



HAL
open science

Histoire de la domestication de *Triticum turgidum*: la capture d'exons au service de l'étude de la diversité génétique

Morgane Ardisson

► To cite this version:

Morgane Ardisson. Histoire de la domestication de *Triticum turgidum*: la capture d'exons au service de l'étude de la diversité génétique. Génétique des populations [q-bio.PE]. 2019. hal-02409617

HAL Id: hal-02409617

<https://ephe.hal.science/hal-02409617>

Submitted on 5 Mar 2020

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.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

MÉMOIRE
Présenté par

Morgane ARDISSON

Pour l'obtention du diplôme de l'École Pratique des Hautes Études

**Histoire de la domestication de *Triticum turgidum* :
La capture d'exons au service de l'étude de la
diversité génétique**

Soutenu le 10 décembre 2019 devant le jury composé de :

Claudie DOUMS	Directeur d'études EPHE, Paris	Présidente
Anne-Céline THUILLET	Chargée de recherche IRD, Montpellier	Rapporteur
Erick DESMARAIS	Ingénieur de recherche CNRS, Montpellier	Examineur
Pierre ROUMET	Chargé de recherche INRA, Montpellier	Tuteur scientifique
Stéphanie MANEL	Directeur d'études EPHE, Montpellier	Tutrice pédagogique

Mémoire préparé sous la direction de :

Dr Pierre ROUMET: UMR AGAP – Equipe Génomique évolutive et gestion des populations
Montpellier / Responsable équipe : Joëlle RONFORT

Dr Stéphanie MANEL: UMR CEFE – Equipe Biogéographie et écologie des vertébrés
Montpellier / Responsable équipe : Claude MIAUD

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

MÉMOIRE
Présenté par

Morgane ARDISSON

Pour l'obtention du diplôme de l'École Pratique des Hautes Études

**Histoire de la domestication de *Triticum turgidum* :
La capture d'exons au service de l'étude de la
diversité génétique**

Soutenu le 10 décembre 2019 devant le jury composé de :

Claudie DOUMS	Directeur d'études EPHE, Paris	Présidente
Anne-Céline THUILLET	Chargée de recherche IRD, Montpellier	Rapporteur
Erick DESMARAIS	Ingénieur de recherche CNRS, Montpellier	Examineur
Pierre ROUMET	Chargé de recherche INRA, Montpellier	Tuteur scientifique
Stéphanie MANEL	Directeur d'études EPHE, Montpellier	Tutrice pédagogique

Mémoire préparé sous la direction de :

Dr Pierre ROUMET: UMR AGAP – Equipe Génomique évolutive et gestion des populations
Montpellier / Responsable équipe : Joëlle RONFORT

Dr Stéphanie MANEL: UMR CEFE – Equipe Biogéographie et écologie des vertébrés
Montpellier / Responsable équipe : Claude MIAUD

Remerciements

Tout d'abord, je remercie chaleureusement les membres du jury : Anne-Céline Thuillet, Erick Desmarais et Claudie Doums, d'avoir accepté d'évaluer mon travail dans le cadre de mon Diplôme de l'École Pratique des Hautes Études en Sciences de la Vie et de la Terre. Merci également à Stéphanie Manel, ma tutrice pédagogique, qui m'a suivie pendant toute cette formation.

Je voudrais ensuite remercier l'ensemble de l'équipe GE2POP sans qui cette aventure n'aurait été possible.

Je pense bien évidemment à Pierre Roumet, je ne te remercierai jamais assez pour le soutien et la confiance que tu m'accordes depuis 11 ans déjà ! Tu m'as soutenue depuis le premier jour où j'ai eu l'idée folle de faire cette formation. Merci pour ton encadrement, tes encouragements dans les moments difficiles et surtout ta patience !

J'espère que ça n'a pas été trop dur pour toi de supporter mes fréquentes interrogations scientifiques mais aussi mes doutes et cela malgré un emploi du temps très chargé. Je te remercie pour le temps que tu m'as accordé, notamment lors de la rédaction de mon rapport. Le chemin fut parfois en « anse de sceau », mais je suis ravie d'avoir partagé cette aventure avec toi. Un grand merci !

Je voudrais ensuite remercier Muriel Tavaud, d'avoir passé des heures entières avec moi à remplir des tableaux blancs sur les concepts de génétique des populations et autant devant R à trouver des stratégies pour affronter cette montagne de données. Impossible de compter le nombre de fois où j'ai débarqué dans ton bureau, avec une liste de questions longue comme le bras ou avec la mine déconfite ! Tu as toujours été là et je t'en remercie sincèrement. Nous aurons finalement réussi à faire la peau à ce salami !

Un grand merci à toi, Jacques David, de m'avoir soutenue tout au long de cette formation et de m'avoir toujours encouragée à me poser des questions. Merci pour ton enthousiasme de chaque instant qui rend la recherche croustillante et excitante.

Merci à toi Sylvain Santoni, « professionnel de la profession », de m'avoir épaulée dans le projet « challenging » de mettre au point un protocole d'enrichissement par capture « de la mort qui tue ». L'idée semblait « capilotractée », mais on l'a fait, car ne l'oublions pas : « il n'y a pas de problème que des solutions » et même si « le diable se niche dans les détails », nous avons fini par tordre le coup à ce génome hors catégorie !

Merci à Vincent Ranwez et Gautier Sarah pour les longues heures à me former à la bio-informatique. Grâce à vous j'ai arrêté d'hyper-ventiler devant cet écran noir rempli d'un langage venu tout droit de l'espace. Un de mes objectifs était d'apprendre à analyser les séquences NGS, c'est chose faite ! Merci pour votre patience, quand les « j'ai juste 2 ou 3 petites questions, je ne vais pas t'embêter longtemps » se transformaient en 3 heures de discussion et de tortillage de scripts. Je tiens également à remercier Johanna Girodolle pour son soutien et sa patience.

Je voudrais aussi remercier Nathalie Chantret, de m'avoir aidée à dompter Egglib, le langage python et l'armée de paramètres de diversité de la génétique des populations. Merci pour ta pédagogie, ta disponibilité et ta patience.

Merci à Hélène Fréville et Joëlle Ronfort d'avoir pris le temps de relire mon mémoire dans la dernière ligne droite !

Je tiens aussi à remercier ou re-remercier Muriel T., Nathalie, Hélène et Joëlle de m'avoir toujours soutenue dans les moments de doute. Vos encouragements et votre bonne humeur m'ont beaucoup aidée. Longue vie au carré de chocolat !

Un grand merci à toutes les personnes qui m'ont aidée au cours de cette formation. Je pense à Hélène Fréville, la team Fst : Laurène Gay, Joëlle Ronfort et Margaux Jullien, Yan Holtz, Nicolas Rode, Fabien Condamine, Nicolas Galtier, mais j'en oublie certainement et je m'en excuse d'avance.

Un merci tout particulier à Muriel Latreille qui a toujours été là, pour le meilleur comme pour le pire. Une collègue mais surtout une amie ! Merci de m'avoir soutenue dans cette aventure, rien n'aurait été pareil sans toi. Tes encouragements et tes petites attentions m'ont permis de tenir quand je commençais à flancher. Un grand pardon pour mon manque de disponibilité, promis je vais maintenant pouvoir sortir de ma grotte !

Un grand merci à vous tous, qui avez fait de ce projet une aventure humaine formidable. L'objectif de cette formation était pour moi de découvrir les mondes de la bio-informatique et de la génétique des populations, me voilà comblée ! Vous m'avez appris à me poser des questions, à analyser, à me dépasser, avec la rigueur qu'exige le travail du scientifique. Merci infiniment pour cette expérience enrichissante, à tout point de vue.

Qu'aurait été cette aventure sans le soutien sans faille de mes amies, Laure, Cinderella, Anne-Laure, Sabine, Virginie, Anaïs, Ophélie, Emilie, Laurie, Delphine, Jessica... et toute la bande ! Merci d'avoir toujours été là pour moi.

Une pensée pour ma famille qui m'a encouragée pendant ces trois ans de formation et qui m'a reboostée dans les moments de fatigue ou de doute.

Ma dernière pensée est pour vous, Jean-charles et Annaé. Vous avez toujours cru en moi et votre soutien m'a été indispensable. J'admets que ça n'a pas dû être facile de vivre avec moi ces derniers temps ! Promis, maintenant je serai plus disponible ... et détendue !

Sommaire

Abréviations	4
Liste des figures	6
Liste des tableaux	8
Liste des encadrés	8
<i>Avant-propos</i>	9
1 SYNTHÈSE BIBLIOGRAPHIQUE	10
1.1 La domestication des plantes et les grandes forces évolutives	11
1.1.1 La domestication des plantes	11
1.1.2 Les grandes forces évolutives	12
1.1.2.1 La mutation	12
1.1.2.2 La dérive génétique	12
1.1.2.3 La sélection	13
1.1.3 L'équilibre mutation-dérive / théorie de neutralité	14
1.1.4 L'impact de la domestication	14
1.2 L'histoire évolutive du blé dur : <i>Triticum turgidum</i>	16
1.2.1 La polyploïdie et l'organisation du génome	16
1.2.2 L'histoire de la domestication du blé dur	17
1.2.2.1 La première phase de domestication : de <i>T. turgidum</i> ssp <i>dicoccoides</i> vers <i>T. turgidum</i> ssp <i>dicoccum</i>	17
1.2.2.2 La deuxième phase de domestication : de <i>T. turgidum</i> ssp <i>dicoccum</i> vers <i>T. turgidum</i> ssp <i>durum</i>	18
1.2.2.3 La sélection moderne et la révolution verte	18
1.2.3 L'impact de la domestication sur la diversité génétique	20
1.2.4 Les traits phénotypiques impactés au cours de la domestication du blé dur	21
1.2.4.1 La solidité du rachis	21
1.2.4.2 La solidité des glumes	22
1.2.4.3 La hauteur des plantes	23
1.2.4.4 Le poids des grains	23
1.2.4.5 La teneur en azote contenu dans la feuille	23
1.3 Les techniques de génotypage au service de l'étude de la diversité génétique	25
<i>Présentation du sujet d'étude</i>	26

2	MATERIEL ET METHODES	27
2.1	Matériel végétal	28
2.2	Outils moléculaires	29
2.2.1	Extraction d'ADN	29
2.2.2	Sondes	30
2.2.3	Librairies génomiques	30
2.2.4	Enrichissement en séquences cibles par capture	32
2.2.5	Séquençage	33
2.3	Outils bio-informatiques	34
2.3.1	Qualité des séquences	34
2.3.2	Démultiplexage	34
2.3.3	Nettoyage	36
2.3.4	Références génomiques	36
2.3.5	Mapping	37
2.3.6	Détection de SNPs	38
2.3.7	Mise en forme des données	40
2.4	Statistiques et génétique des populations	41
2.4.1	Structure génétique de la série de domestication	41
2.4.2	Caractérisation des effets démographiques de la série de domestication	42
2.4.3	Détection de signatures génétiques de sélection liées à la domestication	45
2.4.3.1	Détection sans <i>a priori</i> sur l'ensemble du génome	45
2.4.3.2	Détection au niveau de zones candidates contrôlant des traits du syndrome de domestication	46
3	RESULTATS	48
3.1	Structure génétique de la série de domestication	49
3.1.1	Mise en évidence de groupes génétiques	49
3.1.2	Analyse du niveau d'admixture	50
3.2	Caractérisation des effets démographiques de la série de domestication	53
3.2.1	Diversité au sein des génomes homéologues	53
3.2.2	Diversité au sein des quatre groupes évolutifs de <i>T. turgidum</i>	53
3.2.3	Diversité génétique le long des chromosomes	54
3.2.4	Différenciation entre les quatre groupes évolutifs de <i>T. turgidum</i>	56
3.3	Détection de signatures génétiques de sélection liées à la domestication	58
3.3.1	Détection sans <i>a priori</i> sur l'ensemble du génome	58
3.3.1.1	Détection avec les rapports de π de Tajima	58
3.3.1.2	Détection avec les rapports de θ_s de Watterson	58
3.3.1.3	Détection avec les F_{st}	60

3.3.2	Détection de sélection dans les zones candidates contrôlant des traits du syndrome de domestication	61
3.3.2.1	Mesures morphologiques	61
3.3.2.2	Détection avec les rapports de π de Tajima	62
3.3.2.3	Détection avec les F_{st}	63
4	DISCUSSION ET PERSPECTIVES	65
4.1	Développement technologique	66
4.2	Structure génétique de la série de domestication	67
4.3	Caractérisation des effets démographiques de la série de domestication	69
4.4	Détection de signatures génétiques de sélection liées à la domestication	70
	BIBLIOGRAPHIE	73
	ANNEXES	83
	Annexe 1 : Protocole détaillé pour l'extraction d'ADN	
	Annexe 2 : Protocole détaillé pour la préparation des librairies	
	Annexe 3 : Protocole détaillé pour l'enrichissement par capture	
	Annexe 4 : Analyse de structure par ACP : tableau des valeurs propres	
	Annexe 5 : Analyse de structure par DAPC sur 120 génotypes : graphiques complémentaires	
	Annexe 6 : Analyse de structure par DAPC sur 90 génotypes : graphiques complémentaires	
	Annexe 7 : Evolution des valeurs du D de Tajima le long des chromosomes	
	Annexe 8 : Détection de contigs sous sélection par la méthode de Wright, 2005	
	Annexe 9 : Estimation des F_{st} (WC) sur les contigs qui composent les fenêtres cibles de 6Mb	

Abréviations

ACP : Analyse en Composantes Principales
ADN : Acide DésoxyriboNucléique
ADNc : Acide Désoxyribonucléique complémentaire
AFLP : Amplified Fragment Length Polymorphism
AGAP : Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales
ARN : Acide RiboNucléique
ARNm : Acide RiboNucléique messenger
ATP : Adenosine TriPhosphate
BET : Bromure d’Ethidium
BIC : Bayesian Information Criterion
BLAST : Basic Local Alignment Search Tool
Bp: Base pair (paire de base)
BWA : Burrows-Wheeler Aligner
CTAB: CetylTrimethylAmmonium Bromide
CpDNA : ADN chloroplastique
CYMMIT : Centre international d'amélioration du maïs et du blé
DAPC : Discriminant analysis of principal components
DC : *Triticum turgidum* ssp *dicoccum*
DD : *Triticum turgidum* ssp *dicoccoïdes*
DE : *Triticum turgidum* ssp *durum* « élite »
dNTP : désoxyriboNucléotides Tri-Phosphate
DP : *Triticum turgidum* ssp *durum* « population »
EDTA : Acide EthyleneDiamineTetraacetic
FIS : Indice de fixation
FST : indice de différenciation
GEVES : Groupe d'Etude et de contrôle des Variétés Et des Semences
GE²POP : Génomique évolutive et gestion des populations
GIE : Groupement d’Intérêt Economique
GWAS : Genome-wide association study
ICARDA : Centre international de recherche agricole dans les zones arides
InDel : Insertion ou Délétion d’une base ou de plusieurs bases
INRA : Institut National de la Recherche Agronomique
MAF : Minor Allele Frequency
MSE : Mean Squared Error
NaCl : Chlorure de sodium
NCBI : National Center for Biotechnology Information
NGS : Next Generation Sequencing
PCR : Polymerase Chain Reaction
PMG : Poids de 1000 grains
PVP : PolyVinylpyrrolidone
QTL : Quantitative Trait Loci
RFLP : Restriction fragment length polymorphism

SDS: Sodium Dodecyl Sulfate

sNMF : sparse Non-Negative Matrix Factorization

SNP: Single Nucleotide Polymorphism

TRIS : TRIShydroxyméthylaminométhane

UP : Ultra Pure

UV : Ultra-Violet

USDA : Département de l'Agriculture des États-Unis

WEW : Wild Emmer Wheat

WGS : Whole Genome Shotgun

Liste des figures

- Figure 1 : Les foyers mondiaux de domestication des plantes
- Figure 2 : Les différences morphologiques entre la téosinte et le maïs
- Figure 3 : La dérive génétique
- Figure 4 : Les trois types de sélection naturelle
- Figure 5 : Déséquilibre de liaison
- Figure 6 : Le temps de coalescence
- Figure 7 : Evolution des estimateurs de diversité lors d'un goulot d'étranglement suivi d'une ré-expansion démographique
- Figure 8 : Polyploïdie et organisation du génome de l'espèce *Triticum turgidum*
- Figure 9 : Représentation schématique de l'histoire évolutive des différentes espèces de *Triticum*
- Figure 10 : Origine et la diffusion de l'espèce *Triticum turgidum*
- Figure 11 : Impact de la domestication sur la morphologie des épis, épillets et grains
- Figure 12 : Impact de la domestication et de la sélection sur le niveau de diversité nucléotidique
- Figure 13 : Impact de la domestication et de la sélection sur la taille efficace au sein de l'espèce *Triticum turgidum*
- Figure 14 : Représentation schématique d'un goulot d'étranglement et de son impact au niveau d'un gène neutre et d'un gène sous sélection
- Figure 15 : Représentation schématique du modèle de coalescent utilisé pour caractériser les épisodes successifs de goulot d'étranglement de domestication et de sélection
- Figure 16 : Modifications génétiques et phénotypiques lors du passage de l'amidonner sauvage, *T. turgidum* ssp *dicoccoïde* (Zavitan), à la forme domestiquée, *T. turgidum* ssp *dicoccum* (Svevo)
- Figure 17 : Corrélations entre les traits foliaires pour la définition du « spectre d'économie foliaire »
- Figure 18 : Variations des traits foliaires sur les 4 grandes formes évolutives de blé tétraploïdes
- Figure 19 : Visualisation des ADN sur gel d'agarose après coloration au BET
- Figure 20 : Définition des baits de capture
- Figure 21 : Les différentes étapes du protocole expérimental de la préparation des librairies génomiques
- Figure 22 : Visualisation d'une librairie génomique sur fragment analyser (DNF474).
- Figure 23 : Les différentes étapes du protocole expérimental de l'enrichissement par capture
- Figure 24 : Visualisation de deux mélanges de librairies génomiques, après enrichissement par capture, sur BioAnalyser (DNA7500).
- Figure 25 : Séquençage Illumina de dernière génération (NGS)
- Figure 26 : Les différentes étapes du pipeline d'analyses bio-informatiques
- Figure 27 : Scores de qualité des reads du mélange DEV_Cap009
- Figure 28 : Construction de la référence génomique simplifiée Zavitan_Baits
- Figure 29 : Répartition des 19738 contigs de la référence Zavitan_baits sur les chromosomes 14 chromosomes
- Figure 30 : Impacts de la référence utilisée pour le mapping des reads
- Figure 31 : Impacts des filtres sur le nombre de contigs de la référence Zavitan_baits portant des SNPs
- Figure 32 : Répartition des 683 contigs de la référence Zavitan_baits, sélectionnés pour l'analyse de structure, sur les 14 chromosomes

Figure 33 : Différence entre ACP et DAPC

Figure 34 : Effet du nombre de données manquantes sur l'estimateur de diversité génétique π de Tajima

Figure 35 : Choix de conservation des contigs

Figure 36 : Distribution des valeurs propres de l'Analyse en Composantes Principales (ACP)

Figure 37 : Pourcentage de variance expliquée pour les 10 premiers axes de l'ACP

Figure 38 : Projection des 120 individus sur les axes 1 et 2 de l'ACP

Figure 39 : Projection des 120 individus sur les axes 1 et 3 de l'ACP

Figure 40 : Valeurs du BIC en fonction du nombre de clusters (K), pour l'analyse de la structure des 120 génotypes avec DAPC.

Figure 41 : Projection des 120 génotypes (DD, DC, DP, DE) sur les deux premiers axes discriminants par la méthode DAPC.

Figure 42 : Valeurs du Bayesian Information Criterion (BIC) en fonction du nombre de clusters (K), pour l'analyse de la structure des 90 génotypes (DC, DP et DE) avec DAPC

Figure 43 : Projection des 90 génotypes (DC, DP, DE) sur les deux premiers axes discriminants par la méthode DAPC

Figure 44 : Valeurs de Cross-entropy pour 1 à 10 groupe(s) génétiques (K) par la méthode sNMF

Figure 45 : Niveau d'amixture au sein des quatre groupes évolutifs, obtenu à l'aide de l'outil sNMF

Figure 46 : Distribution du π de Tajima et du θ_s de Watterson, pour les quatre formes évolutives

Figure 47 : Valeurs de l'estimateur θ_s , pour chaque contig, entre les quatre groupes évolutifs deux à deux

Figure 48 : Evolution du niveau de diversité le long des chromosomes avec l'estimateur π

Figure 49 : Evolution du niveau de diversité le long des chromosomes avec l'estimateur θ_s

Figure 50 : Fragment du contig chr1B:1-356313144:340003850-340004569 situé dans la zone proche du centromère du chromosome 1B

Figure 51 : Détection de sélection, sans *a priori*, à l'aide des rapports de π de Tajima

Figure 52 : Détection de sélection, sans *a priori*, à l'aide des rapports de θ_s de Watterson

Figure 53 : Fragment de 50pb du contig chr1B:1-356313144:19395448-19396491 situé sur le chr. 1B

Figure 54 : Niveau de différenciation mesuré par le F_{st} , pour chacune des trois transitions évolutives et à l'échelle de chacun des 10734 contigs

Figure 55 : Mesures morphologiques sur les 120 génotypes appartenant aux quatre groupes évolutifs

Figure 56 : Graphique représentant l'évolution de la diversité à chaque transition de l'histoire évolutive de *T. turgidum* au niveau de six zones cible de 5Mb

Figure 57 : Distribution des rapports de π de Tajima calculés pour chaque contig et à chacune des trois transitions évolutives de *T. turgidum*

Figure 58 : F_{st} calculés pour chaque contig et pour les trois transitions : DD_DC, DC_DP et DP_DE , au niveau des six zones cibles

Figure 59 : La généalogie des blés

Liste des tableaux

Tableau 1 : Données passeport des 120 génotypes

Tableau 2 : Score de qualité – Phred

Tableau 3 : Localisation des gènes TtBtr1-A, TtBtr1-B, Q et Rht-B1b ainsi que les deux QTLs impliqués dans le poids des grains(PMG) et la teneur en azote de la feuille sur la référence ZAVITAN.

Tableau 4 : Répartition des 120 génotypes (DD, DC, DP, DE) dans les cinq clusters DAPC

Tableau 5 : Répartition des 90 génotypes (DC, DP, DE) dans les quatre clusters DAPC

Tableau 6 : Comparaison du niveau de diversité entre les deux génomes A et B.

Tableau 7 : Comparaison du niveau de diversité entre les quatre groupes évolutifs

Tableau 8 : Fst par paire entre les quatre groupes évolutifs

Tableau 9 : Mesures de diversité du contig chr1B:1-356313144:19395448-19396491

Tableau 10 : Perte de diversité nucléotidique entre les formes sauvages et les formes cultivées

Liste des encadrés

Encadré 1 : Déséquilibre de liaison

Encadré 2 : Temps de coalescence

Encadré 3 : Analyse en Composantes Principales (ACP)

Encadré 4 : Analyse Discriminante en Composante Principale (DAPC)

Encadré 5 : Factorisation de matrice non négative (sNMF)

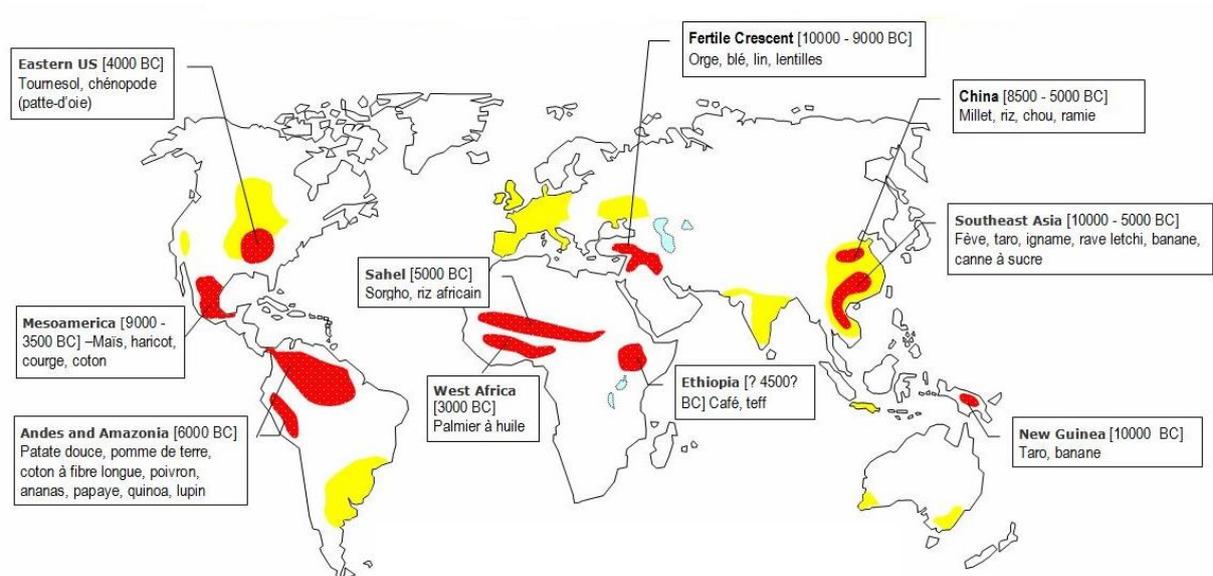
AVANT-PROPOS

La majorité des plantes qui sont cultivées aujourd'hui sont issues d'un processus de domestication, plus ou moins long, à partir d'espèces sauvages ancestrales. Comme pour la plupart des espèces cultivées, la domestication du blé dur a impacté certaines caractéristiques morphologiques (hauteur, taille des grains, etc ...) et s'est accompagnée d'une réduction importante de la diversité génétique. Pour comprendre ce processus de domestication, il est nécessaire de décrire les forces évolutives qui en sont responsables et de les quantifier de façon objective.

Cinquante ans après la révolution verte, le modèle de production actuel n'est plus en adéquation avec la demande sociétale et les défis d'aujourd'hui : améliorer la sécurité alimentaire mondiale tout en s'inscrivant dans une démarche agro-écologique, notamment par la réduction des intrants dans les systèmes de culture. Par ailleurs, les changements climatiques affectent, de façon non négligeable, le cycle de croissance du blé dur et fragilisent ses rendements (Lobell and Gourджи 2012). L'augmentation des températures endommage les feuilles et modifie donc les mécanismes de la photosynthèse, ce qui accélère le processus de sénescence des feuilles et affecte la qualité du grain. Dans ce contexte, il est impératif de proposer un nouveau système de production innovant en réintroduisant, par exemple, de la variabilité génétique dans les variétés cultivées, afin de les rendre plus robustes face aux changements climatiques actuels et à venir.

La caractérisation de l'ensemble de la variabilité génétique disponible dans les différents compartiments qui se sont succédés au cours de la domestication du blé, nous offre un réservoir d'informations susceptible de répondre à cette problématique. Pour cela, il est nécessaire de documenter la diversité génétique pour chacune des formes évolutives et de les comparer, puis d'observer l'évolution, au cours du processus de domestication, de certains traits phénotypiques, susceptibles apporter une réponse à ce besoin de changement de mode de production.

Synthèse bibliographique



Diamond, 2002

Figure1 : Les foyers mondiaux de domestication des plantes

En rouge: les différents foyers de la domestication des plantes

En jaune: les zones agricoles les plus productives du monde moderne

A l'exception de la chine et des Etats-Unis, nous pouvons noter que les foyers de la domestication ne correspondent pas aux plus grandes zones de productions actuelles.

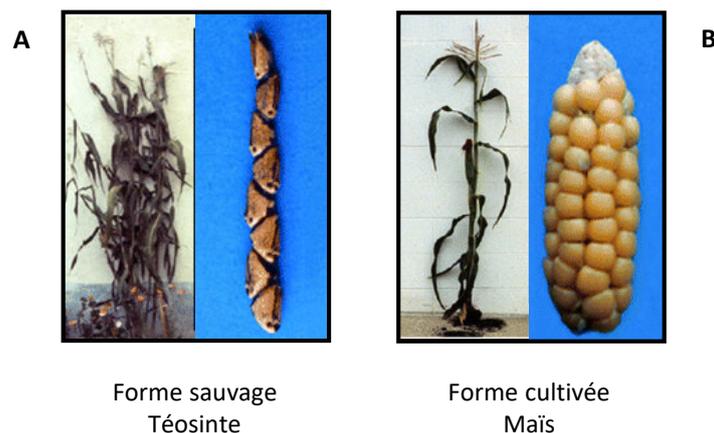


Figure2: Les différences morphologiques entre la téosinte (forme sauvage) et le maïs (forme cultivée)

A: morphologie d'un plant et d'un épi de téosinte: ancêtre sauvage du maïs actuel

B: morphologie d'un plant et d'un épi de maïs cultivé actuel

Les deux différences morphologiques les plus marquantes entre la forme sauvage et la forme cultivée sont le nombre de talles et la taille des épis.

1 Synthèse bibliographique

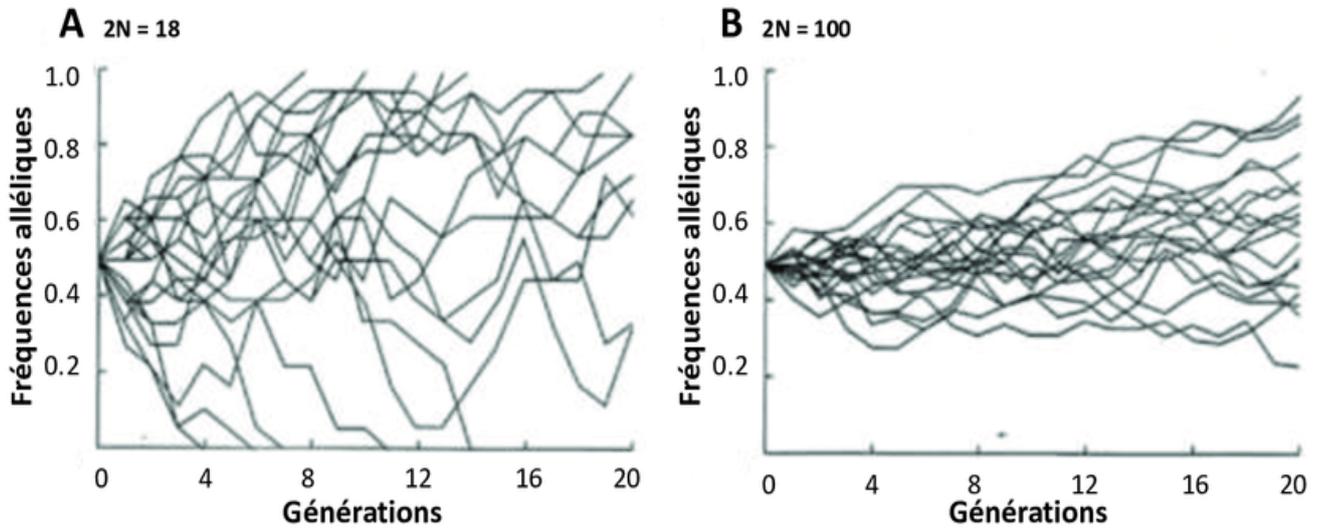
1.1 La domestication des plantes et les grandes forces évolutives

1.1.1 La domestication des plantes

L'agriculture a pris naissance dans plusieurs endroits du monde au Néolithique, il y a environ 12 000 ans et marque une véritable révolution dans l'histoire de l'humanité (Harlan 1992). Cela se caractérise par la transition de tribus et communautés de chasseurs-cueilleurs vers l'agriculture et la sédentarisation. Au cours des millénaires suivants, le développement de l'agriculture transforme les petits groupes de chasseurs-cueilleurs mobiles en sociétés sédentarisées qui modifient radicalement leur environnement au moyen de techniques agricoles adaptées (travail du sol puis irrigation) permettant d'augmenter la production. Ces développements favorisent de grandes densités de population, la division du travail, le commerce, les structures administratives et politiques centralisées, les systèmes de partage des connaissances (écriture).

Ces nouveaux modes de vie ont entraîné la domestication des plantes, des animaux et parfois de micro-organismes (levures, champignons, bactéries). Cette domestication est un exemple unique d'évolution rapide par la sélection, et a été l'élément central dans l'élaboration de la théorie de la sélection naturelle par Darwin (Darwin 1859). Chez les plantes, il a été démontré que cette domestication a eu lieu dans un nombre limité de foyers (Vavilov 1951; Harlan 1971; Diamond 2002; Glémin and Bataillon 2009) (figure 1). Les premières traces de domestication ont été localisées dans la zone du croissant fertile. Le climat (longue saison sèche et courte saison pluvieuse) convient particulièrement aux céréales comme le blé et l'orge (Frankel et al. 1995; Zohary and Hopf 2000). En Afrique, c'est principalement trois zones qui ont été le berceau de la domestication d'espèces comme le café, le palmier à huile, le sorgho et l'igname. D'autres espèces comme le maïs, la tomate ou le tournesol ont été domestiquées en Amérique. En Asie a eu lieu la domestication du riz ou de la canne à sucre.

Le processus de domestication s'accompagne d'une série de changements phénotypiques que l'on appelle **syndrome de domestication**. Les traits phénotypiques concernés peuvent être regroupés en plusieurs catégories et sont communs à plusieurs espèces. Premièrement, il y a les **traits associés aux conditions de récolte** ou de conservation. Dans cette catégorie, un des traits les plus importants, marqueur de la domestication pour les espèces dont les grains sont consommés (comme les céréales), est la non-dispersion des grains à maturité. En effet, afin de permettre la récolte par les agriculteurs puis la propagation des cultures, il est indispensable que les grains ne se dispersent pas à maturité (Harlan et al. 1973; Zohary 2004). D'autre part, chez les plantes sauvages, la production de nombreuses tiges et feuilles de petites tailles assure un succès reproducteur malgré des conditions climatiques éventuellement variables. Au cours de la domestication, la morphologie des plantes s'est souvent modifiée, optimisant la production des organes récoltés. Par exemple les ramifications axillaires ont fortement été réduites entre la téosinte (forme sauvage) et le maïs (forme cultivée) (figure 2). Pour finir, des traits associés au cycle de vie des plantes ont pu être affectés et certaines espèces sauvages pérennes sont devenues annuelles lors de la domestication comme le riz asiatique (Cheng et al. 2003). Deuxièmement, les **traits associés à la compétition des semences** sont aussi caractéristiques du syndrome de domestication. La culture a permis la sélection de plantules plus vigoureuses à travers l'augmentation du poids des grains dont la composition biochimique s'est aussi modifiée : plus importante en glucides et réduite en protéine. La diminution ou la perte totale de dormance et la



<https://www.researchgate.net/figure/Illustration-du-phenomene-de-derive-genetique>

Figure 3 : La dérive génétique.

Évolution au cours de 20 générations de 20 allèles à 20 locus indépendants, dont les fréquences initiales sont de 50% dans (A) une population de 9 individus et (B) une population de 50 individus. L'unique force en jeu est la dérive, dont la force est inversement proportionnelle à la taille de la population (N). Dans cet exemple, on observe que dans la petite population (A) plusieurs allèles se fixent ou disparaissent alors qu'au bout du même intervalle de temps ils sont encore tous présents dans la population de grande taille (B).

réduction des glumes sont également deux traits caractéristiques du syndrome de domestication (Harlan et al. 1973; Doebley 2004).

Troisièmement, les **traits associés à la production**. Au cours de la domestication, un certain nombre de traits visant à augmenter la production, ont été sélectionnés. Par exemple, chez le maïs et le sorgho, des modifications de la structure de l'inflorescence ont été sélectionnées pour produire des rendements plus élevés. Effectivement, l'inflorescence femelle de la téosinte est composée de deux rangées d'épillets simples, alors que le maïs a plusieurs rangs d'épillets appariés (Doebley 2004). Chez le blé, l'orge et le riz, le nombre d'inflorescences a évolué préférentiellement pour être plus dense (Doust 2007) mais aussi pour augmenter la fertilité au niveau de la pointe des épis. Du côté des solanacées, le fruit de la tomate a vu son poids multiplié par cent (Lin et al. 2014).

Toutes les espèces présentes sur notre planète sont caractérisées par une information génétique, contenue dans leur génome, qui leur est propre. Plusieurs grandes forces évolutives contribuent à faire évoluer ces génomes. Le niveau de diversité dans les populations (animales ou végétales) dépend de l'intensité de ces forces. La génétique des populations nous permet de caractériser ces forces et de décrire leur importance relative.

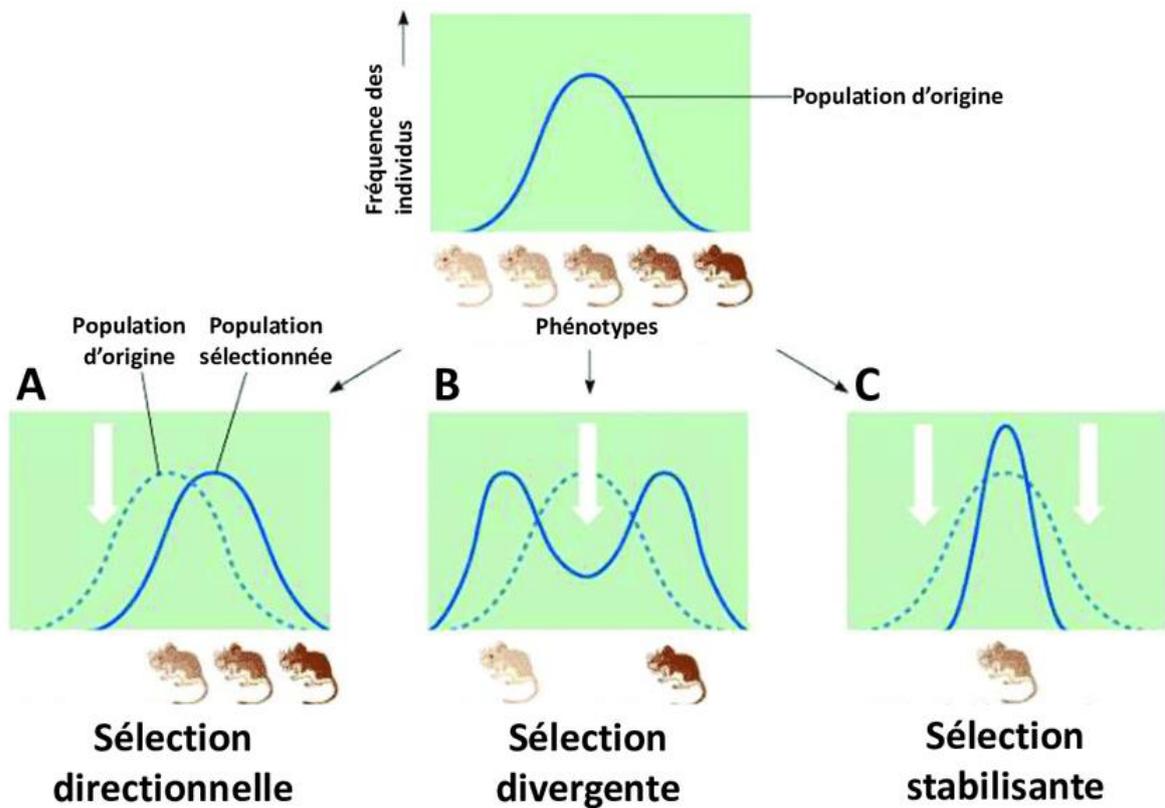
1.1.2 Les grandes forces évolutives

1.1.2.1 La mutation

Au cours de la réplication de l'ADN, processus indispensable à la multiplication cellulaire, des mutations peuvent se produire. Un certain nombre de processus permettent de corriger ces erreurs, spontanées et aléatoires, dans la cellule, mais certaines mutations subsistent. Il existe 3 types de mutations : les mutations ponctuelles par substitution qui correspondent au remplacement d'un nucléotide par un autre, générant un polymorphisme de type SNP (Single Nucleotide Polymorphism), l'insertion ou la délétion d'une base ou de plusieurs bases (InDel). La mutation est héréditaire seulement si elle touche la séquence génomique d'une cellule germinale. Chez les plantes, le taux de mutation ponctuelle (SNP), noté μ , est estimé en moyenne à 1.10^{-9} par base et par génération (Lynch and Walsh 2007), avec une estimation à 7.10^{-9} chez *Arabidopsis thaliana* (Weng et al. 2019). Ces mutations peuvent avoir un effet positif sur le phénotype (mutation avantageuse) ou négatif (mutation délétère) ou ne pas avoir d'effet (mutation silencieuse, ou neutre). Les mutations sont donc à l'origine de la variabilité génétique. Cependant, une fois ces mutations apparues, l'évolution de leur fréquence dans une population dépendra de deux autres forces évolutives : la dérive génétique et la sélection.

1.1.2.2 La dérive génétique

La dérive génétique est un des mécanismes majeurs de l'évolution. Elle caractérise l'évolution, au sein d'une population, de la fréquence des allèles d'un gène, causée par des phénomènes aléatoires résultant du mécanisme de formation (et de rencontre) des gamètes. La dérive génétique est étroitement liée à la notion de taille efficace (ou effectif efficace), notée N_e , qui correspond à l'effectif d'une population idéale (de type Wright-Fisher, de taille démographique constante notamment) pour laquelle les fluctuations des fréquences alléliques sont équivalentes à celle de la population étudiée.



<https://www.researchgate.net/figure/Les-trois-types-de-selection-naturelle>

Figure 4: Les trois types de sélection naturelle.

Une population de souris présentant une variation quantitative de la coloration du pelage peut être affectée par trois types de sélection : (A) directionnelle, (B) divergente, (C) stabilisante. Les flèches blanches représentent les pressions de sélection exercées contre certains phénotypes.

En effet, en fonction de la taille efficace de la population et de la fréquence initiale des allèles, le temps de fixation des allèles dans la population varie (figure 3). Les effets de la dérive génétique sont particulièrement importants sur les populations ayant une taille efficace limitée (Wright 1969) alors que dans une population de taille infinie, en l'absence de sélection et de mutation, les fréquences alléliques sont stables au cours des générations (équilibre d'Hardy-Weinberg).

1.1.2.3 La sélection

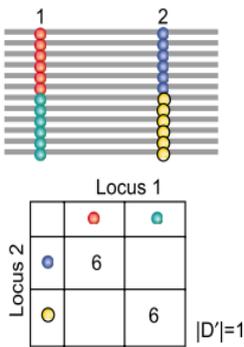
Charles Darwin (1809-1882) est le premier naturaliste à considérer la sélection naturelle comme une force évolutive. Cette sélection est définie par le fait que les individus ayant un caractère avantageux vis-à-vis d'une contrainte environnementale seront plus à même de survivre et de se reproduire que les individus n'ayant pas ce caractère. Pour que la sélection agisse, il est nécessaire d'avoir de la variabilité génétique dans la population, que le caractère favorable confère un avantage reproducteur et qu'il soit héritable (transmissible à la génération suivante). La sélection aura donc comme effet d'augmenter la fréquence des mutations favorables. La fréquence d'une mutation favorable va augmenter dans la population en fonction de l'intensité de son impact sur le phénotype. Il existe plusieurs types de sélection qui engendrent différentes signatures moléculaires (figure 4). La sélection directionnelle ou positive va faire augmenter la fréquence de la mutation avantageuse dans la population, et accroître l'adaptation des individus à leur environnement. Ce type de sélection est souvent rencontré lorsqu'une population subit des changements extrinsèques, par exemple des changements environnementaux, ou si une partie de cette population émigre dans un nouvel habitat. Plus le coefficient de sélection est important, plus la sélection sera rapide. Une forte sélection positive entraîne une réduction de la diversité avec *in fine* la fixation de l'allèle sélectionné, mais également sur les zones avoisinantes qui sont en déséquilibre de liaison avec l'allèle (Charlesworth and Eyre-Walker 2007) (encadré 1 au dos). La sélection divergente, se produit lorsque deux phénotypes extrêmes sont avantageux. Ce type de sélection peut conduire à une spéciation. La sélection stabilisante, quant à elle, élimine les phénotypes extrêmes pour favoriser les intermédiaires. Ceci a pour effet de diminuer la variance du caractère entre les individus de la population. Pour finir, la sélection équilibrante ou balancée, permet à plusieurs allèles de coexister à un locus donné. Dans ce cas, les individus hétérozygotes ont un avantage par rapport aux individus homozygotes. Ce type de sélection favorise le maintien de la diversité et le niveau d'hétérozygotie. Le maintien du polymorphisme peut être nécessaire à la survie dans un environnement présentant une hétérogénéité spatiale ou temporelle (Nielsen 2005).

La domestication est un cas particulier de sélection, causée par les humains depuis l'apparition de l'agriculture. Plus récemment, la sélection moderne pratiquée par les agriculteurs et les sélectionneurs a pour objectif de choisir à chaque génération les individus présentant les « meilleures » caractéristiques pour les faire se reproduire.

Encadré 1: Déséquilibre de liaison

En génétique, on dit qu'il y a déséquilibre de liaison (DL), si les fréquences des allèles à deux loci sont différentes de celles qui résulteraient d'une association aléatoire entre ces allèles. Autrement dit, c'est un signe qu'il y a association préférentielle entre certains allèles. Ce concept est devenu un outil indispensable pour la cartographie génétique et la cartographie de locus quantitatifs (QTL), par l'identification de déséquilibres d'associations entre allèles à un locus marqueur et à un locus impliqué dans la variation d'un caractère quantitatif.

A – Déséquilibre de liaison



B – Equilibre de liaison

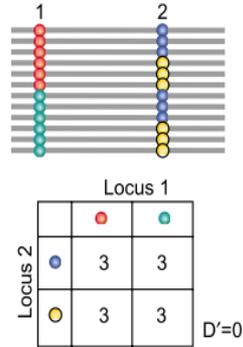


Figure 5: Déséquilibre de liaison

- (A) Le cas où 2 allèles sont en déséquilibre de liaison total. Quand il y a l'allèle rouge, il y a forcément l'allèle bleu.
- (B) Le cas où 2 allèles sont en parfait équilibre de liaison i.e. indépendants. L'association entre l'allèle rouge et l'allèle bleu est aussi fréquente que l'association entre l'allèle rouge et l'allèle jaune

D'après Rafalski, 2002

Avec le temps, le déséquilibre de liaison diminue en raison des recombinaisons. Plus la distance entre deux marqueurs est petite, moins il y a de chances qu'il y ait recombinaison entre les deux et donc, plus le déséquilibre de liaison augmente.

On mesure la force du déséquilibre de liaison entre deux marqueurs à l'aide du coefficient de liaison, D. Soient deux marqueurs, M1 et M2, possédant respectivement les allèles A et a et les allèles B et b. Le coefficient de déséquilibre de liaison entre M1 et M2 correspond à la différence entre la proportion d'haplotypes AB (ou ab) observée sous l'hypothèse d'indépendance. Si les marqueurs sont indépendants, on s'attend à ce que la proportion d'haplotypes AB soit égale au produit des fréquences des allèles A et B, soit :

$$D = p(AB) - p(A) * p(B)$$

Ainsi, plus D est élevé, plus les marqueurs sont en déséquilibre de liaison. Des standardisations de D ont été proposées afin d'avoir des coefficients compris entre -1 et 1. La mieux connue est le D' de Lewontin (1964).

Encadré 2: Temps de coalescence

La théorie de la coalescence est un modèle rétrospectif de génétique des populations. Cette théorie mathématique a été développée par John Kingman (1980). L'objectif est de simuler la généalogie d'une population de gènes jusqu'à un ancêtre commun. Les probabilités de coalescence des gènes sont fonction de la taille efficace (N_e) de la population dont ils proviennent. La théorie de la coalescence suppose que la population évolue selon le modèle Wright-Fisher où il n'y a ni recombinaison, ni sélection naturelle, ni flux de gènes et que la population n'est pas structurée. Ce modèle formalise une population panmictique de taille finie, constante au cours des générations non chevauchantes et dont les gènes d'une génération sont issus d'un tirage avec remise parmi les gènes de la génération précédente.

Dans ce contexte, la probabilité que deux lignées coalescent à la génération précédente est qu'ils aient le même parent. Pour une population diploïde dont la taille reste constante et égale à $2N_e$ copies de chaque locus, il y a $2N_e$ parents potentiels dans la génération précédente. La probabilité que les deux allèles aient le même parent est donc de $1/(2N_e)$. Réciproquement, la probabilité qu'il n'y ait pas coalescence à la génération précédente est de $1 - (1/(2N_e))$.

Par extension, La probabilité que la coalescence arrive à la génération t est de :

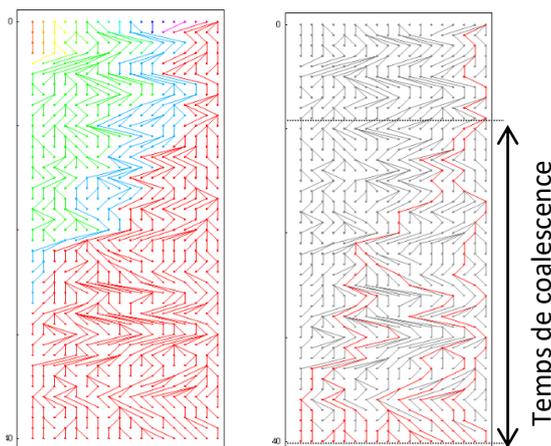


Figure 6 : Temps de coalescence

$$P(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right)$$

Probabilité que 2 allèles n'aient pas coalescé pendant t-1 génération

Probabilité que 2 allèles coalescent à une génération donnée

Adapté d'un schéma de Raphael Lebois

1.1.3 L'équilibre mutation-dérive / théorie de neutralité

La théorie neutraliste de l'évolution moléculaire a été initiée par Motoo Kimura en 1968 (Kimura and Ohta 1971; Kimura 1983). Selon lui, la majorité des mutations a un effet très faible sur la valeur sélective des individus car il s'agit pour la plupart des cas, de mutations synonymes (diversité neutre). La mutation et la dérive génétique permettent d'expliquer l'essentiel du polymorphisme. Si l'on considère cette diversité neutre, le polymorphisme génétique d'une population résulte d'un équilibre dynamique entre la mutation qui est responsable de l'apparition d'un nouvel allèle et la dérive génétique qui entraîne la fixation aléatoire (ou sa disparition) de cet allèle dans une population de taille finie. D'après Kimura, les mutations se produisent à un taux par génération de $2N_e\mu$ où N_e est la taille efficace de la population, et μ le taux de mutation.

Avec l'action de la dérive génétique, chaque mutation a une probabilité de fixation de $\frac{1}{2N_e}$.

Le taux de substitution d'un allèle neutre est donc de : $2N_e\mu * \frac{1}{2N_e} = \mu$.

Dans une population de N individus diploïdes, Kimura a montré qu'il s'écoule en moyenne $4N$ générations entre l'apparition d'un allèle et sa fixation par dérive. Ce temps de $4N$ générations correspond au temps de coalescence de la population (encadré 2). A l'équilibre mutation-dérive, le polymorphisme attendu dans une population dépend donc du produit du taux de mutation et du temps de coalescence de la population, que l'on note : $\theta = 4N_e\mu$ où N_e est la taille efficace de la population, et μ le taux de mutation.

Deux estimateurs de θ sont couramment utilisés :

-Le **π de Tajima** (1983) qui correspond au nombre moyen de différences nucléotidiques entre paires de séquences. Si on tire au hasard deux séquences dans une population, la diversité génétique de ces deux séquences peut être estimée en multipliant le taux de mutation (μ), le temps de coalescence entre ces deux séquences ($2N$) et le nombre de brins d'ADN sur lequel la mutation peut avoir lieu (2). L'espérance de π est donc $4 N_e \mu$.

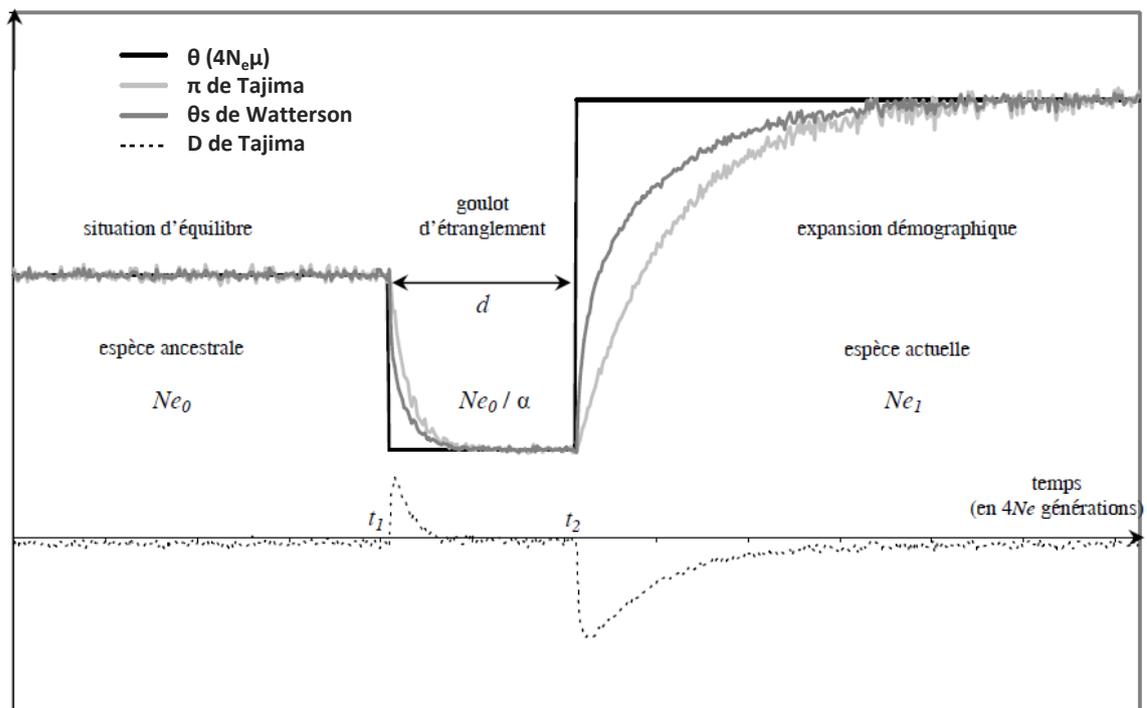
-Le **θ_s de Watterson** (1975) qui correspond au nombre de sites polymorphes S observés dans un échantillon. Il dépend du nombre de mutations apparues sur la longueur totale de la généalogie des gènes de l'échantillon ($4N$). Pour calculer cet estimateur, il est indispensable de standardiser le nombre de sites polymorphes S à la taille de l'échantillon à l'aide de la formule suivante :

$$\theta_s = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

Où S est le nombre de sites polymorphes et n est la taille de l'échantillon

1.1.4 L'impact de la domestication

La théorie neutraliste de l'évolution moléculaire selon Kimura fait abstraction de la sélection, en partant du postulat que la diversité génétique n'est affectée que par la dérive et la mutation. De plus, ce modèle se base sur le fait que les mutations qui ségrégent dans une population sont sélectivement neutres. C'est seulement dans ces conditions, que les estimateurs π de Tajima et θ_s de Watterson sont attendus égaux, à l'équilibre mutation – dérive, et valent $4N_e\mu$. Cependant, certains évènements



Haudry, 2007

Figure 7 : Evolution des estimateurs de diversité lors d'un goulot d'étranglement suivi d'une ré-expansion démographique.

Représentation des résultats de simulations de coalescence modélisant des changements démographiques successifs au cours de l'histoire d'une population. Une population ancestrale de taille efficace N_{e0} à l'équilibre, subit un goulot d'étranglement au temps t_1 . La taille efficace de la population est réduite d'un facteur α (intensité du goulot) pendant une durée $d=t_1-t_2$. Au temps t_2 , la population connaît une forte expansion démographique et reprend une taille efficace N_{e1} .

Le niveau de diversité de la population $\theta (4N_e\mu)$ est en noir, l'estimateur π de Tajima (1983) en gris clair et le θ_s de Watterson (1975) en gris foncé. L'évolution du D de Tajima (1989) au cours du temps est représentée en dessous, en pointillés.

démographiques (expansion démographique ou goulot d'étranglement) entraînent des modifications de l'intensité de la dérive génétique. La sélection, par essence, agit sur des mutations ayant un impact sur le phénotype (non-neutres). Dans ce cas, l'équilibre mutation-dérive n'est plus conservé et la distribution des fréquences alléliques est modifiée. Afin de détecter des signatures moléculaires de la sélection, il est possible de comparer les données théoriques obtenues avec le modèle neutraliste aux données de polymorphisme observées (diversité nucléotidique, nombre d'allèles, distribution de fréquences alléliques, etc...).

Tajima (1989) a proposé un des premiers tests d'écart à l'équilibre mutation-dérive, basé sur la diversité moléculaire d'un échantillon. Il consiste à comparer les deux estimateurs π et θ_s :

$$D_{\text{tajima}} = \frac{(\pi - \theta_s)}{\sqrt{\text{Var}(\pi - \theta_s)}}$$

Où :

π de Tajima correspond au nombre moyen de différences nucléotidiques entre paires de séquences
 θ_s de Watterson correspond au nombre de sites polymorphes observés dans un échantillon

Le modèle de **Wright-Fisher** se base sur une population dont l'effectif efficace ne varie pas au cours du temps. Or, au cours de l'histoire, certains évènements démographiques majeurs peuvent provoquer des expansions ou des réductions importantes de la taille de la population entraînant une baisse du niveau de polymorphisme attendu θ . Dans le cas d'une réduction massive, on parle de **goulot d'étranglement**. Ce type d'évènement est observé dans le cas de la quasi extinction d'une espèce avant une ré-expansion (cas de l'éléphant de mer ou du bison d'Europe). Un goulot d'étranglement a également été observé chez certaines espèces du genre homo (*Homo neanderthalensis*) suite à un épisode de glaciation. La perte de diversité associée à un goulot d'étranglement dépend de deux paramètres : la durée du goulot (d) et son intensité α (Eyre-Walker et al. 1998). Cette dernière est calculée en faisant le rapport de la taille efficace de la population avant le goulot sur la taille efficace pendant le goulot d'étranglement (figure 7). Si on fait l'hypothèse que le taux de mutation, dans une population reste constant au cours du temps, le niveau de diversité attendu θ suit l'évolution de l'effectif efficace N_e . Cependant, les deux estimateurs de diversité moléculaire **π et θ_s** réagissent différemment à ces évènements démographiques que sont les goulots d'étranglement. Lors d'un goulot d'étranglement, la taille efficace de la population étant faible, les mutations ayant une faible fréquence vont plus rapidement disparaître sous l'action de la dérive génétique que les mutations en forte fréquence. Le paramètre **θ_s** étant basé sur le nombre de sites polymorphes, sa valeur va plus rapidement diminuer que la valeur du paramètre **π** qui est particulièrement sensible aux mutations à fréquences équilibrées. L'évolution de ces deux estimateurs s'inverse lors d'une phase d'expansion. En effet, l'apparition de nouvelles mutations, favorisées par l'augmentation de la taille démographique, va impacter **θ_s** plus fortement que **π** . Le D de Tajima est positif lors de la réduction de taille démographique qui a eu lieu lors de la première phase du goulot d'étranglement et il devient négatif lors d'une expansion démographique qui suit. Le test d'équilibre du D de Tajima permet donc de caractériser ces évènements démographiques à l'aide des outils moléculaires. Néanmoins, les signatures moléculaires des évènements démographiques s'atténuent avec le temps et reviennent progressivement à l'équilibre. Par ailleurs, après un goulot d'étranglement les deux estimateurs **π et θ_s** mesurés sur une population actuelle et sa population d'origine permettent d'évaluer l'intensité de la perte de diversité.

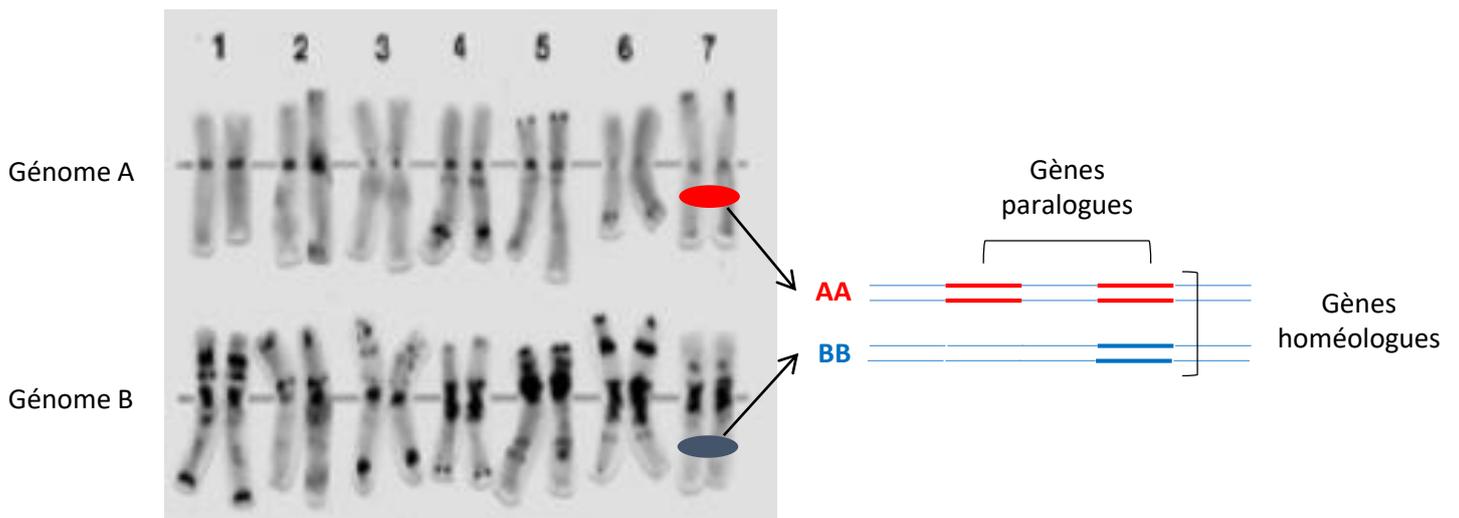


Figure 8 : Polyplôidie et organisation du génome de l'espèce *Triticum turgidum*

Représentation des sept chromosomes de chacun des deux génomes, A et B, de l'espèce *T. turgidum*.

Deux gènes ayant la même localisation sur des chromosomes d'une même paire mais de génomes différents, sont appelés gènes homéologues.

Deux gènes, situés sur un même chromosome, issus d'un même gène ancestral suite à un évènement de duplication, sont appelés gènes paralogues.

1.2 L'histoire évolutive du blé dur : *Triticum turgidum*

1.2.1 La polypléidie et l'organisation du génome

La polypléidie est le résultat de l'association de plusieurs jeux complets de chromosomes dans un noyau et résulte de l'hybridation entre espèces diploïdes (ou de pléidie inférieure) plus ou moins proches, la plupart du temps à la suite de la production de gamètes non réduits. Plusieurs événements de polypléidie ont eu lieu au cours de l'évolution des plantes (certains événements sont très anciens et d'autres beaucoup plus récents) et ont contribué à façonner les génomes tels qu'on les observe aujourd'hui. C'est un facteur important dans l'évolution des génomes des eucaryotes. On distingue l'autopolpléidie (les jeux de chromosomes proviennent de la même espèce) et l'allopolypléidie (les jeux de chromosomes viennent d'espèces différentes, mais suffisamment proches pour s'hybrider). Dans le cas de l'allopolypléidie, lorsque les génomes qui s'hybrident ont le même nombre de chromosomes, les chromosomes d'une même paire mais d'un génome différent sont dits « homéologues » (figure 8). L'espèce tétraploïde *Triticum turgidum* a deux génomes : le génome A (provenant de l'espèce diploïde *Triticum urartu*) et le génome B (provenant d'une espèce proche d'*Aegilops speltoides*), chacun étant constitué de 7 paires de chromosomes, et possèdent donc 14 paires de chromosomes ($2n=4x=28$, AABB). L'homologie entre les chromosomes homéologues est relativement élevée. Il est important de noter que chez les espèces allotétraploïdes, l'hérédité est disomique. Cela signifie, que l'appariement entre chromosomes des génomes A et B, lors de la méiose, est inhibé grâce à un système génétique complexe dont l'élément majeur est le gène *Ph* (Pairing homeologous) (Griffiths et al. 2006). La taille du génome complet du blé dur, *T. turgidum ssp durum*, est de 10.5 Gb, soit cinq fois plus gros que le génome humain et 30 fois plus gros que celui du riz. Le nombre de gènes chez le blé est estimé à 62 813 gènes (Avni et al. 2017). Ces gènes sont répartis de façon à peu près équivalente entre les deux génomes A et B. Cependant, la distribution des gènes le long des chromosomes n'est pas homogène (Choulet et al. 2014; Avni et al. 2017).

Deux gènes ayant la même localisation sur des chromosomes d'une même paire mais de génomes différents sont appelés gènes homéologues. En plus de la complexité génomique liée à la polypléidie, le génome du blé est constitué à 80% de séquences répétées. Ces séquences répétées sont principalement des éléments transposables : des fragments d'ADN présents dans les génomes et qui ont la capacité de transposer, c'est-à-dire de se déplacer de façon autonome. On distingue deux mécanismes principaux de transposition, les retro-transposons (classe I) utilisent un intermédiaire à ARN et transposent par un mécanisme de type « copier – coller » (l'élément augmente alors son nombre de copies au sein du génome). Ces éléments sont les plus abondants dans les génomes des plantes. Le deuxième mécanisme est de type « couper – coller », et caractérise les éléments à ADN (classe II) (McClintock 1950). Par ailleurs, il arrive que les gènes eux-mêmes se dupliquent au sein du même génome. On appelle alors gènes paralogues, des gènes issus d'un même gène ancestral suite à un événement de duplication, par opposition aux gènes orthologues qui sont issus d'un ancêtre commun à la suite d'un événement de spéciation et sont donc dans des génomes différents.

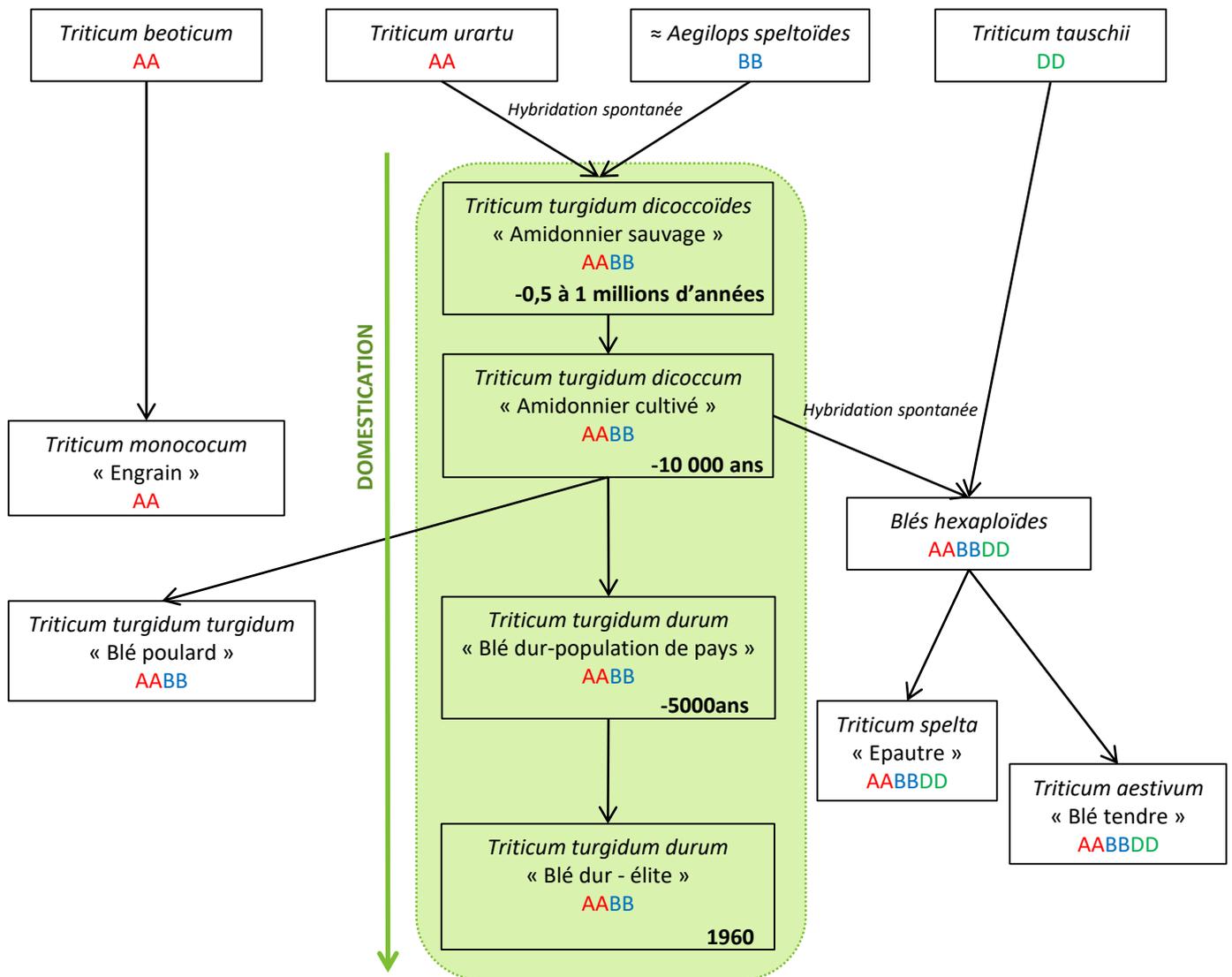


Figure 9 : Représentation schématique de l'histoire évolutive des différentes espèces de *Triticum*. *Triticum turgidum* ssp. *dicocoides* est issu de l'hybridation spontanée entre *Triticum urartu* (AA) et une espèce proche de *Aegilops speltaoides* (BB). La domestication de cette sous-espèce a eu lieu il y a 120000 ans pour donner *Triticum turgidum* ssp. *dicocum*. Plusieurs phases de sélection ont ensuite permis de voir apparaître deux nouvelles sous-espèces: *Triticum turgidum* ssp. *durum* et *Triticum turgidum* ssp. *durum*.



Zohary, 2000

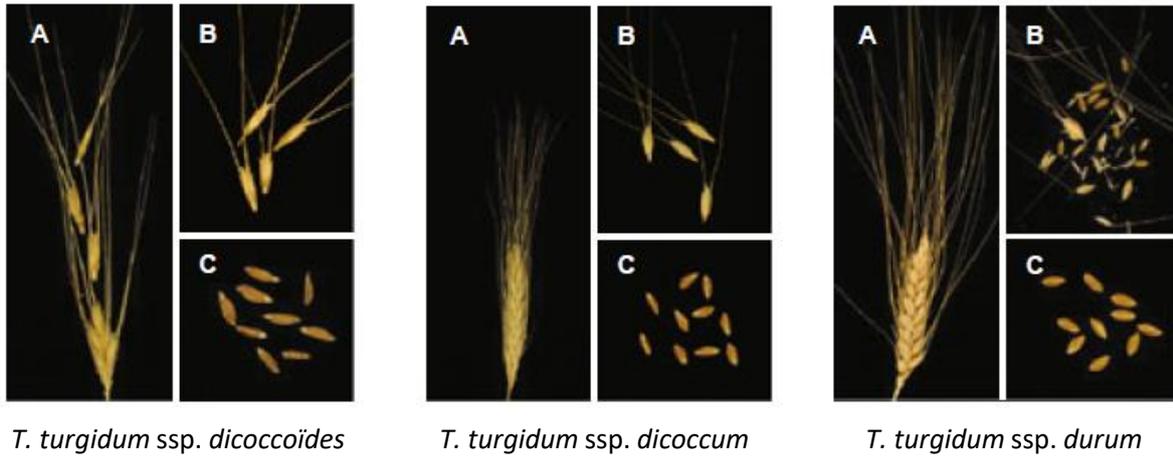
Figure 10 : Origine et la diffusion de l'espèce *Triticum turgidum*

Le centre de domestication de *T. turgidum* ssp. *dicocoides* a eu lieu au niveau du Mont Karaca Dag situé dans le Sud-Est de la Turquie. La diffusion de l'amidonnier cultivé est corrélée à l'expansion de l'agriculture et va s'étendre de la Scandinavie à l'est de l'Afrique.

1.2.2 L'histoire de la domestication du blé dur

1.2.2.1 La première phase de domestication : de *T. turgidum* ssp *dicocoides* vers *T. turgidum* ssp *dicoccum*

Le blé dur a été domestiqué il y a environ 12 000 ans dans la région du croissant fertile (Padulosi et al. 1996; Nesbitt and Samuel 1998; Willcox 1998). L'amidonner sauvage, *Triticum turgidum* ssp *dicocoides*, est l'espèce sauvage identifiée comme la représentante actuelle de l'espèce ancestrale à l'origine de la domestication. Cette espèce est allotétraploïde (génome AABB) ; elle résulte d'une hybridation spontanée, qui se serait produite entre -0.5 et -1 millions d'années (Mori et al. 1995; Huang et al. 2002; Dvorak and Akhunov 2005) entre deux espèces diploïdes : *Triticum urartu* (génome AA) et une espèce inconnue, proche de *Aegilops speltaoides* (génome BB) (figure 9). L'aire de répartition actuelle de *T. turgidum* ssp *dicocoides* couvre le croissant fertile, zone allant du sud Levant, à la Turquie sud-orientale au Nord, et à l'ouest de l'Iran vers l'Est (figure 10). Les données archéologiques disponibles montrent que sa distribution géographique, à l'époque de la domestication, correspond à sa répartition actuelle, mais elles n'excluent pas l'extinction locale d'un certain nombre de populations se traduisant par une diminution de densité (Nesbitt and Samuel 1996). Plusieurs études basées sur différents types de marqueurs moléculaires (AFLP : Özkan et al. 2002; Özkan et al. 2005 / CpDNA : Mori 2003 / RFLP : Luo et al. 2007), ont été nécessaires afin de localiser le centre de domestication de *T. turgidum* ssp *dicocoides* au niveau du Mont Karaca Dag situé dans le Sud-Est de la Turquie. Par ailleurs, les indices archéo-botaniques sont plutôt en faveur de l'hypothèse qu'un seul événement de domestication aurait eu lieu (Willcox 2000 ; Zohary and Hopf 2000). La domestication a eu comme effet majeur de changer la morphologie des plantes sauvages pour les rendre plus adaptées à la pratique de l'agriculture (figure 11). C'est le passage de l'amidonner sauvage, *T. turgidum* ssp *dicocoides* à l'amidonner cultivé, *T. turgidum* ssp *dicoccum*. La composante principale du syndrome de domestication chez le blé dur est la perte de la capacité à la dispersion des graines grâce à un rachis solide, qui ne se rompt pas spontanément à maturité. Parmi les autres évolutions importantes liées à l'adaptation à l'agriculture il faut mentionner la perte du caractère de dormance des grains, la diminution du nombre de talles et d'épis produits par la plante et l'augmentation du poids des grains. Les données archéo-botaniques montrent que la forme domestiquée *T. turgidum* ssp *dicoccum* a mis près de 1500 ans à s'imposer en tant que forme cultivée. Durant cette période, l'amidonner cultivé était encore confiné au niveau du croissant fertile, au contact du compartiment sauvage (Zohary and Hopf 2000). Les premiers mouvements migratoires de la forme cultivée ont eu lieu au sein de cette zone. D'une manière générale, la diffusion des espèces domestiquées est très fortement corrélée à l'expansion de l'agriculture. Cela est dû au fait que l'agriculture ne s'est pas diffusée en tant que savoir-faire ou avec des semences, mais avec la migration des populations humaines vers d'autres régions. Dès le début du 7ème millénaire avant JC, l'amidonner cultivé s'est étendu à l'Est de l'Anatolie, au Nord de l'Irak et Sud-Ouest de l'Iran (Harlan 1955). Au 6ème millénaire avant notre ère, il est cultivé dans les plaines de Mésopotamie et Anatolie occidentale puis atteint le Turkménistan (Harris et al. 1993). Le mouvement migratoire s'est poursuivi en Europe, d'abord diffusé en Grèce et en Bulgarie (-5900 ans) puis dans les Balkans, les montagnes des Carpates et le bassin du Danube (-5500 ans) (Nesbitt and Samuel 1996). La diffusion a continué ensuite vers l'ouest le long de la côte nord de la Méditerranée (-5500 ans) : de l'Italie, au sud de la France et à l'Espagne (Kipfer 2000). L'amidonner cultivé n'atteint l'Allemagne, la Pologne, les îles Britanniques et la Scandinavie que vers 3500 avant JC (Cavalli-Sforza and Ammerman 1984; Barker 1985). *T. turgidum* ssp *dicoccum* est également présent



Faris et al., 2014

Figure 11 : Impact de la domestication sur la morphologie des épis, épillets et grains

Photographies permettant de visualiser les différences morphologiques au niveau des épis (A), épillets (B) et des grains (C) des trois sous-espèces : *T. turgidum* ssp. *dicoccoïdes*, *T. turgidum* ssp. *dicoccum* et *T. turgidum* ssp. *durum*.

Le trait phénotypique caractéristique du passage de *T. turgidum* ssp. *Dicoccoïdes* à *T. turgidum* ssp. *dicoccum* est le rachis solide. Le passage de *T. turgidum* ssp. *dicoccum* à *T. turgidum* ssp. *durum* a permis l'apparition des grains nus.

en Afrique (Ethiopie, Yémen et Maroc), mais l'absence de site archéologique nous permet seulement de supposer que sa première apparition a eu lieu il y a 5000 ans en Ethiopie (Belay and Furuta 2001).

1.2.2.2 La deuxième phase de domestication : de *T. turgidum* ssp *dicoccum* vers *T. turgidum* ssp *durum*

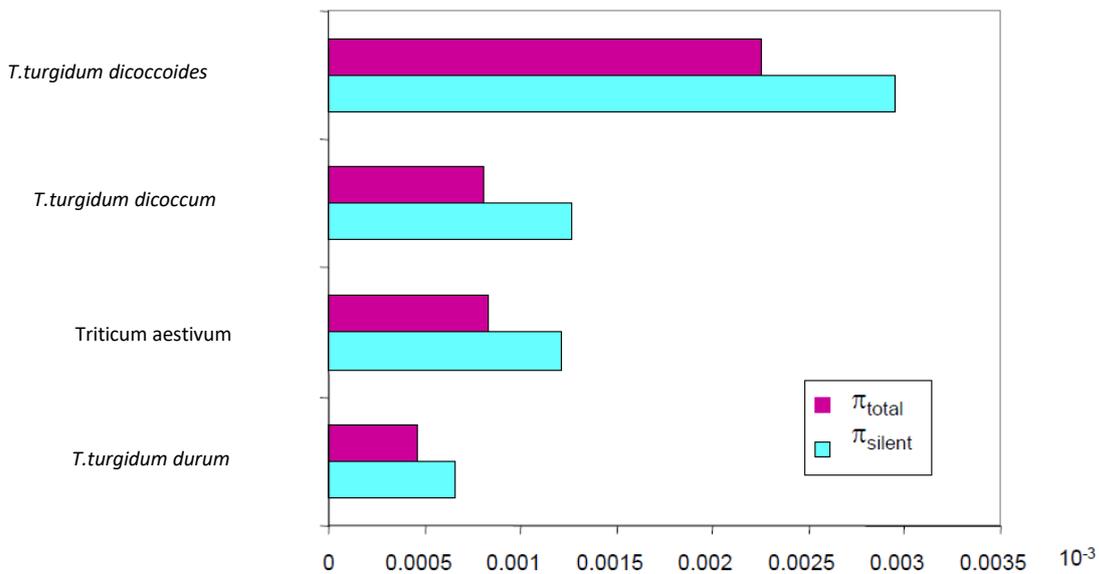
La deuxième étape majeure, plus récente, est le remplacement de l'amidonniér cultivé par une nouvelle forme, le blé dur : *Triticum turgidum* ssp *durum*. La différence morphologique essentielle entre ces deux formes est l'épaisseur des glumes (figure 11). L'amidonniér possède des glumes dures et leur forme carénée les rend solidaires du grain, provoquant le détachement des épillets lors du battage des épis. Il appartient à la catégorie des blés à grains vêtus. Le blé dur possède des glumes plus fines permettant la libération des grains nus lors du battage des épis tout en conservant le rachis intact. La transition progressive des formes cultivées à grains vêtus (*T. turgidum* ssp *dicoccum*) vers des formes cultivées à grains nus (*T. turgidum* ssp *durum*) est documentée par des données archéologiques synthétisées par Zohary et Hopf (2000). Les plus anciennes traces de grains nus ont été retrouvées au Proche-Orient très rapidement après l'apparition de *T. turgidum* ssp *dicoccum*. Ils apparaissent ensuite en Turquie et en Syrie au 7^{ème} millénaire avant JC, puis en Grèce au début du 5^{ème} millénaire avant JC, accompagnant l'installation des sites agraires en Europe méditerranéenne (Italie, sud de la France, Espagne) (Nesbitt and Samuel 1996). Dans ces régions, ainsi qu'au Proche-Orient, la culture du blé dur domine dès l'âge de bronze tardif (-1500 à -1200 ans av. JC). En ce qui concerne l'Afrique, le remplacement de l'amidonniér par le blé dur en Egypte représente une innovation culturelle significative (Zohary and Hopf 2000). Au Maroc, la culture du blé dur est attestée dans le nord dès le Néolithique (Morales, 2013) et en Libye à l'époque romaine (Van der Veen 1995). La zone de culture du blé dur s'est ensuite étendue vers les plaines du nord des Etats-Unis et au Canada à partir du 19^{ème} siècle.

Il semble donc que la propagation de la culture du blé dur (*T. turgidum* ssp *durum*) suit celle de l'amidonniér (*T. turgidum* ssp *dicoccum*) avec un léger retard chronologique. Cela permet de supposer que le blé dur est issu d'un évènement unique de sélection à partir de l'amidonniér du Proche-Orient.

1.2.2.3 La sélection moderne et la révolution verte

De l'Antiquité jusqu'au 19^{ème} siècle, l'évolution des pratiques culturelles a permis de sélectionner des populations permettant un meilleur rendement tout en s'adaptant aux conditions environnementales, aux pratiques culturelles et aux modes de transformations des grains. La culture est en phase d'expansion. Au 19^{ème} siècle, le blé dur est cultivé sous forme de « populations de pays » : une variété semée dans un champ était un mélange de différents génotypes ayant un taux d'homozygotie élevé du fait du système de reproduction autogame du blé dur, mais pas totalement (égal à 1) du fait de l'allogamie résiduelle.

Des programmes de sélection visant à produire des variétés de type « lignées pures » (totalement homozygotes) ont été progressivement mis en place à partir de la fin du 19^{ème} siècle et du début du 20^{ème} siècle principalement dans le bassin méditerranéen (France, Italie, Maghreb) puis aux Etats-Unis. Dans un premier temps, l'essentiel de l'effort de sélection a porté sur la collecte et l'évaluation agronomique de populations locales (populations cultivées dans le bassin méditerranéen, Asie

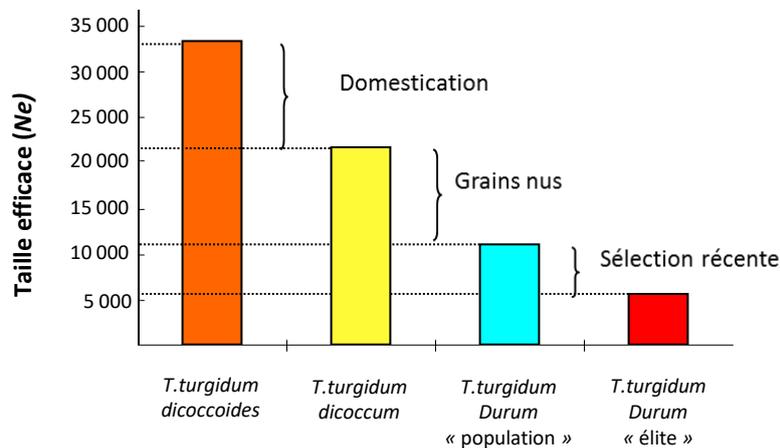


Haudry et al., 2007

Figure 12 : Impact de la domestication et de la sélection sur le niveau de diversité nucléotidique.

La diversité nucléotidique (π) est mesurée sur la base de 21 gènes et exprimée à 10^{-3} pour les quatre groupes : *T.turgidum dicoccoides*, *T.turgidum dicoccum*, *T.turgidum durum*, *Triticum aestivum* (blé tendre).

π_{silent} correspond au polymorphisme nucléotidique qui ne modifie pas la protéine, que l'on appelle « synonyme »
 π_{total} correspond à l'ensemble du polymorphisme : « synonyme » et « non synonyme »



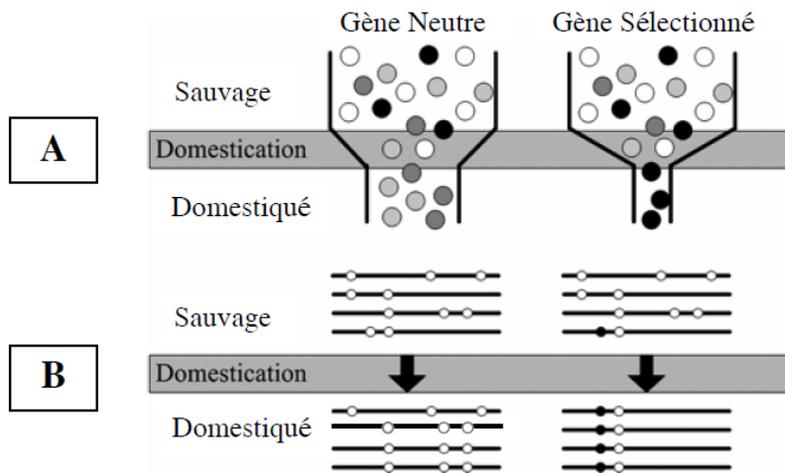
Thuillet et al. 2004

Figure 13 : Impact de la domestication et de la sélection sur la taille efficace au sein de l'espèce *Triticum turgidum*.

Représentation de la variation de la taille efficace chez *T.turgidum dicoccoides*, *T.turgidum dicoccum*, population de pays de *T.turgidum durum* et variétés élites de *T.turgidum durum*.

Les tailles efficaces ont été calculées d'après la relation $\theta=4N_e\mu$. Les estimations se basent sur l'indice de diversité mesuré par l'hétérozygotie de Nei sur 15 marqueurs microsatellites pour lesquels le taux de mutation avait été estimé (Thuillet et al., 2005)

mineure, le Moyen Orient et en Russie) ainsi que sur la fixation de lignées extraites de ce matériel végétal. Puis, des programmes de sélection ont été mis en place, par croisement de ces lignées locales de façon contrôlée, grâce à des schémas de sélection généalogique basés sur les performances de la descendance (Vilmorin 1880). Le but était d'améliorer la productivité (potentiel de rendement, verse), l'adaptation aux conditions pédoclimatiques locales (température, sécheresse, pression parasitaire) et la qualité du grain (couleur ambrée et aspect vitreux). Par exemple, des géniteurs de la sous-espèce *T. turgidum* ssp *dicoccum*, ont été largement utilisés pour améliorer la résistance du blé dur à la rouille noire (*Puccinia graminis*). Avec l'apparition de l'agriculture intensive basée sur la mécanisation et une utilisation importante d'intrants, les fermes spécialisent leurs productions. La création variétale est alors perçue comme un levier important pour améliorer quantitativement et qualitativement les productions, c'est ce que l'on appelle « la révolution verte ». En France, les programmes de recherche et d'amélioration du blé dur s'implantent à Montpellier (après la décolonisation de l'Algérie), dans un contexte réglementaire très encadré obligeant l'utilisation exclusive des grains de blé dur pour la production de semoule et de pâtes. Dans les années 1960, l'INRA débute un programme de création variétale ambitieux, ayant pour but l'obtention de variétés semi-naines à grains vitreux, traits observés sur les variétés d'Afrique du nord, mais avec des rendements de production plus importants et avec un grain de couleur jaune. L'introduction de gènes de nanisme par croisements avec des lignées de blé tendre est un tournant dans l'histoire de l'amélioration variétale. Grâce à sa petite taille (70-90cm), la variété « Durtal » est moins sensible à la verse et donc plus productive. Elle occupe 80% des surfaces cultivées en blé dur en 1975. Cependant, lors de l'introgession des gènes de nanisme, des allèles de blé tendre codant pour des protéines du grain (gliadines) ont également été transférés, ce qui a impacté fortement la qualité des protéines du grain de blé dur et donc la qualité pastière. Ce revers fait chuter drastiquement les surfaces cultivées et la production. Pour remédier à cela, un nouveau programme de sélection a été nécessaire pour retrouver la qualité pastière des grains. Pour prévenir de nouvelles crises, à partir de 1983, l'INRA anime un Groupement d'Intérêt Economique (GIE) rassemblant les sélectionneurs français, les industriels et les instituts techniques liés à la culture du blé dur et à sa transformation. Son objectif est de contribuer au développement et au rayonnement des variétés françaises. Depuis, le blé dur continue d'être amélioré à partir de croisements entre les lignées « élites » existantes pour répondre à la fois aux attentes des agriculteurs et de la filière semoulière. L'ensemble de ces variétés « élites », présentant les meilleures caractéristiques phénotypiques, sont inscrites au catalogue CTPS (Catalogue technique permanent de la sélection). La sélection variétale travaille sur les caractéristiques agronomiques comme la résistance à la verse, la productivité, la résistance à certaines maladies (rouille brune causée par *Puccinia recondita* et rouille jaune causée par *Puccinia striiformis*, fusariose, oïdium et plus récemment septoriose) ainsi que la bonne valorisation de l'azote. Plusieurs variétés sont cultivées en France (Miradoux, Fabulis, Isildur, Sculptur, Tablur, Pescadou, Babylone, Dakter, etc.) en fonction des conditions environnementales variables. Les nouvelles variétés sont aussi sélectionnées sur des critères de qualité alimentaire comme la couleur des grains (indice de jaune), la résistance à la moucheture, la résistance au mitadinage (anomalies de couleur et de texture qui affectent la qualité) et la teneur et la qualité des protéines.

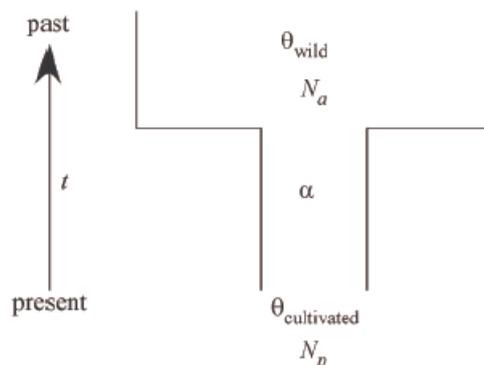


Ross-Ibarra et al., 2007

Figure 14 : Représentation schématique d'un goulot d'étranglement et de son impact au niveau d'un gène neutre et d'un gène sous sélection.

A : La diversité génétique est représentée par des cercles grisés avant et après la domestication au niveau d'un gène neutre et d'un gène sous sélection lors de la domestication. Le goulot d'étranglement réduit le nombre de génotypes dans le compartiment domestiqué par rapport à la population sauvage. Cette réduction de diversité est plus forte pour un gène sous sélection, pour lequel un seul génotype persiste après la domestication.

B : La diversité est représentée par des haplotypes schématisés avant et après la domestication, au niveau d'un gène neutre et d'un gène sous sélection lors de la domestication. La séquence d'ADN est représentée par les traits horizontaux et les sites polymorphes par les ronds. Les ronds blancs représentent les sites neutres alors que les ronds noirs représentent un site impliqué dans le déterminisme génétique d'un trait sous sélection. Tandis que plusieurs haplotypes du gène neutre subsistent à l'étape de domestication, la sélection du locus représenté par le rond noir, entraîne la fixation de cet haplotype uniquement, emportant avec lui un site neutre (« selective sweep » ou « balayage sélectif »).



Haudry et al., 2007

Figure 15 : Représentation schématique du modèle de coalescent utilisé pour caractériser les épisodes successifs de goulot d'étranglement de domestication et de sélection. La population ancestrale est caractérisée par sa taille efficace (N_a) et son niveau de diversité (θ_{wild}). La population cultivée est quant à elle caractérisée par sa taille efficace (N_p) et son niveau de diversité ($\theta_{cultivated}$). Le goulot d'étranglement qui sépare ces deux population est caractérisé par son intensité (α) et le temps de génération (t).

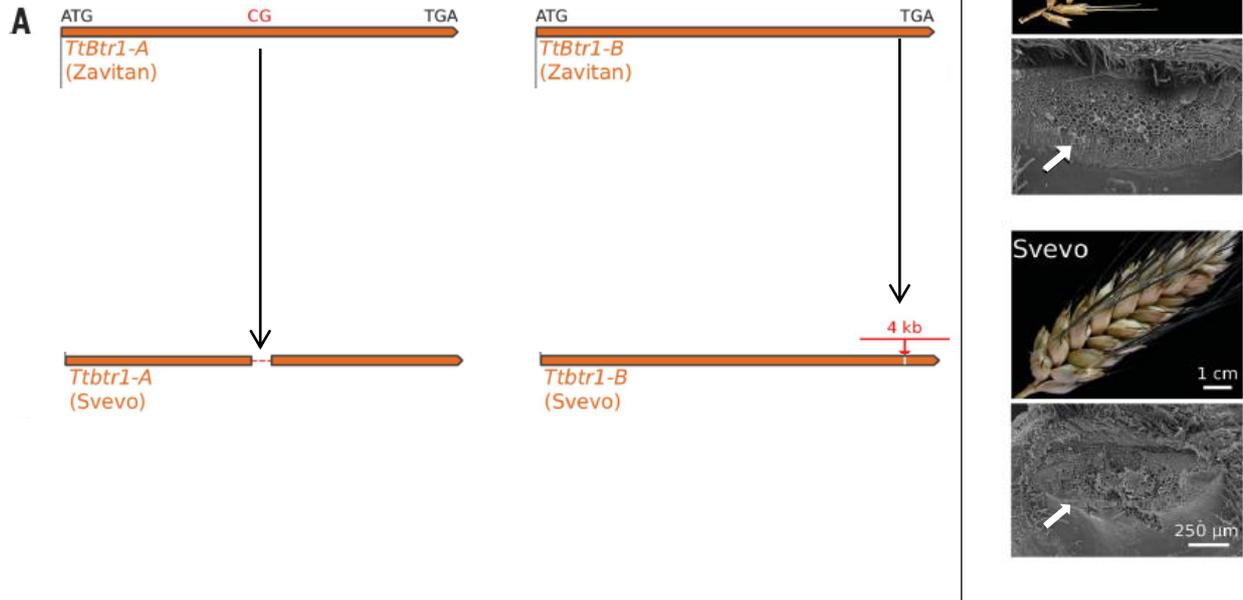
1.2.3 L'impact de la domestication sur la diversité génétique

De nombreuses études ont démontré que la diversité génétique des formes cultivées était moins importante que celle des populations de plantes sauvages dont elles sont issues (Buckler et al. 2001; Thuillet et al. 2005; Haudry et al. 2007). La diversité actuelle des espèces cultivées résulte de différents événements, tels que la domestication, l'expansion de l'agriculture ou la sélection moderne, qui se sont succédés. A partir de la fin des années 1990, l'accès facilité à l'information génétique a permis de préciser les contours de l'histoire évolutive de l'espèce *Triticum turgidum* (figure 12).

Lors de la domestication du blé marquant le passage de l'amidonner sauvage, *T. turgidum* ssp *dicocoides* à l'amidonner cultivé, *T. turgidum* ssp *dicocum*, a eu lieu le premier événement démographique majeur. Au cours de cette phase de domestication, par effet d'échantillonnage, on observe une diminution drastique de la taille efficace entre la forme sauvage et la forme domestiquée (Thuillet et al. 2005) (figure 13) : on parle de goulot d'étranglement de domestication (Eyre-Walker et al. 1998). Grâce aux mouvements migratoires, l'aire de culture de l'amidonner s'est rapidement étendue à l'ensemble des pays du pourtour méditerranéen. Ce goulot d'étranglement de domestication étant relativement récent à l'échelle de l'évolution des espèces, des traces moléculaires détectables persistent encore aujourd'hui, et ce malgré la phase d'expansion démographique qui a suivi. L'information génétique contenue dans les séquences d'ADN nous renseigne sur la diversité présente dans une population. Le fait de disposer des descendants de l'espèce ancestrale permet d'avoir un prédicteur de la situation avant sa domestication. Sous l'hypothèse que la population sauvage était à l'équilibre mutation-dérive (Tajima 1983) et qu'elle l'est encore aujourd'hui, nous pouvons estimer la diversité ($\theta=4N_e\mu$) à l'état initial. La théorie de la coalescence (Hudson 1990) nous permet, à partir des patrons de diversité de différents échantillons représentatifs des étapes décrites ci-dessus, de proposer des scénarios évolutifs pour décrire les transitions entre la forme sauvage (*T. turgidum* ssp *dicocoides*) et la forme domestiquée (*T. turgidum* ssp *dicocum*) sous un modèle neutre de Wright-Fisher.

L'émergence des formes domestiquées s'est également accompagnée d'une sélection, consciente ou inconsciente, sur certains caractères phénotypiques avantageux pour la culture. Cette sélection a laissé des signatures dans le génome, qu'il faut distinguer des effets démographiques, même s'ils participent tous les deux à la diminution de la diversité. La réduction de la diversité génétique liée à la domestication (effet démographique d'échantillonnage) touche l'ensemble du génome (Doebly 1989). Par contre, les gènes codant pour un caractère sur lequel la sélection a eu lieu, subissent une réduction de diversité supplémentaire car ils sont les cibles de la sélection positive lors de la domestication (figure 14). Le niveau de diversité à chaque locus, dans le compartiment cultivé, est potentiellement le fruit de ces deux phénomènes. Les gènes ayant joué un rôle important dans la domestication du blé dur (rachis solide, grain nu, etc...) ont perdu une grande partie de leur variabilité génétique lors de la transition vers les formes cultivées. Par ailleurs, on remarque que les locus neutres liés génétiquement (déséquilibre de liaison) aux locus fortement sélectionnés sont également touchés. Ce phénomène, appelé « selective sweep » ou « balayage sélectif » entraîne la diminution de la diversité neutre dans la zone autour des gènes d'intérêt agronomique.

Une étude menée par Haudry et al., (2007) a proposé un modèle démographique afin de caractériser les épisodes successifs de goulot d'étranglement de domestication (figure 15). L'analyse est basée sur le polymorphisme de séquence de 21 gènes au sein des populations *T. turgidum* ssp *dicocoides*, *T. turgidum* ssp *dicocum*, *T. turgidum* ssp *durum* et *T. aestivum* (blé tendre). La comparaison des



Adapté de Avni et al., 2017

Figure 16 : Modifications génétiques et phénotypiques lors du passage de l'amidonniér sauvage, *T. turgidum* ssp. *Dicoccoïde* (Zavitan), à la forme domestiquée, *T. turgidum* ssp. *Dicoccum* (Svevo).

A : Représentation schématique des séquences codantes des deux gènes homéologues portant les allèles mutés *Ttbtr1-A* et *Ttbtr1-B* chez la forme domestiquée, *T. turgidum* ssp. *Dicoccum* (Svevo). La première mutation, sur *Ttbtr1-A*, est une délétion de 2pb qui change le cadre de lecture et introduit un codon stop, tronquant ainsi la protéine en la raccourcissant de 196 à 97 acides aminés. La deuxième mutation, sur *Ttbtr1-B*, est une insertion de 4Kb qui produit une séquence protéique plus longue et non fonctionnelle chez la forme domestiquée.

B : Images de la zone d'abscission des épillets sur le rachis, obtenues par microscopie électronique à balayage. Chez la forme sauvage, *T. turgidum* ssp. *Dicoccoïde* (Zavitan), la zone cicatrice de la zone d'abscission est lisse alors qu'elle est rugueuse chez la forme domestiquée, *T. turgidum* ssp. *Dicoccum* (Svevo).

données avec des échantillons issus de simulations basées sur la théorie de la coalescence a permis de faire des inférences sur l'histoire évolutive du blé dur. Cela a permis d'estimer plusieurs paramètres : Les effectifs efficaces avant (N_a) et après (N_p) la domestication, l'intensité du goulot d'étranglement (α), le nombre de générations écoulées depuis le goulot d'étranglement (τ) et la diversité génétique (θ) de l'espèce sauvage. Sur les 21 gènes analysés dans cette étude, la perte de diversité entre la population sauvage, *T. turgidum ssp dicocoides* et *T. turgidum ssp dicocum* a été estimée à 69%, et elle s'élève à 80% lorsque l'on compare l'espèce sauvage à *T. turgidum ssp durum*. Par ailleurs, l'estimation des intensités de réduction des effectifs efficaces des populations lors des goulots d'étranglements sont de 3.15 lors du premier épisode de domestication (entre *T. turgidum ssp dicocoides* et *T. turgidum ssp dicocum*) et de 5.83 lors du passage au blé dur (*T. turgidum ssp dicocum* à *T. turgidum ssp durum*).

1.2.4 Les traits phénotypiques impactés au cours de la domestication du blé dur

1.2.4.1 La solidité du rachis

Le trait phénotypique caractéristique de la première phase de domestication, de *Triticum turgidum ssp dicocoides* vers *T. turgidum ssp dicocum*, est la perte de la capacité à la dispersion des graines grâce à un **rachis solide**, qui ne se rompt pas spontanément à maturité. Durant la mise en place de ce trait qui a pris environ 1000 ans (Willcox 2000), d'autres traits ont également été sélectionnés, mais il reste néanmoins le trait phénotypique caractéristique de cette transition. La solidité du rachis est contrôlée par des allèles récessifs au niveau de deux loci majeurs, situés sur les bras courts des chromosomes 3A et 3B (Watanabe et al. 2002). Une étude plus récente (Avni et al. 2017), a démontré que les deux gènes, notés « Btr », étaient des gènes homéologues, dérivant d'un même gène porté par le génome diploïde ancestral. Sur le chromosome 3A, on retrouve les allèles : **TtBtr1-A** et **TtBtr2-A** et sur le chromosome 3B, les allèles TtBtr1-B et TtBtr2-B (figure 16). La comparaison des séquences des allèles des individus « sauvages » (*T. turgidum ssp dicocoides*, génotype « Zavitan ») et « cultivés » (*T. turgidum ssp durum*, génotype « Svevo »), a démontré que les allèles Ttbtr1-A et Ttbtr1-B différaient par des mutations nucléotidiques susceptibles de perturber la structure des protéines. Par ailleurs, aucun polymorphisme n'a été détecté sur les deux autres gènes TtBtr2-A et TtBtr2-B entre la forme sauvage et la forme cultivée. Les deux mutations récessives homozygotes dans les gènes homéologues TtBtr1-A et TtBtr1-B sont nécessaires à la perte de la fonction de dispersion des grains à maturité. L'utilisation de la microscopie électronique à balayage a permis de préciser l'impact de ces mutations sur le phénotype. La modification de la structure des protéines a entraîné à son tour la modification de la structure des sites d'abscission des épillets sur la tige. En effet, dans la forme sauvage (génotype « Zavitan »), les sites d'abscissions sont lisses permettant d'avoir des rachis cassants, alors qu'ils deviennent rugueux dans la forme cultivée (génotype « Svevo »).

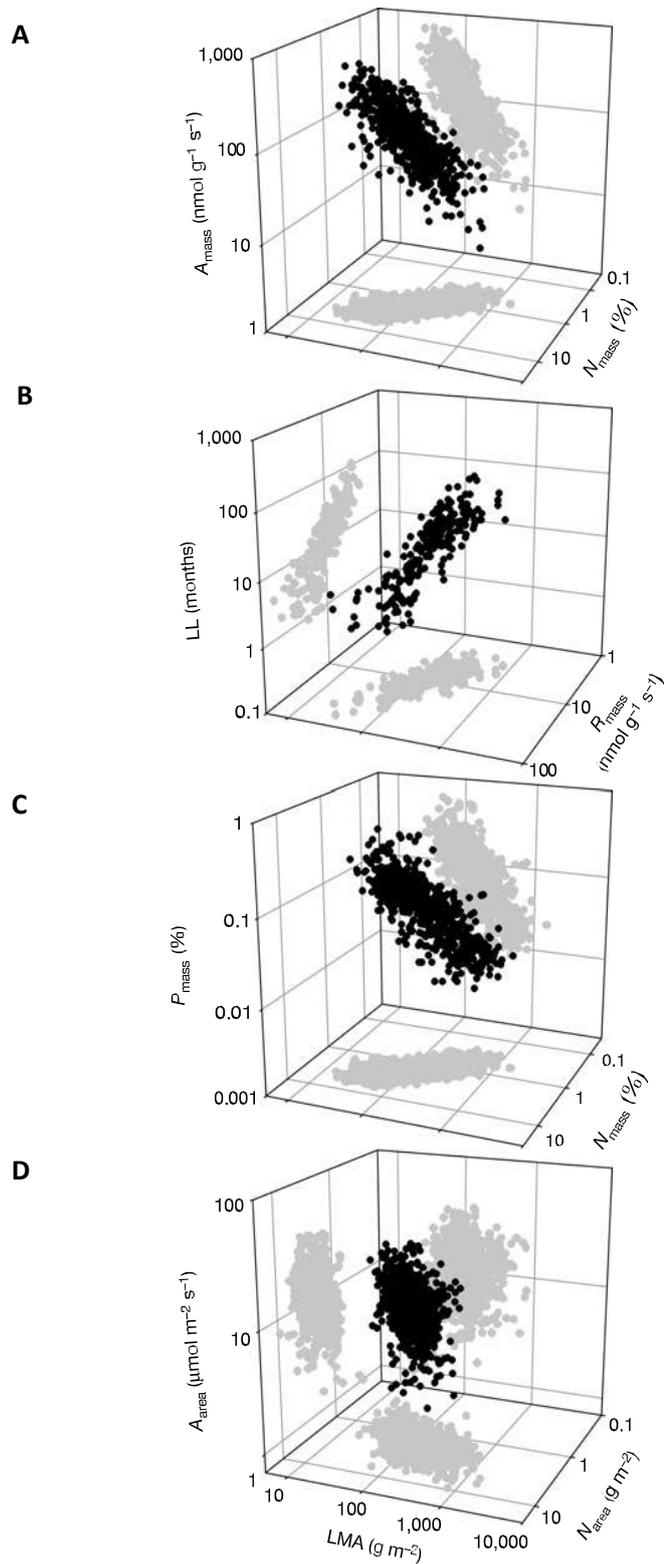
L'analyse génétique à l'aide de paramètres de diversité (π , FST et D de Tajima) des locus liés à ce trait sur les deux formes : *T. turgidum ssp dicocoides* (Zavitan) et *T. turgidum ssp durum* (Svevo), a montré un résultat typique d'un trait phénotypique associé à un goulot d'étranglement de domestication (Avni et al. 2017).

1.2.4.2 La solidité des glumes

Le trait phénotypique caractéristique de cette deuxième phase de domestication, de *T. turgidum* ssp *dicoccum* vers *T. turgidum* ssp *durum* est l'apparition des **grains nus**. Une analyse menée par Peleg *et al.* (2011), des locus de caractères quantitatifs (QTL) à l'aide d'une population de cartographie issue d'un croisement entre un blé dur cultivé et un blé sauvage, a montré que le phénotype de « grain nu » était contrôlé par six QTL. Deux des principaux QTL qui confèrent un battage facile, expliquent la majorité de la variation phénotypique. Le premier, partiellement récessif, est le locus « Tg » (Tenacious glume), situé sur le bras court du chromosome 2B, et contrôle la ténacité de la glume (Simonetti *et al.* 1999; Jantasuriyarat *et al.* 2004). Le deuxième est le locus « Q », partiellement dominant, localisé sur le bras long du chromosome 5A. Le clonage moléculaire du locus « Q » a permis de déterminer qu'il appartient à la famille des facteurs de transcription (AP2) et que les allèles Q et q ne diffèrent que par un seul nucléotide (SNP) (Faris *et al.* 2003; Simons *et al.* 2006). Cette mutation conduit à la substitution de l'acide aminé valine par l'isoleucine et modifie l'expression du microRNA172 situé sur l'exon 10 (Simons *et al.* 2006; Zhang *et al.* 2011). Par ailleurs les analyses montrent que la mutation de l'allèle primitif q conduisant à l'allèle Q ne s'est produite qu'une seule fois au cours de l'histoire évolutive du blé dur et lors du passage de *T. turgidum* ssp *dicoccum* à *T. turgidum* ssp *durum*. La classe de facteurs de transcription (AP2) à laquelle appartient le gène Q est bien connue pour son rôle dans la régulation des méristèmes floraux (Irish and Sussex 1990; Bowman *et al.* 1993), dans la mise en place des organes floraux (Komaki *et al.* 1988; Bowman *et al.* 1989; Kunst *et al.* 1989; Jofuku *et al.* 1994; Yant *et al.* 2010) et dans la régulation spatiale et temporelle de l'expression des gènes impliqués dans l'homéostasie florale chez *Arabidopsis* (Drews *et al.* 1991). Les orthologues de ce facteur de transcription AP2 chez d'autres plantes, tels que le riz (*Oryza sativa* L.) ou le maïs (*Zea mays* L.) fonctionnent de la même manière au niveau du développement de l'inflorescence (Lee *et al.* 2007; Chuck *et al.* 2008). Chez le blé, la mutation de l'allèle q vers l'allèle Q a modifié complètement la morphologie des épis en affectant, de manière pléiotropique la forme et la ténacité des glumes ainsi que de nombreux autres caractères phénotypiques (Simonetti *et al.* 1999; Jantasuriyarat *et al.* 2004; Simons *et al.* 2006). Parmi ces traits, l'allèle Q a modifié la morphologie des glumes et du méristème qui permet la fixation des épillets le long du rachis. En effet, l'allèle q confère un méristème solide et large permettant la non-déhiscence des grains, carénés par des glumes épaisses, à maturité et provoque la désarticulation du rachis lors du battage. L'allèle Q confère, quant à lui, des glumes plus souples et réduit la taille du méristème à une petite zone à la base des glumes, ce qui permet un battage plus facile et un maintien du rachis intact lors du battage (Zhang *et al.* 2011).

L'interaction entre ces deux locus est épistatique car il est nécessaire d'avoir l'allèle tg^{2B} ainsi que l'allèle Q^{5A} pour conférer le phénotype « grains nus » (Matsuoka 2011).

Chez les céréales, les modifications dans l'architecture de l'inflorescence peuvent modifier la fertilité florale, la croissance des graines et, par conséquent, le rendement final en grains. Chez le blé, deux études menées par Xie *et al.*, (2015; 2018) ont examiné les effets de l'allèle Q sur l'augmentation du rendement (poids de mille grains (PMG), le rendement en grains ($y=M/S$) ainsi que le nombre de grains au m^2). La modification de la morphologie des méristèmes permettant la fixation des épillets le long du rachis a permis, de façon pléiotropique, d'augmenter le nombre d'épillets par méristème et donc, le nombre de fleurs par épillets (de 3 à 5). Il agit également sur la forme des grains en augmentant leur taille (rondeur).



Wright et al., 2004

Figure 17 : Corrélations entre les traits foliaires pour la définition du « spectre d'économie foliaire ».

A : relation de traits à trois voies entre : le taux d'assimilation photosynthétique (A_{mass}), Le taux de respiration de la feuille (R_{mass}) et la densité foliaire (LMA)

B : relation de traits à trois voies entre : La durée de vie des feuilles (LL), Le taux de respiration de la feuille (R_{mass}) et la densité foliaire (LMA).

C : relation de traits à trois voies entre : Le taux d'azote dans les feuilles (N_{mass}), le taux de Phosphore dans les feuilles (P_{mass}) et la densité foliaire (LMA).

D : relation de traits à trois voies entre : La quantité d'azote dans les feuilles (N_{area}), l'activité photosynthétique (A_{area}) et la densité foliaire (LMA).

1.2.4.3 La hauteur des plantes

La sélection moderne et la révolution verte sont à l'origine des variétés modernes qui se caractérisent, entre autre, par des plantes courtes. L'introduction du trait phénotypique « **semi-nain** » dans les cultivars de blé est une étape déterminante de la « révolution verte » (Hedden 2003). Les tiges hautes de blé n'étant pas assez solides pour supporter les épis des variétés à rendement élevé, les tiges se couchent à maturité, c'est ce qu'on appelle la verse. Ce phénomène entraîne d'importantes pertes de rendement, ainsi qu'une sensibilité à certaines maladies. Des plantes semi-naines ont été obtenues, dans les années 1960 et 1970, par croisement interspécifique avec des lignées « Norin 10 » de blé tendre, d'origine japonaise, porteuses de deux gènes de nanismes. Les gènes responsables de ce trait phénotypique sont les gènes **Rht**. Parmi eux, Rht1, utilisé très majoritairement pour réduire la taille chez le blé dur provoque une réponse réduite à l'hormone végétale : l'acide gibbérellique (GA). L'allèle Rht-B1b est situé sur le chromosome 4B et agit comme un répresseur de croissance (insensibles à l'AG) en diminuant la hauteur de la plante de 20% (Peng et al. 1999).

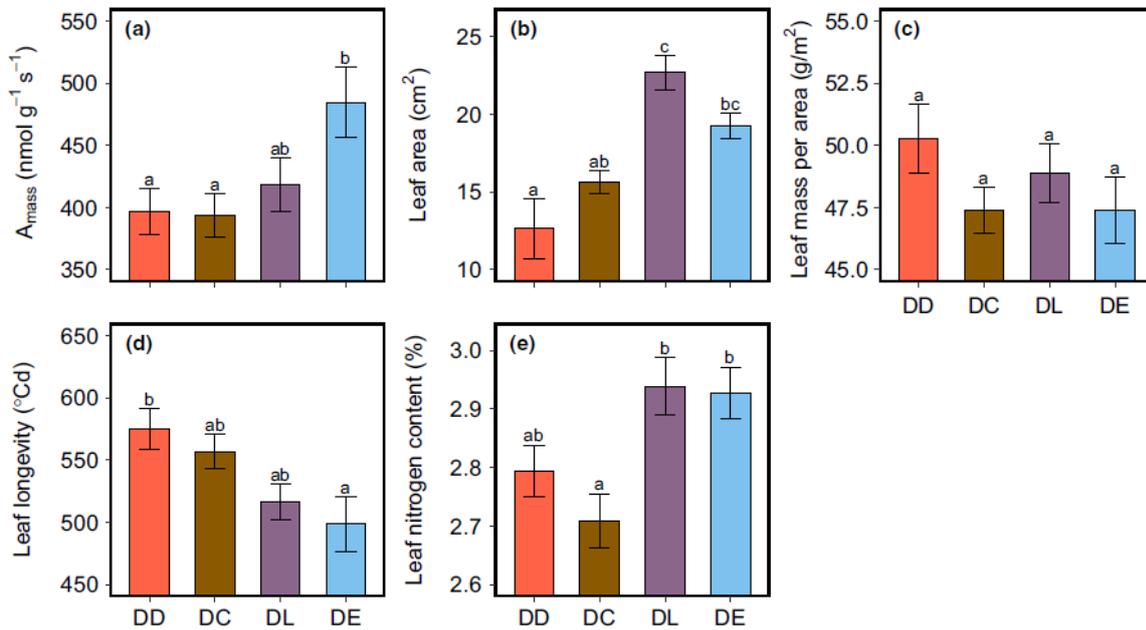
1.2.4.4 Le poids des grains

Le poids des grains est l'un des composants majeurs du rendement chez le blé dur. En effet, il contribue à la vigueur des plantules, pouvant conduire à une augmentation du rendement en grain et donc du rendement en semoule, qui est un paramètre économique très important pour les industriels de la filière blé dur. Il est estimé par le poids de mille grain (noté PMG). Les mesures effectuées sur les céréales anciennes démontrent que le poids des grains de blé sont restées essentiellement les mêmes entre 9 500 et 6 500 avant J.C., avec un PMG à environ 35 grammes (Willcox 2004). L'augmentation du poids des grains a été lente malgré l'héritabilité relativement élevée de ce caractère. Plusieurs QTLs contrôlant l'expression de ce caractère ont été identifiés sur de nombreux chromosomes (Peng et al. 2003). La taille et le poids des grains dépendent également de l'environnement comme la disponibilité en eau et la température pendant le remplissage de grain.

Au cours de la domestication, l'amélioration de ce trait est un processus lent mais continu. A partir de la fin du 19^{ème} siècle, la mise en place d'une sélection généalogique a permis de répondre au besoin de l'industrie semoulière alors en plein essor. Elle infléchit notamment la pression de sélection sur ce caractère favorisant l'émergence de variétés à gros grains dont le rapport amande/enveloppe est un facteur de rentabilité important. Le PMG est actuellement entre 50 et 60 grammes pour les formes « élites » de *T. turgidum ssp durum*.

1.2.1.1 La teneur en azote contenu dans la feuille

Alors que l'impact de la domestication a été très documenté pour certains traits, pour d'autres, l'information disponible sur l'ensemble des traits impliqués dans l'efficacité de l'acquisition des ressources par la plante (feuilles, racines) reste sporadique. C'est le cas des traits foliaires impliqués dans la photosynthèse, tel que le taux d'assimilation photosynthétique ou la teneur en azote foliaire (N). Le « spectre d'économie foliaire » met en évidence la corrélation de différents traits foliaires caractéristiques du système d'acquisition des ressources chez les plantes (Wright et al. 2004) (figure 17).



Roucou et al., 2017

Figure 18 : Variations des traits foliaires sur les quatre grandes formes évolutives de blé tétraploïdes

Cinq traits foliaires ont été mesurés : (a) le taux maximum de photosynthèse, (b) la surface foliaire, (c) la masse foliaire par surface (densité foliaire), (d) la durée de vie des feuilles, (e) la teneur en azote par unité de masse sèche de feuille.

Ces traits ont été mesurés sur 10 accessions de chaque forme évolutive au cours de la domestication. En rouge : *T. turgidum* ssp. *Dicoccoïdes* (DD), en marron : *T. turgidum* ssp. *Dicoccum* (DC), en violet: *T. turgidum* ssp. *Durum* avant la révolution verte, appelés « Landrace » ou populations de pays (DL) et en bleu : *T. turgidum* ssp. *Durum* après la révolution verte, appelés « élite »(DE).

Une étude récente, publiée par Roucou et al. (2018), a documenté des modifications du mode d'acquisition des ressources au cours de la domestication du blé dur. L'étude s'est effectuée sur 40 accessions appartenant aux quatre grandes formes évolutives : *T. turgidum* ssp *dicoccoïdes*, *T. turgidum* ssp *dicoccum*, *T. turgidum* ssp *durum* avant la révolution verte et *T. turgidum* ssp *durum* après la révolution verte. Les mesures de différents traits foliaires (figure 18) et une analyse en composante principale a permis d'inscrire l'ensemble de ces accessions dans le « spectre d'économie foliaire ».

La forme *T. turgidum* ssp *dicoccoïdes* a des feuilles plutôt épaisses, une capacité photosynthétique faible et une capacité de stockage des ressources importante, ce qui correspond dans la théorie du « leaf economy spectrum » à une stratégie d'acquisition lente. Les variétés élites de *T. turgidum* ssp *durum* cultivées actuellement se situent à l'autre extrémité du spectre d'économie foliaire, avec des feuilles fines ayant une forte capacité photosynthétique et une faible capacité de stockage des ressources.

1.3 Les techniques de génotypage au service de l'étude de la diversité génétique

Les techniques de génotypage sont, depuis le début de la biologie moléculaire, des outils de soutien importants pour des programmes de recherche en génétique, d'amélioration des plantes, ou comme ici, d'étude de la diversité génétique. La description de la diversité génétique au sein de l'espèce *Triticum turgidum* a été démarrée, dans l'équipe, avec l'utilisation de 15 marqueurs neutres de type microsatellites (Thuillet et al. 2005), puis à l'aide des séquences de 21 gènes (Haudry et al. 2007). Aujourd'hui, l'évolution des techniques moléculaires et l'apparition des technologies de génotypage et de séquençage haut-débit de nouvelle génération permet de produire des données massives et précises de génomique et de transcriptomique.

Etant donné sa taille et sa complexité, le génome du blé dur peut être étudié plus facilement après lui avoir fait subir une réduction génomique. Actuellement, deux techniques sont utilisées principalement : les puces à ADN et l'enrichissement par capture.

Les puces à ADN (Axiom Affymetrix par exemple), permettent d'analyser le polymorphisme au niveau de plusieurs centaines de milliers de SNPs grâce à une détection de fluorescence issue d'une hybridation. Cette technique est fréquemment utilisée (Balfourier et al. 2019; Maccaferri et al. 2019) mais possède toutefois des limites car il est impossible de détecter un polymorphisme s'il n'est pas présent dans la référence choisie pour cibler des SNPs bi-alléliques. De ce fait, l'utilisation des puces à ADN pour les études de diversité nécessite de connaître préalablement le polymorphisme présent dans la variété ancestrale qui servira de référence afin de ne pas risquer de sous-estimer la diversité génétique.

La deuxième technique est l'enrichissement par capture basée sur le séquençage, dont le principe est d'utiliser des sondes pour capturer, par hybridation, des régions génomiques cibles. Les zones ciblées peuvent être des régions impliquées dans un trait phénotypique d'intérêt agronomique, mais également des zones génomiques enrichies en séquences codantes (Holtz et al. 2017). Les sondes de capture correspondent à des oligonucléotides de 120pb qui sont des molécules d'ARN biotinylées, complémentaires des séquences d'intérêt, permettant un enrichissement rapide et sélectif de la cible pour un séquençage de nouvelle génération (NGS). Pour étudier la diversité génétique, les séquences de 120pb peuvent être définies pour leur capacité à révéler un polymorphisme de type SNP (Holtz et al. 2016). L'avantage de cette technique, par rapport aux puces à ADN, est que le séquençage complet des fragments capturés (300-400pb) permet de révéler d'autres SNPs, que ceux initialement ciblés, qui seront eux, non soumis à des biais (« ascertainment-bias ») créés lors de la constitution des sondes (Cavanagh et al. 2013; McTavish and Hillis 2015; Malomane et al. 2018).

Présentation du sujet d'étude

C'est dans ce contexte que s'inscrit mon étude, qui a pour objectifs (1) d'adapter la technique d'enrichissement par capture aux contraintes génomiques du blé dur, afin de pouvoir (2) affiner nos connaissances sur son histoire évolutive en décrivant la diversité génétique des quatre formes : *T. turgidum* ssp *dicoccoïdes*, *T. turgidum* ssp *dicoccum*, *T. turgidum* ssp *durum* avant la révolution verte et *T. turgidum* ssp *durum* après la révolution verte pour estimer l'impact démographique de la domestication. Nous chercherons ensuite à (3) détecter les signatures génétiques de sélection, sur les gènes impliqués dans le contrôle génétique de traits phénotypiques caractéristiques de la domestication (Br, Q, Rht), mais aussi des QTLs associés à deux autres traits phénotypiques : le poids des grains et la teneur en azote foliaire.

Matériel et méthodes

Tableau 1: Données passeport des 120 génotypes

Ce tableau rassemble, pour chacun des génotypes sélectionnés, la sous-espèce, le groupe évolutif, la collection donc il est issu accompagnée du numéro d'accession dans cette collection, le pays d'origine avec la latitude et la longitude, le numéro d'accession dans la collection de Montpellier et pour finir le code ADN qui sera utilisé lors des analyses génétiques. Pour certains génotypes, les données d'origine ne sont pas disponible.

Sous-espèces	groupes	collections	N° accessions	pays d'origine	longitudes	latitudes	N° acc. Coll. Mtp	Numéros ADN
Triticum turgidum ssp. dicoccoides	DD	ICARDA	113302	Iran			44	Tc2220
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46470	Syrie	36.5	32.42	69	Tc2224
Triticum turgidum ssp. dicoccoides	DD	ICARDA	116172	Turquie			80	Tc2226
Triticum turgidum ssp. dicoccoides	DD	USDA	467014	Israël	35.32	32.52	85	Tc2227
Triticum turgidum ssp. dicoccoides	DD	USDA	428133	Liban			93	Tc2228
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46516	Syrie	36.25	32.48	96	Tc2229
Triticum turgidum ssp. dicoccoides	DD	USDA	487255	Syrie	36.5	33.45	99	Tc2230
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46499	Jordanie	35.44	32.06	48	Tc2385
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46518	Syrie	36.16	33.03	64	Tc2391
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46501	Syrie	36.46	32.48	70	Tc2393
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46244	Turquie	39.27	37.43	78	Tc2398
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46491	Jordanie	35.49	32.31	50	Tc2435
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46323	jordanie			57	Tc2438
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46309	Palestine			62	Tc2442
Triticum turgidum ssp. dicoccoides	DD	ICARDA	119437	Syrie	36.42	35	66	Tc2443
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46453	Syrie			71	Tc2445
Triticum turgidum ssp. dicoccoides	DD	USDA	481499	Israël			87	Tc2451
Triticum turgidum ssp. dicoccoides	DD	USDA	352324	Liban			90	Tc2454
Triticum turgidum ssp. dicoccoides	DD	CIMMYT	cwi19112	Israël			108	Tc2460
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46253	Turquie	40.33	37.55	73	Tc3203
Triticum turgidum ssp. dicoccoides	DD	ICARDA	116179	jordanie			56	Tc3301
Triticum turgidum ssp. dicoccoides	DD	ICARDA	116179	Turquie	37.19	37.15	75	Tc3302
Triticum turgidum ssp. dicoccoides	DD	USDA	428105	Israël	35.32	32.58	82	Tc3303
Triticum turgidum ssp. dicoccoides	DD	ICARDA	117894				95	Tc3305
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46310	Palestine			63	Tc3309
Triticum turgidum ssp. dicoccoides	DD	ICARDA	113301	Iran	46.27	33.37	43	Tc3401
Triticum turgidum ssp. dicoccoides	DD	ICARDA	45963	Jordanie	35.54	31.47	46	Tc3402
Triticum turgidum ssp. dicoccoides	DD	ICARDA	111000	Jordanie	35.45	32.02	49	Tc3403
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46391	Jordanie	35.55	32.1	54	Tc3404
Triticum turgidum ssp. dicoccoides	DD	ICARDA	46528	Liban			58	Tc3405
Triticum turgidum ssp. Dicoccum	DC	CIMMYT	cwi17084	Iraq			208	Tc2208
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45239	Italie	15.06	41.17	130	Tc2211
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45354	Russie	39.52	57.36	137	Tc2212
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45280	Slovaquie			139	Tc2213
Triticum turgidum ssp. Dicoccum	DC	USDA	352365	Allemagne			157	Tc2215
Triticum turgidum ssp. Dicoccum	DC	USDA	415152	Israël			184	Tc2218
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45318	Afghanistan			110	Tc2462
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45377	Herzégovine	18.58	43.23	115	Tc2465
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45245	Italie	15.34	40.32	129	Tc2476
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45441	Syrie			143	Tc2484
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45408	Ukraine			150	Tc2486
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45357	URSS			151	Tc2487
Triticum turgidum ssp. Dicoccum	DC	CIMMYT	cwi17083				154	Tc2489
Triticum turgidum ssp. Dicoccum	DC	CIMMYT	cwi44340				155	Tc2490
Triticum turgidum ssp. Dicoccum	DC	USDA	275999	Espagne	-5.46	43.4	167	Tc2501
Triticum turgidum ssp. Dicoccum	DC	USDA	94617	Russie	47	42	172	Tc2503
Triticum turgidum ssp. Dicoccum	DC	USDA	591868	Georgie	42.37	43	176	Tc2506
Triticum turgidum ssp. Dicoccum	DC	USDA	94623	Iran			180	Tc2510
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45444	Palestine			190	Tc2517
Triticum turgidum ssp. Dicoccum	DC	USDA	362071	Roumanie			193	Tc2520
Triticum turgidum ssp. Dicoccum	DC	USDA	319868	Turquie			202	Tc2528
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45383	Bulgarie			116	Tc3205
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45483	Iran			125	Tc3306
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45351	Iran			128	Tc3314
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45334	Arménie			113	Tc3406
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45095	Arménie			114	Tc3407
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45413	Bulgarie			117	Tc3408
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45332	Bulgarie			118	Tc3409
Triticum turgidum ssp. Dicoccum	DC	ICARDA	45087	Turquie			146	Tc3410
Triticum turgidum ssp. Dicoccum	DC	Gasterleben		Espagne?			Dic2	Td4005

Sous-espèces	groupes	collections	N° accessions	pays d'origine	longitudes	latitudes	N° acc. Coll. Mtp	Numéros ADN
Triticum turgidum ssp. Durum "population"	DP	Montpellier	Senatore Cappelli (21)	Italie			278	Tc2235
Triticum turgidum ssp. Durum "population"	DP	ICARDA	97210	jordanie	35.44	30.11	290	Tc2236
Triticum turgidum ssp. Durum "population"	DP	ICARDA	99313	Liban	35.49	33.37	296	Tc2237
Triticum turgidum ssp. Durum "population"	DP	ICARDA	84849	Liban			299	Tc2238
Triticum turgidum ssp. Durum "population"	DP	Zaharieva M.	LR B6R 5546 / 49 / 4	Bulgarie			585	Tc2246
Triticum turgidum ssp. Durum "population"	DP	Montpellier	kubanka (250)	Russie			233	Tc2248
Triticum turgidum ssp. Durum "population"	DP	ICARDA	84866	Syrie			302	Tc2249
Triticum turgidum ssp. Durum "population"	DP	ICARDA	95920	Syrie	36.28	34.48	328	Tc2250
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82697	Turquie	31.32	41.02	344	Tc2251
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82726	Turquie	31.51	37.25	351	Tc2252
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82715	Turquie	34.51	38.45	354	Tc2253
Triticum turgidum ssp. Durum "population"	DP	Zaharieva M.	B6R	Bulgarie			581	Tc2262
Triticum turgidum ssp. Durum "population"	DP	Montpellier	Mindum (1)	USA			237	Tc2549
Triticum turgidum ssp. Durum "population"	DP	ICARDA	97225	jordanie	35.39	30.4	293	Tc2586
Triticum turgidum ssp. Durum "population"	DP	ICARDA	84871	Syrie			305	Tc2595
Triticum turgidum ssp. Durum "population"	DP	ICARDA	97512	Syrie			312	Tc2601
Triticum turgidum ssp. Durum "population"	DP	ICARDA	97511	Syrie	37.09	36.13	319	Tc2607
Triticum turgidum ssp. Durum "population"	DP	ICARDA	92260	Syrie	36.34	35.15	323	Tc2611
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82762	Turquie	35.35	36.57	341	Tc2627
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82702	Turquie	40.39	37.5	346	Tc2630
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82768	Turquie	37.07	36.43	349	Tc2632
Triticum turgidum ssp. Durum "population"	DP	Zaharieva M.	B6R	Bulgarie			582	Tc2808
Triticum turgidum ssp. Durum "population"	DP	Zaharieva M.	LR B6R 5687 / 109 / 4	Bulgarie			593	Tc2818
Triticum turgidum ssp. Durum "population"	DP	ICARDA	82700	Turquie	40.16	37.22	353	Tc3197
Triticum turgidum ssp. Durum "population"	DP	Montpellier	edmore (495)	Usa			250	Tc3307
Triticum turgidum ssp. Durum "population"	DP	Montpellier	oued zenati (542)	Algerie			227	Tc4265
Triticum turgidum ssp. Durum "population"	DP	Montpellier	Biskri Bouteille (519)	Algerie			238	Tc4280
Triticum turgidum ssp. Durum "population"	DP	Montpellier	mohammed ben bachir	Algerie			256	Tc4300
Triticum turgidum ssp. Durum "population"	DP	Montpellier	Mahmoudi 981 (535)	Algerie			267	Tc4306
Triticum turgidum ssp. Durum "population"	DP	Montpellier	Sentry (419)	Usa			274	Tc4309
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ARDENOIS	France			459	Tc2245
Triticum turgidum ssp. Durum "elite"	DE	GEVES	BRUMAIRE	France			364	Tc2254
Triticum turgidum ssp. Durum "elite"	DE	GEVES	PRIMADUR	France			395	Tc2255
Triticum turgidum ssp. Durum "elite"	DE	GEVES	VILLEMUR	France			434	Tc2258
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ARMET	France			435	Tc2259
Triticum turgidum ssp. Durum "elite"	DE	GEVES	DURENTAL	France			437	Tc2260
Triticum turgidum ssp. Durum "elite"	DE	GEVES	AGATHE	France			362	Tc2641
Triticum turgidum ssp. Durum "elite"	DE	GEVES	DURTAL	France			363	Tc2642
Triticum turgidum ssp. Durum "elite"	DE	GEVES	DURGAMM	France			374	Tc2651
Triticum turgidum ssp. Durum "elite"	DE	GEVES	SAFARI	France			378	Tc2655
Triticum turgidum ssp. Durum "elite"	DE	GEVES	CHANDUR	France			379	Tc2656
Triticum turgidum ssp. Durum "elite"	DE	GEVES	NITA	France			380	Tc2657
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ROMEO	France			382	Tc2658
Triticum turgidum ssp. Durum "elite"	DE	GEVES	DURANDAL	France			388	Tc2664
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ARDENTE	France			393	Tc2668
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ARTENA	France			397	Tc2671
Triticum turgidum ssp. Durum "elite"	DE	GEVES	CARGIVOX	France			400	Tc2674
Triticum turgidum ssp. Durum "elite"	DE	GEVES	DURELLE	France			401	Tc2675
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ARBOIS	France			405	Tc2679
Triticum turgidum ssp. Durum "elite"	DE	GEVES	BIODUR	France			407	Tc2681
Triticum turgidum ssp. Durum "elite"	DE	GEVES	KEOPS	France			422	Tc2693
Triticum turgidum ssp. Durum "elite"	DE	GEVES	MINODUR	France			432	Tc2701
Triticum turgidum ssp. Durum "elite"	DE	GEVES	SYENE	France			444	Tc2709
Triticum turgidum ssp. Durum "elite"	DE	GEVES	ORJAUNE	France			464	Tc2727
Triticum turgidum ssp. Durum "elite"	DE	GEVES	SACHEM	France			485	Tc2746
Triticum turgidum ssp. Durum "elite"	DE	GEVES	MEXIDUR	France			474	Tc3411
Triticum turgidum ssp. Durum "elite"	DE	GEVES	NEODUR	France			411	Td4001
Triticum turgidum ssp. Durum "elite"	DE	GEVES	IXOS	France			430	Td4002
Triticum turgidum ssp. Durum "elite"	DE	Montpellier	Lloyd (747)	Usa			254	Td4003
Triticum turgidum ssp. Durum "elite"	DE	RAGT	Karur	France				Tf1150

2. Matériel et Méthodes

2.1 Matériel végétal

Pour répondre aux objectifs de cette étude, nous avons sélectionné 120 génotypes représentant les trois principales étapes de la domestication du blé tétraploïde *Triticum turgidum* : la forme sauvage *T. turgidum* ssp *dicocoides* (notée **DD**), la première forme domestiquée : *turgidum* ssp *dicoccum* (notée **DC**) et la deuxième forme cultivée à grain nu : *T. turgidum* ssp *durum*. Cette dernière sous-espèce a été divisée en deux sous-groupes en fonction qu'il s'agisse de variétés issues de la période pré ou post-révolution verte. Le premier groupe se compose de lignées issues de variétés locales appelées « population de pays » (noté **DP**) et le second de variétés « élites » enregistrées en Europe après la Révolution verte (de 1970 à 1990) (noté **DE**).

La sélection des 30 génotypes par groupes génétiques, a été effectuée en maximisant la diversité génétique de chaque groupe à l'aide du logiciel MSTRAT (Gouesnard et al. 2001). L'échantillonnage a été effectué dans une core-collection de 600 accessions (INRA-Montpellier) génotypées à l'aide de 15 marqueurs microsatellites non liés, cartographiés sur les 14 chromosomes du blé dur (Röder et al. 1998; Sahri et al. 2014) (tableau 1).

L'ensemble des génotypes, quel que soit le groupe évolutif auquel ils appartiennent, sont issus de collections de semences conservées dans différents centres de stockage tel que l'ICARDA (Centre international de recherche agricole dans les zones arides), le CYMMIT (Centre international d'amélioration du maïs et du blé), l'USDA (Département de l'Agriculture des États-Unis), le GEVES (Groupe d'Etude et de contrôle des Variétés Et des Semences) ou l'INRA de Montpellier. Pour chaque génotype, les semences provenaient d'autofécondations successives afin de limiter l'hétérozygotie résiduelle et de garantir la fixation génétique du matériel. Les semences ont été produites par l'INRA-UMR AGAP à Montpellier.

Par ailleurs, des mesures morphologiques ont été réalisées sur les 120 génotypes qui composent le matériel expérimental de cette étude :

- ✓ La solidité du rachis en donnant une mesure de 1 (solide) à 5 (cassant)
- ✓ La solidité des glumes en donnant une mesure de 1 (grain nu) à 5 (grain vêtu)
- ✓ La hauteur des plantes à maturité
- ✓ Le poids des grains (poids de 1000 grains – PMG)
- ✓ La teneur en azote contenu dans la feuille au cours de la phase de montaison

Les analyses de structure génétique, les estimations de la perte de diversité au cours de la domestication, ainsi que la détection de signatures de sélection ont été réalisés sur ces quatre groupes évolutifs.

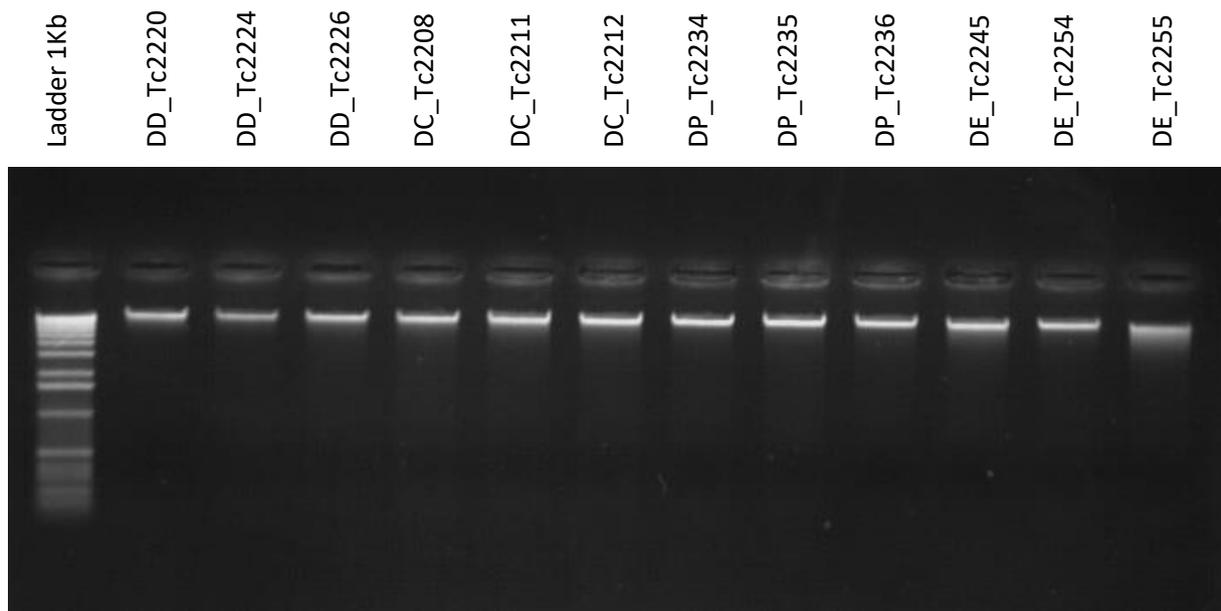


Figure 19: Visualisation des ADN sur gel d'agarose après coloration au BEt
 Vérification de la qualité des ADN après extraction, par dépôt sur une gel d'agarose à 1.5%.
 Les trois géotypes par groupes qui sont présentés ici ne sont pas dégradés car leur poids moléculaire est supérieur à 40Kb.

2.2 Outils moléculaires

Afin de réduire la complexité génomique du blé, nous avons choisi d'utiliser l'enrichissement par capture. Pour cela, il est nécessaire, d'une part, d'extraire l'ADN génomique de chacun des échantillons et de construire des bibliothèques génomiques, et d'autre part, de définir des sondes qui permettent de cibler des régions codantes portant un polymorphisme de type SNP.

Plusieurs expérimentations préliminaires ont été réalisées au laboratoire dans le but d'obtenir un protocole expérimental optimisé, présenté dans cette partie.

2.2.1 Extraction d'ADN

Pour chacun des 120 géotypes, l'ADN génomique a été extrait à partir de 50 mg de matériel végétal. Les prélèvements de tissu foliaire ont été effectués sur les plantes avant l'épiaison, afin d'assurer la bonne viabilité des cellules présentes dans les tissus foliaires. Le matériel végétal prélevé a été conditionné en tube de 2mL et stocké dans un congélateur à -80°C.

Après broyage dans l'azote liquide, les membranes cellulaires ont été lysées à l'aide d'un tampon contenant différents composants (SDS, CTAB, EDTA, TRIS, NaCl, PVP40000) pour extraire les ADN sans risquer de les dégrader. A cette étape, un traitement RNase va dégrader les ARN afin qu'ils n'interfèrent pas avec l'ADN lors des expérimentations suivantes. Pour assurer une lyse optimale des cellules et une dégradation complète des ARN, la réaction a été placée pendant 30 minutes à 65°C. Afin de précipiter les protéines, une forte concentration en sels (Acétate de potassium) a été ajoutée au lysat. Après centrifugation, (10 min à 12500 rpm à 4°C), le culot était composé de l'ensemble des débris cellulaires et protéines précipitées alors que le surnageant contenait l'ADN.

La purification des ADN est réalisée sur un robot KingFisher™ Flex Purification System grâce à des billes métalliques et magnétiques coatées avec de la silice (Billes Perkin-Elmer Chemagen). L'adsorption des molécules d'ADN à la silice s'effectue grâce au chlorure de guanidium et en présence d'éthanol (déshydratation). Cinq lavages successifs sont réalisés à base de perchlorate de sodium, d'éthanol et de Triton. Les ADN purifiés sont finalement élués dans de l'eau ultra-pure (annexe 1).

Après extraction, la qualité des ADN (poids moléculaire supérieur à 40Kb) a été vérifiée par migration sur un gel d'agarose à 1.5 % (polymère linéaire) (figure 19). Sur l'ensemble des 120 ADN extraits, les ratios d'absorbance moyens entre 260/280 nm et 260/230 nm étaient respectivement de 1.89 et 2.12, signifiant que, d'une part, les protéines avaient bien été éliminées et que, d'autre part, l'ADN avait été correctement purifié avec élimination de l'ensemble des solvants, sels et contaminants organiques. Le dosage quantitatif des ADN est réalisé grâce à une molécule fluorescente qui se fixe spécifiquement sur les doubles brins d'ADN : le Hoescht 33258 (famille des bisbenzimidés). L'intensité du spectre d'émission de la fluorescence est mesurée sur un spectro-fluorimètre InfiniteM200™ TECAN. Pour nos échantillons, la moyenne des concentrations est de 122 ng/μL, quantité satisfaisante pour la suite des expérimentations.

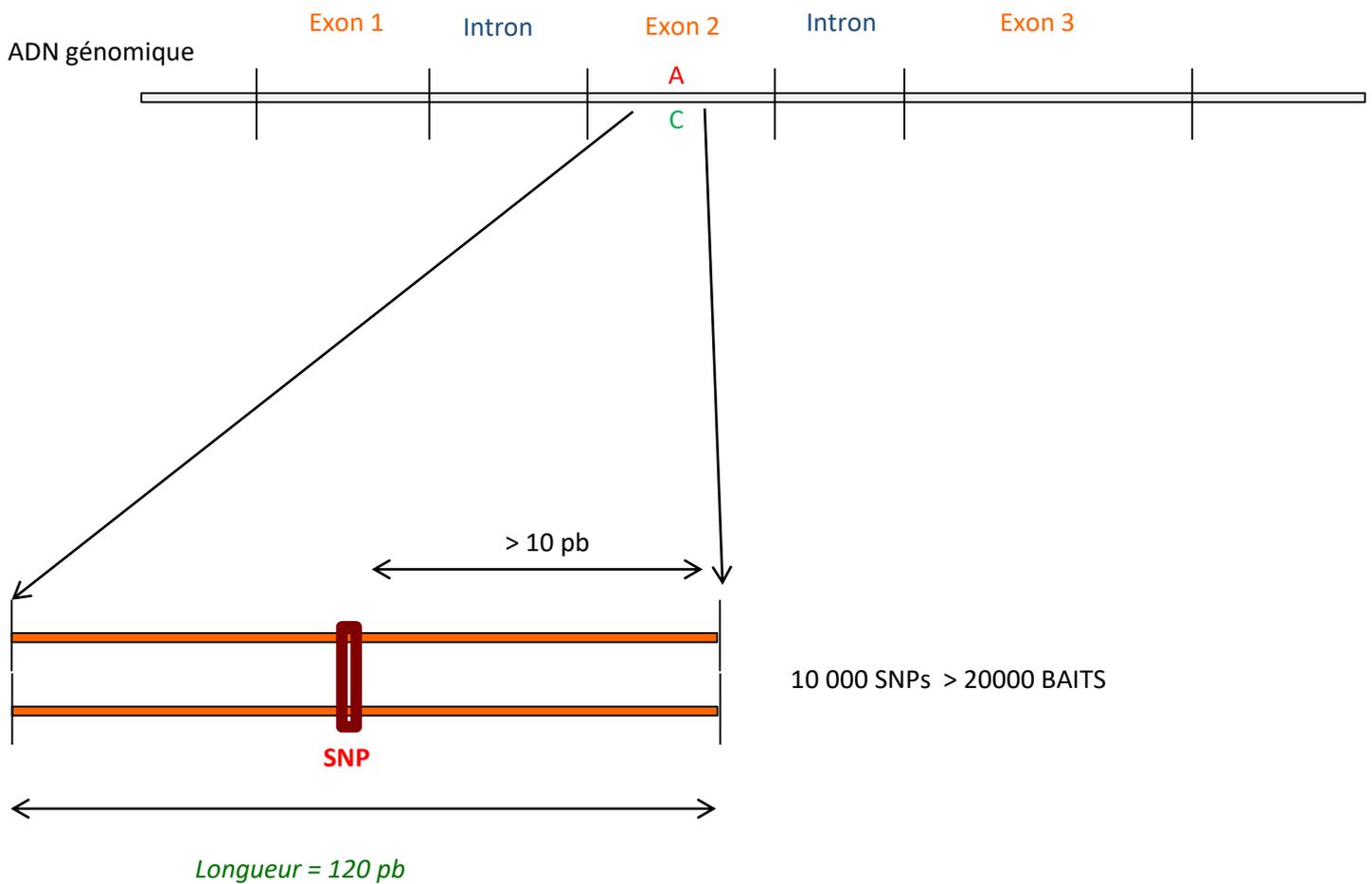


Figure 20: Définition des baits de capture

Les baits qui serviront à la capture sont des séquences de 120pb. 10 000 SNPs sont ciblés dans les régions exoniques (zone codante). Pour chaque SNP, deux baits sont constitués, chacune portant une version de l'allèle. Les séquences des 20 000 baits ont été choisies de manière à ce que le SNP ciblé soit situé à au moins 10pb des extrémités du baits afin d'optimisé sa capture.

2.2.2 Sondes

Les sondes de capture, correspondant à des oligonucléotides de 120pb, ont été choisies pour leur capacité à révéler un polymorphisme de type SNP. Chaque SNP est porté par deux sondes correspondant aux deux variants possibles au SNP.

Nous avons utilisé un lot de sondes défini, au sein de l'équipe, pour un projet antérieur. Tout d'abord, des polymorphismes de type SNP ont été identifiés en séquençant le transcriptome de plusieurs génotypes dont quatre variétés élités de *turgidum ssp durum* : Lloyd, Silur, Pescadou, Soldur et une lignée de *T. turgidum ssp dicocum* : DIC2, très utilisée au laboratoire comme géniteur dans des croisements d'étude. L'étape suivante est de sélectionner les polymorphismes situés dans les parties codantes (exons, UTR exclus) (David et al. 2014; Holtz et al. 2016) (figure 20). Enfin, des séquences de 20 000 sondes de 120pb encadrant 10 000 SNP sélectionnés ont été définies en respectant les règles thermodynamiques pour optimiser les capacités d'hybridation des sondes. Dans la mesure du possible nous avons vérifié que ces séquences ne correspondaient pas à des séquences répétées (familles de gènes paralogues) et ne portaient pas des motifs microsatellites. Le SNP est situé au milieu de la séquence de 120 pb afin de maximiser la probabilité d'hybridation avec la cible à cette position. Ce type de sonde permet une hybridation spécifique avec les cibles, tout en acceptant une certaine non-homologie, estimée entre 4 et 5 %, ce qui correspond à une divergence de 4 à 5 pb entre la cible et la sonde. Cette flexibilité a pour objectif de capturer des régions cibles malgré la variabilité entre les différentes sous-espèces. Les sondes MyBaits™ ont été synthétisées par la plateforme Arbor-Biosciences (Michigan, USA).

2.2.3 Bibliothèques génomiques

Elles ont été préparées à l'aide du protocole de Rohland et al. (2012), avec quelques modifications dans un but d'optimisation technique (schéma récapitulatif : figure 21, protocole détaillé : annexe 2). Pour chaque échantillon, 1µg d'ADN total (dans 100 µL d'eau UP), ont été **fragmentés** mécaniquement par sonication à l'aide du BioRuptor Pico™ - Diagenode (6 cycles de 30sec ON / 90sec OFF). Les fragments ont ensuite été sélectionnés pour ne garder que ceux dont la taille était comprise entre 200pb et 600pb à l'aide de billes magnétiques Agencourt AMPure XP (Beckman Coulter). Pour cela, 1 volume d'AMPure XP est ajouté au 100 µL d'ADN fragmenté et suivi d'une incubation pendant 10 min à température ambiante. Une fois l'ADN fixé aux billes, la plaque a été placée sur un support magnétique et le surnageant éliminé. Les billes ont été lavées deux fois avec de l'éthanol à 80%, avant d'être remises en suspension dans 40 µL d'eau.

Pour **réparer** les extrémités double brin de chaque fragment d'ADN, endommagé lors de la sonication, nous avons utilisé le kit End Repair Module (New England Biolabs) et une incubation de 30 min à 20°C. Par la suite, le tampon et les résidus d'enzyme ont été éliminés avec une purification AMPure XP (1.3V) et l'ADN fragmenté a été repris dans 30µL d'eau UP.

Les bibliothèques allant être mélangées lors de l'étape de capture, chaque génotype est identifié par un **barre-code** (Tag de 6bp) afin d'attribuer a posteriori chaque séquence au génotype dont elle est issue. J'ai amélioré cette phase du protocole en plaçant ce tag de 6pb des **deux côtés du fragment d'ADN** à séquencer, en 3' (P5) et en 5' (P7), afin d'augmenter la capacité de multiplexage en combinant 2 codes-barres différents et surtout de pouvoir éliminer des chimères créées lors de la phase de PCR (Schnell

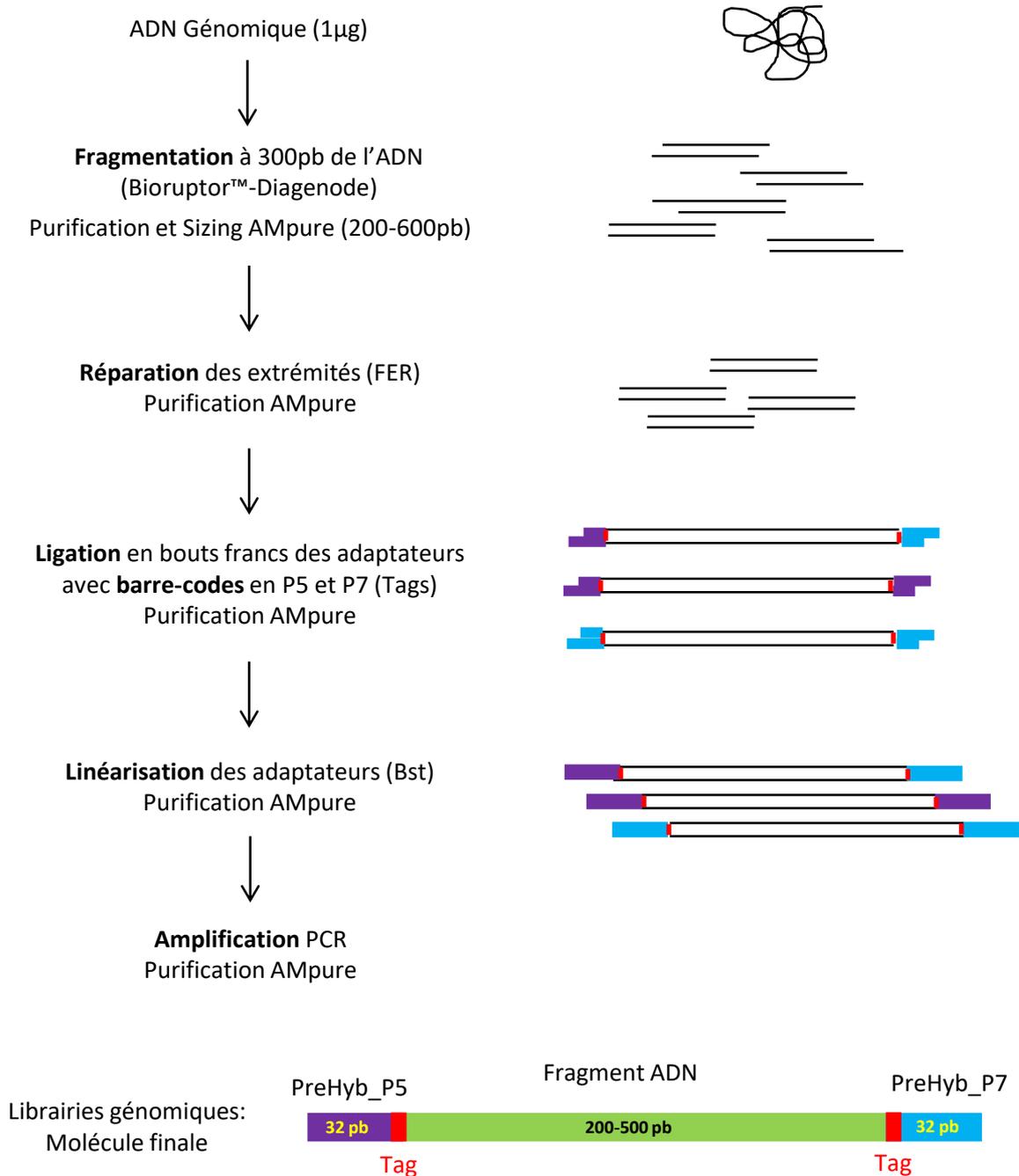


Figure 21: Les différentes étapes du protocole expérimental de la préparation des librairies génomiques. La préparation des librairies génomiques est constituée de plusieurs étapes qui permettent de passer de l'ADN génomique à des molécules composées d'un fragment d'ADN de 200 à 550pb entouré de deux adaptateurs de 32pb portant un tag.

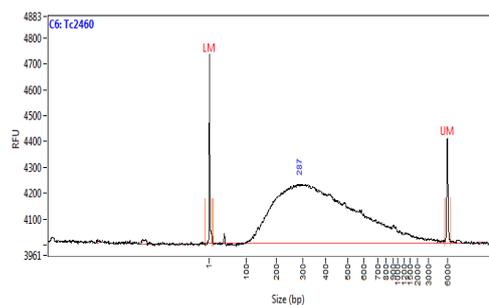


Figure 22: Visualisation d'une librairie génomique sur fragment analyser (DNF474). Après la préparation des librairies génomiques, il est nécessaire de vérifier leur taille sur Fragment Analyser (DNF494). Cela permet également de quantifier les librairies afin de mélanger en équi-proportion.

et al. 2015). En effet, lors de l'étape d'amplification, si une élongation est stoppée dans les premières bases, le fragment produit peut servir d'amorce au cycle suivant et produire au final un fragment avec deux tags différents en P5 et P7 (séquences paired-end). Les séquences obtenues vont alors poser des problèmes lors de l'analyse. Cette étape de **ligation** s'effectue à partir de 10µl de fragments ADN réparé et 8pmol d'adaptateurs de 32pb (PreHyb_PE_P5 barcoded et PreHyb_PE_MP7 Barcoded) portant les barres-codes (tags). La ligation se déroule dans un volume final de 20 µL avec 1 unité d'enzyme T4 DNA ligase (Invitrogen). Une incubation d'une heure à 22°C est suivie d'une étape d'inactivation de l'enzyme à 65°C pendant 10 min. Une purification AMPure XP (1.3V) est nécessaire pour éliminer le tampon et les résidus d'enzyme, et reprendre l'ADN ligué dans 24µL d'eau UP.

L'étape suivante consiste à **linéariser les adaptateurs** en utilisant l'enzyme polymérase Bst (New England Biolabs) car ils sont pour partie simple brin, ce qui permet d'éviter la ligation des adaptateurs entre eux. Le volume réactionnel composé de 24µL de librairie, 16 unités d'enzyme Bst, 1X de tampon ThermoPol et 1µL de dNTP à 10mM (each) est incubé 15 min à 37°C. Une purification AMPure XP (1.6V) est nécessaire pour éliminer le tampon et les résidus d'enzymes, et reprendre l'ADN dans 20µL d'eau UP.

Chaque librairie barre-codée est **amplifiée** individuellement par PCR avec l'enzyme haute-fidélité (HiFi HotStart Ready mix – KAPA) en présence de 5pmol de l'amorce PreHyb-PE_F et 5pmol de l'amorce PreHyb-MPE_R qui sont spécifiques de la partie conservée chez tous les adaptateurs barre-codés. Le programme d'amplification se compose d'une dénaturation initiale de 2 min à 98°C, suivie de 12 cycles avec 20 sec à 98°C de dénaturation, 45 sec à 55°C d'hybridation et 30 sec à 72°C d'élongation et est terminé par une élongation de 10 min à 72°C. Une purification AMPure XP (1.6V) permet d'éliminer le tampon et les résidus d'enzyme, et de reprendre les librairies dans 20µL d'eau UP.

La dernière étape est le **multiplexage des librairies génomiques** appartenant à des génotypes différents, qui subiront l'étape de capture ensemble. J'ai optimisé le protocole à cette étape : dans le protocole initial (Rohland and Reich 2012), le multiplexage était réalisé par un mélange en équivalence après la ligation. Or, il a été montré qu'il y avait une grande variabilité de l'efficacité de la ligation entre les fragments d'ADN et les adaptateurs, pour les différents génotypes (Esling, 2015). Afin de réduire au maximum la variabilité du nombre de séquences capturées entre chacun des génotypes, j'ai choisi d'effectuer le multiplexage des librairies génomiques à la fin de la construction et en **équivalence** de quantité et non de volume, suite à un dosage quantitatif. Pour calculer la concentration en nmol/L pour chaque génotype, les librairies génomiques sont analysées individuellement sur un Fragment Analyzer™ (AATI) avec le kit DNF-474 High Sensitivity (figure 22). La concentration moyenne sur l'ensemble des 120 librairies est de 53nM (min : 17nM, max : 90nM, écart-type : 16). Cette concentration dépendant de la taille des fragments, il est donc nécessaire de vérifier que la taille des fragments est comprise entre 200 pb et 600 pb et que la distribution de la taille des librairies est approximativement identique pour chacun des génotypes. Les 120 génotypes ont été multiplexés pour former deux mélanges de 60 génotypes (DEV_Pol009 et DEV_Pol010). Après multiplexage, les deux mélanges ont été dosés par spectrophotométrie (InfiniteM200™ TECAN) afin d'estimer la concentration en ng/µL et de déterminer la proportion qui sera nécessaire d'utiliser lors de la phase de capture.

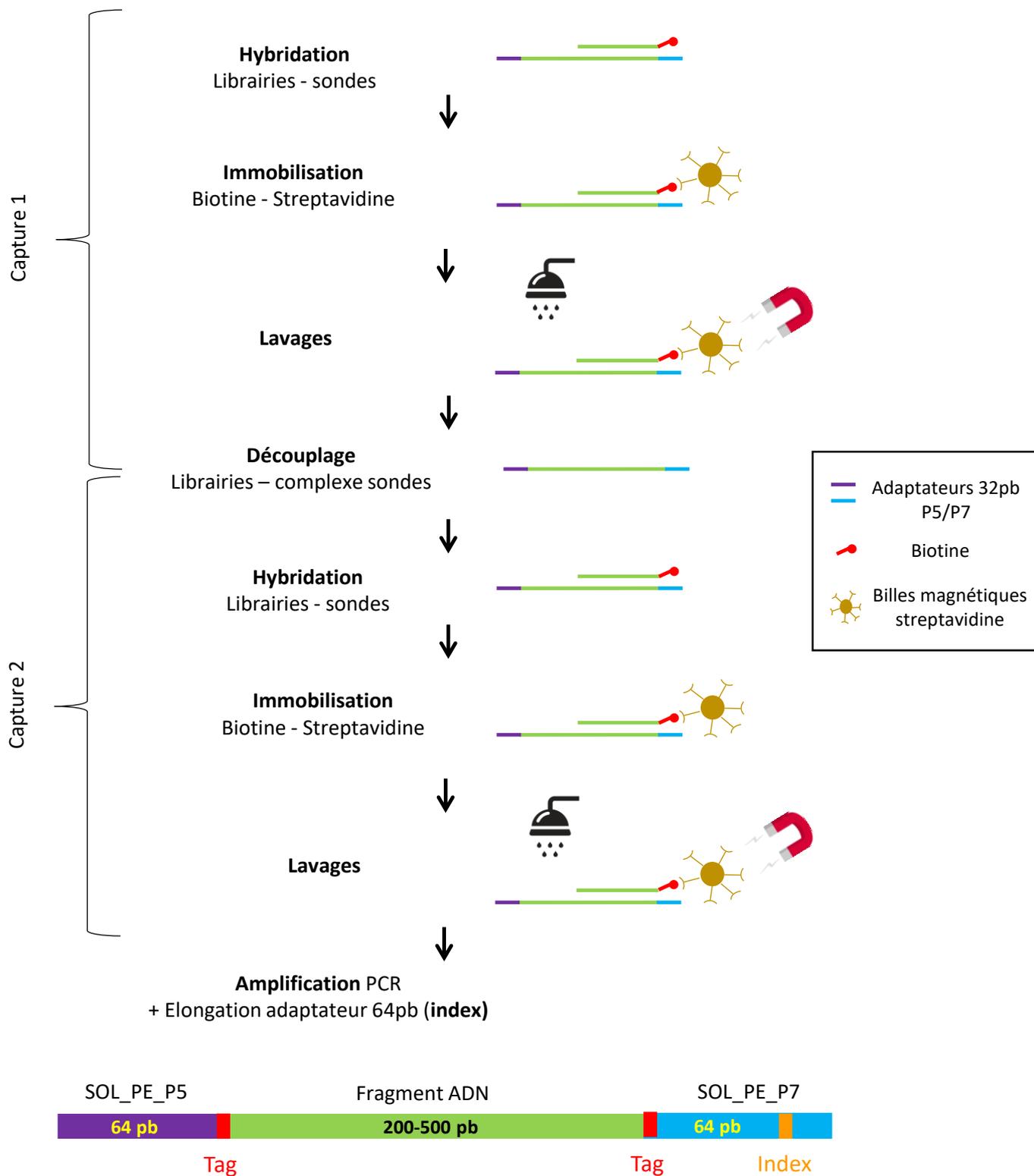
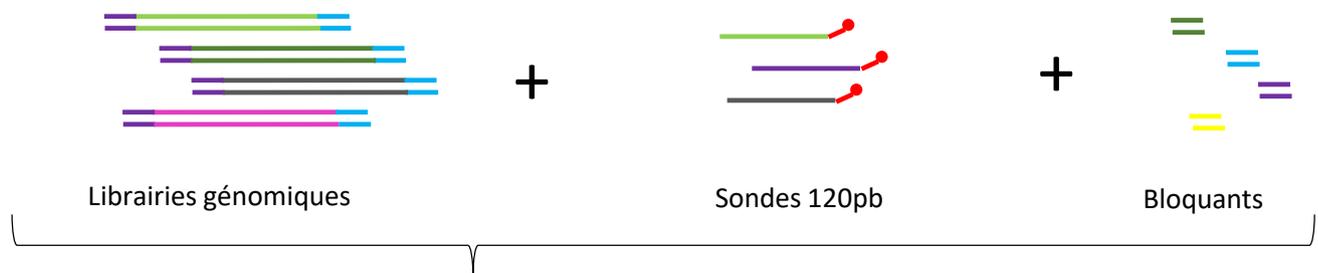


Figure 23: Les différentes étapes du protocole expérimental de l'enrichissement par capture. L'enrichissement par capture permet de cibler des fragments du génome (exons) à l'aide de baits de 120pb.

2.2.4 Enrichissement en séquences cibles par capture

Après avoir défini les sondes de 120pb (MyBaits™) et constitué les bibliothèques génomiques, la phase d'enrichissement par capture a été réalisée grâce au kit SeqCap EZ Hyb and Wash (Nimblegen) (Chen et al. 2015) (schéma récapitulatif : figure 23, protocole détaillé : annexe 3). Cependant, la complexité du génome du blé dur nous oblige à prendre des précautions particulières afin de rendre la capture la plus efficace possible. En effet, la proportion du génome ciblé par les 20 000 sondes MyBaits™ de 120 pb représente seulement 0.02% du génome complet. Le protocole de capture (Holtz et al. 2016) a donc été optimisé sur les deux points suivants :

- ✓ Pour augmenter la spécificité des hybridations et, par conséquent, le taux d'enrichissement de la capture, j'ai réalisé deux phases successives de capture. La dose de sonde de capture préconisée (4,5 µL) est divisée en deux parts inégales : sur la base du ratio entre ADN cible/ADN total, 80% (3.5µL) et 20 % (1µL) de la dose seront utilisés respectivement pour la première et la seconde capture.
- ✓ Afin de saturer les zones très répétées, j'ai utilisé deux types de bloquants : un bloquant commercial (Nimblegen) de séquences répétées conservées chez plusieurs espèces animales et végétales (Sequence Capture Developer Reagent), un second correspondant à des séquences d'oligonucléotides de motifs microsatellites très fréquents dans le génome du blé dur basé sur trois motifs (GAA, GGA, CAA) pour lesquels nous avons fait synthétiser les brins complémentaires (MWG).

La première étape de la phase de capture est la constitution du volume réactionnel de la première **hybridation**. Il est composé de 1000 ng du mélange de bibliothèques génomiques, 10 µL de « sequence Capture Developer Reagent » (Nimblegen), 2µL des bloquants des trois motifs microsatellites (33pmol/L chacun). Afin de limiter les hybridations non-spécifiques au niveau des adaptateurs, 2µL de bloquants des adaptateurs P5/P7 (50pmol/L chacun) ont été ajoutés. Le volume réactionnel a été lyophilisé sur un SpeedVac™ (thermo Fisher), puis repris dans un tampon d'hybridation (kit SeqCap EZ Hyb and Wash, Nimblegen). Les 3.5µL de sonde BAITS ont été dénaturés 10 minutes à 95°C avant d'être ajoutés au volume réactionnel. Une incubation de 64 à 72 heures à une température de 47°C permet une hybridation optimale des cibles et des sondes.

Les hybridations cibles/sondes ont été **immobilisées** à l'aide de billes magnétiques couplées avec de la streptavidine ayant une forte affinité pour la biotine présente sur les sondes. Cela s'effectue par l'ajout de billes Dynabeads MyOne Streptavidin C1™ (thermo Fisher) et une incubation de 45 min à 47°C. Le complexe sonde-séquence cible a ensuite été retenu par aimantation, alors que l'ADN non ciblé a été éliminé par six **lavages** successifs.

L'étape suivante est le **découplage** des cibles par dénaturation (3 min à 95°C) avant de réaliser, en suivant, la deuxième capture. Le volume réactionnel de la deuxième hybridation est constitué des cibles capturées lors de la première hybridation (20µL), auquel ont été ajoutés les deux bloquants, « Sequence Capture Developer Reagent » et microsatellites, à une concentration dix fois plus faible que pour la première hybridation. Les bloquants des adaptateurs P5/P7 ont été, quant à eux, dilués cinq fois. Cette différence entre les deux hybridations est liée au fait que le ratio ADN cible/ADN total a augmenté pour la deuxième hybridation. Une nouvelle fois, le volume réactionnel a été lyophilisé sur un SpeedVac™ (thermo Fisher), puis repris dans un tampon d'hybridation fourni dans le kit SeqCap EZ Hyb and Wash (Nimblegen). Pour cette deuxième hybridation, 1 µL de sonde BAITS a été dénaturé 10

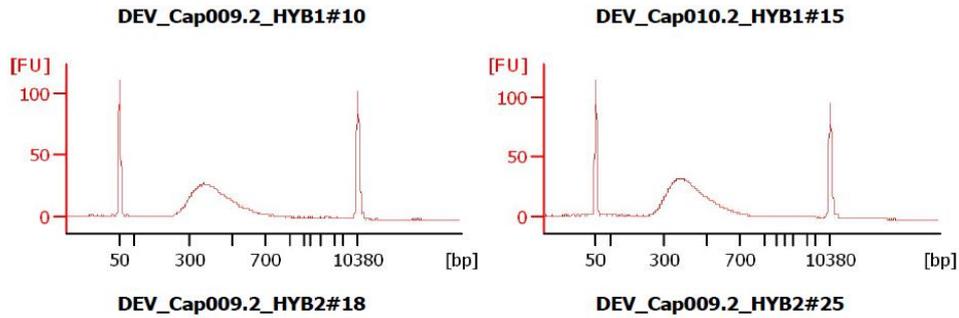


Figure 24: Visualisation de deux mélanges de bibliothèques génomiques, après enrichissement par capture, sur BioAnalyser (DNA7500). Après enrichissement par capture des deux mélanges de bibliothèques (DEV_Cap009 et DEV_Cap010), il est nécessaire de vérifier leur taille sur BioAnalyser (DNA7500). Cela permet également de quantifier les mélanges de bibliothèques enrichies afin de mélanger en équi-proportion avant leur dépôt sur séquenceur haut-débit.

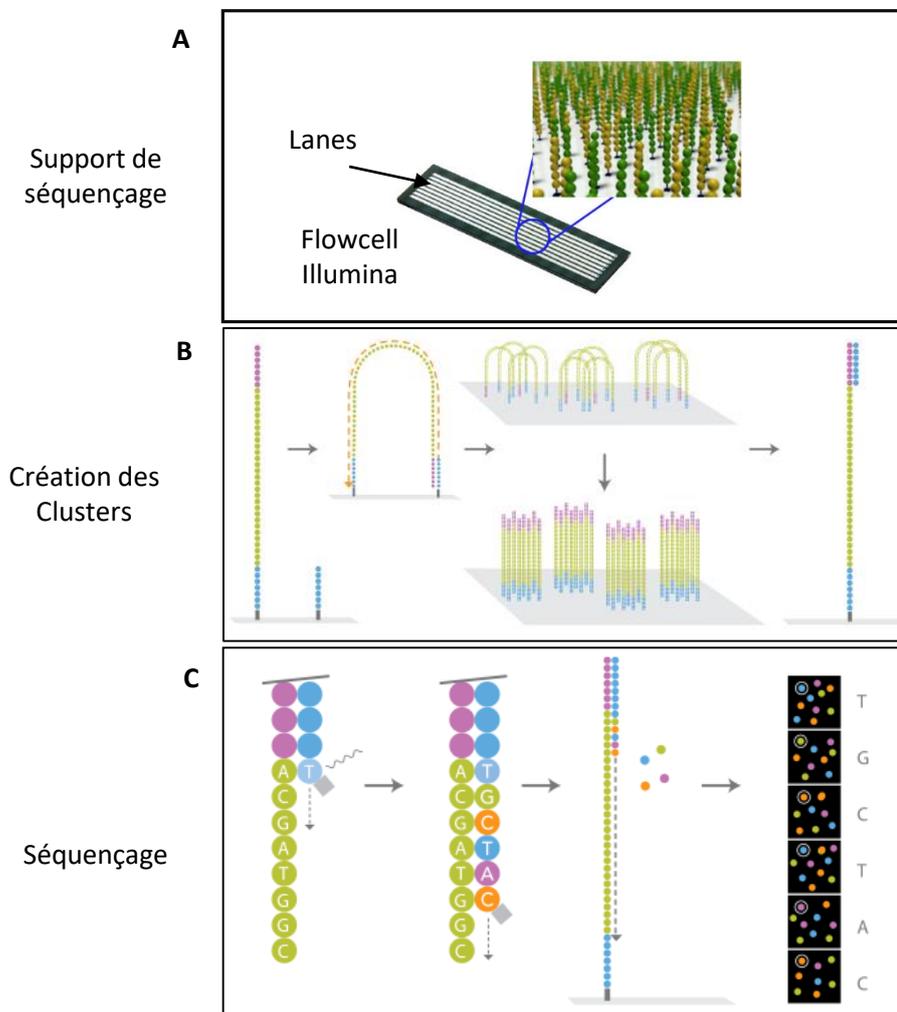


Figure 25: Séquençage Illumina de dernière génération (NGS)

Le séquençage sur Hiseq 3000 s'effectue par dépôt des bibliothèques à séquencer sur une « Flowcell ». Chaque « Flowcell » est composée de 8 « Lanes » : 7 pour de l'analyse d'échantillons et 1 pour un témoin de séquençage (A).

La première étape du séquençage est amplification « en ponts » des fragments d'ADN sur la flowCell afin de créer des colonies d'ADN constituées de la même séquence, appelé « Clusters » (B). Cette étape permettra d'augmenter l'intensité des signaux de fluorescence à chaque cycles. Le séquençage s'effectue par l'incorporation de nucléotides fluorescents. A chaque incorporation et pour chaque cluster, le système optique de l'appareil détecte la fluorescence émise qui sera ensuite transformé en nucléotides afin de produire une séquence de 150pb.

minutes à 95°C avant d'être ajouté au volume réactionnel puis incubé de 17 à 20 heures à 47°C. L'immobilisation par la streptavidine et les étapes de lavage ont été réalisées dans les mêmes conditions.

La dernière étape est une amplification par PCR, qui a plusieurs objectifs. Elle permet le découplage des cibles du complexe sonde-billes magnétiques, l'amplification des fragments capturés afin de rendre possible le séquençage et pour finir, l'allongement des adaptateurs P5/P7 par l'ajout d'un index et de la séquence qui permet la fixation des adaptateurs au support de séquençage. Cette amplification PCR a été effectuée avec le kit d'amplification HiFi ready mix (KAPA) en présence de 15 pmol d'amorce SOL-PE-PCR-F (Rohland, 2012) et de 15 pmol d'amorce SOL-MPE-PCR_INDEX-R spécifiques des adaptateurs présents sur les cibles. Plusieurs amorces SOL-MPE-PCR_INDEX-R sont disponibles, portant chacune une séquence variable de six paires de bases (Index). Cela permet d'ajouter à l'ensemble des génotypes capturés en mélange, un index propre à chaque capture, multipliant ainsi la capacité de multiplexage à l'étape finale de séquençage (Tag/ génotypes + Index / capture). Le programme d'amplification se compose d'une dénaturation initiale de 2 min à 98°C, suivie de 18 cycles avec 20 sec à 98°C de dénaturation, 30 sec à 62°C d'hybridation et 30 sec à 72°C d'élongation et terminé par une élongation finale de 5 min à 72°C. Après une purification AMPure XP (1.8V), les librairies enrichies par capture ont été vérifiées et quantifiées sur BioAnalyser™ (Agilent tech.) avec une puce DNA 7500. La taille des deux mélanges de 60 librairies enrichis par capture (DEV_CAP009 et DEV_CAP010) était comprise entre 400 et 600 pb et leur concentration bien supérieure à 10nM, quantité minimale nécessaire au séquençage (figure 24).

Avant de mélanger les deux captures portant des index différents, un dosage précis par qPCR à l'aide du kit KAPA Library Quantification Kit et d'un appareil LightCycler96™ (Roche) a été réalisé. Le mélange final a ensuite été dilué et dénaturé avant le séquençage.

2.2.5 Séquençage

Le séquençage des librairies obtenues après capture à l'aide des sondes MyBaits™ a été effectué sur un séquenceur haut-débit Illumina de type Miseq™ ou Hiseq3000™. Ces deux systèmes permettent le séquençage en simultané de plusieurs millions de fragments d'ADN. La différence entre ces deux séquenceurs de nouvelle génération est leur productivité : en moyenne 15 millions de séquences de 150 pb, et 300 millions de séquences de 150 pb, par unité de séquençage (appelée « lanes ») respectivement pour des séquenceurs Miseq™ et Hiseq3000™. Pour cette étude, plusieurs « lanes » de Miseq™ ont été utilisées pour la première phase de développement méthodologique (double tag, quantification avant multiplexage, double capture, bloquants). Pour la phase de production de données en haut-débit, le séquençage des deux mélanges de librairies enrichies par capture (DEV_CAP009 et DEV_CAP010) a été effectué sur une « lane » de Hiseq3000™.

Dans les deux cas, le principe de séquençage est identique (figure 25). Les fragments à séquencer sont fixés sur un support solide (FlowCell) grâce aux séquences situées aux extrémités des adaptateurs puis amplifiés « en pont » afin de créer des colonies d'ADN constituées de la même séquence, appelées « Clusters » et de rendre détectable la fluorescence émise à chaque cycle. L'étape suivante est le séquençage de chacun de ces clusters, par l'incorporation de nucléotides fluorescents à chaque cycle et la détection de la fluorescence émise. Des séquençages en « paired-end » ont été effectués, c'est-à-dire, le séquençage de 150 paires de base dans le sens 1 (reads_R1), le séquençage de l'index ajouté lors de la PCR finale après capture, et le séquençage de 150 paires de base dans le sens 2 (reads_R2).

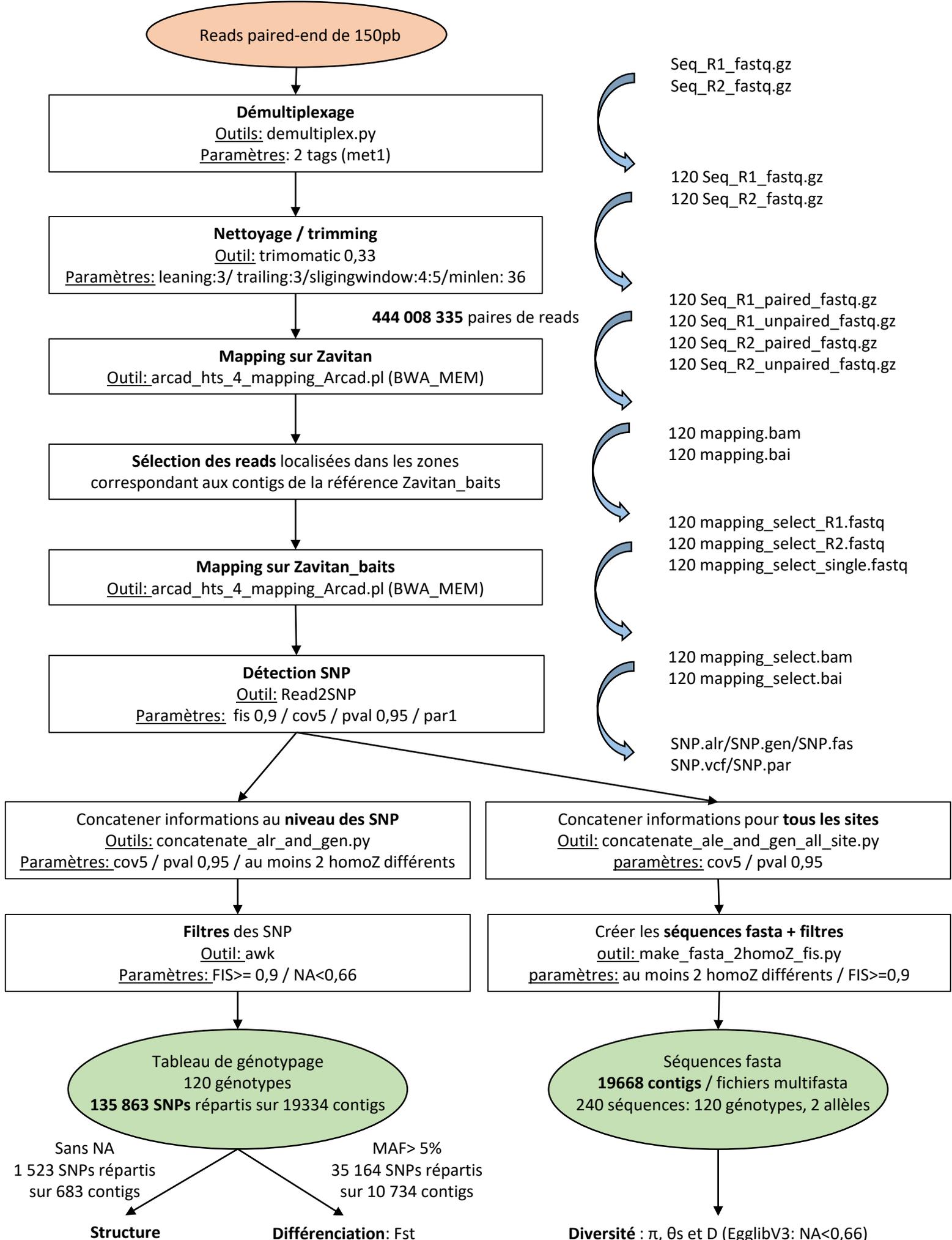


Figure 26: Les différentes étapes du pipeline d'analyses bio-informatiques
 Schéma récapitulatif permettant de visualiser l'ensemble des analyses bio-informatiques qui ont été réalisées pour passer des reads issus du séquenceur aux données de polymorphisme utilisé pour estimer la structure, la différenciation et la diversité génétique sur l'échantillon de 120 génotypes.

2.3 Outils bio-informatiques

Suite au séquençage sur Hiseq3000 des 120 génotypes séparés en deux captures (DEV_Cap009 et DEV_Cap010), nous avons obtenu deux dossiers compressés (.gz) par capture, le premier contenant les reads produits par le séquençage dans le sens 1 (R1) et l'autre avec les reads produits par le séquençage dans le sens 2 (R2). Les outils de bio-informatique ont permis de passer des données brutes issues du séquenceur, à des données nettoyées et utilisables pour les analyses de génétique des populations. Ces analyses s'effectuent en ligne de commande directe ou en utilisant des scripts en langage Bash ou python, sur le cluster de calcul « CC2 » hébergé sur le site du CIRAD de notre UMR. La figure 26 présente les différentes étapes d'analyses qui ont été réalisées.

2.3.1 Qualité des séquences

La qualité des reads a été vérifiée avec l'outil FASTQC (Andrews 2010) sur la base de différents critères, comme le score de qualité de chaque base (tableau 2).

$$Q_{PHRED} = -10 * \log_{10}(Pe)$$

Où Pe est la probabilité estimée d'erreur

Pour nos deux mélanges, les scores de qualité des bases nucléotidiques lues étaient tous au-dessus du seuil Q30 que nous nous étions fixés, soit 1 risque sur 1000 qu'il y ait une erreur de lecture de la base. La figure 27 présente les scores de qualité du premier mélange de 60 génotypes (DEV_Cap009). Nous pouvons observer que les nucléotides qui se trouvent au début et à la fin des reads de 150 pb, sont de moins bonne qualité. Cela est dû au fait que le système optique qui détecte la fluorescence effectue la calibration sur les six à dix premières bases. Par ailleurs, après 130 cycles de séquençage, une fluorescence résiduelle constante apparaît malgré les rinçages à chaque cycle. Cette observation est d'autant plus notable lorsqu'il s'agit de la lecture dans le deuxième sens (R2).

L'outil FASTQC a également permis de contrôler la proportion de chaque base (A, T, G, C) pour les 150 positions, le % en GC des séquences et la longueur des reads.

2.3.2 Démultiplexage

Les séquences de chaque génotype ont été identifiées en utilisant leur barre-code (Tag) spécifique, ajouté lors de la ligation des adaptateurs (cf partie 2.2.3).

Le démultiplexage s'effectue avec l'outil DEMULTIPLEX (Flutre 2014). Les 218 353 055 paires de reads du premier mélange (DEV_Cap009) et les 223 689 152 paires de reads du deuxième mélange (DEV_Cap010) ont été attribuées à chacun des 120 génotypes. Deux options sont nécessaires pour cette étape de multiplexage :

- ✓ « -re ApeKI » spécifie que la ligation des 6pb du tag est en « bouts francs » et que le tag correspond aux six premières bases des reads.
- ✓ « -met 1 » permet de garder les paires de reads (séquençage en « paired-end ») seulement si les 6pb du tag sont identiques des deux côtés (en Read1 et en Read2). Cette option va donc supprimer tous les reads correspondant à des molécules chimériques (cf partie 2.2.3).

Tableau 2: Score de qualité – Phred

Lors du séquençage, un score de qualité est attribué à chaque base à l'aide d'un caractère alpha-numérique. Le logiciel FASTQC transforme ce caractère en score Phred. Par exemple, un score à Q30 signifie que le nucléotide a été attribué avec une probabilité d'une chance sur 1000 d'une identification incorrecte. La précision de l'identification est donc de 99,9%.

Score de qualité - phred	Probabilité d'une identification incorrecte	Précision de l'identification d'une base
Q10	1 pour 10	90 %
Q20	1 pour 100	99 %
Q30	1 pour 1000	99.9 %
Q40	1 pour 10000	99.99 %
Q50	1 pour 100000	99.999 %

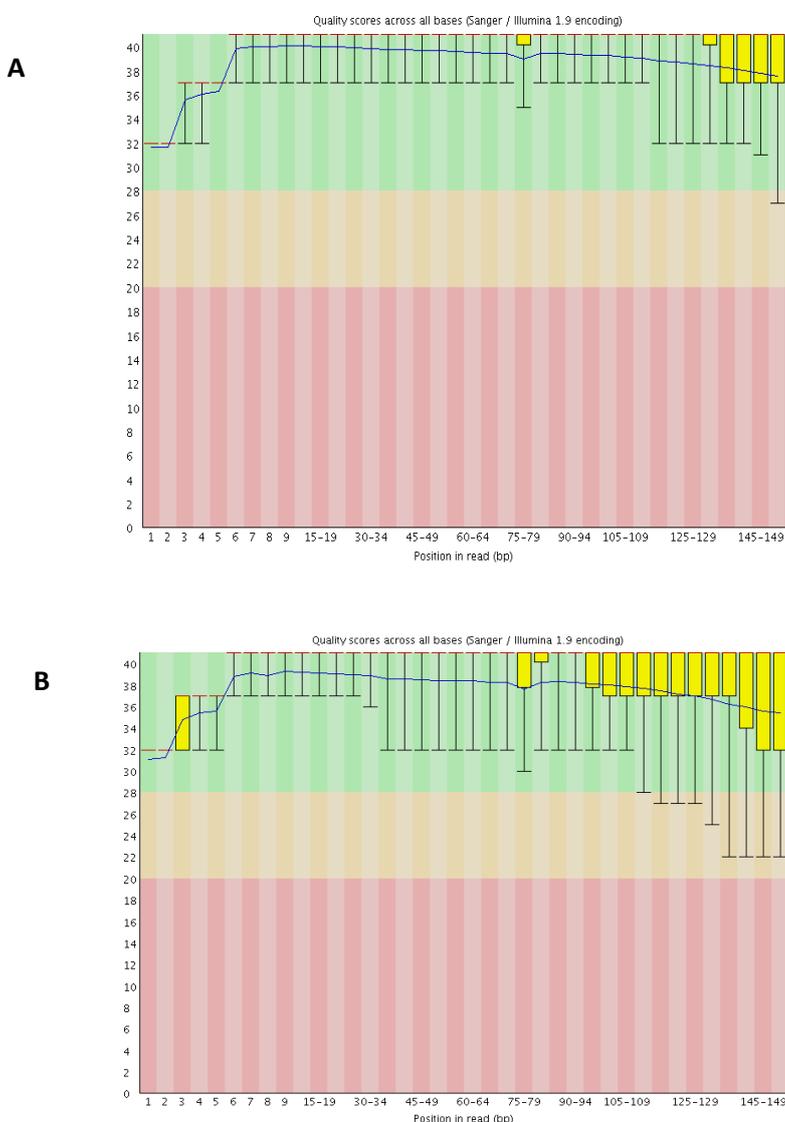


figure 27: Scores de qualité des reads du mélange DEV_Cap009

Les scores de qualité de l'ensemble des reads obtenus pour un mélange de bibliothèques peuvent être visualisés facilement grâce à l'outil FASTQC. Le score de qualité est présenté en ordonné et la longueur des séquences en abscisse. Ce type de graphique est réalisé pour les reads séquencés dans le sens R1 (A) et dans le sens R2 (B).

Cette option impacte le nombre de reads retenus. Si nous prenons l'exemple du premier mélange de 60 génotypes (DEV_Cap009), 8,5% des reads sont éliminés du fait que les tags n'étaient pas identiques sur le read1 et sur le read2. Par comparaison, l'assignation des reads à un génotype sur la base d'un seul des deux tags génère un taux d'élimination de 0.2% des reads. Etant donnée la complexité du génome du blé, nous avons choisi de ne pas prendre le risque de garder des séquences avec des ambiguïtés d'assignations. Suite à cette étape, nous avons conservé 199 690 239 de paires de reads pour le premier mélange (DEV_Cap009) et 202 144 687 de paires de reads pour le deuxième mélange (DEV_Cap010), répartis sur les 120 génotypes.

La répartition des reads entre les différents génotypes a ensuite été observée. Idéalement (cas X), cette distribution du nombre de reads par génotype est uniforme et suit une loi de Poisson de paramètre λ :

$$X \sim P(\lambda) \text{ avec } E(X) = \text{Var}(X) = \lambda$$

où λ et $\text{Var}(X)$ sont la moyenne et la variance des nombres observés de reads par individu.

Dans notre cas, la distribution du nombre de reads par génotype (X) ne suivait pas une loi de Poisson idéale : la variance étant bien plus élevée que la moyenne. Pour rendre compte de cette variance, un paramètre de sur-dispersion (θ) a été introduit :

$$\text{Var}(X) = \lambda + \theta\lambda^2$$

Pour estimer θ , on utilise l'équation :

$$\hat{\theta} = (\widehat{\text{var}(X)} - \hat{\lambda})/\hat{\lambda}^2$$

Pour le premier mélange (DEV_Cap009), la valeur du θ est de 10,1 avec des nombres de reads par génotype compris entre 701 457 reads (génotype Tc2254) et 4 681 622 reads (génotype Tc2251). Pour le deuxième mélange (DEV_Cap010), cette valeur de θ était de 7,3 avec une variation comprise entre 740 644 reads (Tc3433) et 6 184 798 reads (Tc2656). Ces valeurs indiquent que la répartition des reads n'était pas idéale ; une optimisation technique parfaite permettant d'obtenir des banques en équiproportion et donc de faire tendre vers 0 la valeur du paramètre de sur-dispersion (θ) ; toutefois, l'optimisation du protocole a tout de même permis de réduire significativement la valeur du θ , qui était comprise auparavant entre 40 et 50.

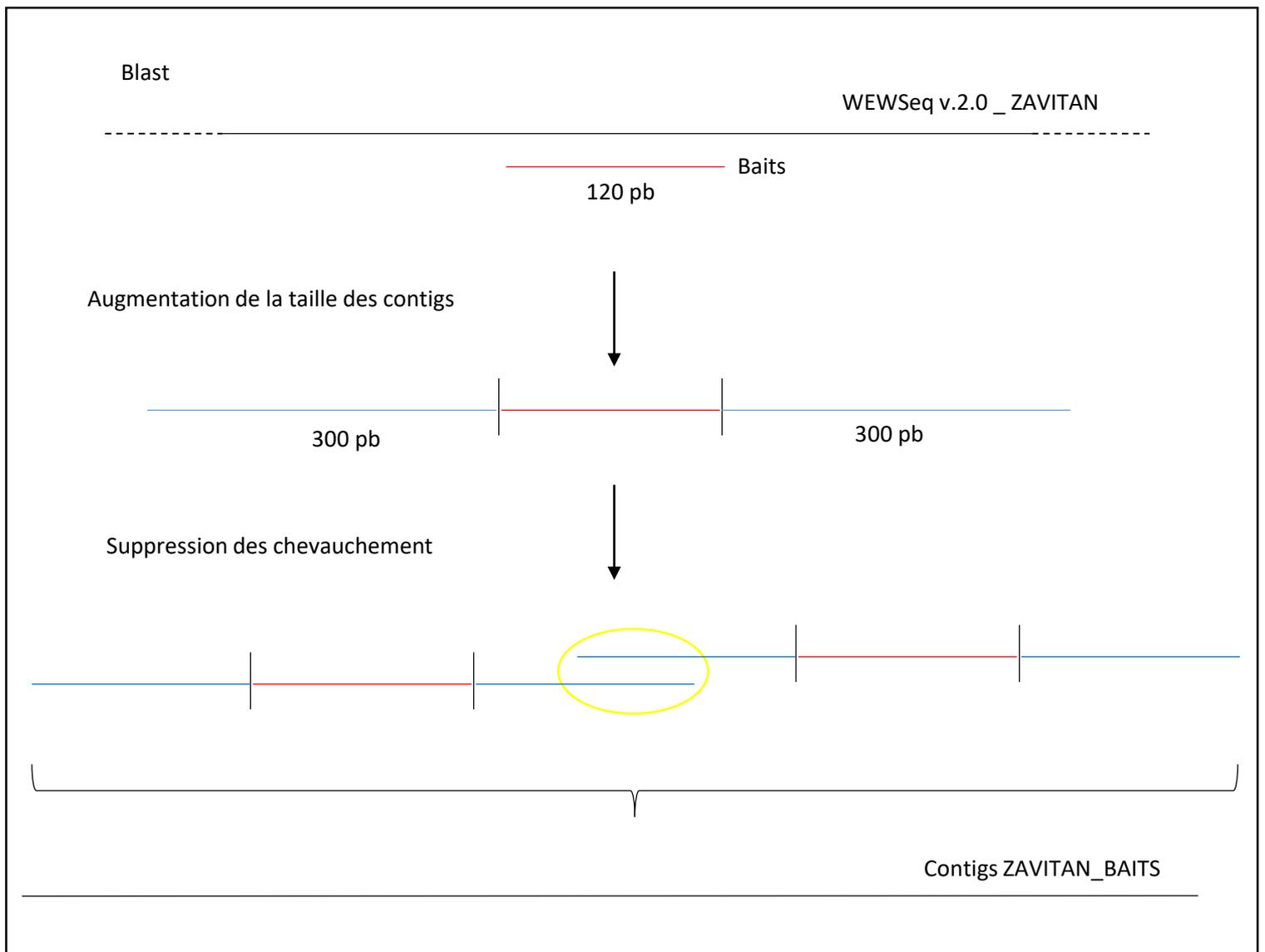


Figure 28: Construction de la référence génomique simplifiée ZAVITAN_BAITS

La construction de la référence ZAVITAN_BAITS s'est effectuée en trois étapes principales: La première est le blast des 20000 baits de 120pb sur la référence génomique complète WEWSeq v.2.0_ZAVITAN. Après sélection des meilleurs blasts, les contigs sont constitués en allongeant la taille de 600pb. Pour finir, lorsque deux contigs sont chevauchants, ils sont regroupés en un seul. La référence ZAVITAN_BAITS est composée de 19 738 contigs de 690 à 4583pb.

2.3.3 Nettoyage

Le nettoyage (« trimming ») des reads a été réalisé à l'aide de seuils grâce à l'outil TRIMMOMATIC (Bolger et al. 2014). Les adaptateurs (P5 et P7) situés de part et d'autre du fragment d'intérêt, les bases aux deux extrémités des reads souvent de moins bonne qualité, ainsi que les séquences inférieures à 36 pb ont été éliminés.

Pour la première capture (DEV_Cap009), 4.6% des reads ont été supprimés car ils ne passent pas les filtres de qualité et le décompte du nombre de reads exploitables s'élève alors à 190 498 521. Pour le deuxième mélange (DEV_Cap010), 4.8% des reads sont éliminés et il reste donc 192 479 453 reads filtrés.

Suite à cette étape, j'ai choisi d'ajouter à cet ensemble de séquences (DEV_Cap009 + DEV_Cap010), celles que j'avais obtenues lors des différentes phases de mise au point, après les avoir démultiplexées et nettoyées dans les mêmes conditions. Au final, j'ai travaillé sur 430 millions de paires de séquences, réparties sur les 120 génotypes.

2.3.4 Références génomiques

Pour identifier les modifications génétiques qui se sont produites au cours de la domestication du blé dur, un génome de référence permet de situer les polymorphismes étudiés.

En 2017, un consortium international a généré une séquence de référence de haute qualité de l'accession « ZAVITAN » issue de *T. turgidum ssp dicocoides* : WEWSeq v.1.0 (Avni et al. 2017). Ce génome de référence a été construit en utilisant une approche de type génome complet (WGS) pour produire des « scaffolds » qui correspondent à un regroupement de séquences proches, mais non contiguës, séparées par des gaps (Brenchley et al. 2012). Ces « scaffolds » sont ensuite assemblés grâce à l'algorithme d'assemblage MAGICTM (NRGene, NesZiona), puis validés à l'aide de la technologie Hi-C (Lieberman-Aiden et al. 2009) et des cartes génétiques à haute densité. Cette méthode a permis la construction d'assemblages à l'échelle chromosomique (pseudo-molécules) ainsi que des analyses du contenu en gènes. Au final, la référence génomique de 10,4 Go est composée de 14 séquences de pseudo-molécules représentant les 14 chromosomes de l'accession « Zavitan » appartenant au groupe *T. turgidum ssp dicocoides*.

En 2019, le même consortium international a généré une nouvelle version de la séquence de référence de l'accession « ZAVITAN » : WEWSeq v.2.0 (Zhu et al. 2019). L'assemblage des « scaffolds » en pseudo-molécules a été amélioré grâce à la cartographie optique (Bionano Genomics). Cette amélioration a permis, non seulement, de réduire le nombre de scaffolds, mais également de corriger leur orientation. Dans la suite de ce mémoire, cette référence sera notée « ZAVITAN ».

En complément de cette référence, j'ai constitué une référence génomique simplifiée correspondant aux zones ciblées par les 20 000 sondes MyBaits™ de 120 pb (figure 28). Pour cela, les séquences des sondes ont été alignées, avec un BLAST, sur la séquence de référence complète WEWSeq v.2.0. Les 65 460 BLASTs ont été filtrés pour ne conserver que ceux pour lesquels il y avait au moins 90 % d'identité entre la séquence des sondes (120pb) et la référence sur au moins 90pb (seuils arbitraires). Après ce filtre, 24 549 BLASTs ont été conservés et 9 919 loci sur 10 000 ciblés représentés. Etant donné que les bibliothèques génomiques sur lesquelles ont été faites les captures étaient constituées de fragments de 400 à 600pb, nous avons fait le choix d'étendre les fragments retenus de 300pb en amont et en aval. Après

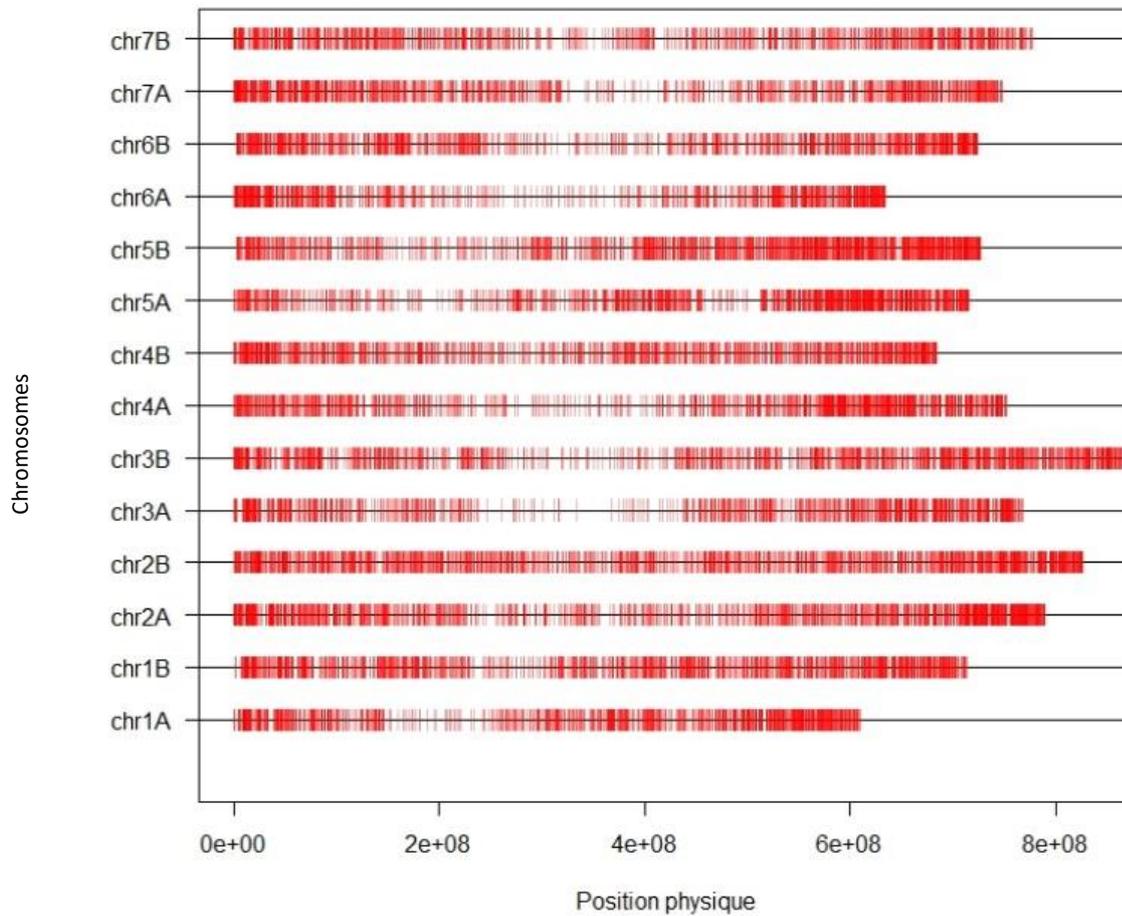


Figure 29: Répartition des 19738 contigs de la référence ZAVITAN_BAITS sur les chromosomes 14 chromosomes

Chaque trait rouge vertical représente un contig de la référence ZAVITAN_BAIT. Ils sont disposés sur les 14 chromosomes en fonction de leur position physique. Les centromères portent moins de contigs car ces zones sont moins polymorphes, contrairement aux télomères.

cet allongement, si deux blasts avaient des coordonnées chevauchantes, nous avons fait le choix de les rassembler pour créer une seule séquence. Ainsi 19 738 séquences de 690 à 4583pb, qu'on appellera « contigs », constituent la référence « ZAVITAN_BAITS ». Ces 19 738 contigs sont plutôt bien répartis sur l'ensemble du génome de la référence génomique complète (figure 29). Cependant, comme attendu, les régions centromériques portent moins de contigs que les régions télomériques (Lukaszewski and Curtis 1993; Dvorak et al. 1998).

2.3.5 Mapping

Les reads (séquences de 150pb « paired-end » issues du séquençage Illumina) ont été positionnées sur les deux références génomiques (ZAVITAN et ZAVITAN_BAITS) en fonction de leur similarité de séquences, c'est ce que l'on appelle le « mapping ». Pour chaque read, on obtient les coordonnées de la région génomique sur laquelle le read s'est aligné avec le meilleur score (fichier .bam).

Pour réaliser le mapping, j'ai utilisé un pipeline composé de différents outils :

- ✓ BWA (Burrows-Wheeler Aligner) (Li and Durbin 2009) : l'algorithme BWA_MEM (paramètres par défaut) recommandé pour des séquences obtenues par les techniques Illumina™ supérieures à 100pb va créer un index de la référence (bwa index) ce qui permettra d'optimiser la phase de mapping.
- ✓ Samtools view et samtools fixmate pour modifier le format des fichiers (.bam).
- ✓ SortSam pour relier les reads1 et reads2 provenant du même fragment d'ADN.
- ✓ Picard tools-MarkDuplicates pour conserver les duplicats PCR (option –normdup) du fait que la proportion génome ciblé / génome total est très faible.

L'outil choisi pour la détection des SNPs (Read2SNP - Gayral et al. 2013) nous a imposé une contrainte, en amont, au niveau de l'étape de mapping. Cet outil ayant été développé pour travailler plutôt sur le transcriptome, il n'est pas adapté lorsque la séquence de référence, nécessaire à la localisation des SNPs, est supérieure à 100Mo. Sachant que la référence complète « ZAVITAN » fait 10,4 Go, il a été indispensable d'utiliser la référence simplifiée ZAVITAN_BAITS qui ne fait, elle, que 16 Mo.

De manière générale, lors d'un mapping, il est préférable d'utiliser une référence la plus complète possible afin que chacun des reads se place de façon optimale sur la référence. Si une référence incomplète est utilisée, les reads pour lesquels la cible n'est pas présente ne devraient pas être positionnés, or il est fréquent qu'une partie de ces reads soient néanmoins positionnés lorsque la référence contient des régions homologues aux régions cibles absentes. Malgré un niveau de similitude très élevé entre le read et la séquence de référence (l'algorithme BWA_MEM), le génome du blé dur étant composé à 80 % de séquences répétées, cette étape reste critique.

En considérant ces deux problématiques, j'ai, dans un premier temps, réalisé un mapping des reads sur la référence génomique complète « ZAVITAN » (99.84% de mapping), puis j'ai sélectionné les reads positionnés spécifiquement dans les zones correspondant à celles de la référence simplifiée ZAVITAN_BAITS. Ces reads ont ensuite été mappés sur la référence simplifiée ZAVITAN_BAITS (99.91% de mapping) afin que les fichiers de sortie du mapping soient homogènes et compatibles avec l'utilisation de l'outil Read2SNP. Cette procédure nous a permis d'éliminer 11.25% de reads qui auraient été positionnés, de façon erronée, sur la référence simplifiée ZAVITAN_BAITS (figure 30).

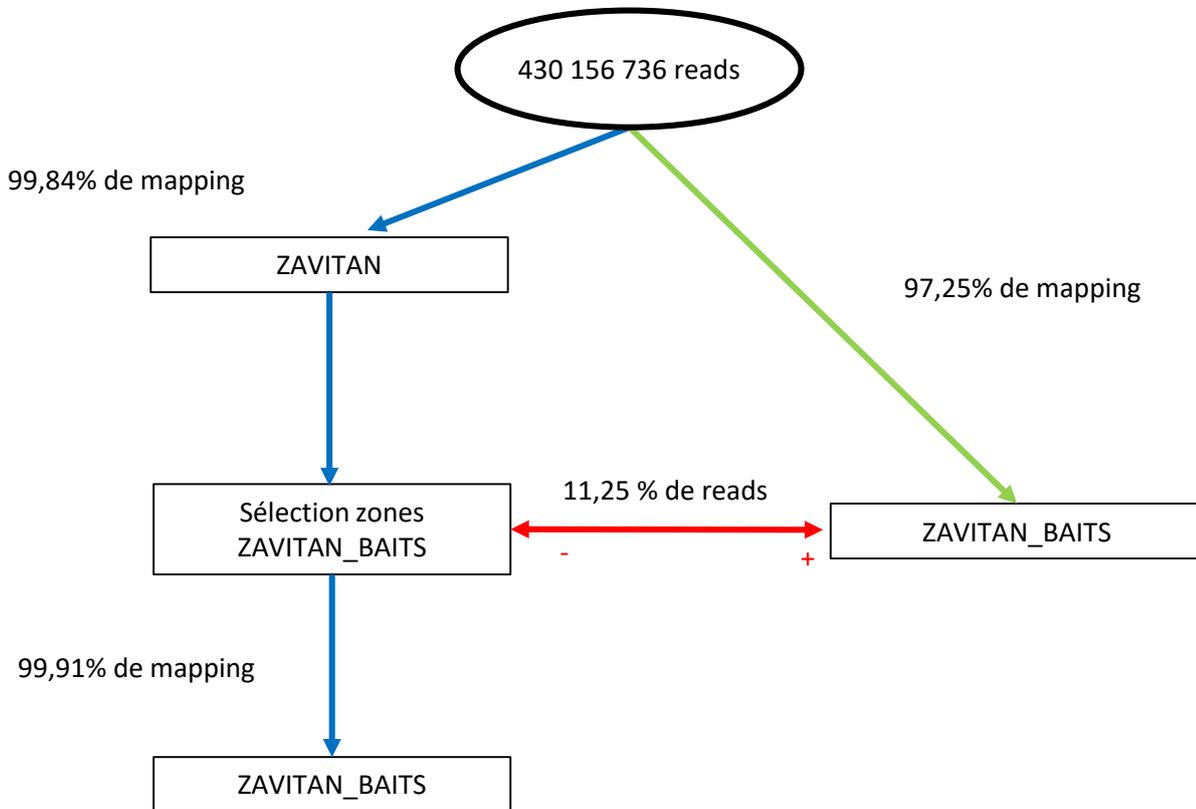


Figure 30: Impacts de la référence utilisée pour le mapping des reads

Lorsque le mapping des 430 millions de paires de reads a été effectué sur la référence génomique « ZAVITAN », 99.84% d’entre-deux trouvent une correspondance (en bleu). Ce qui permet de valider le protocole expérimental (double capture et bloquants). Lorsque le mapping a été réalisé sur la référence « ZAVITAN_BAITS », le taux de mapping passe à 97.25% (en vert).

Si nous comparons le nombre de reads qui se sont mappés sur la référence « ZAVITAN_BAITS » au nombre de reads qui mappaient sur la référence «ZAVITAN» mais seulement dans les zones correspondant aux contigs « ZAVITAN_BAITS », nous avons comptait 11.25% des reads qui se « mappent par défaut » sur la référence « ZAVITAN_BAITS ».

L’utilisation de l’outil Read2SNP pour la détection des SNPs nous impose d’effectuer un mapping sur la référence « ZAVITAN_BAITS », en suivant, afin de réduire la taille de la référence (en bleu).

Effet de la capture sur la spécificité des fragments capturés

Nous avons séquencé sur Hiseq3000, les fragments capturés après la première capture et après les deux captures successives, sur le même mélange de 60 génotypes (DEV_Cap010). Le mapping des deux lots de reads a été effectué sur référence ZAVITAN_BAITS. Les résultats de ces deux mappings indiquent que seulement 50.12% des reads issus de la première capture correspondent aux contigs de la référence ZAVITAN_BAITS, alors qu'une double capture permet d'obtenir 97,06% pour les reads issus des deux captures successives. Nous pouvons donc en conclure, qu'à l'issue du premier enrichissement par capture, les baits entraînent un grand nombre de fragments non spécifiques. Cette situation est due essentiellement à la complexité du génome du blé. La réalisation d'un deuxième enrichissement par capture permet d'éliminer la très grande majorité des fragments issus de capture non spécifique et donc de doubler le nombre de reads correspondant aux zones ciblées par les baits de 120pb. Nous avons donc décidé d'utiliser ce protocole pour l'ensemble des expérimentations.

2.3.6 Détection de SNPs

Les nucléotides polymorphes (SNPs) ont été détectés à l'aide de l'outil « Read2SNP » (Gayral et al. 2013), qui utilise la méthode du maximum de vraisemblance introduit par Tsagkogeorga *et al.* (2012). Nous avons fait le choix de détecter tous les SNPs présents et pas seulement les 10 000 SNPs ciblés. Cela nous a permis d'augmenter considérablement la quantité d'information mais également d'éviter un biais dû au fait que les SNPs avaient été choisis en s'appuyant sur des polymorphismes entre DC et DE et pas sur l'ensemble des 4 formes évolutives : c'est ce que l'on appelle « l'incertainement biais » (Cavanagh et al. 2013; McTavish and Hillis 2015; Malomane et al. 2018).

Les caractéristiques biologiques de la population étudiée ont été utilisées pour définir des attendus sur les données de génotypage permettant de générer les filtres à appliquer aux SNPs.

Le premier filtre se base sur le taux d'autogamie. Les 120 génotypes analysés ont été considérés comme issus d'une population se reproduisant de manière constante avec un fort taux d'autogamie. De manière arbitraire, nous avons fixé un seuil d'autogamie attendu (S) de 95 % minimum. Cela nous permet d'estimer l'indice de fixation : FIS, qui mesure l'écart à la panmixie (déficit d'hétérozygote) dû à la consanguinité du régime de reproduction au sein de la population. Dans notre cas :

$$FIS = \frac{S}{2 - s} = \frac{0.95}{2 - 0.95} = 0.9$$

Une valeur de FIS attendue à 0.9 est passée comme paramètre à « Read2SNP » (- fis 0.9). L'algorithme de READ2SNP estime tout d'abord le taux d'erreur dans le cadre du maximum de vraisemblance, en fonction du FIS déclaré de la population. Puis, il calcule la probabilité de chacun des génotypes possibles connaissant le taux d'erreur, en supposant l'équilibre de Hardy-Weinberg. Dans cette étude, nous avons choisi de valider le génotypage à chacune des positions lorsque cette probabilité (pvalue) est supérieure à 0.95 (- th1 0.95), soit au moins 95% de chance que le génotype annoncé, à une position donnée, soit correct.

Le deuxième filtre se base sur le taux d'hétérozygotie. En effet, les 120 génotypes analysés sont attendus très majoritairement homozygotes. Ce filtre a été nécessaire car les copies de gènes

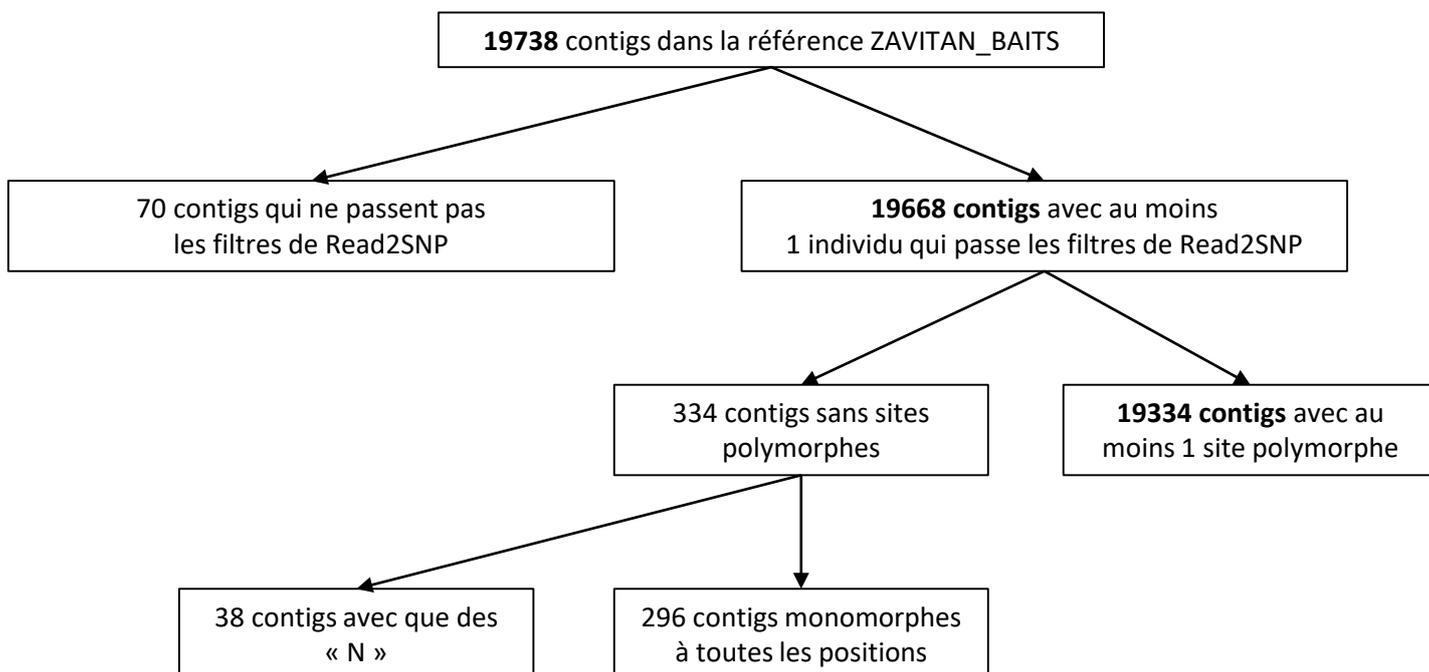


Figure 31: Impacts des filtres sur le nombre de contigs de la référence ZAVITAN_BAITS portant des SNPs
 Différents filtres sont appliqués aux SNPs afin d’augmenter leur fiabilité. Parmi les 19 738 contigs de la référence ZAVITAN_BAITS, 19 334 portent au moins un site polymorphe respectant les filtres.

paralogues sont une source potentielle de SNPs parasites. Pour remédier à ce problème, nous avons utilisé l'option « paraclean » (- par 1) de Read2SNP, qui a pour objectif de filtrer les SNPs détectés sur une paralogie potentielle grâce à un test de rapport de vraisemblance. Ce test fonctionne de la façon suivante : pour un SNP donné, la probabilité des données observées (proportion de A, C, G et T chez chaque individu) a été calculée, d'une part, selon un modèle à un locus et, d'autre part, sous un modèle à deux locus. Le modèle à deux locus suppose que deux loci paralogues contribuent à la lecture de ce SNP, entraînant un excès de génotypes hétérozygotes et un déséquilibre par rapport à la proportion attendue de 50% / 50% entre les deux allèles possibles à ce locus. Les SNPs ont été validés lorsque le modèle à deux locus n'améliorait pas la valeur de la probabilité, de manière significative ($p\text{-val} < 0,001$), par rapport au modèle à un seul locus pour le seuil de FIS retenu.

Par ailleurs, nous avons choisi d'appliquer une profondeur minimum de 5 lectures (5 reads) pour valider le génotype, pour chacune des positions (-min 5).

Lorsque ces trois paramètres ne sont pas respectés, Read2SNP déclare la position en données manquantes (notées « - »). Parmi les 19 738 contigs de la référence ZAVITAN_BAITS, 70 contigs ne respectent pas au moins un des trois filtres ci-dessus. Sur les 19 668 contigs restants, 19 334 portent au moins un site polymorphe qui respecte le filtre sur le FIS et possèdent au moins deux homozygotes différents. Par ailleurs, parmi les 334 contigs qui ne portent pas de sites polymorphes, 296 sont totalement monomorphes pour les 120 génotypes et 38 contigs ne possèdent aucun locus qui respecte les filtres (séquences composées seulement de « N ») (figure 31).

L'outil READ2SNP produit cinq fichiers ayant différents formats, les informations contenues dans deux d'entre eux (.aln et .gen), permettent, une fois rassemblés, d'obtenir le génotypage des 120 génotypes à toutes les positions de la référence génomique simplifiées ZAVITAN_BAITS. La constitution de cette matrice s'effectue grâce à un script en langage python, qui permet également de calculer, pour chaque SNP, différents paramètres comme : le nombre d'individus génotypés, le nombre d'allèles, le nombre de génotypes homozygotes, le nombre de génotypes hétérozygotes ou le FIS. Le calcul de ces différents paramètres permet d'appliquer deux filtres supplémentaires afin d'assurer la fiabilité des données.

Le premier filtre a pour objectif d'éliminer les SNPs pour lesquels on ne retrouve pas, sur l'ensemble des 120 génotypes, les deux allèles possibles à l'état homozygote. Cela se justifie de la façon suivante : si un des deux allèles n'est présent qu'à l'état hétérozygote, cela signifierait qu'il y ait eu une mutation récente au niveau de ce locus/SNP, or les 120 génotypes étudiés sont des lignées fixées depuis au moins six générations. Nous avons donc choisi d'éliminer tous les SNPs ayant ce profil, issus soit d'une erreur de séquençage sur un individu, soit d'une interaction entre locus paralogues non détectés par l'option « paraclean » de Read2SNP.

Le deuxième filtre permet d'éliminer les SNPs pour lesquels la valeur du FIS (taux d'hétérozygotie observé), sur les 120 génotypes, est inférieure à 0.9. Lors de la détection des SNPs par Read2SNP, nous avons annoncé un FIS attendu à 0.9 dans la population et nous avons conservé les SNPs pour lesquels la probabilité (pvalue), tenant compte du FIS attendu, était supérieure à 0.95. Cependant, malgré cela, certains SNPs ont encore des valeurs faibles de FIS. Ce dernier filtre sur le FIS, nous a donc permis d'augmenter, encore une fois, la fiabilité des données.

Encadré 3 : Analyse en Composantes Principales (ACP)

L'ACP est un outil incontournable pour identifier des corrélations entre les variables (SNPS) qui sont le signe de structures génétiques, dans de très grands ensembles de données, en un temps de calcul très réduit et en l'absence d'hypothèse sur le modèle génétique sous-jacent de la population étudiée. Cette méthode statistique consiste à transformer un ensemble de variables possiblement corrélées en un nouvel ensemble de variables orthogonales et donc non corrélées, appelées composantes principales. La première composante principale est définie de façon à représenter au maximum la diversité présente dans le jeu de données. De la même manière, la deuxième composante principale correspond à la droite affine maximisant la variance restante, sous la contrainte d'être orthogonale à la composante principale précédente. Les composantes principales suivantes se déduisent donc des précédentes en suivant ce schéma itératif. L'objectif étant de représenter au mieux la variation existant dans le jeu de données avec un petit nombre de composantes principales nécessaire et suffisant pour représenter au mieux la diversité génétique. Le choix du nombre de composantes principales à conserver se fait grâce à l'histogramme des valeurs propres de chacune des composantes principales et à la proportion de la variance portée par chaque axe (l'inertie cumulée).

Encadré 4 : Analyse Discriminante en Composante Principale (DAPC)

La différenciation entre groupes d'individus peut être étudiée à l'aide d'une Analyse Discriminante en Composante Principale (DAPC) (Jombart, 2010). Sur la base d'un algorithme de K-means cette méthode permet dans un premier temps d'assigner les génotypes à des groupes différents. Pour chaque valeur de nombre de groupe (clusters K) une mesure de l'adaptation du modèle aux données est calculée : c'est le Bayesian Information Criterion (BIC). La deuxième étape débute par la réalisation d'une Analyse en Composantes Principales (ACP) pour construire des variables non corrélées, combinaison linéaire des variables initiales. Le choix du nombre de composantes principales à conserver s'effectue à l'aide d'une validation croisée (cross-validation) qui a pour objectif de capter uniquement la variation générale, qui permet la différenciation des groupes formés. Cette validation croisée consiste à diviser l'échantillon en deux sous-échantillons, le premier dit d'apprentissage et le second dit de test. Le choix du nombre de composantes est réalisé sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'erreur est estimée en calculant un score de performance du modèle pour l'attribution des individus aux groupes sur l'échantillon de test : c'est l'erreur quadratique moyenne. Ici, l'adaptation des données au modèle est estimée par le Mean Squared Error (MSE). L'objectif étant de choisir le nombre de composantes principales qui représente un compromis entre pouvoir de discrimination et flexibilité, dans le but de pouvoir appliquer le modèle à un autre échantillon. La troisième étape est une Analyse Discriminante (DA). Elle a pour objectif de définir de nouvelles variables appelées discriminantes de façon à maximiser la variation inter-groupe et minimiser la variation intra-groupe. Cette méthode permet donc de décrire la discrimination des groupes prédéfinis, en utilisant les composantes principales de l'ACP définis dans l'étape précédente. Le nombre d'axes discriminants dépend du nombre de groupe K déterminé lors de l'étape de regroupement (axes DA= K-1).

Encadré 5 : Factorisation de matrice non négative (sNMF)

la méthode snmf (*sparse non-negative matrix factorization*), qui permet, à l'aide de la factorisation de matrice non négative, d'estimer statistiquement les coefficients d'admixture (métissage génétique) individuels à partir d'un échantillon de génotypes multilocus (Frichot, 2014). Les coefficients d'admixture permettent identifier les proximités génétiques entre individus et définir ainsi la structure de la diversité génétique présente dans notre échantillon. Autrement dit, cette approche permet d'identifier des groupes génétiques homogènes dans l'échantillon et d'identifier des individus d'origine mixte. Cette méthode utilise un codage binaire des SNPs afin de pouvoir estimer séparément la fréquence d'allèle ancestral de chaque génotype. De ce fait, le logiciel sNMF ne nécessite pas que le jeu de données vérifie l'hypothèse d'Hardy-Weinberg et ne requiert pas d'information *a priori* sur les populations. Le choix du nombre de groupes qui décrit au mieux la structure génétique de l'échantillon est faite à l'aide d'un calcul statistique appelée « cross-entropy » basée sur une approche de validation croisée. Pour cela, à chaque nombre de groupe (noté K), l'analyse est faite sur un sous-échantillon des individus disponibles et les fréquences alléliques de chaque groupe sont utilisées pour prédire l'appartenance à chaque groupe des individus restants. La « cross-entropy » mesure donc l'ajustement du modèle aux données (appartenance des individus aux groupes). La valeur de K qui minimise la « cross-entropy » est celle qui décrit mieux les données.

2.3.7 Mise en forme des données

Pour effectuer les analyses de structure génétiques de nos quatre groupes génétiques, ainsi que pour les analyses de différenciation basées sur le F_{st} , il est nécessaire d'avoir les données sous la forme d'un tableau de génotypage : une colonne par génotype et une ligne par SNP. Nous avons, 135 863 SNPs filtrés, répartis sur les 19 334 contigs de la référence ZAVITAN_BAITS. Pour cela, j'ai utilisé un script en langage python, qui transforme le fichier produit par Read2SNP, après les différents filtres, en tableau de génotypage. Les données de génotypages sont ensuite encodées. Dans cette étude, nous avons considéré que pour un locus donné, il n'existait que deux nucléotides possibles, dont l'un sera considéré comme l'allèle de référence et l'autre comme l'allèle alternatif. Pour un locus donné, l'encodage des données de SNPs consiste à attribuer à chaque individu la valeur 0 s'il est homozygote pour l'allèle de référence, la valeur 1 s'il est hétérozygote et la valeur 2 s'il est homozygote pour l'allèle alternatif. Le choix de l'allèle de référence et de l'allèle alternatif peut se faire de façon totalement arbitraire sans que cela n'influe sur les méthodes statistiques utilisées basées sur la variance des allèles.

Par ailleurs, pour effectuer les analyses de diversité génétique, j'avais besoin d'une matrice de génotypage sous forme de fichier séquences au format fasta. Grâce à un script en langage python (coll. Ranwez V., Chantret N., non publié), j'ai reconstitué les séquences des deux allèles, pour chacun des 120 génotypes, à toutes les positions de la référence génomique simplifiées ZAVITAN_BAITS en prenant, à chacune des positions, un des deux allèles possibles. Nous avons donc obtenu, 19 668 fichiers fasta correspondant aux contigs qui contenaient au moins un site qui passait les filtres appliqués par READ2SNPS (couverture, pvalue, paralogie) (sur les 19 738 contigs au total). Chacun de ces fichiers contenait les séquences des deux allèles des 120 génotypes, soit un total de 240 séquences.

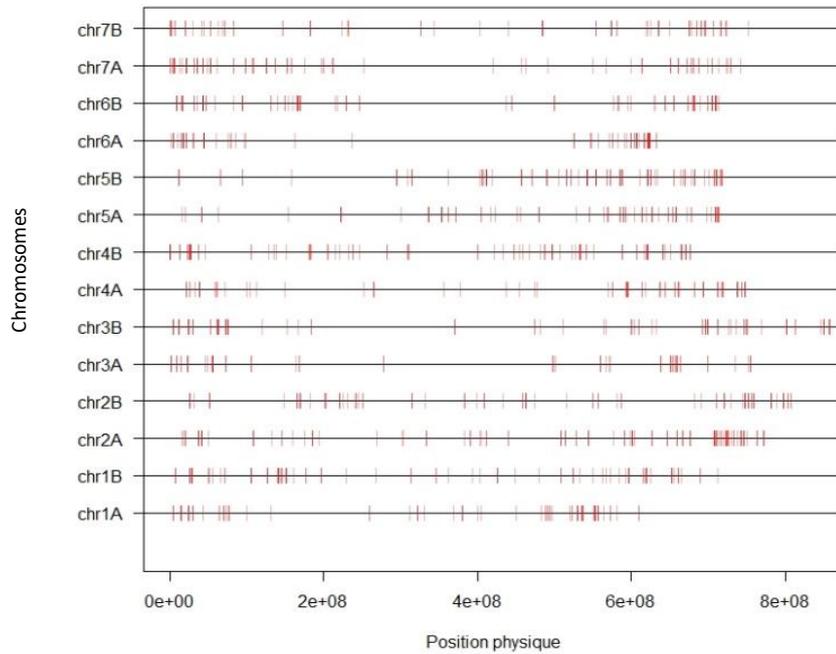


Figure 32: Répartition des 683 contigs de la référence ZAVITAN_BAITS, sélectionnés pour l'analyse de structure, sur les 14 chromosomes

Chaque trait rouge vertical représente un contig de la référence ZAVITAN_BAITS. Ils sont disposés sur les 14 chromosomes en fonction de leur position physique. Les 683 contigs présentés ci-dessus sont ceux qui portent les 1523 SNPs pour lesquels le génotypage est disponible pour les 120 individus étudiés. Ces SNPs sont ceux qui ont servi à l'analyse de la structure génétique de l'échantillon.

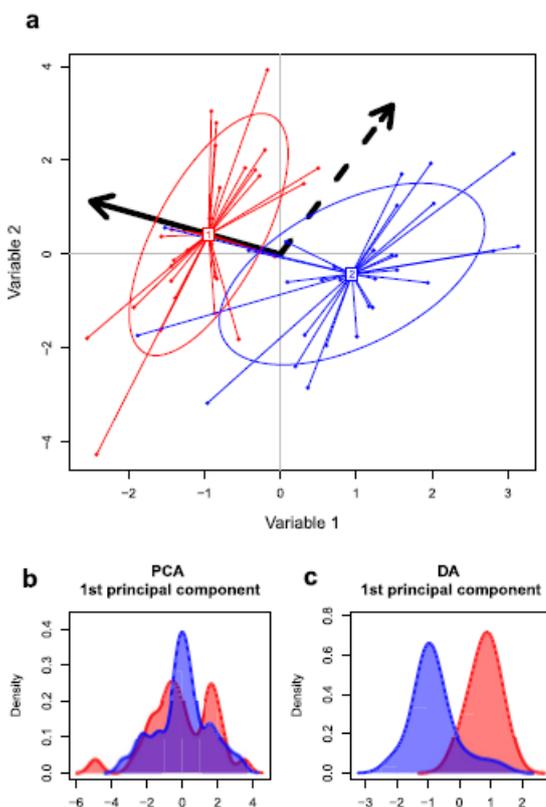


Figure 33: Différence entre ACP et DAPC (Jombart, 2010)
 (a) Le diagramme montre la différence essentielle entre l'Analyse en composantes principales (ACP) et l'Analyse Discriminante (AD).

Les individus (points) et les groupes (couleurs et ellipses) sont positionnés sur le plan en utilisant leurs valeurs pour deux variables. Dans cet espace, l'ACP recherche l'axe montrant la variance totale la plus grande (flèche en pointillés), alors que l'AD maximise la séparation entre les groupes (flèche trait plein) tout en minimisant les variations au sein du groupe. Par conséquent, l'ACP ne parvient pas à discriminer les groupes (b), tandis que DA affiche les différences de groupe (c).

2.4 Statistiques et génétique des populations

2.4.1 Structure génétique de la série de domestication

Pour étudier la structure de la série de domestication, composée des 4 formes : *T. turgidum ssp dicoccoïdes*, *T. turgidum ssp dicoccum*, *T. turgidum ssp durum* « population » et *T. turgidum ssp Durum* « élite », nous avons utilisé différentes méthodes. Pour analyser la structure génétique d'un échantillon, quelques centaines de marqueurs répartis sur l'ensemble du génome sont suffisants. J'ai échantillonné les SNPs sans données manquantes, ce qui évite ainsi l'imputation. Ce faisant, j'ai fait l'hypothèse que les données manquantes étaient réparties aléatoirement entre les locus et que la sélection des SNPs pour lesquels le génotypage a été validé pour les 120 individus étudiés ne générerait pas de biais. Sur cette base, 1523 SNPs répartis sur 683 contigs de la référence ZAVITAN_BAITS ont été utilisés (figure 26). La répartition des SNPs utilisés est assez homogène au niveau du génome, hormis sur les zones centromériques qui sont moins polymorphes et moins échantillonnées par les baits car moins riches en gènes. Par ailleurs, la densité en contigs contenant des SNPs polymorphes est proportionnellement conservée lorsque nous travaillons avec le sous-échantillon de 683 contigs par rapport à l'ensemble des SNPs (figure 32).

L'analyse de la structure a été réalisée à l'aide d'une Analyse en Composante Principale (ACP) (encadré 3) avec la fonction `dudi.pca` du package R « `ade4` » (Chessel et al. 2004). La sélection du nombre de composantes principales a été effectuée sur la base de la visualisation des valeurs propres de chacune des composantes principales et de la proportion de la variance portée par chaque axe (l'inertie cumulée).

La différenciation entre groupes d'individus a été étudiée à l'aide d'une Analyse Discriminante en Composante Principale (DAPC) (encadré 4) avec la fonction `dapc` du package R « `adegenet` » (Jombart et al. 2010). La particularité de cette méthode est de définir de nouvelles variables appelées discriminantes de façon à maximiser la variation inter-groupe et minimiser la variation intra-groupe (figure 33). La validation croisée pour le choix du nombre de composantes principales a été effectuée avec 100 répétitions.

Nous avons ensuite cherché à déterminer la constitution génétique de chacun des 120 génotypes. Pour cela, nous avons choisi d'utiliser la méthode sNMF (encadré 5) avec la fonction `snmf` du package R « `LEA` » (Frichot et al. 2014). La structure génétique a été recherchée sans *a priori*, en demandant à l'algorithme d'évaluer plusieurs configurations possibles en faisant varier le nombre de groupes génétiques (K). Pour chaque valeur de K choisies, nous avons réalisé un barplot où est représentée, pour chaque individu, la proportion de son génome issu des différents groupes génétiques (admixture).

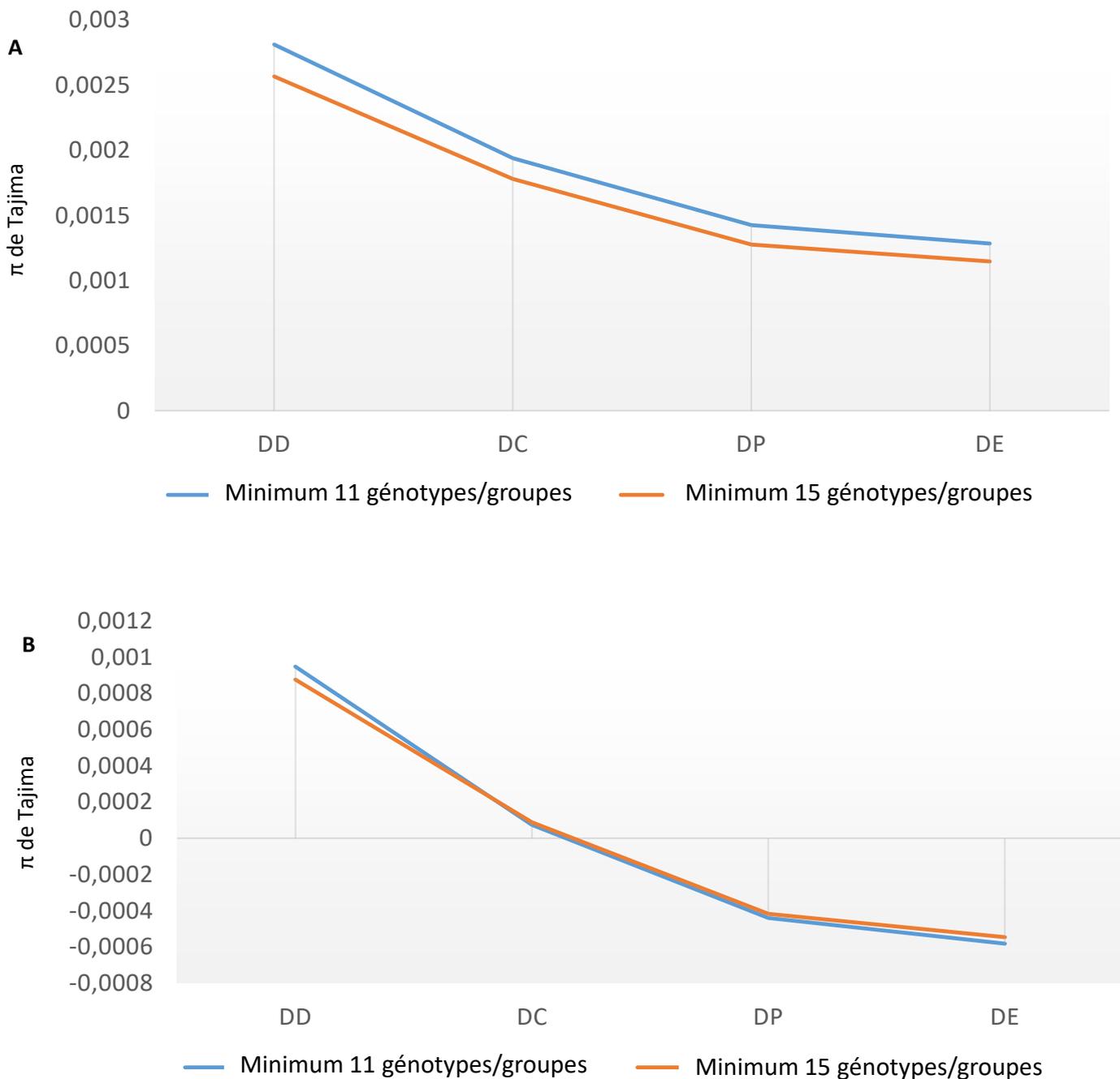


Figure 34: Effet du nombre de données manquantes sur l'estimateur de diversité génétique: π de Tajima
 Pour calculer les différents paramètres de diversité, avec l'outil EglibV3, il est nécessaire de fixer un seuil maximum de données manquantes. Nous avons testé deux seuils différents: minimum 11 génotypes par groupe (en bleu) et minimum 15 génotypes par groupe (en orange). Pour ces deux seuils, les valeurs moyennes du π de Tajima pour chacun des 4 groupes sont calculées (A). Afin de voir si l'évolution de la diversité au cours de la domestication suit la même tendance avec les deux seuils, les valeurs de π ont été centrées sur la moyenne, nous observons alors que les deux courbes se superposent (B).

2.4.2 Caractérisation des effets démographiques de la série de domestication

La diversité génétique d'une population peut être appréhendée par le niveau de polymorphisme génétique qu'elle contient. A l'équilibre mutation dérive, ce polymorphisme génétique θ est attendu fonction du produit entre N_e (taille efficace de la population) et μ (le taux de mutation) :

$$\theta = 4N_e\mu$$

où N_e est la taille efficace de la population, et μ le taux de mutation.

Plusieurs estimateurs de θ peuvent être calculés sur des données de séquences. Deux d'entre eux sont le π de Tajima (1983) qui est la probabilité que deux séquences tirées au hasard soit différentes à un site donné, et le θ_s de Watterson (1975) qui correspond au nombre de sites polymorphes S observés dans un échantillon de séquences d'allèles (présentés en détail dans la partie 1.2.4).

Par ailleurs, le D de Tajima, présenté en détail dans la partie 1.1.4, permet de tester la neutralité (équilibre mutation dérive) et donc de détecter des événements démographiques, sélectifs ou migratoires, tel que les goulots d'étranglements apparus au courant de la domestication. Un D de Tajima positif indique un excès d'allèles à fréquence intermédiaire et donc un goulot d'étranglement, une sélection balancée récente ou une sous structuration. A contrario, un D de Tajima négatif indique au contraire un excès de polymorphisme (un excès de variants rares) et donc une population en expansion ou un balayage sélectif.

Ces trois estimateurs sont calculés, pour chacun des groupes évolutifs de la série de domestication dans le but d'apporter une première estimation sur la perte globale de diversité génétique.

Pour cela, j'ai utilisé l'outil Egglib v3 (3.0.0b15) (De Mita and Siol 2012) intégré dans un script en python qui prend en entrée les alignements contenant la séquence des deux allèles de chacun des 120 génotypes (format multifasta). Le script renvoie un tableau de résultats contenant pour chacun des contigs (19 688) et chacun des quatre groupes (DD, DC, DP et DE) : la longueur de la séquence analysée (l_{seff}), le nombre moyen d'individus génotypés sur cette longueur (n_{seff}), le nombre de polymorphisme (S), le π de Tajima, le θ_s de Watterson et le D de Tajima.

Pour calculer ces estimateurs, il a été nécessaire de fixer le nombre maximum de données manquantes possibles par groupe de 30 génotypes soit pour 60 séquences (2 allèles). Pour cela, deux seuils différents ont été testés afin d'observer l'impact du nombre maximum de données manquantes sur les deux estimateurs :

- ✓ 15 données manquantes soit un minimum de 15 individus génotypés par groupe ($NA < 0.5$)
- ✓ 19 données manquantes soit un minimum de 11 individus génotypés par groupe ($NA < 0.66$)

Pour déterminer le seuil à utiliser pour l'analyse de la diversité génétique, j'ai comparé les moyennes, sur l'ensemble des contigs, du π de Tajima, pour les deux seuils. Nous pouvons voir sur la figure 34 que les courbes des valeurs de π représentant l'évolution de la diversité au cours de la domestication suivent la même tendance avec les deux seuils (figure 34 A). Après avoir centré les données sur la moyenne, nous observons que les deux courbes se superposent (figure 34 B). Ces résultats ainsi que des études antérieures (Clément et al. 2017; Sauvage et al. 2017), nous ont permis de conclure que 11

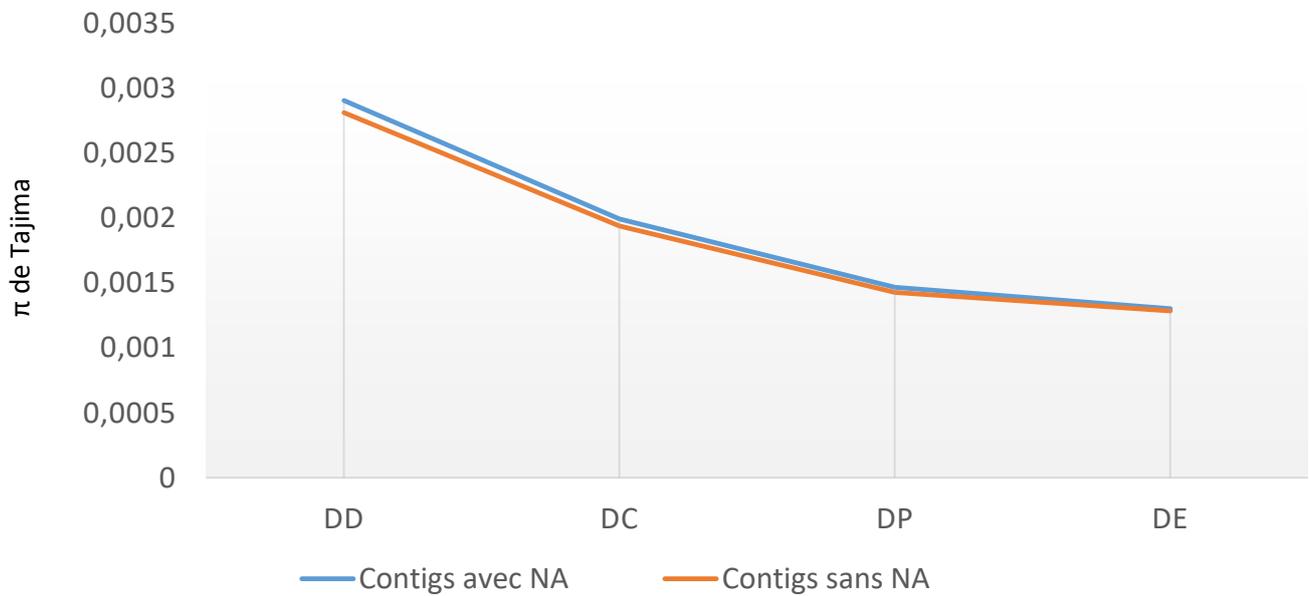


Figure 35: Choix de conservation des contigs

Les filtres appliqués lors de la détection des SNPs impactent la taille de la séquence analysées par contigs et par groupe (l_{seff}). Par ailleurs, lorsque la valeur du l_{seff} est inférieure à 100pb, nous avons fait le choix de ne pas considérer les paramètres génétiques (π , θ_s et D) qui seront notées « NA ».

Pour chaque contig, il y a donc des groupes où les paramètres de diversité ont été calculés et d'autres pour lesquels l'information est manquantes « NA »). Afin de savoir si nous pouvions supprimer l'ensemble des contigs pour lesquels les paramètres de diversité n'avait pas été calculés sur les 4 groupes sans biaiser les données, nous avons comparés les valeurs de l'estimateur π de Tajima sur l'ensemble des contigs, même ceux contenant des NA (en bleu) ou seulement sur les contigs qui ne contenait pas de NA (en orange). Nous pouvons voir que les deux séries de valeurs sont très proches.

génotypes par groupe permettent aussi bien de documenter le goulot d'étranglement, dû à la domestication, de notre jeu de données que 15 génotypes. Cela s'explique par le fait, que le nombre de données exploité est très important, rendant aléatoire les individus génotypés pour chacun des contigs, ne créant ainsi pas de biais d'analyse. Par ailleurs, la variance interlocus pour les différents paramètres étant très forte, il est préférable d'avoir le plus grand nombre de locus analysés possible, même s'ils sont représentés par moins de génotypes par groupe. Nous avons donc décidé de calculer l'ensemble des paramètres de diversité avec un minimum de 11 génotypes par groupe ($NA < 0.66$).

Lors de la création de la référence ZAVITAN_BAITS, les contigs étaient constitués d'un minimum de 720pb, cependant, les reads ne couvrent pas toujours l'ensemble du contig. Par ailleurs, l'ensemble des filtres qui ont été appliqués lors de la détection des SNP (READ2SNP) et des filtres de sélection (couverture, FIS, au moins deux homozygotes différents, etc...) ont ajouté des « N » dans les séquences analysées par l'outil Egglib. De ce fait, la longueur de la séquence analysée (Iseff) est très variable d'un contig à l'autre. Cette variabilité influe sur le nombre de polymorphismes présents par contig (plus la séquence est longue plus il y a de chance qu'il y ait des sites polymorphes) et donc sur la valeur du π de Tajima et du θ s de Watterson. Afin de fiabiliser les données, les estimateurs de diversité n'ont pas été pris en compte si la longueur de la séquence analysée (Iseff) était inférieure à 100pb. Cette longueur est variable pour chacun des quatre groupes, ce qui implique que l'impact de ce filtre varie en fonction des groupes et donc que la valeur des estimateurs n'est plus disponible pour tous les groupes en fonction des contigs (« noté « NA »).

Afin de pouvoir comparer, de façon la plus homogène possible, la diversité au sein des groupes, il nous a semblé nécessaire de conserver seulement les contigs pour lesquels les estimateurs de la diversité étaient disponibles pour tous les groupes. Comme précédemment, pour le choix du seuil de données manquantes à appliquer, j'ai comparé les valeurs du π de Tajima avec les deux configurations : « contigs sans NA » ou « contigs avec NA » créées par le filtre sur la longueur de la séquence analysée. La figure 35 nous permet de voir facilement que les courbes sont quasiment superposées et donc, que nous pouvons travailler seulement sur les contigs pour lesquels les estimateurs sont disponibles pour les quatre groupes, sans créer de biais dans l'analyse. Ce dernier filtre, sur les données, réduit le jeu de données à 16101 contigs analysés pour l'étude de la diversité génétique de notre échantillon. Dans un souci de clarté, nous avons donc décidé, d'appliquer ce filtre à l'ensemble des analyses de diversité.

Dans notre étude, étant donné que la taille des contigs de la référence ZAVITAN_BAITS est variable, nous avons choisi de calculer le π de Tajima et le θ s de Watterson par site et non par contigs. Pour cela, nous avons calculé les valeurs moyennes par site de π de Tajima, θ s de Watterson et D de Tajima, en divisant les valeurs obtenues par contig, par la longueur de la séquence analysée (Iseff).

Pour visualiser le niveau de diversité génétique le long de chacun des chromosomes, j'ai calculé les valeurs moyennes de π et θ s par fenêtre glissante. Pour chaque contig, l'outil Egglib v3 calcule la longueur de la séquence analysée, notée « Iseff » pour calculer les différents paramètres de diversité. J'ai donc utilisé ces valeurs de « Iseff » pour créer des fenêtres glissantes de 20 kb de séquences, ce qui permet d'avoir la même quantité d'informations sur la diversité génétique dans chaque fenêtre et pour chacun des groupes. La valeur du « Iseff » étant différente entre les quatre groupes génétiques, les fenêtres ne sont pas toujours alignées mais cela permet d'avoir tout de même une bonne vision globale afin de comparer la diversité entre les groupes, le long des chromosomes. Par ailleurs, le fait

que ces fenêtres se chevauchent de 10Kb permet de lisser les graphiques et implique que chaque polymorphisme participe au calcul de deux valeurs successives sur le chromosome. En utilisant cette méthodologie pour chacune des quatre formes, nous pouvons observer l'évolution du niveau de diversité, le long des chromosomes, au cours de la domestication.

Pour estimer l'impact de l'histoire démographique sur le niveau de différenciation entre les quatre groupes évolutifs, de façon globale sur l'ensemble du génome, nous avons utilisé l'indice de différenciation **Fst** (Wright 1969). Il a été estimé par la méthode de Weir et Cockerham (1984) qui s'appuie sur une analyse de variance. Cette méthode permet d'estimer les composantes du polymorphisme présent intra-individu (entre allèles d'un même locus : $\sigma^2_{all.}$), entre individus ($\sigma^2_{ind.}$) et entre populations ($\sigma^2_{pop.}$). Ces différentes composantes de la variance sont ensuite combinées selon la formule ci-dessous pour déterminer la part totale de la variabilité moléculaire présente entre les populations et d'estimer ainsi le niveau de différenciation entre populations.

$$Fst_{wc} = \frac{\sigma^2_{pop.}}{\sigma^2_{all.} + \sigma^2_{ind.} + \sigma^2_{pop.}}$$

Le mode de calcul des variances tient compte de la taille de l'échantillon, à chaque locus, ce qui permet de prendre en compte la présence de données manquantes. Le Fst est calculé à partir du tableau de génotypage des 120 génotypes après avoir appliqué le filtre sur les données manquantes (le locus n'est considéré que si la fréquence des NA < 0,66) ainsi qu'un filtre sur la fréquence des allèles mineurs (MAF). Pour éviter d'avoir des locus pour lesquels un allèle n'a été observé que sur un génotype dans l'ensemble de la population, nous avons filtrés les SNPs ayant une valeur de MAF inférieure à 5%. Après avoir appliqué ces deux filtres, le jeu de données final était de 35 164 SNPs répartis sur 10 734 contigs. L'estimation du Fst par la méthode de Weir et Cockerham (1984) a été effectuée entre chacune des formes évolutives (« pairwise ») avec la fonction pairwise.WCfst du package R Hierfstat (Goudet 2005). Les intervalles de confiance autour des valeurs moyennes de Fst ont été calculés avec la fonction boot.pfst en réalisant 1000 bootstraps, technique qui consiste à reconstituer des nouveaux jeux de données (1000) en échantillonnant, avec remise, dans le jeu de données initial.

Si le Fst est égal ou très proche de 0, cela signifie, soit qu'il y a de nombreux flux de gènes entre les deux sous-populations (migration), soit que les tailles efficaces (N_e) des deux sous-populations sont restées élevées depuis leur séparation et donc qu'elles se sont peu différenciées. À l'inverse si le Fst est proche de 1, cela traduit une forte différenciation génétique entre les sous-populations, suggérant soit très peu voire aucun flux de gènes entre les populations, soit que le passage d'une forme à l'autre a été associé à un important goulot d'étranglement. D'après Wright (1978), un Fst compris entre 0 et 0.05 révèle une différenciation faible ; un Fst compris entre 0.05 et 0.15 traduit une différenciation modérée ; un Fst entre 0.15 et 0.25 suggère une différenciation importante et au-delà de 0.25, le Fst illustre une différenciation très importante.

Dans le cas particulier de l'analyse d'une série de domestication, on s'attend à ce que le niveau de diversité soit inversement proportionnel au niveau de différenciation. Entre la forme sauvage et la dernière forme cultivée, la diversité diminue alors que la différenciation est la plus élevée. Le Fst mesure l'ampleur du goulot d'étranglement.

2.4.4 Détection de signatures génétiques de sélection liées à la domestication

2.4.4.1 Détection sans *a priori* sur l'ensemble du génome

Après avoir estimé l'évolution de la diversité et de la différenciation génétique qui a eu lieu au cours de la domestication, du fait de l'histoire démographique de l'espèce *Triticum turgidum* (π , θ_s , D et Fst global), nous avons cherché à détecter des signatures génétiques caractéristiques de la sélection.

Cette détection a été effectuée, dans un premier temps sans *a priori*. Pour cela, j'ai estimé la diminution de l'effectif efficace, N_e , en faisant l'hypothèse que la diversité observée (θ) dans chacun des groupes évolutifs est issue de l'équilibre mutation - dérive, avec une taille efficace différente pour chacun des groupes en utilisant la formule :

$$\theta = 4N_e\mu$$

Si nous faisons l'hypothèse que le taux de mutation (μ) est resté constant au cours du temps, pour un locus donné (Schlotterer 2002), nous pouvons écrire que le rapport entre les effectifs efficaces de deux groupes évolutifs (N_{e1} et N_{e2}) correspond à celui des θ :

$$\frac{\theta_1}{\theta_2} = \frac{4 N_{e1} \mu}{4 N_{e2} \mu} = \frac{N_{e1}}{N_{e2}}$$

Ce rapport peut être calculé avec les deux estimateurs de $\theta=4N_e\mu$: le π de Tajima et le θ_s de Watterson sur chacun des contigs de la référence ZAVITAN_BAITS et pour chaque étape de la domestication (Lin et al. 2014).

$$\frac{\theta_{DD}}{\theta_{DC}}, \frac{\theta_{DC}}{\theta_{DP}} \text{ et } \frac{\theta_{DP}}{\theta_{DE}} \quad || \quad \frac{\theta_{sDD}}{\theta_{sDC}}, \frac{\theta_{sDC}}{\theta_{sDP}} \text{ et } \frac{\theta_{sDP}}{\theta_{sDE}}$$

Plus la valeur du rapport entre les deux groupes étudiés est forte, plus le contig montre des signes de sélection importante au cours de la phase de transition concernée.

Méthodologiquement, pour pouvoir effectuer ces rapports, il est nécessaire que les valeurs de l'estimateur situé au dénominateur soient supérieures à zéro et donc inclure au moins un SNP polymorphe sur le contig concerné. Le nombre de contigs utilisés pour estimer le paramètre N_e va donc varier en fonction des groupes génétiques qui seront comparés.

le nombre de contigs utilisés pour le rapport : $\frac{\theta_{DD}}{\theta_{DC}}$ est de 14058 contigs, 12680 contigs pour le rapport : $\frac{\theta_{DC}}{\theta_{DP}}$ et 12502 contigs pour le rapport : $\frac{\theta_{DP}}{\theta_{DE}}$.

Il est important de noter que ce filtre peut supprimer, soit des contigs avec un polymorphisme très faible à nul pour les quatre formes, soit des contigs pour lesquels la totalité de la diversité a disparu d'une forme à l'autre. Le cas échéant, nous pouvons représenter, à l'aide d'histogrammes de fréquence, la distribution des valeurs des θ du groupe placé au numérateur quand la valeur du groupe placé au dénominateur est nulle.

Afin d'identifier les contigs ayant subi les plus fortes diminutions de taille efficace lors d'une phase de transition, nous avons calculé une valeur seuil correspondant au 95^{ème} centile.

Tableau 3: Localisation des gènes TtBtr1-A, TtBtr1-B, Q et Rht-B1b ainsi que les deux QTLs impliqués dans le poids des grains(PMG) et la teneur en azote de la feuille sur la référence ZAVITAN. Six fenêtres de 5Mb ont été positionner autour de ces localisation et l'ensembles des contigs qui les composent ont été analysés.

Gènes	Chromosomes	Localisations sur ZAVITAN		Coordonnées fenêtres 5Mb		Nombres de contigs
TtBtr1-A	3A	64 025 304	64 025 982	61 500 001	66 500 000	6
TtBtr1-B	3B	103 040 911	103 041 584	100 500 001	105 500 000	1
Q	5A	654 780 701	654 784 432	652 000 001	657 000 000	22
Rht-B1b	4B	30 005 125	30 008 097	27 500 001	32 500 000	24
PMG	2A	157 508 711		155 000 001	160 000 000	6
Teneur Azote	3A	35 222 498		33 500 001	38 500 000	9

Une autre méthode de détection de la sélection, basée sur le niveau de différenciation, a été utilisée. Pour cela, le F_{st} a été estimé, à l'échelle des contigs, par la méthode de Weir et Cockerham (1984), pour les trois paires de populations représentant les différentes transitions de l'histoire évolutive (F_{st} pairwise) : DD_DC, DC_DP, DP_DE. Ce paramètre est calculé à partir du tableau de génotypage (35 164 SNPs) des 120 génotypes pour les 10 734 contigs après sélection des SNPs sur le nombre de données manquantes et une MAF supérieure à 5% (figure 26). Nous avons utilisé la fonction pairwise.WCfst package R Hierfstat (Goudet 2005).

Plus la valeur du F_{st} est forte, plus le contig est susceptible de présenter une trace de sélection au cours de la phase de transition concernée.

Pour chaque paire de groupe évolutif, nous avons calculé un seuil correspondant au 95^{ème} centile, c'est-à-dire la valeur du F_{st} contenant 95% de la distribution sur l'ensemble des contigs. Nous considérons que les valeurs de F_{st} supérieures à cette valeur seuil constituent un signal à analyser comme une trace potentielle de sélection au cours de la phase de transition concernée.

2.4.4.2 Détection au niveau de zones candidates contrôlant des traits du syndrome de domestication

Dans un deuxième temps, nous avons recherché les signatures génétiques associés à la sélection de cinq traits phénotypiques caractéristiques de la domestication de l'espèce *Triticum turgidum* :

- ✓ Les gènes **TtBtr1-A** et **TtBtr1-B**, situés respectivement sur les bras courts des chromosomes 3A et 3B et impliqués dans le caractère « rachis solide », pour renseigner la transition entre DD et DC.
- ✓ Le gène **Q**, situé sur le bras long du chromosome 5A et impliqué dans le trait phénotypique « grain nu », pour renseigner la transition en DC et DP.
- ✓ Le gène **Rht-B1b**, situé sur le bras court du chromosome 4B et impliqué dans le trait phénotypique « plante semi-naine » pour renseigner la transition entre DP et DE.
- ✓ Le QTL impliqué dans le **poids des grains (PMG)**, localisé sur le bras court du chromosome 2A.
- ✓ Le QTL impliqué dans la **teneur en l'azote** dans la feuille, qui reflète la stratégie d'acquisition des ressources par la plante, localisé sur le bras court du chromosome 3A.

La séquence des gènes a été récupérée sur la base de données NCBI en prenant les séquences venant de *Triticum aestivum* (blé tendre) pour les gènes TtBtr1-A, TtBtr1-B et Rht-B1b et de *T. turgidum dicoccoïdes* pour le gène Q. Ces trois séquences ont ensuite été positionnées sur la référence « ZAVITAN » afin d'obtenir, pour chacune d'entre elle, une position physique. Concernant les deux QTLs, des analyses de génétique d'association (GWAS) sont actuellement en cours dans l'équipe et nous ont permis de localiser ces QTLs sur la référence ZAVITAN.

Cependant, les 20 000 baits utilisées pour l'enrichissement par capture n'ayant pas été défini pour cibler ces cinq traits, aucun contig de la référence ZAVITAN_BAITS se situent sur les gènes et QTLs ciblés. De ce fait, nous avons travaillé sur des fenêtres de 5Mb sur le génome de référence autour des séquences des gènes ou QTL ciblés. En fonction du gène ou du QTL concerné, le nombre de contigs situées dans la zone de 5Mb varie, d'un seul contig pour la zone correspondant au gène TtBtr1-B à 24 contigs pour la zone correspondant au gène Rht-B1b (tableau 3).

La détection des signatures de sélection pour ces cinq traits a été effectuée à l'aide de l'estimateur de diversité π . Pour chacune de ces cinq fenêtres de 5Mb, nous avons calculé la valeur moyenne de l'estimateur de diversité π sur les contigs qui les composent. Afin de pouvoir mesurer la perte de diversité à chaque transition évolutive, nous avons calculé des différences de valeurs de π entre les deux formes concernées en prenant soin ensuite de les rapporter à la valeur de π dans le compartiment sauvage (valeur de référence) :

- ✓ Pour la transition entre DD et DC : $(\pi_{DD} - \pi_{DC}) / \pi_{DD}$
- ✓ Pour la transition entre DC et DP : $(\pi_{DC} - \pi_{DP}) / \pi_{DD}$
- ✓ Pour la transition entre DP et DE : $(\pi_{DP} - \pi_{DE}) / \pi_{DD}$

Plus la valeur du rapport de π est grande plus la perte de diversité a été importante lors de la transition concernée. Pour savoir si ces observations sont imputables à la sélection et pas seulement à l'histoire démographique, ces valeurs ont été comparées à celles mesurées sur le reste du génome.

Par ailleurs, la détection des signatures de sélection pour ces cinq traits a également été effectuée à l'aide des valeurs de F_{st} de chacun des contigs qui composent les six zones ciblées, en les comparant aux valeurs de F_{st} obtenues sur l'ensemble des autres contigs de la référence ZAVITAN_BAITS.

Résultats

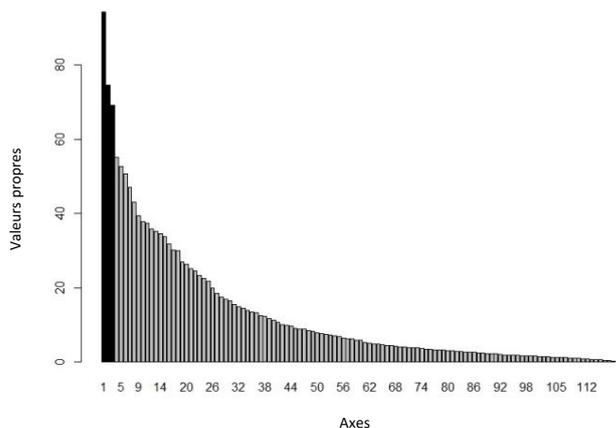


Figure 36: Distribution des valeurs propres de l'Analyse en Composantes Principales (ACP).

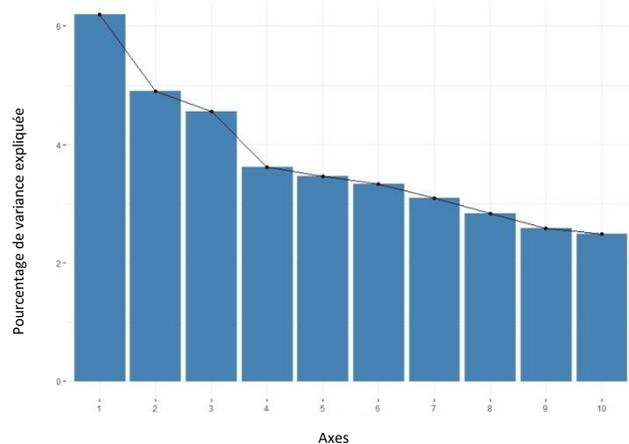


Figure 37: Pourcentage de variance expliquée (axe des Y) pour les dix premiers axes de l'ACP (axe des X).

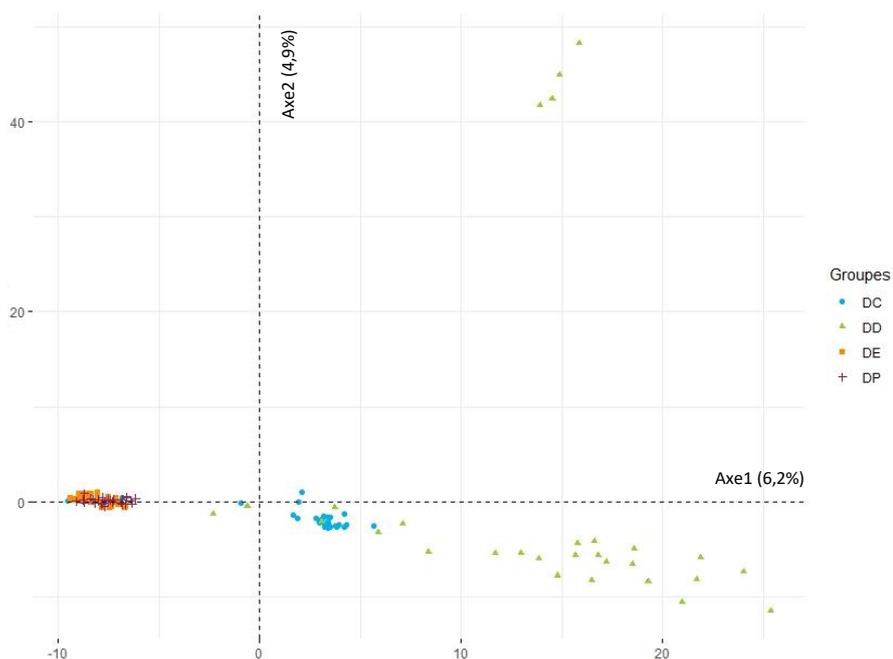


Figure 38: Projection des 120 individus sur les axes 1 et 2 de l'ACP.

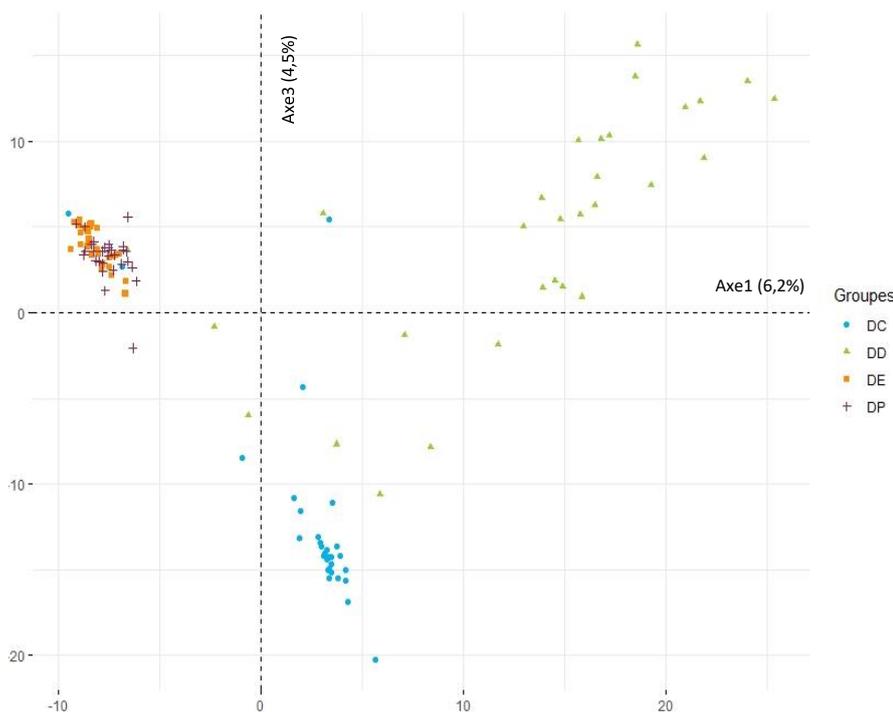


Figure 39: Projection des 120 individus sur les axes 1 et 3 de l'ACP.

3 Résultats

Cette partie présente l'ensemble des résultats obtenus sur la structure génétique de la série de domestication ainsi que sur la caractérisation des effets démographiques et la détection de signatures génétiques de sélection liées à l'histoire de l'espèce *Triticum turgidum*.

3.1 Structure génétique de la série de domestication

La structure génétique de la série de domestication a été étudiée à partir de la matrice de génotypage des 120 génotypes pour les 1523 SNP répartis sur les 683 contigs de la référence ZAVITAN_BAITS.

3.1.1 Mise en évidence de groupes génétiques

Sur la base des différentes valeurs propres (figure 36) des axes de l'Analyse en Composantes Principales (ACP), j'ai fait le choix de ne retenir que les trois premiers axes principaux. Ils expliquent respectivement 6,2 ; 4,9 et 4,5 % de la variabilité (figure 37) soit en cumulé, 15.6% de l'information génétique totale (Annexe 4).

La projection des 120 génotypes sur le plan constitué par les axes 1 et 2 (figure 38) permet d'observer que les différents groupes évolutifs se répartissent le long de l'axe 1 (6.2%) selon le gradient de domestication : les 120 génotypes sont répartis en trois groupes qui se superposent assez bien avec les trois groupes évolutifs analysés: le groupe DD a des coordonnées très positives, le groupe DC des coordonnées positives mais plus proches de 0 et les coordonnées sont négatives pour la forme à grains nus qui regroupe DP et DE. La forme des nuages de points correspondant à chacun des groupes évolutifs est différentes : elle est très compacte pour DP et DE, et plus étalée pour DC et DD. Cela traduit un niveau de diversité intra-groupe plus important pour DD, que pour les autres groupes et notamment DP et DE.

L'axe 2, qui représente 6.2% de la variance totale, met en évidence quatre génotypes DD (Tc2227, Tc2451, Tc2460 et Tc3309), génétiquement très différents des autres génotypes de ce groupe mais aussi des autres génotypes présents dans l'échantillonnage.

Par ailleurs, nous pouvons voir, grâce à la projection des individus sur les axes 1 et 3 de l'ACP (figure 39), que l'axe 3 (4.5% de la variabilité totale) permet de mieux visualiser la diversité au sein du groupe DC et de séparer, plus distinctement, les génotypes des groupes DD et DC.

Les trois premiers axes de l'ACP opposent les groupes DD et DC aux deux formes cultivées plus récentes : DP et DE. Ils permettent également d'observer les différences de variabilité intra-groupes plus importantes pour DD et DC par rapport à DP et DE. Sur la base de ses trois premiers axes, cette analyse en composantes principales ne permet pas de distinguer les groupes DE et DP qui ne forment qu'un seul ensemble compact quel que soit l'axe considéré.

La première étape de l'Analyse Discriminante en Composante Principale (DAPC) a été d'utiliser la valeur du Bayesian Information Criterion (BIC) pour définir la structuration de l'échantillon analysé et donc le nombre le plus probable de groupes, que nous appellerons clusters (K), dans lesquels se répartissent les différentes accessions analysées (figure 40). Le modèle le plus probable, et donc le mieux adapté aux données, est composé de cinq clusters (K=5). La valeur du MSE est la plus faible avec dix composantes principales. Le modèle le plus optimisé est donc constitué de dix composantes principales (MSE) et de quatre axes discriminants (K-1) (Annexe 5).

Figure 40: Valeurs du Bayesian Information Criterion (BIC) en fonction du nombre de clusters (K), pour l'analyse de la structure des 120 génotypes (DD, DC, DP et DE) avec DAPC.

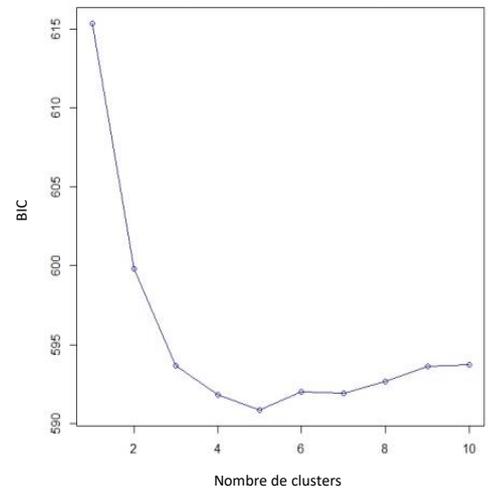


Tableau 4: Répartition des 120 génotypes (DD, DC, DP, DE) dans les cinq clusters DAPC.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
DD	1	20	1	4	4
DC	1	1	1	0	27
DP	25	0	5	0	0
DE	3	0	27	0	0

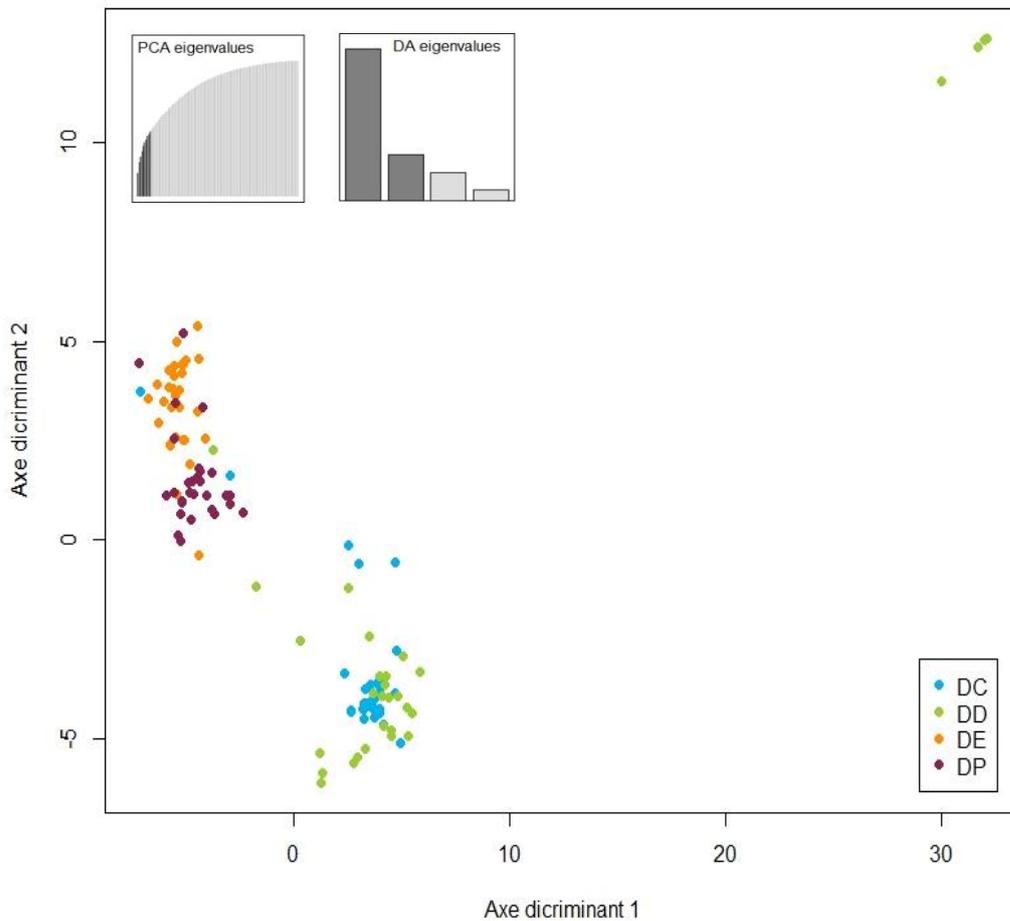


Figure 41: Projection des 120 génotypes (DD, DC, DP, DE) sur les deux premiers axes discriminants par la méthode DAPC.

Figure 42: Valeurs du Bayesian Information Criterion (BIC) en fonction du nombre de clusters (K), pour l'analyse de la structure des 90 génotypes (DC, DP et DE) avec DAPC.

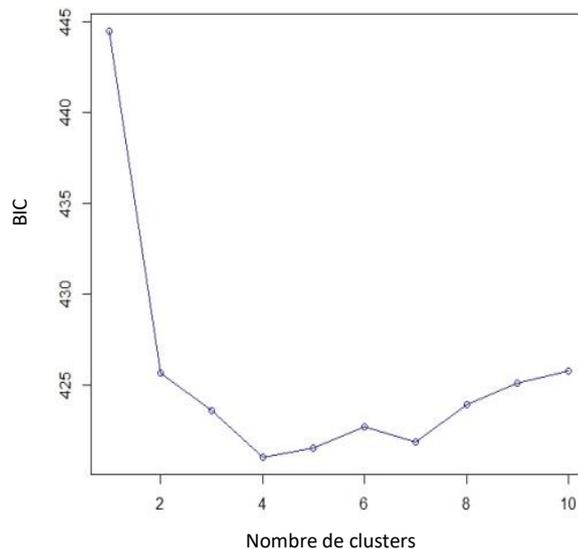


Tableau 5: Répartition des 90 génotypes (DC, DP, DE) dans les quatre clusters DAPC.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
DC	5	2	1	22
DP	0	25	5	0
DE	0	3	27	0

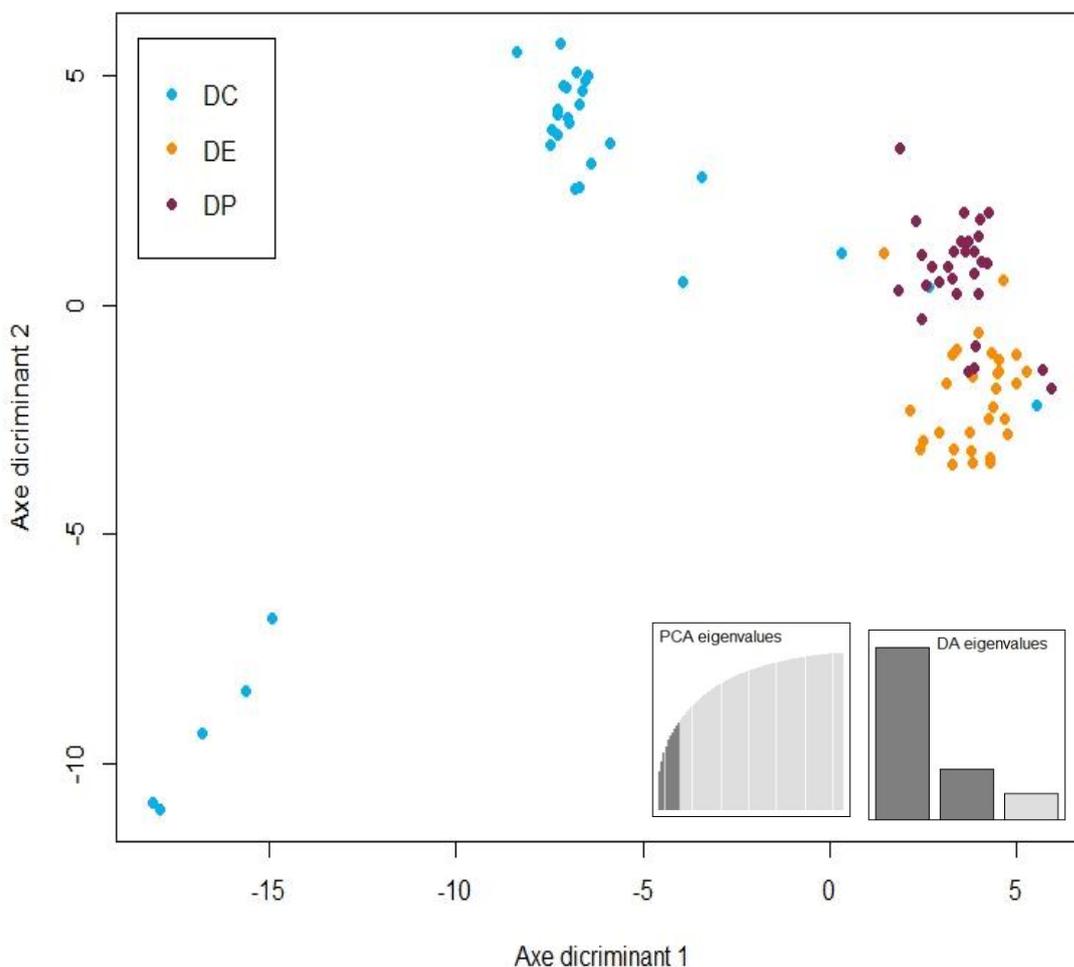


Figure 43: Projection des 90 génotypes (DC, DP, DE) sur les deux premiers axes discriminants par la méthode DAPC.

A l'aide de la composition des cinq clusters (tableau 4), les 120 génotypes ont été projetés sur les deux premiers axes discriminants, maximisant la variance entre les différents groupes de notre échantillon (figure 41).

Parmi les 30 génotypes du groupe DD, 20 sont situés dans le cluster 2 ; les quatre génotypes (Tc2227, Tc2451, Tc2460 et Tc3309) qui se singularisaient dans l'analyse ACP, constituent à eux seuls le cluster 4. Par ailleurs, les six génotypes restants, appartenant au groupe DD, ont été assignés dans trois autres clusters : le génotype Tc3302 se retrouve dans le cluster 1 constitué en majorité de génotypes du groupe DP, le génotype Tc3203 dans le cluster 3 porté par les génotypes du groupe DE et les génotypes Tc2220, Tc2398, Tc2454 et Tc3401 dans le cluster 5, constitué majoritairement des génotypes du groupe DC.

Les génotypes du groupe DC se retrouvent pour l'essentiel dans le cluster 5 (27 / 30 génotypes) ; les trois génotypes restants ont été assignés aux cluster 1 (génotype Tc2462), 2 (génotype Tc2218) et 3 (génotype Tc3205).

Au sein du groupe DP, 25 génotypes se retrouvent dans le cluster 1 et 5 génotypes (Tc2549, Tc2611, Tc2818, Tc 3307 et Tc 4309) dans le cluster 3.

Pour finir, les 30 génotypes du groupe DE sont très majoritairement situés dans le cluster 3 ; seuls trois (Tc2254, Tc2258 et Tc2675) ont été placés dans le cluster 2.

Les barycentres des clusters proposés par DAPC correspondent, approximativement, aux groupes génétiques, à l'exception de DD (DD : clusters 2 et 4, DC : cluster 4, DP : cluster 1 et DE : cluster 3).

Après cette première analyse, nous observons que la diversité génétique intra-groupe et inter-groupe chez la forme DD est bien plus importante que celle des trois autres groupes. De ce fait, lors de la projection des individus sur les axes discriminants, la diversité au sein des groupes DC, DP et DE est difficilement observable. J'ai donc refait l'Analyse Discriminante en Composante Principale (DAPC) sans les 30 génotypes DD en utilisant la même méthodologie qu'avec le jeu de données complet.

Dans ce cas, l'hypothèse de quatre clusters (K=4) elle celle qui décrit le mieux notre nouveau jeu de données (BIC) (figure 42) et la valeur du MSE est la plus faible avec dix composantes principales. Le modèle le plus optimisé est donc constitué de dix composantes principales et de trois axes discriminants (K-1) (Annexe 6).

La composition des quatre clusters (tableau 5), nous a permis de projeter les 90 génotypes sur les deux premiers axes (figure 43).

Parmi les 30 génotypes du groupe DC, 24 sont situés dans le cluster 4, et cinq génotypes (Tc2212, Tc2213, Tc2465, Tc2503 et Tc3408) constituent à eux seuls le cluster 1. Les trois génotypes restants ont été assignés dans deux autres clusters. Les génotypes Tc2218 et Tc2462 se retrouvent dans le cluster 2 en majorité constitué de génotypes du groupe DP, et le génotype Tc3205 dans le cluster 3 constitué majoritairement de génotypes DE. Cette répartition met en évidence le niveau élevé de variabilité au sein du groupe DC.

Les génotypes du groupe DP sont répartis dans deux clusters : 25 génotypes dans le cluster 2 et cinq génotypes (Tc2549, Tc2611, Tc2818, Tc3307 et Tc4309) ont été placés dans le cluster 3.

Enfin, les génotypes du groupe DE se retrouvent majoritairement situés dans le cluster 3, seuls trois génotypes (Tc2254, Tc2258 et Tc2675) étant placés dans le cluster 2.

La répartition des génotypes DP et DE dans deux clusters communs (2 et 3), en proportion quasiment inversée, est cohérente avec l'analyse précédente réalisée sur l'ensemble des génotypes.

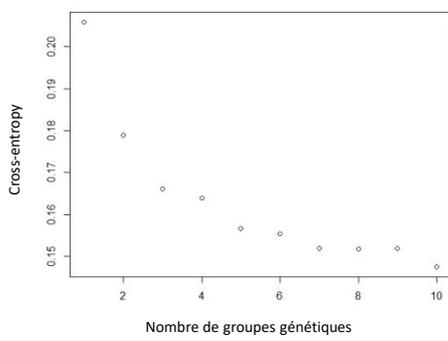


Figure 44: Valeurs de Cross-entropy pour 1 à 10 groupe génétiques (K) par la méthode sNMF.

K=4

K=5

K=6

K=7

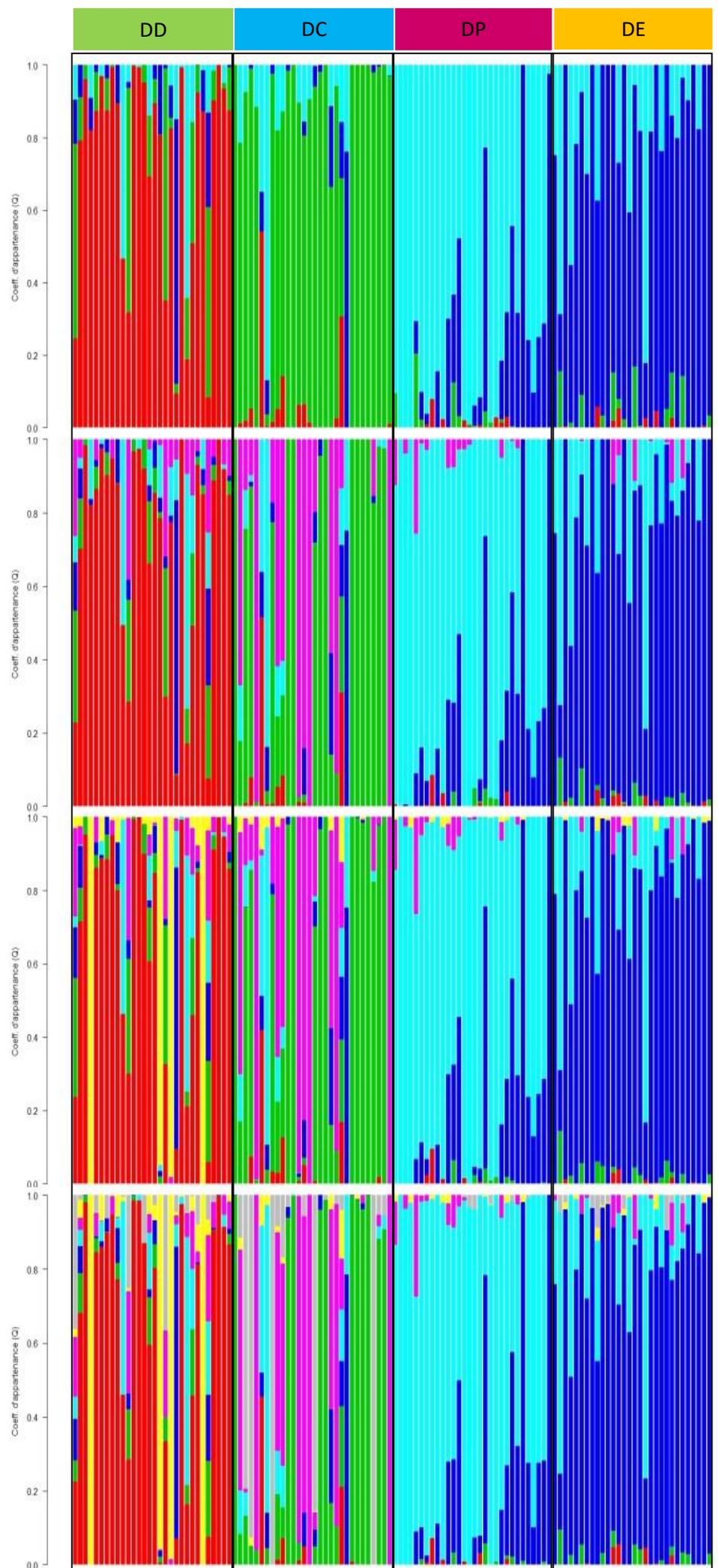


Figure 45: Niveau d'admixture au sein des quatre groupes évolutifs, obtenu à l'aide de l'outil sNMF.

Les quatre graphiques montrent le coefficient d'admixture pour K=4 à K=7. La division des 120 génotypes en quatre groupes évolutifs (DD en vert, DC en bleu, DP en bordeaux et DE e orange) est notée en haut.

Le coefficient d'appartenance des 120 génotypes à chacun des groupes est représenté par des segments de couleurs différentes.

3.1.2 Analyse du niveau d'admixture

Nous avons ensuite cherché à déterminer la constitution génétique de chacun des 120 génotypes, en utilisant la méthode sNMF. La première étape est de calculer le nombre de populations ancestrales qui expliquerait le mieux la diversité génétique de l'échantillon (cross-entropy) (figure 44). Du fait de la particularité de notre échantillon (série de domestication), nous parlerons plutôt, ici, de groupes ancestraux, notés K. J'ai choisi d'analyser le niveau d'admixture (composition génétique) de chaque individu pour plusieurs valeurs de K afin de voir l'impact de l'ajout d'un groupe ancestral sur l'assignation des individus aux différents groupes formés. La figure 45 montre le coefficient d'admixture pour K=4 à K=7. Le coefficient d'appartenance des génotypes à chacun des groupes ancestraux est représenté par des segments de couleurs différentes.

Sur la base de quatre groupes ancestraux (K=4), nous distinguons les quatre formes évolutives : DD en majorité rouge, DC en majorité vert, DP en majorité bleu clair et DE en bleu foncé. Cependant, comme pour les analyses ACP et DAPC, certains génotypes ont des compositions génétiques inattendues.

Dans le groupe DD, les génotypes Tc2220, Tc2398, Tc2454, Tc3401 sont plus proches génétiquement du groupe DC : le segment de couleur verte, représentant le coefficient d'appartenance (Q) au groupe DC est plus important que le segment de couleur rouge (appartenance au groupe DD). Il en est de même pour le génotype Tc3302, dont la proportion du segment bleu foncé est plus importante que celle du rouge, ce qui traduit une composition génétique plus proche des génotypes du groupe DE. Pour finir, les génotypes Tc2393 et Tc 3302 sont plus proches du groupe DP avec une proportion de segments bleu clair supérieurs au rouge.

De la même façon, trois génotypes du groupe DC se retrouvent apparentés génétiquement à d'autres groupes. C'est le cas de la lignée Tc2218 proche génétiquement, des groupes DD et DP, du génotype Tc2462 proche du groupe DP et de la lignée Tc3205 qui se rattache au groupe DE.

Concernant les deux formes de *T. turgidum* ssp *durum* (DP et DE), la distinction est moins évidente. Mais on retrouve les cinq génotypes du groupe DP : Tc2549, Tc2611, Tc2818, Tc3307 et Tc4309 avec une composition génétique plus proche des DE (segments bleus supérieurs à 50%) et les trois génotypes du groupe DE : Tc2254, Tc2258 et Tc2675 majoritairement proches du groupe DP (segments clairs supérieurs à 50%).

Avec cinq groupes ancestraux (K=5), le nouveau groupe génétique formé (en rose) est placé au sein des quatre groupes, mais c'est dans le groupe DC que l'impact de l'ajout de ce groupe est le plus notable. En effet, dix génotypes apparaissent maintenant différents des autres génotypes du groupe DC : Tc2211, Tc2215, Tc2476, Tc2484, Tc 2489, Tc2490, Tc2501, Tc2515, Tc2520 et Td4005 (segments roses supérieurs à 50%).

Lorsque nous regardons le coefficient d'admixture de chacun des individus en considérant six groupes ancestraux (K=6), nous remarquons (en jaune) que les quatre génotypes du groupe DD : Tc2227, Tc2451, Tc2460 et Tc3309 apparaissent maintenant génétiquement différents des autres, comme avec les analyses ACP et DAPC (segments jaunes supérieurs à 50%).

Pour finir, si nous considérons sept groupes ancestraux (K=7), nous voyons apparaître (en gris), comme avec l'analyse DAPC, les cinq génotypes du groupe DC : Tc2212, Tc2213, Tc2465, Tc2503 et Tc3408 comme un groupe génétique différent (segments gris supérieurs à 50%).

Conclusion sur la structure génétique :

L'analyse de la structure génétique de notre jeu de données nous a permis de conclure sur plusieurs aspects :

- ✓ Le groupe DD présente une structure diffuse, visible par la dispersion des points sur les axes principaux de l'ACP et les axes discriminants de l'analyse DAPC, signe d'une importante diversité intra-groupe. Par ailleurs, dans l'analyse sNMF, il s'agit du groupe le plus morcelé avec des coefficients d'admixture importants. Enfin, les trois analyses ont montré que quatre génotypes (Tc2220, Tc2398, Tc2454, Tc3401) étaient génétiquement différents des autres génotypes de ce groupe, ce qui crée une structure forte.
- ✓ Les génotypes du groupe DC présentent une structure marquée en au moins deux groupes bien distincts avec une forte diversité intra-groupe ce qui reflète une diversité génétique encore conséquente. Les génotypes de ce groupe ont des coefficients d'admixture moins importants que ceux du groupe DD, mais il existe au moins deux groupes distincts. En effet, Les trois méthodes ont mis en évidence un groupe de cinq génotypes (Tc2212, Tc2213, Tc2465, Tc2503 et Tc3408) qui semblent différents des autres génotypes de ce groupe.
- ✓ Les deux formes de *T. turgidum ssp durum* (DP et DE) ont une structure nette avec un nuage de points commun, séparé de ceux des autres groupes et un niveau de diversité intra-groupe est faible. Le taux d'admixture de ces deux groupes est faible et interconnecté.

Tableau 6: Comparaison du niveau de diversité entre les deux génomes: A et B. Ces tableaux présentent les valeurs moyennes du π de Tajima (A) et du θ_s de Watterson (B), calculées sur les 7681 contigs situés sur le génome A et les 8420 contigs situés sur le génome B. Le test de Willcoxon, dont les p-value ont été calculées à l'aide du test de Willcoxon

A	π de Tajima			B	θ_s de Watterson		
	génoméA	génoméB	p-value		génoméA	génoméB	p-value
DD	0,00233	0,00326	< 2,2e-16	DD	0,00295	0,00368	< 2,2e-16
DC	0,00166	0,00220	< 2,2e-16	DC	0,00177	0,00215	1,92E-13
DP	0,00124	0,00159	1,98E-07	DP	0,00135	0,00162	2,44E-07
DE	0,00114	0,00141	1,92E-09	DE	0,00125	0,00146	8,36E-09
Moyenne	0,00159	0,00212		Moyenne	0,00183	0,00223	

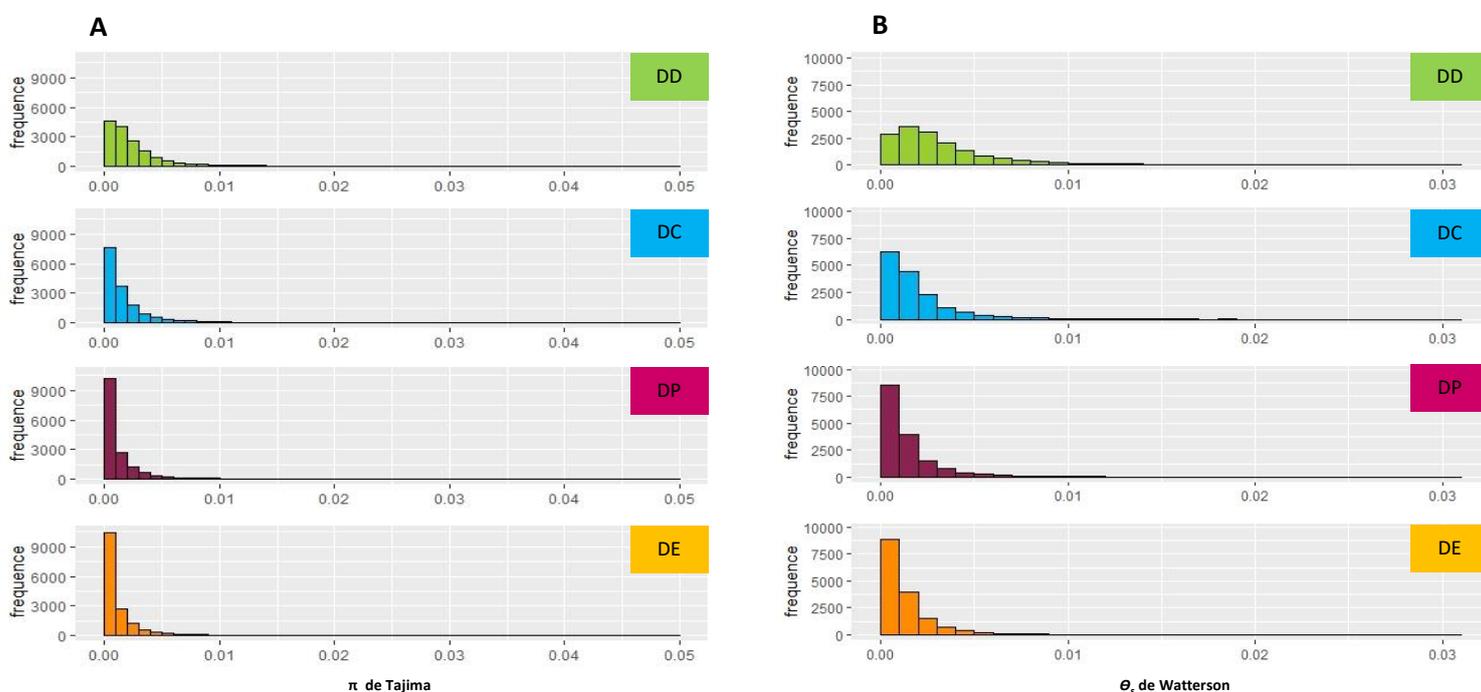


Figure 46: Distribution du π de Tajima (A) et du θ_s de Watterson (B), pour les quatre groupes évolutifs: DD en vert, DC en bleu, DP en bordeaux et DE en orange.

Tableau 7: Comparaison du niveau de diversité entre les quatre groupes évolutifs. Ce tableau présente les valeurs moyennes pour les trois paramètres de diversité: π de Tajima, θ_s de Watterson et D de Tajima., estimés pour les quatre formes évolutives: DD, DC, DP et DE

	π	θ_s	D
DD	0,00281	0,00333	-0,47167
DC	0,00194	0,00197	-0,14582
DP	0,00143	0,00149	-0,25351
DE	0,00128	0,00136	-0,25350

3.2 Caractérisation des effets démographiques associés aux transitions évolutives

3.2.1 Diversité au sein des génomes homéologues

L'amidonnier sauvage (*T. turgidum* ssp *dicoccoides*) est une espèce tétraploïde résultant d'une hybridation spontanée entre deux espèces diploïdes. Nous avons caractérisé le niveau de diversité de chacun des deux génomes homéologues, en utilisant les deux estimateurs, π et θ_s .

L'information génétique disponible est un peu plus importante sur le génome B que pour le génome A : parmi les 19738 contigs de référence ZAVITAN_BAITS, 9324 sont situés sur le génome A et 10414 sont sur le génome B. Cette différence est maintenue au sein des 16101 contigs sélectionnés : 7681 sont situés sur le génome A et 8420 sur le génome B.

Sur les 16101 contigs, le génome B a des valeurs de π et θ_s significativement plus élevées que celles du génome A, quel que soit le groupe évolutif considéré (tableau 6). Cette différence n'est pas due à la différence de quantité d'information génétique disponible, car elle est toujours significative si nous ré-échantillons 7681 contigs, parmi les 8420, sur le génome B.

3.2.2. Diversité au sein des quatre groupes évolutifs de *T. turgidum*

La diversité génétique au sein des quatre groupes évolutifs (DD, DC, DP et DE) a été caractérisée afin de quantifier les effets démographiques au cours de la domestication du blé dur.

La distribution des valeurs des estimateurs π et θ_s , calculées sur les 16101 contigs pour les quatre groupes génétiques, indique une baisse de diversité génétique lors de chaque phase de transition évolutives (Figure 46). Cela se manifeste par l'augmentation de la proportion de contigs ayant des valeurs de π et de θ_s proche de 0.

Afin de quantifier cette perte de diversité, nous pouvons comparer les valeurs moyennes de π et θ_s pour les quatre groupes (tableau 7). Cela nous permet de voir que la diminution de la diversité n'est pas homogène à chaque transition. La plus grosse perte de diversité s'effectue lors du passage de la sous-espèce DD ($\pi=0,00281$, $\theta_s=0,00333$) à la sous-espèce DC ($\pi=0,00194$, $\theta_s=0,00197$). Sur la base du rapport de diversité calculé à l'aide de l'estimateur θ_s (Wright et al. 2005), seulement 59 % de la diversité présente chez le groupe DD se retrouve chez le groupe DC. La perte de diversité est moins importante entre les groupes DC et DP ($\pi=0,00143$, $\theta_s=0,00149$) : le rapport de θ_s nous indique que 76% de la diversité chez DC se retrouve dans le groupe DP. Pour finir, la perte de diversité est faible entre les deux groupes DP et DE ($\pi=0,00128$, $\theta_s=0,00136$) avec 91% de la diversité conservée entre les deux formes de *T. turgidum* ssp *durum*.

Pour chacun des contigs, la relation entre la diversité estimée par θ_s , entre deux groupes, est représentée en figure 47. Le coefficient de la pente correspondant à la régression linéaire, pour les trois transitions évolutives : DD/DC, DC/DP et DP/DE sont respectivement à 0.47, 0.58 et 0.58.

Pour les trois transitions, le coefficient de la pente est inférieur à celui de la bissectrice, ce qui valide la perte de diversité lors des trois transitions.

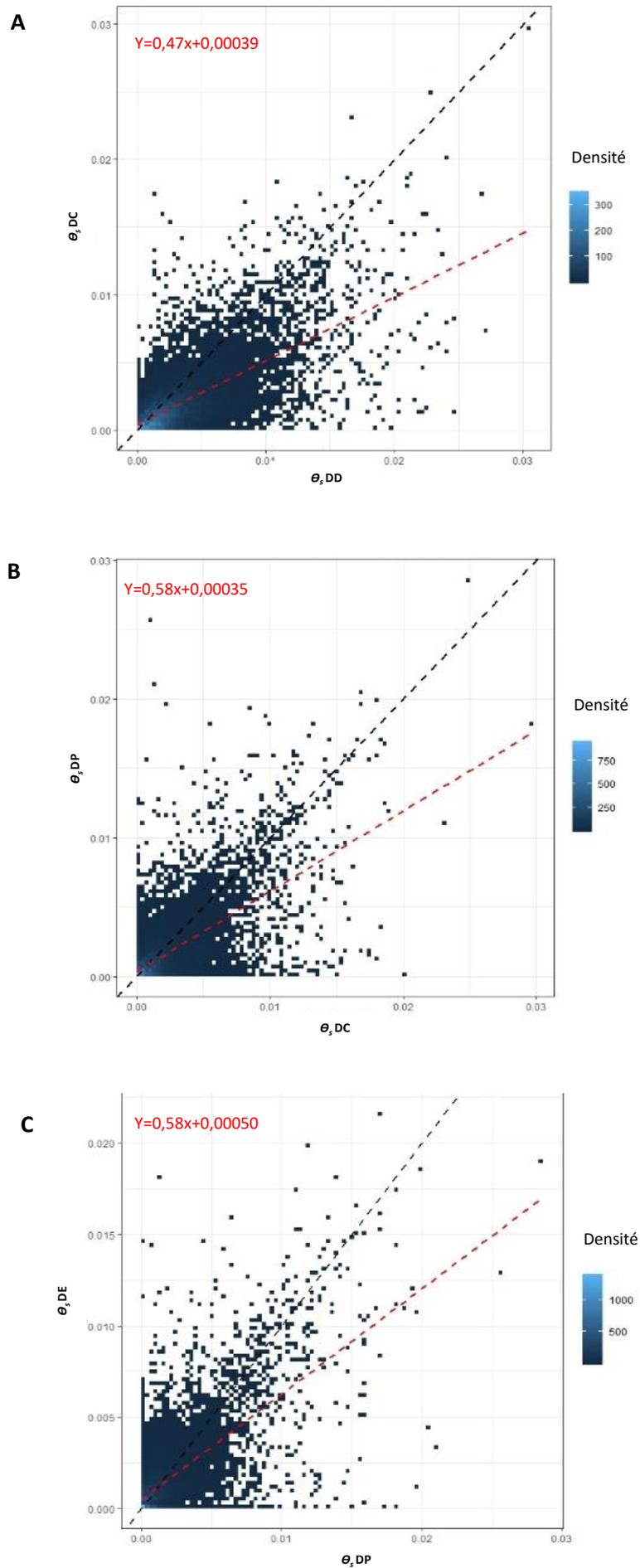
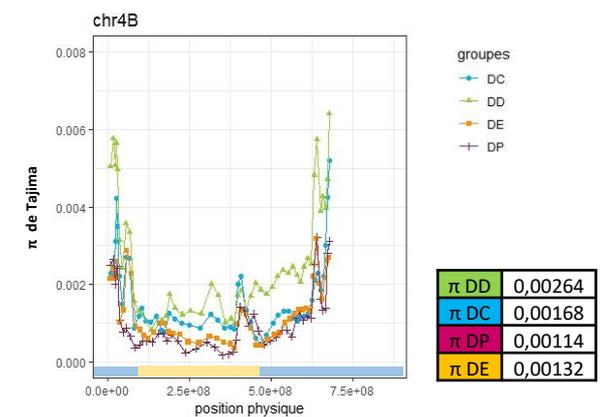
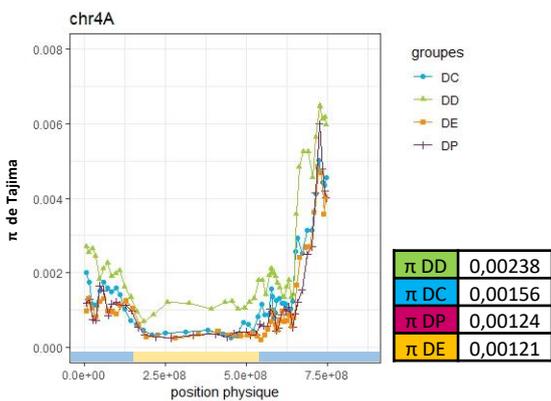
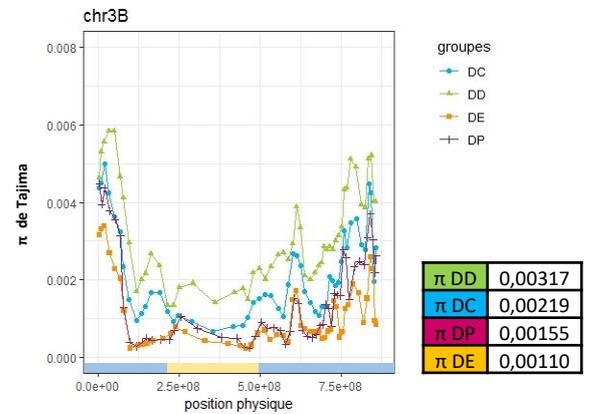
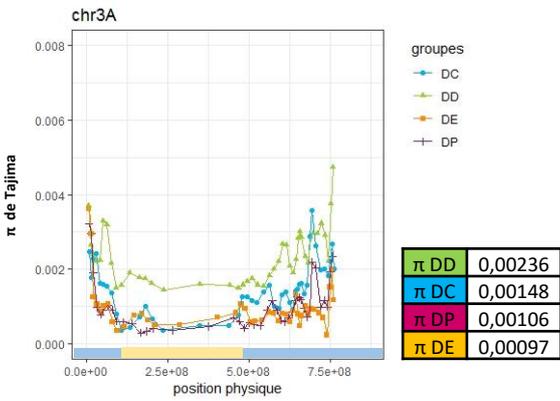
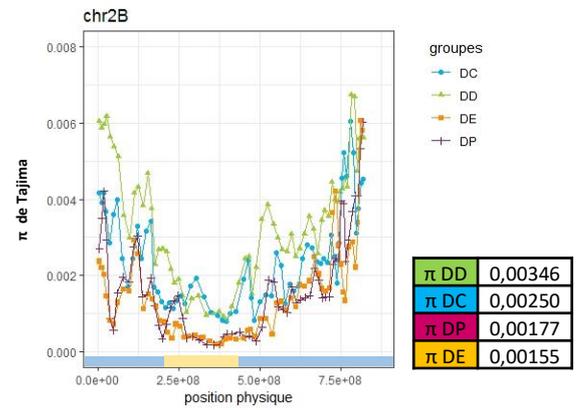
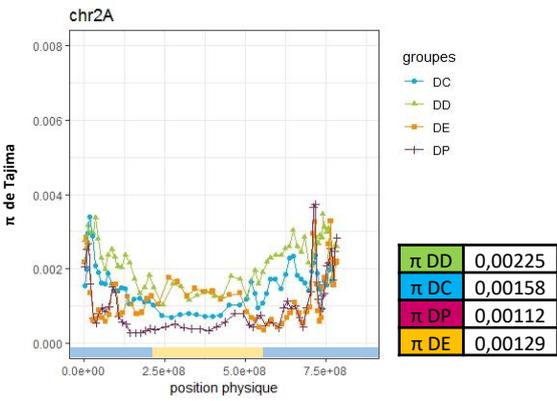
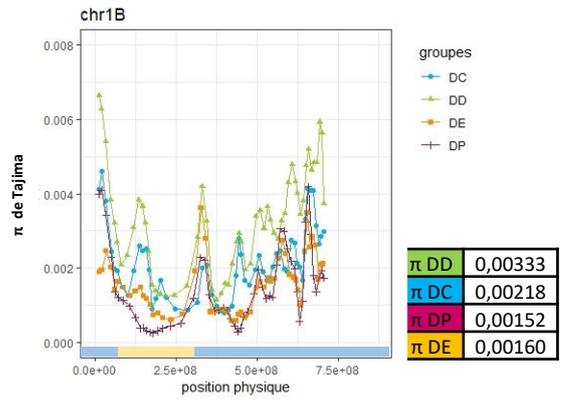
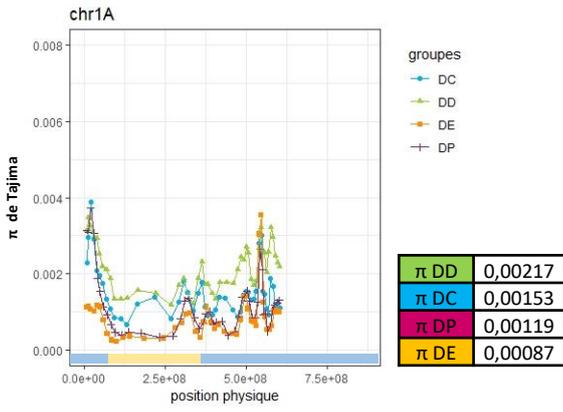


Figure 47: Valeurs de l'estimateur θ_s de Watterson pour chaque contig, entre les quatre groupes évolutifs considérés deux à deux: groupes DD et DC (A), groupes DC et DP (B), et groupes DP et DE (C). La bissectrice est en pointillée noire et la pente correspondant à la régression linéaire par couple de groupes est présenté en pointillée rouge. L'équation de la droite de régression est noté en rouge.



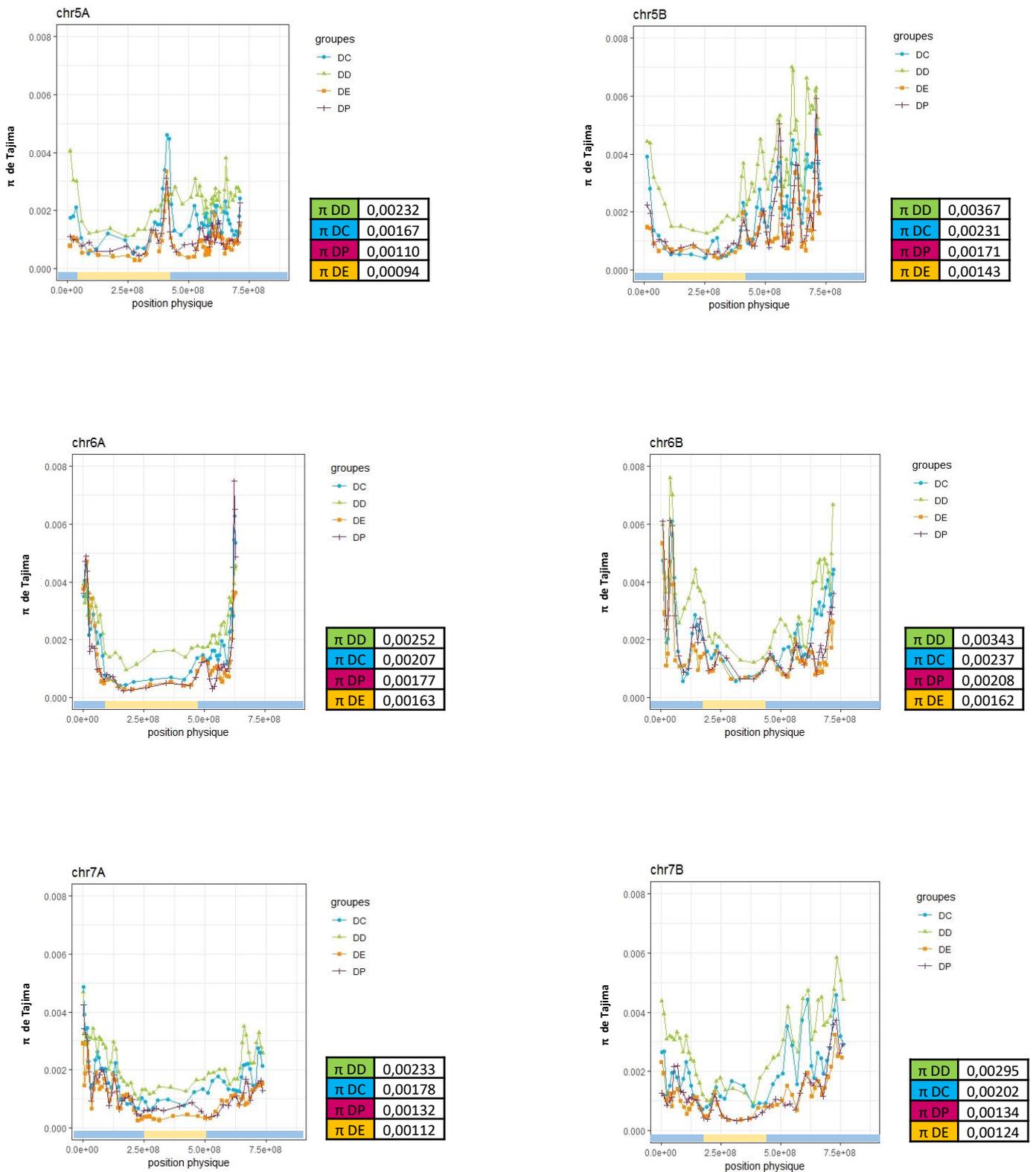
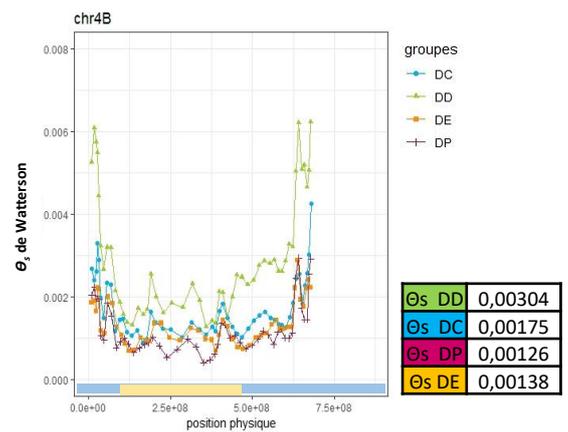
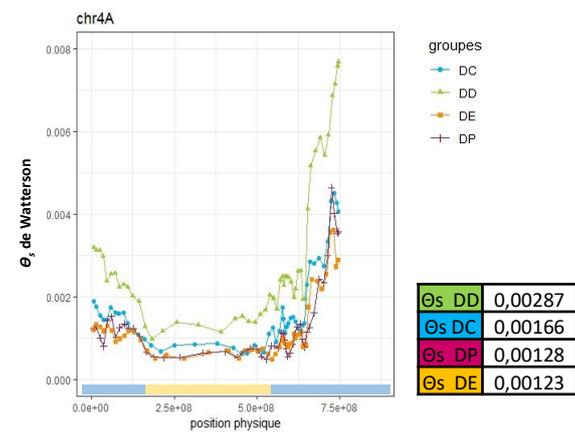
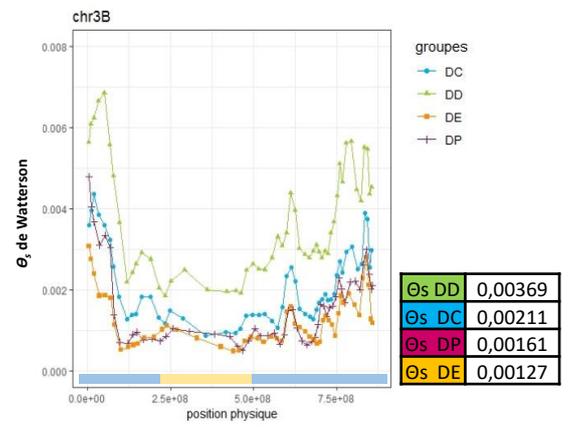
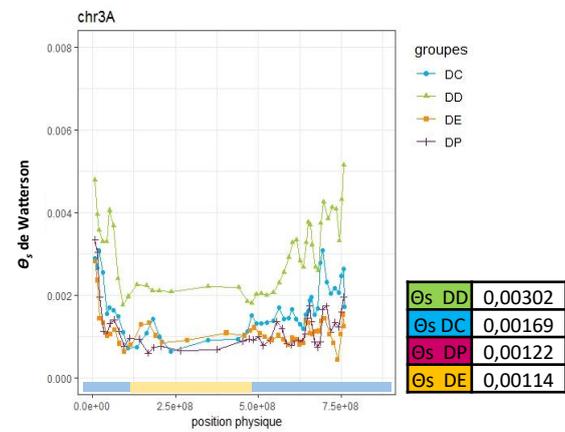
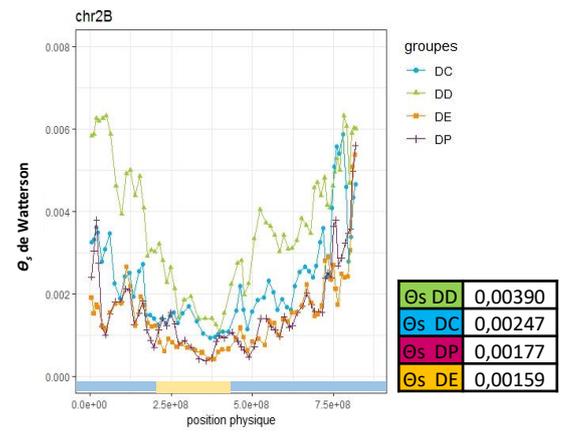
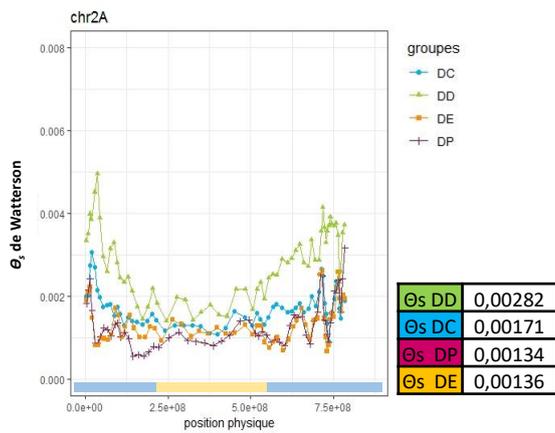
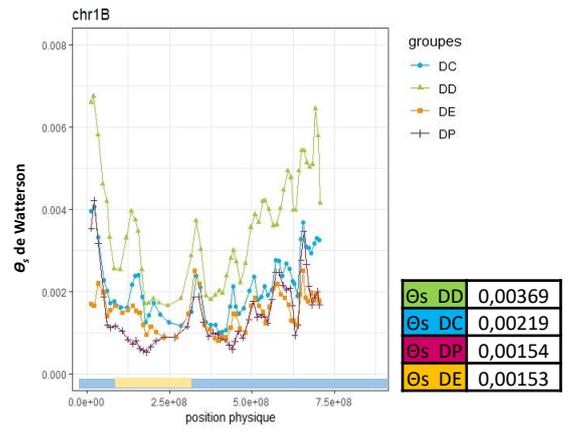
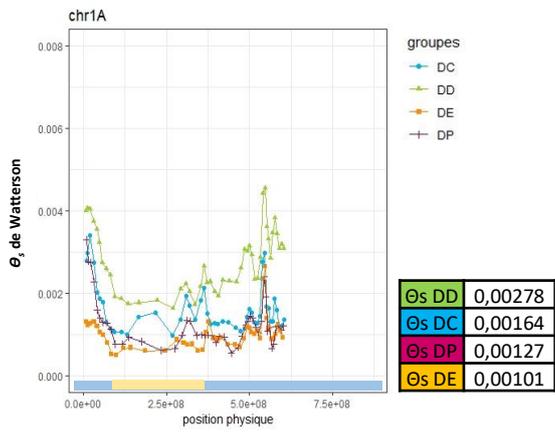


Figure 48: Evolution du niveau de diversité le long des chromosomes avec l'estimateur π de Tajima. calculé sur des fenêtres glissantes de 20 Kb de séquences analysées avec un chevauchement des fenêtres de 10Kb. La diversité de DD est présentée en vert, celle de DC en bleu, celle de DP en bordeaux et celle de DE en orange. Les valeurs moyennes de π ont été calculées par chromosome pour chacun des groupes et présentées dans les tableaux à droite des figures. Par convention, les coordonnées physiques commencent par le bras court du chromosome. Les zones centromériques sont matérialisées par des segments jaunes (Maccaferri et al., 2019)



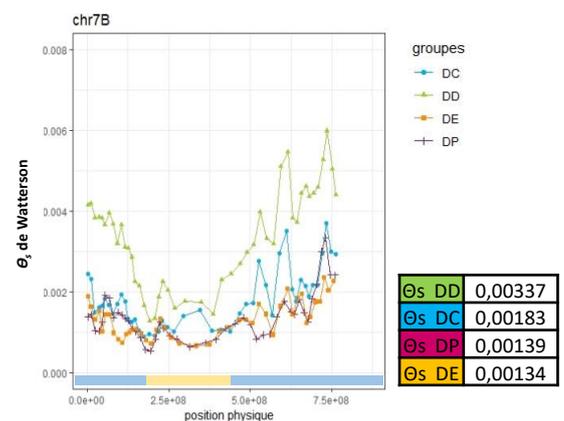
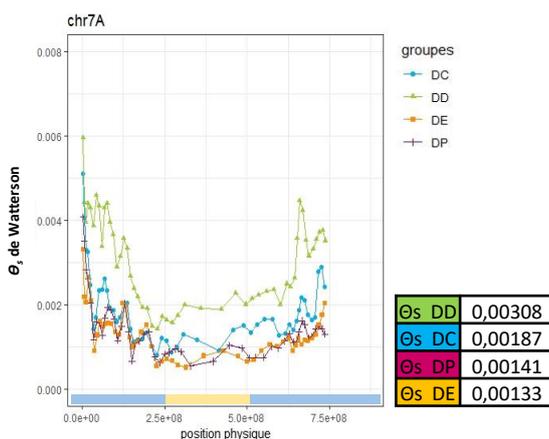
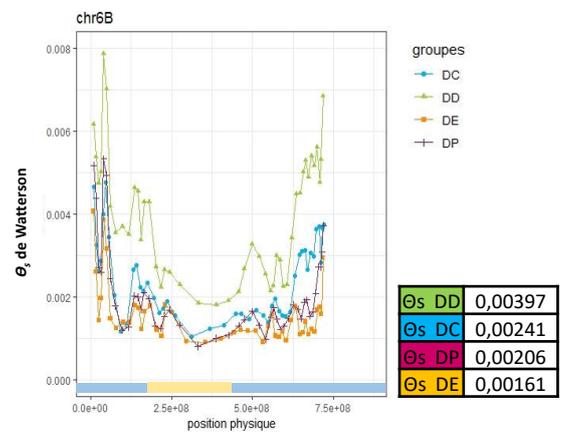
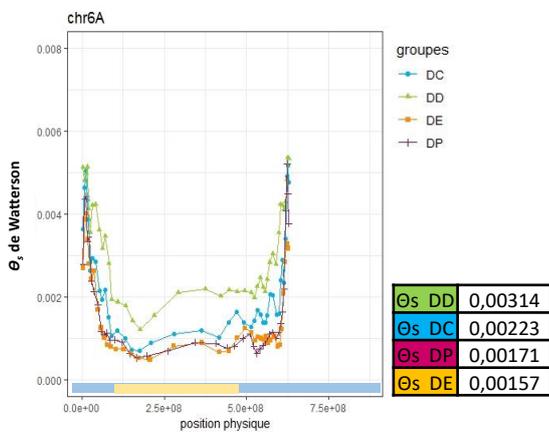
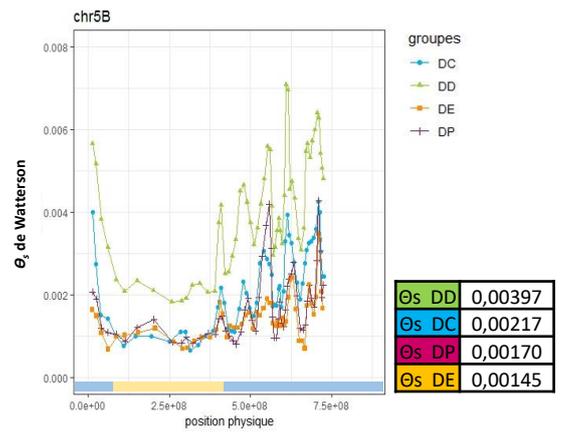
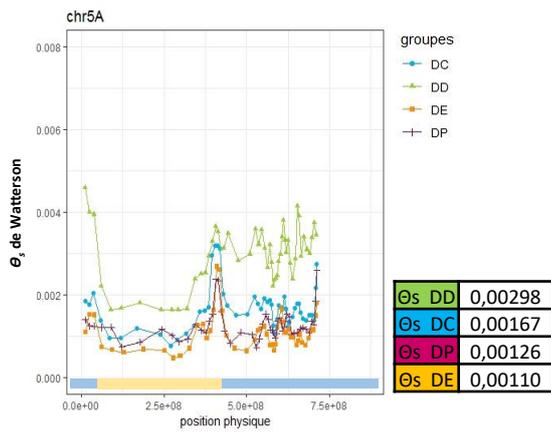


Figure 49: Evolution du niveau de diversité le long des chromosomes avec l'estimateur θ_s de Watterson. calculé sur des fenêtres glissantes de 20 Kb de séquences analysées avec un chevauchement des fenêtres de 10Kb. La diversité de DD est présentée en vert, celle de DC en bleu, celle de DP en bordeaux et celle de DE en orange. Les valeurs moyennes de θ_s on été calculées par chromosome pour chacun des groupes et présentées dans les tableaux à droite des figures. Par convention, les coordonnées physiques commencent par le bras court du chromosome. Les zones centromériques sont matérialisées par des segments jaunes (Maccaferi et al., 2019).

La valeur du D de Tajima augmente en moyenne entre le groupe DD ($D = -0,47167$) et le groupe DC ($D = -0,14582$). Dans un contexte de domestication, une réduction démographique (perte des allèles rares et/ou en faible fréquence), est caractéristique des premières phases d'un goulot d'étranglement sans retour à l'équilibre (figure 7) (Tajima, 1989). Les données obtenues sont en accord avec ce scénario (un goulot d'étranglement), même si le signal est faible. A contrario, la valeur du D de Tajima diminue entre le groupe DC ($D = -0,14582$) et les groupes DP ($D = -0,25351$) et DE ($D = -0,25350$), ce qui est davantage en accord avec un scénario d'expansion démographique (Tableau 7 au dos).

3.2.3 Diversité génétique le long des chromosomes

Pour représenter l'évolution de la diversité le long des chromosomes, nous avons calculé les valeurs moyennes de π de Tajima sur des fenêtres de 20 kb pour chacun des quatre groupes génétiques (figure 48). Les profils de diversité diffèrent selon les chromosomes. Il est tout de même possible de les regrouper en trois catégories :

- ✓ La première catégorie correspond aux chromosomes ayant des profils en « anse de seau ». Dans cette catégorie figurent 10 des 14 chromosomes : 1A, 2A, 2B, 3A, 3B, 4A, 4B, 6A, 6B et 7A. Pour ces chromosomes la diversité génétique est plus importante au niveau des télomères que des centromères. Parmi ces 10 chromosomes, certains ont une courbure plus marquée, avec un niveau de diversité très important au niveau des télomères ; c'est le cas des chromosomes 4A, 4B, 6A et 6B.
- ✓ La deuxième catégorie est constituée par les chromosomes 1B et 5A pour lesquels nous remarquons un niveau de diversité important, pour les quatre formes évolutives, dans la zone centrale du chromosome.
- ✓ Le troisième type regroupe les chromosomes 5B et 7B, pour lesquels nous observons un niveau de diversité très hétérogène le long d'un des deux bras de chromosomes, en l'occurrence ici, le bras long.

En moyenne sur chacun des chromosomes, le niveau de diversité génétique le plus élevé est obtenu avec le groupe DD. Le groupe DC présente un niveau de diversité intermédiaire, et les deux groupes de *T. turgidum* ssp *durum* (DP et DE) ont les niveaux de diversité les plus bas.

Cependant, si on s'intéresse à quelques zones en particulier, cet ordre n'est pas toujours respecté. C'est le cas, par exemple, de la zone située autour de 300 Mb sur le chromosome 1B, où le groupe DE présente presque autant de diversité que le groupe DD alors que les niveaux de polymorphisme des groupes DC et DP sont plus faibles. Nous retrouvons cette situation entre 250Mb et 500Mb sur le chromosome 2A. Nous remarquons également, la zone autour de 400 Mb du chromosome 5A, où le groupe DC présente une diversité bien plus importante que les groupes DP et DE, eux même plus polymorphes que le groupe DD.

L'analyse de la diversité génétique le long des chromosomes a été réalisée également, avec les valeurs de θ_s (figure 49). Les profils sont très comparables à ceux obtenus avec les valeurs de π . De la même façon qu'avec l'estimateur π , le niveau de diversité génétique n'est pas homogène le long des chromosomes et nous observons les trois types de profils présentés précédemment. Cependant, deux points sont à souligner. Premièrement, le niveau de diversité du groupe DD est plus élevé et se différencie plus nettement de ceux des autres groupes. Deuxièmement, certaines zones, comme la

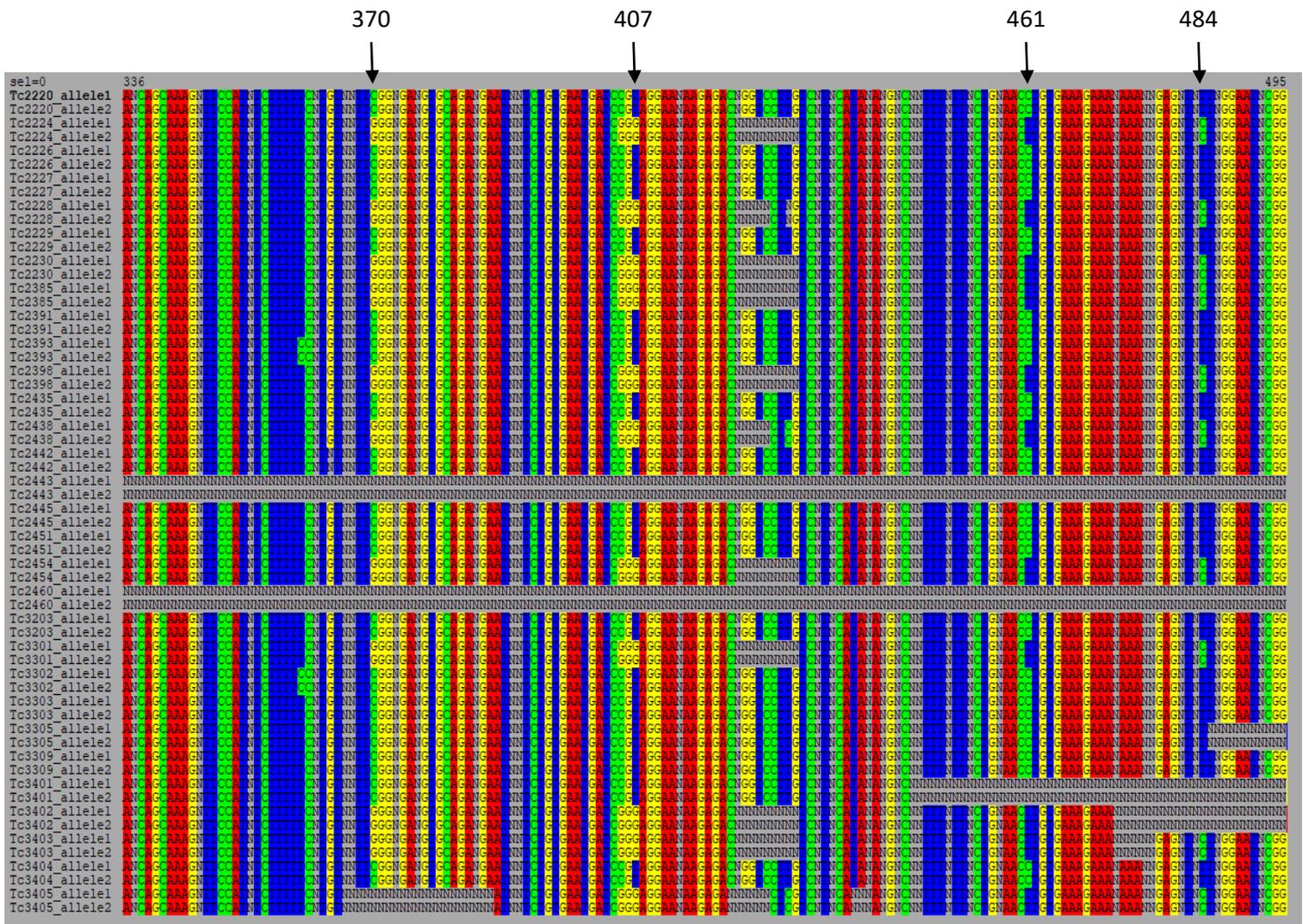


Figure 50: Fragment (336 à 495pb) du contig chr1B:1-356313144:340003850-340004569 situé dans la zone proche du centromère du chromosome 1B.

Les génotypes (2 allèles) sont présentés en lignes et les locus en colonnes. Les nucléotides sont marqués avec 4 couleurs différentes (A en rouge, T en bleu, G en jaune et C en vert) afin de visualiser les SNPs.

Les locus où tous les génotypes sont en données manquantes « N » correspondent au locus supprimés par les filtres de FIS, p-value et « au moins deux homozygotes différents ».

zone autour de 4Mb du chromosome 5A, qui ne présentait pas des pertes de diversité attendues à chaque transition avec l'estimateur π , ont, avec l'estimateur θ_s , un profil en adéquation avec l'histoire démographique. Ces différences s'expliquent par le mode de calcul différent de ces deux paramètres. En effet, le calcul de π se base sur les fréquences alléliques alors que le calcul de θ_s prend en considération le nombre d'allèles ou de sites polymorphes. Ce dernier est donc plus sensible aux allèles rares. Chez la forme sauvage (DD), on s'attend à ce que le nombre d'allèles rares soit plus important et diminue dès les premières phases de domestication lors desquelles une réduction forte de la taille démographique est attendue. Cela explique donc pourquoi la différence de niveau de polymorphisme entre le groupe DD et les trois autres groupes est exacerbée avec l'estimateur θ_s . Le rapport entre ces deux estimateurs, le long des chromosomes peut être calculé grâce au D de Tajima (Annexe 7).

Concernant les profils des chromosomes 1B et 5A pour lesquels nous remarquons un niveau de diversité important dans des zones proches du centromère, ou les profils des chromosomes 5B et 7B, pour lesquels nous observons un niveau de diversité très hétérogène sur le bras long de ces chromosomes, nous pouvons émettre deux hypothèses : un problème au niveau de l'estimation de niveau de diversité ou une mauvaise localisation des contigs de la référence ZAVITAN_BAITS concernés sur la référence génomique ZAVITAN. Pour mieux comprendre ces observations, j'ai étudié le polymorphisme de la région centrale des chromosomes 1B et 5A et, plus particulièrement, les deux fenêtres de 20 Kb de séquences analysées (Iseff calculée par Egglib) où la valeur moyenne de π était la plus importante.

Afin de vérifier la première hypothèse, j'ai vérifié la qualité des séquences consensus analysées et le calcul des paramètres de diversité, par le logiciel Egglib.

La zone proche du centromère du chromosome **1B**, la fenêtre de 20Kb de séquences analysées (Iseff) est représentée par 44 contigs. Pour le groupe DD, la valeur moyenne de π est de 0.0042, impactée fortement par deux contigs qui ont une valeur de $\pi > 0.01$. L'observation des séquences des 30 génotypes du groupe DD pour le contig ayant la valeur de π la plus élevée (chr1B:1-356313144:340003850-340004569), montre que le contig est plutôt bien couvert à l'exception de 2 génotypes (Tc 2443 et Tc2460). Par ailleurs, certains sites sont notés « N » pour l'ensemble des génotypes, car ils n'ont pas passé les filtres de couverture, pvalue, paraclean, FIS ou nombre d'homozygotes (figure 50). Dans cette portion du contig, nous pouvons voir deux haplotypes différents avec des polymorphismes de type SNP. Le premier haplotype porte un C en position 370 et 405, un T en position 407, un C en position 461 et un T en position 484. Le deuxième haplotype porte un G en position 370, 405 et 407, un T en position 461 et un C en position 484. Nous remarquons également qu'ils portent une délétion d'une dizaine de paires de base après la position 420.

En observant les séquences des trois autres groupes, nous pouvons voir que le nombre de polymorphisme diminue chez DC et DP.

Pour la zone proche du centromère du chromosome **5A**, la fenêtre de 20Kb de séquences analysées (Iseff) est représentée par 45 contigs. La valeur moyenne de π , pour le groupe DC est de 0.0046, impactée fortement par six contigs ayant une valeur de $\pi > 0.01$. L'observation des séquences de la totalité des 120 génotypes nous permet, comme pour le chromosome 1B, de valider le polymorphisme. Cette première vérification nous permet d'écarter un problème de qualité de séquences ou de calcul des paramètres de diversité par le logiciel Egglib v3. Par ailleurs, cela permet d'invalider la possibilité qu'il s'agisse de locus paralogues. En effet, si nous étions en présence de paralogues nous aurions des

Tableau 8: Fst par paire entre les quatre groupes évolutifs: DD, DC, DP et DE calculés avec le package R Hierfstat (Goudet 2005). Les intervalles de confiance (entre crochets) ont été estimés par 1000 bootstraps. Toutes les valeurs de Fst sont significatives.

	DC	DP	DE
DD	0,236 [0,232 ; 0,239]	0,353 [0,349 ; 0,357]	0,385 [0,381 ; 0,389]
DC	X	0,347 [0,343 ; 0,351]	0,369 [0,365 ; 0,373]
DP	X	X	0,116 [0,113 ; 0,118]
DE	X	X	X

sites très hétérozygotes qui auraient été supprimés par les filtres et l'ensemble des autres sites seraient homozygotes. Or nous observons des polymorphismes qui respectent les filtres.

Pour tester la deuxième hypothèse, j'ai vérifié que les contigs soient correctement localisés sur la référence génomique « ZAVITAN ». Pour mémoire, les contigs de la référence ZAVITAN_BAITS ont été construits en réalisant un blast des baits de 120 pb sur la référence génomique complète WEWSeq v.2.0 (Zhu et al. 2019). Suite à ce blast, j'ai conservé les localisations pour lesquelles il y avait 90% l'homologie et 90 pb de recouvrement entre les baits et la référence. Ces séquences ont été allongées de 300pb de part et d'autre pour créer les contigs ZAVITAN_BAITS. Nous pouvons donc nous interroger sur une possible correspondance d'un contig ZAVITAN_BAITS à plusieurs endroits du génome, générant de plus fortes valeurs de polymorphismes. J'ai donc testé l'impact du nombre de blast sur le niveau de diversité. Pour évaluer l'occurrence de cette situation, j'ai réalisé un blast des contigs des deux zones à analyser, sur la référence génomique complète WEWSeq v.2.0, sans appliquer de filtres afin d'observer l'ensemble des possibilités d'assignations.

Pour la zone proche du centromère du chromosome **1B**, les 44 contigs ont généré 8202 blasts sur la référence complète. Parmi eux, 32 contigs avaient moins de cinq correspondances et deux contigs plus de 2000 correspondances, sans appliquer de filtres. Les deux contigs pour lesquels les valeurs de π étaient > 0.01 possèdent, pour le premier, une seule correspondance forte sur le 1B, et pour le deuxième, une correspondance forte sur le 1B et 3 correspondances bien plus faibles sur le 1B et le 3B.

Pour la zone proche du centromère du chromosome **5A**, les 32 contigs avaient moins de 5 correspondances et 2 contigs ont plus de 2000 correspondances, sans appliquer de filtres. Les 6 contigs ayant les valeurs de $\pi > 0.01$ avaient entre 2 et 12 correspondances.

Cette deuxième étape permet de confirmer que ni la paralogie ni l'homéologie ne sont à l'origine des fortes valeurs de π .

3.2.4 Différenciation entre les quatre groupes évolutifs de *T. turgidum*

Afin de comparer les niveaux de différenciation génétique entre les six paires de groupes génétiques (« pairwise »), nous avons calculé le Fst de façon globale sur l'ensemble des 35 164 SNPs répartis sur les 10734 contigs (filtre MAF>5%) (tableau 8).

La valeur du Fst entre DD et DC est de 0.236 et passe respectivement à 0.353 et 0.385 entre DD et les deux groupes de *T. turgidum ssp durum*, DP et DE. Le niveau de différenciation évolue donc de la même manière que le niveau de diversité et en cohérence avec le fait que la réduction progressive de la diversité au cours de la domestication s'accompagne d'une augmentation du Fst par rapport au groupe ancestral.

La valeur du Fst entre DC et DP (0.347) est importante, traduisant une différenciation importante entre ces deux groupes alors que celle obtenue entre DP et DE est plus faible (Fst=0.116). Cette faible différenciation peut s'expliquer par l'histoire plus récente de cette dernière transition (5000 et 50 ans respectivement) et/ou par la différence d'intensité des goulots d'étranglements correspondant à ces deux transitions.

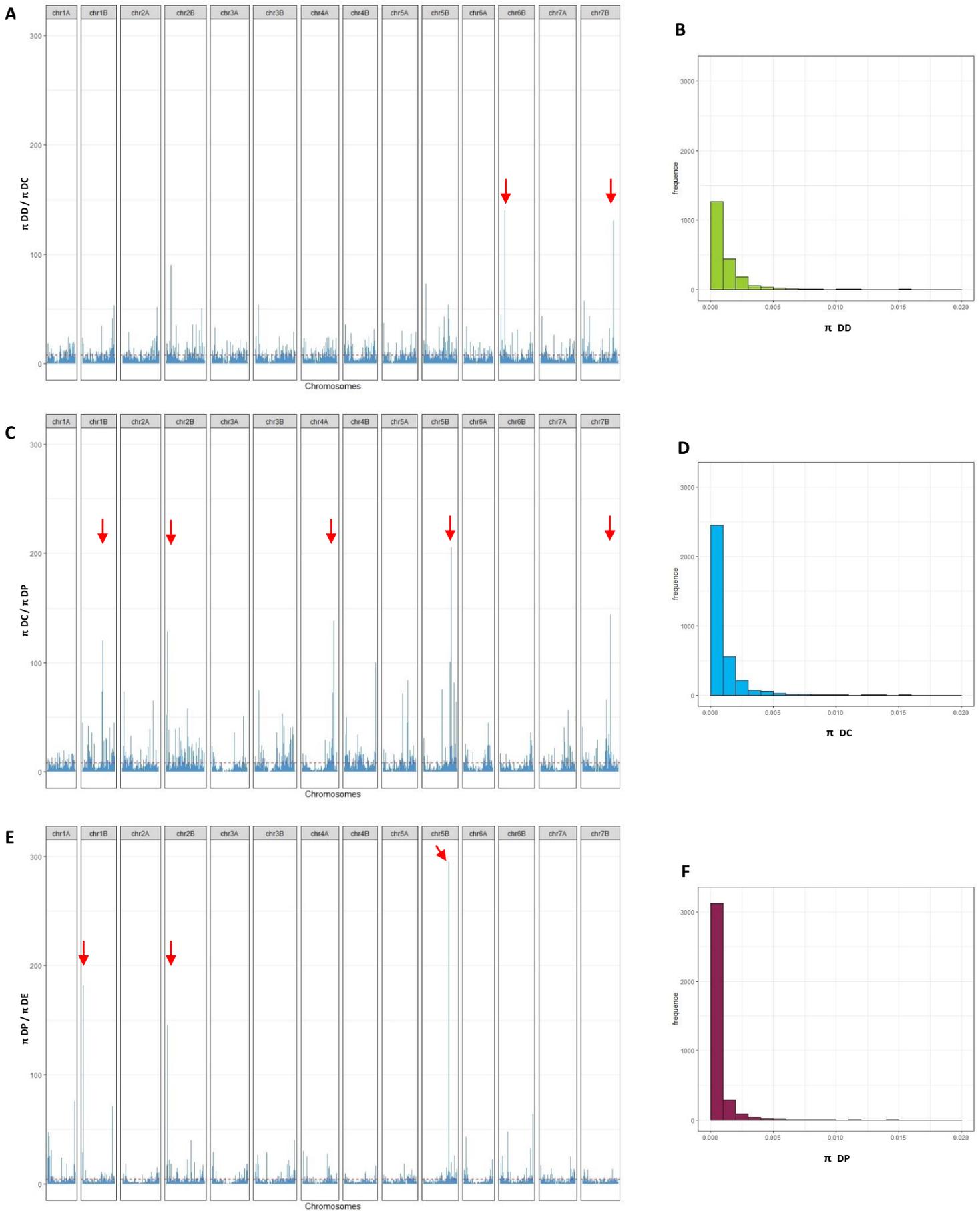


Figure 51: Détection de sélection, sans *a priori*, à l'aide des rapports de π de Tajima pour chacune des trois transitions évolutives de la domestication et au niveau des 16101 contigs situés le long de chaque chromosome.

A: perte de diversité lors de la transition entre DD et DC / B: histogramme des valeurs de π_{DD} pour les 2043 contigs où $\pi_{DC}=0$.
 C: perte de diversité lors de la transition entre DC et DP / D: histogramme des valeurs de π_{DC} pour les 3421 contigs où $\pi_{DP}=0$.
 E: perte de diversité lors de la transition entre DP et DE / F: histogramme des valeurs de π_{DP} pour les 3599 contigs où $\pi_{DE}=0$.
 Pour chaque transition, la ligne en pointillés rouge est le seuil correspondant au 95^{ème} centile.

Conclusion sur la caractérisation des effets démographiques

L'analyse de la diversité génétique de notre échantillon de 120 génotypes répartis en quatre groupes correspondant à des transitions de l'histoire évolutive de *T. turgidum* nous a permis de conclure sur plusieurs aspects :

- ✓ La diversité génétique du génome B est plus importante que celle du génome A
- ✓ Les estimateurs π , θ , et D ont permis de mettre en évidence que la chute de la diversité génétique au cours de la domestication de l'espèce *Triticum turgidum*, n'a pas été homogène. Le goulot d'étranglement le plus marqué s'est produit entre DD et DC, suivi d'un deuxième goulot, plus faible, entre DC et DP/DE.
- ✓ Les estimations de π et θ , le long des 14 chromosomes nous a permis de voir, d'une part, que le niveau de polymorphisme n'est pas homogène le long du génome, et d'autre part, que l'histoire démographique n'a pas impactée uniformément le génome.
- ✓ La différenciation génétique (F_{st}) entre la forme ancestrale sauvage (DD) et chacune des formes issues du processus de domestication (DC, DP puis DE) augmente.

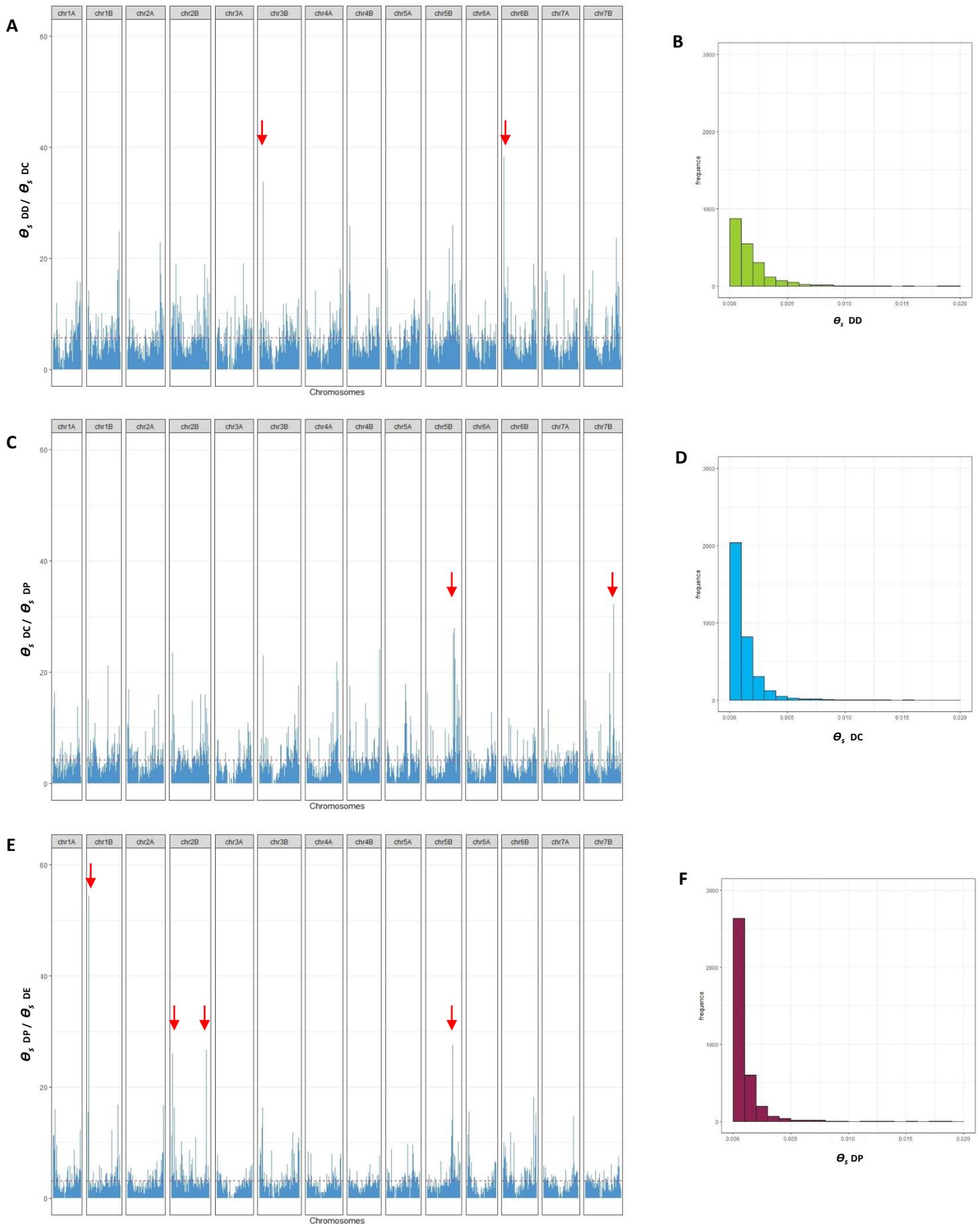


Figure 52: Détection de sélection, sans a priori, à l'aide des rapports de θ_s de Watterson pour chacune des trois transitions évolutives de la domestication et au niveau des 16101 contigs situés le long de chaque chromosome.
 A: perte de diversité lors de la transition entre DD et DC / B: histogramme des valeurs de $\theta_s \text{ DD}$ pour les 2043 contigs où $\theta_s \text{ DC} = 0$.
 C: perte de diversité lors de la transition entre DC et DP / D: histogramme des valeurs de $\theta_s \text{ DC}$ pour les 3421 contigs où $\theta_s \text{ DP} = 0$.
 E: perte de diversité lors de la transition entre DP et DE / F: histogramme des valeurs de $\theta_s \text{ DP}$ pour les 3599 contigs où $\theta_s \text{ DE} = 0$.
 Pour chaque transition, la ligne en pointillés rouge est le seuil correspondant au 95^{ème} centile

3.3 Détection de signatures génétiques de sélection liées à la domestication

Après avoir caractérisé les effets démographiques sur les quatre groupes génétiques, nous allons maintenant essayer de détecter des signatures génétiques de sélection liée à la domestication.

3.3.1 Détection sans *a priori* sur l'ensemble du génome

Nous avons tout d'abord cherché à détecter les locus sous sélection au cours des différentes étapes de domestication, sans *a priori* sur l'ensemble du génome. Pour cela, nous nous sommes appuyés sur différents paramètres.

3.3.1.1 Détection avec les rapports de π de Tajima

La figure 51 présente les rapports avec l'estimateur π : $\frac{\pi_{DD}}{\pi_{DC}}$, $\frac{\pi_{DC}}{\pi_{DP}}$ et $\frac{\pi_{DP}}{\pi_{DE}}$ ainsi que les trois histogrammes de fréquences des valeurs de π_{DD} , π_{DC} et π_{DP} lorsque π_{DC} , π_{DP} et π_{DE} sont respectivement égaux à 0.

Le rapport de : $\frac{\pi_{DD}}{\pi_{DC}}$ met en évidence deux pics forts (>100), correspondant à deux contigs: un sur le chromosome 6BS¹ (102Mb) et l'autre sur le 7BL¹ (677Mb)(figure 51 A). Par ailleurs, parmi des 2043 contigs pour lesquels les valeurs de π_{DC} sont nulles (le rapport ne peut donc pas être calculé), 12 contigs ont des valeurs de π_{DD} supérieure à 0.01 (figure 51 B). Ces résultats confirment le fait que le passage de la forme sauvage à la première forme domestiquée, s'était traduit par une forte baisse de la diversité génétique dans ces zones chromosomiques.

Avec le calcul du rapport $\frac{\pi_{DC}}{\pi_{DP}}$, cinq pics forts (>100) sont identifiés ; ils sont situés sur 5 chromosomes différents : 1BL (445Mb), 2BS (25Mb), 4AL (687Mb), 5BL (612Mb) et 7BL(610Mb) (figure 51 C). Parmi les 3421 contigs pour lesquels les valeurs de π_{DP} sont nulles, 15 contigs ont une valeur π_{DC} supérieure à 0.01 (figure 51 D), traduisant une forte baisse de diversité dans ces régions lors de la transition entre DC et DP.

Avec le calcul du rapport $\frac{\pi_{DP}}{\pi_{DE}}$ correspondant à la transition « révolution verte », trois pics forts (>100) sont détectés : un sur le 1BS (19MB), un sur le 2BS (22Mb) et un sur 5BL (558Mb) (figure 51 E). De la même manière que précédemment, parmi les 3599 contigs pour lesquels les valeurs de π_{DE} sont nulles, huit contigs ont une valeur de $\pi_{DP}>0.01$ (figure 51 F).

Aucune des régions correspondant à ces fortes valeurs pour les ratios de π n'est commune entre les différentes transitions de la domestication.

¹ L et S correspondent respectivement aux bras long et court des chromosomes

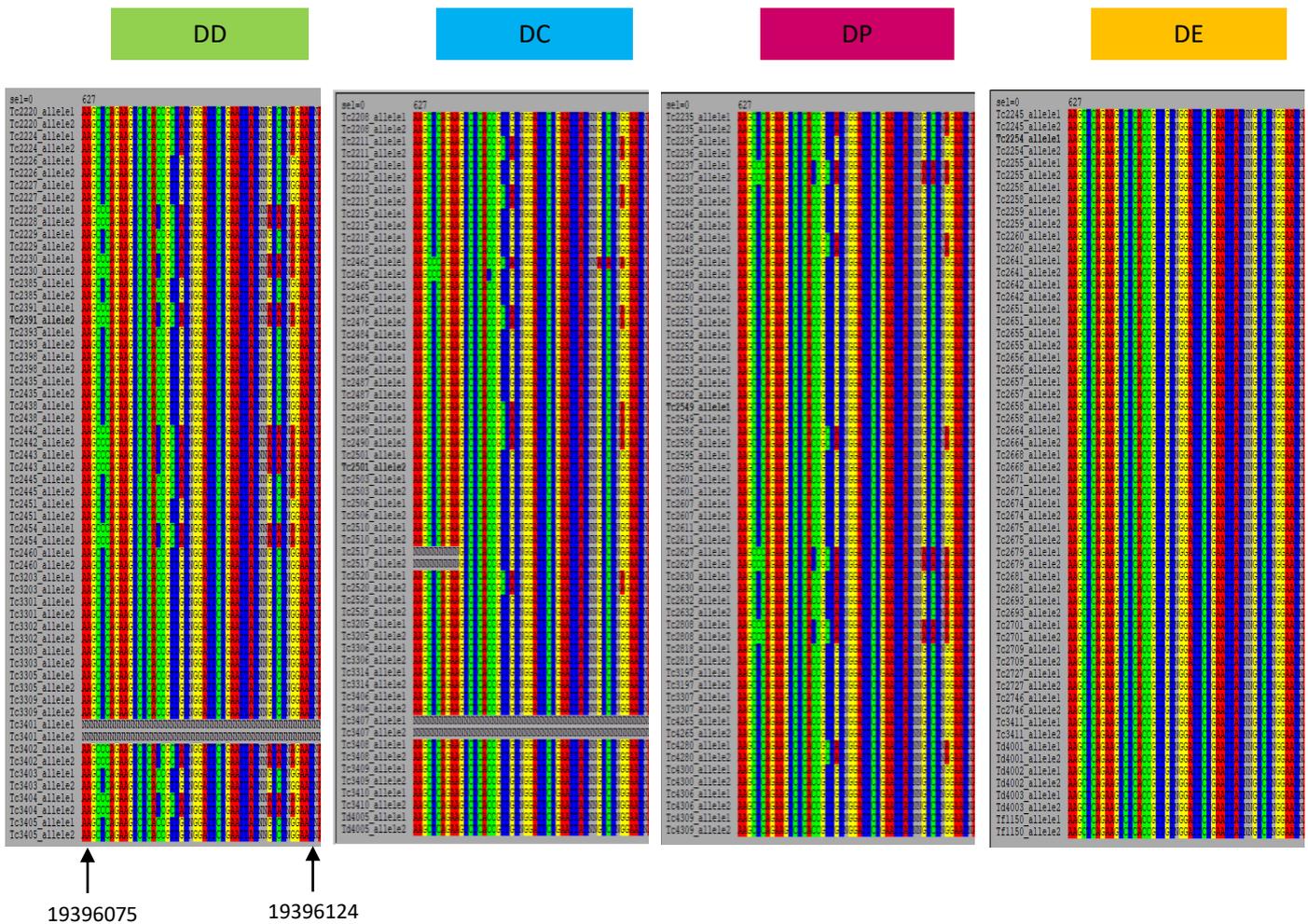


Figure 53: Fragment de 50pb du contig chr1B:1-356313144:19395448-19396491 situé sur le chromosome 1B, présentant une forte diminution de diversité au cours de la transition entre DC et DP. Les génotypes (2 allèles) sont présentés en lignes, regroupés en 4 groupes: DD, DC, DP et DE. Les locus sont en colonnes et les nucléotides sont marqués avec 4 couleurs différentes (A en rouge, T en bleu, G en jaune et C en vert) afin de visualiser les SNPs.

Tableau 9: Mesures de diversité du contig chr1B:1-356313144:19395448-19396491 situé sur le chromosome 1B. Le tableau présente, pour chacune des groupes évolutifs (DD, DC, DP et DE) la taille, en paire de bases, analysée sur l'ensemble du contig (lseff), le nombre moyen de séquences analysées (nseff), le nombre de sites polymorphes (S), les valeurs des estimateurs de diversité θ_s et π .

	DD	DC	DP	DE
lseff	739	807	719	753
nseff	49,03	50,82	49,25	53,03
S	55	57	51	1
θ_s	0,016692	0,015699	0,015908	0,000293
π	0,026898	0,019163	0,017530	0,000097

3.3.1.2 Détection avec les rapports de θ_s de Watterson

L'étude du rapport de $\frac{\theta_w DD}{\theta_w DC}$ met en évidence deux pics forts : un sur le 3BS (71Mb) et l'autre sur le 6BS (15Mb) (figure 52 A). Ces deux régions avaient également de fortes valeurs pour ces ratios de π de Tajima, mais dans des proportions différentes. Lorsque le calcul du ratio n'a pas été possible du fait des valeurs nulles de $\theta_s DC$, l'examen de la distribution des valeurs des $\theta_s DD$ correspondantes (n=2043) a montré que 11 de ces contigs ont une valeur de $\theta_s DD > 0.01$ (figure 52 B).

Le rapport $\frac{\theta_w DC}{\theta_w DP}$ met également en évidence deux pics majeurs : un sur le 5BL (578MB) et l'autre sur le 7BL (610Mb) (figure 52 C), deux régions qui avaient été également identifiées lors de l'étude basée sur le rapport des π de Tajima. Parmi les 3421 contigs pour lesquels les valeurs de $\theta_s DP$ sont nulles, 10 contigs ont une valeur de $\theta_s DC > 0.01$ (figure 52 D).

Le rapport $\frac{\theta_w DP}{\theta_w DE}$ met en évidence quatre pics: un pic très important sur 1BS (19MB), deux pics sur le chromosome 2B (22Mb et 759Mb) et un sur le 5BL (558Mb) (figure 52 E). Parmi ces quatre pics, trois sont commun avec les calculs basés sur les π de Tajima. Seule la zone située sur le 2BL n'avait pas été détectée précédemment. De la même manière, nous notons que parmi les 3599 contigs pour lesquels les valeurs de $\theta_s DE$ sont nulles, huit contigs ont une valeur de $\theta_s DP > 0.01$ (figure 52 F).

Après cette première phase d'observation globale des variations des estimateurs de diversité le long des chromosomes, j'ai choisi à titre exploratoire, deux zones présentant une forte diminution de la taille efficace, signe potentiel d'un événement sélectif, dans le but de déterminer si ces régions sont impliquées dans le contrôle génétique des traits phénotypiques ciblés lors de la domestication et/ou de l'amélioration variétale. Pour cela, j'ai comparé les résultats obtenus avec les deux estimateurs, en prenant soin de réunir les contigs ayant des rapports de diversité élevés et les contigs présentant des valeurs de π et/ou θ_s nulles alors que la forme de départ présentait des valeurs supérieures à 0.01.

Nous pouvons prendre comme premier exemple, un contig situé sur le bras court du chromosome **1B** (chr1B:1-356313144:19395448-19396491), dont la diversité a très fortement diminué lors du passage de DP à DE (figure 51 E et figure 52 E). Si le niveau de polymorphisme est important, dans cette zone, pour les groupes DD, DC et DP, l'ensemble des sites devient monomorphe pour le groupe DE (figure 53). Le tableau 9, nous permet de confirmer cette observation à l'échelle du contig, car il ne reste plus qu'un seul site polymorphe (S) sur le contig chez DE. En consultant les annotations du génome de référence, cette zone correspond à une protéine kinase, enzyme qui catalyse une phosphorylation. La phosphorylation consiste à transférer un groupe phosphate de l'ATP vers un substrat ou une cible spécifique. Les kinases occupent donc une place centrale dans les mécanismes de signalisation cellulaire en agissant sur une protéine en l'activant ou la désactivant.

Nous pouvons prendre comme deuxième exemple le bras long du chromosome **5B** où plusieurs zones sont affectées par une diminution de la diversité.

Une première zone située à environ 492 Mb, présente une diminution du polymorphisme lors du passage de la forme sauvage (DD) à la forme domestiquée (DC), notable avec le ratio de θ_s sur un contig (21,75), et une perte totale de la diversité ($\theta_{DD} > 0.01$ et $\theta_{DC} = 0$) sur deux contigs. Les annotations du génome de référence dans la zone de ces trois contigs, correspondent à différents

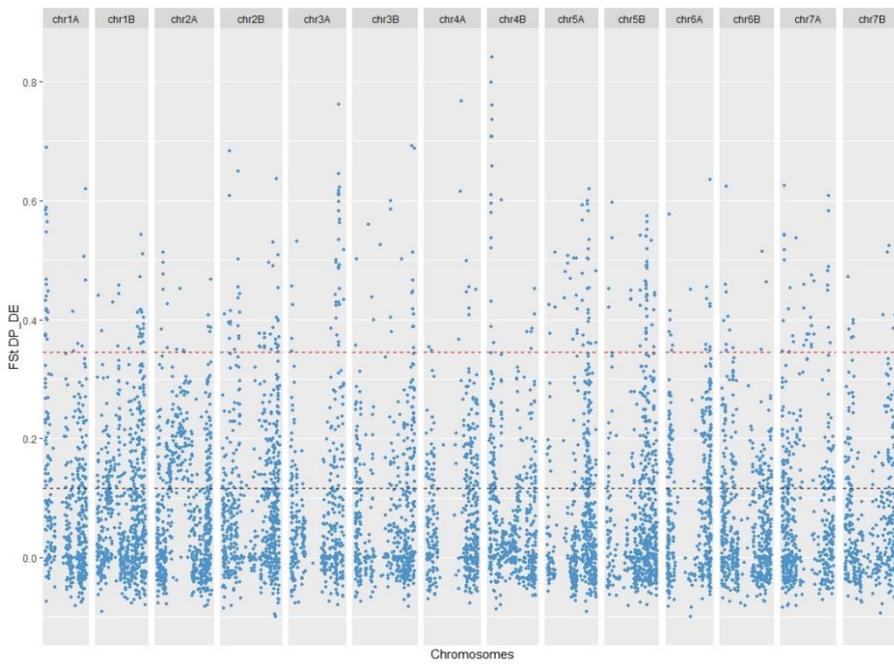
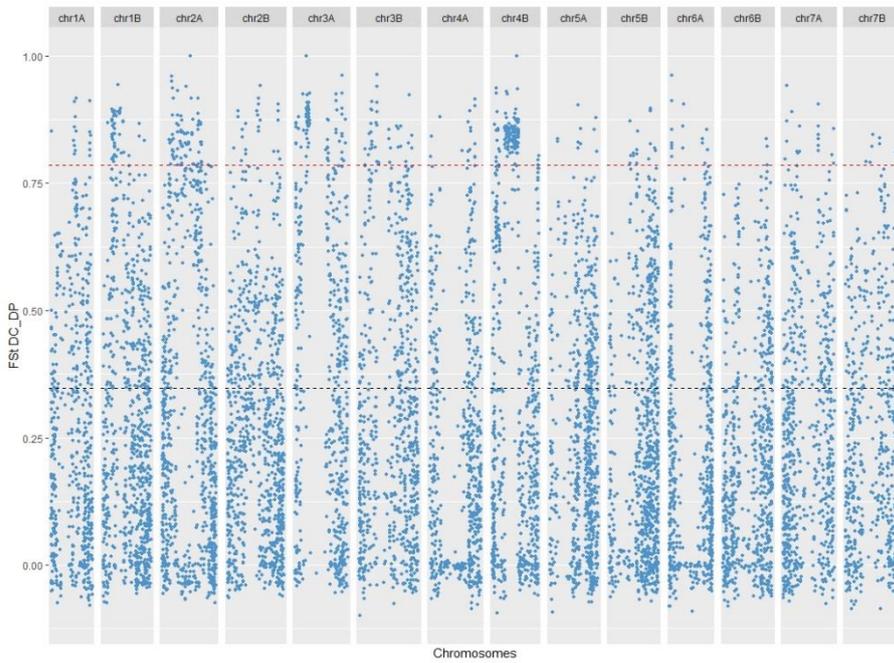
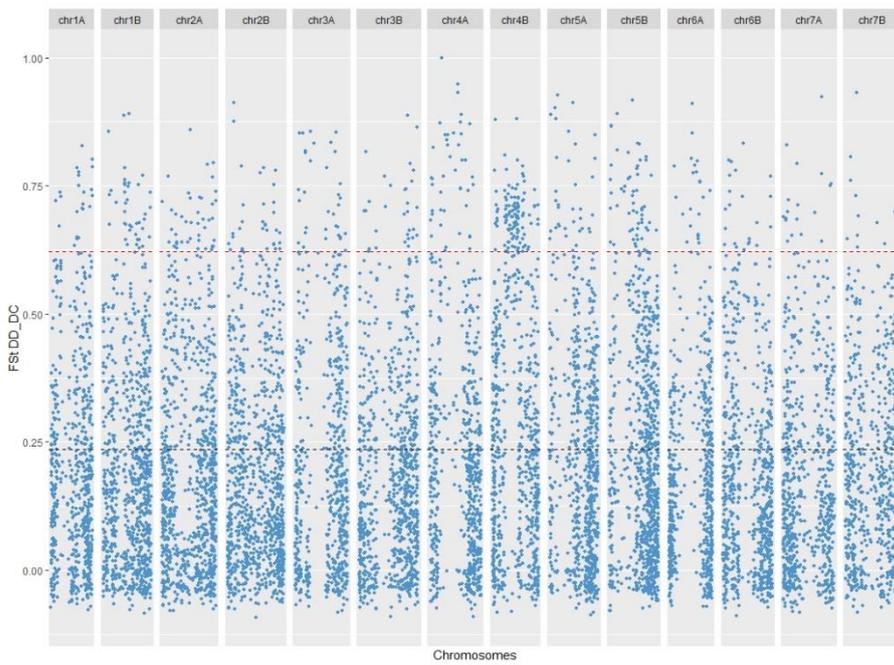


Figure 54: Niveau de différenciation mesuré par le F_{st} , pour chacune des trois transitions évolutives et à l'échelle de chacun des 10734 contigs (filtre $NA < 0.66 + MAF > 5\%$).

Pour chaque transition, la ligne en pointillés rouge est le seuil correspondant au 95^{ème} centile et la ligne en pointillés noire la valeur des F_{st} globaux, calculés entre sous-groupes sur l'ensemble des contigs.

types de gènes dont une protéine kinase, une protéine F-box, un facteur de transcription et des protéines de résistance à des maladies.

La deuxième zone concernée se situe autour de 558 Mb ; le niveau de diversité a fortement diminué lors de la transition DD vers DC, notable avec le ratio de θ_s sur deux contigs (23.25 et 25.95) et une perte totale de la diversité ($\theta_{DD} > 0.01$ et $\theta_{DC} = 0$) sur un contig. Cette région a également été impactée lors de la révolution verte entre les deux groupes DP et DE. Cette diminution de diversité touche le même contig que lors de la première transition (ratio $\theta_s = 27.53$, ratio $\pi = 295.32$) et une perte totale de la diversité, détectable avec les deux estimateurs, sur un autre contig proche. Les fonctions des gènes situés dans cette zone sont des méthyltransférases (Histone), des protéines ubiquitine, des polymérases ARN et des protéines du cytochrome.

La dernière zone remarquable sur cette partie du bras long du chromosome 5B est localisée à 611Mb. Dans cette région, la diversité génétique de trois contigs a fortement diminué au moment du passage entre DC et DP et entre DP et DE, visible avec les deux estimateurs θ_s et π . Les annotations de la référence correspondant à cette zone sont des facteurs de transcription (GRAS) et d'élongation, des glucosidases, des protéines codant pour les ribosomes et les reticulums endoplasmiques et un gène de résistance.

Nous pouvons conclure en disant que cette portion du génome porte un grand nombre de gènes codant pour des protéines de structure nécessaires au fonctionnement des cellules mais également des gènes de résistance. Ces fonctions sont susceptibles d'être affectées par des processus sélectifs au cours de la domestication car elles sont liées au développement et aux résistances aux pathogènes. Néanmoins d'autres études devront être menées pour confirmer l'occurrence de la sélection et préciser quel(s) gène(s) exactement a(ont) été ciblé(s).

3.3.1.3 Détection avec les Fst

L'indice de différenciation **Fst** a été calculé, pour les trois transitions évolutives et pour chacun des 10 734 contigs (filtre $NA < 0.66 + MAF > 5\%$) (Figure 54). Afin d'identifier les contigs ayant les plus fortes valeurs de Fst reflétant potentiellement une trace de sélection, un seuil correspondant au 95^{ème} centile a été calculé et les valeurs des Fst entre groupes sur l'ensemble du génome (présentée dans le tableau 8) ont été reportés sur les figures.

Pour la **transition entre DD et DC**, les valeurs de Fst global et du 95^{ème} centile étaient respectivement à 0.236 et 0.622, sur 10 495 contigs analysés (polymorphes entre les deux groupes évolutifs). Comme pour la diversité, nous remarquons que les contigs avec des valeurs de Fst élevées sont situés aux deux extrémités des chromosomes. De plus, la variance entre les valeurs faibles et les valeurs fortes de Fst est très importante.

Pour la **transition entre DC et DP**, le Fst global avait été calculé à 0.347 et le 95^{ème} centile a été calculé à 0.785, sur 9 700 contigs analysés. Cela traduit une différenciation plus importante entre ces deux groupes que la différenciation entre les deux groupes de la transition précédente, même si visuellement, les représentations graphiques sont comparables.

Pour la **transition entre DP et DE**, le Fst global avait été calculé à 0.116 et le 95^{ème} centile s'élevait à 0.345, sur 7 196 contigs analysés, ce qui traduit cette fois, une différenciation moins importante. Dans ce cas, les contigs très différenciés se distinguent plus facilement.

Nous pouvons voir que certaines zones ont des valeurs de Fst importantes pour plusieurs transitions évolutives. C'est le cas d'un ensemble de contig situé sur le chromosome 4B. Nous avons donc cherché

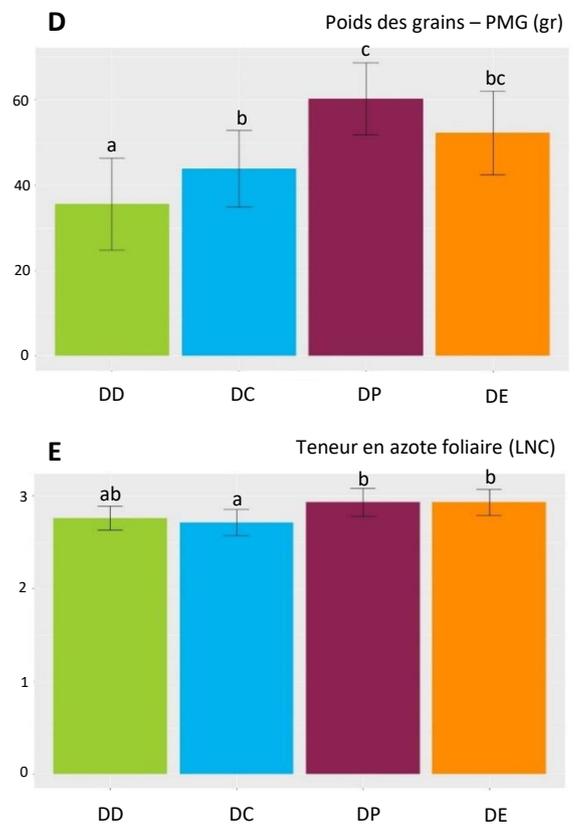
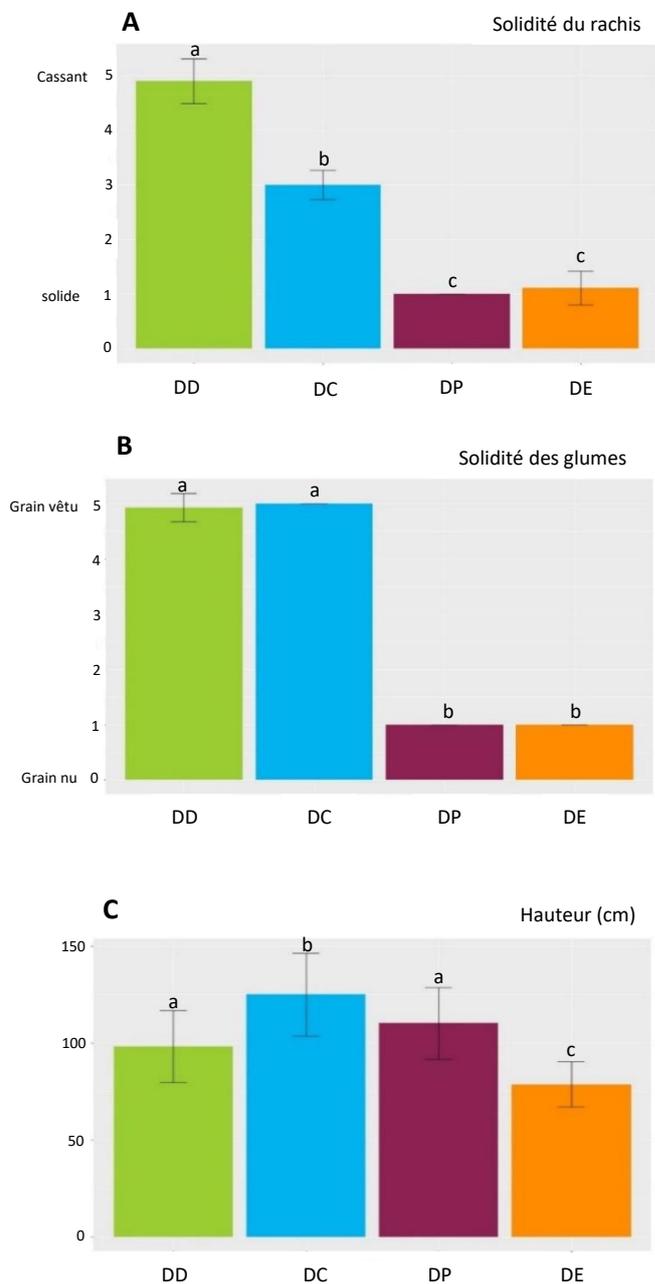


Figure 55: Mesures morphologiques sur les 120 génotypes appartenant aux 4 groupes évolutifs: DD en vert, DC en bleu, DP en bordeaux et DE en orange.

La solidité du rachis (A) est notée à 1 pour solide et à 5 pour cassant

Solidité des glumes (B) est notée de 1, pour un grain nu, à 5, pour un grain vêtu.

La hauteur des plante (C) est mesurée en cm

Le poids des grains (D) est mesurée en pesant le poids de 1000 grains (PMG) et noté en grammes.

La teneur en azote contenue dans la feuille drapeau (LNC) (E) est mesuré par spectrométrie proche infra-rouge (NIRS).

Des modèles linéaires généralisés (solidité du grain et des glumes) ou des modèles linéaires (hauteur, taille des grains et teneur en azote foliaire) ont été réalisés et associés à des tests de Tukey (Tukey, 1949) pour valider la significativité des différences de moyennes, matérialisées par les lettres a, b et c.

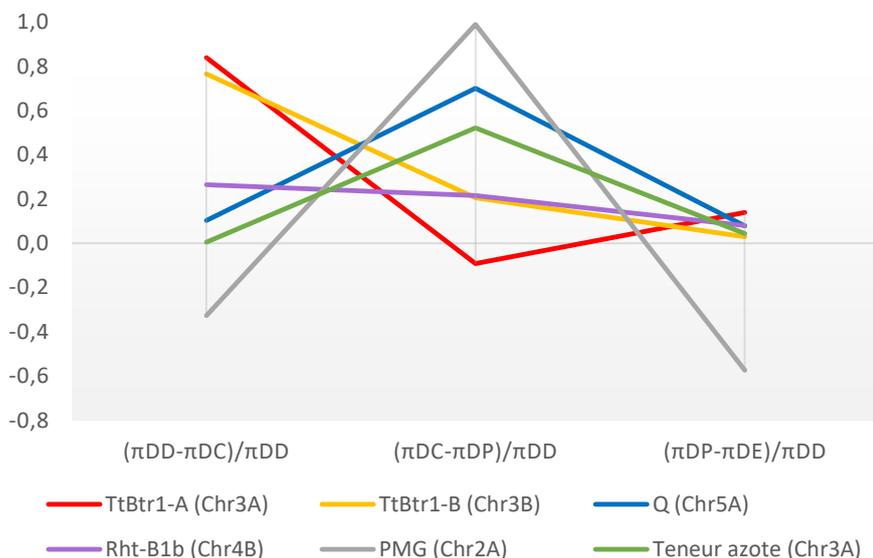


Figure 56: Graphique représentant l'évolution de la diversité à chaque transition de l'histoire évolutive de *T. turgidum* au niveau de six zones de 5Mb portant les gènes TtBtr1-A (en rouge), TtBtr1-B (en orange), Q (en bleu) et Rht-B1b (en violet) ainsi que les QTLs de la taille des grains (PMG) (en gris) et la teneur en azote foliaire (en vert).

L'évolution de la diversité est estimé par le π de Tajima à l'aide des rapports suivants:

Transition entre DD et DC : $(\pi_{DD}-\pi_{DC})/\pi_{DD}$

Transition entre DC et DP: $(\pi_{DC}-\pi_{DP})/\pi_{DD}$

Transition entre DP et DE: $(\pi_{DP}-\pi_{DE})/\pi_{DD}$

à savoir si certains contigs avaient subi des pressions de sélection importantes lors de plusieurs transitions, en comparant les contigs situés au-dessus du seuil correspondant au 95^{ème} centile. Cela nous a permis de constater que 111 contigs avaient une valeur de Fst importante à la fois lors de la transition entre DD et DC et lors de la transition entre DC et DP. Cependant, il ne restait plus que sept contigs qui avaient une valeur de Fst importante à la fois lors de la transition entre DC et DP et lors de la transition entre DP et DE.

3.3.2 Détection de sélection dans les zones candidates contrôlant des traits du syndrome de domestication

La deuxième phase de la détection des signatures génétiques de sélection liées à la domestication a été d'analyser six zones de 5Mb portant les gènes d'intérêt : TtBtr1-A (3A) et TtBtr1-B (3B) impliqués dans le caractère « rachis solide », le gène Q (5A) impliqué dans le trait phénotypique « grain nu » et le gène Rht-B1b (4B) impliqué dans le trait phénotypique « plante semi-naine » ; ainsi que les QTLs impliqués dans le poids des grains (PMG) (2A) et la teneur en azote dans la feuille, qui reflète la stratégie d'acquisition des ressources de la plante (3A).

3.3.2.1 Mesures phénotypiques

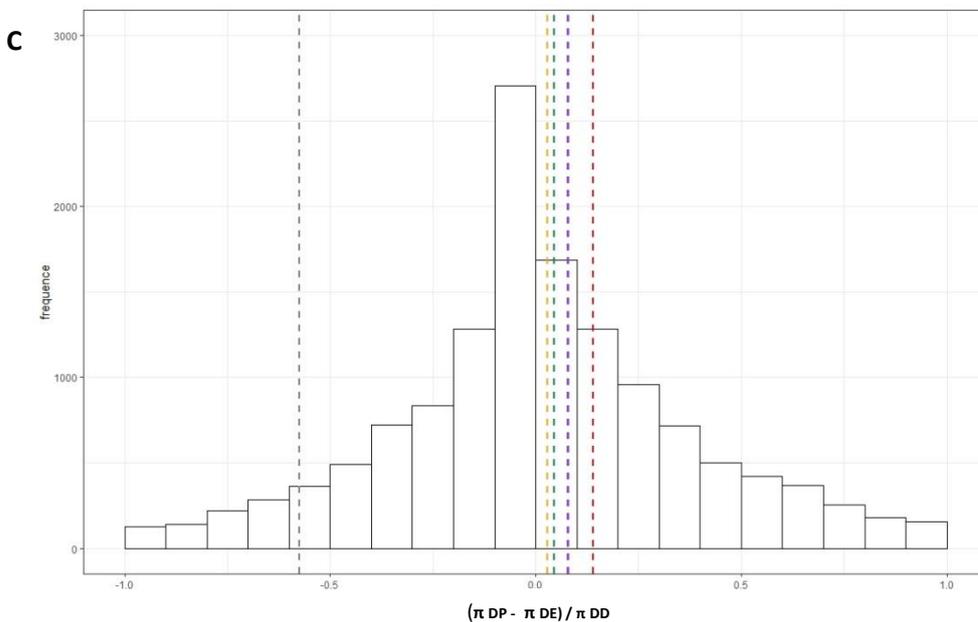
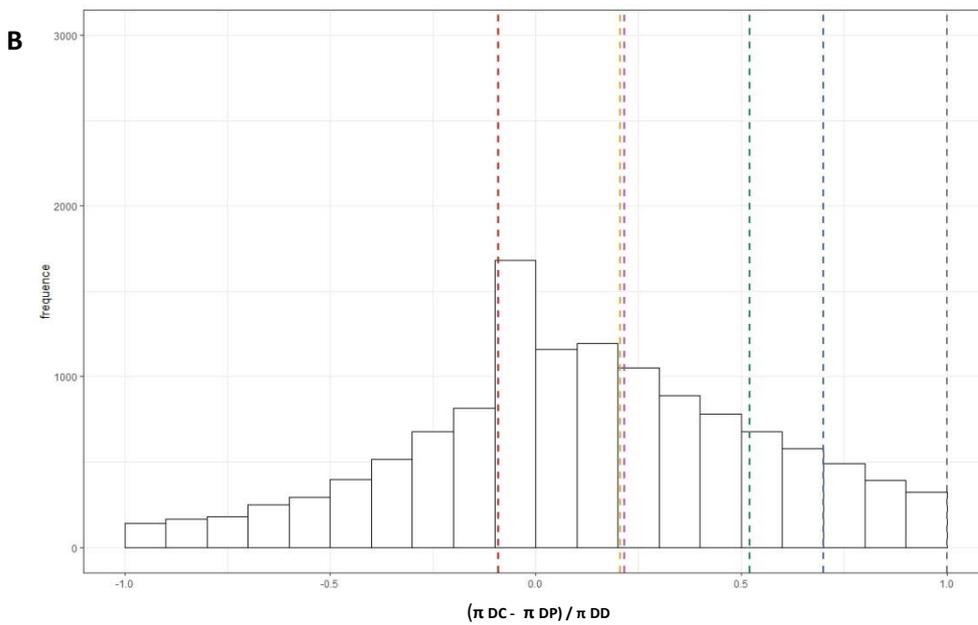
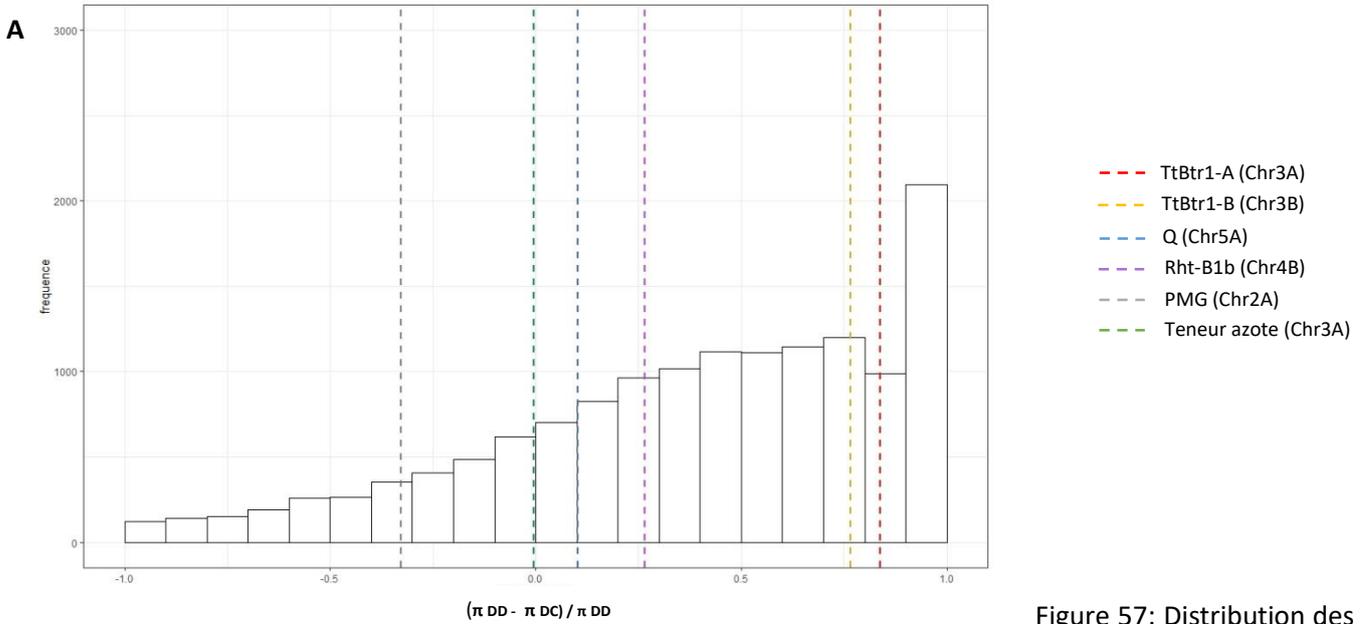
Pour valider les relations génotype-phénotype, nous nous sommes appuyés sur les mesures morphologiques réalisées au laboratoire sur les 120 génotypes qui composent le matériel expérimental de cette étude :

- ✓ La solidité du rachis en donnant une mesure de 1 (solide) à 5 (cassant)
- ✓ La solidité des glumes en donnant une mesure de 1 (grain nu) à 5 (grain vêtu)
- ✓ La hauteur des plantes à maturité
- ✓ Le poids de 1000 grains
- ✓ La teneur en azote contenu dans la feuille drapeau

Ces mesures nous ont permis d'évaluer l'impact de la domestication sur la valeur de ces traits et d'identifier la ou les transitions les plus concernées par cet impact.

Comme attendu, le passage **rachis cassant / rachis solide** s'est effectué pour l'essentiel lors de la première phase de la domestication avec l'apparition du groupe DC (figure 55 A) (p -value $< 1.10^{-3}$). Du fait de la présence d'une variabilité résiduelle pour ce caractère au sein du groupe DC, sans doute due aux difficultés de fixation de ce trait associé à des flux géniques entre les amidonniers sauvages et cultivés, la fixation complète de ce caractère n'est intervenue qu'au sein du groupe DP (p -val $< 1.10^{-4}$). L'apparition des **grains nus** est connue et est utilisée comme un marqueur de la transition entre la première forme domestiquée DC et les formes à grains nues dont DP est l'un des représentants. Les observations faites sur les 120 accessions confirment cet attendu : les notes obtenues entre DC et DP sont significativement différentes (p -value $< 1.10^{-8}$) alors que les observations entre DD et DC, d'une part, et DP et DE, d'autre part, sont comparables (figure 55 B) (p -value NS).

La dynamique temporelle de la **hauteur des plantes** est plus complexe que les précédentes : elle augmente entre DD et DC (p -value $< 1.10^{-7}$), diminue légèrement entre DC et DP (p -value = 1.10^{-2}), avant de diminuer très significativement lors de la transition entre DP et DE (p -value $< 1.10^{-8}$). Ces



observations sont conformes aux différentes observations phénotypiques et notamment à l'introgession des gènes de nanisme qui a marqué l'épisode de la révolution verte (figure 55C).

L'évolution du poids des grains, mesuré par le poids de 1000 grains (PMG) est également plus graduelle. Celle-ci augmente régulièrement entre les groupes DD et DP (p -value = 1.10^{-2}) puis se stabilise avec entre DP et DE (p -value NS) (figure 55 D).

Pour finir, la **teneur en azote** contenue dans la feuille drapeau augmente significativement entre DC et DP (p -value = 1.10^{-3}), alors qu'elle était stable entre les deux transitions DD/DC et DP/DE (p -value NS) (figure 55 E).

3.3.2.2 Détection avec les rapports de π de Tajima

Pour chacune de ces six fenêtres de 5Mb (tableau 3), nous avons mesuré la perte de diversité à chaque transition évolutive à l'aide du rapport entre les valeurs de π des groupes considérés deux à deux (figure 56).

Pour la **solidité du rachis**, la zone de 5Mb contenant le gène TtBtr1-A sur le chromosome 3A a subi une importante chute de diversité entre DD et DC ($(\pi_{DD}-\pi_{DC}) / \pi_{DD} = 0.839$), puis le niveau de diversité s'est stabilisé avec des valeurs de rapport proches de 0 pour les deux transitions suivantes. Pour la zone de 5Mb contenant le gène TtBtr1-B, sur le chromosome 3B, la principale chute de diversité a eu lieu lors de la transition entre DD et DC ($(\pi_{DD}-\pi_{DC}) / \pi_{DD} = 0.765$) mais une deuxième baisse de diversité, bien que plus faible, est intervenue lors de la transition entre DC et DP ($(\pi_{DC}-\pi_{DP}) / \pi_{DD} = 0.206$) pour se stabiliser ensuite.

Concernant la **solidité des glumes et la teneur en azote** dans la feuille : les dynamiques de diversité des deux zones de 5Mb contenant le gène Q, sur le chromosome 5A et le QTL associé à la teneur en azote, sur le chromosome 3A sont remarquablement parallèles : elles sont stables lors de la transition entre DD et DC, puis la diversité dans chacune de ces régions diminue entre DC et DP avec des valeurs de rapport de π , respectivement, à 0.700, et 0.522. Pour finir, la diversité se stabilise entre DP et DE

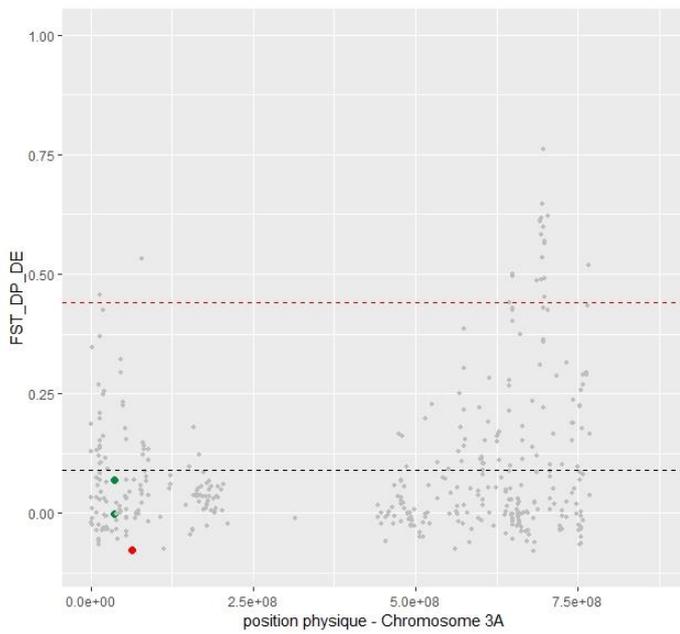
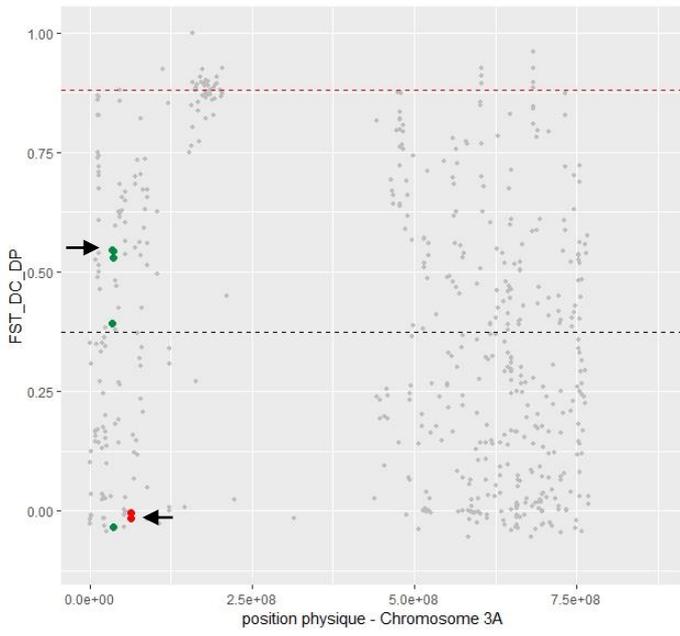
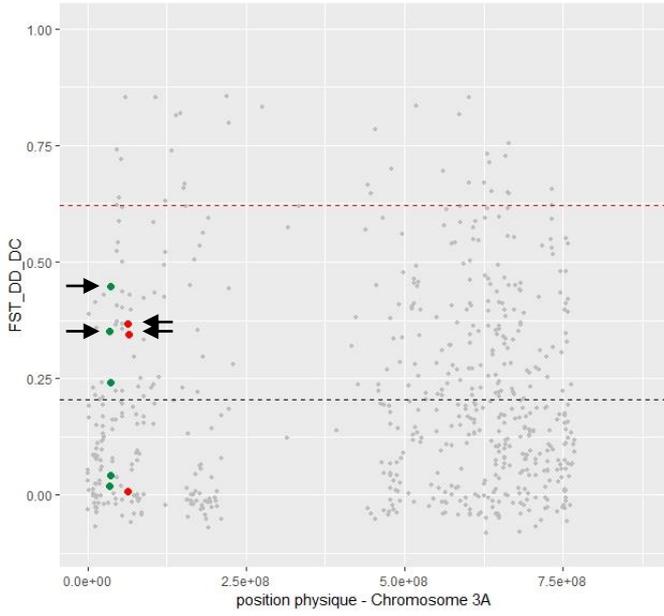
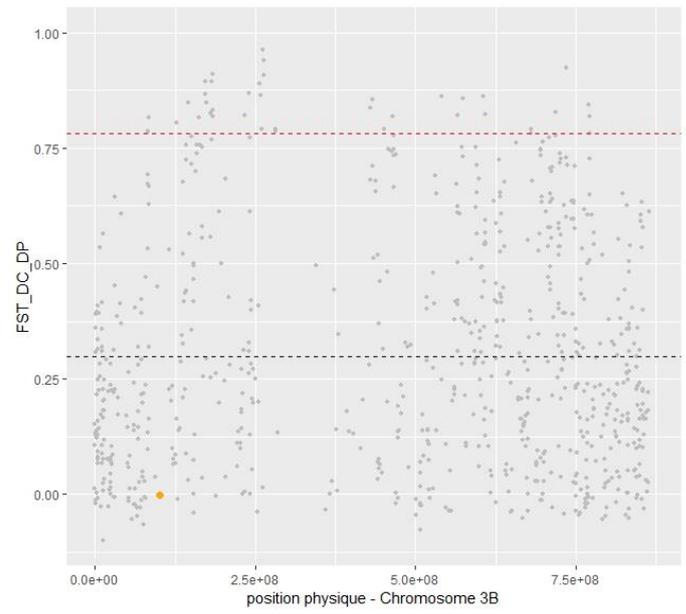
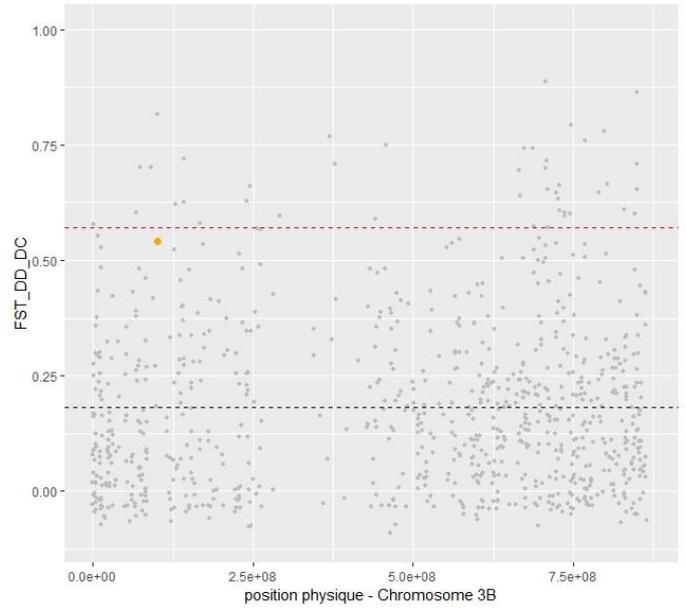
Pour la **hauteur des plantes**, la diversité génétique dans la zone de 5Mb contenant le gène Rht-B1b diminue faiblement et de façon constante, entre les deux premières transitions ($(\pi_{DD}-\pi_{DC}) / \pi_{DD} = 0.265$ et $(\pi_{DC}-\pi_{DP}) / \pi_{DD} = 0.215$) avant de se stabiliser lors de la transition DP/DE ($(\pi_{DP}-\pi_{DE}) / \pi_{DD} = 0,080$).

Concernant le poids des grains (**PMG**) : l'évolution de la diversité de la zone de 5Mb contenant le QTL suit la même tendance que celles de la solidité des glumes et de la teneur en azote, mais dans des proportions plus importantes car le rapport de π lors de la transition entre DC et DP est de 1.002.

Pour savoir si ces observations sont imputables à la sélection et pas seulement à l'histoire démographique, ces valeurs ont été comparées à celles mesurées sur le reste du génome.

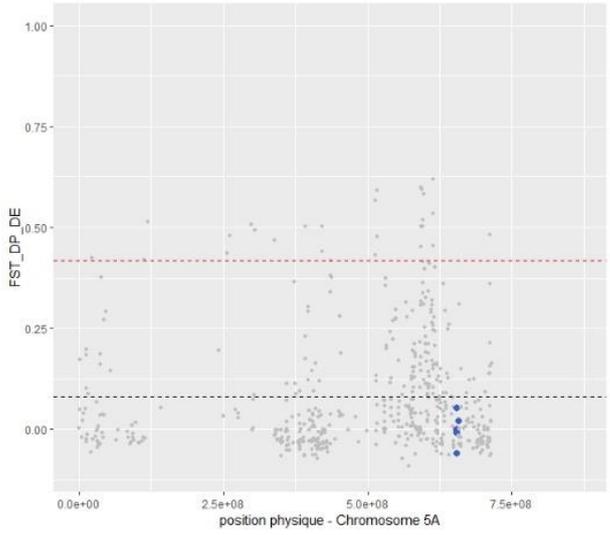
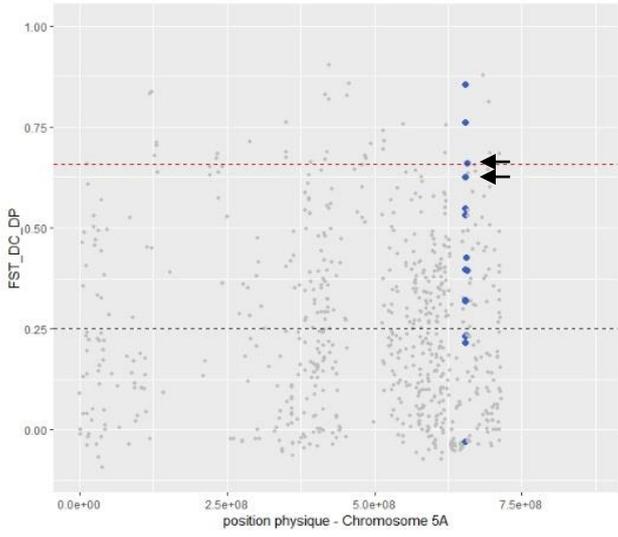
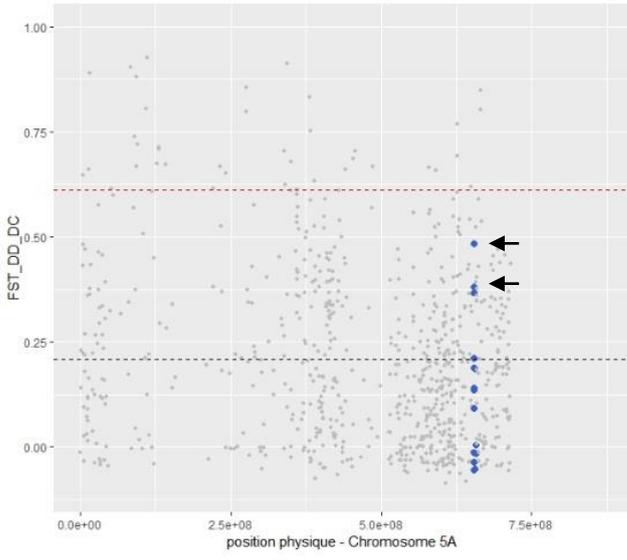
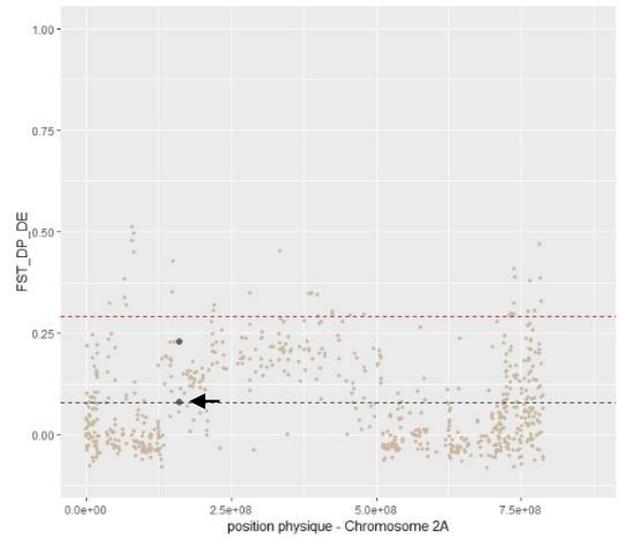
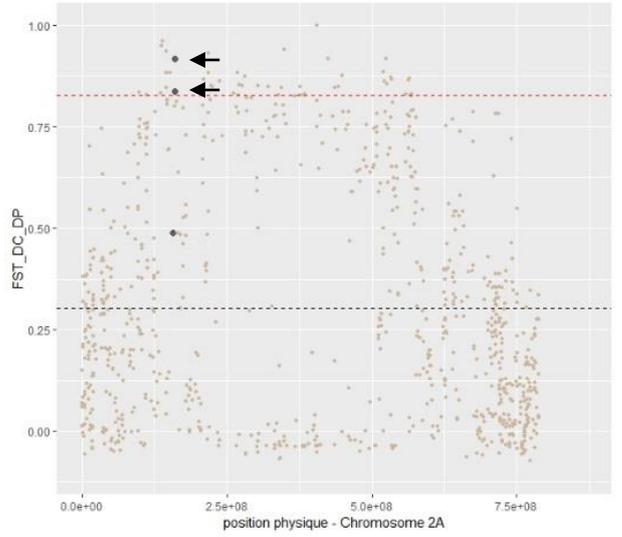
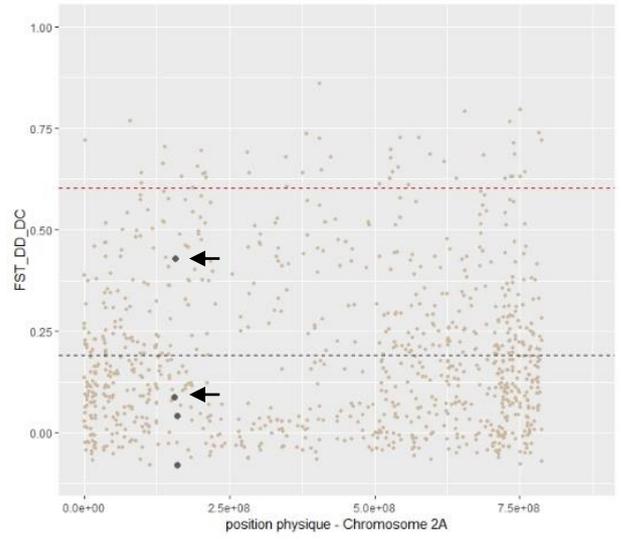
Pour la **transition entre DD et DC** (figure 57 A), 14152 contigs avaient une valeur du rapport de π comprise entre -1 et 1 et parmi eux, 4282 (soit 30 %) une valeur supérieure à 0.7. Les valeurs des rapports de π des zones portant les gènes TtBtr1-A (0,865) et TtBtr1-B (0,732) sont comparables à ce tiers des contigs qui ont subi la plus grosse perte de diversité entre DD et DC. Comme attendu, les quatre autres zones s'inscrivent au même niveau que les contigs dont la diversité n'a pas chuté de façon significative au cours de cette transition.

Lors de la **transition entre DC et DP** (figure 57 B), la valeur du rapport de π de la zone portant le gène Q est comparable au 1/12^{ème} des contigs qui ont subi la plus grosse perte de diversité. La zone portant

A**B**

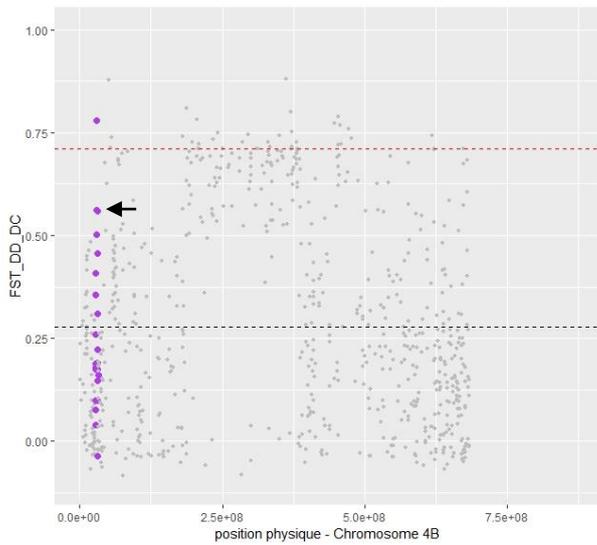
NA

- TtBtr1-A (Chr3A)
- TtBtr1-B (Chr3B)
- Q (Chr5A)
- Rht-B1b (Chr4B)
- PMG (Chr2A)
- Teneur en azote (Chr3A)
- - - Fst moyen sur le chromosome
- - - Seuil 0,95

C**D**

- TtBtr1-A (Chr3A)
- TtBtr1-B (Chr3B)
- Q (Chr5A)
- Rht-B1b (Chr4B)
- PMG (Chr2A)
- Teneur en azote (Chr3A)
- - - Fst moyen sur le chromosome
- - - Seuil 0,95

E



- TtBtr1-A (Chr3A)
- TtBtr1-B (Chr3B)
- Q (Chr5A)
- Rht-B1b (Chr4B)
- PMG (Chr2A)
- Teneur en azote (Chr3A)
- - - Fst moyen sur le chromosome
- - - Seuil 0,95

Figure 58: Fst calculés pour chaque contig et pour les trois transitions : DD_DC, DC_DP et DP_DE , au niveau des six zones cibles.

Le chromosome 3A (A) porte le gène TtBtr1-A (en rouge) et le QTL impliqué dans la teneur en azote foliaire(en vert)

Le chromosome 3B (B) porte le gène TtBtr1-B (en orange)

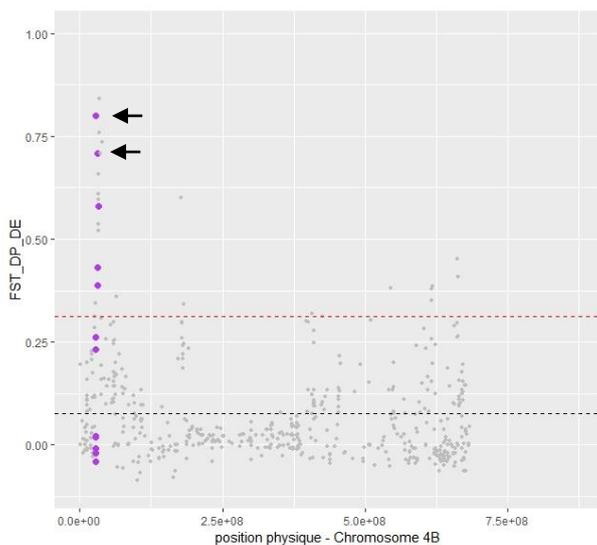
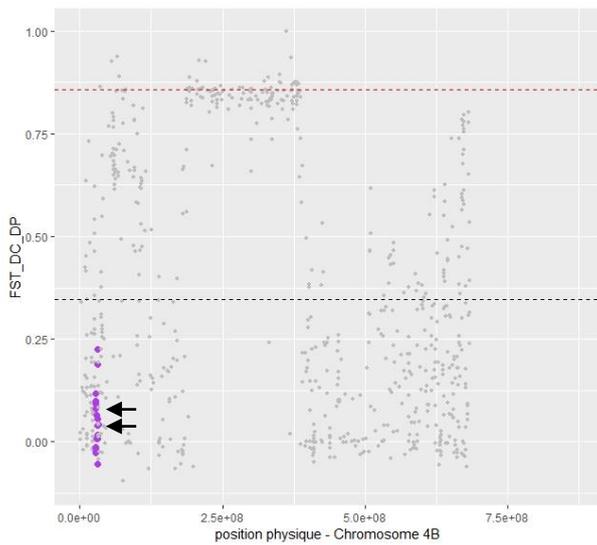
Le chromosome 5A (C) porte le gène Q (en bleu)

Le chromosome 2A (D) porte le QTL impliqué dans la taille des grains « PMG » (en gris)

Le chromosome 4B (E) pour le gène Rht-B1b (en violet).

Pour chaque transition, la ligne en pointillés rouge est le seuil correspondant au 95^{ème} centile et la courbe en pointillés noirs à la valeur moyenne de ces Fst.

Les flèches noires pointent les contigs les plus proches des gènes et QTLs cibles.



le QTL associé à la teneur en azote foliaire est, lui, comparable au 1/6^{ème} de ces contigs. Pour finir, la zone portant le QTL de rendement (PMG) ayant un rapport de π de 1.002 correspond à celles des contigs les plus touchés par la chute de diversité entre DC et DP.

Pour la **transition entre DP et DE** (figure 57 C), les six zones portant les gènes ou QTLs cibles, y compris le gène **Rht-B1b**, s'inscrivent au même niveau que les contigs dont la diversité n'a pas chuté de façon significative au cours de cette transition.

L'analyse individuelle, de chacune des cibles pour les trois transitions successives, permet de mettre clairement en évidence les étapes pendant lesquelles la sélection a eu lieu au cours de la domestication mais aussi de la sélection moderne.

3.3.2.3 Détection avec les Fst

Pour identifier des régions génomiques qui ont été soumises à sélection au cours de l'histoire du blé dur, nous avons également utilisé le paramètre **Fst**. Pour chacune des six fenêtres de 5Mb portant les gènes et QTLs cibles, nous avons comparé les valeurs de Fst des contigs présents dans ces fenêtres, par rapport à l'ensemble des valeurs des Fst des autres contigs présents sur ces chromosomes (figure 58).

Les valeurs de Fst des contigs correspondant aux gènes **TtBtr1-A** et **TtBtr1-B** ont été reportées, respectivement, sur les chromosomes 3A (figure 58 A) et 3B (figure 58 B). Dans les deux cas, le niveau de différenciation le plus important se trouve entre DD et DC, avec des valeurs de Fst supérieures à la moyenne et proche du 95^{ème} centile. Les valeurs de Fst entre DC et DP, puis entre DP et DE sont proches de 0 ce qui correspond à la fixation d'un seul et même allèle à partir de DC. Cette évolution est en accord avec le fait que la solidité du rachis est un trait qui est principalement apparu entre DD et DC. Les valeurs de Fst des contigs correspondant au gène **Q** ont été reportées sur le chromosome 5A (figure 58 C). La différenciation augmente entre DC et DP avec 3 contigs au-dessus du 95^{ème} centile, ce qui correspond à l'apparition du caractère « grain nu » chez DP. La différenciation entre DP et DE est nulle car il y a eu fixation de l'allèle Q chez DP et DE.

Les valeurs de Fst des contigs correspondant au gène **Rht-B1b**, ont été reportées sur le chromosome 4B (figure 58 E). Dans ce cas, cinq contigs ont un Fst supérieur au 95^{ème} centile entre DP et DE, reflétant l'apparition des plantes « semi-naines » chez DE. Nous notons tout de même que la différenciation entre DD et DC est plus importante qu'entre DC et DP.

Si nous observons maintenant les valeurs de Fst des contigs correspondant au **QTL** impliqué dans le poids des grains (**PMG**), reportées sur le chromosome 2A (figure 58 D), nous pouvons voir que la différenciation est maximale entre DC et DP, alors qu'elle est en dessous du 95^{ème} centile entre DD et DC ainsi qu'entre DP et DE.

Pour finir, les valeurs de Fst du **QTL** associé à la **teneur en azote dans la feuille**, ont été reportées sur le chromosome 3A (figure 58 A). Dans ce cas, nous remarquons qu'aucun contig ne passe le 95^{ème} centile, cependant, la différenciation est plus importante entre DD et DC ainsi qu'entre DC et DP.

La comparaison des valeurs de Fst de chaque contig contenu dans les fenêtres de 5Mb au regard de la position des gènes, nous a permis d'observer que dans la très grande majorité des cas, les contigs les plus proches des gènes ont les plus grandes valeurs de Fst (tableau 3, figure 58, Annexe 9). C'est le cas, par exemple, du gène Q, positionné sur la référence ZAVITAN à 30Mb sur le chromosome 4B, pour

lequel les deux contigs les plus proches (28,5Mb et 31Mb) ont les valeurs de Fst les plus importantes (respectivement 0,8 et 0,7) lors de la transition entre DP et DE. Ceci montre que cette méthode de recherche de signatures sélective est adéquate.

Conclusion sur la détection de la sélection :

Pour conclure, une première phase de détection sans *a priori* nous a permis de mettre en évidence des zones potentiellement soumises à sélection en identifiant les valeurs extrêmes du niveau de diversité (π) ou du niveau de différenciation (Fst) au sein du génome. Une deuxième phase d'analyse des zones candidates a permis de mettre en évidence des signatures génétiques lors de chaque transition évolutive. Cette étude souligne la complexité et la chronologie du processus évolutif menant d'une forme sauvage aux formes actuelles utilisées dans l'agriculture moderne.

Par ailleurs, la détection de ces zones semble plus efficace et fiable avec l'indice de différenciation Fst.

Par ailleurs, la recherche de signatures génétiques de sélection des gènes et QTLs impliqués dans les traits caractéristiques de la domestication, a donné les résultats suivants :

- ✓ Les signatures de la sélection associées aux gènes **TtBtr1-A et TtBtr1-B** contrôlant la solidité du rachis ont bien été détectées, à l'aide des niveaux de diversité de différenciation génétique, lors du passage de **DD à DC**.
- ✓ Il en va de même pour le gène **Q** impliqué dans le trait phénotypique « grain nu », dont les signatures de sélection, diminution de la diversité et augmentation de la différenciation, ont été détectées lors du passage de **DC à DP**.
- ✓ La signature de la sélection associée au gène **Rht-B1b**, contrôlant la hauteur des plantes a pu être détectée et se caractérise par une augmentation de la différenciation (Fst) entre **DP et DE**.
- ✓ Les deux QTLs impliqués dans **le poids des grains (PMG)** et **la teneur en azote** ont vu leur diversité diminuer et leur niveau de différenciation augmenter lors du passage de **DC à DP**.

Ces signatures moléculaires sont en accord avec les mesures morphologiques effectuées sur les génotypes des différentes formes évolutives.

Discussion et perspectives

4 Discussion et perspectives

4.1 Développement technologique

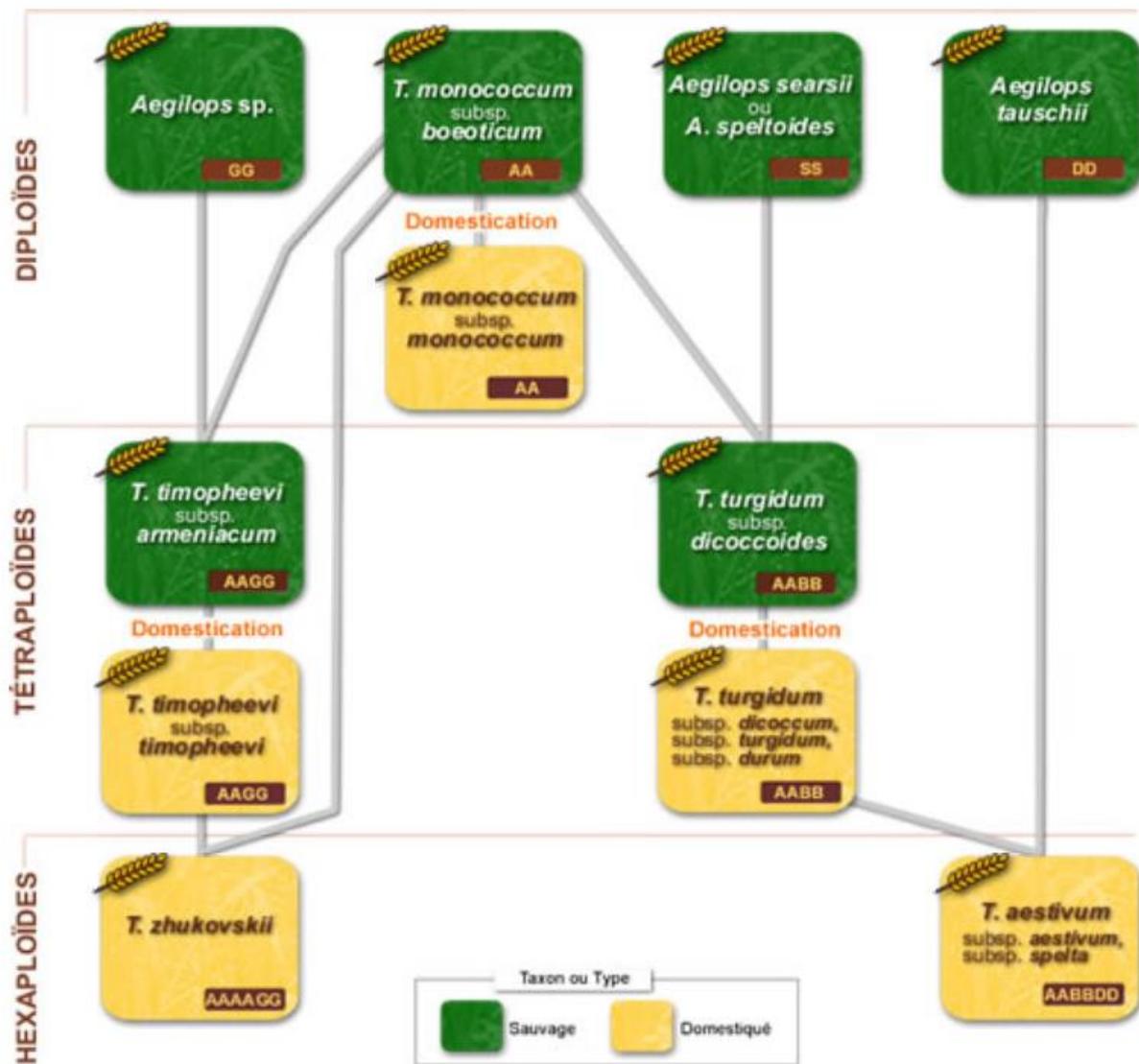
L'enrichissement par capture est une technique basée sur le séquençage de régions spécifiques dans le génome, sans avoir défini au préalable le polymorphisme attendu. Nous avons choisi, 10 000 cibles de 120pb situés dans la partie codante du génome, ce qui nous a permis d'obtenir 135 863 SNPs valides.

Des études ont été menées précédemment dans l'équipe (Holtz et al. 2016) en se basant sur le protocole publié par Rolhand et al. (2012), adapté aux grands nombres d'échantillons. Cependant, la polyploïdie et l'importante proportion de séquences répétées dans le génome du blé dur, nous a imposé de développer un protocole spécifique. La première optimisation a porté sur les modalités de multiplexage des librairies dans le but de réduire la variabilité du nombre de séquences capturées entre chacun des génotypes. Le paramètre de sur-dispersion (θ) permet de mesurer l'écart de la distribution des reads par génotype par rapport à une loi de poisson. Cette sur-dispersion a été réduite par cinq en multiplexant les librairies génomiques en équi-proportions suite à un dosage quantitatif. La seconde optimisation de ce développement technologique été d'augmenter la spécificité des fragments capturés pour augmenter le taux d'enrichissement en régions cibles. L'ajout de bloquants de séquences non-spécifiques ainsi que la réalisation d'une double capture nous a permis de multiplier par deux la spécificité des fragments capturés.

Malgré ces optimisations, il est encore nécessaire de réduire la variabilité du nombre de séquences capturées entre les génotypes afin de pouvoir augmenter davantage le nombre de génotypes par multiplex de capture et donc de réduire les coûts liés à l'expérimentation. La difficulté réside dans le fait que le dosage des librairies par fluorimétrie, utilisé pour le multiplexage des librairies, quantifie l'ensemble des fragments, avec ou sans adaptateurs, or seulement les fragments portant deux adaptateurs seront séquencés. L'amplification PCR, réalisée à la fin de la construction des libraires et avant le dosage, permet d'augmenter la proportion de fragments portant deux adaptateurs (fragments utiles). Il faut trouver un équilibre entre faire beaucoup de cycles PCR pour augmenter la proportion de fragments utiles et peu de cycles pour diminuer la quantité de duplicats PCR.

Une des solutions serait de réaliser un dosage par PCR quantitative en s'appuyant sur la partie commune à tous les adaptateurs (sans le tag), afin de ne doser que les fragments utiles. J'ai testé cette approche en réalisant des qPCR haut-débit en format microgouttes (plateforme qPHD de l'université Montpellier 2). Malgré de bons résultats, le coût de cette procédure, au regard du nombre de génotypes à traiter, n'est pas envisageable.

Une autre option serait de faire en sorte que la proportion de « fragments utiles » soit la plus équilibrée possible entre les génotypes. Pour cela, l'optimisation de la phase de ligation entre les fragments d'ADN et les adaptateurs est indispensable. Dans notre protocole, nous avons utilisé une ligation en bouts francs, pour limiter les coûts. Pour augmenter l'efficacité de celle-ci, il faudrait la remplacer par une ligation en bout cohésifs avec une base adénine débordante (ligation d'A). Cette option va nécessairement augmenter légèrement les coûts de la constitution des librairies, aspect non négligeable si on considère la quantité d'échantillons ; toutefois cette augmentation pourrait être balancée si elle permettait d'augmenter le nombre de génotypes multiplexables.



Agropolis museum

Figure 59: La généalogie des blés

Après avoir analysé les données, nous avons appliqué cette technique d'enrichissement par capture de régions codantes, pour documenter l'impact de la domestication au cours de différentes transitions de l'histoire évolutive du blé dur. La technologie mise au point nous a permis de travailler sur un grand nombre de SNPs et d'estimer correctement, différents paramètres de génétique des populations (π , Θ , D , N_e , F_{st}) à différentes échelles : sur le génome complet, le long de chaque chromosome et sur des régions ciblées contenant des gènes intervenant dans le déterminisme génétique de traits phénotypiques caractéristiques de la domestication.

Les contraintes techniques inhérentes à cette technologie d'enrichissement en séquences cibles par capture à l'aide de baits de 120pb ne garantissent pas une correspondance unique dans le génome entre les baits définis et leur sites d'accrochage, conduisant à un risque d'erreur lors de l'identification des régions orthologues. Les contigs étudiés étant d'une taille moyenne de 720pb, la variance entre eux est élevée. Pour ces raisons, nous avons étudié l'évolution de la diversité sur des statistiques calculées à partir de plusieurs contigs voisins.

4.2 Structure génétique de la série de domestication

Les résultats obtenus lors de l'analyse de la structure génétique des 120 génotypes, à l'aide des trois méthodes ACP, DACP et sNMF, ont permis de mettre en évidence trois éléments particuliers.

Des individus T. turgidum spp dicoccoïdes particuliers

Les trois méthodes utilisées ont montré que le groupe DD présentait la plus importante diversité génétique intra-spécifique, mais aussi que cette forme était la plus distante, génétiquement, des autres. Les analyses de structure (ACP et DAPC) nous ont permis de voir que les 30 génotypes du groupe DD ne formaient pas un nuage de point continu, quatre génotypes (Tc2220, Tc2398, Tc2454, Tc3401) étant génétiquement éloignés des autres génotypes de ce groupe.

La présence de ces génotypes dans l'échantillon n'est pas étonnante dans la mesure où notre méthode d'échantillonnage est basée sur des algorithmes de construction de « core collection » avec MStrat (Gouesnard, 2001) qui maximise la diversité de chaque groupe. Pour expliquer cette diversité importante intra-groupe, nous pouvons avancer deux hypothèses. La première est qu'il existe réellement une structure en deux groupes au sein de la forme sauvage. Une structure basée sur la l'origine géographique a été mise en évidence par Willcox et al. (2004) et Özkan et al. (2011). Les coordonnées géographiques des quatre génotypes (Liban, Iran et Turquie) ne soutiennent pas cette hypothèse car ils proviennent de régions de faible et haute altitude.

La seconde hypothèse serait une erreur de classement. Ce type d'erreur est possible car l'assignation des lignées à une sous-espèce a été réalisée sur quelques traits morphologiques. Ainsi, il est possible que ces quatre génotypes n'appartiennent pas à la sous-espèce *T. turgidum ssp dicoccoïdes* (AABB) mais à une autre sous-espèce tétraploïde à grain vêtu, comme par exemple : *T. timopheevi ssp armeniacum* (AAGG) (figure 59). Pour vérifier cela, nous pourrions ajouter des génotypes appartenant à cette sous-espèces ainsi que des génotypes appartenant aux formes diploïdes et refaire une analyse de structure pour voir où se situent ces quatre génotypes DD. Nous pourrions également effectuer une analyse phylogénétique (Haudry et al. 2007; Glémin et al. 2019; Maccaferri et al. 2019).

Pour estimer le niveau de différenciation entre les 4 génotypes et les autres individus DD, il faudrait calculer le *Fst* entre ces deux sous-ensembles.

Deux origines majeures pour T. turgidum spp dicoccum

L'analyse de structure avec DAPC sans les 30 génotypes DD a mis en évidence la séparation des génotypes DC en deux sous-ensembles. Cinq génotypes : Tc2212, Tc2213, Tc2465, Tc2503 et Tc3408 issus de la région caucasienne (Russie, Slovaquie, Bosnie-Herzégovine et Bulgarie) se différencient clairement des autres génotypes de ce groupe. Cette observation pourrait correspondre aux deux périodes de diffusion de *T. turgidum spp dicoccum* (Zaharieva et al. 2010).

Une étude réalisée précédemment par Sahri *et al.*, (2014) en se basant sur 15 marqueurs moléculaires de type microsatellites avait également mis en évidence une séparation du groupe DC en deux sous-groupes. La répartition des génotypes entre les deux clusters (DAPC) ne correspond pas complètement aux résultats que nous avons obtenus. Nous pouvons émettre l'hypothèse que cela vienne de la différence d'échelle d'analyse entre 15 marqueurs microsatellites et les 1523 SNPs répartis sur 683 contigs situés dans les régions codantes du génome.

Peu de diversité chez T. turgidum spp durum

Les trois méthodes d'analyse de structure utilisées ont montré que les groupes DP et DE sont proches (nuages de points superposés et niveau d'admixture faible et interconnecté). Cette observation conforte l'idée selon laquelle DP et DE sont issus d'un événement de domestication initial à partir d'une seule population fondatrice de DC. Ce scénario devra néanmoins être validé avec des modèles démographiques. La faible différenciation entre les DP et DE peut s'expliquer, d'une part, par le fait que ces deux groupes ne divergent que depuis 50 ans ; et d'autre part, que les croisements récurrents entre les accessions de ces deux groupes ont maintenu un flux de gènes constant, limitant la différenciation génétique. La sélection depuis la « révolution verte » a porté sur un nombre de traits assez réduit permettant d'augmenter le rendement, les résistances aux maladies et, bien sûr, de réduire la taille des plantes. A l'échelle du génome, ces modifications sont très ponctuelles.

Effet de l'échantillonnage sur la portée des résultats

L'échantillonnage de 30 génotypes par groupe analysés a été effectué en utilisant l'information génétique obtenue par 15 marqueurs microsatellites. La question de la représentativité de l'échantillon par rapport à la diversité disponible peut donc être posée. Cet échantillonnage est une phase critique si on considère la forme sauvage (DD) ou la première forme domestiquée (DC), car leur diversité intra et interspécifique est importante.

Les analyses de diversité et la détection des effets de sélection ont été conduites en se basant sur les groupes génétiques définis *a priori*, grâce aux informations fournies par les centres de ressources biologiques. Cette classification est basée sur quelques marqueurs morphologiques. Au vu des résultats obtenus, nous pouvons nous questionner sur la pertinence du périmètre de chaque groupe

Tableau 10 : Perte de diversité nucléotidique entre les formes sauvages et les formes cultivées

La diversité génétique totale a été estimée à l'aide du π de Tajima, sur différentes espèces comme le Maïs (*Zea mays*), la luzerne (*Medicago sativa*), le tournesol (*Helianthus annuus*) ou le blé dur (*Triticum turgidum*). D'après Haudry *et al.*, 2007.

	Diversité π compartiment sauvage (10^{-3})	Diversité π compartiment cultivé (10^{-3})	références
<i>Zea mays</i>	9.7	6.4	Wright <i>et al.</i> (2005)
<i>Medicago sativa</i>	20.2	13.5	Muller <i>et al.</i> (2006)
<i>Helianthus annuus</i>	12.8	5.6	Liu and Burke (2006)
<i>Triticum turgidum</i>	2.7	0.8	Haudry <i>et al.</i> (2007)

(notamment DD et DC). Il serait intéressant, suite à l'analyse de structure des données moléculaires, de redéfinir des groupes *a posteriori* et d'effectuer les analyses de diversité et de détection de sélection sur ces nouveaux groupes. Des analyses complémentaires restent donc à faire pour valider les signatures de sélection obtenues sur la base des groupes *a priori*. Des nouvelles estimations de π , Θ_s et F_{st} tenant compte de la structure *a posteriori* amèneront certainement quelques ajustements à nos conclusions.

4.3 Caractérisation des effets démographiques de la série de domestication

Diversité et différenciation à l'échelle du génome

La première observation que nous avons faite est que le niveau de diversité chez la forme sauvage DD est faible ($\pi=2,8 \times 10^{-3}$) au regard d'autre espèces (tableau 10).

Le niveau de diversité et de différenciation estimés entre les quatre formes évolutives de notre étude a permis de quantifier l'impact de l'histoire démographique de façon globale sur le génome. *A priori* dans le cas d'une série de domestication, nous nous attendons à ce que la diversité diminue et que le niveau de différenciation augmente, entre la forme sauvage et la dernière forme cultivée. C'est bien ce qui est observé ici à l'aide des différents estimateurs : π , Θ_s , D et F_{st} . Le plus fort goulot d'étranglement a eu lieu lors du passage de DD à DC ($\pi_{DD}=2,8 \times 10^{-3}$, $\pi_{DC}=1,9 \times 10^{-3}$), suivit d'un deuxième, de moindre importance, entre DC et DP ($\pi_{DP}=1,4 \times 10^{-3}$). La transition entre DP et DE n'a eu que peu d'impact sur le niveau de diversité à l'échelle globale ($\pi_{DE}=1,3 \times 10^{-3}$). L'estimation du D de Tajima nous permet de visualiser une phase d'expansion démographique correspondant à l'émergence de l'agriculture moderne.

D'autres études sur le blé dur ont estimé l'impact de la domestication sur la diversité génétique et le niveau de différenciation (Thuillet et al. 2005; Haudry et al. 2007; Akhunov et al. 2010; Maccaferri et al. 2019). L'origine des accessions étudiées ainsi que le type de marqueurs utilisés pour estimer la réduction de diversité sont différents pour chaque étude (microsatellites, gènes, SNPs...) (Freville et al. 2001). Il faut également prendre en considération le type de diversité observé : diversité nucléotidique totale ou diversité nucléotidique non-codante ou synonyme, d'autant plus quand il s'agit de documenter l'impact de la domestication. Cet ensemble de paramètres rend difficile une comparaison fine entre les différentes études mais le patron global de l'évolution de la diversité et de la différenciation, au cours de la domestication que nous avons obtenu est comparable aux travaux menés précédemment.

L'érosion de la diversité génétique chez les espèces cultivées par rapport à leurs apparentés sauvages a été mise en évidence sur de nombreuses espèces (Buckler et al. 2001). La proportion de diversité génétique perdue lors du passage de la forme sauvage à la forme cultivée est variable en fonction des espèces. Par exemple, la réduction de la diversité chez le maïs est moins forte que chez le blé dur (Tableau 10). La perte de diversité entre la téosinte (forme sauvage) et la première forme cultivée de maïs, et équivalente à la perte de diversité entre la forme DD et la forme DE ($\Theta_{\text{Maïs}} / \Theta_{\text{téosinte}} = 0.57$) (Wright et al. 2005).

Diversité génétique le long des chromosomes

Je me suis intéressée aux niveaux de diversité et de différenciation le long des chromosomes afin de déterminer si l'histoire démographique avait eu un impact homogène sur l'ensemble du génome.

Quel que soit le groupe considéré, nos résultats (π et θ_s) ont montré que le niveau de polymorphisme était plus important au niveau des télomères que des centromères. De tels résultats avaient déjà été rapportés dans les études antérieures (Choulet et al. 2014; Avni et al. 2017; Maccaferri et al. 2019) et s'explique par le fait que la densité en gènes et le taux de recombinaison sont plus forts au niveau des télomères (Dvorak et al. 1998). Il existe quelques exceptions à ce patron global : certaines zones proches des centromères sont très polymorphes (cas des chromosomes 1B et 5A), alors que d'autres situées sur les parties plus distales ont des patrons de diversité très variables (cas du 5BL).

Après avoir vérifié qu'il ne s'agissait pas d'un problème d'estimation de niveau de diversité ou une mauvaise localisation des contigs de la référence « ZAVITAN_BAITS » sur la référence génomique « ZAVITAN », nous pouvons émettre deux hypothèses.

La première hypothèse serait un mauvais positionnement de ces zones sur la carte physique WEWSeq v.2.0 (Avni et al. 2017; Zhu et al. 2019). Les régions pourraient en réalité être localisées plus éloignées des centromères, ce qui compte tenu de la complexité du génome du blé, ne peut être exclu. J'ai comparé les positions physiques des contigs des trois zones ayant des valeurs importantes de π , obtenus avec la référence WEWSeq v.2.0 à ceux obtenus en utilisant une nouvelle référence génomique complète, d'une variété élite (« Svevo ») appartenant au groupe *T. turgidum ssp durum* (Maccaferri et al. 2019). Les positions sont identiques sur les deux références génomiques, ce qui rend moins probable une erreur d'assignation.

Une deuxième hypothèse serait qu'il y a bien, à ces endroits, des zones très polymorphes. Les études de Avni *et al.* (2017) et Maccaferri *et al.* (2019) ont également observé des niveaux de polymorphisme élevés dans ces trois zones. Cette forte diversité pourrait provenir d'une introgression interspécifique, d'une partie de chromosome, via une hybridation ancienne avec une espèce proche (Balfourier et al. 2019).

Afin d'observer les flux de gènes potentiels propres à ces zones particulières, il serait possible de réaliser une analyse de structure, ciblées sur les locus des zones concernées, en intégrant les espèces et sous-espèces proches. Ces comparaisons sont également faisables après séquençage (banques BAC) des zones concernées. Ce type d'analyse pourrait permettre d'identifier l'origine de ces zones particulières.

4.4 Détection de signatures génétiques de sélection liées à la domestication

Détection des effets sélectifs, sans a priori, sur l'ensemble du génome

Un des objectifs de cette étude était de détecter des signatures de sélection liées à la domestication. En explorant sans *a priori* l'ensemble du génome, les ratios de diversité Θ (rapport Ne) et le F_{st} pour chacune des transitions (DD/DC, DC/DP, DP/DE) mettent en évidence des zones candidates.

Certaines zones cumulent des signaux positifs avec les deux estimateurs, comme sur le bras long du chromosome 5B lors de la transition entre DP et DE. D'autres traces de sélection ne sont détectables qu'avec un seul des deux estimateurs. C'est le cas de la zone centromérique du 4B, où nous observons

des valeurs de F_{st} élevée pour les transitions DD/DC et DC/DP sur une importante fraction du chromosome. Cependant, aucune réduction de diversité notable n'est visible sur cette zone. Cette signature ayant également été observée dans l'étude de Maccafferi et al. (2019), avec différents estimateurs (indice diversité nucléotidique et haplotypique, F_{st} , etc...), nous pouvons nous poser la question de la fiabilité de la détection de sélection en se basant sur les ratios de taille efficace.

Pour augmenter la fiabilité de détection des signaux de sélection, il aurait été intéressant de créer un modèle démographique neutre et par contraste avec ce modèle, repérer les zones soumises à sélection (Haudry et al. 2007). La détection des effets sélectifs peut également être faite à l'aide d'une méthode Bayésienne de « genome-scan » (Beaumont et al. 2002; Foll and Gaggiotti 2008) en se basant sur les F_{st} (Narum and Hess 2011).

L'étude menée par Wright et al. (2005), sur la sélection chez le maïs, utilisait le θ_s de Watterson pour détecter les gènes sous sélection forte. En utilisant le coefficient de la droite de régression, entre les valeurs de θ_s du groupe DD et du groupe DC, ainsi que les résidus du modèle, j'ai simulé des données afin de représenter une enveloppe correspondant à 95% des données (Annexe 8). Cependant, le niveau de diversité présent chez *T. turgidum* est beaucoup plus faible que chez la forme sauvage du maïs (téosinte $\pi=0.21$) et la puissance du test n'est pas suffisante pour détecter les gènes sous sélection. En effet, tous les contigs avec une valeur de θ_s forte pour le groupe DD et faible à nulle pour le groupe DC se situent dans l'enveloppe de 95%.

Détection de la sélection sur les locus ciblés

Nous avons choisi de constituer des fenêtres de 5Mb autour de six zones impliquées dans le syndrome de domestication. Cette taille résulte d'un compromis entre taille de la zone cible et le nombre de contigs présents dans la zone.

Les mesures morphologiques sur la **solidité du rachis** ont montré que l'évolution de ce trait s'est produite en deux étapes successives DD/DC puis DC/DP. Le rapport de π dans la zone portant le gène TtBtr1-B (3B) est en adéquation avec cette observation car la diversité génétique à ce locus diminue encore lors de la transition DC/DP. Cependant, ce résultat est à prendre avec prudence car le rapport de π est basé sur un seul contig. L'évolution du rapport de π de la zone portant le gène TtBtr1-A (3A) présente, lui, une seule diminution massive de diversité entre DD et DC (rapport à 0.839). Les valeurs de F_{st} des contigs concernés reflètent une absence de différenciation, et donc une fixation des allèles, à partir du groupe DC.

L'analyse des données morphologiques met en évidence une transition vers le phénotype « **grain nu** » très significative lors de la transition entre DC et DP. Cette observation est confirmée par une valeur du rapport de π lors de la transition DC/DP (0.700) bien supérieure à la transition DP/DE (0.079). C'est également le cas avec des valeurs de F_{st} supérieure au 95^{ème} centile lors de la transition DC/DP et une fixation de l'allèle Q à partir de la forme DP.

Concernant la **hauteur des plantes**, les mesures morphologiques ont montré que la taille a augmenté significativement entre DD et DC avant de diminuer chez le groupe DE. Le niveau de diversité, à ce locus, évolue peu au cours du temps (rapports de π stables pour les trois transitions). Cela pourrait s'expliquer par une sélection modérée sur ce caractère pour les trois premières formes évolutives suivi d'un apport de nouveaux allèles, venant du blé tendre, lors de l'introgession de l'allèle Rht-B1b. Cependant, les valeurs de F_{st} montrent une différenciation forte entre DD et DC, faible lors de la

transition DC/DP puis de nouveau forte entre DP et DE. Ce patron évolutif nous permet d'émettre l'hypothèse que lors de la transition entre DD et DC, un haplotype « grande taille », au locus Rht a été sélectionné car les premiers blés du groupe DC ont été cultivés pour leur grains mais aussi pour leur paille. Cela peut également venir du fait que dès l'apparition des cultures denses, les plantes d'une même espèce vont rentrer en compétition, pour la lumière par exemple, ce qui confère ainsi un avantage aux plantes de grande taille. La situation est ensuite restée stable jusqu'à la sélection moderne avec l'apparition d'un nouvel haplotype lié à l'introduction de l'allèle *rht1* issu du blé tendre. Cette hypothèse pourrait être validée par l'analyse des haplotypes pour chacune des formes évolutives.

Le **poids des grains** (PMG) est un trait phénotypique qui a évolué de façon plus lente au cours de la domestication. Il augmente significativement lors de la transition entre la forme sauvage DD et la forme cultivée DC, puis de façon encore plus importante entre DC et DP. Cette deuxième augmentation correspond bien à la mise en place de la sélection génalogique dont un des objectifs principaux était le rendement en semoule basé pour l'essentiel sur des paramètres liés à la taille et à la densité du grain. La signature génétique de cette sélection a été détectée, à l'aide des rapports de π (1.002) pour la transition entre DC et DP. Le niveau de diversité chez DD étant faible et légèrement inférieur à celui de DC, le calcul du rapport de π : $(\pi_{DD} - \pi_{DC}) / \pi_{DD}$ engendre une valeur négative. Un biais d'échantillonnage est sûrement à l'origine de la valeur négative pour le rapport de π lors de la transition entre DP et DE. Par ailleurs, la signature génétique a bien été détectée avec les valeurs de *Fst* car les deux contigs les plus proches du QTL ont des valeurs de *Fst* supérieure 95^{ème} centile lors de la transition DC/DP, alors qu'ils sont en dessous pour les deux autres transitions.

Enfin, les mesures de la **teneur en azote** dans la feuille ont permis d'observer une augmentation significative entre DC et DP. Cette observation est confirmée par les valeurs des rapports de π ainsi que par les valeurs de *Fst* dans la zone du QTL. Cette augmentation de la teneur en azote dans les feuilles lors de la transition entre DC et DP traduit une modification de stratégie d'acquisition des ressources par la plante et est en adéquation avec les résultats publiés récemment par Roucou *et al.*, (2018).

Les signatures génétiques associées à des traits phénotypiques contrôlés par un seul locus ont été détectées à l'aide l'indice de différenciation *Fst*, ce qui nous permet de valider cette méthode de détection. Il serait intéressant de rechercher les signatures de sélection de l'ensemble des QTLs impliqués, dans le poids des grains par exemple, afin d'analyser la dynamique de sélection à chaque transition évolutive et pour chacun des locus.

En conclusion, la technique d'enrichissement par capture nous a permis de retrouver un certain nombre d'éléments connus, relatifs à la structure génétique de l'espèce *T. turgidum* et aux conséquences de la domestication sur la diversité génétique. La densité de marqueurs, de type SNPs, nous a permis de visualiser de façon relativement fine, les effets de la domestication le long des chromosomes. Pour finir, nous avons détecté les signatures de sélection aux locus impliqués dans le déterminisme génétique de cinq traits phénotypiques caractéristiques de la domestication.

A la suite de ce travail nous avons formulé différentes propositions pour améliorer l'efficacité de cette technologie, dans le but d'affiner notre connaissance de l'histoire évolutive de l'espèce *T. turgidum*.

Bibliographie

- Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Deal KR, Dubcovsky J, Gill BS, Gu YQ, Hadam J, Heo H, Huo N, Lazo GR, Luo M-C, Ma YQ, Matthews DE, McGuire PE, Morrell PL, Qualset CO, Renfro J, Tabanao D, Talbert LE, Tian C, Toleno DM, Warburton ML, You FM, Zhang W, Dvorak J (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* 11:702 . <https://doi.org/10.1186/1471-2164-11-702>
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357:93–97 . <https://doi.org/10.1126/science.aan0032>
- Balfourier F, Bouchet S, Robert S, De Oliveira R, Rimbert H, Kitt J, Choulet F, Paux E (2019) Worldwide phylogeography and history of wheat genetic diversity. *Science Advances* 5:eaav0536
- Barker R (1985) International Research and Third World Agriculture: Discussion. *American Journal of Agricultural Economics* 67:1085 . <https://doi.org/10.2307/1241377>
- Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. *Genetics* 162:2025–2035
- Belay G, Furuta Y (2001) Zymogram patterns of α -amylase isozymes in Ethiopian tetraploid wheat landraces: insight into their evolutionary history and evidence for gene flow. *Genetic Resources and Crop Evolution* 11:34–42
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120 . <https://doi.org/10.1093/bioinformatics/btu170>
- Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR (1993) Control of flower development in *Arabidopsis thaliana* by APETALA1 and interacting genes. *The Company of Biologists* 721–743
- Bowman JL, Smyth DR, Meyerowitz EM (1989) Genes Directing Flower Development in *Arabidopsis*. *The Plant Cell* 37–52
- Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo M-C, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–710 . <https://doi.org/10.1038/nature11650>
- Buckler ES, Thornsberry JM, Kresovich S (2001) Molecular Diversity, Structure and Domestication of Grasses. *Genet Res* 77: . <https://doi.org/10.1017/S0016672301005158>
- Cavalli-Sforza LL, Ammerman AJ (1984) *The Neolithic transition and the genetics of populations in Europe*. Princeton University Press

- Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences* 110:8057–8062 . <https://doi.org/10.1073/pnas.1217133110>
- Charlesworth J, Eyre-Walker A (2007) The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences* 104:16992–16997 . <https://doi.org/10.1073/pnas.0705456104>
- Chen R, Im H, Snyder M (2015) Whole-Exome Enrichment with the Roche NimbleGen SeqCap EZ Exome Library SR Platform. *Cold Spring Harb Protoc* 2015:pdb.prot084855 . <https://doi.org/10.1101/pdb.prot084855>
- Cheng C, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E (2003) Polyphyletic origin of cultivated rice: based on the interspersed pattern of SINEs. *Molecular Biology and Evolution* 20:67–75
- Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package - I : One-table methods. *R News* 4:6
- Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345:1249721–1249721 . <https://doi.org/10.1126/science.1249721>
- Chuck G, Meeley R, Hake S (2008) Floral meristem initiation and meristem cell fate are regulated by the maize AP2 genes *ids1* and *sid1*. *Development* 135:3013–3019 . <https://doi.org/10.1242/dev.024273>
- Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, Nabholz B, Sabot F, Sauné L, Ardisson M, Bacilieri R, Besnard G, Berger A, Cardi C, De Bellis F, Fouet O, Jourda C, Khadari B, Lanaud C, Leroy T, Pot D, Sauvage C, Scarcelli N, Tregear J, Vigouroux Y, Yahiaoui N, Ruiz M, Santoni S, Labouisse J-P, Pham J-L, David J, Glémin S (2017) Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet* 13:e1006799 . <https://doi.org/10.1371/journal.pgen.1006799>
- Darwin C (1859) *On the Origin of Species by Means of Natural Selection*
- David J, Holtz Y, Ranwez V, Santoni S, Sarah G, Ardisson M, Poux G, Choulet F, Genthon C, Roumet P, Tavaud-Pirra M (2014) Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. *Mol Breeding* 34:1531–1548 . <https://doi.org/10.1007/s11032-014-0179-z>
- De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet* 13:27 . <https://doi.org/10.1186/1471-2156-13-27>
- Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418:700–707 . <https://doi.org/10.1038/nature01019>

- Doebley J (1989) Isozymic evidence and the evolution of crop plants, pp. 165–191 in *Isozymes in Plant Biology*, edited by Soltis DE, Soltis PS. Dioscorides Press, Portland, OR
- Doebley J (2004) The Genetics of Maize Evolution. *Annu Rev Genet* 38:37–59 .
<https://doi.org/10.1146/annurev.genet.38.072902.092425>
- Doust A (2007) Architectural Evolution and its Implications for Domestication in Grasses. *Annals of Botany* 100:941–950 . <https://doi.org/10.1093/aob/mcm040>
- Drews GN, Bowman JL, Meyerowitz EM (1991) Negative regulation of the Arabidopsis homeotic gene AGAMOUS by the APETALA2 product. *Cell* 65:991–1002
- Dvorak J, Akhunov ED (2005) Tempos of Gene Locus Deletions and Duplications and Their Relationship to Recombination Rate During Diploid and Polyploid Evolution in the Aegilops-Triticum Alliance. *Genetics* 171:323–332 . <https://doi.org/10.1534/genetics.105.041632>
- Dvorak J, Luo M-C, Yang Z-L (1998) Restriction Fragment Length Polymorphism and Divergence in the Genomic Regions of High and Low Recombination in Self-Fertilizing and Cross-Fertilizing Aegilops Species. *Genetics* 423–434
- Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences* 95:4441–4446 . <https://doi.org/10.1073/pnas.95.8.4441>
- Faris JD, Fellers JP, Brooks SA, Gill BS (2003) A Bacterial Artificial Chromosome Contig Spanning the Major Domestication Locus Q in Wheat and Identification of a Candidate Gene. *Genetics* 311–321
- Flutre T (2014) Demultiplex. <https://github.com/timflutre/quantgen/blob/master/demultiplex.py>
- Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* 180:977–993 .
<https://doi.org/10.1534/genetics.108.092221>
- Frankel OH, Brown AH, Burdon JJ (1995) *The conservation of plant biodiversity*. Cambridge University Press
- Freville H, Justy F, Olivieri I (2001) Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). *Mol Ecol* 10:879–889 . <https://doi.org/10.1046/j.1365-294X.2001.01249.x>
- Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics* 196:973–983 .
<https://doi.org/10.1534/genetics.113.160572>
- Gayral P, Melo-Ferreira J, Glémin S, Bierre N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, Belkhir K, Cahais V, Loire E, Bernard A, Galtier N (2013) Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. *PLoS Genet* 9:e1003457 . <https://doi.org/10.1371/journal.pgen.1003457>
- Glémin S, Bataillon T (2009) A comparative view of the evolution of grasses under domestication: Tansley review. *New Phytologist* 183:273–290 . <https://doi.org/10.1111/j.1469-8137.2009.02884.x>

- Glémin S, Scornavacca C, Dainat J, Burgarella C, Viader V, Ardisson M, Sarah G, Santoni S, David J, Ranwez V (2019) Pervasive hybridizations in the history of wheat relatives. *Science Advances* 5:eaav9188
- Goudet J (2005) hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol Ecol Notes* 5:184–186 . <https://doi.org/10.1111/j.1471-8286.2004.00828.x>
- Gouesnard B, Bataillon T, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: An Algorithm for Building Germ Plasm Core Collections by Maximizing Allelic or Phenotypic Richness. *Journal of Heredity* 92:93–94 . <https://doi.org/10.1093/jhered/92.1.93>
- Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* 439:749–752 . <https://doi.org/10.1038/nature04434>
- Harlan J (1955) The great plains region (Part 4). *Agric Food Chem* 3:29–31
- Harlan JR (1971) Agricultural origins: centers and noncenters. *Science* 174:468–474
- Harlan JR (1992) Crops and man. American Society of Agronomy
- Harlan JR, de Wet JMJ, Price EG (1973) COMPARATIVE EVOLUTION OF CEREALS. *Evolution* 27:311–325 . <https://doi.org/10.1111/j.1558-5646.1973.tb00676.x>
- Harris DR, Masson V, Berezkin YE, Charles M, Gosden C, Hillman G, Kasparov A, Korobkova G, Kurbansakhatov K, Legge A, others (1993) Investigating early agriculture in Central Asia: new research at Jeitun, Turkmenistan. *Antiquity* 67:324–338
- Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glémin S, David J (2007) Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Molecular Biology and Evolution* 24:1506–1517 . <https://doi.org/10.1093/molbev/msm077>
- Hedden P (2003) The genes of the Green Revolution. *Trends in Genetics* 19:5–9 . [https://doi.org/10.1016/S0168-9525\(02\)00009-4](https://doi.org/10.1016/S0168-9525(02)00009-4)
- Holtz Y, Ardisson M, Ranwez V, Besnard A, Leroy P, Poux G, Roumet P, Viader V, Santoni S, David J (2016) Genotyping by Sequencing Using Specific Allelic Capture to Build a High-Density Genetic Map of Durum Wheat. *PLoS ONE* 11:e0154609 . <https://doi.org/10.1371/journal.pone.0154609>
- Holtz Y, Bonnefoy M, Viader V, Ardisson M, Rode NO, Poux G, Roumet P, Marie-Jeanne V, Ranwez V, Santoni S, Gouache D, David JL (2017) Epistatic determinism of durum wheat resistance to the wheat spindle streak mosaic virus. *Theor Appl Genet* 130:1491–1505 . <https://doi.org/10.1007/s00122-017-2904-6>
- Huang S, Sirikhachornkit A, Su X, Faris J, Gill B, Haselkorn R, Gornicki P (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences* 99:8133–8138 . <https://doi.org/10.1073/pnas.072223799>
- Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology* 7:44

- Irish VF, Sussex Ian M (1990) Function of the *apetala-1* Gene during Arabidopsis Floral Development. *The Plant Cell* 741–753
- Jantasuriyarat C, Vales MI, Watson CJW, Riera-Lizarazu O (2004) Identification and mapping of genetic loci affecting the free-threshing habit and spike compactness in wheat (*Triticum aestivum* L.). *Theor Appl Genet* 108:261–273 . <https://doi.org/10.1007/s00122-003-1432-8>
- Jofuku KD, Den Boer BGW, Van Montagu M, Okamoto JK (1994) Control of Arabidopsis Flower and Seed Development by the Homeotic Gene *APETALA2*. *The Plant Cell* 6:1211–1225
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 11:94 . <https://doi.org/10.1186/1471-2156-11-94>
- Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press
- Kimura M, Ohta T (1971) *Theoretical aspects of population genetics*. Princeton University Press
- Kipfer BA (2000) *Encyclopedic dictionary of archaeology*. Springer Science & Business Media
- Komaki MK, Okada K, Nishino E, Shimura Y (1988) Isolation and characterization of novel mutants of Arabidopsis thaliana defective in flower development. *The Company of Biologists* 195–203
- Kunst L, Klenz JE, Martinez-Zapater J, Haughn GW (1989) *AP2* Gene Determines the Identity of Perianth Organs in Flowers of Arabidopsis thaliana. *The Plant Cell* 1:1195–1208
- Lee D-Y, Lee J, Moon S, Park SY, An G (2007) The rice heterochronic gene *SUPERNUMERARY BRACT* regulates the transition from spikelet meristem to floral meristem. *The Plant Journal* 49:64–78
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760 . <https://doi.org/10.1093/bioinformatics/btp324>
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326:289–293 . <https://doi.org/10.1126/science.1181369>
- Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* 46:1220–1226 . <https://doi.org/10.1038/ng.3117>
- Lobell DB, Gourdji SM (2012) The Influence of Climate Change on Global Crop Productivity. *Plant Physiol* 160:1686–1697 . <https://doi.org/10.1104/pp.112.208298>
- Lukaszewski AJ, Curtis CA (1993) Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. *Theoret Appl Genetics* 86:121–127 . <https://doi.org/10.1007/BF00223816>

- Luo M-C, Yang Z-L, You FM, Kawahara T, Waines JG, Dvorak J (2007) The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor Appl Genet* 114:947–959 . <https://doi.org/10.1007/s00122-006-0474-0>
- Lynch M, Walsh B (2007) *The origins of genome architecture*. Sinauer Associates Sunderland, MA
- Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Gnocchi M, Moscatelli M, Avni R, Deek J, Biyiklioglu S, Frascaroli E, Corneti S, Salvi S, Sonnante G, Desiderio F, Marè C, Crosatti C, Mica E, Özkan H, Kilian B, De Vita P, Marone D, Joukhadar R, Mazzucotelli E, Nigro D, Gadaleta A, Chao S, Faris JD, Melo ATO, Pumphrey M, Pecchioni N, Milanese L, Wiebe K, Ens J, MacLachlan RP, Clarke JM, Sharpe AG, Koh CS, Liang KYH, Taylor GJ, Knox R, Budak H, Mastrangelo AM, Xu SS, Stein N, Hale I, Distelfeld A, Hayden MJ, Tuberosa R, Walkowiak S, Mayer KFX, Ceriotti A, Pozniak CJ, Cattivelli L (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 51:885–895 . <https://doi.org/10.1038/s41588-019-0381-3>
- Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H (2018) Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics* 19:19–22 . <https://doi.org/10.1186/s12864-017-4416-9>
- Matsuoka Y (2011) Evolution of Polyploid Triticum Wheats under Cultivation: The Role of Domestication, Natural Hybridization and Allopolyploid Speciation in their Diversification. *Plant and Cell Physiology* 52:750–764 . <https://doi.org/10.1093/pcp/pcr018>
- McClintock B (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* 36:344–355 . <https://doi.org/10.1073/pnas.36.6.344>
- McTavish EJ, Hillis DM (2015) How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics* 16:266 . <https://doi.org/10.1186/s12864-015-1469-5>
- Mori N (2003) Origins of domesticated emmer and common wheat inferred from chloroplast DNA fingerprinting. *Tenth International Wheat Genetics Symposium, 2003* 25–28
- Mori N, Liu Y-G, Tsunewaki K (1995) Wheat phylogeny determined by RFLP analysis of nuclear DNA. 2. Wild tetraploid wheats. *Theoret Appl Genetics* 90:129–134 . <https://doi.org/10.1007/BF00221006>
- Narum SR, Hess JE (2011) Comparison of FS outlier tests for SNP loci under selection. *Molecular Ecology Resources* 11:184–194 . <https://doi.org/10.1111/j.1755-0998.2011.02987.x>
- Nesbitt M, Samuel D (1998) Wheat domestication: archaeobotanical evidence. *Science* 279:1431–1431
- Nesbitt M, Samuel D (1996) From staple crop to extinction? The archaeology and history of hulled wheat
- Nielsen R (2005) Molecular Signatures of Natural Selection. *Annu Rev Genet* 39:197–218 . <https://doi.org/10.1146/annurev.genet.39.073003.112420>

- Özkan H, Brandolini A, Pozzi C, Effgen S, Wunder J, Salamini F (2005) A reconsideration of the domestication geography of tetraploid wheats. *Theor Appl Genet* 110:1052–1060 .
<https://doi.org/10.1007/s00122-005-1925-8>
- Özkan H, Brandolini A, Schäfer-Pregl R, Salamini F (2002) AFLP Analysis of a Collection of Tetraploid Wheats Indicates the Origin of Emmer and Hard Wheat Domestication in Southeast Turkey. *Molecular Biology and Evolution* 19:1797–1801 .
<https://doi.org/10.1093/oxfordjournals.molbev.a004002>
- Özkan H, Willcox G, Graner A, Salamini F, Kilian B (2011) Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). *Genet Resour Crop Evol* 58:11–53 .
<https://doi.org/10.1007/s10722-010-9581-5>
- Padulosi S, Hammer K, Heller J (eds) (1996) Hulled wheats. IPK, Gatersleben
- Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F, Sudhakar D, Christou P, Snape JW, Gale MD, Harberd NP (1999) ‘Green revolution’ genes encode mutant gibberellin response modulators. *Nature* 400:256–261 .
<https://doi.org/10.1038/22307>
- Peng J, Ronin Y, Fahima T, Roder MS, Li Y, Nevo E, Korol A (2003) Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. *Proceedings of the National Academy of Sciences* 100:2489–2494 . <https://doi.org/10.1073/pnas.252763199>
- Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier M-H, Leroy P, Ganal MW (1998) A Microsatellite Map of Wheat. *Genetics* 149:2007–2023
- Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* 22:939–946 . <https://doi.org/10.1101/gr.128124.111>
- Roucou A, Violle C, Fort F, Roumet P, Ecartot M, Vile D (2018) Shifts in plant functional strategies over the course of wheat domestication. *J Appl Ecol* 55:13 . <https://doi.org/10.1111/1365-2664.13029>
- Sahri A, Chentoufi L, Arbaoui M, Ardisson M, Belqadi L, Birouk A, Roumet P, Muller M-H (2014) Towards a comprehensive characterization of durum wheat landraces in Moroccan traditional agrosystems: analysing genetic diversity in the light of geography, farmers’ taxonomy and tetraploid wheat domestication history. *BMC Evol Biol* 14:264 .
<https://doi.org/10.1186/s12862-014-0264-2>
- Sauvage C, Rau A, Aichholz C, Chadoeuf J, Sarah G, Ruiz M, Santoni S, Causse M, David J, Glémin S (2017) Domestication rewired gene expression and nucleotide diversity patterns in tomato. *Plant J* 91:631–645 . <https://doi.org/10.1111/tpj.13592>
- Schlotterer C (2002) A Microsatellite-Based Multilocus Screen for the Identification of Local Selective Sweeps. *Genetics* 753–763
- Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* 15:1289–1303 .
<https://doi.org/10.1111/1755-0998.12402>

- Simonetti MC, Bellomo MP, Laghetti G, Perrino P, Simeone R, Blanco A (1999) Quantitative trait loci influencing free-threshing habit in tetraploid wheats. *Genetic Resources and Crop Evolution* 46:267–271
- Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, Gill BS, Faris JD (2006) Molecular Characterization of the Major Wheat Domestication Gene Q. *Genetics* 172:547–555 .
<https://doi.org/10.1534/genetics.105.044727>
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Thuillet A-C, Bataillon T, Poirier S, Santoni S, David JL (2005) Estimation of Long-Term Effective Population Sizes Through the History of Durum Wheat Using Microsatellite Data. *Genetics* 169:1589–1599 . <https://doi.org/10.1534/genetics.104.029553>
- Thuillet A-C, Bru D, David J, Roumet P, Santoni S, Sourdille P, Bataillon T (2002) Direct Estimation of Mutation Rate for 10 Microsatellite Loci in Durum Wheat, *Triticum turgidum* (L.) Thell. ssp durum desf. *Molecular Biology and Evolution* 19:122–125 .
<https://doi.org/10.1093/oxfordjournals.molbev.a003977>
- Tsagkogeorga G, Cahais V, Galtier N (2012) The Population Genomics of a Fast Evolver: High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate *Ciona intestinalis*. *Genome Biology and Evolution* 4:852–861 . <https://doi.org/10.1093/gbe/evs054>
- Van der Veen M (1995) Ancient agriculture in Lybia: a review of the evidence. *Acta Palaeobotanica* 85–98
- Vavilov N (1951) The origin, variation, immunity and breeding of cultivated plants
- Vilmorin (1880) Les meilleurs blé description et culture des principales variétés de froments d’hiver et de printemps, Vilmorin-andrieux et cie
- Watanabe N, Sekiya T, Sugiyama K, Yamagishi Y, Imamura I (2002) Telosomic mapping of the homoeologous genes for the long glume phenotype in tetraploid wheat. *Euphytica* 128:129–134
- Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7:256–276 . [https://doi.org/10.1016/0040-5809\(75\)90020-9](https://doi.org/10.1016/0040-5809(75)90020-9)
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370 . <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x>
- Weng M-L, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB (2019) Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in *Arabidopsis thaliana*. *Genetics* 211:703–714 . <https://doi.org/10.1534/genetics.118.301721>
- Willcox G (1998) Archaeobotanical evidence for the beginnings of agriculture in Southwest Asia. *The origins of agriculture and crop domestication* 25–38
- Willcox G (2000) De la cueillette à l’agriculture. *Pour la science*

- Willcox G (2004) Measuring grain size and identifying Near Eastern cereal domestication: evidence from the Euphrates valley. *Journal of archaeological science* 31:145–150
- Wright IJ, Reich PB, Westoby M, Ackerly DD, Baruch Z, Bongers F, Cavender-Bares J, Chapin T, Cornelissen JHC, Diemer M, Flexas J, Garnier E, Groom PK, Gulias J, Hikosaka K, Lamont BB, Lee T, Lee W, Lusk C, Midgley JJ, Navas M-L, Niinemets Ü, Oleksyn J, Osada N, Poorter H, Poot P, Prior L, Pyankov VI, Roumet C, Thomas SC, Tjoelker MG, Veneklaas EJ, Villar R (2004) The worldwide leaf economics spectrum. *Nature* 428:821–827 .
<https://doi.org/10.1038/nature02403>
- Wright S (1978) Genetic variability in natural populations: methods. *Evolution and the Genetics of Populations* 4:
- Wright S (1969) *Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies*, University of Chicago Press
- Wright SI, Bi Irie V, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The Effects of Artificial Selection on the Maize Genome. *Science* 308:1310–1314 .
<https://doi.org/10.1126/science.1107891>
- Xie Q, Li N, Yang Y, Lv Y, Yao H, Wei R, Sparkes DL, Ma Z (2018) Pleiotropic effects of the wheat domestication gene Q on yield and grain morphology. *Planta* 247:1089–1098 .
<https://doi.org/10.1007/s00425-018-2847-4>
- Xie Q, Mayes S, Sparkes DL (2015) Spelt as a Genetic Resource for Yield Component Improvement in Bread Wheat. *Crop Science* 55:2753 . <https://doi.org/10.2135/cropsci2014.12.0842>
- Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, Wollmann H, Chen X, Schmid M (2010) Orchestration of the Floral Transition and Floral Development in *Arabidopsis* by the Bifunctional Transcription Factor APETALA2. *Plant Cell* 22:2156–2170 . <https://doi.org/10.1105/tpc.110.075606>
- Zaharieva M, Ayana NG, Hakimi AA, Misra SC, Monneveux P (2010) Cultivated emmer wheat (*Triticum dicoccon* Schrank), an old crop with promising future: a review. *Genet Resour Crop Evol* 57:937–962 . <https://doi.org/10.1007/s10722-010-9572-6>
- Zhang Z, Belcram H, Gornicki P, Charles M, Just J, Huneau C, Magdelenat G, Couloux A, Samain S, Gill BS, Rasmussen JB, Barbe V, Faris JD, Chalhoub B (2011) Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proceedings of the National Academy of Sciences* 108:18737–18742 .
<https://doi.org/10.1073/pnas.1110552108>
- Zhu T, Wang L, Rodriguez JC, Deal KR, Avni R, Distelfeld A, McGuire PE, Dvorak J, Luo M-C (2019) Improved Genome Sequence of Wild Emmer Wheat Zavitan with the Aid of Optical Maps. *G3* 9:200902.2018 . <https://doi.org/10.1534/g3.118.200902>
- Zohary D (2004) Unconscious selection and the evolution of domesticated plants. *Economic botany* 58:5–10
- Zohary D, Hopf M (2000) *Domestication of plants in the Old World* . 3rd edn . 316pp. New York: Oxford University Press.

Annexes

Annexe 1: Protocole détaillé pour l'extraction d'ADN



Extraction de 20 à 50 mg matière fraîche avec billes magnétiques et les tampons AMM sur le robot KingFisher Flex

Historique des versions

Référence : PEX-EXT-002n		Gestionnaire : Responsable qualité
Version	date de version	Historique des modifications
Version 00	26/11/2015	Création
Version 01	02/10/2018	Révision

Etat du document : A Contrôler

Confidentialité : Public

Sommaire

1. Objectif /Principe	1
2. Consignes de sécurité.....	2
Fiche de visas.....	2

1. Objectif /Principe

L'objectif de ce protocole est d'isoler rapidement et de manière fiable de l'ADN génomique de haute qualité à partir de tissus végétaux. Il est adapté au traitement d'une quantité de départ de 20 à 50 mg de matériel frais. Ce protocole utilise la technologie de « capture » des acides nucléiques sur billes magnétiques et est donc dédié à l'utilisation conjointe d'une robotique adaptée (KingFisher Flex, instruction « **INS-EXT-005-Kingfisher-Flex.pdf** »).

Ce protocole a été spécifiquement mis au point pour purifier l'ADN végétal et minimiser la co-purification avec des polysaccharides et des polyphénols de la cellule. Le tampon de lyse cellulaire utilise à la fois le SDS et le CTAB, les deux détergents les plus efficaces. Nous utilisons des billes métalliques coatées avec de la silice (groupements Silanol) (Billes Perkin-Elmer Chemagen CMG 252-A). L'adsorption des acides nucléique est réalisée par déshydratation en présence d'éthanol. Les tampons de lavage contiennent du perchlorate de sodium. Il s'agit d'un protocole dérivé de celui indiqué dans le brevet portant sur la bille Chemagen (US patent 6 958 372 B2, Magnetic Silanised Polyvinylalcohol based carrier materials)

2. Consignes de sécurité

Si nécessaire, décrire les précautions

- ❖ A la manipulation de produits dangereux
 - Ethanol
 - Perchlorate de sodium en solution
 - CTAB

A l'utilisation de l'équipement (KingFisher Flex, voir **INS-EXT-005-Kingfisher-Flex.pdf**)

3. Matériel utilisé/nécessaire

- ❖ Centrifugeuse (3000 à 5000g)
- ❖ Incubateur-agitateur (56°C à 65°C)
- ❖ Robot de purification des acides nucléiques (KingFisher Flex)
- ❖ Matériel de laboratoire classique

Fiche de visas

	Rédacteur	Testeur	Vérificateur	Approbateur
Nom	TOLLON			
Fonction	Technicienne			
Visa				
Date				

Mode opératoire

BROYAGE (SUR 25 A 50 MG DE MATIERE FRAICHE)

L'échantillon frais (limbe, tige feuillée) est placé dans un tube Eppendorf de 2 ml.

L'échantillon peut être lyophilisé et broyé dans un tube de 2 ml avec le broyeur à bille à T° ambiante.

Il peut être congelé et broyé avec le broyeur à bille en définissant (selon l'échantillon) un protocole de trempage dans l'azote liquide pendant la phase de broyage

Remarques : dans le cas d'un broyage avec des billes :



- Nous recommandons d'utiliser des tubes de 2 ml à fond rond (et non conique) pour une meilleure efficacité du broyage.
- En général, deux billes en acier inoxydable ou une bille de céramique sont utilisées par échantillon.
- Un broyage direct dans le tampon d'extraction entraîne souvent une certaine dégradation de l'ADN

EXTRACTION

La poudre est reprise par 400 µl de tampon d'extraction SDS chaud.

TAMPON D'EXTRACTION : 200mM Tris pH = 8.0, 50mM EDTA, 500mM NaCl, 1.25 % SDS, 1% PVP 40000 + 1 g/100ml NaBisufite (à ajouter extemporanément)

Le tampon est préchauffé vers 50°C.

Ajouter 16 µl de solution de CTAB 12,5 %
Ajouter 2 µl d'ARNase à 10mg/ml
Agiter par inversion et au Vortex.

Incuber à 65°C pendant 30 min dans une étuve (tube à plat ou rangés dans une boîte en carton posée sur la tranche, sous agitation douce, 60rpm).

DEPROTEINISATION

Ajouter 150 µl d'acétate de K (5M/3M) froid (dans la glace). Agiter par inversion et au Vortex. Incuber 5 min dans la glace.

Centrifuger 10 min à 12 500 rpm à 4°C dans la **centrifugeuse Eppendorf avec le rotor cassette**.

Remarque : cette étape de centrifugation à haute vitesse permet d'obtenir un surnageant clair, débarrassé des particules en suspension.

Préparer la plaque de **Binding**

- 15 µl de billes magnétiques Chemagic
- 100 µL de CG 7,8 (Chlorure de guanidium 7,8 M)
- 600 µL d'éthanol à 96°

Transférer 250 µl de surnageant dans la plaque de binding

Le robot KingFisher se chargera du mixing.

Remarque : à cette étape, il est possible de laisser l'échantillon à 4°C pendant quelques heures (temps du repas ou d'une nuit par exemple)

	Instruction Atelier de marquage moléculaire INRA	PEX-EXT-002n 26/11/2015 Page 4 sur 7
	Extraction 20-50 mg avec billes magnétiques et robot KingFisher	

A partir de cette étape c'est le robot KingFisher qui prend en charge la purification.

1. Préparation des plaques pour le robot :

- 1 plaque 96 puits deepwell contenant 600 ul de tampon Wash 1
- 1 plaque 96 puits deepwell contenant 600 ul de tampon Wash 2
- 1 plaque 96 puits deepwell contenant 600 ul de tampon Wash 3
- 1 plaque 96 puits deepwell contenant 600 ul de tampon Wash 4
- 1 plaque 96 puits deepwell contenant 600 ul de tampon Wash 5
- 1 plaque d'éluion 96 puits avec 100 ul de TE 1X

2. Passage sur robot King Fisher en utilisant le programme Chemagic_DNA_Plant_Kit_5 lavages (voir le protocole d'utilisation du robot)

3. Poser la plaque d'éluion sur une plaque aimantée. Transférer les ADN dans les tubes de la DNA Bank (tubes matrix) en prenant le moins de billes restantes possible.

Les ADNs sont ensuite conservés à -20°C.

4. Procéder à l'élimination des déchets et au lavage des plaques suivant le protocole PEX-EXT-010.

4. Réactifs

Réactifs chimiques, éléments de sécurité

	Formule chimique	Toxicité	Protection	Phrase de risque	Phrase de conseil de prudence
Tris (Trisma Base, Tris[hydroxyméthyl]aminométane)	C ₄ H ₁₁ NO ₃	irritant	cutanée	R36/37/38	S26-36
EDTA (Ethylenediaminetetraacetic Acid)	C ₁₀ H ₁₄ N ₂ O ₈ Na ₂ 2H ₂ O	irritant	cutanée	R36	S26
Chlorure de sodium (sodium chloride)	NaCl		cutanée		
SDS (Sodium Dodecyl Sulfate, Sodium lauryl sulfate)	C ₁₂ H ₂₅ O ₄ SNa	irritant	Cutanée, respiratoire (si poudre)	R22-36/37/38	S26-36
Bisulfite de sodium (sodium bisulfite, sodium disulfite, sodium metabisulfite)	Na ₂ S ₂ O ₅	irritant	cutanée	R22-31	S25-46
CTAB cetyltriméthylammonium bromide	C ₁₉ H ₄₂ BrN	irritant	Cutanée, respiratoire (si poudre)	R22-36/37/38	S26-36
Acétate de potassium (potassium acetate)	CH ₃ COOK	irritant	Cutanée, respiratoire	R36/37/38	S26-36
Acide acétique glacial (glacial acetic acid)	CH ₃ COOH	irritant	Cutanée, respiratoire , oculaire	R10-35	S26-36/37/39-45

	Instruction Atelier de marquage moléculaire INRA	PEX-EXT-002n 26/11/2015 Page 5 sur 7
	Extraction 20-50 mg avec billes magnétiques et robot KingFisher	

Polyvinylpyrrolidone PVP (PVP 40000, PVT 40T)	[C ₆ H ₉ NO] _n	Irritant	Cutanée, respiratoire		S22-24/25
Ethanol (alcool éthylique, Ethyl alcool)	CH ₃ CH ₂ OH		cutanée	R11-20/21/22-36/37/38	S7-16-24/25/26
Perchlorate de sodium	NaClO ₄	Irritant			

PREPARATION DES SOLUTIONS

Tampons d'extraction

Tampon d'extraction SDS :

200mM Tris pH = 8.0, 50mM EDTA, 500mM NaCl, 1.25 % SDS, 1% PVP 40000

Préparé à partir de solutions stock disponibles (cf fiches de préparation)

SOLUTIONS/Poudres	100 ml	200 ml	300 ml	400ml
TRIS 1M PH = 8	20ml	40	60	80
EDTA 0,5M	10ml	20	30	40
NAACL 2.5M	20ml	40	60	80
SDS 20%	6,25ml	12,5	19	25
PVP 40000	1 g	2 g	3 g	4 g
NA BISULFITE	1 g	2 g	3 g	4 g
H₂O UP	43.5ml	87.5ml	131ml	175

Ajouter le NaBisulfite au moment de l'utilisation.

Solution aqueuse à 12.5 % de CTAB : 6.25 g de CTAB + qsp 50 ml H₂O

Acétate de K (3M K et 5M Ac)

Pour 100 ml, mélanger 60 ml de KAc 5M, 11.5 ml d'acide acétique glacial (37%) et 28.5 ml d'H₂O.

T5ERNase : 4 ul de RNase 10 mg/ml dans l'eau (solution stock) dans 16 ul de **T5E**

T5E : Tris 50 mM pH 8, EDTA 1 mM.

CG 7.8 Chlorure de guanidium 7,8 M. Dissoudre 37.25 gr de chlorure de guanidium (produit Sigma ref G-3272) dans l'eau jusqu'à un volume final de 50 mL de solution (faire le mélange dans un tube Sarstedt ou Falcon de 50 mL). Vérifier, si possible, l'indice de réfraction au réfractomètre : n = 1.449.

Billes magnétiques Chemagic (Perkin Elmer), ref. CMG-252-A.

Tampons de lavages

Tampons de lavages Home made (wash 1, wash 2, wash 3, wash 4).

Perchlorate de Na 2M : 12.2 g dans 50 ml H₂O dd. !! **Préparer la quantité exacte à utiliser.**
Ne pas stocker de solution mère.

Wash 1 :

Solution finale	Produit/Solution initiale	50 ml	100 ml	500 ml
Ethanol 30%	Ethanol 96%	15 ml	30 ml	150 ml
AcONa 0.15M	Acétate de Na (trihydraté)	1 g	2 g	10 g
Perchlorate de Na 1M	Perchlorate de Na 2M	25 ml	50 ml	250 ml
Chlorure de guanidium 1M	AMMCG 7.8 M	6.5 ml	13 ml	65 ml
Triton X100 1%	Triton X100	0.5 ml	1 ml	5 ml
H2O UP	H2O UP	Qsp 50 ml	Qsp 100 ml	Qsp 500 ml

Wash 2 :

Solution finale	Solution initiale	50 ml	100 ml	500 ml
Ethanol 30%	Ethanol 96%	15 ml	30 ml	150 ml
Perchlorate de Na 1M	Perchlorate de Na 2M	25 ml	50 ml	250 ml
Chlorure de guanidium 1M	AMMCG 7.8 M	6.5 ml	13 ml	65 ml
Triton X100 0.5%	Triton X100	0.25 ml	0.5 ml	2.5 ml
H2O UP	H2O UP	Qsp 50 ml	Qsp 100 ml	Qsp 500 ml

Wash 3 et 5 : éthanol 75 %

Ethanol à 75%. En respectant les proportions de la Table de Gay-Lussac :
100 ml d'éthanol à 96% + 31 ml d'eau.

Wash 4 : AMMLAV/E

Faire une solution aqueuse **AMMLav** (Acétate de potassium 160 mM, Tris HCl ph 8 22,5 mM, EDTA 0,1 mM)



Solutions	100 ml	500 ml	1 litre
Tris 1M pH=8	2.25 ml	11.25 ml	22.5 ml
EDTA 0.5M	20 ul	100 ul	200 ul
Acétate de Potassium 5M	3.2 ml	16 ml	32 ml
H2O UP	Qsp 100 ml	Qsp 500 ml	Qsp 1 000 ml

AMMLAV/E : mélanger 100 ml d'AMMLAV avec 170 ml d'éthanol à 96%

Tampon d'éluion TE1X : Tris 10 mM pH 8, EDTA 1 mM.

Eau UP : eau ultra pure ou eau MilliQ : eau de laboratoire purifiée par le système MilliQ de Millipore. Test de conductivité 18MΩ/cm.

Nettoyage, élimination des déchets

Les tubes plastiques Eppendorf de 2 ml et la plaque DeepWell de binding contenant les résidus d'extraction sont éliminés en poubelle spécifique.

Les solutions contenant du chlorure de guanidium et de l'éthanol sont versées dans le bidon de récupération spécifique prévu à cet usage.

Annexe 2: Protocole détaillé pour la préparation des librairies

	Protocole expérimentale Atelier de marquage moléculaire INRA	PEX-NGS-021 08/08/2018 Page 1 sur 10
	Banque pour Capture LT_1000ng	

Banque pour Capture_LT_1000ng

Historique des versions

Référence : PEX-NGS-021		Gestionnaire : Responsable qualité
Version	Date de la version	Historique des modifications
Version 01	06/07/2017	Création
Version 02	13/09/2017	Révision
Version 03	24/11/2017	Révision
Version 04	08/08/2018	Révision

Sommaire

1. Objectif.....	2
2. Principe.....	2
3. Consignes de sécurité.....	2
4. Matériel utilisé	2
5. Réactifs	2
6. Opérations préalables	2
7. Protocole expérimental.....	2
8. Hygiène et élimination des déchets.....	10
9. Documents associés et bibliographie.....	10
Fiche de visas.....	10

	Protocole expérimentale Atelier de marquage moléculaire INRA	PEX-NGS-021 08/08/2018 Page 2 sur 10
	Banque pour Capture LT_1000ng	

1. Objectif

Protocole permettant de préparer des banques d'ADN pour la capture à l'aide de sonde de différente nature.

Paramètres du protocole : 1 ligation (tags en P5 et MP7) / génotype, LT (1000ng d'ADN, 1 dose de FER, librairies traitées individuellement, normalisation et mélange après PCR PreHyb).

2. Principe

Préparer des banques NGS à Partir d'ADN génomique de bonne qualité afin qu'elles soient compatibles avec la capture avec des sondes.

3. Consignes de sécurité

- ❖ Porter des gants propres et sans poudres
- ❖ Porter une blouse

4. Matériel utilisé

Vortex, centrifugeuse, pipettes...

5. Réactifs

Nom du réactif	Formule chimique	Toxicité	Protection cutanée	Protection oculaire	Protection respiratoire
Azoture de Sodium					

6. Opérations préalables

A- Préparation des adaptateurs et amorces PCR

-Fabrication des Adaptateurs

> préparer une solution :

50mM Tris pH7	500µL d'une solution à 1M
50mM NaCl	200µL d'une solution à 2.5M
	+ 9300µL d'eau mQ RNase Free

Filtrer la solution avec un filtre de 0.22µ



7. Protocole Expérimental

<p>Appareil disponible à l'IRD Voir C. Mariac -</p> <p>Sortir les billes et le HMS 30min avant utilisation + Bien mélanger avant utilisation</p> <p>Programmes dans le dossier : « Ampure et NGS » puis dans : « Mixing »</p> <p>« PurifAmpure »</p> <p>« Mixing »</p> <p>« Transfert »</p> <p>STOP possible à 4°C</p>	<p>1-Dosage et normalisation des ADN</p> <ul style="list-style-type: none">• Doser les ADN au Hoescht• Diluer 1000ng d'ADN génomique dans <u>100µL</u> final d'eau UP <p>2-Fragmentation de l'ADN au Bioruptor/ Diagenode</p> <ul style="list-style-type: none">• Mettre les 1µg d'ADN génomique dilués dans <u>100µL</u> d'eau dans un tube Diagenode de 500µL• Vortexer et centrifuger les tubes et conserver dans la glace• Fragmenter en utilisant le programme : « 300pb » (30sec ON / 90sec OFF) X6• Centrifuger les tubes après fragmentation• Transvaser 95 µL d'ADN fragmenté dans une plaque PCR 96 <p>Purif AMPure XP (1 V) pour éliminer les petits fragments :</p> <p>A LA MAIN :</p> <ul style="list-style-type: none">• Ajouter 1 V de billes : 95µl (25µL de billes et 70µL de HMS) dans les puits contenant 95µl d'ADN• Mélanger par up and down• Incuber la plaque 10min à température ambiante• Placer la plaque sur le support magnétique 2-3min• Enlever et jeter le surnageant• Sans enlever la plaque du support, ajouter 150µl d'éthanol 80% sans perturber les billes• Attendre 30sec et enlever le surnageant• Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol• Laisser la plaque sécher 5 min et enlever la plaque du support magnétique• Ajouter 45 µl d'eau UP mélanger par up and down jusqu'à reprise• Laisser 5min à température ambiante• Placer la plaque sur le support magnétique 2-3min• Transférer 40 µl du surnageant dans une nouvelle plaque <p>AU ROBOT :</p> <ul style="list-style-type: none">• Distribuer manuellement 25µL de billes AMPure puis 70µL de HMS + Pulse• Lancer le programme : « Mixing AMPure » - volume Mixing : 170µL• Incuber la plaque 10min à température ambiante• Faire un Pulse rapide• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Purif AMPure PCR 96 » volume à pipeter : 170µL – volume de départ : 190µL – volume élution : 45µL• Faire un Pulse rapide• Lancer le programme : « Mixing AMPure » - volume Mixing : 35µL• Incuber 5min à TA, puis Faire un Pulse rapide• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Transfert ADN - petit volume -AMPure – 1pl –P50 » volume Transfert : 40µL <p>> Vérifier la fragmentation au BioAnalyser (DNA7500 – dépôt 2 µL) ou au Fragment Analyser (DNF 474 – dépôt 1µL) : 1 à 2 génotype(s) par série de fragmentation</p>
--	--



<p>Sortir buffer à TA - Bien homogénéiser le tampon - TRAVAILLER SUR LA GLACE - Sortir les billes et le HMS 30min avant utilisation + Bien mélanger avant utilisation Programmes dans le dossier : « Ampure et NGS » puis dans : « Mixing » « PurifAmpure » - Sans que le volume touche le film alu - « Transfert »</p>	<p>3- End Repair (Blunt)</p> <ul style="list-style-type: none">• Préparer un mix avec, pour 1 réaction:<ul style="list-style-type: none">○ 5 µl Buffer 10X○ 5 µl d'enzyme <p>Dans les 40µl d'ADN fragmenté/ sizé ;</p> <ul style="list-style-type: none">• Distribuer 10µL du mix (ci-dessus) > volume total : 50µL• Mélanger par up and down et centrifuger (volume total 50µL)• Fermer la plaque avec un film collant en aluminium• Mélanger par up and down OU vortexer DOUCEMENT• Incuber la plaque 30min à 20°C Programme « 30min 20°C » <p>FAIRE IMMEDIATMENT LA PURIFICATION POUR STOPPER LA REACTION</p> <p>Purif AMPure XP (1,3 V) Eliminer enzyme + tampon + petits fragments : A LA MAIN :</p> <ul style="list-style-type: none">• Ajouter 1,3 V de billes : 65 µl (25µL de billes et 40µL de HMS) dans les puits contenant 50 µl d'ADN• Mélanger par up and down• Incuber la plaque 10min à température ambiante• Placer la plaque sur le support magnétique 2-3min• Enlever et jeter le surnageant• Sans enlever la plaque du support, ajouter 150µl d'éthanol 80%• Attendre 30sec et enlever le surnageant• Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol• Laisser la plaque sécher 5 min et enlever la plaque du support magnétique• Ajouter 30µl d'eau UP mélanger par up and down jusqu'à reprise• Laisser 5min à température ambiante• Placer la plaque sur le support magnétique 2-3min• Prendre une nouvelle plaque• Transférer 10 µL de surnageant dans une nouvelle plaque <p>AU ROBOT :</p> <ul style="list-style-type: none">• Distribuer manuellement : 25µL de billes AMPure puis 40µL de HMS + Pulse• Lancer le programme : « Mixing AMPure » - volume Mixing : 100µL• Incuber la plaque 10min à température ambiante• Faire un Pulse rapide• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Purif AMPure PCR 96 » volume à pipeter : 100µL – volume de départ : 115µL – volume élution : 30µL• Vortexer DOUCEMENT (car billes au-dessus du niveau de l'eau) + Pulse• Incuber 5min à TA• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Transfert ADN - petit volume -AMPure – 1pl –P50 » Volume Transfert : 10µL <p>>Contrôler la concentration de <u>quelques</u> génotypes au SPARK (sur volume restant de la plaque source) – concentration attendue entre 15 et 35 ng/µL REALISER LA LIGATION DES ADAPTATEURS EN SUIVANT</p>
---	---

<p>Sortir buffer à TA -</p> <p>TRAVAILLER SUR LA GLACE</p> <p>1 adaptateur ≠ par génotype</p> <p>Sortir les billes et le HMS 30min avant utilisation + Bien mélanger avant utilisation</p>	<p>4- Ligation Blunt des adaptateurs</p> <p><i>Les TAG sont placés des deux côtés des fragments d'ADN, sur l'adaptateur tronqué P5 ET sur le MP7</i></p> <ul style="list-style-type: none"> • Préparer un mix avec, pour 1 réaction: <ul style="list-style-type: none"> ○ 1.5 µL d'H2O UP ○ 4 µl Buffer ligase 5X ○ 0.5 µl T4 DNA Ligase <p>Dans les 10µL d'ADN :</p> <ul style="list-style-type: none"> • Ajouter 2µL ADAPT BARCODED P5 à 4µM (≠ pour chaque génotype) • Ajouter 2µL ADAPT BARCODED MPE-P7 à 4µM (≠ pour chaque génotype) • Distribuer 6µL du mix (ci-dessus) > volume total : 20µL • Mélanger par up and down OU vortexer DOUCEMENT • Sceller la plaque à la thermo-scelleuse • Incuber la plaque 1H à 22°C 10min à 65°C Hold 4°C <p>Programme « Ligation »</p> <p style="text-align: center; color: red;">STOP POSSIBLE A 4°C</p> <p>Purif AMPure XP (1.6V): Eliminer enzyme + tampon A LA MAIN :</p> <ul style="list-style-type: none"> • Ajouter 1.6V de billes : 32µl (20µL de billes et 12µL de HMS) dans les puits contenant les 20µl de ligation • Mélanger par up and down • Incuber le tube 10min à température ambiante • Placer le tube sur le support magnétique 2-3min • Enlever et jeter le surnageant • Sans enlever le tube du support, ajouter 150µL d'éthanol 80% • Attendre 30sec et enlever le surnageant • Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol • Laisser le tube sécher 5 min et enlever le tube du support magnétique • Ajouter 27 µl d'eau UP mélanger par up and down jusqu'à reprise • Laisser 5min à température ambiante • Placer le tube sur le support magnétique 2-3min • Transférer 24 µl de surnageant dans une nouvelle plaque
--	--



<p>« AMPure et NGS » / « Mixing » - « PurifAMPure » - Sans que le volume touche le film alu - « Transfert »</p> <p>Sortir buffer + dNTP à TA</p> <p>TRAVAILLER SUR LA GLACE</p> <p>Sortir les billes et le HMS 30min avant utilisation + Bien mélanger avant utilisation</p>	<p>AU ROBOT :</p> <ul style="list-style-type: none">• Distribuer manuellement : 20µL de billes AMPure puis 12µL de HMS + Pulse• Lancer le programme : « Mixing AMPure » - volume Mixing : 40µL• Incuber la plaque 10min à température ambiante• Faire un Pulse rapide• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Purif AMPure PCR 96 » volume à pipeter : 40µL – volume de départ : 52µL – volume élution : 27µL• Vortexer DOUCEMENT (car billes au-dessus du niveau de l'eau)• Faire un Pulse rapide• Incuber 5min à TA• Placer la plaque sur le support magnétique 2-3min• Lancer le programme : « Transfert ADN - petit volume -AMPure – 1pl –P50 » Volume Transfert : 24µL <p>Pas de dosage possible après cette étape (molécules non linéaires)</p> <p>STOP POSSIBLE A 4°C</p> <p>5- « Overhang » des adaptateurs par Bst</p> <ul style="list-style-type: none">• Préparer un mix avec, pour 1 réaction :<ul style="list-style-type: none">○ 3 µl Buffer Bst 10X○ 1 µl dNTP (10mM)○ 2 µl Enzyme Bst (8U/µL) <p>Dans les puits contenant les 24µL de ligation purifiée :</p> <ul style="list-style-type: none">• Distribuer 6µL du mix (ci-dessus) > volume total : 30µL• Mélanger par up and down OU vortexer DOUCEMENT• Fermer la plaque avec un film collant en aluminium• Incuber pendant 15min à 37°C Hold 4°C <p>Programme « Bst »</p> <p>GARDER DANS LA GLACE</p> <p>FAIRE IMMEDIATMENT LA PURIFICATION POUR STOPPER LA REACTION</p> <p>Purif AMPure XP (1.6V) : Eliminer enzyme + tampon</p> <p>A LA MAIN :</p> <ul style="list-style-type: none">• Ajouter 1.6V de billes : 50 µl (25µL de billes et 25µL de HMS) dans les puits contenant les 30µl de ligation/Bst• Mélanger par up and down• Incuber le tube 10min à température ambiante• Placer le tube sur le support magnétique 2-3min• Enlever et jeter le surnageant• Sans enlever le tube du support, ajouter 150µl d'éthanol 80%• Attendre 30sec et enlever le surnageant• Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol• Laisser le tube sécher 5 min et enlever le tube du support magnétique
--	--



« Ampure et
NGS » /
« Mixing »
-
« PurifAmpure »
-
Sans que le
volume touche le
film alu
-
« Transfert »

- Ajouter 22µl d'H₂O UP, mélanger par up and down jusqu'à reprise
- Laisser **5min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**
- Transférer 19µl du surnageant dans une nouvelle plaque

AU ROBOT :

- Distribuer manuellement : 25µL de billes AMPure puis 25µL de HMS + Pulse
- Lancer le programme : « Mixing AMPure » - volume Mixing : 65µL
- Incuber la plaque **10min** à température ambiante
- Faire un Pulse rapide
- Placer la plaque sur le support magnétique **2-3min**
- Lancer le programme : « Purif AMPure PCR 96 »
volume à pipeter : 65µL – volume de départ : 80µL – volume élution : 22µL
- Vortexer DOUCEMENT (car billes au-dessus du niveau de l'eau)
- Faire un Pulse rapide
- Incuber **5min** à TA
- Placer la plaque sur le support magnétique **2-3min**
- Lancer le programme : « Transfert ADN - petit volume -AMPure – 1pl –P50 »
Volume Transfert : 19µL

>Contrôler la concentration de quelques génotypes au SPARK (sur volume restant de la plaque source) – concentration attendue entre 5 et 15 ng/µL

STOP POSSIBLE A 4°C ou -20°C

6- PCR enrichissement-Préhybridation

Sortir amorces +
kit KAPA à TA

TRAVAILLER SUR
LA GLACE

- Préparer un mix, avec pour une réaction :
 - 0.5 µl Amorce PreHybrid PE-F à 10µM
 - 0.5 µl Amorce PreHybrid MPE-R à 10µM
 - 20 µl 2X KAPA HiFi HotStart Ready mix

Dans les puits contenant les 19 µL de Bst purifié :

- Distribuer 21 µL du mix (ci-dessus) > volume total : 40µL
- Mélanger par up and down OU vortexer DOUCEMENT
- Sceller la plaque à la thermo-scelleuse
- Placer la plaque dans le thermocycler
Programme « PREHYB12 »

98°C 2min
98°C 20sec
55°C 45sec **12 cycles**
72°C 30sec
72°C 10min
4°C hold

STOP POSSIBLE A 4°C ou -20°C



*Sortir les billes et
le HMS 30min
avant utilisation
+
Bien mélanger
avant utilisation*

*Programmes
dans le dossier :
« AMPure et
NGS » puis dans :
« Mixing »*

*« PurifAMPure »
-
Sans que le
volume touche le
film alu
-
« Transfert »*

Purif AMPure XP (1.3V) : Eliminer enzyme + tampon + petits fragments

A LA MAIN :

- Vortexer les billes AMPure
- Ajouter 1.3V de billes : 52 µl (22µL de billes et 30µL de HMS) dans les puits contenant les 40µl de PCR
- Mélanger par up and down
- Incuber le tube **10min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**
- Enlever et jeter le surnageant
- Sans enlever le tube du support, ajouter 150µl d'éthanol 80%
- Attendre **30sec** et enlever le surnageant
- Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol
- Laisser le tube sécher **5min** et enlever le tube du support magnétique
- Ajouter 30µl d'eau UP, mélanger par up and down jusqu'à reprise
- Laisser **5min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**
- Transférer 27µl du surnageant dans une nouvelle plaque

AU ROBOT :

- Distribuer manuellement : 22µL de billes AMPure puis 30µL de HMS + Pulse
- Lancer le programme : « Mixing AMPure » - volume Mixing : 72µL
- Incuber la plaque **10min** à température ambiante
- Faire un Pulse rapide
- Placer la plaque sur le support magnétique **2-3min**
- Lancer le programme : « Purif AMPure PCR 96 »
volume à pipeter : 72µL – volume de départ : 92µL – volume élution : 30µL
- Vortexer DOUCEMENT (car billes au-dessus du niveau de l'eau)
- Faire un Pulse rapide
- Incuber **5min** à TA
- Placer la plaque sur le support magnétique **2-3min**
- Lancer le programme : « Transfert ADN - petit volume - AMPure – 1pl –P50 »
Volume Transfert : 27µL

>Contrôler la concentration de quelques génotypes au SPARK (sur volume restant de la plaque source) – concentration attendue entre 15 et 35 ng/µL

STOP POSSIBLE A 4°C ou -20°C

7- Dosage

- Préparer une dilution 1/3 : 2µL de banques + 4 µL d'eau UP
- Déposer 2 µL de la dilution 1/3 sur fragment Analyser avec le kit DNF 474

8- Pooling des librairies

Le logiciel « PROsize 2.0 » associé au Fragment Analyser permet d'estimer la quantité de chaque banque en nM en tenant compte de la taille des fragments. Pour optimiser le mélange des banques en EQUI-PROPORTION, il est préférable d'utiliser la quantité en nM de chaque banque dans l'intervalle de 100 à 600pb. Le mélange est réalisé dans un tube de 1,5mL.



*Sortir les billes et
le HMS 30min
avant utilisation
+
Bien mélanger
avant utilisation*

Purif AMPure XP (1.6V) : réduire le volume

A LA MAIN :

- Vortexer les billes AMPure
- Ajouter 1.6V de billes (40-60µL de billes et HMS qsp 1.6V)
- Mélanger par up and down
- Incuber le tube **10min** à température ambiante
- Placer le tube sur le support magnétique **5min**
- Enlever et jeter le surnageant
- Sans enlever le tube du support, ajouter 500µl d'éthanol 80% sans perturber les billes
- Attendre **30sec** et enlever le surnageant
- Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol
- Laisser le tube sécher **5min** et enlever le tube du support magnétique
- Ajouter 30µl d'eau UP, mélanger par up and down jusqu'à reprise
- Laisser **5min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**
- Transférer 30µl du surnageant dans un tube à vis « DNA Bank »
- Numéroté le tube avec la notation :

Ex : DEV_Pol001.1 ← N° de l'expérimentation
 ↑ ↑
 3 2

3 grandes lettres pour définir le projet N° du mélange

9- Dosage

- Doser le mélange au SPARK pour obtenir une concentration en ng/µL

>Entre **500 et 1000ng** du mélange de banques sont nécessaire pour la capture

8. Hygiène et élimination des déchets

Rappeler les règles d'hygiène se rapportant à l'équipement ou à l'environnement de travail ainsi que les règles d'élimination des déchets.

9. Documents associés et bibliographie

Fiche de visas

	Rédacteur	Testeur	Vérificateur	Approbateur
Nom	M. ARDISSON			
Fonction	TR			
Visa				
Date	08/08/18			

Annexe 3: Protocole détaillé pour l'enrichissement par capture

	Protocole expérimental Atelier de marquage moléculaire INRA	PEX-NGS-019 08/08/2018 Page 1 sur 10
	Double Capture Sonde MYBAITS/ROCHEx2	

***Double Capture de banque d'ADN
par des sondes MYBAITS / protocole Roche x2***

Historique des versions

Référence : PEX-NGS-019		Gestionnaire : Responsable qualité
Version	Date de la version	Historique des modifications
Version 01	21/12/2016	Création
Version 02	08/08/2018	Révision

Sommaire

1. Objectif	2
2. Principe	2
3. Consignes de sécurité	2
4. Matériel utilisé	2
5. Réactifs	2
6. Opérations préalables	2
7. Protocole Expérimental	3
8. Hygiène et élimination des déchets	10
9. Documents associés et bibliographie	10

	Protocole expérimental Atelier de marquage moléculaire INRA	PEX-NGS-019 08/08/2018 Page 2 sur 10
	Double Capture Sonde MYBAITS/ROCHEx2	

1. Objectif

Protocole permettant d'effectuer une double capture des banques d'ADN à l'aide de sonde produites par MYBAITS et utilisant le protocole de capture Roche puis MYBAITS. Cette double capture permet d'augmenter la spécificité des fragments capturés.

2. Principe

La capture de fragments d'ADN à partir de sondes de 82 ou 120pb (MYBAITS) permet de d'étudier plus spécifiquement certaines parties du génome (SNP, zone d'intérêt...).

3. Consignes de sécurité

- ❖ Porter des gants propres et sans poudre
- ❖ Porter une blouse

4. Matériel utilisé

Vortex, centrifugeuse, pipettes, SpeedVac...

5. Réactifs

Nom du réactif	Formule chimique	Toxicité	Protection cutanée	Protection oculaire	Protection respiratoire
Azoture de Sodium	NaN ₃	T+ très toxique	X		
Formamide	CH ₃ NO	CMR	X		X

6. Opérations préalables

A- Préparation des bloquants d'adaptateurs (si besoin, vérifier les stocks)

Préparer les bloquants d'adaptateurs P5/MP7 pour avoir un mélange à 50µmol/L each >Hybridation1
 20 µL "Multiplex-block-P7" à 100 µM (stock MWG) + 20µL "Univ-block-P5" à 100 µM (stock MWG)
 Préparer une dilution de ces bloquants d'adaptateurs P5/MP7 au 1/5^{ème} > Hybridation 2
 4µL de la solution bloquants d'adaptateurs P5/MP7 à 50µmol/L each + 16 µL d'eau UP

B- Préparation de Bloquants microsatellites (si besoin, vérifier les stocks)

Mélanger les 3 bloquants microsatellites pour avoir un mélange à 33 µmol/L each > Hybridation 1
 20 µl de chaque bloquants microsats : Block(GAA)7, Block(CAA)7, Block(GGA)7 à 100 µM (MWG)
 Préparer une dilution de ces bloquants microsatellites au 1/10^{ème} > Hybridation 2
 2µL de la solution bloquants microsatellites à 33 µmol/L each + 18 µL d'eau UP

C- Dilution du « Sequence Capture Developer Reagent » (si besoin)

Diluer au ½ le « Sequence Capture Developer Reagent » (produit Roche) > Hybridation 2
 20µL de « Sequence Capture Developer Reagent » + 20 µL d'eau UP

D- Préparation du tampon d'éluion (si besoin, vérifier les stocks)

Préparer une solution d'éluion : 10mM Tris HCl, 0.05% Tween 20

7. Protocole Expérimental

JOUR 1	1-Hybridation 1
Commencer en début d'après-midi	Sortir dans la glace les bloquants stockés à -20 °C : -« Sequence Capture Developer Reagent » (tube vert) ROCHE - BLOCK Microsatellites : 33 µmol/L each (optionnel) - BLOCK adaptateurs tronqués « home made » : P5/MP7 à 50µmol/L each
Préparer à TA	<ul style="list-style-type: none"> Percer le capuchon d'un tube de 1.5 ml avec une aiguille stérile(7trous) <i>Attention numéroter le tube sur le bouchon et sur la tranche du tube</i> Ajouter dans le tube (capuchon percé) : <ul style="list-style-type: none"> 10 µl de « Sequence Capture Developer Reagent » (vert) 500ng à 1000ng du mélange de librairies (volume variable) 2 µl de BLOCK adaptateurs tronqués P5/MP7 à 50µmol/Leach 2 µl de BLOCK Microsat : 33 µmol/L each Lyophiliser le contenu du tube au SpeedVac – DNA 120 20 minutes à 43°C (position « MEDIUM ») pour 20-25 µL
Sous la hotte	<p>-->Pendant la lyophilisation:</p> <ul style="list-style-type: none"> Préchauffer le bain marie à sec à 95°C. Sortir dans la glace les tubes 5 et 6 du kit Capture Roche <p>--<</p>
Jeter les déchets dans un sachet > BET	<ul style="list-style-type: none"> Récupérer le tube de 1.5mL avec le culot de lyophilisat Couper le capuchon percé du tube et le remplacer par un nouveau capuchon
Ne pas faire de UP and DOWN	<ul style="list-style-type: none"> Ajouter au lyophilisat : <ul style="list-style-type: none"> 7.5 µL « 2 X Sequence Capture Hybridization Buffer » (tube 5) 3 µL « Hybridization Component A » (tube6) Vortexer 10 secondes puis faire un pulse Dénaturer le mélange 10 minutes à 95°C dans le bain marie à sec.
Placer un des deux blocs sur les tubes > couvercle chauffant	<p>-->Pendant la dénaturation :</p> <ul style="list-style-type: none"> Sortir les BAIITS du congélateur -80°C et les placer dans la glace Mettre dans un tube PCR de 0.2 ml : <ul style="list-style-type: none"> 3.5 µl de BAIITS (78% d'une dose classique de 4.5µL) 1 µL d'eau UP <p>--<</p> <ul style="list-style-type: none"> Faire un pulse du tube dénaturé, puis transférer la totalité du volume dans le tube de 0.2mL contenant les 4.5 µl de BAIITS/eau UP. Mélanger 10 fois par pipetage UP and DOWN et faire un pulse. Incuber le ou les tubes pendant 64 à 72 H à 47 °C Couvercle chauffant à 57°C Programme « HYB-ROCH »



Double Capture
Sonde MYBAITS/ROCHEx2

<p>Toujours sur le programme « HYB-ROCH »</p> <p>Les billes adhèrent aux cônes et aux tubes</p> <p>Wash buffer Préchauffées à 47°C</p> <p>Avec le même cône</p>	<p>4-Immobilisation : Biotine/Streptavidine</p> <ul style="list-style-type: none">Faire un pulse rapide des tubes contenant les hybridationsAjouter les 15 µl d'hybridation au tube contenant les billes streptavidine C1Mélanger par pipetage UP and DOWNIncuber les tubes 45 minutes à 47°C <p>Attention : Mélanger les billes toutes les 15 minutes par pipetage UP and DOWN</p> <p>-->Pendant l'immobilisation :</p> <ul style="list-style-type: none">Allumer le bain marie à sec à 47 °CPréchauffer les solutions Wash buffer 1, 2, 3 et 4 diluées au 1 X au bain marie à sec à 47°C pendant les 45min d'immobilisationPréparer 1 tube numéroté de 1.5mL/capt. + 1 tube numéroté de 0.2mL/capt. <p>--<</p> <ul style="list-style-type: none">Ajouter 100 µl de « 1X Wash Buffer 1 » préchauffé dans le tube contenant les hybridations/billesMélanger par UP and DOWN et transférer la totalité avec le même cône dans un tube de 1.5 mLPlacer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageant (ou conserver dans un tube de 1.5mL)Enlever le tube du support magnétique <p>5-Lavages</p> <ul style="list-style-type: none">Ajouter 200 µl de « 1 X Stringent Wash Buffer » préchaufféMélanger par pipetage et incuber 5 minutes dans le <u>bain marie à 47°C</u>Placer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageantEnlever le tube du support magnétiqueAjouter 200 µl de « 1 X Stringent Wash Buffer » préchaufféMélanger par pipetage et incuber 5 minutes dans le <u>bain marie à 47°C</u>Placer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageantAjouter 200 µl de « 1X Wash Buffer 1 » préchauffé aux billesVortexer <u>2 minutes</u> et faire un pulse très rapidePlacer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageantEnlever le tube du support magnétiqueAjouter 200 µl de « 1X Wash Buffer 2 » préchaufféVortexer <u>1 minute</u> et faire un pulse très rapidePlacer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageantEnlever le tube du support magnétiqueAjouter 200 µl de « 1X Wash Buffer 3 » préchaufféVortexer <u>30 secondes</u> et faire un pulse très rapidePlacer le tube sur le support magnétique 1 minuteEnlever et jeter de surnageantEnlever le tube du support magnétiqueResuspendre dans 25µl de tampon d'éluionTransférer dans un tube de 0.2mL
---	---



<p><i>Etapes à réaliser rapidement pour éviter la renaturation</i></p> <p><i>numéroter sur le bouchon et sur la tranche du tube</i></p> <p>Sous la hotte</p> <p><i>Jeter les déchets dans un sachet > BET</i></p> <p><i>Ne pas faire de UP and DOWN</i></p> <p><i>Placer un des deux blocs sur les tubes > couvercle chauffant</i></p>	<p>6-Découplage / Déhybridation :</p> <ul style="list-style-type: none">• Dénaturer les 25µL d'éluion <u>3 min à 95°C</u>• Centrifuger RAPIDEMENT le tube• Placer le tube de 0.2mL support une plaque magnétique• Transférer, RAPIDEMENT, 21µL de surnageant dans un tube de 1.5mL <p>Dans le tube de 0.2mL avec les billes et les 3-5µl de l'Hybridation 1 + billes :</p> <ul style="list-style-type: none">• Ajouter eau UP qsp 22µL• Conserver à 4°C <p>7-Hybridation 2</p> <p>Sortir dans la glace les bloquants stockés à -20 °C :</p> <ul style="list-style-type: none">- « Sequence Capture Developer Reagent » (ROCHE) <u>dilué au 1/2</u>- BLOCK adaptateurs tronqués <i>P5/MP7</i> à 50µmol/Leach <u>dilué au 1/5^{ème}</u>- BLOCK Microsatellites à 33 µmol/L each <u>dilué au 1/10^{ème}</u> <ul style="list-style-type: none">• Percer le capuchon du tube de 1.5 ml contenant les 21 µL de HYB1 avec une aiguille stérile(7trous) <i>numéroter sur le bouchon + la tranche du tube</i>• Ajouter dans le tube (capuchon percé) :<ul style="list-style-type: none">○ 2 µl de « Sequence Capture Developer Reagent » <u>dilué au 1/2</u>○ 2 µl BLOCK adapt. P5/MP7 à 50µmol/L each <u>dilué au 1/5^{ème}</u>○ 2 µl de BLOCK Microsat à 33 µmol/L each <u>dilué au 1/10^{ème}</u>• Lyophiliser le contenu du tube au SpeedVac – DNA 120 20 minutes à 43°C (position « MEDIUM ») <p>-->Pendant la lyophilisation:</p> <ul style="list-style-type: none">• Préchauffer le bain marie à sec à 95°C.• Sortir dans la glace les tubes 5 et 6 du kit Capture Roche <p>--<</p> <p>Jeter les déchets dans un sachet > BET</p> <ul style="list-style-type: none">• Récupérer le tube de 1.5mL avec le culot de lyophilisat• Couper le capuchon percé du tube et le remplacer par un nouveau capuchon• Ajouter au lyophilisat :<ul style="list-style-type: none">○ 7.5 µL « 2 X Sequence Capture Hybridization Buffer » (tube 5)○ 3 µL « Hybridization Component A » (tube6)• Vortexer 10 secondes puis faire un pulse• Dénaturer le mélange 10 minutes à 95°C dans le bain marie à sec. <p>-->Pendant la dénaturation :</p> <ul style="list-style-type: none">• Sortir les BAITS du congélateur -80°C et les placer dans la glace• Mettre dans un tube PCR de 0.2 ml :<ul style="list-style-type: none">○ 1 µl de BAITS (22% d'une dose classique de 4.5µL)○ 3.5 µL d'eau UP <p>--<</p>
---	---



<p><i>Eteindre le programme puis relancer car >99h</i></p>	<ul style="list-style-type: none">• Faire un pulse du tube dénaturé, puis transférer la totalité du volume dans le tube de 0.2mL contenant les 4.5 µl de BAITS/eau UP.• Mélanger 10 fois par pipetage UP and DOWN et faire un pulse.• Incuber le ou les tubes pendant toute la nuit à 47 °C (17 à 20h) Couvercle chauffant à 57°C Programme « HYB-ROCH »										
<p>JOUR 3</p> <p><i>commercer vers 8h30-9h</i></p>	<p>ATTENTION : Pour les étapes 2 à 5 (immobilisation et lavages) >Traiter au maximum deux captures à la fois</p> <p>Ces étapes sont donc à renouvelées deux fois s'il y a 4 captures à traiter, etc... Pendant le traitement de l'ensemble des captures, les tubes pas encore traités sont gardés à 47°C et les tubes déjà traités sont conservés à 4°C dans le tampon d'éluion.</p> <p>8-Préparation des solutions de lavage</p> <ul style="list-style-type: none">• Sortir à température ambiante :• Les solutions de lavages: Wash Buffer 1, 2, 3, 4 et 7 Kit: SeqCap EZ Hybridization and Wash Kit (stocké à -20°C)• Les billes : DYNABEADS - STREPTAVIDIN – C1 (stockés à 4°C) <ul style="list-style-type: none">• Préparer les dilutions des solutions de lavages dans des tubes de 1.5mL: Volumes pour 1 hybridation: <table border="0"><tr><td>• 10 X Wash Buffer 1 (tube 1)</td><td>35 µL + 315 µl d'eau UP</td></tr><tr><td>• 10 X Wash Buffer 2 (tube 2)</td><td>25 µL + 225 µl d'eau UP</td></tr><tr><td>• 10 X Wash Buffer 3 (tube 3)</td><td>25 µL + 225 µl d'eau UP</td></tr><tr><td>• 10 X Strigent Wash Buffer (tube 4)</td><td>45 µL + 405 µl d'eau UP</td></tr><tr><td>• 2.5 X Bead Wash Buffer (tube 7)</td><td>120 µL + 180 µl d'eau UP</td></tr></table>	• 10 X Wash Buffer 1 (tube 1)	35 µL + 315 µl d'eau UP	• 10 X Wash Buffer 2 (tube 2)	25 µL + 225 µl d'eau UP	• 10 X Wash Buffer 3 (tube 3)	25 µL + 225 µl d'eau UP	• 10 X Strigent Wash Buffer (tube 4)	45 µL + 405 µl d'eau UP	• 2.5 X Bead Wash Buffer (tube 7)	120 µL + 180 µl d'eau UP
• 10 X Wash Buffer 1 (tube 1)	35 µL + 315 µl d'eau UP										
• 10 X Wash Buffer 2 (tube 2)	25 µL + 225 µl d'eau UP										
• 10 X Wash Buffer 3 (tube 3)	25 µL + 225 µl d'eau UP										
• 10 X Strigent Wash Buffer (tube 4)	45 µL + 405 µl d'eau UP										
• 2.5 X Bead Wash Buffer (tube 7)	120 µL + 180 µl d'eau UP										
<p><i>Garder les solutions à TA</i></p> <p><i>Utiliser le portoir magnétique à tube</i></p> <p><i>Les billes adhèrent aux cônes et aux tubes</i></p>	<p>9-Préparation des billes streptavidine type C1</p> <p>Pour une hybridation :</p> <ul style="list-style-type: none">• Bien mélanger les billes et mettre 50 µL de billes dans un tube de 1.5 mL• Placer le tube sur le support magnétique pendant 2min• Enlever et jeter de surnageant.• Enlever le tube du support magnétique• Ajouter 100 µL de «1 X Bead Wash Buffer »• Mélanger par pipetage UP and DOWN• Placer le tube sur le support magnétique 2 minutes• Enlever et jeter de surnageant.• Enlever le tube du support magnétique• Ajouter 100 µL de «1 X Bead Wash Buffer »• Mélanger par pipetage UP and DOWN• Placer le tube sur la plaque magnétique 2 minutes• Enlever et jeter le surnageant.• Enlever le tube du support magnétique• Ajouter <u>50 µL</u> de «1 X Bead Wash Buffer » (1V)• Mélanger par pipetage UP and DOWN• Transférer la totalité avec le même cône dans un tube de 0.2 mL										

Toujours sur le
programme
« HYB-ROCH »

- Placer le tube sur la plaque magnétique 2 minutes
- Enlever et jeter le surnageant.
- Faire un pulse et replacer sur la plaque magnétique pour bien enlever tout le liquide
- Laisser sécher 2 ou 3 minutes à TA

10-Immobilisation : Biotine/Streptavidine

- Faire un pulse rapide des tubes contenant les hybridations
- Ajouter les 15 µl d'hybridation au tube contenant les billes streptavidine C1
- Mélanger par pipetage UP and DOWN
- Incuber les tubes 45 minutes à 47°C

Attention : Mélanger les billes toutes les 15 minutes par pipetage UP and DOWN.

-->Pendant l'immobilisation :

- Allumer le bain marie à sec à 47 °C
- Préchauffer les solutions Wash buffer 1, 2, 3 et 4 diluées au 1 X au bain marie à sec à 47°C pendant les 45min d'immobilisation
- Préparer 1 tube numéroté de 1.5mL/capt. + 1 tube numéroté de 0.2mL/capt.

--<

- Ajouter 100 µl de « **1X Wash Buffer 1** » préchauffé dans le tube contenant les hybridations/billes
- Mélanger par UP and DOWN et transférer la totalité avec le même cône dans un tube de 1.5 mL
- Placer le tube sur le support magnétique 1 minute
- Enlever et jeter de surnageant (ou conserver dans un tube de 1.5mL)
- Enlever le tube du support magnétique

11-Lavages

- Ajouter 200 µl de « **1 X Stringent Wash Buffer** » préchauffé
- Mélanger par pipetage et incuber 5 minutes dans le bain marie à 47°C
- Placer le tube sur le support magnétique 1 minute
- Enlever et jeter de surnageant
- Enlever le tube du support magnétique
- Ajouter 200 µl de « **1 X Stringent Wash Buffer** » préchauffé
- Mélanger par pipetage et incuber 5 minutes dans le bain marie à 47°C
- Placer le tube sur le support magnétique 1 minute
- Enlever et jeter de surnageant
- Ajouter 200 µl de « **1X Wash Buffer 1** » préchauffé aux billes
- Vortexer 2 minutes et faire un pulse très rapide
- Placer le tube sur le support magnétique 1 minute
- Enlever et jeter de surnageant
- Enlever le tube du support magnétique
- Ajouter 200 µl de « **1X Wash Buffer 2** » préchauffé
- Vortexer 1 minute et faire un pulse très rapide
- Placer le tube sur le support magnétique 1 minute
- Enlever et jeter de surnageant
- Enlever le tube du support magnétique



- Ajouter 200 µl de « **1X Wash Buffer 3** » préchauffé
- Vortexer 30 secondes et faire un pulse très rapide.
- Placer le tube sur le support magnétique 1 minute.
- Enlever et jeter de surnageant.
- Enlever le tube du support magnétique.
- Resuspendre dans 22µL d'eau UP
- Transférer les 22µL avec les billes dans un tube de 0.2mL

12- Découplage et Amplification PCR « ON BEADS »

Réaliser la PCR sur les 22µL de HYB1 et les 22µL de HYB2 pour chaque capture
Attention : utiliser des index différents

Dans les 22µL de Capture:

- Distribuer 1.5 µl d'amorce Sol PE-PCR F à 10µM
- Distribuer 1.5 µl d'amorce Sol MPE-PCR R (avec Index) à 10µM
- Distribuer 25µl 2X KAPA HiFi HotStart Ready mix
- Mélanger par up and down (volume total 50µL)
- Lancer le programme « PCR TOT-B »

98°C 2min
98°C 20sec
62°C 30sec **18 cycles**
72°C 30sec
72°C 5min
4°C Hold

- Placer le tube sur le support magnétique **1min**
- Transférer le surnageant dans un tube de 1.5mL

*Sortir les billes
et le HMS
30min avant
utilisation
+
Bien mélanger
avant
utilisation*

Purif AMPure XP (1.8V): Eliminer enzyme + tampon

- Ajouter 1.8V de billes (30µL de billes puis 60 µL de HMS) dans les puits contenant les 50µl de librairies capturées et amplifiées
- Mélanger par up and down
- Incuber le tube **10min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**
- Enlever et jeter le surnageant
- Sans enlever le tube du support, ajouter 150µl d'éthanol 80% sans perturber les billes
- Attendre **30sec** et enlever le surnageant
- Répéter le lavage à l'éthanol 80%, Bien enlever tout l'éthanol
- Laisser le tube sécher **5min** et enlever le tube du support magnétique
- Ajouter 22µl de l'H2O UP mélanger par up and down jusqu'à reprise
- Laisser **5min** à température ambiante
- Placer le tube sur le support magnétique **2-3min**



- Transférer 20µl du surnageant dans un tube à vis « DNA Bank »
- Numéroté le tube avec la notation :

N° de l'expérimentation

Ex : **DEV_Cap001.1** HYB1 ou HYB2 ← N° de l'hybridation

3 grandes lettres N° de la capture
Nom du projet

L'hybridation 1 est seulement un contrôle expérimental,
c'est l'hybridation 2 qui sera séquencée

Sortir le kit
DNA7500 à TA
30min avant

13- Dosage

Déposer 1µL de chacune des deux hybridations par capture sur une puce Agilent DNA 7500 ou sur le Fragment analyser, kit DNF474.

14- pooling des captures (si besoin)

L'estimation la quantité de chaque Capture en nM en tenant compte de la taille des fragments, permet d'optimiser le mélange des captures en EQUI-PROPORTION.

1. Hygiène et élimination des déchets

Rappeler les règles d'hygiène se rapportant à l'équipement ou à l'environnement de travail ainsi que les règles d'élimination des déchets

2. Documents associés et bibliographie

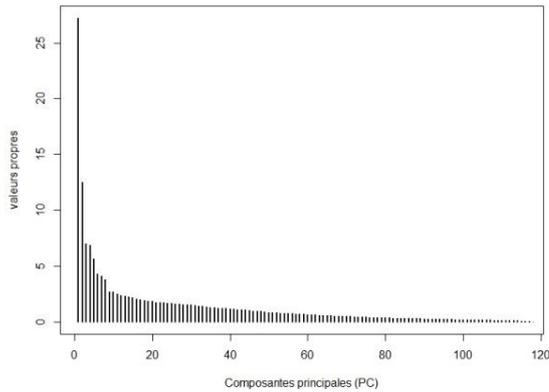
Fiche de visas

	Rédacteur	Testeur	Vérificateur	Approbateur
Nom	M. ARDISSON			
Fonction	TR			
Visa				
Date	08/08/2018			

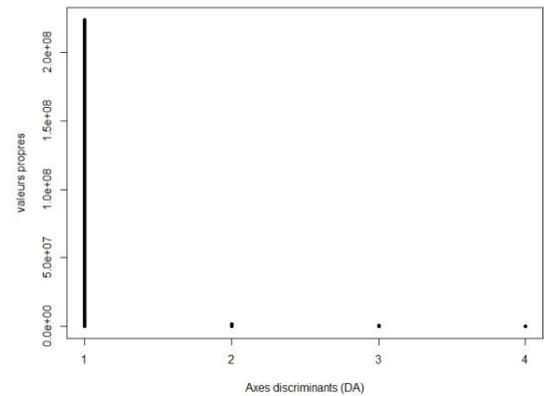
Annexe 4: Analyse de structure par ACP : tableau des valeurs propres et pourcentage d'inertie cumulée, pour les 10 premiers axes de l'ACP.

	valeur propre	% inertie cumulée
Axe 1	94,33	6,19
Axe 2	74,57	11,09
Axe 3	69,30	15,64
Axe 4	55,10	19,26
Axe 5	52,69	22,72
Axe 6	50,69	26,05
Axe 7	47,17	29,14
Axe 8	43,01	31,97
Axe 9	39,35	34,55
Axe 10	37,81	37,03

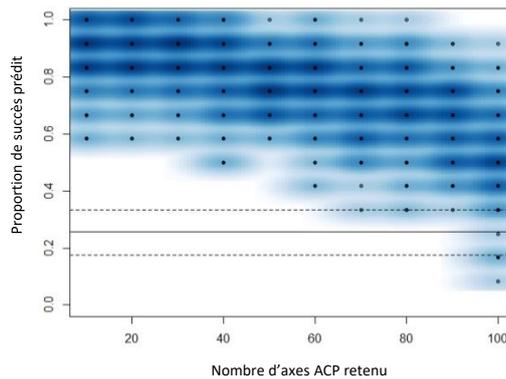
Annexe 5 : Analyse de structure par DAPC sur les 120 génotypes : graphiques complémentaires



Histogramme des valeurs propres des composantes principales du modèle généraliste avec les 120 génotypes (DD, DC, DP, DE) – DAPC.

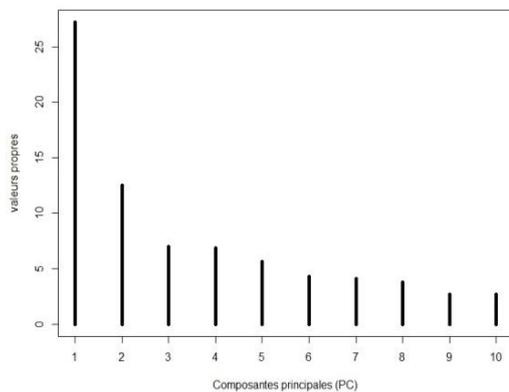


Histogramme des valeurs propres des axes discriminants du modèle généraliste avec les 120 génotypes (DD, DC, DP, DE) – DAPC.

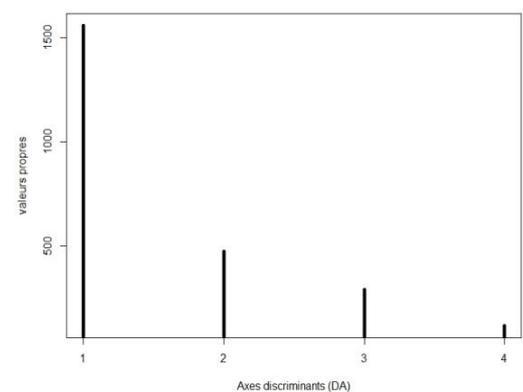


Validation croisée – MSE

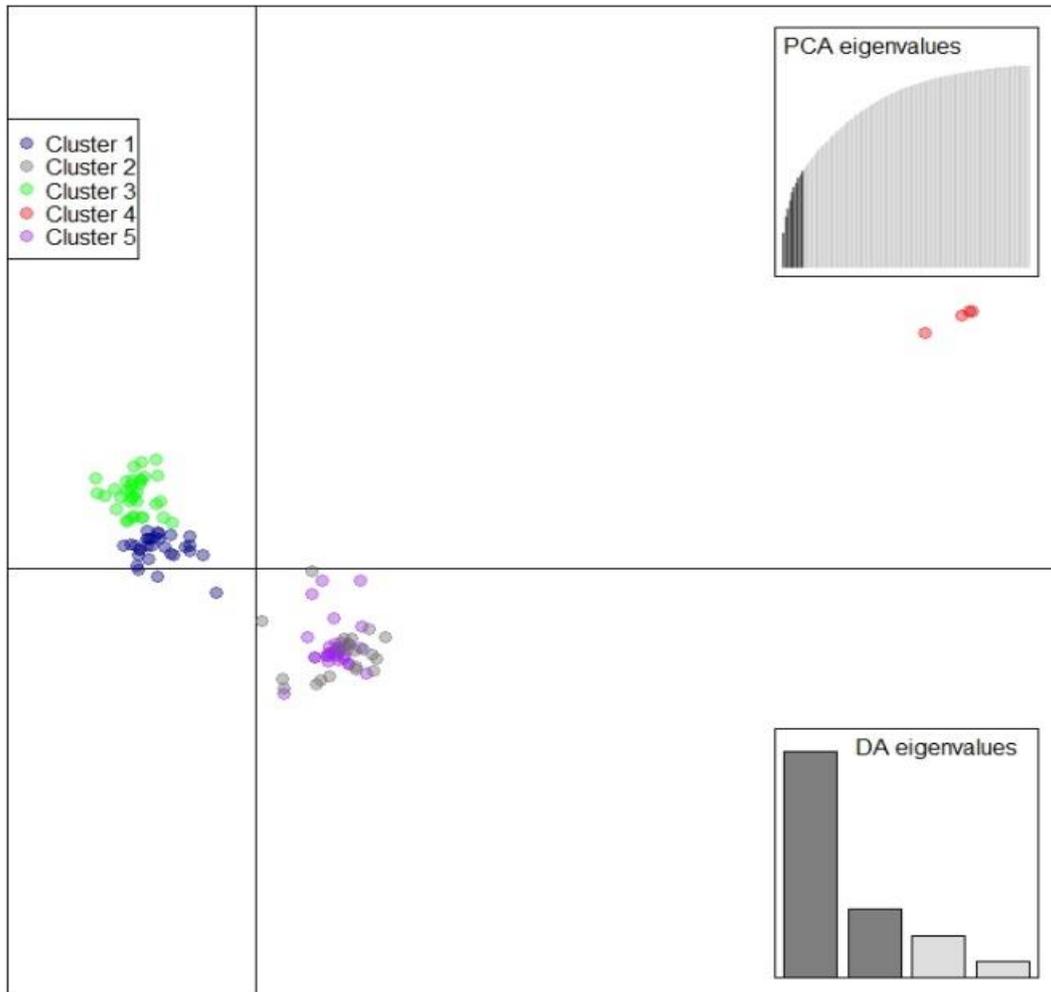
Le graphique montre l'erreur quadratique moyenne (Mean Squared Error : MSE) obtenue par validation croisée (100 répétitions) permettant de sélectionner le nombre de composantes principales à conserver pour l'analyse DAPC avec les 120 génotypes (DD, DC, DP, DE).



Histogramme des valeurs propres des 10 premières composantes principales du modèle optimisé avec les 120 génotypes (DD, DC, DP, DE) – DAPC.



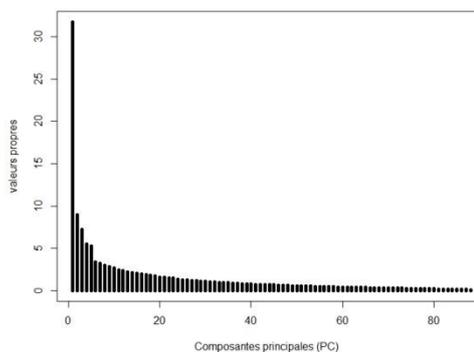
Histogramme des valeurs propres des quatre axes discriminants du modèle optimisé avec les 120 génotypes (DD, DC, DP, DE) – DAPC.



Projection des cinq groupes/clusters (K) sur les deux axes discriminants de l'analyse DAPC (A). Les valeurs propres du modèle optimisé sont présentées en haut à gauche et celles des axes discriminants en bas à droite

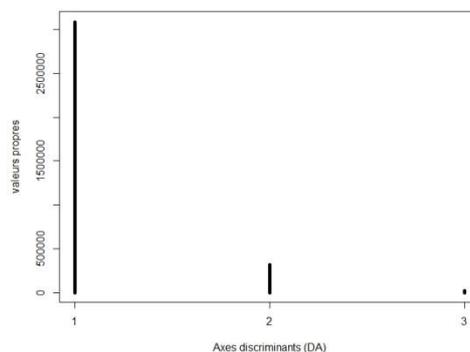
Annexe 6 : Analyse de structure par DAPC sur les 90 génotypes : graphiques complémentaires

Histogramme des valeurs propres - composantes principales (PC)



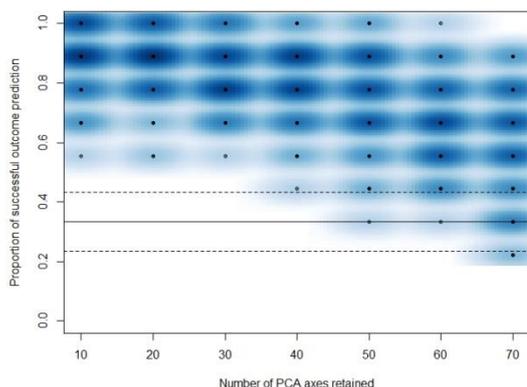
Histogramme des valeurs propres des composantes principales du modèle généraliste avec les 90 génotypes (DC, DP, DE) – DAPC.

Histogramme des valeurs propres - Axes discriminants (DA)



Histogramme des valeurs propres des axes discriminants du modèle généraliste avec les 90 génotypes (DC, DP, DE) – DAPC.

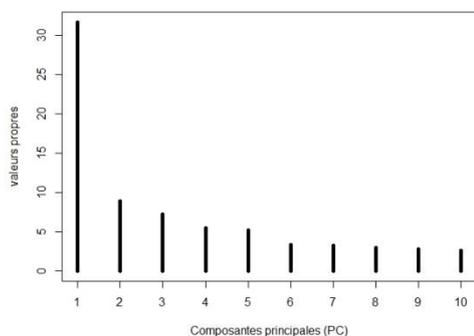
DAPC Cross-Validation



Validation croisée – MSE

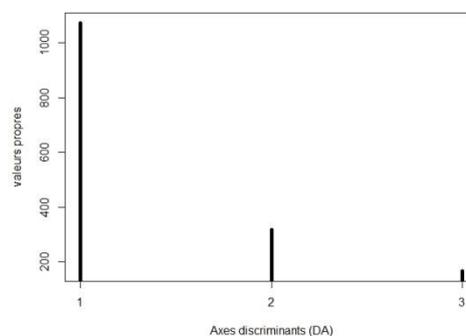
Le graphique montre l'erreur quadratique moyenne (Mean Squared Error : MSE) obtenue par validation croisée (100 répétitions) permettant de sélectionner le nombre de composantes principales à conserver pour l'analyse DAPC avec les 90 génotypes (DC, DP, DE).

Histogramme des valeurs propres - composantes principales (PC)

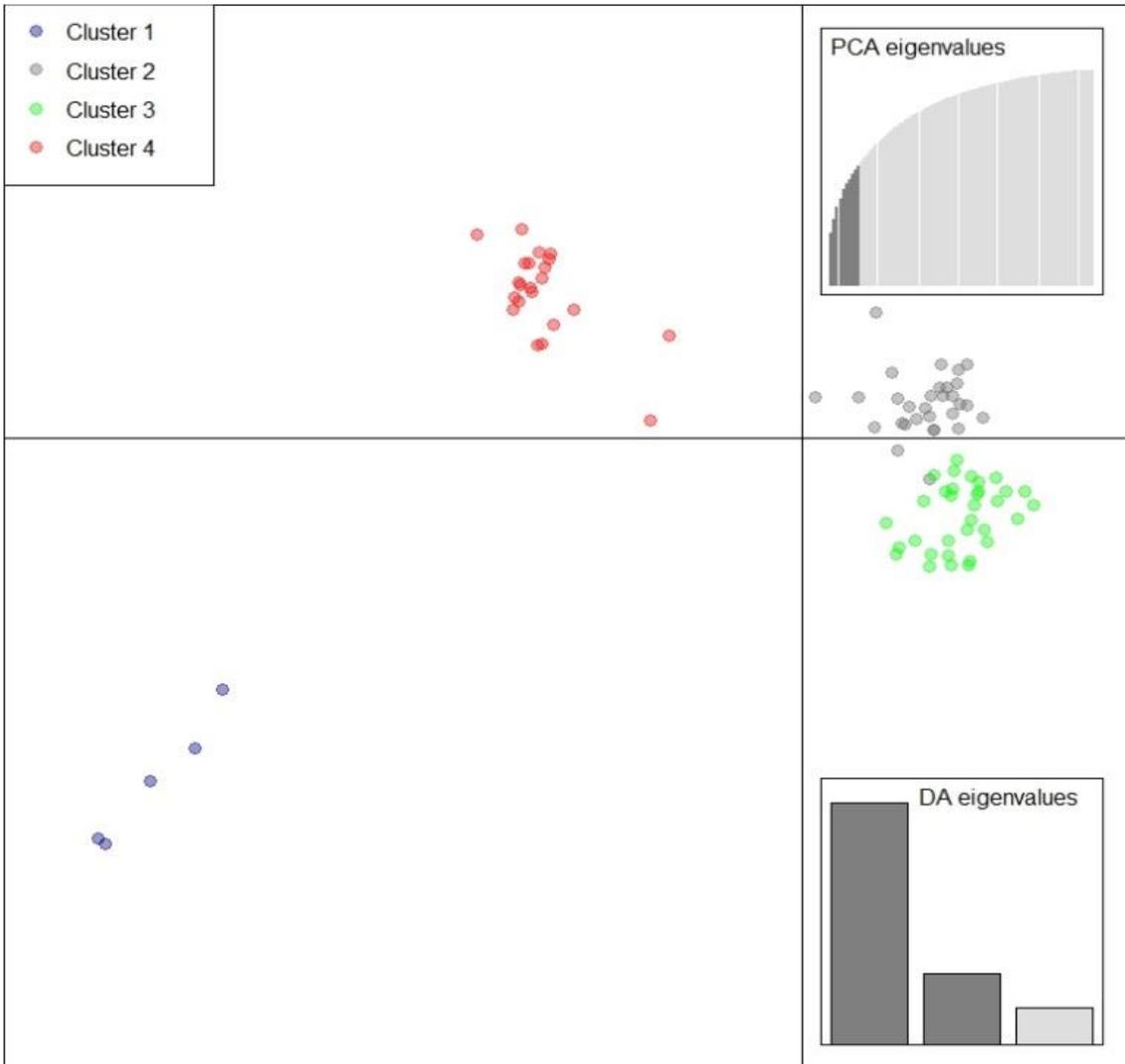


Histogramme des valeurs propres des 10 premières composantes principales du modèle optimisé avec les 90 génotypes (DC, DP, DE) – DAPC.

Histogramme des valeurs propres - Axes discriminants (DA)

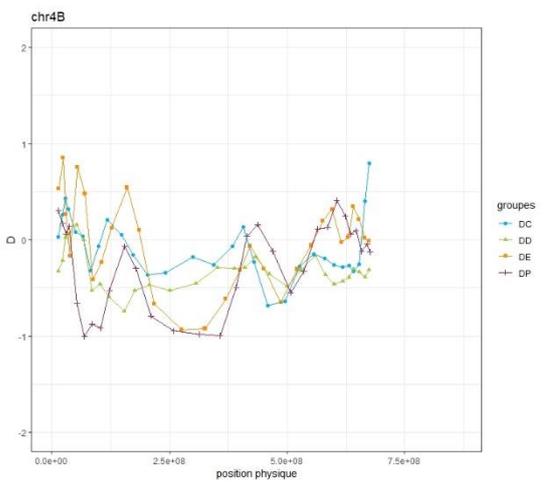
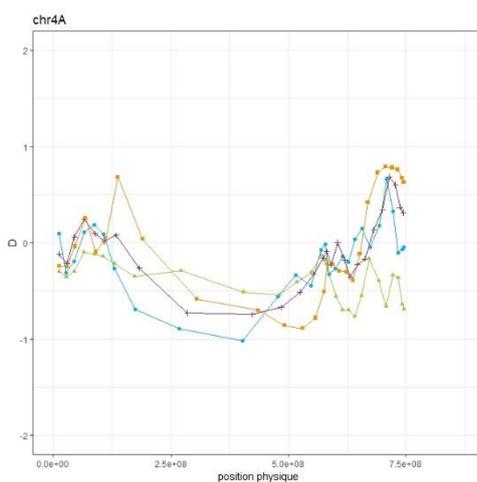
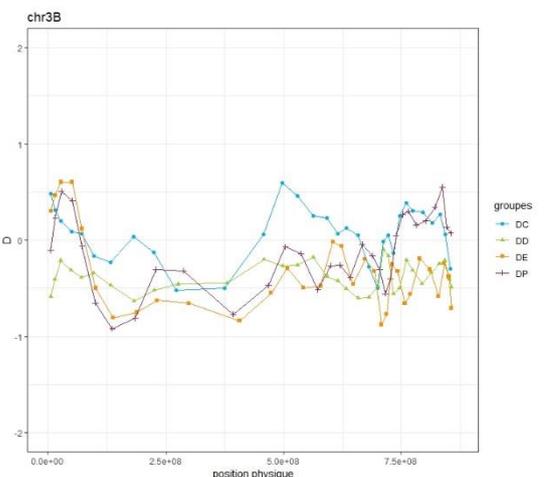
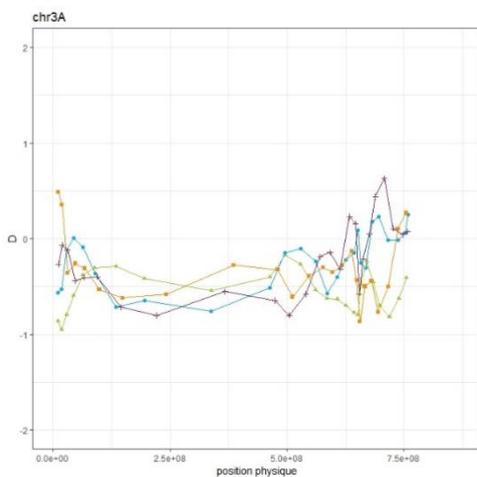
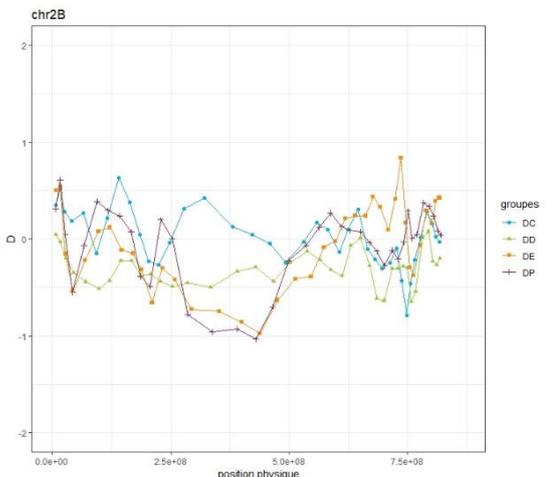
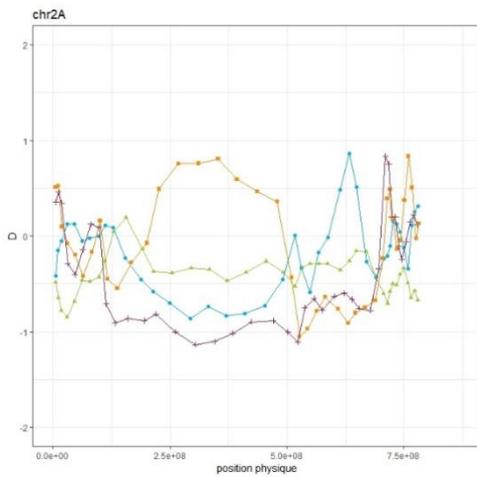
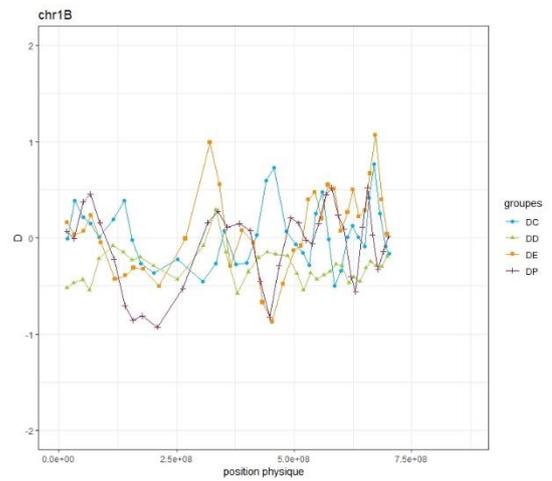
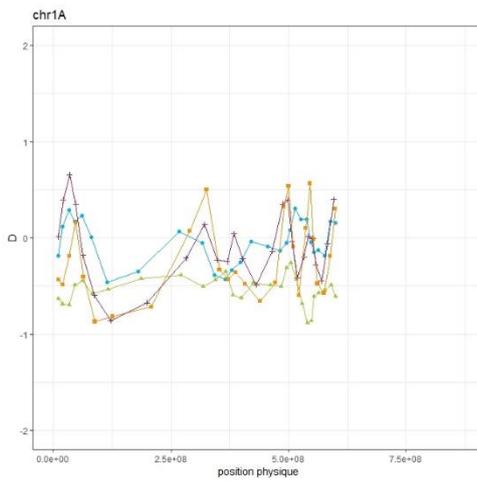


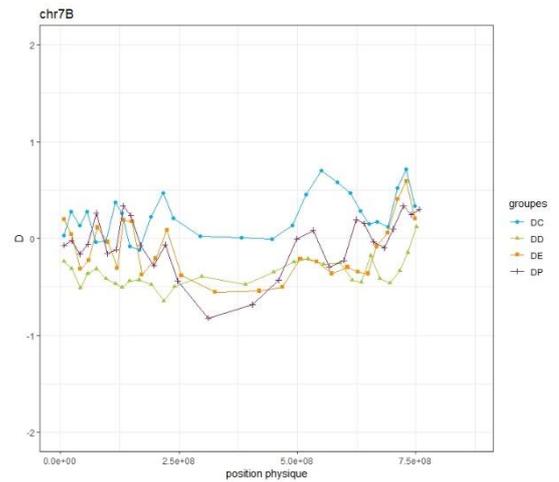
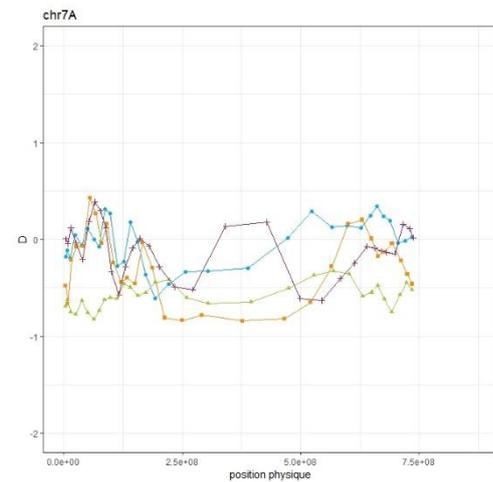
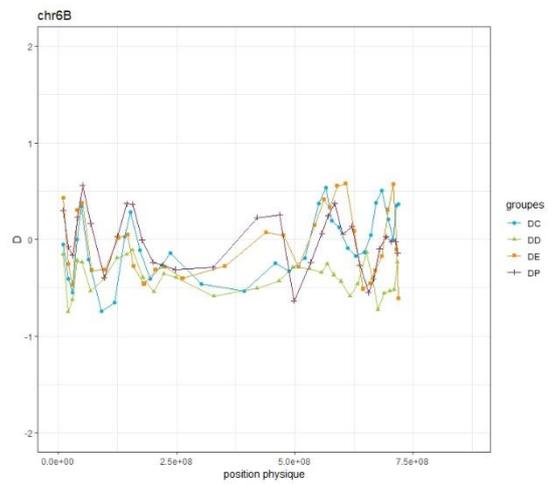
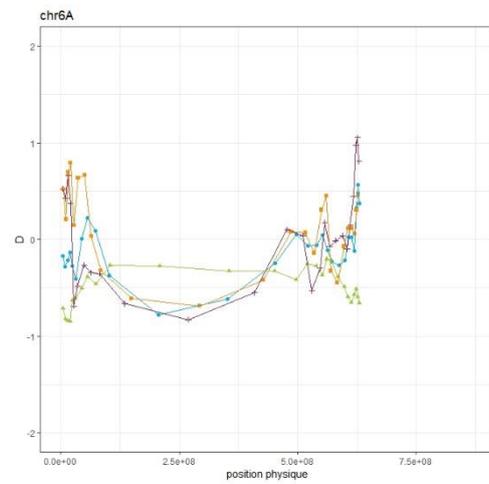
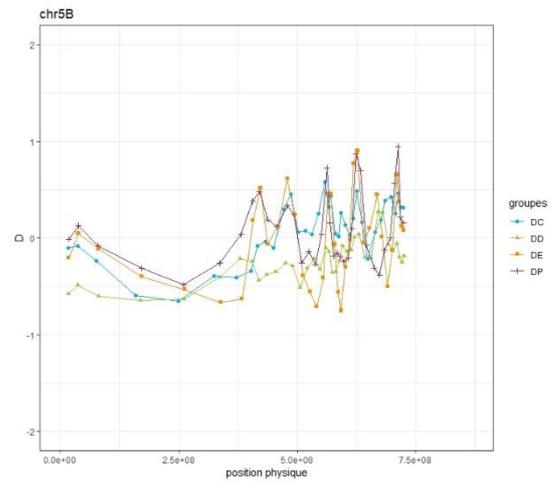
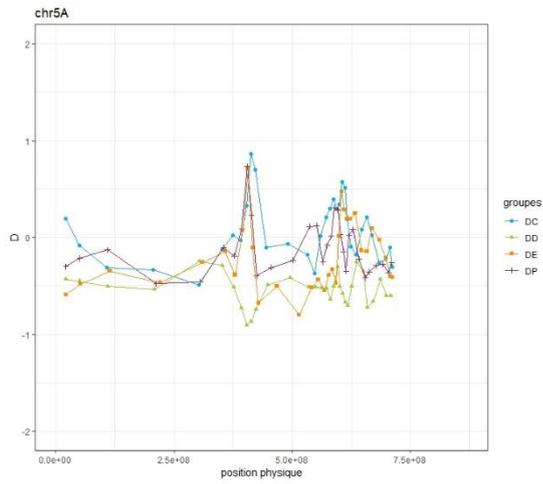
Histogramme des valeurs propres des 4 axes discriminants du modèle optimisé avec les 90 génotypes (DC, DP, DE) – DAPC.



Projection des quatre clusters (K) sur les deux axes discriminants de l'analyse DAPC (A). Les valeurs propres de l'ACP sont présentées en haut à gauche et celles des axes discriminants en bas à droite

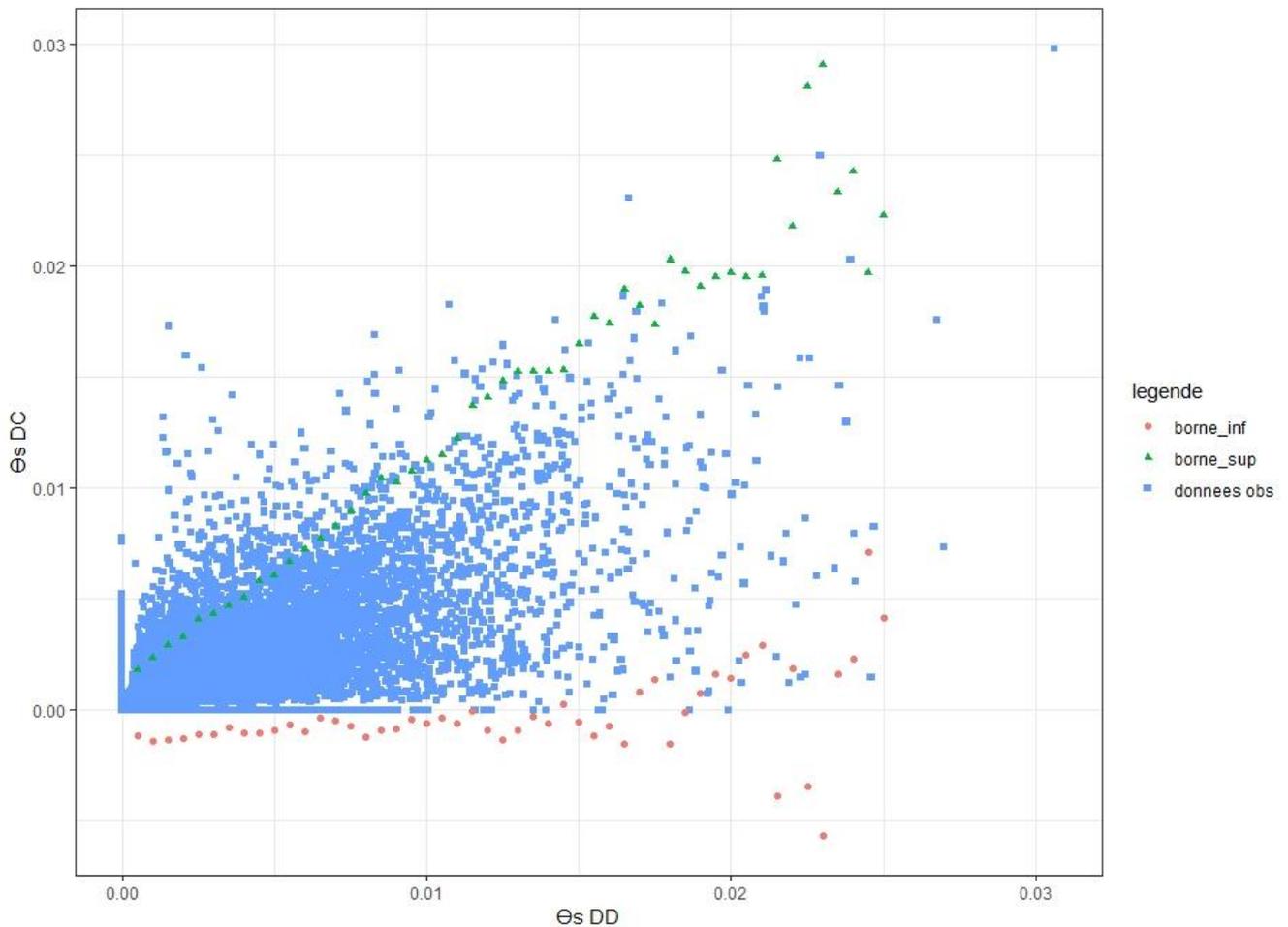
Annexe 7: Evolution des valeurs du D de Tajima le long des chromosomes





Evolution du D de tajima le long des chromosomes, calculé sur des fenêtres glissantes de 20 Kb de séquences analysées avec un chevauchement des fenêtres de 10Kb. La diversité de DD est présentée en vert, celle de DC en bleu, celle de DP en bordeaux et celle de DE en orange.

Annexe 8: Détection de contigs sous sélection par la méthode de Wright, 2005



Représentation graphique de la diversité chez DD en fonction de la diversité chez DC, en utilisant l'estimateur θ_s de Watterson (en bleu).

L'étude menée par Wright et *al.* (2005) sur la sélection chez le maïs, utilisait le θ_s de Watterson pour détecter les gènes sous sélection forte

Ici, le coefficient de la droite de regression, entre les valeurs de θ_s du groupe DD et du groupe DC, ainsi que les résidus du modèle, ont permis de simuler des données afin de représenter une enveloppe correspondant à 95% des données (borne inférieure en rouge et borne supérieure en vert)

le niveau de diversité présent chez *T. turgidum* est beaucoup plus faible que chez la forme sauvage du maïs (téosinte $\pi=0.21$) et la puissance du test n'est pas suffisante pour détecter les gènes sous sélection. En effet, tous les contigs avec une valeur de θ_s forte pour le groupe DD et faible à nulle pour le groupe DC se situe dans l'enveloppe de 95%.

Annexe 9: Estimation des Fst (WC), pour les trois transitions, sur les contigs qui composent les fenêtres de 6Mb , autour des six gènes et QTLs cibles

gène/QTL	chromosome	début_contigs	Contig zavitan_baits	FST_DD_DC	FST_DC_DP	FST_DP_DE
PMG	chr2A	155152645	chr2A:1-394391205:155152645-155153364	0,087	NA	NA
PMG	chr2A	157324478	chr2A:1-394391205:157324478-157325197	0,429	0,489	NA
PMG	chr2A	159371738	chr2A:1-394391205:159371738-159372457	0,041	0,837	0,080
PMG	chr2A	159374995	chr2A:1-394391205:159374995-159375711	-0,079	0,916	0,231
abs_azote	chr3A	35019265	chr3A:1-383808486:35019265-35019984	0,020	0,392	NA
abs_azote	chr3A	35020837	chr3A:1-383808486:35020837-35021556	0,352	0,545	NA
abs_azote	chr3A	36425162	chr3A:1-383808486:36425162-36426555	0,447	0,544	NA
abs_azote	chr3A	36679018	chr3A:1-383808486:36679018-36679737	0,240	0,529	-0,003
abs_azote	chr3A	37085488	chr3A:1-383808486:37085488-37086207	0,041	-0,033	0,069
TtBtr1-A	chr3A	62895787	chr3A:1-383808486:62895787-62896506	0,008	-0,016	-0,076
TtBtr1-A	chr3A	62948044	chr3A:1-383808486:62948044-62948763	0,367	-0,005	NA
TtBtr1-A	chr3A	64407143	chr3A:1-383808486:64407143-64407956	0,345	NA	NA
TtBtr1-B	chr3B	101456687	chr3B:1-432975020:101456687-101457405	0,541	-0,002	NA
Rht-B1b	chr4B	27605560	chr4B:1-342023913:27605560-27606476	0,041	-0,015	0,261
Rht-B1b	chr4B	27610914	chr4B:1-342023913:27610914-27611895	0,355	0,081	-0,019
Rht-B1b	chr4B	27612425	chr4B:1-342023913:27612425-27613122	0,179	-0,013	0,018
Rht-B1b	chr4B	27621685	chr4B:1-342023913:27621685-27622404	0,100	0,099	-0,009
Rht-B1b	chr4B	27622595	chr4B:1-342023913:27622595-27623819	0,175	-0,027	-0,041
Rht-B1b	chr4B	27637182	chr4B:1-342023913:27637182-27638294	0,408	0,091	0,021
Rht-B1b	chr4B	27639188	chr4B:1-342023913:27639188-27639907	0,259	0,118	0,019
Rht-B1b	chr4B	27862499	chr4B:1-342023913:27862499-27863274	0,077	-0,025	0,233
Rht-B1b	chr4B	28409018	chr4B:1-342023913:28409018-28409737	0,190	NA	0,799
Rht-B1b	chr4B	29703055	chr4B:1-342023913:29703055-29703774	0,781	0,007	NA
Rht-B1b	chr4B	29925453	chr4B:1-342023913:29925453-29926169	0,502	NA	NA
Rht-B1b	chr4B	29984076	chr4B:1-342023913:29984076-29984795	0,562	0,066	NA
Rht-B1b	chr4B	30660299	chr4B:1-342023913:30660299-30661018	0,559	0,039	NA
Rht-B1b	chr4B	30784503	chr4B:1-342023913:30784503-30785222	0,457	0,008	NA
Rht-B1b	chr4B	30931689	chr4B:1-342023913:30931689-30932408	0,174	0,018	NA
Rht-B1b	chr4B	30964189	chr4B:1-342023913:30964189-30964908	0,191	NA	NA
Rht-B1b	chr4B	31309326	chr4B:1-342023913:31309326-31310045	0,147	0,188	0,708
Rht-B1b	chr4B	31378673	chr4B:1-342023913:31378673-31379391	0,309	NA	NA
Rht-B1b	chr4B	31942322	chr4B:1-342023913:31942322-31943725	-0,035	0,056	0,430
Rht-B1b	chr4B	31947930	chr4B:1-342023913:31947930-31948649	0,099	-0,054	0,389
Rht-B1b	chr4B	32057400	chr4B:1-342023913:32057400-32058119	0,222	0,226	NA
Rht-B1b	chr4B	32469500	chr4B:1-342023913:32469500-32470907	0,161	0,042	0,580
Q	chr5A	653812697	chr5A:357693102-715386202:296119596-296120429	0,366	0,232	0,053
Q	chr5A	654039677	chr5A:357693102-715386202:296346576-296347818	0,188	0,318	-0,003
Q	chr5A	654043503	chr5A:357693102-715386202:296350402-296351121	-0,014	0,761	-0,008
Q	chr5A	654255515	chr5A:357693102-715386202:296562414-296563133	0,212	0,627	NA
Q	chr5A	654256772	chr5A:357693102-715386202:296563671-296564390	0,092	0,322	0,000
Q	chr5A	654391471	chr5A:357693102-715386202:296698370-296699089	-0,055	-0,029	-0,058
Q	chr5A	654418183	chr5A:357693102-715386202:296725082-296725801	0,141	0,398	NA
Q	chr5A	654419010	chr5A:357693102-715386202:296725909-296726622	0,136	0,216	-0,005
Q	chr5A	654482139	chr5A:357693102-715386202:296789038-296789757	-0,035	0,855	0,000
Q	chr5A	654663478	chr5A:357693102-715386202:296970377-296971095	0,483	0,547	NA
Q	chr5A	654822879	chr5A:357693102-715386202:297129778-297130487	0,380	0,533	NA
Q	chr5A	655136051	chr5A:357693102-715386202:297442950-297443669	0,186	0,426	NA
Q	chr5A	655304635	chr5A:357693102-715386202:297611534-297612253	-0,051	0,235	NA
Q	chr5A	656565140	chr5A:357693102-715386202:298872039-298872758	-0,015	0,395	0,021
Q	chr5A	656620799	chr5A:357693102-715386202:298927698-298928417	0,006	0,661	NA

ECOLE PRATIQUE DES HAUTES ETUDES

Sciences de la Vie et de la Terre

Morgane ARDISSON

10 décembre 2019

Histoire de la domestication de *Triticum turgidum* : La capture d'exons au service de l'étude de la diversité génétique

RESUME

Depuis le début de sa domestication dans le croissant fertile, il y a 12 000 ans, le blé dur, *Triticum turgidum* a subi de nombreux événements démographiques et sélectifs. La première transition caractérisant le passage du blé dur sauvage (amidonnier sauvage) *Triticum turgidum* spp *dicoccoïdes*, à la première forme cultivée (amidonnier cultivé) *Triticum turgidum* spp *dicoccum* est marquée, entre autre, par l'apparition d'un rachis solide. Ce caractère, contrôlé notamment par les gènes Br, a permis aux premiers agriculteurs de récolter, sur les plantes, le grain à maturité. La deuxième transition s'est produite entre l'amidonniér cultivé et *Triticum turgidum* spp *durum*. Elle est caractérisée par l'apparition des grains nus permettant un battage plus facile (caractère contrôlé notamment par le gène Q). La dernière transition majeure intervient lors de la révolution verte, dans les années 1960, et marque le passage des « populations de pays » aux variétés « élites », sélectionnées pour de nombreux caractères morphologiques comme une taille réduite (gène Rht), un rendement plus fort ou une capacité plus grande à absorber beaucoup d'azote sans que les plantes ne soient sujettes à la verse.

Dans le but d'affiner nos connaissances sur l'histoire évolutive de l'espèce *Triticum turgidum*, nous avons produit un jeu de données moléculaires en utilisant la méthode d'enrichissement par capture (10 000 régions de 120pb situées dans la partie codante du génome) pour contourner les difficultés liées à la grande taille de son génome (10.5 Gb). Cette étude a été réalisée sur 120 génotypes, 30 pour chacune des quatre formes évolutives précédemment citées. L'analyse de la diversité nucléotidique obtenue (135 863 SNPs) nous a permis d'estimer la réduction de celle-ci associée à chaque transition. Le goulot d'étranglement le plus important a eu lieu lors du premier événement de domestication avec seulement 59 % de la diversité présente chez *Triticum turgidum* spp *dicoccoïdes* encore présente chez la forme cultivée, *Triticum turgidum* spp *dicoccum*. Pour la deuxième transition, nous avons estimé que 76% de la diversité de *Triticum turgidum* spp *dicoccum* se retrouvait dans le groupe *Triticum turgidum* ssp *durum*. Et pour finir, 91% de la diversité est conservée entre les deux formes de *T. turgidum* ssp *durum*. L'utilisation de la séquence génomique de référence (accession «Zavitan» - Avni et al. 2017; Zhu et al. 2019), nous a également permis d'observer la variation du niveau de diversité le long du génome pour chacune des quatre formes évolutives. Nous avons observé que certaines zones ont vu leur niveau de diversité considérablement diminuer au cours de la domestication, alors que d'autres sont encore très polymorphes. Nous avons également pu détecter des signatures de sélection, sur ces quatre formes, d'abord sans *a priori*, puis en ciblant les locus impliqués dans l'expression de traits phénotypiques caractéristiques de la domestication (locus Br, Q et Rht) ainsi que deux QTLs impliqués dans le poids des grains et la teneur en azote dans la feuille, trait qui reflète la stratégie d'acquisition des ressources de la plante.

Ce travail a permis de proposer de nouvelles perspectives pour le développement d'outils moléculaires et des nouvelles pistes pour affiner notre compréhension de l'histoire évolutive de l'espèce *Triticum turgidum*.

Mots-clés : *Triticum turgidum*, domestication, histoire démographique, sélection, capture, SNP, structure, diversité.