530 (2008).

Xiong, W., Dooner, H. K. & Du, C. Rolling-circle amplification of centromeric *Helitrons* in plant genomes. *The Plant Journal* **88**, 1038–1045 (2016).

Ye, C., Ji, G. & Liang, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific Reports* **6**, (2016).

Yu, C., Zhang, J. & Peterson, T. Genome Rearrangements in Maize Induced by Alternative Transposition of Reversed *Ac/Ds* Termini. *Genetics* **188**, 59–67 (2011).

Zemach, A. *et al.* The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. *Cell* **153**, 193–205 (2013)

Zhong, S. *et al.* Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nature Biotechnology* **31**, 154–159 (2013).

Zytnicki, M., Akhunov, E. & Quesneville, H. Tedna: a transposable element de novo assembler. *Bioinformatics* **30**, 2656–2658 (2014).

Annexes

Annexe 1 : Script python pour le formatage des gènes.

```
#!/usr/local/bin/python
# -*- coding: utf8 -*-

import sys
import os


# Gestion des fichiers
gff3 = sys.argv[1]

gff3_o = open (gff3, "r")
gff3_l = gff3_o.readlines()
gff3_o.close()

gff3_out = open (gff3 + "_out.gff3", "w")


# Traitement des données

genes_list = []

for i in xrange (0, len(gff3_l)):

      parent = gff3_l[i].split("\t")[8].split(";")[1].split(":")[1]

      if parent in genes_list:

            continue

      genes_list.append(parent)

      cmd = "grep \"%s\" %s > CDS.txt" %(parent, gff3)

      os.system (cmd)

      cds = open ("CDS.txt", "r")
      cds_l = cds.readlines()
      cds.close()

      for j in xrange (0, len(cds_l)):

            if j == 0:

                  start = cds_l[j].split("\t")[3]
                  stop = cds_l[j].split("\t")[4]

            else:

                  if cds_l[j].split("\t")[3] < start :
                        start = cds_l[j].split("\t")[3]
                  if cds_l[j].split("\t")[4] > stop :
                        stop = cds_l[j].split("\t")[4]
```

```
      towrite = "\t".join([gff3_l[i].split("\t")[0],
gff3_l[i].split("\t")[1], gff3_l[i].split("\t")[2], start, stop,
gff3_l[i].split("\t")[5], gff3_l[i].split("\t")[6],
gff3_l[i].split("\t")[7], parent])

      if i != len(gff3_l) - 1 :

            towrite = "\t".join([towrite, "\n"])

      gff3_out.write(towrite)

gff3_out.close()
```

Annexe 2 : Script python pour le calcul d'une distance intergénique médiane ou moyenne à partir d'un fichier gff3 ordonné par chromosome et par position.

```
#!/usr/local/bin/python
# -*- coding: utf8 -*-


import sys
import re
import statistics


# Gestion des fichiers
gff3 = sys.argv[1]

gff3_o = open (gff3, "r")


# Traitement des données
start = ""
end = ""
chrom_old = ""
dist = 0
dist_interg = []

for line in gff3_o:

      if re.search ("SL2.50ch", line):

            chrom = line.split("\t")[0]
            start = line.split("\t")[3]
            end = line.split("\t")[4]

            if chrom_old != ""  and chrom_old == chrom :
                  if int(start) > int(end_old):
                        dist = int(start) - int(end_old)
                        dist_interg.append(dist)

            chrom_old = chrom
            start_old = start
            end_old = end

med = statistics.median (dist_interg)
# moy = statistics.mean (dist_interg)
print med
# print moy


gff3_o.close()
```

Annexe 3: Script python pour le comptage des effectifs reels par fenêtre de 625 bases en *upstream* des gènes.

```
#!/usr/local/bin/python
# -*- coding: utf8 -*-

import sys
import re

css = sys.argv[1]

ocss = open (css, "r")

count = {}
for i in xrange(0, 2500, 625):
      count[i] = 0

for line in ocss:

      dist = line.split("\t")[18]

      if re.search ("\n", dist):
            dist = int (dist[0:len(dist)-1])*-1
      else:
            dist = int(dist)*-1

      range_dist = (int (dist / 100)) * 100

      if range_dist <= 625:
            count [0] = int (count [0]) + 1
      elif range_dist <= 1250:
            count [625] = int (count [625]) + 1
      elif range_dist <= 1875:
            count [1250] = int (count [1250]) + 1
      elif range_dist <= 2500:
            count [1875] = int (count [1875]) + 1

ocss.close()

count_out = open (css + "_count.txt", "w")

for i in xrange (0, 2500, 625):

      towrite = str(i) + "\t" + str(i+625) + "\t" + str(count[i])
      count_out.write (towrite + "\n")

count_out.close()
```

Annexe 4: Recherche de nouvelles insertions d'ET suite à leur réactivation chez deux mutants *ddm1*.

Introduction

Chez les plantes, la méthylation des cytosines peut avoir lieu dans tous les contextes nucléotidiques (CG, CHG et CHH, H pouvant être n'importe quel nucléotide sauf G) et a pour cible principale les répétitions du génome (Inagaki, S. & Kakutani, T., 2012). Le maintien de ces méthylations au cours des générations est assuré par des protéines spécifiques. Mais bien que la plupart des méthylations soient transmises au cours des générations, elles peuvent aussi avoir lieu *de novo*, dans tous les contextes, grâce à un mécanisme de méthylation de l'ADN dirigé par les ARN (RNA-directed DNA methylation, RdDm) (Wassenegger, M., Heimes, S., Riedel, L. & Sänger, H. L., 1994) (Law, J. A. & Jacobsen, S. E., 2010) (Haag, J. R. & Pikaard, C. S., 2011) (Dalakouras, A. & Wassenegger, M., 2018). D'autres protéines encore, comme la protéine remodeleuse de chromatine DDM1 (Dicrease in DNA Methylation 1) (Brzeski, J. & Jerzmanowski, A., 2003) (Lyons, D. B. & Zilberman, D., 2017), sont impliquées dans la préservation des caractéristiques de l'hétérochromatine.

L'étude de la protéine DDM1 chez *Arabidopsis* a montré sa fonction essentielle pour le maintien des niveaux globaux de méthylation (Zemach, A. *et al.*, 2013) et elle semble principalement contrôler le *silencing* des éléments transposables, et notamment des longs ET localisés dans l'hétérochromatine, prévenant ainsi leur réactivation et donc leur transcription. Chez ce même organisme, les mutants *ddm1* ont largement été observés comme hypométhylés dans tous les contextes cytosines. Cependant, certains phénotypes révélés dans un contexte *ddm1* ne sont pas liés à des altérations dans la structure du génome, mais sont plutôt associés à des modifications épigénétiques qui influencent l'expression des gènes et génèrent des épiallèles stables.

En agronomie, la tomate est l'une des cultures les plus importantes au monde et la régulation de la méthylation de l'ADN est cruciale pour la maturation du fruit, notamment chez cette espèce. Il a été constaté qu'environ 65% du génome de *S. lycopersicum* est composé de répétitions. Il semble alors que comprendre comment les transposons de la tomate sont contrôlés est d'un intérêt à la fois fondamental et agronomique. Cependant, peu de mutants des voies épigénétiques ont été décrits chez cet organisme, et en particulier, les mutants *ddm1*.

Deux copies du gène ddm1 chez la tomate dont les fonctions respectives restent à analyser. L'inactivation de gènes spécifiques de la tomate a récemment été rendue possible grâce à l'avènement de la technologie CRISPR/Cas9. En utilisant cette voie de transformation génétique, l'équipe de Nicolas Bouché (INRA de Versailles) a obtenu des mutants de chacune des deux copies du gène codant DDM1, fournissant alors les mutants *ddm1a* et *ddm1b*. Leur croisement abouti à un double mutant non viable. Une étude des profils de méthylation et des petits ARN de chacun de ces mutants a alors pu être menée. Pour compléter ces analyses, l'équipe a souhaité savoir si l'inactivation de ces gènes chez la tomate M82 causait la réactivation de certains ET et permettait leur transposition. Afin de répondre à cette question, les génomes des deux simples mutants et d'une référence ont alors été générés par une approche de reséquençage en lecture courte pour évaluer si les mutants montraient une augmentation du taux de transposition. L'équipe de Nicolas Bouchés nous a confié ces séquences afin que nous y analysions les polymorphismes liés aux ET. Alors que le génome de référence de la tomate concerne la variété Heinz, les mutants ddm1 ont été obtenus dans un fond génétique de la variété M82. Nous avons donc commencé par annoter les ET du génome de référence de M82 grâce à l'outil REPET (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011, Quesneville, H. *et al.*, 2005). Ensuite, différents outils, T-lex (Fiston-Lavier, A.-S., Carrigan, M., Petrov, D. A. & González, J., 2011), Jitterbug (Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S., 2015) et MELT (Mobile Element Locator Tool) (Gardner, E. J. *et al*., 2017) ont été évalués pour la détection des insertions et délétions d'ET dans nos deux mutants. Au final, un seul outil, MELT, a été utilisé pour réaliser cette détection.

Résultats

Plant Cell Advance Publication. Published on June 6, 2018, doi:10.1105/tpc.18.00167

RESEARCH ARTICLE

Redistribution of CHH Methylation and Small Interfering RNAs across the Genome of Tomato *ddm1* Mutants

Shira COREM[1], Adi DORON-FAIGENBOIM[1], Ophélie JOUFFROY[2], Florian MAUMUS[2], Tzahi ARAZI[1,*] and Nicolas BOUCHÉ[3,*]

[1] Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, P.O. Box 15159, Rishon LeZion 7505101, Israel
[2] URGI, INRA, Université Paris-Saclay, 78000 Versailles, France
[3] Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, 78000 Versailles, France
[*]Corresponding authors: Tzahi ARAZI (tarazi@volcani.agri.gov.il) & Nicolas BOUCHÉ (Nicolas.Bouche@inra.fr)

Short title: Tomato *ddm1* mutants
One-sentence summary: The production of siRNAs and the CHH methylation mediated by the RdDM pathway are enhanced in heterochromatin when DDM1 is non-functional, at the expense of silencing mechanisms occurring in euchromatin.

The author(s) responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Nicolas Bouché (Nicolas.Bouche@inra.fr) and Tzahi Arazi (tarazi@volcani.agri.gov.il).

ABSTRACT

In plants, cytosine methylation, an epigenetic mark critical for transposon silencing, is maintained over generations by key enzymes that directly methylate DNA and is facilitated by chromatin remodelers, like DECREASE IN DNA METHYLATION 1 (DDM1). Short-interfering RNAs (siRNAs) also mediate transposon DNA methylation through a process called RNA-directed DNA methylation (RdDM). In tomato (*Solanum lycopersicum*), siRNAs are primarily mapped to gene-rich chromosome arms, and not to pericentromeric regions as in *Arabidopsis thaliana*. Tomato encodes two *DDM1* genes. To better understand their functions and interaction with the RdDM pathway, we targeted the corresponding genes via the CRISPR/Cas9 technology, resulting in the isolation of *Slddm1a* and *Slddm1b* knockout mutants. Unlike the single mutants, *Slddm1a Slddm1b* double mutant plants display pleiotropic vegetative and reproductive phenotypes, associated with severe hypomethylation of the heterochromatic transposons in both the CG and CHG methylation contexts. The methylation in the CHH context increased for some heterochromatic transposons and conversely decreased for others localised in euchromatin. We found that the number of heterochromatin-associated siRNAs, including RdDM-specific small RNAs, increased significantly, likely limiting the transcriptional reactivation of transposons in *Slddm1a Slddm1b*. Taken together, we propose that the global production of siRNAs and the CHH methylation mediated by the RdDM pathway are restricted to chromosome arms in tomato. Our data suggest that both pathways are greatly enhanced in heterochromatin when DDM1 functions are lost, at the expense of silencing mechanisms normally occurring in euchromatin.

INTRODUCTION

In plants, DNA methylation occurs in all cytosine contexts, mainly to silence repeats and transposable elements (TEs) found in heterochromatic regions. Methylation is maintained over generations by specific proteins, like DNA METHYLTRANSFERASE 1 (MET1) for CG sites or CHROMOMETHYLASES (CMT2 and CMT3) for CHG and CHH sites (where H is any nucleotide except G). In addition to these well-characterized maintenance pathways, cytosines can be methylated *de novo*, in all contexts, by an RNA-directed DNA methylation (RdDM) mechanism (Matzke *et al.*, 2015). The RdDM pathway occurs through two sequential steps involving the production of small interfering RNAs (siRNAs) and non-coding transcripts generated by the plant-specific DNA-dependent RNA polymerase V (Pol V). In the canonical form of RdDM, 24-nt siRNAs are first produced by the successive actions of another plant specific polymerase, the Pol IV, coupled to the RNA-DEPENDENT RNA POLYMERASE 2 (RDR2), and DICER-LIKE 3 (DCL3) (Herr *et al.*, 2005, Onodera *et al.*, 2005, Kasschau *et al.*, 2007, Zhang *et al.*, 2007, Jia *et al.*, 2009, Law *et al.*, 2011, Haag *et al.*, 2012, Blevins *et al.*, 2015, Li *et al.*, 2015b, Zhai *et al.*, 2015). The biosynthesis of siRNAs also results from alternative RdDM pathways. For instance, 21/22-nt siRNAs are produced by the Pol II, RDR6, DCL4 and ARGONAUTE1 (AGO1), in particular when TEs are transcriptionally reactivated (McCue *et al.*, 2012, Nuthikattu *et al.*, 2013, McCue *et al.*, 2015). The siRNAs then guide either AGO4 or AGO6, by base-pairing association, towards Pol V nascent scaffold transcripts (Wierzbicki *et al.*, 2008, Wierzbicki *et al.*, 2009). Finally, the complex formed by AGOs and siRNAs recruits the DNA methyltransferase DOMAIN REARRANGED METHYLTRANSFERASE 2 (DRM2) to *de novo* methylate the genomic region that remained associated with the Pol V transcript (Cao and Jacobsen, 2002, Zhong *et al.*, 2014). Additional proteins such as DECREASE IN DNA METHYLATION 1 (DDM1), a chromatin remodelling protein that belongs to the SWI2/SNF2 family, are involved in preserving heterochromatic features. Indeed, DDM1 was shown to shift nucleosomes *in vitro* (Brzeski and Jerzmanowski, 2003), assisting enzymes maintaining epigenetic marks on DNA or histones to access condensed heterochromatin (Zemach *et al.*, 2013, Lyons and Zilberman, 2017). In this context, recent data suggest that the function of DDM1 and RdDM are antagonistic (Zemach *et al.*, 2013).

In Arabidopsis, DDM1 is essential to sustain global levels of DNA methylation and *ddm1* mutants are extensively hypomethylated in all cytosine contexts (Vongs *et al.*, 1993, Kakutani *et al.*, 1995, Kakutani *et al.*, 1996, Lippman *et al.*, 2004, Zemach *et al.*, 2013).

Disrupting the mouse *LYMPHOID SPECIFIC HELICASE* (*LSH*) gene, which is the mammalian gene most related to *DDM1*, also leads to demethylation of the genome (Dennis *et al.*, 2001), suggesting an ancient and widespread role for DDM1 in maintaining methylation. DDM1 preferentially controls the silencing of TEs (Lippman *et al.*, 2004), particularly long TEs located in the heterochromatin (Zemach *et al.*, 2013), preventing their reactivation and transposition. Consequently, Arabidopsis self-pollinating *ddm1* lines undergo a burst of uncontrolled retrotransposition events associated with developmental abnormalities gradually acquired over generations (Miura *et al.*, 2001, Singer *et al.*, 2001, Tsukahara *et al.*, 2009). By contrast, some of the phenotypes revealed in a *ddm1* background are not alterations in the structure of the genome, but are rather associated with epigenetic modifications that influence gene expression and generate stable epialleles (Kakutani, 1997, Saze and Kakutani, 2007). Accordingly, the epigenetic recombinant inbred lines derived from a *ddm1* mutant show heritable phenotypic variation (Cortijo *et al.*, 2014). Aside from Arabidopsis, *ddm1* mutants have been isolated in maize (Li *et al.*, 2014) and rice (Tan *et al.*, 2016), both species containing two *DDM1* homologs. In rice, the T-DNA insertion loss-of-function single mutants *Osddm1a* and *Osddm1b* had no distinct phenotype but severe growth defects were observed at the first generation of the double mutant, presenting a major reduction of methylation in all contexts. In maize, two single T-DNA insertion loss-of-function *ddm1* mutants showed a significant reduction of methylation in non-CG contexts. Nonetheless, a double *ddm1* mutant could not be isolated by crossing the two single mutants.

Tomato (*Solanum lycopersicum*) is one of the major crops cultivated worldwide and the regulation of DNA methylation is crucial for fruit ripening in this species (Zhong *et al.*, 2013, Liu *et al.*, 2015, Gallusci *et al.*, 2016, Lang *et al.*, 2017). Tomato pericentromeric regions largely extend beyond centromeres and repeats cover about 65% of the genome (The Tomato Genome Consortium, 2012, Zhong *et al.*, 2013, Jouffroy *et al.*, 2016). Understanding how tomato transposons are controlled and silenced is therefore both of fundamental and agronomic interest. Still, few mutants corresponding to epigenetic pathways have been reported in this model plant (Kravchik *et al.*, 2014a, Gouil and Baulcombe, 2016, Lang *et al.*, 2017), and in particular, *ddm1* mutants are yet to be obtained.

In this study, we characterized the methylome, transcriptome and small RNA content of tomato plants deficient for DDM1. DDM1 is encoded in tomato by two genes for which we generated loss-of-function alleles using the CRISPR/Cas9 technology. We found that the *Slddm1a Slddm1b* mutant had drastic hypomethylation particularly in TEs of heterochromatic regions in both CG and CHG contexts. As a counter-balancing mechanism, the distribution of

both 24-nt siRNAs and CHH methylated sites were strongly modified in this mutant. RNA-seq analyses revealed that the transcriptional reactivation of TEs remained limited in the mutant despite the major changes occurring between heterochromatin and euchromatin.

RESULTS

Generation of *Slddm1* knockout mutants

To identify the tomato *DDM1* genes, the BLASTP program was used to search the tomato protein databases (ITAG 2.40 release), using the Arabidopsis DDM1 protein sequence. We found two proteins showing 81.5% identity to each other and 73% (SlDDM1a, Solyc02g062780) or 70% (SlDDM1b, Solyc02g085390) identity to AtDDM1, suggesting that the tomato genome encodes two *DDM1* genes (Supplemental Figure 1A). The SlDDM1a and SlDDM1b proteins also display domain architectures similar to their respective Arabidopsis and rice orthologs (Supplemental Figure 1B). Analysis of published RNA-seq data (The Tomato Genome Consortium, 2012) reveals that the genes are expressed at relatively low levels at anthesis, and are upregulated following fruit set, reaching maximum levels at immature young fruit (Supplemental Figure 1C). The expression of *SlDDM1a* and *SlDDM1b* gradually decreases during fruit maturation and, upon ripening, the expression of both genes declines (Supplemental Figure 1C).

To functionally characterize SlDDM1, we edited the corresponding genes, using the CRISPR/Cas9 technology, to produce loss-of-function mutants. Specific guide RNAs were designed to target the second exons of the *SlDDM1a* gene and the fourth exon of *SlDDM1b* (Figure 1A). These sgRNAs, together with the plant codon-optimized version of *Cas9* (Li *et al.*, 2013), were transgenically expressed in tomato. Genotyping of regenerated T0 transgenic plants identified individual plants that showed Cas9 activity in *SlDDM1* genes and non-transgenic homozygous *SlDDM1* mutants were successfully isolated among their progenies. Sequencing the *crispr-slddm1a-5* and the *crispr-slddm1b-16* mutant alleles (hereafter called *Slddm1a* and *Slddm1b*, respectively) revealed a large deletion of 131 bp in the 2nd exon of *SlDDM1a* (nt 1109-1239, Figure 1B) and a one bp deletion in the 4th exon of *SlDDM1b* (nt 1605, Figure 1B), both causing frameshift mutations near the N-terminal parts of the corresponding proteins (Figure 1C). Phenotype analysis of *Slddm1a* and *Slddm1b* mutants did not reveal any differences with the wild type (Figure 1D). A *Slddm1a Slddm1b* double mutant was obtained by crossing *Slddm1a* and *Slddm1b* homozygotes. Genotyping of the resulting F2 progeny revealed that while all heterozygous and single homozygous genotypes were

indistinguishable from wild-type plants and segregated in a Mendelian manner, the frequencies of the *Slddm1a Slddm1b* mutant were lower (4%, n=24/581) than the expected theoretical frequency (6%, chi-square test p=0.035). The *Slddm1a Slddm1b* plants exhibited pleiotropic vegetative and reproductive phenotypes. Vegetative phenotypes included variegated cotyledons and leaves and overall smaller size likely due to growth retardation (Figure 1D and 2A). Reproductive phenotypes included smaller floral buds, most of which senesced prematurely except few that produced small flowers displaying partially opened petals and normal anthers and pistils (Figure 2B). Compared to wild-type, mutant anthers produced much less pollen with significantly reduced viability as indicated by Alexander staining (Supplemental Figure 2). Occasionally, mutant flowers could set a small parthenocarpic fruit (Figure 2C) that never produced viable offspring. Altogether, our results indicate that DDM1 is essential for normal vegetative and reproductive tomato development.

The *ddm1* mutations have a limited effect on global CHH methylation levels in tomato

We determined the methylation patterns of *Slddm1* single and double mutants by sequencing their genomes after bisulfite conversion. Genomic DNAs were extracted from leaves of two biological replicates per genotype. The levels of methylation per cytosine confirmed that the biological replicates were closely correlated (Pairwise Pearson correlation values between biological replicates > 0.87 for CGs and CHG and 0.79 for CHH; Supplemental Figure 3) and that the bisulfite conversion rates were >99% because the chloroplast sequences remained unmethylated (Supplemental Table 1). In the *Slddm1a Slddm1b* mutant, the total number of methylated cytosines was decreased by 49%, 64% and 24% in the CG, CHG and CHH contexts, respectively, compared to the wild type (Supplemental Table 1). We observed similar changes when the average levels of methylation were calculated in 1 kb-tiles partitioning the genome (Figure 3A and 3B). Consistent with results obtained in Arabidopsis (Vongs *et al.*, 1993, Kakutani *et al.*, 1995, Kakutani *et al.*, 1996, Lippman *et al.*, 2004, Zemach *et al.*, 2013) and rice (Tan *et al.*, 2016), the complete disruption of *DDM1* genes in tomato led to a drastic hypomethylation of the genome, mainly in the CG and CHG contexts.

Monitoring the global methylation of genes and TEs revealed several interesting features. On average, the CG, CHG and CHH methylation levels of the *Slddm1a Slddm1b* mutant were reduced by 12%, 42% and 21% in gene bodies, respectively (Figure 3C). CG methylation is frequently found in genes of plants and is independent of DDM1 (Stroud *et al.*, 2013). The presence of non-CG methylation in tomato genes is more unusual but could possibly be explained by the large number of TEs associated with gene-enriched regions

(Jouffroy *et al.*, 2016). When TEs were inspected globally, we found that both CG and CHG methylation levels were drastically reduced (by 41% and 51%, respectively) in the *Slddm1a Slddm1b* mutant, in agreement with the role played by DDM1 to maintain heterochromatin. However, the CHH methylation level of TEs was decreased by only 7% (Figure 3C), in sharp contrast with the strong global loss of CHH methylation (40%) observed in Arabidopsis or rice *ddm1* TEs (Zemach *et al.*, 2013, Ito *et al.*, 2015, Panda *et al.*, 2016, Tan *et al.*, 2016). These data suggest that both CG and CHG methylations of TEs are severely altered in *Slddm1a Slddm1b*, unlike CHH methylation, and this was confirmed by measuring the methylation levels for different families of TEs and repeats (Supplemental Figure 4).

CHH methylation changes between euchromatic and heterochromatic *Slddm1* TEs

The regions that were significantly differentially methylated (DMRs) between mutants and wild type were identified. The *Slddm1a Slddm1b* mutant contained a very high number of DMRs hypomethylated (hypoDMRs) in the CG and CHG contexts (190026 and 249593, respectively; Figure 4A and Supplemental Dataset 1) predominantly related to heterochromatic regions (Figure 4B; Supplemental Figure 5, regions in green) and overlapping with TEs (82% for the CG hypoDMRs, 87% for the CHG hypoDMRs; Figure 4C) that were heavily methylated in the wild type (Figure 4D). In parallel, a more limited number of hypermethylated DMRs (hyperDMRs) was identified for CGs and CHGs (719 and 1816 respectively; Figure 4A and Supplemental Dataset 2). 71% (507) of the CG hyperDMRs and 64% (1164) of the CHG hyperDMRs were included in repeat-poor regions (Supplemental Dataset 3) and were therefore mostly localized in euchromatic regions (Figure 4B; Supplemental Figure 5, regions in black). They overlap with genes or genes containing TEs (30% of the CG hyperDMRs and 40% of the CHG hyperDMRs), or with TEs alone (16% of the CG hyperDMRs and 23% of the CHG hyperDMRs) (Figure 4C). CG hyperDMRs corresponded to regions unmethylated in the wild type that become methylated at both CGs and CHGs in *Slddm1a Slddm1b* (Supplemental Figure 6). CHG hyperDMRs were methylated in all contexts in the wild type and gained additional mCHG, and to a lesser extent mCHH (Supplemental Figure 6). Therefore, TEs of the heterochromatin were vastly depleted of CG and CHG methylated sites in *Slddm1a Slddm1b*, and additionally, certain euchromatic regions become methylated in these contexts.

The opposite situation was observed for regions differentially methylated in the CHH context. The density of CHH hypoDMRs (8518 were identified; Figure 4A and Supplemental Dataset 1) was higher at chromosome arms (Figure 4B; Supplemental Figure 5, regions in

green) in regions enriched for genes or localised at the frontiers between euchromatin and heterochromatin (*i.e* in repeat-intermediate regions). Indeed, 30% (2582) of the CHH hypoDMRs were localized in repeat-poor regions, 33% (2802) in repeat-intermediate regions and 39% (3313) in repeat-rich regions, a distribution that differed (chi-square test, $p<10^{-300}$) from the one expected if those DMRs were equally distributed in all regions (26%, 19% and 55%, respectively). Our results were consistent with previous analyses showing a reduction of mCHH in euchromatin of the rice *ddm1* mutants (Tan *et al.*, 2016). The CHH hypoDMRs mostly overlapped with TEs (74% of the CHH hypoDMRs; Figure 4C). A total of 10297 CHH hyperDMRs (Figure 4A and Supplemental Dataset 2) were detected between *Slddm1a Slddm1b* and the wild type, but this time, mostly localized in heterochromatic regions (Figure 4B; Supplemental Figure 5, CHH regions in black). They correspond to TEs (85% of CHH hyperDMRs overlap with TEs; Figure 4C) in which mCHH levels increase by almost three times in *Slddm1a Slddm1b* (Figure 5; CHH hyperDMRs). 0.1% (3) of the hypoCHH DMRs found in repeat-poor regions were close to (within 500 bp) a CG hyperDMR and 0.4% (11) to a CHG hyperDMR. Therefore, in repeat-poor regions, CHH hypomethylation and CG/CHG hypermethylation occurred at different locations. By contrast, 64% (4841) of the CHH hyperDMRs found in repeat-rich areas were close to (within 500 bp) a CG hypoDMR and 60% (4478) to a CHG hypoDMR. Therefore, some heterochromatic TEs were actively re-methylated at CHH sites, while others, localised within gene-enriched regions, became hypomethylated in this context, altogether resulting in limited quantifiable changes in overall CHH methylation of TEs (Figure 3C).

In Arabidopsis, CHH methylation is both maintained by CMT2 (Zemach *et al.*, 2013, Stroud *et al.*, 2014) or the RdDM pathway that depends on Pol IV and Pol V. To examine whether the RdDM pathway is altered in *Slddm1a Slddm1b*, thus compromising the methylation of CHH sites in TEs, we retrieved the methylome sequences of both *Slpol iv* and *Slpol v* tomato *crispr* mutants (Gouil and Baulcombe, 2016). 51482 CHH hypoDMRs were identified between *Slpol iv* and the corresponding wild type and 41016 CHH hypoDMRs for *Slpol v*. 60% of these CHH hypoDMRs were localised in regions enriched for genes and 20% were in repeat-rich regions, confirming that the RdDM is mostly active in euchromatin. Interestingly, we found that 60% (1546) of the CHH hypoDMRs (Supplemental Dataset 1) identified in the euchromatin of *Slddm1a Slddm1b* overlapped with CHH hypoDMRs of *Slpol iv* and 50% (1299) overlapped with those of *Slpol v*. This indicates that euchromatic regions hypomethylated in the CHH context in *Slddm1a Slddm1b* are mainly TEs targeted by the RdDM (Figure 4D). We confirmed this result by dividing the *Slddm1a Slddm1b* CHH

248 hypoDMRs of repeat-poor regions in two groups: the first one (1198 DMRs) depended on
249 RdDM and corresponded to CHH hypoDMRs overlapping between *Slddm1a Slddm1b*, *pol iv*
250 and *pol v*. The second group (943 DMRs) corresponded to RdDM-independent CHH
251 hypoDMRs not overlapping with *pol iv* or *pol v* DMRs. 76% (917) of RdDM CHH
252 hypoDMRs overlapped with short euchromatic TEs (mean length: 238 bp) that had lost
253 almost 13% of mCHG and 52% of mCHH in *Slddm1a Slddm1b* (Figure 5 and Supplemental
254 Figure 7A). 77% (726) of the hypoCHH non-RdDM DMRs corresponded to long
255 heterochromatic TEs (mean length: 751 bp) not targeted by the RdDM in the wild type and
256 losing 39% of mCG, 66% of mCHG and 69% of mCHH in *Slddm1a Slddm1b* (Figure 5 and
257 Supplemental Figure 7B). Thus, tomato euchromatin contains two types of TEs differently
258 controlled by DDM1 and the RdDM.
259
260 **The production of small interfering RNAs increases in heterochromatin and decreases**
261 **in euchromatin of *Slddm1a Slddm1b***
262 The RdDM pathway depends on the production of small RNAs, in particular 24-nt siRNAs;
263 hence, we sequenced the small RNAs of *Slddm1a Slddm1b* and compared their distribution
264 along the genome to that in wild-type tomato. Reproducibility between biological replicates
265 was confirmed by performing a principal component analysis (PCA) to visualise the
266 differences (Supplemental Figure 8A). Mapping the reads revealed that the 24-nt siRNAs of
267 wild-type tomato follow the general patterns of small RNAs along the chromosomes (The
268 Tomato Genome Consortium, 2012), being almost excluded from the large pericentromeric
269 regions, and accumulating in chromosome arms (Figure 6A). Indeed, repeat-poor regions of
270 the wild-type plants contain 7-fold more 24-nt reads, compared to regions highly enriched in
271 repeats (Figure 6B). While the production of these 24-nt siRNAs, depending on both SlPol IV
272 and SlDCL3 (Figure 6A and 6B), was similar to that in Arabidopsis, their genomic
273 distribution differed sharply as Arabidopsis siRNAs are highly prevalent at pericentromeres
274 (Zhang *et al.*, 2007, Mosher *et al.*, 2008, Law *et al.*, 2013, Liu *et al.*, 2014). The distribution
275 of 24-nt siRNAs in the *Slddm1a Slddm1b* mutant was drastically modified compared to the
276 wild type. Their levels decreased by almost two-fold in repeat-poor regions and increased by
277 the same proportion in repeat-rich regions when reads were normalized against total mapped
278 reads (Figure 6B). Similar results were obtained when reads were normalized against miRNA
279 reads (Supplemental Figure 9). Then, 106926 siRNA clusters were defined (see Methods) and
280 we compared their profiles of expression between the wild type and the *Slddm1a Slddm1b*
281 mutant (Figure 6C). We found that 6913 heterochromatic siRNA clusters showed increased

(log2FC(*Slddm1aSlddm1b*/WT) > 2) expression of 24-nt siRNAs compared to the wild type (DESeq2 significance cut-off of 0.01). In repeat-poor regions, the levels of 24-nt siRNAs were decreased for 967 siRNAs clusters and increased for 680 clusters. Furthermore, we determined the levels of 21, 22 and 23-nt siRNAs in repeat-poor and repeat-rich regions containing a significant (DESeq2 significance cut-off of 0.01) number of reads (Figure 7 and Supplemental Figure 10). The production of 23/24-nt siRNAs was inhibited and enhanced in gene-rich and gene-poor regions, respectively and the production of 21/22-nt siRNAs was increased in repeat-rich regions of *Slddm1a Slddm1b*, likely because the RDR6-RdDM pathway was activated. Less than 3% of CG or CHG DMRs overlapped with siRNAs clusters deregulated (log2FC(*Slddm1a Slddm1b*/WT) > 2 or < -2) in *Slddm1a Slddm1b*. The same results were obtained for CHH hypoDMRs found in repeat-poor areas. By contrast, 14% of the CHH hyperDMRs localised in regions enriched for repeats overlapped with 23/24-nt siRNA clusters upregulated (log2FC(*Slddm1aSlddm1b*/WT) > 2) in *Slddm1a Slddm1b* and 6% with 21/22-nt siRNA clusters. CHH hyperDMRs corresponded to heterochromatic TEs that were heavily methylated in the wild type but were not targeted by 24-nt siRNAs and the canonical RdDM pathway (Figure 5; CHH hyperDMRs). In *Slddm1a Slddm1b*, these TEs gained mCHH and became the targets of 24-nt siRNAs, although their levels remained modest (Figure 5; CHH hyperDMRs). Further analyses will help to determine the functional role of these small RNA populations in the *Slddm1a Slddm1b* mutant.

Finally, we examined the distribution of siRNAs around genes. In the wild type, the metaprofile of 24-nt siRNAs revealed a peak at ≈ 500 bp upstream of the transcription start and to a lesser extent downstream of the transcription stop site that corresponded to similar changes in the levels of mCHH (Figure 6D). Interestingly, both peaks were markedly reduced in *Slddm1a Slddm1b*, further indicating that the RdDM pathway is compromised in euchromatin of the mutant.

The number of TEs reactivated in the first generation of *Slddm1a Slddm1b* mutants is limited

To better understand whether disrupting the RdDM in tomato affects gene and TE expression, we performed RNA-seq analyses using three biological replicates for the wild type and four for the *Slddm1a Slddm1b* mutant, including those already used for the sRNA-seq (Supplemental Table 2). We verified that all biological replicates were grouped together (Supplemental Figure 8B). We found a total of 138 genes that were significantly (FDR threshold ≤ 0.01) downregulated (log2FC(*Slddm1a Slddm1b*/WT) < -1.5) in *Slddm1a*

316 *Slddm1b*, compared to the wild type, and 1239 upregulated (log2FC(*Slddm1a Slddm1b*/WT) >
317 1.5) genes (Supplemental Dataset 4) that include 390 TE-genes, corresponding to TEs
318 incorrectly annotated as genes (Jouffroy *et al.*, 2016). Almost 50% of the upregulated genes
319 were overlapping with hypoDMRs in the CG or CHG contexts, and 10% with CHH
320 hypoDMRs (Figure 8A). Moreover, 71% of these different types of hypoDMRs were
321 overlapping between them (Figure 8B), and 50 to 70% overlapped with TEs (Figure 8A). Yet,
322 very few de-regulated genes overlapped with hyperDMRs (Figure 8A). Therefore, our RNA-
323 seq data show that TEs localised near or within genes were derepressed.
324 We determined more precisely whether the TEs were transcriptionally reactivated in
325 the *Slddm1a Slddm1b* mutant by using the annotations obtained for each family (see
326 Methods). To monitor their expression, we used both a multiple-mapping strategy, where
327 reads mapping to different locations with a high score were assigned to all these locations and
328 a unique-mapping strategy (see Methods). The results showed that the transcriptional
329 reactivation of TEs was limited to a small fraction of the annotated TEs in *Slddm1a Slddm1b*
330 mutants. Indeed, in a total of 536643 TE annotations, we found that 2% were upregulated
331 when reads were mapped at unique locations, and 3% when they were mapped multiple times
332 (Table 1). On average, 65% of the derepressed TEs were localised in repeat-rich regions while
333 12% and 23% were localised in repeat-intermediate and repeat-poor regions, respectively
334 (Figure 8C). Only a fraction of the TEs annotated could potentially be both re-activated
335 transcriptionally and detected by our RNA-seq analysis. Although their exact number was
336 difficult to establish, we hypothesised that longer elements (>2 kb) were the most conserved,
337 containing functional internal genes that could be transcribed. We found that 6.6 % of the
338 *Gypsy* elements longer than 2 kb and 9.4 % for the *Copia* were derepressed in *Slddm1a*
339 *Slddm1b*. From these data, we conclude that the fraction of TEs transcriptionally reactivated
340 in the first generation upon loss-of-*SlDDM1* function was restricted. The reactivation of
341 additional TEs in *Slddm1* might take more generations, or alternatively, the RdDM pathway
342 of tomato might be particularly efficient to rapidly re-silence TEs localised in the
343 heterochromatin.
344
345 **DISCUSSION**
346
347 In this study, we used the CRISPR/Cas9 technology to generate loss-of-function *Slddm1*
348 alleles, and utilized them to investigate *SlDDM1* functions in tomato, a model crop with a
349 complex genome rich in TEs. We isolated single *Slddm1a* and *Slddm1b* mutants, as well as

the corresponding *Slddm1a Slddm1b* mutant plants that show severe developmental defects. In plants deficient for DDM1, heterochromatic TEs are depleted of mCGs and mCHGs that normally silence them. We also found that some euchromatic TEs have lost DNA methylation, but in the CHH context. Similarly, the production of siRNAs, including 24-nt, is enhanced in heterochromatin.

The *ddm1* alleles can be propagated in the homozygous state for several generations in Arabidopsis, but not in other plant species. In maize, two *DDM1* orthologs were identified. The corresponding single mutants are viable, but the double mutant could not be recovered despite the screening of a substantial amount of offspring (Li *et al.*, 2014). Rice also contains two *DDM1* genes and the double mutant exhibits severe developmental abnormalities and sterility (Tan *et al.*, 2016). In tomato, we report that the vegetative and reproductive development of *Slddm1a Slddm1b* is drastically altered (Figure 2) and that the plants are completely sterile. Although Arabidopsis seems to be particularly tolerant to genome-wide modifications of methylation patterns, other plant species including crops are much more sensitive. One explanation could be that the number of TEs re-mobilized in a *ddm1* background remains limited in Arabidopsis in contrast to these sensitive species. In addition, the chromatin is structured and organised very differently in plants enriched in TEs. For instance, the genome of rice is partitioned in thousands of topologically associated domains (TADs) that greatly differ from the chromatin packing patterns of Arabidopsis (Liu *et al.*, 2017).

DDM1 is essential for global maintenance of DNA methylation, and consequently, the corresponding tomato mutants are extensively hypomethylated, in particular in heterochromatic regions, as observed in Arabidopsis (Vongs *et al.*, 1993, Kakutani *et al.*, 1995, Kakutani *et al.*, 1996, Lippman *et al.*, 2004, Zemach *et al.*, 2013) and rice (Tan *et al.*, 2016). We also reveal that the *Slddm1a Slddm1b* mutant is drastically hypomethylated in both the CG and CHG contexts (Figure 3A and 3B) in regions corresponding to heterochromatic TEs (Figure 4). Additionally, our results show that hypoDMRs are already detectable in both *Slddm1a* and *Slddm1b* single mutants (Figure 4A), indicating that the absence of one of the two DDM1 proteins is not fully compensated for by the other one. Nevertheless, the two single mutants and the wild type develop similarly (Figure 1D), at least in the first generations. Future work will help to determine whether the heterochromatic hypomethylation observed in the single mutants is stable over generations, whether TEs are

384 derepressed in these backgrounds, and finally if the two tomato DDM1 proteins have specific
385 functions.
386
387 In Arabidopsis, the methylomes of first-generation homozygous *ddm1* plants diverge
388 from those of the progenies obtained after eight rounds of self-propagation (Ito *et al.*, 2015).
389 Whereas the CG methylation continuously decreases over generations, genes tend to gain
390 ectopic methylation at non-CG sites. A possible explanation is that factors normally targeting
391 heterochromatin are released in a *ddm1* background, inducing the spreading of methylation
392 into euchromatic regions in later generations. We observe an ectopic gain of CG and CHG
393 methylation in the euchromatin of the *Slddm1a Slddm1b* mutant (Figure 4B and Supplemental
394 Figure 6) occurring by two different ways. First, some regions that are not methylated in the
395 wild type, or targeted by siRNAs, are *de novo* methylated at both CGs and CHGs sites in
396 *Slddm1a Slddm1b*, by a mechanism independent of RdDM and siRNA targeting
397 (Supplemental Figure 6A; CG hyperDMR). More data are needed to determine whether these
398 regions carry specific features that attract chromatin modifying factors or methyltransferases.
399 Second, the CHG and more slightly the CHH methylations increase in regions that are already
400 methylated and targeted by 23/24-nt siRNAs in the wild type (Supplemental Figure 6A; CHG
401 hyperDMR). These euchromatic regions are likely TEs remethylated by both CMT3 and
402 CMT2 in *Slddm1a Slddm1b*, reinforcing their silencing (an example is given in Supplemental
403 Figure 6B). The CHG hypermethylation changes, detected after multiple generations in
404 Arabidopsis, occur in only one generation for tomato, suggesting that disrupting SlDDM1
405 activities has more immediate consequences. However, we found very few genes deregulated
406 in the *Slddm1a Slddm1b* mutant and associated with hypermethylated CG or CHG-DMRs
407 (Figure 8A) despite the presence of these hyperDMRs in numerous genes (Figure 4C). Thus,
408 the euchromatic CG and CHG methylation gained in one generation in tomato does not
409 significantly alter the transcription of the associated genes. By contrast, we found that almost
410 50% of the upregulated genes in *Slddm1a Slddm1b* are associated with regions
411 hypomethylated (Figure 8A) in both the CG and CHG contexts (Figure 8B). The majority of
412 these genes are overlapping TEs (Figure 8A; *CDS+TE*) that are likely derepressed. Further
413 studies will be required to determine if TE activation within tomato genes controls gene
414 expression.
415
416 Two parallel pathways controlling the methylation at CHH sites of TEs were
417 discovered in Arabidopsis. The first pathway involves CMT2, depends on DDM1, and targets

418 the long TEs predominantly found in the constitutive heterochromatin, like the *Gypsy*
419 elements (Zemach *et al.*, 2013, Stroud *et al.*, 2014). Although CMT2 seems to be absent from
420 maize (Zemach *et al.*, 2013), BLAST analyses reveal the presence of one homolog in the
421 genome of tomato (Gallusci *et al.*, 2016). Direct evidence to confirm the function of this gene
422 is yet to be provided, but indirect evidence exists, such as the methylation profiles of tomato
423 RdDM mutants (Gouil and Baulcombe, 2016). Indeed, the methylation of CHH sites
424 decreases in gene-enriched regions of both *Slpol iv* and *Slpol v crispr* mutants, towards
425 chromosome arms, and remains unchanged in heterochromatin where mCHH most likely
426 depends on SlCMT2 (Gouil and Baulcombe, 2016). Second, the RdDM (Matzke *et al.*, 2015)
427 is active at short TEs and edges of long TEs (Zemach *et al.*, 2013, Stroud *et al.*, 2014), is
428 required to maintain the CHH methylation of TEs located within genic regions, relies on Pol
429 IV-dependent 24-nt siRNAs and is independent of DDM1. In the *Slddm1a Slddm1b* mutant,
430 the CHH methylation decreases in certain TEs that are found in euchromatic regions (Figure
431 4B, 4C, 5 and Supplemental Figure 7) and the global content of 24-nt siRNA also decreases in
432 these regions (Figure 6), indicating that the corresponding pathways are compromised in
433 *Slddm1a Slddm1b* plants. At the same time, the situation is inverted for certain other
434 heterochromatic TEs of *Slddm1a Slddm1b* that gain CHH methylation (Figure 4B, 4C and 5)
435 and become the targets of 24-nt siRNAs (Figure 5). In addition, the heterochromatin of
436 *Slddm1a Slddm1b* is enriched in siRNAs (Figure 6 and Figure 7), indicating that their
437 production is enhanced in regions densely populated by TEs. Altogether, we propose that the
438 homeostasis of the pathways controlling the production of siRNAs and CHH methylation
439 driven by the RdDM is severely compromised in *Slddm1a Slddm1b*, leading to their partial
440 redistribution towards heterochromatin. In euchromatin, this has different consequences on
441 long or short TEs that are independent or dependent of RdDM, respectively, and both are
442 differently affected by the lack of DDM1 enzymes. In *Slddm1a Slddm1b*, the canonical
443 RdDM targets short TEs, but much less efficiently compared to wild type. Long TEs of
444 euchromatin are severely hypomethylated in all contexts in the mutant, implying that
445 SlDDM1 controls their silencing. Few TEs that were targeted by RdDM become targeted by
446 other pathways, likely involving CMTs. In addition, we observed that the distribution of 24-nt
447 siRNAs and mCHH correlates near genes, reaching a peak at ≈ 500 bp upstream of the
448 transcription start site, that decreases in the *Slddm1a Slddm1b* mutant (Figure 6D). Thus,
449 RdDM seems to be particularly active at gene boundaries of wild-type tomato and disrupting
450 the *SlDDM1* genes strongly impairs this control. Whether some of these regions are similar to
451 CHH islands found in maize (Gent *et al.*, 2013, Li *et al.*, 2015a) remains to be determined.

452

453 In tomato, the production of 24-nt siRNAs that depends on both SlPol IV (Gouil and
454 Baulcombe, 2016) and SlDCL3 (Kravchik *et al.*, 2014a) is restricted to gene-enriched regions
455 (Figure 6A and 6B) and follows the pattern previously reported for all categories of small
456 RNAs (The Tomato Genome Consortium, 2012). Likewise, a very similar distribution was
457 observed in other *Solanaceae* such as pepper (*Capsicum*) (Qin *et al.*, 2014) or potato (The
458 Tomato Genome Consortium, 2012). By contrast, Pol IV-dependent siRNAs of Arabidopsis
459 are produced throughout the whole genome, including pericentromeres (Zhang *et al.*, 2007,
460 Mosher *et al.*, 2008, Law *et al.*, 2013, Liu *et al.*, 2014). Therefore, contrary to Arabidopsis,
461 SlPol IV is not efficiently transcribing the heterochromatic regions of tomato, or alternatively,
462 Pol IV transcripts derived from heterochromatin are not efficiently processed, leading to the
463 restriction of Pol IV-siRNA production to chromosome arms. Interestingly, we found that
464 23/24-nt siRNA production, and therefore the Pol IV activity, can be greatly enhanced in the
465 pericentromeric regions of the *Slddm1a Slddm1b* mutant, making tomato plants, and this
466 particular mutant, singular model systems in which to study Pol IV recruitment target sites.
467 We observed a similar increase for 21/22-nt siRNAs (Figure 7 and Supplemental Figure 10),
468 that are likely produced by the RDR6-RdDM pathway following the transcriptional
469 reactivation of TEs (McCue *et al.*, 2012, Nuthikattu *et al.*, 2013, McCue *et al.*, 2015). Future
470 studies will determine whether tomato miRNAs control the production of easiRNAs like in
471 Arabidopsis (Creasey *et al.*, 2014, Borges *et al.*, 2018). Altogether, this indicates that the
472 heterochromatic regions of wild-type tomato are probably much less accessible than those of
473 Arabidopsis to enzymes involved in the production of small RNAs.
474

475 We provide further evidence that *DDM1* genes are essential in plants by using the
476 CRISPR/Cas9 technology to obtain the corresponding tomato mutants. In plants deficient for
477 DDM1, heterochromatic TEs are depleted of the mCGs and mCHGs that normally keep them
478 inactive. By contrast, some of them gain CHH methylation, by a counterbalancing
479 mechanism, that probably limits the number of reactivated TEs. The small RNAs, including
480 the 24-nt siRNAs, and the methylated CHH sites, are both partially redistributed from
481 euchromatic regions towards heterochromatic regions in *Slddm1a Slddm1b*, suggesting that
482 both pathways which are under a tight homeostatic control, are compromised in the mutant.
483 Additionally, this also strongly suggests that the global production of siRNAs is restricted to
484 chromosome arms in the wild type because the compacted heterochromatin is inaccessible to
485 the enzymes responsible for their synthesis.

METHODS

Plant material and growth conditions

Tomato (*Solanum lycopersicum*) cv. M82 plants were grown in a greenhouse, with temperatures ranging from 15 to 30°C, in 4 L pots filled with tuff-peat mix with nutrients. For *in vitro* culture, seeds were surface-sterilized by 3 min treatment with 70% ethanol, followed by 20 min with 2% hypochlorite solution containing 0.1% Tween. After a thorough rinse with sterile distilled water, seeds were sown on sterile solidified medium based on MS (Murashige and Skoog) medium including vitamins (*Duchefa*), with pH adjusted to 5.7 using KOH. Plant Agar (*Duchefa*), was added to a final concentration of 0.8%. Germinated seedlings were grown in a growth chamber at 22°C under a 16/8h light/dark regime (photosynthetic photon flux density: 50–70 µmol m^{-2} s^{-1}; six OSRAM basic T8 cool daylight lamps model L18w/765).

Generation of *ddm1* tomato mutants using the CRISPR/Cas9 technology

To knockout the *SlDDM1a* and *SlDDM1b* genes by CRISPR/Cas9, sgRNAs containing 20-bp target sequences specific to the 5′ coding region of the corresponding genes followed by the NGG protospacer adjacent motif (PAM) were designed. To facilitate mutation detection, target sequences were designed to include a restriction enzyme site (*Pml*I for *SlDDM1a* and *Mlu*CI for *SlDDM1b)* overlapping the predicted cut site of the Cas9 nuclease. Then, corresponding sgRNAs were amplified using specific primers (Supplemental Table 3), digested and cloned into the pRCS binary vector *Sal*I-*Hin*dIII sites under the control of the synthetic Arabidopsis *U6* promoter (Waibel and Filipowicz, 1990) alongside the plant codon-optimized version of *Cas9* (Li *et al.*, 2013) expressed under the constitutive CaMV 35S promoter.

The binary constructs pRCS:Cas9-sgRNA-*SlDDM1a* and pRCS:Cas9-sgRNA-*SlDDM1b* were transformed into tomato by co-cultivation of cotyledons derived from 14-day-old seedlings using *Agrobacterium*-mediated transformation (strain GV3101) followed by regeneration on selective kanamycin-containing media as described previously (Kravchik *et al.*, 2014b). Further validation was performed by PCR of genomic DNA with the primer pair Cas9-Fwd and Cas9-Rev to detect the 35S:Cas9 transgene.

To identify mutant plants, genomic DNA was extracted (Phire, Thermo) from each transgenic T0 plant and used for PCR with specific primers flanking the sgRNA target

sequences (primer sequences are given in Supplemental Table 3). The resulting amplicons were resolved by agarose gel electrophoresis to detect large indels or digested with *Pml*I for *SlDDM1a* and *Mlu*CI for *SlDDM1b* and resolved by agarose gel electrophoresis to detect loss of the corresponding restriction enzyme sites. Then, the progeny of positive T0 plants were screened as above to detect Mendelian segregation of the mutation, to confirm its heritability and the absence of the *35S:Cas9-sgRNA* transgene. The amplicons from identified non-transgenic *Slddm1* homozygous mutants were sequenced to determine the nature of the mutation.

Pollen viability and germination assay

Freshly harvested anthers from 16 anthesis flowers were sliced and incubated for 3 hours in germination medium (10% sucrose, 100 mg/L H_3BO_3, 300 mg/L $CA(NO_3)_2$, 200 mg/L $MgSO_4$, and 100 mg/L KNO_3) at room temperature, followed by 1 hr incubation in Alexander dye (Alexander, 1969). Pollen grains were counted from 18 arbitrarily selected microscopic fields; they were considered viable when active cytoplasm was evident and considered germinated if the tube length was equal or greater than the grain diameter.

Whole-genome bisulfite sequencing and DMR analyses

Both *Slpol iv* and *Slpol v* BS-seq (Gouil and Baulcombe, 2016) are available from the SRA database (accession SRP081115). For *Slddm1* mutants, genomic DNA was extracted using Genomic DNA extraction kit (Macherey-Nagel, Germany) from wild type (cv. M82), *Slddm1a* mutant (*crispr-Slddm1a-5*), *Slddm1b* mutant (*crispr-Slddm1b-16*) and *Slddm1a Slddm1b* mutant leaves (leaves 3 and 4). For the wild type, the *Slddm1a* and the *Slddm1b* single mutants, we combined the DNAs from 20 plants for each genotype. The DNAs of 12 plants were combined for *Slddm1a Slddm1b*. We did two biological replicates per genotype. To sequence the methylomes of *Slddm1* mutants, bisulfite treatments, library preparations and whole-genome sequencings (depth of ± 19X final; Supplemental Table 1) were performed at BGI (China) using the HiSeq technology (*Illumina*) producing 150 bp paired-end reads (Supplemental Table 1). Reads were trimmed and cleaned (short reads < 20 bp removed; paired-end validation) with *Trim Galore!* (Babraham Bioinformatics) version 0.4.2 and *Cutadapt* version 1.8.3 (Martin, 2011). Clean reads were aligned to the wild-type reference genomes (version SL2.50; The Tomato Genome Consortium, 2012) with *Bismark* (Krueger and Andrews, 2011) version 0.14.5 and standard options (*Bowtie2*; 1 non-bisulfite mismatch allowed per read). Reads that were not matching at unique locations were discarded. Identical

554 pairs were collapsed using the script provided with *Bismark*. Subsequent analyses were done
555 using the following R (version 3.3.3) packages: *bsseq* version 1.10 (Hansen *et al.*, 2012) and
556 *DSS* (Dispersion Shrinkage for Sequencing data) version 2.14 (Wu *et al.*, 2015) to call
557 Differentially Methylated Regions (DMRs) based on a Wald test procedure and accounting
558 for both biological variations among replicates and sequencing depths. First, differential
559 methylation statistical tests were performed at each C locus by calling the *DSS DMLtest*
560 function with the parameter smoothing=TRUE. Then, Differentially Methylated Loci (DML)
561 were retained when the difference in mean methylation levels was >0.1 for CGs or CHGs and
562 >0.07 for CHH with a posterior probability > 0.9999. Differentially Methylated Regions
563 (DMRs) were identified by using the *DSS callDMR* function with standard parameters (DMR
564 length > 50 bp, number of DML > 3, more than 50% of C sites with p-value < 0.0001). DMRs
565 closer than 50 bp were merged into longer ones. To define hypo- or hyper-DMRs, we applied
566 an additional cutoff to keep DMRs with at least a 10% change in methylation ratio for CHHs,
567 20% for CHGs and 30% for CGs. We used the MethylKit software (Akalin *et al.*, 2012) to
568 monitor the levels of methylation per cytosine. To determine the bisulfite conversion rates,
569 reads were aligned to the *Solanum lycopersicum* chloroplast sequence (NCBI reference
570 sequence NC_007898.3) with *Bismark* (Krueger and Andrews, 2011). Whole-genome
571 bisulfite sequencing statistics are provided in Supplemental Table 1.
572
573 **Annotation of TEs**
574 Transposable elements were annotated with REPET (Flutre *et al.*, 2011), and the repeat rich,
575 intermediate and poor regions were defined as previously described (Jouffroy *et al.*, 2016),
576 using the SL2.50 version of the tomato genome assembly (The Tomato Genome Consortium,
577 2012). The *gff* file is available on the Sol Genomics Network. 2246 putative TE genes for
578 which over half of the CDS fraction is covered by high confidence TEs were annotated
579 previously (Jouffroy *et al.*, 2016).
580
581 **Analysis of expression by RNA-seq**
582 Libraries preparation were done using the INCPM-RNA-seq. Briefly, polyA fraction (mRNA)
583 was purified from 500 ng of total RNA extracted from leaves, following by fragmentation and
584 generation of *Slddm1a Slddm1b* stranded cDNAs. Then, end repair, A base addition, adapter
585 ligation and PCR amplification steps were performed. Libraries were evaluated by Qubit and
586 TapeStation. Sequencing libraries were constructed with barcodes to allow multiplexing of
587 the samples in one lane. On average, 27 million single-end 60 bp reads were sequenced per

588 sample on Illumina HiSeq 2500 V4 instrument. Raw reads were filtered and cleaned using
589 *Trimmomatic* (Bolger *et al.*, 2014) to remove adapters and the FASTX-Toolkit version
590 0.0.13.2 for (1) trimming read-end nucleotides with quality scores < 30 using
591 fastq_quality_trimmer and (2) removing reads with less than 70% base pairs with quality
592 score ≤ 30 using fastq_quality_filter. Reads were mapped with *bowtie2* (Langmead and
593 Salzberg, 2012) against the tomato coding sequences (SL2.5 release) with the parameters: --
594 no-mixed --no-discordant --gbar 1000 --end-to-end -k 20. The transcript quantification was
595 performed using the Expectation-Maximization method (RSEM) (Li and Dewey, 2011) based
596 on the script align_and_estimate_abundance.pl from Trinity software (Haas *et al.*, 2013) ;
597 https://github.com/trinityrnaseq/trinityrnaseq/wiki). Genes were annotated using the
598 annotation provided by the International Tomato Annotation Group (ITAG, release 2.40).
599 To quantitate the expression of TEs, we used both a unique and a multiple mapping strategy.
600 *Tophat2* (Kim *et al.*, 2013) was used to map the clean-reads onto the tomato genome
601 reference (SL2.5 release) with the option -g 1 for unique mapping and -g 200 that allows for
602 up to 200 reported alignments with the best alignment score, for multiple mapping. The
603 Bedtools version 2.25 multicov program including the -D option (that includes duplicate
604 reads) was used for calculation count table for each TE elements based on the annotations
605 generated by REPET (Flutre *et al.*, 2011). Differential expression analyses were done with
606 *DESeq2* version 1.16.1 (Love *et al.*, 2014) in the R environment. To define transcripts
607 differentially expressed, we used a significance cut-off of 0.01 and a 1.5-fold change relative
608 to wild-type. RNA-seq statistics are listed in Supplemental Table 2.
609
610 **Small RNA analysis**
611 *35S:amiR-SlDCL3* (Kravchik *et al.*, 2014a), *Slpol iv* and *Slpol v* (Gouil and Baulcombe, 2016)
612 sRNA-seq data are available from the SRA database (accession SRP032929 for *35S:amiR-*
613 *SlDCL3* and SRP081115 for both *Slpol iv* and *Slpol v*). All mutants are in the same tomato
614 genetic background (cv. M82) as the *Slddm1* mutants described in this study. In addition,
615 small RNAs were all extracted from leaf samples.
616 *Slddm1* sRNA-seq libraries were prepared from two biological replicates per genotype, using
617 the same RNA as those used for the RNA-seq. Sequences were trimmed and filtered with
618 Trim Galore! and small RNA reads were mapped to the tomato genome (SL2.50 release)
619 using *ShortStack* version 3.8.3 (Johnson *et al.*, 2016) with the option --mmap f (placement
620 guided by all mapped reads). To define siRNA clusters, we used *ShortStack* (distance
621 minimum between alignments: 75 bp and minimum coverage per cluster: 0.5 rpm) with the

622 option --mmap n (placement guided by uniquely mapping reads) to map the reads of 20 to 24-
623 nt corresponding to sRNA-seq data merged from all biological replicates (two *Slddm1a*
624 *Slddm1b* mutants and two wild type). siRNA counts were normalized to the total number of
625 mapped reads, and analyzed with *DESeq2* version 1.16.1 (Love *et al.*, 2014). To define
626 clusters differentially expressed, we used a significance cut-off of 0.01 and a two-fold change
627 relative to wild type.
628
629 **Accession numbers**
630 RNA-seq, BS-seq and sRNA-seq data are available from the ENA database under the
631 accession number PRJEB23761. *35S:amiR-SlDCL3*, *Slpol iv* and *Slpol v* sRNA-seq data are
632 available from the SRA database (accession SRP032929 for *35S:amiR-SlDCL3* and
633 SRP081115 for both *Slpol iv* and *Slpol v*). Sequence data from this article can be found in the
634 Sol Genomics Network (https://solgenomics.net/) under the following accession numbers:
635 Solyc02g062780, *SlDDM1a*; Solyc02g085390, *SlDDM1b*; Solyc08g080210, *SlPol IV* and
636 Solyc01g096390, *SlPol V*. The TE *gff* file is available on the Sol Genomics Network.
637
638
639 **SUPPLEMENTAL DATA**
640 **Supplemental Figure 1.** The tomato *DDM1* genes
641 **Supplemental Figure 2.** Effect of *SlDDM1* knockout on pollen quality
642 **Supplemental Figure 3.** Correlation between BS-sequencing biological replicates
643 **Supplemental Figure 4.** Patterns of methylation in TEs of the *Slddm1* mutants
644 **Supplemental Figure 5.** DMRs identified between the *Slddm1a Slddm1b* mutant and
645 the wild type
646 **Supplemental Figure 6.** Methylation and siRNA profiles of CG and CHG hyperDMRs
647 localised in repeat-poor regions
648 **Supplemental Figure 7**. Methylation profiles of euchromatic TEs
649 **Supplemental Figure 8.** Principal component analysis (PCA) plots of Small-RNA and
650 RNA-seq data
651 **Supplemental Figure 9.** 24-nt siRNA patterns of *Slddm1a Slddm1b* mutants
652 normalized against miRNAs
653 **Supplemental Figure 10.** siRNA patterns of *Slddm1a Slddm1b* mutants along a tomato
654 chromosome
655 **Supplemental Table 1.** Whole-genome bisulfite sequencing statistics

Supplemental Table 2. RNA-seq statistics

Supplemental Table 3. Primers used in this study

Supplemental Dataset 1. hypoDMR between the wild type and the *Slddm1a Slddm1b* mutant

Supplemental Dataset 2. hyperDMR between the wild type and the *Slddm1a Slddm1b* mutant

Supplemental Dataset 3. Regions corresponding to the compartmentalization of the tomato genome in three major regions

Supplemental Dataset 4. Genes deregulated in the *Slddm1a Slddm1b* mutant

ACKNOWLEDGEMENTS

This work was supported by a grant from the Chief Scientist of the Israel Ministry of Agriculture and Rural Development no. 20-10-0039 to T.A. and by project MemoCROP France-Israel joint grant 33583WA to N.B. and T.A. We thank Michal Liberman Lazarovich and Assaf Zemach for critical reading of the manuscript, Filipe Borges and Christine Mézard for the helpful discussions. The Institut Jean-Pierre Bourgin benefits from the support of the LabEx Saclay Plant Sciences-SPS (Project 10-LABX-0040-SPS).

AUTHOR CONTRIBUTIONS

SC, TA and NB designed the experiments. SC, TA and NB performed the experiments. SC, ADF, OJ, FM, TA and NB analyzed the data. SC, TA and NB wrote the paper.

FIGURE LEGENDS

Figure 1. CRISPR/Cas9 knockout of *SlDDM1a* and *SlDDM1b* genes
(A) Schematic illustrations of *SlDDM1a* and *SlDDM1b* gene regions targeted by the guide RNA. Black bars indicate exons and lines introns. The guide RNA-target sequences are shown, PAM sequences are colored in red, black triangles mark the predicted Cas9 cut site, restriction enzyme recognition sites are underlined, black half arrows indicate the location of primers used for mutant genotyping by PCR.
(B) Sequences of isolated mutant *Slddm1a* and *Slddm1b* alleles aligned to their respective wild types. The guide RNA target sequence is highlighted in red.
(C) Scheme of SlDDM1 proteins and predicted mutant protein sequences aligned to their respective wild types below. The DNA-binding domain (green), SNF2 family N-terminal domain (blue), DEAD-like helicases superfamily domain (yellow) and helicase superfamily C-terminal domain (orange) are indicated. Arrowheads indicate the site of mutation.
(D) Representative 25-days-post-germination seedlings of indicated genotypes, all segregated from the same double heterozygote parent plant. Scale bar = 2 cm.

Figure 2. Phenotypes of *Slddm1a Slddm1b* mutant plants

(A) Representative vegetative organs of wild type and *Slddm1a Slddm1b*. Left to right: adaxial and abaxial sides of cotyledons (scale bar = 2 mm), mature expanded leaf (scale bar = 5 cm), terminal leaflet (scale bar = 4 mm), and a whole plant (scale bar = 20 cm).
(B) Representative floral organs of wild-type and *Slddm1a Slddm1b* flowers. Left to right: mature bud (scale bar = 2 cm), an inflorescence (scale bar = 1 cm), flower and ovary at anthesis (scale bars = 2 mm) and a manual cross section of an ovary (scale bars = 0.4 mm).
(C) Representative wild-type and *Slddm1a Slddm1b* red fruits and their respective cross sections (scale bars = 1 cm).

Figure 3. Patterns of methylation in tomato *Slddm1* mutants
(A) Boxplots showing mean methylation content of the *Slddm1* mutants and the corresponding wild type (WT). The tomato genome (SL2.5 release) was partitioned in 1 kb-tiles and methylation levels correspond to the ratios of methylated cytosines over the total number of cytosines. Only cytosines covered by at least five reads were considered. The average methylation levels were determined by combining the two biological replicates for each genotype.
(B) Pairwise comparison of methylation in the wild type (WT) and the *Slddm1a Slddm1b* mutant. Each dot represents a 1-kb window and their methylation levels were determined as in (A). The color scale measures the density of points (red being very dense). The Pearson correlation coefficients between the samples are 0.29 for mCG, 0.55 for mCHG and 0.7 for mCHH.
(C) Patterns of methylation in genes and TEs of the *Slddm1* mutants. The average methylation levels of genes (upper panels) and TEs (lower panels) were determined by dividing the corresponding annotated regions (ITAG 2.40 release) into 100 bp bins. Regions located 1 kb upstream and 1 kb downstream of the gene bodies and TEs are also presented in these metaplot analyses. TE-Genes (see Methods) were removed to perform the analysis.

Figure 4. Number, localization and nature of the Differentially Methylated Regions (DMRs) identified in the tomato *Slddm1* mutants
(A) Total number of DMRs found in the three methylation contexts (CG, CHG and CHH). Hypo- and hypermethylated DMRs are shown. DMRs associated with Chr00 were excluded from the count.
(B) Densities of DMRs between the *Slddm1a Slddm1b* mutant and the wild type, for chromosomes 3 (SL2.5 genome release). Chromosome 6 is shown in Supplemental Figure 5A and all chromosomes are presented in Supplemental Figure 5B. The density of hyperDMRs is represented in black, and the density of hypoDMRs is represented in green. The numbers of TEs (*Gypsy* elements) and genes contained within bins of 100 kb are plotted on histograms. The scale is 0 to 100 elements for TEs and 0 to 30 for genes. Regions were considered to be differentially methylated when the absolute differences of methylation were at least 10% for mCHH, 20% for mCHG or 30% for mCG, between the mutant and the wild type.
(C) Nature of the DMRs identified in the *Slddm1* mutants. *CDS+TE* corresponds to DMRs overlapping with both genes and TEs, *CDS* corresponds to DMRs overlapping with genes and *TE* corresponds to DMRs overlapping with TEs. All other remaining DMRs are classified as *Intergenic*. DMRs associated with chr00 were excluded from the count.
(D) Methylation levels of CG and CHG hypoDMRs. The average methylation levels of the DMRs were determined by dividing the corresponding regions into 100-bp bins. Regions located 1 kb upstream and 1 kb downstream of the DMRs are shown.

Figure 5: Methylation levels and siRNA contents of CHH DMRs

744 The average methylation levels and 24-nt siRNA contents of the DMRs were determined by
745 dividing the corresponding regions into 100 bp bins. Regions located 1 kb upstream and 1 kb
746 downstream of the DMRs are shown. CHH hypoDMRs localised within repeat-poor regions
747 (Supplemental Dataset 3) were divided in two groups: one that depends on RdDM (i.e.
748 corresponding to *Slddm1a Slddm1b*, *Slpol iv* and *Slpol v* overlapping hypoCHH DMRs), and
749 the other one that is RdDM-independent (i.e. corresponding to *Slddm1a Slddm1b* not
750 overlapping with *Slpol iv* or *Slpol v* hypoCHH DMRs). The results correspond to the mean
751 values obtained for the two biological repeats.
752
753 **Figure 6. 24-nt siRNA patterns of *Slddm1a Slddm1b* mutants**
754 (A) Distribution of 24-nt reads along chromosome 11 (chosen as a representative example of
755 all chromosomes). Both sense (blue) and antisense (red) normalized (RPM) reads were
756 plotted. The numbers of TEs (*Gypsy* elements) and genes contained within bins of 100 kb are
757 plotted on histograms. *Sldcl3*, *Slpol iv* and *Slpol v* sRNA-seq data were obtained from
758 previous reports (Kravchik *et al.*, 2014a, Gouil and Baulcombe, 2016). The 24-nt read content
759 of one genomic region comprising between 19 and 26 Mb and corresponding to a repeat-rich
760 region are shown below (reads from both strands were collapsed and the values correspond to
761 5 kb-bins).
762 (B) Average number of 24-nt reads contained within the 106 repeat-rich and 68 repeat-poor
763 regions defined in this study (Supplemental Dataset 3). The numbers of reads were
764 normalized to the total number of mapped reads. Error bars indicate SD (n=2 biological
765 repeats). Asterisks indicate significant differences (Mann-Whitney *U*-test, $p < 0.01$).
766 (C) MA plots for 24-nt siRNA clusters. Each dot represents a siRNA cluster that was
767 identified as significantly differentially expressed at a 1% FDR using the *DESeq2* R package.
768 The clusters differentially expressed by two-fold (log2FC(*Slddm1a Slddm1b*/WT) > 2 or < -2)
769 are in red and their numbers are indicated within each panel. Above the center horizontal line
770 are 24-nt clusters that are more expressed in *Slddm1a Slddm1b* compared to the wild type and
771 below are the ones that are less expressed.
772 (D) Patterns of CHH methylation and 24-nt siRNAs for regions surrounding the transcription
773 start and stop of genes in the wild type and *Slddm1a Slddm1b*. The average mCHH levels and
774 normalized 24-nt siRNA contents were determined by dividing the corresponding regions,
775 into 100 bp bins. Regions located 1 kb upstream and 1 kb downstream of the gene bodies and
776 TEs are shown. TE-Genes (see Methods) were removed to perform the analysis. The results
777 correspond to the mean values obtained for the two biological repeats.
778
779 **Figure 7. siRNA content in repeat-rich and repeat-poor regions of the *Slddm1a Slddm1b***
780 **mutant and the wild type.**
781 Boxplots showing average numbers of 21 to 24-nt reads contained within repeat-rich and
782 repeat-poor regions (*n*) having a significant number of reads (DESeq2 cut-off of 0.01). The
783 numbers of reads were normalized to the total number of mapped reads. Asterisks indicate
784 significant differences and NS nonsignificant differences (Mann-Whitney *U*-test, $p < 0.01$).
785
786 **Figure 8. Methylation patterns and gene expression are correlated in**
787 ***Slddm1a Slddm1b***
788 (A) Number of genes that are significantly (FDR≤0.01) downregulated (log2FC(*Slddm1a*
789 *Slddm1b*/WT) < -1.5) or upregulated (log2FC(*Slddm1a Slddm1b*/WT) > 1.5) and overlapping
790 a DMR in either one of the three contexts (CG, CHG and CHH) are indicated. For the
791 upregulated genes corresponding to CG or CHG hypoDMRs, the proportion of DMRs
792 overlapping genes (*CDS*) or both genes and TEs (*CDS+TE*) is shown.
793 (B) Overlap between hypoDMR localised within the CDS of upregulated genes.

794 (C) Localisation of derepressed TEs. RP, repeat-poor regions; INT, regions containing an
795 intermediate number of repeats; RR, repeat-rich regions.

Table 1: TEs are derepressed in the *Slddm1a Slddm1b* mutant

The number of bases covered genome-wide, by a specific TE family, is indicated, as well as the corresponding proportions compared to all TEs. The average size and the numbers of TEs annotated with REPET (Flutre *et al.*, 2011) are indicated. The results of the two mapping strategies used are shown (see Methods). The number (and %) of TEs that were significantly (FDR<0.01) downregulated (log2FC(*ddm1*/WT) < -1.5) in the *Slddm1a Slddm1b* mutant, compared to the wild type, and upregulated (log2FC(*ddm1*/WT) > 1.5) are shown. *Others* corresponds to TEs that were predicted by REPET, but not classified within the *Gypsy*, *Copia*, DNA elements, MITE or Line families.

TE/repeat type family	Number of base covered	Fraction covered (%)	Mean element size (bp)	Number of annotated TEs	**Unique-mapping strategy**		**Multiple-mapping strategy**	
					TE down	TE up	TE down	TE up
Gypsy	247495673	46.2	1394	177553	40 (0.02%)	4688 (2.6%)	2186 (1.2%)	6047 (3.4%)
Copia	75592636	14.1	1175	64329	33 (0.05%)	1472 (2.3%)	333 (0.5%)	2229 (3.5%)
Line	25233169	4.7	656	38480	6 (0.02%)	401 (1%)	114 (0.3%)	1074 (2.8%)
DNA	20558297	3.8	777	26454	15 (0.06%)	1162 (4.4%)	277 (1%)	1744 (6.6%)
MITE	6758477	1.3	458	14744	4 (0.03%)	73 (0.5%)	54 (0.4%)	223 (1.5%)
Others	159808192	29.8	918	215083	53 (0.02%)	2506 (1.1%)	624 (0.3%)	4857 (2.3%)
TOTAL	535446444	100		536643	151 (0.03%)	10302 (1.9%)	3588 (0.7%)	16174 (3%)

REFERENCES

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, **13**, R87.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Wang, J., Rusch, D., Tang, H. and Pikaard, C.S. (2015) Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. *Elife*, **4**, e09591.

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

Borges, F., Parent, J.S., van Ex, F., Wolff, P., Martinez, G., Kohler, C. and Martienssen, R.A. (2018) Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. *Nat Genet*, **50**, 186-192.

Brzeski, J. and Jerzmanowski, A. (2003) Deficient in DNA methylation 1 (DDM1) defines a novel family of chromatin-remodeling factors. *J Biol Chem*, **278**, 823-828.

Cao, X. and Jacobsen, S.E. (2002) Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr Biol*, **12**, 1138-1144.

Cortijo, S., Wardenaar, R., Colome-Tatche, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J.M., Wincker, P., Roudier, F., Jansen, R.C., Colot, V. and Johannes, F. (2014) Mapping the epigenetic basis of complex traits. *Science*, **343**, 1145-1148.

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C. and Martienssen, R.A. (2014) miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature*, **508**, 411-415.

Dennis, K., Fan, T., Geiman, T., Yan, Q. and Muegge, K. (2001) Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes Dev*, **15**, 2940-2944.

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

Gallusci, P., Hodgman, C., Teyssier, E. and Seymour, G.B. (2016) DNA Methylation and Chromatin Regulation during Fleshy Fruit Development and Ripening. *Front Plant Sci*, **7**, 807.

Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X. and Dawe, R.K. (2013) CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res*, **23**, 628-637.

Gouil, Q. and Baulcombe, D.C. (2016) DNA Methylation Signatures of the Plant Chromomethyltransferases. *PLoS Genet*, **12**, e1006526.

Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L. and Pikaard, C.S. (2012) In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol Cell*, **48**, 811-818.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B. and Lieber, M. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**, 1494-1512.

Hansen, K.D., Langmead, B. and Irizarry, R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, **13**, R83.

855 **Herr, A.J., Jensen, M.B., Dalmay, T. and Baulcombe, D.C.** (2005) RNA polymerase IV directs
856 silencing of endogenous DNA. *Science*, **308**, 118-120.
857 **Ito, T., Tarutani, Y., To, T.K., Kassam, M., Duvernois-Berthet, E., Cortijo, S., Takashima, K.,**
858 **Saze, H., Toyoda, A., Fujiyama, A., Colot, V. and Kakutani, T.** (2015) Genome-wide
859 negative feedback drives transgenerational DNA methylation dynamics in
860 Arabidopsis. *PLoS Genet*, **11**, e1005154.
861 **Jia, Y., Lisch, D.R., Ohtsu, K., Scanlon, M.J., Nettleton, D. and Schnable, P.S.** (2009) Loss of
862 RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and
863 unexpected changes in the expression of transposons, genes, and 24-nt small RNAs.
864 *PLoS Genet*, **5**, e1000737.
865 **Johnson, N.R., Yeoh, J.M., Coruh, C. and Axtell, M.J.** (2016) Improved Placement of Multi-
866 mapping Small RNAs. *G3 (Bethesda)*, **6**, 2103-2111.
867 **Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. and Maumus, F.** (2016) Comprehensive
868 repeatome annotation reveals strong potential impact of repetitive elements on
869 tomato ripening. *BMC Genomics*, **17**, 624.
870 **Kakutani, T.** (1997) Genetic characterization of late-flowering traits induced by DNA
871 hypomethylation mutation in Arabidopsis thaliana. *Plant J*, **12**, 1447-1451.
872 **Kakutani, T., Jeddeloh, J.A., Flowers, S.K., Munakata, K. and Richards, E.J.** (1996)
873 Developmental abnormalities and epimutations associated with DNA
874 hypomethylation mutations. *Proc Natl Acad Sci U S A*, **93**, 12406-12411.
875 **Kakutani, T., Jeddeloh, J.A. and Richards, E.J.** (1995) Characterization of an Arabidopsis
876 thaliana DNA hypomethylation mutant. *Nucleic Acids Res*, **23**, 130-137.
877 **Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A. and**
878 **Carrington, J.C.** (2007) Genome-Wide Profiling and Analysis of Arabidopsis siRNAs.
879 *PLoS Biol*, **5**, e57.
880 **Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L.** (2013) TopHat2:
881 accurate alignment of transcriptomes in the presence of insertions, deletions and
882 gene fusions. *Genome Biol*, **14**, R36.
883 **Kravchik, M., Damodharan, S., Stav, R. and Arazi, T.** (2014a) Generation and
884 characterization of a tomato DCL3-silencing mutant. *Plant Sci*, **221-222**, 81-89.
885 **Kravchik, M., Sunkar, R., Damodharan, S., Stav, R., Zohar, M., Isaacson, T. and Arazi, T.**
886 (2014b) Global and local perturbation of the tomato microRNA pathway by a trans-
887 activated DICER-LIKE 1 mutant. *J Exp Bot*, **65**, 725-739.
888 **Krueger, F. and Andrews, S.R.** (2011) Bismark: a flexible aligner and methylation caller for
889 Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572.
890 **Lang, Z., Wang, Y., Tang, K., Tang, D., Datsenka, T., Cheng, J., Zhang, Y., Handa, A.K. and**
891 **Zhu, J.K.** (2017) Critical roles of DNA demethylation in the activation of ripening-
892 induced genes and inhibition of ripening-repressed genes in tomato fruit. *Proc Natl*
893 *Acad Sci U S A*, **114**, E4511-e4519.
894 **Langmead, B. and Salzberg, S.L.** (2012) Fast gapped-read alignment with Bowtie 2. *Nat*
895 *Methods*, **9**, 357-359.
896 **Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M., Strahl, B.D., Patel, D.J.**
897 **and Jacobsen, S.E.** (2013) Polymerase IV occupancy at RNA-directed DNA
898 methylation sites requires SHH1. *Nature*, **498**, 385-389.
899 **Law, J.A., Vashisht, A.A., Wohlschlegel, J.A. and Jacobsen, S.E.** (2011) SHH1, a
900 homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and

chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet*, **7**, e1002195.

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

Li, J.F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M. and Sheen, J. (2013) Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat Biotechnol*, **31**, 688-691.

Li, Q., Eichten, S.R., Hermanson, P.J., Zaunbrecher, V.M., Song, J., Wendt, J., Rosenbaum, H., Madzima, T.F., Sloan, A.E., Huang, J., Burgess, D.L., Richmond, T.A., McGinnis, K.M., Meeley, R.B., Danilevskaya, O.N., Vaughn, M.W., Kaeppler, S.M., Jeddeloh, J.A. and Springer, N.M. (2014) Genetic perturbation of the maize methylome. *Plant Cell*, **26**, 4602-4616.

Li, Q., Gent, J.I., Zynda, G., Song, J., Makarevitch, I., Hirsch, C.D., Hirsch, C.N., Dawe, R.K., Madzima, T.F., McGinnis, K.M., Lisch, D., Schmitz, R.J., Vaughn, M.W. and Springer, N.M. (2015a) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc Natl Acad Sci U S A*, **112**, 14728-14733.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Zheng, B., Gregory, B.D. and Chen, X. (2015b) Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. *Genome Res*, **25**, 235-245.

Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J.C., Doerge, R.W., Colot, V. and Martienssen, R. (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**, 471-476.

Liu, C., Cheng, Y.J., Wang, J.W. and Weigel, D. (2017) Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants*.

Liu, R., How-Kit, A., Stammitti, L., Teyssier, E., Rolin, D., Mortain-Bertrand, A., Halle, S., Liu, M., Kong, J., Wu, C., Degraeve-Guibault, C., Chapman, N.H., Maucourt, M., Hodgman, T.C., Tost, J., Bouzayen, M., Hong, Y., Seymour, G.B., Giovannoni, J.J. and Gallusci, P. (2015) A DEMETER-like DNA demethylase governs tomato fruit ripening. *Proc Natl Acad Sci U S A*, **112**, 10804-10809.

Liu, Z.W., Shao, C.R., Zhang, C.J., Zhou, J.X., Zhang, S.W., Li, L., Chen, S., Huang, H.W., Cai, T. and He, X.J. (2014) The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. *PLoS Genet*, **10**, e1003948.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

Lyons, D.B. and Zilberman, D. (2017) DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. *Elife*, **6**.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**.

Matzke, M.A., Kanno, T. and Matzke, A.J. (2015) RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. *Annu Rev Plant Biol*, **66**, 243-267.

946 **McCue, A.D., Nuthikattu, S., Reeder, S.H. and Slotkin, R.K.** (2012) Gene expression and
947 stress response mediated by the epigenetic regulation of a transposable element
948 small RNA. *PLoS Genet*, **8**, e1002474.
949 **McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N. and Slotkin, R.K.**
950 (2015) ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the
951 establishment of DNA methylation. *Embo j*, **34**, 20-35.
952 **Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H. and Kakutani, T.** (2001)
953 Mobilization of transposons by a mutation abolishing full DNA methylation in
954 Arabidopsis. *Nature*, **411**, 212-214.
955 **Mosher, R.A., Schwach, F., Studholme, D. and Baulcombe, D.C.** (2008) PolIVb influences
956 RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc*
957 *Natl Acad Sci U S A*, **105**, 3145-3150.
958 **Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N. and Slotkin, R.K.**
959 (2013) The initiation of epigenetic silencing of active transposable elements is
960 triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiol*, **162**,
961 116-131.
962 **Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O. and Pikaard, C.S.** (2005) Plant
963 nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent
964 heterochromatin formation. *Cell*, **120**, 613-622.
965 **Panda, K., Ji, L., Neumann, D.A., Daron, J., Schmitz, R.J. and Slotkin, R.K.** (2016) Full-length
966 autonomous transposable elements are preferentially targeted by expression-
967 dependent forms of RNA-directed DNA methylation. *Genome Biol*, **17**, 170.
968 **Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., Yang,**
969 **Y., Wu, Z., Mao, L., Wu, H., Ling-Hu, C., Zhou, H., Lin, H., Gonzalez-Morales, S.,**
970 **Trejo-Saavedra, D.L., Tian, H., Tang, X., Zhao, M., Huang, Z., Zhou, A., Yao, X., Cui, J.,**
971 **Li, W., Chen, Z., Feng, Y., Niu, Y., Bi, S., Yang, X., Cai, H., Luo, X., Montes-Hernandez,**
972 **S., Leyva-Gonzalez, M.A., Xiong, Z., He, X., Bai, L., Tan, S., Liu, D., Liu, J., Zhang, S.,**
973 **Chen, M., Zhang, L., Zhang, Y., Liao, W., Wang, M., Lv, X., Wen, B., Liu, H., Luan, H.,**
974 **Yang, S., Wang, X., Xu, J., Li, X., Li, S., Wang, J., Palloix, A., Bosland, P.W., Li, Y.,**
975 **Krogh, A., Rivera-Bustamante, R.F., Herrera-Estrella, L., Yin, Y., Yu, J., Hu, K. and**
976 **Zhang, Z.** (2014) Whole-genome sequencing of cultivated and wild peppers provides
977 insights into Capsicum domestication and specialization. *Proc Natl Acad Sci U S A*,
978 **111**, 5135-5140.
979 **Saze, H. and Kakutani, T.** (2007) Heritable epigenetic mutation of a transposon-flanked
980 Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *Embo j*, **26**,
981 3641-3652.
982 **Singer, T., Yordan, C. and Martienssen, R.A.** (2001) Robertson's Mutator transposons in A.
983 thaliana are regulated by the chromatin-remodeling gene Decrease in DNA
984 Methylation (DDM1). *Genes Dev*, **15**, 591-602.
985 **Stroud, H., Greenberg, M.V., Feng, S., Bernatavichute, Y.V. and Jacobsen, S.E.** (2013)
986 Comprehensive analysis of silencing mutants reveals complex regulation of the
987 Arabidopsis methylome. *Cell*, **152**, 352-364.
988 **Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J. and Jacobsen, S.E.**
989 (2014) Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis.
990 *Nat Struct Mol Biol*, **21**, 64-72.
991 **The Tomato Genome Consortium.** (2012) The tomato genome sequence provides insights
992 into fleshy fruit evolution. *Nature*, **485**, 635-641.

Tan, F., Zhou, C., Zhou, Q., Zhou, S., Yang, W., Zhao, Y., Li, G. and Zhou, D.X. (2016) Analysis of Chromatin Regulators Reveals Specific Features of Rice DNA Methylation Pathways. *Plant Physiol*, **171**, 2041-2054.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, **461**, 423-426.

Vongs, A., Kakutani, T., Martienssen, R.A. and Richards, E.J. (1993) Arabidopsis thaliana DNA methylation mutants. *Science*, **260**, 1926-1928.

Waibel, F. and Filipowicz, W. (1990) U6 snRNA genes of Arabidopsis are transcribed by RNA polymerase III but contain the same two upstream promoter elements as RNA polymerase II-transcribed U-snRNA genes. *Nucleic Acids Res*, **18**, 3451-3458.

Wierzbicki, A.T., Haag, J.R. and Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, **135**, 635-648.

Wierzbicki, A.T., Ream, T.S., Haag, J.R. and Pikaard, C.S. (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat Genet*, **41**, 630-634.

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P. and Conneely, K.N. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res*.

Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L. and Zilberman, D. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, **153**, 193-205.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., Liu, W., Henderson, I.R., Meyers, B.C., Ausin, I. and Jacobsen, S.E. (2015) A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell*, **163**, 445-455.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J. and Jacobsen, S.E. (2007) Role of RNA polymerase IV in plant small RNA metabolism. *Proc Natl Acad Sci U S A*, **104**, 4536-4541.

Zhong, S., Fei, Z., Chen, Y.R., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang, J., Shao, Y. and Giovannoni, J.J. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat Biotechnol*, **31**, 154-159.

Zhong, X., Du, J., Hale, C.J., Gallego-Bartolome, J., Feng, S., Vashisht, A.A., Chory, J., Wohlschlegel, J.A., Patel, D.J. and Jacobsen, S.E. (2014) Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell*, **157**, 1050-1060.

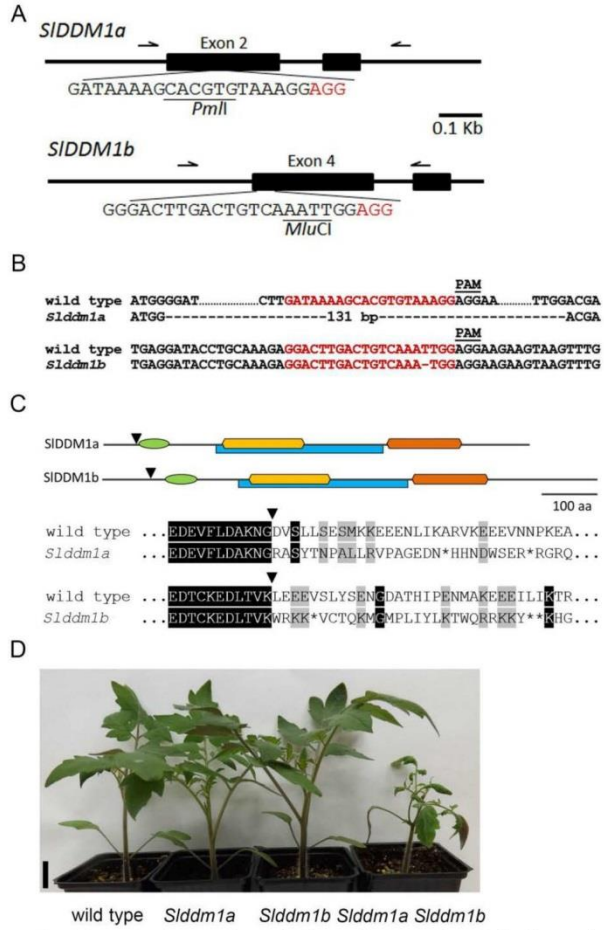


Figure 1. CRISPR/Cas9 knockout of *SlDDM1a* and *SlDDM1b* genes

(A) Schematic illustrations of *SlDDM1a* and *SlDDM1b* gene regions targeted by the guide RNA. Black bars indicate exons and lines introns. The guide RNA-target sequences are shown, PAM sequences are colored in red, black triangles mark the predicted Cas9 cut site, restriction enzyme recognition sites are underlined, black half arrows indicate the location of primers used for mutant genotyping by PCR.

(B) Sequences of isolated mutant *Slddm1a* and *Slddm1b* alleles aligned to their respective wild types. The guide RNA target sequence is highlighted in red.

(C) Scheme of SlDDM1 proteins and predicted mutant protein sequences aligned to their respective wild types below. The DNA-binding domain (green), SNF2 family N-terminal domain (blue), DEAD-like helicases superfamily domain (yellow) and helicase superfamily C-terminal domain (orange) are indicated. Arrowheads indicate the site of mutation.

(D) Representative 25-days-post-germination seedlings of indicated genotypes, all segregated from the same double heterozygote parent plant. Scale bar = 2 cm.

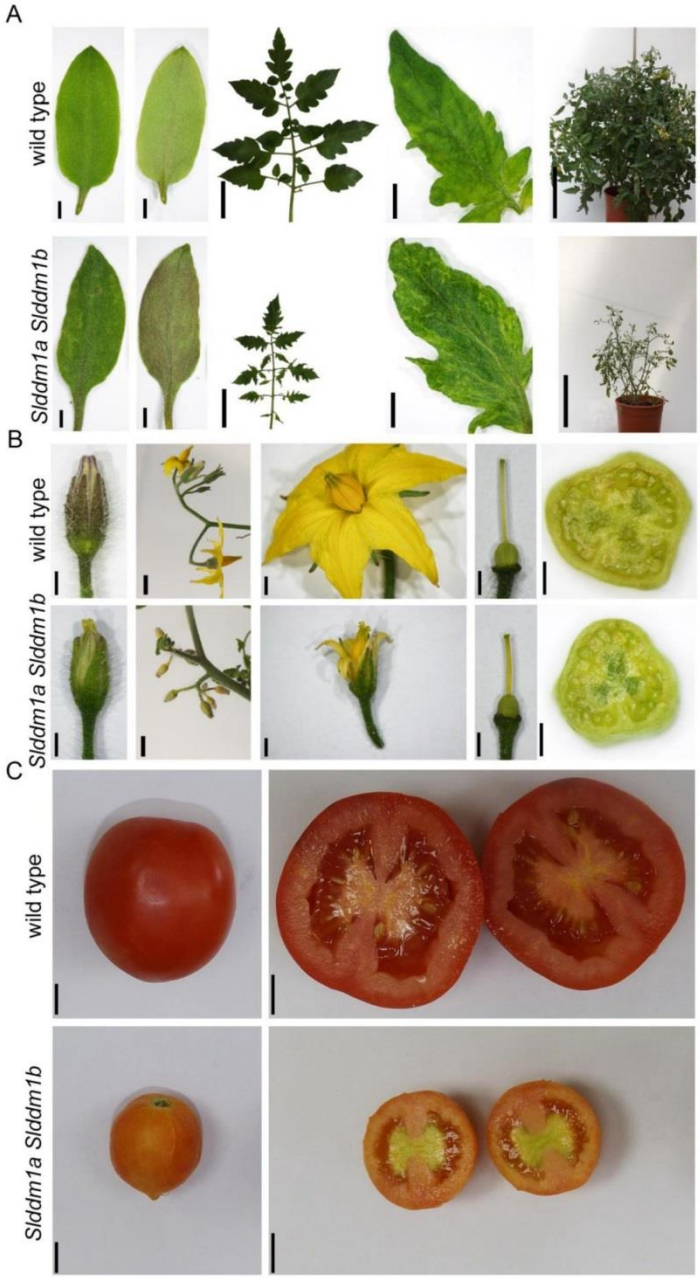


Figure 2. Phenotypes of *Slddm1a Slddm1b* mutant plants

(A) Representative vegetative organs of wild type and *Slddm1a Slddm1b*. Left to right: adaxial and abaxial sides of cotyledons (scale bar = 2 mm), mature expanded leaf (scale bar = 5 cm), terminal leaflet (scale bar = 4 mm), and a whole plant (scale bar = 20 cm).

(B) Representative floral organs of wild-type and *Slddm1a Slddm1b* flowers. Left to right: mature bud (scale bar = 2 cm), an inflorescence (scale bar = 1 cm), flower and ovary at anthesis (scale bars = 2 mm) and a manual cross section of an ovary (scale bars = 0.4 mm).

(C) Representative wild-type and *Slddm1a Slddm1b* red fruits and their respective cross sections (scale bars = 1 cm).

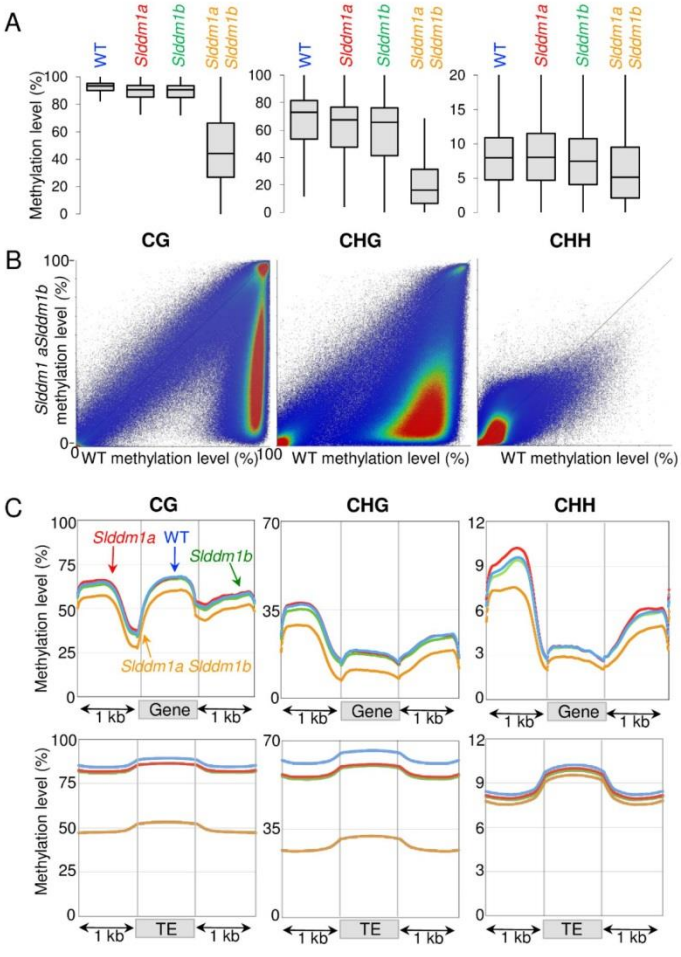


Figure 3. Patterns of methylation in tomato *Slddm1* mutants

(A) Boxplots showing mean methylation content of the *Slddm1* mutants and the corresponding wild type (WT). The tomato genome (SL2.5 release) was partitioned in 1 kb-tiles and methylation levels correspond to the ratios of methylated cytosines over the total number of cytosines. Only cytosines covered by at least five reads were considered. The average methylation levels were determined by combining the two biological replicates for each genotype.

(B) Pairwise comparison of methylation in the wild type (WT) and the *Slddm1a Slddm1b* mutant. Each dot represents a 1-kb window and their methylation levels were determined as in (A). The color scale measures the density of points (red being very dense). The Pearson correlation coefficients between the samples are 0.29 for mCG, 0.55 for mCHG and 0.7 for mCHH.

(C) Patterns of methylation in genes and TEs of the *Slddm1* mutants. The average methylation levels of genes (upper panels) and TEs (lower panels) were determined by dividing the corresponding annotated regions (ITAG 2.40 release) into 100 bp bins. Regions located 1 kb upstream and 1 kb downstream of the gene bodies and TEs are also presented in these metaplot analyses. TE-Genes (see Methods) were removed to perform the analysis.

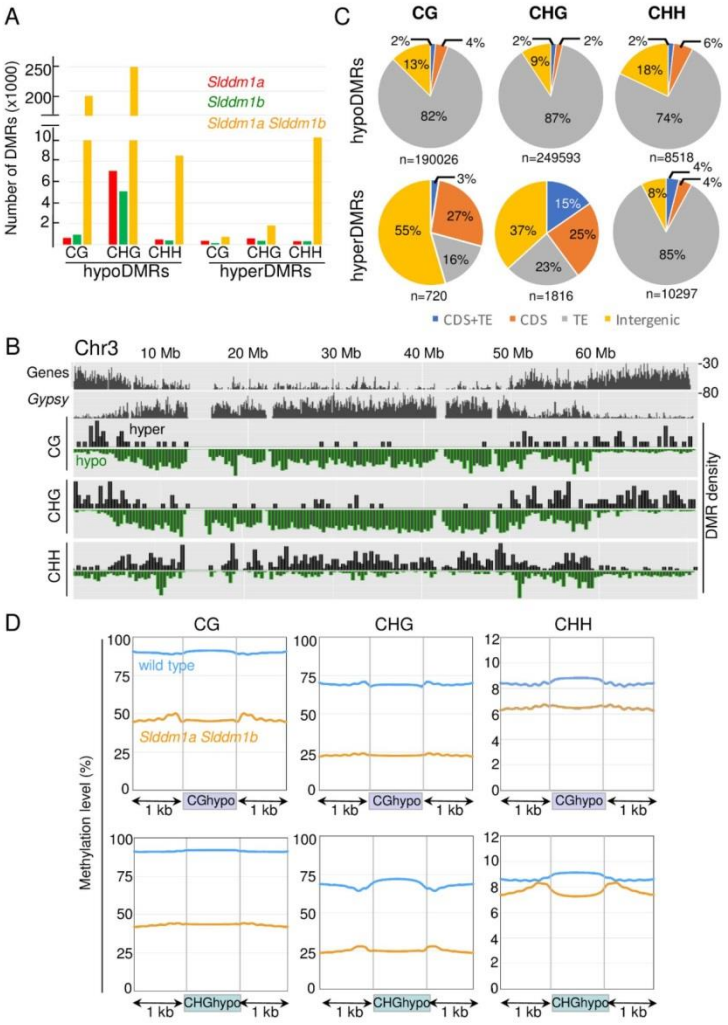


Figure 4. Number, localization and nature of the Differentially Methylated Regions (DMRs) identified in the tomato *Slddm1* mutants

(A) Total number of DMRs found in the three methylation contexts (CG, CHG and CHH). Hypo- and hypermethylated DMRs are shown. DMRs associated with Chr00 were excluded from the count.

(B) Densities of DMRs between the *Slddm1a Slddm1b* mutant and the wild type, for chromosomes 3 (SL2.5 genome release). Chromosome 6 is shown in Supplemental Figure 5A and all chromosomes are presented in Supplemental Figure 5B. The density of hyperDMRs is represented in black, and the density of hypoDMRs is represented in green. The numbers of TEs (*Gypsy* elements) and genes contained within bins of 100 kb are plotted on histograms. The scale is 0 to 100 elements for TEs and 0 to 30 for genes. Regions were considered to be differentially methylated when the absolute differences of methylation were at least 10% for mCHH, 20% for mCHG or 30% for mCG, between the mutant and the wild type.

(C) Nature of the DMRs identified in the *Slddm1* mutants. *CDS+TE* corresponds to DMRs overlapping with both genes and TEs, *CDS* corresponds to DMRs overlapping with genes and *TE* corresponds to DMRs overlapping with TEs. All other remaining DMRs are classified as *Intergenic*. DMRs associated with chr00 were excluded from the count.

(D) Methylation levels of CG and CHG hypoDMRs. The average methylation levels of the DMRs were determined by dividing the corresponding regions into 100-bp bins. Regions located 1 kb upstream and 1 kb downstream of the DMRs are shown.

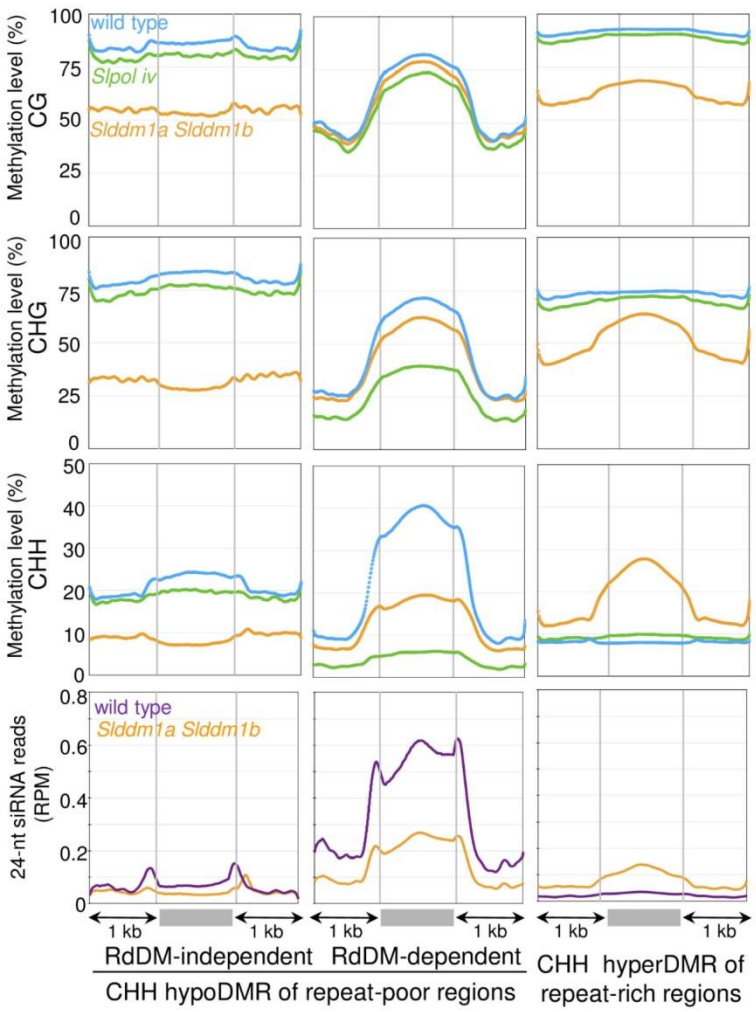


Figure 5: Methylation levels and siRNA contents of CHH DMRs

The average methylation levels and 24-nt siRNA contents of the DMRs were determined by dividing the corresponding regions into 100 bp bins. Regions located 1 kb upstream and 1 kb downstream of the DMRs are shown. CHH hypoDMRs localised within repeat-poor regions (Supplemental Dataset 3) were divided in two groups: one that depends on RdDM (i.e. corresponding to *Slddm1a Slddm1b*, *Slpol iv* and *Slpol v* overlapping hypoCHH DMRs), and the other one that is RdDM-independent (i.e. corresponding to *Slddm1a Slddm1b* not overlapping with *Slpol iv* or *Slpol v* hypoCHH DMRs). The results correspond to the mean values obtained for the two biological repeats.

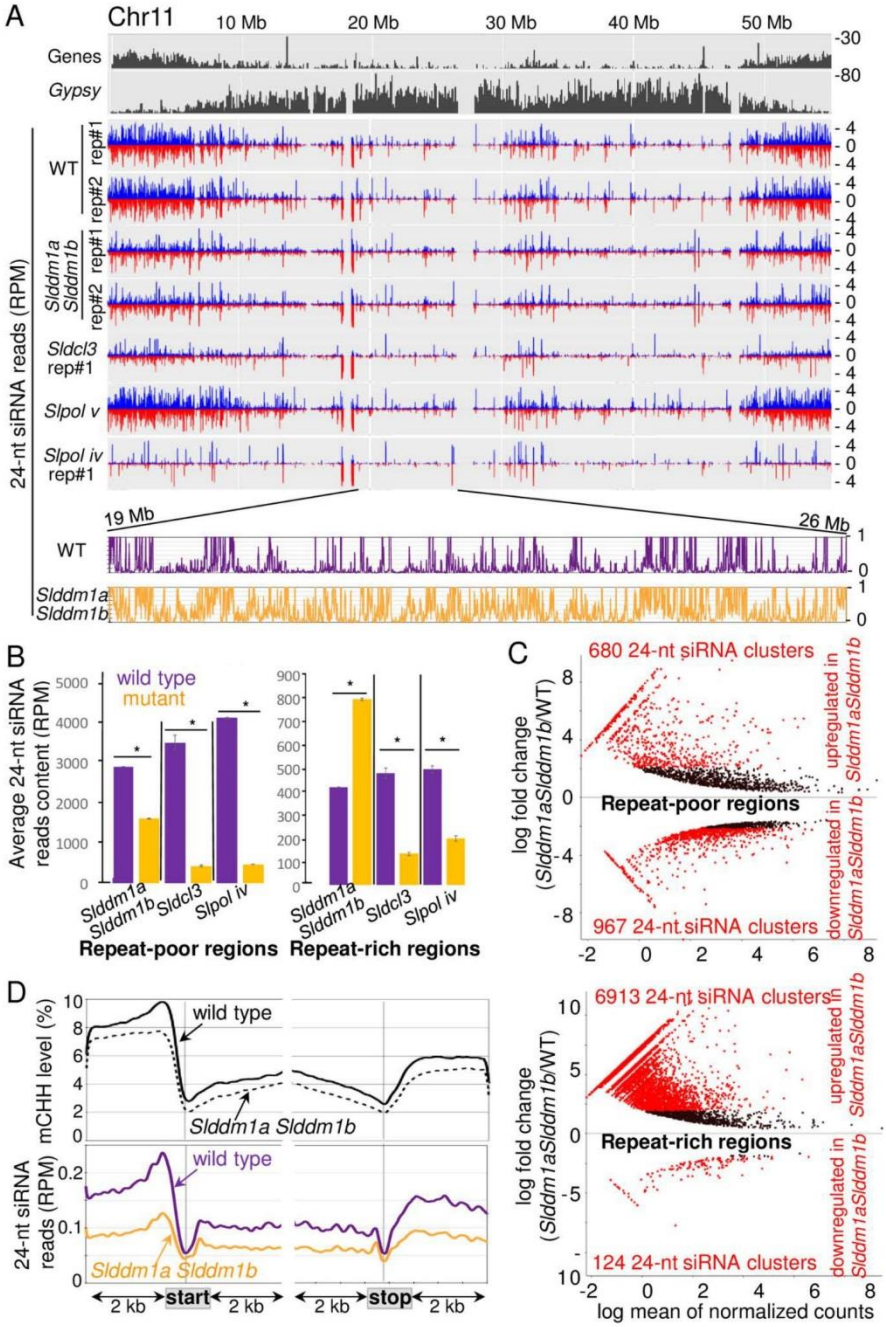

A
Chr11
10 Mb
20 Mb
30 Mb
40 Mb
50 Mb
Genes
Gypsy
WT
rep#1
rep#2
Slddm1a Slddm1b
Sldcl3 rep#1
Slpol v
Slpol iv rep#1
24-nt siRNA reads (RPM)
19 Mb
26 Mb
B
Average 24-nt siRNA reads content (RPM)
wild type
mutant
Slddm1a Slddm1b
Sldcl3
Slpol iv
Repeat-poor regions
Repeat-rich regions
C
680 24-nt siRNA clusters
upregulated in Slddm1aSlddm1b
Repeat-poor regions
downregulated in Slddm1aSlddm1b
967 24-nt siRNA clusters
log fold change (Slddm1aSlddm1b/WT)
6913 24-nt siRNA clusters
Repeat-rich regions
124 24-nt siRNA clusters
log mean of normalized counts
D
mCHH level (%)
wild type
Slddm1a Slddm1b
24-nt siRNA reads (RPM)
2 kb
start
stop

Figure 6. 24-nt siRNA patterns of *Slddm1a Slddm1b* mutants

(A) Distribution of 24-nt reads along chromosome 11 (chosen as a representative example of all chromosomes). Both sense (blue) and antisense (red) normalized (RPM) reads were plotted. The numbers of TEs (*Gypsy* elements) and genes contained within bins of 100 kb are plotted on histograms. *Sldcl3*, *Slpol iv* and *Slpol v* sRNA-seq data were obtained from previous reports (Kravchik *et al.*, 2014a, Gouil and Baulcombe, 2016). The 24-nt read content of one genomic region comprising between 19 and 26 Mb and corresponding to a repeat-rich region are shown below (reads from both strands were collapsed and the values correspond to 5 kb-bins).

(B) Average number of 24-nt reads contained within the 106 repeat-rich and 68 repeat-poor regions defined in this study (Supplemental Dataset 3). The numbers of reads were normalized to the total number of mapped reads. Error bars indicate SD (n=2 biological repeats). Asterisks indicate significant differences (Mann-Whitney *U*-test, $p < 0.01$).

(C) MA plots for 24-nt siRNA clusters. Each dot represents a siRNA cluster that was identified as significantly differentially expressed at a 1% FDR using the *DESeq2* R package. The clusters differentially expressed by two-fold (log2FC(*Slddm1a Slddm1b*/WT) > 2 or < -2) are in red and their numbers are indicated within each panel. Above the center horizontal line are 24-nt clusters that are more expressed in *Slddm1a Slddm1b* compared to the wild type and below are the ones that are less expressed.

(D) Patterns of CHH methylation and 24-nt siRNAs for regions surrounding the transcription start and stop of genes in the wild type and *Slddm1a Slddm1b*. The average mCHH levels and normalized 24-nt siRNA contents were determined by dividing the corresponding regions, into 100 bp bins. Regions located 1 kb upstream and 1 kb downstream of the gene bodies and TEs are shown. TE-Genes (see Methods) were removed to perform the analysis. The results correspond to the mean values obtained for the two biological repeats.

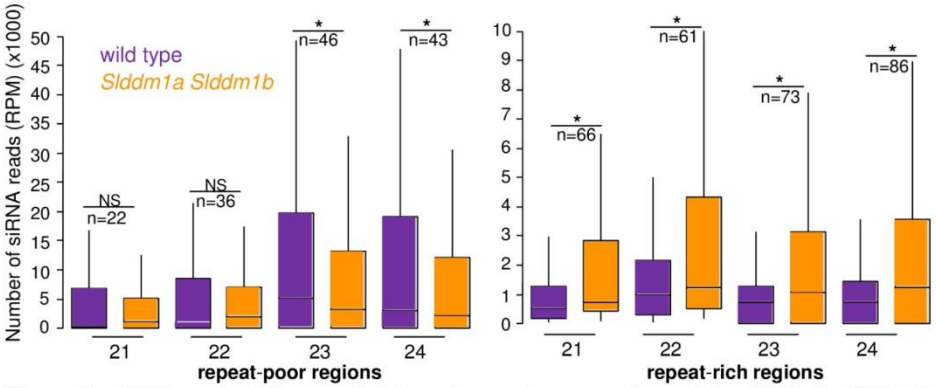


Figure 7. siRNA content in repeat-rich and repeat-poor regions of the *Slddm1a Slddm1b* mutant and the wild type.

Boxplots showing average numbers of 21 to 24-nt reads contained within repeat-rich and repeat-poor regions (*n*) having a significant number of reads (DESeq2 cut-off of 0.01). The numbers of reads were normalized to the total number of mapped reads. Asterisks indicate significant differences and NS nonsignificant differences (Mann-Whitney *U*-test, $p < 0.01$).

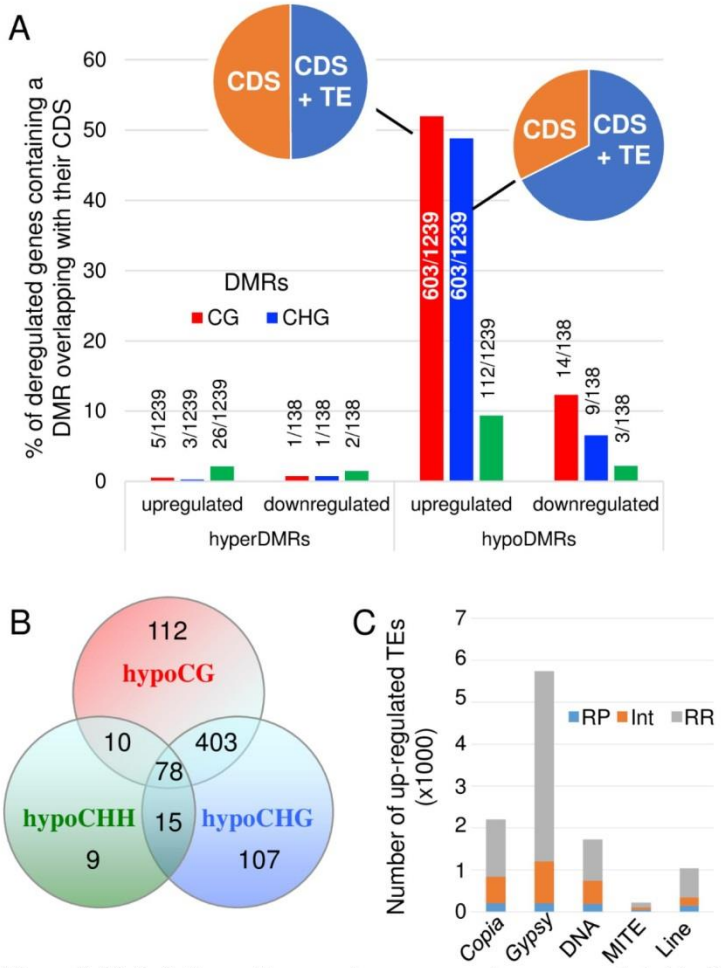


Figure 8. Methylation patterns and gene expression are correlated in *Slddm1a Slddm1b*

(A) Number of genes that are significantly (FDR≤0.01) downregulated (log2FC(*Slddm1a Slddm1b*/WT) < -1.5) or upregulated (log2FC(*Slddm1a Slddm1b*/WT) > 1.5) and overlapping a DMR in either one of the three contexts (CG, CHG and CHH) are indicated. For the upregulated genes corresponding to CG or CHG hypoDMRs, the proportion of DMRs overlapping genes (*CDS*) or both genes and TEs (*CDS+TE*) is shown.

(B) Overlap between hypoDMR localised within the CDS of upregulated genes.

(C) Localisation of derepressed TEs. RP, repeat-poor regions; INT, regions containing an intermediate number of repeats; RR, repeat-rich regions.

Parsed Citations

This work was supported by a grant from the Chief Scientist of the Israel Ministry of Agriculture and Rural Development no. 20-10-0039 to T.A. and by project MemoCROP France-Israel joint grant 33583WA to N.B. and T.A. We thank Michal Liberman Lazarovich and Assaf Zemach for critical reading of the manuscript, Filipe Borges and Christine Mézard for the helpful discussions. The Institut Jean-Pierre Bourgin benefits from the support of the LabEx Saclay Plant Sciences-SPS (Project 10-LABX-0040-SPS).

AUTHOR CONTRIBUTIONS

SC, TA and NB designed the experiments. SC, TA and NB performed the experiments. SC, ADF, OJ, FM, TA and NB analyzed the data. SC, TA and NB wrote the paper.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol, 13, R87.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Wang, J., Rusch, D., Tang, H. and Pikaard, C.S. (2015) Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. Elife, 4, e09591.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114-2120.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Borges, F., Parent, J.S., van Ex, F., Wolff, P., Martinez, G., Kohler, C. and Martienssen, R.A. (2018) Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. Nat Genet, 50, 186-192.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Brzeski, J. and Jerzmanowski, A. (2003) Deficient in DNA methylation 1 (DDM1) defines a novel family of chromatin-remodeling factors. J Biol Chem, 278, 823-828.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Cao, X. and Jacobsen, S.E. (2002) Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. Curr Biol, 12, 1138-1144.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Cortijo, S., Wardenaar, R., Colome-Tatche, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J.M., Wincker, P., Roudier, F., Jansen, R.C., Colot, V. and Johannes, F. (2014) Mapping the epigenetic basis of complex traits. Science, 343, 1145-1148.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C. and Martienssen, R.A. (2014) miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature, 508, 411-415.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Dennis, K., Fan, T., Geiman, T., Yan, Q. and Muegge, K. (2001) Lsh, a member of the SNF2 family, is required for genome-wide methylation. Genes Dev, 15, 2940-2944.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One, 6, e16526.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gallusci, P., Hodgman, C., Teyssier, E. and Seymour, G.B. (2016) DNA Methylation and Chromatin Regulation during Fleshy Fruit Development and Ripening. Front Plant Sci, 7, 807.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X. and Dawe, R.K. (2013) CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res, 23, 628-637.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gouil, Q. and Baulcombe, D.C. (2016) DNA Methylation Signatures of the Plant Chromomethyltransferases. PLoS Genet, 12, e1006526.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L. and Pikaard, C.S. (2012) In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. Mol Cell, 48, 811-818.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B. and Lieber, M. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8, 1494-1512.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hansen, K.D., Langmead, B. and Irizarry, R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol, 13, R83.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Herr, A.J., Jensen, M.B., Dalmay, T. and Baulcombe, D.C. (2005) RNA polymerase IV directs silencing of endogenous DNA. Science, 308, 118-120.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ito, T., Tarutani, Y., To, T.K., Kassam, M., Duvernois-Berthet, E., Cortijo, S., Takashima, K., Saze, H., Toyoda, A., Fujiyama, A., Colot, V. and Kakutani, T. (2015) Genome-wide negative feedback drives transgenerational DNA methylation dynamics in Arabidopsis. PLoS Genet, 11, e1005154.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jia, Y., Lisch, D.R., Ohtsu, K., Scanlon, M.J., Nettleton, D. and Schnable, P.S. (2009) Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. PLoS Genet, 5, e1000737.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Johnson, N.R., Yeoh, J.M., Coruh, C. and Axtell, M.J. (2016) Improved Placement of Multi-mapping Small RNAs. G3 (Bethesda), 6, 2103-2111.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. and Maumus, F. (2016) Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics, 17, 624.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kakutani, T. (1997) Genetic characterization of late-flowering traits induced by DNA hypomethylation mutation in Arabidopsis thaliana. Plant J, 12, 1447-1451.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kakutani, T., Jeddeloh, J.A., Flowers, S.K., Munakata, K. and Richards, E.J. (1996) Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. Proc Natl Acad Sci U S A, 93, 12406-12411.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kakutani, T., Jeddeloh, J.A. and Richards, E.J. (1995) Characterization of an Arabidopsis thaliana DNA hypomethylation mutant. Nucleic Acids Res, 23, 130-137.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A. and Carrington, J.C. (2007) Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. PLoS Biol, 5, e57.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 14, R36.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kravchik, M., Damodharan, S., Stav, R. and Arazi, T. (2014a) Generation and characterization of a tomato DCL3-silencing mutant. Plant Sci, 221-222, 81-89.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kravchik, M., Sunkar, R., Damodharan, S., Stav, R., Zohar, M., Isaacson, T. and Arazi, T. (2014b) Global and local perturbation of the

tomato microRNA pathway by a trans-activated DICER-LIKE 1 mutant. J Exp Bot, 65, 725-739.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 27, 1571-1572.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lang, Z., Wang, Y., Tang, K., Tang, D., Datsenka, T., Cheng, J., Zhang, Y., Handa, A.K. and Zhu, J.K. (2017) Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. Proc Natl Acad Sci U S A, 114, E4511-e4519.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods, 9, 357-359.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M., Strahl, B.D., Patel, D.J. and Jacobsen, S.E. (2013) Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature, 498, 385-389.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Law, J.A., Vashisht, A.A., Wohlschlegel, J.A. and Jacobsen, S.E. (2011) SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. PLoS Genet, 7, e1002195.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, J.F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M. and Sheen, J. (2013) Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. Nat Biotechnol, 31, 688-691.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, Q., Eichten, S.R., Hermanson, P.J., Zaunbrecher, V.M., Song, J., Wendt, J., Rosenbaum, H., Madzima, T.F., Sloan, A.E., Huang, J., Burgess, D.L., Richmond, T.A., McGinnis, K.M., Meeley, R.B., Danilevskaya, O.N., Vaughn, M.W., Kaeppler, S.M., Jeddeloh, J.A. and Springer, N.M. (2014) Genetic perturbation of the maize methylome. Plant Cell, 26, 4602-4616.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, Q., Gent, J.I., Zynda, G., Song, J., Makarevitch, I., Hirsch, C.D., Hirsch, C.N., Dawe, R.K., Madzima, T.F., McGinnis, K.M., Lisch, D., Schmitz, R.J., Vaughn, M.W. and Springer, N.M. (2015a) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A, 112, 14728-14733.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Zheng, B., Gregory, B.D. and Chen, X. (2015b) Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res, 25, 235-245.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J.C., Doerge, R.W., Colot, V. and Martienssen, R. (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature, 430, 471-476.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Liu, C., Cheng, Y.J., Wang, J.W. and Weigel, D. (2017) Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. Nat Plants.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Liu, R., How-Kit, A., Stammitti, L., Teyssier, E., Rolin, D., Mortain-Bertrand, A., Halle, S., Liu, M., Kong, J., Wu, C., Degraeve-Guibault, C., Chapman, N.H., Maucourt, M., Hodgman, T.C., Tost, J., Bouzayen, M., Hong, Y., Seymour, G.B., Giovannoni, J.J. and Gallusci, P. (2015) A DEMETER-like DNA demethylase governs tomato fruit ripening. Proc Natl Acad Sci U S A, 112, 10804-10809.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Liu, Z.W., Shao, C.R., Zhang, C.J., Zhou, J.X., Zhang, S.W., Li, L., Chen, S., Huang, H.W., Cai, T. and He, X.J. (2014) The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS Genet, 10, e1003948.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15, 550.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lyons, D.B. and Zilberman, D. (2017) DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. Elife, 6.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011, 17.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Matzke, M.A., Kanno, T. and Matzke, A.J. (2015) RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol, 66, 243-267.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

McCue, A.D., Nuthikattu, S., Reeder, S.H. and Slotkin, R.K. (2012) Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genet, 8, e1002474.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N. and Slotkin, R.K. (2015) ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. Embo j, 34, 20-35.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H. and Kakutani, T. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature, 411, 212-214.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mosher, R.A., Schwach, F., Studholme, D. and Baulcombe, D.C. (2008) PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. Proc Natl Acad Sci U S A, 105, 3145-3150.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N. and Slotkin, R.K. (2013) The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Physiol, 162, 116-131.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O. and Pikaard, C.S. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell, 120, 613-622.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Panda, K., Ji, L., Neumann, D.A., Daron, J., Schmitz, R.J. and Slotkin, R.K. (2016) Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. Genome Biol, 17, 170.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., Yang, Y., Wu, Z., Mao, L., Wu, H., Ling-Hu, C., Zhou, H., Lin, H., Gonzalez-Morales, S., Trejo-Saavedra, D.L., Tian, H., Tang, X., Zhao, M., Huang, Z., Zhou, A., Yao, X., Cui, J., Li, W., Chen, Z., Feng, Y., Niu, Y., Bi, S., Yang, X., Cai, H., Luo, X., Montes-Hernandez, S., Leyva-Gonzalez, M.A., Xiong, Z., He, X., Bai, L., Tan, S., Liu, D., Liu, J., Zhang, S., Chen, M., Zhang, L., Zhang, Y., Liao, W., Wang, M., Lv, X., Wen, B., Liu, H., Luan, H., Yang, S., Wang, X., Xu, J., Li, X., Li, S., Wang, J., Palloix, A., Bosland, P.W., Li, Y., Krogh, A., Rivera-Bustamante, R.F., Herrera-Estrella, L., Yin, Y., Yu, J., Hu, K. and Zhang, Z. (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci U S A, 111, 5135-5140.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Saze, H. and Kakutani, T. (2007) Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. Embo j, 26, 3641-3652.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Singer, T., Yordan, C. and Martienssen, R.A. (2001) Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-

remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev, 15, 591-602.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Stroud, H., Greenberg, M.V., Feng, S., Bernatavichute, Y.V. and Jacobsen, S.E. (2013) Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell, 152, 352-364.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J. and Jacobsen, S.E. (2014) Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol, 21, 64-72.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

The Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485, 635-641.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tan, F., Zhou, C., Zhou, Q., Zhou, S., Yang, W., Zhao, Y., Li, G. and Zhou, D.X. (2016) Analysis of Chromatin Regulators Reveals Specific Features of Rice DNA Methylation Pathways. Plant Physiol, 171, 2041-2054.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. Nature, 461, 423-426.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Vongs, A., Kakutani, T., Martienssen, R.A. and Richards, E.J. (1993) Arabidopsis thaliana DNA methylation mutants. Science, 260, 1926-1928.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Waibel, F. and Filipowicz, W. (1990) U6 snRNA genes of Arabidopsis are transcribed by RNA polymerase III but contain the same two upstream promoter elements as RNA polymerase II-transcribed U-snRNA genes. Nucleic Acids Res, 18, 3451-3458.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wierzbicki, A.T., Haag, J.R. and Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell, 135, 635-648.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wierzbicki, A.T., Ream, T.S., Haag, J.R. and Pikaard, C.S. (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet, 41, 630-634.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P. and Conneely, K.N. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L. and Zilberman, D. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell, 153, 193-205.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., Liu, W., Henderson, I.R., Meyers, B.C., Ausin, I. and Jacobsen, S.E. (2015) A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. Cell, 163, 445-455.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zhang, X., Henderson, I.R., Lu, C., Green, P.J. and Jacobsen, S.E. (2007) Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci U S A, 104, 4536-4541.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zhong, S., Fei, Z., Chen, Y.R., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang, J., Shao, Y. and Giovannoni, J.J. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat Biotechnol, 31, 154-159.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zhong, X., Du, J., Hale, C.J., Gallego-Bartolome, J., Feng, S., Vashisht, A.A., Chory, J., Wohlschlegel, J.A., Patel, D.J. and Jacobsen, S.E. (2014) Molecular mechanism of action of plant DRM de novo DNA methyltransferases. Cell, 157, 1050-1060.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Conclusion et discussion

Le travail qui m'a été confié avait donc pour objectif d'identifier, chez deux mutants ddm1 de la tomate M82 (*ddm1a* et *ddm1b*), de nouvelles insertions d'éléments transposables qui laisseraient alors supposer une réactivation de certains ET dans ces contextes.

Pour réaliser ce travail, l'annotation en ET du génome de référence de la variété M82 de la tomate a été réalisée grâce au pipeline REPET (Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H., 2011, Quesneville, H. *et al.*, 2005). Par la suite, afin d'identifier les nouvelles insertions présentes dans les génomes des souches mutées, différents outils, dédiés à ce type d'analyses, ont été évalués : T-lex (Fiston-Lavier, A.-S., Carrigan, M., Petrov, D. A. & González, J., 2011), Jitterbug (Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S., 2015) et MELT (Gardner, E. J. *et al*., 2017). La plupart des outils développés pour de telles analyses ont été créés pour travailler sur le génome humain ou sur celui de la drosophile, ce qui rend souvent inutilisables pour les génomes de plantes qui ont leurs propres spécificités et souvent un contenant en ET important et varié. Ainsi, T-lex, développé pour des analyses chez la drosophile ne permettait pas de gérer un génome de la taille de celui de la tomate. Jitterbug, quant à lui, n'a jamais pu être lancé en raison, semble-t-il, de problèmes de versions et de compatibilité d'outils, qui n'ont pu être réglés. Finalement, MELT a pu être utilisé après un important travail de pré-processing des fichiers d'entrée qui a tout de même pu être automatisé. En plus de ne pas être spécialisé pour l'analyse d'un organisme précis, et bien que développé pour des analyses chez les animaux et non chez les plantes, cet outil est capable d'analyser des jeux de données importants de manière relativement rapide et dispose également d'un excellent support en cas de difficulté, celui-ci ayant été sollicité par mes soins pour une mise à jour du code lors de son installation.

MELT, une fois lancé sur nos données fourni une liste d'insertions d'ET potentielles. Cependant, le format des fichiers de sortie, VCF, n'est pas le plus simple à comprendre et exploiter pour définir la présence ou absence d'une copie d'ET. Une liste simplifiée des coordonnées d'insertions potentielles dans chacun des deux mutants a toutefois pu être fournie à l'équipe d'origine afin de réaliser des tests de validation des insertions grâce à des analyses d'amplification PCR dont les amorces sont spécifiques aux régions d'insertion.

Malheureusement, les validations PCR n'ont pas fourni de résultats probants chez la tomate, contrairement à ce qui avait été obtenu chez d'autres organismes comme chez l'homme, le

chimpanzé et d'anciens hominidés. Une analyse *a posteriori* du travail réalisé a révélé un problème possiblement lié à l'échantillonnage, ceux-ci étant composés d'un mélange de plusieurs plantes. Le séquençage effectué présente alors une mosaïque de plusieurs génomes dont chacun peut avoir subi des modifications spécifiques. Cependant, MELT n'a pas été conçu pour analyser ce type de données car il cherche à définir si un échantillon est homozygote ou hétérozygote pour une variation donnée. Il est alors envisagé de réaliser de nouveau ces analyses sur des individus et non des pools d'individus afin de pouvoir tirer profit de MELT pour répondre à la question posée.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Approches *in silico* de l'impact des éléments transposables sur la régulation de l'expression des gènes

Mots clés : Eléments transposables, évolution, épigénétique, sélection

Résumé : Les génomes de plantes sont peuplés de différents types d'éléments répétés, notamment des éléments transposables (ET) et des séquences satellites (simple sequence repeats, SSRs) qui peuvent avoir un impact important sur la taille et la dynamique du génome, ainsi que sur la régulation de la transcription génique. Au moins les deux-tiers du génome de la tomate sont composés de répétitions. Bien que leur impact global sur l'organisation du génome ait été largement révélé par l'assemblage du génome entier, leur influence sur la biologie et le phénotype de la tomate reste largement sans réponse. Plus spécifiquement, les effets et les rôles des répétitions de l'ADN sur la maturation des fruits charnus, processus complexe présentant un intérêt agro-économique essentiel, doivent encore être étudiés de manière approfondie et la tomate est sans aucun doute un excellent modèle pour cette étude. Nous avons réalisé une annotation complète du *repeatome* de la tomate pour explorer son impact potentiel sur la composition du génome de la tomate et la transcription des gènes. Nos résultats montrent que le génome de la tomate peut être fractionné en trois compartiments avec une densité de gènes et de répétitions différente, chaque compartiment présentant une composition répétée et génique contrastée, des associations gènes-répétitions et des niveaux transcriptionnels des gènes différents. Dans le contexte de la maturation des fruits, nous avons constaté que des répétitions sont présentes dans la majorité des régions méthylées différentiellement (*differentially methylated regions*, DMRs) et que des milliers de DMRs associées à des répétitions se trouvent à proximité des gènes, y compris des centaines qui sont différentiellement régulés durant ce processus. De plus, nous avons constaté que des répétitions sont également présentes à proximité des sites de liaison de la protéine clé de maturation RIN. Nous avons également observé que certaines familles de répétitions sont présentes à une fréquence élevée inattendue à proximité des gènes exprimés de manière différentielle au cours de la maturation de la tomate. Compte tenu du lien entre ces différentes entités, nous nous sommes demandé s'il était possible que certains éléments transposables du génome de la tomate aient été sélectionnés au cours de l'évolution pour leur impact sur le génome. Nous avons donc développé une série d'analyses afin d'essayer de détecter *in silico* de tels éléments. Les familles d'éléments ainsi sélectionnées sont alors au nombre de 36, et certaines se trouvent associées à des fonctions de gènes particulières. Des analyses plus fines des séquences pourraient alors potentiellement permettre de mettre en évidence des motifs d'intérêt, notamment pour la régulation transcriptionnelle des gènes.

Title : In silico approaches to the impact of transposable elements on the regulation of gene expression

Keywords : Transposable elements, evolution, epigenetics, selection

Abstract : Plant genomes are populated by different types of repetitive elements including transposable elements (TEs) and simple sequence repeats (SSRs) that can have a strong impact on genome size and dynamic as well as on the regulation of gene transcription. At least two-thirds of the tomato genome is composed of repeats. While their bulk impact on genome organization has been largely revealed by whole genome assembly, their influence on tomato biology and phenotype remains largely unaddressed. More specifically, the effects and roles of DNA repeats on the maturation of fleshy fruit, which is a complex process of key agro-economic interest, still needs to be investigated comprehensively and tomato is arguably an excellent model for such study. We have performed a comprehensive annotation of the tomato repeatome to explore its potential impact on tomato genome composition and gene transcription. Our results show that the tomato genome can be fractioned into three compartments with different gene and repeat density, each compartment presenting contrasting repeat and gene composition, repeat-gene associations and different gene transcriptional levels. In the context of fruit ripening, we found that repeats are present in the majority of differentially methylated regions (DMRs) and thousands of repeat associated DMRs are found in the proximity of genes, including hundreds that are differentially regulated during this process. Furthermore, we found that repeats are also present in the proximity of DNA binding sites of the key ripening protein RIN. We also observed that some repeat families are present at unexpected high frequency in the proximity of genes that are differentially expressed during tomato ripening. Given the link between these different entities, we wondered whether it was possible that some transposable elements of the tomato genome were selected during evolution for their impact on the genome. To address this question, we have developed a series of analyzes to try to detect *in silico* such elements. A total of 36 transposable elements families were found to present empirical properties of selection, and some are associated with particular gene functions. More refined analyzes of the sequences could then potentially make it possible to discover motifs of interest, in particular for the transcriptional regulation of genes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France