



**HAL**  
open science

# Évolution des gènes dupliqués chez le pommier : Identification et caractérisation de la dominance du sous-génome dans le génome de la pomme

Tanguy Lallemand

► **To cite this version:**

Tanguy Lallemand. Évolution des gènes dupliqués chez le pommier : Identification et caractérisation de la dominance du sous-génome dans le génome de la pomme. Sciences agricoles. Université d'Angers, 2022. Français. NNT : 2022ANGE0073 . tel-04081238

**HAL Id: tel-04081238**

**<https://theses.hal.science/tel-04081238>**

Submitted on 25 Apr 2023

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

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

# THÈSE DE DOCTORAT DE

L'UNIVERSITÉ D'ANGERS

ÉCOLE DOCTORALE N° 600  
*Écologie, Géosciences, Agronomie et Alimentation*  
Spécialité : *Génétique, génomique et bio-informatique*

Par

**Tanguy LALLEMAND**

## Évolution des gènes dupliqués chez le pommier

Identification et caractérisation de la dominance du sous-génome dans le génome de la pomme

Thèse présentée et soutenue à l'Université d'Angers, le 15 Novembre 2022

Unité de recherche : Institut de Recherche en Horticulture et Semences - UMR 1345 UA AO INRAE [IRHS]

### Rapporteurs avant soutenance :

Ingrid LAFONTAINE Professeure à Sorbonne Université  
Boulos CHALHOUB Directeur de Recherche à Agroscope

### Composition du Jury :

Président :	Didier PELTIER	Professeur à l'Université d'Angers
Examineurs :	Mathieu ROUSSEAU-GUEUTIN	Chargé de Recherche à l'INRAE
	Amandine CORNILLE	Chercheuse CNRS
	France DENOEUDE	Chercheuse CEA
Dir. de thèse :	Claudine LANDÈS	Professeure à l'Université d'Angers
Co-dir. de thèse :	Jean-Marc CELTON	Maître de Conférences à l'Université d'Angers



# REMERCIEMENTS

---

Pour commencer, je souhaite remercier Ingrid Lafontaine et Boulos Chalhoub pour avoir accepté de rapporter ma thèse. Je remercie aussi France Denoeud, Amandine Cornille, Didier Peltier et Mathieu Rousseau-Gueutin pour leur participation à l'évaluation de mon travail de thèse en tant qu'examinateur.

Je tiens également à remercier Jean-Marc Celton et Claudine Landes pour avoir encadré ce projet pendant ces trois années.

Merci, Jean-Marc, pour tes patientes explications en génétique, tes idées d'analyses et de visualisations toujours novatrices. Merci aussi pour ton implication et ta rigueur dans la relecture de mes différents écrits au cours de ces trois années.

Merci Claudine pour ton suivi au quotidien malgré ton emploi du temps chargé qui m'a permis de ne pas me perdre en route. Merci pour tes présentations des différents algorithmes fondateurs, le chemin est encore long avant que je les connaisse aussi bien ! Pour finir merci beaucoup pour ton aide dans ma découverte de l'enseignement.

Je tiens aussi à remercier Sébastien Aubourg, un troisième encadrant, bien qu'officieux qui s'est beaucoup impliqué dans ce projet de thèse. Tu m'as soutenu tout au long de ses trois années et tu as toujours pris le temps de suivre ce projet,, en t'impliquant dans chacune des analyses, en y apportant systématiquement ton expertise et ta rigueur. Ta méthodologie et tes nombreuses vérifications ont permis l'identification et la correction d'erreurs et de biais. Par ailleurs, tes relectures de mes différentes productions m'ont également beaucoup aidé et permis d'augmenter la qualité de celles-ci. Pour finir, merci pour tes conseils en matière de jeux !

Merci également aux membres de mon CSI, Carène Rizzon, Béatrice Duval, Jean-Pierre Renou et Sébastien Aubourg. Merci pour votre implication dans ma thèse, vos avis et conseils sur mon travail. Ces deux rencontres ont été enrichissantes et ont été des jalons importants de mon travail. En effet, la rédaction des rapports ainsi que les présentations associées ont permis d'avoir un regard extérieur d'experts sur les différentes approches utilisées mais aussi sur la façon de les présenter.

Je tiens à remercier plus spécialement Carène Rizzon pour son implication dans ma thèse pour l'ensemble des analyses et plus particulièrement son expertise pour les analyses



de pression de sélection et d'éléments Transposables. Merci aussi pour ton implication dans la construction et la relecture d'un article de revue, d'un article de recherche ainsi que de ce manuscrit.

Merci à Martin Leduc, le meilleur camarade de bureau. Merci pour tes nombreux conseils et avis toujours pertinents sur mon travail. Nos discussions et tes critiques ont toujours été très enrichissantes. Merci également pour ton implication dans les enseignements et tes conseils pédagogiques. Merci pour ta patience et ton implication dans des beta tests grandeur nature de pipelines. Ils m'ont permis de fiabiliser les différentes implémentations et les rendre le plus généralistes possible. Ceux-ci m'ont aussi permis d'améliorer, voir, de créer les documentations associées.

Je voudrais aussi remercier les membres de l'équipe BIDEfl de m'avoir accueilli et aidé. En premier lieu, je voudrais remercier Sandra Peltier. Merci pour ta passion et ta créativité au quotidien ! De même, je voulais te remercier en particulier pour ton script anaDiff, qui est un exemple de package fiable, documenté et facilement utilisable. Ta prise en main de certains de mes outils m'a permis de les améliorer que ce soit en termes d'implémentations mais également de documentation. Merci également pour tes découvertes d'outils et d'astuces ainsi que ton expertise sur énormément de sujet et notamment la transcriptomique ! Merci aussi à Sylvain Gaillard pour nos discussions techniques et tes connaissances profondes des différents systèmes informatiques et ta vision globale du métier de bioinformaticien. De même, je tiens à remercier Eric Montaudon, pour l'administration de mes données et systèmes et en particulier leur sauvegarde. Merci également pour l'administration des ressources de calculs, clés dans mes analyses à grande échelle. Pour finir, merci à Jean-Pierre Renou pour tes explications et anecdotes toujours passionnantes et enrichissantes et ta vision globale du monde de la recherche toujours inspirante.

Merci aussi aux membres de l'équipe VALEMA de m'avoir intégré et avant tout Amanda Cattani, Patricia Mallegol et Sandrine Balzergue. Merci de m'avoir montré comment fonctionnaient les études sur le terrain mais aussi vos avis sur les études que j'ai mené. Merci à Skander Hatira, pour nos innombrables discussions informatiques et en particulier pour m'avoir fait découvrir Snakemake, un outil qui change définitivement la vie. Bon désolé pour conda, j'étais déjà adepte !

Merci aux doctorants que j'ai pu côtoyer au cours de ces années à l'IRHS et notamment Antoine Bodelot, Marie Charlotte Guilloux, Pierre Bouillon, Loup Tran Van Canh, Xabi Cazenave et Andréa Bouanich. Je tenais à remercier plus particulièrement Marie Charlotte, Loup et Andréa pour leur implication dans la relecture et pour la préparation

---

de la soutenance! Merci également Marie-Charlotte pour ton aide dans mes aventures administratives.

En parlant d'administratif, je voulais remercier aussi l'équipe PAIGE et en particulier Lydie Mauge et Magali Tabuteau qui m'ont aidé à de (trop?) nombreuses reprises, toujours avec le sourire et efficacement à gérer les soucis administratifs qui se sont présentés.

Merci également à l'école doctorale EGAAL et en particulier Karine Couturier et Nicole Lotode pour leur aide administrative côté universitaire, que ce soit pour les inscriptions, les CSI et les contrats DCACE.

Merci à GenOuest pour leur ressources de calculs qui m'ont permis d'apprendre à implémenter des outils à l'épreuve des grilles de calculs mais aussi d'accélérer les différents calculs.

Pour finir, merci à Emmanuelle Lerat de m'avoir donné l'opportunité d'écrire un article de revue.

Merci également à mes amis et ma famille et mes amis qui m'ont soutenus durant ces trois années!

Enfin, merci à ma femme de m'avoir accompagné au cours de ces trois années. Merci à mon petit garçon qui éclairent mes journées et pour ses quelques lettres écrites au clavier dont une partie ont été conservées dans le manuscrit final.



# TABLE DES MATIÈRES

---

<b>Introduction</b>	<b>21</b>
Famille des rosacés . . . . .	21
L’histoire évolutive du pommier . . . . .	21
Les mécanismes de duplications, identification et devenir des gènes dupliqués . .	22
La dominance de sous-génome . . . . .	61
<b>Questions de recherche</b>	<b>65</b>
<b>1 Matériel et méthode</b>	<b>67</b>
1.1 Environnement informatique . . . . .	67
1.2 Statistiques . . . . .	70
1.2.1 Tests de comparaison d’une seule variable . . . . .	70
1.2.2 Tests de comparaison de deux variables . . . . .	76
1.2.3 Tests d’adéquation à une loi . . . . .	77
1.2.4 Tests utilisés dans les méta-analyses . . . . .	78
1.3 Matériel génétique . . . . .	80
<b>2 Préparation des données</b>	<b>81</b>
2.1 Blocs de synténie . . . . .	81
2.1.1 Introduction . . . . .	81
2.1.2 Matériel et méthode . . . . .	82
2.1.3 Résultats . . . . .	88
2.1.4 Discussion . . . . .	90
2.1.5 Conclusion . . . . .	95
2.2 Triplets pommier-pêcher . . . . .	96
2.2.1 Matériel et méthode . . . . .	96
2.2.2 Résultats . . . . .	96
2.2.3 Discussion . . . . .	100
2.2.4 Conclusion . . . . .	102

<b>3</b>	<b>Étude du fractionnement génomique chez le pommier</b>	<b>103</b>
3.1	Introduction . . . . .	103
3.2	Matériel et méthode . . . . .	104
3.3	Résultats . . . . .	106
3.4	Discussion . . . . .	111
3.5	Conclusion . . . . .	113
<b>4</b>	<b>Étude des QTLs entre les fragments synténiques</b>	<b>115</b>
4.1	Introduction . . . . .	115
4.2	Matériel et méthode . . . . .	117
4.3	Résultats . . . . .	121
4.4	Discussion . . . . .	127
4.5	Conclusion . . . . .	129
<b>5</b>	<b>Évaluation de la pression de sélection des gènes ohnologues</b>	<b>131</b>
5.1	Introduction . . . . .	131
5.2	Matériel et méthode . . . . .	137
5.3	Résultats . . . . .	138
5.4	Discussion . . . . .	145
5.5	Conclusion . . . . .	148
<b>6</b>	<b>Analyse de l'expression des gènes ohnologues</b>	<b>151</b>
6.1	Introduction . . . . .	151
6.2	Matériel et méthode . . . . .	154
6.2.1	Récupération des données . . . . .	154
6.2.2	Pseudo alignement et analyse différentielle . . . . .	155
6.2.3	Analyse des niveaux d'expression . . . . .	157
6.3	Résultats . . . . .	158
6.3.1	Récupération des données . . . . .	158
6.3.2	Pseudo alignement et analyse différentielle . . . . .	162
6.3.3	Analyse des niveaux d'expression . . . . .	164
6.4	Discussion . . . . .	169
6.4.1	Récupération des données . . . . .	169
6.4.2	Pseudo alignement et analyse différentielle . . . . .	170
6.4.3	Analyse des niveaux d'expression . . . . .	172

6.5	Conclusion . . . . .	173
<b>7</b>	<b>Analyse des éléments transposables des gènes ohnologues</b>	<b>175</b>
7.1	Introduction . . . . .	175
7.2	Matériel et méthode . . . . .	180
7.3	Résultats . . . . .	181
7.3.1	Comparaison des régions géniques et intergéniques . . . . .	181
7.3.2	Comparaison des gènes ohnologues et des singletons . . . . .	183
7.3.3	Comparaison des types d'ET des segments synténiques . . . . .	184
7.3.4	Comparaison de l'environnement en ET des gènes ohnologues . . .	186
7.4	Discussion . . . . .	193
7.4.1	Comparaison des régions géniques et intergéniques . . . . .	193
7.4.2	Comparaison des gènes ohnologues et des singletons . . . . .	194
7.4.3	Comparaison des types d'ET des segments synténiques . . . . .	195
7.4.4	Comparaison de l'environnement en ET des gènes ohnologues . . .	196
7.5	Conclusion . . . . .	197
<b>8</b>	<b>Étude des méthylation de l'ADN des gènes ohnologues</b>	<b>199</b>
8.1	Introduction . . . . .	199
8.2	Matériel et méthode . . . . .	201
8.2.1	Récupération des données Bisulfite Seq . . . . .	201
8.2.2	Calcul des rapports de méthylation . . . . .	202
8.3	Résultats . . . . .	203
8.3.1	Comparaison des niveaux de méthylation de l'ADN entre les gènes ohnologues . . . . .	203
8.3.2	Comparaison du nombre de positions de cytosines . . . . .	210
8.4	Discussion . . . . .	214
8.5	Conclusion . . . . .	216
<b>9</b>	<b>Discussion générale</b>	<b>219</b>
9.1	Analyse du déséquilibre de QTL . . . . .	219
9.2	La dominance de sous-génome . . . . .	229
	<b>Conclusion</b>	<b>237</b>
	Déséquilibre . . . . .	237
	Perspective . . . . .	238

## TABLE DES MATIÈRES

---

<b>A</b>	<b>Appendices</b>	<b>241</b>
A.1	Étude du fractionnement génomique chez le pommier . . . . .	241
A.2	Évaluation de la pression de sélection des gènes ohnologues . . . . .	248
A.3	Analyse de l'expression des gènes ohnologues . . . . .	252
A.4	Analyse des éléments transposables des gènes ohnologues . . . . .	267
A.5	Étude des méthylations de l'ADN des gènes ohnologues . . . . .	269
A.6	Intégration des résultats . . . . .	275
<b>B</b>	<b>Appendices Articles</b>	<b>282</b>
B.1	Article accepté . . . . .	282
B.2	Article soumis . . . . .	291
	<b>Bibliographie</b>	<b>307</b>

# TABLE DES FIGURES

---

2.1	Schéma de l'algorithme de construction des blocs de synténie par i-ADHoRe	87
2.2	Distribution des tailles des blocs de synténie identifiés avec MCSanX, i-ADHoRe et SynMap . . . . .	89
2.3	Visualisation circulaire des blocs de synténie . . . . .	91
2.4	Diagramme de Venn représentant les proportions de couples de gènes ohnologues identifiés par les différents outils . . . . .	93
2.5	Schéma de la construction de triplets <i>Malus domestica</i> et <i>Prunus persica</i> .	97
2.6	Histogramme du nombre de triplets pommier-pêcher pour les couples de chromosomes ohnologues . . . . .	98
2.7	Histogramme du pourcentage d'identité des gènes du pommier et des gènes du pêcher intégrés dans les triplets . . . . .	99
2.8	Histogramme de la taille des fragments chromosomique synténiques . . . .	100
3.1	Distribution des pourcentages de rétention des gènes ohnologues de <i>M. domestica</i> le long du génome de <i>P. persica</i> . . . . .	107
3.2	Distribution des pourcentages de rétention des gènes ohnologues du pommier le long du chromosome 2 de <i>P. persica</i> . . . . .	108
3.3	Distribution des pourcentages des gènes ohnologues entre les chromosomes ohnologues . . . . .	110
4.1	Schéma des différentes possibilités de localisation de Quantitative Trait Locus (QTL) suivant le marqueur auquel il est associé. . . . .	118
4.2	Ontologie des métadonnées décrivant les traits phénotypiques associés aux QTLs . . . . .	120
4.3	Histogramme du nombre de QTLs associés aux blocs synténiques dans chaque paire de fragments de chromosomes ohnologues. . . . .	122
4.4	Histogramme à barres empilées des proportions de QTLs annotés par les termes de l'ontologie. . . . .	124



TABLE DES FIGURES

---

4.5 Matrice de couleur des résultats des test z de proportion pour les QTLs, filtré sur les différents niveaux de l'ontologie . . . . . 126

5.1 Distribution des valeurs de  $K_a$ ,  $K_s$  et  $\omega$  pour les différentes méthodes de comptages . . . . . 139

5.2 Estimation de la datation de la spéciation de *P. persica* et *M. domestica* . 142

5.3 Distribution valeurs de  $K_a$  (en vert),  $K_s$  (en orange) et  $K_a/K_s$  (en bleu) intraspécifiques . . . . . 143

5.4 Distribution des valeurs d' $\omega$  entre les chromosomes ohnologues. . . . . 144

6.1 Schéma de la construction de l'analyse différentielle des gènes ohnologues . 157

6.2 Ontologie des traitements et des tissus des séries de séquençages RNA-Seq analysés . . . . . 160

6.3 Proportions des métadonnées de tissus, cultivars et traitements associés aux expériences RNA-Seq testées . . . . . 161

6.4 Distribution du nombre de gènes sur-exprimé entre les chromosomes ohnologues . . . . . 166

6.5 Distribution du nombre de gènes sur-exprimé par rapport à son ohnologue le long du fragment chromosomique synténique 1-7 . . . . . 168

7.1 Distribution des proportions des Élément Transposable (ET) associés aux régions intergéniques et géniques au sein des différentes classes, familles et sous-familles . . . . . 182

7.2 Distribution du nombre d'ET et de la couverture en ET entre les gènes non dupliqués et les gènes ohnologues. . . . . 185

7.3 Diagrammes circulaires de la répartition des classes (Figure 7.3A), familles (Figure 7.3B) et sous-familles (Figure 7.3C) des ET associés aux gènes ohnologues . . . . . 187

7.4 Distribution de la taille des ET associés aux gènes ohnologues . . . . . 188

7.5 Distribution de la taille des ET en fonction des sous classes d'ET . . . . . 189

7.6 Distribution de la couverture en ET pour les paires de chromosomes ohnologues étudiés . . . . . 192

8.1 Distribution des rapports de méthylation des cytosines localisées 500 bp en amont des gènes et groupés par paires de fragments chromosomiques synténiques . . . . . 205

8.2	Distribution du pourcentage de gènes non commutants parmi les 25 gènes les plus différents en terme de nombre de position de cytosines dans le contexte CHH . . . . .	213
9.1	Carte thermique regroupant les coefficients de corrélation globaux . . . . .	220
9.2	Cartes thermiques regroupant les coefficients de corrélation . . . . .	221
9.3	Représentation des corrélation sous forme de Modèles Graphiques Gaussiens (GGM) . . . . .	222
9.4	Ensemble des matrices de dispersion des différentes variables et des lignes de régression associées pour les paires 1-7 . . . . .	223
9.5	Schémas des mécanismes post-Whole Genome Duplication (WGD) et la dynamique des ETs. . . . .	232
9.6	Schéma récapitulatif du mécanisme hypothétique mis en œuvre dans le génome de la pomme . . . . .	233
A.1	Pourcentages de rétention des gènes du pommier pour le chromosome 1 de <i>P. persica</i> . . . . .	241
A.2	Pourcentages de rétention des gènes du pommier pour le chromosome 3 de <i>P. persica</i> . . . . .	242
A.3	Pourcentages de rétention des gènes du pommier pour le chromosome 4 de <i>P. persica</i> . . . . .	243
A.4	Pourcentages de rétention des gènes du pommier pour le chromosome 5 de <i>P. persica</i> . . . . .	244
A.5	Pourcentages de rétention des gènes du pommier pour le chromosome 6 de <i>P. persica</i> . . . . .	245
A.6	Pourcentages de rétention des gènes du pommier pour le chromosome 7 de <i>P. persica</i> . . . . .	246
A.7	Pourcentages de rétention des gènes du pommier pour le chromosome 8 de <i>P. persica</i> . . . . .	247
A.8	Distribution des taux de substitution non synonyme ( $K_a$ ) entre les chromosomes ohnologues pour les différents modèles évolutifs . . . . .	248
A.9	Distribution des taux de substitution synonyme ( $K_s$ ) entre les chromosomes ohnologues pour les différents modèles évolutifs . . . . .	249
A.10	Distribution des taux $\omega$ entre les chromosomes ohnologues pour les différents modèles évolutifs . . . . .	249

TABLE DES FIGURES

---

A.11 Distribution des taux de substitution non synonyme entre les chromosomes  
ohnologues . . . . . 250

A.12 Distribution des taux de substitution synonyme ( $K_s$ ) entre les chromosomes  
ohnologues . . . . . 250

A.13 Résumés des différents indicateurs du pipeline de pseudo-mapping pour les  
896 échantillons . . . . . 252

A.14 Distribution de la taille des fragments aligné par Salmon . . . . . 253

A.15 Nombre de *reads* avec certaines longueurs d'adaptateur coupées . . . . . 253

A.16 Estimation du nombre de séquences pour chaque échantillon . . . . . 254

A.17 La valeur de qualité moyenne (Phred score) pour chaque position de base  
dans la lecture . . . . . 255

A.18 Distribution du nombre de *reads* avec des scores de qualité moyens. . . . . 255

A.19 Distribution du pourcentage d'appels de base à chaque position pour la-  
quelle un N a été appelé. . . . . 256

A.20 Distribution du contenu en GC moyen des *reads* . . . . . 256

A.21 La quantité totale de séquences sur-représentées trouvées dans chaque li-  
brairie. . . . . 257

A.22 Ensemble des Distribution des pourcentagess cumulés des séquences adap-  
tatrices . . . . . 259

A.23 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 3-11 . . . . . 261

A.24 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 5-10 . . . . . 262

A.25 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 6-14 . . . . . 263

A.26 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 8-15 . . . . . 264

A.27 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 9-17 . . . . . 265

A.28 Distribution du nombre de gènes sur-exprimé par rapport a son ohnologue  
pour la paire 13-16 . . . . . 266

A.29 Carte thermique des résultats des tests agrégés selon les conditions expéri-  
mentales . . . . . 267

A.30 Répartition des classes, sous-classes, superfamilles des TEs associés au gènes ohnologues ou au gènes non dupliqués par WGD . . . . . 268

A.31 Boîtes à moustache de la distribution des rapport de méthylations des cytosines localisées 100 bp en amont des gènes et groupés par paires de fragments chromosomiques synténiques . . . . . 270

A.32 Boîtes à moustache de la distribution des rapport de méthylations des cytosines localisées en 2 kb en amont et en aval des gènes et groupés par paires de fragments chromosomiques synténiques . . . . . 271

A.33 Boîtes à moustache de la distribution des rapport de méthylations des cytosines localisées sur les régions exoniques des gènes groupés par paires de fragments chromosomiques synténiques . . . . . 272

A.34 Matrices de dispersion et droites de régression associées à la paire 2-15 . . 275

A.35 Matrices de dispersion et droites de régression associées à la paire 3-11 . . 276

A.36 Matrices de dispersion et droites de régression associées à la paire 5-10 . . 277

A.37 Matrices de dispersion et droites de régression associées à la paire 6-14 . . 278

A.38 Matrices de dispersion et droites de régression associées à la paire 08-15 . . 279

A.39 Matrices de dispersion et droites de régression associées à la paire 09-17 . . 280

A.40 Matrices de dispersion et droites de régression associées à la paire 13-16 . . 281

# LISTE DES TABLEAUX

---

1.1	Exemple de table contingence 2 X 2 pour les tests exacts de Fisher . . . .	74
3.1	Résultats des t-test appariés des pourcentages de rétention des gènes . . .	109
5.1	Description des valeurs de $\omega$ , $K_a$ et $K_s$ calculés avec les différentes méthodes de comptages implémentées dans PAML . . . . .	140
5.2	Résultats des tests de Wilcoxon et des tests de Kolmogorov-Smirnov à deux échantillons pour la qualité de l'ajustement des valeurs de $\omega$ entre les fragments synténiques . . . . .	145
6.1	Résultats des tests de Wilcoxon agrégés à l'aide de la méthode de Fisher pour les paires de fragments chromosomiques synténiques. . . . .	165
7.1	Résultats des tests de rang de Mann-Whitney $U$ pour la couverture en ET, la densité en ET et le nombre ET entre les gènes ohnologues et les gènes non dupliqués. . . . .	184
7.2	Proportion des annotations des ET associés au gènes ohnologues . . . . .	186
7.3	Description de la taille des ET en fonction des différentes sous classes associés au gènes ohnologues . . . . .	190
7.4	Résultats des tests de rang de Mann-Whitney $U$ de la taille des ET entre les fragments synténiques . . . . .	190
8.1	Ratios de méthylation sur chaque paire de chromosomes ohnologues dans les différentes régions considérées pour les 3 contextes de méthylation. . . .	210
8.2	Nombre de cytosines associées à chacun des contextes de méthylation 100 bp en amont pour les paires de gènes ohnologues. . . . .	212
8.3	Résultats des tests appariés de Wilcoxon des pourcentages de gènes non commutants pour les 25 gènes présentant la plus grande différence de nombre de positions de cytosines dans le contexte CHH par rapport à 25 gènes sélectionnés au hasard. . . . .	213

---

9.1	Résultats des corrélations de Pearson pour les couples de fragments chromosomiques synténiques ohnologues . . . . .	226
A.1	Résultats des tests de Wilcoxon et des tests de Kolmogorov-Smirnov à deux échantillons pour la qualité de l'ajustement des valeurs de dN et dS entre les fragments chromosomiques synténiques . . . . .	251
A.2	Comptages des cultivars associés aux analyses RNA-Seq utilisées . . . . .	258
A.3	Comptages des tissus associés aux analyses RNA-Seq utilisées . . . . .	259
A.4	Comptages des traitements associés aux analyses RNA-Seq utilisées . . . . .	260
A.5	Comparaison du nombre de cytosines considérés dans les régions exoniques des gènes ohnologues pour les paires de fragments chromosomiques synténiques . . . . .	269
A.6	Comparaison du nombre de cytosines considérés 2 kb en amont et en aval des gènes ohnologues pour les paires de fragments chromosomiques synténiques . . . . .	273
A.7	Comparaison du nombre de cytosines considérés 500 bp en amont des Transcription Start Site (TSS) pour les paires de fragments chromosomiques synténiques . . . . .	274

# LISTE DES ACRONYMES

---

<i>A. arenosa</i>	<i>Arabidopsis arenosa.</i>
<i>A. thaliana</i>	<i>Arabidopsis thaliana.</i>
AFLP	Amplified Fragment Length Polymorphism.
<i>B. juncea</i>	<i>Brassica juncea.</i>
<i>B. napus</i>	<i>Brassica napus.</i>
<i>B. oleracea</i>	<i>Brassica oleracea.</i>
<i>B. rapa</i>	<i>Brassica rapa.</i>
BBH	<i>Best Blast Hit.</i>
<i>C. bursa-pastoris</i>	<i>Capsella bursa-pastoris.</i>
CDS	<i>Coding DNA Sequence.</i>
DAG	Directed Acyclic Graph.
DDC	Duplication-Degeneration-Complementation.
ENA	European Nucleotide Archive.
EST-SSR	Expressed Sequence Tags Simple Sequence Repeats.
ET	Élément Transposable.
<i>G. raimondii</i>	<i>Gossypium raimondii.</i>
<i>G. tuberosa</i>	<i>Genlisea tuberosa.</i>
gbM	gene Body Methylated.
GDR	Genome Database for Rosaceae.
GEO	Gene Expression Omnibus.
GG	Greedy Graph.
GGM	Modèles Graphiques Gaussiens.
GHM	Matrice d'Homologie de Gènes.

GWAS	Genome-Wide Association Study.
LINE	Long Interspersed Nuclear Element.
LTR	Long Terminal Repeat.
<i>M. domestica</i>	<i>Malus domestica</i> .
<i>M. orientalis</i>	<i>Malus orientalis</i> .
<i>M. ringens</i>	<i>Mimulus ringens</i> .
<i>M. sieversii</i>	<i>Malus sieversii</i> .
<i>M. sylvestris</i>	<i>Malus sylvestris</i> .
MCMC	Markov chain Monte Carlo.
<i>N. sylvestris</i>	<i>Nicotiana sylvestris</i> .
<i>N. tabacum</i>	<i>Nicotiana tabacum</i> .
<i>N. tomentosiformis</i>	<i>Nicotiana tomentosiformis</i> .
<i>P. japonica</i>	<i>Paris japonica</i> .
<i>P. persica</i>	<i>Prunus persica</i> .
QTL	Quantitative Trait Locus.
RADP	Random Amplified Polymorphic DNA.
RBBH	<i>Reciprocal Best Blast Hit</i> .
RFLP	Restriction Fragment Length Polymorphism.
RISC	RNA-Induced Silencing Complex.
SAM	Sélection Assistée par Marqueurs.
SINE	Short Interspersed Nuclear Element.
siRNA	Small Interfering Ribonucleic Acid.
SNP	Single Nucleotide Polymorphism.
SRA	Sequence Read Archive.
SSR	Simple Sequence Repeats.
<i>T. aestivum</i>	<i>Triticum aestivum</i> .



## Liste des acronymes

---

TIR	Terminal Inverted Repeat.
TSS	Transcription Start Site.
<i>U. gibba</i>	<i>Utricularia gibba</i> .
URL	Uniform Resource Locator.
<i>V. inequalis</i>	<i>Venturia inequalis</i> .
WGD	Whole Genome Duplication.
<i>Z. maize</i>	<i>Zhea maize</i> .

# INTRODUCTION

---

## Famille des rosacées

La famille des *Rosaceae* est une famille majeure qui présente une diversité de plantes d'intérêt économique, que ce soit d'un point de vue alimentaire ou ornemental. Cette famille regroupe près de 100 genres et 3000 espèces (Soundararajan et al., 2019). Elle présente une diversité importante de type de fruits parmi lesquels on retrouve des drupes, pommes, drupelets et akènes. Les études phylogénétiques des *Rosaceae* basées sur des séquences nucléaires et chloroplastiques ont permis la construction de trois familles à partir de 88 genres (Potter et al., 2007). Ainsi, les *Rosaceae* sont constitués des sous-familles des *Amygdaloideae*, *Rosoideae* et des *Dryadoideae* qui présentent une grande diversité génétique. Les *Amygdaloideae* sont subdivisés en trois groupes dont les *Amygdaloideae* (n=8) qui rassemble des organismes tels que la cerise, le pêcher ou l'abricot du genre *Prunus*. On retrouve aussi les *Spiraeoideae* (n=9) groupant des organismes tels que les Spirées (genre *Spiraea*). Pour finir, les *Amygdaloideae* rassemble les *Maloideae* (aussi appelé *Pomoideae*) qui présentent principalement des espèces n=17 avec des organismes tels que les pommiers (genre *Malus*), poiriers (genre *Pyrus*) et sorbier du genre *Sorbus*. Chez les *Rosaceae* on retrouve aussi la sous-famille des *Rosoideae* (n=7) qui rassemble les fraisiers (genre *Fragaria*), le framboisier (genre *Rubus*) et le rosier du genre *Rosa*. Pour finir, la sous-famille des *Dryadoideae* (n=9) rassemble des espèces du genre *Cercocarpus* et *Dryas*. La famille des *Rosaceae* présente une importante diversité génétique en particulier du fait du grand nombre de duplications connues par cette famille (Clark & Donoghue, 2018; Van de Peer et al., 2017) et en particulier une WGD récente chez les *Maloideae* (Daccord et al., 2017; J. Wu et al., 2013).

## L'histoire évolutive du pommier

Le pommier est une espèce d'importance agronomique qui permet la production de nombreuses denrées alimentaires (Morgan, 2013). Les espèces du genre *Malus* partagent toutes une WGD commune aux *Maloideae*. Cette WGD a été datée selon les dernières



estimations entre 13,5 et 27,1 millions d'années (Su et al., 2021). Des espèces du genre *Malus* ont été retrouvées presque partout dans le monde, principalement en Eurasie (Cornille et al., 2014). Ce genre regroupe une trentaine d'espèces (Robinson et al., 2001). Au sein des *Malus*, différentes espèces sont particulièrement importantes et spécifiquement *Malus pumila* Mill., plus connu sous le nom de *M. domestica* Borkh. Cette espèce est issue de la domestication de *Malus sieversii* et proviendrait d'Asie Centrale, en particulier des montagnes de Tian Shan. La dissémination de l'espèce s'est faite le long de la route de la soie de -3500 av. J.-C. jusqu'au milieu du XVe siècle (Cornille et al., 2019). Cette dissémination a été à l'origine d'hybridations avec d'autres espèces de pommiers sauvages et notamment avec *Malus orientalis* Uglitz. (Cornille et al., 2019) et *Malus sylvestris* Mill. (Cornille et al., 2012). Ainsi, pour le génome du cultivar Gala il est estimé que 23 % du génome provient d'hybridation avec *M. orientalis* et *M. sylvestris* (X. Sun et al., 2020). Une majorité des espèces des variétés de *M. domestica* sont diploïdes ( $2n=34$ ). On retrouve néanmoins 18,5 % de variétés triploïdes (Lassois et al., 2016).

## Les mécanismes de duplications, identification et devenir des gènes dupliqués

Les duplications de gènes sont des processus particulièrement importants dans l'évolution des espèces. Dans cet article de revue, nous nous sommes intéressés aux mécanismes qui aboutissent à la duplication de gènes à différentes échelles. Cet article traite aussi des différents devenir possibles pour ces gènes dupliqués. Par ailleurs, cette revue traite des algorithmes bio-informatiques permettant l'identification des gènes dupliqués suivant leur mode de duplication. Une partie importante traite notamment de l'identification des blocs de synténies dans un contexte de WGD, une question clé pour cette thèse.

Review

# An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice

Tanguy Lallemand <sup>1,†</sup> , Martin Leduc <sup>1,†</sup>, Claudine Landès <sup>1</sup>, Carène Rizzon <sup>2</sup> and Emmanuelle Lerat <sup>3,\*</sup> 

- <sup>1</sup> IRHS, Agrocampus-Ouest, INRAE, Université d'Angers, SFR 4207 QuaSaV, 49071 Beaucouzé, France; tanguy.lallemand@inrae.fr (T.L.); martin.leduc@etud.univ-angers.fr (M.L.); claudine.landes@inrae.fr (C.L.)
- <sup>2</sup> Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, Université Paris-Saclay, UMR CNRS 8071, ENSIE, USC INRAE, 23 bvd de France, CEDEX, 91037 Evry Paris, France; carene.rizzon@univ-evry.fr
- <sup>3</sup> Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France
- \* Correspondence: emmanuelle.lerat@univ-lyon1.fr; Tel.: +3342432918
- † These authors contributed equally to this work.

Received: 30 July 2020; Accepted: 2 September 2020; Published: 4 September 2020



**Abstract:** Gene duplication is an important evolutionary mechanism allowing to provide new genetic material and thus opportunities to acquire new gene functions for an organism, with major implications such as speciation events. Various processes are known to allow a gene to be duplicated and different models explain how duplicated genes can be maintained in genomes. Due to their particular importance, the identification of duplicated genes is essential when studying genome evolution but it can still be a challenge due to the various fates duplicated genes can encounter. In this review, we first describe the evolutionary processes allowing the formation of duplicated genes but also describe the various bioinformatic approaches that can be used to identify them in genome sequences. Indeed, these bioinformatic approaches differ according to the underlying duplication mechanism. Hence, understanding the specificity of the duplicated genes of interest is a great asset for tool selection and should be taken into account when exploring a biological question.

**Keywords:** gene duplication; bioinformatic tools; paralogous genes; genome evolution; synteny

## 1. Introduction

The eukaryotic genome organization is complex and contains different types of sequences with much of them being non-coding sequences that may have an important impact on genome functioning and regulation. Moreover, genomes are highly dynamic with several ongoing processes allowing the creation of genetic novelty necessary for species to evolve and adapt to changing environments. Among the different possibilities, gene duplication is a very important mechanism providing new genetic material and opportunities to acquire new functions [1].

In particular, numerous examples have described the role of duplication in some cases of adaptation to environmental conditions [2]. For example, gene duplication has played a role in nutrient transport under stress conditions, in protection against heat, cold, or salty environments, in the resistance to drugs and pesticides, but also in the adaptation to domestication. Gene duplication can also be involved in speciation, especially via whole genome duplication (WGD) as it is suspected in plants, where a correlation has been observed between WGD and increased rates of speciation or divergence [3]. In particular, this mechanism is thought to have generated the new flowering plant *Mimulus peregrinus*

within the last 140 years [4]. Although less numerous than in plants, some examples also exist in animals such as in *Drosophila* where the hybrid-male sterility gene *Odysseus* was formed by gene duplication [5]. On the other hand, duplication may also have important deleterious effects in humans and can be associated with some diseases [6]. For example, the analysis of human genes linked to diseases made it possible to show that 80% of them have been duplicated in their evolutionary history, the disease-associated mutation being associated with only one of the duplicated copies [7]. Recently, the analysis of the evolution of cancer suppression in mammals revealed that species known to be resistant to cancer contain the most cancer gene copies [8].

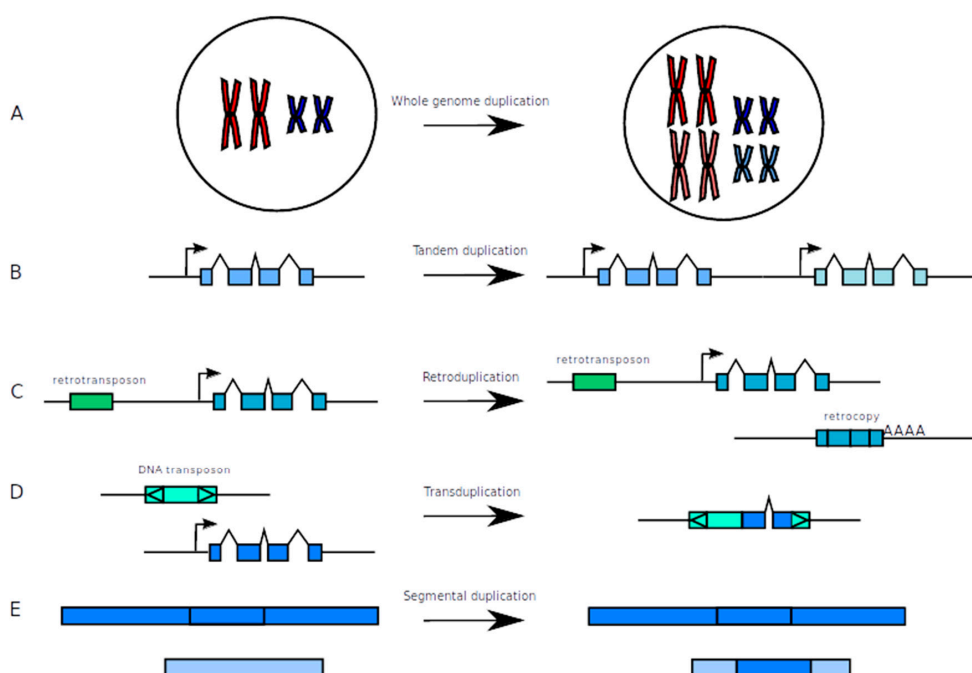
Duplicated genes are also called paralogs in contrast to orthologs, to refer to their homologous relationship, i.e., the fact that they descend from a common ancestor via a duplication event rather than a speciation event. The terminology concerning duplicated genes can be complex and depends upon different factors (for a review, see [9]). In particular, it may be difficult to assess precisely the evolutionary relationships between duplicated genes since duplication is often followed by speciation and gene loss. Several definitions have been proposed to integrate more or less precise ideas concerning the mechanism of formation and the evolutionary relationship among paralogs. For example, ohnologs correspond to paralogs that have been created by WGD [10]. Three new definitions, pro-ortholog, semi-ortholog, and trans-homolog, were proposed to account for situations in which one or both lineages that lead to two present-day genes involve gene duplication [11]. In that respect, a pro-ortholog is a gene that is orthologous to the ancestor of a set of paralogs of the gene under consideration whereas a semi-ortholog is one of the descendants of an ortholog of the gene under consideration, after that gene has duplicated. Trans-homologs can be defined as genes related to each other via two independent duplication events from the same ancestral gene. Moreover, it is also possible to link paralogous relationships to speciation events with the definition of in-paralogs and out-paralogs [12]. When paralogs from a given lineage have evolved by gene duplication that happened after a speciation event, they can be referred to as in-paralogs. On the opposite, paralogous genes which have evolved by duplication events happening before a speciation event, can be referred to as out-paralogs. Many other terms, although less used, have been proposed to take into account chromosomal position retention, the combination of vertical and horizontal transmissions or to highlight paralogous genes appearing to be orthologs due to differential gene loss [9].

In a genome, duplicated genes can thus be formed by various mechanisms and may have different ages and fates. This makes their bioinformatic identification all the more difficult since according to the methods used, different duplicated gene datasets will be identified inside the same organism. In this review, we thus aim at describing the evolutionary processes implicated in the formation and the fate of duplicated genes as well as the different bioinformatic approaches that can be used to identify them in genome sequences. The question of deciphering the evolutionary relationships among duplicated genes will not be discussed in detail, for reviews on the subject see [13,14].

## 2. Evolutionary Processes Leading to the Formation and the Fate of Duplicated Genes

### 2.1. How to Make New from Old: Duplication Mechanisms

Duplicated genes can appear under various forms which are the consequences of the mechanisms that generated them. Some of the mechanisms can be particularly well documented but it is not always the case, at least for some organisms. According to the mechanism, the results concerning the gene content can be different since it can either involve individual genes or all genes on entire chromosomes (Figure 1).



**Figure 1.** The different types of duplications. **(A)** Whole genome duplication which implies complete chromosome duplication. **(B)** Tandem duplications which produce identical adjacent sequences. **(C)** Retroduplication, which produces a retrocopy of a gene devoid of introns and with a polyA tail. **(D)** Transduplication in which a DNA transposon acquires fragments of genes. **(E)** Segmental duplications which correspond to long stretches of duplicated sequences with high identity.

### 2.1.1. Whole Genome Duplication (WGD)

In the first mechanism, duplicated genes arise from the duplication of complete chromosomes, which correspond to what is called whole genome duplication (WGD) (Figure 1A). In that case, all chromosomes from a genome will be duplicated, leading each gene from the genome to exist in two copies. This type of duplication has been well documented in plants and is defined as polyploidization, for which it is possible to distinguish the mechanism of hybridization between different species (allopolyploidization) or inside a given species (autopolyploidization) [15]. Different mechanisms have been shown to produce this outcome such as polyspermy, non-reduced gametes or incomplete mitosis during the early stage of embryo development [16]. Gene duplication, independently of the mechanism of formation, is largely present in plant genomes since on average 64.5% of genes have been recognized as duplicated in an analysis that considered 41 genomes and used the same methodology to build gene families, with a range going from 45.5% in a moss to 84.4% in the apple tree [17]. It is possible to estimate that several WGD events took place during the evolution of plant species, the most ancient happening in the ancestor of all seed plants about 319 million years ago and another more recent before the diversification of angiosperms 192 million years ago [18]. A large number of WGD are also consecutive to recent events. For example, the wheat group has evolved through different complex hybridizations among species from the plant genera *Aegilops* and *Triticum* followed by genome doubling. The most recent event giving birth to allotetraploid wheat (two different diploid parental species) has been proposed to occur about 300,000 to 500,000 years ago, while an allohexaploid wheat (three different diploid parental species) was formed only about 10,000 years ago [19]. Another domesticated plant, the Oilseed crop (*Brassica napus* L.) originated between around 6700 to 51,000 years ago by hybridization between two species, which were themselves

polyploids [20–23]. The consequences of the different types of hybridization, and thus WGD, are that many plants arising from these processes have very large genomes. On the contrary, in other organisms, there are still some debates concerning the occurrence of WGD versus several more local duplications. This is the case in vertebrates in which the “2R hypothesis”, originally proposed by Susumu Ohno [1], assumes the existence of two rounds of WGD in their early evolution. The “2R hypothesis” has been the subject of numerous studies to prove this theory. This has led to numerous works published during the last twenty years either in favor of the “2R hypothesis”, or in favor of only one round of WGD, or rejecting any idea of WGD (see for a review [24]). The main reason explaining the difficulty to determine whether two rounds of WGD happened or not very anciently comes from two phenomena which could blur the signal. Both phenomena make it harder to detect ancient WGD either through the loss of signal (fractionation) or increased complexity (diploidization). The fractionation is characterized by a heavy loss of duplicated genes following WGD [25]. The diploidization refers to the chromosomal rearrangements and segment loss often observed after WGD when the genome goes back to a diploid state [26]. Indeed, a return to diploidization involves the transition to disomic inheritance as it has been proposed in Salmonid species, for example [27]. In a recent work, 61 animal genomes were used to reconstruct the gene order of the ancestral Amniota genome, to identify duplicated genes produced by the 2R in this genome, and to reconstruct the timeline of events conducting a pre-vertebrate genome going from 17 chromosomes to 54 after the occurrence of two successive WGD [28]. Although a lot of arguments seem now to be more in favor of the “2R hypothesis”, the question is still not completely resolved. Very recently, an investigation using phylogenetic approaches and tree topology comparisons of gene families containing at least three members and located on several human chromosomes led to the conclusion that small-scale duplication (SSD) events scattered on all the animal history were more likely to be involved in vertebrate genome evolution rather than WGD [29].

### 2.1.2. Tandem Duplications

At smaller scales, local events called tandem duplication, create a novel copy of a gene next to it producing tandemly arrayed genes (TAGs) (Figure 1B). The molecular mechanism involved consists in unequal crossing overs, which can produce regions containing one or several genes, depending on the position of the breakage on the chromosomes [30]. These unequal crossing overs are either the result of homologous recombination between sequences (on homologous chromosomes or on sister chromatids) or of non-homologous recombination by replication-dependent chromosome breakages [31]. When multiple occurrences of unequal crossovers happen, it might lead to increasing or decreasing copy numbers in gene families. The molecular mechanism allowing the recombination depends on the sequences that promote the exchange between chromosomes or chromatids, which can be long direct repeats (>100 bp) and short ones (>12 bp) [32]. When repeats are long, the tandem duplication can arise via the homologous recombination whereas when they are short, duplication arises by single-strand annealing, template switching, or non-homologous end joining. This type of duplication leads to the formation of clusters of duplicated genes sometimes representing specific gene families. For example, this mechanism has been shown to confer soybean resistance against cyst nematode (*Heterodera glycines*) at *Rhg1*, a quantitative trait locus on chromosome 18, by changing the copy number variation that increases the gene expression [33]. In maize, thousands of tandem gene duplicates were identified that correspond to about 10% of the annotated genes [34]. Some of them may contribute to a phenotypic variation such as the *White Cap locus*, which provided the possibility to select white-grain color [35].

### 2.1.3. Duplications Via the Action of Transposable Elements

Duplicated sequences can also be formed by the action of transposable elements (TEs) according to different ways. TEs are repeated sequences with the ability to move from one position to another along and across chromosomes and which may represent a very large proportion in genomes, going from about 3% in yeast to more than 80% in maize [36,37]. When they are mobilized, some of them can drag

host sequences with them or can target the gene transcript, all of these having the consequence to duplicate the host sequences. There are two mechanisms by which TEs can promote duplication of complete genes or part of genes as a direct consequence of their transposition: The retroposition and the transduplication (Figure 1C,D). The retroposition mechanism consists of the reverse transcription of a messenger RNA from a host gene into a cDNA then inserted in another location of the genome by the action of the enzymes of a retrotransposon [38]. Genes submitted to this mechanism are located in the 3' side of retrotransposons and benefit from a transcription read-through initiated inside the TE [39,40]. This new gene, that is called a retrocopy, has particular features such as the presence of a polyA tail in its 3' end, the loss of introns, and the presence of target site duplication at both extremities which are the signature of its insertion. Retrocopies have been discovered in different organisms such as in mammals, and especially in the human genome where thousands of them have been identified [41,42]. Although less numerous, retrocopies have also been identified in insects such as in *Drosophila melanogaster* [43,44], or in the mosquito *Anopheles gambiae* [45]. Interestingly, it has been observed a bias in the location in the genome of these retrocopies which move from the X chromosome toward the autosomes in the insects [43,45]. In mammals, X chromosomes seem to have generated and recruited more retrocopies than the other chromosomes [46]. This type of duplicated gene is also found in plant genomes. For example, in the rice genome (*Oryza sativa*), between 491 and 1235 retrocopies were identified according to the methodology [47,48]. In *Arabidopsis thaliana*, 271 retrocopies were identified [48]. The other mechanism that involves TEs, the transduplication, happens when DNA transposons incorporate unspliced fragments of different genes, although the true mechanism is still unknown [49]. The gene fragments may still contain introns. First discovered in maize, this mechanism has then been documented only in plants such as *A. thaliana*, Japanese morning glory, soybean, and rice [49–54]. In rice especially, a particular type of DNA transposons called Pack-MULE, which represent about 3000 insertions in the genome, has been shown to contain sequence fragments derived from more than a thousand genes [54].

#### 2.1.4. Segmental Duplications

At a larger scale, segmental duplications, also called “low copy repeats”, correspond to very long stretches of duplicated sequences that can span between 1 to 200 kb and that share a sequence identity higher than 95% (Figure 1E; [55]). They have been first observed in several eukaryotic organisms such as the yeast [56] and humans [57]. These duplications are formed from the replicative transpositions of small portions of chromosomes. However, the exact mechanism is unclear and the fact that these duplications do not generate tandem repeats and that no short direct repeats at junction have been found suggests that neither unequal crossing-overs nor double-stranded breakages followed by repair are involved [55]. It has been proposed that in yeast, the segmental duplications could result from replication accidents [58] and that most of these sequences present a certain level of instability that can be rescued when translocation within another chromosome happens [59]. In *Drosophila*, high enrichment in TEs at segmental duplication extremities have been observed, indicating their possible implication in the duplication formation by homologous repair ends [60]. Similarly in mammals, particular types of TEs were found to be enriched at the junction of segmental duplications [61,62]. In the human genome, the sequence divergence of the duplicated segments has been used to estimate their evolutionary age which corresponds to the divergence between the New and Old World monkeys, 35 million years ago [63]. Segmental duplications account for an average of 13.7% of the total human genome, located in pericentromeric and subtelomeric regions [64]. Moreover, some chromosomes seem to be enriched in duplicated segments of this type such as the Y chromosome where they represent 50.4% of this chromosome [64]. In other mammals such as rat, mouse, or dog, this type of duplication is less abundant [64]. The comparative analysis of several genomes of Lepidoptera species made it possible to determine a large variation in the content of segmental duplications, going from 1.2% in the silkworm (*Bombyx mori*) to 15.2% in the postman butterfly (*Heliconius melpomene*) [65].



### 2.1.5. Differences among Duplication Types

Notable differences depending on the formation mechanisms in terms of function, expression, evolutionary constraints, and protein interactions have been reported. For example, in yeast duplicated genes issued from WGD are associated with different sets of functions when compared to duplicated genes generated by SSDs [66,67]. This has also been shown in plants [68–72]. In *Arabidopsis* and rice, for example, TAGs were found to be enriched with genes that encode membrane proteins and with functions in “abiotic and biotic stress” when compared to other duplicated genes. TAGs were also underrepresented in genes involved in transcription and DNA or RNA binding functions compared to non-TAG duplicated genes [73]. More recently, Acharya et al. [74] reported a higher multifunctionality, estimated by the number of GO and Pfam annotations, for WGD duplicated genes compared to SSD genes in humans. They also observed a significantly higher proportion of essential genes among the WGD genes relative to SSD genes.

It has also been observed that duplicated genes differ in divergence of expression according to the mode of duplication. In *Arabidopsis* and in poplar, for example, WGD genes were found to display a lower divergence of expression than other duplicated genes [71,75]. In a study deciphering more deeply the different types of duplicated genes, Wang et al. [48] observed that in *Arabidopsis* and rice, WGD genes and TAGs displayed a lower divergence of expression than proximal, retrotransposed dispersed, and DNA based transposed duplicated genes.

In a recent study in Angiosperms, WGD duplicated genes were shown to be under stronger constraints to diverge at the sequence and expression level relative to SSDs [76]. It has also been observed that among WGD genes, those that are also involved in local duplications showed higher non synonymous substitution rates (Ka) and selection rates (Ka/Ks) than nonlocally duplicated WGD genes indicating that they evolve faster [77].

When considering protein-protein interactions (PPI) networks, it has been observed that the fraction of shared PPI between paralogous genes was higher when the genes shared the same function and showed a higher co-expression [78]. Among duplicated genes, WGD gene pairs displayed a higher fraction of shared PPIs than other duplicated gene pairs [78]. Arsovski et al. [79] examined the density of *Arabidopsis* DNaseI footprints, which are locations of protein binding sites, in the 1000 bp flanking upstream and downstream sequences of duplicated genes. They found that WGD duplicated genes had more footprints than TAGs. Moreover, WGD duplicated genes formed denser and more complex regulatory networks than TAGs when genome-wide regulatory networks were analyzed.

In summary, mechanisms that can lead to the formation of duplicated genes are various. The fates encountered by the new duplicated genes are also distinct and may depend on several factors.

## 2.2. Evolutionary Fates of Duplicated Genes

### 2.2.1. Pseudogenization and Neo-Functionalization

After their formation, duplicated genes can encounter various fates (for a complete review on this matter, see [80]). The most likely is the pseudogenization or the complete loss of one copy since only one gene copy will continue to be under purifying selective constraints for its current function, leaving the other one free to accumulate deleterious mutations. These pseudogenes can be conserved in the genome. For example, *A. thaliana* and the rice contain thousands of pseudogenes in their genomes [81]. In humans, the olfactory receptor gene families have been shown to be composed of between 60–70% pseudogenes whereas in dogs pseudogenes represent less than 20% in those gene families, explaining the reduced sense of smell in humans [82,83]. Sometimes, however, the process of mutation accumulation can drive to a completely different outcome. Different models of population genetics have been proposed to highlight evolutionary mechanisms explaining the different fates of duplicated genes allowing them to be maintained in organisms (for specific reviews on this subject, see [84,85]). It has been proposed that three main steps are needed for duplicated genes to be maintained: Phase 1 consists of the origin of a genetic change through mutation, phase 2 corresponds

to the fixation period when the mutation segregates in the population, and phase 3 corresponds to the preservation period where the duplication is conserved. Although infrequent, a mutation can provide a new allele giving rise to a new function for the gene copy. If this function is advantageous, it will be subjected to distinct selective constraints leading to its fixation in the population, in a process called neo-functionalization. There are two models to explain this mechanism. The Dykhuzen-Hartl model proposes that the mutations at the duplicated gene are fixed by drift and later, during a change in the environment, the new gene will become advantageous for the organism [86], whereas the “Adaptation model” proposes that an adaptive mutation is fixed at one of the duplicated locus because it is immediately advantageous [85]. Various examples of neo-functionalization have been described. The analysis of the copper transporter gene family, which contains the two genes *Ctr1* and *Ctr2*, suggested that the metazoan *Ctr2* arose several hundred million years ago via a duplication event of the *Ctr1* genomic locus. The resulting *Ctr2* then lost the ability to transport copper but gained the ability to regulate *Ctr1* cleavage [87]. In mammals, the family of retinoic acid receptors (RARs), which play a role in the embryonic development, contains three duplicated genes, *RAR $\alpha$* , *RAR $\beta$* , and *RAR $\gamma$* , with *RAR $\beta$*  having kept the ancestral RAR role, while the two others have diverged both in ligand-binding capacity and in expression patterns suggesting that neo-functionalization occurred at both the expression and the functional levels for these genes [88]. A wide transcriptomic analysis in maize made it possible to determine that 13% of all gene pairs generated by WGD have been submitted to regulatory neo-functionalization in leaves [89]. The analysis of a gene family containing three members in the *D. melanogaster* genome made it possible to show that the family was created by two rounds of tandem gene duplication in the last five million years and that the two new duplicated copies have diverged in function from the parental copy [90].

#### 2.2.2. Sub-Functionalization and Functional Redundancy

Alternatively to the possibilities of pseudogenization and neo-functionalization, the duplicated genes can be submitted to sub-functionalization. In this process, accumulation of mutations drives the subdivision of the ancestral gene function among the duplicated genes. The complementarity can come from a change in the regulatory sequences, leading the two copies to have different expression patterns that will recapitulate the ancestral one when taken together, for example [91]. Several models have been proposed to explain this mechanism. In the first model called duplication–degeneration–complementation (DDC) the two gene copies will acquire complementary functions through independent mutations, which will lead to the need of the two copies to fulfill the original function by drift rather than by selective constraints [91]. Another possibility is described by the “gene sharing” model in which the acquisition of two expression domains could predate the duplication, with each copy losing one of the two afterward [92,93]. A close model corresponds to the “specialization” model [94] which proposes that an ancestral function is split among paralogs that will be expressed in different tissues or developmental stages. These two last models predict that the duplication will be followed by advantageous mutations in all duplicated genes with positive selection patterns detectable in their sequences. Moreover, it is supposed that the ancestral gene is able to fulfill the function of all duplicated genes but not so well. Numerous examples of sub-functionalization have been identified in eukaryotes. For example, in mammals, the Agouti-melanocortin system is represented by the Agouti protein (ASIP) and the Agouti-related protein (AgRP) whose expression patterns with distinct physiologic functions were acquired through sub-functionalization such that the current expression pattern and function of each protein correspond to a subset of the ancestral gene [95]. In tomato, two members of the gene family encoding phytochromes, which are light receptors playing a role in plant development, exhibit both common and non-redundant functions suggesting that they have sub-functionalized since their duplication [96]. Finally, it is also possible for the two copies of a gene to be both maintained in the genome by dosage subfunctionalization, each expressing the ancestral function, leading to a functional redundancy [97,98]. A model proposed to explain this possibility stipulates that expression reduction could help the retention of duplicates and the conservation of

their ancestral function [99]. Several cases have been identified such as, for example, two members of the mammalian *HOX* gene complex, *Hoxa3* and *Hoxd3*, implicated in the embryonic development, that have been shown to display a similar function in mice [100]. In the yeast *S. cerevisiae*, duplicated genes were shown to maintain functional redundancy for several million years [101].

### 2.2.3. The Fates of Duplicated Genes Depend on Different Factors

These different fates can be conditioned by the mechanism that led to the formation of the duplicated genes. Indeed, it was suggested that tandem duplication could more often produce duplicated genes having differential partitioning of regulatory sequences which implies that both genes would be necessary to recapitulate the ancestral expression pattern [102]. In *A. thaliana*, it was proposed that pseudogenes are more often derived from tandem duplications although this could be a bias due to the higher proportion of this type of mechanism compared to others in this organism [70]. The fate of retrocopies is often to become pseudogenes because of the lack of regulatory sequences [38]. However, it is sometimes possible for retrocopies to recruit other regulatory sequences allowing them to develop a new function. The structure of these retrogenes is usually chimeric with coding or regulatory features not present in the original genes [43,103–106]. Moreover, it has been observed in mammals that retrocopies located on the same chromosome than their parental gene have more chance to remain active indicating a role for the genomic context to maintain their expression [107]. In plants, a positive correlation has been observed between the size of gene family and the number of pseudogenes, with large families being more subjected to gene loss [81]. However, the gene function is also an important factor in the fate of duplicated genes. Indeed, in *A. thaliana*, pseudogenes tend to have functional counterparts in disease resistance, specialized metabolism cell wall modification, and protein degradation, whereas transcription factor and receptor-like kinase gene families are devoid of pseudogenes [70,81]. Other factors may also influence the fate of duplicated genes such as the number of protein interactions [76,108] as well as particular structural features [109]. According to the organisms, the outcome and formation mechanism of duplicated genes can also be different. In human and mouse, for example, the relative contributions of two types of duplication mechanisms made it possible to show that tandem duplications contributed more to duplications in the entire genome than retroposition, except for the two-copy gene families, and generated duplicated genes with more chance to be retained [110]. At another scale in primates, recent duplicated genes originated more often from segmental duplication than in other mammals in which the main mechanism to generate them rather corresponds to tandem duplication [111]. WGD in humans was proposed to have generated duplicated genes functionally more divergent but with a higher proportion of essential genes, which is the opposite trend to what was observed in yeast [74]. In *Drosophila*, young duplicated genes were shown to be preferentially subjected to neo-functionalization, implying the retention of almost two-thirds of these duplicated genes [112]. In plants, where most duplicated genes are derived from WGD and tandem duplication, a functional bias can be observed in genes according to their mechanism of formation [70]. Thus, genes involved in responses to environmental stimuli and upregulated in stress conditions are rather generated by tandem duplication, which implies that this mechanism is important for adaptive evolution in changing environments. Recently, a model was proposed to explain the gene retention after WGD in *Paramecium* species by dosage constraints, i.e., the majority of duplicated genes keep their ancestral function and are retained to produce the requested amount of proteins to perform this ancestral function [98].

In the next section, we will present in detail some of the current bioinformatic methodologies available to identify and analyze duplication in genomes with the goal to emphasize their advantages and weaknesses according to the situation.

## 3. Bioinformatic Approaches to Identify Duplications in Genomes

The identification of duplication within or between genomes is a complex process. Many algorithms have been developed for this purpose and different approaches can be used that have different aims

and computation costs. Moreover, some of them are more suitable to search for a particular duplication event, are more optimized for large genomes, can deal with multiple genomes, or can handle genomes that have undergone multiple duplication and rearrangement events. In addition, there may be difficulties in the installation, the configuration, the launch, and the parsing of the results. This means for the user that programming skills may be required to use some of these softwares. There are also variations in the input data and the pre-processing requirements, the computing time or the associated visualizations, all of this making the choice of a tool not easy. Moreover, these tools do not all identify the same type of duplication and may therefore, be more or less adapted according to the biological question investigated. In summary, there is no stand-alone software that can solve all these problems and the choice of the tool will depend on computer skills but also on the genomes being compared and the biological questions being asked. In the following sections, we will present the different types of algorithms highlighting their specificities, advantages and weaknesses, with a focus on some tools that will be presented in more detail.

### 3.1. Paralog Detection

As said before, homologs are genes that share a common ancestry and are divided between orthologs (derived by speciation) and paralogs (derived by duplication). Based on this definition, the search for duplicated genes can be done through the identification of paralogous relationships. Therefore, it can be conducted by either identifying homologous genes in a given genome, which by definition can only be paralogous, or between multiple genomes before distinguishing orthologs from paralogs. Several approaches exist to this aim that we will present below.

#### 3.1.1. Homology Assessment

Homology, even if defined by a few words, is a challenging concept to be detected through bioinformatic tools (for a broad overview, see [13,113]). The only material given to us to infer common ancestry that may have started millions of years ago is the sequences of contemporary organisms. A notable exception to this limit came with the rise of paleogenomics which aims at sequencing genomes of extinct species through preserved elements such as ancient seeds or fossilized body parts [114]. However, even if paleogenomics provides useful information, the amount of material is scarce compared to the number of contemporary species. Two methods are typically used to assess the homology between genes: The sequence similarity and the gene structure. Both methods rely on the idea that common ancestry (i.e., homology) is the most likely explanation when two genes share a strong similarity and/or structure. The limitations of these methods account for the aforementioned problem of inferring history through present traces: Divergence becomes difficult to detect when the distance between species increases. Hence, when two genes share sequence similarity or structure, it is a strong indication of homology, but when two genes do not share those, it hardly says anything about their homology.

The sequence similarity can be tested with a sequence alignment algorithm. The most popular ones such as *BLAST* [115], *Psi-BLAST* [116], and *HMMER3* [117] are heuristic methods. Thus, they might not give the best results, but they drastically save computational time compared to a classical method such as the Smith and Waterman algorithm [118] even though some implementations have tried to make it faster as *PARALIGN* [119] or *SWIMM* [120]. In the case of homology, the alignment is generally performed on protein sequence instead of the gene. This allows a greater sensitivity since amino acid substitutions occur less frequently than nucleotide substitutions allowing silent mutations and because introns generate a lot of noise [121]. With these methods, the homology is tested against a cutoff on three different metrics: E-value, bit-score, or percent identity. The e-value is a statistic representing the expected number of times a given alignment score would occur by chance given the length and number of sequences being aligned. It is the most widely used metric as a first step to assess homology. Since the e-value is dependent on the database size, a potential caveat when setting a cutoff is to apprehend how the results might change for different databases. The bit-score is another metric

measuring the sequence similarity given the raw score and the score system used but independent of the length or the number of sequences being aligned. The bit-score might be preferred in the case of a comparison between alignments since it relies only on the two sequences being aligned. Finally, percent identity is a straightforward metric giving how many amino acids are identical in the local alignment. When assessing homology on a genome-wide scale, the difficulty resides in setting the right cutoff for these metrics. For instance, to capture duplicates that diverged in function, the threshold needs to be relaxed, but with the risk of increasing the number of false positives. Based on empirical results, a 30% identity is generally accepted as a significant cutoff for protein homology [122]. However, countless identified homologs have an identity percentage lower than 30%. The same problem arises when using only e-value or bit-score. To allow better identification, more complex similarity-based metrics were developed. For example, Rost [123] proposed a formula based on the homology-derived secondary structure of proteins (HSSP) curve defined by Sander and Schneider [122] and considering the number  $L$  of aligned residues between two proteins to define a curve to separate true and false positives. Two proteins are then considered homologous if the proportion  $p$  of identical residues over  $L$  aligned residues is higher than the cutoff point defined by the formula. Li et al. [124] proposed a rewording of Rost's formula to define different sets of duplicated genes with different stringencies in human. Since a gold standard cutoff is impossible to determine, a variety of values are used, sometimes combining different metrics leading to different results (Table 1).

**Table 1.** Estimation of the amount of duplicated genes in different species.

Species	No. of Considered Genes	No. of Estimated Duplicated Genes	% Estimated Duplicated Genes	Methodology	Duplicated Gene Types	References
	25,557	11,937	46.7	All-against-all nucleotide sequence similarity searches using <i>BLASTN</i> among the transcribed sequences. Sequences aligned over >300 bp and showing at least 40% identity were defined as pairs of paralogs.	Not specified, all paralogous pairs were searched	[125]
	27,558	12,761	46.3 *	All-against-all protein sequence similarity search using <i>BLASTP</i> (e-value cutoff of e-10). Sequences alignable over a length of 150 amino acids with an identity of 30% were defined as paralogs. Gene families were built through single-linkage clustering.	Not specified, genes families were obtained	[69]
	25,972	10,483–17,406	40.4–67	All-against-all protein sequence similarity search using <i>BLASTP</i> (e-value cutoff of 1.0). For each pair of genes, blast-hits were merged to compute the total length and the global similarity of the aligned regions. Two datasets were constructed with respectively 30 and 50% sequence identity over respectively 70 and 90% protein length. Gene families were built through single-linkage clustering.	Not specified, genes families were all obtained (gene families)	[73]
<i>Arabidopsis thaliana</i>	22,810	21,622	94.8 *	All-against-all protein sequence similarity search using <i>BLASTP</i> (top five non-self protein matches with e-value of 10e-10 were considered). Genes without hits that met a threshold of e-value 10e-10 were deemed singletons. Pairs of WGD duplicates were downloaded from published lists. Single gene duplications were derived by excluding pairs of WGD duplicates from the population of gene duplications. Tandem duplications were defined as being adjacent to each other on the same chromosome. Proximal duplications were defined as non-tandem genes on the same chromosome with no more than 20 annotated genes between each other. Single gene transposed-duplications were searched for from the remaining single gene duplications using syntenic blocks within and between 10 species to determine the ancestral locus. If the parental copy had more than two exons and the transposed copy was intronless, the pair of duplicates was classified as coming from a retrotransposition. Other cases of single gene-transposed duplications were classified as DNA based transpositions. Dispersed duplications corresponded to the remaining duplications not classified as WGD, tandem, proximal, or transposed duplications.	WGD, tandem, proximal, DNA based transposed, retrotransposed, and dispersed duplications	[48]

Table 1. Cont.

Species	No. of Considered Genes	No. of Estimated Duplicated Genes	% Estimated Duplicated Genes	Methodology	Duplicated Gene Types	References
<i>Homo sapiens</i> (human)	33,869->19,727	12,981	65.8	All-against-all protein sequence similarity search using <i>BLASTP</i> with the BLOSUM62 matrix and the SEG filter [126], <i>TribeMCL</i> with the default parameters. Tandem duplications were then searched for among families.	Gene families (tandem duplications searched among families)	[127]
	13,298	11,386	85–97	All-against-all protein sequence similarity search using <i>BLASTP</i> with cutoff expectation <2 and <10 <sup>-3</sup> .	Not specified, distant duplicates	[128]
	31,126	14,473	46.5 *	Ensembl family database and genes >300 nt. Tandem duplications were then searched for among families.	Gene families (tandem duplications searched for among families)	[129]
	20,415	15,569	76.3	Pooling of different datasets from [130] and all-against-all protein sequence similarity search using <i>BLASTP</i> .	WGD and SSD	[131]
	22,447	11,740	52.3 *	Ensembl version 77, >50% sequence identity, and high confidence for paralogy.	WGD and SSD	[74]
<i>Mus musculus</i> (mouse)	21,305	14,043	65.9	All-against-all protein sequence similarity search using <i>BLASTP</i> with the BLOSUM62 matrix and the SEG filter [126], <i>TribeMCL</i> with the default parameters. Tandem duplications were then searched for among families.	Gene families (tandem duplications searched for among families)	[127]
	27,736	16,091	58.01	Ensembl family database and genes >300 nt. Tandem duplications were then searched for among families.	Gene families (tandem duplications were searched for among families)	[129]

Table 1. Cont.

Species	No. of Considered Genes	No. of Estimated Duplicated Genes	% Estimated Duplicated Genes	Methodology	Duplicated Gene Types	References
<i>Rattus norvegicus</i> (rat)	18,468	12,466	67.5	All-against-all protein sequence similarity search using <i>BLASTP</i> with the BLOSUM62 matrix and the SEG filter [126], <i>TribeMCL</i> with the default parameters. Tandem duplications were then searched for.	Gene families (tandem duplications searched for among families)	[127]
	27,194	16,446	60.48 *	Ensembl family database and genes >300 nt. Tandem duplications were then searched for among families.	Gene families (tandem duplications searched for among families)	[129]
	18,562	9149	49.3	All-against-all nucleotide sequence similarity searches using <i>BLASTN</i> were done among the transcribed sequences. Sequences aligned over >300 bp and showing at least 40% identity were defined as pairs of paralogs.	Not specified, all paralogous pairs were searched	[125]
	42,534	8244–19,322	19.4–45.4	All-against-all protein sequence similarity search using <i>BLASTP</i> (e-value cutoff of 1.0). For each pair of genes, blast-hits were merged to compute the total length and the global similarity of the aligned regions. Two datasets were constructed with respectively 30 and 50% sequence identity over respectively 70 and 90% protein length. Gene families were built through single-linkage clustering.	Not specified, genes families were all obtained (gene families)	[73]
<i>Oryza sativa</i> (rice)	27,910	21,461	76.9 *	All-against-all protein sequence similarity search using <i>BLASTP</i> (top five non-self protein matches with e-value of 10e-10 were considered). Genes without hits that met a threshold of e-value 10e-10 were deemed singletons. Pairs of WGD duplicates were downloaded from published lists. Single gene duplications were derived by excluding pairs of WGD duplicates from the population of gene duplications. Tandem duplications were defined as being adjacent to each other on the same chromosome. Proximal duplications were defined as non-tandem genes on the same chromosome with no more than 20 annotated genes between each other. Single gene transposed-duplications were searched for from the remaining single gene duplications using syntenic blocks within and between 10 species to determine the ancestral locus. If the parental copy had more than two exons and the transposed copy was intronless, the pair of duplications was classified as coming from a retrotransposition. Other cases of single gene-transposed duplications were classified as DNA based transpositions. Dispersed duplications corresponded to the remaining duplications not classified as WGD, tandem, proximal, or transposed duplications.	WGD, tandem, proximal, DNA based transposed, retrotransposed, and dispersed duplications	[48]

\* These values have been calculated according to the information provided in the corresponding reference article.



When not working on a genome-scale but on specific sequences, homology imputation can be reinforced by looking at the gene structure. Domains shared by proteins are strong indicators of homology. Conserved domains can be found in databases such as Pfam [132] or InterPro [133] and searched against sequences of interest. This method is also a great tool to unravel complex evolution such as gene splitting and fusion for multi-domain proteins but require a time consuming manual expertise.

When searching for duplicated genes within a genome, assessing homology inside this genome is enough. However, when comparing multiple genomes, a link needs to be made between homologs of the different genomes. This raises the issue of resolving ortholog and paralog relationships. For this, a different kind of method needs to be applied. At first, methods to identify orthologous genes were only constructing orthologous groups because they focused on one-to-one ortholog relationships across multiple species. However, with the addition of one-to-many and many-to-many relationships, paralogs were included. Therefore, it could be argued that these methods are eligible to detect duplicated genes across multiple genomes. They are generally split into two categories: The graph-based methods and the tree-based methods [14,134]. Generally, graph-based methods construct a homology graph then build clusters of genes based on the types of inferred relationships. On the contrary, tree-based methods identify clusters of genes before constructing a tree along which the types of relationships are inferred.

### 3.1.2. Multispecies Graph-Based Methods

In graph-based methods, each gene is a vertex and a homology relationship is depicted by an edge. These edges are first drawn by assessing sequence similarity in the various forms described before. At this step, edges only correspond to potential homology relationships which can be orthology, paralogy, or noise. The noise can be removed by the clustering step. Depending on the clustering method, some paralogous relationships can also be removed. It is important to note that for the resolution of ortholog and paralog relationships, all these methods consider that for a given speciation event, in-paralogs are less diverged than orthologs that are less diverged than out-paralogs.

One of the first proposed clustering methods was the identification of triangle patterns inside a graph where at least three genomes are used [135]. It relies on the idea that two similar genes from two genomes, which are also similar to a third gene from another genome are highly susceptible to be orthologs. Then, triangles sharing similar edges are added to the same group until no other can be added. These groups, called clusters of orthologous groups (COGs) can therefore contain paralogs. However, the nature of the paralogs included in a group is hard to control. Hence, another way to detect paralogs based on graph exploration was proposed with *InParanoid* [136]. Here, two genes from different species with a best reciprocal hit are defined as orthologs and will be used as a seed for the group. Any gene having a better score with the seed gene of the same species than with the seed gene from the other species is included inside the group as an in-paralog relative to the speciation event. Thus, only in-paralogs in regard to the speciation event considered should be added, allowing a better control over the group formation. The method *Hieranoid* expanded this idea with the use of a guiding species tree for a better scalability when using many species [137]. The algorithm enlarges groups by exploring the guiding tree. It first runs *InParanoid* between two closely related species. Then, it creates a pseudo-species where each identified homologous group is represented by either a consensus sequence or a Hidden Markovian Model profile, depending on the number of sequences. *InParanoid* is then used again between the pseudo-species and the next closest neighbor. The process is repeated until all species are included in the analysis. By keeping track of groups formed at each step, it is possible to identify the speciation event encompassing any in-paralog pairs. Acting as a synthesis between *InParanoid* and COGs, both *eggNOG* [138] and *OrthoDB* [139] start by identifying groups of in-paralogs for each species then link them between species using triangulation.

Considering that the *InParanoid* method was reliable to detect “ancient” paralogs but not “recent” ones, Li et al. added steps and proposed another method, *OrthoMCL* [140]. It begins by the same ortholog seed approach but with the constraint that in-paralogs must have a better score with the seed genes from their respective species than with any other sequences from any species. In addition,

a Markovian Cluster algorithm is run to simulate a random walk on the graph with each edge having a transition probability depending on the similarity score. This makes it possible to identify robust subgraphs and notably separate diverged paralogs. Using also a similar approach than *InParanoid*, the method *OrthoInspector* starts by constructing species-wise in-paralogous groups. Inside a species, an in-paralogous group is inferred for each protein [141]. Inside a species, a group of potential in-paralogs is inferred for each protein. When two proteins are potential in-paralogs, only the intersection of their respective potential groups is conserved as the final in-paralogous group. Therefore, if we have three proteins A, B, and C, they will belong to the definitive in-paralogous group (A, B, C) if and only if all three potential in-paralogous groups constructed for each protein give (A, B, C). This stringent method creates groups of lowly diverged in-paralogs. In-paralogous groups or single proteins are then grouped between species based on best-reciprocal hits.

Finally, two other methods add an interesting consideration regarding homologs. Aiming to tackle a well-known problem of sequence alignment, *OrthoFinder* [142] allows a reliable incorporation of short sequences. Indeed, alignment score is correlated with the sequence length, which is a problem for short sequences giving high scores even when not related. *OrthoFinder* proposes a normalization of the alignment score after a grouping according to the sequence lengths into equally sized bins. This normalization makes the score for short and really long sequences less dependent on the sequence size. Another interesting method, *OMA* [143], proposes to detect falsely imputed orthology inferences due to paralogs with differential gene loss. The detection is performed by using a third species containing both paralogs which acts as an evidence of non-orthology. *OMA* is also more permissive in the grouping of paralogs because it takes into account that paralogs may evolve faster than orthologs [144].

When studying genes, especially across species, representing their evolutionary relationships as a tree is easier to analyze. However, constructing such a tree is done at the cost of computational time. In addition, different strategies can be adopted for the tree reconstruction.

### 3.1.3. Multispecies Tree-Based Methods

In tree-based methods, homology is assessed according to the various forms described before, then groups of homologs are constructed across species. Genes from these groups are aligned to build gene trees. Paralog and ortholog relationships are then resolved by the reconciliation of the gene trees and the associated species tree. Therefore, in these methods, the detection of duplicated genes is only performed at the first step. The tree construction only influences the evolutionary history used to explain the appearance of such duplications.

In regards to the homolog grouping strategy, tree-based methods are generally more inclusive than graph-based methods. Indeed, after the group construction, they use all the sequences from all the species to infer paralog and ortholog relationships. Therefore, they can extract more information and are less restricted by false homology prediction and thus are able to capture more diverged homologs. Most of them construct homologous groups by clustering all genes that have a significant alignment score, defined differently according to the method used such as *TreeFam* [145], *BranchClust* [146], *HOGENOM* [147], or *PhylomeDB* [148]. Some tree-based methods use pre-processed homologous groups and are only used to reconcile the gene and species trees such as *Orthostrapper* [149], *Softparsmap* [150], or *LOFT* [151]. Therefore, graph-based methods can be used as an entry-point to combine the power of both methods.

When reconciling the gene and species trees, all these methods use the Maximum Parsimony principle [152]. This is translated by minimizing the number of duplication events, which are assumed to be rare events. A notable exception is *PrIME-GSR* [153] that tries to take into account the duplication and loss of genes through a probabilistic model. Apart from this exception, tree-based methods differ according to the type of species tree they accept, how they root the gene trees, and how tree uncertainty is assessed. Since it does not affect duplication detection, they are not as thoroughly explored as the graph-based method (for a complete review, see [14]).

### 3.2. Detection of Syntenic Blocks (WGD-Segmental Duplications)

A syntenic block can be defined as a region of the genome spanning a number of genes that are orthologous and co-arranged compared to another genome [154]. Two regions of a genome with a number of homologous genes co-arranged with each other can also be defined as a syntenic block. Here, we focus on this second definition because pairs of homologous genes between these pairs of regions correspond to duplicated genes.

It can be interesting to access different databases storing pre-calculated syntenic blocks shared between different species. This makes it possible for an easy and direct access to reliable information without any computation. Nevertheless, these databases cannot include every contemporary species nor information about recently released genomes. This implies that depending on the organism being studied it can be necessary to manually identify syntenic blocks using different tools. To accurately detect homologous chromosomal segments within a genome or between different ones, many approaches and tools are available. The choice of the tool depends on various parameters.

A first important parameter is the degree of preservation of duplicates in the compared genomes. This will influence the level at which the study should be conducted, and thus will impact the choice of the tool since each of them works at a particular level. For closely related genomes, synteny can be studied at the DNA level using tools such as *Satsuma* [155] or *SyMap 3.4* [156]. In the case of more distant genomes, the DNA level cannot be used because the sequences will be too divergent. A solution is to perform analysis at the protein level because coding genes may retain for a longer time enough amino-acid sequence similarity and a similar relative order along chromosomes. Tools such as *MCSanX* [157], *i-ADHoRe* [158], *CYNTENATOR* [159], or *SynChro* [160] search for syntenic blocks using protein sequences and can therefore be adapted to this type of genome comparison. Finally, in the case of more distant genomes, it is more appealing to use tools based on analyses at the protein level and on the construction of profiles, graphs, or statistical models to help manage the evolutionary distance.

Four types of approaches can be applied to search for syntenic blocks. The first one is based on the construction of a sparse matrix of homologous genes. The matrix is investigated to look for dense diagonals which correspond to the syntenic blocks. Tools such as *i-ADHoRe 3.0* [158], *DiagHunter* [161], *FISH* [162], *SyMAP* [163], or *Cinteny* [164] implement this type of approach. The second approach corresponds to different greedy algorithms that will be optimized by dynamic programming at the benefit of computational costs. This type of algorithm operates by constructing chains of collinear gene pairs, called anchoring genes. It is implemented in tools such as *DAGChainer* [165], *MCSanX* [157], or *LineUp* [166]. An important sub-category of this approach consists of algorithms based on aligning sequences using a modified Smith-Waterman algorithm as in *CollinearScan* [167] or *CYNTENATOR* [159]. To continue, the graph approach aims at building graphs allowing the identification of the syntenic blocks. To do this, local collinear alignments are constructed between the input genomes. By combining the local alignments with the blocks, a graph can be constructed which allows, after different analyses, the reconstruction of the syntenic blocks. This approach can be found in tools such as *DRIMM-Synteny* [168] or *Enredo* [169]. Finally, another approach aims at inferring syntenic blocks based on genomic rearrangements. This type of approach can be useful in the reconstruction of ancestral genomes and has been implemented in different tools such as *GASTS* [170] or *PMAG++* [171,172]. This approach is not covered in the present review but has already been discussed in another recent review article [173].

All these tools are able to answer different questions and their use depends on the number of studied genomes as well as the level of divergence among them. Most of them have many critical parameters, sometimes with important pre-processing requirements, which need to be mastered before obtaining reliable results. Most of the tools are presented in a comprehensive format in Table 2. Therefore, the purpose of this section is to examine in detail a representative sample of tools illustrative of each approach.

**Table 2.** Summary of the characteristics of different existing tools for identifying syntenic blocks.

Name	Input	Output Text	Output Plots	Main Algorithm	Specificities	Other Information		Documentation	Programming Language	Interface	References
						Gene Orientation	Genome Number				
<i>i-ADHoRe 3.0</i>	BLASTP output or gene families and list of genes in a gff like format	Tabulated text	Graphical visualization	Custom Greedy Graph	Typical implementation of the collinearity strategy	Yes	N	Complete	C++ Wrapper in Python	Command line interface	[158]
<i>MCSamX-Transposed</i>	BLASTP output and a list of genes on chromosomes	Tabulated text	Graphical visualization	DAGChainer equivalent	Able to detect transposed gene duplications, detection of the type of duplicates	No	N	Incomplete and with errors	C++	Command line interface	[157]
<i>PhylDiag</i>	Species gene list and gene tree	Tabulated text	Graphical visualization	DAGChainer equivalent	Uses gene trees to define gene homologies. Takes into account gene orientations, and tandem duplication blocks	Yes	2	Complete	Python	Command line interface	[174]
<i>SynChro</i>	List of protein-coding genes and their associated amino-acid sequences	Text files containing homology relationships (RBH and non-RBH) and syntenic blocks description	Chromosomal painting representation, genome-wide dotplot	Computes Reciprocal Best-Hits (RBH) to reconstruct the backbones of the syntenic blocks and complete with non-RBH syntenic homologs	Only one parameter: the syntenic block stringency. Use OPSSCAN instead of BLAST due to its optimization to detect RBH	only in visualizations	N	Complete	Python, bash	Command line interface	[160]
<i>Satsuma</i>	Nucleic sequences	Tabulated text	Multiple interactive plots	Cross-correlation, implemented as a fast Fourier transform	Based on a search strategy at a global level and cross-correlation at the local level	Yes	2	Short	C++, on linux	Command line interface	[155]
<i>DAGchainer</i>	Homologous genes and associated E-value	Tabulated text	Dot plot	Identification of chains of ordered gene pairs by searching paths in directed acyclic graph	Use of dynamic programming making it fast and highly reliable. Many softwares are based on this algorithm	No	2	Short	C++, Perl	Command line interface, Graphical user interface	[165]
<i>ColinearScan</i>	Any type of genetic markers (physical or genetic distance between markers, gene numbers)	Tabulated text with syntenic blocks and associated p-value	None	Dynamic programming algorithm based on the Smith-Waterman algorithm	Statistical inference, high computational efficiency, and flexibility of input data types	No	2	Not available	C++, Perl	Command line interface	[167]
<i>CYNTENATOR</i>	Sequences or alignments and an annotation file	Text file gathering alignments	None	Profile-profile alignment setting, which is an extension of the Waterman-Eggert algorithm	Implementing a phylogenetic scoring function	-	N	Complete	C++	Command line interface	[159]

Table 2. Cont.

Name	Input	Output Text	Output Plots	Main Algorithm	Specificities	Other Information	Documentation	Programming Language	Interface	References	
<i>FISH</i>	List of the linear order and orientation of features on each contig and list of the pairwise homologies between features	Text file results	Dot Plot	Dynamic programming algorithm based on the Smith-Waterman algorithm	Modeling of the probability of observing segmental homologies assumed by chance and taking this model into account to parameterize the algorithm and the statistical evaluation of its output	Yes	2	Not available	C++	Command line interface	[162]
<i>DRIMM-Synteny</i>	Set of anchors (e.g., local alignments or pairs of similar genes)	Text file where each genome is represented as a shuffled sequence of the syntenic blocks	Dot Plot	Construction of A-Brujn graph	Graph-based algorithm allowing to identify non-overlapping syntenic blocks	No	N	Not available	C#	Command line interface	[168]
<i>DingHunter</i>	BLAST output	Two text files containing gene names and/or coordinates	Dot Plot	Homology matrix based algorithm	Typical implementation of the colinearity strategy. Identifies large-scale syntenic blocks despite high levels of background noise	No	2	Short	Perl, and requires the BioPerl and GD.pm modules	Command line interface	[161]
<i>OSfinder</i>	Genomic locations of anchor or BLASTP results	Genomic locations of chains and orthologous segments	Dot Plot and a synteny map	Machine Learning and Markov Chains	Use Markov chain models and machine learning techniques. Automatically optimizes the parameters used in the Markov chain models. Scoring scheme based on stochastic models	Yes	N	Complete	C++	Command line interface	[175]
<i>SyMap</i>	Genome sequences in FASTA format and associated GFF files	Homologous genes, diagonals, and identified syntenic blocks.	Visualization available and interactive	DAGChainer	Interactive visualizations. Calculates synonymous and nonsynonymous mutation rates for syntenic gene pairs using <i>CodeML</i> of the <i>PAML</i> package	No	N	Complete	No requirements	Web user interface	[163]

Table 2. Cont.

Name	Input	Output Text	Output Plots	Main Algorithm	Specificities	Other Information	Documentation	Programming Language	Interface	References	
<i>Cinteny</i>	Information about markers and the homologous groups.	Tabulated text	Three interactive visualizations Whole Genome Synteny, Chromosome Level Synteny, Synteny Around a Marker	Ternary search trees (TST)	On-the-fly computations allowing fast parameters adjustments	Yes	N	Complete	No requirements	Web user interface	[164]
<i>MultiSyn</i>	Protein sequences in FASTA format and genome annotation in BED	Output files from <i>MCSamX</i>	Multiple synteny plots	<i>MCSamX</i>	Efficient tool for non-programming skilled users. Precomputed data for 18 plant genomes	No	N	Not available	No requirement	Web user interface	[176]
<i>OrthoCluster</i>	Genome file and a file storing orthologous relationships among genes in all input genomes	Cluster file, with all the syntenic blocks detected, Stat file with information related to the size distribution of the syntenic blocks	One associated plot	Depth-first search method, can also use <i>Cinteny</i> or <i>SyMap</i>	Fast and easy to use. Can be applied using any types of markers as an input as long as their relationships can be established	Yes	N	Complete	C++	Web user interface, Command Line	[177]

N: Theoretically arbitrary number of studied genomes.

### 3.2.1. Approaches Based on the Construction of Homologous Gene Matrices

These approaches correspond to tools based on the search for synteny using clustering of neighboring matching gene pairs. The basic concept is to consider the homology in or between genomes as a sparse matrix. In summary, homologous gene pairs are considered as 1, other cases are encoded as 0. The goal is to detect syntenic regions by searching for dense diagonals of 1 in the matrix. Tandem duplication can also be accounted for by detecting horizontal or vertical lines of 1.

The main advantage of this approach is that it is designed around a formal definition of the syntenic blocks. Moreover, statistical validation can be performed on putative syntenic blocks to filter out false positives. However, several weaknesses exist for this approach. To begin with, the important impact of the parameters requires a good knowledge of the biological question asked. With this type of approach different parameters are critical and need to be finely tuned. An example being the size of the gap allowed between genes considered as belonging to the same block. This parameter can deeply affect the results, and needs to be configured according to the specificity of the study. The size of the gap depends mainly on the density of the matrix, i.e., the density of the pairs of homologous genes between the segments constituting the matrix. On one hand, a small gap value results in many small syntenic blocks that are more difficult to analyze. On the other hand, a high gap value produces long blocks that are easier to analyze but allow for more false positives. Moreover, the metrics used to estimate the distance between genes in a matrix are also an important setting. Two types of metrics are often proposed: The Diagonal Pseudo Distance (used by *i-ADHoRe* and *DiagHunter*), and the Manhattan Distance (used by *FISH*, *SynMAP*, or *Cinteny*). The Diagonal Pseudo Distance promotes genes near the diagonal axis and therefore, the distance inflates rapidly the further away genes are from this diagonal. In contrast, the Manhattan Distance tends to give smaller distances between aligned genes on the vertical or horizontal axis. Other types of distances have been implemented in tools such as *PhylDiag* [174] that uses the Euclidean Distance or the Chebyshev Distance in addition to the others mentioned above. Thus, the choice of the distance is not easy and will impact the results as surely as a wrongly set gap value. A benchmark analysis suggested that the Manhattan Distance gives the best results among these four distances [174]. The importance of the configuration is really to be taken into account when using this approach in an optimal way and makes these algorithms difficult to use without a minimum of expertise on both the tool and the biological question. Moreover, statistical tests to evaluate homologous regions are based on the assumption that the rate of gene loss is balanced between homologous regions. This is not the case for many species. Furthermore, some differences in terms of genome structure, especially the gene density and repetition in chromosomal regions, both locally and at the genome level, are difficult to account for with this approach. Finally, matrices can only compare genomes by pair, which implies that benefits of comparing multiple genomes at once, including WGD studies or diverged synteny blocks, cannot be done. Moreover, this approach cannot resolve multiple relationships between genes, detect inversions, nor non-overlapping syntenic blocks. To finish, not all of these implementations can detect tandem duplication. In the already cited tools, only *i-ADHoRe* and *FISH* can handle them. We will present these tools in more detail below.

#### *i-ADHoRe* (Iterative Automatic Detection of Homologous Regions) 3.0

*i-ADHoRe* [158] is one of the most used programs to find syntenic blocks and can be considered as a state-of-the-art algorithm. In its latest version, *i-ADHoRe* enables the detection of genomic homology through the identification of gene collinearity. This version is well optimized to handle a large number of genomes, taking advantage of parallel computing.

The algorithm begins by assimilating tandem duplicated genes as a single representative. Then, for each pair of genes, a sparse gene homology matrix is constructed. In this matrix, homologous genes are considered as dots, making collinearity zones seen as dense diagonals. Gene clusters with at least three homologous gene pairs are included in diagonals after a statistical validation taking into account the overall background density of the matrix. In the case of multiple clusters found, a correction for multiple testing is performed using the Bonferroni or False Discovery Rate (FDR) method.

This part corresponds to the traditional homology matrix approach. The next part of the algorithm is an optimization by dynamic programming.

Significant collinear regions found during this initial detection are aligned using the progressive Needleman–Wunsch (pNW) algorithm or a greedy graph-based algorithm [178]. The results of this alignment are stored into a profile, which contains the combined content of the two collinear regions and constitutes a more sensitive probe to find new homologous regions including more degenerated ones. Using this newly constructed profile, a search is performed in an iterative way. Thus, this profile is used to search for new sequences that can be aligned with it. If possible, the new matches are added to the profile. As long as new collinear regions can be added to a profile, these two steps are repeated.

The results are provided as text files and two associated plots: A dot plot and a set of graphs representing each final aligned profile.

This tool has many key parameters that directly influence the quality of the results:

- `prob_cutoff`, is used to store the maximum probability for a cluster to be generated by chance. The default value is 0.001.
- `gap_size`, indicating the maximum distance between genes in a cluster. The default value is 15.
- `cluster_gap`, indicating the maximum distance between individual base clusters in a cluster. The default value is 20.
- `q_value`, storing the minimum  $r^2$ -value which measures the quality of the linear regression prediction.
- `anchor_points`, the minimum number of anchor points which is comprised between 3 and 6.

The main advantage of this tool is to allow the computation of multiple genomes thanks to different optimizations including the use of parallel computation, an efficient statistical model to estimate  $p$ -values of diagonals before including them, the use of greedy graph-based alignment algorithms, and the use of ordered gene lists instead of genome sequence. This level of abstraction allows a more efficient detection of collinearity and thus the divergent intergenic sequences will have less impact on the algorithm.

#### OrthoCluster

In this category, *OrthoCluster* [177] appears as particular. It is not based on the classical approach of homology matrix construction although it is using the same philosophy by identifying syntenies via the clustering of neighboring matching gene pairs. This program is based on an algorithm implementing a strategy of tree enumeration to detect orthologous gene clusters. This tool can handle many genomes and makes it possible to overcome some of the weaknesses of the other classical approaches. Indeed, it can detect four types of genome rearrangements including insertions, transpositions, insertions/deletions, and reciprocal translocations via different algorithms. To detect reciprocal translocations (exchange of DNA parts by recombination), *OrthoCluster* merges syntenic blocks to build the longest possible blocks, identifying blocks not broken by duplications, inversions or transpositions. To detect transpositions (regions moved from a chromosome and inserted into a non-homologous chromosome), *OrthoCluster* searches in each adjacent syntenic block for a region between their homologous syntenic block in the other genome. If a fragment of less than 50 genes is found between them, a transposition is identified. Then, the detection of inverted segments in the genome is performed by checking if the order of the genes is the same in each syntenic block. If the gene order is inverted between the two, this region constitutes an inversion. Finally, in order to detect insertion or deletion of genes, *OrthoCluster* compares the pairs of adjacent syntenic blocks in the reference genome. Genes identified between these blocks are considered as insertions/deletions and reported. It can also detect segmental duplications and resolve one-to-many relationships. Moreover, the orientation and the order of genes are taken in to account. Nevertheless, this tool is limited to the orthology detection and can therefore only be applied on closely related organisms.



The fine-tuned configuration of this tool is crucial to obtain reliable results. Eight parameters can be defined by the user to set up the algorithm according to the needs:

- *l* max defining the upper bound on the number of genes in each cluster.
- *l* min defining the lower bound on the number of genes in each cluster.
- *op* maximal percentage of out-map genes allowed.
- *ip* defining the maximal percentage of mismatched in-map genes allowed.
- *op* and *ip* can control the number of genes involved in transpositions in synteny block.
- *i* maximal number of mismatched in-map genes allowed.
- *o* maximal number of out-map genes allowed.
- *r* to find order-preserving clusters.
- *s* to find strandedness-preserving clusters.

### 3.2.2. Algorithms Using Dynamic Programming Optimizations

This type of approach generally implements algorithms more costly in computation than the homology matrix approaches. Some methods benefit from dynamic programming to build a chain of collinear pairwise genes. In these methods, the dynamic programming algorithms are implemented in the search for collinear genes, allowing an exhaustive and fast search. A scoring system is set up allowing to build pairs of adjacent collinear genes, which constitute anchoring genes, and to penalize the distance between them. The main advantage of a multi-alignment of collinear chromosomal regions is its ability to reveal past WGD events and complex chromosomal rearrangement relationships. In this type of approach, the syntenic blocks are composed of anchoring genes that are located at collinear positions and between them non-anchoring genes that are assumed to have undergone mutations. Nevertheless, the user needs to already know what to look for and the characteristics of the genomes and syntenic blocks being studied.

#### MCSanX and MultiSyn

*MCSanX* [154] is one of the most used tools aiming at searching for syntenic blocks and is implemented in the webtool *MultiSyn* [176], allowing biologists with no informatic skills to use this approach. Moreover, this tool produces additional visualizations allowing a simplified analysis.

The *MCSanX* algorithm takes place in three steps. The first step uses the results of an all-against-all comparison using BLASTP [179] to find collinear blocks. BLASTP matches are sorted according to their genomic positions. To handle tandem regions, all consecutive genes with a BLASTP match that are separated by less than five genes, are collapsed into a single representative. Then, the highest scoring chains of collinear gene pairs are searched for using dynamic programming. Non-overlapping chains involving at least five collinear gene pairs are saved. In a pair of collinear blocks, two distinct genomic locations with aligned collinear genes are assigned as anchors.

The second step makes it possible to assign each syntenic block to a gene class. To do that, all genes are first assigned to the singleton class. Genes with BLASTP hits to other genes are assigned to the class “dispersed duplicates”. If the hits are close enough, they are assigned to the class “proximal duplicates”. If the hits are neighboring, they are assigned to the class “tandem duplicates”. To finish, anchored genes are assigned to the WGD/segmental class.

In the last step, twelve downstream analyses can be performed using different scripts and correspond to the computation of the nonsynonymous and synonymous rates (Ka and Ks), the generation of various plots, the construction of gene families with associated analyses, the detection of collinear tandem arrays, the computation of the number of intra- and inter-species collinear blocks at each locus of reference genomes, and the display of statistics on gene numbers at different duplication depths.

To be functional this tool needs to be configured using at least six parameters:

- *match\_score*, defines a threshold used to validate a synteny block. Default value is 50.

- `gap_penalty`, defines the penalty added when opening a gap. Default value is 21.
- `match_size`, defines the number of genes required to consider it as a collinear block. Default value is 5.
- `e_value`, defines the statistical significance of the synteny block alignment. Default value is 1e-10.
- `max_gaps`, maximum number of gaps allowed. Default value is 25.
- `overlap_window` stores the maximum number of genes to collapse BLAST matches. Default value is 5.

The special feature and strength of *MCScanX* is that each chromosome is used as a reference. Thus, all collinear segments in pairs are mapped. This is followed by a multiple alignment procedure of homologous genes, described as “transitive homology” [180]. This approach allows *MCScanX* to match regions that were not initially detected based on their collinearity with the reference.

To conclude, this tool is powerful and allows performing many analyses, if the user has the ability to install and configure it properly. *MultiSyn* eases the configuration step, the initial formatting of the data and the analyses using a graphical interface. As a final advantage, this tool can be deployed locally. As for *i-ADHoRe*, the use of ordered gene lists instead of a genome sequence allows getting more reliable results at lower computational costs.

### SynChro

SynChro [160] is based on Reciprocal Best-Hits (RBH) to construct syntenic blocks. This algorithm is faster and easy to use thanks to its unique parameter ( $\Delta$ ) which represents syntenic block stringency. To go into more details, this algorithm is composed of three simple steps. In the first step, RBH are identified using *Opscan* (<http://www.wabi.snv.jussieu.fr/public/opscan/>), a tool based on the FASTA algorithm [181]. RBH can be defined as two genes whose best hit is mutual. In the second step, the RBH makes it possible to define syntenic blocks using co-localizing RBH (defined by  $\Delta$ ) along the chromosomes of two genomes as anchors. In the third step, syntenic blocks are completed by non-RBH homologs. Genes are defined as non-RBH if they share 30% of similarity and if the ratio of the length of the match between the two sequences and the length of the smallest sequence is greater than 0.5.

This tool provides various graphical outputs including dotplots, chromosome painting, and synteny maps, as well as text results. Therefore, it makes it possible to obtain in a limited computational time very good quality results with a fast handling in an “all in one” manner allowing to easily visualize the results.

### CYNTENATOR

*CYNTENATOR* [159] is a tool aiming at identifying conserved syntenic regions between distant genomes. This tool is based on a progressive multiple gene order alignment. The main advantage of this tool is its scalability allowing it to work on more than 10 genomes contrary to many other approaches. Moreover, it makes it possible to get rid of heavy preprocessing steps due to its high flexibility.

To begin, a progressive pairwise alignment between genomes is performed. These alignments are based on a user-defined phylogenetic tree that directs the order in which the genomes will be compared. Only valid alignments gathering homologous regions of all species are retained for collinearity search in the next genome. This filtering step helps lower the computational costs and allows determining the gene order conservation between distant genomes. This step is followed by a pairwise alignment using a Smith-Waterman local alignment weighted by the phylogenetic distance. The results of these alignments constitute the syntenic blocks. The use of a progressive alignment algorithm makes it possible to conduct studies on several large genomes while taking into account the phylogeny of the studied species. The absence of a heavy pre-processing on the input data, except an all against all homology score, allows to avoid bias.

## SyMap

*SyMap* [163] is a tool based on *DAGChainer* [165]. The advantage of this software is that its interface via a web application allows the user to be free from any configuration and data preparation via the code. In addition, this tool allows retrieving various additional information and in particular the Ka/Ks ratio using *PAML* [182]. The intermediate results can be retrieved and the final results can be visualized in an interactive dot-plot. Once the genomes have been added to the database and the parameters have been defined, the computations are launched. The *SyMap* algorithm works as follows. First, the genomes are aligned using an alignment software. Different tools can be used for this step, including BLAST. Then, different filters are applied and in particular the condensation of tandemly duplicated genes into a single occurrence and filtering out of repeated sequences. The syntenic blocks within this homologous matrix are then searched for using *DAGChainer*. Finally, different visualizations are constructed. The main advantages of this tool are its speed, the ease of use, and the visualizations. However, some parameters are not configurable and it does not allow the study of more than two genomes at the same time.

### 3.2.3. Approaches Based on Graphs

The principle of this type of algorithm is to construct a graph gathering all the pairs of homologous genes shared by the compared genomes. These approaches aim at solving many problems raised by the methods presented before, in particular the possibility of studying several large genomes and to detect non-overlapping syntenic blocks. The previous approaches have difficulties decoding more complex genomic architectures that have undergone phases of significant duplication followed by repleidization. The search for non-overlapping syntenic blocks is of great interest because it makes it possible to focus on rearrangements that happened after the duplication events. However, the search for non-overlapping syntenic blocks is not just about simply decomposing overlapping blocks. Different algorithms have been proposed to meet these needs. The first algorithms as *GRIMM-Synteny* [183] or *MAGIC* [184] were suitable for small sets of genomes, but were not able to handle genomes with large duplications and deletions, and were not able to find non-overlapping blocks. Later, *Enredo* [169] was written to solve this problem. One last problem with the algorithms from the two previous approaches is that as more and more genomes are integrated into comparative studies, the number of genes shared between genomes decreases. This has a strong impact on the algorithms with the risk of rejecting the blocks because they are statistically nonsignificant, as it happens with tools such as *GRIMM-Synteny*.

Typically, the algorithms used in the graph-based approaches follow different steps. First, input genomes are locally aligned and the resulting alignments are used to construct a graph. Then, the different sub-structures (depending on the initial graph structure) are searched for to find the different segmentations of the genomes. The type of graph structure has a major influence on the results. Indeed, some of them handle these problems with more or less success and can therefore not find similar results. Four graph structures are predominant to analyze syntenic blocks.

The first structure corresponds to an alignment graph. The graph contains a vertex referring to each character in the sequence and edges referring to aligned characters. It is then possible to obtain collinear or noncollinear alignments by solving the maximum weight trace problem. Duplicated regions are more easily visible in an alignment graph structure. Nevertheless, this structure does not allow the user to get inversion information.

The second structure corresponds to A-Bruijn graphs that can be found in *DRIMM-Synteny* [168]. The main idea behind this graph is to merge aligned vertices. Thus, A-Bruijn graphs have one vertex for each aligned sequence. The links represent only the sequence. The main problem with this approach is represented by the short cycles, which tend to make local alignments hide a local collinearity. As an alignment graph structure, it does not allow access to inversion information, meaning that it is not possible to differentiate between the tandem repeats and palindromes.

The third structure, known as the Enredo graphs, can be found in *Enredo* [169]. It aims at managing genomes partitioned into segments. The nucleotide alignments are then made. Thus, Enredo graphs

have two vertices per set of aligned segments, a head vertex and a tail vertex. It is then possible to eliminate various substructures from the Enredo graph in order to obtain the final segmentation of the genome. An Enredo structure can help find non-overlapping blocks and is suitable to consider non-overlapping inversions.

To finish, the Cactus graphs [185,186] have also been proposed. They are structures with vertices for adjacencies and undirected edges for genome segments. This type of graph is Eulerian meaning that there is a path that crosses all the nodes only once. This graph is also subdivided into independent units where each edge is part at most of one cycle. The cactus structure is a unique sub structure that allows an easy detection of short cycles.

These different graph structures allow the study of some particular sub-structures to identify syntenic blocks. One of the most important corresponds to the collinear paths. These sub-structures are a set of blocks that appear in genomes consecutively, without breaks and with the same orientation thus representing syntenic blocks. A second sub-structure corresponds to the presence of microblocks within larger regions that tend to introduce breaks into syntenic blocks. A third sub-structure corresponds to the short cycles. They are the mark of rearrangement. Indeed, similarities between sequences make them appear and thus break the collinearity. An important number of short cycles is problematic because they can aggregate into complex networks and hide true collinear blocks. We will detail a little bit more on one algorithm implementing the graph based approach below.

#### DRIMM-Synteny

This algorithm is the update of *GRIMM-Synteny* and aims at solving various problems from this previous version. In particular, the suppression of blocks due to their statistical non-significance in the case of the study of several distant genomes. The syntenic blocks that do not overlap are identified, which allows, in a second step, to bypass the threading problem based on the use of an A-Bruijn graph structure. This type of graph is an Eulerian and undirected multigraph. Edges are weighted by the number of times a gene pair is consecutive in the analyzed genomes.

In *DRIMM-Synteny*, an A-Bruijn graph is constructed by collapsing together identically labeled vertices from all genomes. From this graph, syntenic blocks can be found. In fact, a perfectly repeated block corresponds to a path in the graph. Perfectly repeated regions that do not share genes with other regions in the set of genomes being studied will appear as unconnected paths. These are referred to as the maximum paths in the graph, satisfying the condition that all of their internal vertices have only two neighboring vertices. This algorithm solves some existent problems known for this type of approach using different subroutines. There may be small differences between the different syntenic genes, which leads to short cycles. *DRIMM-Synteny* is able to detect them by computing a shaft at maximum range. A heuristic then allows detecting the links that create them in order to remove them. In addition, the presence of syntenic microblocks separate the long unbranched paths into several subpaths, thus complicating the detection of the blocks. Finally, the short palindromic regions that can be found within syntenic blocks form thornes that have the same effect as the microblocks.

#### 3.3. Detection of Tandemly Arrayed Genes (TAGs)

Specific methods have been developed to handle specifically tandem duplication detection. TAGs are gene family members that are tightly clustered on a chromosome [73]. The vast majority of the methods are home-made pipelines available from the authors and may require programming skills. A few tools, particularly those related to the identification of syntenic blocks, are able to help in the identification of TAGs because they are generally summarized in a single occurrence of the dataset to lower the statistical noise. In general, they are not dedicated methods but more trivial algorithms. However, they have the advantage of being simpler to use. Most of these algorithms rely on protein comparisons, making them dependent on genome annotation. However, there exists very few methods that can deal with genomic sequences to search for long DNA tandem repeats. The advantage of these latest methods is that they can detect pseudogenes that originated from duplication or short ORFs

generally missed by automatic genome annotation. We will first describe TAG detection in the genome at protein level, then at DNA level.

### 3.3.1. Detection at Protein Level

These methods begin with the identification of homologous gene pairs. This can be done using different algorithms, in most cases an all-by-all BLASTP comparison of the proteome against itself or between the proteomes of two species, followed by a filtering using a threshold to retain only homologous pairs. The difference between these approaches lies in the homology assessment and the degree of sophistication to filter out false positives.

The most straightforward, but trivial, way is used in the first step of WGD detection algorithms such as *MCScanX* or *i-ADHoRe* [157,158]. These algorithms take as input homologous gene pairs, the preferred format being the BLASTP output. Then, the program classifies homologous pairs according to their rank along the chromosome. If consecutive BLASTP matches have a common gene and its paired genes are separated by fewer than five genes, these matches (forming a TAG) are collapsed using a representative pair with the smallest BLASTP e-value. The advantage of this approach is its speed but the drawback is that it can miss divergent homologous genes. Moreover, even if few programming skills are required, a parsing step is still necessary to obtain the list of identified TAGs.

To alleviate some problems related to the input (an all-against-all BLASTP), it is possible to use gene families as input. They can be constructed by different algorithms summarized in Table 1. Then, a TAG is defined as a block of adjacent genes belonging to the same family and separated by spacers that are generally genes not belonging to the homologous family. Several definitions can be used for the allowed number of spacers, mostly 0 or 1 but also ranging from 0 to 10 spacers [73,127,129,187]. The construction of gene families allows incorporating more distantly related homologs than the previous approach. The definition of homologous genes can be improved by merging all non-overlapping HSP of one hit [73]. The most widely used clustering algorithms are the single linkage algorithm, and more and more Markov clustering (MCL) and its variants. It is an efficient approach but adjusting the inflation and expansion parameters of MCL is not easy. The inflation parameter controls the flux between groups of classification (i.e., the number of steps in the random walk along the similarity graph). The expansion parameter controls the strength of links by strengthening them inside the clusters and weakening them between clusters.

### 3.3.2. Detection at DNA Level

The vast majority of Tandem Repeat detection methods at DNA level deal with the identification of short highly repeated sequences. They are used to mask sequences corresponding to TEs or/and segments of low complexity before genome annotation or to explore the amplification of short duplications associated with human diseases for example, or copy number variation (CNV) between genotypes. These types of DNA duplication are not the focus of this review and will not be treated in detail. Here, we give a list of some famous short DNA Tandem Repeat detection tools able to deal with large datasets: *DUSTMASKER* [188], *SEGMASKER* [189], *Tandem Repeat Finder (TRF)* [190], *TANTAN* [191] and more recently *ULTRA* [192], *TARDIS* [193], and *dot2dot* [194].

We will now focus on long tandem duplication detection because all studies on TAGs based on protein similarity are biased by the quality of the available genome annotation. They exclude RNA genes or degenerated copies [195]. However, duplicated pseudogenes are an important evolutionary residue of a genome past activity [196]. A genome-wide approach has been proposed to take into account pseudogenes in TAG detection [197]. It scans, using TBLASTN, each protein against its chromosomal regions (the surrounding DNA sequences is three times longer than the CDS) and filter hits according to a refined bit-score, called the BTF score, that takes into account all non-overlapping HSP of less than 20% on the same strand. Then, it looks at CDSs in the ascending order of their chromosomal positions to extract TAGs. This mixed approach (at DNA and protein levels) is implemented in Python

2.4. The scripts are available from the authors and need a step of manual curation to eliminate false positive TAGs, due to the presence of minisatellites.

This previous approach is based on proteins and therefore depends on genome annotation. It has mainly been used on compact genomes [195]. *ReD Tandem* is an alternative method that circumvents this limitation [195]. Indeed, the main problem of detecting TAGs at genomic level is that large duplications despite being close, are far from being contiguous. The authors thus proposed to define tandemly duplicated segments as paralogous segments of size  $l$  with adjacent copies separated by a maximum distance  $T$  (in *A. thaliana*, the parameter values are  $l = 500$  bp,  $T = 150$  kb). The algorithm begins with anchors (paralogous segments of size  $l$ ) and chains then using *DAGchainer* or *OSfinder* [175] into longer duplicated regions (called tandem units). Such alignable units are anchors of length  $l$  and separated by less than  $L$  bases ( $L = 40$  kb for *A. thaliana*). Then, the tandem units are assembled into TAGs (i.e., tandem units separated by less than  $T$  bases, with  $T = 150$  kb for *A. thaliana*). The C++ scripts are available but need some computational skills to be installed. Nevertheless, this elegant approach has allowed the authors to identify in *A. thaliana* several types of TAGs previously undetectable for genome-wide approaches. In decreasing order of importance, these new TAGs correspond to trans-elements genes, pseudogenes, pre-tRNAs, other RNAs, miRNAs, snoRNAs, and unknown genes [195].

### 3.4. Databases Storing Syntenic Block or Homology Information

#### 3.4.1. Syntenic Information

These databases have the advantage to not require computations and therefore no programming skills for the user. Some of them also offer visualizations and search tools. The main disadvantage is that they do not contain information from all organisms. Each of these databases provide particular features but some elements are common. In some cases, it is possible to access all the syntenic blocks between two organisms. The list of organisms is more or less extended depending on the database. Some of them propose to visualize these blocks using various representations such as circular visualizations, chromosome painting, or dot-plots. Some databases allow manually importing genomes to identify blocks of synteny. In this case, different tools may be implemented for the identification and are more or less easy to configure. For example, Ensembl [198] stores different information including syntenic blocks generated by Pecan [169] as a multi-alignment algorithm and *Enredo* to detect syntenic blocks. Synteny portal [199] and Genomicus [200] provide also syntenic blocks generated by *inferCars* [201] for different species but also multiple visualizations. Finally, other databases exist including ECRbase [202] with syntenic blocks generated from the DNA level. OrthoClusterDb [203] is a good example of what can be found inside these databases. Two main possibilities are available. First, it allows online access to the *OrthoCluster* tool [177] and to carry out identification of syntenic blocks on a remote server using a graphical interface facilitating the configuration and the retrieval of the results. Another possibility is to access different pre-computed syntenic blocks by *OrthoCluster* for different species. Pre-computed species belong to different groups (*Mammals*, *Pseudomonaceae*, *Drosophila*, *Plasmodium*, and *Caenorhabditis*) with 54 species available. The syntenic blocks can be visualized on a figure called genome painting which allows visualizing the chromosomes of the compared species with a system of colored segments highlighting the syntenic blocks. It is also possible to retrieve raw output files or to access to syntenic blocks using an online genome browser.

#### 3.4.2. Homology Relationships Databases

Dataset of duplicated genes without specification of the underlying mechanism of duplication can be retrieved from public databases. These databases can be associated or not with a specific methodology with available tools for a local use. The INPARANOID 8 database for example, provides the *InParanoid* tool and proposes orthology analysis between 273 proteomes, mostly eukaryotic. The dataset of orthologous and paralogous relationships between genes can be downloaded by pairs of



species [204]. In HOGENOM, gene families are built from complete genomes from all three domains of life [147]. Its clustering pipeline is based on the *SiLiX* clustering method [205]. Even if this database is regularly updated, users can only retrieve families one by one according to keywords. The total amount of paralogous genes in a species is only available upon request directly to the authors.

Ensembl Compara is a specific section of the Ensembl database providing cross-species resources and analyses, at both the sequence and the gene levels. The main Ensembl database is dedicated to chordate genomes and displays now counterparts for several groups of organisms (Ensembl Genomes, Ensembl Bacteria, Ensembl Protists, Ensembl Fungi, Ensembl Plants, and Ensembl Metazoa). All these databases are associated with the Ensembl Compara system. This system provides access to protein gene families via a Perl API [198]. These families are built using all proteins from Ensembl through a classical process using BLASTP for similarity searches and a MCL clustering with scores as weight for edges in the initial graph. A final step aligns all sequences from a family using *MAFFT* [206]. It is to note that the Ensembl Compara protein families correspond to the most similar proteins compared to its gene tree section, where paralogous relationships are also available but in a tree format. Many other repositories are available but our goal is not to be exhaustive. Among the most generalist, we can cite PhylomeDB [148], OMA [207], OrthoDB [139], *OrthoInspector* [141], eggNOG [138], or the database Homologene from the NCBI portal.

For plant comparative genomics, we can cite the databases PLAZA [208], GreenPhyl [209], and Phytozome [210]. PLAZA 4.0 contains gene family data, phylogenetic trees, and gene colinearity information. It comprises two instances, one for monocots (Monocots PLAZA 4.5) that includes data from 39 species and one for dicots (Dicots PLAZA 4.0) that includes data from 55 species. The latest PLAZA instance offers one or more REST-full APIs, depending on the Platform software version. GreenPhyl 4 contains gene families and phylogenetic trees from 37 species. It has not been updated since 2015 but contains a section of manually annotated families comprising 2956 clusters. Other interesting sections are transcription factors and families specific to species or phylum (family of homologous genes found only in one species or excluding/including one phylum). Finally, the plant database Phytozome13 (last update in May 2019) contains 184 assembled and annotated genomes. Inparanoid pairwise orthology and paralogy groups have been calculated across all Phytozome proteomes and families of related genes representing the modern descendants of putative ancestral genes have been constructed at key phylogenetic nodes. The dataset can easily be downloaded or mined via a dedicated tool named *PhytoMine*.

#### 4. Conclusions

To conclude, when considering duplicated genes inside a given species, it appears clear that they represent very different entities when taking into account their mechanism of formation, their fate, and their age. This is particularly important when it comes to their identification and analysis. It is indeed tempting to only detect all genes that are in several copies without taking into account the evolutionary complexity behind them. This is why it is also important to be aware of the different methodological approaches that can be used because this choice will greatly depend on the investigated biological question.

**Author Contributions:** E.L. and C.R. conceived the review; C.L., E.L., M.L., T.L., and C.R. wrote the different versions of the manuscript; T.L., and M.L. equally contributed to the work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work received no external funding.

**Acknowledgments:** This work was supported by the CNRS, the University Lyon 1, and the Laboratory “Biométrie et Biologie Evolutive”.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Ohno, S. *Evolution by Gene Duplication*; Springer: Berlin/Heidelberg, Germany, 1970; ISBN 978-3-642-86661-6.
- Kondrashov, F.A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.* **2012**, *279*, 5048–5057. [[CrossRef](#)]
- Van de Peer, Y.; Maere, S.; Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **2009**, *10*, 725–732. [[CrossRef](#)]
- Vallejo-Marín, M.; Buggs, R.J.A.; Cooley, A.M.; Puzey, J.R. Speciation by genome duplication: Repeated origins and genomic composition of the recently formed allopolyploid species *Mimulus peregrinus*. *Evolution* **2015**, *69*, 1487–1500. [[CrossRef](#)]
- Ting, C.T.; Tsaur, S.C.; Sun, S.; Browne, W.E.; Chen, Y.C.; Patel, N.H.; Wu, C.I. Gene duplication and speciation in *Drosophila*: Evidence from the Odysseus locus. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 12232–12235. [[CrossRef](#)] [[PubMed](#)]
- Zhang, F.; Gu, W.; Hurles, M.E.; Lupski, J.R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 451–481. [[CrossRef](#)] [[PubMed](#)]
- Dickerson, J.E.; Robertson, D.L. On the origins of Mendelian disease genes in man: The impact of gene duplication. *Mol. Biol. Evol.* **2012**, *29*, 61–69. [[CrossRef](#)] [[PubMed](#)]
- Tollis, M.; Schneider-Utaka, A.K.; Maley, C.C. The Evolution of Human Cancer Gene Duplications across Mammals. *Mol. Biol. Evol.* **2020**. [[CrossRef](#)] [[PubMed](#)]
- Mendivil Ramos, O.; Ferrer, D.E.K. Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *Int. J. Evol. Biol.* **2012**, *2012*, 1–10. [[CrossRef](#)] [[PubMed](#)]
- Wolfe, K. Robustness—it's not where you think it is. *Nat. Genet.* **2000**, *25*, 3–4. [[CrossRef](#)]
- Sharman, A.C. Some new terms for duplicated genes. *Semin. Cell Dev. Biol.* **1999**, *10*, 561–563. [[CrossRef](#)]
- Sonnhammer, E.L.L.; Koonin, E.V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **2002**, *18*, 619–620. [[CrossRef](#)]
- Koonin, E.V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **2005**, *39*, 309–338. [[CrossRef](#)] [[PubMed](#)]
- Altenhoff, A.M.; Glover, N.M.; Dessimoz, C. Inferring Orthology and Paralogy. In *Evolutionary Genomics*; Anisimova, M., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2019; Volume 1910, pp. 149–175. ISBN 978-1-4939-9073-3.
- Van de Peer, Y.; Mizrachi, E.; Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **2017**, *18*, 411–424. [[CrossRef](#)] [[PubMed](#)]
- Ramsey, J.; Schemske, D.W. Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants. *Annu. Rev. Ecol. Syst.* **1998**, *29*, 467–501. [[CrossRef](#)]
- Panchy, N.; Lehti-Shiu, M.; Shiu, S.-H. Evolution of Gene Duplication in Plants. *Plant Physiol.* **2016**, *171*, 2294–2316. [[CrossRef](#)]
- Jiao, Y.; Wickert, N.J.; Ayyampalayam, S.; Chanderbali, A.S.; Landherr, L.; Ralph, P.E.; Tomsho, L.P.; Hu, Y.; Liang, H.; Soltis, P.S.; et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **2011**, *473*, 97–100. [[CrossRef](#)]
- Feldman, M.; Levy, A.A. Genome Evolution Due to Allopolyploidization in Wheat. *Genetics* **2012**, *192*, 763–774. [[CrossRef](#)]
- Chalhoub, B.; Denoeud, F.; Liu, S.; Parkin, I.A.P.; Tang, H.; Wang, X.; Chiquet, J.; Belcram, H.; Tong, C.; Samans, B.; et al. Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **2014**, *345*, 950–953. [[CrossRef](#)]
- Yang, J.; Liu, D.; Wang, X.; Ji, C.; Cheng, F.; Liu, B.; Hu, Z.; Chen, S.; Pental, D.; Ju, Y.; et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **2016**, *48*, 1225–1232. [[CrossRef](#)]
- Sun, F.; Fan, G.; Hu, Q.; Zhou, Y.; Guan, M.; Tong, C.; Li, J.; Du, D.; Qi, C.; Jiang, L.; et al. The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. *Plant J.* **2017**, *92*, 452–468. [[CrossRef](#)]



23. Lu, K.; Wei, L.; Li, X.; Wang, Y.; Wu, J.; Liu, M.; Zhang, C.; Chen, Z.; Xiao, Z.; Jian, H.; et al. Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **2019**, *10*, 1154. [[CrossRef](#)] [[PubMed](#)]
24. Kasahara, M. The 2R hypothesis: An update. *Curr. Opin. Immunol.* **2007**, *19*, 547–552. [[CrossRef](#)] [[PubMed](#)]
25. Wendel, J.F.; Lisch, D.; Hu, G.; Mason, A.S. The long and short of doubling down: Polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* **2018**, *49*, 1–7. [[CrossRef](#)] [[PubMed](#)]
26. Freeling, M.; Scanlon, M.J.; Fowler, J.E. Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* **2015**, *35*, 110–118. [[CrossRef](#)] [[PubMed](#)]
27. Wright, J.E.; Johnson, K.; Hollister, A.; May, B. Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. *Isozymes* **1983**, *10*, 239–260.
28. Sacerdot, C.; Louis, A.; Bon, C.; Berthelot, C.; Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **2018**, *19*, 166. [[CrossRef](#)]
29. Pervaiz, N.; Shakeel, N.; Qasim, A.; Zehra, R.; Anwar, S.; Rana, N.; Xue, Y.; Zhang, Z.; Bao, Y.; Abbasi, A.A. Evolutionary history of the human multigene families reveals widespread gene duplications throughout the history of animals. *BMC Evol. Biol.* **2019**, *19*, 128. [[CrossRef](#)]
30. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298. [[CrossRef](#)]
31. Arguello, J.R.; Fan, C.; Wang, W.; Long, M. Origination of chimeric genes through DNA-level recombination. In *Gene and Protein Evolution*; Karger Publishers: Basel, Switzerland, 2007; Volume 3, pp. 131–146. [[CrossRef](#)]
32. Reams, A.B.; Roth, J.R. Mechanisms of gene duplication and amplification. *Cold Spring Harb. Perspect. Biol.* **2015**, *7*, a016592. [[CrossRef](#)]
33. Cook, D.E.; Lee, T.G.; Guo, X.; Melito, S.; Wang, K.; Bayless, A.M.; Wang, J.; Hughes, T.J.; Willis, D.K.; Clemente, T.E.; et al. Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean. *Science* **2012**, *338*, 1206–1209. [[CrossRef](#)]
34. Kono, T.J.Y.; Brohammer, A.B.; McGaugh, S.E.; Hirsch, C.N. Tandem Duplicate Genes in Maize Are Abundant and Date to Two Distinct Periods of Time. *G3 Genes Genomes Genet.* **2018**, *8*, 3049–3058. [[CrossRef](#)]
35. Tan, B.C.; Guan, J.C.; Ding, S.; Wu, S.; Saunders, J.W.; Koch, K.E.; McCarty, D.R. Structure and Origin of the White Cap Locus and Its Role in Evolution of Grain Color in Maize. *Genetics* **2017**, *206*, 135–150. [[CrossRef](#)] [[PubMed](#)]
36. 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 Res.* **1998**, *8*, 464–478. [[CrossRef](#)] [[PubMed](#)]
37. Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S.; Liang, C.; Zhang, J.; Fulton, L.; Graves, T.A.; et al. The B73 maize genome: Complexity, diversity, and dynamics. *Science* **2009**, *326*, 1112–1115. [[CrossRef](#)] [[PubMed](#)]
38. Brosius, J. Retroposons—Seeds of evolution. *Science* **1991**, *251*, 753. [[CrossRef](#)] [[PubMed](#)]
39. Moran, J.V.; DeBerardinis, R.J.; Kazazian, H.H. Exon shuffling by L1 retrotransposition. *Science* **1999**, *283*, 1530–1534. [[CrossRef](#)]
40. Elrouby, N.; Bureau, T.E. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *J. Biol. Chem.* **2001**, *276*, 41963–41968. [[CrossRef](#)]
41. Zhang, Z.; Harrison, P.M.; Liu, Y.; Gerstein, M. Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome. *Genome Res.* **2003**, *13*, 2541–2558. [[CrossRef](#)]
42. Casola, C.; Betrán, E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? *Genome Biol. Evol.* **2017**, *9*, 1351–1373. [[CrossRef](#)]
43. Betrán, E.; Thornton, K.; Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **2002**, *12*, 1854–1859. [[CrossRef](#)]
44. Bai, Y.; Casola, C.; Feschotte, C.; Betrán, E. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol.* **2007**, *8*, R11. [[CrossRef](#)] [[PubMed](#)]
45. Toups, M.A.; Hahn, M.W. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics* **2010**, *186*, 763–766. [[CrossRef](#)] [[PubMed](#)]

46. Emerson, J.J.; Kaessmann, H.; Betrán, E.; Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **2004**, *303*, 537–540. [[CrossRef](#)] [[PubMed](#)]
47. Wang, W.; Zheng, H.; Fan, C.; Li, J.; Shi, J.; Cai, Z.; Zhang, G.; Liu, D.; Zhang, J.; Vang, S.; et al. High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. *Plant Cell* **2006**, *18*, 1791–1802. [[CrossRef](#)] [[PubMed](#)]
48. Wang, Y.; Wang, X.; Tang, H.; Tan, X.; Ficklin, S.P.; Feltus, F.A.; Paterson, A.H. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PLoS ONE* **2011**, *6*, e28150. [[CrossRef](#)]
49. Juretic, N.; Hoen, D.R.; Huynh, M.L.; Harrison, P.M.; Bureau, T.E. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **2005**, *15*, 1292–1297. [[CrossRef](#)]
50. Le, Q.H.; Wright, S.; Yu, Z.; Bureau, T. Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 7376–7381. [[CrossRef](#)]
51. Yu, Z.; Wright, S.I.; Bureau, T.E. Mutator-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. *Genetics* **2000**, *156*, 2019–2031.
52. Kawasaki, S.; Nitasaka, E. Characterization of Tpn1 family in the Japanese morning glory: En/Spm-related transposable elements capturing host genes. *Plant Cell Physiol.* **2004**, *45*, 933–944. [[CrossRef](#)]
53. Zabala, G.; Vodkin, L.O. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell* **2005**, *17*, 2619–2632. [[CrossRef](#)]
54. Jiang, N.; Bao, Z.; Zhang, X.; Eddy, S.R.; Wessler, S.R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **2004**, *431*, 569–573. [[CrossRef](#)] [[PubMed](#)]
55. Samonte, R.V.; Eichler, E.E. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **2002**, *3*, 65–72. [[CrossRef](#)] [[PubMed](#)]
56. Wolfe, K.H.; Shields, D.C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **1997**, *387*, 708–713. [[CrossRef](#)] [[PubMed](#)]
57. Bailey, J.A.; Gu, Z.; Clark, R.A.; Reinert, K.; Samonte, R.V.; Schwartz, S.; Adams, M.D.; Myers, E.W.; Li, P.W.; Eichler, E.E. Recent segmental duplications in the human genome. *Science* **2002**, *297*, 1003–1007. [[CrossRef](#)]
58. Koszul, R.; Caburet, S.; Dujon, B.; Fischer, G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **2004**, *23*, 234–243. [[CrossRef](#)]
59. Koszul, R.; Dujon, B.; Fischer, G. Stability of large segmental duplications in the yeast genome. *Genetics* **2006**, *172*, 2211–2222. [[CrossRef](#)]
60. Fiston-Lavier, A.-S.; Anxolabehere, D.; Quesneville, H. A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res.* **2007**, *17*, 1458–1470. [[CrossRef](#)]
61. Bailey, J.A.; Liu, G.; Eichler, E.E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **2003**, *73*, 823–834. [[CrossRef](#)]
62. She, X.; Cheng, Z.; Zöllner, S.; Church, D.M.; Eichler, E.E. Mouse segmental duplication and copy number variation. *Nat. Genet.* **2008**, *40*, 909–914. [[CrossRef](#)]
63. Lander, E.S.; Linton, L.M.; Birren, B.; Nussbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [[CrossRef](#)]
64. Bailey, J.A.; Eichler, E.E. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **2006**, *7*, 552–564. [[CrossRef](#)] [[PubMed](#)]
65. Zhao, Q.; Ma, D.; Vasseur, L.; You, M. Segmental duplications: Evolution and impact among the current Lepidoptera genomes. *BMC Evol. Biol.* **2017**, *17*, 161. [[CrossRef](#)] [[PubMed](#)]
66. Hakes, L.; Lovell, S.C.; Oliver, S.G.; Robertson, D.L. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7999–8004. [[CrossRef](#)] [[PubMed](#)]
67. Wapinski, I.; Pfeffer, A.; Friedman, N.; Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **2007**, *449*, 54–61. [[CrossRef](#)]
68. Blanc, G.; Wolfe, K.H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **2004**, *16*, 1679–1691. [[CrossRef](#)]
69. Maere, S.; Bodt, S.D.; Raes, J.; Casneuf, T.; Montagu, M.V.; Kuiper, M.; De Peer, Y.V. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 5454–5459. [[CrossRef](#)]

70. Hanada, K.; Zou, C.; Lehti-Shiu, M.D.; Shinozaki, K.; Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **2008**, *148*, 993–1003. [[CrossRef](#)]
71. Rodgers-Melnick, E.; Mane, S.P.; Dharmawardhana, P.; Slavov, G.T.; Crasta, O.R.; Strauss, S.H.; Brunner, A.M.; DiFazio, S.P. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Res.* **2012**, *22*, 95–105. [[CrossRef](#)]
72. Freeling, M. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **2009**, *60*, 433–453. [[CrossRef](#)]
73. Rizzon, C.; Ponger, L.; Gaut, B.S. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.* **2006**, *2*, e115. [[CrossRef](#)]
74. Acharya, D.; Ghosh, T.C. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genom.* **2016**, *17*, 71. [[CrossRef](#)] [[PubMed](#)]
75. Casneuf, T.; De Bodt, S.; Raes, J.; Maere, S.; Van de Peer, Y. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol.* **2006**, *7*, R13. [[CrossRef](#)] [[PubMed](#)]
76. Defoort, J.; Van de Peer, Y.; Carretero-Paulet, L. The Evolution of Gene Duplicates in Angiosperms and the Impact of Protein–Protein Interactions and the Mechanism of Duplication. *Genome Biol. Evol.* **2019**, *11*, 2292–2305. [[CrossRef](#)] [[PubMed](#)]
77. Wang, Y. Locally duplicated ohnologs evolve faster than nonlocally duplicated ohnologs in *Arabidopsis* and rice. *Genome Biol. Evol.* **2013**, *5*, 362–369. [[CrossRef](#)] [[PubMed](#)]
78. Arabidopsis Interactome Mapping Consortium; Dreze, M.; Carvunis, A.R.; Charlotteaux, B.; Galli, M.; Pevzner, S.J.; Tasan, M.; Ahn, Y.Y.; Balumuri, P.; Barabási, A.L.; et al. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **2011**, *333*, 601–607. [[CrossRef](#)]
79. Arsovski, A.A.; Pradinuk, J.; Guo, X.Q.; Wang, S.; Adams, K.L. Evolution of Cis-Regulatory Elements and Regulatory Networks in Duplicated Genes of *Arabidopsis*. *Plant Physiol.* **2015**, *169*, 2982–2991. [[CrossRef](#)] [[PubMed](#)]
80. Prince, V.E.; Pickett, F.B. Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.* **2002**, *3*, 827–837. [[CrossRef](#)]
81. Zou, C.; Lehti-Shiu, M.D.; Thibaud-Nissen, F.; Prakash, T.; Buell, C.R.; Shiu, S.-H. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiol.* **2009**, *151*, 3–15. [[CrossRef](#)]
82. Rouquier, S.; Taviaux, S.; Trask, B.J.; Brand-Arpon, V.; Van den Engh, G.; Demaille, J.; Giorgi, D. Distribution of olfactory receptor genes in the human genome. *Nat. Genet.* **1998**, *18*, 243–250. [[CrossRef](#)]
83. Quignon, P.; Kirkness, E.; Cadieu, E.; Touleimat, N.; Guyon, R.; Renier, C.; Hitte, C.; André, C.; Fraser, C.; Galibert, F. Comparison of the canine and human olfactory receptor gene repertoires. *Genome Biol.* **2003**, *4*, R80. [[CrossRef](#)]
84. Hahn, M.W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **2009**, *100*, 605–617. [[CrossRef](#)]
85. Innan, H.; Kondrashov, F. The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* **2010**, *11*, 97–108. [[CrossRef](#)] [[PubMed](#)]
86. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983; ISBN 978-0-521-31793-1.
87. Logeman, B.L.; Wood, L.K.; Lee, J.; Thiele, D.J. Gene duplication and neo-functionalization in the evolutionary and functional divergence of metazoan copper transporters Ctr1 and Ctr2. *J. Biol. Chem.* **2017**. [[CrossRef](#)] [[PubMed](#)]
88. Escriva, H.; Bertrand, S.; Germain, P.; Robinson-Rechavi, M.; Umbhauer, M.; Cartry, J.; Duffraisse, M.; Holland, L.; Gronemeyer, H.; Laudet, V. Neofunctionalization in vertebrates: The example of retinoic acid receptors. *PLoS Genet.* **2006**, *2*, e102. [[CrossRef](#)] [[PubMed](#)]
89. Hughes, T.E.; Langdale, J.A.; Kelly, S. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* **2014**, *24*, 1348–1355. [[CrossRef](#)] [[PubMed](#)]
90. Fan, C.; Chen, Y.; Long, M. Recurrent Tandem Gene Duplication Gave Rise to Functionally Divergent Genes in *Drosophila*. *Mol. Biol. Evol.* **2008**, *25*, 1451–1458. [[CrossRef](#)]

91. Force, A.; Lynch, M.; Pickett, F.B.; Amores, A.; Yan, Y.L.; Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **1999**, *151*, 1531–1545.
92. Piatigorsky, J.; Wistow, G. The recruitment of crystallins: New functions precede gene duplication. *Science* **1991**, *252*, 1078–1079. [[CrossRef](#)]
93. Hughes, A.L. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B Biol. Sci.* **1994**, *256*, 119–124. [[CrossRef](#)]
94. Otto, S.P.; Yong, P. The evolution of gene duplicates. *Adv. Genet.* **2002**, *46*, 451–483. [[CrossRef](#)]
95. Jackson, P.J.; Douglas, N.R.; Chai, B.; Binkley, J.; Sidow, A.; Barsh, G.S.; Millhauser, G.L. Structural and molecular evolutionary analysis of Agouti and Agouti-related proteins. *Chem. Biol.* **2006**, *13*, 1297–1305. [[CrossRef](#)] [[PubMed](#)]
96. Carlson, K.D.; Bhogale, S.; Anderson, D.; Zaragoza-Mendoza, A.; Madlung, A. Subfunctionalization of phytochrome B1/B2 leads to differential auxin and photosynthetic responses. *Plant Direct* **2020**, *4*, e00205. [[CrossRef](#)] [[PubMed](#)]
97. Vavouri, T.; Semple, J.I.; Lehner, B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.* **2008**, *24*, 485–488. [[CrossRef](#)]
98. Gout, J.F.; Lynch, M. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. *Mol. Biol. Evol.* **2015**, *32*, 2141–2148. [[CrossRef](#)] [[PubMed](#)]
99. Qian, W.; Liao, B.Y.; Chang, A.Y.F.; Zhang, J. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* **2010**, *26*, 425–430. [[CrossRef](#)]
100. Greer, J.M.; Puetz, J.; Thomas, K.R.; Capecchi, M.R. Maintenance of functional equivalence during paralogous *HOX* gene evolution. *Nature* **2000**, *403*, 661–665. [[CrossRef](#)]
101. Dean, E.J.; Davis, J.C.; Davis, R.W.; Petrov, D.A. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.* **2008**, *4*, e1000113. [[CrossRef](#)]
102. Averof, M.; Dawes, R.; Ferrier, D. Diversification of arthropod *HOX* genes as a paradigm for the evolution of gene functions. *Semin. Cell Dev. Biol.* **1996**, *7*, 539–551. [[CrossRef](#)]
103. Wang, W.; Brunet, F.G.; Nevo, E.; Long, M. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 4448–4453. [[CrossRef](#)]
104. Nisole, S.; Lynch, C.; Stoye, J.P.; Yap, M.W. A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13324–13328. [[CrossRef](#)]
105. Sayah, D.M.; Sokolskaja, E.; Berthou, L.; Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **2004**, *430*, 569–573. [[CrossRef](#)] [[PubMed](#)]
106. Zhang, J. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16246–16250. [[CrossRef](#)] [[PubMed](#)]
107. Machado, J.P.; Antunes, A. The genomic context of retrocopies increases their chance of functional relevancy in mammals. *Genomics* **2020**, *112*, 2410–2417. [[CrossRef](#)] [[PubMed](#)]
108. Makino, T.; McLysaght, A. Positionally biased gene loss after whole genome duplication: Evidence from human, yeast, and plant. *Genome Res.* **2012**, *22*, 2427–2435. [[CrossRef](#)]
109. Jiang, W.; Liu, Y.; Xia, E.; Gao, L. Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. *Plant Physiol.* **2013**, *161*, 1844–1861. [[CrossRef](#)]
110. Pan, D.; Zhang, L. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: A novel strategy to estimate gene duplication rates. *Genome Biol.* **2007**, *8*, R158. [[CrossRef](#)]
111. Marques-Bonet, T.; Girirajan, S.; Eichler, E.E. The origins and impact of primate segmental duplications. *Trends Genet.* **2009**, *25*, 443–454. [[CrossRef](#)]
112. Assis, R.; Bachtrog, D. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 17409–17414. [[CrossRef](#)]
113. Pearson, W.R. An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinforma.* **2013**. [[CrossRef](#)]
114. Shapiro, B.; Hofreiter, M. A paleogenomic perspective on evolution and gene function: New insights from ancient DNA. *Science* **2014**, *343*, 1236573. [[CrossRef](#)]
115. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]

116. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [[CrossRef](#)] [[PubMed](#)]
117. Johnson, L.S.; Eddy, S.R.; Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinform.* **2010**, *11*, 431. [[CrossRef](#)]
118. Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197. [[CrossRef](#)]
119. Saebo, P.E.; Andersen, S.M.; Myrseth, J.; Laerdahl, J.K.; Rognes, T. PARALIGN: Rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.* **2005**, *33*, W535–W539. [[CrossRef](#)] [[PubMed](#)]
120. Rucci, E.; Garcia Sanchez, C.; Botella Juan, G.; Giusti, A.D.; Naiouf, M.; Prieto-Matias, M. SWIMM 2.0: Enhanced Smith-Waterman on Intel’s Multicore and Manycore Architectures Based on AVX-512 Vector Extensions. *Int. J. Parallel Program* **2019**, *47*, 296–316. [[CrossRef](#)]
121. Koonin, E.V.; Galperin, M.Y. *Sequence—Evolution—Function: Computational Approaches in Comparative Genomics*; Kluwer Academic: Boston, MA, USA, 2003; ISBN 978-1-4020-7274-1.
122. Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9*, 56–68. [[CrossRef](#)]
123. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *12*, 85–94. [[CrossRef](#)] [[PubMed](#)]
124. Li, W.H.; Gu, Z.; Wang, H.; Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **2001**, *409*, 847–849. [[CrossRef](#)]
125. Blanc, G.; Wolfe, K.H. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* **2004**, *16*, 1667–1678. [[CrossRef](#)]
126. Wootton, J.C.; Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **1993**, *17*, 149–163. [[CrossRef](#)]
127. Shoja, V.; Zhang, L. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol. Biol. Evol.* **2006**, *23*, 2134–2141. [[CrossRef](#)] [[PubMed](#)]
128. Britten, R.J. Almost all human genes resulted from ancient duplication. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 19027–19032. [[CrossRef](#)] [[PubMed](#)]
129. Pan, D.; Zhang, L. Tandemly arrayed genes in vertebrate genomes. *Comp. Funct. Genom.* **2008**, 545269. [[CrossRef](#)] [[PubMed](#)]
130. Makino, T.; McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 9270–9274. [[CrossRef](#)] [[PubMed](#)]
131. Singh, P.P.; Arora, J.; Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput. Biol.* **2015**, *11*, e1004394. [[CrossRef](#)]
132. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [[CrossRef](#)]
133. Mitchell, A.L.; Attwood, T.K.; Babbitt, P.C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S.D.; Chang, H.Y.; El-Gebali, S.; Fraser, M.I.; et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **2019**, *47*, D351–D360. [[CrossRef](#)]
134. Kuzniar, A.; Van Ham, R.C.H.J.; Pongor, S.; Leunissen, J.A.M. The quest for orthologs: Finding the corresponding gene across genomes. *Trends Genet.* **2008**, *24*, 539–551. [[CrossRef](#)]
135. Tatusov, R.L.; Koonin, E.V.; Lipman, D.J. A genomic perspective on protein families. *Science* **1997**, *278*, 631–637. [[CrossRef](#)]
136. Remm, M.; Storm, C.E.; Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **2001**, *314*, 1041–1052. [[CrossRef](#)] [[PubMed](#)]
137. Schreiber, F.; Sonnhammer, E.L.L. Hieranoid: Hierarchical orthology inference. *J. Mol. Biol.* **2013**, *425*, 2072–2081. [[CrossRef](#)] [[PubMed](#)]
138. Jensen, L.J.; Julien, P.; Kuhn, M.; Von Mering, C.; Muller, J.; Doerks, T.; Bork, P. eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **2008**, *36*, D250–D254. [[CrossRef](#)] [[PubMed](#)]



139. Kriventseva, E.V.; Tegenfeldt, F.; Petty, T.J.; Waterhouse, R.M.; Simão, F.A.; Pozdnyakov, I.A.; Ioannidis, P.; Zdobnov, E.M. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.* **2015**, *43*, D250–D256. [[CrossRef](#)] [[PubMed](#)]
140. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **2003**, *13*, 2178–2189. [[CrossRef](#)]
141. Linard, B.; Thompson, J.D.; Poch, O.; Lecompte, O. OrthoInspector: Comprehensive orthology analysis and visual exploration. *BMC Bioinform.* **2011**, *12*, 11. [[CrossRef](#)]
142. Emms, D.M.; Kelly, S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, 157. [[CrossRef](#)]
143. Train, C.-M.; Glover, N.M.; Gonnet, G.H.; Altenhoff, A.M.; Dessimoz, C. Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **2017**, *33*, i75–i82. [[CrossRef](#)]
144. Dalquen, D.A.; Dessimoz, C. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol. Evol.* **2013**, *5*, 1800–1806. [[CrossRef](#)]
145. Li, H.; Coghlan, A.; Ruan, J.; Coin, L.J.; Hériché, J.K.; Osmotherly, L.; Li, R.; Liu, T.; Zhang, Z.; Bolund, L.; et al. TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **2006**, *34*, D572–D580. [[CrossRef](#)]
146. Poptsova, M.S.; Gogarten, J.P. BranchClust: A phylogenetic algorithm for selecting gene families. *BMC Bioinform.* **2007**, *8*, 120. [[CrossRef](#)] [[PubMed](#)]
147. Penel, S.; Arigon, A.M.; Dufayard, J.F.; Sertier, A.S.; Daubin, V.; Duret, L.; Gouy, M.; Perrière, G. Databases of homologous gene families for comparative genomics. *BMC Bioinform.* **2009**, *10* (Suppl. 6), S3. [[CrossRef](#)] [[PubMed](#)]
148. Huerta-Cepas, J.; Capella-Gutierrez, S.; Pryszcz, L.P.; Denisov, I.; Kormes, D.; Marcet-Houben, M.; Gabaldón, T. PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* **2011**, *39*, D556–D560. [[CrossRef](#)] [[PubMed](#)]
149. Storm, C.E.V.; Sonnhammer, E.L.L. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **2002**, *18*, 92–99. [[CrossRef](#)]
150. Berglund-Sonnhammer, A.C.; Steffansson, P.; Betts, M.J.; Liberles, D.A. Optimal Gene Trees from Sequences and Species Trees Using a Soft Interpretation of Parsimony. *J. Mol. Evol.* **2006**, *63*, 240–250. [[CrossRef](#)]
151. Van der Heijden, R.T.J.M.; Snel, B.; Van Noort, V.; Huynen, M.A. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinform.* **2007**, *8*, 83. [[CrossRef](#)]
152. Goodman, M.; Czelusniak, J.; Moore, G.W.; Romero-Herrera, A.E.; Matsuda, G. Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Syst. Biol.* **1979**, *28*, 132–163. [[CrossRef](#)]
153. Åkerborg, Ö.; Sennblad, B.; Arvestad, L.; Lagergren, J. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 5714–5719. [[CrossRef](#)]
154. Liu, D.; Hunt, M.; Tsai, I.J. Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinform.* **2018**, *19*, 26. [[CrossRef](#)]
155. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [[CrossRef](#)]
156. Haug-Baltzell, A.; Stephens, S.A.; Davey, S.; Scheidegger, C.E.; Lyons, E. SynMap2 and SynMap3D: Web-based whole-genome synteny browsers. *Bioinformatics* **2017**, *33*, 2197–2198. [[CrossRef](#)] [[PubMed](#)]
157. Wang, Y.; Tang, H.; DeBarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*, e49. [[CrossRef](#)] [[PubMed](#)]
158. Proost, S.; Fostier, J.; De Witte, D.; Dhoedt, B.; Demeester, P.; Van de Peer, Y.; Vandepoele, K. i-ADHoRe 3.0—Fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **2012**, *40*, e11. [[CrossRef](#)] [[PubMed](#)]
159. Rödelsperger, C.; Dieterich, C. CYN TENATOR: Progressive Gene Order Alignment of 17 Vertebrate Genomes. *PLoS ONE* **2010**, *5*, e8861. [[CrossRef](#)] [[PubMed](#)]

160. Drillon, G.; Carbone, A.; Fischer, G. SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. *PLoS ONE* **2014**, *9*, e92621. [[CrossRef](#)] [[PubMed](#)]
161. Cannon, S.B.; Kozik, A.; Chan, B.; Michelmore, R.; Young, N.D. DiagHunter and GenoPix2D: Programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol.* **2003**, *4*, R68. [[CrossRef](#)] [[PubMed](#)]
162. Calabrese, P.P.; Chakravarty, S.; Vision, T.J. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **2003**, *19*, i74–i80. [[CrossRef](#)]
163. Soderlund, C.; Nelson, W.; Shoemaker, A.; Paterson, A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* **2006**, *16*, 1159–1168. [[CrossRef](#)]
164. Sinha, A.U.; Meller, J. Cinteny: Flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinform.* **2007**, *8*, 82. [[CrossRef](#)]
165. Haas, B.J.; Delcher, A.L.; Wortman, J.R.; Salzberg, S.L. DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics* **2004**, *20*, 3643–3646. [[CrossRef](#)]
166. Hampson, S.; McLysaght, A.; Gaut, B.; Baldi, P. LineUp: Statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.* **2003**, *13*, 999–1010. [[CrossRef](#)] [[PubMed](#)]
167. Wang, X.; Shi, X.; Li, Z.; Zhu, Q.; Kong, L.; Tang, W.; Ge, S.; Luo, J. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinform.* **2006**, *7*, 447. [[CrossRef](#)]
168. Pham, S.K.; Pevzner, P.A. DRIMM-Synteny: Decomposing genomes into evolutionary conserved segments. *Bioinformatics* **2010**, *26*, 2509–2516. [[CrossRef](#)] [[PubMed](#)]
169. Paten, B.; Herrero, J.; Beal, K.; Fitzgerald, S.; Birney, E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **2008**, *18*, 1814–1828. [[CrossRef](#)]
170. Xu, A.W.; Moret, B.M.E. GASTS: Parsimony Scoring under Rearrangements. In *Algorithms in Bioinformatics, Proceedings of the 11th International Workshop, WABI 2011, Saarbrücken, Germany, 5–7 September 2011*; Przytycka, T.M., Sagot, M.F., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 351–363.
171. Zhou, L.; Feng, B.; Yang, N.; Tang, J. Ancestral reconstruction with duplications using binary encoding and probabilistic model. In *Proceedings of the 7th International conference on Bioinformatics and Computational Biology, Honolulu, HI, USA, 9–11 March 2015*; pp. 97–104.
172. Yang, N.; Hu, F.; Zhou, L.; Tang, J. Reconstruction of Ancestral Gene Orders Using Probabilistic and Gene Encoding Approaches. *PLoS ONE* **2014**, *9*. [[CrossRef](#)]
173. Feng, B.; Zhou, L.; Tang, J. Ancestral Genome Reconstruction on Whole Genome Level. *Curr. Genom.* **2017**, *18*, 306–315. [[CrossRef](#)]
174. Lucas, J.M.; Muffato, M.; Crollius, H.R. PhylDiag: Identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinform.* **2014**, *15*. [[CrossRef](#)]
175. Hachiya, T.; Osana, Y.; Popenndorf, K.; Sakakibara, Y. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics* **2009**, *25*, 853–860. [[CrossRef](#)]
176. Baek, J.H.; Kim, J.; Kim, C.K.; Sohn, S.H.; Choi, D.; Ratnaparkhe, M.B.; Kim, D.W.; Lee, T.H. MultiSyn: A Webtool for Multiple Synteny Detection and Visualization of User’s Sequence of Interest Compared to Public Plant Species. *Evol. Bioinform.* **2016**. [[CrossRef](#)]
177. Zeng, X.; Nesbitt, M.J.; Pei, J.; Wang, K.; Vergara, I.A.; Chen, N. OrthoCluster: A new tool for mining synteny blocks and applications in comparative genomics. In *Advances in database technology, Proceedings of the 11th international conference on Extending database technology, Nantes, France, 25–29 March 2008*; Association for Computing Machinery: New York, NY, USA, 2008; pp. 656–667.
178. Fostier, J.; Proost, S.; Dhoedt, B.; Saeys, Y.; Demeester, P.; Van de Peer, Y.; Vandepoele, K. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **2011**, *27*, 749–756. [[CrossRef](#)]
179. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)] [[PubMed](#)]
180. Simillion, C.; Vandepoele, K.; Saeys, Y.; Van de Peer, Y. Building Genomic Profiles for Uncovering Segmental Homology in the Twilight Zone. *Genome Res.* **2004**, *14*, 1095–1106. [[CrossRef](#)] [[PubMed](#)]
181. Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441. [[CrossRef](#)] [[PubMed](#)]
182. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [[CrossRef](#)]

183. Pevzner, P.; Tesler, G. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Res.* **2003**, *13*, 37–45. [[CrossRef](#)]
184. Swidan, F.; Rocha, E.P.C.; Shmoish, M.; Pinter, R.Y. An Integrative Method for Accurate Comparative Genome Mapping. *PLoS Comput. Biol.* **2006**, *2*. [[CrossRef](#)]
185. Paten, B.; Earl, D.; Nguyen, N.; Diekhans, M.; Zerbino, D.; Haussler, D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **2011**, *21*, 1512–1528. [[CrossRef](#)]
186. Paten, B.; Diekhans, M.; Earl, D.; St. John, J.; Ma, J.; Suh, B.; Haussler, D. Cactus Graphs for Genome Comparisons. In *Research in Computational Molecular Biology, Proceedings of the 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, 25–28 April 2010*; Berger, B., Ed.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 410–425.
187. Zhang, L.; Gaut, B.S. Does Recombination Shape the Distribution and Evolution of Tandemly Arrayed Genes (TAGs) in the *Arabidopsis thaliana* Genome? *Genome Res.* **2003**, *13*, 2533–2540. [[CrossRef](#)]
188. Morgulis, A.; Gertz, E.M.; Schäffer, A.A.; Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **2006**, *13*, 1028–1040. [[CrossRef](#)]
189. Wootton, J.C.; Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **1996**, *266*, 554–571. [[CrossRef](#)]
190. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580. [[CrossRef](#)] [[PubMed](#)]
191. Frith, M.C. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* **2011**, *39*, 1–8. [[CrossRef](#)] [[PubMed](#)]
192. Olson, D.; Wheeler, T. ULTRA: A Model Based Tool to Detect Tandem Repeats. In Proceedings of the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 29 August–1 September 2018; Association for Computing Machinery: Washington, DC, USA, 2018; pp. 37–46.
193. Soylev, A.; Le, T.M.; Amini, H.; Alkan, C.; Hormozdiari, F. Discovery of tandem and interspersed segmental duplications using high-throughput sequencing. *Bioinformatics* **2019**, *35*, 3923–3930. [[CrossRef](#)] [[PubMed](#)]
194. Genovese, L.M.; Mosca, M.M.; Pellegrini, M.; Geraci, F. Dot2dot: Accurate whole-genome tandem repeats discovery. *Bioinformatics* **2019**, *35*, 914–922. [[CrossRef](#)] [[PubMed](#)]
195. Audemard, E.; Schiex, T.; Faraut, T. Detecting long tandem duplications in genomic sequences. *BMC Bioinform.* **2012**, *13*, 83. [[CrossRef](#)]
196. Zheng, D.; Gerstein, M.B. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol.* **2006**, *7* (Suppl. 1), S13. [[CrossRef](#)]
197. Despons, L.; Baret, P.V.; Frangeul, L.; Louis, V.L.; Durrens, P.; Souciet, J.-L. Genome-wide computational prediction of tandem gene arrays: Application in yeasts. *BMC Genom.* **2010**, *11*, 56. [[CrossRef](#)]
198. Herrero, J.; Muffato, M.; Beal, K.; Fitzgerald, S.; Gordon, L.; Pignatelli, M.; Vilella, A.J.; Searle, S.M.J.; Amode, R.; Brent, S.; et al. Ensembl comparative genomics resources. *Database* **2016**, *2016*. [[CrossRef](#)]
199. Lee, J.; Hong, W.; Cho, M.; Sim, M.; Lee, D.; Ko, Y.; Kim, J. Synteny Portal: A web-based application portal for synteny block analysis. *Nucleic Acids Res.* **2016**, *44*, W35–W40. [[CrossRef](#)]
200. Muffato, M.; Louis, A.; Poisnel, C.E.; Croliius, H.R. Genomicus: A database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **2010**, *26*, 1119–1121. [[CrossRef](#)]
201. Ma, J.; Zhang, L.; Suh, B.B.; Raney, B.J.; Burhans, R.C.; Kent, W.J.; Blanchette, M.; Haussler, D.; Miller, W. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **2006**, *16*, 1557–1565. [[CrossRef](#)]
202. Loots, G.; Ovcharenko, I. ECRbase: Database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics* **2007**, *23*, 122–124. [[CrossRef](#)] [[PubMed](#)]
203. Ng, M.P.; Vergara, I.A.; Frech, C.; Chen, Q.; Zeng, X.; Pei, J.; Chen, N. OrthoClusterDB: An online platform for synteny blocks. *BMC Bioinform.* **2009**, *10*, 192. [[CrossRef](#)] [[PubMed](#)]
204. Sonnhammer, E.L.L.; Östlund, G. InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **2015**, *43*, D234–D239. [[CrossRef](#)] [[PubMed](#)]
205. Miele, V.; Penel, S.; Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **2011**, *12*, 116. [[CrossRef](#)] [[PubMed](#)]
206. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]



207. Altenhoff, A.M.; Glover, N.M.; Train, C.-M.; Kaleb, K.; Warwick Vesztrocy, A.; Dylus, D.; De Farias, T.M.; Zile, K.; Stevenson, C.; Long, J.; et al. The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **2018**, *46*, D477–D485. [[CrossRef](#)]
208. Van Bel, M.; Diels, T.; Vancaester, E.; Kreft, L.; Botzki, A.; Van de Peer, Y.; Coppens, F.; Vandepoele, K. PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **2018**, *46*, D1190–D1196. [[CrossRef](#)]
209. Conte, M.G.; Gaillard, S.; Lanau, N.; Rouard, M.; Périn, C. GreenPhylDB: A database for plant comparative genomics. *Nucleic Acids Res.* **2008**, *36*, D991–D998. [[CrossRef](#)]
210. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, D1178–D1186. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

## La dominance de sous-génome

Les duplications complètes de génomes par WGD mènent à l'obtention de deux, ou plus, copies du génome au sein d'un seul organisme. Les WGDs sont prévalentes dans l'histoire des plantes et en particulier des angiospermes (Clark & Donoghue, 2018; Panchy et al., 2016; Qiao et al., 2019; Van de Peer et al., 2021; Van de Peer et al., 2009; Van de Peer et al., 2017; T. E. Wood et al., 2009). Les mécanismes à l'origine de WGDs peuvent être un doublement d'un génome diploïde, on parle dans ce cas d'autopolyploïdie (Tate et al., 2005) ou une hybridation interspécifique définie par le terme d'allopolyploïdie (Wendel, 2015). La redondance des gènes supprime la contrainte sélective et fournit davantage de ressources génétiques, permettant une innovation génétique plus rapide. La WGD entraîne un isolement reproductif et peut constituer un mécanisme majeur de spéciation (Adams & Wendel, 2005; Shimizu-Inatsugi et al., 2017). De plus, les WGDs peuvent être à l'origine de changements de vigueur (hétérosis), de changements des systèmes reproductifs ou de déséquilibres épistatiques (Soltis et al., 2015). Les WGDs, et en particulier les allopolyploïdies, vont être à l'origine d'instabilité importante du génome entre autres à cause de la coexistence de sous-génomes provenant d'espèces progénitrices différentes. Ce processus est décrit par le terme de choc génomique (McClintock, 1984) et va avoir différentes implications.

Le choc génomique va aboutir à une levée des méthylations des ET, qui va être à l'origine de l'activation massive d'ET (Springer et al., 2016) et d'une explosion de transpositions dans le génome dupliqué (An et al., 2014; Parisod et al., 2009). Par ailleurs, une levée de la contrainte sélective du fait de la redondance peut aussi induire une hausse de l'activité d'ET (Ågren et al., 2016). C'est pourquoi, une augmentation du nombre d'ET après une WGD a été constatée chez de nombreuses espèces (Casacuberta & González, 2013; Wendel, 2015); notamment chez le maïs (J. C. Schnable et al., 2011), le framboisier (Edger et al., 2019), la fleur de singe (*Mimulus ringens*) (Edger et al., 2017), le coton (L. Flagel et al., 2008; L. E. Flagel & Wendel, 2010) ou *Brassica rapa* (Cheng et al., 2016). Une explosion du nombre d'ET a été rapportée chez *M. domestica*, et datée à 21 millions d'années (Daccord et al., 2017), ce qui correspond à la période suivant la WGD des *Maloideae*.

Cette activation et accumulation importante d'ET dans le génome dupliqué va avoir une série de conséquences. D'une part, les ET vont impacter les mécanismes de restructuration du génome post-WGD par le biais de recombinaison entre les ET. D'autre part,

l'insertion des ET peut aussi être à l'origine de pertes de gènes et de mutations. C'est impacts massifs sur le génome vont aboutir à une inactivation des ET par un mécanisme épigénétique qui suit leur néo-insertion dans le génome. L'inactivation des ET est réalisée par méthylation de l'ADN (Feschotte et al., 2002; Hollister & Gaut, 2009; Parisod & Senerchia, 2012; J. Zhang et al., 2015) et peut ainsi affecter l'expression des gènes localisés à proximité (Freeling et al., 2015). En effet, la méthylation des ET peut impacter la transcription par divers mécanismes et notamment le débordement de la méthylation et/ou un remodelage de la chromatine qui réduit l'accessibilité des facteurs de transcription (Cheng et al., 2016; Hollister & Gaut, 2009). Ces différents mécanismes sont en partie liés aux familles des ET retrouvés à proximité des gènes (J. Y. Choi & Purugganan, 2018; Eichten et al., 2012; Noshay et al., 2019). Par exemple chez *M. ringens*, le remodelage de la méthylation des ET après l'hybridation interspécifique initiale reflète les niveaux d'expression des gènes voisins (Edger et al., 2017). Par ailleurs, le niveau de méthylation de l'ADN a été relié aux différentes classes d'ET (Yaakov & Kashkush, 2012).

Chez les espèces dont le génome a été dupliqué par allodiploïdisation, le mécanisme d'explosion d'ET et les méthylation de l'ADN en résultant sont biaisées entre les sous-génomes. Cela peut être expliqué par différents processus. Tout d'abord, les deux génomes parentaux possèdent des efficacités de répression des ET différentes, en particulier à cause de proportion d'ET et de localisation d'ET différentes entre les deux sous-génomes (Garsmeur et al., 2014). Cette différence originelle dans la répartition des ET va être le moteur de la différenciation des sous-génomes (Freeling et al., 2015; Woodhouse et al., 2014). En effet, il semblerait que le sous-génome amené à dominer est celui portant le moins de ET originellement. Par exemple chez *B. rapa*, il a été observé que les Small Interfering Ribonucleic Acids (siRNAs) 24 nt ciblent les régions en amont des gènes associés au sous-génome dominé (Woodhouse et al., 2014).

Ainsi, les ET et en particulier leur répression épigénétique, vont aboutir à des méthylation de l'ADN dans leur environnement. Leur répartition biaisée va permettre la mise en place de différences de méthylation de l'ADN et de modifications de chromatine entre les deux sous-génomes. Ces différences épigénétiques peuvent avoir des répercussions sur la chromatine et impacter ainsi l'activité transcriptionnelle des gènes associés (Kashkush et al., 2003). La mise en place de ce processus sera donc d'autant plus rapide que les espèces parentales sont différentes en termes d'ET (Cheng et al., 2016). En conséquence, pour les organismes autoploïdes où il n'y a pas de différences d'ET préexistants, ce type de processus ne se met pas en place. Ainsi, le sous-génome présentant le moins d'ET

méthylés, va présenter un profil transcriptionnel dominant par rapport au sous-génome dominé qui est moins souvent exprimé. Il est à noter que ce mécanisme n'est pas de type on/off. Le sous-génome dominé présente toujours des gènes exprimés et certains ont des niveaux supérieurs à ceux présents dans le sous-génome dominant (Edger et al., 2017). Chez certains organismes comme la myrtille (Colle et al., 2019), le blé (A. Li et al., 2014) et le coton (L. Flagel et al., 2008), des inversions de la dominance du sous-génome ont été observées selon des stades de développement et/ou les organes. Ces éléments seront détaillés dans les chapitres associés à la question des traits phénotypiques et de la transcription.

Après la WGD, afin de rétablir la stabilité du génome, différents mécanismes vont se mettre en place. Ces processus sont définis sous le terme de diploïdisation (Soltis et al., 2015). Ces mécanismes permettent une réduction de la taille du génome dupliqué, ce qui explique que malgré les nombreuses WGDs présentes dans l'histoire des plantes, les génomes restent de taille contenue par rapport aux nombreux événements de duplication (Bennetzen & Kellogg, 1997; Panchy et al., 2016). La diploïdisation est en partie médiée par les ET qui autorisent des recombinaisons et des réarrangements chromosomiques engendrant la perte de fragments chromosomiques et de gènes dupliqués (Vicent et al., 1999). Ceci a été observé chez de nombreux organismes, par exemple *Nicotiana tabacum* (Lim et al., 2007) et le maïs (Bruggmann et al., 2006). Étant donné que les processus précédemment décrits permettent la mise en place de biais dans la répartition des ET, les mécanismes de perte de gènes peuvent aussi être de ce fait biaisés. Par ailleurs, le fait qu'un biais d'expression soit aussi présent va engendrer un biais dans la pression de sélection appliquée aux séquences codantes des gènes dupliqués et une perte de gènes moins impactante sur le phénotype pour le sous-génome dit dominé. Ainsi, chez les espèces où le génome a été doublé par WGD, il a été observé un déséquilibre dans la perte des gènes dupliqués. Ce phénomène, observé pour la première fois chez le maïs, est décrit comme le biais de fractionnement (Woodhouse et al., 2010). Il a depuis été observé chez de nombreuses espèces, notamment *N. tabacum* (Renny-Byfield et al., 2011) et *Arabidopsis thaliana* (Freeling & Thomas, 2006) et le maïs (*Zhea maize*) (Woodhouse et al., 2010).

L'ensemble des éléments aboutissent à un biais de participation au phénotype de l'organisme entre les sous-génomes. Certains sous-génomes vont donc plus participer à la variation phénotypique que d'autres. Ce mécanisme a été observé chez différentes espèces comme le maïs (Renny-Byfield et al., 2017), *B. rapa* (Z. Wang et al., 2022), la myrtille (Colle et al., 2019), le blé (A. Li et al., 2014) et le coton (L. Flagel et al., 2008).

Pour résumer, la dominance sous-génomique consiste en un déséquilibre dans la proportion d'ET après l'explosion d'ET se produisant à la suite de la WGD. Ce déséquilibre va aboutir à un déséquilibre dans la répression épigénétique de l'expression des gènes. Les mécanismes liés au retour à la diploïdie vont alors être biaisés avec une perte préférentielle des gènes du sous-génome dominé, dû à la plus grande proportion d'ET, mais aussi à une participation moindre au phénotype dû au déséquilibre transcriptionnel qui va lever la pression de sélection appliquée sur ce sous-génome. Cet ensemble de mécanismes connus comme la dominance sous génomique a été décrit, à l'heure actuelle, seulement chez des espèces allopolyploïdes.

# QUESTIONS DE RECHERCHE

---

Chez différentes espèces pour lesquelles une WGD récente a été identifiée, il a été observé un déséquilibre dans la répartition du pourcentage d'explication de la variance de nombreux et divers traits phénotypiques entre fragments de chromosomes considérés comme synténiques. Une duplication complète du génome du pommier par autopolyploïdie a permis le passage d'un génome de neuf chromosomes à 17. Cette WGD, récente et bien conservée fait de *M. domestica* un organisme de choix pour l'étude du devenir des gènes dupliqués chez le pommier. Chez *M. domestica*, le nombre important de QTLs identifiés à partir des années 2000 a permis d'observer, de manière empirique, une tendance indiquant que certains chromosomes seraient privilégiés en terme de participation à l'explication de la variance de traits phénotypiques d'importance agronomique. Ainsi au cours de cette thèse, nous avons entrepris de répondre aux deux questions biologiques suivantes :

## Première question

Les proportions de QTLs expliquant une part significative des traits phénotypiques mesurés chez le pommier et localisés sur les fragments chromosomiques ohnologues sont-elles réparties de manières différentes ?

## Deuxième question

Si une telle répartition différentielle des QTLs est observée, quels seraient les mécanismes épi/génétiques et/ou transcriptomiques à l'origine du déséquilibre observé ?



# MATÉRIEL ET MÉTHODE

---

## 1.1 Environnement informatique

L'ensemble des développements informatiques liés à la thèse ont été faits dans une démarche FAIR (*F*indable, *A*ccessible, *I*nteroperable, and *R*e-usable). Cette démarche vise à améliorer la qualité des développements et des données associées et repose sur différents piliers (Wilkinson et al., 2016). Tout d'abord, le code et les données en résultant doivent être accessibles. Ainsi, l'ensemble des scripts écrits sont versionnés à l'aide de git (Chacon & Straub, 2014). Cette gestion de version du code à l'aide de git permet de conserver un historique du développement, facilite le développement collaboratif et le déploiement du code sur les ressources de calculs. Par ailleurs, cette approche est communément utilisée en informatique. L'ensemble des développements sont stockés sur des dépôts sur un serveur gitLab de la forgeMIA, un service de forge logiciel fournit pour les utilisateurs INRAe. Cette forge, en plus de servir de serveur central agrégeant les développements faits, présente l'avantage de proposer des fonctionnalités supplémentaires comme l'intégration continue et de déploiement continu (*Continuous Integration/ Continuous deployment*, CI/CD). Cette approche permet de mettre en place un pipeline de validation et de déploiement du code avec une fréquence plus élevée, permettant la production d'un code plus fiable et de meilleure qualité.

La forgeMia permet aussi de rendre le code simplement accessible, puisque celui-ci est sauvegardé sur un serveur accessible sur internet, dont le contrôle par authentification est possible et présentant un moteur de recherche et une indexation des projets. Cette forge logicielle permet, en plus d'archiver et de mettre à disposition les codes sources relatifs à la thèse de mettre à disposition un wiki et un système de remonter de bogue. De plus, l'ensemble des projets est documenté afin d'offrir un déploiement du code et de ses dépendances facilité. Les codes sources sont commentés selon la norme associée à la bibliothèque *numpy* (Van der Walt & Virtanen, s. d.).

L'interopérabilité des différents outils utilisés ou développés est octroyée par l'utili-



sation des outils au sein de conteneurs Docker (Avram, 2013). Cette technologie permet d’empaqueter et de déployer des logiciels. Ainsi l’ensemble des dépendances nécessaires à l’exécution de l’application sont stockées sous forme d’un ensemble de couches formant une image. Les conteneurs, qui permettent l’exécution de l’application, sont alors une instance déployée d’une image, qui rassemble la pile de logiciels nécessaires à l’outil pour fonctionner de manière isolée du reste du système. Les conteneurs reposent néanmoins sur le noyau du système d’exploitation hôte et s’appuient sur lui pour effectuer certaines actions de base comme la mise en réseau et l’écriture de fichiers. Cette approche permet d’alléger la mise en place du système par rapport à une machine virtuelle qui doit embarquer l’ensemble du système d’exploitation en plus de l’application et de ses dépendances. Les images construites au cours de la thèse sont stockées sur la forge logicielle dans les registres de conteneurs des dépôts associés aux codes sources sous la forme d’images directement déployables. Cette technologie garantit un déploiement sûr, mais présente quelques lourdeurs notamment lors de la communication entre différentes applications déployées dans des conteneurs ou lorsque la persistance des données est nécessaire. C’est pourquoi une partie des analyses ne nécessitant pas de pile de logiciels complexes ont été mises en place au sein d’environnements virtuels gérés avec la distribution *Anaconda* (« Anaconda software distribution », 2020) et la forge de *conda* (Community, 2015). Cette technologie permet d’avoir une version de Python (R et Julia sont aussi possibles) et de ses dépendances isolées du système. Ces technologies permettent aussi de fixer les versions et contrôlent les montées en version des outils et de leur dépendance qui seront consistantes et reproductibles. Ces deux technologies permettent d’empaqueter le code et ses dépendances permettant aussi de garantir que le code produit pendant la thèse sera exécutable même après des mises à jour majeures des outils, des langages ou des systèmes d’exploitation.

Afin de produire un code reproductible, évolutif, hautement parallélisable, réutilisable et facile d’utilisation y compris pour des personnes non-programmeurs, les scripts sont intégrés dans des pipelines à l’aide de Snakemake (Köster & Rahmann, 2012) ayant évolué au cours de la thèse de la version 5 à 7.12.1. Ce système de gestion des flux de travail permet de créer et gérer les différentes étapes des analyses de données via un langage basé sur Python. Les pipelines écrits peuvent être sans effort déployés sur des clusters ou des grilles de calculs. En effet, la description de l’ensemble des outils et de leurs versions associées empaquetées dans des environnements *conda* ou des conteneurs docker permet un déploiement automatique sur tout type d’infrastructure informatique. Ce formalisme

permet aussi d’empaqueter ses propres outils afin de les rendre interopérables et réutilisables entre les projets. Pour finir, cet outil open source et actif m’a permis d’y effectuer des développements intégrés notamment en version 7.3.2. Ces développements visaient à résoudre un ensemble de bogues qui faisaient échouer l’exécution des étapes du pipeline dans des environnements construits dans des conteneurs Singularity.

Pour l’ensemble des développements de cette thèse, les principaux langages utilisés sont le shell, Python en version 3.8 à 3.10 (Oliphant, 2007) et dans une moindre mesure R (Team, 2013) et JavaScript. Le développement en Python est fait en suivant les recommandations PEP 8 mis à part la taille de l’indentation passée à deux espaces pour suivre les règles ayant cours au sein de l’équipe dans laquelle j’évolue qu’est l’équipe *Bioinformatics for plant Defense Investigations* (BiDefI). Différentes bibliothèques de fonctions Python ont été utilisées afin de faciliter et accélérer le développement. La gestion des données structurées et la vectorisation des itérations ont été permises par *pandas* (McKinney, 2010). La montée en charge, la gestion de gros volumes de données et la parallélisation des calculs ont été faites avec *Dask* (Rocklin, 2015). Les bibliothèques *numpy* (van der Walt et al., 2011) et *Scipy* (Virtanen et al., 2020) ont été utilisées pour différents traitements mathématiques et tests statistiques. Les opérations de classification et de machine *learning* ont été faites à l’aide de *scikit-learn* (Pedregosa et al., 2011). La construction des visualisations a été faite à l’aide de *matplotlib* (Hunter, 2007), *seaborn* en version 0.10.1 (Waskom et al., 2020) et Plotly (Inc., 2015) des versions 4.2.1 à 5.10.0. Les requêtes http ont été basées sur la bibliothèque *requests* (Chandra & Varanasi, 2015) et l’analyse des fichiers HTML en résultant a été réalisé avec *Beautiful soup* (Richardson, 2007).

Les développements en R se sont principalement appuyés sur les collections de bibliothèques de *tidyverse* (Wickham et al., 2019) et *easystats* (Lüdecke et al., 2022) en particulier le paquet *correlations* (Makowski et al., 2020). Les développements en JavaScript ont essentiellement des buts de visualisation et reposent sur *D3.js* (Bostock, s. d.) et *Observable* (« Observable - Explore, analyze, and explain data. As a team. » s. d.).

La traçabilité des analyses et la communication des résultats est permise par des *Jupyter notebooks* (Perez & Granger, 2007). Cette application web offre de construire des programmes contenant à la fois du code et du texte au format *markdown*. Les résultats de ces programmes peuvent alors être exportés au format HTML et PDF. De plus, un blog permet de garder une trace et une communication hebdomadaire des développements et analyses conduites avec l’ensemble des membres du projet. Celui-ci est codé en *markdown*, un langage de balises léger ainsi que *Hugo*, un *framework* permettant la compilation de

fichiers *markdown* en site HTML statique.

Différentes analyses ont été menées au cours de cette thèse dont une partie est associée à des technologies omics NGS très consommatrices d'espaces disque comme le transcriptomique et les séquençages au bisulfite. L'ensemble des données stockées et produites au cours de la thèse représente un volume de 7 To. Le stockage des données brutes et traitées a été fait sur des baies de stockages fournies par la plateforme bio-informatique GenOuest (<https://www.genouest.org>) ainsi que les clusters de l'IRHS. Les ressources de calculs ont aussi été fournies par ces infrastructures.

## 1.2 Statistiques

Afin de valider les observations faites au cours de la thèse, différents tests d'hypothèses ont été mis en place. Ces tests statistiques permettent de prendre des décisions entre deux hypothèses à partir des données observées en prévoyant le risque que la conclusion soit erronée. Différents tests peuvent être utilisés suivant les différentes caractéristiques de données observées et des hypothèses posées. Étant donné les design expérimentaux et hypothèses testées au cours de cette thèse, de multiples tests d'hypothèses ont été utilisés. Le détail des tests utilisés ainsi que leurs conditions d'utilisations sont détaillées dans cette section.

### 1.2.1 Tests de comparaison d'une seule variable

Dans de nombreuses analyses, nous avons cherché à observer si le comportement d'une variable présentait des différences, au sein d'une même population ou entre des populations. Pour ces tests, nous avons considéré les chromosomes comme des populations différentes. Si les groupes de gènes étudiés au sein des chromosomes pouvaient être liés entre eux notamment via des homologies ou des positions, nous avons considéré les populations comme appariées. Les tests d'hypothèses non appariés ont été utilisés lorsque les données représentaient une population indépendante par exemple des groupes de gènes entre lesquels aucun lien n'a pu être construit.

#### Test U de Mann-Whitney

Le test de Wilcoxon-Mann-Whitney (Mann & Whitney, 1947), aussi appelé test U de Mann-Whitney, est un test d'hypothèse non paramétrique. Ce test vise à tester l'hypothèse

selon laquelle les distributions de chacun de deux groupes de données sont proches. Il suppose que les données soient indépendantes et ordonnées. Ainsi, pour ce test l'hypothèse nulle,  $H_0$  est définie comme « les distributions associées à chacune des populations sont égales ». La statistique de test est définie telle que présentée dans l'Équation 1.1.

$$U = \sum_{i=1}^n \sum_{j=1}^m S(X_i, Y_j) \quad (1.1)$$

Avec,

$$U = \sum_{i=1}^n \sum_{j=1}^m \begin{cases} 1 & Y_i < X_j \\ \frac{1}{2} & Y_i = X_j \\ 0 & Y_i > X_j \end{cases} \quad (1.2)$$

et,

$X_i$  La valeur de l'échantillon  $i$  dans  $X$

$Y_j$  La valeur de l'échantillon  $j$  dans  $Y$

Un exemple de ce test peut être retrouvé en Section 7.3.2 lorsque l'on souhaite tester si la distribution de la couverture en ET entre une population de gènes dupliqués et des gènes retrouvés en un seul exemplaire est similaire. Ces deux populations peuvent être ordonnées le long des chromosomes et sont indépendantes. Ainsi, un test U de Mann-Whitney peut être utilisé dans ce cadre.

### Test des rangs signés de Wilcoxon

Le test des rangs signés de Wilcoxon (Wilcoxon, 1945) est un test d'hypothèse non paramétrique. Ce test suppose un appariement des données et des données ordonnées. De plus, le calcul de la différence entre les valeurs doit comporter un sens du point de vue des données testées, ici un sens biologique. Ainsi, ce test présente les mêmes conditions que le test des signes et permet de répondre à une question similaire, mais présente une meilleure puissance statistique, c'est pourquoi il a été privilégié quand cela était possible. La statistique du test est définie selon l'Équation 1.3. Pour ce test qui s'intéresse à la médiane de la différence entre les distributions l'hypothèse nulle est définie comme « les médianes des distributions associées à chacune des populations sont égales ».

$$T^+ = \sum_{i=1}^n R_i \psi_i \quad (1.3)$$

Avec,

$T^+$  statistique de test des rangs signés de Wilcoxon

$R_i$  le rang de la valeur absolue de la différence  $Z_i$

$\psi_i$  la fonction indicatrice, i.e. égale à 1 si  $Z_i > 0$  et 0  $Z_i \leq 0$

Un exemple de ce test peut être retrouvé en Section 6.2.3 lorsque l'on souhaite tester si la distribution des niveaux d'expression des gènes ohnologues est similaire. Ces deux populations peuvent être ordonnés appariés par le lien d'ohnologie et la différence de niveau d'expression entre deux gènes à un sens biologique. Ainsi, un test de rangs signés de Wilcoxon peut être utilisé dans ce cadre.

### Student t test

Les t-tests de Student (Student, 1908) regroupent un ensemble de test d'hypothèses qui se basent sur une statistique de test qui suit la distribution  $t$  de Student. Ces tests sont paramétriques. Ils nécessitent la normalité des moyennes de l'échantillon, et la variance de l'échantillon doit suivre une distribution de  $\chi^2$ . De plus, la moyenne et la variance de l'échantillon doivent être statistiquement indépendantes. De plus, la normalité des valeurs de la population testée doit être établie, dans le cas d'échantillon de taille importante, le théorème central limite permet d'estimer la distribution comme normale. À l'image du test des rangs signés de Wilcoxon, les tests  $t$  s'intéressent à tester un paramètre des distributions étudiées, ici c'est la moyenne des distributions qui est testées. Ainsi, l'hypothèse nulle est définie comme « les moyennes des distributions associées à chacune des populations sont égales ». Ces catégories de tests regroupent différents tests. Dans le cadre de cette thèse, nous avons utilisé le test dans sa version adaptée aux données non appariées dont la statistique de test est présentée dans l'Équation 1.4 et la version adaptée pour les échantillons appariés (Équation 1.5). Les tests  $t$  appariés permettent une plus grande puissance statistique notamment en évitant les erreurs de faux négatifs (erreur de type II).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}} \sqrt{\frac{2}{n}}} \quad (1.4)$$

Avec,

$\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$  la déviation standard groupée pour les populations de taille  $n$

$s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2$  les estimateurs non biaisés des variances des deux échantillons

$n$  la taille de la population 1 et 2

$\bar{X}_1$  et  $\bar{X}_2$  les moyennes des deux populations

Ce test a été utilisé dans des conditions similaires au test U de Mann-Whitney.

$$t = \frac{\bar{X}_D - \mu_0}{s_D / \sqrt{n}} \quad (1.5)$$

Avec,

$\bar{X}_D$  la moyenne des différences entre toutes les paires

$s_D$  l'écart type des différences entre toutes les paires

$\mu_0$

Ce test a été utilisé dans la Section 3.2 pour tester si la distribution des pourcentages des gènes ohnologues entre les chromosomes ohnologues était similaire.

### Tests de $\chi^2$

Les tests de  $\chi^2$  (Pearson, 1900) sont un ensemble de tests statistiques qui se basent sur une statistique de test qui suit la distribution  $\chi^2$  sous l'hypothèse nulle. Ce sont des tests non paramétriques. Différentes déclinaisons du test du  $\chi^2$  ont été utilisées au cours des analyses menées pendant la thèse.

Nous avons principalement utilisé le  $\chi^2$  d'homogénéité qui permet de vérifier si deux populations de même taille sont dérivées d'une même loi de probabilité. Pour ce test l'hypothèse  $H_0$  est que « la distribution d'une variable catégorielle est similaire pour l'ensemble des populations ». La statistique de test associée est précisée en Équation 1.6.

$$\sum_{i=1}^J \frac{(O_i - E_i)^2}{E_i} \quad (1.6)$$

$O_i$  valeurs observées pour lesquelles l'échantillon prend la valeur  $i$

$E_i$  valeurs attendues sous l'hypothèse nulle

Le  $\chi^2$  d'homogénéité a notamment été utilisé pour comparer l'environnement en ET des régions géniques et intergéniques en Section 7.3.2

Nous avons aussi utilisé le  $\chi^2$  d'indépendance, un test permettant de vérifier l'indépendance statistique entre deux variables. L'hypothèse nulle est définie comme « les variables

sont indépendantes, il n’y a pas de relation statistique entre les variables catégorielles ». La statistique est décrite en Équation 1.7.

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1.7)$$

$O_{ij}$  valeurs observées pour lesquelles l’échantillon 1 prend la valeur  $i$  et l’échantillon prends la valeur 2

$E_{ij}$  valeurs attendues sous l’hypothèse nulle définie telle que :  $E_{ij} = \frac{O_i \times O_j}{N}$

Le  $\chi^2$  d’homogénéité a notamment été utilisé pour comparer l’environnement en ET des régions géniques et intergéniques et en particulier si certaines familles étaient surreprésentées dans une des deux populations en Section 7.3.2.

Le test exact de Fisher (Fisher, 1936), est un test apparenté au test de  $\chi^2$ , adapté à des échantillons de petite taille, mais reste valide, quel que soit la taille des effectifs. Ainsi, il est moins restrictif que les tests de  $\chi^2$ . Ce test se base sur des tables de contingence de type de 2 X 2. Sous l’hypothèse nulle, la statistique de test suit une loi hypergéométrique. Un exemple est présenté en Table 1.1. La Table de contingence permet un calcul exact de la statistique de test. L’hypothèse de test  $H_0$  est définie comme « les proportions relatives d’une variable sont indépendantes des proportions de la seconde variable ». La statistique de test est présentée en Équation 1.8.

**Table 1.1** – Exemple de table contingence 2 X 2 pour les tests exacts de Fisher

	Proportion 1	Proportion 2	Totaux
Variable 1	a	b	a + b
Variable 2	c	d	c + d
Totaux	a + c	b + d	n

$$p = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{a!b!c!d!n!} \quad (1.8)$$

$a$  valeur associée à la cellule (1,1)

$b$  valeur associée à la cellule (1,2)

$c$  valeur associée à la cellule (2,1)

$d$  valeur associée à la cellule (2,2)

$n$  taille de la population

Le test exact de Fisher a été utilisé pour tester si un groupe présentait une proportion significativement plus importante par rapport à un autre groupe, par exemple en Section 8.2.2.

### Test binomial

Ce test d'hypothèse permet de tester la probabilité de succès d'un événement en se basant sur la distribution binomiale. Cette distribution décrit la probabilité d'observer un événement particulier pour une variable quantitative. Pour être applicable, la variable testée doit suivre les critères de Bernoulli à savoir : i) le nombre d'observations est fixé, ii) les observations sont indépendantes, iii) chaque observation doit être binaire c'est-à-dire avoir l'issue échec ou succès et iv) la probabilité d'un succès est la même pour chaque essai. La statistique de test associée est présentée en Équation 1.9. Pour ce test l'hypothèse nulle est définie comme « la probabilité de succès de l'évènement considéré est similaire à la probabilité théorique ».

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1.9)$$

$\binom{n}{k}$  coefficient binomial

$n$  le nombre d'expériences réalisées

$p$  la probabilité de succès

$k$  le nombre de succès dans une répétition de  $n$  expériences

Ce test a par exemple été utilisé pour tester si la répartition des QTLs était équiprobable entre deux fragments chromosomiques synténiques en Section 4.2.

### Test z de proportion

Si la taille de l'échantillon est importante, la distribution binomiale peut être approximée par une distribution normale. Si la loi normale est approximée, un test Z peut être utilisé. Cette catégorie de test se base sur une statistique de test qui suit une loi normale et permet de tester l'hypothèse nulle d'une égalité des proportions comparées. Afin de tester des différences de proportions, la statistique décrite en Équation 1.10 est utilisée.



$$Z = \sqrt{n} \frac{\bar{X}_n - p_0}{\sqrt{p_0(1 - p_0)}} \quad (1.10)$$

$\pi$  coefficient binomial

$n$  le nombre d'expériences réalisées

$k$  le nombre de succès dans une répétition de  $n$  expériences

Ce test a été utilisé dans un cadre similaire au test binomial avec des populations de taille importante.

### 1.2.2 Tests de comparaison de deux variables

Pour certaines analyses, nous avons cherché à tester s'il existait un lien entre des variables. Afin de tester le lien par paire de variables, nous nous sommes appuyés sur le calcul de corrélations. Étant donné les questions posées, le lien statistique, et les données quantitatives, nous nous sommes appuyés sur des corrélations linéaires. Les corrélations offrent d'établir finement le sens et la force d'un lien entre deux variables. Différentes méthodes permettant la mesure de la corrélation existent. Dans le cadre de la thèse nous avons utilisé la corrélation de Spearman (Spearman, 1904) notée  $\rho$  dont le détail du calcul est présenté dans l'Équation 1.11. Cette méthode est l'équivalent non paramétrique de la corrélation de Pearson. Cette méthode est adaptée pour l'étude de variables dont la relation ne semble pas affine, mais plutôt monotone. Le coefficient de corrélation s'appuie sur les rangs des valeurs plutôt que sur leurs valeurs. Cette approche est similaire au  $\tau$  de Kendal (Kruskal, 1958) qui s'intéresse à la corrélation de rang entre des variables en s'intéressant aux paires discordantes.

$$r_s = \frac{\text{cov}(\text{rg}_X, \text{rg}_Y)}{\sigma_{\text{rg}_X} \sigma_{\text{rg}_Y}} \quad (1.11)$$

$\text{cov}(\text{rg}_X, \text{rg}_Y)$  est la covariance de variables de rang,

$\sigma_X$  et  $\sigma_Y$  définissent les écarts types des variables de rang

Nous avons aussi utilisé la corrélation de Pearson (Pearson, 1895) présenté dans l'Équation 1.12. Cette méthode va permettre un calcul de corrélation particulièrement adaptée aux relations affines. Ainsi, dans le cas d'une relation affine, les deux méthodes auront un comportement similaire. Dans le cas d'une relation monotone, le coefficient de Spearman présentera des valeurs plus hautes que le coefficient de Pearson.

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (1.12)$$

$\text{Cov}(X, Y)$  la covariance des variables X et Y

$\sigma_X$  et  $\sigma_Y$  définissent les écarts types des variables X et Y

Ces tests ont été utilisés pour tester la relation entre les différents mécanismes étudiés au cours de la thèse en Section 9.1.

### 1.2.3 Tests d'adéquation à une loi

Nous avons eu besoin de tester l'adéquation de distribution à des lois particulières, notamment pour tester la normalité de celle-ci, mais aussi de tester si les lois sous-jacentes aux distributions des valeurs observées étaient similaires.

#### Test de Kolmogorov–Smirnov

Afin de comparer les lois sous-jacentes aux distributions, nous avons utilisé le test de Kolmogorov–Smirnov (Smirnov, 1948). Ce test permet de comparer une distribution à une loi connue ou deux distributions entre elles. La statistique associée au pour chacun des échantillons est présentée en Équation 1.13 et est comparée via l'Équation 1.14. Pour ce test l'hypothèse nulle est définie telle que : « les distributions sont issues de la même loi de répartition ».

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i(\omega) \leq x} \quad (1.13)$$

$\omega$  la fonction de répartition empirique

$X_i$  variable ordonnée X de rang i

$n$  taille de la population X

Ce test a été principalement utilisé pour tester si une distribution observée présentait une loi connue.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \quad (1.14)$$

$F_{1,n}$  les fonctions de distribution empirique du premier échantillon

$F_{2,m}$  les fonctions de distribution empirique du second échantillon

Ce test a été utilisé pour tester si des distributions étaient similaires, par exemple en Section 5.2.

### Test de Shapiro Wilk

Le test de Shapiro Wilk (S. S. Shapiro & Wilk, 1965) a été utilisé au cours de la thèse pour tester la normalité d'une population. l'hypothèse nulle associée est « l'échantillon considéré est issu d'une distribution normale ». La statistique de test associée est présentée en Équation 1.15.

$$W = \frac{\left( \sum_{i=1}^n a_i x_{(i)} \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.15)$$

$x_{(i)}$  la statistique d'ordre

$\bar{x}$  la moyenne de l'échantillon

$a_i$  définit tel que  $(a_1, \dots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}$

$m$  les espérances des statistiques d'ordre d'un échantillon de variables suivant une loi normale

$V$  est la matrice de variance-covariance de ces statistiques d'ordre.

Ce test a été principalement utilisé pour tester les conditions d'utilisation des tests nécessitant de vérifier la normalité des données.

### 1.2.4 Tests utilisés dans les méta-analyses

Pour un certain nombre d'analyses, nous avons eu besoin d'agrèger des résultats provenant de différentes expériences. Pour effectuer ces méta-analyses, nous nous sommes appuyés sur des stratégies statistiques d'agrégation de p-values. Ces méthodes permettent de combiner les *p-values* de tests indépendants qui portent sur la même hypothèse. Les résultats des tests combinés doivent être issus de tests produisant des statistiques de tests issues de distributions continues. Dans l'ensemble, ces tests se basent sur l'hypothèse que sous  $H_0$  les p-values sont échantillonnés indépendamment et uniformément dans l'intervalle  $[0, 1]$ . La statistique de test et la p-value associé sont mesurés sur la base de la distribution de cette statistique de test sous l'hypothèse nulle. La statistique de test peut être calculée suivant différentes méthodes. Ces diverses méthodes vont mettre en évidence

différents aspects d'une combinaison de p-value (Heard & Rubin-Delanchy, 2018). Au cours des différentes analyses, nous avons utilisé deux types de méthode de calcul. La méthode Fisher (Fisher, 1970) est définie tel que :

$$\chi_{2k}^2 \sim -2 \sum_{i=1}^k \log(p_i) \quad (1.16)$$

$p_i$  la p-value associée au  $i^{eme}$  test d'hypothèse

$k$  le nombre de tests d'hypothèses, c'est-à-dire le degré de liberté de la distribution de  $\chi^2$

La méthode de Fisher présente le problème de surestimer les résultats lorsque les p-values agrégées proviennent de tests produisant des statistiques de tests discrètes comme des tests de rangs ou basés sur des tables de contingence (Kincaid, 1962). Néanmoins, ce problème s'efface pour des tailles d'échantillons importantes qui permettent aux distributions discrètes d'être approximée par une distribution continue (Mosteller & Fisher, 1948). Cette méthode ne permet pas d'appliquer facilement un poids aux expériences pour leur donner une importance différente contrairement à la méthode de Stouffer (Stouffer et al., 1949). Néanmoins, ce besoin ne s'est pas présenté au cours des analyses menées pendant la thèse. Cette méthode d'agrégation a été utilisée dans l'analyse des expériences de transcriptomiques présentées en Chapitre 6. La méthode de Mudholkar et George (Mudholkar E.O., 1983) se base sur la moyenne de statistiques de test de la méthode de Fisher et la méthode de Pearson. la statistique de test est définie tel que :

$$T_M = -c \sum_{i=1}^k \log\left(\frac{P_i}{1 - P_i}\right) \quad (1.17)$$

$p_i$  la p-value associée au  $i^{eme}$  test d'hypothèse

$k$  le nombre de tests d'hypothèses, c'est-à-dire le degré de liberté de la distribution de  $\chi^2$

Cette méthode a la particularité de produire des résultats extrêmes avec des p-values très proches de 1 ou 0. Cette méthode est aussi particulièrement conservatrice et va permettre de modérer des résultats extrêmes. Cette méthode d'agrégation a été utilisée dans l'analyse des expériences de séquençage au bisulfite présentées en Chapitre 8.

## 1.3 Matériel génétique

Au cours de cette thèse, nous avons étudié différents aspects du génome du pommier. Différents séquençages de cette espèce ont été effectués ces dernières années. À la date du début de la thèse (2019), le dernier séquençage de haute qualité était celui du génome d'un haploïde doublé de *Golden delicious* et avait été réalisé au laboratoire (Daccord et al., 2017). Cette version du génome est connue sous le nom de GDDH#13 et la dernière version est donnée en 1.1. Il résulte d'un séquençage combinant des séquençages de *reads* courts de technologie Illumina et de *reads* longs obtenus par la technologie PacBio. Le *scaffolding* s'est appuyé sur des cartes optiques générées avec des outils BioNano et des cartes génétiques. De plus les séquences obtenues ont été annotées en se reposant sur les données RNA-Seq (incluant les petits ARN) et les ETs. Le génome de *M. domestica* mesure environ 650 Mb, exactement 624 851 326 bp rassemblée en 17 chromosomes et un chromosome chimérique (le chromosome 0) constitué des *scaffolds* qui n'ont pas pu être rattachées à un groupe de liaison. Le chromosome 0, contient 2218 gènes pour une longueur totale de 52 728 359 bp. Le chromosome 0 a été supprimé des analyses menées au cours de la thèse, car ses séquences ne sont pas associées à un chromosome en particulier. Les ETs représentent 372,2 Mb du génome. Le génome complet présente 52 741 séquences annotées comme gènes, dont 45 116 ont été identifiées comme présentant un *Coding DNA Sequence* (CDS). Le génome filtré du chromosome 0 représente 50 522 séquences annotées comme gènes, dont 43 488 annotées comme codant pour une protéine. Ce séquençage a permis de confirmer la présence d'une WGD par les auteurs du premier génome de pommier disponible (Velasco et al., 2010). De plus, une explosion majeure de différents ETs a été identifiée (Daccord et al., 2017). Elle a été datée par estimation du taux de mutation (Lander et al., 2001) à 21 millions d'années (Daccord et al., 2017).

# PRÉPARATION DES DONNÉES

---

Afin d'étudier le devenir des gènes dupliqués par WGD chez le pommier, il convient d'identifier avec fiabilité, les couples de gènes dupliqués par WGD. De même, certaines analyses et en particulier l'analyse portant sur la pression de sélection nécessite de s'appuyer sur les séquences d'une autre espèce, proche, et n'ayant pas subi la dernière WGD.

## 2.1 Identification des blocs de synténie

### 2.1.1 Introduction

Les blocs de synténie se définissent comme la préservation de l'ordre des gènes entre homologues le long des segments chromosomiques (Drillon et al., 2014). Du fait de la duplication complète du génome, les génomes dupliqués par WGD présentent un ensemble de blocs de synténie. Afin d'étudier le devenir des gènes dupliqués par WGD chez le pommier, il convient d'identifier l'ensemble des couples de gènes issus de la duplication par WGD et rassemblés en blocs de synténie. En effet, la synténie permet de confirmer que les gènes ont été dupliqués par des mécanismes de WGD et non des mécanismes de duplication à plus petite échelle. Dans le contexte des gènes dupliqués par WGD, les gènes paralogues peuvent être appelés ohnologues, du nom de leur découvreur, Ohno dans les années 70 (Ohno, 1970). L'ensemble de ces notions a été développées en introduction et en particulier dans l'article de revue. De nombreux outils ont tenté des approches différentes pour résoudre les problèmes inhérents à la détection de blocs de synténie. Dans l'introduction, les différents algorithmes et implémentations ont été détaillés et les avantages et inconvénients des outils associés ont été précisés dans une revue que nous avons publiée en 2020 (Lallemand, Leduc et al., 2020).

## 2.1.2 Matériel et méthode

L'identification des blocs de synténie est un défi algorithmique important. Différents outils visent à résoudre ce problème en s'appuyant sur différents algorithmes, qui vont permettre de répondre à des besoins particuliers comme la recherche de synténies au sein d'un seul génome ou pour un ensemble de génomes. Afin d'étudier le devenir des gènes dupliqués au sein du pommier, nous avons cherché les traces de la dernière WGD commune aux *Maloideae*. L'identification est réalisée à partir du génome de *M. domestica* et en particulier sur le génome du Golden Delicious Double Haploïde #13 (GDDH13) en version 1.1 (Daccord et al., 2017). Le chromosome 0 ayant été supprimé des analyses, l'identification des blocs de synténie est donc faite sur un jeu de données d'un total de 49 921 gènes codants. Ces blocs de synténie étant la base de toutes les analyses qui suivront, la précision et la fiabilité de ceux-ci seront clés. Ainsi nous cherchons ici à identifier des blocs de synténie intraspécifiques issus d'une WGD attendue comme récente (Daccord et al., 2017 ; Velasco et al., 2010). Pour répondre à cette question, à la lumière des tests déjà réalisés (Lallemand, Aubourg et al., 2020), nous avons choisi d'identifier les blocs de synténie avec différents outils que sont SynMap 2, MCScanX et i-ADHoRe 3.0 et d'en comparer les résultats. Cette comparaison permet de s'assurer de la qualité des résultats et privilégier le meilleur.

Une première recherche des blocs de synténie a été effectuée avec SynMap 2 (Haug-Baltzell et al., 2017). SynMap 2 est un outil développé pour effectuer différentes analyses de génome, et permet entre autres l'identification des blocs de synténie. Une présentation complète de cet outil est disponible (Lallemand, Leduc et al., 2020). L'interface avec l'utilisateur est faite par une interface web intuitive adaptée à des utilisateurs non-informaticiens. La plupart des paramètres des outils ne sont pas consultables ou modifiables autorisant une prise en main rapide de l'outil, y compris pour un utilisateur non-expert au détriment d'un paramétrage fin des outils utilisés. Cet outil permet la construction d'une visualisation sous forme de *dot-plot*. Par ailleurs, cet outil autorise le calcul du  $K_a/K_s$ . De plus, il ne nécessite pas de ressources informatiques importantes puisque les calculs sont effectués sur des machines distantes hébergées par les auteurs de l'outil. Ce logiciel permet d'identifier des blocs synténiques en se basant sur la recherche de diagonale dense dans une matrice de gènes homologues. Pour se faire, le génome nucléaire est aligné contre lui-même (*blast-all-against-all*) afin d'identifier les zones d'homologie. Cet alignement peut être effectué par différents outils : (B)LastZ (Kieftbasa et al., 2011), MegaBlast (Morgulis et al., 2008), Megablast discontinu (Morgulis et al., 2008),

BlastN (Altschul et al., 1990), BlastP (States & Gish, 1994) et TblastX (Camacho et al., 2009). TblastX permet une recherche de séquences à partir d'une traduction dans les six cadres de lecture des nucléotides en acide aminé contre une banque de données de nucléotides traduits, et prend donc six fois plus de ressources de calculs par rapport à blastN. TblastX est décrit par les auteurs de SynMap comme peu performant pour la recherche de synténies (Haug-Baltzell et al., 2017). En effet, l'absence de résultats au niveau nucléotidique suggère une structure de génome divergente où sans doute la colinéarité des gènes n'est plus existante. Pour cette analyse, Latz a été utilisé, qui permet une recherche nucléotide-nucléotide plus rapide que blastN (Kiełbasa et al., 2011). Cette implémentation de LastZ permet une parallélisation du calcul et est présentée par les auteurs de SynMap comme le meilleur algorithme en termes de sensibilité et de vitesse (Haug-Baltzell et al., 2017). Pour suivre, différents filtres sont réalisés et notamment : les gènes en tandem sont regroupés en une séquence représentative et les séquences répétées sont supprimées. Ces filtres sont particulièrement importants, car l'identification des blocs de synténie est faite via DAGChainer (Haas et al., 2004). DAGChainer est un outil qui permet d'identifier les paires de gènes homologues qui partagent un ordre conservé en cherchant des chemins au sein d'un graphe orienté acyclique (Directed Acyclic Graph (DAG)). Un DAG rassemble des nœuds reliés par des arêtes orientées et ne comprend pas de cycles. La présence de gènes dupliqués en tandem ou de séquences répétées est donc très mal prise en compte par ce type d'algorithme et introduirait des biais importants. Ainsi, DAGChainer reçoit en entrée un fichier qui rassemble les résultats Blast et les e-value associées, et va construire le DAG. Dans ce DAG, les paires de gènes constituent les nœuds, les distances entre gènes sur le même chromosome constituent les arêtes. Les recherches dans ce DAG permettent de déterminer les blocs de synténie. DAGChainer a été paramétré en suivant les recommandations des auteurs de SynMap 2 (Haug-Baltzell et al., 2017) comme suit : *Relative Gene Order Maximum distance between two matches* (-D) : 20 gènes ; *Minimum number of aligned pairs* (-A) : 5 gènes. Ainsi, une synténie sera donc constituée d'au moins 5 gènes consécutifs et 2 gènes sont considérés comme consécutifs s'ils sont séparés de moins de 20 gènes intercalants.

Les blocs de synténie identifiés par DAGChainer sont alors fusionnés afin de produire des blocs les plus grands possibles. Cette fusion est faite par *Quota Align Merge*, qui s'appuie sur une couverture attendue (appelée quota) pour sélectionner un sous-ensemble de couples de gènes colinéaires avec un score total maximal. Étant donné notre connaissance de l'histoire évolutive du pommier qui suggère la présence d'une WGD (Daccord et al.,



2017) ainsi que les observations sur des *Maloideae* en particulier le poirier (Q. Li et al., 2019), la couverture attendue est 2 :1. L'optimisation des scores de ces blocs de manière globale permet un résultat de qualité. Cette fusion est faite par *Quota Align Merge* avec une distance maximale entre deux blocs (-Dm) de 500 gènes. Cette fusion par *quota* permet de prendre en compte les gènes paralogues provenant de duplications plus anciennes. Les génomes de plantes ayant souvent connu plusieurs WGDs au cours de leur histoire évolutive, cette étape est surtout importante pour une étude portant sur le pommier issu de la famille des Rosacées, une famille qui a connu de nombreuses WGDs. À partir de ces résultats, SynMap produit une visualisation par *dot-plot* et permet l'exportation des fichiers de résultats finaux pour des analyses. Différentes analyses des résultats de synténies ont été mises en place et notamment la construction d'une visualisation circulaire via Circos (Krzywinski et al., 2009). La construction de cette visualisation n'est pas complètement automatique et nécessite un traitement manuel, notamment la construction d'un fichier de configuration, un procédé chronophage.

Les blocs de synténie ont également été identifiés à l'aide d'un deuxième outil, MCS-canX (Y. Wang et al., 2012). Ce package comprend un ensemble d'outils permettant l'identification de blocs de synténie, leur visualisation et peut aussi exécuter différents calculs supplémentaires tels que le calcul de la pression de sélection estimée avec le  $K_a/K_s$  ou la construction de visualisations particulières. Ce programme est principalement écrit en C et en Perl. Il est à noter que la version mise à disposition par les auteurs ne peut pas être compilée telle quelle et nécessite une correction du code. Ceci peut être à l'origine de complications lors du déploiement de l'outil. Ce programme met en œuvre l'algorithme de MCSScan (Tang et al., 2008) dont différentes mises à jour ont permis une optimisation du résultat et des performances. Cet algorithme identifie les blocs colinéaires dans les génomes ou les sous-génomes, puis effectue des alignements multiples de blocs colinéaires en utilisant les gènes colinéaires comme ancres. Cet outil nécessite en entrée les identifiants des cinq meilleurs gènes homologues au niveau protéique de chacun des gènes du génome généré par *blastP-all-against-all* (Camacho et al., 2009) (seuil de e-valeur à  $10 \times 10^{-5}$ ). La taille de cette entrée peut être modifiée, cinq gènes étant la recommandation faite par les auteurs. Pour suivre, si des résultats de blastP consécutifs ont un gène commun et que les gènes appariés sont séparés par moins de cinq gènes, ces correspondances sont regroupées en utilisant la paire représentative avec la plus petite e-value. Ce procédé permet de ne pas biaiser l'algorithme à cause des gènes dupliqués en tandem qui augmenterait artificiellement le score du bloc. Ensuite, par programmation dynamique, le chemin de plus

haut score reliant les différents gènes synténiques entre deux chromosomes est recherché. Le calcul du score est fourni par l'Équation 2.1 (Y. Wang et al., 2012).

$$E = 2P_N^m \prod_{i=1}^{m-1} \left( \frac{l_{1i}}{L_1} \cdot \frac{l_{2i}}{L_2} \right) \quad (2.1)$$

Avec,

$P$  le nombre d'occurrences du  $i$ -ème gène apparié et de ses homologues dans la région chromosomique ohnologue

$N$  le nombre de paires de gènes entre les deux régions chromosomiques

$m$  est le nombre d'ancres entre les deux régions chromosomiques

$L_1$  et  $L_2$  sont les longueurs des deux régions chromosomiques

$l_{1i}$  et  $l_{2i}$  sont les distances entre deux ancres adjacentes dans les deux régions chromosomiques

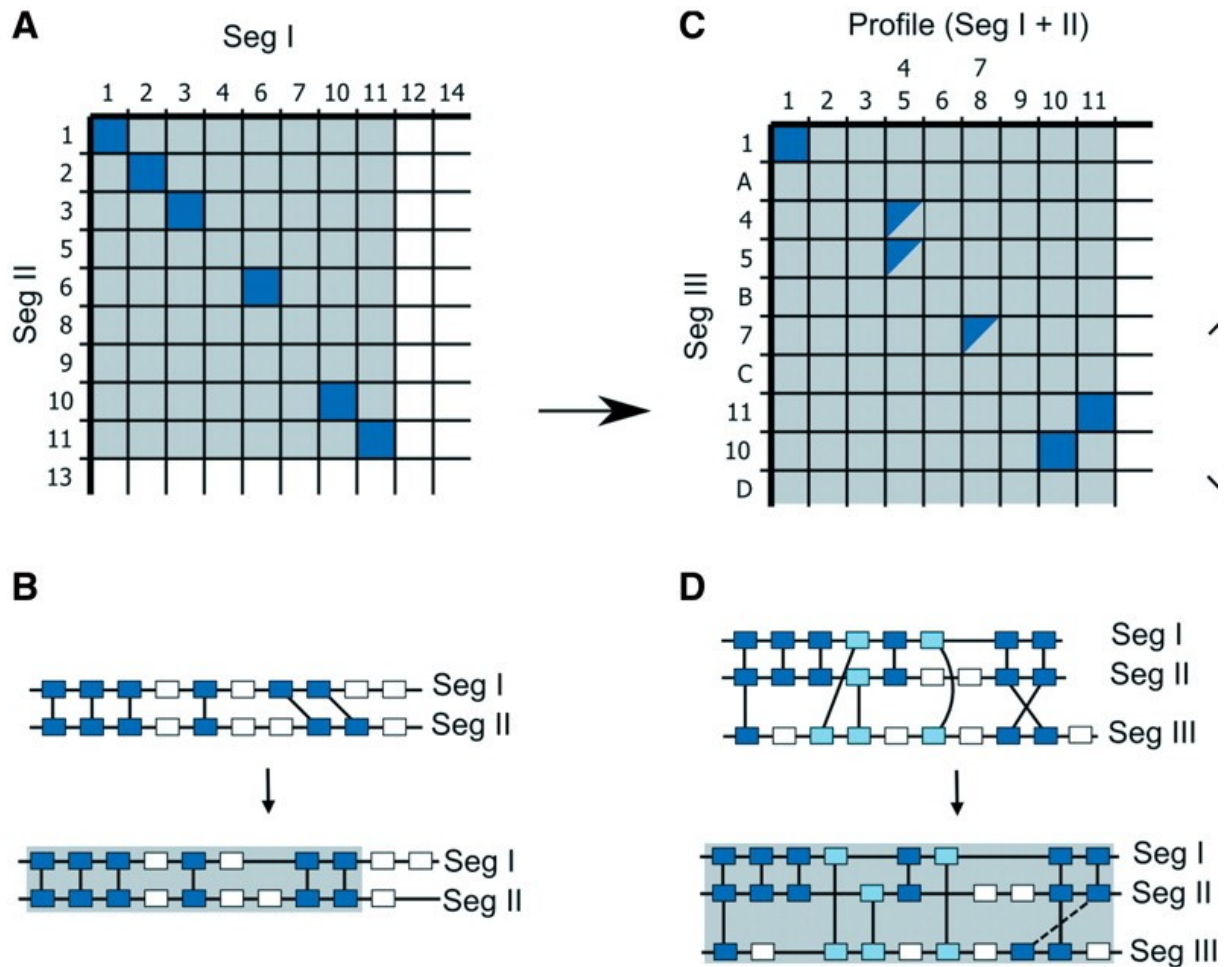
L'équation 2.1 permet de favoriser les gènes colinéaires fortement homologues. Les fragments chromosomiques ciblés par des *loci* ancestraux consécutifs sont peu à peu alignés par rapport aux chromosomes de référence, où chaque génome testé est utilisé comme référence successivement. Cette opération permet de déterminer le nombre de duplications présentes pour une région et d'améliorer ainsi la classification des gènes. Les gènes sont alors classifiés selon leur nombre de copies et leur distribution génomique. Pour l'identification des blocs de synténie chez le pommier, les paramètres utilisés pour MCSCanX étaient ceux proposés par les auteurs après des tests sur différents jeux de données simulées (Y. Wang et al., 2012). Les paramètres utilisés sont les suivants : *match\_score* : 50 ; *match\_size* : 5 ; *gap\_penalty* : -1 ; *overlap\_window* : 5 ; *e\_value* :  $1 \times 10^{-5}$  et *max\_gaps* : 25. Ainsi, seront considérés comme synténiques des régions présentant cinq gènes colinéaires avec un score d'alignement d'au moins 50 pour chacune des paires de gènes, avec moins de 25 gènes non colinéaires et une *e-value* inférieure à  $1 \times 10^{-5}$ .

L'identification des gènes synténiques dans le génome du pommier a aussi été réalisée à l'aide d'un pipeline *ad hoc* basé sur i-ADHoRe 3.0 (Proost et al., 2012). De la même façon que pour MCSCanX, l'entrée d'i-ADHoRe nécessite d'identifier les gènes avec les plus grands scores de similarité pour chacun des gènes du pommier en utilisant un *blastP-all-against-all* (Camacho et al., 2009) avec un seuil de e-valeur à  $1 \times 10^{-5}$ . Comme pour MCSCanX nous avons pris les cinq meilleurs résultats. Ces informations d'homologies sont ensuite utilisées comme données d'entrée pour i-ADHoRe 3.0, permettant la construction

de blocs de synténie basés sur la recherche de diagonales denses dans une matrice creuse de gènes homologues. Le schéma de l'algorithme exact est présenté en Figure 2.1. À l'image des autres outils, les gènes dupliqués en tandem sont regroupés sur un seul représentant. Chaque paire de chromosomes est utilisée pour construire une Matrice d'Homologie de Gènes (GHM) dont une schématisation est présentée en Figure 2.1 A. Dans cette matrice creuse, les paires de gènes homologues sont représentées par des points. Ainsi, les régions colinéaires apparaissent comme des diagonales denses en points. Une évaluation statistique de la signification des diagonales est estimée en tenant compte de la densité globale de la matrice. Les régions colinéaires significatives trouvées lors de cette détection initiale sont converties en profil. Les deux régions colinéaires sont alignées, c'est-à-dire que les gènes homologues ont été placés dans la même colonne en ajoutant des espaces si nécessaire. La construction de l'alignement du profil est présentée en Figure 2.1 B. Cet alignement est effectué par l'algorithme progressif de Needleman-Wunsch (pNW), mais peut être configuré pour utiliser d'autres outils d'alignement protéiques basés sur des stratégies d'alignement elles-mêmes basées sur l'approche par graphe gourmand (Greedy Graph (GG)). En utilisant ce profil aligné, une nouvelle recherche est effectuée (Schéma en Figure 2.1 C). Les régions significatives sont ajoutées à un nouveau profil et la recherche de profil est répétée (Figure 2.1 D). La recherche se poursuit avec le profil suivant jusqu'à convergence de l'algorithme. Cet algorithme itératif permet de retrouver des profils avec une sensibilité importante.

Pour cette analyse, les paramètres utilisés pour i-ADHoRe 3.0 étaient les suivants : *cluster\_type*=colineaire ; *tandem\_gap*=15 ; *prob\_cutoff*= $10 \times 10^{-5}$  ; *gap\_size*= 30 ; *cluster\_gap*= 30 ; *q\_value*=0,75 ; *prob\_cutoff*=0,01 et *anchor\_points*=5. Ainsi, pour cet outil, seront considérés comme des gènes dupliqués en tandem les gènes homologues sur le même chromosome à moins de 15 gènes l'un de l'autre. Par ailleurs, seront considérés comme synténiques les ensembles d'au moins cinq gènes collinéaires espacés de moins de 30 gènes avec une probabilité de génération du bloc de synténie par chance inférieure à  $1 \times 10^{-5}$ .

À partir de l'ensemble des gènes homologues identifiés et d'i-ADHoRe 3.0, nous avons pu reconstruire les fragments de chromosomes synténiques. Ces blocs de synténie peuvent être visualisés à l'aide de diagrammes circulaires construits de la même façon que pour les résultats associés à MCSanX et SynMap. Pour cet outil l'analyse des résultats et la construction de la visualisation via Circos sont réalisées automatiquement par le pipeline y compris la construction du fichier de configuration. Le pipeline *syntenic-blocks-*



**Figure 2.1** – Schéma de l’algorithme de construction des blocs de synténie par i-ADHoRe 3.0. Adapté du schéma associé à la publication des auteurs (Proost et al., 2012). **(A)** Construction d’une matrice d’homologie de gènes avec deux segments colinéaires initiaux. **(B)** Alignement des gènes homologues partagés entre les régions colinéaires pour la construction d’un profil. Le profil construit contient les informations des deux segments. **(C)** Début du processus itératif, la matrice compare le profil créé avec un nouveau segment colinéaire. **(D)** Génération d’un nouveau profil contenant les informations des trois segments. L’itération continue jusqu’à ce que le dernier profil créé ne trouve pas de nouveau segment colinéaire.

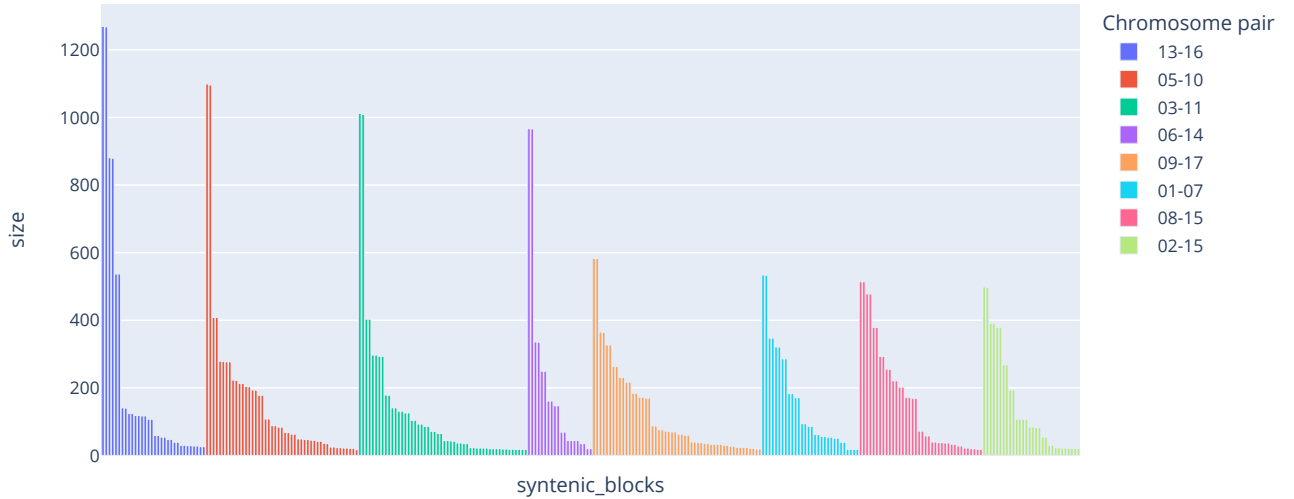
*snakemake* est accessible depuis l'Uniform Resource Locator (URL) suivante : <https://forgemia.inra.fr/tanguy.lallemand/syntenic-blocks-snakemake.git>. Il a été conçu pour être employé avec n'importe quels génomes pour détecter les blocs synténiques intraspécifiques et interspécifiques et peut être exploité dans d'autres contextes que l'étude actuelle. Il a été utilisé uniquement pour identifier les blocs synténiques intraspécifiques dans cette étude, mais a déjà été utilisé avec succès dans d'autres projets et en particulier sur le rosier. Ce pipeline vise de même à reconstruire les triplets, dont les détails de conception sont présentés dans la section suivante.

Les résultats de ces différents outils ont été agrégés et comparés à l'aide d'un ensemble de scripts Python et R afin de comparer les résultats.

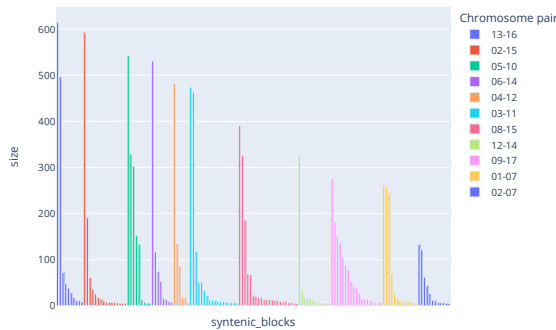
### 2.1.3 Résultats

SynMap a permis l'identification d'un total de 587 blocs de synténie, rassemblant 99,03 % du génome complet. Ces blocs sont composés de 11 266 couples de gènes ohnologues. De plus, 8451 gènes dupliqués en tandem ont été détectés. Chacun des groupes de gènes dupliqués en tandem a été associés à une séquence représentative. La taille moyenne des blocs de synténie construits est de 210 couples de gènes (écart type=334). Le plus petit bloc est composé de 15 couples, ce qui correspond à la taille minimale paramétrée. Le plus grand bloc de synténie rassemble 2538 couples de gènes ohnologues associés à la paire 13-16. La distribution de l'ensemble des tailles de blocs de synténie associées aux principales paires de fragments chromosomiques ohnologues est présentée en Figure 2.2A. On peut observer que pour l'ensemble des paires de chromosomes ohnologues, au moins deux blocs de synténie de taille importante sont identifiés. Pour le reste des blocs, leur taille décroît lentement jusqu'à atteindre la limite minimale de taille des blocs.

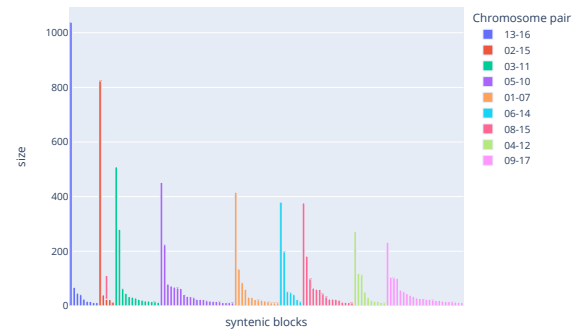
L'utilisation d'un pipeline basé sur MCSanX a permis l'identification de 16 093 couples de gènes ohnologues rassemblés en 567 blocs de synténie. Ces couples sont composés de 23 838 gènes ohnologues, ce qui correspond à 47 % des gènes codants chez le pommier. La taille moyenne des blocs de synténie est de 76 couples de gènes (écart type=133). La taille du plus petit bloc est 6, conformément à la configuration de l'outil. La taille du plus grand bloc est 615 couples de gènes associés à la paire 13-16. La distribution de la taille des blocs de synténie identifiée par MCSanX est présentée dans la figure 2.2B. On peut observer que pour chacune des paires de chromosomes ohnologues un bloc de synténie de taille importante est identifié. Pour le reste des blocs, la taille se réduit rapidement jusqu'à atteindre un palier autour de la valeur basse minimale autorisée. Nous avons pu



(A) Histogramme de la taille (en nombre de gènes) des blocs de synténie pour chaque paire de fragments chromosomiques synténiques identifiés avec SynMap. On peut observer la présence de deux grands blocs de synténie suivis par un ensemble de plus petits blocs dont la taille décroît jusqu'à la taille configurée.



(B) Histogramme de la taille (en nombre de gènes) des blocs de synténie pour chaque paire de fragments chromosomiques synténiques identifiés avec MCSanX. On peut observer la présence d'un grand bloc de synténie suivi par un ensemble de plus petits blocs dont la taille décroît jusqu'à la taille configurée.



(C) Histogramme de la taille (en nombre de gènes) des blocs de synténie pour chaque paire de fragments chromosomiques synténiques identifiés avec i-ADHoRe. On peut observer la présence d'un grand bloc de synténie suivi par un ensemble de plus petits blocs dont la taille décroît jusqu'à la taille configurée.

**Figure 2.2** – Distribution des tailles des blocs de synténie identifiés avec SynMap (2.2A), MCSanX (2.2B) et i-ADHoRe (2.2C). L'axe des ordonnées (*size*) représente le nombre de gènes par bloc de synténie. L'axe des abscisses (*syntenic\_blocks*) présente les différents blocs de synténie identifiés. Les couleurs sont associées aux différents couples de fragments chromosomiques synténiques.

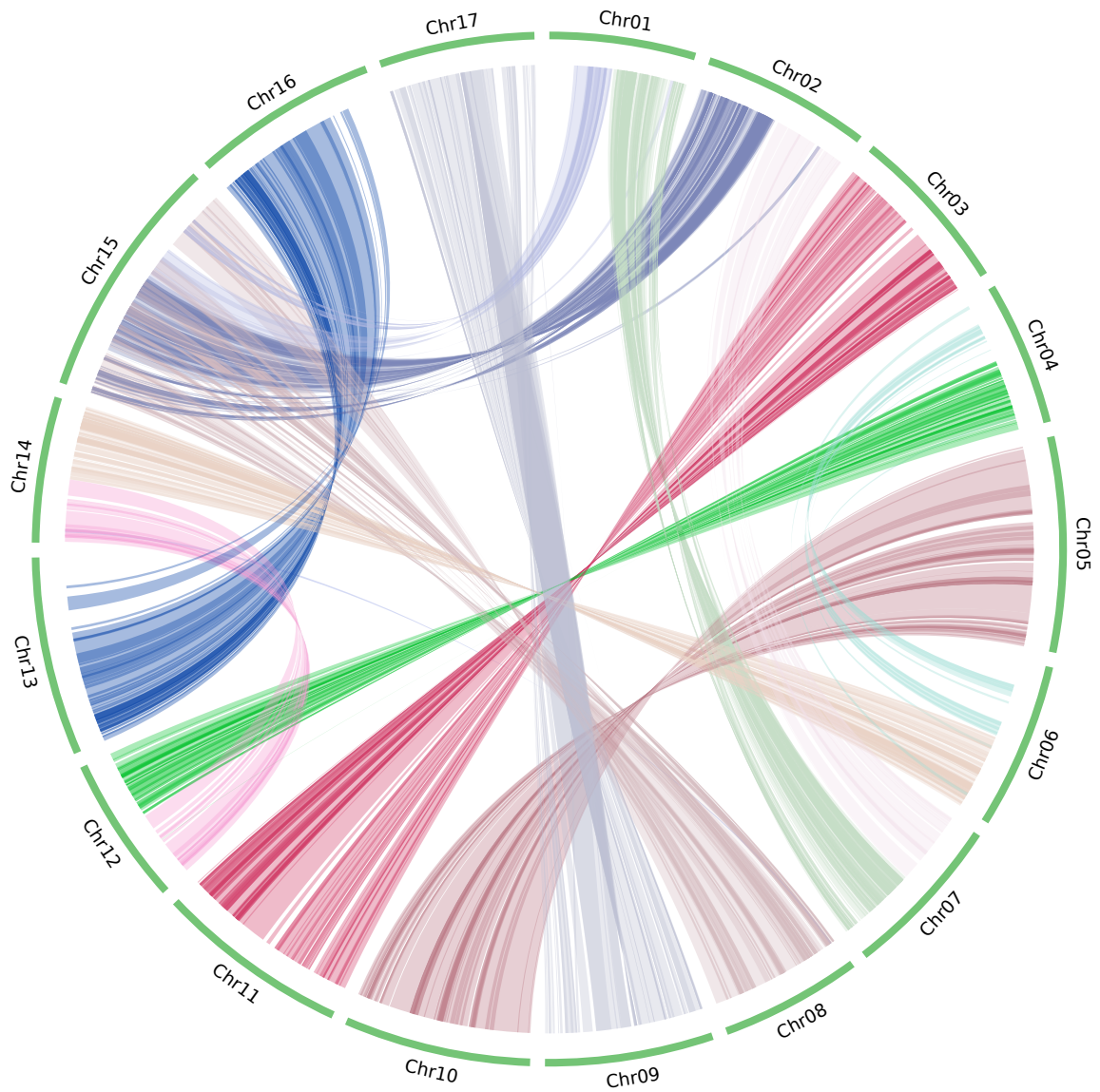
aussi identifier 4914 gènes dupliqués en tandem.

Le pipeline basé sur i-ADHoRe 3.0 a permis l'identification de 865 blocs synténiques. Nous avons identifié 16 779 paires de gènes, dont 10 958 paires de gènes ont été identifiées de façon unique ; 5821 paires de gènes (représentant 6327 gènes uniques) ont été identifiées dans plus d'un bloc de synténie. Ce résultat confirme l'existence de WGD plus ancienne. Nous avons identifié 14 788 gènes dupliqués en tandem (29 % des gènes). Globalement, les blocs de synténie représentent 91,38 % de la longueur totale du génome. La taille moyenne des blocs comprend 44 paires de gènes (écart type=112). Le plus petit bloc présente cinq paires de gènes et le plus grand bloc rassemble 1039 paires de gènes. La distribution de la taille des blocs de synténie identifiés par i-ADHoRe est présentée dans la figure 2.2C. De manière similaire à MCSScanX, les paires de chromosomes présentent chacun un bloc de grande taille puis un ensemble de plus petits blocs.

La visualisation circulaire de l'ensemble des blocs de synténie identifié par i-ADHoRe et reliant les fragments chromosomiques ohnologues entre eux est présentée dans la Figure 2.3. Cette visualisation ainsi que les résultats précédents ont confirmé la bonne conservation de la WGD, puisque très peu de régions ne sont pas synténiques. Nous avons identifié le haut du chromosome 1, le centre du chromosome 4 et une partie du chromosome 6 comme non synténiques. Nous pouvons observer des paires de fragments synténiques à l'échelle du chromosome entier, comme c'est le cas pour les paires 3-11, 5-10, 13-16 et 9-17. Nous avons également identifié des blocs à l'échelle du demi-chromosome tels que les paires 1-7 ou 6-14. Dans la plupart des cas, le point de rupture est localisé près des régions centromériques des chromosomes. Certaines paires synténiques présentent une inversion du sens, c'est notamment le cas de la paire 3-11, 5-10 et 9-17.

#### 2.1.4 Discussion

L'identification des blocs de synténie a été faite avec différentes méthodes qui présentent chacune des approches différentes détaillées dans la section matériel et méthode. Afin de sélectionner l'outil permettant de générer les blocs de synténie de la plus grande qualité qui seront à la base de toutes les analyses suivantes, nous avons cherché à comparer les différents résultats obtenus. Tout d'abord le nombre de gènes ohnologues est relativement semblable entre les différents outils, i-ADHoRe étant l'outil qui a identifié le plus de couples de gènes ohnologues, MCSScanX et SynMap ayant identifié un nombre légèrement inférieur et similaire de couples de gènes ohnologues. Cette différence est avant tout due à l'approche utilisée par i-ADHoRe qui permet la construction de profils de façon



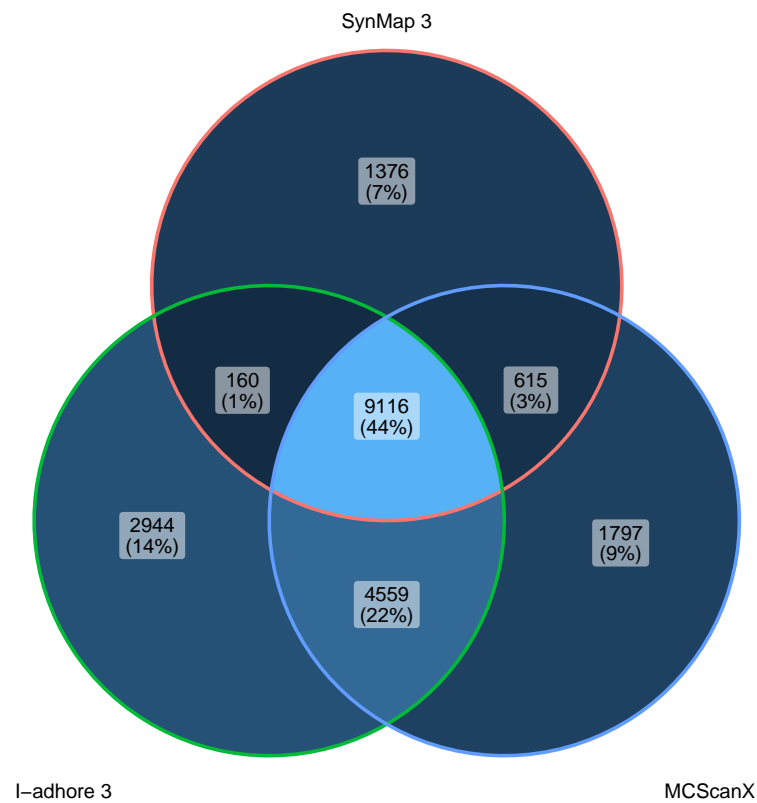
**Figure 2.3** – Visualisation circulaire représentant les régions synténiques au sein du génome de la pomme GDDH13 v1.1. Le cercle externe représente les chromosomes. Les liens entre les gènes représentent les synténies et la couleur associée au lien permet une meilleure visualisation. Des blocs de synténie de petite taille n'ont pas été colorés afin de privilégier une bonne lecture. Des blocs synténiques peuvent être observés à l'échelle du chromosome pour les paires 3-11 ou 5-10. Des blocs demi-chromosomiques peuvent être observés pour les paires 1-7 et 2-7.



incrémentale, permettant une très bonne sensibilité de l’outil et l’identification de couples de gènes ohnologues ayant plus divergé ou issus de WGD plus ancienne.

Ensuite, nous avons comparé la taille des blocs de synténie entre les différents outils et notamment à l’aide de la Figure 2.8 qui présente la taille des plus grandes paires de fragments chromosomiques synténiques. On peut remarquer que globalement les outils ont présenté un résultat semblable avec un grand bloc identifié pour chacun des couples puis une distribution de la taille des blocs qui décroît rapidement. SynMap présente un comportement différent avec plusieurs grands blocs et une décroissance moins rapide. Cette observation indique que pour les blocs de synténie les plus longs, SynMap permet l’obtention de blocs contenant plus de blocs, ce qui est sans doute dû à l’algorithme post-construction des blocs (*quota Align*) qui permet de fusionner les blocs de synténie. La décroissance la plus rapide est observée avec i-ADHoRe, qui présente en plus un grand nombre de petits blocs. Cette observation s’explique peut être par la construction de profils de façon incrémentale permettant une bonne sensibilité allant jusqu’à identifier des couples de gènes plus divergents dont une partie représente probablement des restes de WGD ancestrale. La rapide décroissance du nombre de gènes intégrés dans les blocs suivants est probablement due à l’approche par profil qui permet de rassembler les gènes issus du même évènement de duplication. Étant donné que le pommier a principalement subi une WGD, une grande partie des gènes identifiés comme synténiques sont associés au même bloc de synténie. La différence d’effectifs en termes de nombre de couples de gènes ohnologues identifiés semble être plutôt associée aux petites paires de fragments chromosomiques ohnologues.

Par ailleurs, nous avons voulu comprendre quelles sont les paires identifiées par les différents outils et quelles sont les proportions des couples de gènes en commun. La comparaison des proportions de l’ensemble des couples de gènes ohnologues identifiés par les différents outils est présentée en Figure 2.4. On constate que 44 % des couples de gènes identifiés ont été retrouvés quel que soit l’algorithme utilisé. Cette observation suggère donc des résultats de bonne qualité et robustes. Les proportions de couples de gènes identifiés par deux outils parmi les 3 sont faibles (1 et 3 %) à part pour le couple i-ADHoRe et MCSanX qui a identifié 22 % de couples de gènes que SynMap n’a pas identifié. Cette différence est sans doute due à l’approche de SynMap qui se base sur une information nucléique, ce qui rend la détection d’homologie plus complexe par rapport à une information protéique. En effet, MCSanX et i-ADHoRe identifient plus de gènes ohnologues que SynMap. Ce constat est probablement dû aux séquences plus divergentes qui pré-



**Figure 2.4** – Diagramme de Venn représentant le nombre de couples de gènes ohnologues déterminés par SynMap, i-ADHoRe et MCScanX. Le cercle entouré de vert en bas à gauche représente les couples de gènes ohnologues identifiés par i-ADHoRe. Le cercle au liseré de bleu en bas à droite représente les couples gènes ohnologues identifiés par MCScanX. Le cercle au liseré de rouge en haut représente les couples gènes ohnologues identifiés par SynMap. Les chiffres associés à chacun des cercles représentent le nombre et le pourcentage de paires de gènes uniquement identifiés par cet outil. Les intersections des jeux de données représentent le nombre de couples en commun pour les outils concernés. Ainsi, 44 % des couples de gènes ont été identifiés par l'ensemble des outils, ce qui suggère une bonne qualité du résultat. SynMap fait office d'outsider avec seulement 4 % des gènes en commun avec l'un des deux autres outils, tandis que MCScanX et i-ADHoRe présente 22 % supplémentaires de couples de gènes identifiés en communs.

sentent une homologie significative au niveau protéique lors d'une recherche avec blastP mais pas au niveau nucléaire en utilisant une recherche blastN. Cette différence entre les données nucléique et protéique peut s'expliquer en partie par la taille de l'alphabet et de la dégénérescence du code génétique.

Ainsi, les couples de gènes ohnologues identifiés ont au moins un exon qui est placé dans un bloc de synténie. Néanmoins, cette approche pose problème, car l'intégration d'exons sans tenir compte de l'ensemble du gène peut baisser la fiabilité des blocs de synténie et les paires de gènes ohnologues. De ce fait, les 7% de gènes identifiés seulement par SynMap sont en majeure partie de courts exons que les autres outils n'ont pas identifiés car sont mal annotés ou trop courts pour être significatifs au niveau protéique. De plus, l'ensemble des analyses menées dans cette thèse à partir de ces blocs de synténie sont faites à l'échelle des gènes. L'identification des relations d'homologies à l'échelle génique permet alors de conserver des analyses à la même échelle. De plus, l'intégration de seulement certains exons par SynMap pose une problématique supplémentaire de filtrage des données. Ce résultat n'est donc pas à privilégier bien que présentant l'intérêt de renforcer la fiabilité du résultat des autres outils en le confirmant au moins en partie. En outre, grâce au fait que SynMap travaille sur l'ensemble de la séquence nucléaire cela permet de s'affranchir des biais liés à l'annotation. Ce résultat permet aussi de confirmer le résultat obtenu en 2017 (Daccord et al., 2017) avec SynMap3. i-ADHoRe, est l'outil qui a identifié le plus de paires de gènes ohnologues, de même que le plus de couples de gènes ohnologues que lui seul a identifié. Ceci peut sans doute s'expliquer, au moins en partie, par des problèmes d'identification de blocs de synténie connus chez MCScanX, en particulier dans le cas de synténie intraspécifique (D. Liu et al., 2018). Néanmoins ces outils ont identifié 66% des gènes en commun et les différences entre les deux outils sont principalement faites sur les fragments synténiques de petite taille qui ne seront pas étudiés en profondeur dans les analyses suivantes.

Les bilans entre les différents outils montrent une similarité importante des conclusions, ce qui tend à renforcer la confiance dans le résultat qui semble se montrer répétable dans une certaine mesure entre les différents algorithmes et fichiers d'entrées utilisés. En conséquence, pour des problèmes liés à l'approche, la fiabilité, la présence d'une documentation et la facilité de déploiement (le code source de MCScanX est en C et la version fournie par les auteurs ne compile pas et nécessite une correction manuelle avant compilation) font que l'outil finalement retenu pour la construction des blocs de synténie est i-ADHoRe. Pour renforcer la fiabilité des résultats de cette méthode dont on sait la sen-

sibilité à la configuration, nous avons exécuté l'outil avec différents sets de paramètres et sélectionné le résultat qui nous a permis d'obtenir les blocs de synténie avec le plus gènes ohnologues. Les paramètres finaux ont été précisés dans le section matériel et méthode. Pour la poire, un autre *Maloideae* qui présente cette WGD, un nombre similaire de gènes synténiques (16 509 gènes) a été identifié à l'aide de MCScanX (Q. Li et al., 2019), ce qui suggère que le résultat obtenu ici est fiable.

Le pipeline permettant l'exécution d'i-ADHoRe 3.0 et la construction de la visualisation circulaire ont été pensées pour être généralisées à toutes les conditions possibles. Ainsi ce pipeline peut identifier les blocs de synténie intraspecifics et interspecifics pour d'autres organismes. Il a ainsi été utilisé dans d'autres cadres et a permis la construction et la visualisation des blocs de synténie du rosier et les synténies entre le pommier et le pêcher.

### 2.1.5 Conclusion

Par différentes approches, nous avons pu identifier les blocs de synténie existants au sein de génome du pommier. Les blocs de synténie identifiés à l'aide de SynMap3, i-ADHoRe et MCScanX ont été comparés. Les 16 779 couples de gènes ohnologues identifiés par i-ADHoRe ont été sélectionnés comme les plus pertinents en termes de qualité d'identification. La construction d'un diagramme circulaire des blocs de synténie confirme la présence d'une WGD récente dans l'histoire évolutive du pommier.

## 2.2 Construction de triplets pommier-pêcher

Certaines analyses nécessitent de s'appuyer sur des séquences issues d'un autre génome, les plus proches possibles du génome étudié et non dupliquées par la dernière WGD présente chez tous les *Maloidés*. Nous avons choisi d'utiliser le génome de *P. persica* (pêcher), un organisme de la famille de *Rosaceae*, pour sa proximité phylogénétique avec les *Maloidés*.

### 2.2.1 Matériel et méthode

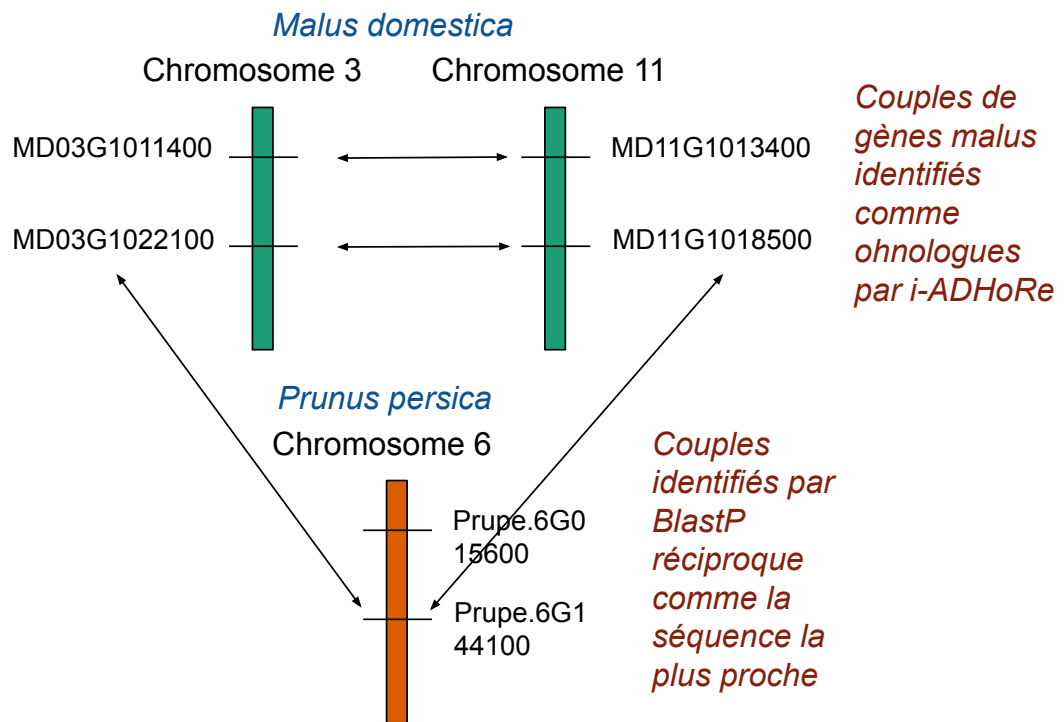
Ainsi, nous avons construit des triplets de gènes constitués de deux gènes ohnologues du pommier et le paralogue le plus proche chez *P. persica*. Les triplets sont construits à partir d'un *Reciprocal Best BlastP-all-against-all* en utilisant les séquences ohnologues de *Malus* identifiées par i-ADHoRe et l'ensemble des séquences protéiques de *Prunus*. Étant donné la dépendance aux paires de gènes ohnologues, cette recherche a été implémentée dans le pipeline permettant l'identification des blocs synténiques (*syntenic-blocks-snakemake*). est accessible depuis l'URL suivante : <https://forgemia.inra.fr/tanguy.lallemand/syntenic-blocks-snakemake.git>).

Les triplets sont conservés si les deux protéines ohnologues de *M. domestica* ont le même meilleur résultat avec une recherche BlastP chez *P. persica*, et réciproquement la séquence chez *Prunus* a comme meilleur résultat l'un des deux séquences *M. domestica*. Un schéma illustrant cette construction est présenté en Figure 2.5. Cette méthode de construction de triplets est très stringente. Cette démarche permet ainsi la construction de triplets fiables et garantit que l'on est bien en présence de séquences qui ont été dupliquées par la dernière WGD pour les gènes du pommier et le plus proche orthologue chez le pêcher.

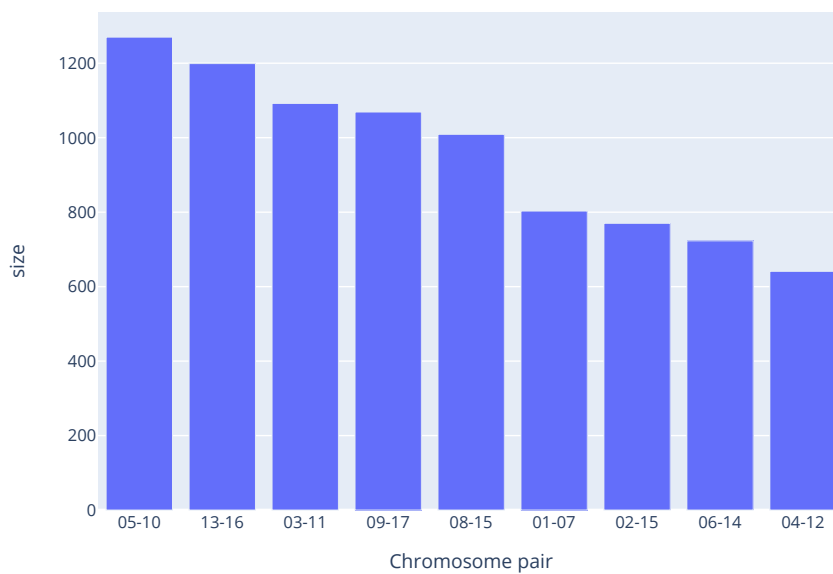
### 2.2.2 Résultats

Nous avons pu identifier 9821 les triplets avec deux gènes ohnologues *Malus* et leur orthologue chez *Prunus*. La numération des triplets associés à chacun des couples de fragments chromosomiques ohnologues est présentée en Figure 2.6. La distribution des tailles des couples *Malus* associés à ces triplets est similaire à celle obtenue pour les blocs de synténie avec des effectifs légèrement inférieurs pour chacun des couples.

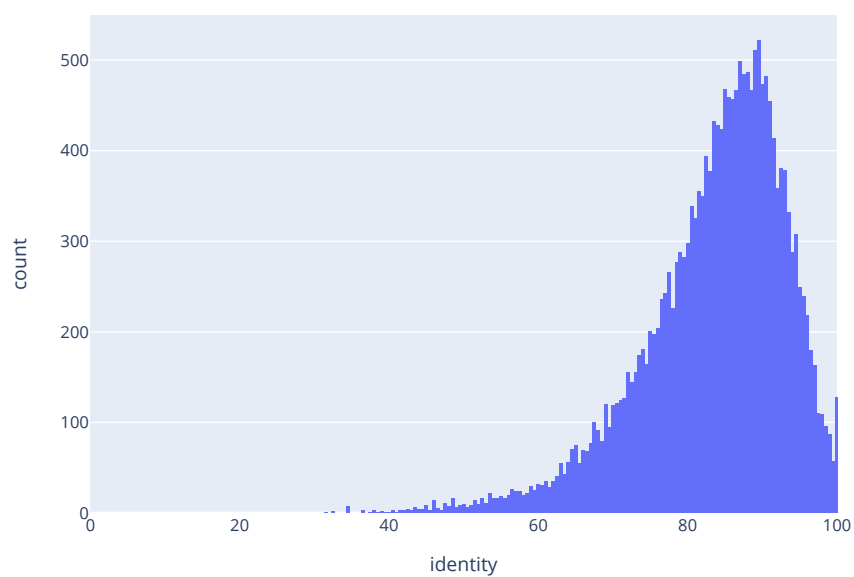
Les pourcentages d'identité entre les gènes du pommier et du pêcher sont représentés



**Figure 2.5** – Schéma de la construction de triplets constitués de deux gènes ohnologues chez *M. domestica* (représentée en vert) et de l'orthologue chez *P. persica* (représentée en orange). À titre d'exemple, pour le triplet MD11G1173000, MD03G1155400, Prupe.6G144100, nous avons ainsi une ohnologie Malus-Malus identifiée par i-ADHoRe (représenté par les doubles flèches entre les chromosomes de *M. domestica*) et une double orthologie entre MD11G1173000 et Prupe.6G144100 et MD03G1155400, Prupe.6G144100 identifiée par blastP réciproque et représenté par les doubles flèches entre les gènes du pommier et du pêcher.



**Figure 2.6** – Histogramme du nombre de triplets pommier-pêcher pour les couples de chromosomes ohnologues. L'axe des ordonnées (*size*) présente le nombre de triplets pour un couple de fragments chromosomiques synténiques. L'axe des abscisses (*Chromosome pair*) présente les différents couples de fragments chromosomiques synténiques. La distribution du nombre de triplets de gènes intégrés dans les couples de fragments chromosomiques ohnologues suit une distribution similaire à celle du nombre de gènes intégrés dans les blocs de synténie.

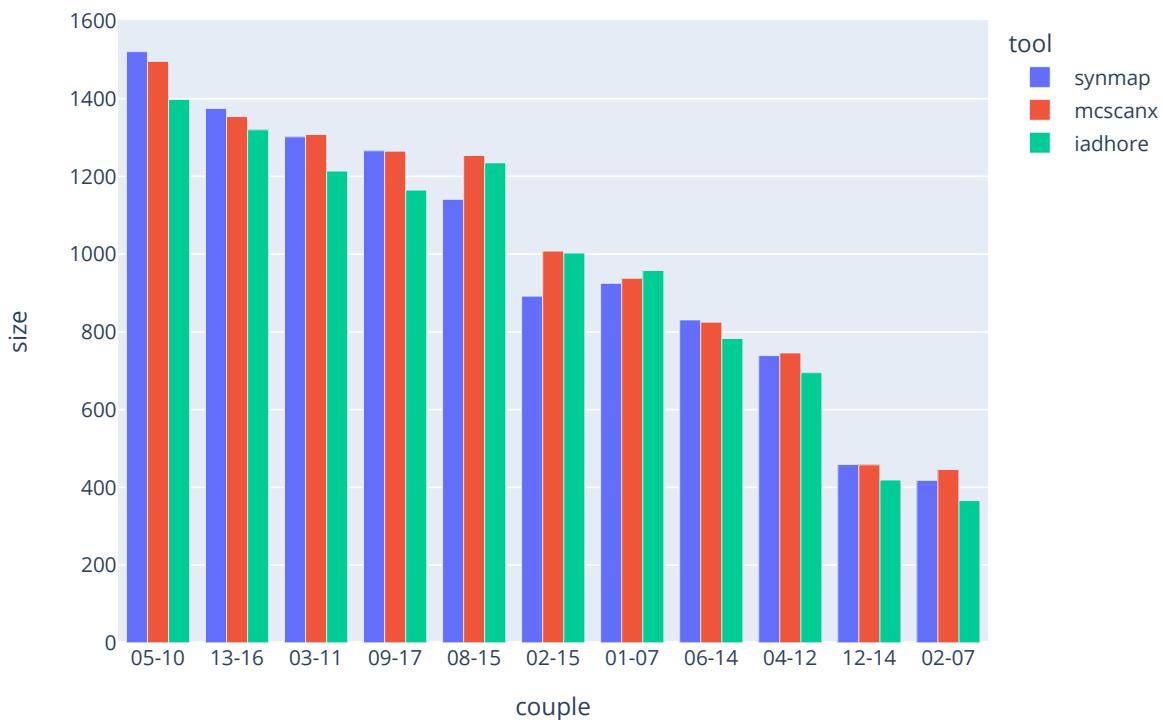


**Figure 2.7** – Histogramme du pourcentage d'identité des gènes du pommier et des gènes du pêcher intégrés dans les triplets. L'axe des ordonnées (*count*) présente le nombre de gènes. L'axe des abscisses (*identity*) présente le pourcentage d'identité, une barre correspond à un intervalle de 0,5. On peut observer que la majorité des triplets sont composés de séquences présentant des pourcentages d'identité entre 80 % et 100 %.



en Figure 2.7. La moyenne de l'identité des séquences conservées est de 83,56 % avec un écart type de 9,64 %. La moyenne de couverture est de 94,58 % (écart type=13,83 %). De plus, la taille moyenne des séquences de requête est de 433 bp (écart type=321), de cible est de 469 bp (écart type=332) et la moyenne de la taille des séquences alignées est de 416 bp (écart type=321).

### 2.2.3 Discussion



**Figure 2.8** – Histogramme de la taille des fragments chromosomiques synténiques. L'axe des ordonnées présente la taille des fragments synténiques en nombre de gènes. L'axe des abscisses présente les différents couples de fragments chromosomiques synténiques. Les barres bleues sont associées aux blocs de synténie construits par SynMap, les barres rouges sont associées à MCScanX et les barres vertes sont associées à i-ADHoRe.

Afin de pouvoir mener les analyses détaillées dans les chapitres suivants et notamment une analyse de la pression de sélection, il était nécessaire de pouvoir accéder à un

ensemble de gènes synténiques, mais aussi à des gènes orthologues chez une autre espèce. C'est pourquoi nous avons choisi de construire des triplets de gènes en s'appuyant sur un génome proche, celui du pêcher. Le choix du pêcher a été fait pour la qualité de l'assemblage de son génome, mais aussi, car cette espèce n'a pas eu de WGD depuis sa divergence d'avec l'ancêtre du pommier et du poirier. De plus, cet organisme est très proche phylogénétiquement et présente une très bonne synténie avec le pommier. Parmi les génomes disponibles chez les Rosacées de qualité suffisante, c'est le plus proche du pommier (exceptés les autres *Maloideae* qui ont aussi subi la WGD récente) notamment par rapport à d'autres espèces comme le fraisier.

Nous avons pu identifier 9821 les triplets composés d'un couple de gènes ohnologues du pommier et d'un gène orthologue chez le pêcher. Ainsi, 6958 paires de gènes identifiés comme ohnologues n'ont pas permis la construction de triplets. Ceci est dû à la perte de certains couples notamment à cause de la réciprocité lors de la recherche par *Reciprocal Best Blast Hit* (RBBH) qui n'est pas respecté. Cette perte de 41 % des paires de gènes lors de la construction des triplets montre la stringence de notre approche qui à nous permis de garder que des gènes proches.

En sachant que le pêcher est proche phylogénétiquement du pommier, mais n'a pas subi la dernière WGD, nous pouvions attendre un nombre de triplets de cet ordre. De plus, une analyse similaire a été faite chez le poirier en s'appuyant sur MCSCanX (Q. Li et al., 2019). Cette étude rapporte l'identification de 824 blocs synténiques entre la poire et la pêche, rassemblant 11 108 gènes de pêche et 16 509 gènes de poire, ce qui représente 61 % du génome du pêcher. Lors de la recherche de synténie 2 :1 entre la poire et la pêche, il a été retrouvé 6203 triplets. Ainsi nous avons identifiés près de 37 % de triplets supplémentaires en relation synténique 2 :1 ce qui montre la pertinence de notre approche par rapport à une recherche simple de synténie.

L'identification des triplets est reproductible, puisqu'en changeant les paramètres de blastP et notamment, la *e-value* ou le nombre de résultats retenus par blast, chacune des exécutions permet l'identification des mêmes triplets, ce qui suggère une bonne robustesse de l'analyse. Cette reproductibilité est permise entre autre par la stringence de la construction par RBBH.

La construction des triplets a été faite pour être très stringente. D'autres approches ont été menées et notamment un *Best Blast Hit* (BBH), ou la construction de famille de 3 gènes via Orthofinder. En comparant les résultats de ces différentes approches, la méthode de construction des triplets par RBBH a permis d'obtenir les meilleurs résultats.

En effet, cette approche a permis l'identification d'un maximum de triplet tout en éliminant les possibilités d'ambiguïtés et de filtrage post-construction. Par ailleurs, le RBBH garantissent, selon l'hypothèse la plus parcimonieuse, que les triplets construits correspondent à des gènes qui ont été dupliqués lors de la dernière WGD chez le pommier.

#### **2.2.4 Conclusion**

Ainsi en nous appuyant sur les 16 779 paires de gènes synténiques construits avec i-ADHoRe 3.0 et par une approche de RBBH nous avons pu construire un ensemble de triplets de gènes. Ces triplets rassemblent deux gènes ohnologues chez le pommier et de leur orthologue chez le pêcher, un organisme proche du pommier, mais dont l'ancêtre commun n'a pas subi la dernière WGD commune aux *Maloideae*. Nous avons pu identifier 9821 triplets qui pourront être utilisés dans les analyses nécessitant des séquences de références.

# ÉTUDE DU FRACTIONNEMENT GÉNOMIQUE CHEZ LE POMMIER

---

## 3.1 Introduction

La WGD est connue pour affecter profondément la structure du génome en doublant, au moins de façon transitoire, le nombre de gènes et la taille du génome. À la suite d'une WGD, différents mécanismes vont restaurer graduellement l'état de ploïdie initiale via un ensemble de processus regroupés sous le terme de diploïdisation. Ainsi la perte progressive d'une partie des gènes dupliqués par pseudogénéisation va être un mécanisme considérable (Lockton & Gaut, 2005) pour un retour à la ploïdie. Ce processus dit de fractionnement (Langham et al., 2004) finit par réduire le nombre de gènes dans le génome à un niveau proche de l'état diploïde d'origine avant la duplication et ainsi éviter le mécanisme de "l'obésité génétique" (Bennetzen & Kellogg, 1997). Les plantes présentant un nombre important de WGDs au cours de leur histoire évolutive, le processus de fractionnement est surtout important pour ces organismes (Bennetzen & Kellogg, 1997).

Chez certaines espèces le fractionnement peut être biaisé, et les deux copies d'une paire de gènes dupliqués n'ont pas toujours la même probabilité d'être perdues, ce mécanisme est nommé biais de fractionnement (Freeling & Thomas, 2006; Freeling et al., 2012; Woodhouse et al., 2010).

Ainsi, une tendance à la suppression préférentielle des gènes de certains sous-génomes a été constatée au sein de plusieurs espèces. Cela a été observé, par exemple, chez *A. thaliana* (Thomas et al., 2006). Chez le maïs, dont la WGD remonte à 10 Ma, après comparaison des taux de rétention des gènes entre le sorgho et les sous-génomes appelés *maize1* et *maize2* pour une fenêtre de 100 gènes, il a été observé une différence significative des taux de rétention entre les deux sous-génomes (J. C. Schnable et al., 2011; M. Zhao et al., 2017). De même, chez *B. rapa* où une triplication par WGD a été datée à 20 Ma (X. Wang et al., 2011), une comparaison avec *A. thaliana* comme référence a

permis de mettre en évidence que l'un des trois sous-génomes conserve systématiquement une fraction plus importante des gènes. Un constat similaire a été fait chez le coton dont la WGD a été datée à 60 Ma (Renny-Byfield et al., 2015). Néanmoins ce biais de fractionnement n'est pas toujours retrouvé chez les espèces dupliquées par WGD, qu'elles soient autopolyploïdes ou allopolyploïdes. Ainsi, il n'a pas été observé de déséquilibre dans les taux de rétention des gènes ohnologues chez le riz (L. Li et al., 2006), le peuplier (Y. Liu et al., 2017), le soja (M. Zhao et al., 2017) ou le bananier (Garsmeur et al., 2014). Chez les Cucurbitacés, des allopolyploïdes depuis 1 Ma, il n'a pas été identifié non plus de déséquilibre de fractionnement notamment ni chez le potiron (H. Sun et al., 2017), ni pour la courge (H. Sun et al., 2017). Pour finir, au sein des *Maloideae* et en particulier le poirier, pas de déséquilibre de pertes de gènes entre les chromosomes ohnologues en comparant avec le génome de *P. persica* comme référence en utilisant une fenêtre glissante de 100 gènes (Q. Li et al., 2019).

Il a été démontré que ce biais dans le fractionnement génomique lors de la rediploïdisation avait un impact important sur l'organisme (Alger & Edger, 2020; J. C. Schnable et al., 2011; M. Zhao et al., 2017). De multiples implications liées à ce mécanisme ont été relevées dans différentes analyses, et le biais de fractionnement génomique semble être un marqueur important associé aux déséquilibres chez les espèces dupliquées. Par exemple, il a été observé un lien fort entre le biais de fractionnement et les déséquilibres transcriptionnels (J. C. Schnable et al., 2011). Ces mécanismes seront discutés dans le chapitre discussion. Ce mécanisme n'est pas systématiquement retrouvé chez les organismes en cours de rediploïdisation. Ce type d'analyse n'ayant pas été conduit sur *M. domestica*, une étude du fractionnement du génome a donc été menée.

## 3.2 Matériel et méthode

Peu d'outils sont mis à disposition pour étudier ce type de phénomène. Cette analyse a été réalisée à l'aide de SynMap (Haug-Baltzell et al., 2017) et, en particulier, de l'outil FractBias (Joyce et al., 2017), un module écrit en Python et accessible depuis l'interface web de SynMap qui permet de calculer le biais de fractionnement des gènes ohnologues et d'estimer le fractionnement par rapport à un génome apparenté. Ici, le génome du pêcher *P. persica* a été choisi. En utilisant une fenêtre glissante de taille donnée le long des chromosomes du pêcher (génome cible), FractBias permet de calculer le biais de fractionnement de gènes homologues sur tous les chromosomes du pommier (génome de

requête). La formule associée au biais de fractionnement est fournie en Équation 3.1 à partir des éléments fournis par les auteurs (Joyce et al., 2017). Cette équation permet le calcul de la proportion de gènes conservés pour une fenêtre glissante le long du chromosome  $T_{a,j}$  et sur le chromosome  $Q_{b,i}$ .

$$f(T_a, Q_b, x_j = \frac{\sum_{i=1}^w S}{w}) \quad (3.1)$$

$$S = t_i \in T_a x_j \cap Q_b \quad (3.2)$$

Avec,

$T$  chromosomes du génome cible T (*Target*)

$Q$  chromosomes du génome de requête Q (*Query*)

$a$  numéro de gène du génome cible

$b$  numéro de gène du génome requête

$w$  taille de la fenêtre glissante

$j$  représente la série de fenêtres =  $1..|Ta| + \omega$

$S$  est évaluée à 1 ou 0 à l'aide de l'Équation 3.2

$|Ta|$  = nombre de gènes sur le a-ième chromosome de T

Pour cette analyse, FractBias a été configuré comme suit : *Syntenic Depth Algorithm* : *Quota Align* ; *Ratio of Coverage Depth* : 1 :2 ; *Sliding Window Size* : 100 , et *Fractionation biais calculation* : *All genes*. Pour les deux dernières variables, différents groupes de paramètres ont été testés : et notamment des fenêtres à 50 ou à 100 et seulement sur les gènes synténiques ou sur l'ensemble des gènes. Ce dernier paramètre va permettre de mener la détermination du taux de rétention des gènes en ne s'appuyant que sur les paires de gènes existants dans les deux génomes, c'est-à-dire dans le cas de cette analyse, présent chez le pommier et le pêcher ou sur tous les gènes des génomes. Il est à noter qu'il est possible d'utiliser FractBias sur un ou plusieurs génomes, mais il est clé qu'un groupe de gènes non-dupliqués soient présents.

Par ailleurs, afin de valider ce résultat nous avons calculé le biais de fractionnement à l'aide d'un pipeline *ad-hoc* accessible via le lien suivant : <https://forgemia.inra.fr/irhs-bioinfo/genome-fractionation-biais>. Ce pipeline repose sur l'identification des gènes homologues intraspécifiques de *M. domestica*, intraspécifiques de *P. persica* et interspécifiques entre *P. persica* et *M. domestica* avec un *blastP all-against-all*. Les blocs de synténie interspécifiques sont alors régénérés avec i-ADHoRe. De la même façon que FractBias, une

fenêtre glissante est alors passée le long des chromosomes de *P. persica* et la formule 3.1 est appliquée.

Concernant les deux pipelines, les analyses en aval sont similaires et sont accessibles sur le dépôt suivant : <https://forgemia.inra.fr/tanguy.lallemand/gene-fractionnement>. De cette manière, les résultats bruts de pourcentage de rétention des gènes ont ensuite été traités pour comparer les biais de fractionnement entre les paires de chromosomes en utilisant un t-test apparié en configuration bilatérale et unilatéral. Le choix de la latéralité des tests a été fait selon les attentes biologiques. Pour les paires 01-07, 03-11, 06-14, 09-17, 13-16, nous avons cherché à savoir si le premier chromosome (ou fragment chromosomique) de la paire était sous-fractionné ; pour les paires 02-15, 05-10, 08-15, nous avons cherché à savoir si le premier chromosome (ou fragment chromosomique) de la paire était sur-fractionné. Un ensemble de visualisations a aussi été généré.

### 3.3 Résultats

En utilisant le génome du pêcher comme référence, nous avons évalué le nombre de gènes paralogues entre les régions chromosomiques synténiques de la pomme et le génome du pêcher à l'aide de FractBias accessible via SynMap 2. Différentes analyses avec des ensembles différents de paramètres ont été testées, les résultats avec l'ensemble des gènes (paramètre *all genes*) pour une taille de fenêtre à 100 gènes présentent les résultats avec la plus grande différence visuelle. La rétention des gènes a été calculée pour 46 225 100 fenêtres glissantes de 100 gènes.

L'ensemble des résultats des pourcentages de rétention sont présentés grâce à un ensemble de courbes en Figure 3.1. Les Figures détaillées sont présentées en annexe dans les Figures A.1, A.2, A.3, A.4, A.5, A.6 et A.7. Sur la figure globale (3.1), on peut tout d'abord confirmer que les paires de fragments chromosomiques synténiques identifiés à l'aide d'i-ADHoRe3.0 sont bien retrouvées par cette approche. En effet, leurs répartitions et les points de rupture sont similaires à ce qui est constaté sur le diagramme circulaire présenté lors de la construction des blocs de synténie en Figure 2.3. Par exemple, on peut observer que le chromosome 1 de Prunus (Pp01) est homologue des couples synténiques 13-16 et 8-15. Le chromosome 2 de Prunus (Pp02) est homologue des couples 1-7 et 2-7. Pp03 est quant à lui homologue des chromosomes 17 et 9 chez le pommier. L'ensemble des autres paires se comportent aussi en conformité avec les paires de fragments chromosomiques ohnologues déjà identifiées. La seule différence est la paire 3-11, qui se retrouve

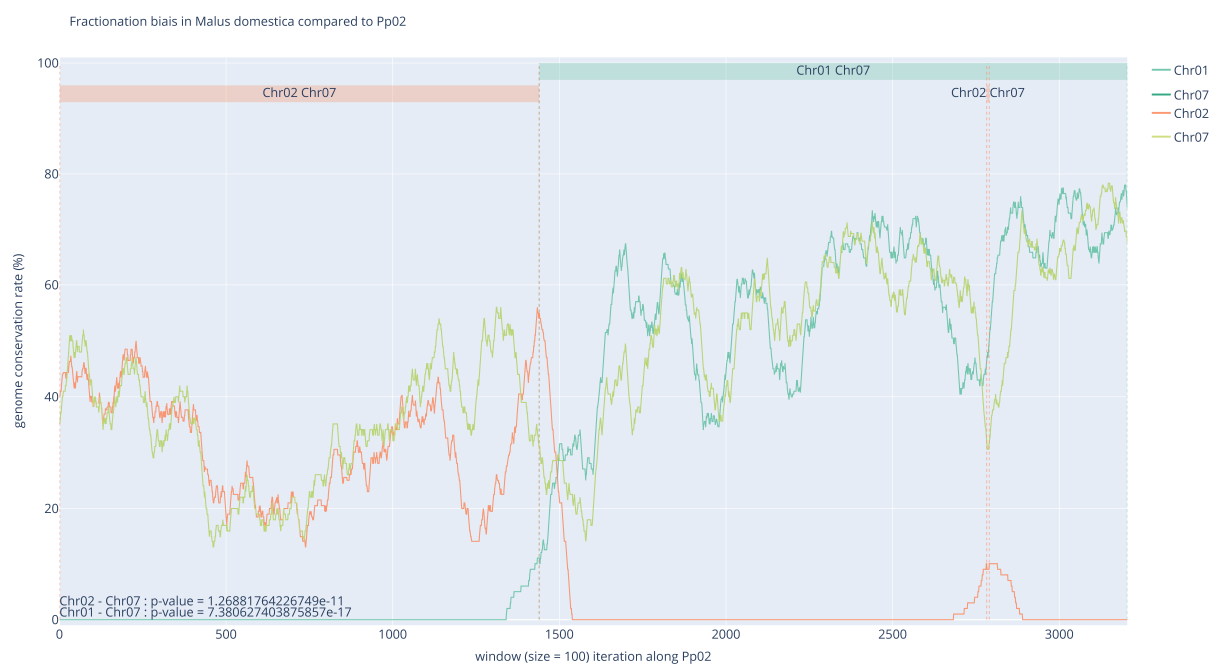
divisée sur plusieurs chromosomes de *Prunus*, Pp04, Pp06 et Pp08. Lors des ruptures de synténies au niveau du chromosome *Prunus* portant différentes paires synténiques de *Malus*, par exemple aux alentours de la fenêtre 2800 sur le chromosome Pp02 (Figure 3.2), on peut observer un croisement des courbes qui est un artefact dû à l'approche par fenêtre glissante. Ce phénomène ne pose pas de problème si ce n'est lors de la visualisation.



**Figure 3.1** — Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes dans le génome de *P. persica*. Les axes des ordonnées représentent le pourcentage de rétention des gènes du pommier par rapport au chromosome du pêcher. Les axes des abscisses exposent le numéro ordonné de la fenêtre glissant le long du chromosome du pêcher. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome de *M. domestica* dont les codes couleurs sont indiqués en légende en haut à droite. Les rectangles sur le haut des graphiques associés aux pointillés verticaux symbolisent les zones synténiques. Un biais de fractionnement peut être observé dans différentes régions comme la fin du chromosome 8 du pêcher (Pp08) ou la région centrale du chromosomes 5 (Pp05) et le premier tiers du chromosomes 3 (Pp03).

À partir de cet ensemble de courbes, plusieurs zones de sur-fractionnement et de sous-fractionnement peuvent être observées parmi les fragments de chromosomes, comme l'extrémité du chromosome 8 de *P. persica* (Pp08, fenêtres numéro 2076 à 2144). Dans cette région, le pourcentage de gènes retenus sur le chromosome 11 de *Malus* (la médiane dans cette région est de 38 %) est supérieur au pourcentage identifié pour le chromosome 3 (la médiane est de 25 %). De même, la zone centromérique du chromosome 5 de *P. persica* (Pp05) présente un sous-fractionnement du chromosome 14 de *Malus* par rapport à son





**Figure 3.2** – Pourcentage de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. L'axe des ordonnées présente le pourcentage de rétention des gènes du pommier par rapport au chromosome du pêcher. L'axe des abscisses présente le numéro ordonné de la fenêtre glissante le long du chromosome du pêcher. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome de *M. domestica* dont les associations sont présentes en légende. Les rectangles sur le haut des graphiques associés aux pointillés verticaux représentent les zones synténiques. Les *p-values* associées au test de Wilcoxon sont présentées dans le coin en bas à droite du graphique. Un biais de fractionnement peut être observé dans différentes régions comme les fenêtres 1000 à 1500 et autour de la fenêtre 2200.

chromosome homologue sur cette portion, qui est le chromosome 6 de Malus. Ainsi, on peut observer des différences locales entre les pourcentages de rétention des gènes.

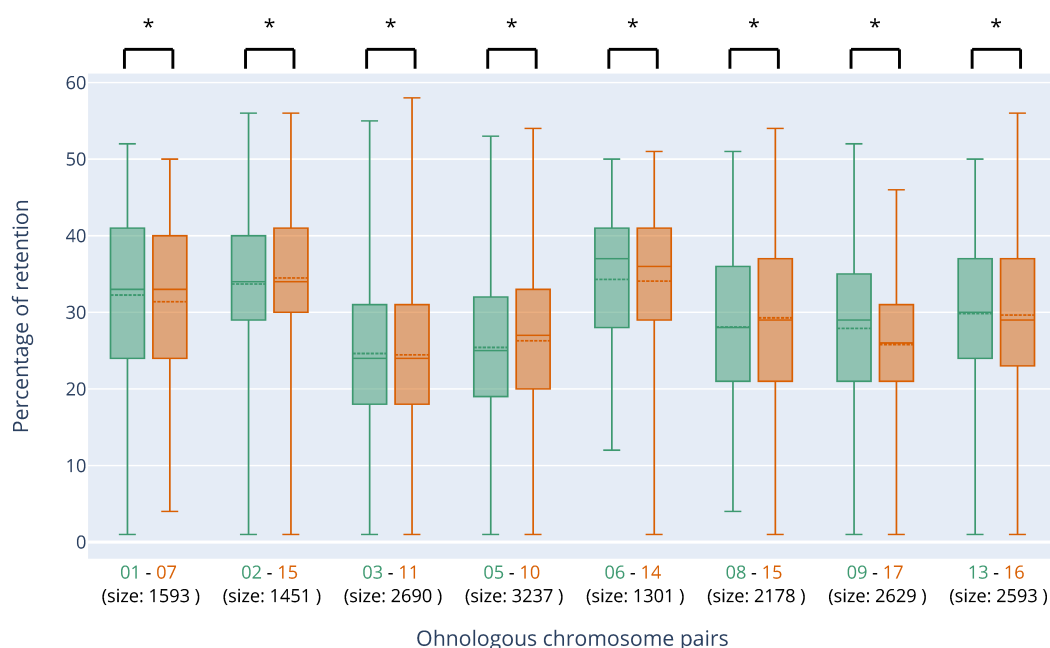
Pour vérifier si le fractionnement est significativement différent à l'échelle des paires de fragments chromosomiques synténiques, les pourcentages de fractionnement des chromosomes synténiques dans toutes les fenêtres pour une paire de chromosomes donnée ont été testés à l'aide d'un t-test apparié. La distribution des pourcentages de rétention est représentée sous forme de diagrammes en boîtes à moustaches dans la Figure 3.3, auquel sont associés les résultats des t-tests appariés. Les résultats précis des t-tests appariés sont quant à eux présentés en Table 3.1. Des t-tests appariés unilatéraux ont aussi été exécutés permettant de confirmer le sens du déséquilibre du fractionnement.

Ainsi pour la paire 1-7, le chromosome 1 est significativement ( $\alpha < 10^{-5}$ ) moins fragmenté avec un pourcentage moyen de rétention de 32,25 % par rapport à son homologue, le chromosome 7 présentant un pourcentage de rétention de 31,37 %. La paire 2-15 présente le chromosome 2 comme moins fractionné que son ohnologue, le chromosome 15 (34,49 % contre 33,69 %). Quant aux paires 3-11, 6-14, 9-17 et 13-16, les premiers chromosomes de la paire (le 3, 6, 9 et 13) sont significativement moins fractionnés que leurs ohnologues. Pour les paires 5-10 et 8-15, ce sont les seconds chromosomes des paires qui sont les moins fractionnés.

À partir de cette analyse statistique du fractionnement des gènes et des visualisations associées, nous avons établi que l'ensemble des principales paires de fragments chromosomiques synténiques sont biaisées dans leur taux de fractionnement des gènes.

**Table 3.1** – Résultats des t-test appariés des pourcentages de rétention des gènes. L'ensemble des principales paires de fragments chromosomiques ohnologues sont significativement différentes en termes de distribution des pourcentages de rétention.

couple	01-07	02-15	03-11	05-10	06-14	08-15	09-17	13-16
Moyenne premier chromosome	32,25	33,69	24,62	25,42	34,29	28,07	27,90	29,82
Moyenne second chromosome	31,37	34,49	24,44	26,29	34,06	29,28	25,77	29,64
Nombre de fenêtres	3187	2902	5380	6475	2603	4357	5258	5186
p-value t-test	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0,0346	$< 10^{-5}$	$< 10^{-5}$	0,0137



**Figure 3.3** – Distribution du pourcentage de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante sur le génome de *P. persica*. Sur l'axe des abscisses, les paires de fragments chromosomiques ohnologues considérées sont présentées. Le nombre de paires de gènes est indiqué entre parenthèses. L'axe des y représente le pourcentage des valeurs de rétention. Le bas de la boîte représente le premier quartile, le trait plein interne à la boîte présente la médiane de la distribution, le trait pointillé la moyenne, et l'extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l'écart interquartile. La moustache inférieure présente le neuvième décile et la moustache supérieure le premier. La première boîte à moustache représente la distribution sur le premier chromosome de la paire (en vert) et la distribution sur le second chromosome de la paire (en orange). \*P-value < 0,05 pour le test t apparié.

## 3.4 Discussion

Nous avons pu calculer les pourcentages de rétention des gènes ohnologues du pommier en nous appuyant sur le génome de *P. persica* comme génome de référence. Ce génome a été sélectionné, car il s'agit d'un génome de haute qualité, proche du génome du pommier, et sans la dernière WGD des *Maloideae*. Ainsi, on s'attend à ce que les synténies pommier-pêcher présentent une synténie 2 :1, c'est-à-dire 2 gènes de pommier et un gène orthologue chez le pêcher. Cette situation fait du pêcher un référentiel de choix pour calculer les pourcentages de rétention des gènes ohnologues et les comparer entre fragments chromosomiques synténiques. Par ailleurs, le génome du pêcher est employé comme référence dans d'autres analyses et notamment chez des espèces de *Pomoidea* comme *Pyrus communis* (Q. Li et al., 2019).

Après différents essais de jeux de paramètres, les résultats liés à tous les gènes avec une taille de fenêtre à 100 gènes semblent produire les résultats de la meilleure qualité. Les résultats de fractionnement (obtenus avec FractBias) associés aux seuls gènes synténiques permettent d'obtenir des visualisations plus claires, mais présentent la difficulté que les paires de gènes synténiques utilisées sont celles qui ont été construites par SynMap. L'outil ne propose pas l'intégration en entrée de blocs de synténie identifiés par d'autres méthodes. Ainsi, pour des problèmes d'homogénéité entre les analyses, nous avons ici choisi d'analyser en profondeur les conclusions associées à tous les gènes. De plus, lors du stage d'Adèle Desmazières (étudiante de L2 en double licence Biologie et Informatique), nous avons pu mettre en place un pipeline produisant des tests de biais de fractionnement à partir de blocs de synténie intra et interspécifiques entre *M. domestica* et *P. persica*. Nous avons pu ainsi effectuer une recherche d'homologies par BlastP sur les séquences protéiques afin d'en déduire les pourcentages de rétention. À des fins exploratoires, ce pipeline a aussi été exécuté à partir des séquences nucléiques. Les conclusions entre les données protéiques et nucléiques sont très comparables. Les analyses statistiques des distributions des pourcentages de rétention via des t-tests appariés ont permis d'obtenir un résultat similaire entre le pipeline *ad-hoc* et le résultat de FactBias, renforçant la fiabilité du résultat obtenu.

Nous avons ainsi mis en évidence un biais de fractionnement entre les segments chromosomiques synténiques du pommier par rapport au pêcher. Concernant une majorité des chromosomes de *P. persica* et notamment les chromosomes 1, 2, 4, 5 et 7 on peut observer qu'ils sont synténiques avec deux paires de chromosomes du pommier, c'est-à-dire quatre

chromosomes de *M. domestica*. La transition entre les paires semble se faire sur la région centromérique, par exemple sur la région des fenêtres 1100 à 1300 du chromosome 7 de *P. persica*. Les chromosomes 3 et 8 quant à eux ne sont synténiques qu'avec une paire de chromosomes de *M. domestica*, les paires 5-10 et 9-17 qui sont des paires ayant subi peu de remaniement en post-WGD. Le chromosome Pp06 présente un cas particulier. En conséquence, ce chromosome montre une synténie avec de multiples chromosomes de *M. domestica* et en particulier une partie de la synténie entre le chromosome 3 et 11 de *Malus*, mais aussi les paires 1-7, 4-12 et 1-15. De plus, dans une moindre mesure, on retrouve un certain nombre de gènes homologues des paires 2 et 15. Ainsi, ce chromosome semble avoir été particulièrement remanié. La paire de chromosomes 3-11 comporte aussi une spécificité. En effet, cette paire est synténique avec trois chromosomes de *P. persica*, le chromosome 4, 6 et 8. Pourtant, les chromosomes 3 et 11 ne semblent pas particulièrement réarrangés, avec une synténie à l'échelle du chromosome au sein du génome de *M. domestica* comme le montre la visualisation circulaire des synténies intraspécifiques présentées en Figure 2.3.

Ainsi on peut observer que le taux de rétention moyen des gènes est différent suivant les paires de fragments chromosomiques synténiques. Ainsi, les paires 6-14 et 2-15 sont les paires présentant les plus grandes rétentions de gènes moyennes avec des valeurs autour de 34%. À l'inverse les paires 8-15 et 9-17 présente des pourcentages de rétention en moyenne plus bas avec des valeurs autour de 27%. Par ailleurs, le biais de fractionnement est dépendant des paires des chromosomes. Ainsi, les paires 1-7, 2-7, 2-15, 6-14 et 9-17 semblent les plus fractionnées. À l'inverse, les paires 3-11 et 13-16 semblent présenter un résultat moins marqué. Ainsi, il semblerait que les paires de *M. domestica* ayant subi le plus de réarrangements, présentent les plus hauts pourcentages de rétention de gènes. Par ailleurs, ils sont aussi ceux présentant les biais de fractionnement les plus accentués.

Ainsi, il semblerait que les paires de chromosomes ayant subi le plus de réarrangements présentent un biais plus marqué. Ceux-ci est probablement dus aux mécanismes à l'origine du biais de fractionnement génomique et notamment les déséquilibres d'insertions d'ETs et de méthylation. Ces mécanismes seront traités dans les chapitres suivants afin de valider cette hypothèse.

Par ailleurs, notre approche ne permet pas de prendre en compte de prendre les pseudogènes. Pourtant, ceux-ci pourraient être la trace de la mise en place du biais de fractionnement. Cependant, l'annotation ne permet pas une étude des pseudogènes. Ainsi, une expertise manuelle sur 1 Mb sur le haut du chromosome 13-16 a été faite. Toutefois

nous n'avons pas identifié de déséquilibre significatif du nombre de pseudogènes entre ces deux régions. Cette zone avait été choisie, car elle présentait un nombre important de QTLs, néanmoins peut être que celle-ci n'était pas propice à cette analyse. De même, la taille de 1 Mb avait été choisie pour des raisons de temps, puisque l'expertise manuelle et l'identification de tous les pseudogènes est chronophage. Néanmoins cette région était peut-être trop petite.

Ce type de résultat a déjà été observé au sein de différentes espèces qui ont connu une WGD récente suivie d'une diploïdisation : par exemple chez *A. thaliana* (Thomas et al., 2006) ou chez le maïs. Le biais de fractionnement a été testé chez le poirier (Q. Li et al., 2019) avec *P. persica* comme référence avec une fenêtre de 100 gènes. Il n'a pas été observé de biais.

### 3.5 Conclusion

Ainsi, à partir du génome du pêcher comme référence, nous avons pu étudier le fractionnement du génome du pommier à l'aide de l'outil FractBias. Nous avons pu mettre en évidence un déséquilibre du fractionnement génomique chez le pommier entre les principales paires de fragments chromosomiques synténiques. Donc les paires 1-7, 2-15, 3-11, 5-10, 6-14, 8-15, 9-17 et 13-16 ont des pourcentages de rétention de gènes significativement différents. Le biais de fractionnement est un mécanisme important dans le processus de rediploïdisation et a un impact dans un ensemble de processus. Il a été observé jusqu'à présent dans des espèces avec des dominances de sous-génomes. Il peut notamment être lié à des pressions de sélection différentes entre les segments ohnologues, des participations à la variation phénotypique différentes, des déséquilibres transcriptionnels, et/ou des déséquilibres dans la répartition des ETs et de méthylation de l'ADN. L'ensemble de ces mécanismes vont donc être analysé au cours des prochains chapitres.



# COMPARAISON DE LA PROPORTION DES QTLs ENTRE LES FRAGMENTS SYNTÉNIQUES

---

## 4.1 Introduction

La duplication complète du génome par WGD a de nombreux impacts sur le génome et le phénotype de l'organisme. Une partie des impacts et du devenir des gènes dupliqués par WGD ont été modélisés par Freeling (Freeling et al., 2012). Ce modèle estime entre autres que la participation des chromosomes à la variation phénotypique de l'organisme est biaisée. En effet, ce modèle prédit un biais dans le niveau d'expression des gènes ohnologues, avec un des sous-génomes étant plus souvent exprimé. Ce déséquilibre transcriptionnel implique que le gène le plus fortement exprimé d'une paire d'ohnologue produit davantage de protéines et contribue donc davantage au phénotype (Renny-Byfield et al., 2017). Ce modèle est soutenu par différentes observations chez plusieurs espèces.

Chez le maïs, il a été montré que la participation à la variation phénotypique était significativement différente entre les gènes ohnologues (Renny-Byfield et al., 2017). De plus, il a été observé que les gènes liés à un phénotype mutant étaient plus probablement localisés sur le sous-génome dominant (J. C. Schnable & Freeling, 2011). De même, chez *B. rapa*, il a été identifié des disparités dans la participation au phénotype entre les différents sous-génomes (Z. Wang et al., 2022). Un constat similaire a été observé chez l'olivier en particulier pour des traits associés à la production de glucosinolates responsables de la production oléagineuse (S. Liu et al., 2014; Unver et al., 2017). Chez le blé, il a été observé une asymétrie de la participation des différents sous-génomes à des traits d'intérêt agronomique ainsi que des traits associés à la résistance aux pathogènes (Feldman et al., 2012).

Néanmoins, ce déséquilibre de participation à la variation phénotypique est un méca-



nisme qui n'est pas toujours systématique et peut dépendre de facteurs spatio-temporels. Par exemple, chez le coton et la myrtille, il a été observé que l'un des sous-génomés était plus souvent exprimé dans certains tissus et sous certaines conditions. Ainsi chez la myrtille, mais aussi chez le coton, un sous-génome semble être lié au phénotype du développement du fruit tandis que l'autre sous génome participe à la variation phénotypique des autres traits (Colle et al., 2019; L. Flagel et al., 2008).

Afin d'étudier la participation à la variation du phénotype des différents segments chromosomiques ohnologues chez le pommier, nous avons choisi de nous appuyer sur les nombreux QTLs publiés. Un QTL correspond à un locus impliqué dans la variation quantitative d'un trait phénotypique et co-localisé avec un ou plusieurs marqueurs génétiques associés (Myles & Wayne, 2008). Par le biais d'études de populations en ségrégation, il est possible de calculer le taux de recombinaison entre *loci* (sous la forme de marqueurs génétiques). Ce taux de recombinaison permet alors de construire des cartes génétiques. Ces cartes autorisent alors à estimer l'association entre des marqueurs génétiques et des variables phénotypiques quantitatives mesurées sur les plantes, et ainsi d'identifier des QTLs *locus* (Members of the Complex Trait Consortium, 2003). Ces QTLs représentent des régions génomiques plus ou moins étendues (de l'ordre de plusieurs centiMorgan - cM) et vont donc être associées à plusieurs gènes ou *loci* impliqués dans l'expression d'un caractère complexe. Les QTLs et marqueurs associés peuvent ensuite être utilisés, entre autres, pour réaliser de la Sélection Assistée par Marqueurs (SAM), pour ne sélectionner que les individus portant les caractères désirés. Les marqueurs utilisés peuvent être de différents types, mais on retrouve principalement des marqueurs Simple Sequence Repeats (SSR), Restriction Fragment Length Polymorphism (RFLP) ou Single Nucleotide Polymorphism (SNP).

Les QTLs peuvent être associés à une valeur numérique appelée le  $R^2$ . Cette valeur représente le pourcentage de variation du trait phénotypique expliqué par le QTL (Myles & Wayne, 2008). On peut ainsi classer de manière empirique les QTLs en deux catégories suivant leur valeur de  $R^2$ . On retrouve en premier lieu des QTLs dits mineurs avec des  $R^2$  inférieur à 30 % et des QTLs majeurs au-dessus de 30 % d'explication de la variation phénotypique.

Les QTLs peuvent être identifiés à partir de deux types de populations. D'une part, ils peuvent être issus d'études conduites sur des individus F1 obtenus à partir de croisements contrôlés biparentaux. Ces QTLs représentent alors qu'une part réduite de la diversité génétique de l'espèce. D'autre part, ils peuvent être issus d'études Genome-Wide Associa-

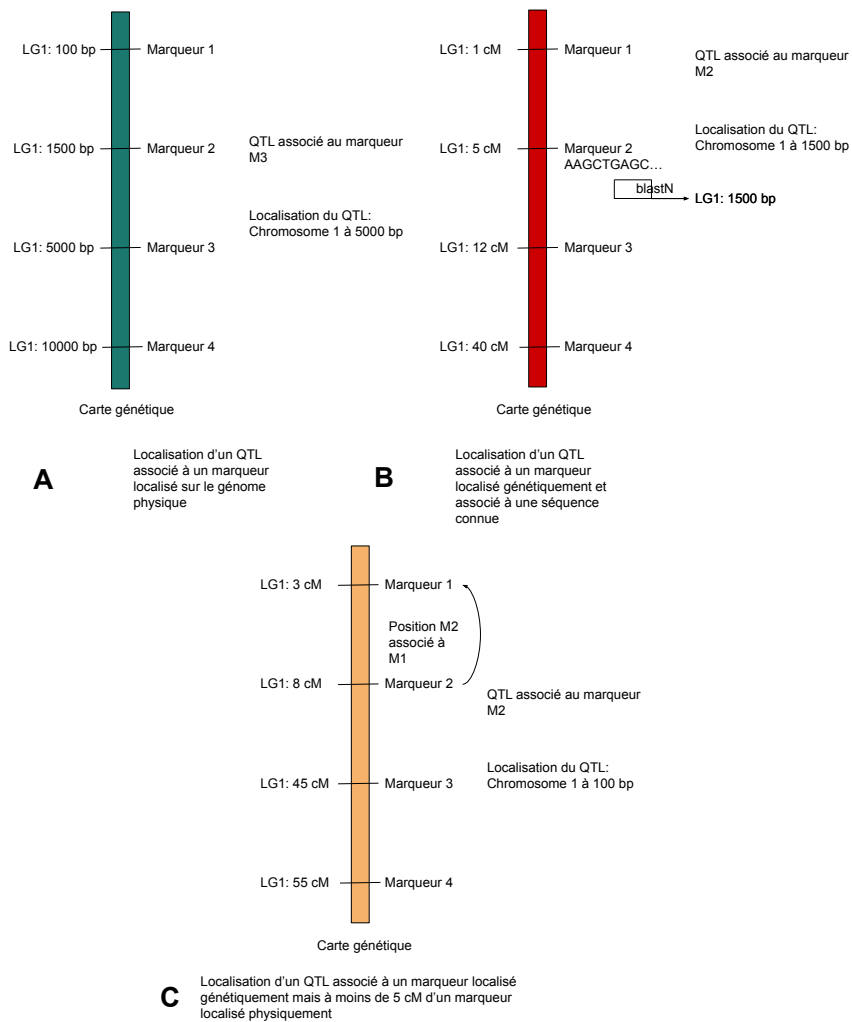
tion Study (GWAS) dans lesquelles les QTLs ont été identifiés dans des *core collections* ou des populations de taille importante qui représentent mieux la diversité génétique de l'espèce.

Le génome du pommier a été dupliqué par une WGD. Nous avons ainsi cherché si un déséquilibre dans la participation à la variation phénotypique pouvait être identifié entre des paires de fragments chromosomiques ohnologues.

## 4.2 Matériel et méthode

Afin de mener une étude aussi exhaustive que possible, nous avons récupéré et localisé un maximum de QTLs. Une partie des QTLs d'origine biparentale a été récupérés dans la base de données Genome Database for Rosaceae (GDR) (Jung et al., 2018). Nous avons aussi récupéré manuellement dans différents articles de la littérature un ensemble de QTLs associés à des études GWAS. Les QTLs récupérés manuellement devaient présenter un intervalle de confiance de la localisation et un trait phénotypique associé pour être intégré au jeu de données. Dans GDR, les QTLs sont localisés par des marqueurs sur des cartes génétiques plus ou moins saturées. Leur localisation sur le génome physique de GDDH13 1.1 n'est pas fournie dans tous les cas. Dans ce but, un ensemble de scripts a été construits pour résoudre cette problématique et placé dans le dépôt accessible à l'adresse suivante : <https://forgemia.inra.fr/tanguy.lallemand/qtl-desequilibrium>.

Dans un premier temps, nous avons supprimé les QTLs qui n'étaient pas associés à des marqueurs génétiques de séquence connue. De même, les QTLs localisés à une distance trop importante de marqueurs associés à une position sur des cartes génétiques comportant une densité trop faible de marqueurs ont été supprimés. En effet, ils ne peuvent pas être associés de manière fiable à une localisation physique. Pour les QTLs restants, nous avons cherché à estimer leur position sur le génome physique à l'aide de différentes approches schématisées en Figure 4.1. Pour les QTLs associés à un marqueur localisé physiquement, nous avons récupéré la localisation des marqueurs génétiques dans la base de données GDR pour GDDH13 1.1 quand celle-ci existe. Cette étape correspond à la Figure 4.1 (A). Pour les marqueurs non localisés physiquement, mais dont les séquences d'amorces ou de sondes sont connues, une recherche par BlastN (Camacho et al., 2009) a été réalisée. Si la séquence peut-être localisée sur le génome physique, le QTL est associé à cette position (Figure 4.1 (B)). Pour les QTLs associés à des marqueurs localisés génétiquement et dont les séquences ne sont pas connues, nous avons utilisé la carte génétique associée pour estimer



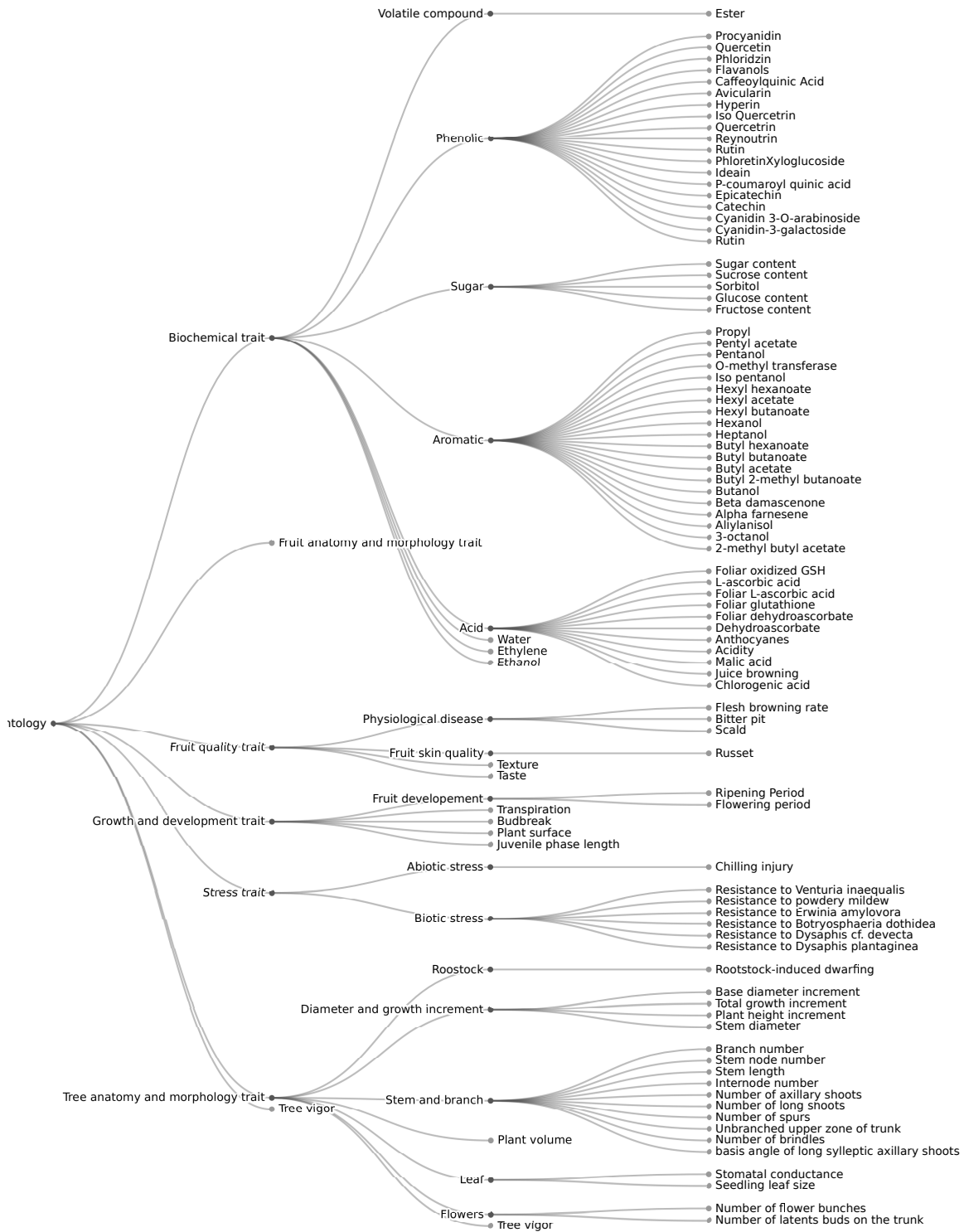
**Figure 4.1** – Schéma des différentes possibilités de localisation de QTL suivant le marqueur auquel il est associé. Le premier schéma (A) avec la carte génétique en vert présente le cas d'un QTL associé à un marqueur localisé physiquement sur GDDH13. Le deuxième schéma (B) avec la carte génétique en rouge présente le cas d'un QTL associé à un marqueur localiser génétiquement et associé à une séquence connue. Il est alors possible de le localiser sur le génome physique à l'aide de Blast. Le troisième schéma (C) avec la carte génétique en orange présente le cas d'un QTL associé à un marqueur localisé génétiquement et non associé à une séquence connue. Il est alors possible de le localisé sur le génome physique si un marqueur proche dans la carte génétique est associé à une position sur le génome physique.

la position du marqueur. Cette approche n'est possible qu'à la condition que la carte génétique comporte une densité de marqueurs suffisante pour estimer leur localisation. En effet, lorsqu'un marqueur dont la localisation est connue est situé à moins de 5 cM du marqueur dont la position est recherchée, ce marqueur est associé à la localisation du marqueur de position connue. La position du QTL associé à ce marqueur nouvellement localisé est associée à cette position estimée. Cette étape est décrite en Figure 4.1 (C).

Tous les QTLs conservés dans le jeu de données sont annotés avec des métadonnées décrivant le trait phénotypique lié au QTL. Dans GDR, cette annotation n'était pas faite par un champ contrôlé, rendant le traitement automatique de l'information difficile. Pour surmonter ce problème, nous avons construit une ontologie et ré-annoté l'ensemble des QTLs avec celle-ci en recourant à un traitement manuel des annotations existantes. Cette ontologie est construite sur trois niveaux composés respectivement de 7, 20 et 59 termes permettant d'atteindre différents degrés de précision d'annotation. L'ensemble des relations de cette ontologie hiérarchique est de type "is a". Le formalisme utilisé est JSON, un format standard dont la conversion vers des formats spécialisés dans le stockage d'ontologie comme le OWL est simplifiée et facilite une utilisation programmatique. Le détail de l'ontologie est présenté sous forme d'arbre hiérarchique en Figure 4.2. La taille des différentes catégories a été équilibrée dans la mesure du possible pour éviter les biais statistiques.

Du fait de l'approche basée sur les QTLs, une redondance des données peut apparaître puisque deux QTLs proches contrôlant l'expression de caractères similaires sont susceptibles d'être associés aux mêmes *loci* ou gènes. Pour éviter ce biais et réduire la redondance, si l'intervalle de confiance du QTL (artificiellement augmenté de 10 cM en amont et en aval) chevauche un QTL associé à un trait de la même catégorie dans l'ontologie, alors le QTL présentant le taux d'explication de la variance,  $R^2$  est le plus bas est supprimé. Pour finir, seuls les QTLs associés à un intervalle de confiance localisé, au moins en partie, dans un bloc synténique ont été conservés dans le jeu de données final.

Enfin, la différence de proportion a été testée à l'aide d'un test z de proportion. Le test a été effectué entre tous les QTLs de chaque paire de segments chromosomiques. De plus, il a été réalisé en filtrant les QTLs au niveau des catégories de l'ontologie sur les trois niveaux pour tester les effets associés à des traits phénotypiques connexes. Ces analyses sont disponibles dans le dépôt <https://forgemia.inra.fr/tanguy.lallemand/statistical-exploration-qt1-desequilibrium>.

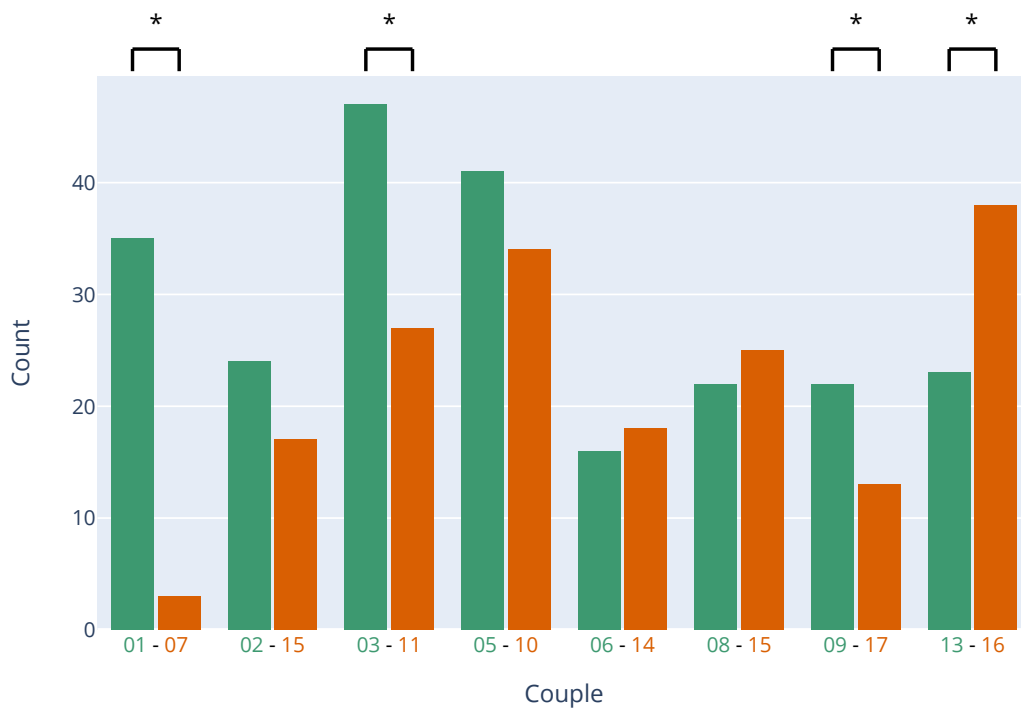


**Figure 4.2** – Arbre hiérarchique de l'ontologie des métadonnées décrivant les traits phénotypiques associés aux QTLs. Cette ontologie de trois niveaux représentée par les trois niveaux de nœuds présentés de gauche à droite. Les branches représentent les relations de type "is\_a" entre les termes de l'ontologie.

## 4.3 Résultats

En utilisant la base de données GDR (Jung et al., 2018), nous avons récupéré 1520 QTLs obtenus à partir de cartes génétiques biparentales. Un premier filtre a permis de supprimer 441 QTLs non associés à des marqueurs génétiques. Aux 1079 QTLs conservés, un total de 135 QTLs dérivés d'études GWAS a ensuite été ajouté manuellement. Pour ces QTLs les intervalles de confiance ont été récupérés dans la littérature. Pour déterminer le positionnement physique sur le génome de la pomme des 1079 QTLs associés à au moins un marqueur et donc à une localisation génétique, nous avons utilisé un ensemble de 515 990 marqueurs disponibles sur GDR. Ces marqueurs sont surtout des GDR (506 047 marqueurs SNP et 8519 marqueurs SSRs). On retrouve aussi un certain nombre d'autres types de marqueurs, dont des marqueurs Random Amplified Polymorphic DNA (RADP), Amplified Fragment Length Polymorphism (AFLP) et divers marqueurs génétiques. Parmi ces 515 990 marqueurs, un total de 513 748 marqueurs ont été localisés sur la carte physique du génome, à partir d'une information déjà existante dans GDR, par recherche de séquences SSR via Blast ou par utilisation de la position de marqueurs proches au sein des cartes génétiques. Ces marqueurs ont permis de localiser 733 QTLs associés à des expériences biparentales sur l'ensemble du génome du pommier en version GDDH13 parmi les 1079 disponibles. À ceux-ci s'ajoutent 135 QTLs provenant d'études GWAS, pour un total de 868 QTLs localisés.

Les métadonnées associées aux QTLs ont été réexpertisées manuellement et associées aux termes de l'ontologie construite pour les décrire. La visualisation sous forme d'arbre hiérarchique de l'ontologie est fournie en Figure 4.2. En raison de la redondance de certaines études de QTLs, de la similarité entre certains traits étudiés, et pour éviter d'artificiallement sur-représenter certains caractères trop similaires présents plusieurs fois, une étape de nettoyage basée sur l'annotation des QTLs par l'ontologie des traits a été effectuée. Ceci a permis d'obtenir un ensemble de données final composé de 541 QTLs (sur les 868 que nous avons réussi à localiser sur le génome du pommier). Ces QTLs sont non redondants d'un point de vue de la localisation et de leur participation au phénotype. Le jeu de données final est composé de 341 QTLs associés à des SNP, 178 associés à des marqueurs SSRs et 22 associés à des marqueurs d'autres types dont des marqueurs RADP, des marqueurs génétiques et AFLP. Dans GDR, les QTLs peuvent être associés à plusieurs marqueurs. Les QTLs retenus dans le jeu de données final sont tous associés à au moins un marqueur, certains à deux ou trois. Enfin comme ultime filtre, seuls les QTLs dont

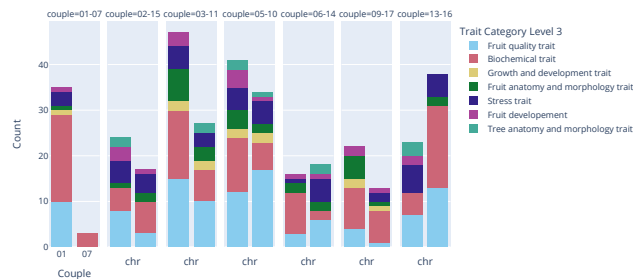


**Figure 4.3** – Histogramme du nombre de QTLs associés aux blocs synténiques dans chaque paire de fragments de chromosomes ohnologues. L’axe des ordonnées est associé au nombre de QTLs portés par un fragment chromosomique ohnologue. L’axe des abscisses présente les paires de fragments chromosomiques ohnologues analysées. La barre verte est associée au premier chromosome de la paire. La barre rouge est associée au deuxième chromosome de la paire. \*P-value < 0,05 pour le test de proportion z

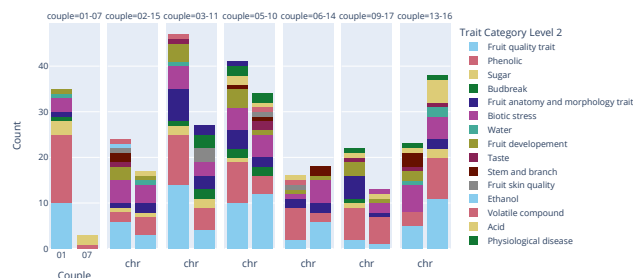
les intervalles de confiance ont été identifiés comme localisés au moins en partie dans un fragment de chromosome considéré comme synténique ont été conservés, ce qui représente un total de 487 QTLs. La distribution du nombre de QTLs localisés sur chaque chromosome des principales paires de chromosomes synténiques est présentée sous la forme d'un histogramme en Figure 4.3. La comparaison des proportions de QTLs localisés sur des fragments synténiques a été réalisée à l'aide d'un test z de proportion et d'un test binomial. Ces tests ont été réalisés par paire de fragments chromosomes en considérant l'ensemble des QTLs. Des tests ont aussi été menés selon l'annotation dans l'ontologie du trait associé. Globalement, les paires 1-7, 3-11, 13-16 et 8-15 ont été identifiées comme significativement déséquilibrées.

Les différentes proportions des différentes annotations des traits phénotypiques associés aux QTLs sont présentées en Figure 4.4. La Figure 4.4A présente les proportions pour le premier niveau de l'ontologie. On peut observer que la plupart des QTLs sont associés à des traits liés à la qualité du fruit ou à la biochimie. La Figure 4.4B présente les proportions des QTLs associés au deuxième niveau de l'ontologie et montre que les QTLs sont plus particulièrement associés à des traits liés aux composés phénoliques, aux sucres, aux stress biotiques et à la morphologie du pommier. Pour finir, la Figure 4.4C présente les proportions pour les traits du dernier niveau de l'ontologie. On peut observer des proportions importantes de QTLs associés à la texture, la résistance aux agents pathogènes, la présence de sucres d'intérêt et la morphologie de l'organisme. Les différences au sein des paires de fragments chromosomiques ohnologues de proportions de QTLs associés aux différents traits ont été testées statistiquement. Les résultats sont présentés sous forme de matrices colorées dans la Figure 4.5. La Figure 4.5A présente les tests pour le premier niveau de l'ontologie, et montre que la majorité des paires sont déséquilibrées pour certains types de QTLs. Néanmoins, les paires 1-7, 3-11, 6-14 et 13-16 sont les plus déséquilibrées dans les proportions de QTLs. Les traits associés sont les traits biochimiques, les caractères associés à la qualité du fruit, au stress et les QTLs associés à l'anatomie et la morphologie du pommier. Le deuxième niveau de l'ontologie dont les résultats sont exposés en Figure 4.5B présente un déséquilibre pour les mêmes paires. Les paires montrent des QTLs plutôt associés aux composés phénoliques et la résistance aux stress biotiques et abiotiques. Pour finir, les tests menés sur le dernier niveau de l'ontologie (Figure 4.5C) montrent un déséquilibre sur des QTLs associés à différents composés phénoliques d'intérêts, des sucres et des pathogènes particuliers. Néanmoins ce dernier niveau comporte des effectifs trop faibles pour effectuer les tests statistiques dans la plupart des

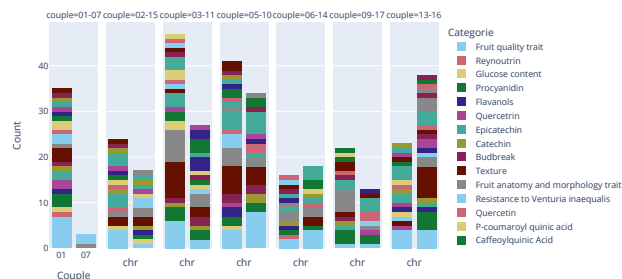




(A) Histogramme à barres empilées des proportions de QTLs annotés par les termes de l'ontologie du premier niveau. Les couleurs des barres sont associées au terme de l'ontologie.



(B) Histogramme à barres empilées des proportions de QTLs annotés par les termes de l'ontologie du deuxième niveau. Les couleurs des barres sont associées au terme de l'ontologie.

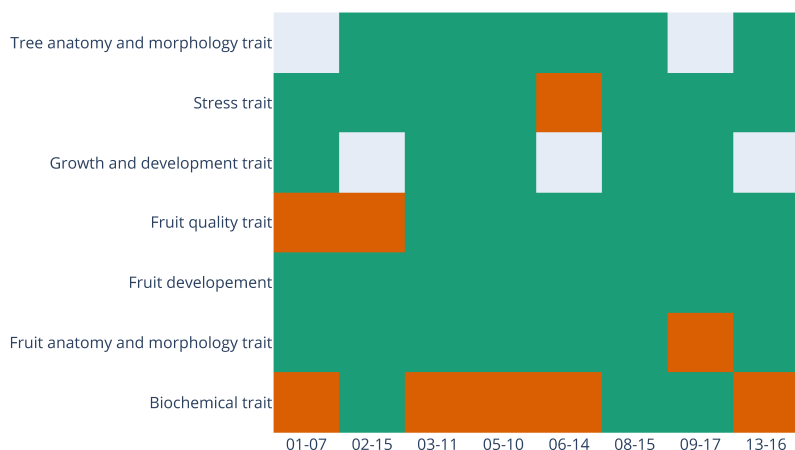


(C) Histogramme à barres empilées des proportions de QTLs annotés par les termes de l'ontologie du dernier niveau. Les couleurs des barres sont associées au terme de l'ontologie.

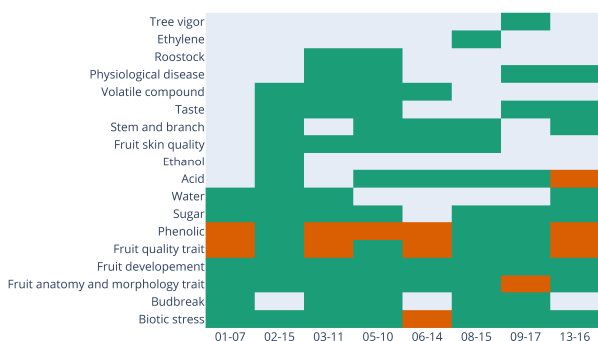
**Figure 4.4** – Histogramme à barres empilées des proportions de QTLs annotés par les termes de l'ontologie. La Figure 4.4A est associée au premier niveau de l'ontologie, La Figure 4.4B est associée au deuxième niveau de l'ontologie, La Figure 4.4C est associée au troisième niveau de l'ontologie. L'axe des ordonnées (*count*) est associé au nombre de QTLs portés par un fragment chromosomique ohnologue. L'axe des abscisses (*couple*) présente les paires de fragments chromosomiques ohnologues analysées. La première barre est associée au premier chromosome de la paire. La seconde barre est associée au deuxième chromosome de la paire. Les couleurs des barres sont associées au terme de l'ontologie.

catégories et présente un intérêt relatif. Il pourrait être intéressant dans le cas d'un jeu de données de taille plus important. Les tests menés à l'échelle des traits phénotypiques associés montrent que les proportions sont différentes surtout pour les traits associés à la production de composés phénoliques, les traits associés à la qualité du fruit et aux stress biotiques, ce qui correspond avant tout à des traits liés au métabolisme spécialisé.

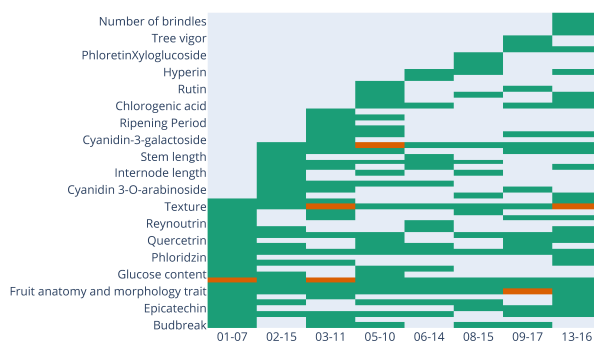
L'ensemble de ces résultats suggère que certains fragments de chromosomes contribuent plus que leurs ohnologues à la variation phénotypique des individus. Ce déséquilibre de QTLs concerne au moins 4 paires de chromosomes synténiques et en particulier les paires 1-7, 3-11, 6-14 et 13-16. Pour la paire 1-7, nous avons identifié le segment chromosomique ohnologue associé au chromosome 1 comme portant une proportion significativement plus importante de QTLs par rapport au segment synténique du chromosome 7. Pour la paire 3-11, c'est le chromosome 3 qui porte une proportion significativement plus importante de QTLs par rapport au chromosome 11.



(A) Matrice de couleur des résultats des tests z de proportion pour les QTLs, filtrés sur le premier niveau de l'ontologie.



(B) Matrice de couleur des résultats des tests z de proportion pour les QTLs, filtrés sur le deuxième niveau de l'ontologie.



(C) Matrice de couleur des résultats des tests z de proportion pour les QTLs, filtrés sur le dernier niveau de l'ontologie.

**Figure 4.5** – Matrice de couleur des résultats des tests z de proportion pour les QTLs, filtrés sur les différents niveaux de l'ontologie. La Figure 4.5A montre le premier niveau, la Figure 4.5B expose le deuxième niveau et la Figure 4.5C présente le troisième niveau. L'axe des ordonnées présente les différents termes de l'ontologie décrivant les traits phénotypiques associés au QTLs. L'axe des abscisses présente les différentes paires de fragments chromosomiques ohnologues. Les cellules colorées en orange représentent des tests significatifs pour un seuil  $\alpha$  à 5%. Les cellules vertes sont associées à des tests non significatifs.

## 4.4 Discussion

Cette étude a été menée sur 1528 QTLs identifiés lors des 20 dernières années chez le pommier et provenant de la base de données GDR et de la littérature. À partir de ce jeu de données initial, les QTLs ont été filtrés afin d'éliminer ceux associés à des traits phénotypiques proches pouvant induire des biais liés à une surreprésentation de ces caractères. En effet les QTLs présentent des localisations sous forme d'intervalles de confiance qui peuvent être larges. De plus, les traits associés aux QTLs peuvent être généraux et pourraient être liés à un même gène/*locus* ou ensemble de gènes/*loci* sous-jacents. Pour éviter d'introduire un biais dans l'analyse en dénombrant plusieurs fois une même information qui est peut-être liée aux mêmes gènes/*loci*, nous avons retiré les QTLs proches physiquement et fonctionnellement. Le jeu de données final rassemble 541 QTLs qui sont donc indépendants et reliés à une diversité de traits phénotypiques. L'étude des métadonnées décrivant les QTLs et en particulier la réannotation, faite à l'aide d'une ontologie, montre que ce jeu de données constitue l'essentiel de la diversité existante en termes de QTLs étudiés chez le pommier et disponible sur des bases de données publiques. L'extraction des données a été faite en 2019, et réexécutée en 2020 pour un résultat similaire. Cette ontologie a été construite pour s'adapter à la description de QTLs liés à des arbres fruitiers et plus particulièrement au pommier. Elle présente des catégories associées à des traits macroscopiques et biochimiques, ainsi que la résistance aux stress biotiques et abiotiques. Le dernier niveau présente des traits précis, plutôt spécifiques au pommier. Dans le cadre d'une généralisation de cette ontologie, il conviendrait d'ajouter des termes à ce dernier niveau afin de correspondre mieux à la diversité des composés produits par les arbres fruitiers.

Cette étude repose sur des différences de comptage de QTLs sur les fragments chromosomiques ohnologues. Les résultats de cette étude représentent ainsi la part de la contribution de chaque fragment chromosomique synténique à la variation phénotypique. En conséquence, l'augmentation de la taille du jeu de données construit représente un intérêt considérable, car il autorise un meilleur pouvoir de résolution statistique ainsi que la possibilité de tester de plus petites paires de fragments synténiques. De plus, les tests utilisés permettent de tester des différences de proportions, mais sont peu résolutifs et nécessitent des effectifs importants pour déceler des différences. L'analyse des proportions de QTLs a permis de mettre en évidence des déséquilibres de QTLs à l'échelle des fragments synténiques, mais aussi pour des QTLs associés à des traits phénotypiques particuliers.

L'analyse en suivant l'ontologie a permis de mettre en évidence un déséquilibre marqué sur le métabolisme secondaire. Néanmoins, cette analyse est incomplète, car un certain nombre de paires ne présentaient pas assez de QTLs associés à certains termes de l'ontologie pour pouvoir les comparer statistiquement. Une augmentation de la taille du jeu de données permettrait de tester plus finement les déséquilibres de QTLs. Cependant, après deux années, l'ensemble de l'analyse a été relancée afin de comparer les résultats, mais le nombre de QTLs stocké dans GDR n'a pas assez évolué pour permettre de nouvelles analyses. L'ajout devrait être fait manuellement à partir de nouvelles données publiées.

Cependant, à l'heure actuelle, seuls les marqueurs SSR dont les séquences d'amorces sont stockées dans GDR peuvent réellement être localisés avec précision. Il conviendrait néanmoins de situer d'autres marqueurs moléculaires et entre autres les marqueurs de type Expressed Sequence Tags Simple Sequence Repeats (EST-SSR) qui rassemblent respectivement 525 occurrences et dont la localisation serait fiable et peu difficile à mettre en place puisque le pipeline actuel pourrait le faire automatiquement à la condition d'obtenir les séquences associées. De plus, les séquences EST-SSRs présentent l'intérêt d'être associées à un gène (EST pour *Expressed Sequenced Tag*, c'est-à-dire associé à un transcrit). Par ailleurs, pour les marqueurs AFLP rassemblant 837 occurrences, leur localisation automatique sera plus difficile. En effet, ces marqueurs sont obtenus par amplification des fragments d'ADN avec des séquences courtes. Ces fragments sont ensuite digérés par des enzymes et migrés sur gel. Ils sont alors utilisés comme marqueurs de taille sur le gel. Ainsi les séquences et la localisation de ces marqueurs présentent de nombreuses difficultés.

Pour le reste des marqueurs, une expertise manuelle des articles associés à ceux-ci pourrait permettre d'en localiser une partie, mais nécessiterait un investissement important pour ne localiser qu'un nombre faible de QTLs. L'approche basée sur l'utilisation des cartes génétiques pour localiser plus de marqueurs permet une augmentation de la taille du jeu de données de marqueurs localisés et donc du nombre de QTLs associés à une position. Néanmoins, le marqueur de position inconnue est localisé à la position du marqueur le plus proche (dans la limite de 5 cM maximum) dans la carte génétique associée à ce marqueur. Ce nouveau marqueur est alors localisé à la même position physique et les QTLs qui y sont associés aussi. Lors de l'étape de filtrage des QTLs chevauchants si ceux-ci sont associés à des traits similaires, ceux-ci vont être supprimés, car ils représentent une donnée redondante et peuvent même sans doute être le reflet des mêmes gènes sous-jacents. Ce filtrage stringent fondé sur la localisation et l'information biologique garantit que les données étudiées sont de bonne qualité, mais limite l'augmentation de la

taille du jeu de données. Il serait donc intéressant de chercher à récupérer le plus possible de localisations de marqueurs afin de placer plus de QTLs. Ces QTLs sont tous associés à des caractères d'intérêt agronomique et ne représentent pas l'ensemble des caractéristiques de la pomme. Néanmoins, afin de limiter les différents biais liés à l'étude des QTLs et notamment la faible précision de localisation, nous avons été rigoureux sur les QTLs retenus et en avons supprimé une grande partie pour le calcul des proportions de QTLs.

L'analyse des proportions tient compte de l'information de localisation et d'annotation du QTL. Néanmoins d'autres informations sont à disposition dans le jeu de données et pourraient être prises en compte en appliquant un poids différent aux QTLs dans les analyses statistiques en suivant différents facteurs. Cependant, les tests utilisés ici ne permettent pas ce type d'analyses. Les facteurs considérés pourraient être la provenance des QTLs qui peuvent être issus, soit d'études biparentales, soit d'études GWAS. Cette information est capitale puisqu'un QTL issu d'une analyse GWAS représente une plus grande diversité génétique que celle d'un QTL issu d'une étude bi-parentale. De même le  $R^2$  est disponible pour une majorité des QTLs, mais n'a pas été pris en compte.

Ce déséquilibre de QTLs entre différentes paires de chromosomes ohnologues constitue un résultat important. En effet, (Freeling et al., 2012) proposent un modèle qui inclut le déséquilibre à la participation à la variation du phénotype (Renny-Byfield et al., 2017). Chez le maïs, il a été montré que la participation à la variation phénotypique pouvait être liée au fractionnement du génome et à des différences de niveau d'expression des gènes ohnologues (Renny-Byfield et al., 2017). Pour le pommier cette étude montre un déséquilibre important du nombre de QTLs portés par les chromosomes de certaines paires synténiques. Ce déséquilibre est observé pour tous les QTLs et se maintient lorsque les QTLs sont comparés au niveau des catégories de l'ontologie. Le sens du déséquilibre n'est pas modifié. Chez certaines espèces comme la myrtille (Y. Wang et al., 2020) ou le blé (Chantret et al., 2005), il a été observé des inversions de la dominance de l'un des sous-génomes dans certains contextes spatio-temporels précis. Un tel mécanisme n'a pas été observé ici.

## 4.5 Conclusion

L'analyse de 1528 QTLs issus de populations F1 biparentales ainsi que de collections représentant un fond génétique plus large (études GWAS), et localisés sur le génome physique du pommier, a permis de mettre en évidence des différences dans les proportions

de QTLs porté par les différents fragments synténiques ohnologues. En conséquence, les paires 1-7, 3-11, 13-16 et 8-15 ont été identifiées comme significativement déséquilibrées dans le nombre de QTLs localisés sur les zones synténiques. L'analyse des QTLs et en particulier de leurs annotations par une ontologie construite pour les décrire, montre que le déséquilibre est en particulier marqué pour des QTLs associés au métabolisme secondaire comme la production de composés phénoliques ou de sucres, à la qualité du fruit et la réponse aux stress. Ces constatations ont été observées chez des espèces allopolyploïdes dans un contexte de sous-dominance génomique. Cette analyse permet d'avoir un premier résultat pour le pommier, un autopolyploïde, suggérant que des mécanismes similaires pourraient aussi être mis en place. Ce contexte de sous-dominance génomique peut s'observer à différentes échelles et sous l'égide de différents mécanismes qu'il convient de tester et notamment en termes de pression de sélection, d'expression des gènes, d'éléments transposables et de méthylation de l'ADN.

# ÉVALUATION DE LA PRESSION DE SÉLECTION DES GÈNES OHNOLOGUES À L'AIDE DU $K_a/K_s$

---

Après identification des gènes ohnologues, nous avons pu mettre en évidence un biais de fractionnement du génome. De plus, nous avons pu identifier un déséquilibre de QTLs entre les fragments chromosomiques ohnologues. Ces déséquilibres peuvent être expliqués par différents facteurs et entre autres des différences d'évolution des séquences codantes. Afin de tester cette hypothèse, nous nous sommes appuyés sur la pression de sélection appliquée sur les gènes ohnologues. Afin de tester s'il y a une différence entre les paires de gènes ohnologues, ces calculs de pression de sélection nécessitent de s'appuyer sur des séquences référence. Celles-ci ont été identifiées par le biais de la construction de triplets de gènes composés de gènes ohnologues du pommier et de séquences de référence de *P. persica*, l'organisme *Rosaceae* le plus proche du pommier avec un génome de bonne qualité et n'ayant pas subi la dernière WGD particulière des *Maloideae*.

## 5.1 Introduction

Afin d'expliquer les déséquilibres observés et les expliquer en suivant un modèle de sous-dominance génomique, nous avons cherché à vérifier si la pression de sélection appliquée aux séquences codantes des gènes ohnologues était similaire. Les technologies de séquençage ont permis l'émergence de plusieurs méthodes pour quantifier les pressions de sélection opérant sur les séquences codant pour des protéines. L'une des approches de référence, reconnue pour sa simplicité et sa robustesse, est le rapport  $dN/dS$  (aussi appelé  $K_a/K_s$  ou  $\omega$ ). Cette mesure permet d'estimer la force et le mode d'action de la pression de sélection appliquée sur des gènes codants pour des protéines. Cette pression de sélection est définie par le rapport entre le taux de substitutions non synonymes et le



taux de substitutions synonymes entre deux séquences, soit  $\omega = dN/dS$ . La construction de ce rapport repose sur les connaissances associées à la génomique de la population. En effet, un des éléments majeurs est que si une mutation apparaît par hasard au sein d'une séquence, deux résultats sont possibles à long terme. D'une part, la mutation peut être perdue, car non transmise à la génération suivante.

D'autre part, la mutation peut être fixée et l'ensemble des individus de l'espèce deviennent à terme porteur de la mutation. Le devenir de la mutation dépend de l'interaction entre la sélection naturelle, la dérive génétique aléatoire, mais aussi de l'effet de la mutation. Globalement, on considère trois effets pour les mutations. Tout d'abord, les mutations peuvent être neutres, c'est-à-dire ayant peu ou pas d'impact sur l'organisme porteur. Les mutations neutres s'accumulent dans la population au rythme du taux de mutation génomique, souvent noté  $\mu$  (Kimura, 1968). Chez les mammifères, il a été estimé à  $1.5e^{-10}$  nucléotide par génération (Kimura, 1968).

Ensuite, elles peuvent être délétères et présenter un désavantage pour l'organisme. Les mutations délétères peuvent être fixées, mais peuvent atteindre au maximum le taux de mutation génomique, tout en s'accumulant dans la population à un taux plus lent puisqu'elles présentent un désavantage évolutif. Pour finir, elles peuvent être avantageuses et présenter un avantage pour l'organisme qui en est porteur. Celles-ci présenteront un taux de mutation supérieur au taux de mutation génomique.

En considérant les mutations qui se produisent au niveau des codons dans les gènes codant pour les protéines, on définit les mutations synonymes comme neutres, car elles ne modifient pas la séquence d'acides aminés de la protéine codée. Ainsi le taux de substitution synonyme ( $\mu S$ ) sera presque égal au taux de mutation génomique  $\mu$ . Parfois, la sélection peut également agir sur des sites synonymes puisque certains codons peuvent être sous-optimaux, ainsi, certaines méthodes et modèles évolutifs modélisent explicitement la sélection de l'usage des codons dans l'estimation de  $\omega$ . En revanche, le taux de substitution non synonyme  $\mu N$  peut être affecté par la sélection. Par conséquent, le rapport  $\omega$  défini tel que  $\omega = \mu N / \mu S$  va être influencé par le mode de sélection agissant sur les sites non synonymes. Néanmoins, l'estimation des taux  $\mu N$  et  $\mu S$  est difficile, ils peuvent être estimés en calculant les divergences non synonymes et synonymes entre les gènes selon les équations suivantes,  $dN = t\mu N$  et  $dS = t\mu S$ . Avec  $t$ , définissant le temps de divergence. Ces distances peuvent être calculées à partir d'un alignement des séquences comparées et permettent d'estimer  $\omega$  telles que  $\omega = dN/dS$ .

Le rapport  $\omega$  peut être utilisé pour décrire le degré et le mode d'action de la « contrainte

sélective » pour un couple de gènes (Hurst, 2002). Ainsi, un rapport  $\omega > 1$  suggère une sélection positive (adaptative ou diversifiant) suggérant que des mutations non synonymes sont intégrées dans la séquence codante aboutissant à une diversification de la séquence (Kryazhimskiy & Plotkin, 2008). Un rapport  $\omega = 1$  indique une évolution neutre avec un ratio de mutation synonyme et non synonyme proche de l’horloge moléculaire (Kryazhimskiy & Plotkin, 2008). Pour finir, un rapport  $\omega < 1$  indique une sélection négative, dite purificatrice, suggérant une pression de sélection contre une modification de séquence, ce qui tend à conserver la séquence identique (Kryazhimskiy & Plotkin, 2008).

Différents outils permettent de calculer  $\omega$ . L’outil de référence depuis plusieurs années est le package Perl PAML (Z. Yang, 2007) qui rassemble de multiples outils permettant d’étudier l’évolution des séquences. Cet outil permet entre autres la comparaison d’arbre phylogénétiques avec BASEML et CODEML, l’estimation des taux de substitution synonymes et non synonymes dans les séquences d’ADN codant pour les protéines (YN00 et CODEML), l’estimation des matrices empiriques de substitution des acides aminés (CODEML) et l’estimation des temps de divergence des espèces dans le cadre de modèles d’horloge moléculaire globale et locale à l’aide de méthodes de vraisemblance (BASEML et CODEML) et bayésiennes (MCMCTREE).

Pour cette analyse, l’outil YN00 a été utilisé. Cet outil se base sur l’analyse des alignements nucléiques afin d’estimer les taux de mutations synonymes et de non-synonymes entre deux séquences d’ADN codant pour des protéines à partir d’un modèle évolutif. Cet outil permet de conceptualiser le processus de substitution pour un site donné via une chaîne de Markov à temps continu avec 61 états possibles, correspondant aux 64 codons auxquels sont soustraits les 3 codons-stop (Z. Yang & Nielsen, 2000). Le modèle de chaîne de Markov qui sous-tend le calcul de  $K_a/K_s$  par PAML ignore explicitement les polymorphismes qui ségrègent au sein d’une population. En effet, le modèle décrit seulement le résultat final de la séquence, mais ne va pas permettre d’étudier en détail le processus par lequel une mutation entre dans une population, change de fréquence et finit par se fixer. Ainsi, le modèle considère les événements de fixation comme se produisant instantanément, et les polymorphismes transitoires au sein de chaque population divergente sont ignorés. Ces estimations vont permettre un calcul robuste tant que les taux de substitution estimés sont calculés entre des espèces ayant divergées avec un temps suffisant. Avec un ensemble de données de séquences divergentes, et en supposant leur relation phylogénétique, PAML estime le paramètre  $\omega$  par maximum de vraisemblance (Z. Yang & Yoder, 2003; Yoder & Yang, 2000). La fonction de vraisemblance est dérivée de la chaîne de

Markov, et suppose que le processus de substitution sur un site est indépendant des processus sur tous les autres sites (B. Shapiro et al., 2006). Par définition,  $\omega$  décrit le taux de mutations sélectionnées par rapport aux fixations de mutations neutres. Par conséquent, il est logique d'estimer  $\omega$  à partir d'un ensemble de données de séquences divergentes, dont les différences représentent des substitutions fixes qui se sont accumulées le long de branches indépendantes (Bofkin & Goldman, 2007).

Cette utilisation de la vraisemblance statistique permet d'intégrer lors du test d'un *locus* les informations issues d'autres *loci*. Cette approche permet une analyse combinée tirant parti des points forts des méthodes basées sur l'utilisation de supermatrices. Les supermatrices sont des matrices interprétables comme ayant été divisées en différents blocs. Ce type de matrice a la capacité à intégrer divers types de données (Dequeiroz & Gatesy, 2007), tout en évitant leurs inconvénients dont la gestion des données manquantes représente le principal problème (Dequeiroz & Gatesy, 2007).

Différentes méthodes de comptage peuvent être intégrées au sein du calcul de PAML afin d'estimer  $K_a$  et  $K_s$  entre deux séquences et de prendre en compte différents phénomènes. Différentes méthodes de comptage sont implémentées au sein de PAML et notamment : NG86 (Nei & Gojobori, 1986), LWL85 (W. H. Li et al., 1985), LPB (Pamilo & Bianchi, 1993), LWL85m, une version modifiée de LWL85 (Z. Yang et al., 2006) et YN00 (Z. Yang & Nielsen, 2000). Ces méthodes diffèrent par des hypothèses et approximations différentes. NG86 fait l'hypothèse qu'il n'y a pas de différence de taux de transition-transversion et pas de biais d'utilisation des codons, il correspond à un modèle de comptage basique. Le calcul des taux est effectué via la formule à un paramètre de Jukes-Cantor (T. H. Jukes et al., 1969) qui permet une correction des substitutions multiples. LWL85, LPB et LWL85m tiennent compte de la différence de taux de transition-transversion, mais ne supposent pas de biais d'utilisation des codons. Cette estimation est valide dans la plupart des cas et en particulier pour les espèces ayant divergées depuis plusieurs millions d'années. La différence des taux de transition-transversion permet de prendre en compte la dégénérescence du code génétique. En effet, pour les sites doublement dégénérés, presque toutes les transitions sont synonymes et presque toutes les transversions sont non synonymes. Néanmoins, on retrouve deux acides aminés qui font exception, qui sont les mutations en deuxième position pour les codons codants pour l'Arginine (CGA, CGG, AGA et AGG) et les troisièmes positions des codons d'Isoleucine (ATT, ATC et ATA). Comme les sites triplement dégénérés sont très peu nombreux dans le tableau universel des codons, si nous considérons les sites triplement dégénérés dans les calculs de  $K_a$  et  $K_s$ , il en résultera des

variations plus importantes des valeurs  $K_a$  et  $K_s$ . Par conséquent, les sites trois fois dégénérés sont considérés comme des sites deux fois dégénérés. LWL85 se base sur la formule de Kimura (Kimura, 1980) et estime  $K_a$  via l'équation 5.2 et  $K_s$  via 5.1. LPB se base aussi sur Kimura (Kimura, 1980) mais propose une correction du biais des sites de comptage en utilisant des formules différentes pour les estimations de  $K_a$  (Équation  $K_a$  5.3 et  $K_s$  5.4). LWL85m est directement dérivé de LWL85.

$$K_s = \frac{L_4 K_4 + L_2 A_2}{L_2/3 + L_4} \quad (5.1)$$

$$K_a = \frac{L_0 K_0 + L_2 B_2}{L_0 + 2L_2/3} \quad (5.2)$$

Avec,

$i$  nombre de sites dégénérés.  $i$  peut prendre les valeurs de 0, 2 et 4

$L_i$  définis comme le nombre de sites  $i$ -fois dégénérés.

$A_i$  et  $B_i$  définissant respectivement les nombres de substitutions transition et transversion par site  $i$ -fois dégénéré

$K_i$  qui est le nombre total de substitutions par site  $i$ -fois dégénéré, c'est-à-dire  $K_i = A_i + B_i$

$$K_a = A_i + \frac{L_0 B_0 + L_2 B_2}{L_0 + L_2} \quad (5.3)$$

$$K_s = B_i + \frac{L_2 A_2 + L_4 A_4}{L_2 + L_4} \quad (5.4)$$

La méthode LPB présente une approche différente. En effet, afin de s'affranchir de l'influence du taux de mutation de transition/transversion, cette méthode va estimer le  $K_a$  et  $K_s$  sans estimer le nombre de site synonymes et non synonymes. Ainsi, cette méthode va approximer que le taux de transition est similaire pour les sites doublement et quadruplement dégénéré et va donc en calculer la moyenne pondéré, ce qui correspond au second membre des Équations 5.3 et 5.4. Pour calculer les nombres moyens attendus  $A_i$  et  $B_i$  définissant respectivement les nombres de substitutions transition et transversion par site  $i$ -fois dégénéré. Ces valeurs correspondent aux probabilités relatives normalisées des différentes voies de substitution de transition et transversion.

Après différentes simulations et comparaisons sur des données réelles, LWL85, LPB et LWL85m ne semblent pas très différents entre eux en termes de résultats (Tzeng et

al., 2004). La méthode LWL85m, qui est une évolution de LWL85 prend en compte une correction supplémentaire pour l'Arginine et présente des valeurs  $K_a$  et  $K_s$  environ 5% plus précises que celles calculées par LWL85 sur des données simulées (Tzeng et al., 2004). Cette correction permet de prendre en compte le fait que les premières positions des codons Arginine (CGA, CGG, AGA et AGG) appartiennent toutes à des sites doublement dégénérés. Afin de prendre en compte certains cas particuliers comme l'alignement de l'un des quatre codons Arginine avec un autre codon qui présente deux ou trois différences. Ces cas peuvent être par exemple retrouvés avec l'alignement des codons CGA (Arg) et ACA (Thr). La règle de base est modifiée pour considérer que toute différence synonyme entre deux séquences sur un site doublement dégénéré est comptée comme une transition, même s'il s'agit en fait d'une transversion.

La méthode la plus aboutie est celle de Yang et Nielsen (Z. Yang & Nielsen, 2000) qui est aussi implémenté dans YN00. Il tient compte à la fois de la différence de taux de transition-transversion, des biais d'utilisation des codons et de la correction pour l'Arginine, permettant ainsi de limiter les biais possibles dans le calcul.

À ces différentes méthodes de comptages sont ajoutées un modèle permettant la modélisation de l'évolution de l'ADN. Différents modèles existent. Le plus basique est le modèle JC69 (T. Jukes & Cantor, 1969). Ce modèle à un paramètre,  $\mu$  estimant le taux de mutation global estime une fréquence des bases équivalente et un taux de mutation égal. Le modèle K80 (dit K2P) à deux paramètres prend en compte le taux de transition ( $\alpha$ ) et transversion (paramétré à 1) (Kimura, 1980). Ce modèle estime une fréquence des bases équivalente. Le modèle K81 (dit K3P) à trois paramètres prends en compte un taux de transition ( $\alpha$ ) et deux taux de transversion ( $\gamma$  et  $\beta$ ) (Kimura, 1981). K3P estime une fréquence des bases équivalente. D'autres modèles existent et notamment celui utilisé par la méthode YN00 dont le calcul est réalisé par estimation du maximum de vraisemblance à l'aide du modèle Hasegawa-Kishino-Yano (HKY85) (Hasegawa et al., 1985). Ce modèle étend le modèle K81 et prend en compte les taux de transitions/transversions à l'aide de  $\alpha$  et  $\beta$  et estime la probabilité d'utilisation des bases comme non équiprobable.

Dans le cadre d'un génome dupliqué par WGD et présentant un profil de dominance génomique, le  $K_a/K_s$  a été étudié chez différentes espèces. Ainsi, chez le maïs, les gènes ohnologues ont été comparés à leurs orthologues chez le sorgho, un organisme proche n'ayant pas eu la dernière WGD. Dans le cadre d'analyses portant sur le génome du maïs où une sous-dominance génomique a été établie, il a été observé une différence significative de la pression de sélection entre les deux sous-génomes (Pophaly & Tellier, 2015).

Ainsi, les gènes dominés semblent avoir une contrainte sélective relâchée et auront tendance à présenter des ratios oméga supérieur à 1 suggérant une pression de sélection négative. Cette pression de sélection aboutit à une accumulation des mutations non synonymes sur le sous-génome dominé aboutissant à une plus forte pseudogénéisation, ce qui participe à la mise en place des mécanismes de la sous-dominance génomique.

## 5.2 Matériel et méthode

Ces analyses nécessitent un génome apparenté dépourvu de la dernière WGD commune aux *Maloidae*. Nous avons donc choisi de réaliser cette étude en nous appuyant sur le génome de *P. persica* (The International Peach Genome Initiative et al., 2013) version 2.0. L'évolution des séquences peut différer entre les paires de gènes ohnologues. Pour tester l'hypothèse d'un biais dans l'évolution des séquences codantes, nous avons utilisé  $K_a/K_s$  comme estimateur de la pression de sélection. Pour évaluer les différences de  $K_a/K_s$  entre séquences codantes ohnologues de la pomme, nous avons d'abord construit des triplets de gènes constitués de deux gènes ohnologues de la pomme et de l'orthologue le plus proche chez *P. persica* puis comparé les ratios *omega* entre les deux paires pommier/poirier. Cette partie a été implémentée dans le pipeline des blocs synténiques décrit dans la section 2.2 relatif à la construction des blocs de synténie et des triplets associés. Les  $K_a/K_s$  de ces triplets ont ensuite été calculés à l'aide d'un pipeline *ad hoc* accessible depuis la forge à l'adresse suivante : <https://forgemia.inra.fr/tanguy.lallemand/ka-ks-snakemake>. Ce pipeline commence par l'alignement des séquences protéiques des triplets en utilisant MUSCLE (Edgar, 2004). L'utilisation de ClustalW (Sievers & Higgins, 2018) est aussi possible par configuration. L'alignement protéique est ensuite converti en alignement nucléaire en utilisant PAL2NAL (Suyama et al., 2006) avec le paramètre *nogap*. Ce paramètre va permettre la suppression de la position pour l'ensemble des séquences si l'une des séquences présente un *gap* sur la position concernée. Ensuite, l'outil YN00 du package PAML (Z. Yang, 2007) est utilisé avec les paramètres suivants : *icode* :0 ; *weighting* :0 ; *commonf3x4* :0. Nous avons calculé les différentes valeurs de  $K_a$ ,  $K_s$  et  $\omega$  pour les cinq méthodes de comptages disponibles (NG86 (Nei & Gojobori, 1986), LWL85 (W. H. Li et al., 1985), LPB (Pamilo & Bianchi, 1993), LWL85m, (Z. Yang et al., 2006) et YN00 (Z. Yang & Nielsen, 2000)). Ces configurations sont gérées par des fichiers de configuration éditables par l'utilisateur. La méthode YN00 a l'avantage de prendre en compte le taux de transition-transversion et les fréquences des nucléotides. L'analyse des résultats de  $K_a/K_s$

a été effectuée seulement sur les résultats de YN00. Ainsi, les différences entre les valeurs de  $K_a$ ,  $K_s$  et  $\omega$  des paires de gènes ohnologues entre fragments de chromosomes synténiques ont été examinées à l'aide du test de Wilcoxon et du test de Kolmogorov-Smirnov à deux échantillons pour la qualité de l'ajustement des valeurs.

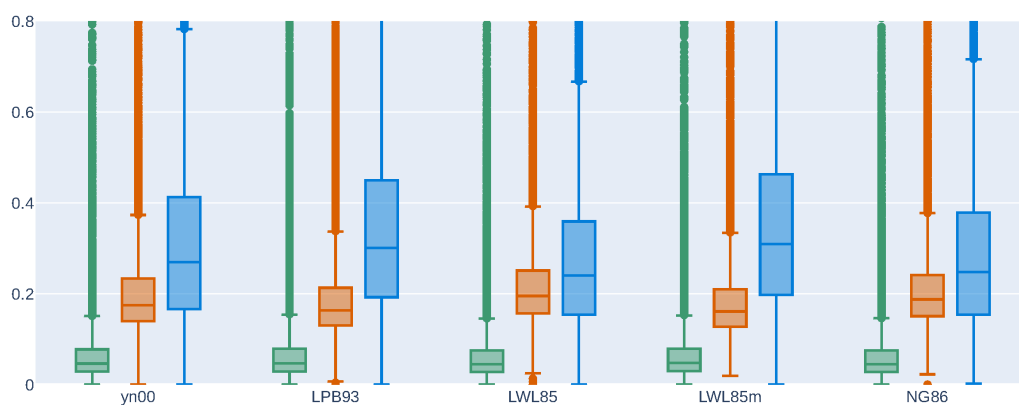
Afin d'estimer la datation de la WGD nous nous sommes appuyés sur une approche d'extrapolation linéaire (Berthelot et al., 2014). En supposant que le taux de substitution est constant au cours du temps, et en ayant une estimation de la date de la spéciation entre le pêcher et le pommier nous avons pu estimer la date de la WGD au moyen des pics de  $K_s$  des distributions des pics de  $K_s$  inter- et intra-spécifique du pommier. La date de spéciation entre *P. persica* et *M. domestica* a été estimée à l'aide de fossiles et de Time-Tree (Kumar et al., 2017). Cet outil permet d'intégrer différents arbres phylogénétiques provenant de la littérature et du contexte géologique et climatique de la Terre afin de construire des modèles d'évolution et de calculer un consensus des dates de spéciation pour les taxons donnés.

L'identification des pics dans les distributions  $K_s$  intraspécifique et interspécifiques a été réalisée automatiquement à l'aide d'une décomposition en ondelettes (Du et al., 2006). Cet algorithme fonctionne via une transformée en ondelettes continues du vecteur de  $K_s$  suivi d'une identification des lignes de crête dans la matrice.

### 5.3 Résultats

En premier lieu, et comme décrit dans le chapitre Préparation des données 2, nous avons cherché à construire des triplets contenant deux ohnologues chez *Malus* et un paralogue chez *Prunus*. Cette construction a été réalisée à l'aide d'un pipeline basé sur un *blastp all-against-all* bidirectionnel. À partir des 16 779 couples de gènes ohnologues, nous avons pu construire 9822 triplets. Les séquences protéiques des gènes composant ces triplets ont par la suite été alignées. Les pourcentages d'identité obtenus varient entre 80 et 100%. La distribution de l'ensemble des pourcentages d'identité est proposée en Figure 2.7. Les alignements protéiques ont ensuite été convertis en alignements nucléiques en supprimant les positions où l'une des séquences présente des *gaps*. La taille médiane des alignements nucléotidiques était de 942 bp avec un écart type de 903 bp. Ces alignements nucléiques permettent le calcul des  $K_a/K_s$  pour 9623 triplets de gènes à l'aide de YN00 et des différentes méthodes de comptages.

Les distributions des valeurs de  $\omega$  calculées à l'aide des différentes méthodes de comp-



**Figure 5.1** — Distribution des valeurs de  $K_a$  (vert),  $K_s$  (orange) et  $\omega$  (bleu) pour les différentes méthodes de comptages sous forme de boîte à moustaches. L'axe des ordonnées présente les valeurs de  $K_a$ ,  $K_s$  et  $\omega$ . Sur les axes des abscisses cinq groupes de boîtes à moustache sont présentés. Le bas de la boîte représente le premier quartile, le trait interne à la boîte présente la médiane de la distribution et l'extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l'écart interquartile. La moustache inférieure présente le neuvième décile et la moustache présente quant à elle le 1<sup>er</sup>. Chaque groupe de boîtes à moustaches est associé aux résultats calculés via une méthode dont le nom est précisé en abscisse. On peut observer que visuellement les résultats sont assez similaires, avec LWL85 et LWL85m qui génère les distributions extrêmes et YN00 qui génère la distribution la plus consensuelle.



tage sont présentées en Figure 5.1. Une description chiffrée des distributions de  $K_a$ ,  $K_s$  et  $\omega$  est présenté dans la Table 5.1. Concernant le  $K_a$ , les valeurs issues des différentes méthodes sont très similaires avec des valeurs médianes présentant un minimum à 0,045 obtenu via la méthode LWL85 et un maximum à 0,048 obtenu avec LWL85m (écart type=0,001). Le  $K_s$  présente des valeurs médianes minimales de 0,16 (LWL85m) et maximales de 0,19 obtenue avec LWL85 pour une erreur standard de 0,01. Pour finir  $\omega$ , présente un minimum à 0,24 calculé avec LWL85 et 0,30 au maximum calculé avec LWL85m (écart type=0,03). Si l'on s'intéresse à la valeur médiane des médianes des distributions calculées avec chacune des méthodes, YN00 est la valeur médiane pour  $K_a$ ,  $K_s$  et  $\omega$ . Ces résultats montrent que globalement les méthodes de comptages présentent des valeurs similaires avec LWL85 et LWL85m générant les valeurs extrêmes et YN00 qui est un résultat médian des cinq méthodes. Par ailleurs cette méthode est l'une des plus utilisées dans la littérature (Hurst, 2002 ; Tzeng et al., 2004 ; Z. Zhang & Yu, 2006). Ainsi, pour les analyses suivantes, seules, les données issues du calcul avec YN00 sont considérées.

**Table 5.1** – Description des valeurs de  $\omega$ ,  $K_a$  et  $K_s$  calculés avec les différentes méthodes de comptages implémentées dans PAML.  $K_a$  fournis des valeurs similaires avec un minimum de médiane à 0,045 (LWL85) et un maximum à 0,048 (LWL85m). Le  $K_s$  présente des valeurs médianes minimales de 0,16 (LWL85m) et maximales de 0,19 (LWL85).  $\omega$  présente un minimum à 0,24 (LWL85) et 0,30 au maximum calculé avec LWL85m.

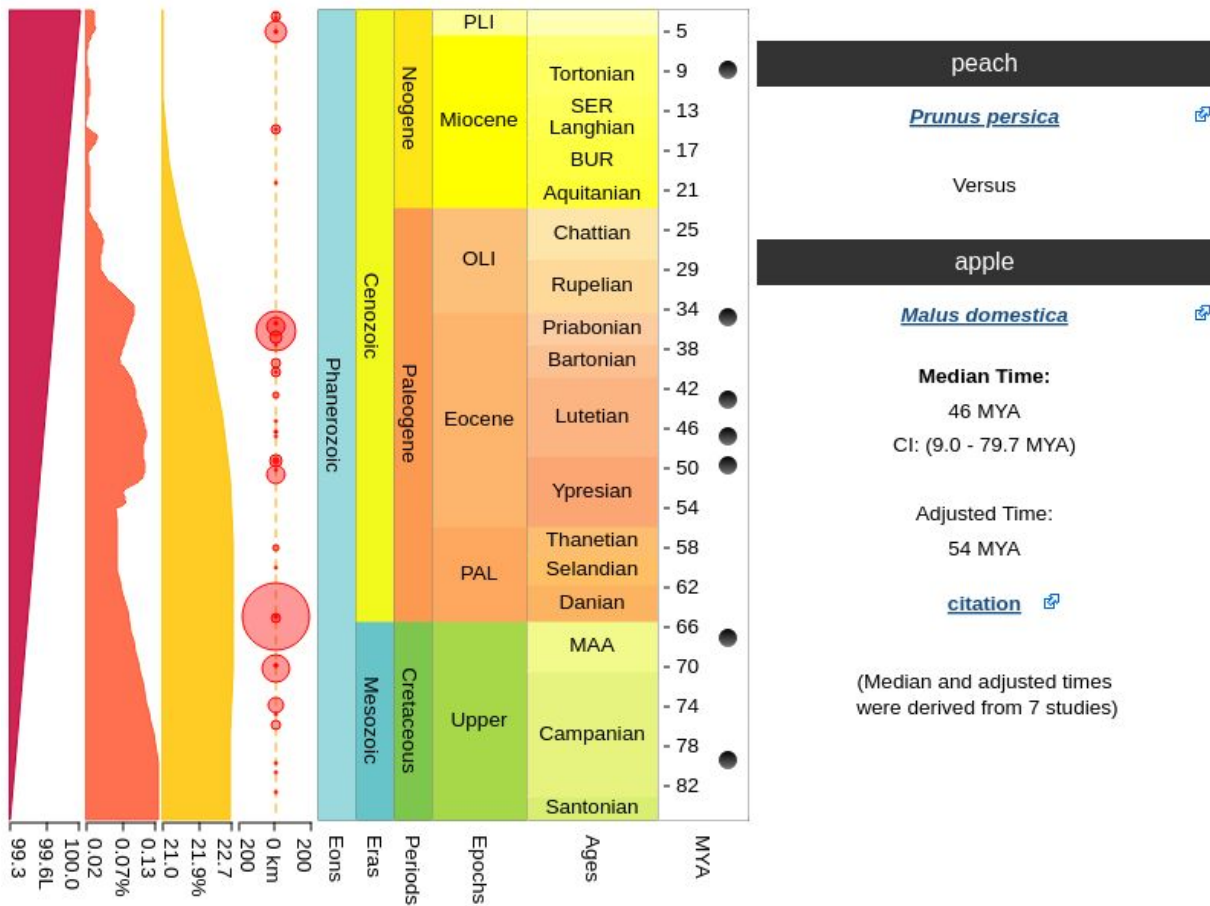
Moyenne	Écart type	Médiane	Valeur	Modèle évolutif
0,3444	0,2328	0,3011	$\omega$	LPB93
0,0784	0,1043	0,0470	$K_a$	LPB93
0,2199	0,2785	0,1630	$K_s$	LPB93
0,2802	0,1926	0,2402	$\omega$	LWL85
0,0758	0,1027	0,0450	$K_a$	LWL85
0,2562	0,3665	0,1950	$K_s$	LWL85
0,3539	0,2237	0,3096	$\omega$	LWL85m
0,0779	0,1007	0,0480	$K_a$	LWL85m
0,2202	0,3352	0,1610	$K_s$	LWL85m
0,2783	0,2371	0,2480	$\omega$	NG86
0,0756	0,1019	0,0451	$K_a$	NG86
0,2455	0,2799	0,1874	$K_s$	NG86
0,3450	1,7381	0,2696	$\omega$	yn00
0,0776	0,1048	0,0463	$K_a$	yn00
0,2604	0,3708	0,1749	$K_s$	yn00

À partir des résultats issus des calculs à l'aide du modèle YN00, les distributions

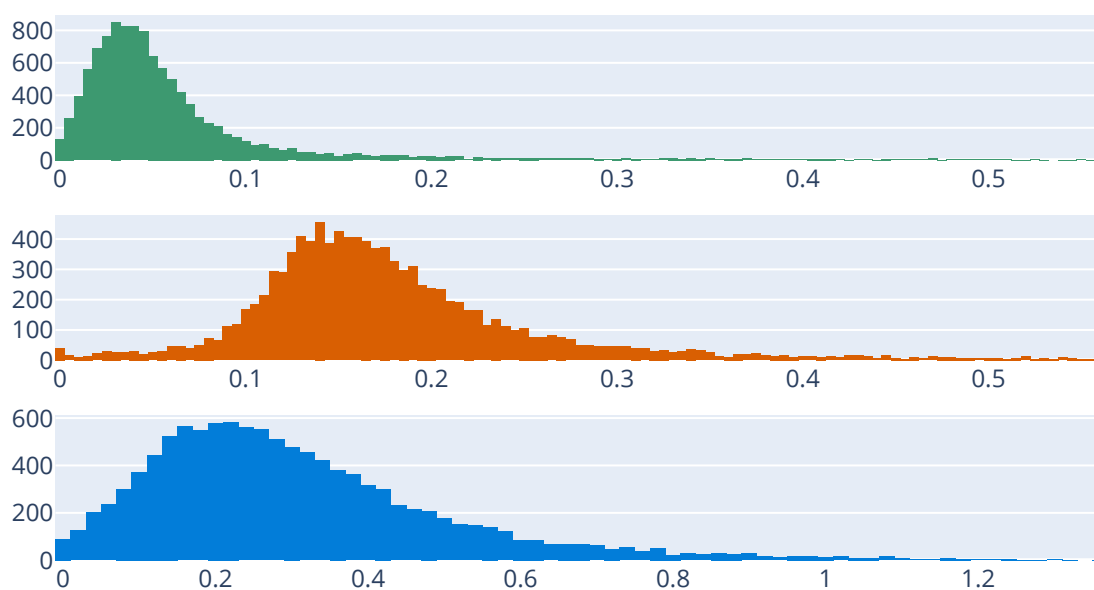
de  $K_a$ ,  $K_s$  et  $K_a/K_s$  sont générées et présentées dans la figure 5.3. La distribution du  $K_s$  intraspécifique présente un seul pic, ce qui suggère la présence d'une seule et récente WGD. À partir de l'identification des valeurs de pics intraspécifique et interspécifique et par extrapolation, nous avons pu estimer la date de la WGD. La spéciation entre *P. persica* et *M. domestica* a été datée à partir de TimeTree dont une visualisation du résultat est présentée en Figure 5.2. À l'aide d'un ensemble d'arbres phylogénétiques issus de la littérature (Davies et al., 2013 ; C.-H. Huang et al., 2016 ; Naumann et al., 2013 ; Pouget et al., 2016 ; Töpel et al., 2012 ; Vanneste et al., 2014) et de preuves fossiles (Wolfe & Wehr, 1988) l'outil permet d'estimer un âge corrigé de la spéciation fixé ici à 54 Ma. La valeur des pics  $K_s$  dans les distributions intraspécifiques a été identifiée à 0,151 et interspécifiques à 0,293. La date de spéciation et les pics dans les distributions des  $K_s$  permet d'extrapoler la datation de la WGD à 27 Ma.

La distribution de  $\omega$  présente une majorité des valeurs en dessous de 1 ce qui suggère une sélection purificatrice des gènes ohnologues. Néanmoins, nous avons identifié 222 couples de gènes avec des ratios  $\omega$  supérieurs à 1 indiquant une sélection positive. Leur distribution suit la répartition de la taille des paires de fragments chromosomiques synténiques, et aucune paire de chromosomes ne semble déséquilibrée.

Afin d'étudier un potentiel déséquilibre dans la pression de sélection appliquée sur les séquences ohnologues, des tests de Wilcoxon ont été exécutés sur les valeurs de  $\omega$ , de  $K_a$  et de  $K_s$  des paires de fragments synténiques. La distribution des ratios de  $\omega$  pour chaque chromosome, regroupé par paires de segments synténiques, est représenté dans la Figure 5.4. Les résultats des tests associés aux valeurs de  $\omega$  sont placés en Table 5.2. Le tableau et la série de boîtes à moustaches montrent qu'il n'y a pas de différence significative dans la pression de sélection des séquences codantes entre les paires de chromosomes ohnologues. Néanmoins, les paires 1-7, 2-7, 2-15 et 5-10 semblent montrer une différence, bien que non significative au seuil fixé de  $\alpha$  à 5%. En s'appuyant sur les moyennes des distributions, dans l'exemple de la paire 1-7, le fragment synténique associé au chromosome 1 aurait une pression de sélection plus forte que côté chromosome 7. Pour la paire 2-7, le fragment synténique associé au chromosome 7 serait le plus conservé. Pour la paire 2-15, le chromosome le plus conservé serait le chromosome 2. Pour la paire 5-10 le chromosome 5 présenterait une pression de sélection plus importante. Par ailleurs des tests de Kolmogorov-Smirnov à deux échantillons pour la qualité de l'ajustement ont permis de tester si les distributions de  $\omega$  étaient similaires entre les paires de fragments chromosomiques synténiques. Les résultats sont présentés en Table 5.2. Ainsi, nous n'avons pas

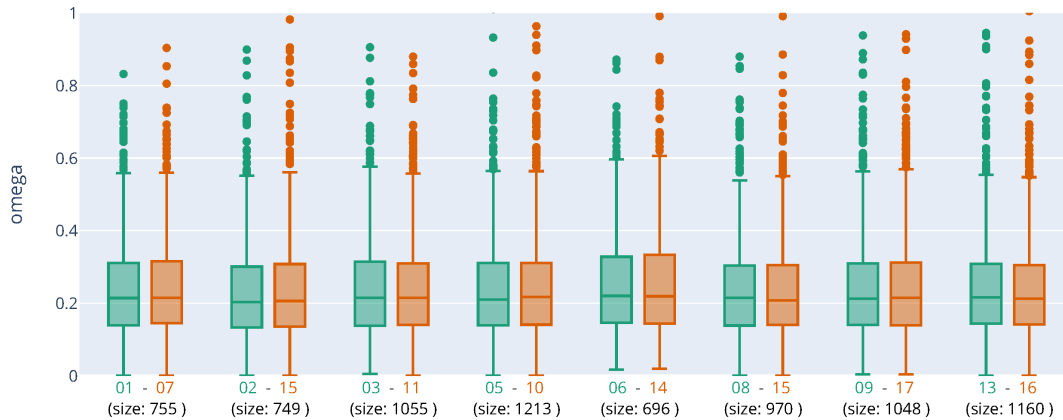


**Figure 5.2** – Estimation de la datation de la spéciation de *P. persica* et *M. domestica*. Cette visualisation a été construite à l’aide de TimeTree. L’axe des ordonnées présente le temps en millions d’années. Sur l’axe des abscisses un ensemble de sous-figures sont présentes. La première partie de la Figure présente des éléments climatiques et géologiques importants. Ainsi la première courbe rouge, présente l’évolution de la luminosité solaire à travers le temps. La seconde, en orange, présente le niveau de CO<sub>2</sub>. La courbe jaune présente le niveau d’O<sub>2</sub>. La Figure suivante présente un ensemble de cercles qui représentent les impacts météoriques importants. Suit alors un ensemble de frises géologiques rappelant les éons, ères, périodes, époques et âges géologiques. Le long de l’axe des ordonnées, des cercles présentent des estimations de la datation de la spéciation entre le pêcher et le pommier.



**Figure 5.3** – Distribution des valeurs estimées de pression de sélection appliquée aux séquences codantes. L’axe des ordonnées (*omega*) présente le nombre de séquences présentant une valeur particulière. L’axe des abscisses présente les valeurs de  $K_a$ ,  $K_s$  ou  $\omega$ . La distribution verte présente les valeurs  $K_a$  intraspécifiques. La distribution orange les valeurs de  $K_s$  intraspécifiques. La distribution bleu les valeurs d’ $\omega$  intraspécifiques. La distribution de  $K_s$  présente un seul pic suggérant une seule WGD. La distribution de  $K_a/K_s$  présente une majorité des valeurs inférieures à 1 suggérant une pression de sélection purificatrice.

identifier de différences significatives entre les distributions de  $\omega$  des paires fragments chromosomiques synténiques.



**Figure 5.4** – Diagrammes en boîte des valeurs d' $\omega$  entre les paires de fragments synténiques chromosomiques. L'axe des y représente l'ensemble des valeurs d' $\omega$ . L'axe des abscisses indique les paires de chromosomes considérées. Pour chaque paire de chromosomes (ou segments de chromosomes), les diagrammes en boîte indiquent la distribution sur le premier chromosome de la paire (en vert) et la distribution sur le second chromosome de la paire (en orange). \* $\alpha < 0,05$

Nous avons également comparé les ratios de mutations synonymes ( $K_s$ ) et les ratios non synonymes ( $K_a$ ). Les distributions des  $K_a$  représentées sous forme d'une série de boîtes à moustaches pour chaque chromosome et regroupées par paires de segments synténiques sont présentées dans la Figure A.11. Nous avons pu identifier les paires 2-15 et 8-15 comme significativement déséquilibrées. Quant aux ratios non synonymes ( $K_a$ ) représentés en Figure A.12, la paire 8-15 est significativement déséquilibrée. Le détail des tests de Wilcoxon associés aux  $K_a$  et  $K_s$  est présenté en Table A.1. Pour finir, les distributions ont été testées via Kolmogorov-Smirnov (Table A.1) et montrent que l'on ne peut pas rejeter l'hypothèse que les distributions sont similaires.

Des tests similaires ont été menés sur des ensembles de gènes filtrés à partir d'informations provenant d'autres analyses menées au cours de cette thèse et notamment des ensembles de gènes particuliers identifiés lors des analyses de transcriptomes et d'éléments transposables. Les résultats étaient similaires avec des différences visibles en termes de moyennes, mais non significatives.

**Table 5.2** – Résultats des tests de Wilcoxon et des tests de Kolmogorov-Smirnov à deux échantillons pour la qualité de l’ajustement des valeurs de  $\omega$  entre les fragments synténiques

couple	p-value de Wilcoxon	p-value de Kolmogorov-Smirnov	Moyenne du premier chromosome	Moyenne du second chromosome
01-07	0,1902	0,7622	0,2373	0,2422
02-07	0,1709	0,6776	0,2578	0,2566
02-15	0,1408	0,6722	0,2330	0,2397
03-11	0,1935	0,7528	0,2420	0,2383
05-10	0,0832	0,8789	0,2381	0,2412
06-14	0,3475	0,6719	0,2475	0,2523
08-15	0,4853	0,5164	0,2366	0,2348
09-17	0,4343	0,9625	0,2410	0,2428
13-16	0,4471	0,9505	0,2388	0,2378

## 5.4 Discussion

Les calculs de pression de sélection appliquée aux séquences des gènes ohnologues ont été réalisés à partir de triplets composés de deux gènes ohnologues du pommier et d’un orthologue du pêcher. Chez *A. thaliana*, il a été observé que l’utilisation de triplets permet une bonne résolution dans l’estimation de  $K_a/K_s$  associés à différentes WGD (Ren et al., 2018). Les auteurs montrent que triplets obtenus par RBBH (consistant à garder les triplets de gènes présentant réciproquement le meilleur résultat en blastP sur le génome complet) permettent une meilleure résolution que l’utilisation de triplets construits via une approche de BBH (consistant à garder les triplets de gènes présentant le meilleur résultat en blastP sur le génome complet sans relation de réciprocité). Cette étude a également montrée que l’utilisation d’un arbre phylogénétique et d’un plus grand nombre d’espèces aurait permis une meilleure résolution surtout dans le cas où plusieurs WGD ont été étudiés. Dans le cas de l’étude d’une unique et récente WGD comme chez le pommier, l’utilisation de RBBH *Blastp all-against-all* permet une bonne résolution et un calcul robuste de la pression de sélection.

Les distributions de valeurs de  $K_a$ ,  $K_s$  et  $\omega$  estimées avec les différentes méthodes de comptages sont présentées sous la forme de boîtes à moustaches en Figure A.8. On peut constater que les distributions sont globalement proches. D’un point de vue théorique, la méthode NG sous-estime souvent le rapport des taux de transition/transversion ( $\chi$ ) et donc le nombre de sites synonymes (S), ce qui entraîne une surestimation de  $K_s$  et

une sous-estimation de  $\omega$ . En effet, les transitions ont été observées comme pouvant être jusqu'à deux fois plus fréquentes que les transversions (Collins & Jukes, 1994). Cette sous-estimation a été confirmée par des simulations (Z. Zhang & Yu, 2006) et des observations sur des données réelles (Z. Yang & Nielsen, 2000). Les méthodes LWL85 et LWL85m génèrent des distributions extrêmes. Ce constat peut s'expliquer par le fait qu'ils ne tiennent pas compte de biais d'utilisation de codons, mais que LWL85m dérivé de LWL85 prend en charge une correction liée à l'Arginine. La méthode YN00 adopte des corrections des principaux biais connus et notamment le biais du taux de transition/transversion ( $\kappa$ ) et le biais de la fréquence des bases. De plus, cette méthode apparaît comme un résultat médian des cinq méthodes en termes de distribution des valeurs. De plus, cette méthode est la plus utilisée dans la littérature (Kreplak et al., 2019; Mao, 2019; Sohpal, 2021; J. Yang et al., 2016). C'est pourquoi la méthode YN00 a été choisi pour les analyses suivantes.

L'estimation de l'âge de divergence entre les paralogues a souvent été réalisée par l'intermédiaire du calcul du taux de substitution synonyme lorsque de nombreux paralogues présentent des valeurs de  $K_s$  dans un intervalle proche (Blanc & Wolfe, 2004; Jiao et al., 2011). Un unique pic est retrouvé dans la distribution du  $k_s$  intraspécifique, ce qui suggère une seule WGD récente comme attendu vis-à-vis de la littérature existante (Daccord et al., 2017; Velasco et al., 2010). La localisation des pics dans les distributions de  $k_s$  est une étape cruciale. Elle a été réalisée via un algorithme de décomposition par ondelette. Cet algorithme détecte les pics en appliquant une correspondance de motifs dans la matrice de valeur, récupérant ainsi des informations supplémentaires comme la forme des pics, ce qui peut améliorer considérablement la détection. Cette méthode peut ainsi détecter à la fois les pics importants et faibles tout en présentant des résultats sensibles avec un taux faible d'erreur de première espèce, comme le montrent les résultats de l'évaluation comparative (Du et al., 2006). Après extrapolation linéaire à partir des pics de  $k_s$ , l'estimation de la datation de la WGD a été faite. Celle-ci est cohérente avec la dernière datation faite (Su et al., 2021) qui confirme une WGD plus récente que les datations précédentes faites via le pommier (Velasco et al., 2010; Xiang et al., 2017) et le poirier (J. Wu et al., 2013). La datation de la WGD a été affinée grâce au génome de haute qualité GDDH13, mais aussi grâce aux séquençages d'espèces supplémentaires proches des *Maloideae*. Ainsi, le séquençage en 2021 de *Gillenia trifoliata* (Su et al., 2021) et du néflier japonais (Su et al., 2021), des Rosacées à 9 chromosomes plus proches des *Maloideae* que *P. persica* a permis une nouvelle datation de la WGD des *Maloideae* entre 13,5 et 27,1 Millions d'an-

nées (Su et al., 2021). Cette estimation a été faite au moyen d'un arbre phylogénétique de 8 espèces de Rosacées et d'un algorithme basé sur une Markov chain Monte Carlo (MCMC). Par extrapolation linéaire et seulement à l'aide de *P. persica* nous avons pu faire une estimation similaire.

On peut observer que la plupart des valeurs de  $\omega$  sont inférieures à 1 avec des médianes des distributions  $\omega$  pour chaque chromosome d'environ 0,20. Ce résultat suggère une sélection purificatrice des gènes ohnologues sur tous les chromosomes. Les espèces dupliquées par WGD sont attendues comme présentant une sélection purificatrice (Cheng et al., 2018; Harikrishnan et al., 2015; Pophaly & Tellier, 2015). De plus, les organismes ayant subi une WGD récente comme le poirier (Q. Li et al., 2019), le maïs (Pophaly & Tellier, 2015), le saule (Harikrishnan et al., 2015) ou la carpe (J.-T. Li et al., 2015) présentent des valeurs de  $\omega$  dans les mêmes ordres de grandeur.

Néanmoins, de par l'approche utilisée pour construire les triplets pommier-pêcher, nous n'avons étudié que des gènes très similaires ayant été conservés au cours de l'évolution. En effet, les pourcentages d'identités des alignements de triplets, compris entre 80 et 100 % montrent que l'on a analysé avant tout des protéines proches d'un point de vue de la séquence protéique. La perte de 199 triplets au cours de l'exécution du pipeline s'explique par le fait que pour certains triplets l'alignement n'a pas fourni des séquences assez longues rendant le calcul du  $K_a/K_s$  impossible, en effet celui-ci nécessite un nombre suffisant de mutations synonymes et non synonymes. Or, PAL2NAL supprime tous les codons contenant un *gap* dans l'alignement, ce qui peut raccourcir beaucoup des séquences qui ne s'alignent pas bien. Pourtant ces séquences sont les plus divergentes et sont peut-être en partie responsables de la différence entre les segments ohnologues. Les pseudogènes n'ont pas été non plus inclus dans cette analyse, car ils ne sont souvent pas ou seulement partiellement détectés par les logiciels de prédiction de gènes, or ils peuvent être en partie à l'origine des divergences observées entre paires de chromosomes ohnologues.

Afin d'expliquer le déséquilibre de QTLs et étudier plus précisément le possible mécanisme de sous-dominance génomique, nous avons testé la présence d'un biais dans la pression sélective entre les gènes ohnologues. Nous nous sommes appuyés sur le rapport  $K_a/K_s$  entre les gènes ohnologues du pommier par rapport aux orthologues du pêcher. Les différences de valeurs de  $\omega$  ont été testées à l'aide du test de Wilcoxon. Les résultats montrent que le taux d'évolution des séquences codantes ohnologues n'est pas significativement différent pour toutes les paires testées. Des analyses sur le taux de mutations synonymes et non-synonymes présentent des résultats similaires. Ainsi, il semblerait que



les gènes maintenus après le premier cycle de pseudogénéisation post-WGD sont soumis à une pression de sélection similaire. Ce constat est valide, quel que soit le degré de réarrangement chromosomique des paires considérées. Ce résultat ne peut donc pas expliquer le déséquilibre de QTL qui ne semble donc pas lié à des différences de séquences codantes entre les gènes ohnologues. Chez les espèces présentant des déséquilibres de QTL et des sous-dominances génomiques, il a été montré un lien entre le déséquilibre de QTL et la pression de sélection. Par exemple chez le maïs, il a été montré que le gène d'une paire donnée qui présente un phénotype mutant et donc porteur du gène majeur présente un ratio  $\omega$  plus faible, ce qui indique qu'une sélection purificatrice plus importante s'exerce sur lui (Pophaly & Tellier, 2015).

Pour le pommier nous n'avons pas observé de différences significatives. Il est possible que cette absence de différence s'explique par la stringence de la méthode de construction des triplets que l'on a utilisé qui nous a permis de regrouper que les gènes les plus similaires, qui ont donc accumulé moins de différences. L'intégration de triplets plus divergents pourrait peut-être amener des différences plus fortes et potentiellement déséquilibrées. Néanmoins, ces différences, non significatives, pourraient présenter un intérêt notamment en les intégrant à d'autres sources d'informations.

Ainsi, les déséquilibres de QTL observés ne peuvent pas être expliqués par des différences globales dans l'évolution des séquences codantes des gènes. Des résultats similaires ont été observés chez différentes espèces, y compris des espèces à dominance sous-génomique comme le maïs (Woodhouse et al., 2010).

## 5.5 Conclusion

Dans cette analyse, 9623 triplets de gènes composés de deux gènes ohnologues pommier et de leur orthologue chez le pêcher ont été testés. Après calcul des valeurs de  $K_a$ ,  $K_s$  et  $\omega$  associés à ces séquences via cinq différentes méthodes de comptages. Nous avons pu estimer que la méthode YN00 présente les résultats les plus consensuels. En comparant la pression de sélection appliquée aux gènes ohnologues entre les différentes paires synténiques, nous n'avons pas identifié de déséquilibre significatif entre les principales paires de fragments chromosomiques synténiques. Les différences de moyennes des distributions de  $\omega$  bien que non significatives suggèrent un déséquilibre entre les paires de chromosomes ohnologues qui serait en adéquation avec ce qui a été observé lors de l'évaluation du fractionnement génomique et de l'étude de la distribution des QTLs. Néanmoins, les

différences de pressions de sélections ne semblent pas expliquer de manière significative les déséquilibres observés. Ainsi, il convient de s'intéresser à des mécanismes dont l'évolution est suspectée comme étant plus rapide, et notamment la transcription des gènes dont le déséquilibre a été associé à la sous génome dominance et pourrait expliquer le déséquilibre de fractionnement et de distribution de QTL.



# ANALYSE DE L'EXPRESSION DES GÈNES OHNOLOGUES

---

Nous avons pu mettre en évidence un déséquilibre dans la proportion de QTL portés par les différentes paires de fragments chromosomiques ohnologues. Ce déséquilibre ne s'explique pas par des différences de pression de sélection entre les séquences des gènes ohnologues, comme discuté dans le chapitre 5. Afin d'expliquer le déséquilibre de QTL, nous avons choisi de nous intéresser à la comparaison du niveau de transcription des gènes ohnologues. En effet, le niveau de transcription est susceptible d'évoluer plus rapidement que les séquences codant pour des protéines et pourrait ainsi avoir un lien avec le différentiel de QTL pour différentes paires de chromosomes. Par ailleurs, nous avons pu identifier un biais dans le fractionnement du génome qui tend à privilégier certains chromosomes. Ainsi nous avons pu observer que le pourcentage de rétention de gènes entre le pommier et le pêcher est significativement différents pour différentes paires. Le biais de fractionnement génomique est un mécanisme connu pour impacter l'équilibre transcriptionnel des organismes dupliqués par WGD.

## 6.1 Introduction

Une duplication complète du génome par WGD présente de très nombreuses implications pour l'organisme. En effet, le génome est rendu instable par une augmentation importante de sa taille et du nombre de gènes. Ainsi, après une WGD, un ensemble de mécanismes, notamment des remaniements chromosomiques et des pertes de gènes, vont se mettre en place afin de rétablir la ploïdie de l'organisme (Wendel, 2015). Les mécanismes de diploïdisation vont amener, dans certains cas, des déséquilibres dans le génome. Chez de nombreuses espèces polyploïdes (allopolyploïdes et autopolyploïdes), il a notamment été observé des déséquilibres dans les niveaux de transcription entre les différents segments synténiques. Ce déséquilibre transcriptionnel a été étudié chez de nombreuses

espèces. Il a été montré que l'un des mécanismes liés à la mise en place de ce déséquilibre est le fractionnement du génome. Ces mécanismes s'initient rapidement après la WGD et ont été retrouvés chez des organismes avec des WGD très récentes comme *Brassica napus* (Xiong et al., 2011). Il semble que ces effets perdurent de nombreuses générations après la WGD puisque l'on a retrouvé ce mécanisme chez des organismes avec des WGD plus anciennes comme *A. thaliana* (Maere et al., 2005) et *Z. maize* (Woodhouse et al., 2010). La perte de gène est théoriquement attendue comme aléatoire et ne privilégie pas de chromosomes ou de sous-génome en particulier (Panchy et al., 2016). Néanmoins, il a été constaté chez différentes espèces et notamment *M. domestica* dans ce manuscrit (chapitre 3) que le fractionnement génomique pouvait être biaisé.

Cette rétention préférentielle des gènes dupliqués par WGD sur certaines parties du génome va avoir différents impacts, entre autres, sur le niveau de transcription. Ainsi, le génome ayant perdu le moins de gènes (dit sous fractionné) a été observé comme étant le sous-génome avec une expression en général plus importante par rapport au sous-génome plus fractionné. Cet ensemble de mécanismes de fractionnement génomique associé à un déséquilibre transcriptionnel a été décrit dans la littérature par le terme de dominance sous-génomique (*subgenome dominance*). Il est à noter que le biais d'expression induit n'est pas total et constitue un modèle valide à l'échelle globale. En effet, localement, le sous-génome dit sous exprimé contient toujours des gènes qui sont exprimés. Une partie même des gènes du sous génome « dominé » peuvent être surexprimés par rapport à leurs ohnologues issus du sous génome dit dominant. Par ailleurs, il a été identifié pour certaines espèces, et en particulier le blé, une sous-dominance génomique qui ne fonctionne pas à l'échelle globale, mais dont la présence a été identifiée au niveau local (Harper et al., 2016). Néanmoins, l'observation la plus généralement faite, et uniquement chez plusieurs espèces allopolyploïdes, est un déséquilibre transcriptionnel global. Ainsi, chez *B. rapa*, de nombreuses études ont analysé les niveaux de transcription des gènes synténiques.

Au sein de *B. rapa*, dans 3 expériences RNA-Seq associées à trois tissus, il a été observé que les trois sous-génomes (LF pour *Less Fractioned subgenome*, MF1 *More Fractioned subgenome 1* et MF2 *More Fractioned subgenome 1*) présentaient des niveaux de transcriptions des gènes synténiques différents, et que l'un des sous-génomes (LF) présentait significativement plus souvent des niveaux d'expression supérieurs aux autres sous-génomes. Chez le maïs, à partir de 4 échantillons RNA-Seq et 1750 paires de gènes synténiques, il a été observé qu'une majorité des gènes surexprimés avec un ratio  $\log_2 > 4$  étaient significativement (distributions binomiales cumulatives) préférentiellement locali-

sés sur le sous-génome 1 plutôt que le sous-génome 2 (J. C. Schnable et al., 2011). Ce différentiel d'expression a été observé dans l'ensemble des tissus testés. Par ailleurs, des cas plus complexes de dominance sous-génomiques avec une dimension spatiotemporelle ont été observés notamment chez la myrtille, le blé hexaploïde (blé tendre) (Eckardt, 2014) et le coton. Ainsi chez la myrtille allotétraploïde, l'un des sous-génomes présente des niveaux d'expression plus importants que ses homologues dans la majorité des organes et des stades de développement testés. Néanmoins, le second sous-génome a été identifié comme dominant lors du développement du fruit (Colle et al., 2019). De même, chez le coton, des dominances sous-génomiques variables ont été identifiées à différents stades de développement et dans différents tissus. Ainsi, le sous-génome D a été identifié comme étant dominant et plus fortement exprimé dans les pétales (L. Flagel et al., 2008). Tandis que le sous-génome A a été observé comme le plus fortement exprimé dans les tissus ovulaires (Samuel Yang et al., 2006). Ces biais locaux suggèrent qu'un sous-ensemble de voies métaboliques pourrait être contrôlé par un sous-génome en particulier, tandis que l'autre ou les autres sous-génomes contrôlent le reste des voies. Cela entraînerait alors le partage des traits phénotypiques et des QTLs entre différents sous-génomes, comme il a été observé pour la myrtille (Colle et al., 2019), le coton (L. Flagel et al., 2008) et le blé (Eckardt, 2014). Cette observation pourrait être une source de lien important entre un déséquilibre de QTL observé au sein du pommier et un potentiel déséquilibre transcriptionnel dont ce chapitre détaille la mise en place et les résultats associés.

Par ailleurs, un certain nombre d'organismes allopolyploïdes ne présentent pas de sous-génome dominant identifié. C'est le cas d'organismes de la famille des cucurbitacées (H. Sun et al., 2017) et *Capsella bursa-pastoris* (Douglas et al., 2015). De plus, il n'a pas encore été identifié de déséquilibre transcriptionnel chez des espèces dupliquées par autopolyploïdie. En effet, ce type d'analyse a été menée entre autres chez le poirier (Q. Li et al., 2019) où il n'a pas démontré de déséquilibre. Enfin, l'émergence d'un déséquilibre transcriptionnel après polyploïdisation semble rapide. En effet, après des tests sur des allopolyploïdes synthétiques notamment du coton, il a été démontré que le biais d'expression pouvait être observé dès l'hybride F1 (L. Flagel et al., 2008). Le biais s'amplifie par la suite au cours des générations suivantes avant de se stabiliser (L. Flagel et al., 2008). Une des hypothèses pouvant expliquer ce mécanisme pourrait être un facteur environnemental qui pourrait influencer le biais pour sélectionner le sous-génome qui serait le plus adapté à l'environnement.

Afin d'expliquer le déséquilibre de QTL, qui ne semble dû ni à un différentiel de

sélection des gènes ni à un déséquilibre de fractionnement du génome, nous nous attachons dans le chapitre suivant à tester si un déséquilibre transcriptionnel pouvait être identifié chez le pommier. Pour ce faire nous avons mené une analyse exhaustive de l’expression des gènes ohnologues chez le pommier à partir de toutes les données RNA-seq publiques disponibles, générées sur une large diversité de tissus et de conditions.

## 6.2 Matériel et méthode

Afin de tester un déséquilibre transcriptionnel en sachant que ce type de déséquilibre pourrait s’appliquer à l’échelle locale, globale ou présenter des profils différents suivants les tissus et/ou conditions testés, nous avons étudié l’expression des gènes dans un maximum d’organes et de traitements. Ainsi, nous nous sommes appuyés sur les données accessibles publiquement. La première étape de l’analyse est donc d’identifier et de télécharger les données brutes de RNA-Seq associées à *M. domestica* sur les différentes bases de données publiques.

### 6.2.1 Récupération des données de séquençage RNA-Seq

Afin de mener une analyse exhaustive de l’expression des gènes ohnologues chez le pommier, nous avons cherché à récupérer un maximum de données de séquençage RNA-Seq de haute qualité. Pour ce faire, nous avons construit un pipeline Snakemake (pipeline *get-rna-seq-data*) (<https://forgemia.inra.fr/tanguy.lallemand/get-rna-seq-data>) permettant la recherche d’expériences transcriptomiques, leur annotation afin de les filtrer suivant la qualité désirée, et le téléchargement des données brutes associées. Ce pipeline permet la recherche d’identifiants d’études (SRP), d’expériences (SRX) et d’échantillons (SRR) ainsi que l’ensemble des métadonnées associées à chacun des échantillons. La recherche est faite sur les principales bases de données publiques rassemblant des données transcriptomiques et notamment, Sequence Read Archive (SRA) (Leinonen et al., 2011), European Nucleotide Archive (ENA) (Toribio et al., 2017) et Gene Expression Omnibus (GEO) (Barrett et al., 2012).

L’interrogation de ces bases de données s’appuie sur la librairie pySRADB (Choudhary, 2019) en version 0.11.1 qui permet la récupération des identifiants d’expérience transcriptomique ainsi qu’une partie des nombreuses métadonnées associées. Les métadonnées ne pouvant être récupérées à l’aide de cette librairie ont été récupérées par des analyses pro-

grammatiques des pages web associées. C'est par cette approche qu'ont été récupérés la taille des fichiers téléchargés afin de fournir une estimation fiable de l'espace disque requis ainsi que la longueur des *reads*.

Afin de garantir une bonne homogénéité de qualité du jeu de données, un filtre suivant des critères de tissus, traitements, organisme, technologie utilisée, mais aussi de qualité des échantillons et/ou du séquençage peuvent être précisés par l'utilisateur. Pour cette analyse, nous avons gardé des données issues de *M. domestica* et respectant les critères suivants : données pairées ; au moins trois réplicats biologiques ; un séquençage d'au moins  $2 \times 10^7$  *reads* par échantillon et une longueur de *reads* minimale de 100 bp.

Le téléchargement des fichiers fastq bruts, des expériences conservées, depuis les bases de données publiques est effectué par le pipeline à l'aide des outils SRA-toolkit, en particulier fastq-dump et Aspera. Les fichiers sont stockés en construisant une architecture de dossier de type SRP/SRX/SRR permettant de faciliter la navigation et l'entrée dans les pipelines suivants. Les métadonnées rattachées aux échantillons ont été agrégées et traitées manuellement afin de reconstituer les plans d'expériences, les tissus testés, les traitements associés et les liens de téléchargement manquants.

Le téléchargement de l'ensemble de ces données brutes construit un jeu de données exhaustif des expériences de séquençage d'ARN chez le pommier. Il convient ensuite de traiter ces données brutes afin de produire des alignements fiables et ainsi permettre une analyse de l'expression des gènes ohnologues chez *M. domestica*.

## 6.2.2 Traitement des données brutes, comptages et analyse différentielle

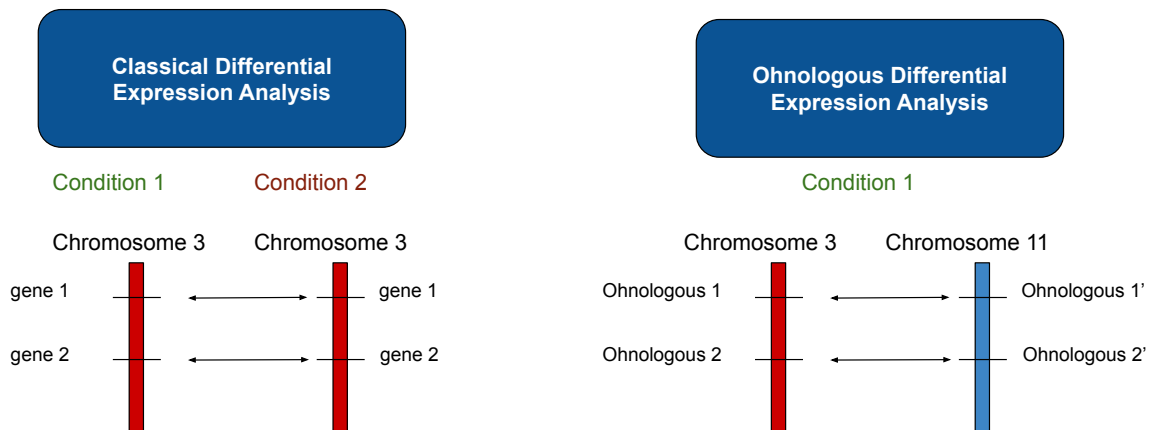
Les données brutes de RNA-Seq nécessitent un ensemble de traitements et en particulier un alignement (*mapping*) des *reads* sur le transcriptome de référence, qui rassemble l'ensemble des gènes potentiellement transcrits dans le génome. Les données brutes ont été traitées à l'aide d'un pipeline de *pseudomapping ad hoc* (<https://forgemia.inra.fr/tanguy.lallemand/rna-seq>). Ce pipeline produit les étapes de suppression des adaptateurs (*trimming*), l'évaluation de la qualité de séquençage et l'alignement des *reads* sur le transcriptome de référence de GDDH 13 #11. Le *trimming* peut être réalisé *via* un des deux principaux outils de référence, cutadapt (Martin, 2011) ou trimmomatic (Bolger et al., 2014). Pour cette analyse c'est cutadapt qui a été utilisé. Les étapes de contrôles de qualité des données brutes sont faites à l'aide de FastQC (Andrews et al., 2010). Le pseudoalignement



ment des *reads* sur le transcriptome de référence et la quantification de l’expression des transcrits dans chacune des conditions analysées ont été réalisés à l’aide de Salmon (Patro et al., 2017). L’utilisation de STAR (Dobin et al., 2013) a aussi été implémentée. L’indexation du transcriptome de référence a été effectuée avec l’outil d’indexation de Salmon en version 4. La configuration de l’outil a été faite comme suit : un échantillonnage dense ; non-conservation des doublons et une taille de k-mer fixé à 31 en suivant la recommandation des auteurs (Patro et al., 2017) pour l’alignement de *reads* de taille supérieur à 75. Ainsi un total de 60 172 826 k-mers est généré. Le pseudoalignement a quant à lui été réalisé avec Salmon en version 1.3.0 avec une bibliothèque adapté aux données pairées, de type *Inward* (l’orientation relative), *Stranded* (le protocole est spécifique au brin), *Forward* (la lecture 1 provient du brin avant) (ISF). De plus, Salmon a été paramétré pour éliminer les alignements ambigus afin de limiter les biais dus aux séquences dupliquées qui sont la cible de l’analyse. Toutes les métriques des étapes du pipeline ont été agrégées à l’aide de MultiQC (Ewels et al., 2016) permettant ainsi la construction d’un rapport unique et complet de la qualité des données et des différents traitements appliqués lors de l’analyse. Après les étapes d’alignement, toutes les expériences pour lesquelles au moins un des réplicats présente un taux global d’alignement inférieur à 70 % ont été retirées de l’analyse. De même, les expériences présentant des métadonnées de mauvaise qualité n’ont pas été conservées. Ceci est permis par le grand nombre d’expériences accessibles qui autorisent d’être particulièrement stringent et de fournir ainsi une analyse de la meilleure qualité possible.

Après alignement et évaluation de la qualité, le nombre de *reads* pour chaque couple de gènes ohnologues est normalisé en fonction de la longueur du gène le plus long de la paire de gènes ohnologues. Enfin, une analyse différentielle entre les gènes ohnologues au sein des mêmes conditions expérimentales a été réalisée. Un schéma présentant l’approche utilisée lors de la construction de cette analyse différentielle des gènes ohnologues est présenté en Figure 6.1. Pour mener les analyses différentielles, nous avons utilisé le script anaDiff (Pelletier, 2022), un package R basé sur DESeq2 (Love et al., 2014) et edgeR (McCarthy et al., 2012). Pour ces comparaisons, et après avoir testé avec les deux outils, DESeq2 est utilisé pour effectuer les analyses différentielles en tenant compte des réplicats biologiques. Les analyses différentielles que nous avons menés dans le cadre de cette thèse sont différentes des analyses classiquement menées. En effet, on ne cherche pas à comparer le niveau d’expression d’un même gène entre deux conditions expérimentales, mais plutôt à comparer le niveau d’expression des couples de gènes ohnologues au sein

de la même condition expérimentale. Pour chaque condition nous avons divisé en deux le génome pour séparer les paires de gènes ohnologues, l'analyse d'expression différentielle est alors faite sur ces deux demi-génomes pour chaque condition (en prenant en compte les trois réplicats) séparément.



**Figure 6.1** — Schéma de la construction de l'analyse différentielle des gènes ohnologues. La partie gauche du schéma résume succinctement le fonctionnement d'une analyse d'expression différentielle où le niveau d'expression d'un gène dans une condition est comparé au niveau d'expression de ce même gène dans une condition expérimentale différente. La partie droite du schéma présente notre approche d'analyse d'expression différentielle des gènes ohnologues consistant en la comparaison du niveau d'expression des paires de gènes ohnologues au sein de la même condition expérimentale.

### 6.2.3 Analyse des niveaux d'expression

Afin de valider des différences de niveaux d'expression des gènes au sein de paires de fragments chromosomiques ohnologues, une analyse statistique a été implémentée pour tester le nombre de gènes ohnologues différentiellement exprimés entre les paires de chromosomes ohnologues. Elle est stockée sur le dépôt : <https://forgemia.inra.fr/tanguy.lallemand/rna-seq-downstream-analysis>. L'analyse statistique est faite à l'aide du test de Wilcoxon qui teste la distribution du nombre moyen de *reads* alignés pour les 3 réplicats pour chacun des gènes ohnologues.

L'analyse a été menée pour chaque condition à part. Nous avons réalisé des tests de Wilcoxon sur les gènes significativement différentiellement exprimés et avec au moins le

double de *reads* alignés sur l’un des deux gènes ( $\log_{2}FC > 2$ ). Cette règle est décrite dans la littérature comme la *two-fold rule* (Renny-Byfield et al., 2015 ; J. C. Schnable et al., 2011 ; Woodhouse et al., 2014)). Cette série de tests a de la même façon été réalisée avec tous les gènes significativement différentiellement exprimés, quelle que soit la valeur du  $\log_{2}FC$ . Cette approche est décrite dans la littérature comme *horse race rule* (Renny-Byfield et al., 2017 ; J. C. Schnable et al., 2011 ; Woodhouse et al., 2014). Ainsi, pour chacune des conditions testées, une analyse d’expression différentielle a été faite et deux analyses statistiques ont été produites. Les résultats des tests de chacune des conditions ont alors été agrégés pour chaque paire de chromosomes ohnologues en appliquant la méthode d’agrégation des p-values de Fisher, dont la méthode est détaillée en section 1.2.4. Cette analyse permet l’obtention d’un résultat unique pour chaque paire de chromosomes. Ce résultat global va déterminer si au travers de l’ensemble des traitements et tissus analysés, le nombre de gènes différentiellement exprimés est significativement différent entre les paires de fragments chromosomiques ohnologues. En outre, pour chacun des gènes et pour chacune des expériences, le nombre de fois où un gène donné était surexprimé par rapport à son ohnologue a été comptabilisé. Ceci permet de vérifier si certains gènes ont des comportements similaires pour l’ensemble des conditions du jeu de données, quels que soient les tissus et les traitements analysés.

## 6.3 Résultats

### 6.3.1 Récupération des données de séquençage RNA-Seq

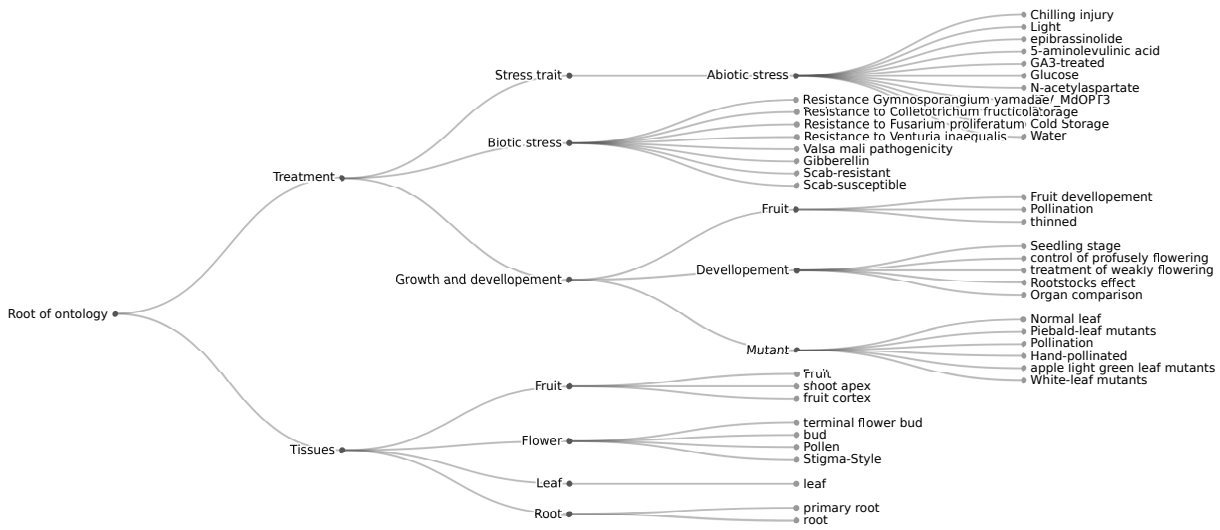
À partir du pipeline *get-rna-seq-data* nous avons pu rechercher toutes les expériences de RNA-Seq disponibles sur les bases de données publiques (SRA, ENA, GEO). Nous avons récupéré 9578 séries associées à la sous-famille *Malus*. En ne conservant que les analyses de séquençage ARN d’une taille d’au moins 20 000 000 de paires de bases, une longueur de lecture d’au moins 100 bp et uniquement ceux associés à l’espèce *M. domestica*, nous avons pu récupérer 2084 identifiants. Parmi eux, 820 sont des échantillons pairés et sont associés à au moins trois réplicats. Ces 820 séries de séquençage ARN (échantillons) sont réparties en 54 expériences réunissant 798 conditions. Nous avons aussi ajoutés 1080 échantillons regroupant des identifiants associés à différentes expériences dont les annotations sur les bases de données publiques étaient insuffisantes pour rendre leur identification automatique simple. Elles ont été expertisées manuellement, et présentaient une grande

proportion de duplicata avec les expériences déjà identifiées. Après nettoyage du jeu de données, notamment la déduplication et la suppression d'expériences dont les métadonnées sont présentes sur les bases de données publiques, mais pas les fichiers bruts, nous avons pu récupérer un total de 633 séries de séquençage RNA-Seq associées à 629 conditions rassemblées en 40 études. Le jeu de données brutes téléchargées sous forme de fichiers compressés nécessite 2176 Go d'espace disque. Pour contrôler la fiabilité des résultats et dans l'optique d'effectuer une analyse la plus exhaustive possible, nous avons exécuté de nouveau ce pipeline une année plus tard (2021). Nous avons récupéré plus d'expériences, pour un total de 755 séries de séquençages RNA-Seq réunis en 250 groupes de 3 réplicats décrivant 751 conditions d'expérimentations rassemblées en 45 expériences pour un espace disque de 2907 Go. C'est sur ce dernier jeu de données, plus récent, que les analyses présentées ici ont été menées.

Après téléchargement des 1510 fichiers de données brutes et annotations manuelles de l'ensemble des expérimentations, nous avons exécuté le pipeline de pseudoalignement pour l'ensemble de ces échantillons. Une étape de filtration a été réalisée afin de supprimer les échantillons dont au moins un des réplicats présente un taux d'alignement inférieur à 70 %. Nous avons aussi éliminé de l'analyse les échantillons présentant des métadonnées de mauvaise qualité et dont l'annotation manuelle à l'aide des publications originelles n'est pas possible avec fiabilité. Nous avons conservé un total de 24 expériences, 44 séries de séquençages RNA-Seq rassemblés en 148 groupes de 3 réplicats (soit 444 échantillons pairés) totalisant 1789 Go.

À partir des annotations existantes et de l'enrichissement manuel des métadonnées rattachées aux expériences, nous avons pu construire une ontologie décrivant les métadonnées des tissus et traitements testés. Afin de permettre de potentielles comparaisons, cette ontologie a été pensée pour se calquer sur le modèle de l'ontologie utilisée dans l'étude QTL. Une visualisation de l'ontologie complète sous la forme d'un arbre hiérarchique est présentée en Figure 6.2. Cet arbre montre la grande diversité des traitements et des tissus dans le jeu de données étudié. La diversité des traitements est en majorité relative à des stress biotiques (résistances à des pathogènes en majorité) ou abiotiques (très divers). On retrouve aussi des études plutôt liées à des notions de développement de l'arbre, du fruit ou à des mutants particuliers. Concernant la diversité des tissus testés, elle rassemble plutôt des fruits, des fleurs, et dans une moindre mesure des feuilles et racines.

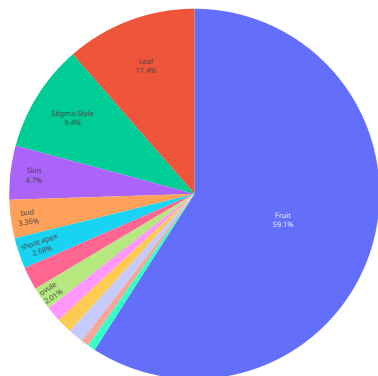
Cette ontologie a permis l'annotation manuelle de l'ensemble des expériences RNA-Seq



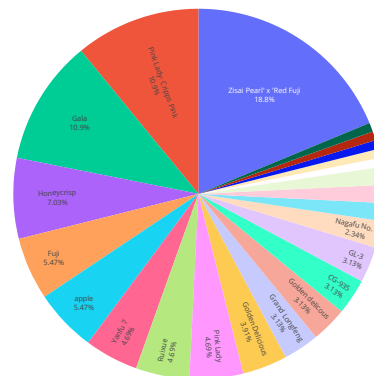
**Figure 6.2** – Arbre hiérarchique de l'ontologie décrivant les traitements et les tissus associés aux séries de séquençages RNA-Seq analysées. Cette ontologie contient quatre niveaux représentés par les quatre niveaux de nœuds présentés de gauche à droite. Les branches représentent les relations de type « *is\_a* » entre les termes de l'ontologie.

analysées. La visualisation des proportions des différents traits dans le jeu de données est présentée en Figure 6.3. Les détails des proportions sont placés en annexe. Les nombres des traitements au niveau trois de l'ontologie sont présentées en Table A.4. Ainsi, on peut observer une grande diversité dans les traitements étudiés, ce qui était suggéré par la représentation de l'arbre de l'ontologie. On retrouve principalement des stress abiotiques (47 %) ou liés au développement du fruit (34 %) 6.3C. Concernant, les tissus, la majorité provient de fruits (59 %) et dans une moindre mesure de feuilles (11 %) 6.3A. Les nombres des tissus sont présentées en Table A.3. Pour finir, les cultivars présentent une grande diversité avec une majorité représentée par Zisai Pearl x Red Fuji (18 %), Gala (10 %) et Pink Lady x Cripps Pink (10 %) 6.3B. Les nombres des tissus sont présentées en Table A.2.

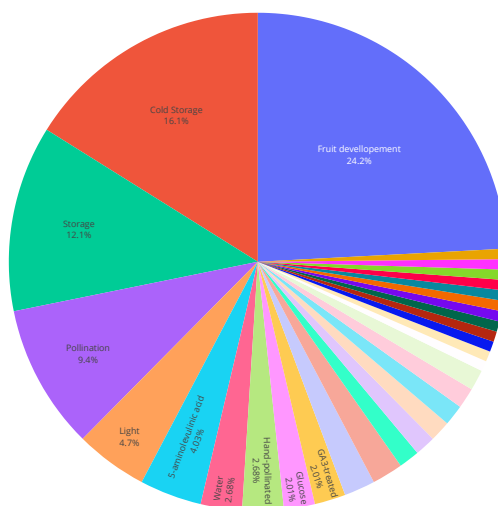
Ainsi, le pipeline développé a permis la recherche d'expérience de séquençage RNA-Seq associées à *M. domestica*. Ces expériences ont été filtrées selon des critères qualité pour assurer une haute qualité des données analysées. Les métadonnées associées à ces expériences ont été récupérées, ré-annotées manuellement et associées à une ontologie permettant la description des tissus et des traitements associés à chacune des conditions analysées. La grande diversité de tissus, traitements et variétés dans ce jeu de données en fait un ensemble le plus exhaustif possible à ce jour pour des données transcriptomiques



(A) Proportions des métadonnées associées aux tissus de pommier analysés



(B) Proportions des métadonnées associées aux cultivars de pommier analysés



(C) Proportions des métadonnées associées aux traitements de pommier analysés

**Figure 6.3** — Proportions des différentes métadonnées décrivant les tissus, cultivars et traitements associés aux expériences RNA-Seq testées. La couleur permet de différencier les secteurs associés aux différentes métadonnées. La taille des secteurs permet de représenter visuellement la proportion d'expériences associées à cette métadonnée. Le pourcentage exact est aussi ajouté dans chacun des secteurs. La proportion des tissus testée présentée en Figure 6.3A montre une majorité de fruits testés. Les cultivars (6.3B) sont quant à eux représentés en majorité par Zisai Pearl x Red Fuji, Gala et Pink Lady x Cripps Pink. Les traitements (6.3C) examinés sont majoritairement liés au développement du fruit et des stress abiotiques.

associées au pommier. Les données brutes doivent ensuite être traitées via un pipeline d'alignement afin de réaliser des comptages pour mener les analyses différentielles d'expression.

### 6.3.2 Traitement des données brutes, comptages et analyse différentielle

Nous avons utilisé le pipeline de *pseudomapping* sur 755 séries de séquençage RNA-Seq pairés, ce qui représente 1510 fichiers de données brutes rassemblées en 250 groupes de 3 réplicats.

Les différents indicateurs récupérés à chacune des étapes du pipeline pour chacun des échantillons ont été agrégés et représentés à l'aide de MultiQC. Une visualisation globale résumant l'ensemble des indicateurs est fournie en annexe (Figure A.13). De par le nombre important d'échantillons analysés, cette visualisation globale en *beeswarm* ne permet pas un grand niveau de détail, mais permet néanmoins d'observer une qualité générale satisfaisante pour l'ensemble des échantillons filtrés.

L'évaluation de la qualité des données brutes associées aux échantillons a été réalisée avec FastQC dont les données ont ensuite été agrégées et visualisées à l'aide de MultiQC. Concernant le compte de séquences fournies en annexe (Figure A.16) on retrouve les valeurs attendues, avec un pourcentage de *reads* uniques entre 40 et 60%. Néanmoins, sept échantillons semblent montrer une anomalie avec des pourcentages de *reads* uniques à près de 90%. Par ailleurs, l'estimation de la qualité des séquences basée sur le score issu des fichiers FastQ montre que l'ensemble des échantillons présentent des scores de qualité (Phred score) entre 35 et 40, avec une baisse légère sur la fin des *reads*. Ce comportement est attendu et confirme ici la très bonne qualité des séquences brutes. Ce constat est confirmé par la médiane des Phred scores qui est supérieurs à 28, et la fin des *reads* ne comporte pas de scores inférieurs à 20. Néanmoins, sept échantillons présentent à une position de 19 bp un Phred score égal à 10. Cette observation est confirmée par la Figure A.19, détaillant la distribution des pourcentages d'utilisation de la base « N » (*any base*) à chaque position et qui montre un fort appel au N pour sept échantillons pour la base 19, suggérant un manque de qualité de séquençage pour cette position sur les sept échantillons concernés. Néanmoins le score de qualité par séquence (Figure A.18), montre que ces *reads* de mauvaise qualité à 19 bp représentent une partie négligeable de l'échantillon. Les échantillons en question sont ceux identifiés comme problématiques par

le compte de séquences. Après expertise manuelle de ces échantillons, nous avons choisi de les supprimer. La distribution du contenu en adaptateur (Figure A.22) montre qu'il ne présente pas d'anomalie par rapport aux valeurs attendues si ce n'est pour les sept échantillons déjà identifiés comme problématiques. Nous n'avons pas non plus remarqué la présence de séquences sur-représentées au-delà des seuils acceptés dans la littérature (Figure A.21). La distribution de la taille des *reads* après *trimming* est présentée en Figure A.15. On peut observer la présence de pics essentiellement au début des *reads*, puis à 120 bp et 140 bp ce qui est en conformité avec la longueur minimale des *reads* fixés à 100 bp dans la sélection des échantillons. De même, on peut observer, dans les distributions des longueurs de fragments alignés présentées en Figure A.14, une absence de valeurs en dessous de 100 bp. La longueur des fragments alignés montre ici une bonne homogénéité de l'ensemble des échantillons alignés avec une majorité des *reads* alignés dont la longueur est située entre 150 et 500 bp. Cette distribution est en adéquation avec la distribution de la taille des transcrits primaire pour le transcriptome de référence utilisé qui présente un premier quartile à 191 bp et un troisième quartile à 621 bp. La qualité des échantillons a aussi été évaluée sur le contenu en GC moyen. L'ensemble des distributions du contenu en GC est proposé en Figure A.20. La distribution du contenu en GC est attendue comme suivant une loi normale. Ainsi on retrouve ici 58 échantillons où une partie des *reads* ont des pourcentages moyens en GC légèrement élevé. Ceci ne constitue pas un signal suffisant pour supprimer ces échantillons, mais constitue un point de vigilance.

Malgré la grande similarité des séquences codantes des gènes ohnologues observée précédemment par l'étude  $K_a/K_s$ , le nombre de lectures ambiguës lors de l'alignement est resté minime avec une moyenne globale de 0,048 %. Seuls 596 gènes avaient une valeur médiane supérieure à une lecture ambiguë dans toutes les expériences. Ce résultat montre la fiabilité de Salmon qui malgré le haut degré de duplication du génome du pommier permet d'obtenir un résultat d'alignement très fiable. De plus, le taux d'alignement des expériences sur lesquelles l'analyse différentielle a été appliquée varie entre 72 % et 91 % avec une médiane de 85 %. Ce résultat confirme la bonne qualité des données initiales ainsi qu'une utilisation optimale des outils.

À partir de l'ensemble des résultats d'alignements calculé pour les 148 expériences rassemblant un total de 444 séries de séquençages RNA-Seq nous avons cherché à déterminer s'il existait des couples de gènes ohnologues différentiellement exprimés dans certaines expériences à l'aide d'une analyse d'expression différentielle. Globalement, pour



la somme de toutes les expériences, nous avons identifié 1 684 796 couples de gènes différentiellement exprimés parmi les 2 299 446 couples de gènes testés à travers l’ensemble des 148 expériences. Parmi ceux-ci, 1 032 881 couples de gènes ohnologues présentent un logFC (en valeur absolue) supérieur à 2. Ainsi, sur un total de 16 779 paires de gènes ohnologues, nous avons identifié une médiane de 11 469 (écart-type=1369) couples de gènes ohnologues significativement ( $\alpha < 5\%$ ) différentiellement exprimés par expérience dont 6691 (écart-type=401) avec un logFC supérieur à 2. Le nombre de gènes différentiellement exprimés ne semble pas lié à des traitements ou des tissus particuliers.

### 6.3.3 Analyse des niveaux d’expression

Nous avons effectué une analyse du différentiel d’expression des gènes ohnologues sur les 148 expériences (444 échantillons) RNA-Seq de haute qualité. Nous avons pu identifier dans chacune des expériences de nombreuses paires de gènes ohnologues différentiellement exprimés. Nous avons ensuite voulu tester si ces gènes différentiellement exprimés étaient répartis aléatoirement ou non. Nous avons ainsi vérifié si la présence de gènes différentiellement exprimés était équilibrée entre les paires de chromosomes ohnologues.

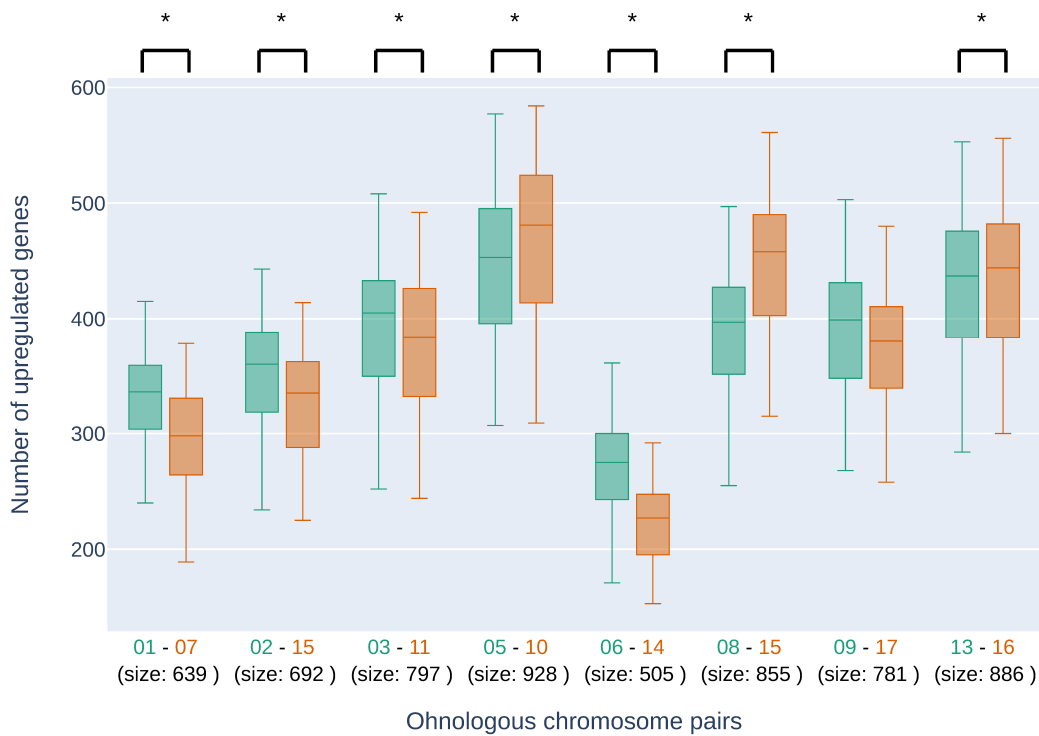
Pour tester un déséquilibre des niveaux d’expression des gènes ohnologues, nous avons compté le nombre de gènes significativement surexprimés pour un chromosome (ou un segment de chromosome) par rapport à son ohnologue dans chacune des 148 expériences RNA-Seq considérées. Des tests de Wilcoxon ont été mis en œuvre pour comparer le nombre de *reads* moyen entre les réplicats biologiques de tous les gènes différentiellement exprimés sur chaque paire de chromosomes. Ces tests ont été menés sur chacune des 148 expériences. Les *p-values* ont été agrégées à l’aide de la méthode de Fisher afin d’obtenir un résultat unique pour chaque paire. Afin de tester si certaines conditions expérimentales présentent des déséquilibres d’expression pour certaines paires de fragments chromosomiques, nous avons agrégés, lors du test de Fisher, les expériences suivant leur condition expérimentale au niveau deux de l’ontologie. La Figure A.29 présente ces résultats sous forme de carte thermique. On peut observer que globalement les résultats pour les paires 1-7, 6-14, 8-15 et 2-15 présentent des résultats similaires à l’agrégation de l’ensemble des expériences suggérant un déséquilibre transcriptionnel pour un ensemble de tissus et/ou de conditions expérimentales. Pour les paires 13-16, 3-11 et 5-10, le déséquilibre est particulièrement marqué pour quelques conditions expérimentales, principalement celles des stress abiotiques comme la lumière et des stress biotiques comme la résistance à des pathogènes.

La distribution du nombre de gènes surexprimés dans toutes les expériences est présentée dans la Figure 6.4. Cette figure montre que dans plusieurs cas, le nombre de gènes surexprimés dans un chromosome ou un segment de chromosome est significativement différent de celui de sa paire ohnologue. Toutes les paires exceptée la paire 9-17 sont significativement déséquilibrées en ce qui concerne le nombre de gènes surexprimés dans chacune des expériences. Le détail des p-values agrégées est présenté en Table 6.1. Ainsi, les paires 1-7, 2-15, 3-11, 5-10, 6-14, 8-15 et 13-16 présentent un déséquilibre dans les niveaux d'expression des gènes ohnologues analysés pour un ensemble de 148 expériences RNA-Seq associées à une diversité de tissus et de conditions expérimentales. Le chromosome 1 est significativement plus souvent surexprimé par rapport à son fragment synténique sur le chromosome 7 avec une médiane de gènes significativement surexprimés au travers des 148 expériences de 336 gènes pour le chromosome 1 contre 298 pour le chromosome 7. Pour les paires 2-15, 3-11 et 6-14 c'est le premier fragment chromosomique (chromosome 2, 3 et 6) qui sont le plus souvent surexprimés par rapport à leur fragment synténiques. À l'inverse, pour les paires 5-10, 8-15 et 13-16, c'est le second fragment chromosomique synténique qui est dominant d'un point de vue de l'expression.

**Table 6.1** — Résultats des tests de Wilcoxon agrégés à l'aide de la méthode de Fisher pour les paires de fragments chromosomiques synténiques.

Couple	p-value Wilcoxon	Médiane nombre de gènes surexprimés pour le premier chromosome	Médiane nombre de gènes surexprimés pour le second chromosome
08-15	$2,2796 \times 10^{-122}$	397	458
01-07	$1,3405 \times 10^{-75}$	336	298
06-14	$2,2879 \times 10^{-2}$	275	227
02-15	$2,4565 \times 10^{-24}$	360	335
03-11	$4,4138 \times 10^{-3}$	405	384
09-17	$4,0946 \times 10^{-1}$	399	380
13-16	$7,2743 \times 10^{-3}$	437	454
05-10	$3,0120 \times 10^{-7}$	453	484

Par ailleurs, nous avons aussi représenté sous la forme d'un histogramme (Figure 6.5) le nombre de fois où, pour un couple de gènes ohnologues, un des gènes est surexprimé par rapport à son ohnologue. On peut observer qu'un grand nombre de gènes sont surexprimés par rapport à leur ohnologue dans de nombreuses expériences. Globalement, pour l'ensemble des paires de chromosomes, un gène ohnologue est différentiellement exprimé



**Figure 6.4** – Diagrammes en boîte de la distribution du nombre de gènes surexprimés par rapport à son ohnologue dans les 148 conditions pour les deux chromosomes dans chacune des paires d'ohnologues. L'axe des y (*Number of upregulated genes*) constitue l'ensemble des valeurs du nombre de gènes surexprimés. L'axe des x (*Ohnologous chromosome pairs*) indique les paires de chromosomes considérées. Le bas de la boîte représente le premier quartile, le trait interne à la boîte présente la médiane de la distribution et l'extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l'écart interquartile. La moustache inférieure présente le neuvième décile et la moustache présente quant à elle le 9<sup>e</sup>. Pour chaque paire de chromosomes (ou segments de chromosomes), les diagrammes en boîte indiquent la distribution sur le premier chromosome de la paire (en vert) et la distribution sur le second chromosome de la paire (en orange). \* $\alpha < 5\%$

en moyenne dans 48 expériences. Nous n'avons pas noté de différence significative de cette moyenne selon les paires de chromosomes testés. De plus, certaines zones de quelques dizaines de couples de gènes semblent être associées à une surexpression d'un même chromosome par rapport à son ohnologue. Cette observation est visible en Figure 6.5, où certaines présentent des régions où les couples de gènes sont significativement surexprimés sur le même chromosome dans un grand nombre d'expériences. Ces zones d'intérêt pourraient permettre une compréhension des mécanismes permettant ce déséquilibre transcriptionnel entre les gènes ohnologues.

Pour finir, la Figure 6.5, présente deux lignes horizontales fixées sur la valeur de 125. Elles permettent de visualiser les paires de gènes ohnologues qui présentent un déséquilibre dans le même sens dans 85% des cas. Ces gènes sont appelés ici des *gènes non commutants*. Nous avons identifié une liste de 2247 paires de gènes non commutants dans lesquels l'un des deux gènes est systématiquement surexprimé par rapport à son ohnologue. À l'inverse nous avons identifié 2011 paires de gènes qui sont différentiellement exprimés dans moins de 20 expériences parmi les 148. 633 appartiennent à l'une des 8 principales paires. Ainsi, sur la Figure 6.5 décrivant la paire 1-7, les couples de gènes dont l'une des barres dépasse le seuil de 125 expériences dans le même sens sont désignés comme des gènes non commutants. Leur répartition le long du chromosome suggère une distribution aléatoire le long du chromosome. Les figures associées aux autres paires sont présentées en Annexe en Figure A.23 à A.28 et suggèrent un résultat similaire. Nous avons identifié 1387 paires de gènes non commutants appartenant aux 8 paires principales de fragments chromosomiques synténiques. De plus, nous n'avons pas identifié de différences significatives dans la proportion de gènes non commutants entre les différentes paires de chromosomes ohnologues, à l'exception de la paire 8-15 et 9-17 (test de proportion  $z$ ,  $\alpha = 0.05$ ).



**Figure 6.5** — Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couples*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 1 et associés au fragment synténique 1-7. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives, associées à une barre verte, représentent le nombre de fois où c'est le gène associé au chromosome 1 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 7. À l'inverse si la valeur est négative, associée à une barre orange, c'est le gène associé au chromosome 7 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 1. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisé pour définir les gènes non commutants.

## 6.4 Discussion

Dans ce chapitre a été détaillée une analyse de l'expression différentielle des gènes ohnologues chez le pommier. Cette analyse a été menée à grande échelle en prenant en compte un maximum de données de la meilleure qualité possible. Elle vise à identifier à l'aide d'un ensemble d'analyses d'expression différentielle entre les gènes ohnologues au sein de la même condition si certains couples de gènes ohnologues sont différentiellement exprimés. Dans le cas de couples différentiellement exprimés, nous avons cherché si globalement leur répartition était équilibrée entre les paires de chromosomes ohnologues.

### 6.4.1 Récupération des données de séquençage RNA-Seq

Nous avons récupéré 755 séries de séquençages RNA-Seq pairés soit 1510 fichiers de données brutes passant les critères de qualité fixés. Nous avons donc pu récupérer un nombre important d'expériences avec la plus grande diversité de traitements et de tissus possible. Après alignements des fichiers bruts, un filtre sur le pourcentage d'alignement a été mis en place pour garder les résultats les plus fiables possibles en termes d'alignement. Ainsi, les expériences conservées pour l'ensemble de l'analyse sont principalement liées à des traits et des variétés d'importance agroéconomique. Les conditions testées présentent une majorité de conditions associées à des stress abiotiques et au développement du fruit. De même, la majorité des tissus examinés sont originaires de fruits. Du point de vue des variétés, les trois cultivars les plus représentés sont issus de variétés avec un fort intérêt économique. De plus, ces trois variétés constituent des variétés proches de Golden Delicious dont le transcriptome de référence est issu, ce qui en principe favorise de hauts pourcentages d'alignements. Le filtre sur le pourcentage de mapping a ainsi pu retirer une partie des variétés et/ou espèces de *Malus* plus éloignées de *Golden delicious*. Les expériences analysées ne représentent donc pas totalement toute la diversité génétique du pommier que ce soit en termes de traits, de tissus et de variétés étudiées. Néanmoins, les traits et cultivars étudiés sont en adéquation avec les traits étudiés lors de l'analyse QTL et n'introduisent donc pas de biais qui serait lié à l'étude de traitements ou de cultivars particuliers. De plus, cette étude est la plus exhaustive possible et rassemble l'ensemble des données de haute qualité présentes sur les bases de données publiques à la date du 9 août 2021. L'importance du jeu de données collecté dans cette étude permet d'obtenir une puissance statistique très importante et permet d'étudier des mécanismes généraux et à grande échelle. De plus, des espèces comme le coton ou le blé présentent des

dominances sous génomiques qui dépendent des tissus et/ou des traitements. Ces diversités d'expériences permettent d'étudier si de telles observations pourraient être faites chez le pommier.

Les données transcriptomiques disponibles sur les bases de données publiques sont en pleine explosion. La production d'un transcriptome de référence de qualité ainsi que l'abaissement des coûts de séquençage sont à l'origine d'une importante hausse de la production de séquençages RNA-Seq. Ainsi, entre le premier téléchargement de données et le second, près de trois mois, sont passés. Ces trois mois de délais ont permis l'obtention de 150 échantillons supplémentaires passant les critères de qualité fixés. Ce pipeline, pensé pour être réutilisable, pourrait ainsi permettre d'à plusieurs reprises augmenter la taille du jeu de données disponibles afin d'augmenter la diversité des expériences, améliorant ainsi le résultat de cette analyse, mais aussi ouvrant la porte à des analyses supplémentaires. Ainsi il conviendrait de pouvoir relancer ces pipelines fréquemment afin de vérifier si de nouveaux éléments pourraient être identifiés. Le fait que les différentes étapes soient empaquetées dans un pipeline Snakemake ainsi que les différentes routines d'historisation des exécutions permettent d'archiver facilement les différentes exécutions et ainsi garder une traçabilité des différentes expériences ajoutées ainsi que leur effet sur les résultats. Néanmoins l'archivage d'un tel jeu de données nécessite un investissement important. Pour terminer, la construction d'une ontologie et la réannotation manuelle des métadonnées associées ont offert une annotation fiable des expériences. Ceci permet de tester avec fiabilité si certains tissus ou traitements présentaient des anomalies parmi le jeu de données analysé. Pour finir, ce jeu de données ainsi que la réannotation manuelle des métadonnées a été archivé et pourrait être une source de données importantes pour de futures analyses sur la transcription chez le pommier.

#### **6.4.2 Traitement des données brutes, comptages et analyse différentielle**

Après annotations des métadonnées associées et alignement, nous avons pu filtrer les fichiers suivants des critères de qualité de la donnée brute, de l'alignement et de la qualité de l'annotation. Étant donné la masse importante de données disponibles, les filtres ont été relativement stringents afin de ne conserver que les données de la meilleure qualité avec une information fiable. L'évaluation manuelle des rapports de qualité agrégés par MultiQC suggère des données de bonne qualité limitant des biais liés aux données

brutes et confirmant que les filtres mis en place ont permis la sélection des meilleures données possibles. Nous n'avons pas mis en évidence de données contaminées ou avec des problèmes techniques divers qui auraient été conservées dans le jeu de données final. C'est pourquoi l'alignement de *reads*, présente un ensemble de résultats de bonne qualité. Les taux d'alignement sont satisfaisants, de plus, les taux d'alignement ambigus sont bas malgré la similarité des séquences ohnologues au niveau protéique, ce qui constitue un élément rassurant sur la qualité et l'interprétation des résultats.

Le nombre médian important de gènes significativement différentiellement exprimés pour chacune des expériences 11 469 (écart-type=1369) (68 %) sans filtre sur le logFC 6691 (écart-type=401) (40 %) avec un filtre sur les logFC supérieur à 2 indique des différences de niveau d'expression des gènes ohnologues importante. Cette différence constitue un résultat inattendu. En effet, l'analyse différentielle a été effectuée sur les gènes ohnologues au sein de la même condition expérimentale. L'attendu serait donc plutôt des gènes différentiellement exprimés à la marge, puisque ceux-ci codent pour les mêmes protéines. Or le nombre médian de gènes significativement différentiellement exprimés au travers des 148 expériences est de 11 469 gènes, ce qui suggère que la plupart des couples de gènes ohnologues sont significativement différentiellement exprimés dans l'ensemble des conditions analysées.

Des résultats similaires ont été observés chez différentes espèces avec des WGD récentes. Ainsi, un certain nombre de gènes différentiellement exprimés ont été retrouvés chez le maïs (J. C. Schnable et al., 2011). Un gène était considéré comme différentiellement exprimé si son RPKM était plus grand que son ohnologue. Des filtres supplémentaires de type *horse race rule* ou *two-fold rule* ont aussi été mis en place. En testant sur 4 échantillons RNA-Seq de faibles qualités, des ratios autour de 60 % de gènes ohnologues ont été considérés comme différentiellement exprimés par rapport à leurs ohnologues. De même, chez *B. rapa*, autour de 50 % des gènes ohnologues ont été identifiés comme différentiellement exprimés (Cheng et al., 2016 ; Woodhouse et al., 2014). Ces différences d'expression sont bien retrouvées entre les trois tissus testés (racine, feuille et tige). Néanmoins comme pour le maïs, ce résultat est moins fiable que celui présenté dans ce chapitre, car la détermination des couples différentiellement exprimés a été réalisée seulement sur le logFC et n'a pas intégré une analyse différentielle à proprement parler qui intègre une validation statistique des différences d'expression en tenant notamment en compte le bruit de fond et en appliquant un certain nombre de corrections. De même l'analyse a été menée sur un ensemble réduit de 3 expériences associées à trois tissus avec des qualités de séquençage



moindres c'est à dire avec des *reads* brutes *single-end* et une longueur minimum de 36 bp. Un constat similaire est fait chez le saule (Harikrishnan et al., 2015), avec des proportions de gènes différentiellement exprimés similaires, 76 % avec la *horse race rule* et 40 % pour la *two-fold rules*.

### 6.4.3 Analyse des niveaux d'expression

Les 148 analyses différentielles menées sur les couples de gènes ohnologues au sein de la même condition ont mis en évidence un grand nombre de gènes ohnologues différentiellement exprimés. De plus, nous avons pu mettre en évidence que les gènes différentiellement exprimés ne paraissent pas être répartis au hasard entre les paires de chromosomes ohnologues. Ainsi, il semble exister un déséquilibre dans les niveaux d'expression des gènes ohnologues à l'échelle des fragments de chromosomes. Ce déséquilibre a été observé à l'échelle des 148 expériences, mais aussi lors de l'agrégation des expériences selon les conditions expérimentales. Globalement, les paires 1-7, 2-15, 3-11, 5-10, 6-14, 8-15 et 13-16 sont significativement déséquilibrées. Le sens du déséquilibre est dans le même sens que le déséquilibre observé au niveau des QTLs pour les paires 1-7, 2-15 et 3-11. Les déséquilibres sont dans le sens inverse pour la paire 6-14.

La forte prévalence de couples de gènes ohnologues différentiellement exprimés, ainsi que l'impact biologique important d'un tel déséquilibre, suggère que ce mécanisme revêt une importance fonctionnelle clé. Dès lors, on peut supposer que les couples de gènes différentiellement exprimés sont néofonctionnalisés et/ou sous-fonctionnalisés en suivant le modèle Duplication-Degeneration-Complementation (DDC) (Force et al., 1999). L'annotation fonctionnelle des gènes ancestraux de l'ancêtre du pommier pré-WGD n'étant pas connue, il sera difficile de faire la part entre les gènes néo-fonctionnalisés et sous fonctionnalisés. L'absence de différences de  $K_a/K_s$  entre les gènes ohnologues suggère un découplage entre les mécanismes associés à l'expression des gènes et la divergence des séquences codantes des gènes ohnologues. Ceci suggère que les substitutions (synonymes et non synonymes) des séquences codantes ont peu d'impact sur la différence d'expression, suggérant que des modifications des régions régulatrices des gènes ohnologues jouent un rôle dans l'expression différentielle des gènes ohnologues. Des phénomènes de différences d'expression résultant de la sous-fonctionnalisation et/ou de la néofonctionnalisation a déjà été observée chez diverses espèces végétales, par exemple chez le coton (*Gossypium raimondii*) (Renny-Byfield et al., 2014), le maïs (T. E. Hughes et al., 2014), *A. thaliana* (Duarte et al., 2006; Hou et al., 2018) ou le saule (Harikrishnan et al., 2015). Il

est à noter que ce déséquilibre transcriptionnel n'a pas été observé chez le poirier qui a probablement subi la même WGD ancestrale que le pommier (Q. Li et al., 2019). Cette observation suggère que les modifications de la régulation des gènes ohnologues ont eu lieu après la WGD.

Par ailleurs, nous avons identifié un ensemble de 2287 couples de gènes au comportement particulier. En effet, pour ces couples de gènes l'un des gènes est systématiquement surexprimé par rapport à son ohnologue et ceci quelle que soit la condition ou le tissu considéré sur un total de 148 études différentielles. Pour la plupart des couples de gènes, les deux gènes du couple sont bien exprimés, mais l'un l'est systématiquement (dans les 148 expériences) plus que l'autre. Ce phénomène de gènes non commutants n'a pas été retrouvé dans d'autres analyses et constitue un sujet qu'il conviendrait d'étudier plus en détail.

## 6.5 Conclusion

Dans cette analyse, nous avons pu récupérer et analyser 148 conditions de RNA-Seq de haute qualité avec 3 réplicats biologiques. Ces données représentent une diversité importante de variétés de pommiers et de traitements et représentent de façon exhaustive les données RNA-Seq associées au pommier à la date de l'étude. Après traitements et alignements des fichiers bruts, nous avons pu mettre en place des analyses d'expression différentielles des gènes ohnologues au sein de chacune des 148 conditions. De nombreux gènes ohnologues ont été identifiés comme significativement différentiellement exprimés dans les différentes conditions testées. Ces différences de niveaux d'expressions sont probablement liées à une régulation différente des gènes ohnologues. Après analyses statistiques, nous avons pu mettre en évidence que les paires sont significativement différentes sur le plan du nombre de gènes surexprimés à travers l'ensemble des expériences. Ce résultat pourrait expliquer, au moins en partie, le déséquilibre de QTL observé.



# ANALYSE DES ÉLÉMENTS TRANSPOSABLES DES GÈNES OHNOLOGUES

---

## 7.1 Introduction

Les ET sont définis comme des éléments génétiques mobiles et moyennement répétés présents dans les génomes. Ils composent une partie importante des génomes d'eucaryotes, et particulièrement pour les génomes de plantes. Les ET constituent en effet une partie importante du génome des plantes et ont par ailleurs été identifiés pour la première fois chez le maïs (McClintock, 1950). Les ET sont un constituant majeur des génomes et en particulier des génomes de plantes, par exemple 85 % du génome du maïs est constitués d'ET (Jiao et al., 2017; P. S. Schnable et al., 2009).

Avec l'avènement des technologies de séquençage à haut débit, de nombreuses plantes ont été séquencées ces dernières années. La diversité des génomes de plantes terrestres est considérable et présente des niveaux de ploïdie, d'hétérozygotie et de taille du génome différentes. Ainsi, des plantes terrestres présentant des tailles de génome allant de 65 Mb pour *Genlisea tuberosa* à 149 Gb pour *Paris japonica* ont été séquencées. Les ET sont des acteurs majeurs de cette diversité, en permettant une variation de la taille du génome et des taux de recombinaison. En effet, les ET, y compris lorsque ceux-ci sont inactivés vont permettre la génération de régions homologues dispersées dans le génome par des ET issus des même familles. Les familles d'ET représentent les séquences d'ET issues les unes des autres par transposition (Britten & Kohne, 1968). De cette origine résulte le fait que ce sont des séquences très similaires entre elles, même après inactivation. La construction des familles est le plus souvent basée sur la règle des 80-80-80 (Wicker et al., 2007). Cette règle implique de garder les ET qui ont : au moins 80 % d'identité avec la cible, 80 % de couverture de la cible et un minimum de 80 bases (Wicker et al., 2007). Ces régions vont

permettre d'augmenter localement les taux de recombinaison (Bennetzen & Wang, 2014; Deininger et al., 2003). Chez les plantes, de nombreuses familles d'ET ont été identifiées. La proportion d'ET est très diverse allant de 3 % dans le génome de *Utricularia gibba* (Ibarra-Laclette et al., 2013) à 80 % chez le blé (Gardiner et al., 2019) et près de 85 % du génome pour le maïs (Stitzer et al., 2021).

Les ET sont impliqués dans un grand nombre de mécanismes génétiques. Par ailleurs, différentes classes d'ET existent portant chacune des propriétés différentes. La classification des ET a été longtemps débattue et continue d'être sujette à modification. La classification de Wicker (Wicker et al., 2007) est l'une des plus abouties et reconnue à ce jour. Elle se présente sous la forme de deux grandes classes d'ET dont la nomenclature se base sur le mécanisme de transposition. Les éléments de classes I rassemblent les rétrotransposons. Cette classe d'ET se transpose par un processus de « copier-coller » par lequel un intermédiaire d'ARN est rétrotranscrit en une copie d'ADN complémentaire qui est incorporée ailleurs dans le génome (Boeke et al., 1985). Les ET de classe I sont ensuite subdivisés en sous-classes en se basant sur le mécanisme d'intégration dans le génome. Pour les ET de classe I on retrouve les rétrotransposons à longue répétition terminale (Long Terminal Repeat (LTR)) qui s'intègrent dans le génome de la même façon que les rétrovirus, par clivage et transfert du brin. Ce processus est catalysé par l'intermédiaire d'une intégrase (Brown et al., 1987). Les ET de classe I rassemblent aussi les rétrotransposons sans LTR qui s'intègrent par transcription inverse via un mécanisme de transcription inverse amorcée par la cible (*target-primed reverse transcription*) (Luan et al., 1993).

La classe II regroupe des ET appelés transposons à ADN, qui se transposent par « couper-coller » (Greenblatt & Alexander Brink, 1963). Deux sous-classes sont principalement connues, les *Helitrons* s'intégrant dans le génome à l'aide d'un intermédiaire d'ADN circulaire (Grabundzija et al., 2016) et les Terminal Inverted Repeat (TIR) agissant à l'aide d'un intermédiaire ADN non circulaire (Greenblatt & Alexander Brink, 1963).

Chaque sous-classe d'ET est divisée en superfamilles. Chez les ET de classes I, les rétrotransposons LTR sont subdivisés en différentes superfamilles, les deux plus communes étant les Ty3/gypsy et Ty1/copia qui sont retrouvés chez l'ensemble des eucaryotes (Malik & Eickbush, 2001). Pour les rétrotransposons sans LTR, ils comprennent les éléments nucléaires intercalés longs, les Long Interspersed Nuclear Element (LINE) (Singer, 1982) et courts, les Short Interspersed Nuclear Element (SINE) (Singer, 1982).

Pour les ET de classe II, les *Helitrons* n'ont pas été découpés en superfamille. Quant aux ET s'intégrant via un intermédiaire ADN, ils sont découpés en deux superfamilles :

hAT et mariner. La classification de Wicker présente un niveau de granularité inférieur appelé famille qui se base sur la phylogénie, et groupe les ET qui ont un ancêtre commun unique. Ce niveau est encore sujet à des évolutions majeures et ne représentait pas d'intérêt pour l'étude menée ici. De plus, l'annotation de génome utilisée est très lacunaire pour le niveau famille.

Une annotation des ET a été réalisée pour le génome de GDDH13 1.1 (Daccord et al., 2017). Pour ce faire, les séquences consensus des ET ont été construites avec TEdenovo du package REPET (Flutre et al., 2011). Pour affiner cette annotation, deux itérations du pipeline TEannot du package REPET ont été utilisées. Les ET identifiés représentent 374,2 Mb, ce qui correspond à 59,5 % de l'assemblage final de GDDH13.

Les ET jouent un rôle majeur lors de la polyploïdisation et des duplications par WGD. Ainsi, il a été observé que le nombre de copies d'ET est plus élevé chez les polyploïdes que chez leurs espèces diploïdes apparentées (Casacuberta & González, 2013; Wendel, 2015). C'est le cas pour l'allopolyploïde *N. tabacum* et ses géniteurs diploïdes *Nicotiana sylvestris* et *Nicotiana tomentosiformis* où le rétrotransposon Tnt1 existe en plus de copies chez le polyploïde et l'allotétraploïde (Petit et al., 2010). Un constat similaire est fait chez *Triticum aestivum* avec la famille Alu de type SINE non autonome (Ben-David et al., 2013). L'abondance d'insertions d'ET postWGD a été expliquée par différentes hypothèses.

Tout d'abord, l'hypothèse de choc génomique (McClintock, 1984). Cette explosion du nombre d'ET causée par des transpositions induites par des instabilités génomiques liées au doublement du génome et en particulier une levée de la répression épigénétique liée aux méthylations des ET (Parisod et al., 2009; Springer et al., 2016). Des observations soutenant cette hypothèse ont été faites dans de nombreuses espèces allopolyploïdes comme le tabac (Petit et al., 2010) ou *B. napus* (Sarilar et al., 2013). Néanmoins, d'autres hypothèses peuvent expliquer l'accumulation d'ET postWGD et en particulier l'hypothèse de la redondance (Matzke & Matzke, 1998).

Cette hypothèse repose sur le fait que les espèces dupliquées par WGD connaissent une réduction de la pression sélective, due à une redondance des gènes, ce qui permet un masquage polysomique. Ainsi les mutations délétères liées à l'insertion d'ET, sont moins sélectionnées négativement et peuvent s'accumuler plus librement. Cette hypothèse est soutenue par des observations chez des plantes autopolyploïdes comme *Arabidopsis arenosa* (Baduel et al., 2019). Cette observation suggère que la duplication du génome à elle seule, lorsqu'elle n'est pas couplée à l'hybridation comme c'est le cas chez les allopoly-

ploïdes, ne suffit pas à déclencher de choc génomique suffisant pour mettre en place les mécanismes de rafales de transpositions (Parisod & Senerchia, 2012). Pour finir, l’hypothèse du goulot d’étranglement (Lynch, 2007) suggère que la formation de polyploïdes implique en général une réduction importante de la taille effective de la population, ce qui peut favoriser la fixation d’insertions neutres ou peu délétères d’ET, par des effets stochastiques. Ces hypothèses coexistent et sont soutenues par différentes observations. Ainsi, chez *B. rapa*, une hausse importante d’activité des ET a été observée lors de l’allopolyploïdisation entre *B. rapa* et *Brassica oleracea* (An et al., 2014) et soutient l’hypothèse d’explosions d’ET. Chez l’allotétraploïde *C. bursa-pastoris*, il a été observé une augmentation de l’abondance des ET dans les régions géniques, liées à une sélection relâchée (Ågren et al., 2016). Ces hypothèses diffèrent par les mécanismes mis en jeu, mais aboutissent à un résultat relativement similaire et diffèrent par la vitesse d’accumulation des ET.

Cette accumulation d’ET va jouer un rôle prépondérant dans la diploïdisation qui va induire des pertes de gènes, des mutations génétiques et une restructuration du génome. Ces mécanismes ont été constatés chez différentes espèces dupliquées comme le tabac (Lim et al., 2007) et le maïs (Bruggmann et al., 2006). Pendant la diploïdisation, il est possible que l’un des génomes parentaux perde plus de gènes que l’autre. Ce phénomène de biais de fractionnement a été observé chez de nombreuses espèces et notamment le tabac (Renny-Byfield et al., 2011), *Arabidopsis* (Freeling & Thomas, 2006), le maïs (Woodhouse et al., 2010) et *M. domestica* dont le biais de fractionnement a été mis en évidence au cours de cette thèse. Ce phénomène peut être expliqué, au moins en partie, par le biais des insertions d’ET lors de la comparaison des sous-génomes. En effet, il a été proposé qu’un contenu différent en ET entre les deux génomes parentaux puisse conduire à la dominance, et à la rétention préférentielle des gènes, du génome ayant la plus faible charge en ET (Woodhouse et al., 2014). De plus les ET ont été associés à l’origine de la production de nouveaux allèles qui pourraient permettre une évolution par une sous-fonctionnalisation ou une néofonctionnalisation. Ainsi, chez *B. napus*, l’insertion d’un élément *Helitrons* non autonome dans le promoteur du gène déterminant le mâle de l’auto-incompatibilité BnSP11-1 permettant à *B. napus* d’être autocompatible, a eu un grand impact sur la reproduction de l’espèce (Gao et al., 2016). De plus, différents événements de recombinaison impliquant des ET ont entraîné la délétion du locus qui contrôle la dureté du grain, dans différents sous-génomes de diverses espèces de blé polyploïdes (Chantret et al., 2005).

Des biais d’expression entre les sous-génomes de certains allopolyploïdes ont été obser-

vés dans différentes espèces ayant eu des WGD récentes. Les biais d'expression entre les sous-génomes ont été reliés dans la littérature à différents mécanismes et notamment aux ET (C. M. Vicient & Casacuberta, 2017). Ainsi, il a été observé que le sous-génome présentant des proportions d'ET les plus importantes était le sous-génome dominé du point de vue de l'expression des gènes (Freeling et al., 2012; C. M. Vicient & Casacuberta, 2017). Ainsi, il a été proposé que les sous-génomes sous-fractionnés et surexprimés sont corrélés à une densité en ET plus faible (Woodhouse et al., 2014). De plus, chez *B. rapa* et *A. thaliana* (Thomas et al., 2006), les auteurs ont signalé que les siRNAs ciblent la région en amont des gènes située préférentiellement dans le sous-génome sous-régulé (Woodhouse et al., 2014). Cette hybridation en amont des séquences des gènes peut provoquer une augmentation du niveau de méthylation, et ainsi modifier le niveau d'expression de la séquence. Par ailleurs les siRNAs peuvent s'hybrider aux ARN messagers via le complexe RNA-Induced Silencing Complex (RISC), et provoquer leur dégradation, ce qui entraîne une baisse du niveau d'expression du gène.

De plus, les ET peuvent influencer l'expression des gènes voisins par des effets épigénétiques. En effet, les ET sont la cible principale des mécanismes de répression génique qui maintiennent leur activité sous un seuil pour éviter de compromettre la viabilité du génome. En conséquence, les ET sont généralement fortement méthylés et sont associés à des marques épigénétiques hétérochromatiques (Freeling et al., 2012; Ito & Kakutani, 2014; West et al., 2014). En effet, l'insertion d'un ET à proximité d'un gène peut générer des marques épigénétiques dans son voisinage et modifier son niveau de transcription (Contreras et al., 2015). Ce mécanisme a été observé chez *A. thaliana* avec le gène régulateur de la floraison FWA (Kinoshita et al., 2007). Pour finir, les ET sont à l'origine de beaucoup de microRNA (miRNAs) (Y. Li et al., 2011; Piriyaopongsa & Jordan, 2008), des répresseurs importants de l'expression des gènes.



## 7.2 Matériel et méthode

Les ET ont été associés à la dominance du sous-génome dans différentes études (Free-ling et al., 2012 ; Woodhouse et al., 2014). Afin de tester cette hypothèse pour le génome du pommier, un pipeline clé en main a été écrit sur la base d’une méthode précédemment décrite (Correa et al., 2021). Ce pipeline est accessible à l’aide du lien suivant : [https://forgemia.inra.fr/tanguy.lallemmand/te\\_analysis](https://forgemia.inra.fr/tanguy.lallemmand/te_analysis). Les ET annotés prédits via REPET (Flutre et al., 2011 ; Quesneville et al., 2005) pour le génome de GDDH13 1.1 ont été filtrés en utilisant la règle 80-80-80. Les séquences annotées en partie comme un ET ont aussi été supprimées. A l’aide de l’annotation des gènes, des régions introniques et exoniques ainsi que celles des ET, une base de données SQL a été construite et interrogée via gffutils. Le pipeline identifie tous les ET filtrés de qualité associés à chaque gène dans les 2 kb en amont et en aval de la séquence du gène ainsi que dans les introns (excluant ainsi les ET exoniques). La densité, la couverture et le nombre d’ET associés à chacun des gènes ohnologues sont ensuite compilés. Cette méthode définit la densité des ET selon l’équation 7.1 et la couverture des ET selon l’équation 7.2.

$$D_g = \frac{N}{L_g + (2 \times L_f) - L_{\text{exonic}} - L_{\text{TEintrinsic}} - L_{\text{TEflanking}}} \quad (7.1)$$

$$C_g = \frac{L_{\text{TEintrinsic}} + L_{\text{TEflanking}}}{L_g + (2 \times L_f) - L_{\text{exonic}}} \times 10^2 \quad (7.2)$$

Où sont définis,

$N$  le nombre d’ET

$L_g$  la taille du gène

$L_f$  la taille des régions flanquantes (2 kb)

$L_{\text{exonique}}$  la taille des régions exoniques du gène considéré

$L_{\text{TEintrinsic}}$  la taille des ET chevauchant les régions introniques

$L_{\text{TEflanking}}$  la taille des ET chevauchant les régions flanquantes en amont et aval des gènes

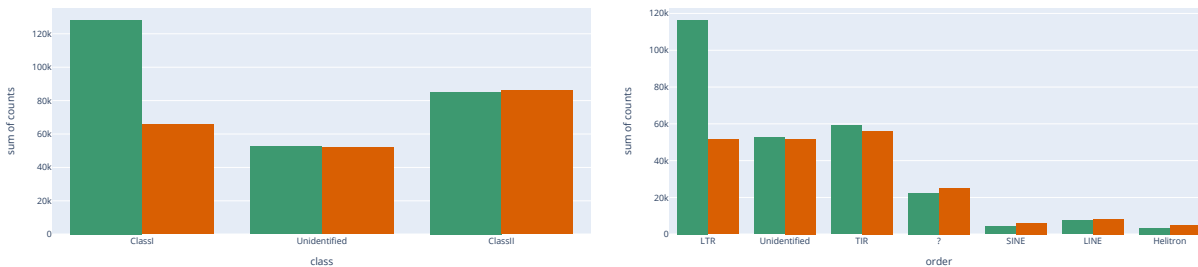
Ces valeurs ont ensuite été comparées à l’aide d’un test de Wilcoxon. Les gènes ont été regroupés selon leur valeur de couverture et de densité à l’aide de différentes méthodes de *clustering* et notamment la méthode de *k-medoid*, *DBSCAN* et *mini-k-batch*.

## 7.3 Résultats

Pour évaluer plus avant si un déséquilibre entre les blocs synténiques pouvait être détecté dans le génome de la pomme, nous avons étudié la distribution des ET au sein du génome du pommier. Pour le génome GDDH13 v1.1, 469 607 ET ont été identifiés. Nous avons associé 265 973 ET à des régions intergéniques, c'est-à-dire non associées à un gène. Nous avons identifié 203 634 associés à des gènes (ohnologues ou non dupliqués par WGD) ou à leur environnement (2 kb en amont ou en aval des gènes).

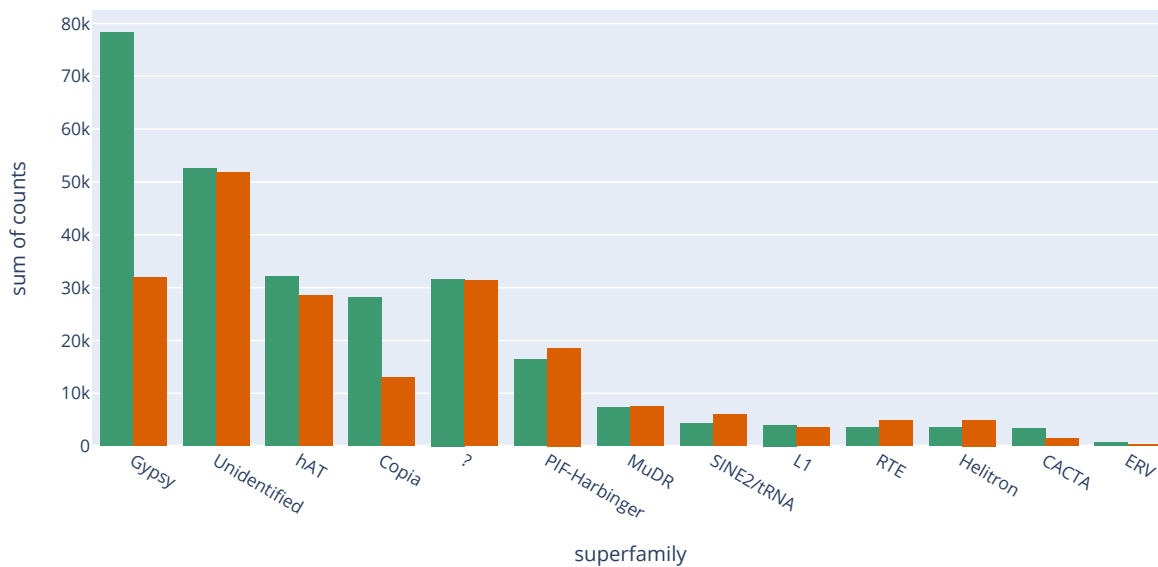
### 7.3.1 Comparaison de l'environnement en ET des régions géniques et intergéniques

Les ET sont annotés et associés à une classe, sous-classe, superfamille et famille d'ET. La Figure 7.1, montre les divers comptages d'ET dans les différentes classes, sous-classes et superfamilles en différenciant les régions intergéniques et géniques. À propos des classes d'ET identifiés au sein de GDDH13, les régions intergéniques présentent significativement plus d'ET de classe I (128 387 contre 65 829, test du  $\chi^2$ , seuil  $\alpha = 10^{-5}$ ) que les régions comportant un gène. Le nombre d'ET non annotés et d'ET de classe II est similaire entre les deux groupes. Au niveau de la sous-classe, les proportions d'ET entre les régions intergéniques et géniques sont significativement déséquilibrées (test du  $\chi^2$ , seuil  $\alpha = 10^{-5}$ ), en particulier en termes de LTR, une sous-classe d'ET de classe I. Au niveau des superfamilles, des différences significatives (test du  $\chi^2$ , seuil  $\alpha = 10^{-5}$ ) ont été observées au niveau des ET appartenant à la sous-classe des LTR. Plus précisément il s'agit d'ET associés aux *gypsy* (78 309 ET associés aux régions intergéniques contre 31 858 associés aux gènes) et *copla* (28 185 contre 13 017). Les ET LINE et SINE sont présents en plus grandes proportions chez les gènes que dans les régions intergéniques. Ainsi nous avons identifié que les régions intergéniques et géniques avaient des proportions d'ET significativement différentes. Les régions intergéniques présentant des proportions plus importantes d'ET de classe I. Afin d'identifier de possibles modifications des régions régulatrices pouvant expliquer les déséquilibres de QTL et d'expression observées dans les chapitres précédents, l'étude d'ET porte sur les ET associés aux régions géniques.



(A) Répartition des classes des ET reliés aux gènes ou aux régions intergéniques. Une différence significative du nombre d’ET de Classe I associés aux régions intergéniques est retrouvée.

(B) Répartition des familles des ET associés aux gènes ou aux régions intergéniques. Une différence significative du nombre d’ET de LTR associés aux régions intergéniques est retrouvée.



(C) Répartition des groupes des ET associés aux gènes ou aux régions intergéniques. Une différence significative du nombre d’ET de *gypsy* et *copia* associés aux régions intergéniques est retrouvée.

**Figure 7.1** – Distribution des proportions des ET associés aux régions intergéniques (en vert) et géniques (en orange) au sein des différentes classes (Figure 7.1A), familles (Figure 7.1B) et sous-familles (Figure 7.1C). L’axe des ordonnées présente le nombre d’ET associés à la catégorie. L’axe des abscisses présente les différentes catégories. Une différence significative de nombre d’ET de classe I intergéniques est retrouvée.

### 7.3.2 Comparaison de l'environnement en ET des gènes ohnologues et des non dupliqués par WGD

Nous avons identifié 203 634 ET associés à des gènes au sein du génome du pommier. Parmi eux 109 069 ET sont associés aux exons de gènes ohnologues et à leur environnement génomique (2 kb en amont et en aval des gènes) et 150 927 ET sont associés aux gènes non dupliqués par WGD et à leur environnement génomique. La distribution du nombre d'ET associé à chacun des groupes de gènes analysés, c'est-à-dire les gènes pour lesquels nous n'avons pas identifié d'ohnologues dans le génome rassemblant les *singletons* et le groupe contenant les gènes ohnologues est présentée en Figure 7.2A. Les distributions du nombre d'ET entre les gènes non dupliqués et les gènes ohnologues ont été testées via le test de rang de Mann-Whitney  $U$ . Globalement nous n'avons pas identifié de différences avec des médianes à 5 ET par gènes pour les deux groupes. À l'échelle des chromosomes, nous avons identifié une différence significative ( $\alpha = 10^{-5}$ ) pour les chromosomes 1, 2, 4, 5, 10 et 15. Néanmoins, on peut observer une différence significative (tests de rang de Mann-Whitney  $U$ ,  $\alpha = 10^{-5}$ ) de couverture en ET entre les deux groupes avec une médiane de couverture d'ET pour les gènes non dupliqués par WGD de 0,40 et 0,24 pour les gènes ohnologues.

En regardant à l'échelle des chromosomes en utilisant des tests de rang de Mann-Whitney  $U$ , pour l'ensemble des chromosomes on observe une différence significative en termes de distribution des ET entre les deux groupes de gènes. Les distributions de la couverture en ET pour les 17 chromosomes du pommier pour le groupe des gènes non dupliqués et celui des gènes ohnologues sont présentées en Figure 7.2B. Un résultat similaire a été observé pour la densité des ET (Figure 7.2C) et la longueur d'ET (Figure 7.2A). L'ensemble des résultats des tests de rang de Mann-Whitney  $U$  est présenté dans la Table 7.1. Ces résultats montrent que les gènes présents en une seule copie et les gènes pour lesquels nous avons pu identifier un ohnologue présentent un nombre globalement proche d'ET insérés. Néanmoins, les ET associés aux gènes non dupliqués sont en général plus longs que ceux associés aux gènes ohnologues.

Nous avons comparé les proportions d'ET associés aux différentes classes, sous-classes et superfamilles entre les gènes ohnologues et les gènes présents en une seule copie. La Figure A.30 en annexe présente ces différentes proportions. Les proportions d'ET associées aux classes (Figure A.30A) sont significativement différentes (test du  $\chi^2$ , seuil  $\alpha = 5\%$ ), les gènes présents en une seule copie étant associés à plus d'ET de Classe I et II,

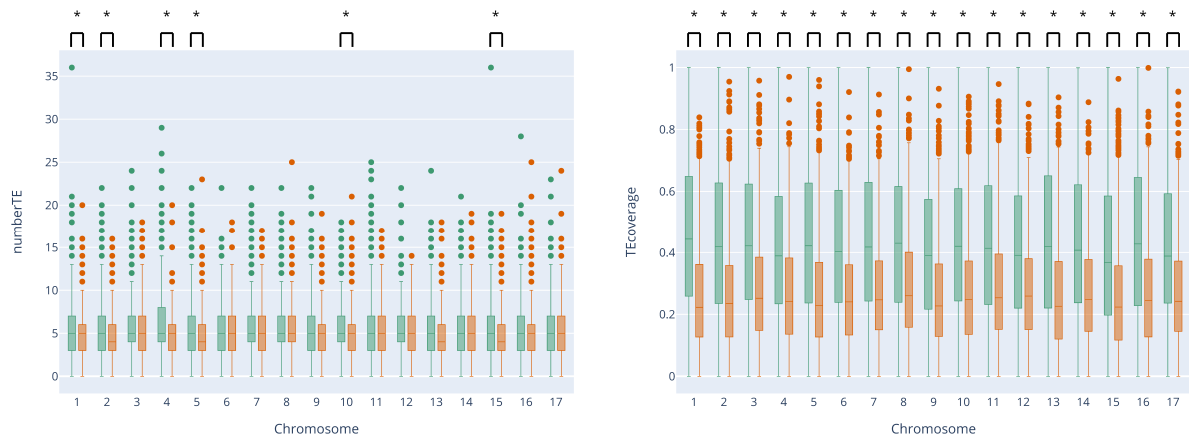
que ce soit globalement ou à l'échelle des chromosomes. La distribution des proportions d'ET dans les différentes sous-classes (Figure A.30B) est aussi significativement différente, particulièrement pour les ET de type TIR, *Helitrons* et LTR. Quant aux superfamilles (Figure A.30C), leurs proportions sont significativement différentes (test du  $\chi^2$ , seuil  $\alpha = 5\%$ ), plus particulièrement pour les ET hAT, *Helitrons*, Gypsy et Copia.

**Table 7.1** – Résultats des tests de rang de Mann-Whitney  $U$  pour la couverture en ET, la densité en ET et le nombre ET et leur longueur entre les gènes ohnologues et les gènes non dupliqués.

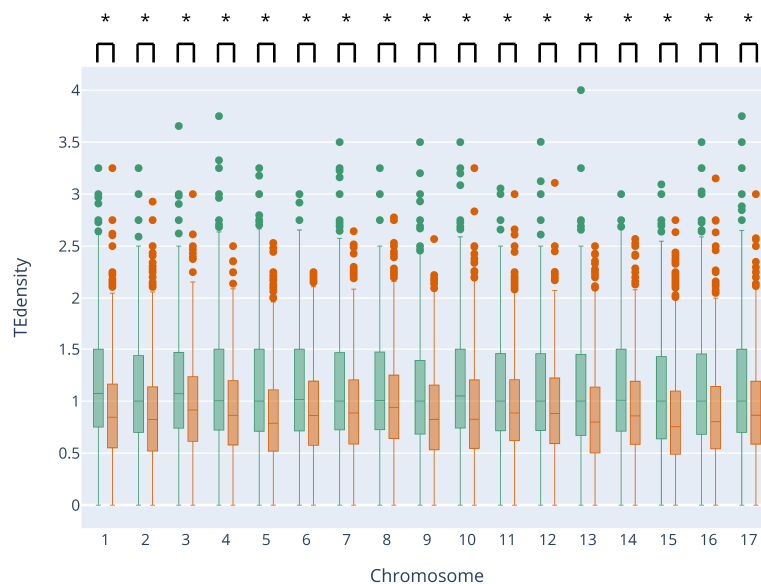
Chromosome	p-value Couverture ET	p-value Densité ET	p-value Nombre ET	p-value Longueur ET
1	$1,52 \times 10^{-87}$	$5,50 \times 10^{-28}$	$2,78 \times 10^{-7}$	$6,70 \times 10^{-55}$
2	$1,33 \times 10^{-91}$	$2,11 \times 10^{-30}$	$3,14 \times 10^{-9}$	$6,99 \times 10^{-59}$
3	$1,26 \times 10^{-76}$	$3,67 \times 10^{-21}$	$7,86 \times 10^{-2}$	$7,34 \times 10^{-38}$
4	$7,06 \times 10^{-51}$	$1,25 \times 10^{-20}$	$4,73 \times 10^{-12}$	$7,10 \times 10^{-37}$
5	$3,23 \times 10^{-99}$	$2,25 \times 10^{-42}$	$1,01 \times 10^{-9}$	$4,30 \times 10^{-55}$
6	$2,98 \times 10^{-60}$	$4,17 \times 10^{-18}$	$2,32 \times 10^{-4}$	$1,73 \times 10^{-37}$
7	$1,96 \times 10^{-80}$	$2,76 \times 10^{-18}$	$9,26 \times 10^{-4}$	$1,40 \times 10^{-48}$
8	$1,23 \times 10^{-52}$	$1,65 \times 10^{-9}$	$4,96 \times 10^{-1}$	$9,51 \times 10^{-28}$
9	$2,22 \times 10^{-62}$	$4,93 \times 10^{-19}$	$9,19 \times 10^{-3}$	$2,51 \times 10^{-32}$
10	$5,95 \times 10^{-88}$	$1,33 \times 10^{-34}$	$4,15 \times 10^{-8}$	$1,22 \times 10^{-50}$
11	$9,89 \times 10^{-63}$	$8,17 \times 10^{-17}$	$1,65 \times 10^{-1}$	$2,06 \times 10^{-30}$
12	$1,39 \times 10^{-48}$	$7,31 \times 10^{-16}$	$8,94 \times 10^{-2}$	$4,73 \times 10^{-24}$
13	$4,02 \times 10^{-77}$	$2,87 \times 10^{-21}$	$2,32 \times 10^{-4}$	$2,37 \times 10^{-47}$
14	$5,08 \times 10^{-61}$	$6,10 \times 10^{-18}$	$8,77 \times 10^{-3}$	$1,72 \times 10^{-31}$
15	$5,79 \times 10^{-88}$	$3,16 \times 10^{-38}$	$8,05 \times 10^{-12}$	$7,09 \times 10^{-54}$
16	$4,46 \times 10^{-73}$	$2,36 \times 10^{-24}$	$3,85 \times 10^{-4}$	$3,36 \times 10^{-40}$
17	$4,85 \times 10^{-66}$	$6,16 \times 10^{-20}$	$3,58 \times 10^{-3}$	$1,12 \times 10^{-35}$

### 7.3.3 Comparaison des types d'ET entre les fragments synténiques ohnologues

Les annotations des ET associés aux gènes ohnologues ont été étudiées. Leurs proportions sont représentées sous la forme de trois diagrammes en secteurs en Figure 7.3. Le premier diagramme représente les classes d'ET (Figure 7.3A). Les ET de classe II sont majoritaires et représentent 40,4% de tous les ET (68 832 ET), les ET de classe I représentent 27,7% (47 203 ET), tandis que la dernière classe est constituée d'ET « inconnus » et représente 31,9% (54 378 ET). Les sous-classes sont présentées en Figure 7.3B. Elles



(A) Distribution du nombre d'ET associé aux gènes ohnologues (distribution orange) et les gènes non dupliqués (distribution verte) par WGD. (B) Distribution de la couverture en ET associés aux gènes ohnologues (distribution orange) et les gènes non dupliqués (distribution verte) par WGD.



(C) Distribution de la densité en ET associés aux gènes ohnologues (distribution orange) et les gènes non dupliqués (distribution verte) par WGD.

**Figure 7.2** – Diagrammes en boîte de la distribution du nombre d'ET et de la couverture en ET entre les gènes non dupliqués et les gènes ohnologues pour chacun des chromosomes. L'axe des y représente l'ensemble des valeurs du nombre d'ET ou la couverture en ET associés aux catégories de gènes. L'axe des x indique les chromosomes considérées. Le bas de la boîte représente le premier quartile, le trait interne à la boîte présente la médiane de la distribution et l'extrémité supérieur de la boîte présente le troisième quartile. La taille de la boîte représente l'écart interquartile. La moustache inférieure présente le neuvième percentile et la moustache présente quant à elle le 91<sup>e</sup>. Pour chaque chromosome, les diagrammes en boîte indiquent la distribution pour les gènes non dupliqués (boîte verte) et la distribution pour les gènes ohnologues (boîte orange). \* tests de rang de Mann-Whitney  $U$ ,  $\alpha$  inférieur à  $10 \times 10^{-5}$

sont principalement représentées par des ET TIR (classe II) et des LTR (classe I). Les familles sont représentées en Figure 7.3C et sont composées en majorité par hAT (14,5 %, classe II), *gypsy* (11,3 %, classe I), PIF-Harbinger (10,8 %, classe II) et SINE2/tRNA (4,8 %, classe I). Les comptages exacts d’ET sont présentés dans la Table 7.2.

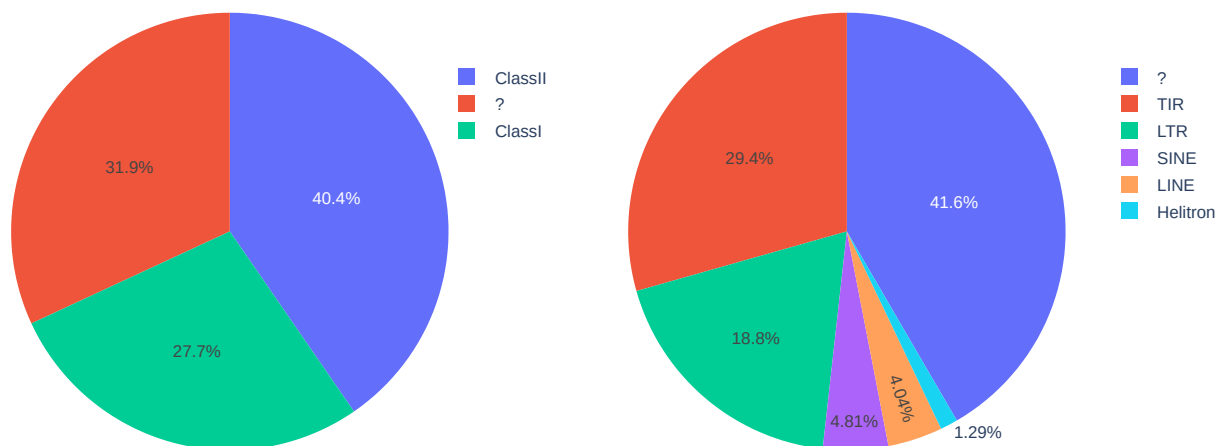
**Table 7.2** – Proportion des annotations des ET associés aux gènes ohnologues

Classe ET	Famille ET	Groupe ET	Nombre ET
ClassI	LINE	L1	2555
ClassI	LINE	RTE	4330
ClassI	LTR	Copia	6607
ClassI	LTR	ERV	52
ClassI	LTR	Gypsy	19 282
ClassI	SINE	SINE2/tRNA	8205
ClassII	Helitron	Helitron	2204
ClassII	TIR	CACTA	664
ClassII	TIR	MuDR	6293
ClassII	TIR	PIF-Harbinger	18 427
ClassII	TIR	hAT	2468

Les classes, sous-classes et superfamilles d’ET présentent de grandes différences dans le fonctionnement notamment dans les modes de transpositions détaillés en introduction de ce chapitre (Boeke et al., 1985 ; Brown et al., 1987 ; Greenblatt & Alexander Brink, 1963 ; Luan et al., 1993). Ainsi, nous avons testé si les proportions d’ET associées à chacune des familles étaient similaires à l’aide de tests de  $\chi^2$ . Au niveau des classes, les paires 9-17 et 8-15 sont significativement différentes. La paire 8-15 présente une proportion significativement différente d’ET de classe I sur le chromosome 15 par rapport au chromosome 8. Pour la paire 9-17, le chromosome 17 a significativement plus d’ET de classe II. Les paires 2-7 et 2-15 sont significativement différentes au niveau des sous-familles d’ET. Nous avons observé des proportions plus élevées de *copia* et de PIF-Harbinger pour le chromosome 2 par rapport au chromosome 7. Pour la paire 2-15, le chromosome 2 a accumulé une proportion significativement plus élevée de PIF-Harbinger que le chromosome 15.

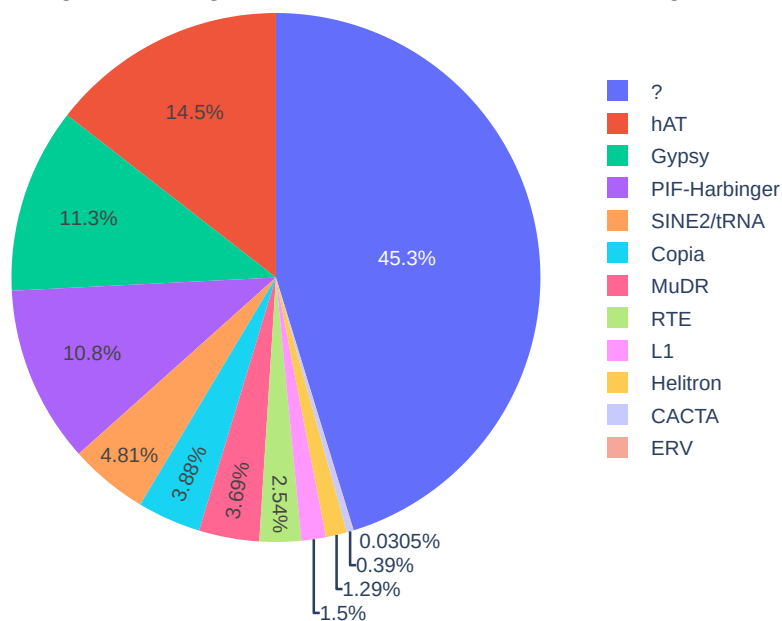
### 7.3.4 Comparaison de l’environnement en ET entre les fragments synténiques ohnologues

Nous avons comparé les environnements en ET des fragments synténiques. La moyenne de la taille d’ET associés aux gènes ohnologues est de 572 bp avec un écart type de 1345 bp.



(A) Diagramme circulaire de la répartition des classes des ET associés aux gènes ohnologues

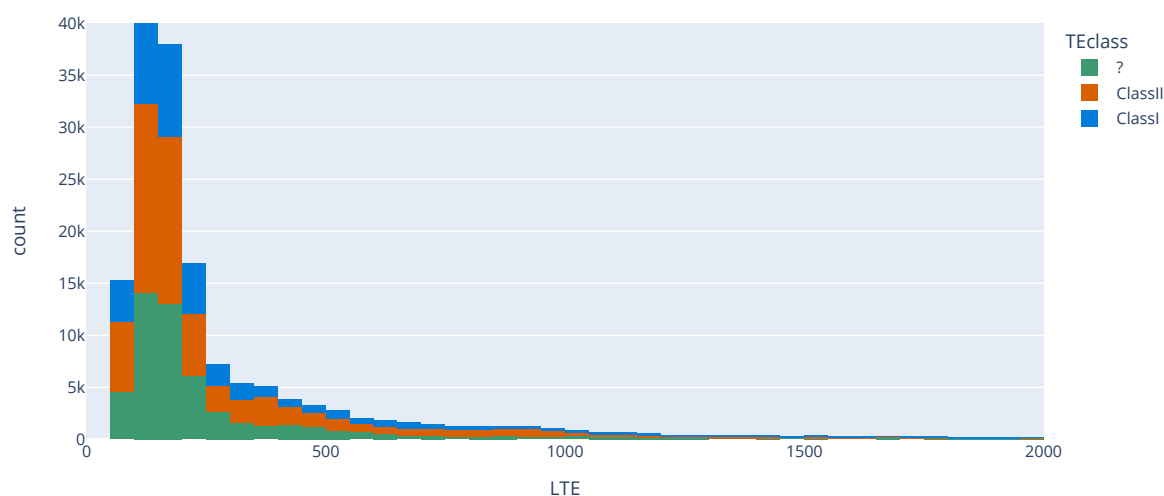
(B) Diagramme circulaire de la répartition des familles des ET associés aux gènes ohnologues



(C) Répartition des sous-familles des ET associés aux gènes ohnologues

**Figure 7.3** – Diagrammes circulaires de la répartition des classes (Figure 7.3A), familles (Figure 7.3B) et sous-familles (Figure 7.3C) des ET associés aux gènes ohnologues. La notation " ? " correspond à des ET pour lesquels aucune classification n'a été identifié. La couleur permet de différencier les secteurs associés aux différentes méta-données. La taille des secteurs permet de représenter visuellement la proportion d'expériences associées à cette métadonnée. Le pourcentage exact est ajouté dans chacun des secteurs.





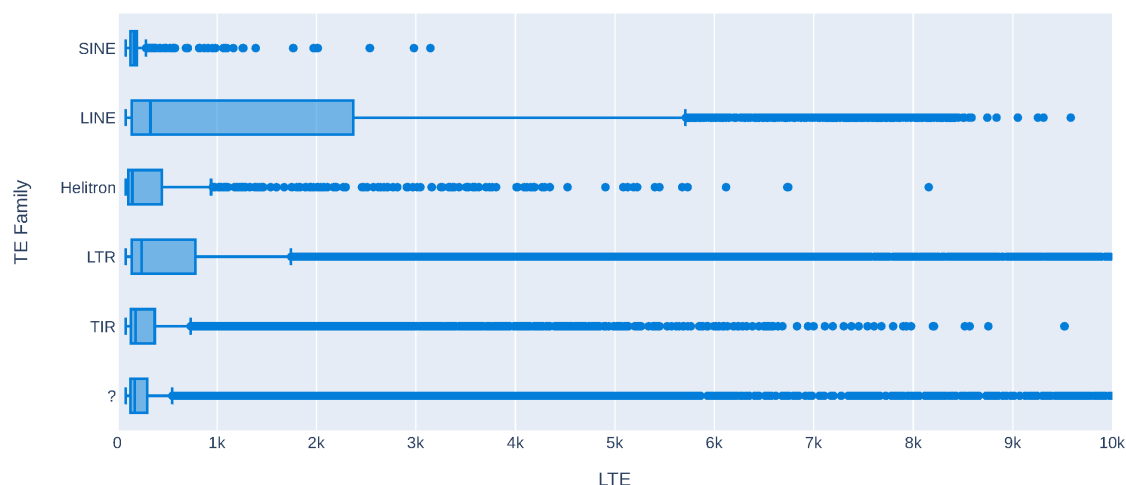
**Figure 7.4** – Distribution de la taille des ET associés aux gènes ohnologues. L’axe des abscisses présente la longueur de l’ET en nombre de bases et l’axe des ordonnées, le nombre d’ET associés à cette longueur. La couleur est associée à la classe d’ET considérée.

La valeur minimale est de 81 bp ce qui correspond à notre filtre de la règle des 80, et le ET le plus grand mesure 46 331 bp. La distribution de la taille des ET associés aux gènes ohnologues en fonction de la classe des ET est présentée en Figure 7.4. Les ET de classe II sont plutôt regroupés autour de la médiane de la distribution tandis que les ET de classe I sont plutôt sur les extrêmes de la distribution.

Outre le mode de duplication des ET qui diffère, l’une des différences importantes est la taille des ET de ces différentes familles. La visualisation de la distribution de la taille des ET en fonction des sous-classes d’ET associés est présentée en Figure 7.5. Le détail des valeurs décrivant la taille des ET est présenté en Table 7.3. Les ET appartenant à la famille des LINE sont les plus grands ET retrouvés comme associés à des gènes ohnologues avec une taille médiane de 331 bp. Viennent ensuite les ET appartenant au LTR (médiane de 242 bp) puis TIR (182 bp), SINE (166 bp), *Helitrons* (149 bp). La taille de LTR présente une médiane petite puisque la taille attendue est plutôt entre 250 et 500 bp ce qui suggère qu’un certain nombre de LTR identifiés ont déjà divergé. Les ET non annotés en termes de sous-classes présentent une taille médiane de 173 bp. Nous avons testé si la distribution de la taille des ET entre les fragments synténiques ohnologues était similaire à l’aide de tests de rang de Mann-Whitney  $U$ . Les résultats sont présentés en Table 7.4. On a identifié

les paires 02-15, 06-14, 01-07 et 08-15 comme significativement différentes en termes de distribution taille des ET associés à ces paires.

Pour la paire 2-15, le chromosome 15 présente les ET les plus longs avec une moyenne de 633 bp contre 481 bp pour le chromosome 2. Pour la paire 8-15, c'est le chromosome 8 qui présente les ET les plus longs (moyenne de 609 bp contre 552 bp). Pour la paire 6-14, le chromosome 14 présente les ET les plus longs (568 bp contre 545 bp). Concernant la paire 1-7, c'est le chromosome 7 qui présente une moyenne de taille d'ET plus importante. Ainsi, nous avons identifié un déséquilibre dans la taille des ET associés aux chromosomes. Ce déséquilibre dans la taille des ET est probablement lié, au moins en partie, au déséquilibre de proportion des familles d'ET associées aux paires de fragments chromosomiques synténiques.



**Figure 7.5** — Boîtes à moustache représentant la distribution de la taille des ET en fonction des sous-classes d'ET. L'axe des ordonnées présente les différentes sous-classes de TEs considérées. L'axe des abscisses présente la distribution de la longueur des TEs en nombre de bases. Le bas de la boîte présente le premier quartile, le trait interne à la boîte présente la médiane de la distribution et l'extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l'écart interquartile. La moustache inférieure présente le neuvième percentile et la moustache supérieure présente le 91<sup>e</sup>.

Pour tester les différences d'ET dans les régions exoniques et l'environnement des gènes ohnologues (2 kb en amont et en aval), nous avons testé les distributions de couverture en ET (Équation 7.2) entre les gènes ohnologues pour les principales paires de fragments de chromosomes synténiques en utilisant un test de Wilcoxon. Les résultats sont décrits

**Table 7.3** – Description de la taille des ET en fonction des différentes sous-classes associées aux gènes ohnologues

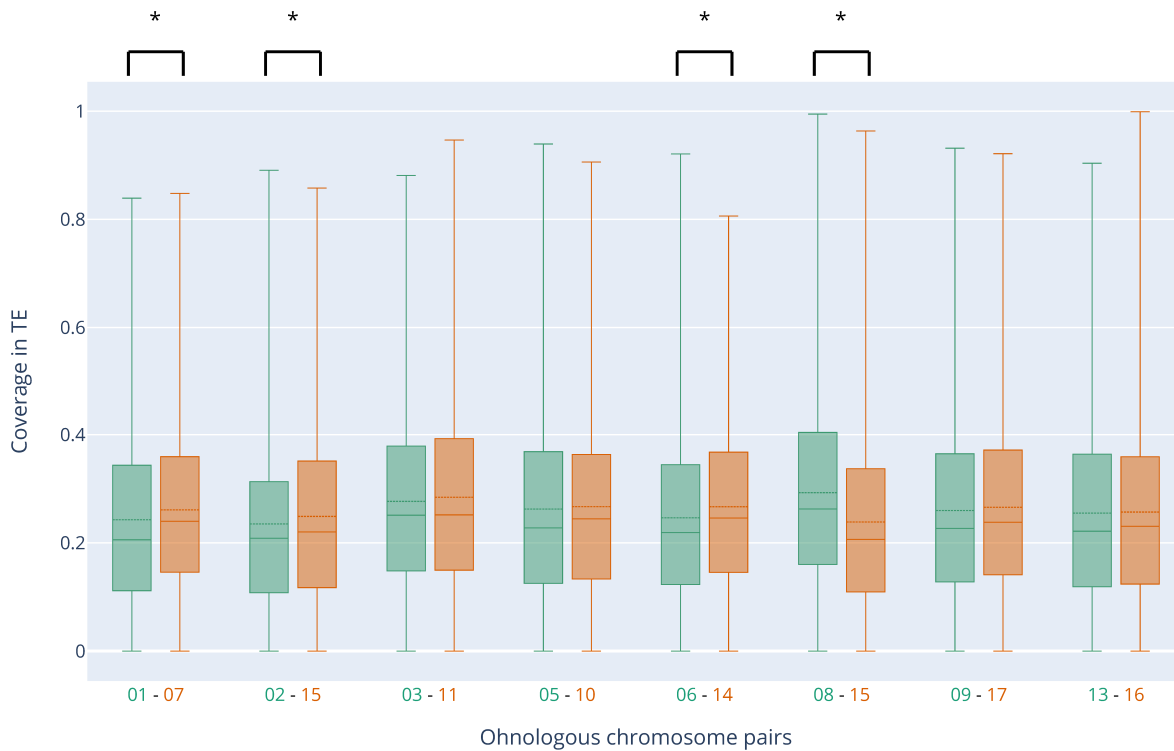
Famille d'ET	Nombre de ET	Moyenne	Écart type	min	25 %	50 %	75 %	max
?	70 915	418,26	1028,29	81	130	173	298	46 331
Helitron	2204	474,50	994,0055	81	107	149	447	15 826
LINE	6885	1759,91	2564,41	81	144	331	2370	21 135
LTR	32 113	1061,50	2060,82	81	144	242	784	36 412
SINE	8205	175,65	240,97	81	129	166	192	10 943
TIR	50 091	392,99	742,70	81	133	182	374	26 916

**Table 7.4** – Résultats des tests de rang de Mann-Whitney  $U$  de la taille des ET entre les fragments synténiques

couple	p-value Mann Whitney $U$	Taille moyenne premier chromosome	Taille moyenne second chromosome
02-15	<0,001	481	633
06-14	0,001	545	568
01-07	0,012	529	534
08-15	0,030	609	552
09-17	0,126	612	533
03-11	0,038	537	601
05-10	0,165	613	587
13-16	0,249	606	603
04-12	0,495	541	549

sous forme de boîtes à moustaches en Figure 7.6. On peut observer que les paires 1-7, 2-15, 6-14 et 8-15 sont significativement déséquilibrées ( $\alpha = 5\%$ ) en termes de couverture en ET. Pour la paire 1-7 on observe une accumulation plus importante d'ET dans le fragment synténique associé au chromosome 7 avec une couverture médiane de 0,21 pour le chromosome 1 contre 0,24 pour le chromosome 7. Pour la paire 2-15, c'est le chromosome 15 qui est enrichi en ET avec une médiane 0,22 contre 0,21 pour le chromosome 2. Pour la paire 6-14, le chromosome 14 a accumulé plus d'ET avec une couverture d'ET de médiane de 0,25 contre 0,22 pour le chromosome 14, et la paire 8-15 où le chromosome 8 présente une médiane 0,26 contre 0,21 pour le chromosome 15. Des résultats similaires ont été observés pour la densité en ET.

En résumé, cette analyse nous a permis d'identifier un ensemble d'indicateurs liés aux ET associés aux gènes ohnologues. Pour les paires 1-7, 2-15, 6-14 et 8-15, nous avons pu montrer que la couverture et la densité en ET sont significativement différentes. Ce biais semble lié aux ET dans leur ensemble plutôt qu'à une classe ou une sous-famille particulière.



**Figure 7.6** – Box plot de la distribution de la couverture en ET pour chaque paire de chromosomes ohnologues sur l’axe des x, les paires de chromosomes ohnologues considérées sont présentées. Le nombre de paires de gènes est indiqué entre parenthèses. L’axe des y représente les valeurs de couverture en ET. Le bas de la boîte présente le premier quartile, le trait interne à la boîte présente la médiane de la distribution, le trait en pointillé la moyenne et l’extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l’écart interquartile. La moustache inférieure présente le neuvième percentile et la moustache supérieure présente quant à elle le 91<sup>e</sup>. La première boîte à moustaches représente la distribution de la couverture ET sur le premier chromosome de la paire (en vert), tandis que la seconde boîte à moustaches représente la distribution de la couverture ET sur le second chromosome de la paire (en orange). \*P-value < 0,05.

## 7.4 Discussion

Nous avons mené un ensemble d'analyses basé sur les ET annotés pour le génome de GDDH13 1.1 avec le logiciel REPET. Cette annotation bien que de bonne qualité présente 78 943 séquences identifiées comme des ET (avec une p-value inférieure au seuil de 5 % selon REPET), mais pour lesquelles le logiciel n'est pas arrivé à assigner la classe d'ET. Les ET présentent une grande diversité de mécanisme avec des implications différentes pour le génome hôte suivant leur classe, sous-classe et superfamille. De cette façon il a été observé chez le riz (J. Zhang et al., 2015) et le blé tendre (Yaakov & Kashkush, 2011) que les déséquilibres de méthylation des ET étaient reproductibles (Yaakov & Kashkush, 2011) et dépendaient de la famille des ET. Ainsi les rétrotransposons ont été identifiés chez ces deux espèces comme liés à des hypométhylations tandis que les éléments de classe II étaient associés à une hyperméthylation. Or l'absence d'information sur les classes pour 29 % des ET dans nos données rend plus difficile une interprétation fine de nos résultats. Malheureusement, une réannotation des ET afin d'en améliorer le résultat ne peut être faite sans une grande expertise puisque REPET nécessite une curation manuelle importante. Toutefois, la taille des ET inconnus, indique qu'il s'agit de séquences principalement courtes, qui ne présentent peut-être pas suffisamment d'informations pour présenter une annotation fiable. Par ailleurs, le pipeline d'analyse des ET utilisé (disponible à l'adresse suivante : [https://forgemia.inra.fr/tanguy.lallemand/te\\_analysis](https://forgemia.inra.fr/tanguy.lallemand/te_analysis)) a été pensé pour être utilisable sur différents organismes et a été utilisé chez le rosier.

### 7.4.1 Comparaison de l'environnement en ET

La comparaison de l'environnement en ET des gènes par rapport à celui des séquences intergéniques a montré un déséquilibre des ET de classe I. Ainsi, les régions intergéniques sont enrichies en LTR, en particulier des ET de la superfamille des *gypsy* et des *copia*. Les proportions d'ET associés aux autres classes, sous-classes et superfamilles n'ont pas montré un déséquilibre aussi marqué. Cet enrichissement en ET de classe I en intergénique peut s'expliquer par son mode de duplication par « copier — coller » qui permet une croissance rapide du nombre de copies dans le génome, plus particulièrement dans les zones avec peu de pression de sélection comme les régions intergéniques. De plus, ce mode de duplication laisse les copies existantes intactes lors de la génération de nouvelles copies, ce qui permet une détection facilitée de ces copies par les outils d'annotations, au moins pour les copies complètes. Dans notre analyse nous avons identifié une médiane de taille des

LTRs plus courte qu’attendue, ce qui indique un ensemble de copies dégradées et nuance ainsi l’existence d’un potentiel biais lié à ce mode de duplication. L’enrichissement d’ET de classe I est en adéquation avec les modèles d’évolution des ET (Kidwell & Lisch, 1997) qui décrivent deux types d’insertions, des ET s’insérant loin des gènes et des ET s’insérant à proximité de gènes. L’environnement de recombinaison dans lequel se trouve un ET peut avoir un impact sur l’efficacité de la sélection naturelle sur l’ET, car une recombinaison élevée peut dissocier la variation délétère des mutations adaptatives (Hill & Robertson, 1966). Ainsi, les rétrotransposons LTR ont été retrouvés chez le maïs plus souvent dans les régions à faible recombinaison (Stitzer et al., 2021), et seulement pour 3,5 % d’entre eux au sein d’un transcrit, à une distance médiane d’un gène synténique de 31,9 kb. Un constat similaire est fait chez le pommier où il a été constaté un enrichissement en ET de classe I (principalement *copia* et *gypsy*) dans les régions intergéniques, présentant de faibles taux de recombinaison (Wright et al., 2003). À l’inverse, chez le maïs, les *Helitrons*, les éléments TIR et rétrotransposons non LTR comme les LINE et les SINE se trouvent plus souvent dans les régions à forte recombinaison (Stitzer et al., 2021). Chez le pommier il a été observé que les ET s’insérant à proximité de gènes sont plutôt des ET de Classe II appartenant aux sous-groupes LINE et SINE et *Helitrons*, et aux superfamilles de PIF-Harbringer, SINE2 et *Helitrons* (Daccord et al., 2017).

#### 7.4.2 Comparaison de l’environnement en ET des gènes ohnologues et des non dupliqués par WGD

La distribution du nombre d’ET dans le voisinage des gènes ohnologues et les gènes présents en une seule copie est significativement différente pour six chromosomes, à savoir les chromosomes 1,2,4,5,10 et 15. En testant la distribution de la longueur des ET, de la couverture et de la densité en ET entre les gènes ohnologues et les gènes *singletons*, nous avons identifié l’ensemble des chromosomes comme significativement différents. Par ailleurs, nous avons identifié une différence de longueur d’ET associée aux gènes ohnologues et aux gènes non dupliqués. La différence de longueur d’ET est probablement due, en partie, à une différence dans la nature des ET associés à ces groupes de gènes. En effet, nous avons identifié un déséquilibre dans les proportions des classes d’ET associés aux gènes ohnologues par rapport aux gènes non dupliqués. Les gènes non dupliqués ont accumulé plus d’ET de classe I, en particulier des LTR appartenant à la superfamille des *copia* et *gypsy*. Or ces sous-classes d’ET appartiennent aux ET les plus longs, comme ob-

servé en Figure 7.5. Des résultats similaires ont été trouvés dans d'autres organismes tels que *A. thaliana* (A. L. Hughes et al., 2003) ou *Homo sapiens* (Correa et al., 2021). Ces observations suggèrent l'importance des ET dans la structuration du génome post-WGD. En effet, les proportions d'ET vont générer des zones de méthylation élevées ou faibles, et des zones hétérochromatiques avec des taux de recombinaison plus faibles (C. M. Vicient & Casacuberta, 2017). De plus, il semble que la différence de distribution des ET entre les gènes non dupliqués et les gènes ohnologues puisse avoir un impact sur le devenir des gènes dupliqués et en particulier sur leur rétention (C. M. Vicient & Casacuberta, 2017). Ainsi ces différentes observations pourraient expliquer en partie la différence de pression de sélection appliquée à ces séquences. En effet, les ET sont moins insérés dans les régions sous pression de sélection (Correa et al., 2021 ; Simons et al., 2006). Néanmoins, étant donné l'approche utilisée pour calculer les ratios  $\omega$ , cette hypothèse ne peut être testée par les données générées au cours de l'étude. Outre, cette pression de sélection, nous n'avons pas identifié une grande différence dans le nombre d'ET insérés entre les gènes ohnologues et le groupe des gènes pour lesquels la paire n'a pas pu être reconstruite. La différence pourrait aussi d'être liée à la longueur des ET insérés, qui va augmenter la couverture sans avoir de différence majeure dans le nombre d'insertions. Cette différence résulte ainsi d'une différence significative de la densité en ET pour l'ensemble des chromosomes du pommier ;

### 7.4.3 Comparaison des types d'ET entre les fragments synténiques ohnologues

Au sein du génome de la pomme, nous n'avons pas observé de différences importantes dans la distribution des ET associés à certaines classes. Celles-ci ne sont significatives que pour quelques paires. Néanmoins, ces différences pourraient expliquer en partie la différence de taille d'ET entre les paires de fragments chromosomiques synténiques. En effet, les paires de chromosomes 02-15, 06-14, 01-07 et 08-15 sont significativement différentes en termes de taille des ET associés à ces paires. Une partie de ces paires et notamment, les paires 2-15 et 8-15 présentent aussi un déséquilibre dans les proportions d'ET associés à la classe I. Pour ces paires déséquilibrées, l'impact est important. En effet, il a été montré chez le maïs que dans l'ensemble des tissus testés, les gènes proches des éléments TIRs et non LTR ont une expression médiane plus élevée que les gènes proches des LTRs et des *Helitrons*. Cette observation s'accroît si les ET sont proches du gène (<1 kb) (Stit-



zer et al., 2021). Au sein de ces paires, ces différences pourraient expliquer en partie le déséquilibre d'expression observé.

Cependant, dans différentes espèces, telles que *A. thaliana* (M. Chen et al., 2008), la différence d'ET entre les sous génomes n'est pas liée à la différence de proportions au sein de classes particulières d'ET, mais à la différence dans toutes les classes d'ET. Dans la littérature, des différences dans la distribution de certaines classes d'ET ont été observées, notamment dans les organismes dont le génome a été affecté par la WGD et en particulier les génomes à dominance sous-génomique comme le maïs, avec une dynamique particulière des LTR (Classe I) et en particulier des *gypsy*, *copia* (Estep et al., 2013) et *B. rapa* avec une dynamique de différents types d'ET (An et al., 2014). Ces différences entraînent des différences de méthylation de l'ADN dans les sous-génomes.

#### 7.4.4 Comparaison de l'environnement en ET entre les fragments synténiques ohnologues

Nous avons identifié un ensemble de paires de fragments synténiques ohnologues comme significativement différent en termes de couverture en ET. Une analyse similaire a été menée sur la densité en ET et présente des résultats similaires. Ainsi, les paires 1-7, 2-15, 6-14 et 8-15 apparaissent comme significativement différentes en termes de couverture en ET. Cette différence d'accumulation d'ET entre les segments synténiques a été observée chez de nombreuses espèces (Garsmeur et al., 2014; Woodhouse et al., 2014). Le déséquilibre a été décrit comme étant un mécanisme clé pour la mise en place de la dominance de sous-génome (Alger & Edger, 2020; C. M. Vicent & Casacuberta, 2017). En effet, l'activation des ET post-WGD par levée de la méthylation des ET permet une explosion du nombre d'ET. La répartition des ET est alors biaisée lors de l'explosion aboutissant à un ensemble de processus menant à la mise en place d'une dominance de sous-génome. En effet, la présence élevée d'ET peut entraîner une méthylation élevée de l'ADN dans ces régions. Ainsi, une distribution biaisée des ET pourrait provoquer des déséquilibres dans la méthylation de l'ADN. Un mécanisme similaire semble se mettre en place chez le pommier à l'échelle des chromosomes. En effet la couverture en ET et la densité en ET des paires de gènes ohnologues est significativement différente pour un ensemble de paires de chromosomes ohnologues. Pour ces paires le déséquilibre est dans le sens inverse des déséquilibres observés dans les chapitres précédents. Ce constant est similaire à celui reporté dans la littérature à savoir le sous-génome ayant accumulé le plus d'ET est le

sous-génomme dominé.

## 7.5 Conclusion

À l'aide d'un pipeline *ad-hoc* nous avons pu filtrer et relier les ET à l'ensemble des gènes du pommier et calculer différents indicateurs dont la couverture et la densité. De même, les annotations des ET en termes de classe, superfamille et famille ont été associés aux différents gènes. En comparant les ET associés aux gènes par rapport aux séquences intergéniques, nous avons pu observer un déséquilibre principalement pour les ET de classe I et notamment les LTR, en particulier les *gypsy* et les *copia*. Ces ET ont été associés dans différentes études à des processus de maintenance structurelle des génomes dupliqués et en particulier, la construction des génomes post-WGD. Par ailleurs, en testant la distribution de la longueur des ET, de la couverture et de la densité en ET entre les gènes ohnologues et les gènes dont la paire synténique n'a pu être reconstruite, nous avons identifié l'ensemble des chromosomes comme significativement différents. Ces différences sont en partie liées à un déséquilibre significatif des ET de classe I, en particulier des LTR appartenant à la superfamille des *copia* et *gypsy*. En comparant les paires de fragments synténiques, nous avons identifié une dynamique biaisée des ET qui semble liée aux ET dans leur ensemble plutôt qu'à une classe ou une sous-famille particulière. Ainsi, les paires 1-7, 2-15, 6-14 et 8-15 apparaissent comme significativement différentes en termes de couverture en ET. La présence élevée d'ET peut entraîner une méthylation élevée de l'ADN dans ces régions. Ainsi, une distribution biaisée des ET pourrait provoquer des déséquilibres dans la méthylation de l'ADN. Par ailleurs, pour l'ensemble des paires significativement déséquilibrées, le déséquilibre est cohérent avec le déséquilibre observé en termes de QTLs, à savoir le chromosome dominant du point de vue de la participation à la variation phénotypique présente une couverture significativement moins importante de ET par rapport à son fragment chromosomique ohnologue.



# ÉTUDE DES MÉTHYLATIONS DE L'ADN DES GÈNES OHNOLOGUES

---

## 8.1 Introduction

Chez les organismes eucaryotes, différents mécanismes permettent la régulation du niveau d'expression des gènes. Les modifications épigénétiques sont un des acteurs de cette régulation. Ces modifications sont principalement retrouvées sous la forme de modifications post-traductionnelles des histones (Vaillant & Paszkowski, 2007), ainsi que sous la forme de méthylation de l'ADN (Bártová et al., 2008).

La régulation de la transcription des gènes par la chromatine est réalisée via les nucléosomes. Ces octamères protéiques sont composées de 8 histones : 2 H2 A, 2 H2 B, 2 H3 et 2 H4. Les modifications post-traductionnelles des histones peuvent avoir un impact sur le niveau de compactage de l'ADN. On retrouve quatre grands types de modifications chimiques : les acétylations, les méthylations, les phosphorylations et les ubiquitinylation. Ces modifications, désignées sous le nom de code histone (Jenuwein & Allis, 2001), constituent un signal dont les effets sont divers (Roudier et al., 2011 ; Sequeira-Mendes et al., 2014 ; X. Zhang et al., 2009) . Ces mécanismes épigénétiques sont très conservés chez les eucaryotes (Feng et al., 2010) et peuvent avoir un impact sur le niveau d'expression des gènes (Razin & Cedar, 1991). Chez les plantes, on retrouve deux catégories de modifications d'histones, la première formant l'euchromatine qui favorise l'accessibilité des facteurs de transcription et l'hypométhylation des gènes favorisant ainsi l'expression des gènes (Richards & Elgin, 2002). C'est le cas par exemple des modifications H3K4me3 et H3K27ac (Jenuwein & Allis, 2001). La présence d'ADN sous forme d'hétérochromatine a été reliée à des modifications comme H3K27me3 et H3K9me2 (Maison et al., 2002). La présence d'ADN sous forme d'hétérochromatine va favoriser une hyperméthylation de l'ADN et une répression des gènes (Volpe et al., 2002).

La méthylation de l'ADN consiste en l'ajout, par le biais de méthyl-transférases, de

groupement méthyl sur des cytosines dans un contexte nucléotidique particulier. Différentes méthyl-transférases ont été identifiées chez les eucaryotes (Feng et al., 2010; Zemach et al., 2010), dont certaines sont spécifiques à des organismes en particulier (Oliveira et al., 2020; R. J. Wood et al., 2007). Certaines méthyl-transférases vont être responsables de la méthylation *de novo*, d'autres du maintien de la méthylation (Zemach et al., 2010). Chez les plantes, la méthylation *de novo* pour tous les contextes est permise par DRM2 (Domains Rearranged Methyltransferase 2) (H. Zhang et al., 2018). Le maintien des méthylations est permis par DRM2 ainsi que MET1 (méthyl-transférase 1) et CMT3 (chromo-méthylase 3) (Cao & Jacobsen, 2002).

Les méthylations de l'ADN, et avant tout des cytosines, peuvent intervenir dans le contexte du dinucléotide CpG (également noter CG) (A. Bird, 2002; Finnegan & Kovac, 2000). Cette modification de l'ADN est partagée par de très nombreuses espèces de vertébrés, plantes et champignons (Feng et al., 2010). Les plantes présentent des contextes supplémentaires pouvant être méthylés et notamment les trinucleotides CHG et CHH, H pouvant représenter une Adénine, Thymine ou Guanine (A. Bird, 2002). À l'image des modifications de la séquence d'ADN, les différences de méthylation de l'ADN, peuvent avoir un impact sur la stabilité du génome (Miura et al., 2001), l'expression des gènes (Razin & Cedar, 1991) et la variation phénotypique (Soppe et al., 2000). En effet, il a été observé que des hypométhylations de l'ADN, quel que soit le contexte, et plus spécifiquement sur les cytosines en amont des gènes, pouvaient aboutir à des activations massives de d'éléments transposables (Tsukahara et al., 2009) et des désordres transcriptionnels (Kankel et al., 2003; L. Zhang et al., 2021) dont certains peuvent impacter le phénotype (Soppe et al., 2000). Les méthylations des cytosines présentes dans le corps des gènes et plus particulièrement sur les exons présentent des effets différents encore mal compris. Néanmoins, il semblerait que celles-ci soient liées à la transcription (Bewick & Schmitz, 2017; J. Choi et al., 2020).

Les modifications épigénétiques ont été étudiées au sein de différentes espèces polyploïdes. Pour certaines espèces où ont été observés des déséquilibres d'expression des gènes ou de participation à la variation phénotypique, il a été également observé des déséquilibres dans la répartition des méthylations de l'ADN. Par exemple, chez *Mimulus peregrinus*, il a été observé que le sous-génome le moins exprimé présentait des taux de méthylation globaux en contexte CHH plus importants que le sous-génome le plus exprimé (Edger et al., 2017). De même, il a été observé chez *A. thaliana* que les hybrides intraspécifiques avec des caractéristiques épigénétiques (épihybrides) différentes présentaient des

différences dans les profils de méthylation de l'ADN (Rigal et al., 2016). L'hybridation de ces organismes est à l'origine d'un choc épigénomique, immédiatement après hybridation, qui va être caractérisée par un déséquilibre de la méthylation de l'ADN associé à un déséquilibre transcriptionnel. Chez le coton, le sous-génome D, considéré comme dominant, présente une activité importante au niveau des histones ainsi qu'un niveau plus bas, dans l'ensemble, de méthylation de l'ADN par rapport au sous-génome A (Zheng et al., 2016). De même, chez le coton, un déséquilibre épigénétique a pu être associé à un biais d'expression et de phénotype entre les paires de chromosomes ohnologues (Song et al., 2015). Chez *B. napus*, un biais dans la méthylation a été observé et relié au biais d'expression des gènes (J. Wang et al., 2004) et de phénotype (Comai et al., 2000 ; Lukens et al., 2006 ; Xu et al., 2009). Une observation similaire a en outre été faite chez le blé (Y. Li et al., 2019).

Au cours de ce travail de thèse, nous avons observé chez le pommier un certain nombre de déséquilibres, particulièrement au niveau de la participation à la variation phénotypique et de l'expression des gènes ohnologues. Dans ce chapitre, nous avons cherché à identifier s'il existait aussi un déséquilibre de méthylation de l'ADN entre les gènes ohnologues.

## 8.2 Matériel et méthode

### 8.2.1 Récupération des données Bisulfite Seq

Le pipeline de recherche et de téléchargement des expériences de RNA-Seq développé pour l'analyse du déséquilibre transcriptionnel détaillé en Chapitre 6 a été dérivé afin de produire un pipeline utilisable dans le cadre de l'étude du méthylome. Ce pipeline, *get-bisulfite-seq* Snakemake est accessible sur la forge logicielle de l'INRAE en suivant l'URL : <https://forgemia.inra.fr/tanguy.lallemand/get-bisulfite-seq>. Ce pipeline nous a permis de télécharger l'ensemble des *runs* de séquençage d'ADN traités au bisulfite pour *M. domestica* disponibles dans les banques publiques en appliquant les mêmes filtres de qualité que pour l'analyse de l'expression des gènes ohnologues c'est à dire : des données pairées ; au moins trois réplicats biologiques ; un séquençage d'au moins  $2 \times 10^7$  *reads* par échantillon et une longueur de *reads* minimale de 100 bp). Les fichiers bruts au format fastq compressés ont été téléchargés avec SRA-toolkit et Aspera. De la même façon que pour les données transcriptomiques, les fichiers ont été stockés en construisant une architecture de dossier de type SRP/SRX/SRR.

### 8.2.2 Calcul des rapports de méthylation

Dans le contexte de la sous-dominance du génome, les régulations épigénétiques sont également suspectées de jouer un rôle, en particulier les méthylations de l’ADN (Alger & Edger, 2020). Pour étudier cette hypothèse, nous avons mis en place une étude permettant de rechercher un déséquilibre des ratios de méthylation dans les contextes CG, CHG et CHH (H pour A, C ou T). Les données brutes téléchargées par le pipeline *get-rna-seq-data* ont été traitées à l’aide de BiSePS en version 10 (Hatira, 2022), un pipeline d’analyse de séquençage bisulfite développé au laboratoire permettant l’alignement et la génération de rapports de méthylation.

Cet outil basé sur un pipeline Snakemake associé à une interface graphique déployée en Électron permet d’examiner les *reads* traités au bisulfite. Ainsi ce pipeline évalue la qualité du séquençage avec FastQC (Andrews et al., 2010) et localise les sites de méthylation dans les différents contextes à l’aide de Bismark (Krueger & Andrews, 2011). L’ensemble des métriques fournies par BiSePS sont agrégées avec MultiQC (Ewels et al., 2016) et exportées au format HTML pour permettre un contrôle de qualité des données brutes et traitées. Les résultats de Bismark peuvent être traités pour être ajoutés à une instance Jbrowse 2 (Buels et al., 2016) pour en assurer leur visualisation ou exportés sous la forme de fichiers bruts pour des analyses complémentaires.

Les fichiers de sortie de Bismark, et en particulier les fichiers de rapport de méthylation, ont ensuite été traités dans un pipeline *ad hoc* (*methylome-imbalance*, accessible à l’aide du lien suivant : <https://forgemia.inra.fr/tanguy.lallemand/methylome-imbalance>). Ce pipeline vise à filtrer les résultats puis applique un traitement statistique et agrège l’ensemble des expériences en un résultat unique pour chacune des paires de fragments chromosomiques synténiques. Ainsi, les positions des cytosines associées à l’un des trois contextes (CG, CHG ou CHH) ont été associées à leurs positions dans les gènes. Les cytosines associées à des régions introniques ont été supprimées. Chaque séquençage a été traité séparément, de même que les contextes de méthylation. Quatre analyses ont ensuite été réalisées sur les cytosines associées à la région exonique du corps des gènes ohnologues, ou à l’environnement des gènes ; l’environnement étant défini soit comme 2 kb en amont et en aval, soit comme 500 bp ou 100 bp en amont des sites d’initiation de la transcription (TSS) des gènes ohnologues. Pour chaque région et contexte, un test de Mann — Whitney  $U$  a été mis en œuvre entre les paires de fragments chromosomiques ohnologues afin de comparer la distribution des pourcentages de cytosines méthylées de toutes les positions pour chacun des échantillons considérés. Les résultats de chaque séquençage

ont ensuite été agrégés à l'aide de la méthode de Mudholkar et George (Mudholkar E.O., 1983) afin d'obtenir un résultat global pour chaque paire de segments chromosomiques synténiques. Cette méthode d'agrégation de *p-values* permet de faire la moyenne des statistiques de la méthode de Fisher et de Pearson. Cette méthode a ainsi la caractéristique de fournir des résultats avec des *p-values* extrêmes proches de 0 ou de 1. Elle est, de plus, moins sensible à de très fortes puissances statistiques.

Par la suite, nous avons examiné le nombre de contextes associés à chacun des gènes ohnologues en amont des gènes. Pour ce faire, nous avons identifié le nombre de cytosines associées à chacun des contextes en amont des gènes ohnologues. Ensuite, nous avons compté le nombre de cytosines associées dans chaque contexte pour une taille de fenêtre allant de 50 bp à 2 kb en amont. Pour chacune des paires de chromosomes ohnologues, nous avons comparé le nombre de cytosines et étudié les 25 gènes les plus différents pour chaque taille de fenêtre. Nous avons ensuite examiné le pourcentage de ces gènes qui appartiennent au groupe de gènes non commutants. Les gènes non commutants, identifiés lors de l'étude de l'expression des gènes ohnologues regroupent des gènes qui ont un déséquilibre du niveau d'expression par rapport à son ohnologue, similaires dans au moins 85 % des expériences de transcriptomiques analysées. Comme contrôle négatif, nous avons pris 50 gènes au hasard à chaque fois. Un test de Wilcoxon permet alors de déterminer si la distribution des pourcentages de gènes non commutants présents dans les fenêtres sont similaires entre les groupes pour chacune des paires de fragments chromosomiques synténiques.

## 8.3 Résultats

### 8.3.1 Comparaison des niveaux de méthylation de l'ADN entre les gènes ohnologues

Pour approfondir la recherche d'un déséquilibre entre les blocs synténiques du génome de la pomme, nous avons testé le déséquilibre des marques de méthylation de l'ADN dans le corps des gènes ou autour des gènes situés dans des fragments chromosomiques synténiques.

Nous avons identifié un ensemble de 61 séquençages Bisulfite-Seq de *M. domestica* disponibles publiquement sur les bases de données. Après filtrages des expériences selon les mêmes critères de qualité que lors de l'analyse RNA-Seq, nous avons pu conserver 36 séries



de séquençage au bisulfite pour une taille totale de 242 Go de fichiers compressés. À partir des méta-données associées à ces expériences dont le pipeline permet une récupération automatique, et après ré-annotation en employant l'ontologie précédemment utilisée, nous avons constaté que les séquençages sont issus d'expériences associées à des feuilles ou des fruits.

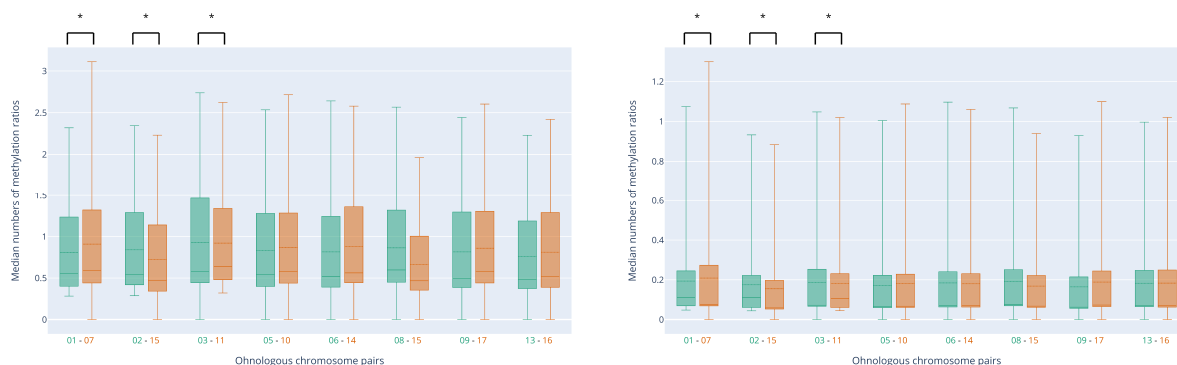
Les étapes d'alignement des *reads* et de construction des rapports de méthylation de chacune des expériences ont été exécutées via BiSePS. Après alignement des 36 expériences conservées, le jeu de données représente 700 Go supplémentaires. Les taux de *mapping* de différents échantillons ont permis de filtrer les expériences dont le *mapping* présente un résultat considéré comme insuffisant, c'est-à-dire présentant un taux inférieur à 60%. Ainsi, 4 échantillons dont le pourcentage de *mapping* était inférieur à 60% ont été supprimés. Pour les 32 échantillons restants, le pourcentage médian de *mapping* était de 75,44% avec un écart type de 4. Le nombre médian de *reads* localisés est de 10 896 005 dans chacune des 32 expériences. Par ailleurs, nous avons identifié un nombre médian de 1 025 086 *reads* ambigus.

Les séquençages conservés et alignés sont associés à des expériences Bisulfite-Seq surtout résultantes de feuilles avec 68% des expériences qui en sont issues (22 expériences) et de fruits qui représentent 31% du jeu de données avec 10 expériences. Les échantillons proviennent essentiellement de tissus issus de GDDH13 ou GDDH18. Ces expériences sont associées à des analyses portant sur le développement, notamment du fruit et les stress abiotiques, en particulier des stress hydriques.

Le traitement de la sortie brute de Bismark a été effectué en passant par un pipeline *ad hoc*. Les données finales traitées représentent 762 Go. Ces données ont permis de chercher à identifier si un déséquilibre dans les méthylation d'ADN des gènes ohnologues pouvait être identifié. Les rapports de méthylation ont permis de calculer les pourcentages de méthylation des cytosines dans les différents contextes pour les différentes régions d'intérêts. Ces régions d'intérêts sont les régions exoniques du gène (corps du gène) et l'environnement du gène (les régions amont et aval des gènes). Les pourcentages de méthylation ont été calculés pour les trois contextes de cytosine.

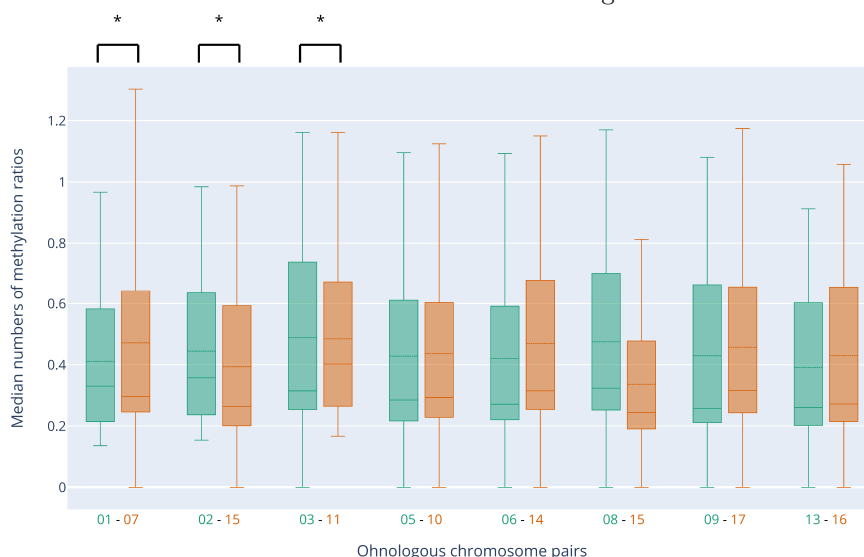
Globalement, les taux de méthylation du contexte CG sont en moyenne plus élevés que ceux des contextes CHG et CHH (moyenne autour de 0,80 contre 0,43 et 0,21). Le contexte CHH montre des valeurs plus élevées que le contexte CHG avec des valeurs moyennes de 0,43 contre 0,21.

Pour chaque expérience et chaque paire de fragments de chromosomes ohnologues,



(A) Boîtes à moustaches de la distribution des rapports de méthylation des cytosines situées 500 bp en amont des gènes pour le contexte CG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

(B) Boîtes à moustaches de la distribution des rapports de méthylation des cytosines situées 500 bp en amont des gènes pour le contexte CHH et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.



(C) Boîtes à moustaches de la distribution des rapports de méthylation des cytosines localisées 500 bp en amont des gènes pour le contexte CHG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

**Figure 8.1** — Boîtes à moustaches de la distribution des rapports de méthylation des cytosines localisées 500 bp en amont des gènes et groupées par paires de fragments chromosomiques synténiques. L'axe des abscisses (*Ohnologous chromosome pairs*) comporte les paires de fragments chromosomiques synténiques. L'axe des ordonnées (*Median numbers of methylation ratio*) décrit les médianes des pourcentages des ratios de méthylation pour les différentes expériences analysées. Le bas de la boîte représente le premier quartile, le trait interne à la boîte présente la médiane de la distribution et l'extrémité supérieure de la boîte présente le troisième quartile. La taille de la boîte représente l'écart inter-quartile. La moustache inférieure présente le neuvième décile et la moustache supérieure présente quant à elle le 91<sup>e</sup>. La première boîte à moustaches représente la distribution des rapports de méthylation sur le premier chromosome de la paire (en vert), tandis que la seconde boîte à moustaches représente la distribution des rapports de méthylation sur le second chromosome de la paire (en orange). Les paires 1-7, 2-15 et 3-11 présentent une différence significative pour l'ensemble des contextes.  $*\alpha < 5\%$

nous avons vérifié si les distributions des taux de méthylation étaient similaires à l'aide d'un test de Mann—Whitney  $U$ . Les 32  $p$ -values de chaque expérience, pour chaque paire, ont ensuite été agrégées par la méthode de Fisher et sont présentées dans le tableau 8.1. Par ailleurs, les distributions des pourcentages de méthylation sont représentées par un ensemble de boîtes à moustaches comportant les rapports de méthylation des cytosines.

Les visualisations associées aux cytosines localisées dans les 500 bp en amont des gènes sont quant à elles présentées en Figure 8.1. En comparant les pourcentages de méthylation des cytosines localisées 500 bp en amont des gènes ohnologues, nous avons identifié une différence significative dans le taux de méthylation entre les gènes ohnologues des paires de fragments chromosomiques synténiques 1-7, 2-15 et 3-11 dans les trois contextes.

Ainsi, pour le contexte CG, les distributions de taux de méthylation pour 500 bp en amont des gènes sont proposées en Figure 8.1A. Pour la paire 1-7, une différence significative a été observée pour les taux de méthylation entre ces fragments chromosomiques synténiques. Ainsi, le chromosome 1 montre un taux médian en contexte CG de 0,55 contre 0,59 pour le chromosome 7. Pour la paire 2-15, le chromosome 2 présente un taux de méthylation médian en contexte CG de 0,54 contre 0,47 pour le chromosome 15. Pour finir, la paire 3-11 sur le chromosome 11 présente un taux de méthylation médian en contexte CG de 0,57 contre 0,64 pour le chromosome 3.

Les distributions associées au contexte CHG sont présentées en Figure 8.1C. Un déséquilibre similaire est observé pour les cytosines 500 bp en amont des gènes dans le contexte CHG. Pour la paire 1-7, le chromosome 7 présente un taux de méthylation significativement différent de celui du fragment synténique associé au chromosome 1 (moyenne de taux de méthylation médian de 0,41 pour le fragment chromosomique 1 contre 0,47 pour le fragment chromosomique 7). Pour la paire 2-15, le chromosome 2 présente un taux de méthylation moyen significativement différent de celui du fragment synténique associé au chromosome 15 (moyenne de taux de méthylation médian de 0,44 pour le fragment chromosomique 2 contre 0,39 pour le fragment chromosomique 15). Quant à la paire 3-11, c'est le chromosome 11 (0,40) qui est significativement plus méthylé que le chromosome 3 (0,31). Les autres paires de fragments chromosomiques synténiques ne sont pas significativement différentes en termes de taux de méthylation des cytosines dans le contexte CHG.

Les distributions associées au contexte CHH sont fournies en Figure 8.1B et présentent des conclusions similaires, avec un déséquilibre de la méthylation en faveur du chromosome 7 pour la paire 1-7 (0,19 pour le chromosome 1 contre 0,20 pour le chromosome 7).

Pour la paire 2-15, le déséquilibre est en faveur du chromosome 2 (0,17 contre 0,15 pour le chromosome 15). La paire 3-11 présente un déséquilibre en contexte CHH en faveur du chromosome 11 (0,07 contre 0,1 pour le chromosome 11).

Les distributions pour les cytosines localisées à 100 bp en amont des gènes pour les principales paires synténiques dans les trois contextes sont présentées en Figure A.31. Elles montrent des résultats similaires, bien que moins marqués. Pour le contexte CG (Figure A.31A), on observe un déséquilibre significatif des paires 1-7, 2-15 et 3-11 dans le même sens que pour les comparaisons faites sur les cytosines localisées dans la région 500 bp en amont des gènes ohnologues. Pour les cytosines en contexte CHG (Figure A.31B) nous avons observé un déséquilibre similaire, significatif pour les paires de fragments synténiques 1-7, 2-15 et 3-11. Concernant le contexte CHH (Figure A.31C), un déséquilibre similaire bien que moins marqué est observé pour les paires 1-17, 2-15 et 3-11.

En ce qui concerne le corps du gène exonique, les distributions des pourcentages de cytosines méthylées sont résumées dans la Table 8.1 et les visualisations associées sont proposées en Figure A.33. Le contexte CG présente des niveaux de méthylation significativement différents pour les paires 1-7, 2-15, et 3-11. Les paires 5-10, 6-14, 8-15, 9-17 et 13-16 présentent des différences peu ou pas significatives. Pour la paire 1-7, le chromosome 1 est significativement hyperméthylé en termes de cytosines associées au contexte CG sur le corps du gène par rapport au chromosome 7 (médiane de 0,32 contre 0,18). Pour la paire 2-15, on observe un chromosome 2 hyperméthylé par rapport à son ohnologue avec un taux médian de méthylation de 0,29 pour le chromosome 2 contre 0,16 pour le chromosome 15. La paire 3-11 montre un chromosome 11 hyperméthylé par rapport au chromosome 3 (avec des médianes de respectivement 0,38 contre 0,20). L'analyse des contextes CHG et CHH indique des différences similaires.

Couple	Contexte	Région	Médiane des ratios de méthylations
01-07	CG	Corps du gène	0,525/0,532*
		Environnement génique	0,784/0,805*
		Amont du gène (500 bp)	0,809/0,909*
	CHG	Corps du gène	0,032/0,034*
		Environnement génique	0,358/0,365*
		Amont du gène (500 bp)	0,411/0,472*
CHH	Corps du gène	0,012/0,012*	
	Environnement génique	0,119/0,119*	

**Table 8.1 suite de la page précédente**

Couple	Contexte	Région	Médiane des ratios de méthylations	
02-15	CG	Amont du gène (500 bp)	0,194/0,209*	
		Corps du gène	0,487/0,457*	
		Environnement génique	0,745/0,742*	
	CHG	Amont du gène (500 bp)	0,841/0,722*	
		Corps du gène	0,033/0,038*	
		Environnement génique	0,326/0,353*	
	CHH	Amont du gène (500 bp)	0,444/0,394*	
		Corps du gène	0,012/0,012*	
		Environnement génique	0,112/0,102*	
	03-11	CG	Amont du gène (500 bp)	0,175/0,155*
			Corps du gène	0,553/0,617*
			Environnement génique	0,876/0,932*
CHG		Amont du gène (500 bp)	0,93/0,921*	
		Corps du gène	0,033/0,041*	
		Environnement génique	0,419/0,454*	
CHH		Amont du gène (500 bp)	0,489/0,485*	
		Corps du gène	0,012/0,013*	
		Environnement génique	0,114/0,12*	
05-10		CG	Amont du gène (500 bp)	0,187/0,181*
			Corps du gène	0,572/0,555
			Environnement génique	0,843/0,846
	CHG	Amont du gène (500 bp)	0,83/0,867	
		Corps du gène	0,045/0,041	
		Environnement génique	0,383/0,396	
	CHH	Amont du gène (500 bp)	0,428/0,436	
		Corps du gène	0,012/0,012	
		Environnement génique	0,108/0,111	
	CG	Amont du gène (500 bp)	0,172/0,181	
		Corps du gène	0,523/0,516	
		Environnement génique	0,795/0,852	
06-14	CG	Amont du gène (500 bp)	0,816/0,878	

Table 8.1 suite de la page précédente

Couple	Contexte	Région	Médiane des ratios de méthylations
	CHG	Corps du gène	0,037/0,028
		Environnement génique	0,35/0,394
		Amont du gène (500 bp)	0,42/0,47
	CHH	Corps du gène	0,012/0,011
		Environnement génique	0,114/0,112
		Amont du gène (500 bp)	0,184/0,18
	CG	Corps du gène	0,484/0,495
		Environnement génique	0,799/0,697
		Amont du gène (500 bp)	0,864/0,662
08-15	CHG	Corps du gène	0,038/0,025
		Environnement génique	0,393/0,316
		Amont du gène (500 bp)	0,475/0,336
	CHH	Corps du gène	0,013/0,01
		Environnement génique	0,118/0,108
		Amont du gène (500 bp)	0,191/0,168
	CG	Corps du gène	0,555/0,574
		Environnement génique	0,85/0,852
		Amont du gène (500 bp)	0,816/0,86
09-17	CHG	Corps du gène	0,045/0,035
		Environnement génique	0,401/0,394
		Amont du gène (500 bp)	0,43/0,457
	CHH	Corps du gène	0,012/0,011
		Environnement génique	0,109/0,113
		Amont du gène (500 bp)	0,164/0,189
	CG	Corps du gène	0,451/0,467
		Environnement génique	0,795/0,802
		Amont du gène (500 bp)	0,759/0,812
13-16	CHG	Corps du gène	0,03/0,03
		Environnement génique	0,385/0,394
		Amont du gène (500 bp)	0,391/0,43
	CHH	Corps du gène	0,011/0,011

**Table 8.1 suite de la page précédente**

Couple	Contexte	Région	Médiane des ratios de méthylation
		Environnement génique	0,117/0,116
		Amont du gène (500 bp)	0,183/0,183

**Table 8.1** – Ratios de méthylation sur chaque paire de chromosomes ohnologues dans les différentes régions considérées pour les 3 contextes de méthylation. Les pourcentages de méthylation des paires 1-7, 2-15 et 3-11 apparaissent comme significativement différents pour les cytosines associées aux corps des gènes, en amont des gènes et dans l'environnement proche des gènes. \*p-value < 0.05

En ce qui concerne le corps du gène exonique, les distributions des pourcentages de cytosines méthylées sont résumées dans la Table 8.1 et les visualisations associées sont présentées en Figure A.33. Le contexte CG présente des niveaux de méthylation significativement différents pour les paires 1-7, 2-15, et 3-11. Les paires 5-10, 6-14, 8-15, 9-17 et 13-16 présentent des différences peu ou pas significatives. Pour la paire 1-7, le chromosome 1 est significativement hyperméthylé en termes de cytosines associées au contexte CG sur le corps du gène par rapport au chromosome 7 (médiane de 0,32 contre 0,18). Pour la paire 2-15, on observe un chromosome 2 hyperméthylé par rapport à son ohnologue avec un taux médian de méthylation de 0,29 pour le chromosome 2 contre 0,16 pour le chromosome 15. La paire 3-11 montre un chromosome 11 hyperméthylé par rapport au chromosome 3 (avec des médianes de respectivement 0,38 contre 0,20). L'analyse des contextes CHG et CHH indique des différences similaires.

Pour les cytosines localisées dans l'environnement génique défini comme 2 kb en amont et en aval des gènes ohnologues, les résultats sont présentés en Table 8.1 et en Figure A.32. Nous avons identifié des différences significatives pour les paires 1-7, 2-15 et 3-11 dans les trois contextes. Les profils de méthylation dans CG, CHH et CHG indiquent une tendance similaire aux résultats du corps du gène.

### 8.3.2 Comparaison du nombre de positions de cytosines entre les gènes ohnologues

Nous nous sommes intéressés au nombre de cytosines dans les différents contextes pour les différentes régions analysées. Les comptages du nombre de positions entre les fragments chromosomiques synténiques ont été comparés avec un test binomial et sont présentés en Table 8.2 pour la région 100 bp en amont des gènes ohnologues. La région 100 bp en amont des gènes, présente des différences significatives du nombre de cytosines

en contexte CHH pour les paires 1-7, 2-15, 3-11, 5-10, 6-14, 8-15 et 13-16. La paire 9-17 ne présente pas de différence du nombre de cytosines en contexte CHH avec 25 282 cytosines associées 100 bp en amont des gènes contre 25 114 cytosines pour le chromosome 17. Les contextes CHG et CG présentent des effectifs inférieurs en termes de nombre et les paires précédentes montrent néanmoins un déséquilibre significatif du nombre de cytosines à part les paires 2-15, 3-11 et 9-17. Les Tables A.5, A.6 et A.7 présentent respectivement les résultats associés aux régions exoniques, la région 2 kb en amont et en aval des gènes et 500 bp des gènes ohnologues. Concernant la région correspondant à l'environnement génique (2 kb en amont et en aval des gènes), l'ensemble des couples pour tous les contextes présentent des nombres significativement différents de cytosines. La différence est la plus marquée pour le contexte CHH. La paire 9-17 semble aussi moins déséquilibrée que les autres, et présente des différences moins importantes. Un constat similaire est fait pour les régions exoniques et 500 bp en amont des gènes.

Afin de détailler plus avant cette observation du nombre différent de cytosines, nous avons cherché à comprendre si cette différence pouvait être associée à des déséquilibres observés dans d'autres analyses et en particulier les gènes non commutants qui constituent un résultat fort. Afin de tester si un lien pouvait être établi entre ces deux observations, nous avons comparé le nombre de cytosines dans le contexte CHH, qui présente le signal le plus fort entre les gènes ohnologues. Sur les 25 paires de gènes ohnologues présentant le plus de différences dans le nombre de cytosines associées au contexte CHH, 52 % ont été identifiées comme appartenant au groupe de gènes dont l'expression était comme non commutée dans toutes les expériences RNA-Seq. En prenant des paires de gènes ohnologues au hasard, seulement 36 % sont des couples de gènes non commutés. En outre, le test apparié de Wilcoxon a permis de confirmer que les 25 paires de gènes ohnologues présentant le plus de différences dans le nombre de cytosines associées au contexte CHH étaient significativement enrichies en gènes non commutants par rapport à une sélection aléatoire de gènes. Le détail des résultats sont fournis en Table 8.3. La Figure 8.2 présente les distributions du pourcentage de gènes non commutants parmi les 25 gènes les plus différents en termes de nombre de position de cytosine dans le contexte CHH. Cette figure montre que certaines zones et surtout, la zone autour de 150 bp en amont du TSS ainsi que celle entre 1500 bp et 2000 bp semblent importantes. En effet, les gènes les plus différents pour ces zones sont associés au plus hauts pourcentages de gènes aussi associés aux gènes non commutants.

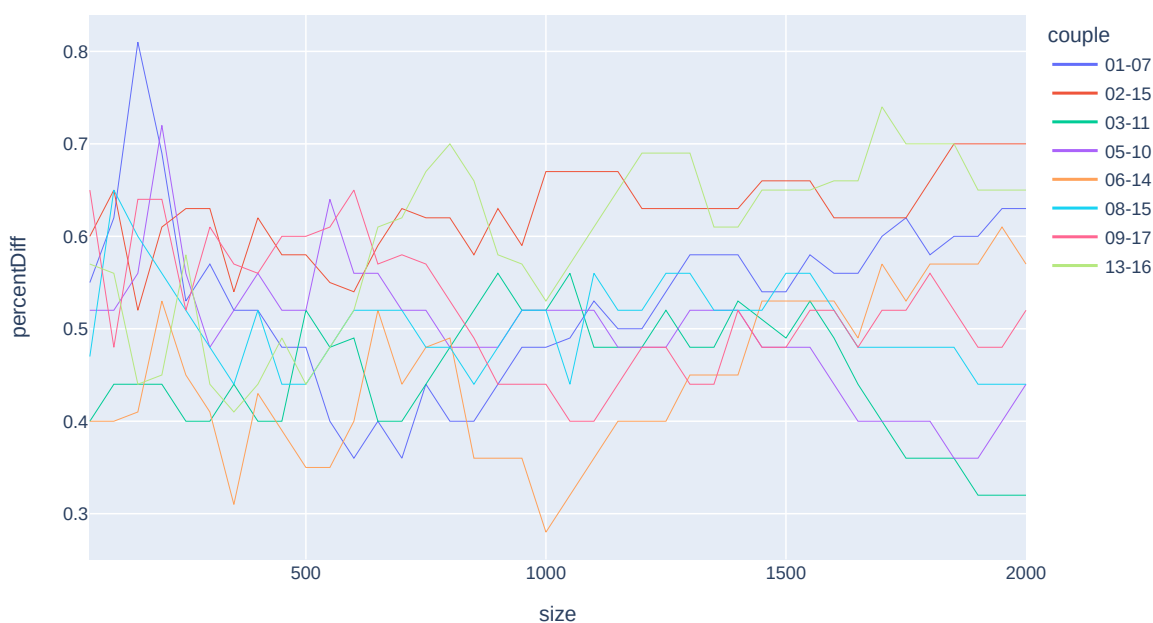


**Table 8.2** – Nombre de cytosines associées à chacun des contextes de méthylation 100 bp en amont pour les paires de gènes ohnologues. Un test binomial a été exécuté pour tester si les nombres de cytosines sont significativement différents.

Couple	Nombre de cytosines sur le premier chromosome	Nombre de cytosines sur le second chromosome	Contexte	p-value test Binomial
01-07	22 123	16 695	CHH	$1,4606 \times 10^{-167}$
02-15	22 754	20 995	CHH	$4,2402 \times 10^{-17}$
03-11	28 930	25 418	CHH	$2,7222 \times 10^{-51}$
05-10	33 904	30 532	CHH	$2,9089 \times 10^{-40}$
06-14	19 401	14 486	CHH	$1,5558 \times 10^{-157}$
08-15	28 697	23 371	CHH	$1,1756 \times 10^{-120}$
09-17	25 282	25 114	CHH	$4,5693 \times 10^{-1}$
13-16	34 219	31 451	CHH	$3,4742 \times 10^{-27}$
01-07	4139	3238	CHG	$9,4483 \times 10^{-26}$
02-15	4330	4216	CHG	$2,2157 \times 10^{-1}$
03-11	5361	4930	CHG	$2,2417 \times 10^{-5}$
05-10	6541	5679	CHG	$6,5998 \times 10^{-15}$
06-14	3722	2554	CHG	$2,1654 \times 10^{-49}$
08-15	5278	4361	CHG	$9,9180 \times 10^{-21}$
09-17	4855	4651	CHG	$3,7330 \times 10^{-2}$
13-16	6610	6074	CHG	$2,0239 \times 10^{-6}$
01-07	4562	3414	CG	$7,0574 \times 10^{-38}$
02-15	4613	4553	CG	$5,3773 \times 10^{-1}$
03-11	5334	4888	CG	$1,0722 \times 10^{-5}$
05-10	6580	5714	CG	$5,9730 \times 10^{-15}$
06-14	4093	2849	CG	$1,3653 \times 10^{-50}$
08-15	5884	4861	CG	$5,8033 \times 10^{-23}$
09-17	5131	5015	CG	$2,5358 \times 10^{-1}$
13-16	6770	6390	CG	$9,5320 \times 10^{-4}$

**Table 8.3** – Résultats des tests appariés de Wilcoxon des pourcentages de gènes non commutants pour les 25 gènes présentant la plus grande différence de nombre de positions de cytosines dans le contexte CHH par rapport à 25 gènes sélectionnés au hasard pour une fenêtre de taille 50 bp allant de 2 kb en amont du TSS jusqu’au TSS.

Couple	Pourcentage des gènes les plus différents	Pourcentage des gènes pris au hasard	p-value test Wilcoxon
01-07	0,53 %	0,37 %	$9,0949 \times 10^{-12}$
02-15	0,63 %	0,39 %	$1,8190 \times 10^{-12}$
03-11	0,45 %	0,37 %	$3,0630 \times 10^{-8}$
05-10	0,50 %	0,30 %	$1,8190 \times 10^{-12}$
06-14	0,45 %	0,34 %	$8,1309 \times 10^{-10}$
08-15	0,50 %	0,40 %	$1,8190 \times 10^{-12}$
09-17	0,52 %	0,32 %	$1,8190 \times 10^{-12}$
13-16	0,60 %	0,40 %	$1,8190 \times 10^{-12}$



**Figure 8.2** – Distribution du pourcentage de gènes non commutants parmi les 25 gènes les plus différents en termes de nombre de positions de cytosines dans le contexte CHH. L’axe des abscisses (*size*) présente la position en paire de base relativement au TSS. L’axe des ordonnées (*percentDiff*) représente le pourcentage moyen de différence de nombre de positions CHH entre les paires de gènes les plus différentes en termes de nombre de cytosines associées au contexte CHH. La couleur est associée à la paire de fragments chromosomiques synténiques.

## 8.4 Discussion

Les méthylations de l'ADN ont été étudiées dans différentes régions du génome et dans différents contextes. Tout d'abord, les niveaux de méthylation de l'ADN dans les régions géniques, et en particulier les exons des gènes ohnologues ont été analysés dans les trois contextes. Dans la littérature, ces méthylations et, en particulier les méthylations dans le contexte CG, ont été reliées à des changements dans l'expression des gènes. En effet, pour les gènes dits méthylés sur le corps du gène (gene Body Methylated (gbM)), l'hyperméthylation des îlots CpG localisés dans ces gènes a été associée dans différentes espèces de vertébrés et typiquement l'humain (Arechederra et al., 2018; Jjingo et al., 2012; Kulis et al., 2012; Maunakea et al., 2010; Varley et al., 2013; Y. Wang et al., 2021) et de plantes (Bewick & Schmitz, 2017; Muyle et al., 2022) à une sur expression des gènes. Dans le contexte de la dominance sous génomique, il a été observé chez certaines espèces et notamment *B. napus*, que les méthylations, pour les trois contextes, sur le corps du gène étaient significativement plus importantes sur les gènes localisés sur le génome dominant transcriptionnellement par rapport à leurs ohnologues localisés sur les fragments chromosomiques synténiques (K. A. Bird et al., 2021; Chalhoub et al., 2014). Cette observation est surtout marquée pour le contexte CG (K. A. Bird et al., 2021).

Chez le pommier, nous avons observé un résultat similaire. En effet, les méthylations associées au contexte CG sont significativement différentes pour les paires 1-7, 2-15 et 3-11. Les fragments chromosomiques ayant accumulé un taux de méthylation significativement plus élevé sont les fragments chromosomiques qui sont significativement plus exprimés que leur ohnologue. Ces fragments chromosomiques sont aussi ceux participant le plus à la variation phénotypique du pommier avec un nombre significativement plus important de QTLs associés. Ce résultat est à pondérer par la méthode d'agrégation de *p-value* utilisée. En effet, la méthode de Mudholkar et George est une méthode qui génère des *p-values* extrêmes, très proches de 0 ou de 1 et qui a tendance à être conservatrice. Ainsi, une agrégation avec la méthode de Fisher permet par exemple d'observer un déséquilibre statistiquement significatif pour l'ensemble des paires de fragments ohnologues excepté la paire 9-17. Un résultat similaire, bien qu'avec une différence moins marquée, sont observés pour les méthylations en contexte CHH et CHG. Ces observations peuvent s'expliquer en partie par le fait que pour ces deux contextes de cytosines, le nombre de positions observées est beaucoup plus important. Ainsi, dans l'ensemble de ces positions, un plus grand nombre de cytosines avec des ratios de méthylation à zéro vont avoir tendance

à abaisser les valeurs médianes des distributions des taux de méthylation. De plus, la distribution des médianes avec des populations d'effectifs plus importants ont pu amener à des différences lissées entre les distributions.

Ensuite les méthylation de l'ADN en amont des gènes ont été analysées. Ces méthylation ont été associées chez les cellules eucaryotes, pour les trois contextes, à une répression de l'expression des gènes associés (Anastasiadi et al., 2018). Ainsi, chez le peuplier, il a été observé un lien entre une hyperméthylation en amont des gènes et une baisse du niveau de transcription du gène (L. Liang et al., 2019 ; Y. Zhang et al., 2020). Chez *M. domestica*, nous avons observé un déséquilibre significatif pour les paires 1-7, 2-15 et 3-11. La paire 1-7, dont le chromosome 1 présente le plus grand nombre de QTLs et de gènes surexprimés, présente un pourcentage significativement moins important de cytosines méthylées sur le chromosome 1. Cette observation est aussi valable pour les cytosines localisées dans la région 500 bp et 100 bp en amont du TSS. Pour la paire 2-15 dont le chromosome 2 a été associé à un nombre significativement plus important de gènes surexprimés par rapport à leurs ohnologues, c'est le chromosome 2 qui présente un taux de méthylation plus important en amont des gènes (500 bp et 100 bp). Pour la paire 3-11, dans laquelle le chromosome 3 porte significativement plus de QTLs et présente un nombre significativement plus important de gènes différentiellement exprimés, nous avons identifié que le chromosome 11 est significativement plus méthylé pour les trois contextes et pour les régions 500 bp et 100 bp.

Enfin, nous avons étudié les méthylation de l'ADN dans l'environnement génique qui a été défini comme 2 kb en amont et en aval des gènes ohnologues, à l'image de ce qui avait été fait pour l'analyse de ETs. En effet, le lien entre les ETs et les méthylation de l'ADN a été fait chez différentes espèces, notamment végétales comme l'allopolyploïde, *A. thaliana* (M. Chen et al., 2008 ; Madlung et al., 2002) mais aussi *B. rapa* (Cheng et al., 2016), *B. napus* (Xu et al., 2009), *Oryza minuta* (Sui et al., 2014), *T. aestivum* (Shaked, 2001 ; Yaakov & Kashkush, 2011) ou le maïs (Eichten et al., 2013) et le riz (J. Zhang et al., 2015). Les mécanismes qui permettent ce lien ont été associés avec le mécanisme de la sous-dominance génomique. Ils seront décrits dans le chapitre discussion générale. Pour le pommier, nous avons ici observé une différence significative pour les paires 1-7, 2-15 et 3-11. Pour les paires 1-7 et 2-15, ce sont les fragments chromosomiques synténiques qui portent le plus de gènes surexprimés et de QTLs qui présentent un pourcentage significativement plus grand de cytosines de méthylées. Pour la paire 3-11, le constat est inverse, puisque c'est le chromosome 11, porteur de significativement moins de QTLs

et de gènes surexprimés, qui présente les pourcentages les plus importants de cytosines méthylées.

Nous avons testé si le nombre de position de cytosines dans les différents contextes étaient similaires entre les fragments chromosomiques des paires synténiques considérées. Tout d'abord, nous avons observé que les cytosines en contexte CHH étaient les plus fréquentes, devant les cytosines en contexte CHG puis en contexte CG, ce qui a été observé chez différentes espèces et notamment le peuplier (L. Liang et al., 2019). L'ensemble des paires sont significativement déséquilibrées pour les trois contextes et pour les différentes régions à part les paires 9-17 et 2-15 qui présentent des différences moins marquées voir pas de différences. Les chromosomes portant le plus de cytosines sont les chromosomes qui sont dominants du point de vue de l'expression et, le cas échéant, du point de vue de la participation au phénotype. Cette observation a été reliée en partie avec les gènes non commutants. En effet, nous avons observé que les 25 gènes les plus différents en termes de nombre de cytosines sont significativement enrichis en gènes non commutants par rapport aux autres couples de gènes des paires testées. À notre connaissance, ce type d'observation n'a pas été faite dans d'autres organismes présentant des déséquilibres de méthylations de l'ADN. Une des sources d'explications pourrait être des différences dans le contenu en GC de l'ADN de certains chromosomes qui serait différent de leur chromosome ohnologue. De plus, le tri-nucléotide CHH est le seul contexte à être non symétrique. Cette non symétrie est relative au fait que le contexte CG sur l'un des brins d'ADN va présenter un brin complémentaire avec le même motif. Le contexte CHG va présenter un motif symétrique sur son brin complémentaire aussi. Pour le contexte CHH, ce n'est pas le cas, ce qui pourrait aussi être une source d'explication.

## 8.5 Conclusion

Nous avons identifié 61 séquençages Bisulfite de *M. domestica* associés à différents tissus. Après filtrage des séquençages suivant des critères de qualité, 36 séquençages complets au bisulfite ont été téléchargés et alignés sur le génome de référence. Ces alignements ont permis de calculer pour chacun des séquençages les rapports de méthylation pour chacune des positions de cytosines, dans les contextes CHH, CHG et CG. Les rapports de méthylations ont été traités séparément afin de tester si les fragments chromosomiques synténiques présentent des taux de méthylations similaires dans les régions exoniques des gènes ohnologues, en amont des gènes ohnologues ainsi que dans l'environnement des gènes

ohnologues (2 kb en amont et en aval). Après agrégation des résultats via la méthode de Mudholkar et George, les paires 1-7, 2-15 et 3-11 ont été identifiées comme significativement déséquilibrées en termes de ratio de cytosines méthylées pour les trois contextes et l'ensemble des régions analysées. Le sens des déséquilibres est corrélé avec le déséquilibre de QTLs observé pour ces paires. Nous avons ici observé un nombre significativement différents de positions de cytosines dans les trois contextes pour l'ensemble des principales paires de fragments chromosomiques synténiques. Pour les cytosines associées au contexte CHH, présentant le plus grand différentiel, les 25 gènes avec le plus grand différentiel sont associés à significativement plus de gènes non commutants qu'un nombre équivalent de gènes sélectionnés au hasard. Cette observation pourrait expliquer en partie l'important et stable différentiel d'expression de ces gènes non commutants.



# DISCUSSION GÉNÉRALE

---

## 9.1 Analyse du déséquilibre de QTL

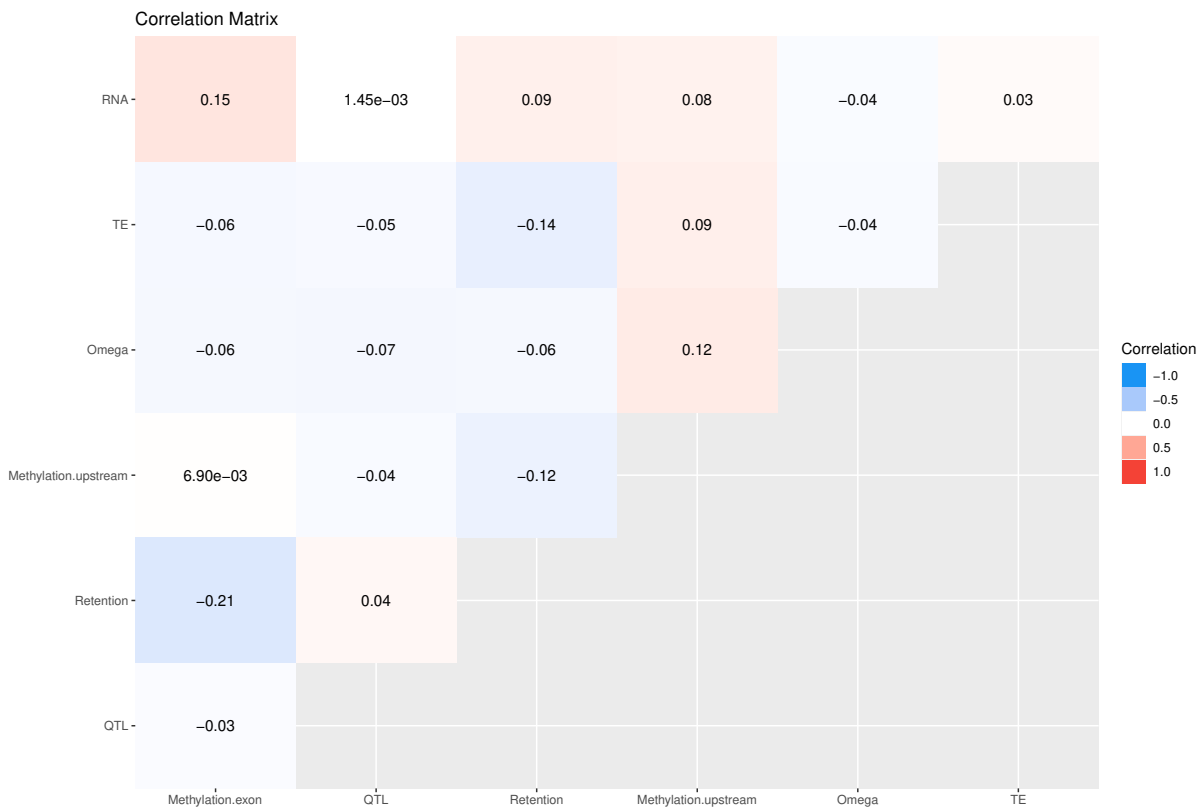
Nous avons analysé un ensemble de 1528 QTLs associés à une grande diversité de traits phénotypiques. De plus le jeu de données QTLs a été construits afin de limiter l'introduction de biais notamment de sur-représentation lié a des QTLs dont les localisations seraient chevauchantes. Nous avons pu mettre en évidence des différences dans les proportions de QTLs porté par les différents fragments synténiques ohnologues. Ainsi, les paires 1-7, 3-11, 13-16 et 8-15 ont été identifiées comme significativement déséquilibrées dans le nombre de QTLs localisés sur les segments synténiques.

Les mécanismes à l'origine du déséquilibre des QTLs, paraissent multiples. En effet, nous avons pu cerner différents déséquilibres qui semblent participer, au moins en partie, à la mise en place du déséquilibre de QTLs. Afin d'intégrer l'ensemble des analyses menées au cours de la thèse, nous avons calculé un ensemble de corrélations entre les différents indicateurs identifiés au cours de la thèse. L'intégration de l'ensemble des résultats a été placée dans le dépôt <https://forgemia.inra.fr/tanguy.lallemand/integrate-rna-seq-te>. Pour calculer les coefficients de corrélations, nous avons suivi le processus suivant. Pour une paire de fragments synténiques, nous avons calculé la différence entre les valeurs associées au premier fragment et au second fragment, gène à gène pour chacune des valeurs extraites des analyses précédentes. Les valeurs considérées sont les valeurs de pourcentages de rétention, de pression de sélection sous la forme du rapport  $\omega$ , de nombre de *reads* moyens associés au gène dans les expériences de transcriptomiques, de couverture en ET et le nombre moyen de cytosines méthylées associées aux gènes ohnologues dans les différents contextes de cytosines pour les expériences de bisulfite-Seq. Les corrélations ont été calculées par la voie de la méthode de Pearson, puisque les données analysées sont continues et une relation linéaire entre nos variables est attendue.

Nous avons cherché si des liens de corrélations pouvaient être identifiés à l'échelle du génome entre les indicateurs étudiés. Les corrélations globales sont présentées en Figure



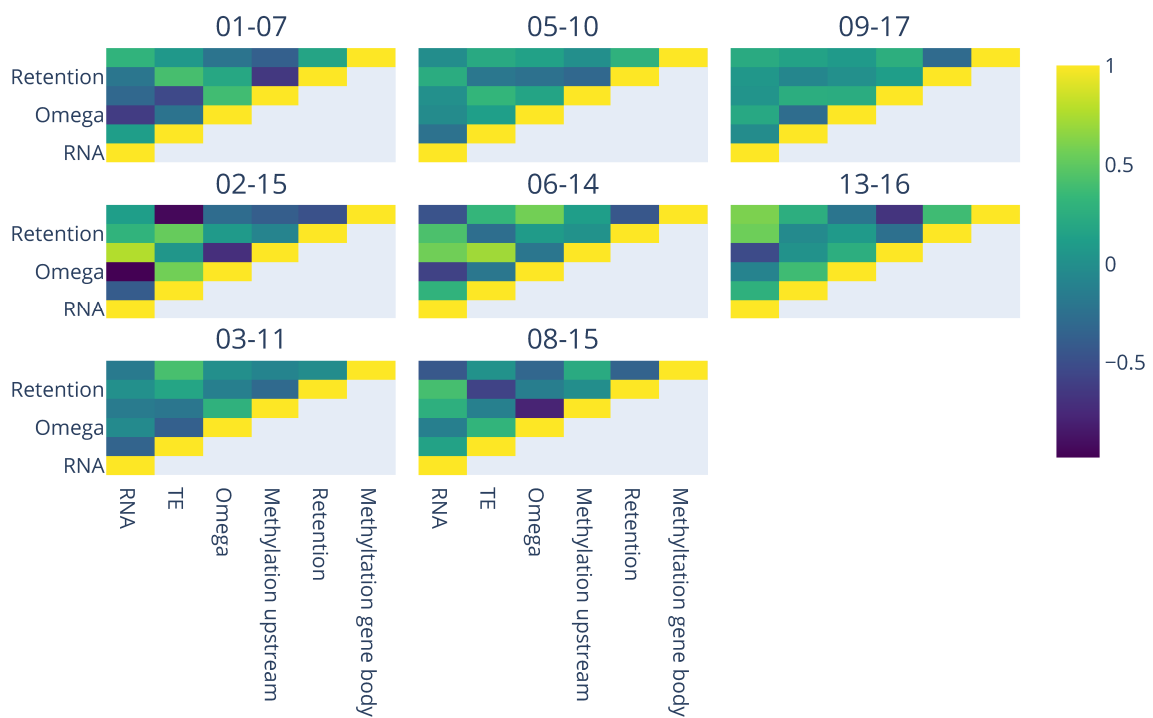
9.1. Globalement, le niveau d'expression est corrélé positivement avec le niveau de méthylation des exons. Le fractionnement biaisé du génome est corrélé négativement avec le niveau de méthylation de l'ADN en amont des gènes et sur les régions exoniques, ainsi que la couverture en ET. Pour finir, le niveau de méthylation en amont des gènes est corrélé négativement avec la pression de sélection. Toutefois, les niveaux de corrélations sont faibles, probablement à cause de la taille du jeu de données qui est importante et d'un phénomène observé qui est de l'ordre de la tendance.



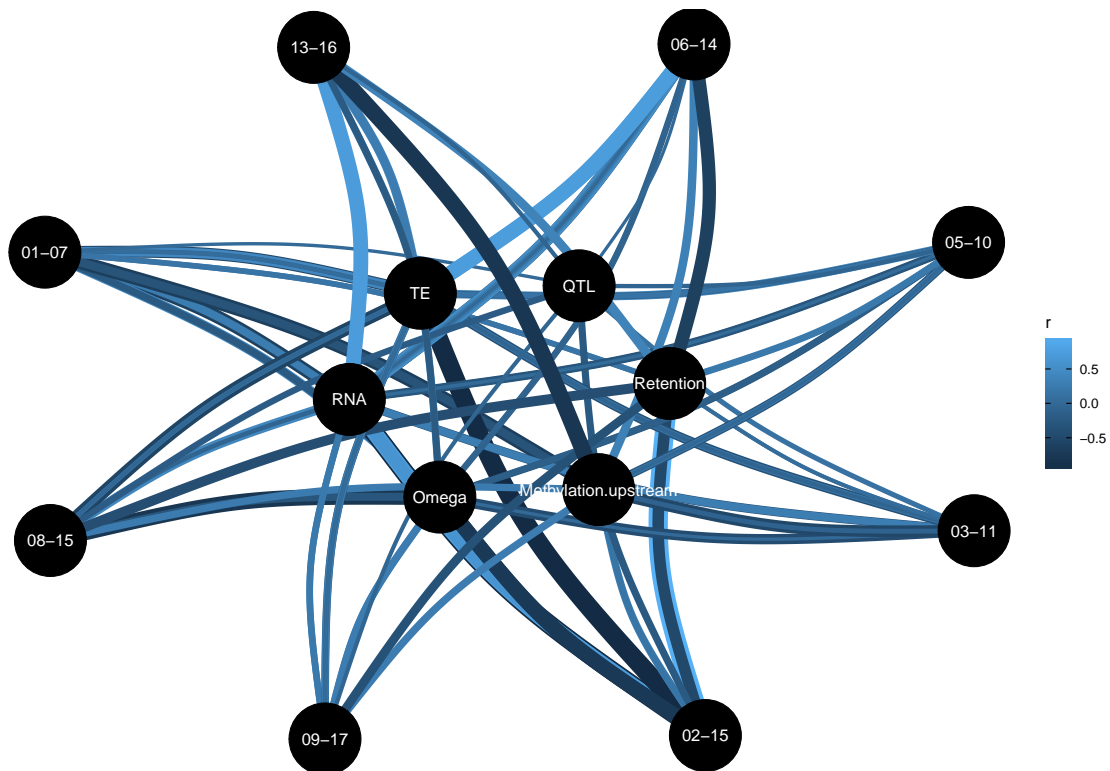
**Figure 9.1** – Carte thermique regroupant les coefficients de corrélation globaux entre les différents indicateurs étudiés au cours de la thèse. Les corrélations sont faibles entre différents indicateurs. Néanmoins, on peut observer que le niveau d'expression semble corrélé positivement avec la méthylation des exons. Le fractionnement biaisé du génome est corrélé négativement avec la méthylation de l'ADN en amont des gènes et sur les régions exoniques, ainsi que la couverture en ET. Pour finir, la méthylation en amont des gènes est corrélée négativement avec la pression de sélection.

Dans le but de préciser quels sont les liens entre les divers indicateurs en fonction des paires de fragments chromosomiques synténiques, nous nous sommes intéressés aux corrélations au niveau des paires. De plus, les différentes analyses ont montré que les paires de fragments chromosomiques synténiques n'étaient pas toutes déséquilibrées, et pas au

même degré. La Table, 9.1, présente les résultats numériques des corrélations entre les différentes variables pour chacune des principales paires de fragments synténiques ohnologues. La Figure 9.2 intègre sous forme d'un ensemble de cartes thermiques l'ensemble des résultats de corrélation associés aux principales paires analysées. Les matrices de dispersion associées à la régression linéaire sont présentées à titre d'exemple, pour la paire 1-7, en Figure 9.4. Les autres matrices sont présentées en Figures supplémentaires (Figure A.34, Figure A.35, Figure A.36, Figure A.36, Figure A.37, Figure A.38, Figure A.39, Figure A.40). Les résultats des corrélations ainsi que les droites de régression, calculés pour chaque paire de blocs chromosomiques synténiques, suggèrent l'existence de différents liens entre les variables. Pour finir, la Figure 9.3 présente sous forme de GGM les relations entre différentes variables. Un GGM permet de présenter les différentes variables sous forme de nœuds et les corrélations sont représentées par les liens entre les nœuds. L'épaisseur et la couleur du lien représentent le degré et le sens de la corrélation.



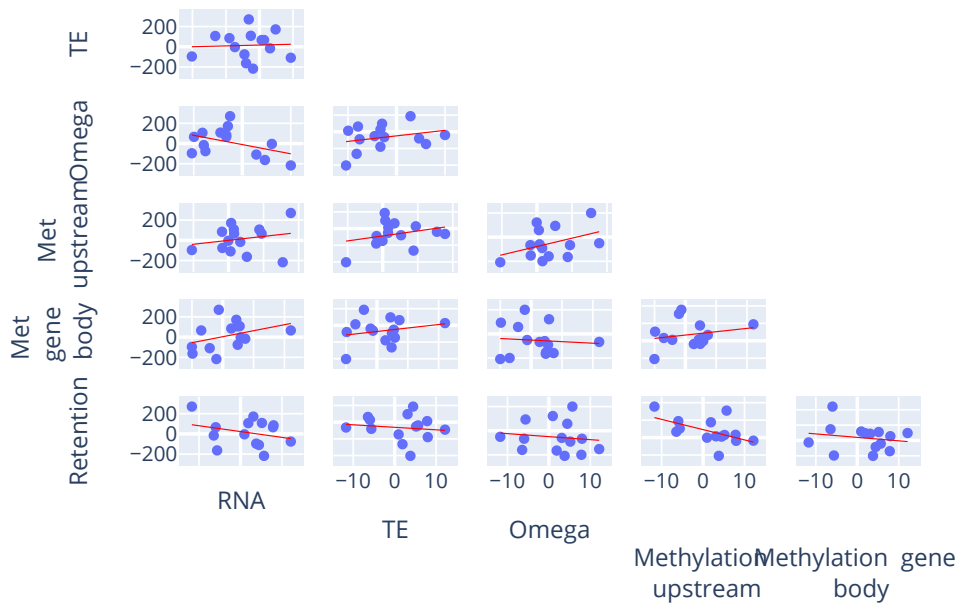
**Figure 9.2** – Ensemble de cartes thermiques regroupant les coefficients de corrélation entre les différents indicateurs pour toutes les paires de fragments chromosomiques ohnologues analysées.



**Figure 9.3** – Représentation des corrélations sous forme de GGM. Les différentes variables sont représentées sous la forme de nœuds et les corrélations sont représentées par les liens entre les nœuds. L'épaisseur et la couleur du lien présentent la force et le sens de la corrélation.

Afin de mieux comprendre la relation entre le déséquilibre de QTLs et les différents mécanismes étudiés au cours de la thèse le plus fidèlement possible, nous nous sommes attachés à décrire les corrélations en nous basant sur le niveau de transcription de gènes. En effet, les QTLs présentent des effectifs faibles et tous les blocs de synténie ne sont pas associés à des QTLs, ce qui rend l'analyse par corrélation difficile. Néanmoins, j'ai choisi ici de m'appuyer sur l'hypothèse que le déséquilibre d'expression des gènes ohnologues est la cause du déséquilibre d'identification de QTLs sur les fragments chromosomiques ohnologues, comme le montrent les analyses de corrélation menées dans le chapitre 6 et les données de la littérature (Renny-Byfield et al., 2017).

Tout d'abord, on peut observer le lien entre le déséquilibre transcriptionnel et le déséquilibre de méthylation des cytosines dans le contexte CG dans les régions exoniques. Ainsi pour les paires de chromosomes 1-7, 2-15, 3-11, 9-17, 13-16 nous avons constaté une



**Figure 9.4** — Ensemble des matrices de dispersion des différentes variables et des lignes de régression associées pour les paires 1-7. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.

corrélation positive et forte entre ces deux variables. Cette tendance suggère que pour un bloc synténique donné, la présence d'un déséquilibre en faveur d'un des chromosomes dans la méthylation sur le corps du gène pour le contexte CG peut être lié à un déséquilibre du niveau d'expression en faveur de ce même chromosome. Chez les paires 5-10, 6-14 et 08-15, nous n'avons pas noté de corrélations ayant un sens biologique, avec des corrélations très faibles en termes de valeurs ( $< 0.07$ ). Cette étude a été réalisée chez d'autres organismes et notamment l'humain (S. Li et al., 2018) et le peuplier (L. Liang et al., 2019), un organisme autoploïde chez qui une corrélation positive a été observée. Elle pourrait s'expliquer par le fait que le niveau de méthylation du corps des gènes est associé à leur niveau de transcription. Le niveau de méthylation serait ainsi un marqueur de la transcription des gènes (Bewick & Schmitz, 2017; J. Choi et al., 2020). À l'inverse, chez *B. rapa*, il a été constaté une anti-corrélation entre ces deux mécanismes (X. Chen et al., 2015).

De même, on peut constater un lien de corrélation entre le déséquilibre du niveau d'expression des gènes et le déséquilibre des taux de méthylation des cytosines en contexte CHH en amont des gènes ohnologues. Ainsi pour les paires 1-7, 3-11 et 13-16, nous avons observé une corrélation négative entre ces deux variables. Ce résultat suggère que plus le déséquilibre d'expression génique médian du bloc de synténie augmente pour un fragment chromosomique donné, moins le déséquilibre de méthylation des cytosines en amont des gènes ohnologues est en faveur de ce fragment chromosomique. Les paires 2-15, 8-15 et 6-14 montrent un constat inverse. Les paires 5-10 et 9-17 ne montrent pas de corrélations ayant un sens biologique entre le niveau d'expression des gènes et le déséquilibre des taux de méthylation des cytosines en contexte CHH en amont des gènes ohnologues. Les méthylations en amont des gènes sont suspectées comme ayant un effet répresseur de la transcription du gène pour tous les contextes (Anastasiadi et al., 2018; Cokus et al., 2008; Lister et al., 2008; H.-Y. Zhang et al., 2016). Il semble ici présenter un profil peu reproductible avec trois paires corrélées positivement, trois négativement et deux non corrélées, ce qui suggère un lien peu fiable et dont l'interprétation biologique est compliquée. Néanmoins, la différence du nombre de cytosines en contexte CHH en amont de paires de gènes ohnologues a pu être associée à un enrichissement en gènes non commutants. Ce groupe de gènes présentant un déséquilibre transcriptionnel très fort est ainsi lié à la méthylation en amont des gènes en contexte CHH et suggère que ce sont plutôt le nombre de positions en amont des gènes qui influe sur l'expression des gènes.

Par ailleurs, nous avons aussi observé un lien entre le déséquilibre dans le niveau

d'expression et la pression de sélection. En effet, nous avons observé pour les paires 1-7, 2-15, 6-14, 8-15, 13-16 une forte corrélation négative entre le niveau d'expression des gènes et la pression de sélection. Les paires 3-11 et 5-10 ne présentent pas de corrélations, et la paire 9-17 présente une modeste corrélation positive. Cette observation suggère qu'en général plus le niveau global d'expression des gènes présents dans le bloc de synténie est important, plus la valeur d' $\omega$  est petite. Ce type d'observation a été constaté chez d'autres espèces de plantes et notamment le maïs (Pophaly & Tellier, 2015) et *Brassica juncea* (J. Yang et al., 2016). Elles pourraient s'expliquer par l'hypothèse que les séquences les plus exprimées sont les plus soumises à une pression de sélection allant vers une conservation de la séquence afin de préserver la fonction.

Concernant le lien entre les ETs et le niveau de transcription, il semble qu'une corrélation négative soit retrouvée chez les paires 2-15, 3-11 et 5-10. Ceci suggère que les fragments chromosomiques d'une paire présentant les plus hauts niveaux d'expression de gènes présentent des densités et des couvertures en ET en général plus faible. Les paires 1-7, 8-15, 6-14, 9-17 et 13-16 ne présentent pas de corrélations ou une petite corrélation positive. Les ETs sont attendus comme ayant différents effets sur le génome et entre autres un effet répresseur sur la transcription (Chuong et al., 2017). L'effet répresseur est en particulier mis en place par une répression épigénétique via de la méthylation de l'ADN et des modifications chimiques des histones (Lippman et al., 2004 ; Rebollo et al., 2011).

Concernant la relation entre le déséquilibre du taux de rétention des gènes et le déséquilibre transcriptionnel, nous avons globalement observé une corrélation positive. Ainsi cette observation suggère que les fragments chromosomiques ohnologues comportant les gènes les plus exprimés sont les moins fractionnés. Ce type de relation a été observé au sein de différentes espèces (Z. Liang & Schnable, 2018) et notamment *B. rapa* (Cheng et al., 2012) et *C. bursa-pastoris* (Z. Liang & Schnable, 2018). Chez *M. domestica*, nous avons montré dans cette thèse, le lien entre le déséquilibre du taux de rétention des gènes et le déséquilibre transcriptionnel pour les paires 2-15, 5-10, 6-14, 08-15 et 13-16. Cette relation semble importante avec des corrélations positives et fortes. Les paires 1-7, 3-11 et 9-17 présentent de légères corrélations négatives ou pas de corrélations.

De l'ensemble de ces constatations, on peut déduire que globalement le déséquilibre transcriptionnel est corrélé positivement avec le taux de rétention des gènes et la méthylation de l'ADN dans le contexte CG des régions exoniques. À l'inverse, les méthylation de l'ADN en amont des gènes en contexte CHH, la pression de sélection et la couverture en ET sont inversement corrélés. Il est important de noter que toutes les observations

décrites ici ne sont pas valables pour toutes les paires. Pour certaines paires, notamment celles qui ne sont pas ou peu déséquilibrées notamment en termes de QTLs, comme les paires 5-10 ou 9-17, nous n'avons pas identifié de corrélations entre les indicateurs ou celles-ci sont trop faibles pour constituer une réalité biologique ou statistique. De même, comme précisées, certaines paires déséquilibrées peuvent présenter des comportements différents de ceux globalement attendus.

**Table 9.1** – Résultats des corrélations de Pearson pour les couples de fragments chromosomiques synténiques ohnologues

couple	Variable	Expression	ET	Omega	Méthylation		
					en amont des gènes	Rétention	Méthylation exon
01-07	Expression	1,00	0,12	-0,62	-0,32	-0,20	0,30
	ET	0,12	1,00	-0,23	-0,54	0,41	0,07
	Omega	-0,62	-0,23	1,00	0,39	0,20	-0,22
	Méthylation en amont des gènes	-0,32	-0,54	0,39	1,00	-0,65	-0,37
	Rétention	-0,20	0,41	0,20	-0,65	1,00	0,18
	Méthylation exon	0,30	0,07	-0,22	-0,37	0,18	1,00
	Expression	1,00	-0,41	-0,98	0,77	0,30	0,12
02-15	ET	-0,41	1,00	0,57	0,06	0,53	-0,94
	Omega	-0,98	0,57	1,00	-0,71	0,09	-0,28
	Méthylation en amont des gènes	0,77	0,06	-0,71	1,00	-0,09	-0,39
	Rétention	0,30	0,53	0,09	-0,09	1,00	-0,48
	Méthylation exon	0,12	-0,94	-0,28	-0,39	-0,48	1,00
	Expression	1,00	-0,35	-0,04	-0,16	0,01	-0,16
	ET	-0,35	1,00	-0,38	-0,20	0,18	0,41
03-11	Omega	-0,04	-0,38	1,00	0,29	-0,13	<0,001

Table 9.1 continued from previous page

couple	Variable	Expression	ET	Omega	Méthylation en amont des gènes	Rétention	Méthylation exon
05-10	Méthylation en amont des gènes	-0,16	-0,20	0,29	1,00	-0,31	-0,08
	Rétention	0,01	0,18	-0,13	-0,31	1,00	-0,02
	Méthylation exon	-0,16	0,41	<0,001	-0,08	-0,02	1,00
	Expression	1,00	-0,23	-0,03	0,01	0,24	-0,02
	ET	-0,23	1,00	0,12	0,32	-0,20	0,22
	Omega	-0,03	0,12	1,00	0,17	-0,25	0,15
	Méthylation en amont des gènes	0,01	0,32	0,17	1,00	-0,32	<0,001
	Rétention	0,24	-0,20	-0,25	-0,32	1,00	0,29
	Méthylation exon	-0,02	0,22	0,15	<0,001	0,29	1,00
	Expression	1,00	0,30	-0,59	0,57	0,43	-0,47
06-14	ET	0,30	1,00	-0,19	0,72	-0,27	0,33
	Omega	-0,59	-0,19	1,00	-0,21	0,10	0,57
	Méthylation en amont des gènes	0,57	0,72	-0,21	1,00	0,02	0,12
	Rétention	0,43	-0,27	0,10	0,02	1,00	-0,44
	Méthylation exon	-0,47	0,33	0,57	0,12	-0,44	1,00
08-15	Expression	1,00	0,16	-0,13	0,27	0,41	-0,44
	ET	0,16	1,00	0,31	-0,11	-0,58	0,03
	Omega	-0,13	0,31	1,00	-0,78	-0,13	-0,33



**Table 9.1 continued from previous page**

couple	Variable	Expression	ET	Omega	Méthylation en amont des gènes	Rétention	Méthylation exon
09-17	Méthylation en amont des gènes	0,27	-0,11	-0,78	1,00	-0,01	0,23
	Rétention	0,41	-0,58	-0,13	-0,01	1,00	-0,35
	Méthylation exon	-0,44	0,03	-0,33	0,23	-0,35	1,00
	Expression	1,00	-0,02	0,21	0,04	0,06	0,23
	ET	-0,02	1,00	-0,28	0,25	-0,07	0,16
	Omega	0,21	-0,28	1,00	0,24	0,01	0,09
	Méthylation en amont des gènes	0,04	0,25	0,24	1,00	0,12	0,26
	Rétention	0,06	-0,07	0,01	0,12	1,00	-0,30
	Méthylation exon	0,23	0,16	0,09	0,26	-0,30	1,00
	13-16	Expression	1,00	0,28	-0,10	-0,52	0,56
ET		0,28	1,00	0,38	0,02	-0,04	0,26
Omega		-0,10	0,38	1,00	0,26	0,09	-0,21
Méthylation en amont des gènes		-0,52	0,02	0,26	1,00	-0,25	-0,68
Rétention		0,56	-0,04	0,09	-0,25	1,00	0,38
Méthylation exon		0,60	0,26	-0,21	-0,68	0,38	1,00

Par ailleurs, d'autres corrélations ont pu être observées entre les différents indicateurs. Ainsi, le relâchement de la contrainte sélective est corrélé positivement avec l'augmentation de la couverture en ET notamment pour les paires 1-7, 2-15, 5-10, 6-14, 8-15. Les paires 3-11, 9-17 et 13-16 présentent peu de corrélation entre ces variables ou une corréla-

tion négative. Cette observation peut s'expliquer d'un point de vue biologique. En effet, l'intégration d'ETs dans la séquence d'ADN va aboutir à un  $\omega$  qui augmente ou à l'inverse une levée de la pression de sélection sur la séquence va favoriser l'intégration d'ETs. La relation entre ces deux mécanismes a été observée dans la littérature, en particulier chez les plantes (J. Huang et al., 2015 ; Pontarotti, 2015 ; D. Zhao et al., 2016) mais a aussi été observé chez d'autres organismes et notamment l'humain (Jin et al., 2012). Nous avons aussi identifié le taux de rétention comme considérablement anti-corrélé avec la méthylation en amont des gènes pour les paires 1-7, 2-15, 5-10, 6-14, 8-15, 9-17 et 13-16. Cette corrélation est partagée par l'ensemble des paires analysées. Ce type de lien a été constaté en amont des gènes chez les sous-génomes du maïs (Renny-Byfield et al., 2017) et dans le génome de *B. rapa* (X. Chen et al., 2015). De même, le taux de rétention est fortement anti-corrélé avec la méthylation exonique des gènes pour les paires 1-7, 2-15, 6-14, 8-15 et 9-17. Les paires 3-11 et 13-16 ne présentent pas de corrélations. De telles observations ont été faites en particulier chez *B. oleracea* (Parkin et al., 2014).

Ainsi nous avons pu observer des corrélations importantes et partagées par une majorité des paires de chromosomes synténiques entre les divers indicateurs analysés au cours de la thèse. Ces corrélations, notamment avec le déséquilibre de transcription, suggèrent que le déséquilibre de QTLs observé chez le pommier pourrait être expliqué, au moins en partie, par ces différents processus. Ainsi, nous pouvons émettre l'hypothèse d'un mécanisme expliquant le déséquilibre de QTLs qui reposerait sur un déséquilibre de l'expression des gènes. Ce déséquilibre transcriptionnel serait causé par des déséquilibres dans la couverture et la densité en ET, à l'origine de différences épigénétiques qui modifieraient l'expression des gènes associés. À cela s'ajouterait une pression de sélection relâchée sur les gènes les moins transcrits qui aboutirait à la perte préférentielle de ces gènes.

## 9.2 La dominance de sous-génome

Nous avons observé un déséquilibre en QTLs entre régions synténiques pour le génome du pommier, qui peut être expliqué par un ensemble de déséquilibres au sein du génome, transcriptome et épigénome. L'ensemble des observations faites et les liens entre les différents indicateurs testés sont globalement en accord avec la littérature dans le contexte de dominance d'un sous-génome par rapport à l'autre (Eichten et al., 2013 ; Renny-Byfield et al., 2017 ; Woodhouse et al., 2014 ; M. Zhao et al., 2017). Le mécanisme de la dominance de sous génome a été observé chez différents organismes allopolyploïdes et les mécanismes

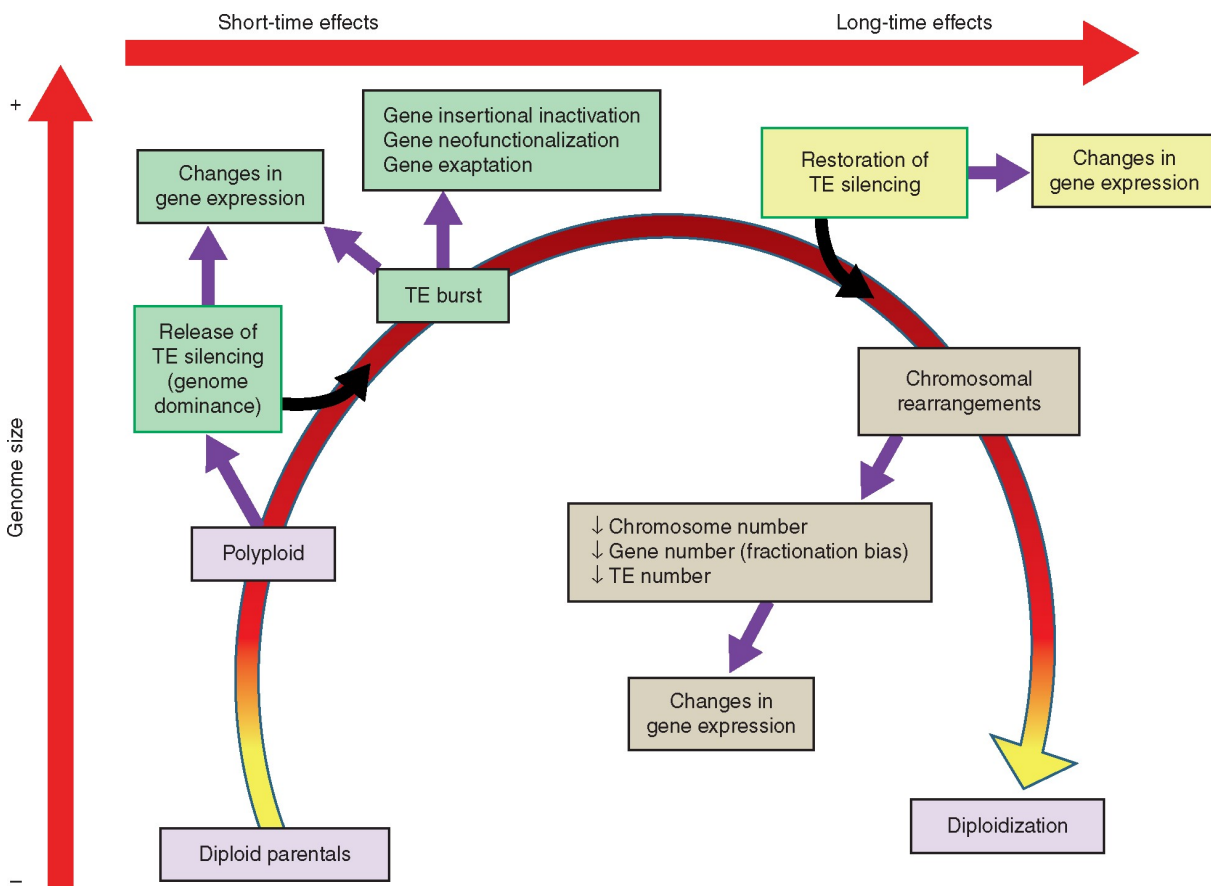
régissant ce processus ont été modélisés. Le schéma présenté en Figure 9.5 présente le mécanisme consensus. Le choc génomique dû à la duplication complète du génome serait à l'origine d'une activation massive des ETs qui aboutirait à une explosion du nombre d'ETs insérés *de novo*. Chez les allopolyploïdes, les différences de proportion d'ETs entre les sous-génomes avant hybridation pourraient être à l'origine des différences de répartition des ETs dans le génome. Les ETs étant fortement reliés à la dynamique de méthylation de l'ADN dans les régions voisines de leur lieu d'insertion, les différences d'ETs vont ainsi être retrouvées au niveau de la méthylation de l'ADN (C. M. Vicient & Casacuberta, 2017). Ce déséquilibre épigénétique peut alors impacter l'équilibre transcriptionnel entre les couples de gènes synténiques (Contreras et al., 2015). Ce changement dans l'équilibre transcriptionnel de l'organisme va avoir des conséquences diverses et en particulier va être à l'origine des différences de participation au phénotype de l'organisme et de la pression de sélection appliquée aux séquences codantes ohnologues. Dès lors, les différences de pression de sélection appliquée, une participation moindre à la production de protéines et au phénotype de l'organisme, ainsi que des proportions plus importantes d'ETs vont aboutir à une différence de rétention de gènes entre les deux sous-génomes post-WGD. Ce biais de fractionnement du génome va renforcer les processus précédemment décrits et mettre en place une domination durable d'un des sous-génomes par rapport à l'autre. Ces différentes assertions sont supportées par des éléments observés chez différentes espèces avec des WGDs. Ainsi, la relation causale entre le déséquilibre transcriptionnel et le biais de fractionnement a été modélisée (J. C. Schnable & Freeling, 2011 ; J. C. Schnable et al., 2011). Ces modèles suggèrent que la présence de l'un des deux biais suppose la présence de l'autre biais. De plus, ce modèle suggère que les couples de gènes équilibrés en termes de niveau d'expression sont moins sujets au fractionnement. En effet, la perte de l'un des gènes entraîne la perte de 50 % de l'expression totale de la paire de gènes. Dans le cas de paires où l'expression est biaisée, la perte du gène moins exprimé n'entraînera pas de conséquences importantes d'un point de vue de la production protéique (Garsmeur et al., 2014 ; Woodhouse et al., 2014). Ainsi, l'étude du fractionnement du génome pourrait permettre en première approche de tester si l'organisme considéré pourrait être sujet au phénomène de dominance de sous-génome. De même, le lien entre l'évolution de séquences, estimée par le  $K_s$ , le biais de fractionnement et le déséquilibre transcriptionnel a été observé et une tendance semble se dégager. Le  $K_s$ , représentant le temps de divergence, ne semble pas influencer sur l'existence ou non d'une dominance de sous-génome (Garsmeur et al., 2014). Ainsi, il a été observé des dominances de sous-génome chez des organismes

avec des WGDs anciennes et récentes. Néanmoins, le lien entre le fractionnement biaisé du génome et la dominance de sous-génome a été observé chez de nombreuses espèces avec un ratio de fractionnement à partir de 1,17 chez *A. thaliana* (Thomas et al., 2006), *Sorghum bicolor* (J. C. Schnable et al., 2012), *B. rapa* (Cheng et al., 2012; X. Wang et al., 2011) et le maïs (J. C. Schnable et al., 2011; Woodhouse et al., 2010)

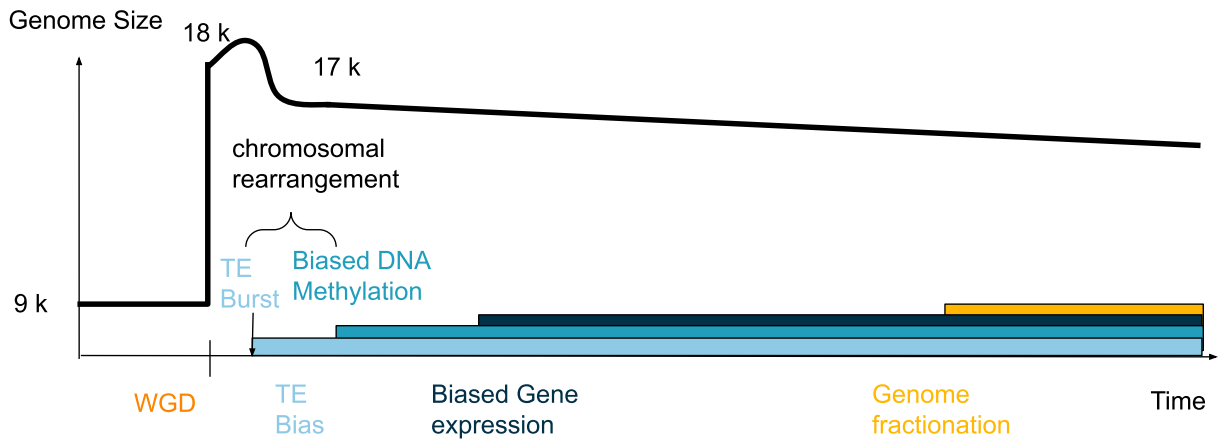
Jusqu'à présent, à notre connaissance, le mécanisme de dominance de sous-génome n'a été observé que chez des espèces allopolyploïdes. La dominance de sous-génome n'a jamais été observée chez des espèces autopolyploïdes. Par exemple, chez le poirier qui partage avec le pommier la WGD commune aux *Pomoïdae*, il n'a pas été identifié de déséquilibre dans le fractionnement, le niveau d'expression des gènes et la méthylation des gènes et de leur environnement génique proche (Q. Li et al., 2019). Ce constat est néanmoins sujet à caution, car l'étude a été conduite à l'échelle du génome et non à l'échelle des paires de chromosomes comme fait dans les analyses menées au cours de cette thèse. Cette méthode présente le risque de lisser les différences entre les zones synténiques les plus différentes. La dominance de sous-génome n'a pas été identifiée chez d'autres autopolyploïdes comme le peuplier (Y. Liu et al., 2017).

Néanmoins, au cours de nos analyses, nous avons pu identifier des déséquilibres significatifs dans ces différents processus chez le pommier. Ainsi, nous avons notamment pu identifier un déséquilibre dans les proportions de QTLs porté par les fragments chromosomiques synténiques, le fractionnement du génome, l'expression des gènes, la couverture et la densité en ETs, la méthylation de l'ADN dans les régions exoniques des gènes ohnologues et la méthylation en amont des gènes ohnologues. Ces déséquilibres ne concernent pas systématiquement toutes les paires de fragments chromosomiques synténiques testées. Ainsi, il semblerait que le pommier présente un processus proche de celui de la sous-dominance génomique, mais qui s'appliquerait à l'échelle des paires de fragments chromosomiques synténiques. Ce processus pourrait être décrit par le terme de dominance sous-chromosomique. Ce type d'observation n'a encore jamais été faite chez un autopolyploïde.

Afin d'expliquer ces différents déséquilibres, nous proposons un modèle de mécanisme qui pourrait rassembler toutes ces observations sur le pommier. Ce mécanisme est résumé dans la figure 9.6. Ainsi, le pommier aurait connu une phase d'explosion des ETs qui a suivi la WGD. Cette explosion a été identifiée chez le pommier (Daccord et al., 2017) et datée à 21 millions d'années (Daccord et al., 2017). Ce type de processus est retrouvé chez différentes espèces dans lesquelles une augmentation du nombre d'ET a été observée



**Figure 9.5** – Schémas des mécanismes post-WGD et de la dynamique des ETs. Issu de (C. M. Vicent & Casacuberta, 2017). La WGD est suivie par une levée de la répression des ETs. Cette levée va mobiliser les ETs et induire des changements dans la régulation des gènes proches d’ETs. L’explosion du nombre d’ET va produire de nouvelles insertions qui vont aussi modifier la régulation des gènes. La répression épigénétique des ETs est rétablie après quelques générations. Néanmoins, la régulation et l’expression des gènes proches des ETs restent modifiées. De même les ETs ont un rôle important dans la diploïdisation en particulier via le mécanisme de recombinaison qui aboutit à des délétions et des réarrangements chromosomiques.



**Figure 9.6** – Schéma récapitulatif du mécanisme hypothétique mis en œuvre dans le génome de la pomme. Après la WGD, le génome de l’ancêtre commun des *Maloidae* est passé de 9 à 18 chromosomes. Rapidement après la WGD, une explosion d’ET s’est produite, entraînant une augmentation de la taille du génome. Cette explosion d’ET a probablement été biaisée et suivie d’importantes méthylations de l’ADN. Ces mécanismes permettent d’établir un réarrangement chromosomique qui entraînera à un génome de 17 chromosomes. De plus, ces mécanismes mettent en place des biais qui vont conduire à un biais d’expression des gènes entre les différentes paires de chromosomes, une perte préférentielle des gènes dupliqués sur certains chromosomes par rapport à leurs ohnologues et une participation différente des fragments chromosomiques ohnologues aux variations phénotypiques.

en post-WGD (McClintock, 1984). Cette explosion d’ET s’est faite de façon inégale et certains chromosomes ont vu l’insertion de plus d’ET que leur ohnologue. Ce biais dans les proportions d’ETs porté par les chromosomes ohnologues a entraîné un déséquilibre de la méthylation de l’ADN entre les régions ohnologues. Les méthylations de l’ADN sont associées au niveau de transcription des gènes proches par le biais d’un processus de répression de l’expression des gènes. Ainsi, le déséquilibre des niveaux de méthylations est impliqué dans la mise en place d’un déséquilibre de l’expression des gènes. Ces différences transcriptionnelles pourraient alors conduire à un biais dans la méthylation de l’ADN du corps du gène et à des variations dans la participation au phénotype de l’organisme entre les segments chromosomiques ohnologues. Ce déséquilibre peut s’observer notamment à l’aide du déséquilibre de QTLs. On peut également imaginer que ces différences d’expression entre les gènes ohnologues, que l’on attend comme présentant des fonctions et des niveaux d’expressions similaires, puissent être à l’origine de différences de pression de sélection sur les gènes puisque les gènes les moins transcrits sont moins susceptibles d’être sous pression de sélection. De plus, l’insertion d’ETs associée à une levée de la pression sélective pourrait aboutir à une différence dans la perte de gènes entre les régions synténiques. Nous n’avons pas encore constaté un tel phénomène au sein du pommier, bien que

des corrélations négatives entre le rapport  $\omega$  et la méthylation en amont des gènes ainsi que l'expression aient été observées. Cela pourrait être dû à l'âge récent de la WGD qui n'a pas encore permis la mise en place d'un tel mécanisme. Contrairement à ce qui a été observé chez les espèces allopolyploïdes, le mécanisme dont nous faisons l'hypothèse ici serait beaucoup plus lent à se mettre en place, car les différences d'ET originelles entre fractions de chromosomes ohnologues sont beaucoup moins marquées voir inexistantes comparées aux chez les espèces issues d'une polyploïdisation par allopolyploïdie.

D'un point de vue de la temporalité, notre modèle présume d'une explosion des ETs dans un temps proche suivant la WGD afin que les différents processus découlant du déséquilibre supposé d'ET se mettent en place. Les WGD sont une source de grands changements dans la structure du génome dupliqué, ce qui peut être à l'origine d'une levée de la répression épigénétique qui mène à des activations massives d'ETs. Une explosion d'ETs a été observée chez le pommier. Celle-ci a été datée par une approche se basant sur un taux estimé de mutation des ETs (Lander et al., 2001). L'explosion d'ETs a ainsi été datée à 21 millions d'années (Daccord et al., 2017). Par ailleurs, notre estimation de la datation de la WGD par extrapolation linéaire du  $K_s$  estime la WGD à 27 millions d'années. Cette estimation est confirmée par les estimations les plus récentes qui suggèrent une WGD ayant eu lieu entre 13 et 27 millions d'années (Su et al., 2021). Ces différentes échelles de temps semblent cohérentes avec le modèle présenté ici.

De cette façon, nous avons pu identifier un ensemble de relation entre les différentes analyses menées au cours de cette thèse. Le déséquilibre de QTLs observé entre certains fragments chromosomiques synténiques semble être lié à un ensemble de déséquilibres. Ainsi, un déséquilibre transcriptionnel a été observé. Celui-ci est corrélé positivement avec la surfractionnement du génome et un taux important de méthylation de l'ADN des régions exoniques en contexte CG. À l'inverse le déséquilibre des niveaux d'expression des gènes est inversement corrélé avec une forte couverture en ETs, une méthylation de l'ADN en amont des gènes importants et une pression de sélection relâchée. Ces observations suggèrent la présence d'une sous-dominance génomique chez le pommier. Le processus de mise en place de cette dominance est sans doute différent de celle telle quelle est conçue pour les organismes allopolyploïdes. C'est ainsi que nous proposons un modèle de mise en place de ce déséquilibre pour le pommier, un organisme autoploïde.

La polyploïdie est un mécanisme qui coïncide avec des périodes importantes de stress comme des changements importants du climat ou des événements géologiques importants (Koenen et al., 2021 ; Novikova et al., 2018 ; Van de Peer et al., 2021 ; Van de Peer

et al., 2017; S. Wu et al., 2020). Ainsi la polyploïdie semble être un mécanisme clé dans la domestication et l'évolution de la réponse au stress (Renny-Byfield et al., 2014), et un avantage sélectif par rapport aux stress biotiques et abiotiques (Van de Peer et al., 2017).

La relation entre la polyploïdie et la résistance aux pathogènes est complexe (King et al., 2012) et semble liée à des réponses de type de "*gene dosage*" et surtout associée à des caractères monogéniques. Néanmoins, il a été constaté chez différentes espèces que les polyploïdes répondaient différemment aux stress biotiques par rapport aux diploïdes. Par exemple, chez le tétraploïde de *Glycine tabacina*, il a été observé que celui-ci résiste plus efficacement à *Phakopsora pachyrhizi* par rapport à son diploïde (Burdon & Marshall, 1981). De même, le tétraploïde *Trifolium pratense* est plus résistant à *Sclerotinia trifoliorum* que son progéniteur diploïde. De plus, chez des néo-polyploïdes synthétiques d'un cultivar de pomme résistant monogénique, il a été observé une résistance accrue à la tavelure causée par *Venturia inaequalis* par rapport aux cultivars diploïdes (Hias et al., 2018). Cette version polyploïde plus résistante au pathogène *Riv6* de *V. inaequalis* présente un intérêt important dans la culture du pommier. Néanmoins, la résistance au *Riv6* dans les vergers a permis l'émergence de deux populations de *V. inaequalis* (Leroy et al., 2013).

Par ailleurs, il a été confirmé que *V. inaequalis* était un pathogène de l'ancêtre du pommier, *M. sieversii* (Gladieux et al., 2010). Ainsi, il a été observé que globalement il existait trois grandes populations de *V. inaequalis*, dont l'une est retrouvée en Europe, une en Asie centrale et une dans les montagnes orientales du Kazakhstan (Gladieux et al., 2010). Les deux premières populations peuvent infecter les populations de *Malus* sauvages (*M. orientalis* et *M. sylvestris*) et domestiquées (*M. domestica*). La dernière population peut infecter les *Malus sieversii*. L'étude de ces populations suggère, une séparation il y a 24 000 ans, suivie d'une co-évolution du pathosystème (Gladieux et al., 2010). Par ailleurs, l'étude du génome de certaines souches *V. inaequalis* a permis l'identification d'une explosion d'ETs. Cependant, cette invasion d'ETs n'a pas été datée, la concomitance des évènements ne peut donc être établie avec certitude. Il est néanmoins intéressant de noter que les souches de *V. inaequalis* infectant les poiriers ne sont pas concernées par l'explosion d'ETs (Gladieux et al., 2010). Étonnamment, il n'a pas été retrouvé non plus de sous dominance génomique chez cette espèce (Q. Li et al., 2019). On peut alors se demander si cette augmentation de la résistance aux pathogènes chez les organismes polyploïdes ne pourrait pas être un moteur sélectionnant chez les pathogènes des organismes polyploïdes, où la génération de matériel génétique supplémentaire pourrait être source d'innovations génétiques plus rapides permettant une réponse aux résistances des hôtes? Dans le cas de



*V. inequalis*, cette co-évolution du pommier et de son pathogène très tôt après la WGD, ainsi que les explosions d'ETs identifiées chez les deux espèces pourraient elles être la trace d'une co-évolution de ce pathosystème ?

Nous avons donc identifié une dominance sous-chromosomique. Cette dominance a des implications phénotypiques, en particulier via un déséquilibre de QTLs chez différentes paires de fragments chromosomiques synténiques. Ainsi, on peut se demander si l'identification de QTLs au sein du génome du pommier peut s'expliquer seulement par une différence allélique de gènes, d'autant plus que nous n'avons pas identifié de différence significative d' $\omega$  à l'échelle des fragments chromosomiques synténiques. De plus, il est difficile de localiser finement les QTLs, rendant les analyses à l'échelle du gène compliquée. Ainsi, on peut imaginer qu'une partie des QTLs résulte de la sous dominance sous-chromosomique qui conduit à des différences d'environnements en ETs qui aboutit à une différence dans le niveau de transcription des gènes, avec pour résultat final un effet significatif sur le phénotype des plantes.

# CONCLUSION

---

## Déséquilibres

Au cours de cette thèse, nous avons pu construire les blocs de synténie présents chez le pommier à partir de différentes approches afin de générer un ensemble de paires de gènes ohnologues les plus fiables possibles. À l'aide d'i-ADHoRe 3, nous avons pu identifier 16 779 couples gènes ohnologues répartis en 865 segments synténiques.

Nous avons alors pu récupérer 1528 QTLs et localiser un ensemble de 589 QTLs sur le génome de haute qualité de GDDH13. Ces QTLs sont associés à une diversité de traits phénotypiques non redondants. Nous avons alors pu identifier un déséquilibre dans les proportions de QTLs porté par certaines paires de fragments chromosomiques synténiques et notamment, les paires 1-7, 3-11, 13-16 et 8-15.

Dans l'intention d'expliquer ce déséquilibre de QTLs, nous avons mené différentes analyses afin de dresser un état des lieux du génome, transcriptome et épigénome du pommier.

En nous appuyant sur le génome du pêcher, nous avons réussi à étudier le fractionnement du génome du pommier via deux méthodes. Nous avons pu déceler un déséquilibre significatif des pourcentages de rétention de gènes pour les paires 1-7, 2-15, 3-11, 5-10, 6-14, 8-15, 9-17 et 13-16.

Nous avons construit par RBBH un ensemble de triplets composés de couples d'ohnologues de *M. domestica* et de leur plus proche orthologue *P. persica*. Ces triplets ont permis de calculer les ratios  $\omega$  associés à ces séquences codantes. Nous n'avons pas identifié de déséquilibre dans la pression de sélection au sein des paires de fragments chromosomiques synténiques.

Afin de tester l'équilibre transcriptionnel du pommier, nous avons récupéré et analysé 148 conditions de RNA-Seq de haute qualité avec 3 réplicats biologiques. Par le biais d'analyses d'expression différentielles des gènes ohnologues au sein de chacune des 148 conditions, nous avons identifié de multiples gènes ohnologues comme significativement différentiellement exprimés dans les différentes conditions testées. Après analyses statistiques, nous avons pu mettre en évidence que les paires de segments chromosomiques

---

ohnologues sont significativement différentes sur le plan du nombre de gènes surexprimés à travers l'ensemble des expériences.

Nous avons identifié un ensemble de paires de fragments synténiques ohnologues comme significativement différent en termes de couverture en ET. En particulier, les paires 1-7, 2-15, 6-14 et 8-15 apparaissent comme différentes en termes de couverture en ET. Cette dynamique biaisée semble liée aux ET dans leur ensemble plutôt qu'à une classe ou une sous-famille particulière. La présence élevée d'ET peut entraîner une méthylation élevée de l'ADN dans ces régions. Ainsi, une distribution biaisée des ET pourrait provoquer des déséquilibres dans la méthylation de l'ADN. Nous avons donc testé cette hypothèse.

À partir de 36 séquençages complets au bisulfite (de haute qualité) et associés à une diversité de tissus, nous avons cherché à étudier la méthylation de l'ADN des segments synténiques. Nous avons calculé les rapports de méthylation pour chacune des positions de cytosine, dans les contextes CHH, CHG et CG. Nous avons identifié les paires 1-7, 2-15 et 3-11 comme significativement déséquilibrées en termes de ratio de cytosines méthylées pour les trois contextes et l'ensemble des régions analysées. En analysant et en comparant les comptages de positions de cytosines associées au contexte CHH entre les paires de gènes ohnologues, nous avons identifié un enrichissement significatif en gènes non commutables pour les gènes les plus différents en termes de nombre de positions CHH. Cette observation pourrait être une source d'explication à la présence de cet ensemble de gènes particulier. Les gènes non commutants, identifiés lors de l'analyse de l'expression des gènes ohnologues, regroupent des gènes qui ont un déséquilibre du niveau d'expression par rapport à son ohnologue, similaire dans au moins 85 % des expériences de transcriptomiques analysées.

## Perspectives

Nous avons mis en évidence un ensemble de déséquilibre génétique, transcriptomique et épigénétiques dans le génome du pommier et en particulier le génome de *Golden Delicious* double haploïde. Afin d'aller plus loin, il pourrait être intéressant d'étudier si ces déséquilibres peuvent aussi être identifiés chez d'autres génomes séquencés de pommier domestiqué (*M. domestica*), et notamment le génome homozygote HFTH1 de *Hanfu* issu d'un croisement entre *Dongguang* x *Fuji* ou le génome de Gala (X. Sun et al., 2020). Ces analyses pourraient être particulièrement intéressantes, notamment dans le cas de HFTH1, où d'importantes variations génomiques ont été observées, principalement liées aux ET (L. Zhang et al., 2019). Par ailleurs, il pourrait être intéressant de vérifier si

---

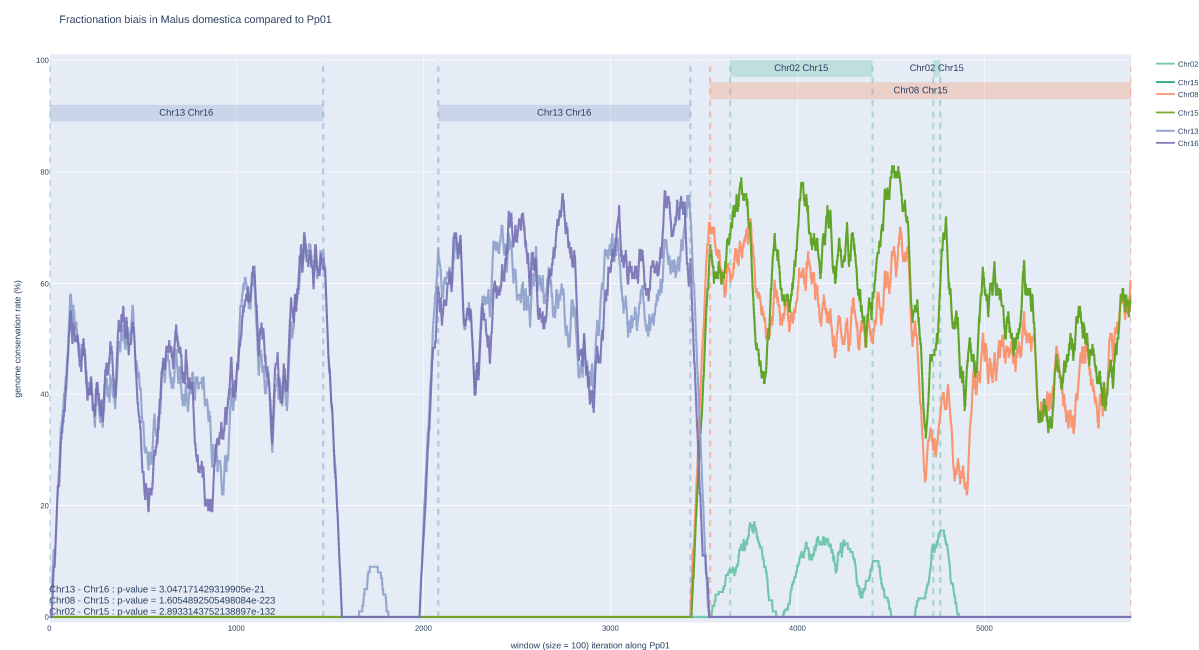
ces mécanismes sont aussi existants chez des pommiers sauvages comme *M. sieversii*. En effet, cet organisme n'a pas subi de sélection pour des traits d'intérêt agronomique. Il pourrait ainsi être intéressant de voir si un déséquilibre associé au niveau de QTLs ou d'autres mécanismes existe aussi. En générant un assemblage *de novo* à l'aide des informations transcriptomique générée en parallèle d'un séquençage, il pourrait être possible de produire le séquençage et les études génomiques et transcriptomiques similaires à celles produites pendant cette thèse.

Par ailleurs, il pourrait être intéressant de vérifier si un tel mécanisme pourrait être retrouvé chez le poirier. En effet, il n'a pas été identifié de sous-génome dominance dans une étude récemment publiée (mais utilisant une méthodologie différente de la nôtre), le fait d'identifier une sous-dominance chromosomique pourrait confirmer les différences mécanistiques entre les dominances associées aux autopolyploïdes et les autopolyploïdes. Pour finir, nous avons exploré un certain nombre de déséquilibres qui sont connus comme ayant un rôle dans la sous-dominance génomique. Néanmoins, certains d'entre eux et en particulier, les modifications d'histones, n'ont pas été analysés. En effet, il a été identifié chez le coton des réorganisations importantes de la dynamique tridimensionnelle de la chromatine associée à des différences dans le profil des modifications des histones (M. Wang et al., 2018). De même, nous avons identifié des ensembles de gènes particuliers et présentant des caractéristiques étonnantes. Ainsi, les gènes non-commutant ainsi que leur lien avec les gènes les plus différents en termes de nombre de positions de cytosines en contexte CHH font de cet ensemble de gènes une piste intéressante à investiguer.



# APPENDICES

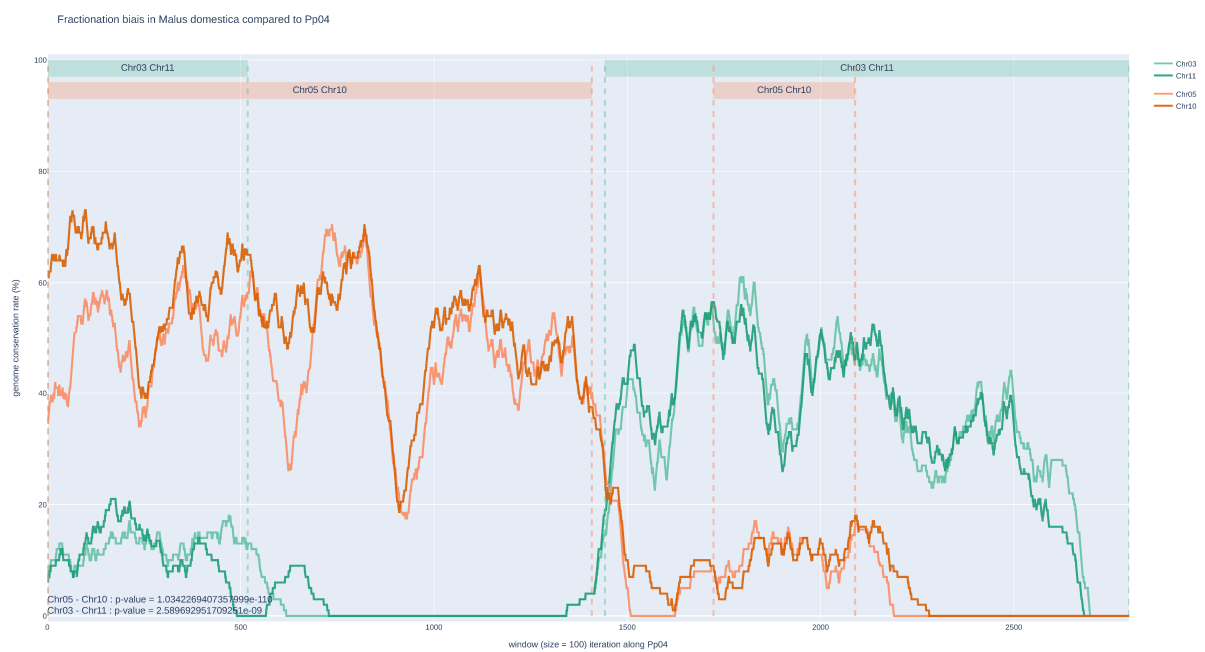
## A.1 Étude du fractionnement génomique chez le pommier



**Figure A.1** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.



**Figure A.2** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.



**Figure A.3** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.





**Figure A.4** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.





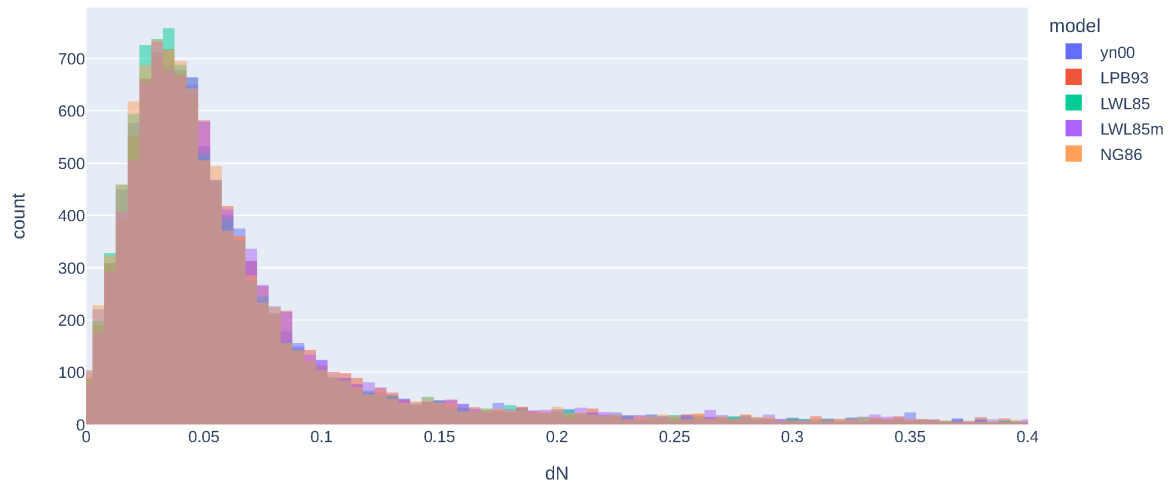
**Figure A.6** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.



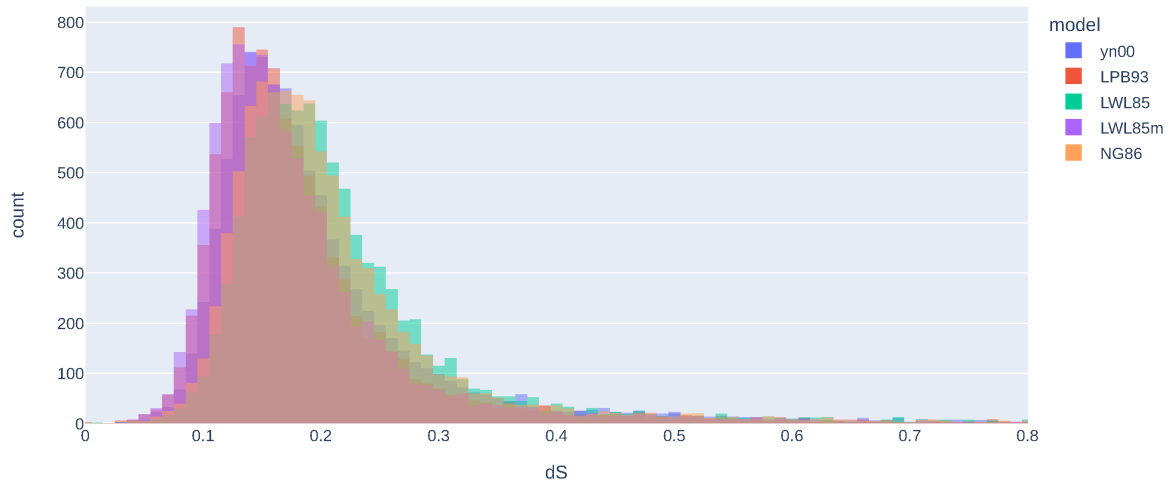
**Figure A.7** – Distribution des pourcentages de rétention des gènes ohnologues de *M. domestica* en utilisant une fenêtre glissante de 100 gènes le long du chromosome 2 de *P. persica*. Chaque ligne colorée représente le pourcentage de rétention d'un chromosome du pommier.

---

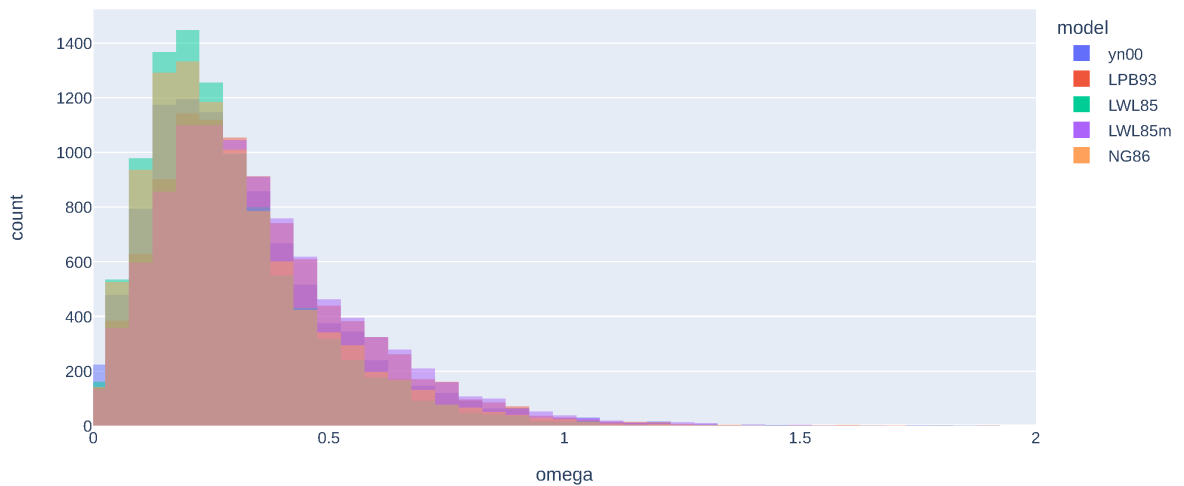
## A.2 Évaluation de la pression de sélection des gènes ohnologues



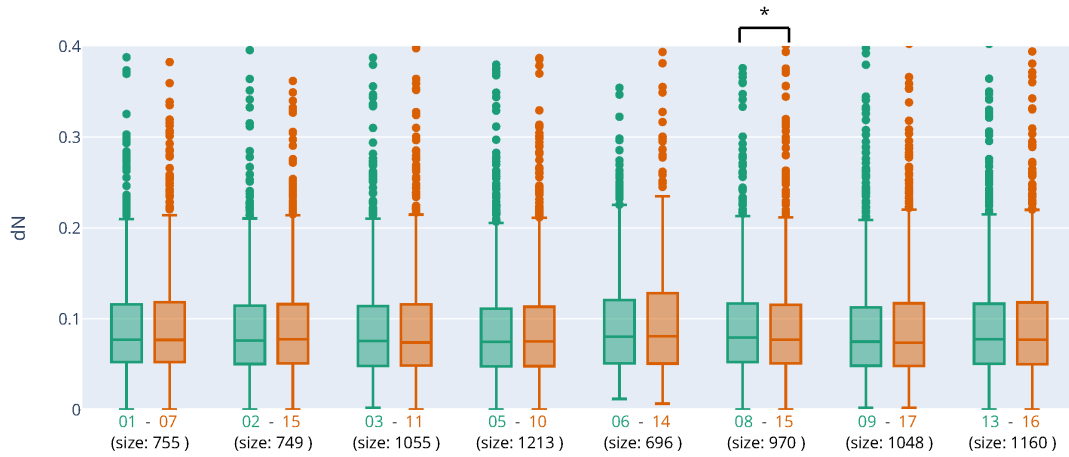
**Figure A.8** – Distribution des taux de substitution non synonyme entre les chromosomes ohnologues pour les différents modèles évolutifs implémentés dans PAML



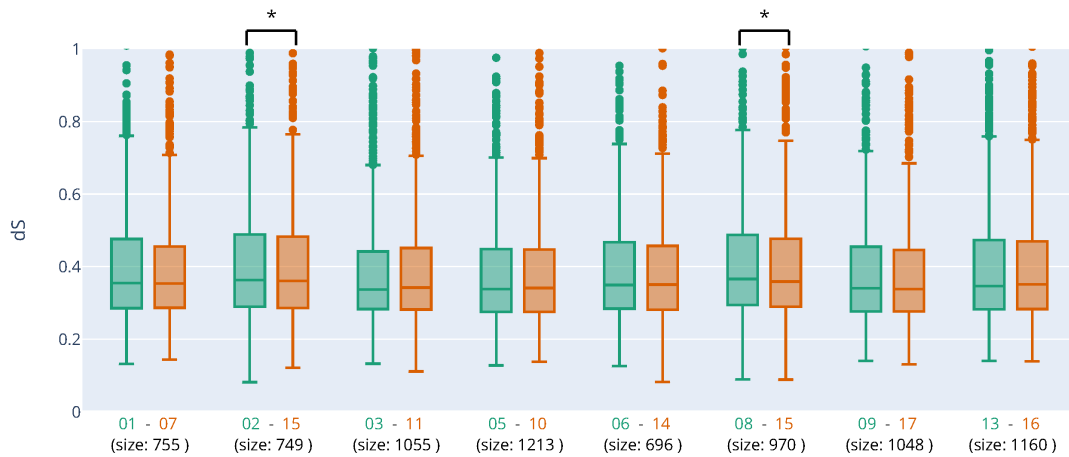
**Figure A.9** – Distribution des taux de substitution synonyme entre les chromosomes ohnologues pour les différents modèles évolutifs implémentés dans PAML



**Figure A.10** – Distribution des taux  $\omega$  entre les chromosomes ohnologues pour les différents modèles évolutifs implémentés dans PAML



**Figure A.11** – Distribution des taux de substitution non synonyme ( $K_a$ ) entre les chromosomes ohnologues



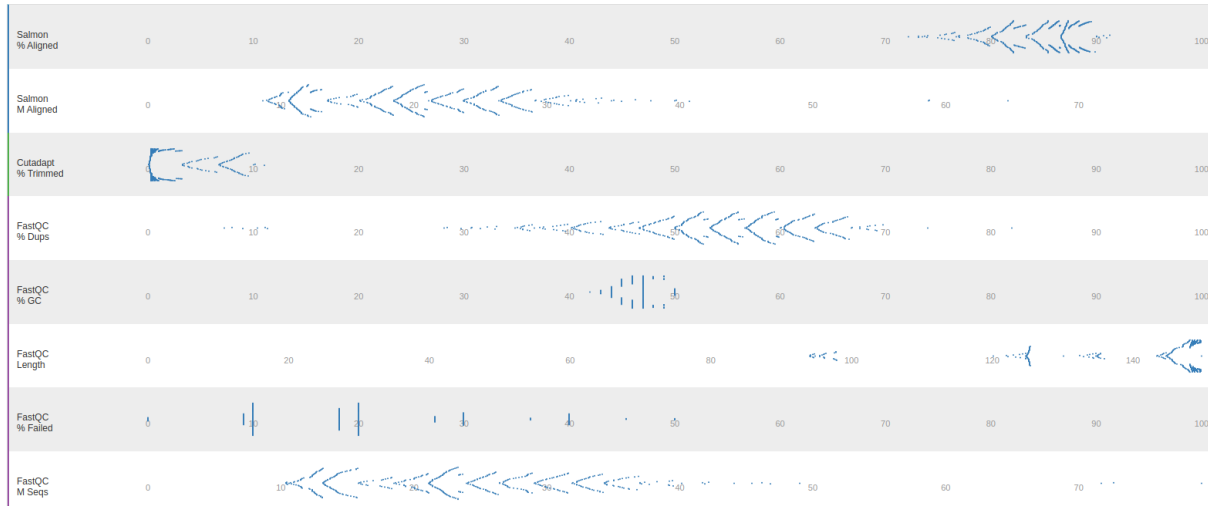
**Figure A.12** – Distribution des taux de substitution synonyme ( $K_s$ ) entre les chromosomes ohnologues

**Table A.1** – Résultats des tests de Wilcoxon et des tests de Kolmogorov-Smirnov à deux échantillons pour la qualité de l’ajustement des valeurs de  $K_s$  (dS) et  $K_a$  (dN) entre les fragments chromosomiques synténiques

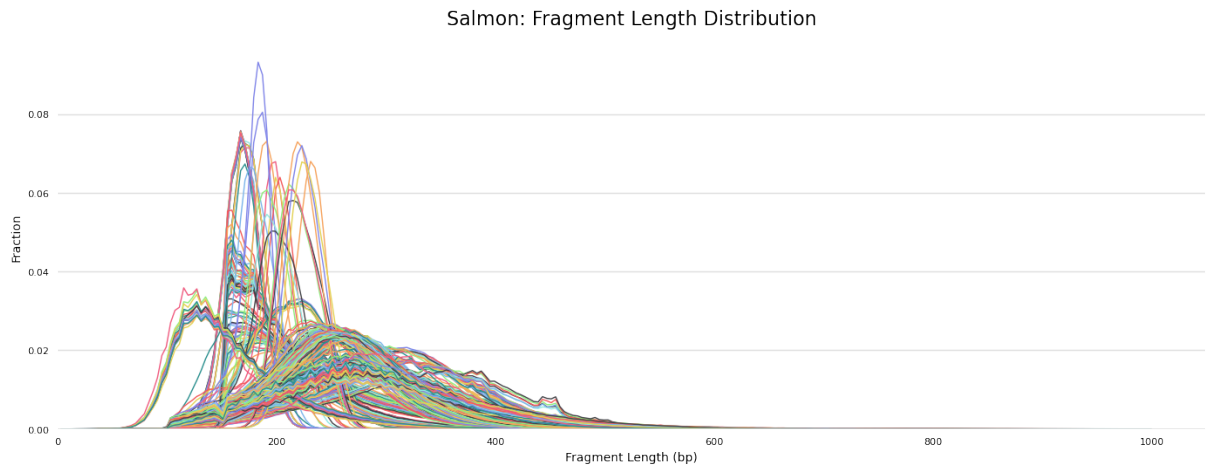
couple	p-value de Wilcoxon	p-value de Kolmogorov-Smirnov	Moyenne du premier chromosome	Moyenne du second chromosome	value
01-07	0,5138	0,8026	0,4144	0,4151	$K_s$
02-07	0,4152	0,6101	0,3765	0,3869	$K_s$
02-15	0,0425	0,9034	0,4212	0,4206	$K_s$
03-11	0,2526	0,9287	0,3919	0,3961	$K_s$
04-12	0,1517	0,8252	0,4193	0,4174	$K_s$
05-10	0,6242	0,9934	0,3881	0,3901	$K_s$
06-14	0,2185	0,9579	0,4101	0,4054	$K_s$
08-15	0,0171	0,6287	0,4213	0,4190	$K_s$
09-17	0,2596	0,9039	0,3898	0,3905	$K_s$
12-14	0,1457	0,8994	0,4023	0,4057	$K_s$
13-16	0,2671	0,9120	0,4116	0,4225	$K_s$
01-07	0,5609	0,9324	0,0942	0,0954	$K_a$
02-07	0,1268	0,9499	0,0917	0,0938	$K_a$
02-15	0,9030	0,9822	0,0920	0,0956	$K_a$
03-11	0,3412	0,9636	0,0931	0,0914	$K_a$
04-12	0,8816	0,8252	0,0964	0,0936	$K_a$
05-10	0,3131	0,9434	0,0888	0,0916	$K_a$
06-14	0,9676	0,7607	0,0960	0,0981	$K_a$
08-15	0,0078	0,5531	0,0953	0,0936	$K_a$
09-17	0,6978	0,8778	0,0907	0,0918	$K_a$
12-14	0,9353	0,8543	0,0943	0,0962	$K_a$
13-16	0,8163	0,9852	0,0935	0,0960	$K_a$



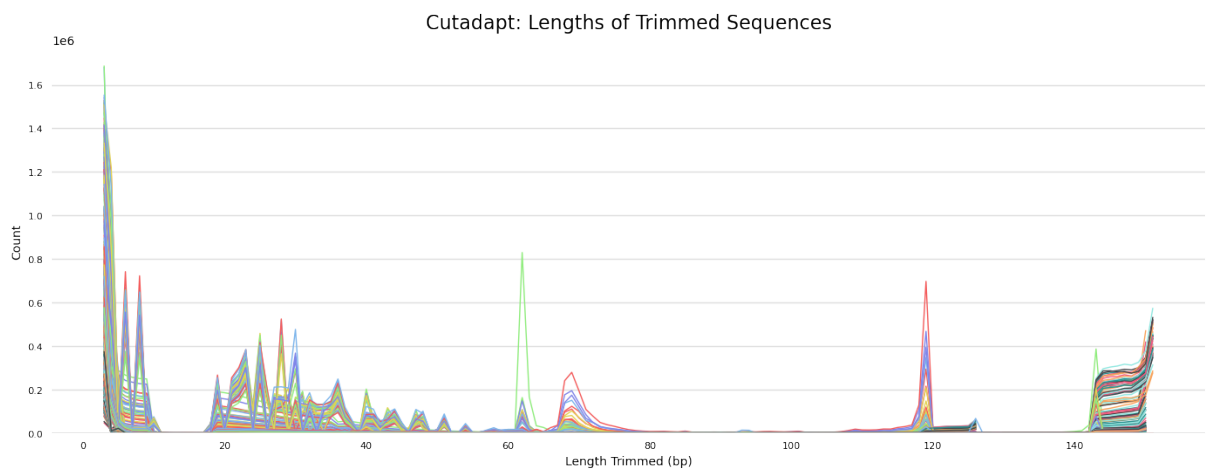
### A.3 Analyse de l'expression des gènes ohnologues



**Figure A.13** – Résumés des différents indicateurs du pipeline de pseudomapping pour les 896 échantillons (148 séries de séquençages RNA-Seq pairé avec au moins 3 répliques). Ce graphe *beeswarm* présente une vue globale de l'ensemble des principaux indicateurs de qualités liés au pipeline de pseudo alignement. Globalement l'ensemble des indicateurs assemble nominaux avec des taux d'alignements supérieurs à 70 %, un nombre de *reads* alignés compris principalement entre 10 et 30 millions, un *trimming* bien effectué par cutadapt, des pourcentages de GC autour de 50 %, des longueurs de *reads* entre 100 bp et 140 bp.

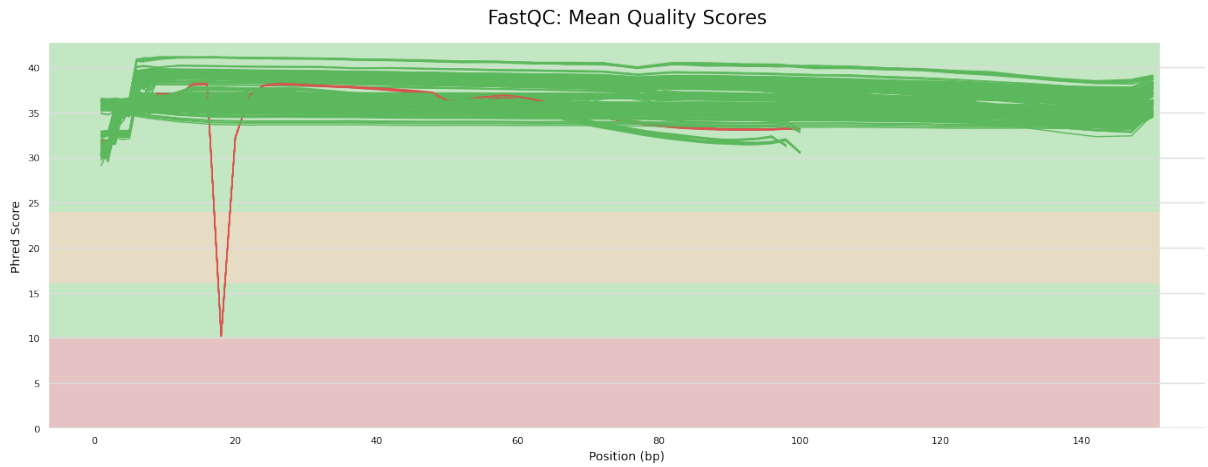


**Figure A.14** – Approximation de la distribution de la longueur du fragment observée.

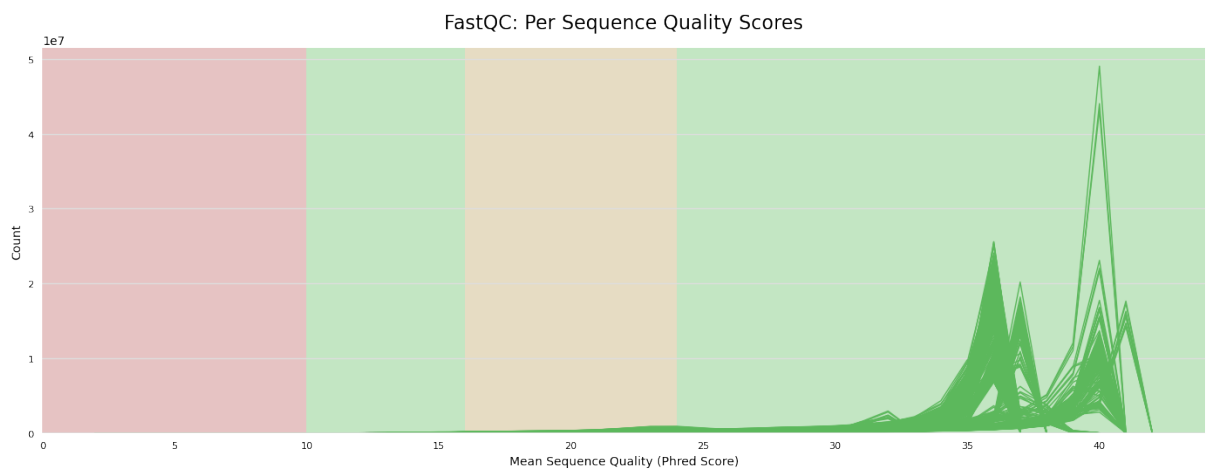


**Figure A.15** – Ce graphique montre le nombre de lectures avec certaines longueurs d'adaptateur coupées. La présence de pic défini peut être liée à la longueur de l'adaptateur. Ici, on peut observer la présence de pics en début de séquence puis à 60, 100, 120 et 140 bp ce qui correspond à des valeurs attendues avec un filtre sur des tailles de *reads* à 500bp.

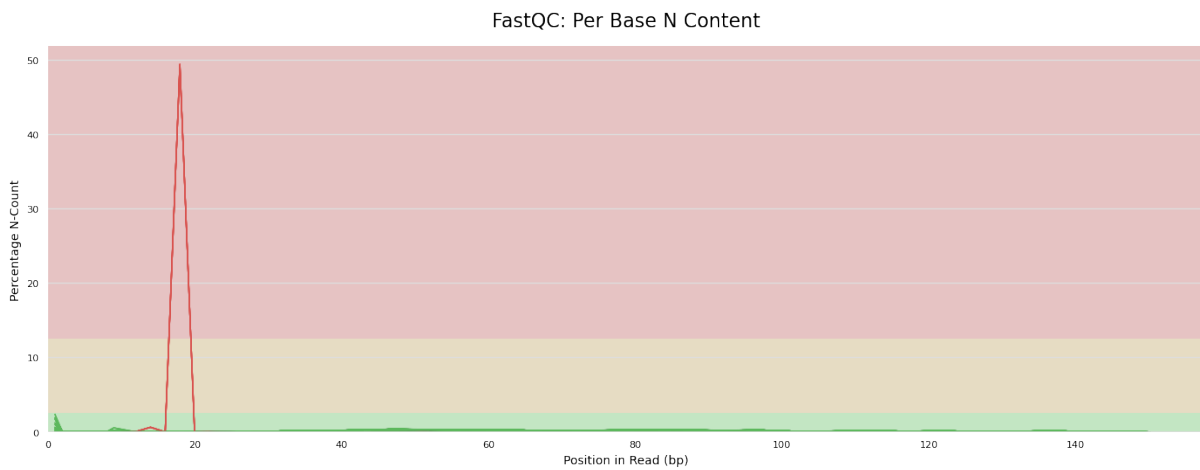




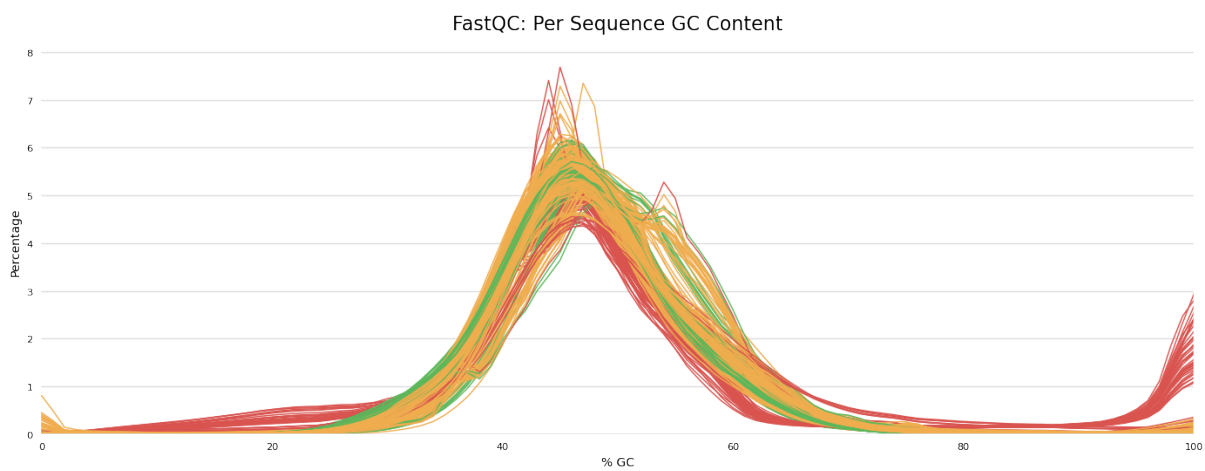
**Figure A.17** – Pour permettre une visualisation globale et lisible, seuls les scores de qualité moyens sont tracés. En y est présenté le score de qualité. Le fond du graphique précise les valeurs attendues pour considérer le séquençage de l'échantillon comme bon (en vert), moyen (orange) ou de mauvaise qualité (rouge). Il est attendu que la qualité des *reads* se dégrade au fur et à mesure du séquençage, laissant les quelques dernières bases tombées dans une qualité moyenne. Ce phénomène n'est pas alarmant.



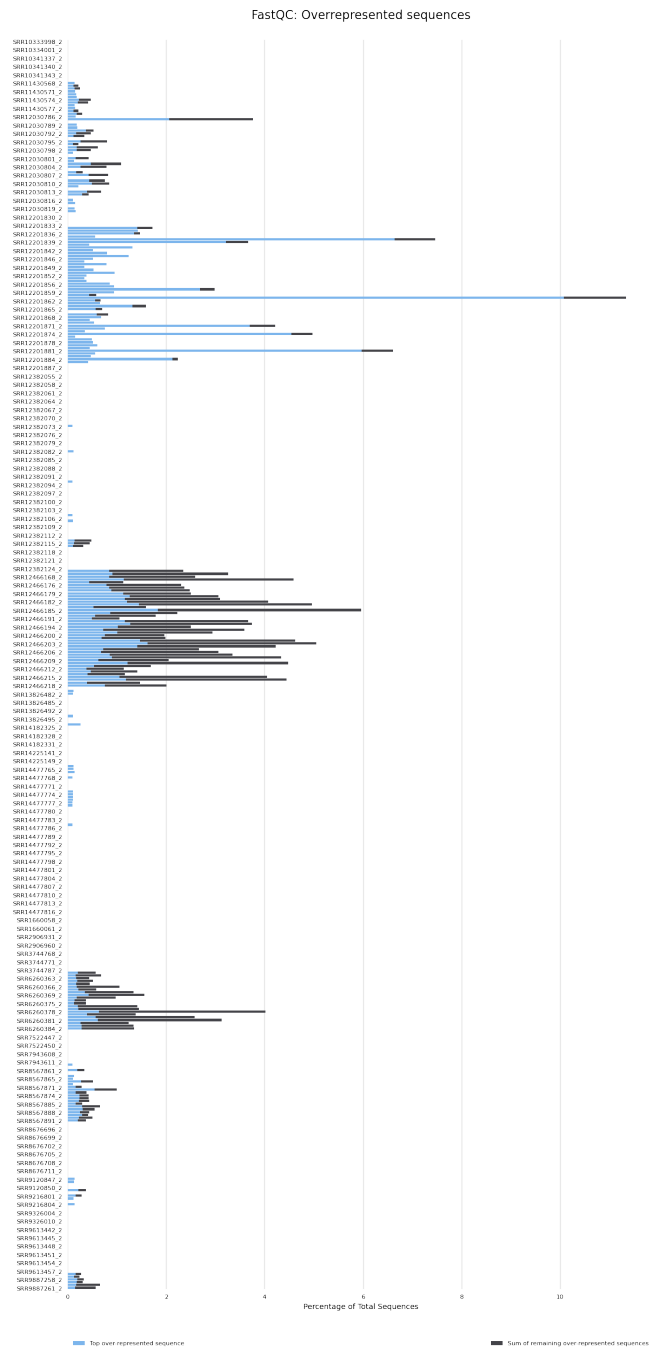
**Figure A.18** – Distribution du nombre de *reads* avec des scores de qualité moyens. Ce graphique va permettre d'estimer si un sous-ensemble de *reads* est de mauvaise qualité.



**Figure A.19** – Distribution des pourcentages d’appels de base à chaque position pour laquelle un N a été appelé. La présence de N n’est pas anormale, mais si le pourcentage dépasse 15 %, cela suggère un problème lors du séquençage de l’échantillon.



**Figure A.20** – Distribution du contenu en GC moyen des *reads*. Pour une librairie aléatoire, la valeur attendue est la distribution normale du contenu GC. Afin de s’adapter au contenu en GC du génome, étudier, la distribution attendue est construite à partir d’un sous-ensemble de donnée prise dans le fichier brut.

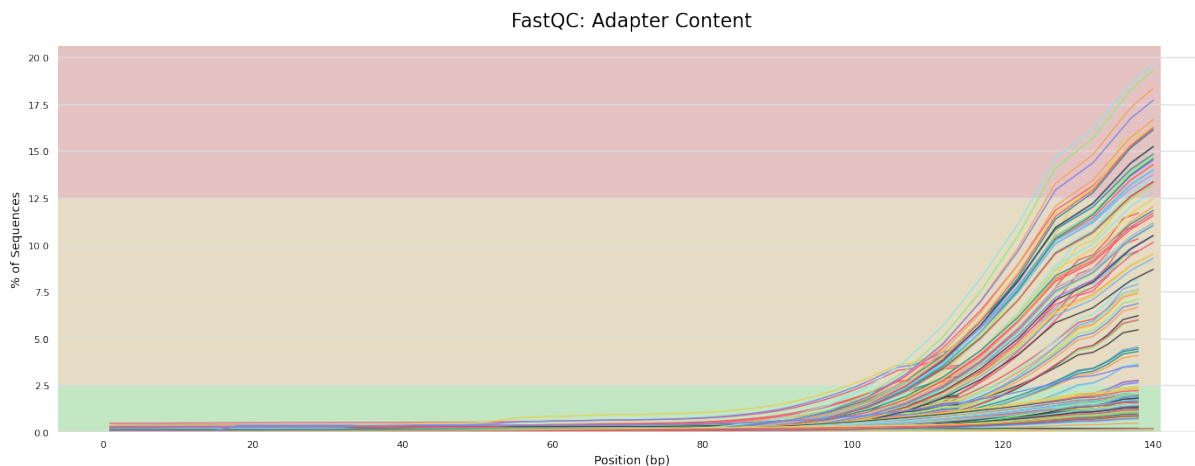


**Figure A.21** – La quantité totale de séquences surreprésentées trouvées dans chaque librairie. FastQC permet de calculer les séquences surreprésentées dans les fichiers FastQ. En effet, une seule séquence peut représenter un grand nombre de lectures dans un ensemble de données. Ainsi sur ce graphique la première barre va représenter mes *reads* surreprésentés de la séquence la plus surreprésentée. La seconde barre va montrer le nombre total pour toutes les autres séquences considérées comme surreprésentées. Comme précédemment, ces chiffres sont des estimations puisque seules les séquences qui apparaissent dans les 100 000 premières séquences sont suivies jusqu'à la fin du fichier.

---

**Table A.2** – Comptages des cultivars associés aux analyses RNA-Seq utilisées

Cultivar	Nombre	Pourcentage
Zisai Pearl' x 'Red Fuji	24	18,7500 %
Pink Lady- Cripps Pink	14	10,9375 %
Gala	14	10,9375 %
Honeycrisp	9	7,0313 %
Fuji	7	5,4688 %
apple	7	5,4688 %
Yanfu 7	6	4,6875 %
Ruixue	6	4,6875 %
Pink Lady	6	4,6875 %
Golden Delicious	5	3,9063 %
Grand Longfeng	4	3,1250 %
Golden delicious	4	3,1250 %
CG-935	4	3,1250 %
GL-3	4	3,1250 %
Nagafu No. 2	3	2,3438 %
Longfeng	2	1,5625 %
M9T337	2	1,5625 %
Rainbow 1	2	1,5625 %
M26	1	0,7813 %
Florina	1	0,7813 %
Vista Bella	1	0,7813 %
Qinguan	1	0,7813 %
Fulford Gala	1	0,7813 %



**Figure A.22** – Ensemble des distributions des pourcentages cumulés de la proportion de la librairie qui a vu chacune des séquences adaptatrices à chaque position. Il est à noter que cette distribution est cumulative, ce qui implique que la croissance du pourcentage en fin des séquences est attendue. De plus la fin des *reads* présente de façon normale des adapteurs.

**Table A.3** – Comptages des tissus associés aux analyses RNA-Seq utilisées

Tissus	Nombre	Pourcentage
Fruit	88	59,0604 %
Leaf	17	11,4094 %
Stigma-Style	14	9,3960 %
Skin	7	4,6980 %
Bud	5	3,3557 %
Shoot apex	4	2,6846 %
Apple Rootstock	3	2,0134 %
Ovule	3	2,0134 %
Hypanthium	2	1,3423 %
Stem apex	2	1,3423 %
Ovary wall	2	1,3423 %
Flower	1	0,6711 %
Stem	1	0,6711 %



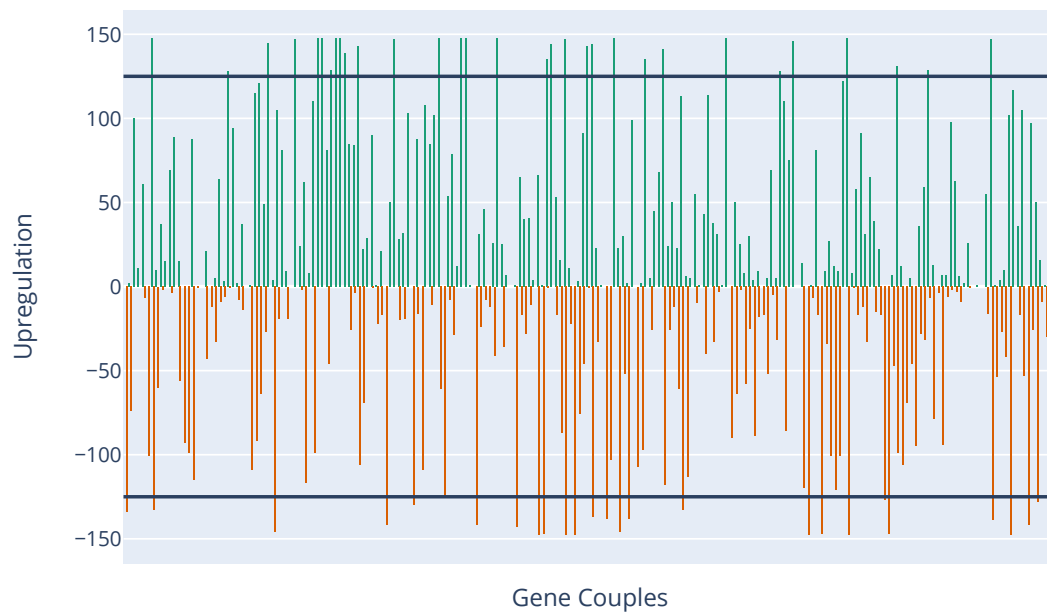
---

**Table A.4** – Comptages des traitements associés aux analyses RNA-Seq utilisées

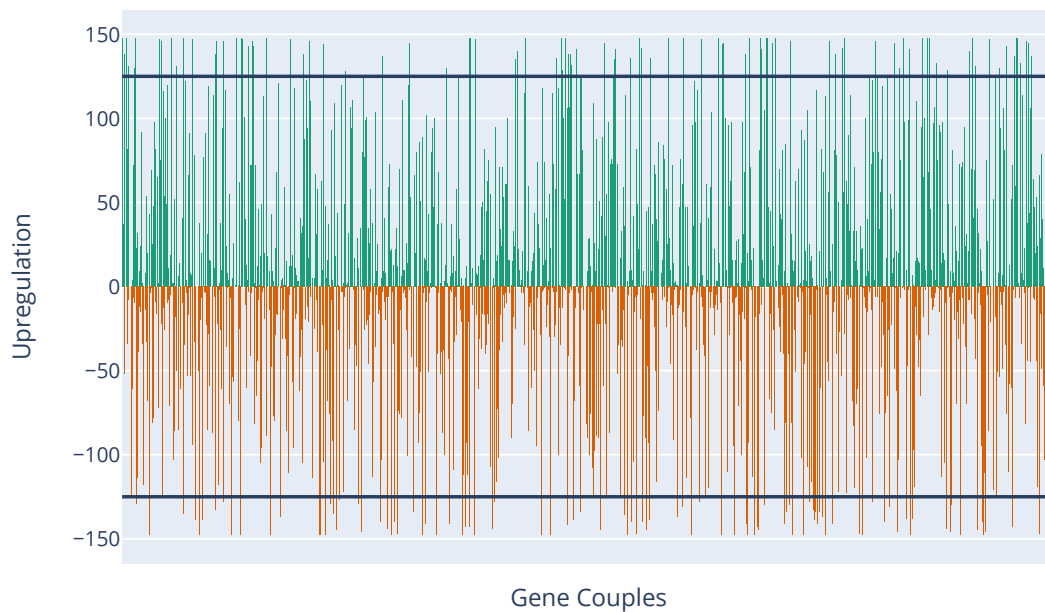
Traitement	Nombre	Pourcentage
Fruit developement	36	24,1611 %
Cold Storage	24	16,1074 %
Storage	18	12,0805 %
Pollination	14	9,3960 %
Light	7	4,6980 %
5-aminolevulinic acid	6	4,0268 %
Water	4	2,6846 %
Hand-pollinated	4	2,6846 %
Glucose	3	2,0134 %
GA3-treated	3	2,0134 %
Organ comparison	3	2,0134 %
Resistance to <i>Fusarium proliferatum</i>	3	2,0134 %
Resistance <i>Gymnosporangium yamadae</i>	2	1,3423 %
Seedling stage	2	1,3423 %
Rootstocks effect	2	1,3423 %
epibrassinolide	2	1,3423 %
N-acetylaspartate	2	1,3423 %
Resistance to <i>Colletotrichum fructicola</i>	2	1,3423 %
White-leaf mutants	1	0,6711 %
apple light green leaf mutants	1	0,6711 %
TRV_MdOPT3	1	0,6711 %
Normal leaf	1	0,6711 %
treatment of weakly flowering	1	0,6711 %
Scab-susceptible	1	0,6711 %
Gibberellin	1	0,6711 %
Piebald-leaf mutants	1	0,6711 %
thinned	1	0,6711 %
TRV_MdOPT3 – Control	1	0,6711 %
control of profusely flowering	1	0,6711 %
Scab-resistant	1	0,6711 %



**Figure A.23** — Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 3 et associés au fragment synténique 3-11. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associées ont une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 3 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 11. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 11 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 3. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



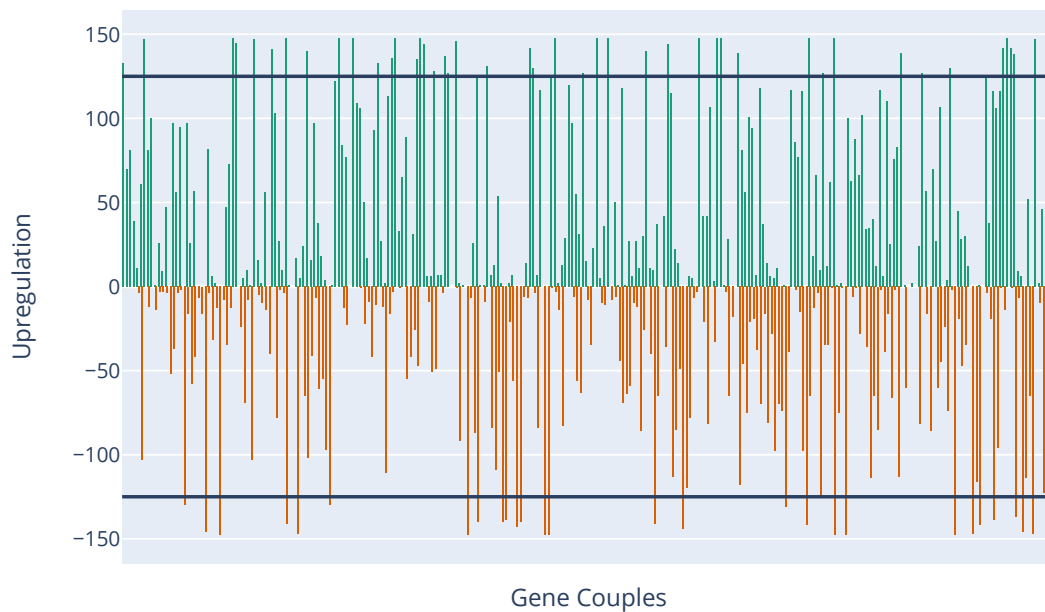
**Figure A.24** – Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 5 et associés au fragment synténique 5-10. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associés a une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 5 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 10. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 10 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 5. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



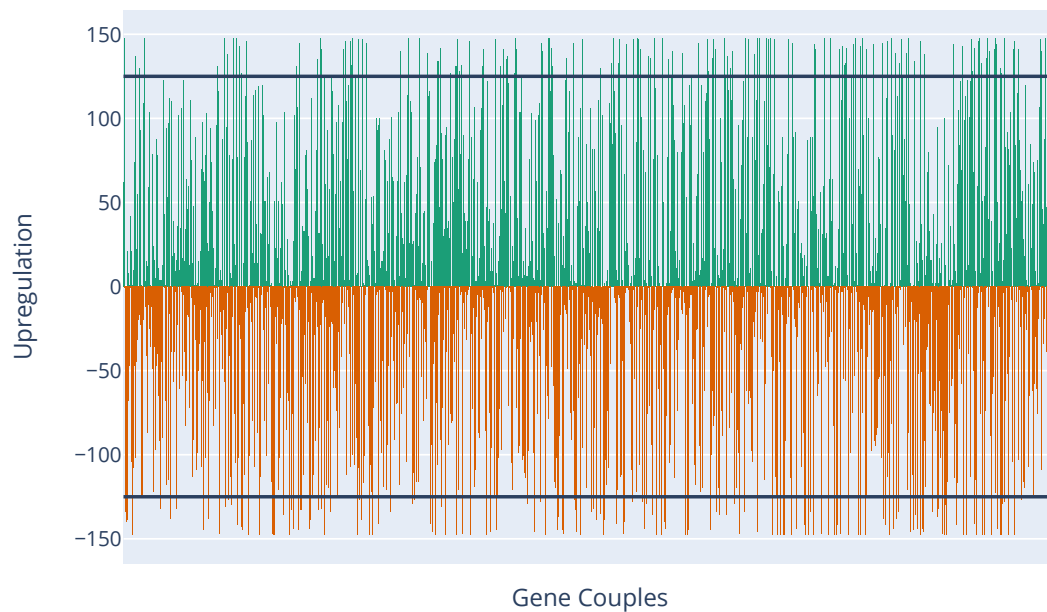
**Figure A.25** — Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 6 et associés au fragment synténique 6-14. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associées à une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 6 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 14. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 14 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 6. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



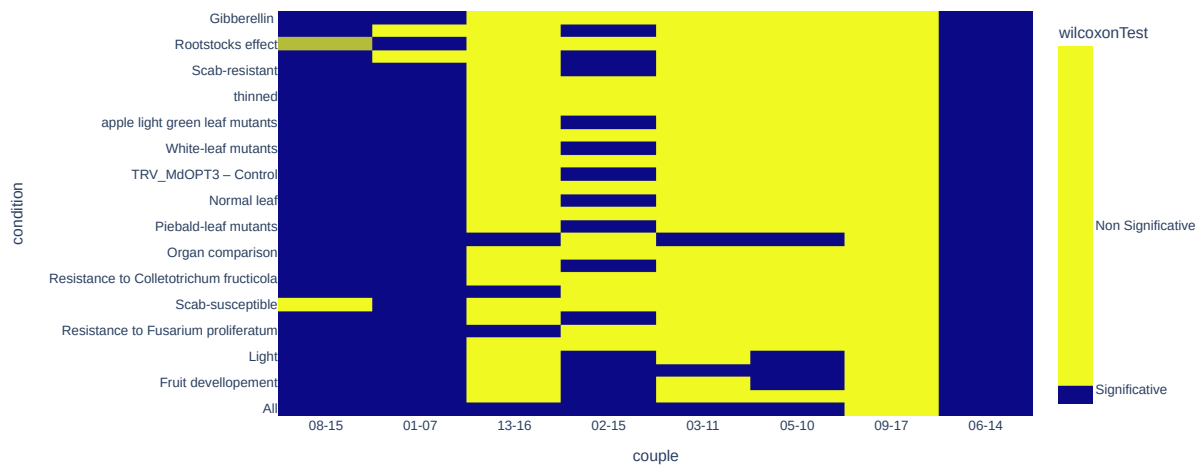
**Figure A.26** – Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 8 et associés au fragment synténique 8-15. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associés a une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 8 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 15. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 15 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 8. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



**Figure A.27** – Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 9 et associés au fragment synténique 9-17. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associés à une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 9 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 17. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 17 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 9. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



**Figure A.28** – Histogramme de la distribution du nombre de fois où un gène est sur exprimé par rapport à son ohnologue. L'axe des abscisses (*Gene couple*) présente l'ensemble des couples de gènes ordonnés le long du chromosome 13 et associés au fragment synténique 13-16. L'axe des ordonnées (*Upregulation*) présente le nombre de fois où ce couple a été identifié comme significativement différentiellement exprimé au sein d'une expérience. Les valeurs positives et associés a une barre verte qui présente le nombre de fois où c'est le gène associé au chromosome 13 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 16. À l'inverse si la valeur est négative et associée à une barre orange c'est le gène associé au chromosome 16 qui a été identifié comme significativement surexprimé par rapport à son ohnologue sur le chromosome 13. Les traits horizontaux noirs représentent le seuil de 125 expériences utilisées pour le définir en tant que gène non commutant.



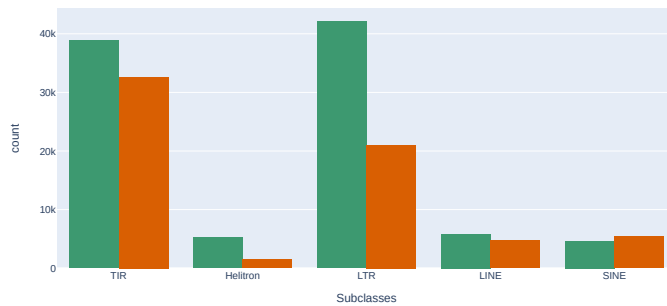
**Figure A.29** – Carte thermique des résultats des tests agrégés selon les conditions expérimentales. L'axe des ordonnées présente les différentes conditions expérimentales testées. L'axe des abscisses présente les différentes paires de fragments chromosomiques synténiques testées. La couleur de la cellule est associée à la significativité du test agrégé par la méthode de Fisher. La cellule est colorée en bleu si la p-value agrégée est significatif au seuil  $\alpha = 5\%$ .

## A.4 Analyse des éléments transposables des gènes ohnologues

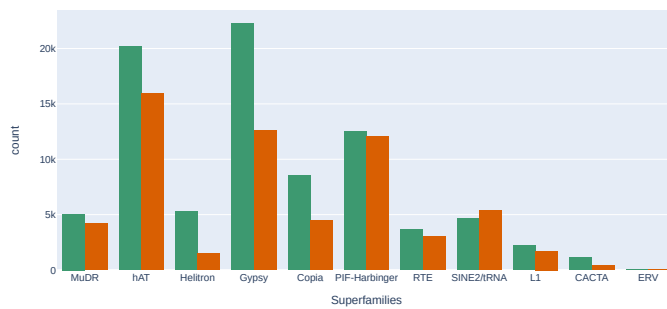




(A) Répartition des classes des TEs associés aux gènes ohnologues (en orange) ou aux gènes non dupliqués par WGD (en vert)



(B) Répartition des sous-classes des TEs associés aux gènes ohnologues (en orange) ou aux gènes non dupliqués par WGD (en vert)



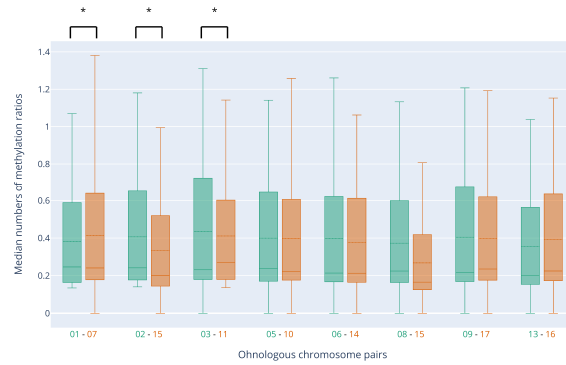
(C) Répartition des superfamilles des TEs associés aux gènes ohnologues (en orange) ou aux gènes non dupliqués par WGD (en vert)

**Figure A.30** — Répartition des classes, sous-classes, superfamilles des TEs associés aux gènes ohnologues (en orange) ou aux gènes non dupliqués par WGD (en vert)

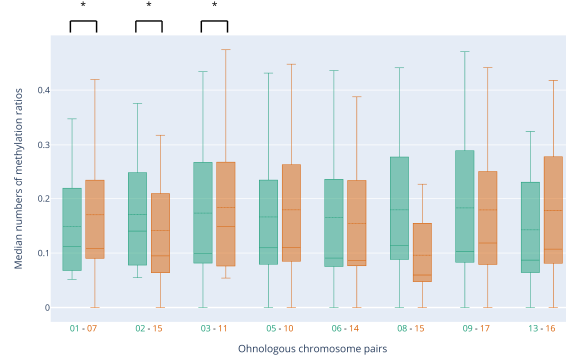
## A.5 Étude des méthylations de l'ADN des gènes ohnologues

**Table A.5** – Comparaison du nombre de cytosines considérées dans les régions exoniques des gènes ohnologues pour les paires de fragments chromosomiques synténiques selon les contextes.

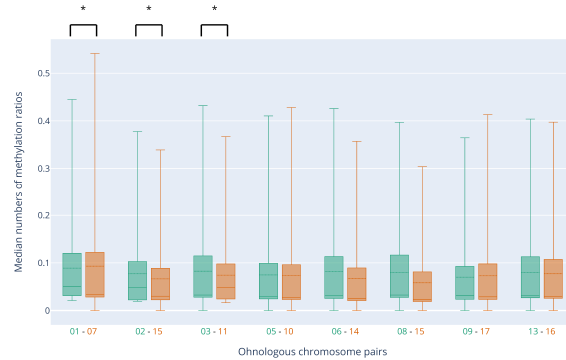
Couple	Nombre de cytosines sur le premier chromosome	Nombre de cytosines sur le second chromosome	Contexte	p-value test Binomial
01-07	436 155	329 786	CHH	$2,4703 \times 10^{-323}$
02-15	451 710	400 394	CHH	$3,9525 \times 10^{-323}$
03-11	559 023	490 656	CHH	$3,9525 \times 10^{-323}$
05-10	652 452	582 568	CHH	$3,9525 \times 10^{-323}$
06-14	380 386	273 662	CHH	$2,4703 \times 10^{-323}$
08-15	554 676	470 903	CHH	$3,9525 \times 10^{-323}$
09-17	498 775	486 059	CHH	$1,3920 \times 10^{-37}$
13-16	656 421	602 036	CHH	$3,9525 \times 10^{-323}$
01-07	122 932	92 819	CHG	$1,4822 \times 10^{-323}$
02-15	129 009	114 944	CHG	$1,9373 \times 10^{-178}$
03-11	153 717	133 273	CHG	$6,9375 \times 10^{-319}$
05-10	179 728	160 769	CHG	$1,1619 \times 10^{-231}$
06-14	106 951	76 890	CHG	$1,4822 \times 10^{-323}$
08-15	156 274	132 985	CHG	$1,4822 \times 10^{-323}$
09-17	139 106	135 280	CHG	$2,8297 \times 10^{-13}$
13-16	186 565	170 969	CHG	$5,3560 \times 10^{-150}$
01-07	94 528	71 872	CG	$2,4703 \times 10^{-323}$
02-15	103 207	90 509	CG	$3,9627 \times 10^{-183}$
03-11	115 062	98 598	CG	$4,1953 \times 10^{-278}$
05-10	132 792	120 038	CG	$5,6567 \times 10^{-142}$
06-14	82 835	60 234	CG	$<0,001$
08-15	125 359	103 692	CG	$1,4822 \times 10^{-323}$
09-17	101 199	99 262	CG	$1,5317 \times 10^{-5}$
13-16	144 638	131 811	CG	$1,7779 \times 10^{-131}$



(A) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées 100 bp en amont des gènes pour le contexte CG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

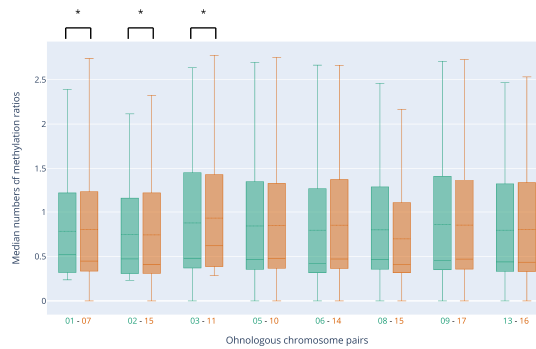


(B) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées 100 bp en amont des gènes pour le contexte CHG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

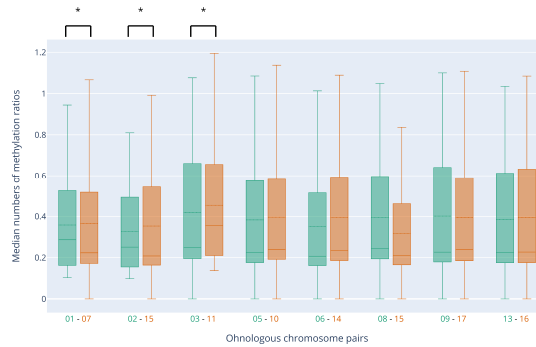


(C) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées 100 bp en amont des gènes pour le contexte CHH et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

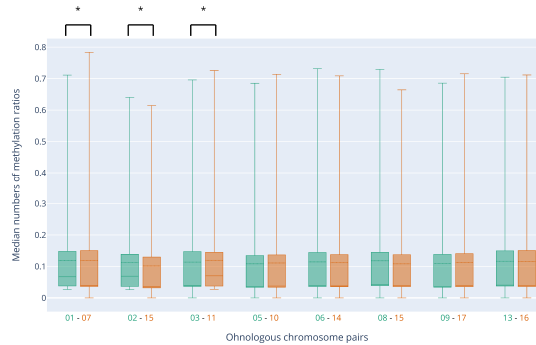
**Figure A.31** – Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées 100 bp en amont des gènes et groupées par paires de fragments chromosomiques synténiques. L'axe des abscisses présente les paires de fragments chromosomiques synténiques. L'axe des ordonnées présente les médianes des pourcentages des ratios de méthylation pour les différentes expériences analysées. Les paires 1-7, 2-15 et 3-11 présentent une différence significative pour l'ensemble des contextes.  $*\alpha < 5\%$



(A) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées en 2 kb en amont et en aval des gènes pour le contexte CG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

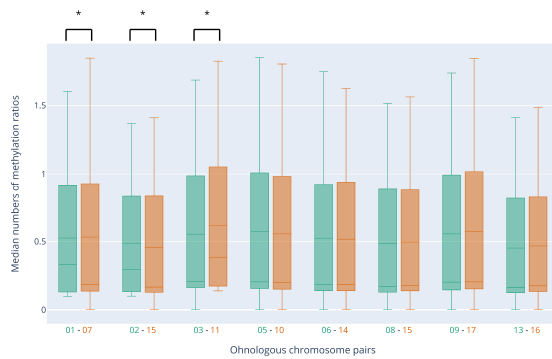


(B) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées en 2 kb en amont et en aval des gènes pour le contexte CHG et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

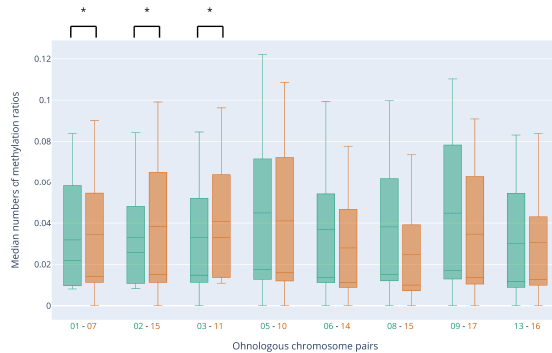


(C) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées en 2 kb en amont et en aval des gènes pour le contexte CHH et groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

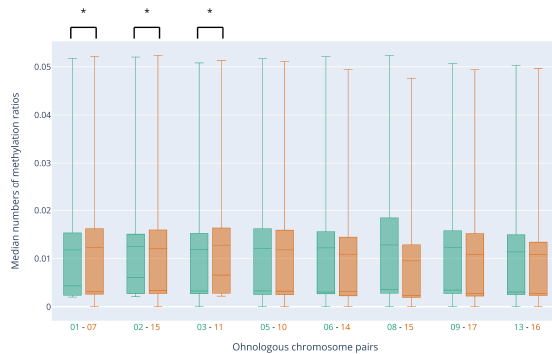
**Figure A.32** — Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées en 2 kb en amont et en aval et groupées par paires de fragments chromosomiques synténiques. L'axe des abscisses présente les paires de fragments chromosomiques synténiques. L'axe des ordonnées présente les médianes des pourcentages des ratios de méthylation pour les différentes expériences analysées. Les paires 1-7, 2-15 et 3-11 présentent une différence significative pour l'ensemble des contextes.  $\alpha < 5\%$



(A) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées sur les régions exoniques des gènes pour le contexte CG groupés par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.



(B) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées sur les régions exoniques des gènes pour le contexte CHG groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.



(C) Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées sur les régions exoniques des gènes pour le contexte CHH groupées par paires de fragments chromosomiques synténiques. Les paires 1-7, 2-15 et 3-11 présentent une différence significative.

**Figure A.33** – Boîtes à moustache de la distribution des rapports de méthylation des cytosines localisées sur les régions exoniques des gènes groupées par paires de fragments chromosomiques synténiques. L'axe des abscisses présente les paires de fragments chromosomiques synténiques. L'axe des ordonnées présente les médianes des pourcentages des ratios de méthylation pour les différentes expériences analysées. Les paires 1-7, 2-15 et 3-11 présentent une différence significative pour l'ensemble des contextes.  $*\alpha < 5\%$

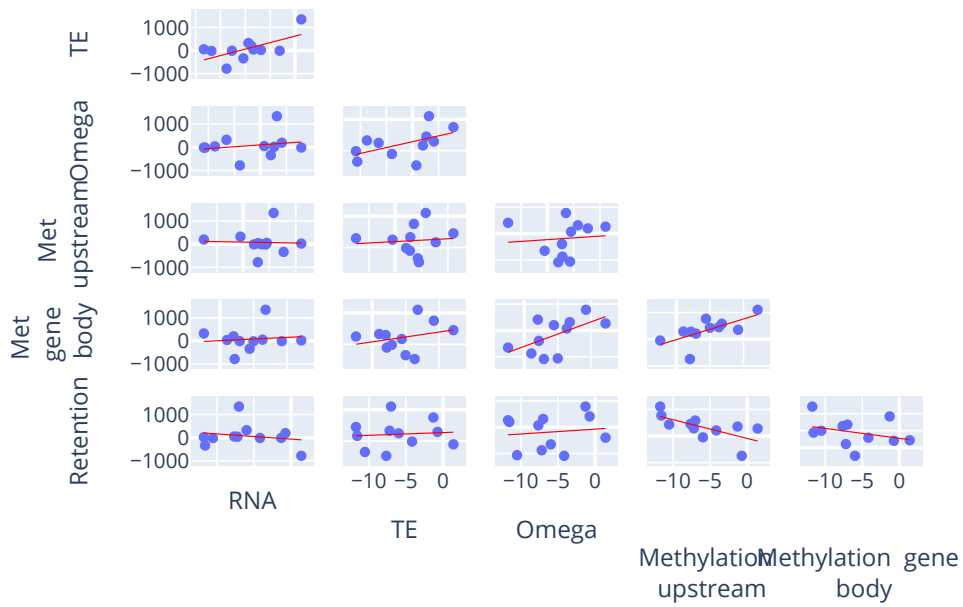
**Table A.6** – Comparaison du nombre de cytosines considérées 2kb en amont et en aval des gènes ohnologues pour les paires de fragments chromosomiques synténiques selon les contextes.

Couple	Nombre de cytosines sur le premier chromosome	Nombre de cytosines sur le second chromosome	Contexte	p-value test Binomial
01-07	670 899	508 097	CHH	$3,9525 \times 10^{-323}$
02-15	700 376	627 279		$2,4703 \times 10^{-323}$
03-11	874 532	767 030		$3,9525 \times 10^{-323}$
05-10	1 022 640	903 507		$4,9407 \times 10^{-323}$
06-14	587 980	420 963		$2,4703 \times 10^{-323}$
08-15	883 354	701 639		$3,9525 \times 10^{-323}$
09-17	770 524	744 945		$6,8479 \times 10^{-96}$
13-16	1 027 082	946 469		$4,9407 \times 10^{-323}$
01-07	142 843	108 623	CHG	$1,4822 \times 10^{-323}$
02-15	150 843	137 700		$3,0948 \times 10^{-132}$
03-11	176 690	153 407		$2,4703 \times 10^{-323}$
05-10	210 318	185 847		$2,4703 \times 10^{-323}$
06-14	124 855	88 990		$2,4703 \times 10^{-323}$
08-15	183 585	148 243		$1,4822 \times 10^{-323}$
09-17	160 520	154 313		$1,9441 \times 10^{-28}$
13-16	219 008	200 572		$3,2118 \times 10^{-178}$
01-07	133 870	102 580	CG	$1,4822 \times 10^{-323}$
02-15	141 183	126 943		$1,5287 \times 10^{-166}$
03-11	163 842	142 887		$1,7541 \times 10^{-313}$
05-10	190 983	171 267		$1,9051 \times 10^{-235}$
06-14	117 752	84 088		$2,4703 \times 10^{-323}$
08-15	177 432	142 186		$2,4703 \times 10^{-323}$
09-17	145 535	142 584		$3,8870 \times 10^{-8}$
13-16	205 488	189 433		$5,5156 \times 10^{-144}$

**Table A.7** – Comparaison du nombre de cytosines considérées 500 bp en amont des TSS pour les paires de fragments chromosomiques synténiques selon les contextes.

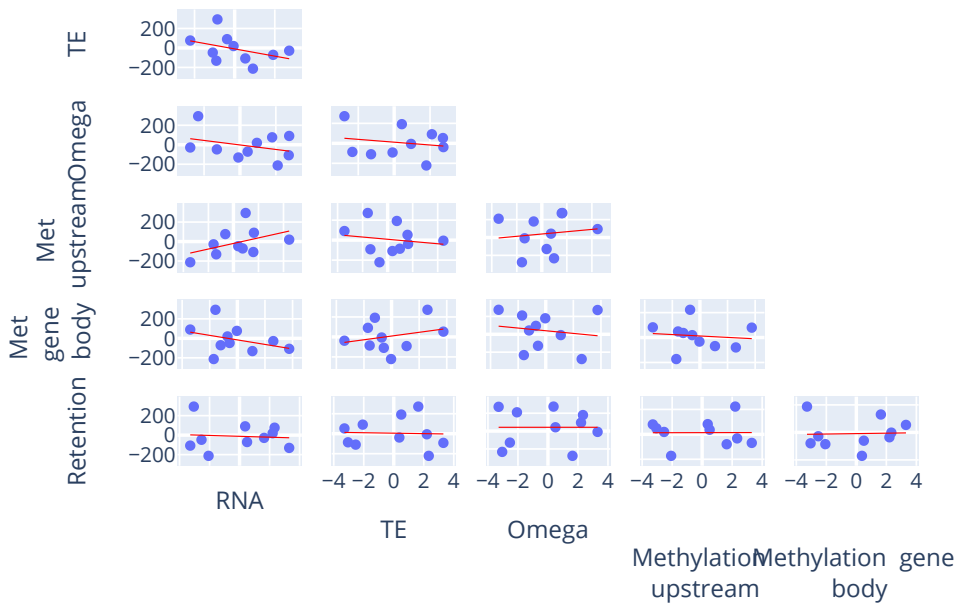
Couple	Nombre de cytosines sur le premier chromosome	Nombre de cytosines sur le second chromosome	Contexte	p-value test Binomial
01-07	107 556	81 022	CHH	$1,4822 \times 10^{-323}$
02-15	112 213	102 536	CHH	$7,5616 \times 10^{-97}$
03-11	141 525	123 623	CHH	$5,3266 \times 10^{-265}$
05-10	163 550	147 633	CHH	$3,9797 \times 10^{-179}$
06-14	94 062	68 621	CHH	$1,4822 \times 10^{-323}$
08-15	137 080	114 307	CHH	$1,4822 \times 10^{-323}$
09-17	124 768	121 978	CHH	$1,9688 \times 10^{-8}$
13-16	165 638	154 098	CHH	$1,3941 \times 10^{-92}$
01-07	18 448	14 405	CHG	$1,9569 \times 10^{-110}$
02-15	19 459	18 351	CHG	$1,2451 \times 10^{-8}$
03-11	23 616	20 719	CHG	$4,5125 \times 10^{-43}$
05-10	28 040	24 618	CHG	$2,7015 \times 10^{-50}$
06-14	16 208	11 530	CHG	$2,4476 \times 10^{-174}$
08-15	23 849	19 749	CHG	$6,3086 \times 10^{-86}$
09-17	21 503	20 600	CHG	$1,1022 \times 10^{-5}$
13-16	28 577	26 482	CHG	$4,4520 \times 10^{-19}$
01-07	20 377	15 753	CG	$5,1569 \times 10^{-131}$
02-15	21 061	19 606	CG	$5,5588 \times 10^{-13}$
03-11	24 855	21 594	CG	$9,9050 \times 10^{-52}$
05-10	29 023	26 160	CG	$3,6877 \times 10^{-34}$
06-14	18 190	12 662	CG	$1,7097 \times 10^{-218}$
08-15	26 215	21 657	CG	$1,7535 \times 10^{-96}$
09-17	22 585	21 760	CG	$9,1134 \times 10^{-5}$
13-16	30 795	28 597	CG	$1,9504 \times 10^{-19}$

## A.6 Intégration des résultats

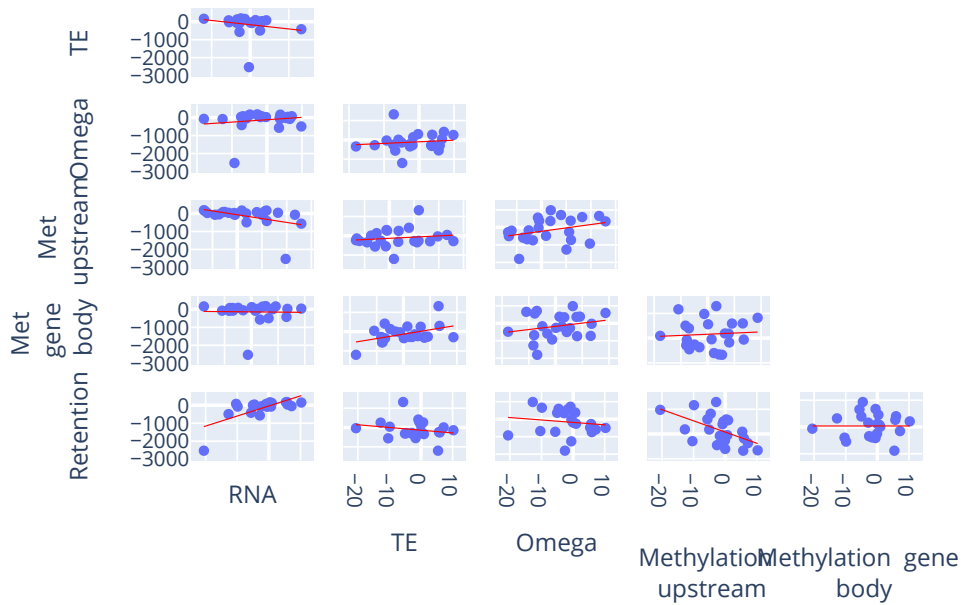


**Figure A.34** – Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 02-15. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.

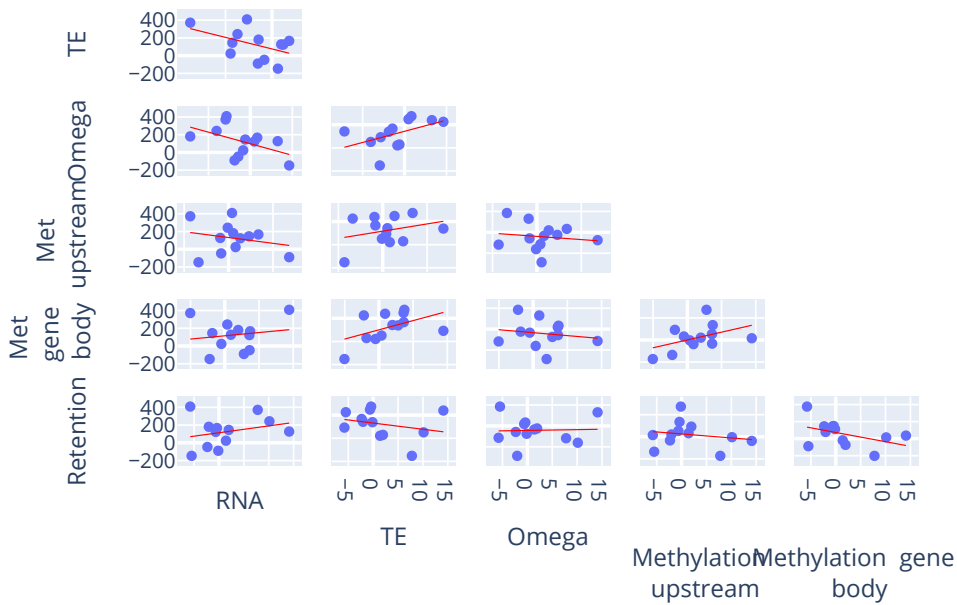




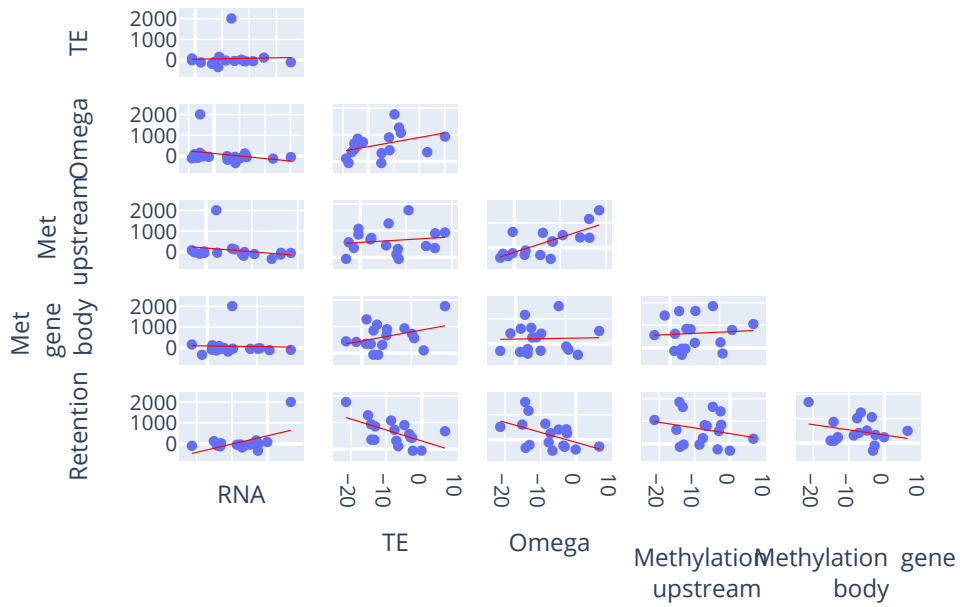
**Figure A.35** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 3-11. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.



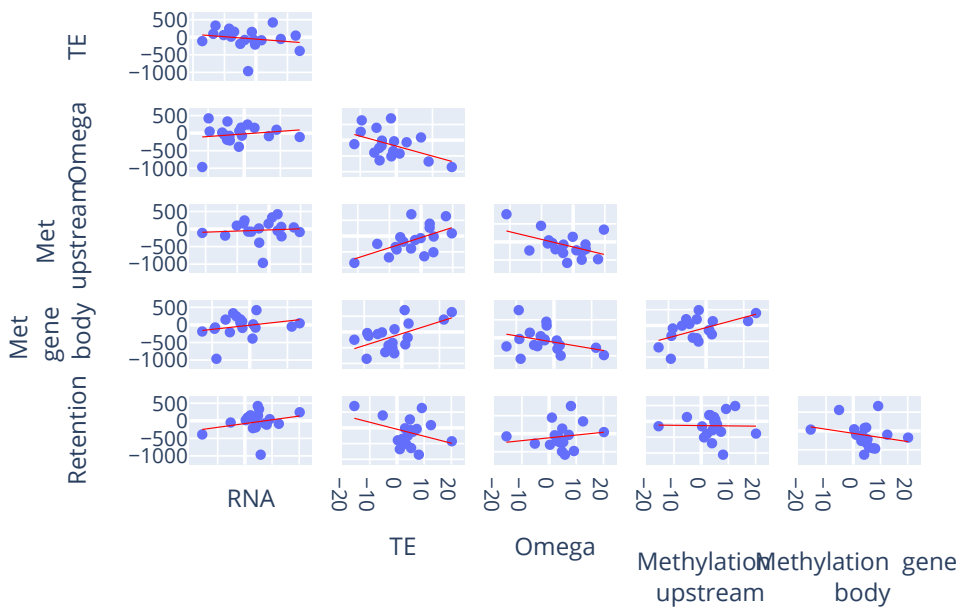
**Figure A.36** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 5-10. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.



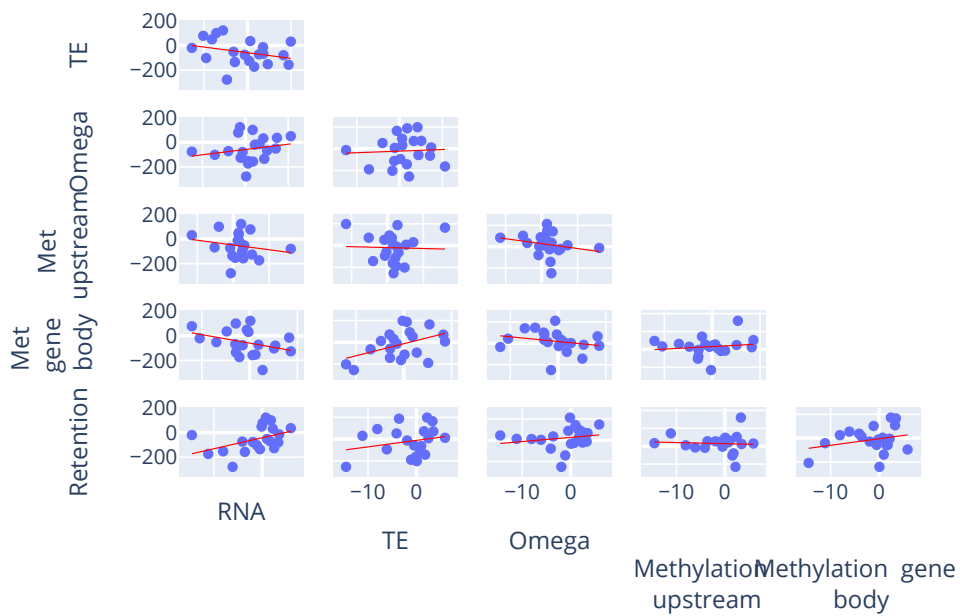
**Figure A.37** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 6-14. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.



**Figure A.38** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 08-15. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.



**Figure A.39** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 09-17. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.



**Figure A.40** — Ensemble des matrices de dispersion des différentes variables et des droites de régression associées pour les paires 13-16. L'ensemble des valeurs médianes associées à chacun des blocs de synténie sont représentées par un point. La droite de régression associée à la distribution est représentée par une droite rouge.

# APPENDICES ARTICLES

---

## B.1 Article accepté

Cet article accepté pour publication dans *Acta Horticulturae* a pour but d'accompagner la présentation orale faite à l'*International Horticultural Congress 2022* (IHC). Dans cet article de conférence, nous avons décrit l'étude sur la pression de sélection appliquée sur les couples de gènes ohnologues et l'analyse de l'expression des gènes ohnologues. Nous avons choisi de soumettre ces deux parties de la thèse car elles nous permettaient, en introduisant l'identification du déséquilibre de QTLs, de montrer un résultat fort, intégré au sein d'un ensemble d'analyse reliée entre elles. De plus, cela nous a permis d'ouvrir la discussion sur ces analyses avec un public expert de ces questions afin d'affiner nos analyses.

# Chromosomal dominance in apple after Whole Genome Duplication

Tanguy Lallemand<sup>1a</sup>, Sébastien Aubourg<sup>1</sup>, Jean-Marc Celton<sup>1</sup> and Claudine Landès<sup>1</sup>

<sup>1</sup> Université d'Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France.

## Abstract

Whole genome duplication (WGD) is a massive mechanism consisting in the duplication of the entire genome. By increasing the size of the genome and removing the selective constraint, WGD is an important driver of genetic innovation. The *Rosaceae* family has experienced at least 3 WGD in its evolutionary history. For apple (*Malus domestica*), which belongs to this family, an additional WGD occurred recently (50 mYA). This WGD is well conserved and not found outside the *Maloideae* subfamily. Apple is therefore an organism of choice to study the fate of duplicated genes after WGD. In a previous analysis, a QTL disequilibrium was observed between ohnologous chromosome pairs. The aim of this project is to identify the molecular mechanism driving this observed disequilibrium. Using a turnkey Snakemake pipeline, our first step was to compute the Ka/Ks ratio to estimate the selective pressure on ohnologous gene pairs. Secondly, we compared the expression levels between ohnologous genes using a large number of biological samples. Primary analyses indicate that while Ka/Ks ratio does not differ between ohnologous genes, significant transcriptional differences were identified at the chromosome level. This subgenome dominance could explain in part the observed disequilibrium in the number of QTL carried by the ohnologous chromosome pairs.

**Keywords:** Whole Genome Duplication, apple, bioinformatics, RNA-Seq, Ka/Ks, genome evolution

## INTRODUCTION

Different mechanisms leading to gene duplication have been described. The most massive mechanism is the Whole Genome Duplication (WGD). This mechanism results in two or more copies of the entire genome. WGD is relatively widespread in the eukaryotic lineages and in particular in plants (Adams and Wendel, 2005; Soltis et al., 2004; Van de Peer et al., 2021), and leads to an increase of genome size and number of genes. Since gene redundancy removes selective constraints, WGD is an important evolutionary driver (Soltis and Soltis, 2009). In the WGD context, paralogous genes are called ohnologous genes in reference to Ohno (Ohno, 1970). Ohnologous genes are found in preserved order in the duplicated segments, forming syntenic blocks that can be identified. Following WGD, different mechanisms come into play to allow a return to ploidy (Freeling et al., 2015; Vicient and Casacuberta, 2017), specifically chromosomal rearrangements and gene conversion, and can involve transposable elements (TE). Moreover, many cases of pseudogenization can eliminate one of the two copies of the gene for most of the duplicated gene pairs (Lockton and Gaut, 2005). This evolution can be biased toward one subgenome and referred to as fractionation bias (Woodhouse et al., 2010). Moreover, gene expression can also be biased and is defined as subgenome dominance. These biases have been found in different species, such as *Brassica rapa* (Wang et al., 2011; Cheng et al., 2012). From a mechanistic point of view, subgenome dominance can be observed by comparing the expression levels of ohnologous genes.

---

<sup>a</sup>E-mail: tanguy.lallemand@inrae.fr



Furthermore, it has been shown in different species that differences in TE density can be related to differential gene expression. In *B. rapa*, an enrichment of TE in the under-expressed subgenome was observed (Woodhouse et al., 2014). This mechanism is not found in all species and does not seem to explain the mechanism of gene fractionation or subgenome dominance. Similarly, differences in DNA methylation, mainly in regulatory regions and in the three cytosine contexts (CG, CHH, CHG, with H standing for A, T or C have been linked to subgenome dominance in some plant species (Zhao et al., 2017).

Other gene duplication mechanisms acting at different scales have been described. Some of them lead to the duplication of a single gene. We can mention tandem duplication, which results from an unequal crossing over (Zhang, 2003). There is also duplication by replicative transposition, mediated by TE (Juretic et al., 2005), and retroduplication obtained by reverse transcription of mRNAs (Brosius, 1991). Other mechanisms result in the duplication of several genes. Besides the WGD, we can find segmental duplication, leading to the duplication of small chromosome segments mediated by LINE sequences (Samonte and Eichler, 2002).

Duplicated genes can have different fates. A large majority of them will be pseudogenized and lost (Prince and Pickett, 2002) while others will be conserved without having developed new functions designed as functional redundancy (Cui et al., 2006). Other genes will be conserved due to sub-functionalization (Otto and Yong, 2002). Finally, some genes will be neo-functionalized and acquire a new function (Hahn, 2009).

Quantitative Trait Loci (QTL) are loci related to the expression of phenotypic traits. In a previous analysis (Lallemand et al., 2020) we located 541 QTL derived from the GDR database (Jung et al., 2019) and associated with a diversity of phenotypic traits, on a physical map of the apple genome. We observed that their proportion was statistically unbalanced within the pair of ohnologous chromosomes. In this project, we tested whether differences in ohnologous coding sequences or ohnologous gene transcription could be found and associated with the observed QTL disequilibrium.

## **MATERIALS AND METHODS**

All scripts are written in Python (Oliphant, 2007) or R (Team, 2013). Statistical analyses use Scipy (Virtanen et al., 2020). Workflows are handled with Snakemake (Köster and Rahmann, 2012). Visualizations are produced with Plotly (Plotly Technologies Inc., 2015).

### **Comparative evolution of ohnologous gene coding sequences: computation of Ka/Ks**

To test for the biased coding sequence evolution hypothesis, we used the Ka/Ks ratio which estimates the selective pressure applied on gene coding sequences (CDS). To test for differences in the selective pressure applied to ohnologous genes, we calculated the Ka/Ks between apple ohnologous genes against their closest orthologs in peach (*Prunus persica*), a species that presents no additional WGD and a high quality genome sequence and annotation (The International Peach Genome Initiative et al., 2013). Calculation of Ka/Ks was performed using an *ad hoc* turnkey Snakemake pipeline. The first step of the pipeline consisted in an alignment of the protein sequences using MUSCLE (Edgar, 2004). Then, PAL2NAL (Suyama et al., 2006) was used to convert the obtained protein alignments to a nucleic alignment. Nucleic alignment was then used by the YN00 program from the PAML package (Yang and Nielsen, 2000) to compute Ka/Ks values of ohnologous genes. This program computes the ratio of synonymous (Ks) and non-synonymous (Ka) mutation rates, using an evolutionary model, and computes the selective pressure applied on the genes. We used the Yang and Nielsen model (Yang and Nielsen, 2000) since it considers the difference in transition-transversion rates and nucleotide observed frequencies.

The difference between Ka/Ks values between gene pairs at the chromosome level was

tested using the Wilcoxon signed-rank test from Scipy (Virtanen et al., 2020). This nonparametric paired test does not make assumptions regarding the normality of the distribution of differences, calculates the difference between each set of matched pairs.

### **Transcriptomic disequilibrium**

To test the subgenome dominance hypothesis, we compared the expression levels of the ohnologous genes. For this analysis, we retrieved a set of publicly available RNA-Seq experiments for *Malus domestica* (9578 biological samples). We applied quality filters and downloaded the set of experiments passing our selection criteria : paired data, at least three biological replicates, a total size of at least  $2 \times 10^7$  reads per sample and a minimum read size of 100 nt. This selection and download were performed via an *ad hoc* turnkey pipeline written in Snakemake. It is based on a download using the SRA-toolkit package retrieving the data from the NCBI server or, if available on ENA, a high-speed download via Aspera when possible.

The processing of the raw transcriptomic data was performed using an *ad hoc* pseudo-mapping pipeline. This pipeline, written in Snakemake produces the trimming and quality checks steps. Quantification of the transcript expression was performed using Salmon (Patro et al., 2017). All experiments with global mapping rates below 70% were not used. The number of reads for each pair of ohnologous genes was then normalized according to the length of the longest gene of the pair. To finish, a differential analysis between ohnologous genes within the same experimental conditions was conducted applying a set of *ad hoc* scripts based on DESeq2 (Love et al., 2014).

To test for difference between transcript expression levels within pairs of ohnologous chromosomes, a statistical analysis based on the Wilcoxon test was implemented for each experiment. The outcomes of the tests were aggregated for each of the ohnologous chromosome pairs using the Fisher method. This analysis provided a final result for each pair of chromosomes, allowing the determination of the presence or absence of a transcriptional disequilibrium using all the available RNA-Seq experiments.

## **RESULTS AND DISCUSSION**

In order to explain the observed QTL disequilibrium between pairs of ohnologous chromosomes (or chromosome segments), we first explored a potential differential evolution of coding sequences between syntenic blocks. To test this assumption, we calculated the Ka/Ks of the ohnologous apple genes against the peach ortholog. A second hypothesis is the evolution of regulatory sequences that would have an impact on gene transcription. This was tested by investigating for a potential disequilibrium in ohnologous gene transcription between syntenic blocks.

### **Estimating Ka/Ks of ohnologous genes**

This analysis aims to identify differences in the evolution of the coding sequences between apple ohnologous genes, using the peach orthologs as a reference. Our results indicate that the distribution of the percentage of identity of the protein alignments obtained from the ohnologous gene pairs and a peach reference varies between 80% and 100%. This distribution indicates that sequences of constructed orthogroups are relatively similar. The size distribution of the alignments has a median of 942 bp, suggesting good size alignments and similar proteins. This allowed for an optimal Ka/Ks calculation and resulted in Ka/Ks values computed for 9,821 gene pairs.

Ks distribution has one main peak with a value of 0.28, confirming a single WGD. Additional smaller peaks representing traces of more ancient WGD can be found. From the distribution of Ka/Ks ratios (Figure 1), we observe that most Ka/Ks values are below 1, implying a negative selection in all chromosomes. The medians of Ka/Ks distributions (around 0.20) are similar to the literature and in particular to plants having had a recent

WGD (Li et al., 2015; Wang et al., 2020; Yang et al., 2020; Yu et al., 2017b).

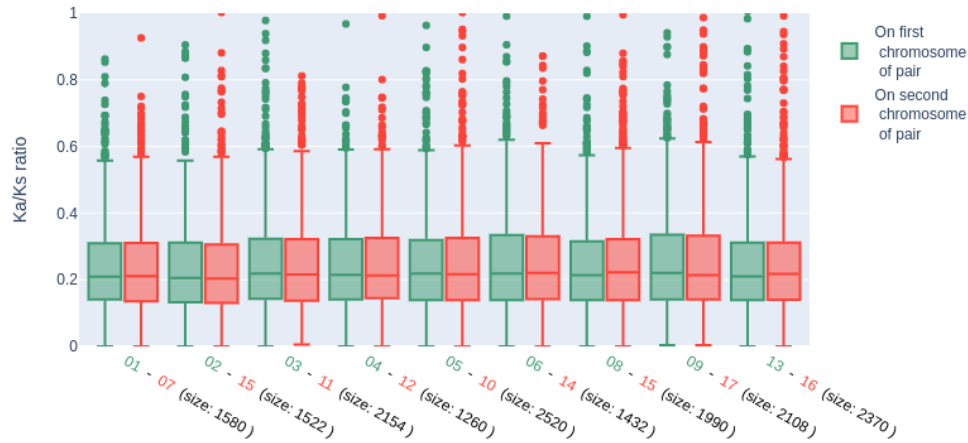


Figure 1. Box plot of the distribution of Ka/Ks values for apple chromosomes of each of the ohnologous pairs. On the x-axis are presented the pairs of ohnologous chromosomes considered. The number of gene pairs is indicated in brackets. For each pair, the box plot representing the distribution of Ka/Ks on the first chromosome of the pair (in green) and the distribution of Ka/Ks on the second chromosome of the pair (in red) is shown.

In order to identify a disequilibrium in the estimated selection pressure on the pairs of ohnologous sequences, we compared their Ka/Ks values computed against a peach reference. The outcomes of the Wilcoxon test are shown in Table 1. Our results show that no pairs are significantly unbalanced regardless of the degree of chromosomal rearrangement, indicating that the evolution rate of ohnologous coding sequences are similar. Separate investigations of Ks and Ka suggest that the coding sequences of the ohnologous genes did not evolve differently and cannot explain QTL disequilibrium between syntenic regions. The genes that were maintained after the first round of post-WGD pseudogenization thus appear to be under comparable selective constraints. Because of the approach used for ohnologous genes identification and the method that selected for analyzing them against a related genome, we investigated genes that are related and therefore quite conserved during evolution. Pseudogenized genes were not included in this analysis (because they are not or only partially detected by gene modelers) and may be part of the source of the divergences. Thus, the QTL disequilibrium cannot be explained by differences in the evolution of coding sequences of genes with an identified ohnolog. Similar results have been found in different plant species such as willows (Harikrishnan et al., 2015) or sesame (Yu et al., 2017a).

Table 1. Results of the Wilcoxon two-tailed test. Results are sorted by p-value in ascending order.

couple	08-15	13-16	02-15	06-14	09-17	05-10	03-11	01-07
P-value	0.2009	0.3137	0.3139	0.3808	0.6726	0.7750	0.8939	0.9752

#### Differential ohnologous gene expression

Genome dominance may be explained by differences in expression levels between

subgenomes. To test this hypothesis in light of the observed QTL disequilibrium, a differential analysis was conducted using DESeq2 on 444 RNA-Seq runs associated with a large diversity of tissues and stresses grouped in 148 experiments of 3 biological replicates. From a total of 16,779 ohnologous gene pairs, we identified on average 7,000 differentially expressed ohnologous genes per experiment. Despite the proximity of the coding sequences of the ohnologous genes observed previously by the Ka/Ks analysis, the number of ambiguous reads during the mapping remained very low, since only 600 genes had a median value superior to one ambiguous read over all experiments. Moreover, the percentage of mapping varies between 72% and 91% with a median of 85%.

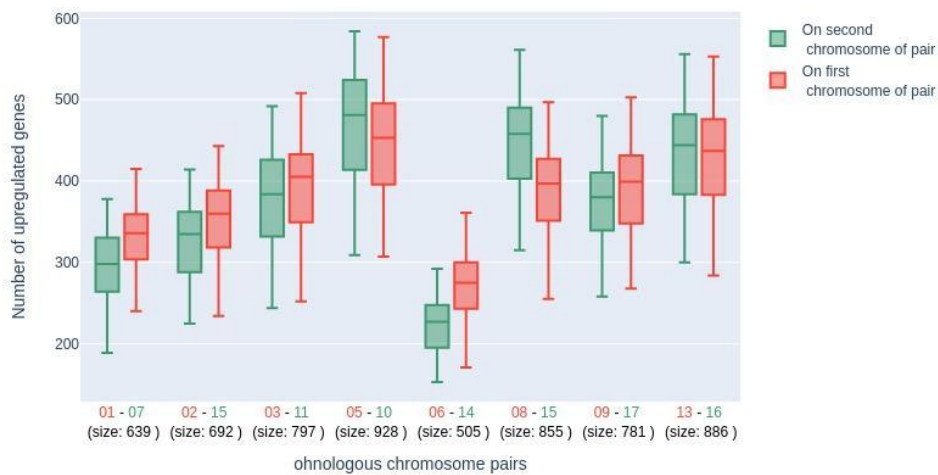


Figure 2. Box plot of the distribution of the number of up-regulated genes for a gene relative to its ohnologue for both of the chromosomes in each of the ohnologue pairs. The y-axis represents the set of up-regulated gene count values. The x-axis indicates the pairs of chromosomes considered. For each pair of chromosomes (or chromosome segments) the box plots indicate the distribution on the first chromosome of the pair (in green) and the distribution on the second chromosome of the pair (in red).

To test for disequilibrium between ohnologous gene expression levels, we counted the number of upregulated genes for a chromosome (or chromosome segment) against its counterpart in each of the considered experiments. The distribution of the number of up-regulated genes is shown in Figure 2. Several chromosomes or chromosome segments seem to be transcriptionally unbalanced with a number of up-regulated genes on one chromosome compared to the other which is, for some pairs, very different. In order to statistically validate this observation, we tested the number of average reads of all differentially expressed genes of each chromosome pair with a Wilcoxon test. All results were aggregated using Fisher's method to obtain a unique result for each pair. The results are presented in Table 2. From those results, we can observe a disequilibrium in the expression level of different pairs of ohnologous chromosomes. In most cases, transcriptional disequilibrium and QTL disequilibrium are oriented toward the same direction. This observed subgenome dominance has already been detected in different species duplicated by WGD and in particular *B. rapa* (Cheng et al., 2012) or sesame (Harikrishnan et al., 2015).

Table 2. Results of aggregated Wilcoxon two-tailed test.

Couple	08-15	01-07	06-14	02-15	13-16	05-10	03-11	09-17
Aggregated P-value	0	0	0	0	0	0	0.001	0.956

## CONCLUSION

In a previous study we observed a disequilibrium in the proportions of QTL carried by ohnologous chromosomes. In order to explain this, we investigated the evolution of coding sequences and expression level of ohnologous gene pairs. Of the 45,116 genes of the apple genome, 33,558 were identified with ohnologs, indicating that this genome has already undergone significant evolution. The Ka/Ks analysis indicated that (i) ohnologous genes are under the influence of strong negative selection (synonymous mutations are more numerous than non-synonymous ones, resulting in the conservation of both coding sequences) and (ii) the coding sequences of the ohnologous genes have not evolved significantly differently. Analysis of transcriptomic data indicated a global disequilibrium in the expression level of ohnologous genes among ohnologous chromosome pairs. Data integration indicated that upregulated genes often colocalize with genomic regions rich in QTL. Thus, it appears that post-WGD evolution affects mostly regulatory sequences rather than coding sequences. As a next step, it would be interesting to extend these analyses to other mechanisms known to play a role in rediploidization such as TE or cytosine methylation densities and to determine if similar disequilibrium trends can also be identified. Integration of these various datasets could in the future allow us to better understand the observed QTL disequilibrium, the evolution of duplicated genes and their genomic environment following a WGD event.

## ACKNOWLEDGEMENTS

The authors want to thank all their colleagues for their assistance in the research for this paper and in particular Sandra Pelletier for her *ad hoc* script used for differential expression analysis, Charles-Eric Durel for his advice on QTL analysis, Eric Montaudon for system administration and Carène Rizzon from Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) for helpful discussion on Ka/Ks.

## Literature cited

- Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* *8*, 135–141. <https://doi.org/10.1016/j.pbi.2005.01.001>.
- Brosius, J. (1991). Retroposons--seeds of evolution. *Science* *251*, 753. <https://doi.org/10.1126/science.1990437>.
- Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE* *7*, e36442. <https://doi.org/10.1371/journal.pone.0036442>.
- Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res* *16*, 738–749. <https://doi.org/10.1101/gr.4825606>.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* *32*, 1792–1797. <https://doi.org/10.1093/nar/gkh340>.
- Freeling, M., Scanlon, M.J., and Fowler, J.E. (2015). Fractionation and subfunctionalization following genome duplications mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* *35*, 110–118. <https://doi.org/10.1016/j.gde.2015.11.002>.
- Hahn, M.W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* *100*, 605–617. <https://doi.org/10.1093/jhered/esp047>.
- Harikrishnan, S.L., Pucholt, P., and Berlin, S. (2015). Sequence and gene expression evolution of paralogous genes in willows. *Sci. Rep.* *5*, 18662. <https://doi.org/10.1038/srep18662>.

Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. <https://plot.ly>.

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S.P., Gasic, K., Scott, K., et al. (2019). 15 years of GDR new data and functionality in the genome database for rosaceae. *Nucleic Acids Res* 47, D1137–D1145. <https://doi.org/10.1093/nar/gky1000>.

Juretic, N., Hoen, D.R., Huynh, M.L., Harrison, P.M., and Bureau, T.E. (2005). The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* 15, 1292–1297. <https://doi.org/10.1101/gr.4064205>.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. <https://doi.org/10.1093/bioinformatics/bts480>.

Lallemant, T., Aubourg, S., Hunault, G., Celton, J.-M., and Landès, C. (2020). Evolution of ohnologous chromosomes following Whole Genome Duplication in apple. In *Journées Ouvertes de Biologie, Informatique et Mathématique*, Montpellier

Lockton, S., and Gaut, B.S. (2005). Plant conserved non-coding sequences and paralogous evolution. *Trends Genet. TIG* 21, 60–65. <https://doi.org/10.1016/j.tig.2004.11.013>.

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. <https://doi.org/10.1186/s13059-014-0550-8>.

Ohno, S. (1970). *Evolution by Gene Duplication* (Springer Science & Business Media).

Oliphant, T.E. (2007). Python for scientific computing. *Comput. Sci. Eng.* 9, 10–20. <https://doi.org/10.1109/MCSE.2007.58>.

Otto, S.P., and Yong, P. (2002). 16 - The evolution of gene duplicates. In *Advances in Genetics*, J.C. Dunlap, and C.-ting Wu, eds. (Academic Press), pp. 451–483.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. <https://doi.org/10.1038/nmeth.4197>.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* 3, 827–837. <https://doi.org/10.1038/nrg928>.

Samonte, R.V., and Eichler, E.E. (2002). Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* 3, 65–72. <https://doi.org/10.1038/nrg705>.

Soltis, D.E., Soltis, P.S., and Tate, J.A. (2004). Advances in the study of polyploidy since plant speciation. *New Phytol.* 161, 173–191. <https://doi.org/10.1046/j.1469-8137.2003.00948.x>.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. <https://doi.org/10.1093/nar/gkl315>.

Team, R.C. (2013). R: A language and environment for statistical computing.

The International Peach Genome Initiative., Verde, I., Abbott, A. et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet* 45, 487–494 (2013). <https://doi.org/10.1038/ng.2586>

Van de Peer, Y., Ashman, T.-L., Soltis, P.S., and Soltis, D.E. (2021). Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* 33, 11–26. <https://doi.org/10.1093/plcell/koaa015>.

Vicent, C.M., and Casacuberta, J.M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann Bot* 120, 195–207. <https://doi.org/10.1093/aob/mcx078>.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>.

Wang, Y., Nie, F., Shahid, M.Q., and Baloch, F.S. (2020). Molecular footprints of selection effects and whole genome duplication (WGD) events in three blueberry species: detected by transcriptome dataset. *BMC Plant Biol.* 20, 250. <https://doi.org/10.1186/s12870-020-02461-w>.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 8, e1000409. <https://doi.org/10.1371/journal.pbio.1000409>.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5283–5288.

<https://doi.org/10.1073/pnas.1402475111>.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* *17*, 32–43. <https://doi.org/10.1093/oxfordjournals.molbev.a026236>.

Yang, F.-S., Nie, S., Liu, H., Shi, T.-L., Tian, X.-C., Zhou, S.-S., Bao, Y.-T., Jia, K.-H., Guo, J.-F., Zhao, W., et al. (2020). Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* *11*, 5269. <https://doi.org/10.1038/s41467-020-18771-4>.

Yu, J., Wang, L., Guo, H., Liao, B., King, G., and Zhang, X. (2017a). Genome evolutionary dynamics followed by diversifying selection explains the complexity of the *Sesamum indicum* genome. *BMC Genomics* *18*, 257. <https://doi.org/10.1186/s12864-017-3599-4>.

Yu, Y., Xiang, Q., Manos, P.S., Soltis, D.E., Soltis, P.S., Song, B.-H., Cheng, S., Liu, X., and Wong, G. (2017b). Whole-genome duplication and molecular evolution in *Cornus* L. (Cornaceae) – Insights from transcriptome sequences. *PLoS ONE* *12*, e0171361. <https://doi.org/10.1371/journal.pone.0171361>.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* *18*, 292–298. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8).

Zhao, M., Zhang, B., Lisch, D., and Ma, J. (2017). Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* *29*, 2974–2994. <https://doi.org/10.1105/tpc.17.00595>.

## B.2 Article soumis

Cet article a été soumis. Dans cet article nous avons présenté l'intégralité des résultats marquants obtenus pendant la thèse. Cet article fait le résumé de l'ensemble des analyses. Nous avons voulu orienter cet article selon la question de la dominance sous-génomique. En effet, ce type de mécanisme n'a jamais été identifié chez un organisme autoploïde. Ainsi nous avons décrit les analyses montrant les déséquilibres de participation à la variation phénotypique, le biais de fractionnement, l'absence de déséquilibre pour la pression de sélection et les déséquilibres transcriptionnels, de TE et de méthylation. Après intégration des résultats nous cherché à construire un modèle expliquant les processus hypothétiques conduisant le pommier, un organisme autoploïde à présenter une sous-dominance chromosomique.



# Identification and Characterization of Subgenome dominance in the Apple Genome

Tanguy Lallemand<sup>1,\*</sup>, Sébastien Aubourg<sup>1</sup>, Martin Leduc<sup>1</sup>, Carène Rizzon<sup>2</sup>, Jean-Marc Celton<sup>1</sup> and Claudine Landès<sup>1\*</sup>

<sup>1</sup> Université d'Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France. <sup>2</sup> LaMME

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

## ABSTRACT

**A Whole Genome Duplication (WGD) event occurred in apple (*Malus domestica*) about 40 MYA ago. This well conserved and complete genome duplication makes apple an organism of choice to study the early evolutionary phases of DNA sequences from duplicated chromosomal fragments. WGD events have a profound impact on genomes and are known to be major drivers of evolution. In this study, we investigated the evolution of duplicated chromosomal fragments, with a particular focus on gene sequences and expression, Transposable Elements (TE) density, and DNA methylation level. Overall, we identified 16,779 gene pairs in the apple genome, confirming the relatively recent WGD. We identified several imbalances in QTL localization among duplicated chromosomal fragments, and characterized various biases in genome fractionation, gene transcription, TE densities, and DNA methylation. Our results suggest a subgenome dominance in this autopolyploid, a phenomenon that has been only described so far in allopolyploid.**

## INTRODUCTION

Gene duplication is an important driver of genetic innovation (41, 57). Multiple mechanisms lead to gene duplication at different scales. On the lowest scale, the mechanisms associated with tandem duplication lead to local duplication of a gene or a part of it, and are mediated by unequal crossing over (76), duplication by replicative transposition (30) with Transposable Elements (TE) as key role players, and retroduplications obtained by reverse transcription of an mRNA followed by insertion inside the genome (6). On a larger scale, the segmental duplication mediated by LINE sequences (51) leads to the duplication of several genes. Finally, the most massive duplication event is described as the Whole Genome Duplication (WGD) and results in the duplication of the entire genome (2). In the context of WGD, paralogous genes can be referred to as ohnologous

genes in reference to Ohno (41). Ohnologous genes in conserved order between two paralogous genomic regions form syntenic blocks (34). Once duplicated, genes can follow different evolutionary routes. Most genes will be eliminated through pseudogenization (46) while others will keep their functions, adding functional redundancy (12). Genes with multiple functions will sub-functionalize (42). Finally, some duplicated genes will acquire a new function through neofunctionalization (21).

WGD can originate from crosses between individuals of the same species (autopolyploidy) or from interspecific hybridization (allopolyploidy) (67). WGD is known to profoundly affect the genome structure by doubling, at least transiently, the number of genes and the size of the genome. Furthermore, gene redundancy removes the selective constraint and provides more genetic resources, enabling faster genetic innovation. WGD results in reproductive isolation and may be an important speciation mechanism (1, 53). Following a WGD, different mechanisms called diploidization restore gradually ploidy status. TE are suggested as main actors of diploidization (18, 64). In fact, WGD is followed by an increase in TE content, known as a transposition burst (64). This burst leads to numerous transposition events and recombination that could be at the origin of gene losses, genome restructuring and neofunctionalization, as observed in maize (7). The progressive loss of one of the duplicated genes by pseudogenization is also an important mechanism (35). Gene loss can be biased in favour of some chromosomes and is called fractionation bias (19, 72). While the exact mechanism has not been elucidated yet, TE insertion bias between subgenomes is suspected to play a role. This bias, known as subgenome dominance, can also be found at the level of gene expression and DNA methylation, and participates in the overall phenotypic variation (19, 20, 45, 50). Subgenome dominance has been observed in different species, including *Arabidopsis thaliana* (19), *Brassica rapa* (66) and *Zea mays* (52). Subgenome dominance appears to be associated with ancient polyploids, suggesting that the mechanisms necessary to establish these subgenomes operate over many generations to promote such evolution (71).

\*To whom correspondence should be addressed. Tel: +332 41 22 56 99; Email: tanguy.lallemand@inrae.fr

© YYYY The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Association between biased gene loss and expression bias (50) has been identified in several species. It appears that the more fractionated a subgenome tends to be, the less expressed are the genes within this subgenome, which results to a decreasing phenotypic variation (52, 64). The origin of this biased expression appears to be related to strict selective limitations, increased methylation rates, and greater TE coverage (20), leading to higher fractionation. Thus, it has been proposed that underfractionated and overexpressed subgenomes are correlated with lower TE density (71). In fact, in *Brassica rapa* and *Arabidopsis thaliana* (61), the authors reported that small interference RNA (siRNA) targets the upstream region of genes preferentially located in the down-regulated subgenome (71). siRNAs inserting upstream of the gene will interfere with gene expression by recruiting the RISC complex (RNA-induced silencing complex) it will allow a rapid degradation of the genes with a complementary sequence. To explain these observations, it has been proposed that TEs are important targets of silencing mechanisms. High methylation level of TEs can be associated with heterochromatic epigenetics marks, therefore affecting expression levels of nearby genes (20, 68). Moreover, unbalances in DNA methylation in duplicated genes environments in the three cytosine contexts (CG, CHH, CHG) have been associated to subgenome dominance in some plant species (64, 77).

TEs are the main target of silencing mechanisms which keep their activity under a threshold to avoid compromising genome viability. As a consequence, TEs are usually heavily methylated and are associated with heterochromatic epigenetic marks (26). The insertion of a TE close to a gene can generate epigenetic marks in its vicinity and modify its transcription. WGD are relatively common in the history of eukaryotes and especially in plants (1, 56, 62). It is reported that at least one WGD occurred in the evolutionary history of all flowering plants (27). In the *Rosaceae* family, which include apple, many WGD events occurred and may have contributed to the wide diversity of fruit types in this clade (73). One recent WGD, dated to 50 MYA (63), is shared among *Pyraeae*. In apple and pear, chromosomes already undergone rearrangement with a return to diploid state, but the conservation of this event is still really strong with an almost one-to-one chromosome synteny. Given the available high quality genome, this makes apple an organism of choice to study the early evolutionary phases of duplicated DNA fragments and genes. In this study, we investigated imbalances between duplicated chromosomal fragments, with a focus on QTL localization in syntenic blocks, studied the evolution of ohnologous gene sequences and transcription, conducted gene retention quantification, estimated the selection pressure, and analysed jointly TE density and DNA methylation levels. Finally, we explored the possibility of a subgenome dominance in this autopolyploid genome.

## MATERIALS AND METHODS

### Identification of Syntenic Blocks

All analyses on *Malus domestica* genomes were conducted on the apple genome Golden Delicious Doubled Haploid #13 (GDDH13) version 1.1 (13). The identification of syntenic

genes within the apple genome was performed using the i-ADHoRe *ad hoc* pipeline. This pipeline allows for the identification of all homologous genes in the Apple genome using all-versus-all blastP (e-value  $10^{-5}$ ) (8). This homology information was then used as input for i-ADHoRe 3.0 (47), allowing the construction of syntenic blocks based on the search of dense diagonals in sparse matrix of homologous genes. For this analysis, the parameters used for i-ADHoRe 3.0 were: cluster\_type=colinear; tandem\_gap=15; prob\_cutoff=0.001; gap\_size= 30; cluster\_gap= 30; q\_value=0.75; prob\_cutoff=0.01 and anchor\_points=5. The identified set of homologous genes allowed reconstruction of syntenic fragments that can be visualized using circular plots constructed by Circos (32), a tool for visualization of data in circular layouts. This pipeline has been designed for use in any other genome to detect intraspecific and interspecific syntenic blocks and can be used in other contexts than the current study. It was used only to identify intraspecific syntenic blocks in this study. This pipeline also aims at reconstructing triplets, which construction details are presented in the section associated with Ka/Ks.

### Genome Fractionation

Differential loss of duplicate genes from homologous genomic regions is known as biased fractionation. Genome fractionation has an important role in the diploidization mechanism and in subgenome dominance. This analysis was carried out using SynMap (23) and, in particular, the FractBias (28) tool, a module that allows the calculation of the percentage of retention of ohnologous genes and estimates fractionation compared to a related genome. Here the peach genome *Prunus persica* has been chosen, because it is the closest high quality genome to apple genome without the WGD and thus show a 2:1 synteny, i.e. 2 *malus* genes and one paralog gene in *Prunus*. Using a sliding window of 100 genes along the peach chromosomes (target genome), FractBias allows the computation of the percentage of homologous genes on all the apple chromosomes (query genome). The raw results were then processed separately to compare the percentages of gene retention between pairs of chromosomes using the one-sided paired t-test.

### Quantitative Trait Loci

To investigate whether ohnologous chromosome fragments contribute equally to phenotypic variability, we relied on previously published Quantitative Traits Loci (QTL) studies. QTLs are defined as loci statistically associated with variation in phenotypic characters. To conduct a study as exhaustive as possible, we retrieved all publicly available biparental QTLs for apples from the GDR database (29) and added manually curated QTL results derived from GWAS studies. In GDR, QTLs are genetically located using molecular markers, but their location on the physical map, in particular in GDDH13 1.1, is not provided systematically. As a first step, we deleted QTLs that were not associated with molecular markers of known sequence, as they cannot reliably be associated to a physical location. When a marker of known sequence was located with 5 cM of a QTL peak, that marker was considered as the reference marker for the associated QTL. For the other QTLs, we retrieved the localization of known genetic

markers from the GDR database. Then, using primer or probe sequences and Blastn, markers were localized on the physical map of GDDH13. All QTLs are annotated with metadata describing the phenotypic trait linked to the QTL. In GDR (29), this annotation is not done by a controlled field making an automatic processing of the information complicated. To overcome this problem, we built an ontology and re-annotated the set of QTLs with it using a manual processing of the existing annotations. This ontology is built on three levels of respectively 7, 20 and 59 terms allowing it to go to the needed degree of precision. The size of the different categories has been balanced when possible to avoid statistical bias. An approach based on QTLs can generate data redundancy since two close QTLs controlling the expression of similar traits are likely to be associated to the same loci or genes. To reduce redundancy, if the confidence interval of the QTL (artificially increased by 10 cM upstream and downstream) overlapped with a QTL associated with a trait of the same ontology category, then one of the two QTLs was removed. Only QTLs associated with a locus associated with a syntenic block was kept. Finally, the proportion of QTLs present on each ohnologous pairs of chromosomes was tested using the proportion z test. The test was also performed for all QTLs of each chromosome segment pair and performed independently using the ontology to test the effects associated with related phenotypic traits.

#### Estimation of Evolutionary Patterns Using Ka/Ks

This analyses requires a related genome lacking the third WGD and were performed using the *Prunus persica* (60) genome version 2.0. Sequence evolution may differ among ohnologous gene pairs. To test the hypothesis of bias in the evolution of the coding sequences, we used Ka/Ks, an estimator of selection pressure. To calculate the Ka/Ks of apple ohnologous coding sequences, we first constructed a triplet of genes consisting of two ohnologous apple genes and the closest ortholog in *Prunus persica*. Triplets were constructed from reciprocal all-against-all BlastP using the Malus ohnologous sequences and Prunus proteins. Triplets were kept if both apple ohnologous proteins had the same Prunus best blastp hit, allowing the construction of reliable triplets. This part was implemented in the syntenic blocks pipeline described above. The Ka/Ks of these triplets were then computed using an *ad hoc* pipeline. This pipeline starts by aligning the protein sequences of the triplets using MUSCLE (14). The protein alignment is then converted to a nucleic alignment using PAL2NAL (58) with nogap parameter. Then the YN00 tool from PAML (75) (icode=0; weighting=0; commonf3x4=0) uses this nucleic alignment to calculate the Ka/Ks two by two using the Yang and Nielsen model (74). This evolutionary model has the advantage of taking into account the transition-transversion rate and the nucleotide frequencies. To finish Ka/Ks, Ka and Ks values of ohnologous genes pairs between syntenic chromosome fragments were tested using the Wilcoxon test.

#### RNA-Seq

In the context of subgenome dominance, gene expression is often imbalanced. To test for potential apple transcriptional imbalance among ohnologous genes located within syntenic

blocks, we gathered all publicly available RNA-Seq experiments from *Malus domestica*. Raw datasets were quality filtered (paired data; at least three biological replicates; a total size of at least  $2 \times 10^7$  reads per sample and a minimum read size of 100 nt) and downloaded. Selection and download were performed using a turnkey pipeline *ad hoc*, based on the SRA toolkit and Aspera. The raw data were processed using a pseudo-mapping pipeline *ad hoc*. This pipeline produces the trimming (using cutadapt (37), trimmomatic (5) is usable) and quality check (using FastQC (4)) steps. Pseudo-mapping and quantification of transcripts were performed using Salmon (43). The index was constructed from the GDDH13 1.1 reference transcriptome, with Salmon's indexing tool in version 4 set with a dense sampling, k-size at 31 for a total of 60,172,826 k-mers. Duplicates are not kept. The pseudo-mapping was done with Salmon version 1.3.0 with an ISR library and elimination of ambiguous reads. All experiments with global mapping rates below 70% were removed from analysis. The number of reads for each pair of ohnologous genes was then normalized according to the length of the longest gene of the pair. Finally, a differential analysis between ohnologous genes within the same experimental conditions was conducted applying anaDiff (44), a set of scripts based on DESeq2 (36) and edgeR (38). All pipeline step metrics were aggregated using MultiQC (17). To test for differences between transcription levels within pairs of ohnologous chromosome segments, a statistical analysis was performed to test for the number of differences in mapped reads between ohnologous genes based on the Wilcoxon test for each experiment. This set of tests was conducted with all differentially expressed genes regardless of the logFC (described as horseshoe rule in the literature (49, 52, 71)) but also on differentially expressed genes with a logFC of at least 2 in absolute value (described as 2-fold rule (49, 52, 71)). The results of the tests were aggregated for each pair of ohnologous chromosomes by applying the Fisher method. This analysis provided a final result for each pair of chromosomes, allowing the determination of the presence or absence of a transcriptional imbalance using all the available quality filtered RNA-Seq experiments. Furthermore, for each of the genes and for each of the 148 experiments, we counted the number of times that a given gene was overexpressed in relation to its counterpart in order to verify if certain genes had similar behaviors regardless of the tissue and the growth conditions.

#### Transposable Elements

TEs have been associated to subgenome dominance in different studies (20, 71). In order to tackle this hypothesis, a turnkey pipeline was written based on a previously described method (11). TE annotation was derived from the GDDH13 genome annotation v1.1 (13). The predicted annotated TEs were filtered using the 80-80-80 rule: at least 80 percent identity with the target, 80 percent coverage of the target, and a minimum of 80 bases (69). The pipeline identifies all the quality filtered TEs associated with each ohnologous gene within 2 kb upstream and downstream of the gene sequence as well as in introns (hence, excluding exonic TEs). The density  $D_g$ , and the coverage  $C_g$  of TEs associated with each ohnologous gene  $g$  are computed using Equation 1 and

Equation 2 respectively. With  $N$  defined as the number of TEs,  $L_g$  the gene size,  $L_f$  the size of the flanking regions (2 kb),  $L_{exonic}$  the size of exonic regions of considered gene,  $L_{TEintrinsic}$  the size of TE overlapping intronic regions, and  $L_{TEflanking}$  the size of TEs overlapping upstream and downstream flanking regions. These values were then compared using a Wilcoxon test.

$$D_g = \frac{N}{L_g + (2 \times L_f) - L_{exonic} - L_{TEintrinsic} - L_{TEflanking}} \quad (1)$$

$$C_g = \frac{L_{TEintrinsic} + L_{TEflanking}}{L_g + (2 \times L_f) - L_{exonic}} \times 10^2 \quad (2)$$

### Methylome

In the context of genome subdominance, epigenetics regulations are also suspected to play a role, in particular DNA methylations. To investigate this hypothesis, we have set up an analysis that allows to search for imbalance of methylation ratios in the CG, CHG, and CHH contexts (H as A, C, or T). First, we identified Bisulfite-Seq samples that passed the same quality criteria as those defined for RNA-Seq runs. Raw data were then processed using BiSePS (55), a tool for bisulfite sequencing analyses. This tool allows mapping using Bismark (31), a software built to map bisulfite treated reads and perform methylation site calls. Metrics provided by BiSePS are aggregated with MultiQC for quality control. The Bismark output files, and in particular the methylation report files, were then processed in a pipeline, allowing parsing, filtering, and statistical processing. Specifically, the positions of the cytosines associated with one of the three contexts was associated with their positions in the genes. Intronic cytosines were removed. The samples were treated separately, as well as the methylation contexts. Four analyzes were then performed on the cytosines associated with the exonic region of gene bodies of ohnologous genes, gene environment 2 kb upstream and downstream, 500 bp, or 100 bp upstream of the ohnologous gene sequences. For each case, a Mann-Whitney  $U$  test was implemented between pairs of ohnologous chromosomes to compare the distribution of methylation percentages of all positions for each of the considered samples. The results were then aggregated using Fisher's method to obtain an overall result for each pair of syntenic chromosomal segment. To follow, we looked at the number of contexts associated with each of the ohnologous genes in upstream. To do this, we first compared the number of cytosines associated with each of the upstream contexts of the ohnologous genes. Then, we looked at the number of associated cytosines in each context for a window size ranging from 50 bp to 2 kb upstream. For each of the ohnologous chromosome pairs, we compared the number of cytosines and analyzed the 50 most different genes for each window size. We then looked at the percentage of these genes that were annotated as non-switching gene in RNA-Seq analysis. As a negative control, we took 50 genes at random each time.

### Code development and statistics

All scripts were written in Python (65) or R (59). Data analysis and manipulation took advantage of Pandas (39), statistical analyses used Scipy (65) and NumPy (22). Visualizations were produced with Plotly (25). The workflows were handled with Snakemake (40), a workflow management tool that allows reproducible and highly scalable analyses. All pipelines and produced data are provided in associated GitHub repositories.

## RESULTS AND DISCUSSION

### Syntenic blocks

Using the apple GDDH13 1.1 reference genome (49,921 coding genes), the pipeline associating a blastP-all-against-all and i-ADHoRe 3.0 allowed the identification of 865 syntenic blocks. We identified 16,779 pairs of genes, of which 10,958 pairs were identified in only one block of synteny. The remaining 5,821 pairs of genes (representing 6,327 unique genes) were identified in more than one block of synteny, demonstrating the existence of older WGD. Overall, syntenic blocks represent 91.38% of the genome length. The median size of blocks comprised 9 gene pairs. The circular layout visualization is presented in Figure 1, showing all identified gene pairs. This analysis confirmed the good conservation of the WGD, since very few regions were not syntenic (*i.e.* the extremity of chromosome 13 is not syntenic with any other section of the sequenced genome). We can observe syntenic fragment pairs at the scale of the whole chromosome, such as found for pairs 3-11, 5-10, 13-16 and 9-17. We also identified blocks at the half-chromosome scale such as pairs 1-7 or 6-14. In most cases, the breaking point was found close to the centromeric regions of the chromosomes. This result is similar to what has been found in a previous study (13). Syntenic blocks within the apple genome were also constructed with MCScanX and SynMap3 and produced similar results (results not shown). The independant construction of Malus-Prunus triplets for the calculation of Ka/Ks confirmed the consistency of the results. In the pear that had this WGD, a similar number of syntenic genes (16,509 genes) were identified using MCScanX (33).

### Gene Fractionation

Using the peach genome as reference, we computed the number of paralogous genes between the syntenic apple chromosomal regions and the peach genome. Gene retention was calculated for 46,225,100 sliding windows of 100 genes. Statistics were computed using the one-sided paired t-test with the following side choice: for pairs 01-07, 03-11, 06-14, 09-17, 13-16 we investigated if the first chromosome (or chromosome segment) of the pair was under-fractionated; for pairs 02-15, 05-10, 08-15, we investigated if the first chromosome (or chromosome segment) of the pair was over-fractionated. Visualization of the percentage of homologous genes observed within the 100-gene window of the peach genome compared to the apple genome is shown in Figure 2. From this set of curves, several areas of over-fractionation and under-fractionation can be observed among chromosome fragments, such as the end of chromosome 8 of *Prunus persica* (Pp08, windows 4153 to 4288). In this region, the percentage



**Figure 1.** Circos plot representing the syntenic regions within the apple genome GDDH13 v1.1. Links between ohnologous gene pairs are represented. Syntenic blocks can be observed at the chromosome scale for pairs 3-11 or 5-10. Half-chromosome blocks can be observed for pairs 1-7 and 2-7.

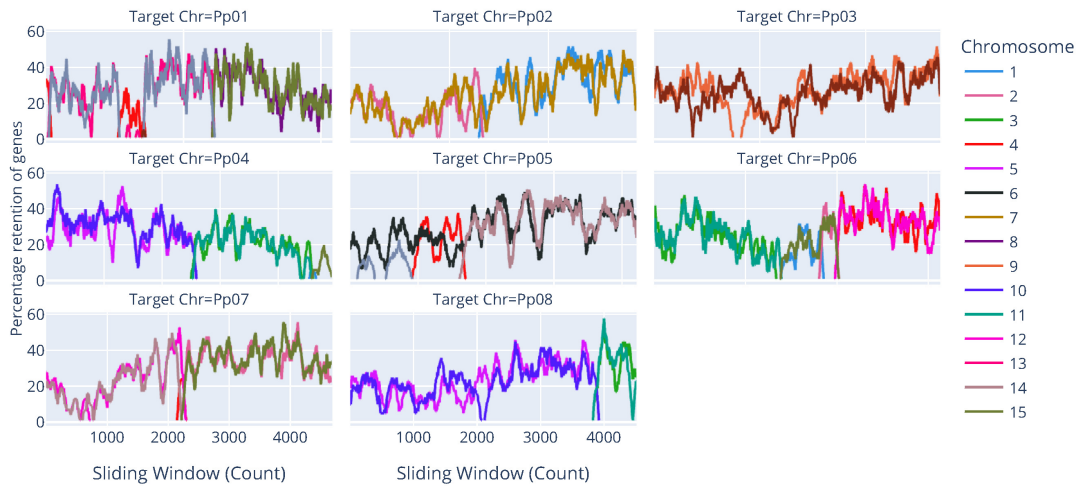
of genes retained on chromosome 11 of *Malus* (median in this region is 38%) is above the percentage identified for chromosome 3 (median is 25%). To test whether fractionation was significantly different among syntenic chromosomes, the percentages of fractionation of syntenic chromosomes in all windows were tested using an one-side paired t-test. The results of the appaired t-test is summarized in the Table 1 and the percentage distribution is represented as a group box plot in Figure 3. From this gene fractionation statistical analysis and the associated visualizations, we established that all major syntenic chromosome fragment pairs are biased. If we look in detail at the retention percentage curves of genes along the chromosomes of *Prunus*, some regions seem to be more biased than others. As an example, the center of chromosome 5 of *Prunus persica* (Pp05) presents an under-fractionation of chromosome 14 in *Malus* compared to its homologous chromosome on this portion, which is chromosome 6 of *Malus*. The significant fractionation bias observed here suggests that a phenomenon of subgenome dominance occurs within the apple genome. In fact, in different species that have experienced recent WGD and have undergone diploidization, biased fractionation has been associated with phenotypic, transcription, TE environment, and methylation imbalances. This has been observed, for example, in maize (52, 77) with a WGD dated to 10 MYA, *Brassica rapa* (66) with a WGD dated to 20 MYA and cotton (49) (60 MYA). The observed biased fractionation in apple confirms that gene fractionation, which usually occurs in the early phases following WGD, seems chromosome dependent and stable.

### Quantitative Trait Loci

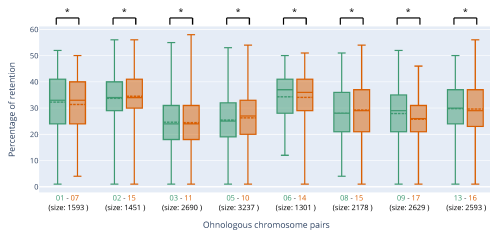
Using the GDR database (29), we retrieved 1,520 QTLs obtained from biparental genetic maps. A first screen based on the identification of the closest markers resulted in the selection of 1,079 QTLs. A total of 135 QTLs derived from GWAS studies were then manually added. To determine their physical location on the apple genome, we used a set of 515,990 markers available from GDR. These markers are mainly SNPs (506,047) and SSR sequences (8,519). This allowed us to localize a total of 513,748 markers on the physical map of the genome. These markers were used to locate 868 QTL among the 1079 on the GDDH13 1.1 genome. Due to the redundancy of some QTL studies and the similarity of some traits studied, and to avoid the representation of similar information multiple times, a cleaning step based on trait ontology was performed. This resulted in a final data set of 541 QTLs, 178 being associated with SSR markers, 341 associated with SNPs, and 22 QTLs associated with markers of other types. Finally, only QTLs associated with a locus located within a chromosome fragment considered as syntenic were kept, representing a total of 487 QTLs (figure 4). Number of QTLs carried by each chromosome of main ohnologous chromosome pairs are shown in Figure 4. The comparison of the proportions of QTLs located on syntenic fragments was performed using a proportion z-test and a binomial test. These tests were carried out on all QTLs, and on QTLs according to the associated trait ontology. Globally, the pairs 1-7, 3-11, 13-16 and were found as significantly imbalanced. If we look at the associated phenotypic traits, the proportions are different especially for traits associated with phenolic compound production, traits associated with fruit quality and biotic stresses. This is particularly true for pairs 1-7, 2-15, 3-11 and 6-14. These results suggest that some chromosome fragments contribute more than their ohnologs to the phenotypic variation of individuals. This study was conducted on 1,528 QTLs identified in apple which represents the essence of all known QTL in apple. These QTL are all associated with agronomic interest traits and do not represent the whole range of apple characteristics. Nevertheless, in order to limit the different biases related to the study of QTL and notably the low precision of localization, we have been stringent on the retained QTL and deleted a large part for the calculation of QTL proportions.

### Estimation of Evolutionary Patterns Using Ka/Ks

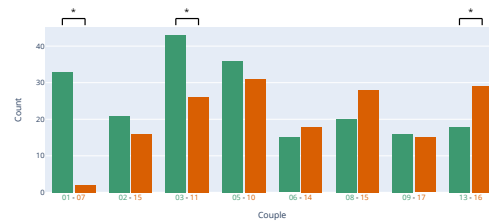
To investigate the observed QTL imbalance, we examined a possible difference in the evolution of ohnologous genes using Ka/Ks ratio, an estimator of selection pressure. By a reciprocal best-blast-hit approach between the peach and apple genomes, we constructed a set of gene triplets composed of two ohnologous genes from apple and their putative orthologous peach gene. This approach resulted in the identification of a set of 10,403 triplets. The protein sequences corresponding to these triplets were then aligned. The identity percentages obtained varied between 80 and 100% at the proteic level, indicating a close similarity between members of the triplets. The median size of the nucleotide alignments was 942 bp, allowing the calculation of the Ka/Ks for 9,821 gene triplets. For 582 triplets, no Ka/Ks could be estimated due to the



**Figure 2.** Percentage retention of ohnologous genes of *Malus domestica* using a sliding window of 100 genes in the *Prunus persica* genome. Each colored line represents the percentage of retention of a *Malus* chromosome. Fractionation bias can be observed in different regions as the end of Pp08, centers of Pp03 and Pp05.



**Figure 3.** Distribution of the percentage of retention of ohnologous genes of *Malus domestica* using a sliding window on the *Prunus persica* genome. On the x-axis, the pairs of ohnologous chromosomes considered are presented. The number of gene pairs is indicated in parentheses. The y-axis represents the percentage of retention values. The first box plot represents the distribution on the first chromosome of the pair (in green) and the distribution on the second chromosome of the pair (in orange). \*P-value < 0.05 for the paired t-test.



**Figure 4.** Histogram of the number of QTLs associated with syntenic blocks in each pair of ohnologous chromosome fragments. The green bar is associated with the first chromosome of the pair. The red bar is associated with the second chromosome of the pair. \*P-value < 0.05 for z-proportion test

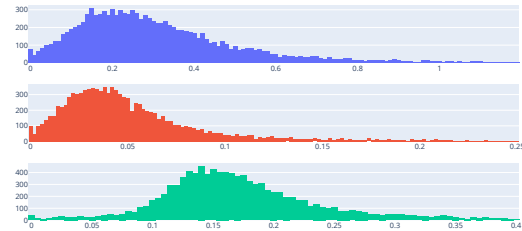
too short alignment size resulting from the gapped regions generated by PAL2NAL.

Distributions of  $K_a$ ,  $K_s$  and  $K_a/K_s$  are shown in Figure 5. The  $K_s$  distribution is clustered into a single peak, suggesting, as expected, a single and recent WGD. The value of  $K_s$  peaks in intraspecific ( $K_s=0.151$ ) and interspecific ( $K_s=0.293$ )

**Table 1.** Results of the paired one-sided t test for gene retention rates between the major ohnologous chromosomes

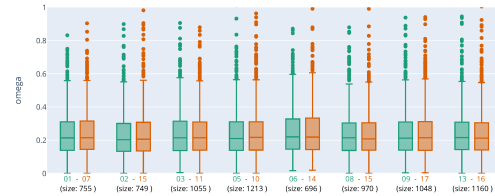
Couple	01-07	02-15	03-11	05-10	06-14	08-15	09-17	13-16
P-value	<10 <sup>-5</sup>	<10 <sup>-5</sup>	<10 <sup>-5</sup>	<10 <sup>-5</sup>	0.0346	<10 <sup>-5</sup>	<10 <sup>-5</sup>	0.0137

From this set of paired t-test, all main syntenic chromosomes have significantly different values.



**Figure 5.** Distribution of Ka/Ks, Ka and Ks values in the apple genome. The blue distribution represents the Ka/Ks, the red curve represents the Ka and the green curve represents the Ks. The Ks curve shows a single peak, confirming that there is only one recent WGD.

distributions, as well as speciation dating based on fossil evidences (70) suggest an approximate WGD dating back to 41 MYA, which is consistent with previous datings (63). The Ka/Ks distribution for each chromosome and grouped by syntenic segment pairs are plotted in Figure 6. We observe that most values are below 1, suggesting a negative selection of ohnologous genes in all chromosomes. The median of Ka/Ks distributions for each chromosome is around 0.20, which is consistent with the literature and, in particular, with plants having undergone a recent WGD (33). We tested for a bias in the selective pressure among ohnologous genes using the Ka/Ks ratio against the peach gene orthologs as an estimator and using the Wilcoxon test. The results presented in Table 2 indicate that the rate of evolution of the ohnologous coding sequences is not significantly different for all tested pairs. This result is similar regardless of the degree of chromosome rearrangement of the pairs considered. Similar analyses on the rate of synonymous (Ks) and non-synonymous (Ka) mutations suggest similar results. Therefore, our results suggest that genes that were maintained after the first round of post-WGD pseudogenization are subjected to similar selection pressure. Due to the approach used to construct the Malus-Prunus triplets, we studied similar genes that were conserved during evolution. Pseudogenes were not included in this analysis, as they are often not or only partially detected by gene modellers, and may be part of the source of the observed discrepancies. It has been shown that the use of triplets obtained by reciprocal best hits of all-against-all Blastp matches allows a better resolution than the use of blastp-all-against-all especially in the estimation of Ka/Ks associated with the different WGD that occurred during the evolutionary history of *Arabidopsis thaliana* (48). This study also showed that the use of a phylogenetic tree and more species would have allowed a better resolution especially in the case where several WGD were studied. In the case of the study of a recent WGD like apple WGD, the use of reciprocal best hits of all-against-all Blastp matches allows a sufficient resolution. Thus, the previously observed QTL imbalances cannot be explained by differences in the evolution of gene coding sequences. Similar results were observed in different species, including species with subgenome dominance such as maize (72).



**Figure 6.** Grouped box plots of the distribution of Ka/Ks values for ohnologous chromosome pairs. On the x-axis, the pairs of ohnologous chromosomes considered are presented. The number of gene pairs is indicated in parentheses. The y-axis represents the Ka/Ks values. The box plot represents the distribution of Ka/Ks. The first chromosome of the pair is presented in green. The second chromosome of the pair is colored in orange. \*P-value < 0.05.

**RNA-Seq**

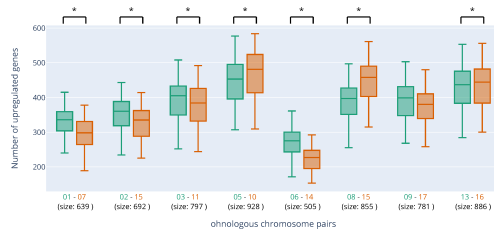
In the case of subgenome dominance, imbalances in expression levels between subgenomes were observed in various species, such as maize (52) and *Brassica rapa* (10). In this study, a differential ohnologous gene expression analysis was performed using DESeq2 using 444 publicly available RNA-Seq runs. The samples from which the RNA was sequenced represent a wide range of organs and conditions. Samples were grouped in 148 experiments of 3 biological replicates. For each sample, RNA-Seq short reads were trimmed and mapped on the GDDH13 reference genome. The percentage of mapping ranged from 72% to 91% with a median of 85%. The proportion of ambiguous reads in the mapping step remained low with a global mean of 0.048%, despite the high similarity between ohnologous gene coding sequences, as shown by the Ka/Ks analysis. Indeed only 596 genes had a median value greater than one ambiguous read across all experiments. From a total of 16,779 pairs of ohnologous genes compared, we identified an average of 11,307 significant differentially expressed ohnologous genes in each experiment of which 6,932 with at least two-fold higher value.

To test for an imbalance between ohnologous gene transcription levels, we counted the number of genes significantly upregulated for a chromosome (or chromosome segment) relative to its ohnolog in each of the 148 RNA-Seq

**Table 2.** Summary of the results of the Wilcoxon test between the Ka/Ks ratio of Malus ohnologous genes against their reference in Prunus

couple	pValue
01-07	0.1900
02-15	0.1408
03-11	0.1936
05-10	0.0830
06-14	0.3472
08-15	0.4850
09-17	0.4346
13-16	0.4472





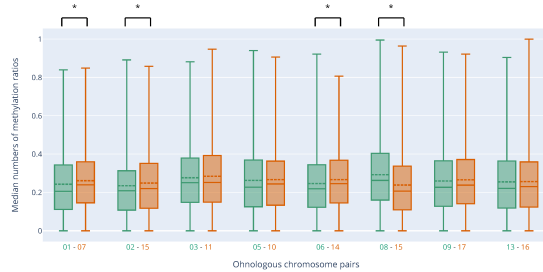
**Figure 7.** Box plot of the distributions of the number of upregulated genes constructed from the set of ohnologous gene differential analyses performed for the 148 experiments. In contrast to what is classically done, here the differential analysis has been done between the ohnologous genes within the same condition; and not for each gene between two conditions. The x-axis represents the set of values of the upregulated genes in each of the 148 experiments. The y-axis represents the set of values for the number of upregulated genes. The x-axis indicates the chromosome pairs considered. For each chromosome pair (or chromosome segments), the box plots show the distribution on the first chromosome of the pair (green) and the distribution on the second chromosome of the pair (orange). \*Wilcoxon p value < 0.05.

experiments considered. Wilcoxon tests were implemented to compare the number of average reads between biological replicates of all differentially expressed genes on each pair of chromosomes and for each of the 148 experiments separately. The results were aggregated using Fisher's method to obtain a single result for each pair. The distribution of the number of up-regulated genes across all experiments is presented in Figure 7. This figure shows that globally the number of genes upregulated in one chromosome or chromosome segment is significantly different compared to its ohnologous gene. All pairs are significantly unbalanced, except for pair 9-17. We identified ohnologous genes pairs that behave systematically in the same way in the 148 experiments. These genes are called here non-switching genes. This list of 2,247 gene pairs seems to be randomly distributed along the chromosomes, and does not favor some genomic regions over others. 1,387 belong to the 8 main pairs of syntenic blocks. Furthermore, we did not identify significant differences in the proportion of non-switching genes between the different pairs of ohnologous chromosomes, except for pair 8-15 (z proportion test,  $\alpha = 0.05$ ).

Thus, we were able to identify an imbalance between ohnologous genes expression levels for all major ohnologous chromosomal fragment pairs except pair 9-17. This imbalance is confirmed for the most exhaustive set of experience of high-quality RNA-Seq experiments available in public databases that represent a diversity of tissues and treatments.

### Transposable Elements

TE have been associated to subgenome dominance (64). To further evaluate whether imbalance between syntenic blocks could be identified in the apple genome, we investigated the distribution of TE. From this analysis, we were able to associate 109,069 TEs with ohnologous gene exons and their genomic environment (2 kb upstream and downstream of the genes), and 150,927 TEs with singletons and their



**Figure 8.** Box plot of the distribution of TE coverage for each pair of ohnologous chromosomes. On the x-axis, the pairs of ohnologous chromosomes considered are presented. The number of gene pairs is indicated in parentheses. The y-axis represents TE coverage values. The first box plot represents the distribution of the TE coverage on the first chromosome of the pair (in green), while the second box plot represents the distribution of TE coverage on the second chromosome of the pair (in orange). \*P-value < 0.05.

genomic environment. The median number of TEs per gene was similar between the singleton group and the ohnologous genes groups, with a median of 5. However, singletons were found to have longer TE median coverage compared to genes with ohnologs, with 0.40 and 0.24, respectively. Using the Mann-Whitney *U* rank tests, we confirmed a different TE distribution coverage between the ohnologous genes and the singletons group (p-value  $\leq 10^{-5}$ ). Similar results have been found in other organisms such as *Arabidopsis thaliana* (24) or *Homo sapiens* (11), and confirm the importance of TEs in structuring the post-WGD genome by creating zones of high or low methylation, and heterochromatic zones with lower recombination rates (64).

Moreover, it seems that the TE distribution difference between singletons and ohnologous genes may have an impact on the fate of duplicated genes and in particular on their retention (64). Indeed, this mechanism could explain in part the difference in selection pressure applied to these sequences. In fact, TEs are less inserted in regions under selection pressure (54).

TEs associated with ohnologous genes are divided into three classes. Class II TE represents 40.4% of all TE (68,832), class I TE represents 27.7% (47,203), while the last class is made up of TE 'unknown' and represents 31.9% (54,378). The subfamilies are mainly represented by TIR (Class II) and LTR (Class I), and the subgroups are represented by hAT (14.5%, Class II), Gypsy (11.3%, Class I), PIF-Harbinger (10.8%, Class II) and SINE2/tRNA (4.8%, Class I). Globally, the distributions of the different TE classes, order and type among pairs of chromosomes are not significantly different ( $\chi^2$  test, threshold  $\alpha = 5\%$ ) except for pairs 9-17 and 8-15 for TE classes and 2-7 and 2-15 for TE subfamilies. For the TE class distribution, in pair 8-15, we observed a significantly different proportion of class I TEs on chromosome 15 compared to chromosome 8. For pair 9-17, chromosome 17 has significantly more class II TEs. For the distribution of the TE subfamilies, we observed higher proportions of Copia and PIF-Harbinger for chromosome 2 compared to chromosome 7. For pair 2-15, chromosome 2 accumulated a significantly higher proportion of PIF-Harbinger than chromosome 15.



In the literature, differences in the distribution of some TE classes have been observed, especially in the organism with genome impacted by WGD and in particular genomes with subgenome dominance such as maize, with a particular LTR dynamic (Class I) (16) and *Brassica rapa* with a dynamic in different types of TEs (3). These differences lead to DNA methylation differences in subgenomes. Within the apple genome, these differences are weak and only significant for a few pairs. However, in different species, such as *Arabidopsis thaliana* (9), the difference is not related to the difference in proportions within particular classes of TEs, but to the difference in all TE classes. To test for differences in TEs in exonic regions and gene environment (2 kb up and downstream), we tested the TE coverage distributions between ohnologous genes for the main pairs of ohnologous chromosome fragments using a Wilcoxon test. The results indicate that pairs 1-7, 2-15, 6-14, and 8-15 are significantly imbalanced ( $\alpha = 5\%$ ).

From this analysis we were able to identify a set of indicators related to TE associated with the ohnologous genes. For the pairs 1-7, 2-15, 6-14, and 8-15 we were able to show that the coverage and density of TE are significantly different. This biased dynamic seems to be related to the TEs as a whole rather than to a particular class or subfamily. The high presence of TEs may cause high DNA methylation in these regions. Thus, a skewed distribution of TEs could cause imbalances in DNA methylation.

### Methylome

To further investigate an imbalance between syntenic blocks of the genome in the apple genome, we tested for imbalance of methylation marks in gene bodies or in the upstream region of genes located in syntenic chromosomal fragments. We retrieved 36 publicly available bisulfite sequencing runs derived from various experiments. The alignment and methylation ratio identification steps were executed via BiSePS pipeline. A total of 4 samples with a mapping percentage below 60% were removed. The median mapping percentage for the 32 remaining samples was 75.44%, for a median of 10,896,005 mapped reads and 1,025,086 ambiguous reads.

The methylation ratio value in ohnologous pairs of chromosomes was calculated in different regions (exonic regions of the gene *i.e.* gene body, environment of the gene and upstream region) and for the three cytosine contexts. For all regions, the CG context methylation rates are higher than the CHG and CHH contexts (average around 0.80 compared to 0.43 and 0.21). The CHH context presents higher values than the CHG context with average values 0.43 against 0.21. For each experiment and each pair of ohnologous fragments of chromosomes, we tested if the methylation rates were similar using a Mann-Whitney *U* test. The p-values of each experiment, for each pair, were then aggregated using Fisher's method and are presented in Table 3. Regarding the exonic gene body, the CG context presents significant different levels of methylation for pairs 1-7, 2-15, and 3-11. Pairs 5-10, 6-14, 8-15, 9-17, and 13-16 show little or no significant differences. For pairs 1-7, chromosome 1 is significantly hypermethylated on the gene body compared to chromosome 7. For pair 2-15, we observe an hypermethylated chromosome 2 compared

to its ohnologous chromosome 15. The pair 3-11 shows an hypermethylated chromosome 11 against the chromosome 3. Analysis of the CHG and CHH contexts indicates similar differences between overall methylation and gene expression on the pairs 1-7, 2-15 and 3-11. For the 2 kb (up and downstream) gene environment, we identified significant differences for pairs 1-7, 2-15, and 3-11 in all three contexts. The methylation profiles in CG, CHH, and CHG shows a similar trend to the gene body results. For the 500 bp and 100 bp upstream gene environment, we identified a significant difference in methylation rate between ohnologous genes from chromosome pairs 1-7, 2-15 and 3-11 in all three contexts.

We also observed the number of cytosines associated with each of the methylation contexts 2 kb upstream for all ohnologous gene pairs. For the CG and CHG contexts, the numbers of positions are not significantly different. However, for the CHH context, the number of cytosines was significantly different in all ohnologous chromosome fragments. Out of the 50 pairs of ohnologous genes displaying the most differences in the number of cytosines associated with the CHH context, 52% were found to belong to the group of genes which expression was as non-switching in all RNA-seq experiments against 36% for randomly selected genes. A z-proportion test further indicated that the 50 pairs of ohnologous genes displaying the most differences in the number of cytosines associated with the CHH context were significantly enriched in non-switching genes compared to a random selection of genes (Table 4).

### Chromosomal dominance

Looking globally at the results from all analyses for each syntenic chromosomal block pairs, we can observe a trend. In fact, for pairs 1-7, 3-11, 8-15, 6-14 in the 500 and 100 bp upstream window, we found that the chromosome of the pair with the least expressed ohnologous genes, the lowest number of QTLs and the highest TE coverage was the most methylated in all contexts. Pair 8-15 has the particularity of being the only one with a significant difference in selection pressure with a higher selection pressure for pair 8 than for pair 15, which is consistent with the other results for this pair. Pair 2-15 shows an opposite pattern, with the chromosome 2 having the highest gene expression, the highest number of QTLs and the lowest TE coverage but the highest methylation rate.

We constructed the scatter matrix of the average differential value of results from gene expression, TE coverage, selection pressure and methylation analysis along the syntenic blocks for each pair of syntenic chromosomal fragments. For each pair of variables, associated linear regression line was also added. The pair 1-7 is presented in Figure 10. All remaining syntenic pairs are presented in supplementary data. Moreover, the Pearson correlation coefficients were calculated between the different variables for all the pairs of chromosomes. They are presented in the form of a set of heatmaps in Figure 9. From these scatter matrices and heatmaps, we can define a set of trends. To begin with, by looking at the correlation coefficients and regression lines, the expression level differences seem to be mainly positively correlated with the methylation rate in CG context on the gene body. This trend suggests that for a given syntenic block the presence of an imbalance in favour of one of the chromosomes in

methylation on the gene body in the CG context may be linked with an imbalance in the expression level in favour of the same chromosome. In addition, TE coverage, upstream methylation and selective pressure are mostly inversely correlated with expression levels. This trend suggests that for a given syntenic block the presence of an imbalance in favour of one of the chromosomes in methylation on upstream of genes or in selective pressure or in TE coverage may be linked with an imbalance in the expression level in favour of the opposite chromosome. In addition, upstream methylation of genes is inversely correlated with methylation on the gene body. This can be explained by the fact that methylation on the gene body is more associated with gene expression whereas upstream methylation is more related to gene silencing. Moreover, TE coverage is inversely correlated with CG methylation on the gene body and positively correlated with upstream methylation. It is important to note that all the observations described here are not valid for all pairs. For some pairs, especially those that are not or only slightly unbalanced, such as pairs 5-10 or 9-17, we have not identified correlations between the indicators or they are too weak to represent a biological reality. Similarly, some pairs that are unbalanced show behaviors that are different from those presented here. Thus chromosomal dominance seems to be a trend but is not always a well-differentiated mechanism.

These observations are supported by various elements in the literature. Indeed, it has been observed that a higher proportion of TEs may lead to higher DNA methylations, particularly in 5' region of the genes (64, 71). This phenomenon has been observed in different species with subgenome dominance and in particular in maize (15) and *Brassica rapa* (71). Such biased methylations could affect the gene transcription as observed in maize (77), TE repartition has also been described in *Brassica rapa* (71). Those bias could also have an impact on the selection pressure of genes and fractionation as observed in maize (52) as well as their participation in the variation of the phenotype of the organism (50). Furthermore, gene expression biases have been associated with biased genome fractionation.

All these observations are consistent with observations made in allopolyploid species in genome dominance (15, 50, 71, 77). In these species, the mechanism described is a pre-WGD difference in the proportions of transposable elements between the two genomes that will cause differences in methylation, expression of selection pressure genes, retention of genes and participation in the phenotype of the organism between the two post-WGD subgenomes. Such observations have not been made in many autopolyploid species except *Arabidopsis thaliana* (19). Thus, in pear, a species which ancestor is common to apple and has undergone the same WGD, no imbalances in fractionation, level of gene expression, and methylation of genes and their environment were identified (33). Nevertheless, in the course of our analyses we were able to identify significant imbalances in these different indicators in apple. These analyses are as exhaustive and stringent as possible and allow us to reliably identify an imbalance in the proportion of QTLs, genome fractionation, gene expression, TEs, exon methylation, and upstream gene methylation. We propose a model summarized in Figure 11 to gather all these observations on the apple tree. Here we hypothesize that the observed imbalances in QTL, gene expression, TE environment, and DNA methylations

could originate from a bias during the TE burst that followed the WGD. This TE bias resulted in an imbalance of methylation in gene upstream region. The methylation bias would have implication in gene expression and has led to bias in gene expression. These differences in gene expression could then lead to bias in DNA methylation of gene body and to variations in the participation in the phenotype of the organism and thus to the observed QTL imbalance. We can also imagine that these differences in expression could be at the origin of differences in selection pressure on the genes. We have not observed such a phenomenon in apple yet. This could be due to a lack of evolutionary time that has not yet allowed the establishment of such a mechanism. Contrary to what has been observed in allopolyploid species, the mechanism we hypothesize here would be much slower to set up because the original TE differences are much less marked than in species resulting from a WGD by allopolyploidy.

## CONCLUSION

We have shown that QTLs are significantly more frequently associated with a chromosome (or chromosome fragment) than with their ohnologous counterpart. This QTL imbalance is often found to be associated with gene fractionation imbalance among syntenic blocks. To explain the observed imbalance, we investigated the selection pressure applied to the coding sequences of the ohnologous genes through Ka/Ks. Our analysis indicates that (i) ohnologous genes are under strong negative selection and (ii) the coding sequences of the ohnologous genes, not involved in pseudogeneization process, did not evolve significantly differently. These results suggest that gene sequence evolution cannot explain by itself the observed QTL imbalance. We then investigated the transcription profiles of ohnologous genes and found an overall imbalance in their expression levels. Integration of QTLs and transcriptomic data indicated that upregulated genes are often colocalized with genomic regions richer in QTLs compared to their ohnologous counterparts. In a next step, the analyses could be completed by adding functional analysis and, in particular, GO terms enrichment to check if particular gene functions are more affected than others. Moreover, integration of all those results can be useful to build a model of WGD-duplicate genes fate. In addition, our differential expression analysis revealed that in the 148 RNA-seq analyses considered, 1,387 genes were systematically overexpressed compared to their ohnolog. This observation prompted us to analyse the TE data and epigenomic context surrounding the ohnologous genes. Our results indicated a genome-wide imbalance in TE density and methylation patterns among whole chromosome pairs and smaller syntenic blocks. The observed imbalance was even more pronounced when only the previously mentioned group of genes was considered to be systematically overexpressed. In fact, out of the 1,387 genes systematically overexpressed compared to their ohnolog, 60% were found to have a lower TE density within their 2 kb upstream region, 60% were found to be significantly more methylated in CG, CHG and CHH context on gene body, and 57% were found to be significantly less methylated in CHH context on gene upstream. Integration of the various results indicates that in most instances, the chromosomal regions displaying a significantly higher

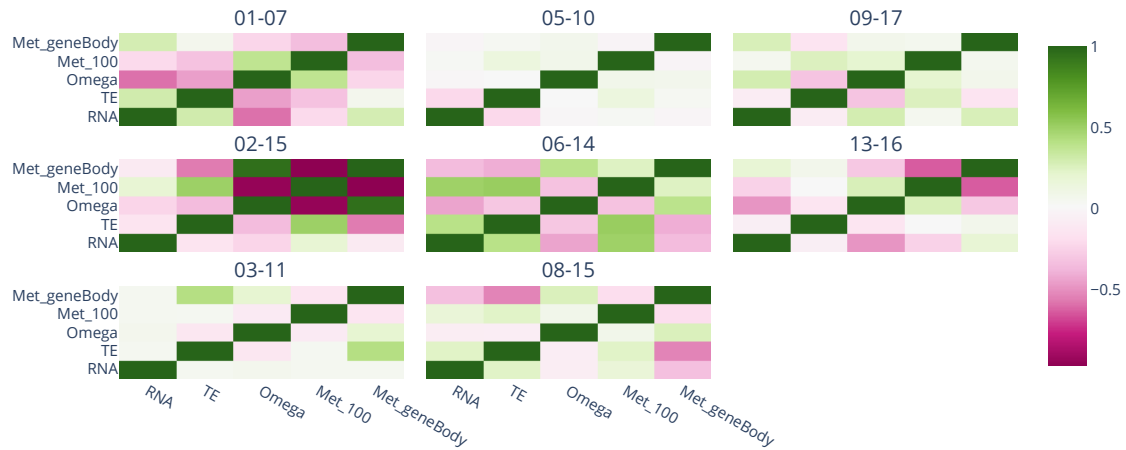


Figure 9. Set of heatmap gathering the correlation coefficients between the different indicators for all the pairs of ohnologous chromosomal fragments analysed

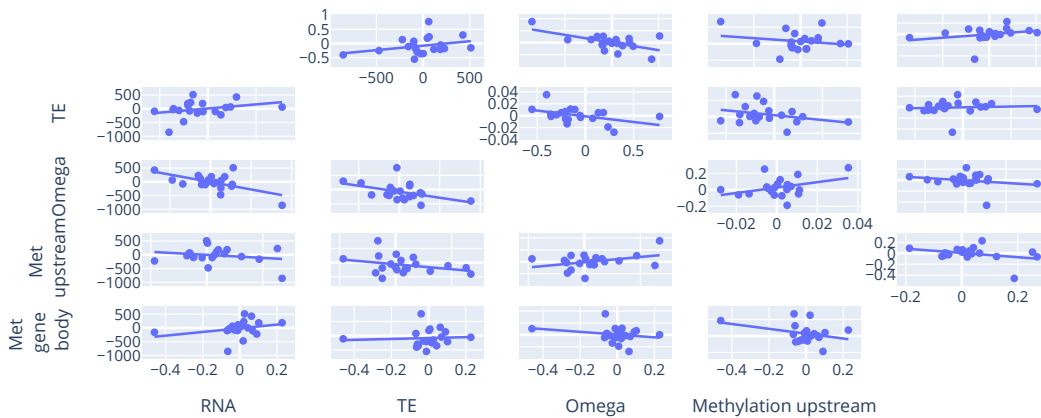
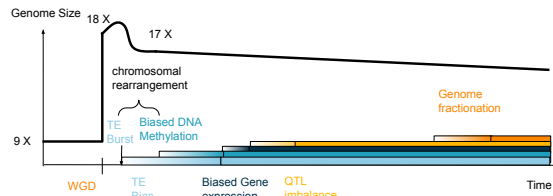


Figure 10. Set of scatter matrix of the different variables and the associated regression lines for the pairs 1-7

number of mapped QTLs, and hence contributing more to the phenotypic variability, compared to their ohnolog, also display a greater percentage of overexpressed genes, and a lower gene fractionation. In addition, these chromosomal regions were found to have an overall lower TE density and to display higher gene body methylation percentages

in all contexts as well as a lower methylation percentages in gene upstream regions. This set of indicators suggest a result similar to the overall observation made earlier but for a particular set of genes. These genes, which expression are systematically higher than their counterpart, represent a set that could be of interest for further study.



**Figure 11.** Summary diagram of the mechanism implemented in the apple genome. After WGD, the genome of the common ancestor of the *Maloidae* increased from 9 to 18 chromosomes. Quickly after WGD a TE burst occurred leading to an increase of genome size. This TE burst was probably biased and followed by important DNA methylations. These mechanisms allow to establish chromosomal rearrangement that will lead to a genome with 17 chromosomes. Moreover, these mechanisms set up biases that will lead to a bias of gene expression between the different pairs of chromosomes, a preferential loss of duplicated genes on some chromosomes compared to their ohnologues and a different participation of ohnologous chromosomal fragments to phenotypic variations.

Altogether, our results indicate that post-WGD evolution affects primarily gene regulatory sequences rather than coding sequences. This mechanism seems to be associated with differences in the distribution of TEs with direct impact on DNA methylation. This differential methylation affects the transcription of genes resulting in transcriptional imbalances between the different pairs of chromosomes. We could not observe any imbalance in selection pressure as a result of this transcriptional difference, but large-scale gene losses were observed with important differences in genome fractionation between different chromosome pairs. Dominant subgenomes have been identified in many allopolyploid species. In this analysis we were able to observe similar bias in an autopolyploid species. In contrast to allopolyploid species where dominance originates from differences in TE in the pre-WGD subgenomes, here we hypothesize that the observed bias may originate from a post-WGD TE burst, leading to a slower genome evolution compared to allopolyploids.

## ACKNOWLEDGEMENTS

We acknowledge the GenOuest bioinformatics core facility (<https://www.genouest.org>) for providing the computing and storage infrastructure. This work was supported by funding from the Régions Pays de la Loire and INRAe.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Keith L Adams and Jonathan F Wendel. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, 8(2):135–141, April 2005.
2. Karine Alix, Pierre R. Gérard, Trude Schwarzacher, and J. S. (Pat) Heslop-Harrison. Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Annals of Botany*, 120(2):183–194, August 2017.
3. Z. An, Z. Tang, B. Ma, A. S. Mason, Y. Guo, J. Yin, C. Gao, L. Wei, J. Li, and D. Fu. Transposon variation by order during allopolyploidisation between *Brassica oleracea* and *Brassica rapa*. *Plant Biology*, 16(4):825–835, 2014.
4. Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012.
5. Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.
6. J. Brosius. Retroposons—seeds of evolution. *Science*, 251(4995):753, February 1991.
7. Rémy Bruggmann, Arvind K. Bharti, Heidrun Gundlach, Jinsheng Lai, Sarah Young, Ana C. Pontaroli, Fusheng Wei, Georg Haberer, Galina Fuks, Chunguang Du, Christina Raymond, Matt C. Estep, Renyi Liu, Jeffrey L. Bennetzen, Agnes P. Chan, Pablo D. Rabinowicz, John Quackenbush, W. Brad Barbazuk, Rod A. Wing, Bruce Birren, Chad Nusbaum, Steve Rounsley, Klaus F.X. Mayer, and Joachim Messing. Uneven chromosome contraction and expansion in the maize genome. *Genome Res*, 16(10):1241–1251, October 2006.
8. Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: Architecture and Applications. *BMC Bioinformatics*, 10:421, December 2009.
9. Meng Chen, Misook Ha, Erika Lackey, Jianlin Wang, and Z Jeffrey Chen. RNAi of met1 Reduces DNA Methylation and Induces Genome-Specific Changes in Gene Expression and Centromeric Small RNA Accumulation in Arabidopsis Allopolyploids. *Genetics*, 178(4):1845–1858, April 2008.
10. Feng Cheng, Jian Wu, Lu Fang, Silong Sun, Bo Liu, Ke Lin, Guusje Bonnema, and Xiaowu Wang. Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of *Brassica rapa*. *PLoS ONE*, 7(5):e36442, May 2012.
11. Margot Correa, Emmanuelle Lerat, Etienne Birmelé, Franck Samson, Bérengère Bouillon, Kévin Normand, and Carène Rizzon. The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality. *Genome Biology and Evolution*, 13(5), May 2021.
12. Liying Cui, P. Kerr Wall, James H. Leebens-Mack, Bruce G. Lindsay, Douglas E. Soltis, Jeff J. Doyle, Pamela S. Soltis, John E. Carlson, Kathiravetpilla Arumuganathan, Abdelal Barakat, Victor A. Albert, Hong Ma, and Claude W. dePamphilis. Widespread genome duplications throughout the history of flowering plants. *Genome Res*, 16(6):738–749, June 2006.
13. Nicolas Daccord, Jean-Marc Celton, Gareth Linsmith, Claude Becker, Nathalie Choisne, Elio Schijlen, Henri van de Geest, Luca Bianco, Diego Micheletti, Riccardo Velasco, Erica Adele Di Piero, Jérôme Gouzy, D Jasper G Rees, Philippe Guérif, Hélène Muranty, Charles-Eric Durel, François Laurens, Yves Lespinasse, Sylvain Gaillard, Sébastien Aubourg, Hadi Quesneville, Detlef Weigel, Eric van de Weg, Michela Troggio, and Etienne Bucher. High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development. *Nat Genet*, 49(7):1099–1106, July 2017.
14. Robert C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.
15. Steven R. Eichten, Roman Briskine, Jawon Song, Qing Li, Ruth Swanson-Wagner, Peter J. Hermanson, Amanda J. Waters, Evan Starr, Patrick T. West, Peter Tiffin, Chad L. Myers, Matthew W. Vaughn, and Nathan M. Springer. Epigenetic and Genetic Influences on DNA Methylation Variation in Maize Populations. *The Plant Cell*, 25(8):2783–2797, August 2013.
16. M. C. Estep, J. D. DeBarry, and J. L. Bennetzen. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity*, 110(2):194–204, February 2013.
17. Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Källér. MultiQC: Summarize analysis results for multiple tools and samples in

- a single report. *Bioinformatics*, 32(19):3047–3048, October 2016.
18. Michael Freeling, Michael J. Scanlon, and John E. Fowler. Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr Opin Genet Dev*, 35:110–118, December 2015.
  19. Michael Freeling and Brian C. Thomas. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.*, 16(7):805–814, July 2006.
  20. Michael Freeling, Margaret R Woodhouse, Shabarimuth Subramaniam, Gina Turco, Damon Lisch, and James C Schnable. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology*, 15(2):131–139, April 2012.
  21. Matthew W. Hahn. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity*, 100(5):605–617, September 2009.
  22. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
  23. Asher Haug-Baltzell, Sean A Stephens, Sean Davey, Carlos E Scheidegger, and Eric Lyons. SynMap2 and SynMap3D: Web-based whole-genome synteny browsers. *Bioinformatics*, 33(14):2197–2198, July 2017.
  24. Austin L Hughes, Robert Friedman, Vikram Ekollu, and John R Rose. Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. *Molecular Phylogenetics and Evolution*, 29(3):410–416, December 2003.
  25. Plotly Technologies Inc. Collaborative data science. <https://plot.ly>, 2015.
  26. Hidetaka Ito and Tetsuji Kakutani. Control of transposable elements in Arabidopsis thaliana. *Chromosome Res*, 22(2):217–223, June 2014.
  27. Yuannian Jiao, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, Yi Hu, Haiying Liang, Pamela S. Soltis, Douglas E. Soltis, Sandra W. Clifton, Scott E. Schlarbaum, Stephan C. Schuster, Hong Ma, Jim Leebens-Mack, and Claude W. dePamphilis. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100, May 2011.
  28. Blake L Joyce, Asher Haug-Baltzell, Sean Davey, Matthew Bomhoff, James C Schnable, and Eric Lyons. FractBias: A graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics*, 33(4):552–554, February 2017.
  29. Sook Jung, Taein Lee, Chun-Huai Cheng, Kathryn Buble, Ping Zheng, Jing Yu, Jodi Humann, Stephen P Ficklin, Ksenija Gasic, Kristin Scott, Morgan Frank, Sushan Ru, Heidi Hough, Kate Evans, Cameron Peace, Mercy Olmstead, Lisa W DeVetter, James McFerson, Michael Coe, Jill L Wegrzyn, Margaret E Staton, Albert G Abbott, and Dorrie Main. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Research*, 47(D1):D1137–D1145, October 2018.
  30. Nikoleta Juretic, Douglas R. Hoen, Michael L. Huynh, Paul M. Harrison, and Thomas E. Bureau. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.*, 15(9):1292–1297, September 2005.
  31. Felix Krueger and Simon R. Andrews. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, April 2011.
  32. Martin Krzywinski, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. Circos: An Information Aesthetic for Comparative Genomics. *Genome Research*, 19(9):1639–1645, September 2009.
  33. Qionghou Li, Xin Qiao, Hao Yin, Yuhang Zhou, Huizhen Dong, Kaijie Qi, Leitong Li, and Shaoling Zhang. Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). *Horticulture Research*, 6, March 2019.
  34. Dang Liu, Martin Hunt, and Isheng J. Tsai. Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics*, 19(1):26, January 2018.
  35. Steven Lockton and Brandon S. Gaut. Plant conserved non-coding sequences and paralogous evolution. *Trends Genet*, 21(1):60–65, January 2005.
  36. Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.
  37. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011.
  38. Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, May 2012.
  39. Wes McKinney. Data Structures for Statistical Computing in Python. In *Python in Science Conference*, pages 56–61, Austin, Texas, 2010.
  40. Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake, April 2021.
  41. Susumu Ohno. *Evolution by Gene Duplication*. Springer Science & Business Media, December 1970.
  42. Sarah P. Otto and Paul Yong. 16 - The Evolution of Gene Duplicates. In Jay C. Dunlap and C. ting Wu, editors, *Advances in Genetics*, volume 46 of *Homology Effects*, pages 451–483. Academic Press, January 2002.
  43. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon: Fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods*, 14(4):417–419, April 2017.
  44. Sandra Pelletier. Spelletier-1345/anadiff: Pegasus production. Zenodo, April 2022.
  45. Saurabh D. Pophaly and Aurélien Tellier. Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. *Molecular Biology and Evolution*, 32(12):3226–3235, December 2015.
  46. Victoria E. Prince and F. Bryan Pickett. Splitting pairs: The diverging fates of duplicated genes. *Nat Rev Genet*, 3(11):827–837, November 2002.
  47. Sebastian Proost, Jan Fostier, Dieter De Witte, Bart Dhoedt, and Piet Demeester. I-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research*, 40(2):11, 2012.
  48. Ren Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, and Ji Qi. Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. *Molecular Plant*, 11(3):414–428, March 2018.
  49. Simon Renny-Byfield, Lei Gong, Joseph P. Gallagher, and Jonathan F. Wendel. Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution. *Molecular Biology and Evolution*, 32(4):1063–1071, April 2015.
  50. Simon Renny-Byfield, Eli Rodgers-Melnick, and Jeffrey Ross-Ibarra. Gene Fractionation and Function in the Ancient Subgenomes of Maize. *Molecular Biology and Evolution*, 34(8):1825–1832, August 2017.
  51. Rhea Vallente Samonte and Evan E. Eichler. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*, 3(1):65–72, January 2002.
  52. James C. Schnable, Nathan M. Springer, and Michael Freeling. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A*, 108(10):4069–4074, March 2011.
  53. Rie Shimizu-Inatsugi, Aika Terada, Kyosuke Hirose, Hiroshi Kudoh, Jun Sese, and Kentaro K. Shimizu. Plant adaptive radiation mediated by polyploid plasticity in transcriptomes. *Molecular Ecology*, 26(1):193–207, January 2017.
  54. Cas Simons, Michael Pheasant, Igor V. Makunin, and John S. Mattick. Transposon-free regions in mammalian genomes. *Genome Res*, 16(2):164–172, February 2006.
  55. SkanderHatira. SkanderHatira/biseps: Release 10.
  56. Douglas E. Soltis, Pamela S. Soltis, and Jennifer A. Tate. Advances in the study of polyploidy since Plant speciation. *New Phytologist*, 161(1):173–191, 2004.
  57. Pamela S. Soltis and Douglas E. Soltis. The role of hybridization in plant speciation. *Annu Rev Plant Biol*, 60:561–588, 2009.
  58. M. Suyama, D. Torrents, and P. Bork. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(Web Server):W609–W612, July 2006.
  59. R. Core Team. R: A language and environment for statistical computing. 2013.
  60. The International Peach Genome Initiative, Ignazio Verde, Albert G

- Abbott, Simone Scalabrin, Sook Jung, Shengqiang Shu, Fabio Marroni, Tatyana Zhebentyayeva, Maria Teresa Dettori, Jane Grimwood, Federica Cattonaro, Andrea Zuccolo, Laura Rossini, Jerry Jenkins, Elisa Vendramin, Lee A Meisel, Veronique Decroocq, Bryon Sosinski, Simon Prochnik, Therese Mitros, Alberto Policriti, Guido Cipriani, Luca Dondini, Stephen Ficklin, David M Goodstein, Pengfei Xuan, Cristian Del Fabbro, Valeria Aramini, Dario Copetti, Susana Gonzalez, David S Horner, Rachele Falchi, Susan Lucas, Erica Mica, Jonathan Maldonado, Barbara Lazzari, Douglas Bielenberg, Raul Pirona, Mara Miculan, Abdelali Barakat, Raffaele Testolin, Alessandra Stella, Stefano Tartarini, Pietro Tonutti, Pere Arús, Ariel Orellana, Christina Wells, Dorrie Main, Giannina Vizzotto, Herman Silva, Francesco Salamini, Jeremy Schmutz, Michele Morgante, and Daniel S Rokhsar. The High-Quality Draft Genome of Peach (*Prunus Persica*) Identifies Unique Patterns of Genetic Diversity, Domestication and Genome Evolution. *Nature Genetics*, 45(5):487–494, May 2013.
61. Brian C. Thomas, Brent Pedersen, and Michael Freeling. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res*, 16(7):934–946, July 2006.
  62. Yves Van de Peer. A mystery unveiled. *Genome Biology*, 12(5):113, May 2011.
  63. Riccardo Velasco, Andrey Zharkikh, Jason Affourtit, Amit Dhingra, Alessandro Cestaro, Ananth Kalyanaraman, Paolo Fontana, Satish K. Bhatnagar, Michela Troglio, Dmitry Pruss, Silvio Salvi, Massimo Pindo, Paolo Baldi, Sara Castelletti, Marina Cavaiuolo, Giuseppina Coppola, Fabrizio Costa, Valentina Cova, Antonio Dal Ri, Vadim Goremykin, Matteo Komjanc, Sara Longhi, Pierluigi Magnago, Giulia Malacarne, Mickael Malnoy, Diego Micheletti, Marco Moretto, Michele Perazzolli, Azeddine Si-Ammour, Silvia Vezzulli, Elena Zini, Glenn Eldredge, Lisa M. Fitzgerald, Natalia Gutin, Jerry Lanchbury, Teresita Macalma, Jeff T. Mitchell, Julia Reid, Bryan Wardell, Chinnappa Kodira, Zhoutao Chen, Brian Desany, Faheem Niazi, Melinda Palmer, Tyson Koepke, Derick Jiwan, Scott Schaeffer, Vandhana Krishnan, Changjun Wu, Vu T. Chu, Stephen T. King, Jessica Vick, Quanzhou Tao, Amy Mraz, Aimee Stormo, Keith Stormo, Robert Bogden, Davide Ederle, Alessandra Stella, Alberto Vecchietti, Martin M. Kater, Simona Masiero, Pauline Lasserre, Yves Lespinasse, Andrew C. Allan, Vincent Bus, David Chagné, Ross N. Crowhurst, Andrew P. Gleave, Enrico Lavezzo, Jeffrey A. Fawcett, Sebastian Proost, Pierre Rouzé, Lieven Sterck, Stefano Toppo, Barbara Lazzari, Roger P. Hellens, Charles-Eric Durel, Alexander Gutin, Roger E. Bumgarner, Susan E. Gardiner, Mark Skolnick, Michael Egholm, Yves Van de Peer, Francesco Salamini, and Roberto Viola. The Genome of the Domesticated Apple (*Malus × Domestica* Borkh.). *Nat. Genet.*, 42(10):833–839, October 2010.
  64. Carlos M. Vicent and Josep M. Casacuberta. Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, 120(2):195–207, August 2017.
  65. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Courmapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, António H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3):261–272, 2020.
  66. Xiaowu Wang, Hanzhong Wang, Jun Wang, Rifei Sun, Jian Wu, Shengyi Liu, Yinqi Bai, Jeong-Hwan Mun, Ian Bancroft, Feng Cheng, Sanwen Huang, Xixiang Li, Wei Hua, Junyi Wang, Xiyin Wang, Michael Freeling, J. Chris Pires, Andrew H. Paterson, Boulos Chalhouh, Bo Wang, Alice Hayward, Andrew G. Sharpe, Beom-Seok Park, Bernd Weisshaar, Binghang Liu, Bo Li, Bo Liu, Chaobo Tong, Chi Song, Christopher Duran, Chunfang Peng, Chunyu Geng, Chushin Koh, Chuyu Lin, David Edwards, Desheng Mu, Di Shen, Eleni Soumpourou, Fei Li, Fiona Fraser, Gavin Conant, Gilles Lassalle, Graham J. King, Guusje Bonnema, Haibao Tang, Haiping Wang, Harry Belcram, Heling Zhou, Hideki Hirakawa, Hiroshi Abe, Hui Guo, Hui Wang, Huizhe Jin, Isobel A. P. Parkin, Jacqueline Batley, Jeong-Sun Kim, Jérémy Just, Jianwen Li, Jiaohui Xu, Jie Deng, Jin A. Kim, Jingping Li, Jingyin Yu, Jinling Meng, Jimpeng Wang, Jiumeng Min, Julie Poulain, Jun Wang, Katsunori Hatakeyama, Kui Wu, Li Wang, Lu Fang, Martin Trick, Matthew G. Links, Meixia Zhao, Mina Jin, Nirala Ramchiary, Nizar Drou, Paul J. Berkman, Qingle Cai, Quanfei Huang, Ruiqiang Li, Satoshi Tabata, Shifeng Cheng, Shu Zhang, Shuijiang Zhang, Shunmou Huang, Shusei Sato, Silong Sun, Soo-Jin Kwon, Su-Ryun Choi, Tae-Ho Lee, Wei Fan, Xiang Zhao, Xu Tan, Xun Xu, Yan Wang, Yang Qiu, Ye Yin, Yingrui Li, Yongchen Du, Yongcui Liao, Yongpyo Lim, Yoshihiro Narusaka, Yupeng Wang, Zhenyi Wang, Zhenyu Li, Zhiwen Wang, Zhiyong Xiong, Zhonghua Zhang, and Brassica rapa Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*, 43(10):1035–1039, August 2011.
  67. Jonathan F. Wendel. The wondrous cycles of polyploidy in plants. *American Journal of Botany*, 102(11):1753–1756, November 2015.
  68. Patrick T. West, Qing Li, Lexiang Ji, Steven R. Eichten, Jawon Song, Matthew W. Vaughn, Robert J. Schmitz, and Nathan M. Springer. Genomic Distribution of H3K9me2 and DNA Methylation in a Maize Genome. *PLOS ONE*, 9(8):e105267, August 2014.
  69. Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhouh, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12):973–982, December 2007.
  70. Jack Wolfe and Wesley Wehr. Rosaceous Chamaebatiaria-Like Foliage from the Paleogene of Western North America. *Aliso*, 12(1):177–200, 1988.
  71. Margaret R. Woodhouse, Feng Cheng, J. Chris Pires, Damon Lisch, Michael Freeling, and Xiaowu Wang. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A*, 111(14):5283–5288, April 2014.
  72. Margaret R. Woodhouse, James C. Schnable, Brent S. Pedersen, Eric Lyons, Damon Lisch, Shabarinath Subramaniam, and Michael Freeling. Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. *PLoS Biol*, 8(6):e1000409, June 2010.
  73. Yezi Xiang, Chien-Hsun Huang, Yi Hu, Jun Wen, Shisheng Li, Tingshuang Yi, Hongyi Chen, Jun Xiang, and Hong Ma. Evolution of rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. 34(2):262–281.
  74. Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17(1):32–43, January 2000.
  75. Ziheng Yang. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–1591, August 2007.
  76. Jianzhi Zhang. Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, 18(6):292–298, June 2003.
  77. Meixia Zhao, Biao Zhang, Damon Lisch, and Jianxin Ma. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *The Plant Cell*, 29(12):2974–2994, December 2017.

**Table 3.** Results of methylation ratios on each pair of ohnologous chromosomes in the different regions considered in the 3 methylation contexts. \*p-value < 0.05

couple	context	region	median of ratios
	CG	gene body	0.525/0.532*
	CG	gene environment	0.784/0.805*
	CG	gene upstream 500bp	0.809/0.909*
01-07	CHG	gene body	0.032/0.034*
	CHG	gene environment	0.358/0.365*
	CHG	gene upstream 500bp	0.411/0.472*
	CHH	gene body	0.012/0.012*
	CHH	gene environment	0.119/0.119*
	CHH	gene upstream 500bp	0.194/0.209*
	CG	gene body	0.487/0.457*
	CG	gene environment	0.745/0.742*
	CG	gene upstream 500bp	0.841/0.722*
02-15	CHG	gene body	0.033/0.038*
	CHG	gene environment	0.326/0.353*
	CHG	gene upstream 500bp	0.444/0.394*
	CHH	gene body	0.012/0.012*
	CHH	gene environment	0.112/0.102*
	CHH	gene upstream 500bp	0.175/0.155*
	CG	gene body	0.553/0.617*
	CG	gene environment	0.876/0.932*
	CG	gene upstream 500bp	0.93/0.921*
03-11	CHG	gene body	0.033/0.041*
	CHG	gene environment	0.419/0.454*
	CHG	gene upstream 500bp	0.489/0.485*
	CHH	gene body	0.012/0.013*
	CHH	gene environment	0.114/0.12*
	CHH	gene upstream 500bp	0.187/0.181*
	CG	gene body	0.572/0.555
	CG	gene environment	0.843/0.846
	CG	gene upstream 500bp	0.83/0.867
05-10	CHG	gene body	0.045/0.041
	CHG	gene environment	0.383/0.396
	CHG	gene upstream 500bp	0.428/0.436
	CHH	gene body	0.012/0.012
	CHH	gene environment	0.108/0.111
	CHH	gene upstream 500bp	0.172/0.181
	CG	gene body	0.523/0.516
	CG	gene environment	0.795/0.852
	CG	gene upstream 500bp	0.816/0.878
06-14	CHG	gene body	0.037/0.028
	CHG	gene environment	0.35/0.394
	CHG	gene upstream 500bp	0.42/0.47
	CHH	gene body	0.012/0.011
	CHH	gene environment	0.114/0.112
	CHH	gene upstream 500bp	0.184/0.18
	CG	gene body	0.484/0.495
	CG	gene environment	0.799/0.697
	CG	gene upstream 500bp	0.864/0.662
08-15	CHG	gene body	0.038/0.025
	CHG	gene environment	0.393/0.316
	CHG	gene upstream 500bp	0.475/0.336
	CHH	gene body	0.013/0.01
	CHH	gene environment	0.118/0.108
	CHH	gene upstream 500bp	0.191/0.168
	CG	gene body	0.555/0.574
	CG	gene environment	0.85/0.852
	CG	gene upstream 500bp	0.816/0.86
09-17	CHG	gene body	0.045/0.035
	CHG	gene environment	0.401/0.394
	CHG	gene upstream 500bp	0.43/0.457
	CHH	gene body	0.012/0.011
	CHH	gene environment	0.109/0.113
	CHH	gene upstream 500bp	0.164/0.189
	CG	gene body	0.451/0.467
	CG	gene environment	0.795/0.802
	CG	gene upstream 500bp	0.759/0.812
13-16	CHG	gene body	0.03/0.03
	CHG	gene environment	0.385/0.394
	CHG	gene upstream 500bp	0.391/0.43
	CHH	gene body	0.011/0.011
	CHH	gene environment	0.117/0.116
	CHH	gene upstream 500bp	0.183/0.183

**Table 4.** Summary of the results of non-switching genes between the most different genes in terms of position number in the CHH context in 2kb upstream and random genes

couple	percentDiff	percentNorm	wilcoxonPval
01-07	0.53	0.37	4.398138e-08
02-15	0.63	0.39	3.310433e-08
03-11	0.45	0.37	8.827114e-07
05-10	0.50	0.30	2.814601e-08
06-14	0.45	0.34	1.646733e-07
08-15	0.50	0.40	2.828986e-08
09-17	0.52	0.32	3.309252e-08
13-16	0.60	0.40	3.514233e-08

---

## Références

- Adams, K. L., & Wendel, J. F., (2005), Polyploidy and genome evolution in plants, *Current Opinion in Plant Biology*, 82, 135-141, <https://doi.org/10.1016/j.pbi.2005.01.001>
- Ågren, J. A., Huang, H.-R., & Wright, S. I., (2016), Transposable element evolution in the allotetraploid capsella bursa-pastoris, *American Journal of Botany*, 1037, 1197-1202, <https://doi.org/10.3732/ajb.1600103>
- Alger, E. I., & Edger, P. P., (2020), One subgenome to rule them all : underlying mechanisms of subgenome dominance, *Current Opinion in Plant Biology*, 54, 108-113, <https://doi.org/10.1016/j.pbi.2020.03.004>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J., (1990), Basic local alignment search tool, *Journal of Molecular Biology*, 2153, 403-410, [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- An, Z., Tang, Z., Ma, B., Mason, A. S., Guo, Y., Yin, J., Gao, C., Wei, L., Li, J., & Fu, D., (2014), Transposon variation by order during allopolyploidisation between brassica oleracea and brassica rapa, *Plant Biology*, 164, 825-835, <https://doi.org/10.1111/plb.12121>
- Anaconda software distribution*, (2020), Anaconda Inc.
- Anastasiadi, D., Esteve-Codina, A., & Piferrer, F., (2018), Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species, *Epigenetics & Chromatin*, 111, 37, <https://doi.org/10.1186/s13072-018-0205-1>
- Andrews, S., et al., (2010), FastQC : a quality control tool for high throughput sequence data.
- Arechederra, M., Daian, F., Yim, A., Bazai, S. K., Richelme, S., Dono, R., Saurin, A. J., Habermann, B. H., & Maina, F., (2018), Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer, *Nature Communications*, 91, 3164, <https://doi.org/10.1038/s41467-018-05550-5>
- Avram, A., (2013), Docker : Automated and consistent software deployments, *InfoQ*. Retrieved, 08-09.
- Baduel, P., Quadrana, L., Hunter, B., Bomblies, K., & Colot, V., (2019), Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation, *Nature Communications*, 101, 5818, <https://doi.org/10.1038/s41467-019-13730-0>



- 
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A., (2012), NCBI GEO : archive for functional genomics data sets—update, *Nucleic Acids Research*, *41*, D991-D995, <https://doi.org/10.1093/nar/gks1193>
- Bártová, E., Krejčí, J., Harnicarová, A., Galiová, G., & Kozubek, S., (2008), Histone modifications and nuclear architecture : a review, *The Journal of Histochemistry and Cytochemistry : Official Journal of the Histochemistry Society*, *56* 8, 711-721, <https://doi.org/10.1369/jhc.2008.951251>
- Ben-David, S., Yaakov, B., & Kashkush, K., (2013), Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat, *The Plant Journal : For Cell and Molecular Biology*, *76* 2, 201-210, <https://doi.org/10.1111/tpj.12285>
- Bennetzen, J. L., & Kellogg, E. A., (1997), Do Plants Have a One-Way Ticket to Genomic Obesity?, *The Plant Cell*, *9* 9, 1509-1514, <https://doi.org/10.1105/tpc.9.9.1509>
- Bennetzen, J. L., & Wang, H., (2014), The contributions of transposable elements to the structure, function, and evolution of plant genomes, *Annual Review of Plant Biology*, *65* 1, 505-530, <https://doi.org/10.1146/annurev-arplant-050213-035811>
- Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G. H., ... Guiguen, Y., (2014), The Rainbow Trout Genome Provides Novel Insights into Evolution after Whole-Genome Duplication in Vertebrates, *Nat Commun*, *5* 1, 3657, <https://doi.org/10.1038/ncomms4657>
- Bewick, A. J., & Schmitz, R. J., (2017), Gene body DNA methylation in plants, *Current opinion in plant biology*, *36*, 103-110, <https://doi.org/10.1016/j.pbi.2016.12.007>
- Bird, A., (2002), DNA methylation patterns and epigenetic memory, *Genes & development*, *16* 1, 6-21, <https://doi.org/10.1101/gad.947102>
- Bird, K. A., Niederhuth, C. E., Ou, S., Gehan, M., Pires, J. C., Xiong, Z., VanBuren, R., & Edger, P. P., (2021), Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid brassica napus, *New Phytologist*, *230* 1, 354-371, <https://doi.org/10.1111/nph.17137>

- 
- Blanc, G., & Wolfe, K. H., (2004), Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes[W], *The Plant Cell*, *16* 7, 1667-1678, <https://doi.org/10.1105/tpc.021345>
- Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R., (1985), Ty elements transpose through an RNA intermediate, *Cell*, *40* 3, 491-500, [https://doi.org/10.1016/0092-8674\(85\)90197-7](https://doi.org/10.1016/0092-8674(85)90197-7)
- Bofkin, L., & Goldman, N., (2007), Variation in Evolutionary Processes at Different Codon Positions, *Molecular Biology and Evolution*, *24* 2, 513-521, <https://doi.org/10.1093/molbev/msl178>
- Bolger, A. M., Lohse, M., & Usadel, B., (2014), Trimmomatic : a flexible trimmer for Illumina sequence data, *Bioinformatics*, *30* 15, 2114-2120, <https://doi.org/10.1093/bioinformatics/btu170>
- Bostock, M., (s. d.), *D3.js - Data-Driven Documents*. Récupérée 7 septembre 2022, à partir de <https://d3js.org/>
- Britten, R. J., & Kohne, D. E., (1968), Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms, *Science (New York, N.Y.)*, *161* 3841, 529-540, <https://doi.org/10.1126/science.161.3841.529>
- Brown, P. O., Bowerman, B., Varmus, H. E., & Bishop, J. M., (1987), Correct integration of retroviral DNA in vitro, *Cell*, *49* 3, 347-356, [https://doi.org/10.1016/0092-8674\(87\)90287-X](https://doi.org/10.1016/0092-8674(87)90287-X)
- Bruggmann, R., Bharti, A. K., Gundlach, H., Lai, J., Young, S., Pontaroli, A. C., Wei, F., Haberer, G., Fuks, G., Du, C., Raymond, C., Estep, M. C., Liu, R., Bennetzen, J. L., Chan, A. P., Rabinowicz, P. D., Quackenbush, J., Barbazuk, W. B., Wing, R. A., ... Messing, J., (2006), Uneven chromosome contraction and expansion in the maize genome, *Genome Research*, *16* 10, 1241-1251, <https://doi.org/10.1101/gr.5338906>
- Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elisk, C. G., Lewis, S. E., Stein, L., & Holmes, I. H., (2016), JBrowse : a dynamic web platform for genome visualization and analysis, *Genome Biology*, *17* 1, 66, <https://doi.org/10.1186/s13059-016-0924-1>
- Burdon, J. J., & Marshall, D. R., (1981), Inter- and Intra-Specific Diversity in the Disease-Response of Glycine Species to the Leaf-Rust Fungus *Phakopsora Pachyrhizi*, *Journal of Ecology*, *69* 2, 381-390, <https://doi.org/10.2307/2259674>

- 
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L., (2009), BLAST+ : Architecture and Applications, *BMC Bioinformatics*, *10*, 421, <https://doi.org/10.1186/1471-2105-10-421>
- Cao, X., & Jacobsen, S. E., (2002), Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes, *Proceedings of the National Academy of Sciences*, *99*, 16491-16498, <https://doi.org/10.1073/pnas.162371599>
- Casacuberta, E., & González, J., (2013), The impact of transposable elements in environmental adaptation, *Molecular Ecology*, *22* 6, 1503-1517, <https://doi.org/10.1111/mec.12170>
- Chacon, S., & Straub, B., (2014), *Pro git*, Apress.
- Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Da Silva, C., Just, J., Falentin, C., Koh, C. S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., ... Wincker, P., (2014), Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome, *Science (New York, N.Y.)*, *345* 6199, 950-953, <https://doi.org/10.1126/science.1253435>
- Chandra, R. V., & Varanasi, B. S., (2015), *Python requests essentials*, Packt Publishing Ltd.
- Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.-F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P., & Chalhoub, B., (2005), Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (*Triticum* and *Aegilops*), *The Plant Cell*, *17* 4, 1033-1045, <https://doi.org/10.1105/tpc.104.029181>
- Chen, M., Ha, M., Lackey, E., Wang, J., & Chen, Z. J., (2008), RNAi of met1 Reduces DNA Methylation and Induces Genome-Specific Changes in Gene Expression and Centromeric Small RNA Accumulation in Arabidopsis Allopolyploids, *Genetics*, *178* 4, 1845-1858, <https://doi.org/10.1534/genetics.107.086272>
- Chen, X., Ge, X., Wang, J., Tan, C., King, G. J., & Liu, K., (2015), Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in *Brassica rapa* suggests that epigenetic modifications play a key role in polyploid genome evolution, *Frontiers in plant science*, *6*, 836, <https://doi.org/10.3389/fpls.2015.00836>

- 
- Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M. R., Liang, J., Cai, C., Freeling, M., & Wang, X., (2016), Epigenetic regulation of subgenome dominance following whole genome triplication in brassica rapa, *New Phytologist*, *211* 1, 288-299, <https://doi.org/https://doi.org/10.1111/nph.13884>
- Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X., (2018), Gene retention, fractionation and subgenome differences in polyploid plants, *Nature Plants*, *4* 5, 258-268, <https://doi.org/10.1038/s41477-018-0136-7>
- Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., & Wang, X., (2012), Biased gene fractionation and dominant gene expression among the subgenomes of brassica rapa (S.-H. Shiu, Éd.), *PLoS ONE*, *7* 5, e36442, <https://doi.org/10.1371/journal.pone.0036442>
- Choi, J. Y., & Purugganan, M. D., (2018), Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice, *Molecular Biology and Evolution*, *35* 2, 365-382, <https://doi.org/10.1093/molbev/msx284>
- Choi, J., Lyons, D. B., Kim, M. Y., Moore, J. D., & Zilberman, D., (2020), DNA methylation and histone h1 jointly repress transposable elements and aberrant intragenic transcripts, *Molecular Cell*, *77* 2, 310-323.e7, <https://doi.org/10.1016/j.molcel.2019.10.011>
- Choudhary, S., (2019), pysradb : A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive, *F1000Research*, *8*, 532, <https://doi.org/10.12688/f1000research.18676.1>
- Chuong, E. B., Elde, N. C., & Feschotte, C., (2017), Regulatory activities of transposable elements : from conflicts to benefits, *Nature reviews. Genetics*, *18* 2, 71-86, <https://doi.org/10.1038/nrg.2016.139>
- Clark, J. W., & Donoghue, P. C., (2018), Whole-Genome Duplication and Plant Macroevolution, *Trends in Plant Science*, *23* 10, 933-945, <https://doi.org/10.1016/j.tplants.2018.07.006>
- Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., & Jacobsen, S. E., (2008), Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning, *Nature*, *452* 7184, 215-219, <https://doi.org/10.1038/nature06745>
- Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., Wisecaver, J. H., Yocca, A. E., Alger, E. I., Tang, H., Xiong, Z., Callow, P., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., Shiue, L., Song, G.-q., Childs, K. L., . . . Edger, P. P., (2019),

- 
- Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry, *GigaScience*, 83, <https://doi.org/10.1093/gigascience/giz012>
- Collins, D. W., & Jukes, T. H., (1994), Rates of transition and transversion in coding sequences since the human-rodent divergence, *Genomics*, 203, 386-396, <https://doi.org/10.1006/geno.1994.1192>
- Comai, L., Tyagi, A. P., Winter, K., Holmes-Davis, R., Reynolds, S. H., Stevens, Y., & Byers, B., (2000), Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids, *The Plant Cell*, 129, 1551-1568.
- Community, C.-F., (2015), The conda-forge Project : Community-based Software Distribution Built on the conda Package Format and Ecosystem, <https://doi.org/10.5281/ZENODO.4774216>
- Contreras, B., Vives, C., Castells, R., & Casacuberta, J. M., (2015), The impact of transposable elements in the evolution of plant genomes : from selfish elements to key players, In P. Pontarotti (Éd.), *Evolutionary biology : biodiversification from genotype to phenotype* (p. 93-105), Springer International Publishing, [https://doi.org/10.1007/978-3-319-19932-0\\_6](https://doi.org/10.1007/978-3-319-19932-0_6)
- Cornille, A., Antolín, F., Garcia, E., Vernesi, C., Fietta, A., Brinkkemper, O., Kirleis, W., Schlumbaum, A., & Roldán-Ruiz, I., (2019), A multifaceted overview of apple tree domestication, *Trends in Plant Science*, 248, 770-782, <https://doi.org/10.1016/j.tplants.2019.05.007>
- Cornille, A., Giraud, T., Smulders, M. J. M., Roldán-Ruiz, I., & Gladieux, P., (2014), The domestication and evolutionary ecology of apples, *Trends in genetics : TIG*, 302, 57-65, <https://doi.org/10.1016/j.tig.2013.10.002>
- Cornille, A., Gladieux, P., Smulders, M. J. M., Roldán-Ruiz, I., Laurens, F., Cam, B. L., Nersesyan, A., Clavel, J., Olonova, M., Feugey, L., Gabrielyan, I., Zhang, X.-G., Tenailon, M. I., & Giraud, T., (2012), New insight into the history of domesticated apple : secondary contribution of the european wild apple to the genome of cultivated varieties, *PLOS Genetics*, 85, e1002703, <https://doi.org/10.1371/journal.pgen.1002703>
- Correa, M., Lerat, E., Birmelé, E., Samson, F., Bouillon, B., Normand, K., & Rizzon, C., (2021), The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality, *Genome Biology and Evolution*, 135, <https://doi.org/10.1093/gbe/evab062>

- 
- Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisine, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., ... Bucher, E., (2017), High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development, *Nat Genet*, *49*, 1099-1106, <https://doi.org/10.1038/ng.3886>
- Davies, T. J., Wolkovich, E. M., Kraft, N. J. B., Salamin, N., Allen, J. M., Ault, T. R., Betancourt, J. L., Bolmgren, K., Cleland, E. E., Cook, B. I., Crimmins, T. M., Mazer, S. J., McCabe, G. J., Pau, S., Regetz, J., Schwartz, M. D., & Travers, S. E., (2013), Phylogenetic conservatism in plant phenology, *Journal of Ecology*, *101*, 1520-1530, <https://doi.org/10.1111/1365-2745.12154>
- Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H., (2003), Mobile elements and mammalian genome evolution, *Current Opinion in Genetics & Development*, *13*, 651-658, <https://doi.org/10.1016/j.gde.2003.10.013>
- Dequeiroz, A., & Gatesy, J., (2007), The supermatrix approach to systematics, *Trends in Ecology & Evolution*, *22*, 34-41, <https://doi.org/10.1016/j.tree.2006.10.002>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R., (2013), STAR : ultrafast universal RNA-seq aligner, *Bioinformatics (Oxford, England)*, *29*, 15-21, <https://doi.org/10.1093/bioinformatics/bts635>
- Douglas, G. M., Gos, G., Steige, K. A., Salcedo, A., Holm, K., Josephs, E. B., Arunkumar, R., Ågren, J. A., Hazzouri, K. M., Wang, W., Platts, A. E., Williamson, R. J., Neuffer, B., Lascoux, M., Slotte, T., & Wright, S. I., (2015), Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*, *Proceedings of the National Academy of Sciences*, *112*, 2806-2811, <https://doi.org/10.1073/pnas.1412277112>
- Drillon, G., Carbone, A., & Fischer, G., (2014), SynChro : A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes (C. Fairhead, Éd.), *PLoS ONE*, *9*, e92621, <https://doi.org/10.1371/journal.pone.0092621>
- Du, P., Kibbe, W. A., & Lin, S. M., (2006), Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, *Bioinformatics*, *22*, 2059-2065, <https://doi.org/10.1093/bioinformatics/btl355>

- 
- Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N., & dePamphilis, C. W., (2006), Expression Pattern Shifts Following Duplication Indicative of Subfunctionalization and Neofunctionalization in Regulatory Genes of Arabidopsis, *Molecular Biology and Evolution*, *23*2, 469-478, <https://doi.org/10.1093/molbev/msj051>
- Eckardt, N. A., (2014), Genome Dominance and Interaction at the Gene Expression Level in Allohexaploid Wheat, *The Plant Cell*, *26*5, 1834, <https://doi.org/10.1105/tpc.114.127183>
- Edgar, R. C., (2004), MUSCLE : multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, *32*5, 1792-1797, <https://doi.org/10.1093/nar/gkh340>
- Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D. L., Wai, C. M., Alger, E. I., Bird, K. A., Yocca, A. E., Pumphlin, N., Ou, S., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., ... Knapp, S. J., (2019), Origin and Evolution of the Octoploid Strawberry Genome, *Nat Genet*, *51*3, 541-547, <https://doi.org/10.1038/s41588-019-0356-4>
- Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., Bowman, M. J., Childs, K. L., Washburn, J. D., Schmitz, R. J., Smith, G. D., Pires, J. C., & Puzey, J. R., (2017), Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower, *The Plant Cell*, *29*9, 2150-2167, <https://doi.org/10.1105/tpc.17.00010>
- Eichten, S. R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P. J., Waters, A. J., Starr, E., West, P. T., Tiffin, P., Myers, C. L., Vaughn, M. W., & Springer, N. M., (2013), Epigenetic and Genetic Influences on DNA Methylation Variation in Maize Populations, *The Plant Cell*, *25*8, 2783-2797, <https://doi.org/10.1105/tpc.113.114793>
- Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C.-T., Gent, J. I., Guo, L., McGinnis, K. M., Zhang, X., Schnable, P. S., Vaughn, M. W., Dawe, R. K., & Springer, N. M., (2012), Spreading of heterochromatin is limited to specific families of maize retrotransposons, *PLOS Genetics*, *8*12, e1003127, <https://doi.org/10.1371/journal.pgen.1003127>

- 
- Estep, M. C., DeBarry, J. D., & Bennetzen, J. L., (2013), The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution, *Heredity*, *110*2, 194-204, <https://doi.org/10.1038/hdy.2012.99>
- Ewels, P., Magnusson, M., Lundin, S., & Källér, M., (2016), MultiQC : summarize analysis results for multiple tools and samples in a single report, *Bioinformatics*, *32*19, 3047-3048, <https://doi.org/10.1093/bioinformatics/btw354>
- Feldman, M., Levy, A. A., Fahima, T., & Korol, A., (2012), Genomic asymmetry in allopolyploid plants : wheat as a model, *Journal of Experimental Botany*, *63*14, 5045-5059, <https://doi.org/10.1093/jxb/ers192>
- Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., & Jacobsen, S. E., (2010), Conservation and divergence of methylation patterning in plants and animals, *Proceedings of the National Academy of Sciences*, *107*19, 8689-8694, <https://doi.org/10.1073/pnas.1002720107>
- Feschotte, C., Jiang, N., & Wessler, S. R., (2002), Plant transposable elements : where genetics meets genomics, *Nature Reviews Genetics*, *3*5, 329-341, <https://doi.org/10.1038/nrg793>
- Finnegan, E., & Kovac, K., (2000), Plant DNA methyltransferases, *Plant Gene Silencing*, 69-81, [https://doi.org/10.1007/978-94-011-4183-3\\_5](https://doi.org/10.1007/978-94-011-4183-3_5)
- Fisher, R. A., (1936), Design of Experiments, *British Medical Journal*, *1*3923, 554.
- Fisher, R. A., (1970), *Statistical methods for research workers* (14th ed., revised and enlarged), Oliver ; Boyd.
- Flagel, L., Udall, J., Nettleton, D., & Wendel, J., (2008), Duplicate gene expression in allopolyploid gossypium reveals two temporally distinct phases of expression evolution, *BMC Biology*, *6*1, 16, <https://doi.org/10.1186/1741-7007-6-16>
- Flagel, L. E., & Wendel, J. F., (2010), Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation, *The New Phytologist*, *186*1, 184-193, <https://doi.org/10.1111/j.1469-8137.2009.03107.x>
- Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H., (2011), Considering transposable element diversification in de novo annotation approaches (Y. Xu, Éd.), *PLoS ONE*, *6*1, e16526, <https://doi.org/10.1371/journal.pone.0016526>
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J., (1999), Preservation of duplicate genes by complementary, degenerative mutations, *Genetics*, *151*4, 1531-1545, <https://doi.org/10.1093/genetics/151.4.1531>



- 
- Freeling, M., Scanlon, M. J., & Fowler, J. E., (2015), Fractionation and subfunctionalization following genome duplications : mechanisms that drive gene content and their consequences, *Current Opinion in Genetics & Development*, *35*, 110-118, <https://doi.org/10.1016/j.gde.2015.11.002>
- Freeling, M., & Thomas, B. C., (2006), Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity, *Genome Research*, *16* 7, 805-814, <https://doi.org/10.1101/gr.3681406>
- Freeling, M., Woodhouse, M. R., Subramaniam, S., Turco, G., Lisch, D., & Schnable, J. C., (2012), Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants, *Genome studies molecular genetics*, *15* 2, 131-139, <https://doi.org/10.1016/j.pbi.2012.01.015>
- Gao, C., Zhou, G., Ma, C., Zhai, W., Zhang, T., Liu, Z., Yang, Y., Wu, M., Yue, Y., Duan, Z., Li, Y., Li, B., Li, J., Shen, J., Tu, J., & Fu, T., (2016), Helitron-like transposons contributed to the mating system transition from out-crossing to self-fertilizing in polyploid *Brassica napus* L, *Scientific Reports*, *6*, 33785, <https://doi.org/10.1038/srep33785>
- Gardiner, L.-J., Joynson, R., & Hall, A., (2019), Next-generation sequencing enabled genetics in hexaploid wheat. In *Applications of genetic and genomic research in cereals* (p. 49-63), Elsevier, <https://doi.org/10.1016/B978-0-08-102163-7.00003-X>
- Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., & Freeling, M., (2014), Two Evolutionarily Distinct Classes of Paleopolyploidy, *Molecular Biology and Evolution*, *31* 2, 448-454, <https://doi.org/10.1093/molbev/mst230>
- Gladieux, P., Zhang, X.-G., Róldan-Ruiz, I., Caffier, V., Leroy, T., Devaux, M., Van Glabeke, S., Coart, E., & Le Cam, B., (2010), Evolution of the population structure of *venturia inaequalis*, the apple scab fungus, associated with the domestication of its host, *Molecular Ecology*, *19* 4, 658-674, <https://doi.org/10.1111/j.1365-294X.2009.04498.x>
- Grabundzija, I., Messing, S. A., Thomas, J., Cosby, R. L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E. J., Dyda, F., Izsvák, Z., & Ivics, Z., (2016), A helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes, *Nature Communications*, *7* 1, 10716, <https://doi.org/10.1038/ncomms10716>

- 
- Greenblatt, I. M., & Alexander Brink, R., (1963), Transpositions of modulator in maize into divided and undivided chromosome segments, *Nature*, *197*4865, 412-413, <https://doi.org/10.1038/197412a0>
- Haas, B. J., Delcher, A. L., Wortman, J. R., & Salzberg, S. L., (2004), DAGChainer : A tool for mining segmental genome duplications and synteny, *Bioinformatics*, *20*18, 3643-3646, <https://doi.org/10.1093/bioinformatics/bth397>
- Harikrishnan, S. L., Pucholt, P., & Berlin, S., (2015), Sequence and gene expression evolution of paralogous genes in willows, *Scientific Reports*, *5*1, 18662, <https://doi.org/10.1038/srep18662>
- Harper, A. L., Trick, M., He, Z., Clissold, L., Fellgett, A., Griffiths, S., & Bancroft, I., (2016), Genome distribution of differential homoeologue contributions to leaf gene expression in bread wheat, *Plant Biotechnology Journal*, *14* 5, 1207-1214, <https://doi.org/10.1111/pbi.12486>
- Hasegawa, M., Kishino, H., & Yano, T.-a., (1985), Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, *22*2, 160-174, <https://doi.org/10.1007/BF02101694>
- Hatira, S., (2022, juillet 15), *SkanderHatira/bisepts : Release 10* (Version 10), Zenodo, <https://doi.org/10.5281/ZENODO.6839274>
- Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., & Lyons, E., (2017), SynMap2 and SynMap3d : web-based whole-genome synteny browsers (J. Hancock, Éd.), *Bioinformatics*, *33* 14, 2197-2198, <https://doi.org/10.1093/bioinformatics/btx144>
- Heard, N. A., & Rubin-Delanchy, P., (2018), Choosing between methods of combining-values, *Biometrika*, *105* 1, 239-246, <https://doi.org/10.1093/biomet/asx076>
- Hias, N., Svara, A., & Keulemans, J. W., (2018), Effect of polyploidisation on the response of apple (*malus × domestica* borkh.) to venturia inaequalis infection, *European Journal of Plant Pathology*, *151* 2, 515-526, <https://doi.org/10.1007/s10658-017-1395-2>
- Hill, W. G., & Robertson, A., (1966), The effect of linkage on limits to artificial selection, *Genetical Research*, *8* 3, 269-294, <https://doi.org/10.1017/S0016672300010156>
- Hollister, J. D., & Gaut, B. S., (2009), Epigenetic silencing of transposable elements : a trade-off between reduced transposition and deleterious effects on neighboring gene expression, *Genome Research*, *19* 8, 1419-1428, <https://doi.org/10.1101/gr.091678.109>

- 
- Hou, J., Shi, X., Chen, C., Islam, M. S., Johnson, A. F., Kanno, T., Huettel, B., Yen, M.-R., Hsu, F.-M., Ji, T., Chen, P.-Y., Matzke, M., Matzke, A. J. M., Cheng, J., & Birchler, J. A., (2018), Global impacts of chromosomal imbalance on gene expression in Arabidopsis and other taxa, *Proceedings of the National Academy of Sciences*, *115*48, E11321-E11330, <https://doi.org/10.1073/pnas.1807796115>
- Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., Edger, P. P., Pires, J. C., Tan, D.-Y., Zhong, Y., & Ma, H., (2016), Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution, *Molecular Biology and Evolution*, *33*2, 394-412, <https://doi.org/10.1093/molbev/msv226>
- Huang, J., Gao, Y., Jia, H., Liu, L., Zhang, D., & Zhang, Z., (2015), Comparative transcriptomics uncovers alternative splicing changes and signatures of selection from maize improvement, *BMC Genomics*, *16*1, 363, <https://doi.org/10.1186/s12864-015-1582-5>
- Hughes, A. L., Friedman, R., Ekollu, V., & Rose, J. R., (2003), Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana, *Molecular Phylogenetics and Evolution*, *29*3, 410-416, [https://doi.org/10.1016/S1055-7903\(03\)00262-8](https://doi.org/10.1016/S1055-7903(03)00262-8)
- Hughes, T. E., Langdale, J. A., & Kelly, S., (2014), The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize, *Genome Research*, *24*8, 1348-1355, <https://doi.org/10.1101/gr.172684.114>
- Hunter, J. D., (2007), Matplotlib : A 2D Graphics Environment, *Computing in Science Engineering*, *9*3, 90-95, <https://doi.org/10.1109/MCSE.2007.55>
- Hurst, L. D., (2002), The ka/ks ratio : diagnosing the form of sequence evolution, *Trends in Genetics*, *18*9, 486-487, [https://doi.org/10.1016/S0168-9525\(02\)02722-1](https://doi.org/10.1016/S0168-9525(02)02722-1)
- Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A. J., Juárez, M. J. A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S. C., Himmelbauer, H., Minoche, A. E., Xu, S., Lynch, M., ... Herrera-Estrella, L., (2013), Architecture and evolution of a minute plant genome, *Nature*, *498*7452, 94-98, <https://doi.org/10.1038/nature12132>
- Inc., P. T., (2015), *Collaborative data science*. <https://plot.ly>

- 
- Ito, H., & Kakutani, T., (2014), Control of transposable elements in arabidopsis thaliana, *Chromosome Research*, 222, 217-223, <https://doi.org/10.1007/s10577-014-9417-9>
- Jenuwein, T., & Allis, C. D., (2001), Translating the histone code, *Science*, 2935532, 1074-1080, <https://doi.org/10.1126/science.1063127>
- Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., ... Ware, D., (2017), Improved maize reference genome with single-molecule technologies, *Nature*, 5467659, 524-527, <https://doi.org/10.1038/nature22971>
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., & dePamphilis, C. W., (2011), Ancestral polyploidy in seed plants and angiosperms, *Nature*, 4737345, 97-100, <https://doi.org/10.1038/nature09916>
- Jin, P., Qin, S., Chen, X., Song, Y., Li-Ling, J., Xu, X., & Ma, F., (2012), Evolutionary rate of human tissue-specific genes are related with transposable element insertions, *Genetica*, 14010, 513-523, <https://doi.org/10.1007/s10709-013-9700-2>
- Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., & Jordan, I. K., (2012), On the presence and role of human gene-body DNA methylation, *Oncotarget*, 34, 462-474, <https://doi.org/10.18632/oncotarget.497>
- Joyce, B. L., Haug-Baltzell, A., Davey, S., Bomhoff, M., Schnable, J. C., & Lyons, E., (2017), FractBias : a graphical tool for assessing fractionation bias following polyploidy, *Bioinformatics*, 334, 552-554, <https://doi.org/10.1093/bioinformatics/btw666>
- Jukes, T., & Cantor, C., (1969), Evolution of protein molecules. In 'Mammalian protein Metabolism'.(Ed. HN munro.) pp. 21–132, *Academic Press, New York*, 1, 504-511.
- Jukes, T. H., Cantor, C. R., Munro, H., et al., (1969), Mammalian protein metabolism.
- Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S. P., Gasic, K., Scott, K., Frank, M., Ru, S., Hough, H., Evans, K., Peace, C., Olmstead, M., DeVetter, L. W., McFerson, J., Coe, M., ... Main, D., (2018), 15 years of GDR : New data and functionality in the Genome Database for Rosaceae, *Nucleic Acids Research*, 47, D1137-D1145, <https://doi.org/10.1093/nar/gky1000>
- Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddloh, J. A., Riddle, N. C., Verbsky, M. L., & Richards, E. J., (2003), Arabidopsis MET1

- 
- Cytosine Methyltransferase Mutants, *Genetics*, 1633, 1109-1122, <https://doi.org/10.1093/genetics/163.3.1109>
- Kashkush, K., Feldman, M., & Levy, A. A., (2003), Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat, *Nature Genetics*, 331, 102-106, <https://doi.org/10.1038/ng1063>
- Kidwell, M. G., & Lisch, D., (1997), Transposable elements as sources of variation in animals and plants, *Proceedings of the National Academy of Sciences*, 94 15, 7704-7711, <https://doi.org/10.1073/pnas.94.15.7704>
- Kielbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C., (2011), Adaptive seeds tame genomic sequence comparison, *Genome Research*, 21 3, 487-493, <https://doi.org/10.1101/gr.113985.110>
- Kimura, M., (1980), A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution*, 16 2, 111-120, <https://doi.org/10.1007/BF01731581>
- Kimura, M., (1981), Estimation of evolutionary distances between homologous nucleotide sequences., *Proceedings of the National Academy of Sciences*, 78 1, 454-458, <https://doi.org/10.1073/pnas.78.1.454>
- Kimura, M., (1968), Evolutionary rate at the molecular level, *Nature*, 217 5129, 624-626, <https://doi.org/10.1038/217624a0>
- Kincaid, W. M., (1962), The Combination of Tests Based on Discrete Distributions, *Journal of the American Statistical Association*, 57 297, 10-19, <https://doi.org/10.1080/01621459.1962.10482147>
- King, K. C., Seppälä, O., & Neiman, M., (2012), Is more better? Polyploidy and parasite resistance, *Biology Letters*, 8 4, 598-600, <https://doi.org/10.1098/rsbl.2011.1152>
- Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W. J., Koornneef, M., & Kikutani, T., (2007), Control of FWA gene silencing in arabidopsis thaliana by SINE-related direct repeats, *The Plant Journal*, 49 1, 38-45, <https://doi.org/10.1111/j.1365-313X.2006.02936.x>
- Koenen, E. J. M., Ojeda, D. I., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., Pennington, R. T., Herendeen, P. S., Bruneau, A., & Hughes, C. E., (2021), The Origin of the Legumes is a Complex Paleopolyploid Phylogenomic Tangle Closely Associated with the Cretaceous–Paleogene (K–Pg) Mass Extinction Event, *Systematic Biology*, 70 3, 508-526, <https://doi.org/10.1093/sysbio/syaa041>

- 
- Köster, J., & Rahmann, S., (2012), Snakemake—a scalable bioinformatics workflow engine, *Bioinformatics*, *28* 19, 2520-2522, <https://doi.org/10.1093/bioinformatics/bts480>
- Kreplak, J., Madoui, M.-A., Cápál, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., d'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., . . . Burstin, J., (2019), A reference genome for pea provides insight into legume genome evolution, *Nature Genetics*, *51* 9, 1411-1422, <https://doi.org/10.1038/s41588-019-0480-1>
- Krueger, F., & Andrews, S. R., (2011), Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics (Oxford, England)*, *27* 11, 1571-1572, <https://doi.org/10.1093/bioinformatics/btr167>
- Kruskal, W. H., (1958), Ordinal measures of association, *Journal of the American Statistical Association*, *53* 284, 814-861, <https://doi.org/10.1080/01621459.1958.10501481>
- Kryazhimskiy, S., & Plotkin, J. B., (2008), The population genetics of dN/dS (T. Gojobori, Éd.), *PLoS Genetics*, *4* 12, e1000304, <https://doi.org/10.1371/journal.pgen.1000304>
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A., (2009), Circos : An Information Aesthetic for Comparative Genomics, *Genome Research*, *19* 9, 1639-1645, <https://doi.org/10.1101/gr.092759.109>
- Kulis, M., Heath, S., Bibikova, M., Queirós, A. C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., Barberán-Soler, S., Papasaikas, P., Jares, P., Beà, S., Rico, D., Ecker, S., Rubio, M., Royo, R., Ho, V., . . . Martín-Subero, J. I., (2012), Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia, *Nature Genetics*, *44* 11, 1236-1242, <https://doi.org/10.1038/ng.2443>
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B., (2017), TimeTree : a resource for timelines, timetrees, and divergence times, *Molecular Biology and Evolution*, *34* 7, 1812-1819, <https://doi.org/10.1093/molbev/msx116>
- Lallemand, T., Aubourg, S., Hunault, G., Celton, J.-M., & Landès, C., (2020), Evolution of ohnologous chromosomes following Whole Genome Duplication in apple.
- Lallemand, T., Leduc, M., Landès, C., Rizzon, C., & Lerat, E., (2020), An overview of duplicated gene detection methods : why the duplication mechanism has to be accounted for in their choice, *Genes*, *11* 9, 1046, <https://doi.org/10.3390/genes11091046>

- 
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., . . . International Human Genome Sequencing Consortium, (2001), Initial sequencing and analysis of the human genome, *Nature*, *409* 6822, 860-921, <https://doi.org/10.1038/35057062>
- Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., & Freeling, M., (2004), Genomic duplication, fractionation and the origin of regulatory novelty., *Genetics*, *166* 2, 935-945.
- Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrand-Saint-Oyant, L., Poncet, C., Lasserre-Zuber, P., Feugey, L., & Durel, C.-E., (2016), Genetic diversity, population structure, parentage analysis, and construction of core collections in the french apple germplasm based on SSR markers, *Plant Molecular Biology Reporter*, *34* 4, 827-844, <https://doi.org/10.1007/s11105-015-0966-7>
- Leinonen, R., Sugawara, H., & Shumway, M., (2011), The Sequence Read Archive, *Nucleic Acids Research*, *39*, D19-D21, <https://doi.org/10.1093/nar/gkq1019>
- Leroy, T., Lemaire, C., Dunemann, F., & Le Cam, B., (2013), The genetic structure of a *Venturia inaequalis* population in a heterogeneous host population composed of different Malus species, *BMC Evolutionary Biology*, *13* 1, 64, <https://doi.org/10.1186/1471-2148-13-64>
- Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., et al., (2014), mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat, *The Plant Cell*, *26* 5, 1878-1900, <https://doi.org/10.1105/tpc.114.124388>
- Li, J.-T., Hou, G.-Y., Kong, X.-F., Li, C.-Y., Zeng, J.-M., Li, H.-D., Xiao, G.-B., Li, X.-M., & Sun, X.-W., (2015), The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp ( *Cyprinus carpio* ), *Scientific Reports*, *5* 1, 8199, <https://doi.org/10.1038/srep08199>
- Li, L., Wang, X., Stolt, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J., & Deng, X. W., (2006), Genome-wide transcription analyses in rice using tiling microarrays, *Nature Genetics*, *38* 1, 124-129, <https://doi.org/10.1038/ng1704>
- Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K., Li, L., & Zhang, S., (2019), Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.), *Horticulture Research*, *6*, <https://doi.org/10.1038/s41438-018-0110-6>

- 
- Li, S., Zhang, J., Huang, S., & He, X., (2018), Genome-wide analysis reveals that exon methylation facilitates its selective usage in the human transcriptome, *Briefings in Bioinformatics*, *19* 5, 754-764, <https://doi.org/10.1093/bib/bbx019>
- Li, W. H., Wu, C. I., & Luo, C. C., (1985), A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes., *Molecular Biology and Evolution*, *2* 2, 150-174, <https://doi.org/10.1093/oxfordjournals.molbev.a040343>
- Li, Y., Li, C., Xia, J., & Jin, Y., (2011), Domestication of transposable elements into MicroRNA genes in plants, *PLOS ONE*, *6* 5, e19212, <https://doi.org/10.1371/journal.pone.0019212>
- Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., & Gao, X., (2019), Deep Learning in Bioinformatics : Introduction, Application, and Perspective in the Big Data Era, *Methods*, *166*, 4-21, <https://doi.org/10.1016/j.ymeth.2019.04.008>
- Liang, L., Chang, Y., Lu, J., Wu, X., Liu, Q., Zhang, W., Su, X., & Zhang, B., (2019), Global Methyloomic and Transcriptomic Analyses Reveal the Broad Participation of DNA Methylation in Daily Gene Expression Regulation of *Populus trichocarpa*, *Frontiers in Plant Science*, *10*, <https://doi.org/10.3389/fpls.2019.00243>
- Liang, Z., & Schnable, J. C., (2018), Functional divergence between subgenomes and gene pairs after whole genome duplications, *Molecular Plant*, *11* 3, 388-397, <https://doi.org/10.1016/j.molp.2017.12.010>
- Lim, K. Y., Kovarik, A., Matyasek, R., Chase, M. W., Clarkson, J. J., Grandbastien, M. A., & Leitch, A. R., (2007), Sequence of events leading to near-complete genome turnover in allopolyploid *nicotiana* within five million years, *New Phytologist*, *175* 4, 756-763, <https://doi.org/10.1111/j.1469-8137.2007.02121.x>
- Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M. W., Dedhia, N., Richard McCombie, W., Lavine, K., Mittal, V., May, B., Kasschau, K. D., Carrington, J. C., Doerge, R. W., Colot, V., & Martienssen, R., (2004), Role of transposable elements in heterochromatin and epigenetic control, *Nature*, *430* 6998, 471-476, <https://doi.org/10.1038/nature02651>
- Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R., (2008), Highly integrated single-base resolution maps of the epigenome in *arabidopsis*, *Cell*, *133* 3, 523-536, <https://doi.org/10.1016/j.cell.2008.03.029>



- 
- Liu, D., Hunt, M., & Tsai, I. J., (2018), Inferring synteny between genome assemblies : a systematic evaluation, *BMC Bioinformatics*, *19* 1, 26, <https://doi.org/10.1186/s12859-018-2026-4>
- Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T.-J., Hu, Q., Wang, X., Yue, Z., ... Paterson, A. H., (2014), The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes, *Nature Communications*, *5*, 3930, <https://doi.org/10.1038/ncomms4930>
- Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N., Sun, S., Zhang, L., & Wang, X., (2017), Two Highly Similar Poplar Paleo-subgenomes Suggest an Autotetraploid Ancestor of Salicaceae Plants, *Frontiers in Plant Science*, *8*, 571, <https://doi.org/10.3389/fpls.2017.00571>
- Lockton, S., & Gaut, B. S., (2005), Plant conserved non-coding sequences and paralogue evolution, *Trends in genetics : TIG*, *21* 1, 60-65, <https://doi.org/10.1016/j.tig.2004.11.013>
- Love, M. I., Huber, W., & Anders, S., (2014), Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology*, *15* 12, 550, <https://doi.org/10.1186/s13059-014-0550-8>
- Luan, D. D., Korman, M. H., Jakubczak, J. L., & Eickbush, T. H., (1993), Reverse transcription of r2bm RNA is primed by a nick at the chromosomal target site : a mechanism for non-LTR retrotransposition, *Cell*, *72* 4, 595-605, [https://doi.org/10.1016/0092-8674\(93\)90078-5](https://doi.org/10.1016/0092-8674(93)90078-5)
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Wiernik, B. M., & Makowski, D., (2022), easystats : Framework for Easy Statistical Modeling, Visualization, and Reporting, *CRAN*.
- Lukens, L. N., Pires, J. C., Leon, E., Vogelzang, R., Oslach, L., & Osborn, T., (2006), Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized Brassica napus Allopolyploids, *Plant Physiology*, *140* 1, 336-348, <https://doi.org/10.1104/pp.105.066308>
- Lynch, M., (2007), *The origins of genome architecture*, Sinauer Associates.
- Madlung, A., Masuelli, R. W., Watson, B., Reynolds, S. H., Davison, J., & Comai, L., (2002), Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids, *Plant Physiology*, *129* 2, 733-746, <https://doi.org/10.1104/pp.003095>

- 
- Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., & Van de Peer, Y., (2005), Modeling gene and genome duplications in eukaryotes, *Proceedings of the National Academy of Sciences*, *102* 15, 5454-5459, <https://doi.org/10.1073/pnas.0501102102>
- Maison, C., Bailly, D., Peters, A. H., Quivy, J.-P., Roche, D., Taddei, A., Lachner, M., Jenuwein, T., & Almouzni, G., (2002), Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component, *Nature genetics*, *30* 3, 329-334, <https://doi.org/10.1038/ng843>
- Makowski, D., Ben-Shachar, M., Patil, I., & Lüdecke, D., (2020), Methods and Algorithms for Correlation Analysis in R, *Journal of Open Source Software*, *5* 51, 2306, <https://doi.org/10.21105/joss.02306>
- Malik, H. S., & Eickbush, T. H., (2001), Phylogenetic analysis of ribonuclease h domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses, *Genome Research*, *11* 7, 1187-1197, <https://doi.org/10.1101/gr.185101>
- Mann, H. B., & Whitney, D. R., (1947), On a test of whether one of two random variables is stochastically larger than the other, *The Annals of Mathematical Statistics*, *18* 1, 50-60, <https://doi.org/10.1214/aoms/1177730491>
- Mao, Y., (2019), GenoDup pipeline : a tool to detect genome duplication using the dS-based method, *PeerJ*, *7*, e6303, <https://doi.org/10.7717/peerj.6303>
- Martin, M., (2011), Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, *17* 1, 10-12, <https://doi.org/10.14806/ej.17.1.200>
- Matzke, M. A., & Matzke, A. J. M., (1998), Polyploidy and transposons, *Trends in Ecology & Evolution*, *13* 6, 241, [https://doi.org/10.1016/S0169-5347\(98\)01390-1](https://doi.org/10.1016/S0169-5347(98)01390-1)
- Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., ... Costello, J. F., (2010), Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters, *Nature*, *466* 7303, 253-257, <https://doi.org/10.1038/nature09165>
- McCarthy, D. J., Chen, Y., & Smyth, G. K., (2012), Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, *Nucleic Acids Research*, *40* 10, 4288-4297, <https://doi.org/10.1093/nar/gks042>
- McClintock, B., (1984), The significance of responses of the genome to challenge, *Science (New York, N.Y.)*, *226* 4676, 792-801, <https://doi.org/10.1126/science.15739260>

- 
- McClintock, B., (1950), The origin and behavior of mutable loci in maize, *Proceedings of the National Academy of Sciences*, *36* 6, 344-355, <https://doi.org/10.1073/pnas.36.6.344>
- McKinney, W., (2010), Data structures for statistical computing in python, 56-61, <https://doi.org/10.25080/Majora-92bf1922-00a>
- Members of the Complex Trait Consortium, (2003), The nature and identification of quantitative trait loci : a community's view, *Nature Reviews Genetics*, *4* 11, 911-916, <https://doi.org/10.1038/nrg1206>
- Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., & Kakutani, T., (2001), Mobilization of transposons by a mutation abolishing full DNA methylation in arabidopsis, *Nature*, *411* 6834, 212-214, <https://doi.org/10.1038/35075612>
- Morgan, J., (2013), *The new book of apples*, Random House.
- Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A., (2008), Database indexing for production MegaBLAST searches, *Bioinformatics (Oxford, England)*, *24* 16, 1757-1764, <https://doi.org/10.1093/bioinformatics/btn322>
- Mosteller, F., & Fisher, R. A., (1948), Questions and answers, *The American Statistician*, *2* 5, 30-31.
- Mudholkar E.O., G. G., (1983), On the Convolution of Logistic Random Variables., *Metrica*, *30*, 1-14, <https://doi.org/10.1007/BF02056895>
- Muyle, A. M., Seymour, D. K., Lv, Y., Huettel, B., & Gaut, B. S., (2022), Gene Body Methylation in Plants : Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes, *Genome Biology and Evolution*, *14* 4, evac038, <https://doi.org/10.1093/gbe/evac038>
- Myles, C., & Wayne, M., (2008), Quantitative trait locus (QTL) analysis, *Nature Education* *1* (1), 208.
- Naumann, J., Salomo, K., Der, J. P., Wafula, E. K., Bolin, J. F., Maass, E., Frenzke, L., Samain, M.-S., Neinhuis, C., dePamphilis, C. W., & Wanke, S., (2013), Single-copy nuclear genes place haustorial Hydnoraceae within piperales and reveal a cretaceous origin of multiple parasitic angiosperm lineages, *PloS One*, *8* 11, e79204, <https://doi.org/10.1371/journal.pone.0079204>
- Nei, M., & Gojobori, T., (1986), Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions., *Molecular Biology and Evolution*, *3* 5, 418-426, <https://doi.org/10.1093/oxfordjournals.molbev.a040410>

- 
- Noshay, J. M., Anderson, S. N., Zhou, P., Ji, L., Ricci, W., Lu, Z., Stitzer, M. C., Crisp, P. A., Hirsch, C. N., Zhang, X., Schmitz, R. J., & Springer, N. M., (2019), Monitoring the interplay between transposable element families and DNA methylation in maize, *PLOS Genetics*, *15* 9, e1008291, <https://doi.org/10.1371/journal.pgen.1008291>
- .Novikova, P. Y., Hohmann, N., & Van de Peer, Y., (2018), Polyploid arabidopsis species originated around recent glaciation maxima, *Current Opinion in Plant Biology*, *42*, 8-15, <https://doi.org/10.1016/j.pbi.2018.01.005>
- Observable - Explore, analyze, and explain data. As a team.* [Observable], (s. d.). Récupérée 7 septembre 2022, à partir de <https://observablehq.com/@d3>
- Ohno, S., (1970, décembre 11), *Evolution by gene duplication*, Springer Science & Business Media.
- Oliphant, T. E., (2007), Python for Scientific Computing, *Computing in Science Engineering*, *9* 3, 10-20, <https://doi.org/10.1109/MCSE.2007.58>
- Oliveira, P. H., Ribis, J. W., Garrett, E. M., Trzilova, D., Kim, A., Sekulovic, O., Mead, E. A., Pak, T., Zhu, S., Deikus, G., Touchon, M., Lewis-Sandari, M., Beckford, C., Zeitouni, N. E., Altman, D. R., Webster, E., Oussenko, I., Bunyavanich, S., Aggarwal, A. K., . . . Fang, G., (2020), Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis, *Nature Microbiology*, *5* 1, 166-180, <https://doi.org/10.1038/s41564-019-0613-4>
- Pamilo, P., & Bianchi, N. O., (1993), Evolution of the Zfx and Zfy genes : rates and interdependence between the genes., *Molecular Biology and Evolution*, *10* 2, 271-281, <https://doi.org/10.1093/oxfordjournals.molbev.a040003>
- Panchy, N., Lehti-Shiu, M., & Shiu, S.-H., (2016), Evolution of Gene Duplication in Plants, *Plant Physiology*, *171* 4, 2294-2316, <https://doi.org/10.1104/pp.16.00523>
- Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., & Ainouche, M., (2009), Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in spartina, *New Phytologist*, *184* 4, 1003-1015, <https://doi.org/10.1111/j.1469-8137.2009.03029.x>
- Parisod, C., & Senerchia, N., (2012), Responses of transposable elements to polyploidy, In M.-A. Grandbastien & J. M. Casacuberta (Éd.), *Plant transposable elements : impact on genome structure and function* (p. 147-168), Springer, [https://doi.org/10.1007/978-3-642-31842-9\\_9](https://doi.org/10.1007/978-3-642-31842-9_9)

- 
- Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., et al., (2014), Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*, *Genome biology*, *15* 6, 1-18, <https://doi.org/10.1186/gb-2014-15-6-r77>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C., (2017), Salmon : fast and bias-aware quantification of transcript expression using dual-phase inference, *Nature methods*, *14* 4, 417-419, <https://doi.org/10.1038/nmeth.4197>
- Pearson, K., (1895), Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London Series I*, *58*, 240-242.
- Pearson, K., (1900), X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50* 302, 157-175, <https://doi.org/10.1080/14786440009463897>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E., (2011), Scikit-learn : Machine Learning in Python, *Journal of Machine Learning Research*, *12*, 2825-2830.
- Pelletier, S., (2022, avril), *spelletier-1345/anadiff : pegasus production* (Version v4.3), Zenodo, <https://doi.org/10.5281/zenodo.6477918>
- Perez, F., & Granger, B. E., (2007), IPython : a system for interactive scientific computing, *Computing in Science & Engineering*, *9* 3, 21-29, <https://doi.org/10.1109/MCSE.2007.53>
- Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, Q. T., Lim, K. Y., Kovarik, A., Leitch, A. R., Grandbastien, M.-A., & Mhiri, C., (2010), Mobilization of retrotransposons in synthetic allotetraploid tobacco, *The New Phytologist*, *186* 1, 135-147, <https://doi.org/10.1111/j.1469-8137.2009.03140.x>
- Piriyapongsa, J., & Jordan, I. K., (2008), Dual coding of siRNAs and miRNAs by plant transposable elements, *RNA*, *14* 5, 814-821, <https://doi.org/10.1261/rna.916708>
- Pontarotti, P., (2015), *Evolutionary biology : biodiversification from genotype to phenotype*, Springer.
- Pophaly, S. D., & Tellier, A., (2015), Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize, *Molecular Biology and Evolution*, *32* 12, 3226-3235, <https://doi.org/10.1093/molbev/msv191>

- 
- Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., Kerr, M., Robertson, K. R., Arsenault, M., Dickinson, T. A., & Campbell, C. S., (2007), Phylogeny and classification of rosaceae, *Plant Systematics and Evolution*, *266* 1, 5-43, <https://doi.org/10.1007/s00606-007-0539-9>
- Pouget, M., Youssef, S., Dumas, P. .-, Baumberger, T., San Roman, A., Torre, F., Affre, L., Médail, F., & Baumel, A., (2016), Spatial mismatches between plant biodiversity facets and evolutionary legacy in the vicinity of a major mediterranean city, *Ecological Indicators*, *60*, 736-745, <https://doi.org/10.1016/j.ecolind.2015.07.017>
- Proost, S., Fostier, J., Witte, D. D., Dhoedt, B., & Demeester, P., (2012), I-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets, *Nucleic Acids Research*, *40* 2, 11.
- Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H., (2019), Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants, *Genome Biology*, *20* 1, 38, <https://doi.org/10.1186/s13059-019-1650-2>
- Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabehere, D., (2005), Combined evidence annotation of transposable elements in genome sequences, *PLOS Computational Biology*, *1* 2, e22, <https://doi.org/10.1371/journal.pcbi.0010022>
- Razin, A., & Cedar, H., (1991), DNA methylation and gene expression, *Microbiological Reviews*, *55* 3, 451-458, <https://doi.org/10.1128/mr.55.3.451-458.1991>
- Rebollo, R., Karimi, M. M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., Goyal, P., Keane, T. M., Jones, S., Hirst, M., Lorincz, M. C., & Mager, D. L., (2011), Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms, *PLOS Genetics*, *7* 9, e1002301, <https://doi.org/10.1371/journal.pgen.1002301>
- Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., & Qi, J., (2018), Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms, *Molecular Plant*, *11* 3, 414-428, <https://doi.org/10.1016/j.molp.2018.01.002>
- Renny-Byfield, S., Chester, M., Kovarik, A., Le Comber, S. C., Grandbastien, M.-A., Deloger, M., Nichols, R. A., Macas, J., Novak, P., Chase, M. W., & Leitch, A. R., (2011), Next generation sequencing reveals genome downsizing in allotetraploid *nicotiana tabacum*, predominantly through the elimination of paternally derived

- 
- repetitive DNAs, *Molecular Biology and Evolution*, 28 10, 2843-2854, <https://doi.org/10.1093/molbev/msr112>
- Renny-Byfield, S., Gallagher, J. P., Grover, C. E., Szadkowski, E., Page, J. T., Udall, J. A., Wang, X., Paterson, A. H., & Wendel, J. F., (2014), Ancient Gene Duplicates in *Gossypium* (Cotton) Exhibit Near-Complete Expression Divergence, *Genome Biology and Evolution*, 6 3, 559-571, <https://doi.org/10.1093/gbe/evu037>
- Renny-Byfield, S., Gong, L., Gallagher, J. P., & Wendel, J. F., (2015), Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution, *Molecular Biology and Evolution*, 32 4, 1063-1071, <https://doi.org/10.1093/molbev/msv001>
- Renny-Byfield, S., Rodgers-Melnick, E., & Ross-Ibarra, J., (2017), Gene Fractionation and Function in the Ancient Subgenomes of Maize, *Molecular Biology and Evolution*, 34 8, 1825-1832, <https://doi.org/10.1093/molbev/msx121>
- Richards, E. J., & Elgin, S. C. R., (2002), Epigenetic codes for heterochromatin formation and silencing : rounding up the usual suspects, *Cell*, 108 4, 489-500, [https://doi.org/10.1016/S0092-8674\(02\)00644-X](https://doi.org/10.1016/S0092-8674(02)00644-X)
- Richardson, L., (2007), Beautiful soup documentation, *April*.
- Rigal, M., Becker, C., Pélissier, T., Pogorelcnik, R., Devos, J., Ikeda, Y., Weigel, D., & Mathieu, O., (2016), Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in *Arabidopsis thaliana* F1 epihybrids, *Proceedings of the National Academy of Sciences*, 113 14, E2083-E2092, <https://doi.org/10.1073/pnas.1600672113>
- Robinson, J. P., Harris, S. A., & Juniper, B. E., (2001), Taxonomy of the genus *malus* mill. (rosaceae) with emphasis on the cultivated apple, *malus domestica* borkh., *Plant Systematics and Evolution*, 226 1, 35-58, <https://doi.org/10.1007/s006060170072>
- Rocklin, M., (2015), Dask : Parallel computation with blocked algorithms and task scheduling, *Proceedings of the 14th python in science conference*, <https://doi.org/10.25080/Majora-7b98e3ed-013>
- Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., ... Colot, V., (2011), Integrative epigenomic mapping defines four main chromatin states in *arabidopsis* : organization of the *arabidopsis* epigenome, *The EMBO Journal*, 30 10, 1928-1938, <https://doi.org/10.1038/emboj.2011.103>

- 
- Samuel Yang, S., Cheung, F., Lee, J. J., Ha, M., Wei, N. E., Sze, S.-H., Stelly, D. M., Thaxton, P., Triplett, B., Town, C. D., & Jeffrey Chen, Z., (2006), Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton, *The Plant Journal*, *47*5, 761-775, <https://doi.org/10.1111/j.1365-313X.2006.02829.x>
- Sarilar, V., Palacios, P. M., Rousselet, A., Ridel, C., Falque, M., Eber, F., Chèvre, A.-M., Joets, J., Brabant, P., & Alix, K., (2013), Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in re-synthesized *Brassica napus* allotetraploids, *The New Phytologist*, *198*2, 593-604, <https://doi.org/10.1111/nph.12156>
- Schnable, J. C., & Freeling, M., (2011), Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize, *PLOS ONE*, *6*3, e17855, <https://doi.org/10.1371/journal.pone.0017855>
- Schnable, J. C., Freeling, M., & Lyons, E., (2012), Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses, *Genome Biology and Evolution*, *4*3, 265-277, <https://doi.org/10.1093/gbe/evs009>
- Schnable, J. C., Springer, N. M., & Freeling, M., (2011), Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss, *Proceedings of the National Academy of Sciences of the United States of America*, *108*10, 4069-4074, <https://doi.org/10.1073/pnas.1101368108>
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kurchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., . . . Wilson, R. K., (2009), The B73 Maize Genome : Complexity, Diversity, and Dynamics, *Science*, *326*5956, 1112-1115, <https://doi.org/10.1126/science.1178534>
- Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S. E., Bastolla, U., & Gutierrez, C., (2014), The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States, *The Plant Cell*, *26*6, 2351-2366, <https://doi.org/10.1105/tpc.114.124578>
- Shaked, H., (2001), Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat, *THE PLANT CELL ONLINE*, *13*8, 1749-1759, <https://doi.org/10.1105/tpc.13.8.1749>



- 
- Shapiro, B., Rambaut, A., & Drummond, A. J., (2006), Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences, *Molecular Biology and Evolution*, *23* 1, 7-9, <https://doi.org/10.1093/molbev/msj021>
- Shapiro, S. S., & Wilk, M. B., (1965), An analysis of variance test for normality (complete samples), *Biometrika*, *52* 3, 591-611, <https://doi.org/10.1093/biomet/52.3-4.591>
- Shimizu-Inatsugi, R., Terada, A., Hirose, K., Kudoh, H., Sese, J., & Shimizu, K. K., (2017), Plant adaptive radiation mediated by polyploid plasticity in transcriptomes, *Molecular Ecology*, *26* 1, 193-207, <https://doi.org/10.1111/mec.13738>
- Sievers, F., & Higgins, D. G., (2018), Clustal omega for making accurate alignments of many protein sequences, *Protein Science*, *27* 1, 135-145, <https://doi.org/https://doi.org/10.1002/pro.3290>
- Simons, C., Pheasant, M., Makunin, I. V., & Mattick, J. S., (2006), Transposon-free regions in mammalian genomes, *Genome Research*, *16* 2, 164-172, <https://doi.org/10.1101/gr.4624306>
- Singer, M. F., (1982), SINEs and LINEs : Highly repeated short and long interspersed sequences in mammalian genomes, *Cell*, *28* 3, 433-434, [https://doi.org/10.1016/0092-8674\(82\)90194-5](https://doi.org/10.1016/0092-8674(82)90194-5)
- Smirnov, N., (1948), Table for estimating the goodness of fit of empirical distributions, *The Annals of Mathematical Statistics*, *19* 2, 279-281, <https://doi.org/10.1214/aoms/1177730256>
- Sohpal, V. K., (2021), Comparative study : nonsynonymous and synonymous substitution of SARS-CoV-2, SARS-CoV, and MERS-CoV genome, *Genomics & Informatics*, *19* 2, e15, <https://doi.org/10.5808/gi.20058>
- Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E., (2015), Polyploidy and genome evolution in plants, *Current opinion in genetics & development*, *35*, 119-125, <https://doi.org/10.1016/j.gde.2015.11.003>
- Song, Q., Guan, X., & Chen, Z. J., (2015), Dynamic roles for small RNAs and DNA methylation during ovule and fiber development in allotetraploid cotton, *PLOS Genetics*, *11* 12, e1005724, <https://doi.org/10.1371/journal.pgen.1005724>
- Soppe, W. J. J., Jacobsen, S. E., Alonso-Blanco, C., Jackson, J. P., Kakutani, T., Koornneef, M., & Peeters, A. J. M., (2000), The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene, *Molecular Cell*, *6* 4, 791-802, [https://doi.org/10.1016/S1097-2765\(05\)00090-0](https://doi.org/10.1016/S1097-2765(05)00090-0)

- 
- Soundararajan, P., Won, S. Y., & Kim, J. S., (2019), Insight on Rosaceae Family with Genome Sequencing and Functional Genomics Perspective, *BioMed Research International*, 2019, 7519687, <https://doi.org/10.1155/2019/7519687>
- Spearman, C., (1904), The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, 151, 72, <https://doi.org/10.2307/1412159>
- Springer, N. M., Lisch, D., & Li, Q., (2016), Creating Order from Chaos : Epigenome Dynamics in Plants with Complex Genomes, *The Plant Cell*, 282, 314-325, <https://doi.org/10.1105/tpc.15.00911>
- States, D. J., & Gish, W., (1994), Combined use of sequence similarity and codon bias for coding region identification, *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 11, 39-50, <https://doi.org/10.1089/cmb.1994.1.39>
- Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J., (2021), The genomic ecosystem of transposable elements in maize, *PLOS Genetics*, 1710, e1009768, <https://doi.org/10.1371/journal.pgen.1009768>
- Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams Jr., R. M., (1949), *The American soldier : Adjustment during army life. (Studies in social psychology in World War II), Vol. 1*, Princeton Univ. Press.
- Student, (1908), The probable error of a mean, *Biometrika*, 1-25.
- Su, W., Jing, Y., Lin, S., Yue, Z., Yang, X., Xu, J., Wu, J., Zhang, Z., Xia, R., Zhu, J., An, N., Chen, H., Hong, Y., Yuan, Y., Long, T., Zhang, L., Jiang, Y., Liu, Z., Zhang, H., ... Liu, Z., (2021), Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe, *Proceedings of the National Academy of Sciences*, 11820, e2101767118, <https://doi.org/10.1073/pnas.2101767118>
- Sui, Y., Li, B., Shi, J., & Chen, M., (2014), Genomic, regulatory and epigenetic mechanisms underlying duplicated gene evolution in the natural allotetraploid *Oryza minuta*, *BMC Genomics*, 151, 11, <https://doi.org/10.1186/1471-2164-15-11>
- Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., Zhang, J., Zhang, H., Gong, G., Jia, Z., Zhang, F., Tian, J., Lucas, W. J., Doyle, J. J., Li, H., Fei, Z., & Xu, Y., (2017), Karyotype stability and unbiased fractionation in the paleo-allotetraploid cucurbita genomes, *Molecular Plant*, 1010, 1293-1306, <https://doi.org/10.1016/j.molp.2017.09.003>
- Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., Zhong, G.-Y., & Fei, Z., (2020), Phased diploid genome assemblies

- 
- and pan-genomes provide insights into the genetic history of apple domestication, *Nature Genetics*, *52* 12, 1423-1432, <https://doi.org/10.1038/s41588-020-00723-9>
- Suyama, M., Torrents, D., & Bork, P., (2006), PAL2nal : robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Research*, *34*, W609-W612, <https://doi.org/10.1093/nar/gkl315>
- Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., & Paterson, A. H., (2008), Unraveling Ancient Hexaploidy through Multiply-Aligned Angiosperm Gene Maps, *Genome Research*, *18* 12, 1944-1954, <https://doi.org/10.1101/gr.080978.108>
- Tate, J. A., Soltis, D. E., & Soltis, P. S., (2005, janvier 1), Polyploidy in plants, In T. R. Gregory (Éd.), *The evolution of the genome* (p. 371-426), Academic Press, <https://doi.org/10.1016/B978-012301463-4/50009-7>
- Team, R. C., (2013), R : A language and environment for statistical computing.
- The International Peach Genome Initiative, Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., Zuccolo, A., Rossini, L., Jenkins, J., Vendramin, E., Meisel, L. A., Decroocq, V., Sosinski, B., Prochnik, S., ... Rokhsar, D. S., (2013), The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution, *Nature Genetics*, *45* 5, 487-494, <https://doi.org/10.1038/ng.2586>
- Thomas, B. C., Pedersen, B., & Freeling, M., (2006), Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes, *Genome Research*, *16* 7, 934-946, <https://doi.org/10.1101/gr.4708406>
- Töpel, M., Antonelli, A., Yesson, C., & Eriksen, B., (2012), Past climate change and plant evolution in Western North America : a case study in Rosaceae, *PloS One*, *7* 12, e50358, <https://doi.org/10.1371/journal.pone.0050358>
- Toribio, A. L., Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R., Goodgame, N., ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R., Liu, X., Martínez-Villacorta, J., Pakseresht, N., Rajan, J., Reddy, K., Rosello, M., ... Cochrane, G., (2017), European Nucleotide Archive in 2016, *Nucleic Acids Research*, *45*, D32-D36, <https://doi.org/10.1093/nar/gkw1106>
- Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., & Kakutani, T., (2009), Bursts of retrotransposition reproduced in Arabidopsis, *Nature*, *461* 7262, 423-426, <https://doi.org/10.1038/nature08351>

- 
- Tzeng, Y.-H., Pan, R., & Li, W.-H., (2004), Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions, *Molecular Biology and Evolution*, *21* 12, 2290-2298, <https://doi.org/10.1093/molbev/msh242>
- Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F. J., Llorens, C., Roig, F. J., Parmaksiz, I., Dundar, E., Xie, F., Zhang, B., Ipek, A., Uranbey, S., Erayman, M., ... Van de Peer, Y., (2017), Genome of wild olive and the evolution of oil biosynthesis, *Proceedings of the National Academy of Sciences of the United States of America*, *114* 44, E9413-E9422, <https://doi.org/10.1073/pnas.1708621114>
- Vaillant, I., & Paszkowski, J., (2007), Role of histone and DNA methylation in gene regulation, *Current Opinion in Plant Biology*, *10* 5, 528-533, <https://doi.org/10.1016/j.pbi.2007.06.008>
- Van de Peer, Y., Ashman, T.-L., Soltis, P. S., & Soltis, D. E., (2021), Polyploidy : an evolutionary and ecological force in stressful times, *The Plant Cell*, *33* 1, 11-26, <https://doi.org/10.1093/plcell/koaa015>
- Van de Peer, Y., Maere, S., & Meyer, A., (2009), The Evolutionary Significance of Ancient Genome Duplications, *Nature Reviews. Genetics*, *10* 10, 725-732, <https://doi.org/10.1038/nrg2600>
- Van de Peer, Y., Mizrachi, E., & Marchal, K., (2017), The evolutionary significance of polyploidy, *Nature Reviews Genetics*, *18* 7, 411-424, <https://doi.org/10.1038/nrg.2017.26>
- Van der Walt, S., & Virtanen, P., (s. d.), *numpy/numpydoc : Numpy's Sphinx extensions*.
- van der Walt, S., Colbert, S. C., & Varoquaux, G., (2011), The NumPy Array : A Structure for Efficient Numerical Computation, *Computing in Science Engineering*, *13* 2, 22-30, <https://doi.org/10.1109/MCSE.2011.37>
- Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y., (2014), Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary, *Genome Research*, *24* 8, 1334-1347, <https://doi.org/10.1101/gr.168997.113>
- Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E., Absher, D. M., Wold, B. J., & Myers, R. M., (2013), Dynamic DNA methylation across diverse human cell lines and tissues, *Genome Research*, *23* 3, 555-567, <https://doi.org/10.1101/gr.147942.112>

- 
- Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., . . . Viola, R., (2010), The Genome of the Domesticated Apple (*Malus × Domestica* Borkh.), *Nat. Genet.*, *42*10, 833-839, <https://doi.org/10.1038/ng.654>
- Vicient, C. M., & Casacuberta, J. M., (2017), Impact of transposable elements on polyploid plant genomes, *Annals of Botany*, *120*2, 195-207, <https://doi.org/10.1093/aob/mcx078>
- Vicient, Suoniemi, Anamthawat-Jónsson, Tanskanen, Beharav, Nevo & Schulman, (1999), Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus *Hordeum*, *The Plant Cell*, *11*9, 1769-1784, <https://doi.org/10.1105/tpc.11.9.1769>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Contributors, S. 1. 0., (2020), SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, *17*, 261-272, <https://doi.org/10.1038/s41592-019-0686-2>
- Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I., & Martienssen, R. A., (2002), Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi, *Science (New York, N. Y.)*, *297*5588, 1833-1837, <https://doi.org/10.1126/science.1074973>
- Wang, J., Tian, L., Madlung, A., Lee, H.-S., Chen, M., Lee, J. J., Watson, B., Kagochi, T., Comai, L., & Chen, Z. J., (2004), Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids., *Genetics*, *167*4, 1961-1973, <https://doi.org/10.1534/genetics.104.027896>
- Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., & Zhang, X., (2018), Evolutionary dynamics of 3D genome architecture following polyploidization in cotton, *Nature Plants*, *4*2, 90-97, <https://doi.org/10.1038/s41477-017-0096-3>
- Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., . . . Brassica rapa Genome Sequencing Project Consortium, (2011), The genome of the mesopolyploid crop species *Brassica rapa*, *Nature Genetics*, *43*10, 1035-1039, <https://doi.org/10.1038/ng.919>

- 
- Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H., (2012), MCScanX : a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Research*, *40* 7, e49-e49, <https://doi.org/10.1093/nar/gkr1293>
- Wang, Y., Nie, F., Shahid, M. Q., & Baloch, F. S., (2020), Molecular footprints of selection effects and whole genome duplication (WGD) events in three blueberry species : detected by transcriptome dataset, *BMC Plant Biology*, *20* 1, 250, <https://doi.org/10.1186/s12870-020-02461-w>
- Wang, Y., Dai, A., Chen, Y., & Tang, T., (2021), Gene Body Methylation Confers Transcription Robustness in Mangroves During Long-Term Stress Adaptation, *Frontiers in Plant Science*, *12*.
- Wang, Z., Yang, J., Cheng, F., Li, P., Xin, X., Wang, W., Yu, Y., Zhang, D., Zhao, X., Yu, S., Zhang, F., Dong, Y., & Su, T., (2022), Subgenome dominance and its evolutionary implications in crop domestication and breeding, *Horticulture Research*, *9*, uhac090, <https://doi.org/10.1093/hr/uhac090>
- Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Rooter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., ... Brian, (2020, avril), *mwaskom/seaborn : v0.10.1 (April 2020)* (Version v0.10.1), Zenodo, <https://doi.org/10.5281/zenodo.3767070>
- Wendel, J. F., (2015), The wondrous cycles of polyploidy in plants, *American Journal of Botany*, *102* 11, 1753-1756, <https://doi.org/10.3732/ajb.1500320>
- West, P. T., Li, Q., Ji, L., Eichten, S. R., Song, J., Vaughn, M. W., Schmitz, R. J., & Springer, N. M., (2014), Genomic distribution of h3k9me2 and DNA methylation in a maize genome, *PLOS ONE*, *9* 8, e105267, <https://doi.org/10.1371/journal.pone.0105267>
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H., (2007), A unified classification system for eukaryotic transposable elements, *Nature Reviews Genetics*, *8* 12, 973-982, <https://doi.org/10.1038/nrg2165>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H., (2019),

- 
- Welcome to the Tidyverse, *Journal of Open Source Software*, 443, 1686, <https://doi.org/10.21105/joss.01686>
- Wilcoxon, F., (1945), Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, 16, 80, <https://doi.org/10.2307/3001968>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B., (2016), The FAIR guiding principles for scientific data management and stewardship, *Scientific Data*, 31, 160018, <https://doi.org/10.1038/sdata.2016.18>
- Wolfe, J., & Wehr, W., (1988), Rosaceous chamaebatiaria-like foliage from the paleogene of western north america, *Aliso*, 121, 177-200, <https://doi.org/10.5642/aliso.19881201.14>
- Wood, R. J., Maynard-Smith, M. D., Robinson, V. L., Oyston, P. C., Titball, R. W., & Roach, P. L., (2007), Kinetic analysis of yersinia pestis DNA adenine methyltransferase activity using a hemimethylated molecular break light oligonucleotide (S. Fugmann, Éd.), *PLoS ONE*, 28, e801, <https://doi.org/10.1371/journal.pone.0000801>
- Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., & Rieseberg, L. H., (2009), The frequency of polyploid speciation in vascular plants, *Proceedings of the National Academy of Sciences of the United States of America*, 106 33, 13875-13879, <https://doi.org/10.1073/pnas.0811575106>
- Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., & Wang, X., (2014), Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids, *Proceedings of the National Academy of Sciences of the United States of America*, 111 14, 5283-5288, <https://doi.org/10.1073/pnas.1402475111>
- Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., & Freeling, M., (2010), Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs, *PLoS Biology*, 86, e1000409, <https://doi.org/10.1371/journal.pbio.1000409>
- Wright, S. I., Agrawal, N., & Bureau, T. E., (2003), Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana, *Genome Research*, 138, 1897-1903, <https://doi.org/10.1101/gr.1281503>

- 
- Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M. A., Tao, S., Korban, S. S., Wang, H., Chen, N. J., Nishio, T., Xu, X., Cong, L., Qi, K., Huang, X., Wang, Y., Zhao, X., Wu, J., ... Zhang, S., (2013), The genome of the pear (*Pyrus bretschneideri* rehd.), *Genome Research*, *23*2, 396-408, <https://doi.org/10.1101/gr.144311.112>
- Wu, S., Han, B., & Jiao, Y., (2020), Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms, *Molecular Plant*, *13*1, 59-71, <https://doi.org/10.1016/j.molp.2019.10.012>
- Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., & Ma, H., (2017), Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication, *Molecular Biology and Evolution*, *34*2, 262-281, <https://doi.org/10.1093/molbev/msw242>
- Xiong, Z., Gaeta, R. T., & Pires, J. C., (2011), Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*, *Proceedings of the National Academy of Sciences*, *108*19, 7908-7913, <https://doi.org/10.1073/pnas.1014138108>
- Xu, Y., Zhong, L., Wu, X., Fang, X., & Wang, J., (2009), Rapid alterations of gene expression and cytosine methylation in newly synthesized brassica napus allopolyploids, *Planta*, *229*3, 471-483, <https://doi.org/10.1007/s00425-008-0844-8>
- Yaakov, B., & Kashkush, K., (2011), Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid, *Genome*, *54*1, 42-49, <https://doi.org/10.1139/G10-091>
- Yaakov, B., & Kashkush, K., (2012), Mobilization of stowaway-like MITEs in newly formed allohexaploid wheat species, *Plant Molecular Biology*, *80*4, 419-427, <https://doi.org/10.1007/s11103-012-9957-3>
- Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D., Ju, Y., Yao, P., Li, X., Xie, K., Zhang, J., Wang, J., Liu, F., Ma, W., Shopan, J., Zheng, H., ... Zhang, M., (2016), The genome sequence of allopolyploid brassica juncea and analysis of differential homoeolog gene expression influencing selection, *Nature Genetics*, *48*10, 1225-1232, <https://doi.org/10.1038/ng.3657>
- Yang, Z., & Nielsen, R., (2000), Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, *Molecular Biology and Evolution*, *17*1, 32-43, <https://doi.org/10.1093/oxfordjournals.molbev.a026236>



- 
- Yang, Z., et al., (2006), *Computational molecular evolution* (T. 284), Oxford University Press Oxford.
- Yang, Z., (2007), PAML 4 : phylogenetic analysis by maximum likelihood, *Molecular Biology and Evolution*, *24* 8, 1586-1591, <https://doi.org/10.1093/molbev/msm088>
- Yang, Z., & Yoder, A. D., (2003), Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species (J. Thorne, Éd.), *Systematic Biology*, *52* 5, 705-716, <https://doi.org/10.1080/10635150390235557>
- Yoder, A. D., & Yang, Z., (2000), Estimation of Primate Speciation Dates Using Local Molecular Clocks, *Molecular Biology and Evolution*, *17* 7, 1081-1090, <https://doi.org/10.1093/oxfordjournals.molbev.a026389>
- Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D., (2010), Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation, *Science*, *328* 5980, 916-919, <https://doi.org/10.1126/science.1186366>
- Zhang, H.-Y., Zhao, H.-X., Wu, S.-H., Huang, F., Wu, K.-T., Zeng, X.-F., Chen, X.-Q., Xu, P.-Z., & Wu, X.-J., (2016), Global Methylation Patterns and Their Relationship with Gene Expression and Small RNA in Rice Lines with Different Ploidy, *Frontiers in Plant Science*, *7*.
- Zhang, H., Lang, Z., & Zhu, J.-K., (2018), Dynamics and function of DNA methylation in plants, *Nature reviews Molecular cell biology*, *19* 8, 489-506, <https://doi.org/10.1038/s41580-018-0016-z>
- Zhang, J., Liu, Y., Xia, E.-H., Yao, Q.-Y., Liu, X.-D., & Gao, L.-Z., (2015), Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression, *Proceedings of the National Academy of Sciences*, *112* 50, E7022-E7029, <https://doi.org/10.1073/pnas.1515170112>
- Zhang, L., Zhao, J., Bi, H., Yang, X., Zhang, Z., Su, Y., Li, Z., Zhang, L., Sanderson, B. J., Liu, J., & Ma, T., (2021), Bioinformatic analysis of chromatin organization and biased expression of duplicated genes between two poplars with a common whole-genome duplication, *Horticulture Research*, *8* 1, 1-12, <https://doi.org/10.1038/s41438-021-00494-2>
- Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., ... Cong, P., (2019), A High-Quality Apple Genome

- 
- Assembly Reveals the Association of a Retrotransposon and Red Fruit Colour, *Nat Commun*, 10, <https://doi.org/10.1038/s41467-019-09518-x>
- Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M., & Jacobsen, S. E., (2009), Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*, *Genome Biology*, 10(6), R62, <https://doi.org/10.1186/gb-2009-10-6-r62>
- Zhang, Y., Liu, C., Cheng, H., Tian, S., Liu, Y., Wang, S., Zhang, H., Saqib, M., Wei, H., & Wei, Z., (2020), DNA methylation and its effects on gene expression during primary to secondary growth in poplar stems, *BMC Genomics*, 21(1), 498, <https://doi.org/10.1186/s12864-020-06902-6>
- Zhang, Z., & Yu, J., (2006), Evaluation of Six Methods for Estimating Synonymous and Nonsynonymous Substitution Rates, *Genomics, Proteomics & Bioinformatics*, 4(3), 173-181, [https://doi.org/10.1016/S1672-0229\(06\)60030-2](https://doi.org/10.1016/S1672-0229(06)60030-2)
- Zhao, D., Ferguson, A. A., & Jiang, N., (2016), What makes up plant genomes : the vanishing line between transposable elements and genes, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(2), 366-380, <https://doi.org/10.1016/j.bbagr.2015.12.005>
- Zhao, M., Zhang, B., Lisch, D., & Ma, J., (2017), Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants, *The Plant Cell*, 29(12), 2974-2994, <https://doi.org/10.1105/tpc.17.00595>
- Zheng, D., Ye, W., Song, Q., Han, F., Zhang, T., & Chen, Z. J., (2016), Histone Modifications Define Expression Bias of Homoeologous Genomes in Allotetraploid Cotton, *Plant Physiology*, 172(3), 1760-1771, <https://doi.org/10.1104/pp.16.01210>





---

**Titre :** Évolution des gènes dupliqués chez le pommier

**Mot clés :** Duplication complète du génome, Bioinformatique, dominance de sous génome, génomique, transcriptomique, épigénétique

**Résumé :** Un événement de duplication du génome entier (WGD) s'est produit chez l'ancêtre du pommier (*Malus domestica*). Les événements de WGD ont un impact profond sur les génomes et sont connus pour être des moteurs majeurs de l'évolution. Cette WGD est relativement récente (27 Millions d'années) et fait du pommier un organisme de choix pour étudier le devenir des gènes dupliqués par autopolyploïdisation. Dans cette étude, nous avons examiné l'évolution des fragments chromosomiques dupliqués, sous le prisme d'analyses génomiques, transcriptomiques et épigénétiques. Nous avons identifié 16 779 paires de gènes dupliqués dans le génome

du pommier, confirmant le caractère récent de la WGD. Les gènes au sein des paires ohnologues ne semblent pas soumis à des pressions de sélection différentes. Nous avons montré plusieurs déséquilibres dans la proportion de QTLs cartographiés entre fragments chromosomiques dupliqués, et caractérisé divers biais dans le fractionnement du génome, le niveau d'expression des gènes, la couverture en éléments transposables et la méthylation de l'ADN. Nos résultats suggèrent une dominance sous-chromosomique dans cet autopolyploïde, un phénomène proche de la sous-dominance génomique décrite jusqu'à présent uniquement chez les allopolyploïdes.

---

**Title:** Evolution of duplicated genes in apple

**Keywords:** Whole Genome Duplication, Bioinformatics, subgenome dominance, genomics, transcriptomics, epigenetics

**Abstract:** A whole-genome duplication (WGD) event occurred in the ancestor of the apple tree (*Malus domestica*). WGD events have a profound impact on genomes and are known to be major drivers of evolution. This WGD is relatively recent (27 million years) and makes the apple tree a prime organism to study the fate of genes duplicated by autopolyploidization. In this study, we examined the evolution of duplicated chromosomal fragments through the lens of genomic, transcriptomic, and epigenetic analyses. We identified 16 779 pairs of duplicated genes in the

apple genome, confirming the recent nature of WGD. Genes within ohnologous pairs do not appear to be subject to different selection pressures. We showed several imbalances in the proportion of QTLs mapped between duplicated chromosomal fragments, and characterized various biases in genome fractionation, gene expression level, transposable element coverage, and DNA methylation. Our results suggest a subchromosomal dominance in this autopolyploid, a phenomenon close to the genomic subdominance described so far only in allopolyploids.