Apport de la recombinaison dans l'optimisation des plans de croisements de blé tendre

Alice Danguy Des Deserts

Université Clermont Auvergne

Ecole Doctorale des Sciences de la Vie et de la Santé, Agronomie & Environnement

Thèse de Doctorat

Présentée à l'Université Clermont Auvergne pour l'obtention du grade de

Docteur d'Université

Spécialité : Biologie végétale

Apport de la recombinaison dans l'optimisation des plans de croisements de blé tendre

*Présentée et soutenue le **15 décembre 2021** par*

Alice DANGUY DES DESERTS

Composition du jury

Pascale LE ROY Directeur de recherche, INRAE UMR Pégase, Saint-Gilles	Rapporteur
Olivier MARTIN Directeur de recherche, Institut des Plantes de Paris Saclay, Gif-sur-Yvette	Rapporteur
Christina LEHERMEIER Head of Statistical Genetics Unit, RAGT 2n, Druelle	Examinatrice
Friedrich LONGIN Außerplanmäßiger Professor, University of Hohenheim, Stuttgart	Examinateur
Sylvie NORRE Professeur des Universités, Université Clermont Auvergne, laboratoire LIMOS, Clermont-Ferrand	Examinatrice
Pierre SOURDILLE Directeur de recherche, INRAE UMR GDEC, Clermont-Ferrand	Directeur de thèse
Sophie BOUCHET Chargée de recherche, INRAE UMR GDEC, Clermont-Ferrand	Invitée
Ellen GOUDEMAND DUGUE Head of Genetics and Biometrics Unit, Florimond Desprez, Cappelle-en-Pévèle	Invitée

UMR 1095 INRAE-UCA « Génétique, Diversité et Ecophysiologie des Céréales »

LABORATOIRE D'ACCUEIL

GDEC - Génétique Diversité et Ecophysiologie des Céréales
UMR 1095 INRAE-UCA
5 chemin de Beaulieu
63 000 Clermont-Ferrand
France

Thèse financée par le métaprogramme INRAE SELGEN et par Florimond Desprez

Remerciements

Avant tout, je tiens à remercier mon encadrante Sophie Bouchet et mes directeurs de thèse, Gilles Charmet et Pierre Sourdille, pour avoir partagé avec moi toute leur érudition et leur ingéniosité, pour avoir répondu avec bienveillance à mes nombreuses interrogations et pour m'avoir conseillée dans la rédaction de ce manuscrit.

Je remercie également Jean-Michel Elsen et Bertrand Servin pour leur grande implication dans ce travail de thèse. Je vous suis très reconnaissante d'avoir partagé avec moi votre culture et votre expertise scientifiques.

Merci à Florimond Desprez d'avoir co-financé cette thèse. Merci à Ellen Goudemand Dugué et aux sélectionneurs Delphine Tailleu et Phillipe Lonnet d'avoir pris le temps de m'expliquer les schémas de sélection et d'avoir répondu à mes questions.

Je remercie aussi sincèrement les membres du jury, Christina Lehermeier, Pascale Le Roy, Friedrich Longin, Olivier Martin et Sylvie Norre, d'avoir accepté d'évaluer ce travail.

Je souhaite également remercier les membres de mon comité de thèse, Frédérique Choulet, Benoît Darrier, Sylvain Glémin, Tristan Mary-Huard et Renaud Rincent pour avoir alimenté la réflexion autour de ce travail.

Merci aussi à Jean-Marc Alliot, Nicolas Durand et Daniel Ruiz, merci à vous pour votre contribution sur le développement des outils informatiques.

Je pense aussi à mes collègues doctorants/post-doctorants/jeunes chercheurs, aux échanges enrichissants que nous avons eus. Merci donc à Amir, Audrey, Claire, Diane, Justin, Sarah et Vanille. Je vous admire beaucoup et vous souhaite le meilleur pour la suite.

Enfin, encore merci à mes proches, qui m'ont soutenue et encouragée tout au long de mes études. Merci beaucoup à Maman et Alexis.

Publications dans des revues avec comité de lecture

- **Alice Danguy des Déserts**, Sophie Bouchet, Pierre Sourdille, et Bertrand Servin. 2021. « Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat ». Genome Biology and Evolution 13 (8). https://doi.org/10.1093/gbe/evab152.

- Peter Civan, Renaud Rincent, **Alice Danguy des Déserts**, Jean-Michel Elsen, et Sophie Bouchet. « Population Genomics Along With Quantitative Genetics Provides a More Efficient Valorization of Crop Plant Genetic Diversity in Breeding and Pre-Breeding Programs ». In , 1-64. Population Genomics. Cham: Springer International Publishing. https://doi.org/10.1007/13836_2021_97.

Liste des communications orales dans des séminaires ou congrès internationaux

- **Alice Danguy des Déserts**, Bertrand Servin, Ludovic Duvaux, Sophie Bouchet et Pierre Sourdille. « Historical recombination maps estimated with two diverging Euopean and Asian population on bread wheat chromosome 3B" Gordon Research Conference - Quantitative Genetics and Genomic 2019, Lucca, Italie [**Poster**]

- **Alice Danguy des Déserts**, Bertrand Servin, Ludovic Duvaux, Sophie Bouchet et Pierre Sourdille. « Historical recombination maps estimated with two diverging Euopean and Asian population on bread wheat chromosome 3B» International Wheat Conference 2019, Saskatoon, Canada [**Poster**]

- **Alice Danguy des Déserts**, Bertrand Servin, Sophie Bouchet et Pierre Sourdille. « Historical recombination maps estimated with two diverging Euopean and Asian population on bread wheat chromosome 3B" Approche Interdisciplinaire de l'Evolution Moléculaire 2019, Toulouse [**Poster**]

Liste des Abréviations

ADN : Acide désoxyribonucléique

BLUP : Best Linear Unbiaised Prediction

BPS : Blé Panifiable Supérieur

COV : Certificat d'Obtention Végétal

CSC : Critères de Sélection de Croisement

DL : Déséquilibre de Liaison

EMBV : Expected Maximum Breeding Value (espérance des K meilleurs descendants parmi D descendants d'un croisement

FAO : Organisation des Nations Unies pour l'alimentation et l'agriculture

F_{ST} : Indice de différenciation

G : matrice de relation génomique

GEBV : Genomic Estimated Breeding Value

GS : Genomic Selection

h^2 : héritabilité (au sens large)

HDs : Haploides doublés

IWGSC : International Wheat Genome Sequencing Consortium

OCS : Optimal Contribution Selection ou Optimal Cross Selection

OHV : Optimal Haploid Value (meilleur gamète dérivable d'un croisement, après avoir défini des blocs haplotypiques supposés être transmis suite à la méiose)

PROBA : Probabilité pour un couple de produire un descendant de valeur génétique supérieur à la meilleure lignée parentale du programme de sélection

QTL : Quantitative Trait Locus

R^2 : Coefficient de Détermination

RILs : Recombinant inbred Lines

SNP : Single Nucleotide Polymorphism

TBV : True Breeding Value

UC : Usefulness Criterion (espérance des q (%) meilleurs descendants d'un croisement)

UCPC : Usefulness Criterion-based Parental Contribution

λ : déviation du taux de recombinaison historique d'un intervalle entre deux SNPs au taux de recombinaison historique de base d'un segment chromosomique chevauchant l'intervalle

ρ : taux de recombinaison historique

Φ : matrice d'apparentement

Table des Matières

Table des Figures

Liste des Tableaux

Introduction générale

Introduction générale

Pour quasiment toutes les espèces de grande culture, les rendements stagnent depuis une vingtaine d'années. Ce déficit de production est particulièrement notable pour le blé tendre (*Triticum aestivum* L.), nourriture de base pour plus de 2.5 milliards de personnes, représentant en moyenne 20% des calories de l'alimentation humaine, et pour lequel il faudrait doubler les gains de rendements annuels afin de subvenir aux besoins futurs (FAOstat, 2013). La stagnation des rendements, qui oscillent entre 30 à 90% du rendement potentiel, concerne plus du tiers des pays producteurs, y compris les plus grands comme la Chine, les Etats-Unis ou la France (Ray et al. 2012). Cette stagnation est partiellement causée par les stress biotiques et abiotiques qui se développent avec le changement climatique (Wang et al. 2003; Olesen et al. 2011).

L'un des principaux leviers pour augmenter la production agricole est l'amélioration génétique des variétés (Calderini et Slafer 1998). L'objectif est d'améliorer le matériel de sélection à chaque génération sur un maximum de caractères (gain de rendement, de qualités nutritionnelles ou technologiques, moins d'exigences en intrants et meilleure tolérance aux stress biotiques et abiotiques) et sur des marchés de niche (blé de force…), sans cependant augmenter la pression sur l'environnement (utilisation d'intrants ou augmentation des surfaces cultivées) en produisant des variétés résilientes aux stress.

L'amélioration génétique des variétés est réalisée par des sélectionneurs privés et publics. Il existe moins d'une trentaine d'entreprises de sélection pour le blé tendre en France (Perronne et al. 2017). On peut citer Florimond Desprez, qui co-finance cette thèse, ainsi que le programme de sélection Agri Obtentions-INRAE, dont la base de données génotypiques et phénotypiques a été utilisée pour cette thèse.

Pour qu'une variété de blé soit commercialisée en France, il faut qu'elle soit inscrite au catalogue officiel français ou européen des variétés. L'inscription est conditionnée par différents tests opérés par le GEVES durant 2 ans dans un réseau d'essai, dont le test de Distinction, Homogénéité et Stabilité, qui s'assure que les variétés présentent un phénotype stable et différent des variétés déjà inscrites, ainsi qu'un test sur la valeur agronomique des variétés (VATE) (https://www.geves.fr/expertises-varietes-semences/). Différents phénotypes sont évalués, répartis en 4 catégories : le rendement (première caractéristique recherchée par les agriculteurs, la valeur technologique des farines (teneur en protéines, force boulangère…), les caractéristiques physiologiques (précocité, hauteur, résistance à la verse…) et les résistances aux bioagresseurs (France AgriMer 2015). Les performances sur chaque caractère sont intégrées dans une note, qui doit dépasser une note minimum pour permettre l'inscription. Il faut aussi que la variété remplisse certaines exigences minimums, par exemple avoir un rendement supérieur à 102% du rendement des variétés témoins du CTPS (Comité Technique Permanent de la Sélection des plantes cultivées), ce seuil pouvant varier selon la catégorie de variétés. Dans l'Union européenne et en

France, la production et la commercialisation de semences des variétés peuvent être réservées à l'obtenteur grâce à un titre de propriété dénommé "Certificat d'Obtention Végétale" (COV). Contrairement à un brevet, utilisé dans certains pays, le COV permet l'utilisation des variétés inscrites comme géniteurs par les concurrents, sans droit de suite pour l'obtenteur. On parle de l'exception du sélectionneur en France pour le blé.

L'amélioration génétique d'une espèce repose sur les croisements entre variétés possédant des caractéristiques intéressantes et complémentaires et l'identification dans la descendance d'individus avec des qualités supérieures aux témoins de référence. Au sein d'un programme de sélection, le choix des croisements est donc déterminant pour le progrès génétique, mais il se heurte à deux difficultés principales.

La première difficulté est le compromis entre plusieurs objectifs non indépendants. En effet, plusieurs caractères d'intérêt peuvent être corrélés négativement, ce qui nécessite soit d'établir des pondérations sur chaque caractère, soit de définir des classes pour chaque caractère et d'améliorer les variétés à l'intérieur de chaque classe (par exemple, sélection spécifique pour la qualité). Pour assurer un gain génétique sur le long terme, des contraintes de maintien de la diversité génétique peuvent également faire partie des objectifs mais sont négativement corrélées avec le progrès variétal à court terme.

La deuxième difficulté est la nécessité d'un retour sur investissement rapide et régulier. Les ressources (budget, temps, main d'œuvre) doivent être investis de manière à assurer un progrès variétal significatif pour inscrire de nouvelles variétés tous les ans. Pour s'assurer d'avoir une bonne descendance, le plus sûr est de croiser des parents avec les meilleures performances pour le caractère d'intérêt, ce qui garantit de maximiser l'espérance de la descendance (Lupton 1961; Melchinger et al. 1998; Souza et Sorrells 1991; Kotzamanidis et al. 2008; Utz, Bohn, et Melchinger 2001). Le risque est alors de croiser des parents génétiquement très proches, et donc produire des descendants trop similaires aux parents et qui ne représentent pas un progrès variétal significatif. Par ailleurs, croiser des parents génétiquement trop proches représente aussi un risque de perdre de la diversité rapidement dans le matériel, surtout dans un programme de sélection fermée, sans introduction de lignées parentales extérieures à la population habituellement travaillée par le sélectionneur. Une méthode alternative pourrait être de choisir les croisements qui ont le plus de chance de produire des descendants transgressifs, c'est-à-dire supérieurs aux variétés parentales ou aux témoins pour l'inscription. Mais si les effectifs sont limités dans la descendance, il y a un risque de ne pas obtenir les individus transgressifs espérés et d'observer une descendance moins performante que si l'on avait optimisé la moyenne des valeurs génétiques des parents. En effet, moyenne et variance peuvent être corrélées négativement chez les populations de plantes (Mohammadi et al. 2015, Lado et al. 2017, Yao et al. 2018). Il s'agit donc d'optimiser la prise de risque et le retour sur investissement pour un budget (taille du programme) donné. La prédiction de la distribution des performances des descendants de chaque croisement candidat, qui *permet in fine* d'avoir une estimation de la probabilité d'obtenir des descendants extrêmes ou une

estimation des valeurs génétiques des meilleurs descendants d'un croisement, a fait l'objet de nombreuses recherches. La moyenne de cette distribution, autrement appelée espérance des descendants, est simple à prédire puisqu'il s'agit de la moyenne des valeurs génétiques des parents. Par contre, l'estimation de la variance est plus compliquée. Grâce au génotypage haut-débit et aux méthodes de prédiction génomique, il est désormais possible d'estimer l'effet de chaque allèle sur un caractère. Le développement de cartes génétiques permet ensuite d'estimer la variabilité des génotypes des descendants d'un couple, et par extension la variance de la descendance pour un caractère. Ainsi, on peut prédire en amont les croisements intéressants (utiles), c'est-à-dire impliquant des parents porteurs d'allèles favorables à différents loci, et dont les allèles favorables ont une probabilité élevée d'être cumulés dans certains descendants.
L'élaboration du patrimoine génétique des descendants est le résultat de deux processus : la formation des gamètes haploïdes (avec un seul stock de chromosomes homologues) durant la méiose et la fusion des gamètes maternels et paternels lors de la fécondation pour restaurer un statut diploïde (avec deux stocks de chromosomes homologues) dans l'embryon. Ces deux processus ont pour conséquence de redistribuer les allèles parentaux dans la descendance, ce qui permet de faire émerger de nouvelles combinaisons alléliques inédites et potentiellement plus adaptées à un environnement.

Ce phénomène de brassage génétique est aléatoire, et est à l'origine de la variabilité des descendants d'un couple et plus généralement de l'évolution des espèces. Cependant, tous les génotypes possibles des descendants ne sont pas équiprobables. En effet, la probabilité pour une série d'allèles aux loci portés par un même chromosome parental de co-ségréger dans la descendance dépend de la recombinaison homologue, un processus biologique qui intervient lors de la formation des gamètes. La recombinaison homologue est le résultat d'un échange systématique, réciproque et obligatoire de larges portions d'ADN entre les chromosomes homologues, appelé crossing-over (ou CO), conduisant à un réarrangement des combinaisons alléliques parentales portées par un même chromosome.

La recombinaison homologue est un processus particulièrement important en amélioration variétale car c'est l'un des mécanismes qui permet de dissocier les allèles favorables des allèles délétères dans le génome des descendants. Par contre, la recombinaison homologue est indésirable lorsqu'elle conduit à dissocier les combinaisons d'allèles favorables. La recombinaison homologue est un phénomène assez rare, avec 1 à 3 CO par chromosome et par méiose (Mercier et al. 2015). De plus, la probabilité d'observer un CO le long des chromosomes n'est pas uniforme. Chez le blé tendre, la distribution des CO le long des chromosomes est particulièrement déséquilibrée car seulement 20% des segments chromosomiques reçoivent 80% des CO. Les télomères montrent une forte fréquence de CO, alors que les régions péri-centromériques sont quasiment dépourvus de CO (Choulet et al. 2014) . Chez de nombreuses espèces, il a été observé que les CO se forment préférentiellement dans certaines régions du génome appelées points chauds de recombinaison (synthèse dans Choi et Henderson 2015). Au contraire, d'autres zones

chromosomiques sont complètement dépourvues de CO, ce sont des zones froides de recombinaison. Le profil de recombinaison est défini comme le vecteur des fréquences de CO le long du génome. A ce jour, chez le blé, peu de choses sont connues sur le déterminisme génétique de la position des CO (autrement appelé déterminisme du profil de recombinaison). Chez plusieurs espèces de plantes, le profil de recombinaison semble être hautement variable entre groupes génétiques ou sous-espèces (Marand et al. 2019; Schwarzkopf et al. 2020). Darrier et al. (2017) ont étudié la variabilité du profil de recombinaison du blé tendre sur deux régions d'environ un mégabase (Mb) chacune pour trois populations de blé tendre : une population d'accessions européennes, une population d'accessions asiatiques et une population de lignées recombinantes issues du croisement entre deux lignées parentales Chinese Spring et Renan. Le profil de recombinaison était globalement conservé entre les trois populations mais montrait des variations locales entre les trois populations. Une estimation du profil de recombinaison sur l'ensemble du génome pour plusieurs populations divergentes de blé tendre semblait donc nécessaire pour apprécier la variabilité génétique du profil de recombinaison à l'échelle du génome.

Mieux connaître la variabilité génétique du profil de recombinaison est nécessaire pour pouvoir prédire plus précisément la variance de la descendance en sélection. Par ailleurs, comprendre le déterminisme du profil de recombinaison représenterait une opportunité de faciliter les recombinaisons intéressantes, en utilisant par exemple des biotechnologies (Ru et Bernardo 2019a). Par exemple, chez le blé tendre, les régions péri-centromériques sont des région froides très étendues (environ un tiers des chromosomes) mais contiennent 60% des gènes (Choulet et al. 2014). L'absence de recombinaison dans ces régions limite l'accumulation d'allèles favorables dans ces régions, mais favorise aussi l'accumulation de mutations délétères (Rodgers-Melnick et al. 2015; Renaut et Rieseberg 2015), ce qui limite le progrès variétal.

L'objectif de cette thèse est double :

- D'une part, il s'agit d'estimer et de comparer les profils de recombinaison spécifiques des principaux groupes génétiques du blé tendre, afin de mieux caractériser la variabilité du profil de recombinaison mais aussi de mieux caractériser le déterminisme de la position des CO. Cette partie a fait l'objet d'une publication dans le journal Genome Biology Evolution (Danguy des Déserts et al. 2021). Les profils de recombinaison ont été estimés à partir des patrons de déséquilibre de liaison observé dans des landraces des quatre principaux groupes génétiques du blé tendre. A notre connaissance, cette méthode a été utilisée chez d'autres espèces de plantes (Marand et al. 2019; Hellsten et al. 2013; Choi et al. 2013; Fuentes et al. 2021; Schwarzkopf et al. 2020) et quelques segments du chromosome 3B du blé tendre (Darrier et al. 2017), mais pas encore pour estimer et comparer la recombinaison sur l'ensemble du génome. Ces approches basées sur le déséquilibre de liaison ont l'avantage d'estimer des variations fines du taux de recombinaison le long du génome grâce à une forte densité de SNP polymorphes, ce qui est pertinent pour la comparaison du profil de recombinaison entre populations. Le profil de recombinaison estimé dans cette première partie a été testé pour prédire la variance de la descendance dans un programme de sélection.
- D'autre part, il s'agit d'étudier les bénéfices de la prise en compte de la recombinaison dans l'optimisation du plan de croisements en termes de gain et de diversité génétiques dans le cadre d'un programme de sélection variétale de blé tendre d'hiver. En effet, plusieurs auteurs soulignent que les bénéfices d'une prise en compte de la variance de la descendance lors du choix des croisements dépendent de la variance génétique et de la structure génétique de la population parentale (Zhong et Jannink 2007; Bijma et al. 2020). A notre connaissance, ces bénéfices n'ont pas encore été étudiés pour une population de blé tendre élite d'hiver, ce que nous nous proposons de faire.

Chapitre I : Synthèse bibliographique

Chapitre I : Synthèse bibliographique

I.1 Amélioration variétale du blé tendre

I.1.1 Domestication et différenciation génétique du blé tendre

La sélection exploite la variabilité génétique du blé tendre, déterminée par son histoire évolutive. Le blé tendre est une espèce qui s'est formée suite à deux hybridations interspécifiques (Charmet 2011; Heun et al. 1997; Gill et al. 2004; Glémin et al. 2019), **Figure 1**). La première hybridation a eu lieu il y a environ 800 000 ans entre l'espèce *Triticum Urartu* (dont le génome diploïde est appelé A^uA^u) et une espèce inconnue proche d'*Aegilops speltoides* (génome BB). Cette première hybridation a produit l'espèce sauvage *Triticum diccocoides* puis la forme domestiquée *T. turgidum* (AABB) (Marcussen et al. 2014). La deuxième hybridation a eu lieu entre 8 000 et 10 000 ans, très probablement entre un *Triticum turgidum* cultivé et une espèce sauvage *Aegilops tauschii* (génome DD), générant ainsi le blé tendre (AABBDD).

En conséquence de cette double hybridation, le génome du blé tendre est composé de 3 sous-génomes homéologues A, B et D présentant des similarités au niveau de la séquence d'ADN **(**https://www.wheatgenome.org/**,** (IWGSC 2014; 2018). Chaque sous-génome contient 7 paires de chromosomes homologues, soit 42 chromosomes, pour une taille totale de 16 gigabases. Les trois sous génomes se comportent comme autant de chromosomes différents lors de la méiose grâce au gène *Ph1* (*Paring Homologous 1*), qui empêche l'appariement entre chromosomes homéologues (Riley et Chapman 1958; Sears 1958). Le blé tendre est donc considéré fonctionnellement comme une espèce diploïde à 21 paires de chromosomes.

Depuis le bassin de domestication dans le croissant fertile, le blé tendre a été diffusé par les migrations humaines, ce qui s'est accompagné d'une différenciation génétique progressive et indépendante des populations de blé tendre de part et d'autre de l'axe eurasiatique. La diversité mondiale actuelle du blé tendre distingue donc deux groupes génétiques majeurs, regroupant les accessions européennes d'une part et les accessions asiatiques d'autre part (**Figure 2**). Chaque groupe est composé de plusieurs sous-groupes spécifiques résultant d'une différenciation spatio-temporelle (Balfourier et al. 2019). Balfourier et al. (2019) font la distinction entre les landraces (variétés locales cultivées dans leur zone d'origine), des variétés traditionnelles (inscrites avant 1960) et des variétés modernes (inscrites après 1960). Les variétés inscrites présenteraient un enrichissement en variations structurales, notamment causées par des introgressions de segments chromosomiques provenant d'espèces apparentées. Certaines comme les gènes *Pch1* issu *d'Aegilops ventricosa* ou *BYDV* issu de *Thinopyrum intermedium* confèrent des tolérances aux maladies, respectivement au piétin-verse ou la jaunisse nanisante (Huguet-Robert et al. 2001; Jahier et al. 2009).

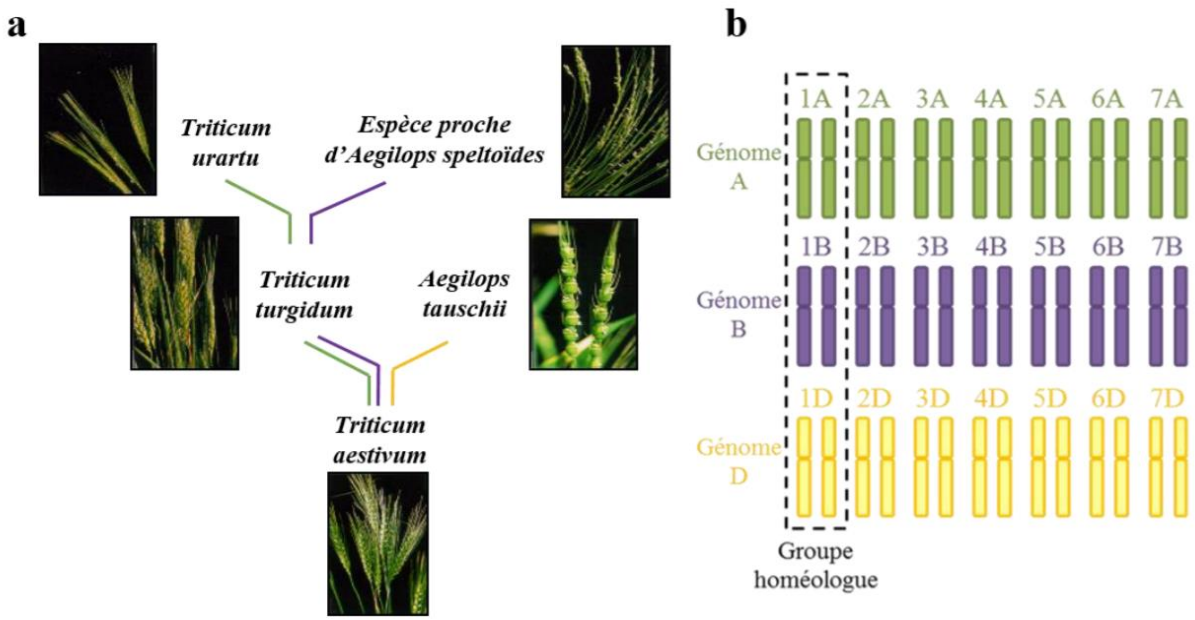


Figure 2 : Représentation schématique de l'histoire phylogénétique du blé tendre et de la structure de son génome

a. *Les événements d'hybridation et les génomes A (vert), B (violet) et D (orange) sont représentés sur la Figure (d'après Gill et al. 2004).*

b. *Le génome du blé tendre est structuré en trois sous-génomes A (vert), B (violet) et D (orange) diploïdes constitués de 7 paires de chromosomes.*

Ben Sadoun (2020)

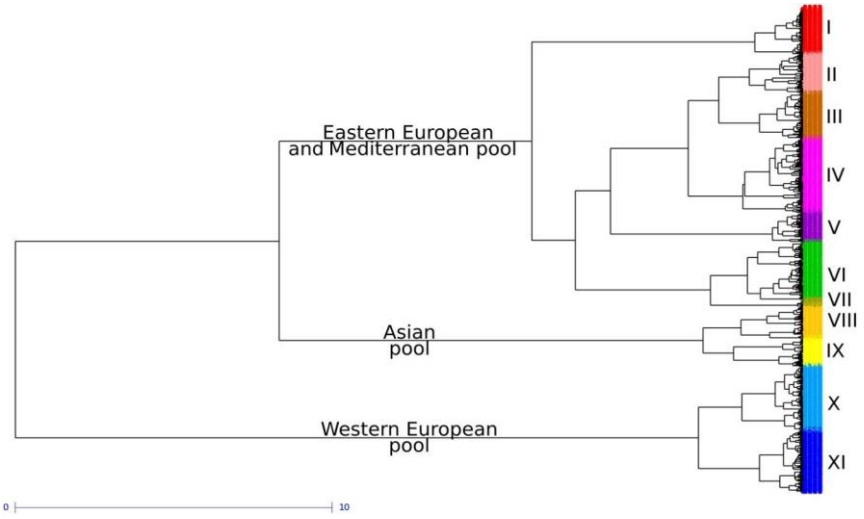


Figure 1 : Relations phylogénétiques entre 4403 accessions de blé tendre échantillonnées

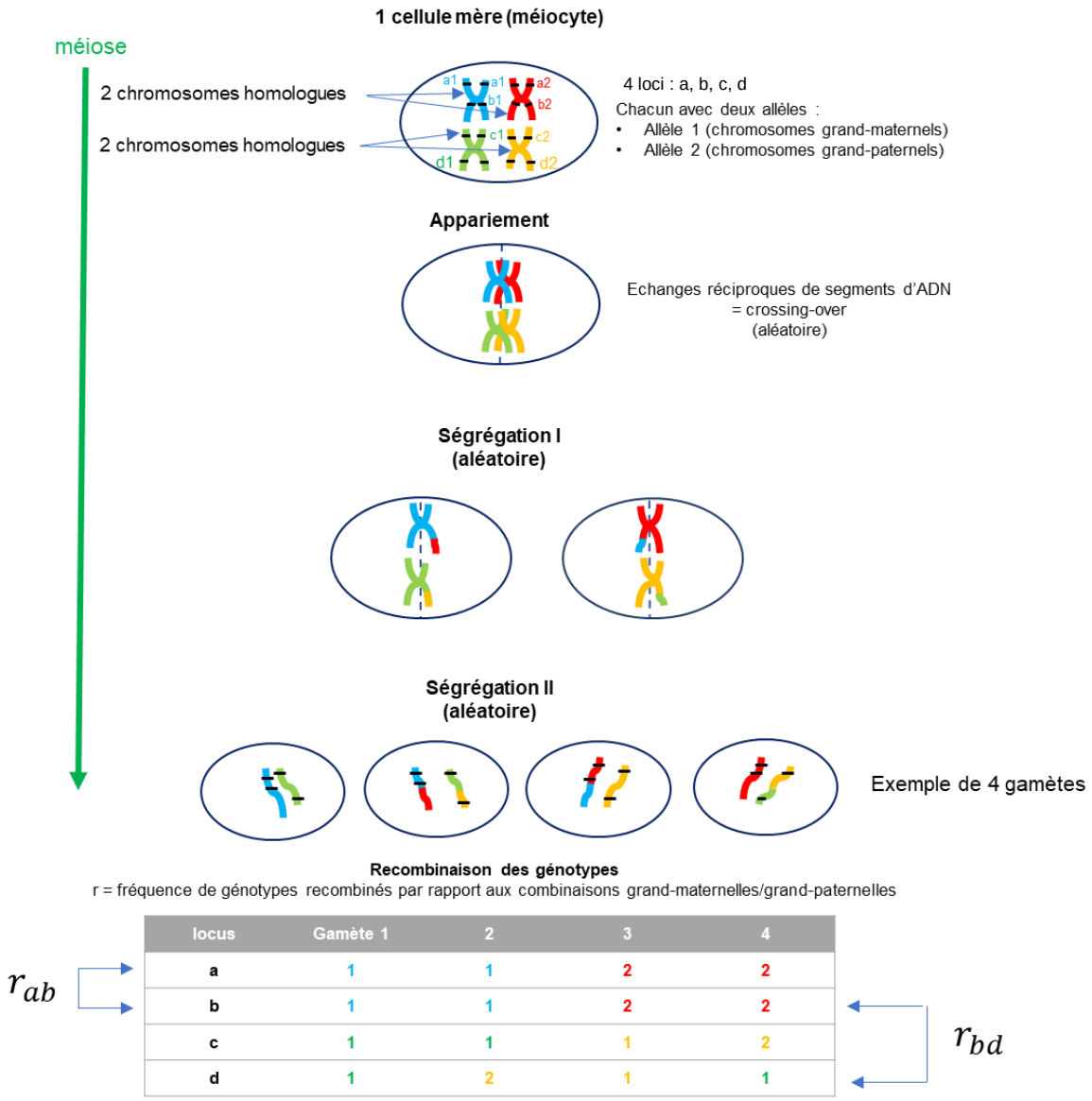
Ces 4403 accessions ont été échantillonnées pour représenter la diversité génétique mondiale géographique et historique.

Balfourier et al. (2019)

I.1.2 Principes généraux de l'amélioration variétale

I.1.2.1 La recombinaison comme moteur du progrès variétale

L'amélioration variétale du blé tendre consiste à proposer de nouvelles variétés élites à chaque génération en croisant des variétés ou des lignées du programme de sélection ou de concurrents. La différence entre les génotypes parentaux est les génotypes des nouvelles lignées est le résultat de deux processus : la recombinaison des génomes parentaux lors de la formation des gamètes durant la méiose et la fusion des gamètes lors de la fécondation. La recombinaison des génomes parentaux est de deux ordres : échantillonnage aléatoire des chromosomes homologues dans les gamètes et recombinaison entre chromosomes homologues (**Figure 3**).



locus	Gamète 1	2	3	4
a	1	1	2	2
b	1	1	2	2
c	1	1	1	2
d	1	2	1	1

Figure 3 : Recombinaison des génomes parentaux lors de la formation des gamètes

La lignée parentale possède deux paires de chromosomes homologues pour l'exemple : bleu/rouge et vert/jaune. Les chromosomes bleu et vert lui ont été transmis par sa mère, et les chromosomes rouge et jaune par son père. L'ADN de ces chromosomes est dupliqué en amont de la formation des gamètes, donc chaque chromosome possède deux brins identiques (chromatide sœurs). Les chromosomes homologues s'apparient en phase précoce de la méiose et les chromatides s'échangent de longues portions d'ADN (crossing-over, ou recombinaison homologue). Les chromatides (recombinées ou non recombinées) ségrégent ensuite indépendamment dans les gamètes durant la suite de la méiose, aux (étapes de ségrégation I et II). Au final, une cellule mère diploïde (méiocyte) se différencie en quatre gamètes haploïdes. La rencontre aléatoire des gamètes lors de la fécondation rétablit la diploïdie du nouvel embryon. Le terme « r » désigne la fréquence de génotypes recombinés par rapport aux combinaisons alléliques présentes sur les chromosomes du parent.

Le blé tendre est une espèce autogame avec plus de 95% des nouveaux individus issus d'autofécondation (Doré et Varoquaux 2006). La forte autogamie conduit traditionnellement à créer des variétés « lignées pures », c'est-à-dire homozygotes à 100% des loci. Le croisement entre lignées parentales homozygotes donne donc un hybride F1 qui est donc hétérozygote aux loci polymorphes. Il y a deux méthodes pour obtenir de nouvelles lignées homozygotes à partir d'une F1 (**Figure 4**). La plus classique (RILs) consiste à obtenir de nouvelles lignées par autofécondation de la F1 et de ses descendants pendant au moins six générations. L'hétérozygotie diminuant de moitié à chaque génération, les lignées recombinantes (RILs ; Recombinant Inbred Lines) sont quasiment fixées. Une alternative (HDs) consiste à regénérer une plante haploïde à partir d'un gamète (grain de pollen ou ovule) de la F1, puis de doubler le stock chromosomique de ses cellules avec de la colchicine. Ce processus est appelé haplo-diploïdisation et forme des descendants 100% homozygotes appelés haploïdes doublés (HD).

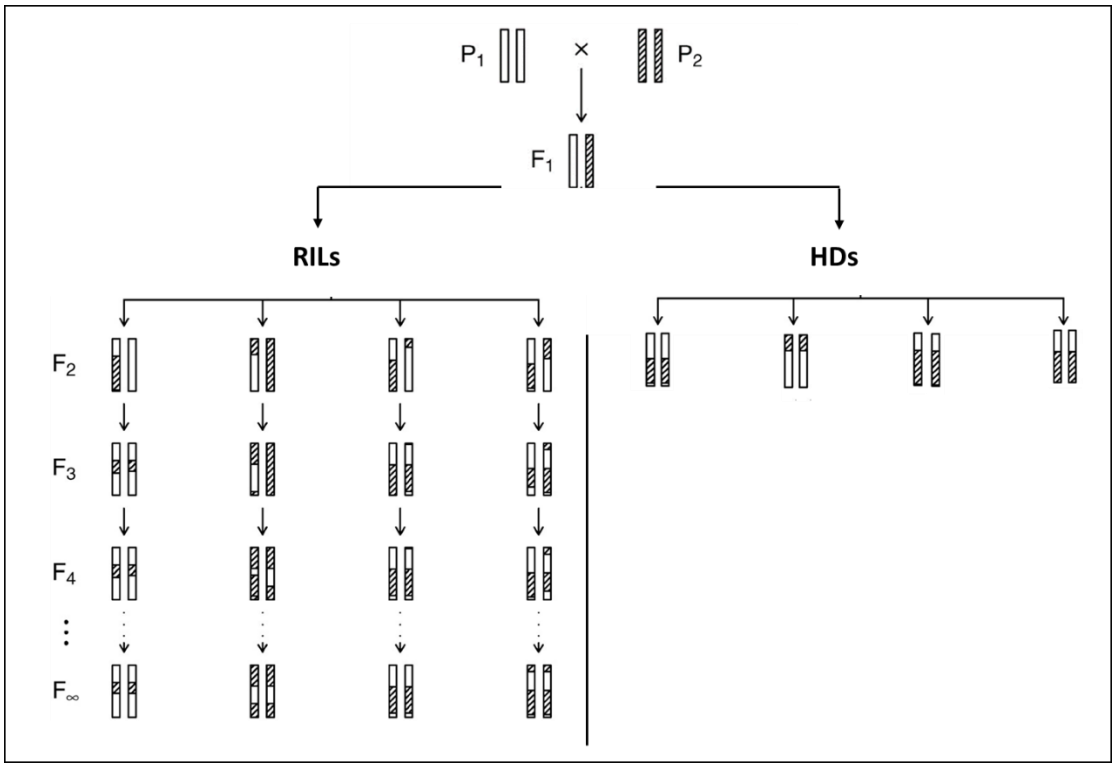


Figure 4 : Obtention de nouvelles lignées à partir d'un croisement

I.1.2.2 Choix des croisements

I.1.2.2.1 Progrès variétal

Les descendants (RILs ou HDs) montrent une recombinaison des patrimoines génétiques parentaux suite aux méioses successives. Or, certaines combinaisons alléliques sont plus avantageuses que d'autres. La valeur génétique additive de la nouvelle lignée l (RILs ou HDs) dépend de la valeur génétique de ses parents P_i et P_j, mais aussi de la recombinaison des génomes parentaux lors de la transmission du patrimoine génétique :

$$g_l = \frac{1}{2} g_{P_i} + \frac{1}{2} g_{P_j} + MS_l \qquad \textbf{[Equation 1]}$$

Où g_{P_i} et g_{P_j} sont les vraies valeurs génétiques additives (ou TBV pour True Breeding Values) des lignées parentales P_i et P_j, qui valent par définition la moitié de l'espérance de leur descendance (Jinks et Pooni 1976). Le terme $PM_{ij} = \frac{1}{2} g_{P_i} + \frac{1}{2} g_{P_j}$ donne donc l'espérance de la descendance.

La génétique quantitative a théorisé trois sources de variance génétique : la variance additive, la variance de dominance et la variance épistatique. La variance additive est causée par des allèles à effets additifs sur le caractère. Les variances de dominance et d'épistasie sont causées par des effets d'allèles en interaction. Pour la dominance, il s'agit d'une interaction entre allèles au même locus, et pour l'épistasie d'interaction entre allèles à différents loci. Il faut noter que dans le cadre de la sélection en lignée homozygote, les effets de dominance ne sont pas à considérer. Les effets additifs sont effectivement largement majoritaires dans l'élaboration de la valeur génétique du rendement du blé tendre (Nanda, Singh et Gill 1982; Singh, Bhullar et Gill 1986; Singh 1978). Plusieurs auteurs rapportent cependant que les effets épistatiques sont non négligeables chez le blé (Carl Friedrich Horst Longin et al. 2013) et que la séparation statistique des effets additifs et épistatiques reste difficile (Mackay 2014; Ignacy Misztal et al. 2021; de los Campos, Sorensen, et Toro 2019).

Le terme MS (pour Mendelian Sampling) de **l'équation 1** fait référence à la déviation entre la valeur génétique de la nouvelle lignée par rapport à celle de ses parents. Ce bonus (ou malus) par rapport aux parents est le résultat de la recombinaison unique des allèles parentaux suite au brassage génétique induit par la production des gamètes lors de la méiose. La variance du MS, aussi appelée variance de la descendance ${\sigma_{ij}}^2$, dépend de la complémentarité allélique des parents, mais aussi de la probabilité de recombiner les allèles parentaux dans la descendance.

Pour un trait polygénique, gouverné par un grand nombre de gènes à petits effets additifs, les valeurs génétiques des descendants (RILs ou HDs) d'un croisement entre les parents P_i et P_j, se distribuent selon une loi Normale centrée sur la moyenne des valeurs génétiques additives des

parents PM_{ij} et de variance ${\sigma_{ij}}^2$: N(PM_{ij}, ${\sigma_{ij}}^2$). Avoir accès à PM et σ² pour chaque croisement candidat permet d'identifier les croisements qui sont les plus prometteurs. Nous verrons qu'il y a de nombreuses façons de décrire l'utilité d'un croisement, mais globalement, les meilleurs croisements montrent une forte espérance PM et une forte variance σ² (**Figure 5**). Avoir accès à la variance d'un croisement est donc fondamentale pour évaluer son utilité et générer le plus grand progrès variétal.

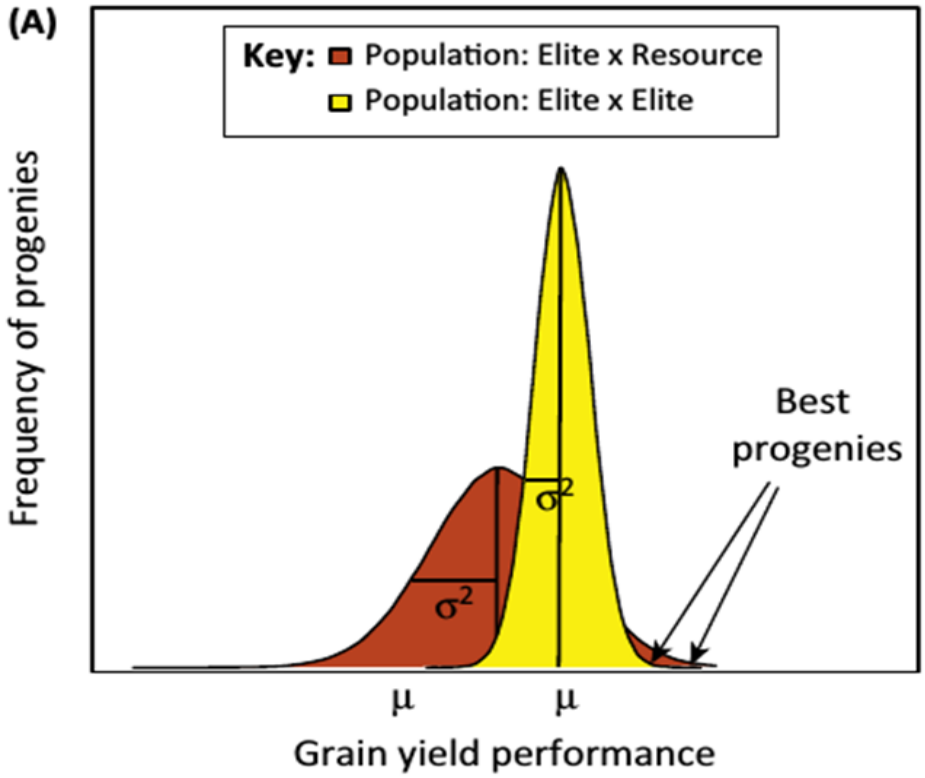


Figure 5 : Comparaison des distributions des descendants prédites pour deux croisements

Le croisement jaune montre une forte espérance (μ, ou PM) mais une faible variance σ², alors que le croisement rouge montre une faible espérance et une forte variance.

Longin et Reif (2014)

I.1.2.2.2 Amélioration des lignées parentales

Pour inscrire des variétés élites (« Best progenies » sur la **Figure 5**), le sélectionneur cherche donc à produire des descendants avec les couples ayant la meilleure combinaison d'espérance et de variance.

Ainsi, afin de proposer des variétés élites de plus en plus performantes au cours des générations, le programme de sélection cherche aussi à améliorer la valeur génétique des lignées parentales. Pour ce faire, le programme de sélection pratique la sélection récurrente, c'est-à-dire que les meilleurs descendants d'une génération sont recyclés comme nouveaux parents à la prochaine génération. Pour un caractère d'intérêt, la valeur génétique de la nouvelle population de parents est donnée par l'équation du sélectionneur (Lynch et Walsh 1998) :

$$G_{t+1} - G_{t+1} = (i r_{g,\hat{g}}\ \sigma_{A\,t})/L \qquad \textbf{[Equation 2]}$$

- Avec G_{t+1} = espérance des valeurs génétiques additives de la nouvelle population de parents
- G_t= moyenne des valeurs génétiques additives de la population actuelle de parents.
- i = intensité de sélection, qui augmente quand la proportion de descendants sélectionnés comme nouveaux parents diminue.
- $r_{g,\hat{g}}$ est la précision des estimateurs des valeurs génétiques additives des descendants (notées g)
- $\sigma_{A\,t}$= variance génétique (additive) dans la population actuelle de parents, qui se calcule comme la somme des variances additives à chaque locus et des covariances entre chaque paire de loci (Bulmer 1971), et reflète donc notamment le polymorphisme de la population parentale.
- L est le nombre d'années nécessaires pour obtenir la nouvelle population parentale, aussi appelé intervalle de génération.

Le programme de sélection doit donc ajuster ces 4 paramètres pour l'ensemble des caractères d'intérêt en fonction de son budget afin de maximiser la valeur génétique des nouvelles variétés élites. L'ajustement de ces paramètres est réalisé à deux étapes : le choix des croisements puis l'évaluation des descendants.

I.1.2.2.3 Amélioration multi-caractères

L'amélioration variétale cherche aussi à améliorer plusieurs caractères simultanément dans la nouvelle génération de variétés, avec une stratégie qui dépend des objectifs du sélectionneur et des attentes du marché. Le rendement fait partie des caractéristiques les plus importantes chez une variété pour les agriculteurs (France AgriMer 2015), mais d'autres phénotypes sont recherchés. Par exemple, les résistances aux stress biotiques (nématodes, piétin-verse), la qualité boulangère de la farine ou le taux de protéines dans la farine.

Cependant, certains caractères d'intérêt montrent une corrélation génétique négative, c'est-à-dire que l'amélioration génétique du caractère 1 s'accompagne d'une dégradation du caractère 2. C'est notamment le cas de la corrélation négative entre le rendement et le taux de protéines (Simmonds 1995), mais aussi entre le rendement et la qualité boulangère. Les blés qui ont une note de panification supérieure à 250 sont classés BPS (Blé Panifiable Supérieur) et doivent faire 102% du rendement des témoins pour permettre l'inscription au catalogue. Ceux qui ont une note comprise entre 220 et 250 sont BP (Blé Panifiable) et doivent faire 104% du rendement des témoins. Ceux dont la note est inférieure à 220 sont BAU (Blé Autre Usage) et doivent faire 107% du rendement des témoins (GEVES).

Le choix des croisements nécessite donc de définir une stratégie de sélection pour améliorer ces caractères négativement corrélés. Une première stratégie consiste à quantifier l'importance économique relative de chaque caractère d'intérêt et de définir un index de sélection, c'est-à-dire sélectionner les croisements sur un nouveau phénotype qui est une combinaison pondérée de chaque caractère d'intérê (Hazel 1943; Hazel et Lush 1942; Williams 1962; Wray et Goddard 1994). Une autre stratégie est de partitionner le plan de croisements, et de sélectionner dans un premier temps un ensemble de croisements susceptibles d'améliorer et/ou stabiliser un caractère, et dans un deuxième temps de sélectionner parmi cette liste de croisements candidats ceux qui permettent d'améliorer et/ou stabiliser un deuxième caractère (Jean et al. 2021).

I.1.2.3 Maximiser le progrès variétal lors du choix des croisements

Le plan de croisements est défini comme la répartition des descendants entre les différents croisements. Les croisements effectivement sélectionnés se voient attribuer un nombre de descendants non nul, alors que les croisements non sélectionnés se voient attribuer un nombre de descendants nul. Le choix du plan de croisements résulte de compromis entre différents objectifs.

I.1.2.3.1 Choisir les croisements pour produire de nouvelles variétés élites

Afin de maximiser l'espérance des nouvelles lignées (PM), les meilleures lignées sont traditionnellement croisées (Lupton 1961; Melchinger et al. 1998; Souza et Sorrells 1991; Kotzamanidis et al. 2008; Utz et al. 2001).

La supériorité des futurs descendants du croisement entre les lignées i et j par rapport à ces lignées parentales (mesurée par leur MS) dépend de la variance génétique de cette descendance $\sigma_{ij}{}^2$. Ainsi, de nombreuses études se sont attachées à prédire la variance de la descendance d'un croisement (compilées dans Osthushenrich (2019)).

Dans un premier temps, la génétique quantitative a permis de décomposer la variance de la descendance d'un couple en variance additive, de dominance et d'épistasie, estimables à partir des moyennes et des variances des phénotypes d'individus apparentés (lignées parentales, F1, F2, rétrocroisements …) (Jinks et Pooni 1976; Gallais 1989). Cependant les estimations arrivent tardivement dans le programme de sélection, puisque des croisements ont déjà été réalisés. Ceci explique que ces estimations aient été peu exploitées dans les programmes de sélection (Gallais 2011). Une alternative est proposée par Lian et al. (2015), qui suggèrent d'estimer la variance d'un couple à partir des variances moyennes observées préalablement dans la descendance de croisements impliquant l'un des deux parents. Si les parents ont été régulièrement utilisés par le passé dans le programme de sélection, c'est une estimation de la variance de la descendance à moindre coût.

Dans le cadre de la sélection animale, où la valeur génétique d'un reproducteur est souvent estimée à partir des performances de ses descendants, et non pas à partir de ses performances propres (exemple : valeur génétique pour la production laitière chez les taureaux laitiers estimés à partir des productions laitières de leurs filles), Van Raden et al. (1984) et Woolliams et Meuwissen (1993) ont proposé de sélectionner les reproducteurs avec une faible précision sur leurs valeurs génétiques estimées, cette faible précision traduisant éventuellement une faculté à produire une descendance très variable. Cependant, les bénéfices de cette méthode, et donc son application, étaient limités (Bijma et al. 2020).

Les méthodes précédemment décrites se basent sur le phénotypage des parents et des premiers descendants, et ne permettent donc pas de prédire précisément la variance de la descendance avant la réalisation du croisement. Elles ne s'appliquent donc pas aux croisements entre nouvelles lignées élites jamais testées en croisement, qui sont pourtant des candidats très prometteurs pour maximiser la valeur génétique des nouvelles variétés élites. Puisque la variance de la descendance dépend de la complémentarité génétique des parents, plusieurs études ont cherché à mettre en évidence une corrélation entre la variance de la descendance et une mesure de la distance génétique entre les parents. Dans un premier temps, certains auteurs ont tenté de corréler les variances des descendants avec les distances phénotypique (Utz et al. 2001). L'avènement des

marqueurs moléculaires a permis aussi de calculer des distances génotypiques entre lignées (Bohn et al. 1999). Cependant, les corrélations entre distances des parents et variance de la descendance sont faibles et dépendent du caractère (Osthushenrich 2019). Comme l'illustrent Beckett et al. (2019) (**Figure 6**), la différence génétique semble nécessaire mais pas suffisante pour générer beaucoup de variance dans la descendance. Dans cet exemple, le R^2 de la régression entre variance de la descendance et similarité génétique des parents n'est que de 0.16.

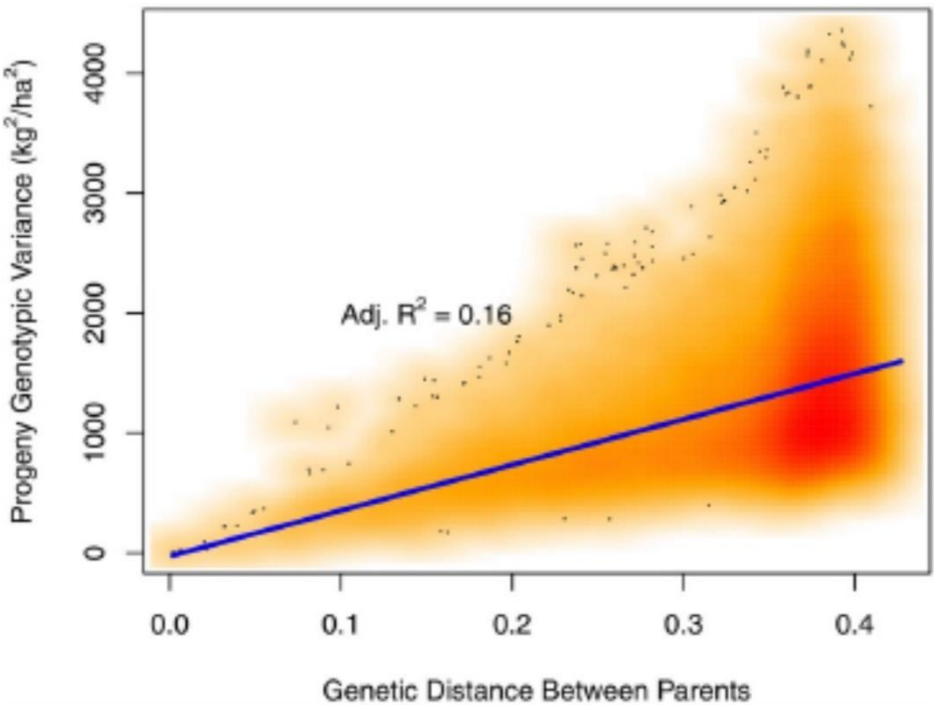


Figure 6 : Relation entre la différence génétique des parents et la variance de la descendance pour une population de maïs

Beckett et al. (2019)

Quoi qu'il en soit, croiser des reproducteurs génétiquement distants, voire issus de programmes concurrents ou banques de diversité génétique, limite le risque de diminution de variance génétique dans le programme de sélection et assure aux descendants de valider le test de Distinction (DHS) lors de l'inscription au catalogue.

I.1.2.3.2 Gestion de la variance génétique dans la population de lignées parentales

I.1.2.3.2.1 Perte de variance génétique dans les programmes de sélection

La sélection des lignées parentales qui vont contribuer à la prochaine génération s'accompagne d'une perte de diversité génétique. Dans une population en sélection fermée, c'est-à-dire dont les nouveaux parents sont uniquement choisis parmi les descendants, l'action combinée de la sélection des reproducteurs et de la dérive entraine une fixation progressive des allèles au fur et à mesure des cycles de sélection. D'une part, la sélection fait augmenter la fréquence des allèles favorables, mais cela peut entrainer la fixation de régions avoisinantes en déséquilibre de liaison par un phénomène d'entrainement (ou autostop génétique) (Smith et Haigh 1974). D'autre part, en échantillonnant une poignée de reproducteurs à chaque génération, la sélection réduit progressivement la taille efficace de la population (Sánchez et al. 2006; Santiago et Caballero 1998). De manière simplifiée, la taille efficace d'une population peut être définie comme celle d'une population théorique non soumise à la sélection, sans migration, dont les variations des fréquences alléliques et l'augmentation de la consanguinité entre générations sont similaires à celles observées dans la (vraie) population d'étude. Ces variations de fréquences alléliques sont uniquement dues à une variance du succès reproducteur, indépendamment d'effets sélectifs. Plus la taille efficace d'une population est petite, plus les variations de fréquences alléliques entre deux générations sont fortes, ce qui augmente le risque de faire converger les fréquences alléliques vers 0 ou 1. Or, la variance génétique additive se calcule comme la somme des variances additives à chaque locus et deux fois la somme des covariances additives entre locus (Bulmer 1971). Quand les fréquences alléliques tendent vers 0 ou 1, la variance génétique devient donc progressivement nulle, et par conséquent le gain génétique cumulé croit de moins en moins vite et finit par atteindre un plateau.

Ce n'est pas la perte de diversité génétique en soi qui est délétère pour l'amélioration du caractère d'intérêt, car toute sélection a pour objectif d'augmenter la fréquence des allèles favorables et donc de réduire la variabilité génétique. La perte de diversité génétique est indésirable lorsqu'elle est trop rapide et conduit à fixer des mutations délétères ou à perdre des allèles favorables (Woolliams et al. 2015), ce qui conduirait à limiter les possibilités de gain génétique. Par ailleurs, la réduction de la taille efficace diminue aussi les opportunités de voir apparaître des recombinaisons entre les allèles parentaux et des mutations intéressantes (Hill 1982b; 1982a). Enfin, la perte de diversité génétique concerne aussi des caractères non sélectionnés ou non prioritaires, limitant ainsi la possibilité de résilience face à de nouveaux objectifs de sélection, par exemple pour surmonter des stress biotiques et abiotiques émergents (McCouch et al. 2013). En conséquence, la perte de diversité génétique chez les plantes représente un risque pour le programme de sélection, mais aussi un ralentissement du gain génétique.

D'après ce modèle théorique précédemment décrit, où le caractère est déterminé par un nombre limité de variants causaux additifs, et en l'absence de mutation, ou de migration source de diversité, la variance génétique diminue progressivement jusqu'à devenir nulle. En conséquence, la courbe de croissance du progrès génétique ralentit puis finit par atteindre un plateau. Cette prédiction théorique est observée par simulation chez plusieurs auteurs (Jannink et al. 2010; Allier et al. 2019a; Goiffon et al. 2017; Gorjanc et al. 2018; De Beukelaer et al. 2017) (exemple dans la **Figure 7**). Dans ces simulations, la variance génétique est quasiment réduite à néant en un demi-siècle, ce qui signifie que le progrès génétique devient nul. En pratique, et notamment chez les plantes, une baisse du gain génétique n'a pas encore été observée sur des programmes de sélection relativement anciens (Duvick 2005; Tadesse et al. 2019; Dudley et Lambert 2004), probablement parce que des sources de variance non négligeables dans la réalité ne sont pas modélisées dans ces simulations. Une source de variance génétique non négligeable repose sur l'introduction régulière de géniteurs extérieurs (Allier, et al. 2019b). Une autre source de variance négligée sont les mutations ou l'épistasie (Hill 2017) , c'est-à-dire l'interaction d'allèles à différents locus (Carlborg et al. 2006). Ces simulations supposent aussi que le trait est déterminé par un nombre relativement restreint de variants causaux, aussi appelés QTLs (quelques milliers au plus). Or la variance génétique diminue plus lentement quand le nombre de QTLs augmente (Crow et Kimura 1969). Dans un cadre théorique où le nombre de QTLs et la population sont de taille infinie, la variance génétique finit même par se stabiliser au bout de quelques générations après la sélection (Bulmer 1971). Par ailleurs, le choix des reproducteurs vise à améliorer simultanément plusieurs caractères, éventuellement négativement corrélés, ce qui limite l'intensité de sélection sur un caractère, et donc freine la perte de diversité génétique par rapport à ce qui est observé dans les simulations.

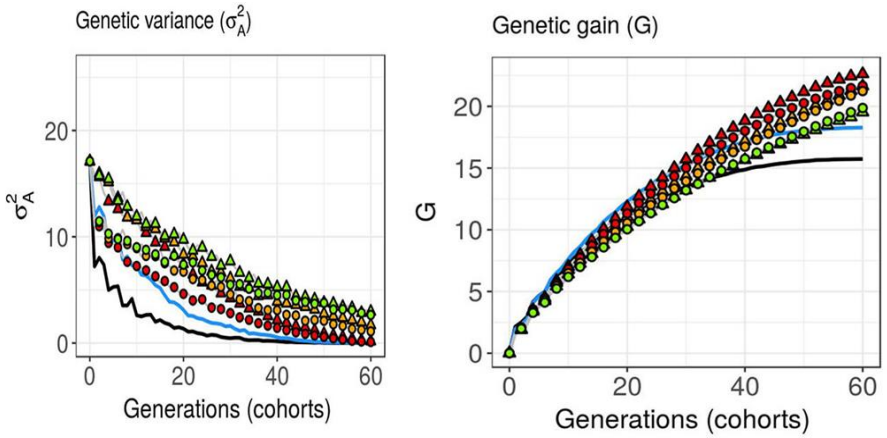


Figure 7 : Diminution de la variance génétique et de l'accroissement du gain génétique au cours de la vie d'un programme de sélection théorique

***A gauche** : Diminution de la variance génétique additive au cours des générations de sélection*

***A droite** : Réduction progressive du gain génétique*

Les différentes couleurs et formes de points correspondent à différentes méthodes pour contrôler l'apparentement des reproducteurs sélectionnés. En noir, choix des croisements sur PM sans contrôle de l'apparentement.

Allier et al. 2019a

I.1.2.1.3.2 Stratégies pour réintroduire de la variance génétique ou limiter la perte de variance génétique

Les programmes de sélection gèrent la diversité génétique de deux manières : la (ré)introduction de diversité génétique *ex situ* ou la gestion des croisements *in situ*.

I.1.2.1.3.2.1 Introduction de diversité ex situ

La (ré)Introduction d'une diversité utile consiste en général à utiliser des géniteurs extérieurs provenant de programmes de sélection concurrents en général, et plus rarement des banques de ressources génétiques, dans le plan de croisements (jusqu'à 50% du plan de croisements, Florimond Desprez, communication personnelle). Les ressources génétiques désignent des accessions aux propriétés intéressantes, de diverses origines géographiques ou cultivées à différentes époques, voire des accessions sauvages apparentées. Il existe plusieurs banques de ressources génétiques, telles que le Centre de Ressource Biologiques INRAE à Clermont-Ferrand, l'IPK (Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung) en Allemagne ou le CIMMYT au Mexique (International Maize and Wheat Improvement Center, https://www.cimmyt.org/). Ces banques de diversité représentent un grand réservoir d'allèles favorables utiles pour l'amélioration des plantes. Des gènes de résistance à la rouille (Kuraparthy et al. 2007; Steffenson et al. 2007; Ellis et al. 2014) et à la jaunisse nanisante (Jahier et al. 2009) ont été introgressés à partir d'espèces apparentées. Des gènes majeurs ayant une influence sur des caractères plus polygéniques, tels que la teneur en protéines, ont également été introgressés (Uauy et al. 2006).

Cependant, l'usage de ressources génétiques est laborieux chez le blé. L'un des premiers problèmes est que ces ressources génétiques présentent des écarts de rendement et de caractéristiques phénologiques trop importants avec les objectifs de sélection. Elles ne peuvent donc pas être utilisées directement comme géniteurs de variétés élites. Leur patrimoine génétique doit d'abord être adapté à l'agriculture moderne (pré-breeding). L'un des objectifs des programmes de pré-breeding en sélection est notamment de mettre à niveau ces ressources génétiques en les croisant avec des variétés élites sur plusieurs générations, et de trier les descendants. Par exemple, le programme de sélection du CIMMYT met à disposition de nouveaux géniteurs dérivés de ressources génétiques. Cependant, ce matériel est adapté pour du blé de printemps, et nécessite un travail supplémentaire de conversion pour être cultivé en tant que blé d'hiver en France.

Un autre problème est la difficulté d'évaluer ces ressources génétiques à cause de la présence d'allèles délétères pour des gènes majeurs contrôlant la phénologie (hauteur, date d'épiaison) et provoquant la verse au champ, de la stérilité ou des difficultés à terminer leur cycle à temps. La caractérisation des ressources génétiques reste encore un grand défi pour la communauté scientifique et les programmes de sélection (Wang et al. 2017). Pour pallier à ce problème, Longin

et Reif (2014) proposent d'évaluer ces accessions sous forme hybride, c'est-à-dire croisées avec des lignées élites qui possèdent les allèles favorables dominants à ces loci.

Un troisième problème est la perte des allèles favorables du parent « ressource génétique » lors des étapes de croisements-sélection durant le pré-breeding. Le fond génétique du parent ressource contient en effet en général peu d'allèles favorables et beaucoup d'allèles délétères. Après croisement et sélection, le risque est de ne retenir que les « meilleurs descendants » du croisement, et donc de revenir rapidement au génome du parent élite et de finalement ne réintroduire que très peu de diversité (Gorjanc et al. 2018; Fradgley et al. 2019). De plus, il faut s'assurer que cette diversité favorable « retenue » ne soit pas déjà présente dans le matériel de sélection (Yang et al. 2020).

En conséquence, l'introduction de diversité par pré-breeding et bridging est une tâche laborieuse et coûteuse, d'autant plus lorsque les ressources génétiques présentent un écart de performance important avec le matériel élite, ce qui est le cas en blé. Ce travail de longue haleine est peut-être du ressort de la recherche publique pour assurer un gain génétique sur le long terme.

Pour un gain génétique à court terme, les sélectionneurs s'intéressent donc à gérer la diversité génétique de leur matériel de sélection propre. Chez le blé tendre, cela implique éventuellement d'utiliser des variétés commerciales développées par les concurrents comme lignées parentales. La conversion optimale de la diversité génétique en progrès génétique nécessite cependant de freiner la perte de diversité génétique spécifique du matériel de sélection à chaque cycle de sélection.

I.1.2.1.3.2.2 Gestion de la diversité in situ

Le gain génétique, c'est-à-dire l'augmentation de la valeur génétique moyenne de la population parentale, dépend de la variance génétique de la population parentale (**Equation 2**, paragraphe I.1.2.2.2). Limiter la perte de variance génétique implique de contrôler/restreindre l'apparentement moyen des reproducteurs (F) (Meuwissen 1997; Wray et Goddard 1994; Woolliams et al. 2015). La pertinence d'un contrôle de F pour limiter la diminution de la variance génétique $\Delta\sigma^2_A$ s'appuie sur la relation négative entre l'augmentation de l'apparentement ΔF et la diminution de la variance génétique $\Delta\sigma^2_A$ (Falconer et Mackay 1966):

$$\Delta\sigma^2_A = -\Delta F^*\sigma^2_{A\,0}$$

L'augmentation ΔF à long terme est déterminée par la somme des contributions (au carré) de chaque reproducteur à la population de descendants : $\Delta F = ¼ \sum_i r_i^2$ (Wray et Thompson 1990), où les r_i donnent les contributions à long terme des individus. Ainsi, limiter ΔF nécessite de répartir la descendance entre les reproducteurs, mais aussi de limiter l'apparentement des parents sélectionnés (Meuwissen 1997). L'apparentement peut se calculer à partir des pedigrees des

parents, avec la quantité F = c Φ c' où c représente le vecteur de la contribution de chaque parent au plan de croisements et Φ est la matrice d'apparentement entre les parents, initialement calculée à partir du pedigree.

Concrètement, un contrôle simple de l'augmentation de l'apparentement consiste à éviter de croiser les parents trop apparentés, et de répartir les descendants entre plusieurs parents. Chez les animaux, en particulier les bovins laitiers, des méthodes plus sophistiquées pour contrôler l'apparentement de l'ensemble des reproducteurs ont été développées. Ces approches ont été récemment adaptées à la sélection variétale (Allier et al. 2019a; Akdemir et al. 2019; Gorjanc et al. 2018).

Ces méthodes supposent que chaque plan de croisements possible est associé à un gain génétique ΔG et à une variation de l'apparentement ΔF (Akdemir et al. 2019). L'ensemble des plans de croisements qui forment les meilleurs compromis de ΔG et ΔF, pour chaque niveau de ΔG et ΔF, forment le front de Pareto, c'est-à-dire l'ensemble des points pour lesquels il n'existe pas de solution alternative permettant d'améliorer un objectif sans en dégrader un autre. Un exemple de Front de Pareto est donné à la **Figure 8**, pour l'optimisation de trois objectifs : progrès

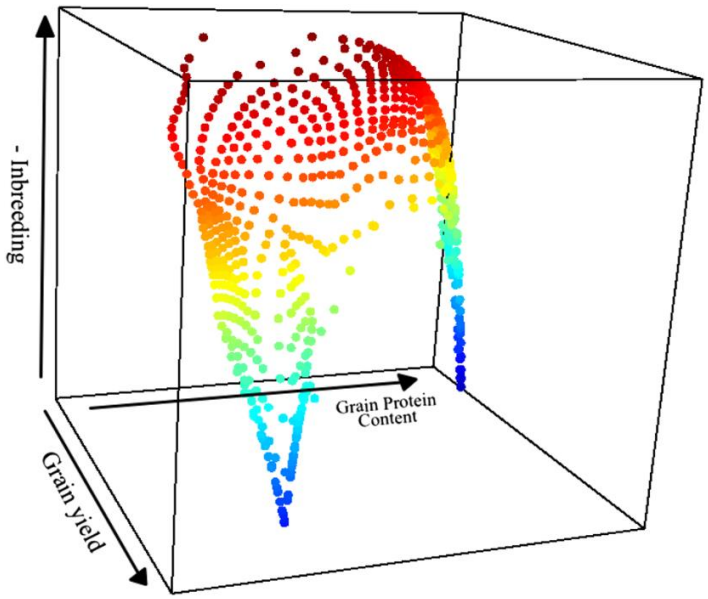


Figure 8 : Exemple de surface de Pareto lors du choix du plan de croisements de blé tendre

Un point = un plan de croisements = un progrès génétique en termes de rendement, de teneur en protéines ou d'augmentation de l'apparentement. Ces trois caractères montrent une covariance négative. Les plans de croisements sur cette surface sont les meilleurs compromis sur ces trois caractères. Tout point en dehors de la surface est impossible ou sous-optimal pour au moins l'un des trois caractères.

Akdemir et al. (2019)

génétique sur le rendement, sur le taux de protéines et sur l'augmentation de l'apparentement. Il s'agit ensuite de choisir un plan de croisements parmi le front de Pareto.

Choisir un plan de croisements sur le front de Pareto n'est pas trivial. En effet, une décision à la génération t a des conséquences difficilement prévisibles aux générations suivantes, que ce soit en termes de gain ou de variance génétique. En théorie, on souhaite choisir les plans de croisements à chaque génération de manière à maximiser le gain génétique sur l'ensemble de la vie du programme de sélection. On peut définir le gain commercial comme la somme des gains génétiques à chaque génération. Il est possible de nuancer ce gain commercial en pondérant les gains successifs de manière à privilégier le gain à court terme, indispensable pour la survie économique du programme de sélection (Allier et al. 2019a; Chakraborty et al. 2002; Dekkers, Birke et Gibson 1995).

Une méthode couramment employée pour choisir le plan de croisements est de fixer une trajectoire de F au cours des générations jusqu'à l'horizon du programme de sélection : $c_t \, \Phi_t \, c_t' < F_t$, puis de maximiser en c_t, vecteur des contributions de chaque parent au plan de croisements, le gain $\Delta G = c_t g_t$ sous cette contrainte, où g_t est la valeur génétique des reproducteurs (Meuwissen 1997; Woolliams et al. 2015; Allier et al. 2019a; Gorjanc et al. 2018). Cette méthode initialement appelée Optimal Contribution Selection (OCS) a été développée chez les bovins laitiers afin d'optimiser les contributions des parents à la descendance. Une extension, l'Optimal Cross Selection, permet de répartir les descendants entre les croisements sachant la contrainte sur F. Dans ce manuscrit, ces méthodes sont communément nommées OCS. La difficulté des OCS est de déterminer le seuil F_t à chaque génération. Par exemple, la FAO suggère que l'augmentation de l'apparentement par génération ΔF_t ne doit pas excéder 1% chez les animaux (FAO 1998), ce qui permet de maintenir une taille de population efficace de l'ordre de Ne=50. Une alternative est de choisir la trajectoire de F_t grâce à des simulations sur l'évolution du programme de sélection (Allier et al. 2019a).

Fixer un seuil F_t admissible à chaque génération présente aussi l'avantage de simplifier énormément le problème calculatoire (supplementary material de Akdemir et al. 2019). En effet, la frontière de Pareto est éventuellement très dense et requiert de moyens calculatoires importants pour identifier chaque point (voire **Aparté**).

Aparté : comment optimiser les plans de croisements à l'aide d'outils informatiques ?

Plusieurs algorithmes permettent d'identifier le plan de croisements qui maximise le gain génétique tout en minimisant l'apparentement des reproducteurs, ou toute autre combinaison d'objectifs.

Un problème d'optimisation classique comporte 1) des variables à ajuster 2) une fonction Objectif à maximiser ou à minimiser, dont la valeur dépend des variables à ajuster et 3) des contraintes sur les variables à ajuster.

Dans le cadre de ce manuscrit, les variables à ajuster sont la proportion de descendants alloués à chaque parent ou à chaque croisement (c_t), la fonction Objectif inclue deux objectifs, à savoir maximiser le gain génétique et minimiser l'apparentement des reproducteurs : $\alpha * c_t g_t - \gamma^* c_t \, \Phi_t c_t'$ où α et γ sont les pondérations données à chaque Objectif. Enfin, des contraintes s'appliquent aux c_t (par exemple, le nombre de descendants par reproducteur est limité).

Dans les OCS, la fonction Objectif est simplifiée par $c_t G_t$, et l'apparentement est relégué dans les contraintes : $c_t \ \Phi_t \, c_t' < F_t$ (méthode de simplification nommée ε-constraint, (supplementary material de Akdemir et al. 2019). Ce système d'équations étant linéaire pour les variables à ajuster, il est possible d'utiliser la programmation linéaire (Jansen et Wilton 1985; Weigel et Lin 2000), qui garantit de trouver le maximum global pour la fonction objectif sachant les contraintes.

Si la fonction Objectif inclue plusieurs objectifs, ou plus globalement si le système d'équations n'est pas linéaire ou quadratique pour les variables à ajuster, il est possible d'utiliser des algorithmes heuristiques pour trouver une solution de bonne qualité (ou un ensemble de solutions), sans garantie d'optimalité. Les algorithmes génétiques (GA) font partie des algorithmes heuristiques inspirés de la sélection naturelle telle que décrite par Darwin (Goldberg 1989). Les algorithmes génétiques commencent par définir une population de solutions candidates. Dans notre exemple, une solution est un plan de croisements, caractérisée par un vecteur c_t respectant les contraintes et associée à une valeur numérique pour la fonction Objectif. Les solutions associées aux meilleures fonctions Objectifs sont conservées (sélectionnées) pour l'itération suivante. Une certaine proportion de ces solutions sélectionnées sont « recombinées », c'est-à-dire qu'elles échangent une partie de leur vecteur c_t pour créer de nouvelles solutions. Une autre proportion de solutions sélectionnées sont mutées, c'est-à-dire qu'une partie de leur vecteur c_t est modifié. Cet ensemble de solutions sélectionnées, et éventuellement recombinées ou mutées, forment une nouvelle population qui est de nouveau sélectionnée. Au fur et à mesure des itérations, la population de solutions s'améliore. Afin de limiter les risques de converger vers des optimums locaux, autrement dit afin de mieux explorer l'espace des possibles, les valeurs Objectifs des solutions peuvent être recalculées suivant un processus de « sharing » qui offre un bonus aux solutions isolées dans l'espace des possibles (Xiaodong Yin et Germay 1993).

I.1.2.4 Production de nouvelles lignées

Une fois le plan de croisements établi, les lignées parentales sont croisées pour former des descendants. Pour rappel (**Figure 4**), il existe deux méthodes principales pour produire de nouvelles lignées à partir du croisement de deux lignées parentales : les RILs (autofécondation du F1 et de ses descendants) ou les HDs (doublement chromosomique d'un gamète de F1, ou haplo-diploïdisation).

Chaque alternative a ses avantages et inconvénients. L'obtention de nouvelles lignées par la méthode RILs nécessite de 6 à 10 générations d'autofécondations pour obtenir les descendants quasiment homozygotes, alors que l'haplo-diploïdisation produit des descendants homozygotes en deux ans. Les RILs sont individuellement moins chères à produire que les HDs, mais les années supplémentaires d'autofécondation augmentent leur coût final, si bien qu'une investigation complète serait nécessaire pour comparer le coût économique des deux méthodes. Par ailleurs, le nombre de recombinaisons efficaces est approximativement deux fois plus grand chez les RILs que chez les HDs (Haldane 1919), ce qui a une conséquence sur la variance de la descendance et par extension sur le gain génétique. Par contre, certains génotypes répondent mal à l'haplo-diploidisation. En général, les sélectionneurs utilisent les deux méthodes, et la méthode RILs est appliquée aux croisements récalcitrants à l'haplo-diploïdisation.

I.1.2.5 Évaluation des lignées par phénotypage

Dans le modèle de Fisher (1930) largement utilisé en amélioration génétique, le phénotype (P) d'un individu est le résultat de sa valeur génétique (G), d'effet environnementaux (E) et d'interactions entre le génotype et l'environnement (GxE).

$$P = G + E + GxE$$

La sélection vise à améliorer le phénotype en améliorant la valeur génétique des variétés. Pour prendre en compte les interactions GxE, la sélection est organisée par grandes zones géographiques et parfois pour des objectifs ou des environnements très précis (TPE : Target Population Environnements, Cooper et al. 1997). Comme ces environnements sont de mieux en mieux caractérisés, des modèles sont proposés pour prédire le GxE en les classifiant ou en utilisant des modèles écophysiologiques permettant de prendre en compte les différents niveaux de stress.

La part relative des effets génétiques et environnementaux dans le phénotype est mesurée par l'héritabilité (au sens large) :

$$h^2 = V(G)/V(P)$$

Puisque les phénotypes résultent d'effets environnementaux et d'effets génétiques, estimer la valeur génétique des nouveaux individus nécessite de les tester dans plusieurs environnements

(combinaisons lieux/années) avec un protocole expérimental qui limite la confusion entre valeur génétique, effets environnementaux et interactions entre génotype et environnement. Ce protocole expérimental est appelé réseau d'essais (ou MET pour Multi Environnement Trialing) et contient des environnements représentatifs des TPE. Les lignées de blé tendre sont multipliées par autofécondation, et le réseau d'essais contient plusieurs répétitions du même génotype au sein d'un même essai.

L'héritabilité, le nombre de réplicats, le protocole expérimental ainsi que le modèle statistique utilisé font parties des facteurs qui améliorent la précision des estimateurs des valeurs génétiques (Lynch et Walsh 1998; Cullis et al. 2020) et donc la précision de la sélection $r_{g,\hat{g}}$ qui intervient dans l'équation du sélectionneur (**Equation 2**, **Chapitre I.1.2.2.2**). Dans le cadre de descendants RILs, la sélection et la multiplication des meilleurs individus s'opère à chaque génération. Les premières générations (F1 - F4) sont en général conduites en serre ou en pépinières dans un seul lieu, ce qui permet d'opérer une sélection drastique sur la tolérance aux maladies, sur la hauteur et la précocité. Le taux de sélection moyen est de 10 à 20 % par famille, une famille regroupant l'ensemble des descendants d'un couple. Certaines familles sont éliminées. La quantité de semences (nombre de clones par lignée) est multipliée à chaque génération pour pouvoir faire des parcelles d'essais. Aux générations suivantes (F5 et au-delà), les descendants sélectionnés (peu nombreux) sont étudiés dans de plus en plus de lieux pour caractériser les interactions GxE. Le rendement et le taux de protéines sont mesurés. Les caractères encore plus coûteux à phénotyper comme la note de panification (Ben Sadoun et al. 2020) sont mesurés le plus tard possible avant les essais CTPS, quand le nombre d'individus est fortement réduit suite à la sélection. Les meilleurs individus seront utilisés comme parents dans le cycle suivant et certains sont proposés à l'inscription au catalogue.

Estimer des valeurs génétiques précises pour les candidats à l'inscription implique de les tester dans un réseau d'essais composé de nombreux environnements. La démultiplication des lieux et des phénotypes testés engendre un coût supplémentaire élevé. Ces efforts sont donc concentrés sur les meilleurs descendants, ce qui nécessite de pouvoir appliquer une sélection très forte dès les premières années. Les modèles de prédiction génomique permettent de prédire la valeur génétique des descendants sur la base de marqueurs, à un stade relativement précoce (F5-F6) et sans nécessité de les phénotyper.

I.2 Apport de la prédiction génomique

1.2.1 Evaluation de la valeur génétique des lignées

1.2.1.1 Modélisation du déterminisme génétique

Le développement des marqueurs moléculaires permet désormais de sélectionner les descendants sur la base de leur génotype (synthèse dans Heffner et al. (2009). Dans un premier temps, la sélection assistée par marqueur (SAM) a permis de sélectionner les individus sur leur génotypage à quelques loci impliqués dans l'élaboration de la valeur génétique du caractère. Le principe de base de la SAM est d'exploiter le phénomène, de co-transmission entre allèles aux marqueurs (mesurables) et variants causaux proches (QTLs non accessibles), due aux liaisons génétiques. Ces QTLs nécessitent donc d'avoir été détectés préalablement par voie expérimentale, par exemple sur des populations de lignées recombinantes dont les parents montrent une grande variabilité phénotypique. Cependant, la détection de QTLs pour les caractères polygéniques tels que le rendement est difficile. En effet, le rendement est considéré comme un caractère obéissant au modèle génétique infinitésimal, c'est-à-dire contrôlé par un grand nombre de QTLs à petits effets (Fisher 1915), ce qui limite la puissance de détection. Par ailleurs, les QTLs détectés dans des populations expérimentales montrent une faible valeur prédictive dans d'autres populations (exemple dans Moreau et al. 2004), notamment à cause des fonds génétiques différents et des effets d'interaction génotypes-environnements différents entre la population expérimentale et la population de sélection (Heffner et al. 2009; Millet et al. 2016). Ainsi, la SAM ne semble pas adaptée à l'amélioration de caractères quantitatifs (Dekkers et Hospital 2002) tels que le rendement.

Les modèles de prédiction génomique proposés par Meuwissen et al. (2001) et Whittaker, et al. (2000) exploitent le phénomène de déséquilibre de liaison (DL) populationnel (et non plus seulement intra famille) entre loci proches. Ce DL est l'association statistique d'allèles à différents loci : certaines combinaisons d'allèles sont plus fréquentes que d'autres dans la population. Le DL dans une population est le produit des forces évolutives et de la recombinaison (**Chapitre I.3.1**). Comme dans la SAM on utilise le modèle proposé par Fisher (1915) dans lequel la ressemblance phénotypique entre individus résulte d'un partage d'allèles en des QTLs affectant le caractère. Les positions des QTLs et les effets de leurs allèles sur le phénotype sont généralement inconnus, mais en génotypant un grand nombre de marqueurs on peut exploiter ce déséquilibre de liaison (DL) entre allèles causaux et allèles aux marqueurs en « capturant » l'effet des allèles aux QTLs dans des modèles statistiques appropriés.

Les modèles de prédiction génomique classiques modélisent la variance génétique additive, sans prendre en compte les interactions entre allèles (dominance, épistasie). Plusieurs auteurs

rapportent cependant que la séparation statistique des effets additifs et épistatiques reste difficile. La confusion entre ces effets peut aller dans les deux sens. Certains effets perçus comme additifs sont en fait épistatiques (Mackay 2014), ou à l'inverse un déséquilibre de liaison imparfait entre allèles additifs et marqueurs peut être interprété comme un effet épistatique (Misztal et al. 2021; de los Campos et al. 2019). Par ailleurs, les modèles de prédiction génomique basés sur les effets additifs sont beaucoup moins coûteux en temps de calcul que des modèles incluant les interactions de dominance ou d'épistasie car ils estiment un moins grand nombre de paramètres (Jiang et Reif 2015), et peuvent donc s'ajuster sur des plus petits jeux de données (Varona et al. 2018).

1.2.1.2 Modèles de prédiction génomique

Le modèle de prédiction génomique le plus couramment utilisé est le GBLUP (Genomic Best Linear Unbiaised Prediction ; Habier et al. 2013), qui dérive historiquement du BLUP « modèle animal » (Henderson 1975). Ce modèle linéaire mixte permet d'estimer les valeurs génétiques additives des individus en exploitant une matrice d'apparentement basée sur les génotypes des individus :

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{X\alpha} + \boldsymbol{g} + \boldsymbol{e}$$

Avec

- **Y** le vecteur des phénotypes de la population d'entrainement de longueur n, où n est le nombre d'observations qui peut être supérieur au nombre de lignées i dans le cas de réplicats (exemple : le rendement d'une lignée testé sur plusieurs parcelles)
- **μ** le phénotype moyen
- $\boldsymbol{\alpha}$ le vecteur des effets fixes, typiquement des facteurs environnementaux, avec **X** la matrice d'incidence associée
- **g** le vecteur des valeurs génétiques, modélisé tel que **g** ~ N(0, Gσ_g^2) avec σ_g^2 la variance génétique additive, G la matrice d'apparentement. Le plus souvent, G est calculée comme **WW**'/2$\sum_{l=1}^{L} p_l(1-p_l)$ avec p_i la fréquence de l'allèle au locus i et W la matrice d'incidence des allèles, de dimension I*L (L est le nombre de marqueurs), avec $\boldsymbol{W_{il}}$ le génotype d'un individu i au marqueur l, codé 0, 1 ou 2 selon le nombre d'allèles minoritaires au marqueur, auquel est retranché l'espérance du génotype moyen au marqueur (moyenne calculée sur tous les individus) (VanRaden 2008). D'autres propositions ont été faites pour tenir compte du DL entre marqueurs (Speed et al. 2012), ou encore de combiner la matrice G avec une matrice d'apparentement calculée sur pedigree (Legarra et al. 2009).
- **e** le vecteur des résidus, c'est-à-dire la part de variation non expliquée par le modèle, par exemple due à des effets environnementaux non contrôlés ou à des effets génétiques non additifs. Les résidus e se modélisent tel que e ~ N(0, Iσ$_e^2$).

Les valeurs génétiques estimées $\hat{\mathbf{g}}$ des individus sont appelées GEBV (Genomic Estimated Breeding Values) et peuvent être obtenues directement à partir des méthodes de résolution des modèles mixtes.

Un autre modèle très utilisé est le Ridge Regression BLUP (RRBLUP) qui s'écrit :

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{X\alpha} + \boldsymbol{W\beta} + \boldsymbol{e}$$

Où β est un vecteur des effets des SNPs, considérés comme aléatoire centrés sur 0 et de même variance σ_β^2, et $\boldsymbol{W}$ la matrice d'incidence des allèles des individus.

Les deux modèles, en considérant que les GEBV s'écrivent $\hat{g}_i = \sum_j w_{ij}\hat{\beta}_j$ sont équivalents (Goddard et al. 2011). Dans le modèle RRBLUP, les estimateurs des effets SNPs sont calculés comme

$$\hat{\boldsymbol{\beta}} = (\mathbf{W'W} + \boldsymbol{\lambda}\mathbf{I})^{-1}\,\mathbf{W'(Y - 1}\hat{\boldsymbol{\mu}}\mathbf{ - X}\hat{\boldsymbol{\alpha}}\mathbf{)}$$

Où $\boldsymbol{\lambda}$ est le paramètre de shrinkage. Dans le cas du GBLUP, les estimations des effets des marqueurs $\hat{\boldsymbol{\beta}}$ peuvent aussi être obtenus par une procédure dite de « backsolving » impliquant les GEBV, la matrice **G**, l'estimateur de la variance génétique additive $\widehat{\sigma_g^2}$ ainsi de la matrice d'incidence des allèles **W** (Legarra et al. 2014; Wang et al. 2012).

1.2.1.3 Efficacité de la prédiction génomique

La prédiction génomique permet donc de prédire la valeur génétique d'individus génotypés mais non phénotypés (population candidate), à partir des génotypes et des phénotypes d'individus génotypés et phénotypés (population d'entrainement) (**Figure 9**).

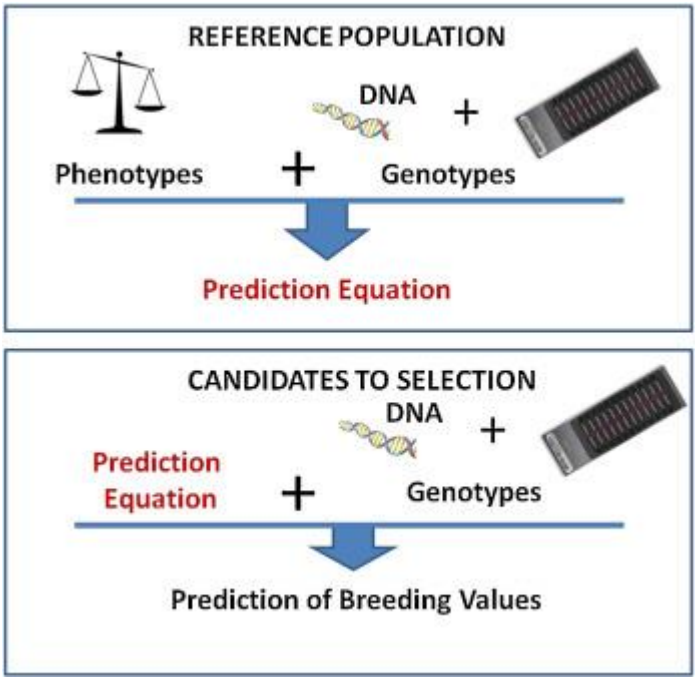


Figure 9 : Fonctionnement de la prédiction génomique

Dans un premier temps, le modèle de prédiction génomique est calibré à partir des phénotypes et génotypes (obtenus via le génotypage sur puce par exemple).
Dans un deuxième temps, les paramètres estimés servent à prédire les valeurs génétiques (GEBV) des candidats à la sélection.

Boichard et al. (2016)

L'intérêt de remplacer le phénotypage par la prédiction génomique dépend notamment de la précision de prédiction du modèle (accuracy), c'est-à-dire la corrélation entre les valeurs génétiques réelles (TBV ; True Breeding Values) et les GEBV. Cependant, les TBV sont inconnues, sauf pour des données simulées. Ainsi, on calcule généralement la capacité de prédiction (predictive ability) comme la corrélation entre les GEBV et les moyennes ajustées des phénotypes, qui sont tous deux des estimateurs des TBV (Bassi et al. 2016). Globalement, la predictive ability pour le rendement en blé tendre est de l'ordre de 0.4 (de los Campos et al. 2009; Heslot et al. 2012; Rincent et al. 2018; Crossa et al. 2010; Michel et al. 2016; He et al. 2016). Historiquement, cette « prédictive ability » est calculée à partir des données d'une population dite de validation, génotypée et phénotypée, sur laquelle on applique les équations de prédiction génomique établies

à partir des données de la population d'entrainement, dont les résultats sont alors confrontés aux observations phénotypiques. En l'absence de population de validation, des validations croisées sont réalisées avec la population d'entraînement en utilisant 4/5ème de cette population pour construire le modèle et 1/5ème pour le valider. Par ailleurs, les simulations se sont aussi avérées utiles pour comparer les estimateurs phénotypiques et génomiques. Par exemple, Zhong et al. (2009) ont simulé un programme de sélection et les TBV des lignées. La précision de la prédiction génomique ou du phénotypage sur les TBV étaient similaires.

Ainsi, la prédiction génomique présente une capacité de prédiction intéressante chez les plantes pour éliminer les individus les moins prometteurs, et ne requiert que le génotypage de la population candidate et l'utilisation d'un modèle statistique approprié. Son intérêt économique est donc lié au coût du génotypage, qui est en constante diminution (Poland et Rife 2012). Par exemple, (Bassi et al. 2016) indiquent que le coût du génotypage d'un individu est compris entre 12 et 50$/lignée pour génotyper entre 10k et 20k marqueurs polymorphes. En comparaison, le coût du phénotypage est de l'ordre de 10-100$/lignée, ceci dépendant du stade d'avancement, autrement dit des phénotypes mesurés et du nombre d'environnements et de répétitions.

Dans la mesure où les valeurs génétiques des lignées sont correctement prédites grâce à la génomique, cette technique permet d'accélérer l'évaluation des candidats à la sélection. Pour rappel, le gain génétique annuel (ΔG) se calcule avec l'équation du sélectionneur (**Equation 2, Chapitre I.1.2.2.2**) et dépend notamment de l'intervalle de génération L (nombre d'années entre les croisements et la production de nouvelles lignées parentales). Heffner et al. (2010) ont montré par simulation que le gain de temps permis par la prédiction génomique, en comparaison de l'évaluation phénotypique classique, permet de doubler le gain génétique annuel dans un programme de sélection blé, en accélérant les dernières années d'évaluation et le renouvellement de la population parentale. La prédiction génomique permet aussi augmenter l'intensité de sélection à coût équivalent, c'est-à-dire tester un plus grand nombre de lignées, et appliquer une plus forte pression de sélection (Crossa et al. 2017), ce qui augmente le gain génétique (**Equation 2**).

Ainsi, de façon alternative à l'évaluation par phénotypage, la prédiction génomique permet en théorie d'économiser des ressources, ou de les redéployer pour augmenter le gain génétique. Cependant, l'implémentation pratique de la prédiction génomique dans les schémas de sélection blé tendre reste débattue (Bassi et al. 2016; Juliana et al. 2018) car l'avantage économique est fortement déterminé par l'organisation du schéma de sélection (R2D2 Consortium 2021). Par ailleurs, depuis très récemment, la prédiction des valeurs génétiques peut désormais être réalisée par prédiction phénomique à très faible coût (de l'ordre de l'euro). La prédiction phénomique utilise les spectres proche infrarouge (NIRS) mesurés sur des organes de la lignée candidate (grain, feuille) pour établir des matrices d'apparentement et prédire les valeurs génétiques de lignées candidates (Rincent et al. 2018). Le développement de la prédiction phénomique pourrait remettre

en cause le développement de la prédiction génomique dans les programmes de sélection dans la mesure où cette technique est moins chère et présente une predictive ability du même ordre.

1.2.1.4 Limites de la prédiction génomique

Tout comme les estimations phénotypiques classiques, la précision des GEBV dépend de l'héritabilité, du protocole expérimental (nombre d'environnements, nombre de répétitions), de la qualité du phénotypage et du modèle statistique utilisé pour dissocier les effets environnementaux des effets génétiques (Daetwyler et al. 2010; Cullis et al. 2020). A ceci s'ajoutent des facteurs spécifiques à l'approche génomique qui peuvent impacter son efficacité (Desta et Ortiz 2014).

La taille de la population d'entrainement et sa proximité génétique avec la population candidate sont deux facteurs primordiaux pour la précision des GEBV (Habier et al. 2007; Bassi et al. 2016). La capacité prédictive du modèle diminue quand la divergence génétique augmente entre la population d'entrainement et la population candidate. Pour compenser cette diminution, les effets des marqueurs doivent être réestimés à chaque cycle de sélection (Heffner et al. 2010). Plusieurs hypothèses ont été avancées pour expliquer la dégradation de la capacité prédictive avec la différenciation des populations. La première est que le DL entre marqueurs et QTLs est spécifique de chaque population, puisque résultant de recombinaisons et des fréquences alléliques spécifiques de chaque population. Le DL entre marqueurs et QTLs diminue notamment à cause de la recombinaison au fur et à mesure des cycles de sélection. Une autre hypothèse est que les effets additifs estimés intègrent des effets épistatiques, et que la différenciation s'accompagne d'une évolution des interactions épistatiques (Mackay 2014). Il existe des outils qui permettent d'optimiser la composition génétique de la population d'entrainement, prenant par exemple en compte la proximité génétique entre population d'entrainement et population de validation (Rincent et al. 2012) ou l'architecture génétique du caractère (Mangin et al. 2019) ou encore la diversité des interactions entre valeurs génétiques et environnements (Bustos-Korts et al. 2016).

La précision des prédictions augmente avec la densité de marqueurs moléculaires couvrant le génome avant d'atteindre un plateau. On suppose que cela est dû à une liaison plus forte entre les marqueurs et les QTLs. Le nombre de marqueurs optimum dépend de l'étendue du DL dans le génome et de la taille de la population d'entrainement (de los Campos et al. 2013). Cependant, Speed et al. (2012) ont montré que l'hétérogénéité du DL entre marqueurs et QTLs le long du génome est préjudiciable à l'estimation de l'héritabilité, ce qui peut vraisemblablement dégrader les estimateurs. Afin d'améliorer l'estimation de la variance génétique, Speed et al. (2012) proposent de pondérer les loci selon l'intensité locale du DL lors de la construction de la matrice d'apparentement génomique G.

Enfin, la précision des estimations dépend de la cohérence entre l'architecture génétique supposée par le modèle et l'architecture génétique réelle (Daetwyler et al. 2010). Les modèles les plus

classiques GBLUP et RR-BLUP supposent que le déterminisme génétique est composé de nombreux QTLs à petits effets additifs, et que ces effets se distribuent selon une même loi Normale. Plusieurs modèles ont été proposés prenant en compte un mélange de distributions pour modéliser les effets, autorisant notamment une partie des marqueurs à avoir des effets nuls. De nombreux auteurs (dont Heslot et al. 2012; Hofheinz et Frisch 2014) montrent que le choix du modèle de prédiction génomique n'impacte pas beaucoup la précision des GEBV lorsque le caractère est polygénique. Par contre, le choix du modèle de prédiction génomique est déterminant si l'on s'intéresse aux estimations des effets des marqueurs plutôt qu'aux GEBV, par exemple pour étudier le déterminisme génétique ou calculer la variance d'un croisement. Hofheinz et Frisch (2014) montrent par simulation que les modèles de prédiction génomique basés sur une distribution normale des effets aux marqueurs produisent de bonnes estimations des GEBV mais sous-estiment fortement les effets aux QTLs. Ils en concluent que la capacité prédictive de ces modèles ne s'appuie pas tellement sur les estimations des effets des marqueurs en soi, mais plutôt sur l'estimation des combinaisons de marqueurs. A l'inverse, les modèles qui supposent que les effets aux marqueurs forment un mélange de distributions montrent une précision supérieure pour l'estimation des effets aux QTLs, mais sont plus laborieux à ajuster car ils nécessitent de choisir une distribution *a priori* des effets.

I.2.2 Choix des croisements avec la prédiction génomique

Pour rappel, les valeurs génétiques des descendants d'un croisement se distribuent selon $N(PM, \sigma^2)$, où PM est l'espérance de la descendance et σ^2 est la variance de la descendance.

L'espérance PM se calcule comme la moyenne des valeurs génétiques additives des parents, et peut être estimée par phénotypage ou par prédiction génomique. Le véritable apport de la prédiction génomique pour le choix des croisements repose sur la prédiction de la variance de la descendance σ^2.

Nous avons vu au **Chapitre I.1.2.3.1** que les prédictions de la variance de la descendance d'un croisement à l'aide de la distance phénotypique ou génotypique entre parents sont de mauvaise qualité, ce qui suggérait que le modèle était incomplet.

En effet, la distribution des valeurs génétiques des descendants dépend de la distribution des allèles favorables et défavorables chez les parents, de leurs effets et de la fréquence des recombinaisons (r) entre loci lors de la formation des gamètes (**Figure 3**).

Pour rappel, la formation des gamètes implique deux types de recombinaison des génomes parentaux : l'échange réciproque de larges portions d'ADN entre chromosomes homologues (crossing-over) et la ségrégation aléatoire des chromosomes homologues dans les gamètes. Pour une paire de loci, on définit r la fréquence de génotypes recombinés dans la population de gamètes.

Cette fréquence r est en théorie comprise entre 0 (aucun génotype recombiné) et 0.5 (il y autant de génotypes recombinés que de génotypes non recombinés). Les loci positionnés sur deux chromosomes homologues différents (par exemple, les chromosomes 1A et 1B, ou 1A et 2A) ont nécessairement un r de 0.5 car la répartition des chromosomes est aléatoire dans les gamètes. Par contre, les loci positionnés sur un même chromosome ont un r éventuellement inférieur à 0.5, qui augmente avec le nombre de crossing-over entre ces deux loci. Le **Tableau 1** présente la diversité des génotypes des gamètes pour deux loci a et b de la **Figure 3**.

Tableau 1 : Diversité génotypique à deux loci en fonction de la fréquence de recombinants

Génotype gamète	Valeur génétique	Fréquence observée	Modèle de la fréquence
a1b1	0	n_{a1b1}	(1-r_{ab})/2
a2b2	$\beta_a + \beta_b$	n_{a2b2}	(1-r_{ab})/2
a1b2	β_b	n_{a1b2}	r_{ab}/2
a2b1	β_a	n_{a2b1}	r_{ab}/2

Les loci sont nommés a et b, et montrent chacun deux allèles : a1 a2 et b1 b2. Les allèles 1 sont portés par les chromosomes grand-maternels et les allèles 2 par les chromosomes grand-paternels. Dans cet exemple, les loci a et b peuvent être positionnés sur un même groupe homologue, ou sur des chromosomes appartenant à deux groupes homologues différents. Les termes $\beta_a$et β_b désignent les effets additifs des allèles a2 et b2 (on suppose que ces effets valent 0 pour a1 et b1).

Pour une paire de loci, la distribution des valeurs génétiques des gamètes issus d'un individu i peut donc se calculer à partir des effets des allèles et de la fréquence des génotypes recombinés. En généralisant à un nombre quelconque de loci (Lehermeier et al. 2017; Santos et al. 2019) :

$$\sigma_i^{gamete} = \sum_{l=1}^{L} \beta_l^2 \, p_{i\,l}(1 - p_{i\,l}) + 2 \sum_{l<m} \beta_l \beta_m D_{i\,lm}(1-2r_{lm}) \qquad \textbf{[Equation 3]}$$

Avec L le nombre total de loci impliqués dans le déterminisme génétique et $p_{i\,l}$ la fréquence allélique au locus l (0 si le parent est homozygote, 0.5 sinon). Le terme r_{lm} est la fréquence des génotypes recombinés entre les loci i et j. Le terme $D_{i\,lm}$ désigne la covariance de chaque paire

d'allèles. Dans le cas où deux loci A et B ont chacun deux allèles (A+/A- et B+/B-), l'un favorable (+) et l'autre défavorable (-), trois cas de figures sont possibles :

(1) si les deux allèles favorables sont portés par un des parents (A+/B+) et les deux allèles défavorables sont portés par l'autre parent (A-/B-), alors les allèles sont dits en couplage. Dans ce cas, la covariance entre les allèles est positive et D = 0.25

(2) si chaque parent porte un allèle favorable et un allèle délétère (A+/B- et A-/B+), alors les allèles sont dits en répulsion. Dans ce cas, la covariance entre allèles est négative et D = -0.25

(3) si l'un des deux loci A ou B est homozygote chez au moins l'un des deux parents, D vaut 0.

Parmi les différences notables entre les deux méthodes pour obtenir des nouvelles lignées (RILs ou HDs, **Figure 4**), le nombre de recombinaisons dans une population RILs est deux fois plus élevé que dans une population HDs (Haldane et Waddington 1931). Le génome des RILs a connu plusieurs méioses depuis le croisement entre lignées parentales, alors que le génome des HDs une seule. Par contre, chez les RILs, les méioses génèrent de moins en moins de recombinaisons visibles (dites efficaces) au fur et à mesure que l'hétérozygotie diminue.

Ainsi, la variance d'une descendance RILs et d'une descendance HDs se calcule avec deux formules différentes (Lehermeier et al. 2017) :

$$\boldsymbol{\sigma_{ij}^{RILs\ Fk}} = \mathbf{4} * \left(\sum_{l=1}^{L} \boldsymbol{\beta_l^2}\, \mathbf{p}_{ij\,l}(\mathbf{1} - \mathbf{p}_{ij\,l}) + \mathbf{2}\sum_{l<m} \boldsymbol{\beta_l \beta_m} \mathbf{4D}_{ij\,lm}\,(\mathbf{1} - \mathbf{2}\boldsymbol{r_{lm}}^{(k)} - (\mathbf{0.5(1\text{-}2}\boldsymbol{r_{lm}}))^k\right)$$

Avec ${r_{lm}}^{(k)} = \frac{2r_{lm}}{1+2r_{lm}}$(1-0.5$^k(1-2r_{lm})$ l'espérance de la fréquence de génotypes recombinés après k générations d'autofécondation.

Pour la variance d'une descendance HDs :

$$\boldsymbol{\sigma_{ij}^{HDs}} = \mathbf{4} * \left(\sum_{l=1}^{L} \boldsymbol{\beta_l^2}\, \mathbf{p}_{ij\,l}(\mathbf{1} - \mathbf{p}_{ij\,l}) + \mathbf{2}\sum_{l<m} \boldsymbol{\beta_l\, \beta_m} \mathbf{4D}_{ij\,lm}\,(\mathbf{1\text{-}2}\boldsymbol{r_{lm}})\right) \qquad \textbf{[Equation 4]}$$

La variance des HDs donnée ici vaut 4 fois la variance gamétique (**équation 3**) lorsque le HD est dérivé d'un gamète de F1, mais il est possible de prendre en compte que le gamète soit dérivé d'un stade plus tardif (F2…).

Une alternative aux formules analytiques est la simulation *in silico* de descendants recombinés (Mohammadi et al. 2015). L'avantage de la simulation *in silico* est que l'on peut tenir compte d'un déterminisme génétique plus complexe qu'un déterminisme génétique purement additif, de la covariance entre caractères et de la taille finie de l'effectif de la descendance sans modélisation mathématique. Par contre, la taille limitée de la descendance simulée n'offre pas un estimateur précis de la variance de la descendance (Lehermeier et al. 2017), à moins de répéter chaque simulation un grand nombre de fois. De plus, estimer la variance des descendants pour un grand nombre de couples peut se révéler très lourd d'un point de vue calculatoire.

I.2.3 Précision de la prédiction de la variance des descendants

La prédiction analytique de la distribution des descendants d'un croisement telle que présentée précédemment requiert 1) le génotype des parents candidats 2) les effets des QTLs ou les effets estimés des marqueurs 3) une carte génétique pour l'ensemble des QTLs ou marqueurs. Or avoir une connaissance, même imprécise, de l'ensemble de ces éléments nécessite des technologies assez récentes. Par exemple, la prédiction génomique a été proposée en 2001 (Meuwissen et al. 2001), et les puces de génotypage ne sont disponibles que depuis quelques années.

Nous avons vu précédemment que la prédiction de la variance intra-famille à partir des distances génotypiques des parents ne fonctionnait pas très bien. Par contre, les estimations des effets des allèles et leur co-ségrégation apportent un gain de précision non négligeable sur la variance de la descendance. Globalement, la comparaison entre prédictions et observations dans la réalité révèle que la précision sur l'espérance des descendants PM est correcte, mais la précision sur la variance de la descendance σ^2 est irrégulière. Tiede et al. (2015) trouvent une corrélation de 0.61 entre la variance prédite et la variance observée de 40 croisements comprenant environ 122 descendants phénotypés pour la tolérance à la fusarium de l'épi (Fusarium Head Blight ; FHB) chez l'orge (h^2=0.4) (**Figure 10**).

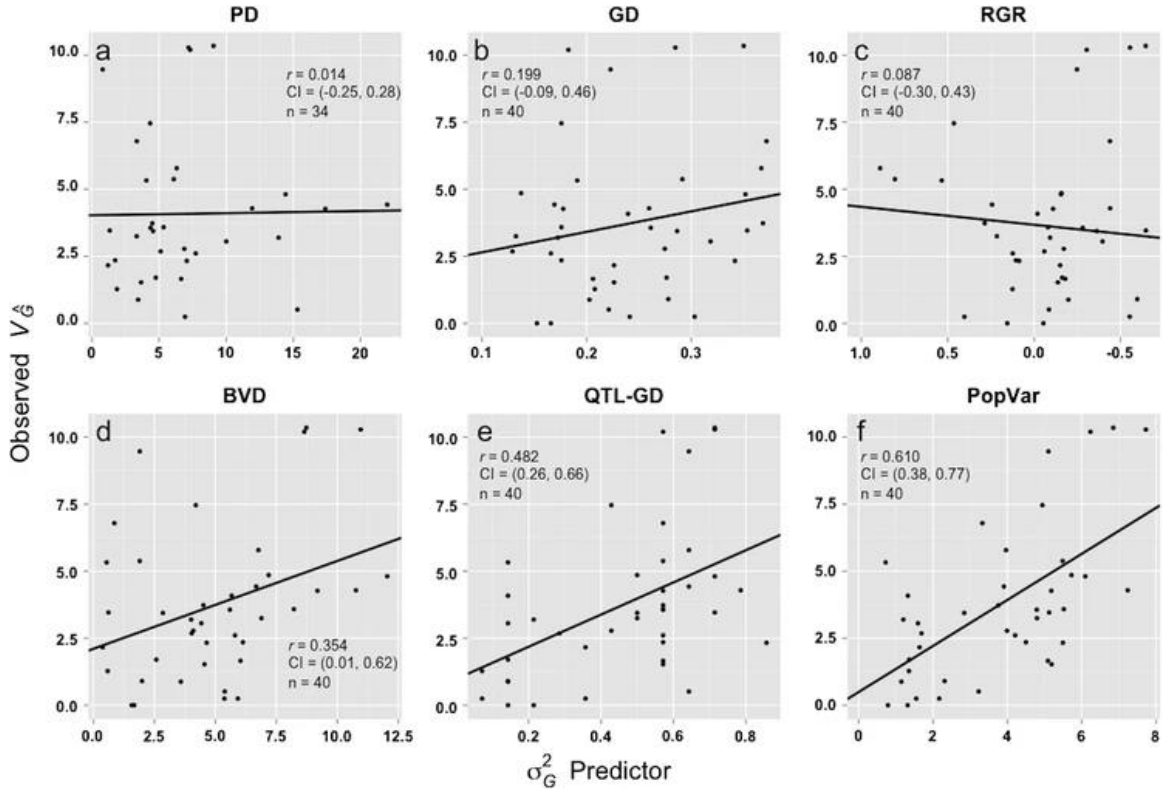


Figure 10 : Précision de différents estimateurs de la variance de la descendance

Sur l'axe des X : variance de la descendance RILS F5 obtenue sur de véritables descendances de croisements biparentaux d'orge, phénotypées pour la fusariose de l'épi. Sur l'axe des Y, différents estimateurs de la variance de la descendance. PD = valeur absolue de la différence phénotypique entre les deux lignées parentales ; GD = Proportion de marqueurs différents ; RGR = coefficients de la matrice d'apparentement ; BVD = valeur absolue de la différence des GEBV ; QTL-GD = proportion de marqueurs différents dont la valeur absolue des effets estimés dépassent un seuil minimum ; PopVar = variance de la descendance obtenue en simulant des descendants in silico et en calculant leur GEBV à partir de leur génotype. Les indications dans la partie supérieure des cadres donnent la corrélation entre la variance observée de la descendance et la mesure en x, ainsi que les intervalles de confiance sur cette corrélation et le nombre de points (nombre de croisements biparentaux).

Tiede et al. (2015)

Toujours chez l'orge, Neyhart et Smith (2019) trouvent une corrélation moyenne de 0.54 pour PM, mais 0.29 pour σ^2 pour 27 croisements comprenant en médiane 90 descendants phénotypés sur la tolérance à la FHB, la date d'épiaison et la hauteur. Dans leurs jeux de données, la qualité de prédiction de σ^2 augmente avec l'héritabilité : 0.01 pour le caractère le moins héritable (tolérance FHB), et jusqu'à 0.48 celui le plus héritable (date d'épiaison). Chez le maïs, Adeyemo et Bernardo (2019) ont étudié 8 croisements, comprenant au moins 120 descendants par croisements, phénotypés pour trois caractères (hauteur, date d'épiaison et taille de l'épi). Les résultats montrent une bonne précision sur la prédiction de PM (corrélation moyenne ≥ 0.8) mais la corrélation entre variance prédite et variance observée varie entre -0.24 et 0.14, sachant que l'héritabilité varie, elle, entre 0.4 et 0.8. Chez le manioc, Wolfe et al. (2021) étudient quatre caractères, et observent une corrélation entre 0.1 et 0.8 entre la moyenne et l'espérance des descendants, et une corrélation entre 0 et 0.4 pour la variance de la descendance. Chez les bovins laitiers, Santos et al. (2019) observent une corrélation qui croît avec le nombre de descendants phénotypés et qui dépend du caractère (production laitière, teneur en protéines, teneur en acides gras). Pour une centaine de taureaux avec plus de 600 descendants chacun, la corrélation entre variance gamétique prédite et observée varie entre 0.30 et 0.97.

La prédiction de la variance de la descendance σ^2 avec les modèles de prédiction génomique repose sur de nombreuses hypothèses simplificatrices : effets purement additifs estimés par prédiction génomique, estimateur des fréquences de recombinaison à valeur prédictive pour la population de sélection, modélisation simple du processus méiotique, pas de biais sélectif. Les simulations ont permis de comprendre les facteurs qui peuvent améliorer l'estimation de σ^2 et cela passe notamment par l'optimisation des modèles de prédiction génomique. Par simulation, Lehermeier et al. (2017) et Santos et al. (2019) ont montré que la taille de la population d'entraînement et l'héritabilité améliorent la prédiction de la variance de la descendance. Ces deux facteurs sont déjà déterminants dans la précision des modèles de prédiction génomique. Cependant, les estimateurs de σ^2 sont sensibles à des facteurs supplémentaires par rapport aux estimateurs des valeurs génétiques (et donc de PM). Plusieurs auteurs notent que la variance de la descendance prédite sous-estime grandement la variance réelle (Santos et al. 2019; Neyhart et Smith 2019; Adeyemo et Bernardo 2019; Tiede et al. 2015). Il existe deux causes probables à cette sous-estimation : la distribution supposée des effets marqueurs lors de l'ajustement du modèle de prédiction génomique et le déséquilibre de liaison entre QTLs.

L'utilisation de différents modèles de prédiction génomique caractérisés par différentes distributions pour les effets des marqueurs donnent un niveau de précision similaire sur l'estimation des valeurs génétiques, à l'exception des phénotypes gouvernés par un petit nombre de variants causaux (Heslot et al. 2012; Daetwyler et al. 2010). Par contre, la précision de l'estimateur de σ^2 est impactée par le modèle, et plus précisément par la distribution des effets marqueurs. Santos et al. (2019) ont estimé la variance gamétique d'un phénotype simulé (20 ou 200 QTLs, h^2=0.1, 0.3 ou 0.5) avec un modèle GBLUP classique (où la distribution des effets marqueurs suit une même

loi Normale) et un modèle Bayesian Lasso (où la distribution des effets est une loi double exponentielle avec une masse à zéro). La corrélation entre variance gamétique observée (simulée) et prédite avec le modèle Lasso est supérieure d'environ 0.1 point à la corrélation avec les variances prédites par le modèle GBLUP (Table 2 de Santos et al. 2019). Santos et al. 2019 estiment que le modèle Lasso permet une meilleure précision sur les marqueurs, ce qui est d'autant plus important que les effets sont élevés au carré dans la prédiction de la variance gamétique et qu'ils sont attachés à des segments chromosomiques précis qui ségrégent dans la descendance. Dans la même optique, Adeyemo et Bernardo (2019) suggèrent d'utiliser les effets des marqueurs non shrinkés ou dérégressés pour prédire la variance de la descendance. Cependant, en blé, (Yao et al. 2018) n'ont pas vu une amélioration de l'estimation de σ en utilisant différents modèles (dont les modèles RRBLUP et Bayesian Lasso). Lehermier et al. (2017) proposent de prendre en compte l'incertitude sur les effets estimés des marqueurs en ajustant un modèle de prédiction génomique Bayésien nommé PMV (Posterior Mean Variance model). Le modèle PMV consiste à estimer la distribution à posteriori des effets des marqueurs avec une Markov Chain Monte Carlo, puis à calculer l'estimateur de la variance comme la moyenne de la distribution à posteriori de la variance de la descendance :

$$\sigma^2{}_{ij} = \frac{1}{S}\sum_{s} \widehat{\beta_s}\ V(X)_{ij}\widehat{\beta_s}'$$

Avec $\sigma^2{}_{ij}$ la variance de la descendance du croisement entre $P_i\ et\ P_j$, S le nombre d'échantillons de la distribution postérieure des effets aux marqueurs β et $V(X)_{ij}$ la matrice de variance-covariance des génotypes des descendants du croisements. Par simulation, le modèle PMV montre un biais proche de 0 et une corrélation proche de 1 avec la vraie valeur de la variance de la descendance (calculée à partir des QTLs) lorsque la taille de la population d'entrainement dépasse la centaine d'individus. Ces modèles Bayésien Lasso ou PMV sont donc les plus adéquats pour la prédiction de la variance sur des données réelles.

L'une des questions que l'on peut se poser est la pertinence de modéliser la variabilité des taux de recombinaison (profil de recombinaison) le long du génome lors de la prédiction de la variance. En effet, la prise en compte de la recombinaison complexifie beaucoup le calcul de la variance. Dans la formule décrite par Lehermier et al. (2017) (**Equation 4**, **Chapitre I.2.2**), pour calculer la variance de la descendance, la prise en compte de la recombinaison exige de calculer M*(M-1)/2 termes où M est le nombre de marqueurs. Une manière de tester la pertinence de la prise en compte de la variabilité de la recombinaison est de comparer les estimateurs « complets » de variance obtenus en tenant compte de la recombinaison (**Equation 4**), avec les estimateurs « simplifiés » obtenus en remplaçant toutes les fréquences de recombinants par 0.5 dans l'**Equation 4**, ce qui revient à rendre nuls les M*(M - 1)/2 termes mentionnés précédemment et accélère grandement le calcul. Cependant, plusieurs auteurs (Lado et al. 2017; Santos et al. 2019; Tiede et al. 2015) rapportent que les estimateurs complets sont préférables aux estimateurs simplifiés de la variance de la

descendance. Par exemple, Tiede et al. (2015) montrent que la corrélation entre variance prédite simplifiée et variance observée dans une descendance réelle est de 0.48 et passe à 0.61 avec les estimateurs complets (**Figure 9**). Par simulation, Santos et al. (2019) et Lado et al. (2017) montrent que les estimateurs simplifiés ont une corrélation intermédiaire avec les estimateurs complets, qui sont eux de très bon proxy de la vraie variance de la descendance (**Figure 11**).

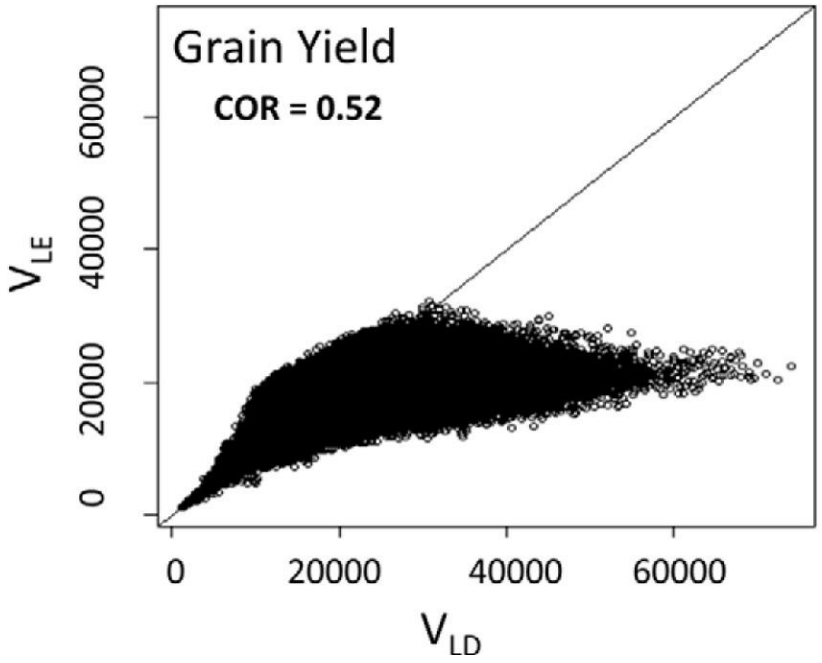


Figure 11 : Variance de la descendance calculée avec ou sans profil de recombinaison dans un programme de blé tendre

Un point = la variance de la descendance d'un croisement. VLE = Variance de la descendance calculée en simplifiant le profil de recombinaison (r = 0.5 pour toutes les paires de loci).
VLD = Variance de la descendance calculée avec le profil de recombinaison

Lado et al. (2017)

I.2.4 Sélection des croisements tenant compte de la variance génétique des descendants

Pour rappel, le plan de croisement est un compromis entre plusieurs objectifs : maximiser la valeur génétique des nouvelles variétés mais aussi améliorer la population parentale en limitant la perte de diversité génétique.

I.2.4.1 Les Critères de Sélection de Croisement (CSC)

La prédiction de la distribution de la descendance permet de trier les croisements les plus prometteurs et donc de déterminer la liste des F1 à produire. On définit l'utilité d'un croisement comme son classement relatif par rapport à d'autres croisements sur un critère précis. Il existe plusieurs critères d'utilité de croisements (ou CSC), qui exploitent différemment la distribution prédite de la descendance.

Les CSC basés sur la variance de la descendance exploitent les propriétés des distributions prédites des descendants d'un croisement selon N(PM, σ^2), sachant que σ^2 est désormais accessible grâce aux modèles de prédiction génomique.

Le critère UC (Usefulness Criterion) proposé par Schnell et Utz (1975) donne l'espérance des q (exprimé en %) meilleurs descendants (RILs ou HDs) d'un croisement et se calcule comme

$$UC = PM + i^q \, h \, \sigma$$

avec l'intensité de sélection i^q correspondant à un taux de sélection de q. Cette quantité i^q se calcule comme l'espérance du quantile supérieur q d'une loi Normale N(0,1). Le terme h fait référence à la racine carrée de l'héritabilité. Dans la pratique, il est systématiquement remplacé par 1 dans la formule (Zhong et Jannink, 2007; Lehermeier et al. 2017; Yao et al. 2018; Bijma et al. 2020). Pour tenir compte des effectifs limités dans la descendance, Müller, Schopp, et Melchinger (2018) ont proposé le critère EMBV (Expected Maximum Breeding Value) qui prédit l'espérance du meilleur descendant parmi une descendance d'une taille précise D. Ce critère EMBV peut se calculer comme

$$EMBV = PM + Int^{1/D} \ \sigma$$

où $Int^{1/D}$ est l'espérance de la plus grande statistique d'ordre dans un échantillon de D variables aléatoires tirées dans une loi Normale N(0,1). A l'origine, l'EMBV est obtenu en simulant plusieurs fois la descendance, puis en calculant l'espérance des meilleurs statistiques sur chaque simulation. Cependant, il existe des approximations de l'espérance de la plus grande statistique d'ordre $Int^{1/D}$ (Burrows 1972) qui permettent de calculer analytiquement l'EMBV.

Si les critères UC et EMBV donnent des estimateurs des valeurs génétiques des meilleurs descendants, ils n'informent pas directement sur la capacité d'un couple à produire des descendants transgressifs. Dans cette perspective, Bijma et al. (2020) et Wellmann (2019) ont suggéré de quantifier l'utilité d'un croisement comme la proportion de descendants supérieurs à un seuil. Cette proportion se calcule à partir de la fonction de répartition de la loi Normale $F_{PM, \sigma}$ (λ). Toute la difficulté réside dans le choix du seuil et sa pertinence pour assurer un gain génétique intéressant à la prochaine génération, tout en ne compromettant pas le gain à long terme (**Chapitre I.1.2.1.3.2.2**). Un seuil relativement simple à définir serait la valeur génétique de la meilleure variété du programme de sélection, afin d'assurer un progrès génétique minimal. Wellmann et al. (2019)

suggèrent de définir le seuil en fonction du gain génétique moyen par génération, extrapolé à partir de données historiques. Bijma et al. (2020) proposent de définir un seuil de troncature à partir de l'ensemble des distributions de descendants d'une population de sélection, afin d'évaluer pour chaque reproducteur la probabilité de produire des descendants supérieurs à ce seuil.

Le critère OHV (Optimal Haploid Value) proposé par (Daetwyler et al. 2015) évalue l'utilité d'un couple comme la valeur génétique du meilleur descendant qu'il peut produire, peu importe le nombre de méioses que cela nécessiterait. Le génome est découpé en blocs haplotypiques, supposés transmis intacts à la descendance (**Figure 12**). Chaque bloc chevauche un certain nombre de marqueurs moléculaires. La combinaison d'allèles à l'intérieur du bloc forme un haplotype. L'effet d'un haplotype est calculé comme la somme des effets de chaque allèle qu'il contient.

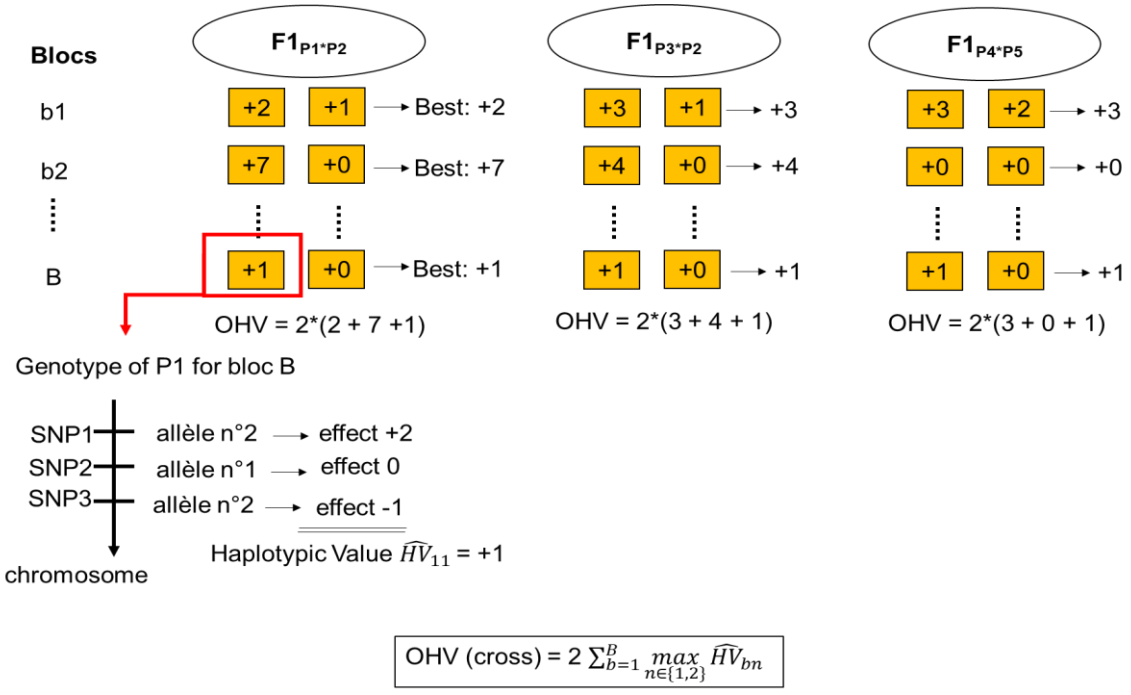


Figure 12 : Exemples de calcul de l'Optimal Haploid Value

Le nombre et la délimitation des blocs sont très importants dans le calcul du critère. Les blocs doivent être judicieusement choisis afin de refléter au mieux la transmission du patrimoine génétique aux descendants. Cela signifie que la délimitation des blocs doit tenir compte du profil de recombinaison. Par exemple, les centromères forment des blocs haplotypiques dans la mesure où ce sont de grandes régions froides de recombinaison. Les blocs haplotypiques sont systématiquement transmis à l'identique aux descendants. Ils peuvent donc être définis en fonction de la position des points froids où points chauds de recombinaison, qui seront définis plus tard dans ce manuscrit comme des segments du génome associés à un faible ou à un fort taux de recombinaison. Bien que l'OHV soit défini comme la valeur espérée du meilleur descendant de première génération uniquement, il s'avère (d'après les simulations de Daetwyler et al. (2015) que le nombre de blocs détermine la pertinence du critère pour un gain génétique long terme ou court terme. En effet, un grand nombre de blocs nécessite un grand nombre de méioses successives pour être cumulés dans la descendance. Un petit nombre de blocs est donc plus pertinent dans le cadre d'une stratégie privilégiant le gain génétique court terme. Dans la mesure où il est très peu probable d'obtenir l'OHV, c'est-à-dire le meilleur descendant potentiel, ce CSC est malgré tout corrélé avec la valeur génétique du meilleur descendant observé, et cette corrélation devrait augmenter en fonction de la complémentarité haplotypique des parents.

La **Figure 13** résume en quoi les CSC exploitent différemment la prédiction de la distribution de la descendance. Ces différents CSC peuvent être vus comme différents objectifs de sélection, ou comme des CSC plus ou moins robustes à l'estimation des effets des marqueurs, ou plus ou moins axés sur la complémentarité allélique des parents pour privilégier le gain long terme au gain court terme.

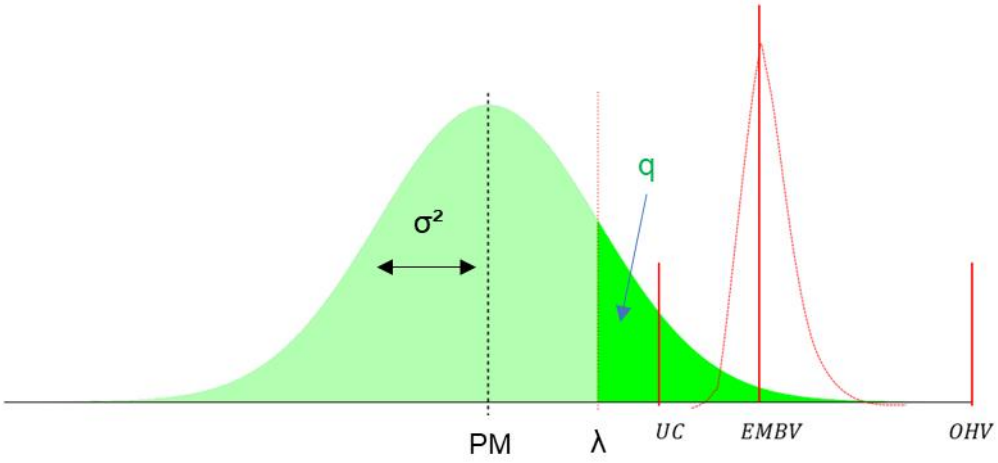


Figure 13 : Résumé des différents critères de sélection de croisements (CSC)

La distribution normale représente la distribution des valeurs génétiques des descendants pour un couple, N(PM, σ²). On peut dénombrer 5 CSC différents. PM = espérance des valeurs génétiques des descendants ; UC = espérance des q meilleurs descendants, avec q qui doit être défini par l'utilisateur ; λ = seuil de troncature permettant de calculer la proportion de descendants supérieurs à ce seuil ; EMBV = espérance du meilleur descendant lorsqu'on attribue D descendants au couple ; OHV = estimateur de la valeur génétique de la meilleure lignée que pourrait produire le couple.

I.2.4.2 Efficacité des CSC sur le gain génétique

Le gain génétique permis par ces différents CSC (UC, EMBV, OHV) a été comparé par plusieurs auteurs sur la base de simulations (Lehermeier et al. 2017; Mohammadi et al. 2015; Yao et al. 2018; Neyhart et Smith 2019; Müller et al. 2018), qui ont l'avantage d'être peu coûteuses en comparaison des tests au champs. La plupart des simulations utilisent des génotypes réels de lignées parentales et un caractère d'intérêt simulé à partir de QTLs échantillonnés au hasard le long du génome parmi les marqueurs disponibles. La plupart des études se placent dans un cadre pseudo-réaliste, où les effets et positions des QTLs sont considérés comme inconnus et doivent être estimés par un modèle de prédiction génomique. Ce modèle est généralement ajusté sur les phénotypes de la population de lignées parentales. Le phénotype d'une lignée parentale est calculé comme la somme de sa TBV à laquelle est rajouté un bruit, de manière à atteindre l'héritabilité désirée. L'efficacité des CSC est testée en comparant la distribution des TBV des descendants simulés *in silico*.

Suivant cette procédure, Mohammadi et al. 2015 ont montré que chez le maïs l'espérance des 10% meilleurs descendants d'un couple était fortement corrélée à la moyenne des valeurs génétiques des parents (r^2 = 0.82 pour le rendement,), mais que la prise en compte de la variance de la descendance permettait d'améliorer la prédiction (r^2 = 0.995). Toujours chez le maïs, Lehermeier et al. (2017) ont comparé le gain génétique permis par différents CSC (**Figure 14**). Le déterminisme du caractère d'intérêt a été simulé en attribuant un effet additif à 300 marqueurs échantillonnés au hasard. Les effets des marqueurs étaient supposés inconnus et ont été estimés par un modèle de prédiction génomique à partir des phénotypes parentaux, avec une héritabilité de 0.2 ou 0.6. Les croisements ont été choisis soit sur la moyenne des valeurs génétiques des parents (PM), soit sur l'espérance d'une fraction supérieure de leur descendance (UC), soit sur l'OHV. Ils montrent que les croisements sélectionnés sur l'UC ou l'OHV apportent un gain génétique supérieur à ceux choisis sur la base de PM. Le gain augmente avec l'héritabilité du caractère d'intérêt car une plus grande héritabilité permet d'estimer plus précisément les effets des allèles (Wimmer et al. 2013) et donc la variance de la descendance.

Chez le blé tendre, la comparaison de CSC a été étudiée par (Yao et al. 2018; Lado et al. 2017). L'étude de Yao et al. 2018 porte sur une population de lignées comprenant des accessions appartenant aux deux principaux groupes génétiques du blé tendre, européen et asiatique. Les phénotypes sont simulés à partir de 30 ou 50 ou 100 QTLs et une héritabilité de 0.3 ou 0.5 ou 0.8. Les croisements sont choisis soit sur la moyenne des valeurs génétiques des parents (PM), soit sur l'espérance d'une fraction supérieure de leur descendance (**Figure 15**).

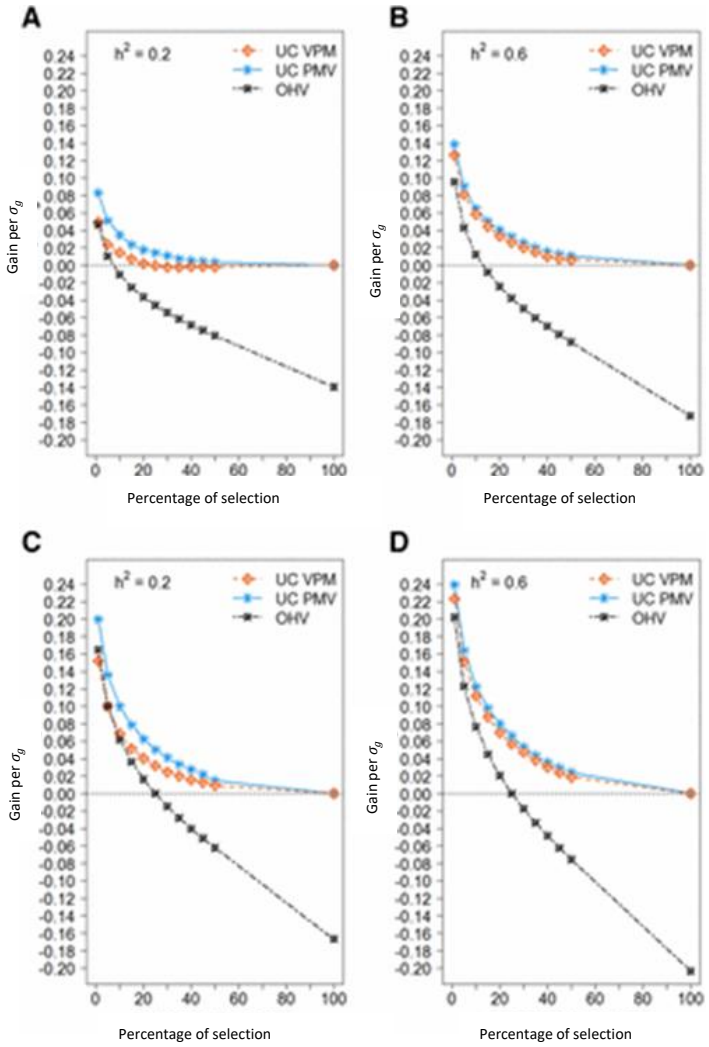


Figure 14 : Gain génétique supplémentaire permis par le choix des croisements sur l'UC ou l'OHV en comparaison de PM dans une programme de sélection simulée de maïs

Le gain génétique est calculé comme la différence entre les valeurs génotypiques moyennes des des descendants issus des croisements sélectionnées sur l'UC ou l'OHV avec les valeurs génotypiques moyenne des descendants issus d'une sélection sur PM. Le gain est standardisé par l'écart-type génétique σ_g de la population d'entraînement. Deux estimateurs de la variance de la descendance sont utilisés pour estimer l'UC (modèle Variance of Posterior Mean VPM et modèle Posterior Mean Variance PMV). Les résultats sont donnés pour des croisements pré-selectionnés sur une différence génétique minimum entre parents et une héritabilité de 0.2 (A) et 0.6 (B), ainsi que pour des croisements non pré-selectionnés, et une héritabilité de 0.2 (C) et 0.6 (D). Les lignées parentales sont issues d'une population NAM réelle.

Lehermeier et al. (2017)

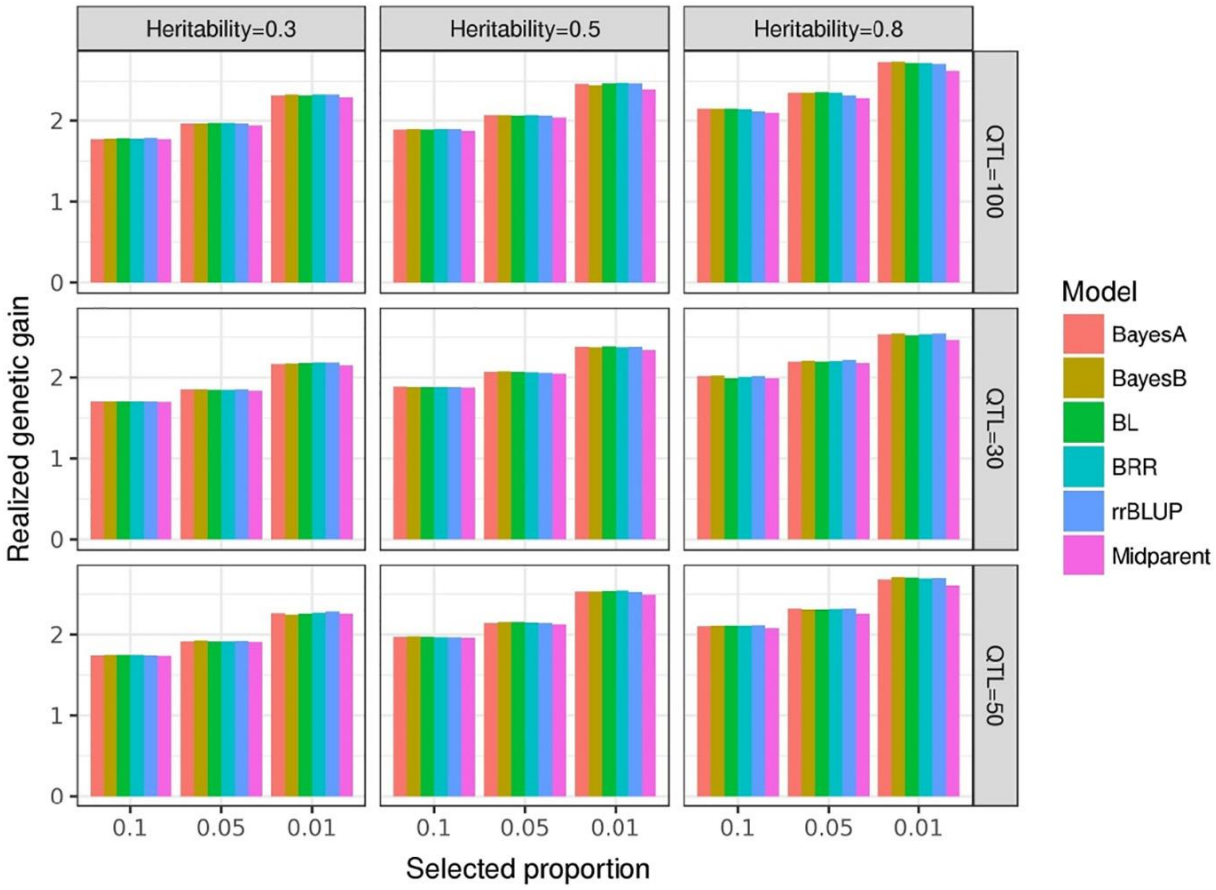


Figure 15 : Gain génétique permis par le choix des croisements sur l'UC ou sur la valeur génétique des parents dans une population simulée de blé tendre

Les colonnes rouges, jaunes, vertes, bleues donnent le gain pour l'UC, les colonnes roses donnent le gain pour PM (unité en écart-type génétique σ_g). Plusieurs modèles de prédiction génomique ont été testés pour estimer l'UC, notamment le Bayesian Lasso (BL) et Bayes Ridge Regression (BRR). Trois architectures génétiques simulées ont été testées (100, 50 ou 30 QTLs). Les lignées parentales sont composées de 57 lignées parentales réelles, issues de programmes de sélection chinois, américains et australiens.

Yao et al. (2018)

Le gain génétique permis par l'UC quel que soit le modèle, est faiblement supérieur au gain génétique permis par PM. Tout comme les résultats de Lehermeier et al. (2017), le gain génétique dépend de l'héritabilité du phénotype qui a permis d'ajuster le modèle de prédiction génomique. Les résultats de Lado et al. (2017) montrent aussi que la sélection des croisements sur PM et UC est très similaire dans une population de lignées de blé tendre du CIMMYT, ce qui amène les auteurs à conclure que le gain génétique supplémentaire permis par l'UC devait être est faible.

Chez les bovins laitiers, Bijma et al. (2020) ont montré que prendre en compte la variance gamétique lors du choix d'un taureau reproducteur permettrait permet d'augmenter la valeur

génétique des nouveaux taureaux de 3.6%, en comparaison du choix basé sur les GEBV des reproducteurs uniquement. On peut donc se demander quels sont les facteurs qui expliquent cette variation dans les bénéfices de l'UC.

1.2.4.3 Apport de la recombinaison dans la prédiction des meilleurs croisements

Les études précédentes montrent que l'intérêt de l'UC est variable selon les populations étudiées. Zhong et Jannink (2007) expliquent que le bénéfice des critères basés sur la variance de la descendance augmente quand le ratio entre la variance des écarts-types des descendances var(σ) et la variance des espérances var(PM) augmente. Ce ratio var(σ)/var(PM) est donc fondamental pour la supériorité des critères d'utilité qui exploitent la variance de la descendance (Zhong et Jannink 2007) Bijma et al. 2020, Utz et al. 2001, Lehermeier et al. 2017, Lado et al. 2017).

Les croisements peuvent être classés selon l'espérance de leur descendance (PM, aussi calculée comme la moyenne de valeurs génétiques des parents), ou selon une CSC basée sur la variance de la descendance, par exemple UC = PM + i* σ, ou encore EMBV = PM + INT* σ. Lorsque la variabilité des écart-types des descendants var(σ) est nulle ou faible devant la variabilité des espérances var(PM), cela signifie que la valeur génétiques des meilleurs descendants des différents couples est surtout expliquée par PM. Concrètement, cela signifie soit que les classements entre PM et les CSC basés sur la variance de la descendance sont très similaires (exemple dans la **Figure 16**), soit que les croisements choisis par PM ou par un CSC basé sur la variance auront des distributions chevauchantes. En conséquence, le gain génétique supplémentaire dû à la prédiction de la variance sera faible.

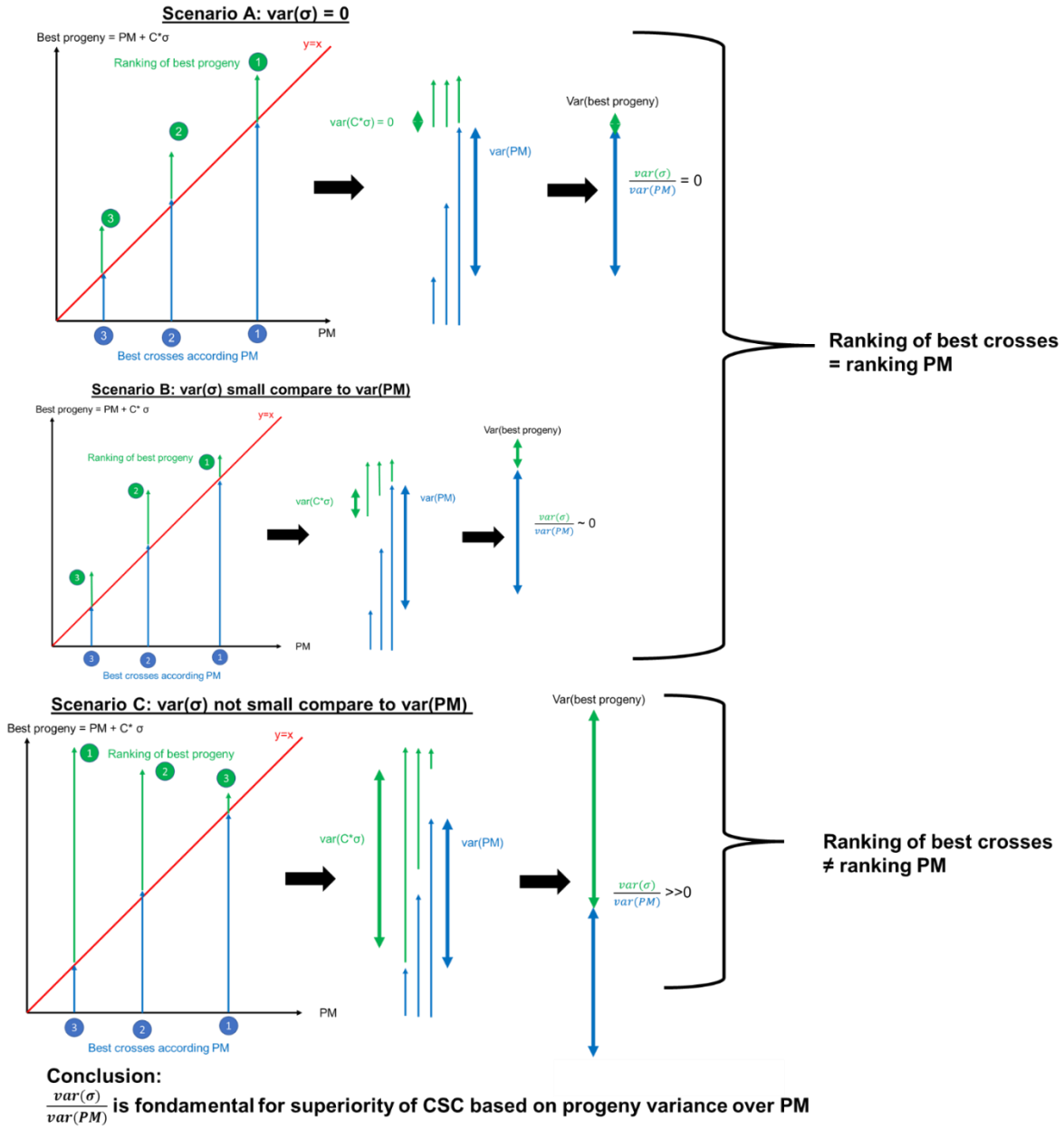


Figure 16 : Importance fondamentale du ratio var(σ)/var(PM) dans la supériorité des critères de choix de croisements basés sur la variance de la descendance.

Dans le scénario A, la variance des descendants n'est pas négligeable devant PM, mais la variance des descendants varie peu entre croisements. En conséquence, la prise en compte de la variance n'apporte aucune différence.

Dans le scénario B, la variance des variances n'est pas nulle, mais insuffisante pour inverser les classements des meilleurs croisements par rapport à PM

Dans le scénario C, la variance des variances n'est pas négligeable devant la variance de PM, ce qui s'accompagne d'une modification des classements des meilleurs croisements

Bijma et al. (2020) proposent même une table du gain génétique supplémentaire permis par le choix des taureaux reproducteurs sur un CSC basé sur la variance gamétique par rapport au choix des taureaux basés sur leur valeur génétique (**Tableau 2**). Cette table dépend du ratio entre l'écart-type des écart-types gamétiques SD($\sigma_{gametique}$) et la variance génétique des gamètes $0.5\sigma_A$. Elle dépend aussi de l'intensité de sélection.

Tableau 2 : Table du gain génétique supplémentaire permis par la sélection des taureaux reproducteurs sur leur variance gamétique

	CV of the SD of gametic GEBV $\left(SD\left(\sigma_{\hat{g}}\right)/0.5\sigma_{\hat{A}}\right)$			
p	**0.05**	**0.10**[a]	**0.15**	**0.20**
0.5	−0.2	−0.2	0.0	0.0
0.2	0.3	0.6	1.3	2.1
0.1	0.3	0.9	1.9	3.6
0.05	0.3	1.3	2.9	5.2
0.01	0.6	2.2	5.1	9.9
0.005	0.6	2.5	6.2	12.1
0.001	0.8	3.6	9.1	18.2

Augmentation du GEBV moyen des descendants sélectionnés (p) en fonction du ratio entre la variabilité des variances gamétiques et la variance génétique, lorsque les taureaux reproducteurs sont choisis sur leur GEBV et leur variance gamétique ou sur leur GEBV seul. Exemple pour des populations qui ont déjà été sélectionnées sur le caractère d'intérêt.

Bijma et al. 2020

Puisque le ratio var(σ)/var(PM) est fondamental lorsque l'on s'intéresse au bénéfice de modéliser la variance de la descendance, il est naturel de s'interroger sur les facteurs qui déterminent ce ratio. Comme souligné par Bijma et al. (2020), la composition génétique et l'histoire de la population parentale va déterminer la valeur de ce ratio.

La variance des espérance var(PM) diminue dans une population qui a déjà été sélectionnée sur le caractère d'intérêt. Cette population voit en effet sa variance génétique réduite par l'instauration de covariances négatives entre QTLs (Bulmer, 1971). Cependant, la variabilité des variances des descendants n'est pas censée être réduite, au moins dans un modèle infinitésimal, c'est-à-dire lorsque le nombre de QTLs et la taille de la population sont infinis. D'une manière générale, une réduction de la variance génétique (via une pré-sélection des croisements par exemple) augmentera le bénéfice des CSC basés ou basé sur la variance de la descendance, à condition que cela n'impacte pas aussi la variabilité des variances des descendants (contre-exemple ci-dessous, **Figure 17**). Une réduction de la variance génétique s'accompagne d'une réduction de la variance de l'espérance des descendants, et donc d'une augmentation du ratio var(σ)/var(PM).

La variabilité des variances des descendants augmente avec deux facteurs : la variabilité du polymorphisme des parents et la variabilité des covariances entre QTLs. Les populations structurées montrent une plus grande variabilité de polymorphisme entre les parents, donc une variabilité accrue des variances des descendants. Des populations structurées impliquant des lignées de bonnes valeurs (lignée H) et des lignées de moins bonnes valeurs (lignée L) bénéficient d'un double bonus pour la variabilité des variances de la descendance : une plus grande variabilité du polymorphisme entre les parents et une forte variabilité des covariances entre QTLs. Par exemple, les croisements entre lignées élites et ressources (H*L) montrent une forte variance de la descendance mais une espérance intermédiaire, alors que les croisements impliquant des lignées élites (H*H) ou ressources (L*L) montrent une faible variance de la descendance mais respectivement une forte et une faible espérance. Il en résulte notamment une covariance négative entre espérance et variance. Cette situation est fréquemment décrite chez les plantes : maïs (Bernardo 2014; Mohammadi, Tiede, et Smith 2015), blé tendre (Lado et al. 2017), et orge (Abed et Belzile 2019; Neyhart et Smith 2019) mais semble absente chez les bovins (Segelke et al. 2014).

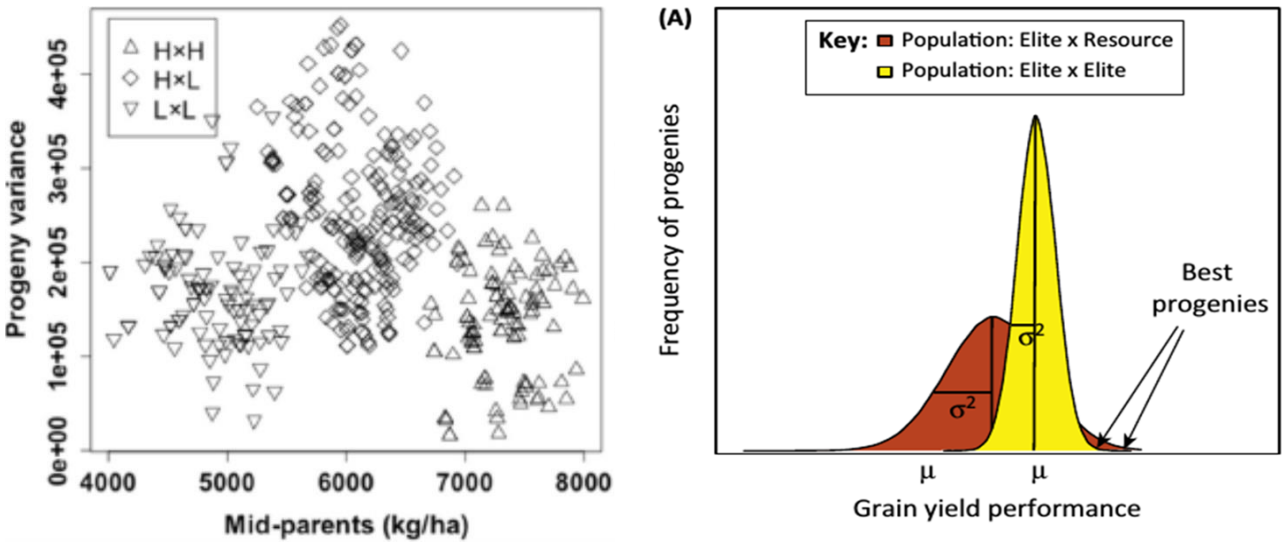


Figure 17 : Bénéfice de la prise en compte de la variance de la descendance dans le cadre d'une covariance négative entre la variance de la descendance et la moyenne des valeurs génétiques des parents.

*A gauche, figure extraite de Mohammadi et al. (2015) montrant la relation triangulaire entre variance de la descendance et espérance des valeurs génétique des parents, pour le rendement, dans une population de maïs. Les couples sont regroupés de trois manières : H*H pour les croisements de deux lignées élites (High), L*H pour les croisements entre une lignée élite et une lignée ressource (L) et H*H pour les croisements entre deux lignées ressources. En ne conservant que la partie droite du graphe (troncature à partie de 6 000kg/ha), on observe une covariance négative entre espérance et variance de la descendance.*

*A droite, figure extraite de Longin et Reif, 2014 montrant les distributions théoriques des descendants de deux couples, un couple formé de deux lignées élites (jaune) et un couple formé d'une lignée élite et d'une lignée ressource (rouge). Le couple élite montre une faible variance σ² mais une forte espérance μ, alors que le couple élite*ressource (L*H) montre une forte variance mais une faible espérance. Ces paramètres sont cohérents avec une covariance négative entre variance et espéance de la descendance. On voit que la covariance négative permet au couple élite*ressource de produire des descendants supérieurs au couple élite*élite.*

Dans le cas d'une covariance négative entre PM et σ, les critères exploitant la variance de la descendance permettent d'identifier des croisements prometteurs, alors que l'information sur les GEBV uniquement ne le permettrait pas. Cette covariance négative associée à une forte variance des variances laissent supposer que les CSC basés sur la variance de la descendance sont particulièrement pertinents dans les programmes de pré-breeding, lorsqu'on souhaite réaliser des croisements entre lignées élites et des ressources génétiques (**chapitre I.1.2.1.3.2.1**). Par contre, la covariance négative ne suffit pas en soi pour assurer la supériorité des CSC basés sur la

variance de la descendance. Il faut tout de même s'assurer que la variance des variances des descendants n'est pas négligeable devant la variance des espérances.

En conclusion, le ratio var(σ)/var(PM) est corrélé au bénéfice supplémentaire apporté par la modélisation de la variance, et par extension de la prise en compte de la recombinaison.

I.2.4.3 Evolution des méthodes pour gérer la perte de diversité génétique in situ

Plusieurs études rapportent que la prédiction génomique accélère la perte de diversité génétique par rapport à la sélection phénotypique (Jannink et al. 2010; Lin et al. 2016), y compris chez le blé tendre (Rutkoski et al. 2016). D'une part, la prédiction génomique diminue l'intervalle de générations, et donc accélère le processus de fixation des allèles. D'autre part, la prédiction génomique favorise aussi la sélection des individus les plus apparentés à la population d'entrainement (et donc entre eux) parce qu'ils sont prédits avec une plus grande précision (Clark et al. 2011; Pszczola et al. 2012). Le déploiement de la prédiction génomique doit donc s'accompagner de mesures pour limiter la perte de diversité génétique.

L'avantage de la prédiction génomique est de pouvoir discriminer les allèles favorables des allèles délétères. En se basant sur les effets des allèles, il est possible de choisir les reproducteurs de manière à augmenter la fréquence des allèles favorables, et symétriquement, de diminuer la fréquence des allèles délétères. On distingue deux catégories de méthodes.

La première méthode consiste à recalculer les GEBV en accordant un bonus aux reproducteurs porteurs d'allèles favorables en fréquence faible dans la population (Jannink et al. 2010; Goddard 2009; Hayes et al. 2009). Ces GEBV « boostées » accordent une plus grande chance aux porteurs d'allèles favorables rares d'être sélectionnés comme reproducteurs à la prochaine génération. Par ailleurs, le bonus est recalculé à chaque génération, ce qui permet de limiter les risques de perdre les allèles favorables. Le choix des pondérations est cependant différent entre les auteurs. Dans un premier temps, Goddard (2009) propose de pondérer les effets des allèles de manière à faire augmenter la fréquence des allèles rares pour maintenir la variance génétique au cours des générations et ainsi générer une réponse à la sélection constante. Jannink et al. (2010) simplifie les pondérations suggérées par Goddard (2009) et propose d'accorder un bonus croissant avec l'effet de l'allèle favorable, de manière à ne pas trop avantager les allèles à faibles effets dont l'avantage sélectif est moins robuste. Dans la prédiction génomique pondérée (WGS ; Weighted Genomic Selection) suggérée par Jannink et al. (2010), les GEBV de chaque individu i se calculent comme :

$$\widehat{g_i^{WGS}} = \sum_j w_{ij}\, \hat{\beta}_j p_j^{-0.5}$$

où w_{ij} est le génotype de l'individu i au SNP j, $\widehat{\beta_j}$ est l'estimateur de l'effet du SNP j et p_j donne la fréquence de l'allèle favorable au SNP j.

La deuxième méthode qui permet d'éviter la perte de diversité utile est décrite dans (Goiffon et al. 2017; Kemper et al. 2012). Ici, la diversité génétique utile est maintenue en maximisant la diversité à l'échelle de la population des parents. La valeur de la population de reproducteurs sélectionnés est nommée OPV (Optimal Population Value) pour Goiffon et al. (2017) ou GB (Genome Building) pour Kemper et al. (2012). La valeur d'une population se calcule comme la valeur génétique de l'individu idéal qui cumulerait tous les allèles les plus favorables apportés par les parents. Ces critères permettent de construire à l'aide d'un algorithme heuristique une population de reproducteurs qui comprend au moins un allèle favorable pour un maximum de loci. La différence entre l'OPV et le GB est le nombre d'allèles favorables différents considérés à chaque locus. Pour l'OPV, imaginé dans le cadre de la sélection végétale de plantes homozygotes, seul un allèle favorable par locus est considéré, alors que dans le GB, on admet que plusieurs haplotypes favorables sont conservés. L'OPV et le GB sont très similaires à l'OHV précédemment présenté, à la différence que l'OHV se calcule pour un couple alors que l'OPV et le GB se calculent pour une population. De plus, ces critères sont présentés par leurs auteurs comme supérieurs pour le gain génétique long terme, et sont donc des critères de sélection de populations plus adaptés au pré-breeding qu'à la production de lignées élites. Dans Goiffon et al. (2017), l'utilité d'une population p est quantifiée par la valeur du meilleur individu théorique homozygote sur tous les blocs. La formule est donc presque la même (exemple pour l'OPV) :

$$\text{OPV(p)} = 2 \sum_{b=1}^{B} \max \widehat{HV_{bn\,(n\,\epsilon\,p)}}$$

où *B* est le nombre de blocs (≤ au nombre de marqueurs), et $HV_{bn\,(n\,\epsilon\,p)}$ est la valeur du meilleur haplotype au bloc b porté par le reproducteur n de la population p.

L'inconvénient des méthodes WGS, OPV, GB est qu'elles sont basées sur des effets aux marqueurs souvent mal estimés (Hofheinz et Frisch 2014). Ces estimations varient au cours des cycles de sélection, à cause de la variation du DL et de la différenciation du fond génétique.

Ainsi, le contrôle de l'apparentement via les méthodes OCS (présentées au **chapitre I.1.2.1.3.2.2**), qui ne présente pas ces faiblesses, reste une approche de choix. Il a pris un nouvel essor chez les animaux avec les marqueurs moléculaires haut-débits qui ont permis de calculer une similarité génétique réelle plus précise que l'apparentement attendu d'après le pedigree (Sonesson et al. 2012). Chez les bovins, le deuxième apport de la prédiction génomique aux OCS est de pouvoir attribuer une valeur génétique à des individus non phénotypés sur la base de leur génotype, et non plus des relations d'apparentement (méthode du Pedigree-BLUP), ce qui permet d'être plus précis quant au gain génétique attendu (Woolliams et al. 2015).

I.2.4.4 Bénéfices des critères d'utilité pour la gestion de la diversité génétique

Le développement de la prédiction génomique s'est accompagné de nombreuses études pour mesurer son impact sur le gain génétique à long terme, avec ou sans contraintes sur la diversité.

Comme décrit précédemment, il existe une variété de critères pour sélectionner les reproducteurs (GEBV, WGS), les croisements (PM, UC, EMBV, OHV), ou encore le plan croisements (OCS, OPV, GB). Les gains génétiques à long terme associés à ces différents critères ont été comparés. Par exemple, Goiffon et al. (2017) montrent que le gain génétique permis par la sélection des reproducteurs sur leur GEBV ou sur des GEBV pondérées (WGS) atteint plus rapidement un plateau que par la sélection de reproducteurs sur les critères OPV, OHV et GB.

Allier et al. (2019c) ont aussi proposé une extension aux OCS nommée UCPC (UC based Parental Contribution). Dans les OCS classiques, le gain génétique associé à un plan de croisements est calculé à partir des valeurs génétiques des reproducteurs (ou des croisements) sélectionnés, pondérées par leurs contributions à la descendance : $c_t g_t$. Or, la valeur génétique d'un reproducteur (ou d'un croisement) n'informe pas sur la valeur génétique des descendants qui seront sélectionnés comme reproducteurs à la génération suivante (t+1). Ces descendants sélectionnés ont plus de chances de faire partie de la fraction haute de la descendance de chaque croisement sélectionné. Autrement dit, le gain génétique dépend de l'espérance de la fraction supérieure de la descendance de chaque croisement, c'est-à-dire leur UC. La fonction Objectif à maximiser devient désormais $c_t^q UC_t$ où c_t^q est la proportion de descendants alloués à chaque couple qui seront sélectionnés (taux de sélection intra-famille q, constant entre les familles).

Par ailleurs, Allier et al. (2019a) proposent aussi de modifier la prise en compte de l'apparentement dans l'OCS. Dans les OCS classiques, l'augmentation de l'apparentement à chaque génération est contrôlée en limitant la quantité $c_t \Phi_t c_t'$ associée au plan de croisements. Or, là encore, cette quantité ignore la sélection qui va être opérée dans la descendance. Allier et al. 2019c montrent que dans un croisement, le génome des descendants supérieurs est généralement majoritairement issu de l'un des deux parents, ce qui est cohérent avec les observations en sélection (Fradgley et al. 2019). Pour prendre en compte l'apparentement des futurs reproducteurs, ils proposent de calculer l'UCPC pour chaque croisement, c'est-à-dire la proportion du génome des descendants supérieurs (sélectionnés) issue de l'un ou l'autre des parents. La diversité génétique d'un plan de croisements est désormais calculée à partir des UCPC et des similarités génétiques entre parents : $c_t^q \Phi c_t^{q\prime}$. Allier et al. 2019a montrent que cette quantité est reliée à l'indicateur de diversité génétique He, qui donne l'hétérozygotie moyenne attendue dans une population panmictique issue de croisements entre les descendants. La diminution de l'hétérozygotie He d'une population au cours d'une génération est inversement proportionnelle à la taille efficace de la population : $He_{t+1} = He_t (1 - 1/2Ne)$ (Falconer et Mackay 1966).

Ainsi, contrôler la baisse du He entre deux générations permet de limiter la diminution de taille efficace et par conséquence la dérive génétique. Les gains génétiques à long terme en utilisant différentes méthodes de sélection des croisements, dont l'UCPC-He, l'OCS, ou encore les critères PM et UC, sont meilleurs avec l'UCPC-He et augmentent lorsque la contrainte sur $c_t^q \Phi c_t^{q\prime}$ est plus sévère.

Conclusion partielle du Chapitre 1.2

Le plan de croisements résulte d'un ensemble de compromis, notamment entre la maximisation du gain génétique à la génération suivante et la sauvegarde de la diversité génétique pour assurer un progrès génétique à long terme.

Il est possible de prédire la distribution de la descendance d'un grand nombre de couples candidats à partir des données de génotypage des parents, d'une carte génétique et d'un modèle de prédiction génomique ajusté sur les phénotypes et génotypes d'une population d'entrainement. Ces éléments sont de plus en plus disponibles dans les programmes de sélection. Pour cette raison, la prédiction de la distribution de la descendance ne représente aucun coût supplémentaire. Par contre, cette prédiction a l'avantage d'explorer le potentiel d'un grand nombre de couples jamais observés auparavant, et ainsi d'identifier en amont les couples les plus prometteurs (forte moyenne et forte variance de la descendance). Enfin, c'est une technique relativement précise, puisqu'elle prend en compte la co-ségrégation des allèles lors de la transmission du patrimoine génétique des parents.

Plusieurs CSC peuvent être définis pour maximiser le gain génétique. Les critères les plus prometteurs prennent en compte la variance de la descendance, qui peut désormais être prédite grâce aux modèles de prédiction génomique. La variance de la descendance dépend notamment de la diversité génotypique des descendants, qui est le résultat d'une recombinaison des génomes parentaux. La fréquence de recombinants entre chaque paire de loci est notée r dans les équations de prédiction de la variance. Ainsi, il est primordial d'optimiser l'estimation du profil de recombinaison r le long du génome et sa variabilité dans différents groupes génétiques.

I.3 Cartes de recombinaison

I.3.1 Recombinaison intrachromosomique permise par les crossing-over

Prédire la fréquence des recombinants dans la descendance d'un croisement nécessite préalablement d'estimer la fréquence des recombinants r entre chaque paire de marqueurs.

Le blé tendre a 21 paires de chromosomes homologues. Au sein de chaque paire, un chromosome a été transmis par le père (grain de pollen) et un chromosome a été transmis par la mère (ovule) de l'individu. Lorsque cet individu (par exemple un F1 chez le blé) produit des gamètes, chaque gamète échantillonne au hasard un chromosome par paire homologue, soit un chromosome paternel soit un chromosome maternel.

Si l'on considère deux loci positionnés sur des chromosomes différents, la fréquence de génotypes non recombinants (paternel/paternel ou maternel/maternel) est égale à la fréquence de génotypes recombinants (paternel/maternel ou maternel/paternel), ces deux fréquences valent donc 0.5.

Par contre, la fréquence de recombinants entre locus positionnés sur une même paire de chromosomes homologues est plus compliqué à prédire. Le processus qui permet d'obtenir de nouvelles combinaisons d'allèles situés sur un même chromosome est la recombinaison homologue, un processus intervenant lors de la prophase I de méiose. La succession des différentes étapes de la méiose est présenté dans la **Figure 18.**

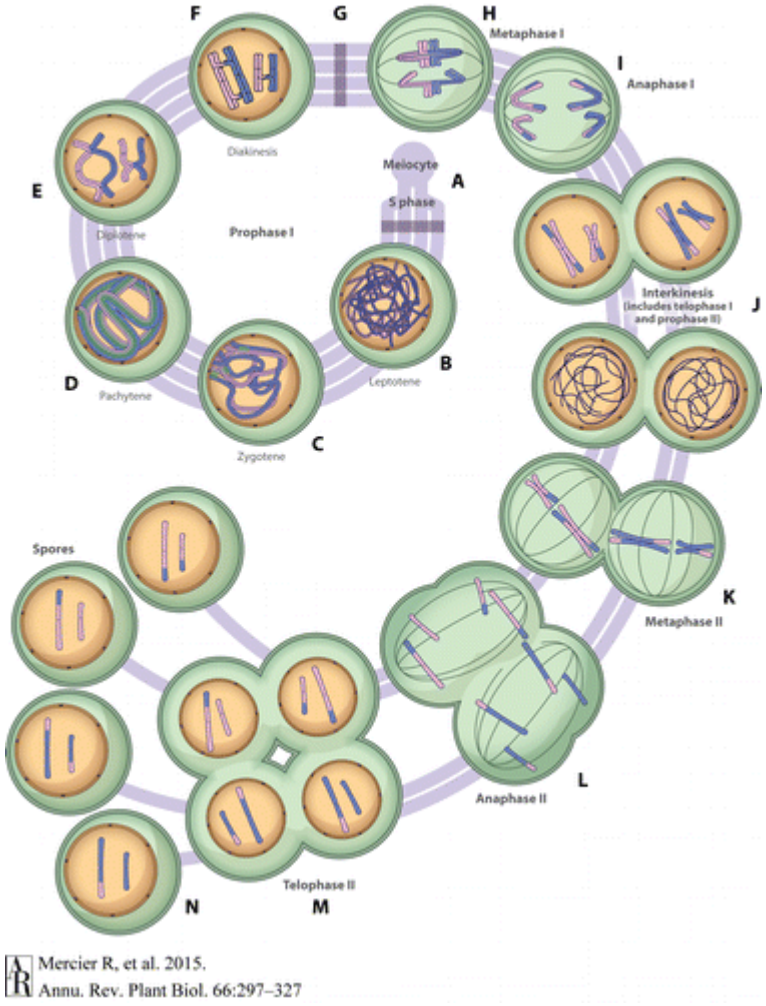


Figure 18 : Schéma résumant les principales étapes de la méiose

Stade pré-méiotique (schéma A) : La cellule de départ est appelée méiocyte, et va se différencier en 4 gamètes. Le méiocyte est diploïde et contient une paire de chromosomes par groupe homologue. Les chromosomes initiaux sont constitués d'une seule chromatide. L'ADN de chaque chromatide est dupliqué à l'identique (sauf rares erreurs de duplication) et les chromosomes sont alors constitués de 2 chromatides sœurs.

Stade prophase I (B-F) : L'ADN se condense. Les chromosomes homologues sont physiquement liés par paire via l'élaboration d'un complexe synaptonémal, et forment des bivalents à quatre chromatides. Suite à des coupures locales de l'ADN par la nucléase SPO11, deux chromatides de chaque chromosome homologue échangent réciproquement de larges portions d'ADN, depuis le point de rupture jusqu'à l'extrémité des chromosomes. Ces échanges sont appelés crossing-over. Les chromatides sont désormais recombinées.

Stade métaphase I (H) : les chromosomes s'alignent sur le plan équatorial de la cellule. La répartition des chromosomes homologues de part et d'autre du plan est aléatoire.

Stade anaphase I (I) : Les liaisons entre chromosomes sont rompues et chaque chromosome homologue migre vers un pôle de la cellule.

Interkinesis (J) : Formation de deux cellules filles haploïdes, chacune contenant un lot de chromosomes, avec un exemplaire par groupe homologue.

Métaphase II (K) : les chromosomes se placent sur le plan équatorial de la cellule, le positionnement de leurs chromatides de part et d'autre du plan étant aléatoire.

Anaphase II (L) : les chromatides sont séparés et migrent vers les pôles opposés de la cellule.

Télophase II et fin de la différenciation (M-N) : Formation des gamètes haploïdes, avec un chromosome à une seule chromatide par groupe homologue.

Mercier et al. (2015)

La recombinaison homologue implique l'intervention successive de nombreuses protéines spécifiques de la méiose largement conservée entre eucaryotes (Gerton et Hawley 2005; Lam et Keeney 2015). La recombinaison homologue est le résultat d'un échange réciproque de segments d'ADN entre chromosomes homologues durant la formation des gamètes. Ces échanges de segments d'ADN sont appelés crossing-over (CO). La formation des CO est initiée par une cassure double brin (DSB, Double Strand Break) de l'ADN sur l'une des deux chromatides d'un chromosome, catalysée par la nucléase SPO11. Seule une petite fraction des DSB sera réparée en CO. Environ 250 DSB sont observés par méiose chez *Arabidopsis*, et seulement 8 sont résolus en CO (Mercier et al. 2015). Il existe deux voies conduisant à la formation de CO. La voie principale produit des CO de type « interférents » (ou type I) qui empêchent la formation d'un autre CO à proximité. La deuxième voie produit des CO « non interférents » (type II) qui sont indépendant de la proximité d'autres CO. Quand les coupures ne se résolvent pas en CO, soit le brin d'ADN cassé se réforme à partir du modèle de la chromatide sœur (celle du même chromosome), soit la coupure se résout en Non-Crossing-Over (NCO), c'est-à-dire un transfert de séquence unilatérale entre deux chromatides homologues non sœurs. Les NCO sont rares ou difficiles à détecter car courts (quelques paires de bases).

Le rôle supposé des CO est d'assurer le placement correct des bivalents sur le plan équatorial de la cellule pour permettre une ségrégation des chromosomes dans les gamètes (Hassold et Hunt 2001). En effet, l'aneuploïdie (nombre incorrect de chromosomes par gamète) est délétère voire fatal pour le futur individu.

I.3.2 Cartes génétiques établies à partir de ségrégations familiales

La fréquence des nouvelles combinaisons d'allèles parentaux dans les gamètes dépend de la probabilité d'occurrence des CO entre locus. Cette probabilité doit être estimée à partir de données expérimentales permettant la production de *cartes génétiques*. Il existe plusieurs méthodes pour obtenir une carte génétique.

La méthode la plus classique pour estimer la fréquence des CO le long des chromosomes est d'observer la transmission du patrimoine génétique entre des parents et des descendants génotypés. Les associations d'allèles inédites dans la descendance par rapport aux parents traduisent l'occurrence d'un CO durant la méiose.

Lorsque la position physique des marqueurs le long des chromosome est inconnue, l'analyse des fréquences de recombinaison permet aussi d'obtenir une carte du génome, c'est-à-dire regrouper les marqueurs par chromosome et les ordonner le long des chromosomes. Cependant cette application est difficilement réalisable lorsque le nombre de marqueurs est important du fait du nombre exponentiel d'ordres possibles (Peñalba et Wolf 2020). Désormais, de nombreuses espèces, dont le blé tendre, disposent d'un génome de référence (séquence d'ADN pour les

différents chromosomes) sur lequel les marqueurs peuvent être positionnés, donc ordonnés (Choulet al. 2014, Rimbert et al. 2018, IWGSC 2014, IWGSC 2018) et la production de cartes génétiques ne nécessite que l'estimation des taux de recombinaison entre marqueurs adjacents.

La fréquence de recombinants n'est pas l'unité habituelle des cartes génétiques. On lui préfère la distance génétique qui est l'espérance du nombre de CO par méiose entre deux locus et a l'avantage d'être additive. L'unité couramment utilisé pour décrire une distance génétique est le centimorgan (cM). Une distance génétique de 1 cM entre deux locus signifie qu'en moyenne 1 CO pour 100 méioses a lieu dans l'intervalle.

Le passage de la fréquence des recombinants entre deux locus à la distance génétique (ou la réciproque) est permis par des fonctions de cartographie (**Figure 19**). Il en existe plusieurs, qui modélisent différemment la fréquence des doubles CO (Zhao et Speed 1996; Crow 1990). La fonction la plus simple est la fonction de Morgan, dans laquelle chaque recombinaison observée n'est le produit que d'un CO. C'est une fonction valable pour des petits intervalles, où la probabilité d'observer deux CO successifs est faible.

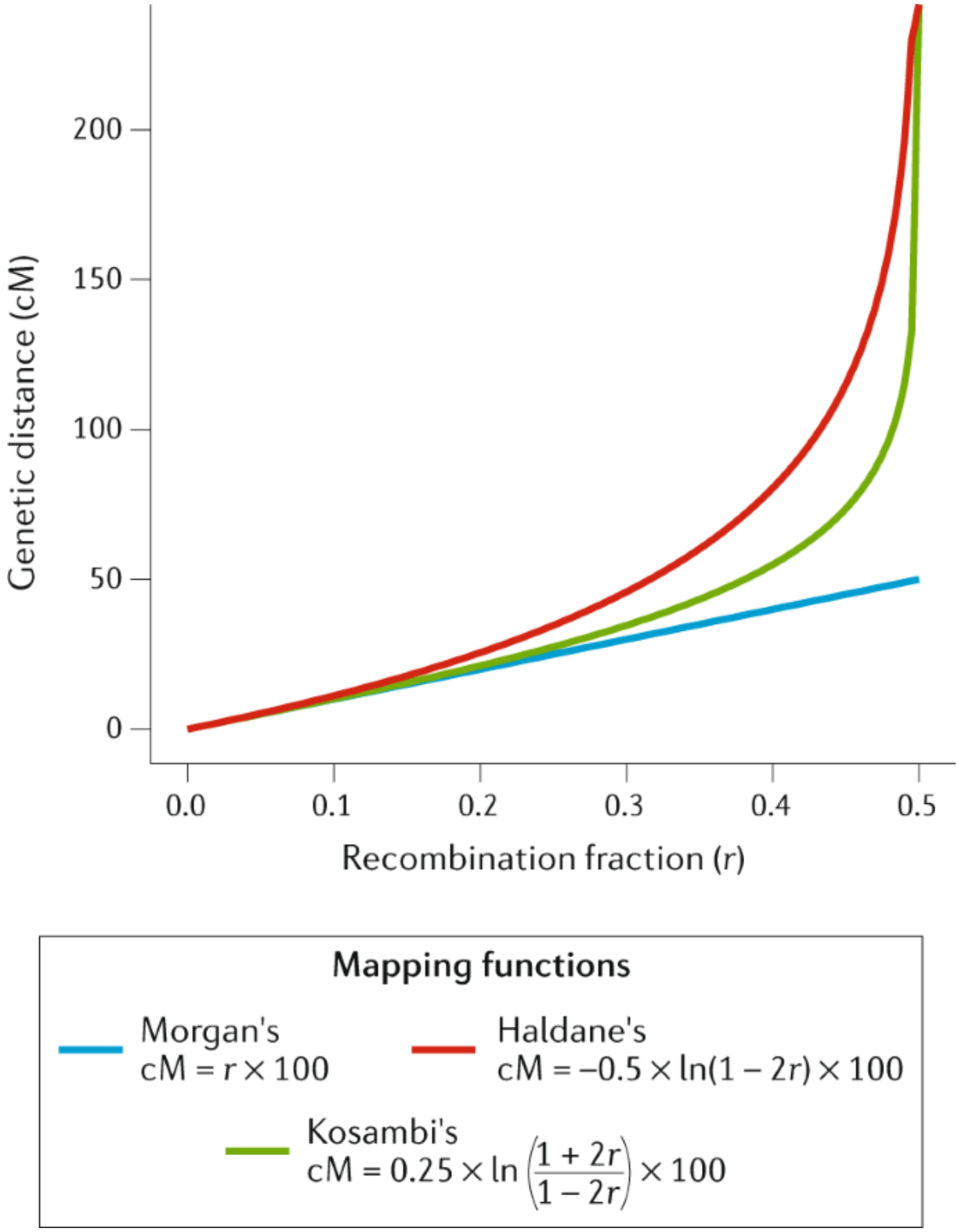


Figure 19 : Relation entre la fréquence des recombinants r et l'espérance du nombre de crossing-over par méiose, pour trois fonctions de cartographie

Peñalba et Wolf, 2020

La position génétique relative entre marqueurs forme une carte génétique. Le ratio entre la distance génétique d'un intervalle et sa taille physique donne le taux de recombinaison (en centiMorgans/Mégabase). Le profil de recombinaison est le vecteur des taux de recombinaison le long du génome. Ce profil obtenu à partir de populations familiales en ségrégation est aussi appelé « profil méiotique » puisqu'il représente directement le taux de recombinaison des méioses produisant les gamètes.

Chez les plantes, le profil de recombinaison méiotique est généralement estimé dans une population de RILs dérivés du croisement de deux parents, ou du croisement entre plusieurs parents (Rimbert et al. 2018; Gardiner et al. 2019; Guo et al. 2020; Vaissayre et al. 2012; Jordan et al. 2018). Cependant, l'estimation des fréquences de recombinaison par cette méthode classique présente trois inconvénients majeurs. Le premier inconvénient est une faible précision dans l'estimation des taux de recombinaisons due au petit nombre de méioses échantillonnées. Un deuxième inconvénient relève de la faible densité de marqueurs polymorphes dans des croisements bi-parentaux. Un plus faible polymorphisme s'accompagne de plus grands blocs IBD entre parents (Tiret et Hospital 2017), et donc une difficulté à positionner précisément les CO. Enfin, le profil de recombinaison d'une population donnée est peu représentatif d'autres populations. Chez les plantes, les variations structurales sont fréquentes, et en particulier l'introgression de segments génomiques issue d'espèces apparentées Par exemple, dans la population biparentale dérivée du croisement entre les variétés Chinese Spring et Renan (Choulet et al. 2014, Rimbert et al. 2018), la variété Renan présente deux introgressions du génome d'une autre espèce sur les chromosomes 2A et 7D ce qui induit un blocage de la recombinaison dans ces deux zones. Les variations structurales entrainent une suppression locale des CO, à l'origine de variations du profil de recombinaison entre individus (Bauer et al. 2013; Rowan et al. 2019).

Un moyen de contourner ces problèmes est de produire de très grandes populations dérivées d'un grand nombre de parents très diversifiés, mais cela à un coût prohibitif. Une alternative moins couteuse consiste à exploiter la relation entre les patrons de DL d'un panel de diversité et la variation locale du taux recombinaison (Stumpf et McVean 2003).

I.3.2 Patrons de déséquilibre de liaison

Un haplotype est définit comme une succession d'allèles portés par un chromosome. Les recombinaisons qui ont eu lieu au cours de l'histoire (de la généalogie) d'une population permettent d'expliquer en partie la diversité des haplotypes observables dans une population.

Pour cela, on définit une population de Wright-Fisher comme une population idéale, affectée par aucune force évolutive, de taille constante et sans génération chevauchante et se reproduisant en panmixie.

Dans une population de Wright-Fisher, remonter la généalogie d'une paire d'haplotypes (haplotypes aux mêmes loci, portés par deux chromosomes/individus différents) non recombinants aboutit à un évènement de coalescence qui correspond à la génération à laquelle ces deux haplotypes sont issus du même individu ancestral. Cet individu est appelé l'ancêtre commun le plus récent (Most Recent Common Ancestor MRCA en anglais). Le même processus pour l'ensemble des haplotypes de la population (mêmes locus, différents chromosomes/individus) produit un arbre de coalescence, dont les branches sont les différents épisodes de coalescence et dont la racine est l'ancêtre commun à tous les haplotypes contemporains.

Dans une région recombinante, différentes parties des haplotypes ont des MRCA différents mais leurs arbres de coalescence sont corrélés (se ressemblent) (**Figure 20**). L'intensité de cette corrélation dépend de la transmission conjointe de ces blocs haplotypiques au cours de la généalogie de la population, qui dépend du nombre de recombinaisons dans la généalogie, et donc du taux de recombinaison méiotique.

Puisque la recombinaison influence la généalogie des haplotypes observés dans une population, il a été suggéré d'utiliser les observations sur les corrélations entre allèles pour estimer les taux de

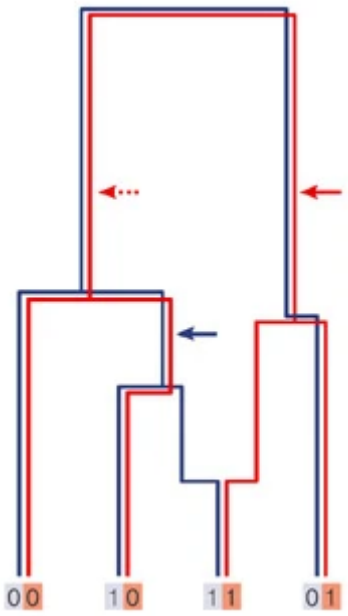


Figure 20 : Arbre généalogique ancestral d'un haplotype

Cet arbre donne l'histoire évolutive d'une population constituée de 4 haplotypes (bas du graphique) dérivé d'un même ancêtre commun. Les lignes bleues et rouges indiquent la généalogique des loci de l'ancêtre commun (bleu pour le locus 1, rouge pour le locus 2) Les allèles de l'ancêtre commun sont notés 0. Les flèches en trait plein indiquent les événements de mutation sur l'un des deux locus. Les locus mutés portent désormais l'allèle 1. La dissociation des lignes bleu et rouge indique une recombinaison.

Stumpf et McVean (2003)

recombinaison. Par exemple, l'évolution de la covariance entre allèles, notée D, au cours du temps peut s'écrire : $E(D_{t+1})$ = (1-c)(1-1/2Ne)D_t (Hill et Robertson 1966), avec c la distance génétique.

Cependant, le DL n'est pas homogène le long du génome, ce qui signifie qu'il y a des segments où les haplotypes ont été moins souvent recombinés que d'autres, autrement dit que la fréquence des recombinaisons varie le long du génome. Plus les recombinaisons ont été fréquentes dans un intervalle au cours de l'histoire de la population, plus les associations d'allèles au niveau de ce segment sont équilibrées, autrement dit le déséquilibre de liaison est faible. A contrario, un segment où les CO ont été rares par le passé montre encore un fort déséquilibre de liaison

Il existe plusieurs mesures du DL. Parmi les plus courantes, il y a la covariance entre allèles (D, **chapitre I.2.2**) et le R^2, qui se calcule comme la corrélation entre paire d'allèles à différents loci pour chaque chromosome de chaque individu, au carré. Plus la valeur du R^2 est élevée, plus le déséquilibre de liaison est élevé, et inversement. On visualise ainsi les variations du DL le long du génome (**Figure 21**).

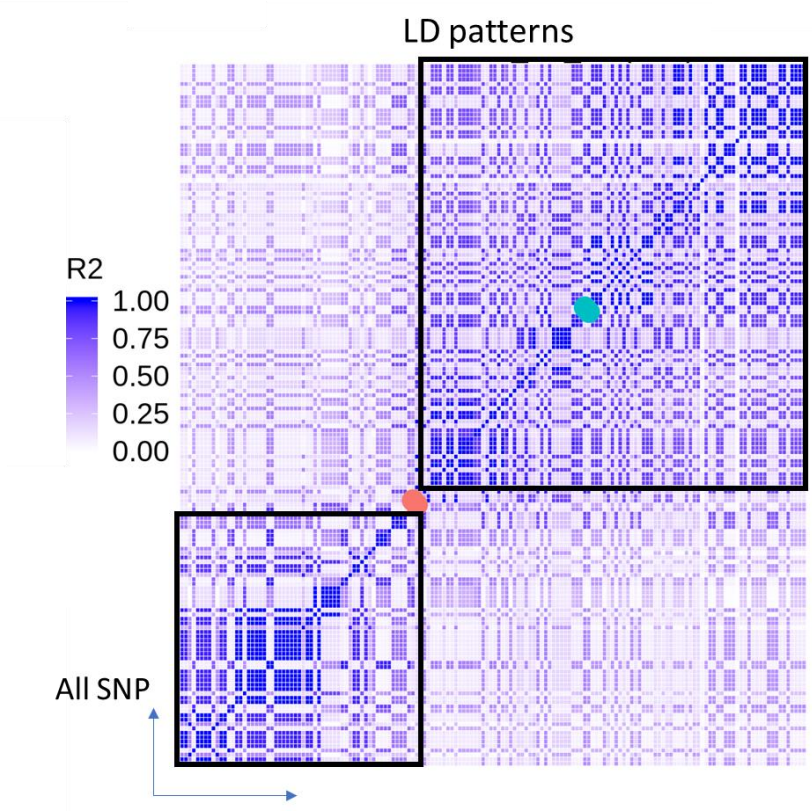


Figure 21 : Variation du DL dans une fenêtre du chromosome 3B dans une population européenne de landraces de blé tendre.

Chaque case donne la valeur du R^2 pour une paire de locus. On remarque deux segments génomiques, matérialisés par des cadres noirs, où le DL est fort à l'intérieur du segment, mais faible entre les segments, ce qui suggère que les recombinaisons intra-segment ont été rares par le passé, mais fréquentes entre les segments.

Cependant, les mesures simples du DL ne fournissent pas un estimateur du taux de recombinaison sous-jacent. Parmi les modèles qui infèrent le profil de recombinaison à partir des patrons de DL, on peut citer celui de Li et Stephens (2003) implémenté dans le logiciel PHASE, ainsi que les programmes LDhat et LDhelmet (Chan et al. 2012) (l'ensemble des logiciels est présenté dans Peñalba et Wolf (2020).

A la différence de l'estimation du profil de recombinaison sur des populations familiales en ségrégation, ces approches historiques ne permettent pas d'estimer directement le taux de recombinaison par génération. En effet, les patrons de DL résultent de l'ensemble des recombinaisons et mutations qu'a subi une population depuis sa coalescence. En l'absence d'information sur l'histoire évolutive de la population, les patrons de DL ne suffisent pas pour estimer directement le taux de recombinaison par génération. A la place, ces algorithmes cherchent à estimer le taux de recombinaison historique ρ, qui peut être vu comme le nombre de recombinaisons efficaces nécessaires pour expliquer le DL dans la population compte tenu de sa

diversité génétique (Ptak et al. 2005). C'est pourquoi ces profils sont qualifiés de profils historiques ou populationnels. Pour une population théorique se reproduisant en panmixie, de taille constante, non soumise aux forces évolutives, le modèle de Wright-Fisher décrit que ρ = 4*Ne*c. Le terme c correspond au taux de recombinaison par génération et le terme Ne correspond à la taille efficace de la population. Chez les espèces autogames comme le blé tendre, Nordborg (2000) propose ρ= 4Nec (2-s)/2 où s est le taux d'individus issus d'autofécondation par génération (proche de 95% chez le blé, s doit être <1 dans ce modèle).

L'approche que j'ai utilisée dans le **chapitre II** du manuscrit pour estimer le taux de recombinaison à partir de données de diversité est le modèle de Li et Stephens (2003). Ce modèle se base sur un principe d'approximation de la vraisemblance du paramètre ρ. Etant donnée une collection d'haplotypes H = $\{h_1, h_2, \ldots h_n\}$, la vraisemblance P(H | ρ) peut s'écrire :

$$P(H \mid \rho) = p(h_1 \mid \rho)\, p(h_2 \mid h_1, \rho)\, p(h_3 \mid h_1, h_2, \rho) \ldots P(h_n \mid h_1, h_2, \ldots h_{n-1}, \rho)$$

Individuellement les termes de ce produit ne sont pas calculables. Le modèle de Li et Stephens propose de les approximer en modélisant un haplotype comme une mosaïque d'haplotypes connus (**Figure 22**). Ceci aboutit à modéliser un haplotype comme par un modèle de Markov caché dont les états cachés sont les haplotypes connus. Un état caché désigne l'appartenance (l'origine commune) d'un marqueur à un autre haplotype de la population. Cet état est dit « caché » car sa vraie appartenance est inconnue. En effet, il y a un grand nombre de possibilités lorsqu'on souhaite reconstruire un haplotype comme une mosaïque d'autres haplotypes. Par ailleurs, on suppose qu'on n'observe pas directement les « vrais haplotypes » des individus, car ceux-ci peuvent avoir été altérés par mutation au cours de l'évolution. La probabilité de transitionner (changer d'appartenance) entre deux états cachés dépend du taux de recombinaison. La probabilité d'observer les vraies données sachant les états cachés est appelée probabilité d'émission. La vraisemblance d'une séquence d'états cachés est le produit des probabilités de transitions et des probabilités d'émissions. La vraisemblance d'un haplotype sachant les autres haplotypes et sachant le taux de recombinaison est la somme des vraisemblances de chaque état caché. La vraisemblance totale de tous les haplotypes est maximale lorsque le taux de recombinaison permet de reconstruire les haplotypes de chaque individu sans faire intervenir trop souvent la mutation ou la recombinaison. Un *a priori* est placé par l'utilisateur du modèle sur la distribution des valeurs vraisemblables de taux de recombinaison ρ. Pour identifier les gammes de valeur vraisemblables du taux de recombinaison parmi cette distribution à priori, le modèle utilise une chaîne de Monte Carlo, qui consiste à tester successivement différents taux de recombinaison, et à les accepter avec une probabilité qui dépend de l'augmentation/diminution de la vraisemblance associée aux nouvelles valeurs du taux de recombinaison. Sous réserve d'attendre suffisamment longtemps, la

chaine de Markov couplée au processus de Monte Carlo (MCMC) permet d'obtenir des distributions à posteriori vraisemblables pour le taux de recombinaison.

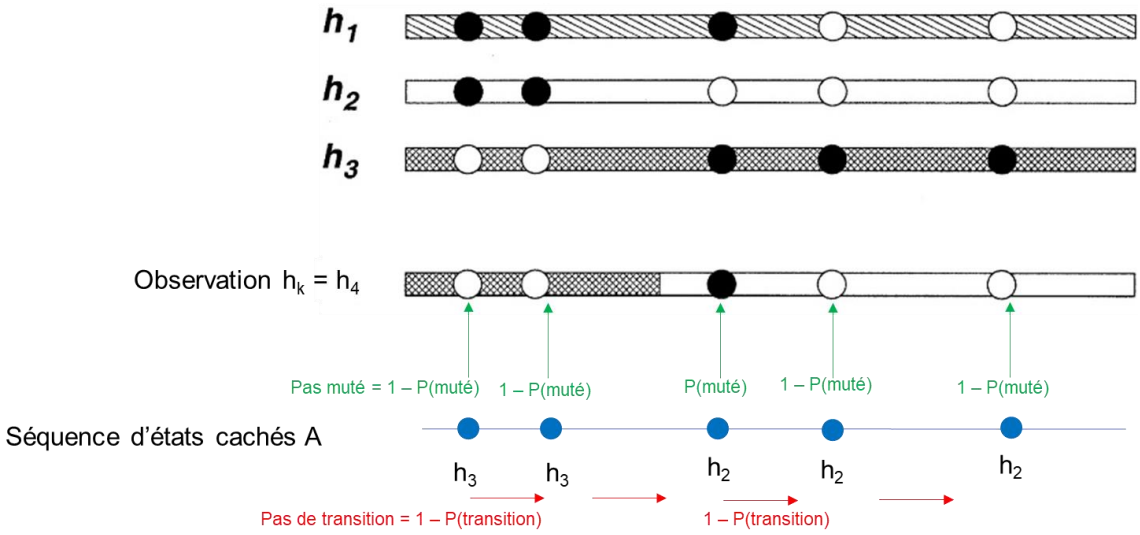


Figure 22 : Exemple de la reconstruction d'un haplotype comme une mosaïque d'haplotypes des autres individus

Un exemple de séquence d'état cachés (indicé « séquence A ») est donné.

Inspiré de Li et Stephens, 2003

L'estimation des taux de recombinaison à partir des patrons de DL présente plusieurs avantages par rapport aux populations familiales en ségrégation. D'une part, les patrons de DL sont observés dans un panel de diversité. Le polymorphisme est donc beaucoup plus élevé, ce qui autorise à obtenir des profils de recombinaison résolutifs. Par ailleurs, les patrons de DL permettent d'accéder au profil de recombinaison moyen de la population qui est moins influencé par les variations individuelles. Enfin, les patrons de DL résultent de l'ensemble des recombinaisons depuis la coalescence de la population, soit potentiellement un grand nombre de générations, et par conséquence de méiose. Ainsi, en théorie, ils permettent d'accéder à des estimateurs beaucoup plus précis et à des variations beaucoup plus fines du taux de recombinaison entre intervalles.

Cependant, l'inférence à partir des patrons de DL présente aussi des inconvénients. Les modèles qui infèrent la recombinaison à partir des patrons de DL supposent que seules la mutation et la recombinaison sont à l'origine des patrons de DL observés. Or, les patrons de DL dépendent des fréquences alléliques qui sont sensibles aux forces évolutives (migration, sélection, dérive) et aux évènements démographiques (goulot d'étranglement, expansion) (Dapper et Payseur 2018; Chan, et al. 2012; Charlesworth et Charlesworth 2010). En conséquence, les taux de recombinaison historiques doivent être analysés avec prudence.

Les profils de recombinaison obtenus à partir de données familiales ne sont pas influencés par ces biais. Ainsi, les articles qui travaillent sur les profils historiques comparent généralement les variations des profils historiques avec des profils méiotiques (Petit et al. 2017; Singhal et al. 2015; Marand et al. 2019; Coop et al. 2008; Fuentes et al. 2021; Rodgers-Melnick et al. 2015; Kong et al. 2010). Une adéquation globale est attendue, mais il est difficile de conclure sur les variations locales du taux de recombinaison estimé entre les deux approches. Dans certains cas, les déviations nettes entre profils historiques et méiotiques ont pu être attribuées à des gènes sous sélection (Petit et al. 2017).

I.3.4 Hétérogénéité du profil de recombinaison

Les cartes méiotiques ou historiques du taux de recombinaison ont permis d'apprécier la variation du taux de recombinaison le long du génome. Chez les espèces à gros génome, comme le blé et l'orge, la recombinaison est élevée au niveau des télomères et faible voire inexistante au niveau des régions péri-centromériques (**Figure 23**). Choulet et al. (2014) ont partitionné les chromosomes de blé tendre en 5 grandes régions en fonction du taux de recombinaison moyen mais aussi du contenu génomique de chaque région : les télomères R1 et R3 très recombinants, les péricentromères R2a et R2b peu recombinants et les centromères C où la recombinaison est quasiment absente. Chez le blé, environ 80% des CO ont lieu dans 20% du génome.

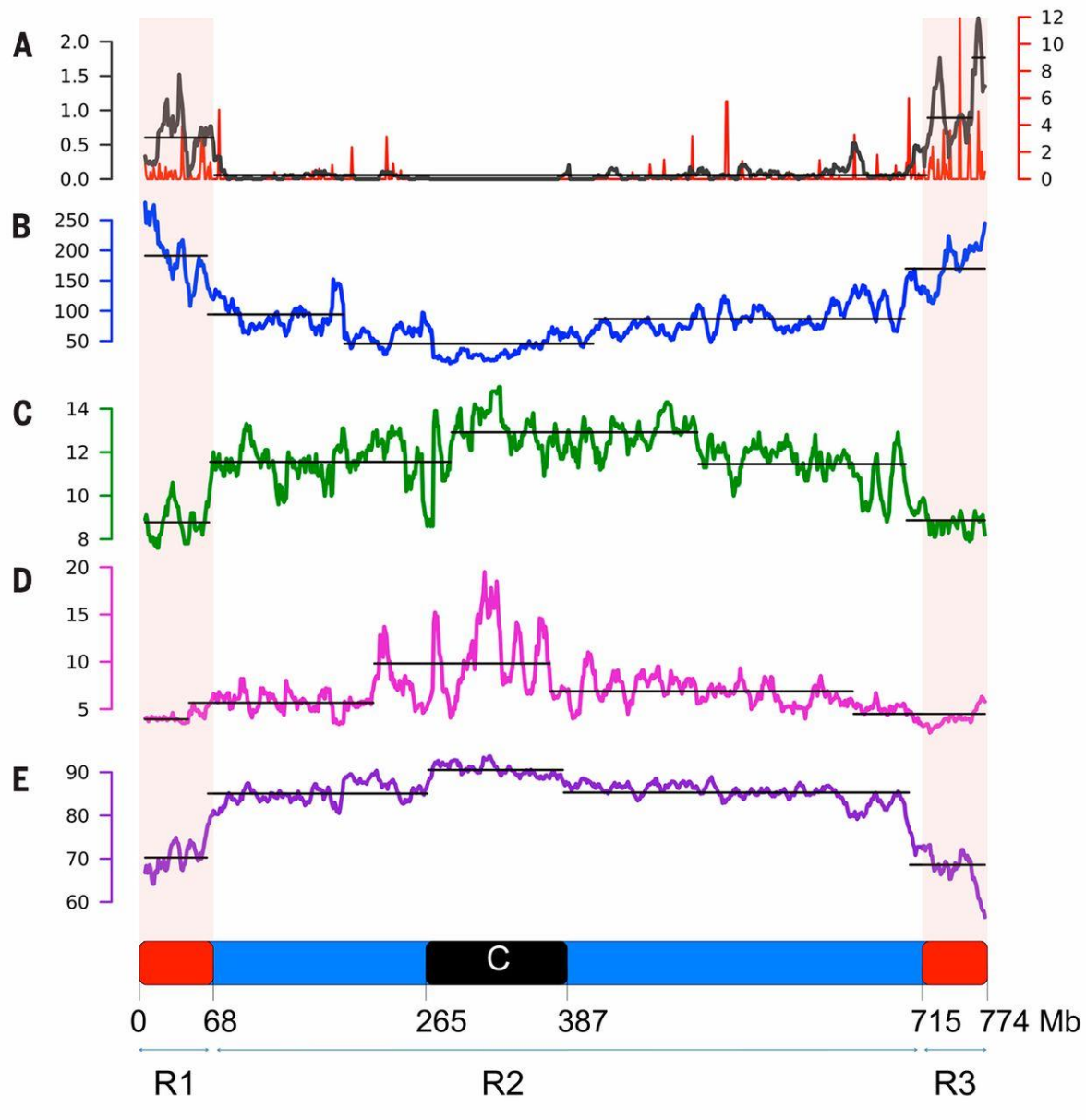


Figure 23 : Partionnement des chromosomes de blé tendre

Exemple du chromosome 3B.

***A** : Taux de recombinaison (cM/Mb) moyenné par fenêtre de 1 Mb (rouge) ou 10 Mb (noir).*

***B** : Densité de gènes.*

***C et D** : Mesures d'expression*

***E** : Densité en éléments transposables*

Choulet et al. (2014)

Chez l'orge, la formation préférentielle de CO au niveau de télomères s'expliquerait par la chronologie des mécanismes méiotiques. Le complexe synaptonémal se forme d'abord aux télomères puis s'étend progressivement au reste du chromosome, ce qui autorise une formation précoce des CO aux régions télomériques, qui interfèrent avec les CO positionnés aux parties plus proximales (Higgins et al. 2012). Cependant, l'influence de la distance au centromère sur le positionnement des CO est encore débattue. Des expériences d'inversions de bras de chromosome de blé suggèrent que la position des CO le long des chromosomes s'expliquerait par le contenu génomique spécifique (Lukaszewski et al. 2012).

A une échelle plus réduite, on observe des points chauds de recombinaison (et des zones froides de recombinaison), définis comme des intervalles génomiques où le taux de recombinaison est significativement plus élevé (respectivement plus faible) que le taux de recombinaison moyen de la région génomique incluant l'intervalle. Les points chauds se comptent par milliers le long du génome, leur nombre et leur taille dépendent vraisemblablement de l'espèce, de la méthodologie de détection et de la résolution en marqueurs. Par exemple, 3 000 points chauds ont été mis en évidence chez *Mimmulus guttatus* (Hellsten et al. 2013), plus de 14 000 chez le riz (Marand et al. 2018) et de l'ordre d'une dizaine de millier chez le cacao (Schwarzkopf et al. 2020). Leur taille fait en général une dizaine de kb, bien qu'elle dépende de la résolution en SNP pour les détecter (de 500 à 23kb chez les plantes en général, d'après la synthèse bibliographique de Choi et Henderson (2015) ; moins de 2kb chez le riz d'après Marand et al. 2019, moins de 6kb chez le cacao d'après Schwarzkop et al. (2020).

I.3.5 Variation des cartes de recombinaison entre individus

La pertinence d'une carte génétique construite sur une population pour prédire la fréquence de recombinants dans une autre population dépend de la variabilité génétique du profil de recombinaison. La variation des cartes de recombinaison peut être analysée sous l'angle de deux phénotypes : le taux moyen de recombinaison, et la localisation des CO.

1.3.5.1 Variabilité du taux moyen de recombinaison

Chez 80% des espèces, dont le blé tendre, le nombre de CO entre chromosomes homologues par méiose est généralement compris entre un et trois, indépendamment de la taille du chromosome (Mercier et al. 2015). Un minimum d'un CO par méiose est nécessaire pour assurer le placement correct des bivalents sur le plan équatorial de la cellule et assurer une ségrégation correcte des chromosomes dans les gamètes (Hassold et Hunt 2001). Un nombre anormal de chromosomes par gamète (aneuploïdie) est fortement délétère voire fatal pour le futur individu. La stabilité

évolutive du nombre de CO par méiose suggère qu'il existe un déterminisme génétique qui contrôle cette quantité. Ce déterminisme est probablement une combinaison de mécanismes. Par exemple, l'interférence réduit le nombre de CO de type I (Berchowitz et Copenhaver 2010). Par ailleurs, le taux de recombinaison moyen est aussi contrôlé par le dosage de différentes voies métaboliques qui inhibent ou catalysent les processus métaboliques impliquées dans la formation des CO. D'après Fernandes et al. 2017, trois processus moléculaires impliqués dans l'inhibition des CO ont été mis en évidence, impliquant plusieurs protéines dont l'hélicase atFANCM (Crismani et al. 2012; Girard et al. 2014), les hélicases du groupe RECQ4 (Séguéla-Arnaud et al. 2015) et la protéine FIDGETIN-LIKE-1 (Girard et al. 2015). A la différence des voies métaboliques inhibant la formation des CO, la ligase HEI10 semble agir comme un catalyseur de la formation des CO (Serra et al. 2018).

Ce déterminisme qui maintient un faible nombre de CO pourrait être le résultat d'un processus sélectif. Il a été supposé qu'un trop grand nombre de CO engendrerait une instabilité mécanique de l'appariement à l'origine d'une ségrégation incorrecte des chromosomes (Mercier et al. 2015). Une autre hypothèse est qu'il existe un niveau optimal de recombinaison pour assurer la perpétuité de l'espèce. En effet, la recombinaison facilite l'adaptation grâce à la création de nouvelles combinaisons alléliques, mais peut aussi dissocier les combinaisons favorables, réduisant la valeur sélective de la descendance (Otto 2009).

Bien que le nombre de CO par méiose soit fortement contraint, il existe tout de même une légère variabilité génétique du nombre de CO par méiose (Stapley et al. 2017), y compris chez le blé tendre (Esch et al. 2007; Wingen et al. 2017; Jordan et al. 2018; Gardiner et al. 2019). Par exemple, Jordan et al. (2018), ont identifié 40 QTLs qui augmentent le nombre de CO par méiose de 7% en moyenne. Par ailleurs, Dreissig et al. (2019) ont montré que le taux de recombinaison moyen chez l'orge dépendait des conditions environnementales (température, sécheresse, luminosité, précipitations), ce qui suggère une certaine plasticité ou une variabilité génétique naturelle du nombre moyen de CO.

I.3.5.2 Variation du profil de recombinaison entre individus

A l'échelle chromosomique, le profil de recombinaison est bien conservé entre groupes génétiques apparentés.

La comparaison fine de cartes génétiques révèle des différences dans la localisation des points chauds entre groupes apparentés. Chez le riz, Marand et al. 2019 ont montré que 80% des points chauds n'étaient pas partagés entre deux sous-espèces de riz, O.*sativa japonica* et *O.s. indica*. Chez le cacao, Schwarzkopf et al. (2020) montrent que 55% des points chauds ne sont pas partagés entre 8 sous-espèces. Pour le blé tendre, Darrier et al. (2017) ont comparé le profil de recombinaison de trois populations (une population Européenne, une population asiatique et une

population dérivée d'un croisements bi-parentale) sur deux régions d'environ 1 Mb chacune. Le profil de recombinaison montre des similarités frappantes, mais aussi des divergences. L'extrapolation de ce résultat à l'ensemble du génome requiert une estimation du profil de recombinaison complet pour plusieurs groupes génétiques.

L'existence de points chauds montre que le profil de recombinaison admet un certain déterminisme, et leur variabilité d'une population à l'autre suggère que ce déterminisme évolue. Les facteurs qui gouvernent précisément le profil de recombinaison chez les plantes et leur variabilité génétique restent globalement incompris. Le statut décompacté de la chromatine semble être le facteur principal qui détermine la position des CO. Ceci peut s'expliquer par le fait que la formation d'un CO résulte de l'action d'un cortège protéique volumineux nécessitant une chromatine décompactée pour accéder à l'ADN (Pan et al. 2011). Puisque la décompaction de la chromatine est le déterminant principal de la localisation des CO, chez une majorité d'espèces les points chauds de recombinaison se retrouvent à proximité des régions codantes, décompactées pour permettre la transcription, en particulier dans les régions 5' des gènes chez le blé tendre (Darrier et al. 2017).

Chez certains mammifères (humain, souris, chimpanzés, bovins, chevaux) (Ptak et al. 2005; Beeson et al. 2019; Sandor et al. 2012; Brunschwig et al. 2012), les points chauds sont localisés à distance des gènes. Chez ces espèces, la protéine PRDM9 repère des motifs de séquence d'ADN (CCNCCNTNNCCNC) et dépose à proximité une marque de méthylation des histones (H3K4me3). Cette marque est à son tour ciblée par des protéines impliquées dans la formation des CO. Ainsi, les portions d'ADN contenant les motifs reconnus par PRDM9 deviennent des points chauds de recombinaison. Chez ces espèces, les allèles de PRDM9 présentent de fortes signatures de sélection positive associées à une diversification des séquences cibles, et donc une forte variabilité du profil de recombinaison entre groupes génétiques (Baudat et al. 2009). Cette sélection diversifiante est une solution possible au paradoxe des points chauds (Lam et Keeney 2015). Ce paradoxe consiste en une inadéquation entre l'observation de points chauds et une disparition prévisible des cibles de PRDM9. En effet, la recombinaison a des conséquences délétères sur les séquences locales. Elle est notamment facteur de mutations (Tiemann-Boege et al. 2017). De plus, au niveau des segments ayant subi des CO ou NCO, il peut y avoir des transferts unidirectionnels de petits segments d'ADN entre chromatides homologues non sœurs, appelés conversions géniques. Chez la majorité des eucaryotes, on observe que ces conversions géniques augmentent la fréquence des nucléotides C et G vis-à-vis des nucléotides A et T (« conversion génique biaisée »). Ainsi, les segments d'ADN où les CO ont été fréquents par le passé sont généralement associés à un fort taux de nucléotides GC (Halldorsson et al. 2016; Jensen-Seaman et al. 2004; Galtier et al. 2001; Meunier et Duret 2004; Duret et Arndt 2008).

Ainsi, les séquences cibles de PRDM9 sont progressivement érodées par mutation, et devraient à terme perdre leur affinité avec PRDM9. En l'absence d'un mécanisme de régénération des séquences cibles ou d'une mutation de PRDM9, les points chauds devraient donc disparaître. La sélection diversifiante sur PRDM9 permet ainsi d'expliquer qu'il soit possible d'observer des points chauds de recombinaison (Oliver et al. 2009). De plus, l'avantage évolutif supposé de PRDM9 est de détourner les CO des régions codantes vers des régions non codantes, afin de limiter l'impact mutagène de la recombinaison sur les séquences fonctionnelles.

Chez les plantes, PRDM9 n'existe pas (Zhang et Ma 2012) et le modèle dominant veut que la recombinaison soit concentrée dans les régions décompactées « par défaut », c'est -à-dire les régions exprimées. Cependant, il est possible que la formation des DSB, et par conséquent des CO, résulte d'un mécanisme plus complexe et que la décompaction locale de l'ADN n'explique pas tout. Les forts taux de recombinaison sont associés à certaines signatures génomiques, comme la présence de formes particulières de méthylation des histones (Choi et al. 2013; Underwood et al. 2018), ou bien un enrichissement en éléments transposables : Mu chez le maïs, TIR-Mariner chez le blé ; s*towaway* and *P instability factor (PIF)/Harbinger* chez le riz et la pomme de terre (Marand et al. 2019; 2017). Darrier et al. (2017) ont identifié chez le blé un motif de séquence (TCCCTCC, dérivé des séquences de la famille des transposons TIR-MARINER) préférentiellement associé aux intervalles fortement recombinants et ont pointé la similarité avec le motif ciblé par PRDM9 (lui aussi dérivé d'une séquence de transposons), suggérant un mécanisme de recombinaison universel. Cependant, l'association entre transposons et chromatine est peut-être issue d'une simple préférence commune pour les régions de l'ADN où la chromatine est décompactée.

Par ailleurs, certains résultats suggèrent que la position des points chauds pourrait résulter de processus sélectifs (Stapley et al. 2017). C'est le cas de Tock et al. (2021) qui rapportent que des clusters de gènes impliqués dans l'immunité sont associés à des points chauds de recombinaison chez le blé tendre. A l'inverse, la recombinaison est réduite à proximité de super-gènes, qui sont des séquences d'ADN présentant une série d'allèles co-adaptés. Par exemple, le centromère des chromosomes est généralement rapporté comme une grande région froide (Fernandes, et al. 2019). Dans le cas du blé, les régions péricentromériques et centromériques sont particulièrement étendues (80% des chromosomes), très peu recombinantes et contiennent pourtant 60% de gènes (Choulet et al. 2014). Le verrouillage de la recombinaison dans ces régions centromériques s'explique par une plus forte condensation de la chromatine, mais la cause de cette condensation reste encore inconnue. Cette condensation permet peut-être de désactiver les nombreux rétrotransposons présents dans cette région (Kent et al. 2017) ou encore de supprimer la recombinaison entre loci co-adaptés ou bien de protéger des gènes structuraux.

Enfin, les conditions pédoclimatiques influencent le profil de recombinaison, au moins à des échelles chromosomiques. Par exemple, chez l'orge, Higgins et al. (2012) montrent que les taux

de recombinaisons aux télomères diminuent et que ceux aux péricentromères augmentent lorsque la température dépasse 30°C.

Puisque la recombinaison dans les régions péricentromériques est un évènement rare, l'amélioration génétique sur ces morceaux de chromosomes dans le matériel de sélection requiert donc de produire beaucoup d'individus recombinants (Rodgers-Melnick et al. 2015). Mieux comprendre le déterminisme génétique de la recombinaison permettrait de développer des technologies pour déverrouiller ces régions riches en gènes.

Conclusion partielle du Chapitre I.3

Le seul mécanisme qui permet de recombiner les allèles parentaux favorables situés sur les mêmes chromosomes homologues est la recombinaison méiotique, via la formation de crossing-over entre chromosomes homologues. Le profil de recombinaison du blé est très hétérogène le long des chromosomes, avec des télomères beaucoup plus recombinants que les régions proximales des chromosomes. A fine échelle génomique, il est possible d'identifier des points chauds de recombinaison, c'est-à-dire des segments chromosomiques où se concentrent les crossing-over. Le déterminisme de la position des crossing-over reste globalement inconnu chez les plantes. La stabilité ou la variabilité du profil de recombinaison chez le blé tendre n'a pas encore été étudié sur l'ensemble du génome. Comparer les profils de recombinaisons à fine-échelle entre groupes génétiques apparentés de blé tendre permettrait de mieux comprendre le déterminisme, notamment s'il est sujet à évolution.

Chapitre II : Estimation et variabilité du profil de recombinaison chez le blé tendre

Chapitre II : Estimation et variabilité du profil de recombinaison chez le blé tendre

II.1 Préambule

Ce chapitre est présenté sous la forme d'un article scientifique publié dans le journal **Genome Biology Evolution** (https://doi.org/10.1093/gbe/evab152).

Cet article a pour but d'estimer les profils de recombinaison des quatre principaux groupes génétiques du blé tendre à partir de leurs patrons respectifs de déséquilibre de liaison et d'évaluer leur similarité entre groupes génétiques.

Dans la mesure où les profils de recombinaison dérivés des patrons de déséquilibre de liaison peuvent être biaisés par les forces évolutives, une comparaison est réalisée avec une carte méiotique obtenue sur une population de RILs dérivée du croisement entre les variétés Chinese Spring et Renan, cette carte méiotique étant supposée dépourvue de ces biais.

Deux aspects sont mesurés pour évaluer la similarité du profil : le partage de points chauds de recombinaison et la corrélation des profils le long du génome.

1) La finesse des profils de recombinaison dérivés des patrons de déséquilibre de liaison permet d'identifier des intervalles fortement recombinants, où le taux de recombinaison présente une augmentation significative par rapport à la région génomique avoisinante. Ces intervalles font quelques dizaines de kilobases, et sont donc supposés chevaucher un ou plusieurs points chauds de recombinaison. Les positions de ces intervalles fortement recombinants sont comparées entre groupes génétiques. Afin d'évaluer si la co-localisation de ces intervalles entre groupes génétiques pourrait être due au hasard (préférence commune pour les régions chromosomiques peu condensées comme les télomères), un test statistique a été mis en place pour évaluer la distribution des taux de chevauchement attendus par hasard. Dans la mesure où la détection des intervalles recombinants implique de définir un seuil arbitraire, nous avons aussi regardé l'augmentation locale de la recombinaison au niveau de chacun de ces intervalles dans chacune des populations.

2) Les profils de recombinaison sont aussi corrélés le long de chacune des grandes régions chromosomiques du blé (télomères R1 et R3, péri-centromères R2a et R2b, centromères exclus). Pour limiter l'impact des forces évolutives sur la variation du taux de recombinaison entre deux intervalles, nous avons corrélé l'inflation ou la déflation locale du taux de recombinaison de chaque intervalle par rapport au taux de recombinaison de base d'une fenêtre chromosomique de 2 cM chevauchant les intervalles. Notre hypothèse est que les forces évolutives n'influencent pas ces déviations locales du taux de recombinaison dans des intervalles de petite taille.

GBE

Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat

Alice Danguy des Déserts[1], Sophie Bouchet[1], Pierre Sourdille [1,*], and Bertrand Servin [2,*]

[1]INRAE-Université Clermont-Auvergne, UMR1095, Génétique Diversité Ecophysiologie des Céréales, Clermont-Ferrand, France
[2]2INRAE, Université de Toulouse, GenPhySE, Castanet-Tolosan, France

*Corresponding authors: E-mails: bertrand.servin@inrae.fr; pierre.sourdille@inrae.fr.

Accepted: 24 June 2021

Abstract

Reciprocal exchanges of DNA (crossovers) that occur during meiosis are mandatory to ensure the production of fertile gametes in sexually reproducing species. They also contribute to shuffle parental alleles into new combinations thereby fueling genetic variation and evolution. However, due to biological constraints, the recombination landscape is highly heterogeneous along the genome which limits the range of allelic combinations and the adaptability of populations. An approach to better understand the constraints on the recombination process is to study how it evolved in the past. In this work, we tackled this question by constructing recombination profiles in four diverging bread wheat (*Triticum aestivum* L.) populations established from 371 landraces genotyped at 200,062 SNPs. We used linkage disequilibrium (LD) patterns to estimate in each population the past distribution of recombination along the genome and characterize its fine-scale heterogeneity. At the megabase scale, recombination rates derived from LD patterns were consistent with family-based estimates obtained from a population of 406 recombinant inbred lines. Among the four populations, recombination landscapes were positively correlated between each other and shared a statistically significant proportion of highly recombinant intervals. However, this comparison also highlighted that the similarity in recombination landscapes between populations was significantly decreasing with their genetic differentiation in most regions of the genome. This observation was found to be robust to SNPs ascertainment and demography and suggests a relatively rapid evolution of factors determining the fine-scale localization of recombination in bread wheat.

Key words: recombination, evolution, bread wheat.

Significance

Recombination is the fundamental biological process that shuffles chromosomes during meiosis to create new allelic combinations in gametes. It has been shown to be controlled by genetic factors in some species but those remain unknown in many. Understanding the genetic determinism of recombination can help to better describe its underlying biological make up, its constraints and understand its evolvability. One approach to study this question is to evaluate if the recombination process differs between groups of individuals that are genetically distinct (populations). Here we propose methods to implement this approach and investigate the recombination process in bread wheat, one of the most widespread crops. We show that recombination patterns between two populations are increasingly correlated when their genetic differentiation decreases. This suggests that recombination in bread wheat can evolve rapidly possibly associated to an underlying modification of its genetic determinism.

Introduction

Meiotic recombination (or crossover; CO) is the obligate genetic exchange between homologous chromosomes that occurs during the production of gametes in sexually reproducing species. Besides its role in ensuring proper segregation of chromosomes in gametes, it also impacts evolution by breaking linkage between advantageous and deleterious alleles and by creating novel combinations of alleles (Barton

1995; Charlesworth and Barton 1996; Otto 2009). Recombination rates are highly variable between species and also at different genomic scales. At the chromosomal level, COs are not evenly distributed depending on either the size of the chromosomes, the region of the chromosomes or on interference. Interference was first observed in *Drosophila* (for review, see Berchowitz and Copenhaver 2010) and is defined as the impossibility for a type I CO (i.e., COs that are submitted to interference contrary to type II COs that are not) to occur in the vicinity of another CO from the same type. Type I COs are thus more regularly spaced along chromosomes than expected from random (Zickler and Kleckner 2015). Within chromosomes, some regions are also deprived of COs, such as centromeres of all species studied so far. Moreover, in many species, distribution of COs is skewed toward telomeres (Haenel et al. 2018). In wheat (*Triticum aestivum* L.), for example, more than 80% of the recombination events occur in the terminal regions of the chromosomes representing less than 20% of the genome (Saintenac et al. 2009; Choulet et al. 2014; Darrier et al. 2017; International Wheat Genome Sequencing Consortium IWGSC 2018). The main hypothesis is that the initiation of synapsis responsible for recombination occur in the telomeric regions as shown in barley (Higgins et al. 2012; Dreissig et al. 2019). In species with small chromosomes such as *Arabidopsis thaliana* or rice (*Oryza sativa*), recombination events are more evenly distributed along the chromosomes with the exception of the centromeres (Choi et al. 2013; Drouaud et al. 2013; Marand et al. 2019). In all studied species, the number of COs per chromosome and per meiosis is rarely superior to three (Mercier et al. 2015).

At a local scale, in most species including yeast, birds, snakes, fishes, mammals, and plants, COs mainly occur in small regions of a few kilobases (kb) called hotspots (Myers et al. 2005; Mancera et al. 2008; Choi and Henderson 2015; Singhal et al. 2015; Shanfelter et al. 2019; Schield et al. 2020). In some mammals, these hotspots are determined by PRDM9, an SET-domain protein with a zinc-finger array that binds DNA (Boulton et al. 1997; Oliver et al. 2009; Baudat et al. 2010; Myers et al. 2010). PRDM9 recognizes specific DNA motifs and deposits an epigenetic landmark (histone H3 trimethylated on lysine 4: H3K4me3) that is further recognized by the machinery forming double-strand breaks that initiates COs (Murakami et al. 2020). However, many if not most species (e.g., birds, plants, yeast, snakes, and fishes) do not exhibit a PRDM9 derived mechanism. Recombination hotspots are often found in accessible chromatin regions and mainly driven by chromatin features (Auton et al. 2013; Choi and Henderson 2015; Singhal et al. 2015; Marand et al. 2017, 2019) although intermediate situations exist (Schield et al. 2020).

The determinism of local recombination rate considering the distribution of CO hotspots remains unknown for many organisms. One approach to better understand this determinism is to characterize the evolution of the recombination landscape and evidence its conservation or lack thereof. This can be achieved by contrasting recombination landscapes in closely related species (Stapley et al. 2017) or in differentiated populations of the same species (Kong et al. 2010; Salomé et al. 2012; Petit et al. 2017). For example, in rice, less than 20% of the CO hotspots are common between the two subspecies *Oryza sativa ssp. japonica* and *O. s. ssp. indica* (Marand et al. 2019) although they diverged relatively recently [440,000—86,000 years ago (YA); Ma and Bennetzen 2004; Vitte et al. 2004; Zhu and Ge 2005; Tang et al. 2006]. Similarly, in the cocoa-tree (*Theobroma cacao*), only little overlap of recombination hotspots was observed across ten diverging populations, with less divergent populations showing higher level of overlap (Schwarzkopf et al. 2020). Note that recombinations tend to cluster in more distal regions in domesticated barley (*H. vulgare*) compared with wild barley (*Hordeum vulgare ssp. spontaneum*) (Dreissig et al. 2019) while domestication began approximately 10,000 YA (Badr et al. 2000). A finer-scale analysis among subpopulations of wild barley revealed that recombination rate varied according to environmental conditions (temperature, aridity, solar radiation, annual precipitations), suggesting that environmental factors might explain part of these differences (Dreissig et al. 2019).

High-density genotyping SNP arrays as well as new generation sequencing (NGS) approaches now allow to analyze large collections of wild/domesticated, ancient/modern populations of both animals and plants. Such a large amount of accurate data permits to better decipher the recombination landscape from patterns of linkage disequilibrium (LD) (Li and Stephens 2003; Auton and McVean 2007; Chan et al. 2012). The advantages of using this approach stem from the large number of meiosis that occurred during the evolution of sampled populations compared with bi-parental or multi-parental experimental populations. First, as LD-based recombination inference is based on recombination happening in many different individuals it should consequently be less sensitive to individual specific variation, which might occur in the presence of structural variation (e.g., Bauer et al. 2013; Rowan et al. 2019). Second, LD-based recombination rate estimates are more resolutive as genetic diversity is higher compared with experimental segregating populations that typically involve few parents. However, the drawback of this approach is that the recombination landscapes obtained have to be interpreted cautiously as they can be affected by evolutionary forces such as selection and demography that can also impact local patterns of LD (Charlesworth and Charlesworth 2010; Auton and McVean 2012; Choi and Henderson 2015).

Despite these limitations, the LD-based approach was successfully applied at the whole-genome level in many species including birds (Singhal et al. 2015; Smeds et al. 2016), yeast (Tsai et al. 2010), Arabidopsis (Choi et al. 2013), rice (Marand et al. 2019), and barley (Dreissig et al. 2019). In bread wheat this approach was used to study recombination pattern on chromosome 3B (Darrier et al. 2017), the only chromosome presenting a sufficiently high-standard reference sequence at that time

(Choulet et al. 2014; IWGSC 2014). The analysis of two collections representative of the Asian and European genetic pools revealed a high similarity between their recombination profiles. These LD-based profiles were also shown to be consistent with a meiotic recombination profile derived from a bi-parental population (Chinese Spring × Renan; Choulet et al. 2014). This result suggested that recombination rate estimation through a LD-based approach could be even more informative and resolutive along the whole genome using the last gold-standard reference sequence available (IWGSC 2018), as well as high-density genotyping of large wheat collections.

The complexity and huge size (16 gigabases) of the wheat genome have long hampered the development of high throughput genomic tools as well as the establishment of a whole-genome sequence. Bread wheat is an allo-hexaploid species (AABBDD; $2n = 6x = 42$) derived from two successive interspecific crosses involving three diploid species (for details, see https://www.wheatgenome.org/; IWGSC 2014, 2018): *T. monococcum ssp. urartu* (AA genome), a yet-unknown species related to the *Sitopsis* section (SS genome related to the wheat BB genome) and *Aegilops tauschii* (DD genome). However, international efforts combined with appropriate and original strategies using chromosome sorting, chromosome-specific BAC libraries, paired-end short-read sequencing and relevant assembly approaches, led to the publication of a high-standard, annotated, oriented and anchored sequence of the wheat genome (IWGSC 2018). At the same time and despite the presence of a high proportion of transposable elements (85%; Wicker et al. 2018), high-density SNP arrays have been successfully developed and used for marker-assisted selection (Sun et al. 2020) and for the characterization of collections (Winfield et al. 2016; Balfourier et al. 2019). In the study of Balfourier et al. (2019), the genetic structuration of 4,506 bread wheat landraces and cultivars representative of the worldwide diversity was described using the TaBW280K SNP chip. These LD data offer the opportunity to extend previous work on bread wheat by analyzing recombination along the whole genome and across more populations. We compared the ancestral recombination profiles of four populations with the meiotic recombination observed in a biparental population of recombinant inbred lines (RILs; Chinese Spring × Renan; CsRe). We developed specific statistical models to evaluate and minimize the influence of evolutionary forces on the comparison of recombination landscapes between populations.

Results

Bread Wheat Landraces Are Structured in Four Main Populations

Establishing LD-based recombination maps requires samples of unrelated chromosomes from a homogeneous population. We extracted a subset of 371 landraces representative of the worldwide diversity from Balfourier et al. (2019), forming four distinct and mostly homogeneous genetic populations (see Materials and Methods; fig. 1) that were named according to the geographical origins of their members: The West-European population (WE), composed of 127 accessions originating from France (52 accessions), Spain (10), Germany (8) and from 30 other Western European, Mediterranean countries and Iberian peninsula; the East-European population (EE), composed of 70 accessions originating from France (9), the Russian Federation (7), Ukraine (5) and from 27 other Eastern European countries; the West-Asian population (WA), composed of 97 accessions originating from Afghanistan (8), Pakistan (8), Turkey (8) and from 33 other of Caucasian and Central Asia countries and Indian peninsula; the East-Asian population (EA) composed of 77 accessions originating from China (61), Japan (7), the Republic of Korea (4) and from five other South East Asian countries (supplementary file S1, Supplementary Material online).

The genetic differentiation of the four populations confirmed an increasing genetic divergence along an Eurasian gradient (fig. 1), consistent with isolation by distance, selection, and differentiation that occurred during the initial independent spreads of bread wheat from the Cradle of Agriculture and Wheat in the Fertile Crescent toward Europe on the one hand and Asia on the other hand during the Neolithic period (Balfourier et al. 2019). WE and EE are the most related groups ($F_{ST} = 0.015$), whereas WE and EA are the more divergent ones ($F_{ST} = 0.085$) and also the most geographically distant. The WA population is the closest population to the tree root possibly because it includes accessions that were collected not far from the center of domestication of bread Wheat (Fertile Crescent: Turkey, Iraq, Iran; Caucasus and Caspian Sea: Armenia, Georgia, Kazakhstan, Turkmenistan). The EA population appears as a very differentiated and homogenous population. WE and EE are less differentiated because they separated more recently from each other (Balfourier et al. 2019).

The genetic composition of the four populations appeared quite distinct between populations but homogenous within populations when described by the $K = 4$ admixture analysis of Balfourier et al. (2019) (fig. 1). WE, EE, WA, and EA have almost all their members belonging to the same specific dominant group (respectively, named by Balfourier et al. (2019) as North West European, South East European, Central Asian and African and South East Asian groups) with a high membership coefficient: 0.74 on average for WE (standard deviation = 0.16), 0.81 for EE (0.16), 0.73 for WA (0.17), and 0.93 for EA (0.14). The WE and WA populations appear to be more admixed than EE and EA at $K = 8$ (supplementary fig. S1, Supplementary Material online). In order to analyze groups that are large enough to estimate relevant statistics, we split landraces into four populations, although there is some sub-structuration within populations. This was motivated by the fact that the model we used to estimate LD-

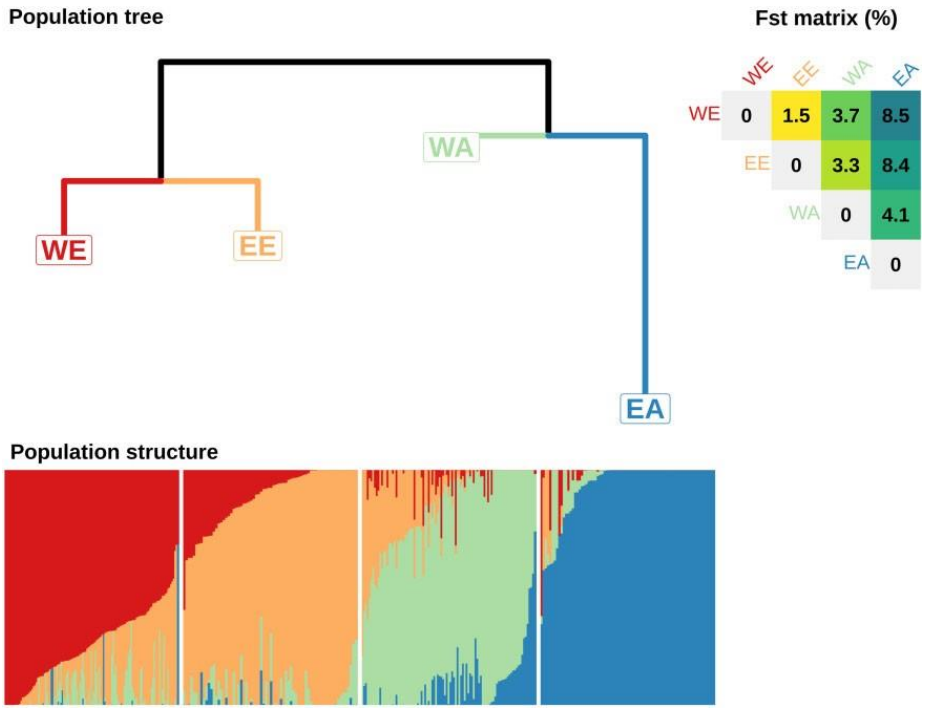


FIG. 1.—Bread wheat landrace genetic divergence and structuration. Population tree: Neighbor Joining tree built with pairwise Reynold distance matrix computed on SNP alleles and rooted by HAPFLK software (Bonhomme et al. 2010; Fariello et al. 2013). WE, West Europe; EE, East Europe; WA, West Asia; EA, East Asia. Fst matrix (%) Weir and Cockerham pairwise F_{ST} computed with simple matching distance of haplotypic alleles. Population structure: Admixture coefficients for $K = 4$ from Balfourier et al. (2019) using STRUCTURE software and haplotypic alleles.

based recombination rates was shown to be robust to moderate levels of structuration (Li and Stephens 2003).

Recombination Patterns Are Broadly Conserved across Populations

Robust Meiotic Recombination Map of a Population of RILs

In order to obtain a view of recombination patterns that is not influenced by evolutionary forces, we established a meiotic recombination map from recombination events observed in a population of 406 F6 RILs (termed CsRe in the following). This population is derived from a cross between two bread wheat varieties: Chinese Spring and Renan belonging, respectively, to the EA and WE gene pools. The CsRe population was previously genotyped for the same set of SNPs as the landraces (Rimbert et al. 2018). Recombination rates in CsRe were derived from the observed proportion of recombinants in each of the 79,543 intervals defined by SNPs that were polymorphic in the cross. The distribution of recombinants in these intervals led to extremely contrasted situations. On one hand, 60% of these intervals harbored no recombinant among the 406 offspring. On the other hand, a few recombinants were observed in very small intervals. Using a frequentist statistical approach to estimate recombination rates from these observations produces extreme differences in recombination rates that are highly influenced by the limited sample size available. In order to produce more reliable estimates that better account for sample size and uncertainty, we fitted a Bayesian Poisson Gamma model on the observed recombinant counts (see Materials and Methods). With this model, the estimates of recombination rates in the RIL population ranged from almost 0 to 78 cM/Mb among intervals. Compared with the frequentist estimates that ranged up to 2,806 cM/Mb this approach has the advantage of shrinking extreme values that are unrealistic and solely due to the limited number of RILs available. Consistent with the Bayesian model correcting for the effect of sample size, the correlation between frequentist and Bayesian estimates increases with the number of observed recombinants per intervals (supplementary fig. S2, Supplementary Material online), that is, the two approaches converge to the same inference when the data is informative enough.

Validation of LD-Based Recombination Maps on CsRe Meiotic Recombination Map

LD-based recombination maps were inferred from patterns of LD between polymorphic SNPs for each landrace population independently using PHASE (Li and Stephens 2003; Crawford et al. 2004). As LD is strongly related to meiotic recombination but can also result from evolutionary forces, those maps were

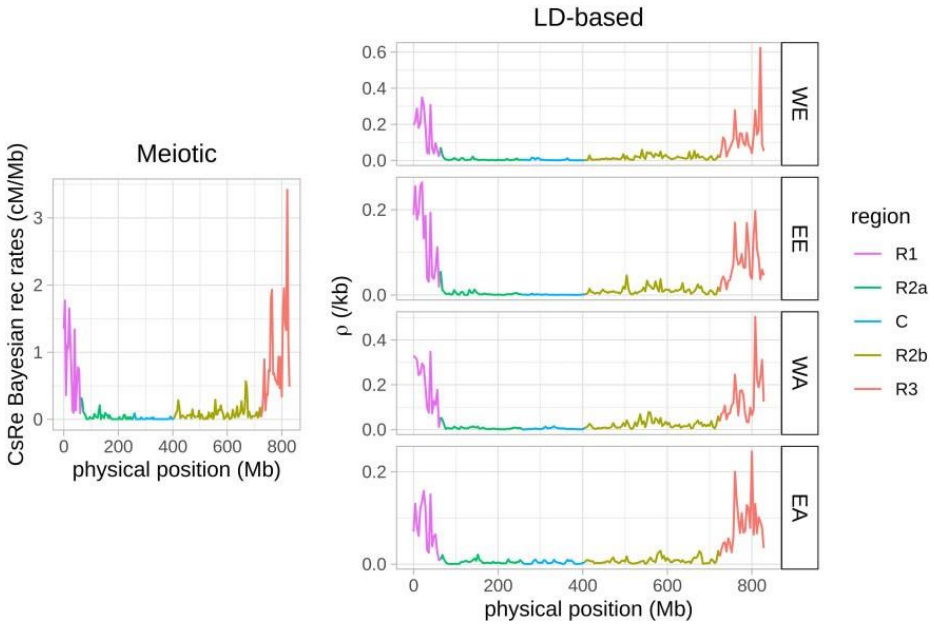


FIG. 2.—Meiotic and LD-based recombination profiles in 4 Mb windows along chromosome 3B in the CsRe segregating population (left) and in the four West European (WE), East European (EE), West Asian (WA), and East Asian (EA) populations (right). Each color corresponds to genomic regions defined by Choulet et al. (2014): highly recombining telomeres R1 (magenta) and R3 (red); low recombining pericentromeres R2a (dark green) and R2b (light green); and centromere C (blue) where recombination rates are close to 0. LD-based recombination profiles at $\log_{10}$ scale are present in supplementary figure S4, Supplementary Material online.

compared with the meiotic CsRe recombination map described above.

Before estimating LD-based recombination rates, SNPs were filtered out on Minor Allele Frequency with a minimum value of 3% within each population, yielding to 170,509 SNPs for WE, 161,137 for EE, 171,901 for WA, and 131,585 for EA. The average marker density was 11 SNPs/Mb with most of the SNPs located at telomeres (25 SNPs/Mb) whereas centromeres were depleted in SNPs (3 SNPs/Mb, supplementary fig. S3, Supplementary Material online). SNP density was almost three times higher on the A and B genomes compared with the D genome (respectively, 14, 14, and 5 SNPs/Mb). This is consistent with the lower rate of polymorphism of the wheat D genome (IWGSC 2018).

Both LD-based and meiotic recombination profiles showed the same global patterns at the chromosome scale (fig. 2; supplementary file S2, Supplementary Material online). In both approaches, the telomeric regions R1 and R3 of chromosomes showed recombination rates (average LD-based recombination rate in WE = 1e-2/kb; average CsRe Bayesian recombination rate = 0.8 cM/Mb) around ten times higher than the pericentromeric regions R2a and R2b (2e-3/kb; 0.1cM/Mb) and one hundred times higher than the centromeric regions C (2e-4/kb; 0.01 cM/Mb). Recombination rates on the D genome (5e-3/kb; 0.3 cM/Mb) were around 25% higher than recombination rates in the A and B genomes (both 4e-3/kb; 0.2 cM/Mb). The chromosomes from the D-genome are 20% shorter than those from the A or B genomes (IWGSC 2018) while they receive the same number of crossovers (supplementary fig. S5, Supplementary Material online), leading to high global recombination rates. IWGSC (2018) study also showed that the D-genome was twice-less polymorphic than the A or B genomes (18%, 40%, and 41% for the D, A, and B genomes, respectively; IWGSC 2018). It has been demonstrated in maize, sorghum and Arabidopsis that recombination rates are higher in chromosome regions showing higher similarity because a lower genetic diversity facilitates homologous pairing and recombination during meiosis (Rodgers-Melnick et al. 2015; Bouchet et al. 2017; Serra et al. 2018). We can therefore speculate that the high recombination rates we observe on the D-chromosomes are due to their reduced physical size associated with a low diversity favoring recombination.

The genome-wide correlation of LD-based recombination profiles and CsRe Bayesian meiotic recombination profile was quite high for the four populations (≤0.7, table 1) but slightly higher for European populations [pairwise significant differences according Zou's test (Zou 2007), R cocor package]. These high correlations between CsRe meiotic recombination profile and LD-based recombination profiles are explained by the strong partitioning of the recombination profile along chromosomes present in all bread wheat populations, that is, low recombination rates in centromeres and high recombination rates in telomeres. As computing correlation

Table 1
Correlation of the LD-Based Recombination Profiles of the Four Populations of Landraces with CsRe Bayesian Meiotic Recombination Profile

	WE	EE	WA	EA
Genome-wide corr. with CsRe	0.76	0.75	0.74	0.70
Average on 84 genomic regions (R1, R2a, R2b, R3 of chr 1A-7D)	0.58±0.22	0.55±0.28	0.55±0.27	0.50±0.29
Average on 21 C regions (chr 1A-7D)	0.32±0.33	0.30±0.34	0.20±0.34	0.19±0.36

N_{OTE}.-Recombination rates were averaged in 4-Mb windows.

coefficient using whole-genome recombination profile artificially inflates the value of correlation, we rather performed correlation within each genomic region. The within-region correlation coefficients were lower, but still significantly positive (1AR1-7DR3, fig. 3; supplementary file S3, Supplementary Material online). In telomeres R1 and R3 and pericentromeres R2a and R2b, the average correlation ranged between 0.50 in EA and 0.58 in WE (table 1), with an average of 0.56 across all populations.

The recombination rates in centromeric regions showed much lower consistency: The correlation of centromeric LD-based recombination rates and CsRe recombination rates ranged from 0.19 in EA to 0.32 in WE. Considering the low correlation but also the low SNP density and the fact that centromere sequence assemblies are challenging because of the presence of numerous repeated sequences such as transposons and retro-transposons (IWGSC 2018; Wicker et al. 2018), centromeres were no longer included in the analyses.

Among the genomic regions considered, 7DR3 exhibited a strikingly low and negative correlation between LD-based and meiotic recombination rates in all populations (≤-0.19, fig. 3). This result is due to a low recombination rate in part of this region in the CsRe biparental genetic map that is not observed in LD-based maps (supplementary fig. S6, Supplementary Material online). This low recombination rate can be explained by the fact that Renan (one parent of the CsRe biparental population) carries an inter-specific introgression of 28 Mb on chromosome 7D around the eyespot resistance gene *Pch1* coming from *Aegilops ventricosa* (tetraploid species; DDNN) (Maia 1967). This introgression does not recombine in the CsRe cross as this was previously evidenced in another background (Worland et al. 1988). Interestingly the Renan line carries another 20 Mb introgression from *Aegilops ventricosa* in 2AR1 region around the *Lr37/Sr38/Yr17* resistance gene cluster. However, in this region, contrary to 7DR3, the LD patterns are also consistent with a locally low recombining segment in landraces at position of introgression. Because the introgression in region 2AR1 suppresses recombination in an already low recombining segment, this explains why the correlation coefficient with LD-based profiles does not stand out particularly (supplementary fig. S6, Supplementary Material online).

Both CsRe and LD-based maps show a high heterogeneity in the distribution of recombination rates along chromosomes: On average 36% (±1%) of physical distance represents 80% of genetic distance in all our populations.

To further study the distribution of chromosome sites cumulating historical crossovers, we defined highly recombining intervals (HRIs) in the four landrace populations as intervals with an LD- based recombination rate exceeding four-times the back- ground recombination rate ($\lambda \geq 4$, see Materials and Methods). Combining all four populations, this resulted in 8,713 HRIs, with a median deviation to background recombination rate $\lambda = 6.5$ (range: $\lambda = 4$ to $\lambda = 511$). Note that we avoid here the term LD- based recombination *hotspot* as functional hotspots typically span much smaller genomic regions (size < 5 kb; Marand et al. 2019) than our defined HRIs (median size = 20 kb). Therefore, we cannot be sure that an HRI harbors a single recombination hotspot. The repartition of HRIs along the genome was heterogeneous. Most HRIs (73%) were located in telomeric R1 or R3 regions, and the other HRIs (27%) in pericentromeric R2a or R2b regions. As HRIs corresponded to, respectively, 2% and 1% of intervals in those regions, telomeres were significantly enriched in HRIs compared with pericentromeres (significant chi-square test, *P*-value < 2.2e-16). These HRIs represented 15% of LD-based genetic distance (from 12% in EA to 18% in WA) and around 9% of the physical distance (from 6% in EA to 10% in WE). On average, in all genomic regions, the 8,713 HRIs tend to highly co-localize with open-chromatin features compared with non-HRIs intervals. For example, the proportion of HRIs overlapping genes was 80%, but this proportion dramatically decreased to 53% when considering non-HRIs intervals (supplementary fig. S7, Supplementary Material online). The density of HRIs is also positively associated with the CsRe meiotic recombination rate averaged in 4 Mb windows in each genomic region R1, R2a, R2b, and R3 (*P*-value < 2.2e-16). The proportion of CsRe crossovers overlapping HRIs ranged from 20% in EA to 37% in WE. Most HRIs (82%) overlapped at least one CsRe crossover.

Despite high similarities between LD-based and meiotic recombination profiles within genomic region, there is still the possibility that LD-based recombination rates might be locally influenced by evolutionary forces, such as positive selection, as shown by Petit et al. (2017) in sheep for example. To evaluate the potential effects of positive selection on the LD-based maps, we studied whether a set of genes known to be involved in domestication [e.g., brittle rachis (*Brt*), tenacious glume (*Tg*), homoeologous pairing (*Ph*), or nonfree-threshing character (*Q*)] or recent crop improvement (Pont et al. 2019) were found in regions outliers for the ρ/CsRe

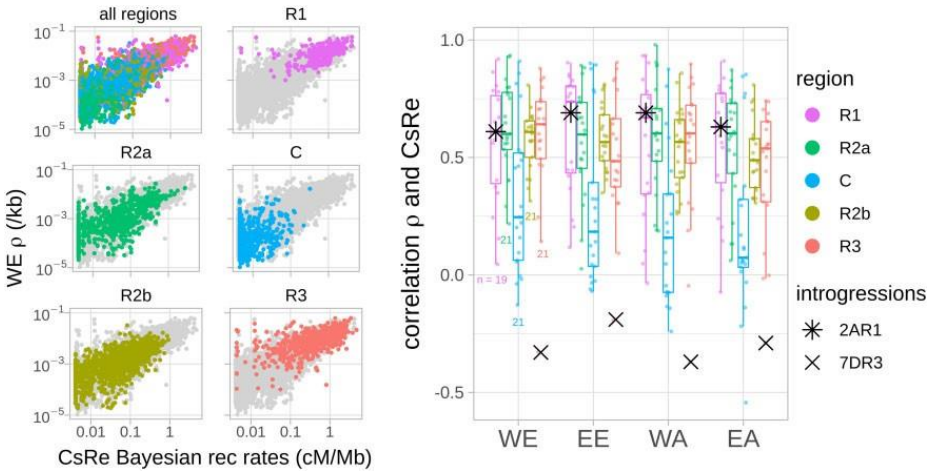


FIG. 3.-Similarity between LD-based recombination rates and CsRe meiotic recombination rates. Left: Genome-wide relationship between the CsRe biparental population meiotic recombination rates and the LD-based recombination profile of a Western European (WE) bread wheat population. Dots represent the recombination rates averaged within 4 Mb windows. Graphs R1, R2a, C, R2b, and R3 gather recombination rates within the five chromosomic regions defined by Choulet et al. (2014) (R1 and R3 are telomeric regions, R2a and R2b are pericentromeric regions, and C are centromeric regions) of all of the 21 chromosomes (1A, 1B, 1D… 7A, 7B, 7D) of bread wheat. *Right*: Correlation of LD-based and CsRe recombination rates for each landrace population within each genomic region (1AR1…7DR3). Dots represent correlation coefficients of recombination profiles (once averaged within 4 Mb windows) per genomic region and population. Small colored numbers indicate the number of correlation coefficients per boxplot. In principle, each boxplot should contain 21 dots (as many as chromosomes). However, two R1 genomic regions smaller than 20 Mb are not included (4DR1 and 7BR1), because of low robustness of their correlation coefficients (computed on less than five data points). Stars (* and x) represent genomic regions including well documented introgressions in CsRe population.

ratio. The results showed no evidence of reduced recombination around these genes (supplementary fig. S8, Supplementary Material online). Although this does not rule out potential effects on other genes or through other selection pressures (e.g., background selection), it indicates that strong selective sweeps do not seem to affect recombination inference and justify converting LD-based maps on the meiotic recombination scale (cM/Mb). Considering that LD-based recombination rates are proportional to meiotic ones, they can be rescaled by computing the scaling factor from the CsRe Bayesian average recombination rate in each genomic region (supplementary protocol S1, Supplementary Material online). This produced scaled LD-based maps specific to each landrace population (supplementary file S4, Supplementary Material online).

Significant Differences between LD-Based Population-Specific Recombination Maps

Our results reveal that the average LD-based recombination rates vary in a 2-fold range between populations: WE has the highest rate and EA the lowest (WE: ρ = 0.004/kb; WA: ρ = 0.004/kb; EE: ρ = 0.003/kb; EA: ρ = 0.002/kb; excluding centromeres). This ranking between populations could be explained by genetic diversity levels (fig. 1) as well as by different average meiotic recombination rates. The fact that WE and WA are more admixed populations than EE and EA favored a more important contribution of diversity levels compared with a real difference on average recombination intensity. To eliminate the systematic effect of diversity and demography on recombination rate estimates, we chose to compare the population recombination profiles in terms of the deviation from their local background recombination rates. Specifically, the Li and Stephens's model (2003) estimates an interval specific recombination parameter (λ) that measures the relative rate of recombination of an interval compared with its neighbors in a 2 cM window (see Materials and Methods). We therefore expect population-specific effects (other than local variation in recombination) to affect the background recombination rate but not the relative intensities of intervals measured by the parameter λ.

The similarity of λ profiles along the genome was evaluated by fitting a linear mixed model on the variations of $log_{10}(\lambda)$ within each genomic region, specifying a variance-covariance matrix with different or common correlation coefficients for each pair of populations. In almost all genomic regions (79 out of 84), a lowest BIC was obtained for the model with correlation coefficients that are different between pairs of population (see Materials and Methods). This indicates that local variations of recombination rates are significantly different between populations.

The average correlation of local variations of recombination rates across genomic regions was twice higher for the highest correlated pair WE-EE (0.47 6 0.11) than for the lowest one EE-EA (0.20 6 0.11), with an average value of 0.32 (fig. 4). The unevenness of the distribution of genetic distance along

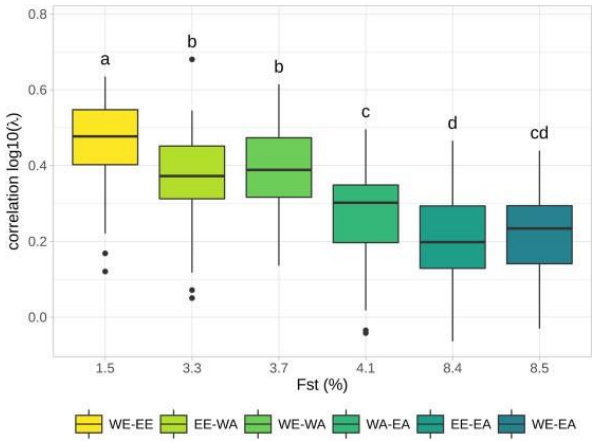


FIG. 4.—Relationship between pairwise correlation of LD-based recombination intensity λ and F_{ST}. Each boxplot contains 84 correlation coefficients corresponding to the 84 genomic regions (1AR1…7DR3, excluding centromeres). Letters indicate whether two pairs have significant different average correlations (Bonferroni corrected *P*-value < 0.05).

chromosomes between two different populations was measured using a Gini coefficient (Gini 1936). We compared the distribution of recombination in one population with the genetic map of the other. A Gini coefficient of 0 corresponds to a uniform distribution and a coefficient of 1 corresponds to the case where the distribution is a single point mass. In our case, a Gini coefficient of 0 corresponds to identical recombination profiles and the more divergent the distribution in recombination profiles is, the higher Gini coefficient is. The pairwise Gini coefficients increased along the Eurasian gradient, with lower values for closely related population (around 0.43 for WE-EE) and higher values in distant populations (0.77 for WE-EA), meaning that similarity in distribution of LD-based genetic distance along chromosomes decreases along the Eurasian gradient (supplementary fig. S9, Supplementary Material online).

In light of these significant differences in the local repartition of recombination events, we investigated whether this could be explained by difference in the localization of crossover hotspots by comparing that of the HRIs (see above). We first defined "hot windows" as genomic regions that harbor an HRI in at least one population. Figure 5*A* represents the proportion of the 5,881 resulting hot windows including HRIs that are population specific (HRI in one population only) or shared by two, three or all four populations. Around 66% of these windows are population-specific and 34% are shared by two populations or more. The proportion of hot windows shared by three or four population drops to 12% and 2%, respectively. Location of shared HRIs along the genome followed the density of HRIs per genomic region. Most (76%) shared windows were located in telomeric regions R1 and R3 and the rest (24%) in pericentromeric regions R2a and R2b (chi-square test *P*-value = 0.06). To check if such an overlap across populations can be explained by chance alone, we compared the observed repartition of hot windows with a simulated distribution obtained by a random assignment of HRIs corresponding to the null hypothesis of the absence of HRI population sharing (see Materials and Methods). The proportion of common hot windows under this random assignment is represented by gray boxplots in figure 5*A*. The observed proportion (colored points) was always significantly different to the expected proportion under random assignment of HRIs. On average, 95% of hot windows are population-specific if assigned randomly, much more than the 66% we observed. In addition, four-population overlaps were rare in the simulations (8.1% of our simulations) and when they happened, they concerned only one or two windows whereas we found 139 windows where HRIs are shared between the four landrace populations. HRIs shared by more populations tend to be more intense. For example, 55% of WE HRIs ($\lambda \leq 4$) colocalize with HRIs of other populations ($\lambda \leq 4$), but this proportion rises to 78% when subsampling WE HRIs with a higher threshold of $\lambda \leq 20$. The intensity of recombination in a hot window increases when it is shared by more populations: The median of λ is 10.7, 8.1, and 6.9 when shared by 4, 3, and 2 populations, respectively, and is only 5.9 for population-specific hot windows. This approach to compare HRIs between populations depends on the threshold to claim HRIs and our ability to detect them, which can vary between populations. To make up for these effects, we looked at the recombination intensity (λ) observed in one population around HRIs detected in another population (supplementary file S5, Supplementary Material online). Figure 5*B* presents this average recombination intensity for HRIs detected in each of the four populations. It shows that the local intensity at an HRI position in the other populations is almost twice the background intensity defined as the intensity measured at 100 kb from the HRI center (average λ at HRI positions: 29%; average background λ: 13%). This further shows that HRIs tend to be shared across populations. We evaluated whether this sharing could be explained by assembly errors that would lead to inflated recombination rates in all populations. Indeed, we found that 13% of hot windows shared by the four populations were associated with scaffold boundaries, which is a higher probability than expected by chance (odds ratio = 8.1, *P*-value < 2e-16). In addition, the probability for a hot window to be associated with scaffold boundaries decreased with the level of sharing (odds ratio ranges from 1.3 for population-specific hot windows, 2.2 for hot windows shared by only two populations, 3.7 when shared only by 3 populations and 8.1 for hot windows shared by 4 populations, *P*-values < 0.03). However, these enrichments are not sufficient to explain the patterns of sharing described above. Hence, a significant amount of sharing of HRIs could be due to an underlying partial conservation of recombination hotspots.

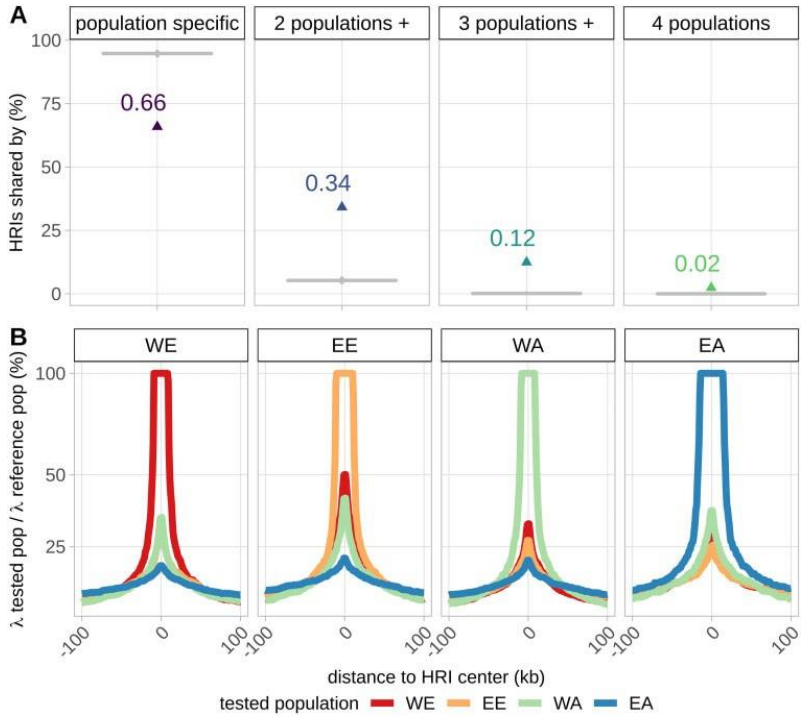


FIG. 5.-Conservation of highly recombining intervals (HRIs) across landrace populations. (*A*) Proportion of colocalizing HR (colored points) and simulated colocalizing values under random assignment of HRIs (gray boxplots). (*B*) LD-based recombination intensity in each of the four populations WE, EE, WA, and EA around HRIs specific to one population.

Further examination of the increase in recombination intensity in figure 5*B* reveals that HRI intensities tend to be more similar when populations are more related. For example, around WE HRIs, the recombination intensity increases in all populations, but slightly less in EA which is the most genetically distant population to WE. To study this further, we studied quantitatively the relationships between the similarity in recombination profiles and the genetic divergence of populations. To do so, we fitted a linear regression to estimate the effect of the local differentiation index (F_{ST}) on the similarity of recombination profiles (measured by their correlation) for all genomic regions (R1, R2a, R2b, and R3) on all chromosomes (1A to 7D) (fig. 6). We found that most F_{ST} effects (slopes) were negative, revealing a striking pattern where the similarity in recombination intensity decreases proportionally with genetic divergence: Almost all genomic regions (67 among 84) had a negative slope estimate significantly different from 0 and others genomic regions (15 among 84) had negative but nonsignificant slope estimates different from 0. Note that the similarity in recombination profiles is based on the relative local recombination intensity (parameter *k*) that should not be affected by the evolutionary history of populations. F_{ST} were calculated from haplotypes rather than single SNPs to avoid an ascertainment effect. But results based on F_{ST} calculated from SNPs showed the same pattern (supplementary fig. S10, Supplementary Material online). To further evaluate if the decreasing similarity of recombination patterns could be explained by the varying proportion of shared polymorphisms between population pairs, that is, SNPs ascertainment, we carried out all our analyses on a subset of 100,381 SNPs that are polymorphic in all four populations. We found that the decreasing similarity of recombination intensities with genetic divergence still hold using this common SNP data set (supplementary fig. S11, Supplementary Material online), even if the absolute values of slope estimates were smaller (supplementary fig. S12, Supplementary Material online). We also found no effect of prior distribution parameters in PHASE and sample size on inferences of recombination profile intensity (supplementary protocol S2 and figs. S13—S15, Supplementary Material online). Finally, these results demonstrate that the similarity in recombination profiles of bread wheat populations is strongly negatively associated with their genetic divergence and highlight that recombination landscapes in bread wheat have been evolving during the establishment of the current genetic structure of wheat populations.

To test whether meiotic genes could be associated to the divergence in the recombination profiles of populations, we assessed if pairwise genetic differentiation between populations at these genes (measured by F_{ST}) was particularly

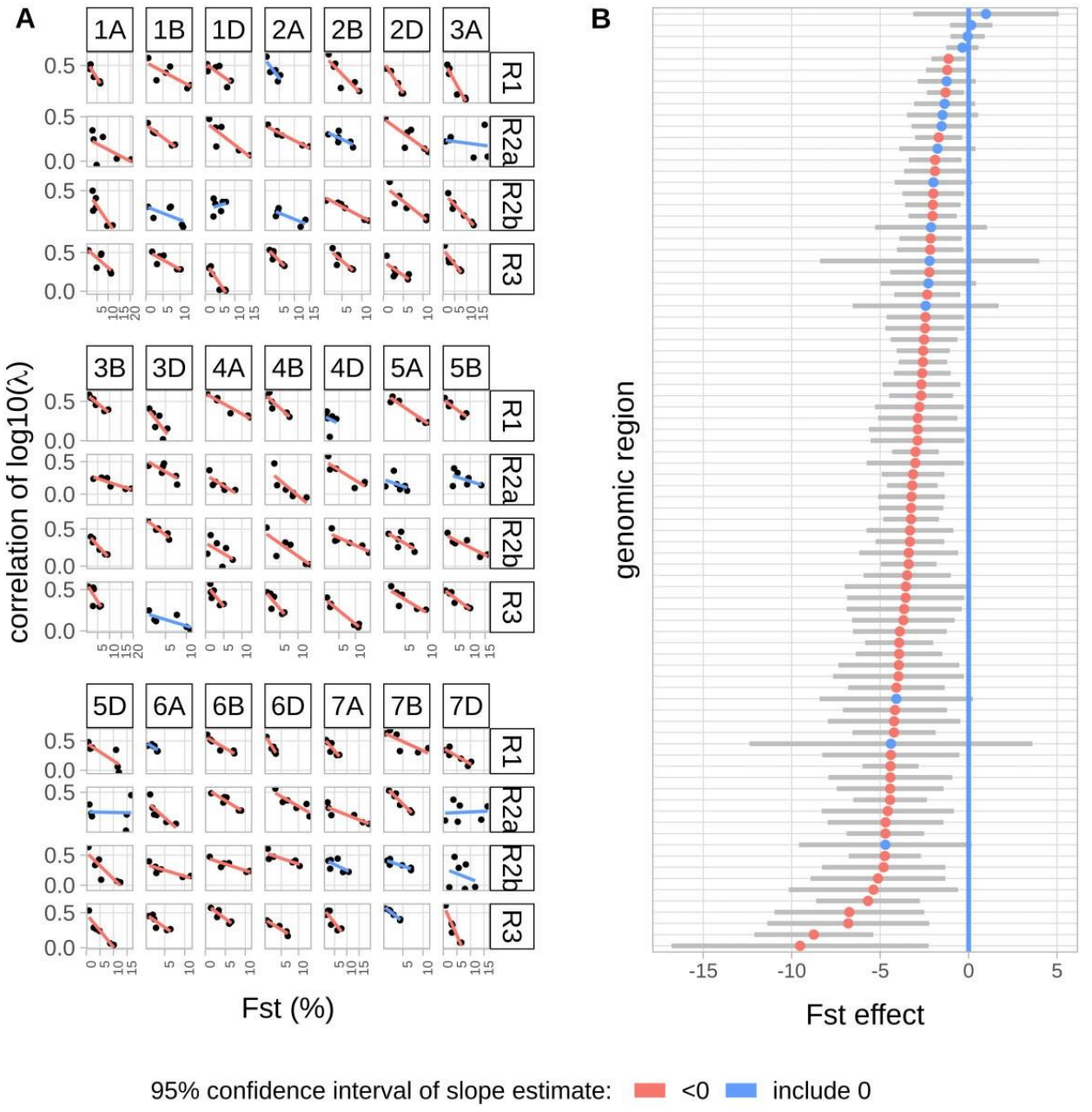


FIG. 6.-Relationships between correlation of local recombination intensity and F_{ST} per genomic region. (*A*) Relationship per genomic region. The slopes values are estimated by linear regression and gives the F_{ST} effects on the correlation of recombination profiles. (*B*) Ranked slope estimates (colored points) and their 95% confidence interval (gray bar). Blue color represents slopes with a confidence interval overlapping 0 and red color confidence interval not overlapping 0.

associated to the pairwise correlation of recombination profiles between populations. We computed F_{ST} around 54 genes known to be involved in the meiosis process (supplementary protocol S3 and file S6, Supplementary Material online) and fitted a specific regression of the F_{ST} around the gene on the genome-wide correlation of recombination profiles (i.e., medians in fig. 4 boxplots, one measure per pair of population, identical for every gene and every genomic region). As the basal level of differentiation depends on the genomic region (fig. 6), we tested whether meiotic genes showed an increased level of differentiation compared with their own genomic region (i.e., a significant negative slope). To control for region-specific effects, meiotic genes were contrasted to "control genes" not involved in meiosis and in the

same genomic region. The number of control genes per genomic region ranged from 9 in 7AC to 733 in 5AR3 regions (median number: 223). Overall, meiotic genes did not show a significantly different slope compared with control genes (*P*-value =0.97). Only *asy4*, located in the 4AR3 genomic region, showed a significantly more negative slope than control genes in its genomic region (False Discovery Rate < 0.01%) (supplementary fig. S16, Supplementary Material online).

Discussion

Fine Scale Genome-Wide Recombination Landscape of Bread Wheat

In our study, we estimated LD-based recombination rates for the first time at the whole-genome scale in bread wheat. Previous studies were done at local scale only (Darrier et al. 2017) but suggested that this approach could be applied genome-wide. We used four diverging populations of landraces representative of the four main worldwide genetic groups (Balfourier et al. 2019). For all maps, 80% of the genetic distance was found in 36% (61%) of the physical distance. This is less concentrated than what was previously observed on single chromosome 3B (80% in less than 20%; Saintenac et al. 2009; Darrier et al. 2017). This discrepancy is likely due, on one hand to the higher SNP density in previous studies on chromosome 3B that allowed to precisely delimit recombination hotspots on this particular chromosome, and on the other hand likely because classical frequentist estimates of recombination rates in biparental maps let most of the genome depleted of recombination. However, and as expected, historical crossovers tend to accumulate in distal sub-telomeric regions of the chromosomes (namely R1 and R3 regions). In most organisms, pairing initiation between homologues occurs in many places along the chromosomes but tends to be favored by a meiosis-specific organization called “bouquet” where telomeres are gathered on the internal nuclear envelope at the Leptotene stage, just before synapsis (Zickler and Kleckner 2015). The bouquet would then facilitate alignment between homologues and pairing would be simultaneously favored through the repair of double-strand breaks including crossovers (Zickler and Kleckner 1998; reviewed in Scherthan 2001 or Harper et al. 2004). In bread wheat, distal crossovers would then be predominant because of the bouquet and be limited in R2a and R2b regions because of interference (Saintenac et al. 2009).

At a fine scale, LD-based maps revealed that 1–2% of intervals of telomeric and pericentromeric regions (depending on the population) exhibited especially high recombination rate (HRIs), suggesting that these intervals overlapped recombination hotspots. The accumulation of crossovers in recombination hotspots was already observed in bread wheat (Saintenac et al. 2011; Darrier et al. 2017) and seems to be a common phenomenon across many species (for a review see Stapley et al. 2017). Recombination hotspots are usually found to be associated with open-chromatin signatures (for a review, see Dluzewska et al. 2018). In previous study in bread wheat, recombination hotspots were found to locate nearby gene promotors and terminators. Our results are consistent with this finding, as most (80%) of our HRIs are located nearby gene features.

LD-based Recombination Maps Correlate Well with the Biparental Genetic Map

In principle, LD-based recombination maps should be suited to study the similarity of recombination profiles of diverging populations. In our study, they allowed to compare recombination rates of four populations with about twice more SNPs than the densest genetic maps currently available (131 - 170k SNPs in EA and WA, respectively, versus 80k SNPs in Rimbert et al. 2018; 55k markers in Liu et al. 2020; 50k SNPs in Jordan et al. 2018). Moreover, LD-based maps are representative of a whole population and less susceptible to individual specific variation, for example, introgressions which are known to prevent local formation of COs between the introgressed chromatid and the native chromatid. Introgressions from wild relative species are frequent in bread wheat species, representing from 4% to 32% of bread wheat genome (Zhou et al. 2020).

The limitation of LD-based maps relies on the fact that they can be affected by evolutionary patterns, which in turn can hinder their usefulness to study the evolution of recombination rate. Indeed, to the extent that evolutionary forces and past demographic events (bottlenecks, population expansions, hidden structuration) affect LD patterns they can also affect recombination rate estimates (Chan et al. 2012; Dapper and Payseur, 2018). To measure to which extent LD-based recombination rates differ from meiotic ones, we compared LD-based maps with the CsRe meiotic map. This revealed that, genome-wide, the correlation between the two approaches was very high (>0.7; table 1). Although part of this correlation is explained by the large differences in recombination rate between chromosomal regions (R1, R2a, R2b, R3, and C), our results also indicate a substantial high correlation within each of these regions. The correlation between LD-based and the CsRe genetic map ranged from 0.50 on average in EA, 0.55 in WA and EE and 0.58 in WE at 4 Mb per genomic region considering all populations but only telomeres and pericentromeres (table 1). This value is consistent with correlation values obtained in the literature for other plant species. For example, the correlation between LD-based and meiotic recombination map was found to be 0.3 in rice (Marand et al. 2019), 0.81 in barley (Dreissig et al. 2019), and 0.44–0.55 in Arabidopsis (Choi et al. 2013). Besides, the correlation values we report are likely to be underestimates of the true values. To compute these correlations, we used estimates of recombination rates. Like any statistical estimates they come with measurement errors of the true parameters. Hence the correlation

between estimates, providing these errors are independent, are necessarily smaller than the true correlation (Fisher 1915). Apart from this statistical effect, we could also explain some of the differences between LD-based maps and the meiotic map by genomic rearrangements (introgressions on chromosome 7D and 2A in Renan) that are specific to the CsRe population: In these regions, the CsRe recombination profile is not representative of the landraces recombination profiles.

The overall similarity between the meiotic map and LD-based maps shows that LD-based recombination patterns offer a robust representation of the distribution of recombination along the bread wheat genome.

Robustness of LD-Based Recombination Maps

Despite good concordance with the meiotic map, LD-based recombination maps can still be locally affected by demographic effects, and thus result in bias when interpreting differences or similarities between populations. For example, Kim and Nielsen (2004) and Chan et al. (2012) showed that selective hard-sweeps can produce LD patterns that mimic those of recombination hotspots. Dapper and Payseur (2018) showed that demographic events can decrease the power to detect hotspots leading to an under estimation of the colocalization of LD-based recombination hotspots when using LDhat (Auton and McVean 2007). Here, we used PHASE (Li and Stephens 2003; Crawford et al. 2004), a software to infer recombination rates from LD patterns that implements a quite different methodological approach than LDhat but it is possible that its inference is also affected by such effects. In particular, there were twice many HRIs detected in WE (2,739) and WA (2,743) than in EE (1,968) and EA (1,253), representing a significant variation from 1% of intervals in EA (122,490 SNPs once centromeres removed) to 2% of intervals in WE (161,953 SNPs once centromeres removed) (significant chi-square test, *P*-value < 2.2e-16). Although this varying number of HRIs per population could result from a variation in recombination patterns, it is likely also due to differences in the power to detect HRIs in each population which would be consistent with results from Dapper and Payseur (2018). Indeed, as the proportion of HRIs per population follows the levels of admixture and SNPs density (both higher for WE and WA than for EE and EA), this favors a possible contribution of a different detection power to the variation of HRIs per population. However, we did not observe any atypical LD-based estimate for intervals located nearby genes known to be involved in domestication (e.g., brittle rachis, tenacious glume, homoeologous pairing or nonfree-threshing character) or in recent crop improvement. To further reduce the potential influence of demographic forces on our inference, we performed the comparison between population maps, not on LD-based recombination rates themselves (ρ) but on the relative rate (λ) of recombination in an interval compared with its neighbors in windows of 2 centi-Morgans. Using relative rates should clean our inference from any local effect of demographic forces, especially selection that could tend to be more shared between closely related populations than distant ones.

Results were not much affected by SNPs ascertainment or the method used to calculate the F_{ST} index. The decreasing similarity of recombination rates with genetic differentiation still hold when estimating LD-based recombination rates on a population specific SNPs data set or a common SNPs data set. The co-localization of HRIs was also not influenced by the SNPs data set (supplementary fig. S17, Supplementary Material online). The estimation of F_{ST} index, using either haplotypic or SNPs alleles, provided also consistent results. Overall, these results strongly support the idea that the decrease of similarity in LD-based recombination profiles is not an artifact of demographic forces or biases due to SNPs ascertainment but that the underlying recombination profile is linked to the divergence of populations.

Evolution of the Recombination Landscape in Bread Wheat

Our results are consistent with previous reports. Gardiner et al. (2019) showed that closely related bread wheat parental lines lead to RIL populations with more similar crossover profiles. Darrier et al. (2017) compared LD-based recombination profiles of a European and an Asian population, the two main ancestral bread wheat genetic pools, on two scaffolds of 1.2 and 2.5 Mb on chromosome 3B. They found that LD-based recombination profiles are broadly conserved, but highlighted that hot intervals in LD-based recombination profiles were not necessarily shared between these two European and Asian populations. Similar results were observed in other plant species such as rice (*Oryza sativa*; Marand et al. 2019) and cocoa-tree (*Theobroma cacao*; Schwarzkopf et al. 2020). Other plant studies hint at a possible decreasing similarity of fine-scale recombination profiles over evolutionary time measured by F_{ST}, such as maize (*Zea mais*, Rodgers-Melnick et al. 2015), poplar (*Populus* species, Wang et al. 2014, 2016), cotton (*Gosypium hirsutum*, Shen et al. 2019), and barley (*Hordeum vulgare*, Dreissig et al. 2019).

Several hypotheses can be formulated to explain the differences in recombination profiles between populations. First, this can be due to environmental effects. This is the case in barley, where recombination rates vary along the genome and are affected by environmental conditions as well as by domestication (Dreissig et al. 2019). For example, high temperatures are known to affect meiosis and above 35 °C, this may lead to complete failure and severe sterility (Loidl 1989; Higgins et al. 2012). Interestingly, within a range of 22 – 30℃, highest temperatures may modify the recombination profile. In barley, it was shown that at 30°C, distal recombination events are reduced whereas interstitial events became more frequent revealing thus a slight shift and a modification of the global recombination profile (Higgins et

al. 2012). However, in our case, this hypothesis is not the most likely as we were using populations from the same hemisphere and latitudes, with landraces from different countries. Environment is thus certainly very different between all the origins of our landraces and temperature should vary a lot in each location and is not stable enough to affect durably and maintain a different recombination profile between the four populations. Moreover, it was recently shown that increased temperature up to 28°C for 3 weeks during wheat meiosis has only a limited impact on recombination distribution (Coulton et al. 2020).

Secondly, differences in recombination profiles can be explained by differences in the chromatin accessibility landscape during meiosis between populations. Many studies showed that chromatin status is the main feature that drives recombination in plants. DNA is partitioned in blocks of heterochromatin and euchromatin which are dispersed along the chromosomes. In bread wheat, heterochromatin preferentially locates in pericentromeric regions whereas euchromatin-rich DNA is more frequent in distal subtelomeric regions of the chromosomes (IWGSC 2018). In *Arabidopsis*, it was shown that crossovers are enriched in euchromatin and mainly occur close to gene promoters and terminators (Choi et al. 2013; Drouaud et al. 2013). Meiotic recombination profile in this species is also shaped by H2A.Z nucleosome occupancy, DNA methylation or epigenetic marks such as Histone 3 Lysine 9 di-methylation (H3K9me2; Choi et al. 2013; Underwood et al. 2018). This led to our second hypothesis that chromatin status has evolved between our four populations, rather than an evolution of the recombination determinism itself. Divergence in chromatin status could be explained by genetic drift on one hand or by selection pressure around different genomic regions depending on geographical area on the other hand. This selection pressure could therefore contribute to the deposition of histone landmarks to regulate gene activity such as H3K4me3, H3K9ac, and H3K27ac that are associated with transcriptional activation (Roth et al. 2001; Howe et al. 2017) or on the contrary H3K27me3 and H3K9me3 associated with transcriptional suppression (Saksouk et al. 2015). Interestingly, in some mammals, recombination is directed by the zinc-finger protein PRDM9 that possesses a set domain that catalyzes the trimethylation of lysine 4 of H3 to produce H3K4me3 (for review see Grey et al. 2017). Similar mechanisms involving histone 3 modifications such as methylation or acetylation that could affect recombination profile afterward are thus likely in plants. We tested differentiation of 54 meiotic genes along evolution of recombination profile. In average, these 54 meiotic genes did not show a higher or lower differentiation level than control genes of their own genomic region. Only ASY4 located in 4AR3 genomic region, showed a significant higher level of differentiation than control genes. In *Arabidopsis*, the *asy4* protein is involved in the formation of the axis between the two sister chromatids (Chambon et al. 2018). Mutation of *Atasy4* significantly reduces the number of crossovers and induces a shift toward the distal parts of the chromosomes. This could explain why we found this gene associated with a difference in recombination rates between populations.

Another factor that may explain the difference of recombination patterns between the populations could be the natural introgression of alien DNA fragments from wheat relatives during the evolution process. Introgressions from wild-species have been widely used and more than 50 alien germplasms have been used to improve wheat varieties (Wulff and Moscou 2014). For example, Renan possesses two introgressed fragments from *Aegilops ventricosa* conferring resistance to leaf, yellow, and stem rusts (*Lr37*/*Yr17*/*Sr38*) on chromosome 2A (2A/2N translocation) and to eye-spot (*Pch1*) on chromosome 7D (7D/7Dv translocation; Maia 1967; Helguera et al. 2003). These introgressions repress recombination (Worland et al. 1988) and this resulted in a poor correlation between CsRe genetic map and our LD-based maps for genomic region 7DR3 in our analysis. It was recently shown that natural or artificial introgressions of wheat wild-relatives DNA contributed to up to 710 and 1580 Mb in wheat landraces and varieties, respectively (Cheng et al. 2019), and represent from 4% to 32% of bread wheat varieties genome (Zhou et al. 2020). A similar analysis used exome capture to evaluate introgression in 890 hexaploid and tetraploid wheats (He et al. 2019). The results also suggest that introgressions of DNA fragments from wheat relatives contributed significantly to improve the diversity of current wheat cultivars. Because natural introgressions are frequent in wheat landraces and because they contribute to modify the recombination profile, we could hypothesize that these introgressions are different in our four collections, which would result in different recombination profiles as well. Only an extensive sequencing of our accessions would allow to bring the answer.

Conclusion

This study demonstrates the evolution of the recombination profile at a genome-wide scale in closely related wheat populations with increasing genetic divergence. Based on recombination landscapes robust to demographic events, the comparison of the four landrace populations revealed a clear signal of a decreasing similarity between fine-scale recombination landscapes with increasing genetic divergence. Specifically, we found 1) that HRIs were more shared between closely related populations, 2) recombination intensities at HRIs detected in one population decreased in the other populations with their genetic divergence, and 3) the correlation of recombination landscapes between pairs of population decreases with their local genetic differentiation as measured by F_{ST}. Our results, interpreted in the light of previous findings in bread wheat and other species, clearly shows that recombination landscapes in wheat change with genetic divergence

between populations. Being based on closely related populations that recently diverged (no more than 10,000 YA), this study further shows that this divergence can be quite fast. Reasons for this divergence remain to be found but our results can hint at some possibilities. Further analyses are needed to settle this question, which should greatly help developing original approaches useful for wheat improvement and breeding.

Materials and Methods

Plant Material

A collection of 632 bread wheat landraces (Balfourier et al. 2019) was genotyped on the TaBW410k SNPs including 280k SNPs from the Axiom Affymetrix TaBW280k SNPs array (Rimbert et al. 2018). Besides, a population of 406 F6 RILs derived from the cross between the Asian variety Chinese Spring and the European variety Renan (CsRe), were also genotyped on the TaBW280k SNPs array (Rimbert et al. 2018). After quality filtering including control of missing data rate (10% maximum), heterozygosity rate (5% maximum), excluding off-target variants, 578 landraces genotyped with 200,062 SNPs were kept for the population-based anal- ysis and 79,564 polymorphic SNPs were successfully mapped on the CsRe population.

The physical positions of SNPs on the 21 bread wheat chromosomes were determined using Basic Local Alignment Search Tool (Blast; Altschul et al. 1990) of context sequences on the International Wheat Genome Sequencing Consortium RefSeq v1.0 genome assembly (IWGSC 2018). Position of high confidence genes, exon, 5′-UTR and 3′-UTR were extracted from RefSeq V1.0 annotation.

Robust Estimation of the Meiotic Recombination Profile

Due to the relatively low number of meiosis sampled in the CsRe data, a Bayesian model inspired from Petit et al. (2017) was used to obtain robust estimates of recombination rates. We modelled the probability distribution of the recombination rates observed in RILs (C_i) given the number of observed recombination events (y_i) as:

$$P(C_i \mid y_i) = \frac{P(y_i \mid C_i)\, P(C_i)}{P(y_i)}$$

The likelihood $P(C_i|y_i)$ is modelled as a Poisson distribution, its parameter being the expected number of recombination events in an interval and computed as: $\mathrm{E}(y_i) = C_i \times L_i \times M$ (where L_i is the physical size (in megabases, Mb) of the interval and M the total number of RILs.Thus, the likelihood of the recombination rate C is:

$$P(y_i|C_i) \sim \mathrm{Poisson}(C_i \times L_i \times M)$$

To specify a prior distribution of $P(C_i)$ we considered that the wheat recombination landscape varies widely along a chromosome. According to the nomenclature of Choulet et al. 2014, each of the wheat chromosomes can be segmented into five chromosomic regions associated with different global recombination rates and genomic content: Two highly recom- bining telomeric regions (R1 and R3), two low-recombining pericentromeric regions (R2a and R2b) and one centromeric region (C) where recombination is almost completely sup- pressed. The small arm of each chromosome is composed of R1 and R2a whereas the long arm is composed of R2b and R3. The physical size of these regions ranges between 10 Mb for the smallest telomere and 321 Mb for the largest pericentromere (supplementary file S7, Supplementary Material online). To account for the specific range of recom- bination rate variation in each region in our model, the prior distribution of the recombination rates in each of these regions was a specific Gamma distribution:

$$P\left(C_{i(r)}\right) \sim \Gamma(\alpha_r, \beta_r),$$

where r denotes the region, α_r/β_r gives the mean of the Gamma distribution and α_r/β_r^2 gives the variance. The Gamma distribution being a conjugate prior to the Poisson distribution, the posterior distribution of C_i is also a Gamma distribution:

$$P(C_i|y_i) \sim \Gamma(y_i + \alpha_r; ML_i + \beta_r)$$

The posterior mean of C_i (in M/Mb) is then:

$$C_{i(r)}^{bay} = \frac{y_i + \alpha_r}{ML_i + \beta_r}$$

The parameters α_r and β_r of the prior Gamma distribution were set using an empirical Bayes approach (i.e., estimating prior distribution directly from data), independently for each of the five r regions (supplementary fig. S18, Supplementary Material online). A Gamma distribution was fitted (R MASS package, Venables and Ripley 2002) over the distribution of frequentist recombination rates observed in RILs. This latter was computed as:

$$C_i^{freq} = \frac{y_i}{ML_i}$$

Note that null recombination rates were replaced by the lowest non-null estimate of recombination rates of the region to allow fitting the Gamma distribution. We derived the meiotic recombination rates from the RILs recombination rates

using the Haldane and Waddington formula (Haldane and Waddington 1931) and the Morgan mapping function (cM = frequency of recombinants × 100). Indeed, the size of intervals (median ¼ 5 kb) were small enough to consider that interference is very strong within and thus one recombination in one individual result from only one crossover (and not from coincidence of several crossovers). We thus obtained the Bayesian meiotic recombination rate c^{bay}_{CsRei} (cM/Mb).

Considering Uncertainty in Crossover Locations

For estimation of recombination rates, it was necessary to count the number of recombinants in CsRe intervals (y_i). Missing data on genomic segments with no parental allele switch at segment extremities were imputed. A crossover was counted at each parental allele switch, yielding 26,239 crossovers. Due to the presence of missing data in RILs genotypes, a number of switches did not occur between pairs of immediately adjacent markers. In such cases, the crossover cannot be assigned with certainty to a single interval of two successive SNPs. For example, an RIL genotype AA/–/BB identifies a switch between the first and third marker but cannot discriminate a recombination in the first versus the second interval. In such cases, we accounted for the uncertainty in crossover location following the sampling procedure of Petit et al. (2017). Briefly, each crossover is overlapped by a set of one or more intervals. A sampling procedure assigned each crossover to a particular interval with a probability computed as the size of the interval divided by the size of the crossover area (physical distance between the two closest SNPs showing different parental alleles). Repeating 1,000 times the sampling procedure yields 1,000 estimates of y_i per interval, which can then be converted into recombination rates and averaged.

LD-Based Recombination Profiles of Four Diverging Populations of Landraces from Patterns of LD

Identification of Four Diverging Populations of Landraces Representative of Bread Wheat Worldwide Diversity

We defined four populations from a data set of 632 landraces representative of worldwide genetic diversity of bread wheat and previously described in Balfourier et al. (2019). The constitution of populations followed a three steps procedure that we briefly described (more details in supplementary protocol S4 and figs. S1 and S19, Supplementary Material online):

i. From Balfourier et al. (2019) *K*¼ 4 admixture analysis of the 632 landraces, we kept only 534 low admixed land- races to maximize differentiation between future four populations.

ii. The 534 landraces were gathered into 4 groups using a hierarchical clustering on the pairwise distance matrix estimated in Balfourier et al. (2019). The four populations were named as West Europe (WE), East Europe (EE), West Asia (WA), and East Asia (EA) from the geographical origin of their members.

The pairwise matrix distance gave the pro- portion of mismatched haplotypic alleles along the genome, computed using 8,741 haplotypic blocks containing up to 20 alleles per block (figure 1 of Balfourier et al. [2019]).

iii. We discarded closely related individuals within each population to avoid over representing family specific recombination events. Pairs of individuals exhibiting a very low genetic difference were discarded, keeping a total of 371 landraces.

Evolutionary Distance between Populations Measured by FST

Pairwise differentiation indexes (F_{ST}) of the four populations were computed within each genomic region (chromosomal region within a chromosome, e.g., 1AR1) using alleles of 8,741 haplotypic blocks (Weir and Cockerham distance, R hierfstat package, function pairwise. WCfst, Goudet and Jombart 2015) or SNPs (Reynolds distance, HAPFLK software, Bonhomme et al. 2010; Fariello et al. 2013) (supplementary file S8, Supplementary Material online).

Inferences of LD-Based Recombination Rates from LD Patterns

LD-based recombination rates were estimated using PHASE software V2.1.1 (Li and Stephens 2003; Crawford et al. 2004; Stephens and Scheet 2005). PHASE inputs were successive windows of SNPs along the genome, constituted of one central part and two flanking parts overlapping the previous and the next windows to avoid border effect in PHASE inferences. Central and flanking parts spanned on average 1 cM and 0.5 cM, respectively, based on the CsRe genetic map (supplementary protocol S5, Supplementary Material online). PHASE was run for each window with default options, except for two parameters of the Markov Chain Monte Carlo (MCMC), following recommendations of the documentation on estimating recombination rates. The number of sampling iterations was increased to obtain larger posterior samples (option -X10) and the algorithm was run ten times independently (option –x10) to better explore combinations of parameters and keep the run with the best goodness of fit. The sampling stage of the MCMC yielded 1,000 samples of the posterior distribution of:

- The background recombination rate of the window w: q_w
- The ratio k_i between the background recombination rate of the window q_W and the LD - based recombination rate in each interval i of two successive SNPs q_i so that q_i ¼ k_im $q_{W\eth i\Þ}$ where $w(i)$ identifies the window which interval i belongs to. The parameter k_i can be seen as a measure of local recombination intensity compared with genomic background (inflation or deflation).

PHASE samples jointly ρ_w and λ_i in their posterior distribution at each iteration, so their product yields 1,000 samples of the posterior distribution of LD-based recombination rate ρ_i (/kb) (supplementary file S9, Supplementary Material online). We assessed the sensibility of PHASE results to prior distribution parameters and population sample size and we found that our inference was robust to modifications of the prior distribution or the down-sampling of the WE population (supplementary protocol S2, Supplementary Material online).

Correlation of LD-Based Recombination Profiles

To compare LD-based recombination profiles, it was necessary to obtain a common set of intervals across the four populations (WE, EE, WA, EA), as polymorphic SNPs sets were different. We defined smaller intervals formed of successive markers that were polymorphic in at least one population (supplementary fig. S20, Supplementary Material online). For each population, the recombination estimates in smaller intervals were considered to be the same as the estimates belonging to population specific intervals overlapping them, assuming that recombination rates are constant within intervals. We removed intervals not overlapped by all populations on chromosome extremities. This process yielded a complete factorial data set of 194,409 intervals with no missing data and a set of 1,000 values sampled from the posterior distribution for each parameter ρ_{pi} and λ_{pi} per interval i and per population p. The similarity between LD-based recombination profiles was measured by correlating the $\log_{10}$ of median of λ_{pi} (noted $\log_{10}(\overline{\lambda_{pi}})$) of all intervals between different populations. The median of posterior distribution of λ_{pi} was chosen as it is robust to outliers in the posterior distribution, as recommended (Li and Stephens 2003) and using the log scale is natural when comparing intensities across groups. To obtain correlation coefficients, a linear model including a full unstructured variance – covariance matrix was fitted on $log_{10}(\overline{\lambda_{pi}})$, so that each population had its own range of variation of local recombination intensity and each pair of population has a specific covariance parameter:

$$Y_{pi} = log_{10}(\overline{\lambda_{pi}})$$

$$Y_{pi} = \boldsymbol{\mu} + E_p$$

$\vec{E} \sim MVN(0, I_n \otimes \Sigma_{4*4})$, where Σ_{4*4} is a variance–covariance matrix from which we extract correlation coefficients.

The model was applied independently to each genomic region (from 1AR1 to 7DR3, except centromeric regions, results in supplementary file S10, Supplementary Material online). The total number of intervals n per genomic region ranged from 154 to 8,131. The differences of recombination intensity profiles across the four populations of landraces were assessed by model comparison. We compared the Bayesian Information Criterion (BIC) of a model with a full variance—covariance matrix with a simpler model with a variance—covariance matrix including only one correlation parameter for all pairs of populations. The complex model was deemed to be a better model if its BIC was inferior to the BIC of the simpler model. The models were fitted with ASReml-R V3 (Butler et al. 2009).

Colocalization of HRIs between Populations

Intervals with a LD-based recombination rate exceeding four- times or more the background recombination rate ($\lambda \geq 4$) figuring as outliers in λ distribution (supplementary fig. S21 and file S11, Supplementary Material online), were defined as HRIs and adjacent HRIs within a population were merged. Due to strong heterogeneity of HRI's size, we discarded too small or too wide HRIs (supplementary protocol S6, Supplementary Material online). For each HRI in each population, the over- lapping HRIs in other populations were recorded. A set of HRIs intervals was considered as co-localizing in two, three or four populations if all HRIs overlapped each other (i.e., they formed a clique in network terminology). Note that this implies that a wide HRI can potentially be involved in more than one clique. For each group of colocalizing HRIs (each clique), we defined a hot window as the smallest common overlapped area (sup- plementary file S12, Supplementary Material online). Population specific HRIs, that is, HRIs which did not overlap any other HRIs, also formed hot windows whose frontiers were defined by the upper and lower limit of HRIs. Each hot window thus included HRIs of one, two, three or four populations. The proportion of HRIs shared by two popula- tions or more (e.g., WE and EE) was computed as the number of hot windows including HRIs of each population (hot windows including both WE's HR and EE's HR) divided by total number of hot windows (including either WE, EE, WA or EA's HRIs) (supplementary fig. S22, Supplementary Material on-line). Dividing by the total number of hot windows is more convenient to compare the proportion of HRIs population-specific, or shared by two, three or four populations.

To test for the hypothesis that the observed proportion of HRIs shared by populations is due to chance, an empiric range of plausible values of co-localization due to chanc was estimated by simulation. In 1,000 simulations, each HF of each population was assigned to a random interval with the genomic region it belongs (1AR1 to 7DR3) and th proportion of shared hot windows was computed (supple tary file S13, Supplementary material online).

Comparison of the LD-Based Recombination Rates and the CsRe Meiotic Recombination Rates

The comparison between meiotic (CsRe) and LD-based recombination rates were done on windows of 4 Mb (~1 cM on average, wide enough to accurately estimate intrinsic recombination rate) along the genome. Meiotic recombination rates were estimated using the Bayesian

model described above, the attribution of crossover to windows being done using the (Petit et al. 2017) approach (see above). To compute the LD-based recombination rate in 4 Mb windows, the total LD-based genetic distance per window of 4 Mb was divided by the total physical distance and averaged over the 1,000 samples of the posterior distribution:

$$\rho_{pw_{4Mb}} = \frac{1}{1,000}\sum_{j=1}^{1,000}\frac{\sum_{i \epsilon w_{4Mb}}(\rho_{pij} * L_i)}{\sum_{i \epsilon w_{4Mb}} L_i}$$

with *i* the interval and *j* one posterior distribution value among 1,000 (supplementary file S14, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors would like to thank the GenoToul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing support, computing, and storage resources. The authors thank Hélène Rimbert for her help in identifying SNP positions on refSeq v1.0 and providing CsRe data files, Ingrid David for her help with the ASReml software and Gilles Charmet, Sylvain Glémin, Susan Johnston and Jean-Michel Elsen for their comments on a previous version of the manuscript. Doctoral work of ADDD was funded by the INRAE metaprogram SELGEN and Florimond Desprez (Cappelle-en-Pévèle, France). Genotyping was supported by the Breedwheat grant (ANR-10-BTBR-03).

Data Availability

Genotyping data set of landraces on TaBW410k (Kitt et al. 2021) after quality control is available as a Zeonodo repository (https://doi.org/10.5281/zenodo.4518374). Genotyping data set of RILs (Kitt et al. 2018) is available as a Zenodo repository (https://doi.org/10.5281/zenodo.4486612). Supplementary figures and protocols can be found in the main supplementary file. Supplementary data including PHASE outputs and population-specific meiotic recombination maps, are available in a Zenodo repository (https://doi.org/10.5281/zenodo.4486586). Computer code and scripts needed to reproduce all results are available on Github (https://github.com/aldanguy/Bread_wheat_recombination).

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17(8):1219–1227.

Auton A, McVean G. 2012. Estimation rates from genetic variation in humans. In: Anisimova M, editor. Evolutionary genomics: statistical and computational methods, Vol. 2. Totowa (NJ): Humana Press. p. 217–237.

Auton A, et al. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. PLOS Genet. 9(12):e1003984.

Badr A, et al. 2000. On the origin and domestication history of Barley (*Hordeum vulgare*). Mol Biol Evol. 17(4):499–510.

Balfourier F, BreedWheat Consortium, et al. 2019. Worldwide phylogeography and history of wheat genetic diversity. Sci Adv. 5(5):eaav0536.

Barton NH. 1995. A general model for the evolution of recombination. Genet Res. 65(2):123–144.

Baudat F, et al. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327(5967):836–840.

Bauer E, et al. 2013. Intraspecific variation of recombination rate in maize. Genome Biol. 14(9):R103–17.

Berchowitz LE, Copenhaver GP. 2010. Genetic interference: dont stand so close to me. Curr Genomics 11(2):91–102.

Bonhomme M, et al. 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics 186(1):241–262.

Bouchet S, et al. 2017. Increased power to dissect adaptive traits in global Sorghum diversity using a nested association mapping population. Genetics 206(2):573–585.

Boulton A, Myers RS, Redfield RJ. 1997. The hotspot conversion paradox and the evolution of meiotic recombination. Proc Natl Acad Sci U S A. 94(15):8058–8063.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ. 2009. ASReml-R reference manual. State Qld Dep Prim Ind Fish Brisb.

Chambon A, et al. 2018. Identification of ASYNAPTIC4, a component of the meiotic chromosome axis. Plant Physiol. 178(1):233–246.

Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLOS Genet. 8(12):e1003090.

Charlesworth B, Barton NH. 1996. Recombination load associated with selection for increased recombination. Genet Res. 67(1):27–41.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics.
USA: Roberts and Company Publishers Greenwood Village, CO.

Cheng H, et al. 2019. Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol. 20(1):136.

Choi K, Henderson IR. 2015. Meiotic recombination hotspots – a comparative view. Plant J. 83(1):52–61.

Choi K, et al. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat Genet. 45(11):1327–1336.

Choulet F, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science 345(6194):1249721.

Coulton A, Burridge AJ, Edwards KJ. 2020. Examining the effects of temperature on recombination in wheat. Front Plant Sci. 11:230.

Crawford DC, et al. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet. 36(7):700–706.
Danguy des Déserts A, Bouchet S, Sourdille P, Servin B. 2021. Supplementary files from: evolution of recombination landscapes in diverging populations of bread wheat. Zenodo. https://doi.org/10.5281/zenodo.4486586. Accessed February 1, 2021.
Dapper AL, Payseur BA. 2018. Effects of demographic history on the de- tection of recombination hotspots from linkage disequilibrium. Mol Biol Evol. 35(2):335–353.
Darrier B, et al. 2017. High-resolution mapping of crossover events in the hexaploid wheat genome suggests a universal recombination mecha- nism. Genetics 206(3):1373–1388.
Dluzewska J, Szymanska M, Ziolkowski PA. 2018. Where to cross over? Defining crossover sites in plants. Front Genet. 9:609–609.
Dreissig S, Mascher M, Heckmann S. 2019. Variation in recombination rate is shaped by domestication and environmental conditions in Barley. Mol Biol Evol. 36(9):2029–2039.
Drouaud J, et al. 2013. Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. PLOS Genet. 9(11):e1003922.
Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hier-archically structured populations. Genetics 193(3):929–941.
Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10(4):507–521.
Gardiner L-J, et al. 2019. Analysis of the recombination landscape of hexa-ploid bread wheat reveals genes controlling recombination and gene conversion frequency. Genome Biol. 20(1):69.
Gini C. 1936. On the measure of concentration with special reference to income and statistics. Colo Coll Publ Gen Ser. 208:73–79.
Goudet J, Jombart T. 2015. hierfstat: estimation and tests of hierarchical F-statistics. R Package Version 004-22. 10.
Grey C, et al. 2017. In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites. Genome Res. 27(4):580–590.
Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol Ecol. 27(11):2477–2497.
Haldane JB, Waddington CH. 1931. Inbreeding and linkage. Genetics 16(4):357–374.
Harper L, Golubovskaya I, Cande WZ. 2004. A bouquet of chromosomes. J Cell Sci. 117(Pt 18):4025–4032.
He F, et al. 2019. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet. 51(5):896–904.
Helguera M, et al. 2003. PCR assays for the Lr37-Yr17-Sr38 cluster of rust resistance genes and their use to develop isogenic hard red spring wheat lines. Crop Sci. 43(5):1839–1847.
Higgins JD, et al. 2012. Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in Barley. Plant Cell 24(10):4096–4109.
Howe FS, Fischl H, Murray SC, Mellor J. 2017. Is H3K4me3 instructive for transcription activation? Bioessays 39(1):e201600095.
IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345(6194):1251788.
IWGSC. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403):eaar7191.
Jordan KW, et al. 2018. The genetic architecture of genome-wide recom-bination rate variation in allopolyploid wheat revealed by nested asso-ciation mapping. Plant J. 95(6):1039–1054.
Kitt J, Darrier B, Rimbert H, Sourdille P, Paux E. 2018. Genotyping of the Chinese Spring × Renan mapping population with the TaBW280K SNP array. Zenodo. 10.5281/zenodo.4486612.
Kitt J., Danguy Des Déserts A., Bouchet S., Servin B., Rimbert H., De Oliveira R., Choulet F., Balfourier F., Sourdille P, and PauxE. 2021. Genotyping of 4,506 bread wheat accessions with the TaBW410K SNP array. Zenodo. 10.5281/zenodo.4518374.
Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167(3):1513–1524.
Kong A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467(7319):1099–1103.
Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165(4):2213–2233.
Liu J, et al. 2020. A novel, major, and validated QTL for the effective tiller number located on chromosome arm 1BL in bread wheat. Plant Mol Biol. 104(1-2):173–185.
Loidl J. 1989. Effects of elevated temperature on meiotic chromosome synapsis in *Allium ursinum*. Chromosoma 97(6):449–458.
Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A. 101(34):12404–12410.
Maia N. 1967. Obtention de blés tendres résistants au piétin-verse par croisements interspécifiques blé x Aegilops [Obtaining soft wheat re- sistant to eyespot Cercosporella herpotrichide by wheat x Aegilops interspecific crosses]. Comptes Rendus L'Académie D'Agriculture Fr. 53:149–155.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High- resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454(7203):479–485.
Marand AP, et al. 2017. Meiotic crossovers are associated with open chro- matin and enriched with Stowaway transposons in potato. Genome Biol. 18(1):203.
Marand AP, et al. 2019. Historical meiotic crossover hotspots fueled pat- terns of evolutionary divergence in rice. Plant Cell 31(3):645–662.
Mercier R, Mézard C, Jenczewski E, Macaisne N, Grelon M. 2015. The molecular biology of meiosis in plants. Annu Rev Plant Biol. 66(1):297–327.
Murakami H, et al. 2020. Multilayered mechanisms ensure that short chromosomes recombine in meiosis. Nature 582(7810):124–128.
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–324.
Myers S, et al. 2010. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. Science 327(5967):876–879.
Oliver PL, et al. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse Metazoan Taxa. PLOS Genet. 5(12):e1000753.
Otto SP. 2009. The evolutionary enigma of sex. Am Nat. 174(S1):S1–S14.
Petit M, et al. 2017. Variation in recombination rate and its genetic deter-minism in sheep populations. Genetics 207(2):767–784.
Pont C, et al.; Wheat and Barley Legacy for Breeding Improvement (WHEALBI) Consortium. 2019. Tracing the ancestry of modern bread wheats. Nat Genet. 51(5):905–911.
Rimbert H, et al. 2018. High throughput SNP discovery and genotyping in hexaploid wheat. PLOS One 13(1):e0186329.
Rodgers-Melnick E, et al. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc Natl Acad Sci U S A. 112(12):3823–3828.
Roth SY, Denu JM, Allis CD. 2001. Histone acetyltransferases. Annu Rev Biochem. 70(1):81–120.
Rowan BA, et al. 2019. An ultra high-density *Arabidopsis thaliana* cross- over map that refines the influences of structural variation and epige- netic features. Genetics 213(3):771–787.

Saintenac C, et al. 2009. Detailed recombination studies along chromo- some 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). Genetics 181(2):393–403.
Saintenac C, et al. 2011. Variation in crossover rates across a 3-Mb contig of bread wheat (Triticum aestivum) reveals the presence of a meiotic recombination hotspot. Chromosoma 120(2):185–198.
Saksouk N, Simboeck E, Déjardin J. 2015. Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin 8(1):3.
Salomé PA, et al. 2012. The recombination landscape in *Arabidopsis thaliana* F2 populations. Heredity (Edinb) 108(4):447–455.
Scherthan H. 2001. A bouquet makes ends meet. Nat Rev Mol Cell Biol. 2(8):621–627.
Schield DR, et al. 2020. Snake recombination landscapes are concentrated in functional regions despite PRDM9. Mol Biol Evol. 37(5):1272–1294.
Schwarzkopf EJ, Motamayor JC, Cornejo OE. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. BMC Genomics 21(1):1–16.
Serra H, et al. 2018. Interhomolog polymorphism shapes meiotic crossover within the *Arabidopsis* RAC1 and RPP13 disease resistance genes. PLOS Genet. 14(12):e1007843.
Shanfelter AF, Archambeault SL, White MA. 2019. Divergent fine-scale recombination landscapes between a freshwater and marine population of threespine stickleback fish. Genome Biol Evol. 11(6):1552–1572.
Shen C, et al. 2019. Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. Plant J. 99(3):494–505.
Singhal S, et al. 2015. Stable recombination hotspots in birds. Science 350(6263):928–932.
Smeds L, Mugal CF, Qvarnstro¨ m A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an Avian pedigree. PLOS Genet. 12(5):e1006044.
Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc B Biol Sci. 372:1736.
Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 76(3):449–462.
Sun C, et al. 2020. The wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. Plant Biotechnol J. 18(6):1354–1360.
Tang T, et al. 2006. Genomic variation in rice: genesis of highly polymorphic linkage blocks during domestication. PLOS Genet. 2(11):e199.
Tsai IJ, Burt A, Koufopanou V. 2010. Conservation of recombination hotspots in yeast. Proc Natl Acad Sci U S A. 107(17):7847–7852.
Underwood CJ, et al. 2018. Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. Genome Res. 28(4):519–531.
Venables WN, Ripley BD. 2002. Modern applied statistics in S. 4th ed. New York: Springer. p. 1–498.
Vitte C, Ishii T, Lamy F, Brar D, Panaud O. 2004. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol Genet Genomics 272(5):504–511.
Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. Genetics. 202(3):1185.
Wang Z, et al. 2014. Phylogeny reconstruction and hybrid analysis of Populus (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. PLOS One 9(8):e103645.
Wicker T, et al.; International Wheat Genome Sequencing Consortium. 2018. Impact of transposable elements on genome structure and evo- lution in bread wheat. Genome Biol. 19(1):103.
Winfield MO, et al. 2016. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. Plant Biotechnol J. 14(5):1195–1206.
Worland AJ, Law CN, Hollins TW, Koebner RMD, Giura A. 1988. Location of a gene for resistance to eyespot (*Pseudocercosparella herpotri- choides*) on chromosome 7D of bread wheat. Plant Breed 101(1):43–51.
Wulff BBH, Moscou MJ. 2014. Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. Front Plant Sci. 5:692.
Zhou Y, et al. 2020. Triticum population sequencing provides insights into wheat adaptation. Nat Genet. 52(12):1412–1422.
Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. New Phytol. 167(1):249–265.
Zickler D, Kleckner N. 1998. The leptotene-zygotene transition of meiosis.
Annu Rev Genet. 32(1):619–697.
Zickler D, Kleckner N. 2015. Recombination, pairing, and synapsis of homologs during meiosis. Cold Spring Harb Perspect Biol. 7(6):a016626.
Zou GY. 2007. Toward using confidence intervals to compare correlations.
Psychol Methods 12(4):399–413.

Associate editor: Brandon Gaut

II.3 Conclusion sur l'évolution du profil de recombinaison chez le blé tendre

II.3.1 Estimation du taux de recombinaison à partir des patrons de déséquilibre de liaison

L'objectif de cet article est d'évaluer la variabilité du profil de recombinaison chez le blé tendre. Cela implique d'estimer et de comparer le profil de recombinaisons de plusieurs populations divergentes.

Nous avons estimé le profil de recombinaison à partir des patrons de DL (« LD-based ») spécifiques des quatre principaux groupes génétiques du blé tendre déterminés à partir d'un génotypage dense à l'aide de plus de 280 000 marqueurs SNP. Ces profils de recombinaison historiques ont deux avantages par rapport aux cartes génétiques classiques obtenues sur des populations expérimentales en ségrégation :

1) une plus grande densité de marqueurs polymorphes pour positionner précisément les évènements de recombinaison ancestraux, grâce à la grande diversité de lignées parentales impliquées (jusqu'à 127 pour la population WE, après avoir supprimé les individus trop apparentés) alors que les populations expérimentales en ségrégation en blé tendre incluent rarement plus d'une vingtaine de parents (Rimbert et al. 2018; Gardiner et al. 2019; M. Lin et al. 2020; Liu et al. 2020; Jordan et al. 2018) .

2) un échantillonnage d'un grand nombre de méioses, puisque les patrons de DL sont en théorie le produit de l'ensemble des méioses efficaces depuis la coalescence de la population, qui date approximativement de la diffusion du blé tendre depuis le Croissant fertile, soit environ 10 000 ans. A l'inverse, les populations expérimentales échantillonnent au plus quelques milliers de méioses, selon le nombre de lignées recombinantes étudiées. Cependant, le nombre de méioses efficaces responsables des patrons de DL n'est pas aussi élevé que pourraient le laisser penser ces échelles de temps, car le blé est une espèce majoritairement autogame (taux d'autofécondation de 95%), donc la majorité des recombinaisons sont invisibles. (Nordborg 2000) démontre effectivement que les populations qui pratiquent l'autogamie ont un rythme de coalescence plus rapide que les populations allogames. Par ailleurs, les profils historiques ont aussi l'avantage d'être peu sensibles aux variations individuelles du profil de recombinaison, causées par exemple par des variations structurales : variation du nombre de copies (Copy Number Variation, CNV) ou présence/absence (PAV) d'une copie. Les variations structurales sont très fréquentes chez le blé tendre (Cheng et al. 2019; He et al. 2019; Zhou et al. 2020). Ainsi, la méthode historique a permis d'obtenir pour la première fois des profils de recombinaison représentatifs des quatre principaux groupes génétiques du blé tendre avec une résolution inégalée sur l'ensemble du génome.

Ces variations fines du taux de recombinaison historique ont permis de confirmer la grande hétérogénéité de la recombinaison fine-échelle, qui avait déjà été observée dans des régions de

quelques mégabases (Saintenac et al. 2011; Darrier et al. 2017). Ces profils ont permis d'identifier ≈ 8k intervalles avec une forte inflation locale du taux de recombinaison. Ces intervalles chevauchent vraisemblablement des points chauds de recombinaison, où les crossing-over (CO) ont été particulièrement fréquents au cours de l'histoire évolutive des populations. Ces intervalles fortement recombinants (appelés HRIs) sont relativement rares (entre 1 à 2% des intervalles) mais représentent plus de 15% de la distance génétique historique. Conformément aux observations chez les autres espèces, la plupart des points chauds chez le blé tendre se situent dans les régions télomériques et à proximité des gènes.

II.3.2 Impact des forces évolutives

Si les profils de recombinaison historiques présentent de nombreux avantages, ils doivent être interprétés avec prudence. La plupart des modèles utilisés pour inférer le taux de recombinaison historique suppose que la recombinaison et la mutation sont les seuls facteurs causaux des patrons de DL (Li et Stephens 2003; Auton et McVean 2007). Selon ces modèles simples, la variation du taux historique entre deux intervalles du génome ne peut donc être due qu'à une variation du taux de recombinaison sous-jacent. Or, toutes les forces évolutives sont susceptibles d'affecter les patrons de DL : dérive, migration et sélection (Chan et al. 2012; Baird 2015; Dapper et Payseur 2018). Dans une population théorique de Wright-Fisher, qui implique notamment la panmixie, le taux de recombinaison historique est ρ= 4Nec, ou Ne est la taille efficace de la population et c le taux de recombinaison par génération. Si le modèle de Wright-Fisher n'est pas adapté au blé tendre, espèce autogame qui ne se reproduit pas en panmixie, le taux de recombinaison historique ρ reste néanmoins sensible aux facteurs qui modifient la taille efficace Ne d'une population, c'est-à-dire les forces évolutives et les évènements démographiques (goulot d'étranglement, expansion). Cependant, la dérive et la migration ainsi que les évènements démographiques affectent le génome tout entier, et donc ne sont pas sources de variation entre les taux de recombinaison historique de deux intervalles. Par contre, la sélection est la seule force évolutive susceptible d'affecter localement les patrons de DL. On peut notamment citer les balayages sélectifs (Kim et Nielsen 2004), où une mutation causale confère un avantage sélectif, ou bien la sélection d'arrière-plan, où une mutation délétère est éliminée de la population. L'impact de la sélection sur le DL dépend de l'âge de la mutation. Par exemple, dans le cas d'un balayage sélectif, il faut distinguer les balayages sélectifs forts (hard sweeps) et faibles (soft sweeps). Dans le cas d'un hard sweep, une mutation avantageuse apparait dans la population, et fait augmenter la fréquence des haplotypes liés à cette mutation dans la population, ce qui augmente ainsi le DL. Dans le cas d'un soft sweep, une mutation ancienne devient soudainement avantageuse dans la population. Cette mutation étant déjà associée à une variété d'haplotypes, l'impact sur le DL est faible (Baird 2015). Pour résumer, les hard sweeps sont particulièrement susceptibles d'affecter les patrons de DL.

Une façon de s'assurer de la robustesse des profils de recombinaison historiques aux forces évolutives est de les comparer à des cartes génétiques classiques obtenues en comptant les CO dans une population expérimentale en ségrégation (Coop et al. 2008; Kong et al. 2010; Rodgers-Melnick et al. 2015; Singhal et al. 2015; Darrier et al. 2017; Petit et al. 2017; Marand et al. 2019; Fuentes et al. 2021). Pour obtenir une carte de recombinaison méiotique, nous avons utilisé une population de 406 RILs (F6) dérivées du croisement entre les variétés Chinese Spring et Renan (CsRe). Le nombre total de méioses échantillonnées dans cette population étant faible (moins de 5 000 au total), les estimateurs du taux de recombinaison par maximum de vraisemblance varient dans une gamme de valeurs peu crédibles (de 0 cM/Mb jusqu'à plus de 2 000 CM/Mb). Pour améliorer la robustesse des estimateurs du taux de recombinaison méiotique, nous avons utilisé un modèle Bayésien Poisson Gamma pour modéliser la distribution du taux de recombinaison de chaque intervalle. Ce modèle implique notamment de définir une distribution *a priori* sur le taux de recombinaison. Dans notre cas, les paramètres de la distribution *a priori* correspondent à la moyenne et à la variance de la distribution des taux de recombinaison estimés par maximum de vraisemblance. Cette distribution à priori a donc eu pour conséquence d'utiliser un taux de recombinaison moyenné sur l'ensemble des intervalles pour « corriger » l'estimateur de taux de recombinaison dans des intervalles où le taux de recombinaison est difficile à obtenir avec une taille d'échantillon limitée (petits intervalles, faible taux de recombinaison).

Globalement, les profils historiques et méiotiques montrent une corrélation de 0.6 à l'échelle de 4Mb, en moyenne pour chacune des régions de chaque chromosome (centromères exclus). Une région en particulier (7DR3) montre une corrélation négative, due à la présence d'une introgression d'un segment génomique d'une espèce apparentée (*Ae. ventricosa*) chez Renan, qui empêche la formation de CO. A noter que même si le profil historique était dépourvu de toute influence des forces évolutives, nous n'attendrions pas une corrélation parfaite entre les taux de recombinaison historiques et méiotiques. D'une part, la comparaison des profils de recombinaison historiques des quatre populations divergentes a permis de montrer de la variabilité dans le profil de recombinaison. D'autre part, les estimateurs du taux de recombinaison sont issus de deux méthodes très différentes et ont chacun leur erreur d'estimation. Chez les oiseaux, où le taux de recombinaison est décrit comme particulièrement stable à l'échelle évolutive, la corrélation entre carte méiotique et historique à l'échelle de 5Mb vaut 0.9 (Singhal et al. 2015).

Nous avons aussi comparé les profils de recombinaison historiques et méiotiques à proximité de gènes impliqués dans la domestication ou l'amélioration agronomique du blé tendre. Cependant, nous n'avons pas pu mettre en évidence de déviation anormale du profil historique et du profil méiotique dans ces zones, ce qui suggère que les profils de recombinaison historiques sont relativement robustes aux forces évolutives. Cette comparaison des profils historiques et méiotiques a été plus fructueuse pour une étude portant sur les moutons Lacaune (Petit et al. 2017). Les auteurs ont identifié 10 intervalles où le ratio entre le taux de recombinaison historique et le taux de recombinaison méiotique est significativement bas ou élevé par rapport au reste du

génome. Parmi ces 10 intervalles, 4 d'entre eux sont associés à des gènes sous sélection (gène ABCG2 impliqué dans la production laitière, LCORL impliqué dans la détermination de la taille, RXFP2 associé aux phénotypes de corne, ASIP associé à la couleur de robe). Reed and Tishkoff (2006) ont estimé que la persistance dans le temps des patrons de DL engendrée par le biais sélectif est du même ordre que la taille efficace de la population. Ainsi, une population de petite taille efficace telle que le blé tendre ne montrerait peut-être pas ce type de patron à proximité des loci sous sélection.

Afin de limiter l'impact des forces évolutives sur les variations du profil de recombinaison historique, nous avons utilisé la déviation locale du taux de recombinaison historique de chaque intervalle (λ) par rapport au taux de recombinaison historique de base estimé sur une fenêtre de 2 cM. Ces déviations sont supposées être majoritairement gouvernées par une inflation ou une déflation du taux de recombinaison dans l'intervalle et pas par une variation des forces évolutives ou des effets démographiques. Nous avons utilisé ces déviations pour détecter environ 8 000 points chauds de recombinaison ($\lambda \geq 4$) et montré que des populations génétiquement plus proches partageaient plus de points chauds. Cependant, Dapper et Payseur (2018) ont montré que la puissance de détection des points chauds ainsi que le nombre de points chauds faux positifs dépendent des évènements démographiques rencontrés par la population au cours de son histoire évolutive. Ainsi, nous avons aussi mesuré la variabilité du profil de recombinaison entre populations grâce à la corrélation des profils de déviation sur l'ensemble du génome, sans appliquer de seuil de détection. Dans la plupart des régions chromosomiques (1AR1…7DR3), la relation entre la corrélation du profil et la divergence génétique est significativement négative. Cela signifie que non seulement le profil de recombinaison présente de la variabilité, mais que cette différence est accentuée par la différenciation génétique.

II.3.3 Hypothèses pour expliquer la variabilité du profil de recombinaison

Nous avons émis trois hypothèses pour expliquer la différenciation croissante du profil de recombinaison avec la divergence génétique.

La première hypothèse est que cette différenciation résulte d'effets environnementaux. En effet, plusieurs études suggèrent que les conditions environnementales ont un impact sur le taux de recombinaison. Par exemple, Dreissig et al. (2019) montrent une association entre le taux de recombinaison moyen et différents facteurs environnementaux tels que la luminosité ou la température. Par ailleurs, plusieurs auteurs (Loidl 1989; Higgins et al. 2012; Phillips et al. 2015; Lloyd et al. 2018) montrent qu'une augmentation de la température entraine une perturbation de la méiose et modifie le profil de recombinaison. Cette hypothèse semble la moins vraisemblable car les aires de répartition des populations sont très grandes, et les conditions pédo-climatiques sont vraisemblablement très différentes à l'intérieur de chaque population et décorrélées entre populations géographiquement proches.

La deuxième hypothèse est que la différenciation du profil s'explique par des variations structurales de la séquence ADN sous-jacente. Le génome du blé tendre comprend une part importante d'introgressions génomiques dérivées d'espèces apparentées (Cheng et al. 2019; He et al. 2019; Zhou et al. 2020). Les variations structurales sont souvent associées à une perturbation du profil de recombinaison. Par ailleurs, les SNP utilisés dans notre article ont été positionnés sur l'assemblage IWGSC RefSeq V1.0 construit à partir du séquençage de la variété Chinese Spring (IWGSC 32018). La présence d'introgressions chez les populations peut conduire à une séquence différente de celle de Chinese Spring. Ainsi, l'ordre, la position et la distance physique associés à chaque marqueur ne sont peut-être pas totalement transposables entre les populations. Or ces paramètres sont déterminants dans l'estimation du taux de recombinaison historique (Li et Stephens 2003). Un séquençage complet de plusieurs accessions d'origines différentes permettrait d'apprécier la conservation de la séquence physique entre les quatre groupes génétiques.

La dernière hypothèse suppose que la différenciation du profil de recombinaison résulte de la modification du paysage chromatinien lors de la méiose. Le profil chromatinien est déterminé par l'état de compaction de la chromatine ainsi qu'un ensemble de marques épigénétiques avec éventuellement des fonctions inhibitrices ou activatrices de l'expression des gènes. De nombreuses études rapportent une corrélation entre la position des points chauds et un statut de la chromatine ouvert et présentant des marques épigénétiques favorables à l'expression des gènes (Auton et al. 2013; Choi et al. 2013a; Choi et Henderson 2015a; Marand et al. 2017; Underwood et al. 2018; Marand et al. 2019). Cette variation du statut chromatinien entre populations peut résulter d'un phénomène de dérive, ou bien de processus sélectifs qui peuvent entrainer l'activation ou l'inhibition de gènes via le dépôt de marques épigénétiques (Roth et al. 2001; Saksouk, et al. 2015; Howe et al. 2017). Il est aussi possible que la variation du profil chromatinien soit déterminée par des variants génétiques. Par exemple, chez certains mammifères, les points chauds de recombinaison sont déterminés par une protéine à doigts de zinc (PRDM9) qui dépose une marque de méthylation à proximité de séquences cibles d'ADN. Ces marqueurs épigénétiques entrainent l'intervention du cortège protéique impliqué dans la formation des CO (Baudat et al. 2010). La protéine PRDM9 montre une évolution diversifiante relativement rapide (Oliver et al. 2009). En conséquence, les différents allèles de PRDM9 ne ciblent pas les mêmes motifs de séquence ADN, ce qui conduit à observer une grande variabilité dans le profil de recombinaison (Myers et al. 2010).

Nous avons testé la différenciation génétique d'un échantillon de 54 gènes impliqués dans la méiose chez le blé tendre. Le locus ASY4 dans la région R3 du chromosome 4A est significativement plus différencié entre les 4 populations que le reste du génome, ce qui suggère une évolution particulièrement rapide de ce locus. Chez Arabidopsis, la protéine ASY4 est impliquée dans la formation du complexe synaptonémal. Les mutants montrent effectivement une modification de leur profil de recombinaison, avec une augmentation des taux de recombinaison aux télomères (Chambon et al. 2018).

Chapitre III :

Comparaison des bénéfices de plusieurs critères de sélection de croisements dans un programme de sélection de blé tendre d'hiver

Chapitre III : Comparaison des bénéfices de plusieurs critères de sélection de croisements dans un programme de sélection de blé tendre d'hiver

III.1 Préambule

Ce chapitre est présenté sous la forme d'un article scientifique en cours de préparation. Cet article n'a pas encore été relu par les pairs.

L'objectif est de comparer les bénéfices de plusieurs critères de choix de croisements décrits dans la littérature, basés ou non sur la distribution de la valeur génétique de la descendance, dans le cadre d'un programme de sélection de blé tendre simulé. Nous proposons aussi un nouveau critère qui classe les croisements selon leur proportion de descendants supérieurs à la valeur génétique de la meilleure lignée parentale du programme de sélection.

Nous avons simulé un programme de sélection à partir de 835 génotypes de lignées élites issues des programmes de sélection INRAE et Agri-Obtentions. Le déterminisme du trait est composé de 300 QTLs aléatoirement choisis parmi les 16k marqueurs.

La comparaison des critères est opérée dans plusieurs scénarios. Une première catégorie de scénarios porte sur la précision des estimations des effets aux marqueurs. L'espérance et la variance de la descendance sont calculées à partir des effets QTLs ou à partir des effets des marqueurs estimés par prédiction génomique. Une deuxième catégorie de scénarios consiste à tester l'impact de la composition génétique de la population parentale. Les lignées parentales peuvent n'avoir jamais été sélectionnées sur le trait, ou au contraire être issues d'une sélection récurrente depuis plusieurs générations.

Les plans de croisements ont été optimisés sur chaque critère de sélection de croisements, en prenant en compte des contraintes de diversité simples sur les croisements et les effectifs des descendances, inspirées des pratiques d'un programme de sélection réel.

Les descendants RILs de chaque plan de croisements sont simulés. La distribution des valeurs génétiques des descendants est comparée à deux niveaux : 1) la proportion de descendants supérieurs à la meilleure lignée du programme de sélection, supposés être des bons candidats pour l'inscription de variétés élites. 2) la valeur génétique d'une fraction supérieure des descendants est mesurée, supposée être utilisés comme nouvelles lignées parentales au cycle suivant.

Nous avons mesuré la différence de recrutement des parents et des couples entre les différents critères, en comparent les classements des croisements, la similarité génétique des parents et des couples et la diversité génétique dans la descendance sélectionnée pour le cycle suivant.

III.2 Article "Comparison of cross-selection criteria to optimize mating plans in a winter bread wheat breeding program" (in prep)

This article is a draft and has not yet been peer-reviewed.

- **Abstract**

A crucial step in plant breeding is the choice of the mating design to both derive highly performing varieties and also maintain a competitive breeding population to obtain genetic gain in next generations. Traditionally, the mating plan relies on crosses involving the best parental lines, to ensure a high mean in progeny performance. Henceforth, genomic prediction models allow to predict progeny distribution and thus to select crosses based on the predicted value of their superior progeny. To our knowledge, this is the first comparison of cross-selection criteria (CSC) in terms of genetic gain and genetic diversity in a winter bread wheat breeding program aiming to both derive highly performing genotypes and maintain genetic gain at next generations. This necessitated to define constraints on parental contributions to ensure sufficient genetic diversity. In our study, we compared benefits of five different CSC described in literature, some of them based on progeny distribution estimation, in a simulated winter bread wheat breeding program, with a starting population containing 835 elite genotypes from the French INRAE and Agri-Obtentions breeding program. We also propose a new CSC based on the proportion of progeny superior to the best parental line of the breeding population. Mating plans were optimized for each CSC, taking into account constraints on crosses and parental contributions inspired from real bread wheat breeding program practices. Our results show that CSC based on progeny distribution provides superior benefits for the proportion of transgressive progeny, but also the genetic gain and genetic diversity in the selected population when markers effects estimation is accurate. In that case, CSC based on progeny distribution increase the genetic value of superior progenies up to 3-5% compared to CSC based on parental genetic values only. These conclusions for different scenarios and the open-source pipeline to optimize mating plans can be useful for breeders.

❖ Introduction

Plant breeders have two main objectives: derive highly performing varieties at each cycle and improve the mean genetic value of their germplasm in order to provide superior varieties in future generations. The design of the mating plan, *i.e.,* the choice of parental lines to cross and the progeny size from each cross, is critical to maintain both short and long-term genetic gain. However, the number of candidate crosses is putatively very high while the number of crosses and progenies that can be tested is often limited.

To maximize the genetic value of the (top) progeny, breeders can rank crosses according to cross-selection criteria (CSC) that estimate the ability of each cross to produce superior progeny for a

trait of interest. The expected mean genetic value of the progeny is a simple criterion, which can be estimated by the mean additive genetic value of the parental lines (criteria named PM for Parental Mean) (Jinks et Pooni 1976). However, the PM criteria does not provide any information about the variance associated with the progeny of a cross, and thus its potential to contribute to genetic gain by producing transgressive progenies, *i.e.,* superior to the best parent. The progeny variance associated with each cross depends on the complementarity of favorable alleles in parents and their probability to recombine during meiosis (Zhong et Jannink 2007).

Traditionally, breeding programs estimate progeny genetic value from phenotypic observations. More recently, Genomic Prediction (GS) methods were developed to estimate progeny genetic values from their genotype (Genomic Estimated Breeding value, GEBV). GS uses a training population which is phenotyped and genotyped to estimate the effects of the segregating genomic variants (markers). Assuming additivity of marker effects, the GEBV of one individual is the sum of its allele effects at every marker. Compared to phenotyping, GS can reduce generation interval in crops by using rapid cycles (two or three per year depending on the species) in greenhouses and replacing phenotyping by genotyping. Depending on the species and the quality of the training population used to build the prediction model, it could also increase the accuracy of predictions and selection intensity.

With GS, the value (usefulness) of a cross can be estimated using parental genotypes and marker effects. Daetwyler et al. (2015) defined a CSC named Optimal Haploid Value (OHV) corresponding to the genetic value of the best theoretical gamete to pass on to the next generation. They defined haplotypic blocks along the genome, considering as an approximation that recombination only occurs between blocks. Note however that the probability to get this best gamete is very low. The correlation of this CSC with the genetic value of the best observed progenies will highly depend on progeny sizes and the number of haplotypic blocks. It is fast to implement and has been shown to provide superior genetic gain, compared to selection based on PM (Daetwyler et al. 2015; Lehermeier et al. 2017), but it may not be the most suitable CSC to maximize short term genetic gain. Another strategy is to predict the distribution of progeny breeding values and exploit the properties of this distribution to calculate a CSC. For a trait determined by a very high number of variants with small effects, the distribution of progeny breeding values is expected to be Gaussian. It is centered on the expected mean of the progeny, which can be estimated from the mean of additive parental genetic values, either estimated from phenotypes or GS (expected mean of progeny is thus called PM for Parental Mean). However, the expected variance is much more difficult to predict and the recent methodologies (Bernardo and Charcosset 2006; Zhong and Jannink 2007; Lehermeier et al. 2017; Allier et al. 2019a; Santos et al. 2019) rely on estimated marker effects. The most recent formula to predict RILs progeny variance derived from a cross between two inbred lines was provided by (Lehermeier et al. 2017). Formulas were also derived to estimate three and four-way cross variances (Allier et al. 2019a) and to predict gametic variability in an animal breeding context (Santos et al. 2019). Formulas explicitly include the vectors of

recombination rate between markers that are polymorphic between parents, marker effects and linkage phase between every pair of alleles. Indeed, when alleles are in coupling phase (*i.e.,* one parent carries the two beneficial alleles while the other carries deleterious alleles), recombination decreases progeny variance, while in repulsion phase, recombination increases progeny variance. Moreover, progeny variance increases with polymorphism between parents. Alternatively, progeny distribution can be estimated by simulating progeny in silico (stochastic simulations) based on parental genotypes by randomly producing crossing-over along chromosomes according to a recombination map (Bernardo and Charcosset 2006, Mohammadi et al. 2015).

Once the distribution of progeny is predicted or obtained by simulation, many CSC can be suggested, depending on the breeding target. Schnell and Utz (1975) suggested to rank crosses based on the expected mean of an upper fraction q of their progeny. This CSC was named Usefulness Criterion (UC) and can be computed from UC = PM + i*h*σ where i is the selection intensity corresponding to the fraction q of selected progenies, h is the square root of heritability and σ is the parental genetic standard deviation. Considering that progeny size is generally limited, another CSC named EMBV was suggested by Müller et al. (2018). EMBV predicts the value of a cross as the expected mean of the K top progenies among D allocated to the cross. Wellmann (2019) and Bijma et al. (2020) suggested to compute the value of a cross as the probability to produce a progeny superior to a given threshold. This threshold can be extrapolated from historical genetic gains observed in the breeding program (Wellmann 2019) or it can correspond to the superior quantile q of the predicted distribution of the progeny at the following generation (Bijma et al. 2020). It can also simply be the genetic value of the best parental line.

Several studies compared the efficiency of those CSC on the short-term selection response (Zhong et Jannink 2007; Lehermeier et al. 2017; Yao et al. 2018; Bijma et al. 2020). They showed that CSC based on progeny variance (or gametic variance) estimation can actually increase genetic gain, even if parental genetic values and progeny variance are not accurately estimated. Zhong and Jannink (2007) and Bijma et al. (2020) showed that the relative benefits of CSC based on progeny variance estimation depends on the ratio between the variance of progeny standard deviations var(σ) and the variance of progeny means var(PM) in the list of candidate crosses. When the variance of progeny means var(PM) among crosses is highly superior to the variance of progeny standard deviations var(σ), PM alone is enough to predict the rank of crosses in that case.

According to the breeder's equation, genetic gain is proportional to genetic diversity and to the selection intensity in the breeding program (Falconer et Mackay 1966). In a closed breeding program, i.e with no importation of external genitors, the more efficient the selection, the faster the diversity decreases. So, another objective of the mating plan's design is to maintain sufficient genetic diversity to ensure long-term genetic gain. Breeders empirically avoid crossing the most related genitors (Wartha et Lorenz 2021) and ensure that a minimum number of parental lines contribute to the next generation. Several more advanced methods were designed to balance

expected genetic gain and expected genetic diversity at following generations when selecting genitors and/or crosses (Toro et Perez-Enciso 1990; Meuwissen 1997; Jannink et al. 2010; Akdemir et al. 2019; Allier et al. 2019a). In any case, the desirable balance between expected genetic gain and expected genetic diversity is not trivial to define. It depends on the objective (short or long-term genetic gain in a breeding or pre-breeding program).

Thus, mating plans should not only be optimized to maximize a CSC, but also include diversity constraints. From a computational point of view, optimization problems usually involve variables to adjust (progeny sizes of each candidate cross in our case), an objective function to maximize (the sum of the products of CSC values by progeny sizes in our case) whose numeric value depends on the variables to adjust and also constraints on variables to adjust. When the equation system is linear for the variables to adjust, one can use linear programming to find the set (s) of variables that maximize the objective function. Otherwise, for more complex problems, one can use heuristic algorithms such as Genetic Algorithms (GA) to obtain a good (but not necessarily the best) solution to the problem.

To our knowledge, no study compared the relative benefits of CSC within a European winter bread wheat breeding program. The objective of this article was to compare several CSC in terms of production of transgressive lines for commercial purpose, but also genetic gain and diversity in the selected progeny after mating plan optimization and one generation of selection. The starting breeding population included 835 genotyped parental lines from the INRAE-AO breeding program. We tested several CSC from the literature (PM, OHV, EMBV, UC) and adapted a new one from Bijma et al. (2020) and Wellmann (2019). This last one, named PROBA, consists in ranking crosses based on the expected proportion of progeny superior to the best breeding line of the breeding program. We compared genetic gain and diversity levels in the selected progeny, when QTLs effects and positions are supposed known, and also when marker effects are estimated by a GBLUP model using observed parental phenotypes. Diversity constraints on parental contributions, *i.e.* minimal and maximal number of parents, crosses and progenies were chosen according to breeding companies' practices.

❖ Material and Methods

Benefits of CSC were compared using different simulated scenarios and one cycle of selection. The selection cycle started either from a population of 835 lines from INRAE-AO breeding program or from a derived population obtained after three cycles of selection of the initial population for the simulated trait. CSC were computed for all possible crosses. In order to attribute an optimal number of progenies to all crosses, the mating plan was optimized using a linear algorithm for most CSC. Corresponding F5 RILs were simulated according to a reference recombination map (from Danguy des Déserts et al. 2021). CSC were ranked according to their ability to provide competitive commercial varieties, *i.e* the proportion of progenies superior to the best parental line. To evaluate

the relevance of each CSC for recurrent selection, we also compared genetic gain and genetic diversity of the 7% best progenies that theoretically become the putative parents of the next breeding cycle. The procedure is summarized in the **supplementary Figure S1** (see appendices of this thesis). The following sections describe the different simulated scenarios and the optimization process.

- ***Initial genotypes***

The parental genotypes used to simulate progenies were 835 F_8-F_9 winter-type bread wheat lines developed and phenotyped between 2000 and 2016 by breeders from Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) and its subsidiary breeding company Agri-Obtentions (AO) (Ben Sadoun et al. 2020).

They were genotyped with 35K SNPs (Ben Sadoun et al. 2020) representative of the TaBW280K array (Rimbert et al. 2018). For this analysis, the markers were filtered by missing data rate (< 5%), heterozygosity rate (< 5%) and minor allele frequency (> 10%) yielding 16 429 SNPs. Missing genotypes were imputed using the Beagle v4.1 software (Browning et Browning 2007; Browning et Browning 2016); implemented in the R-package synbreed, Wimmer et al. 2012).

The genetic values for yield of these 835 lines were estimated by a GBLUP model. The variance of GEBVs for grain yield was equal to 14 dt/ha (with 1dt = 0.1 ton) and was used as a reference genetic variance in phenotype simulations.

- ***Simulation of populations that have never been selected for a trait of interest***

Using starting genotypes of the 835 breeding lines from INRAE/AO breeding program, we simulated 20 genetic architectures of a trait controlled by 300 QTLs randomly picked among the 16k SNPs, with effects drawn from a Gaussian law N(0,1). The favorable allele was attributed at random to one of the two SNP alleles, so that coupling and repulsion associations also occur at random. QTLs effects were adjusted to provide a variance of True Breeding values (TBV) of 14, as the parental GEBVs obtained using experimental data. TBV were calculated as the cross product between QTLs effects and allelic doses.

The first-generation heritability (h^2_0) was set to 0.4. The lines phenotypic values were obtained by adding a normally distributed noise to the TBV. The corresponding environmental variance was 21 (h^2 = 14/(14+21) = 0.4). Using the 20 simulated genetic architectures, two categories of scenarios have been compared. One category of scenarios called TRUE (n=20) in which QTLs effects are known (and so the TBV), and one category called ESTIMATED (n=20) where marker effects were estimated by a GBLUP model from simulated phenotypes and GEBV were computed from estimated marker effects as the cross product between estimated marker effects and allelic doses.

These populations made of real genotypes but simulated genetic architecture are called “unselected populations”

- ***Simulation of a population under selection***

Due to the Bulmer effect (Bulmer 1971) selected populations should present negative covariances between QTLs. In our “unselected population” simulations, this phenomenon is not taken into account as QTLs and effects were randomly sampled along the genome. To take into account the Bulmer effect, we derived “selected populations” from three cycles of truncation selection from the 835 starting genotypes. For each TRUE scenario (n=20), these three preliminary selection cycles used TBV to select superior lines. For each ESTIMATED scenario (n=20) selection used phenotypic values to select superior lines. Phenotypic values of new lines were obtained by adding a normally distributed noise of variance 21 to the TBV.

In the first cycle, 300 crosses were made at random from the 300 best lines. Each cross produced 11 F5 RILs (total progeny = 3 300), simulated with the R package MOBPS (Pook et al. 2020). Only one progeny per cross was selected based on TBV (TRUE scenarios) or phenotypes (ESTIMATED scenarios). Cycles 2 and 3 started by the random mating of the 300 selected progenies in the previous cycle. In the 3rd cycle, the three best progenies per cross were kept leading to a final population of 900 parental lines called “selected population”. Note that the preliminary selection rate of 900/3300 at the 3rd cycle forms a pre-selection of parental lines. The pedigree was recorded at each selection cycle and used to compute probability of Identity by Descent (IBD) with the R package synbreed (Wimmer et al. 2012).

To sum up, we simulated 20 different genetic architectures * 2 levels of prediction accuracy (TRUE and ESTIMATED) * 2 types of populations (unselected populations, made of the 835 starting genotypes, and selected populations, made of 900 genotypes derived from selection), for a total of 80 simulations. Mating plans were optimized from this n=80 simulated populations. Results for the “selected populations” and TRUE scenarios are presented in the main document, the others in supplemental material.

- ***Estimation of genetic values and marker effects***

For the ESTIMATED scenarios, we used a GBLUP model to estimate parental lines genetic values and marker effects, following the model:

$$Y_i = \mu + \alpha_i + e_i$$

where i denotes the name of the parental line (n=835 parental lines in “unselected populations” and INRAE/AO dataset and n=900 in the “selected populations”), Y is the vector of phenotypes, μ is the average phenotype, α is the vector of genetic values and e is the vector of residual effects. The genetic values were supposed to follow N(0, $G^{(1)}\sigma_a^2$), where $G^{(1)}$ is the genomic relationship matrix computed as ZZ’/2$\sum_l p_l(1-p_l)$, Z being the centered genotyping matrix and p_k the allelic frequency at locus l. Residuals effects were supposed to follow N(0, $I\sigma_e$). The parameters were estimated

with the AIREMLf90 software (Misztal 2008). Marker effects were estimated by back-solving the GBLUP model using PostGSf90 software (Wang et al. 2012; Aguilar et al. 2014). Note that QTLs were removed from the genotyping matrix before estimating marker effects.

- ***Prediction of progeny variance***

The expected variance of progeny was computed using Lehermeier et al. (2017) formula for biparental RILs progeny self-fecundated for 4 generations (F5 RILs). For each cross between P_i*P_j the formula for expected variance of progeny was

$$\sigma^{2\,RILs\,F5}_{i,j} = 4 * (\sum_{l=1}^{L} \beta_l^2\, p_l(1-p_l) + 2\sum_{k<l} \beta_k \beta_l 4D_{kl}\,(1-2r_{kl}{}^5 - (0.5(1\text{-}2r_{kl}))^5))$$

Where β are either QTLs effects for TRUE scenarios (length of β = 300), either estimated marker effects for ESTIMATED scenarios (length of β = 16 429 - 300), p_l is the allelic frequency at locus l (either 0 if parents are monomorphic at this locus, either 0.5 if parents are polymorphic), D_{kl} is the linkage disequilibrium (LD) between alleles at loci l and k (either 0 if parents are monomorphic at locus l or at locus k, or – 0.25 if alleles are in repulsion phase or 0.25 if alleles are in coupling phase) and r_{kl} is the expected proportion of recombinant progenies between locus l and k compared to parental haplotypes. The expected proportion of recombinants was computed from the Western European recombination map published by Danguy des Déserts et al. (2021), using Haldane mapping function (Haldane et Waddington 1931): r_{kl} = 0.5*(1 − $e^{-2d_{kl}}$) where d_{kl} is the genetic distance (in Morgans) between loci k and l (Haldane, 1919).

Estimation of progeny variance for a high number of crosses (~400k in our study) and a high number of simulations (n=80) is highly time consuming. So we speeded up this estimation by identifying repetitive or null terms across the 400k crosses, as described in **supplementary Protocol S1**.

- ***Mating plan constraints***

Constraints on mating plans were inspired from the bread wheat breeding program of the private company Florimond Desprez (personal communication).

A mating plan is defined by a vector giving the number of progenies D_{ij} allocated to each possible cross between parents P_i*P_j. The constraints were defined as follow:

- C1: The total number of progenies has to be equal to D = 3 300
- C2: The number of progenies allocated to a cross can vary between D_{min}= 5 and D_{max} = 60
- C3: The number of crosses can vary between K_{min}=200 and K_{max}= 300
- C4: The number of progenies derived from one parent cannot exceed C_{max} = 250
- C5: The number of recruited parents can vary between P_{min}=100 and P_{max} = 132
- C6: Too much related parental lines cannot be crossed. In the “selected populations”, Crosses involving a pair of parental lines with an IBD ≥ 0.25 from pedigree data were removed from the

list of candidate crosses, which actually represented ~1% of the candidate crosses. In the "unselected populations", we used the LDAK software (Speed et al. 2012) to obtain a genomic relationship matrix $G^{(2)}$ in which SNPs are weighted according to local LD. We used LDAK software because LD is very heterogenous in bread wheat, strongly increasing from telomeres to centromeres. This variance-covariance matrix was computed as WW' where W was obtained by centering and scaling each column of the genotyping matrix Z such as $W_i = w_i * (Z_i - p_i)/\sqrt{p_i(1-p_i)}$ where p_i is the allelic frequency of the marker i and w_i is the weight estimated by LDAK software considering local intensity of LD. Crosses involving a pair of parental lines showing a covariance superior to the value of the 99% quantile of covariances were removed from the list of candidate crosses (1% of the candidate crosses).

We compared scenarios with and without constraints, respectively called "CONSTRAINTS" and "NO CONSTRAINTS". In the "NO CONSTRAINTS" scenarios, only constraints C1 and C2 have been considered. Note that parental lines and estimates of marker effects and GEBV are the same in CONSTRAINTS and NO CONSTRAINTS scenario. All results are given for CONSTRAINTS scenarios in the main document, and NO CONSTRAINTS in supplemental material.

- ***CSC and their corresponding objective function***

One mating plan is defined by a set of crosses and their respective number of progenies. For each CSC, the mating plan maximized an objective function specific to the CSC, under the constraints C1 to C6 for the "CONSTRAINTS" scenarios, C1 to C2 for the "NO CONSTRAINTS" scenarios.

CSC n°1: PM

For each cross $P_i * P_j$, the usefulness is the predicted expected mean of progeny, computed as

$$\boldsymbol{PM_{ij}} = \frac{\boldsymbol{\alpha_i + \alpha_j}}{\mathbf{2}}$$

Where α are either TBV in TRUE scenarios or GEBV in ESTIMATED scenarios. The objective function to maximize is

$$\boldsymbol{F_1 = \sum_{i,j} D_{ij} * PM_{ij}}$$

CSC n°2: $UC1$

UC stands for Usefulness Criterion. This CSC is the expected mean of the 7% best progeny of a cross computed as

$$\boldsymbol{UC1_{ij} = PM_{ij} + i^{q=7\%} * \sigma_{ij}}$$

Where $i^{q=7\%}$ ~ 1.91 is the selection intensity corresponding to a selection rate of 7% and σ_{ij} is the expected standard deviation of progeny. Note that the progeny standard deviation σ_{ij} can be

computed either with QTLs effects in TRUE scenarios or estimated allelic effects in ESTIMATED scenario. The objective function to maximize is

$$\boldsymbol{F_2 = \sum_{i,j} D_{ij} * UC1_{ij}}$$

CSC n°3: UC2

This CSC is the expected mean of the 0.01% best progeny of a cross computed as ;

$$\boldsymbol{UC2_{ij} = PM_{ij} + i^{q=0.01\%} * \sigma_{ij}}$$

Where $i^{q=0.1\%} \sim 4$ is the selection intensity corresponding to a selection rate of 0.1%, so twice the selection intensity of the UC1 criteria. Although this selection rate of 0.1% is not realistic considering the small progeny size (Dmax = 60 per cross), the objective is to select crosses with higher expected genetic variance compared to the UC1 criteria. The corresponding objective function to maximize is:

$$\boldsymbol{F_3 = \sum_{i,j} D_{ij} * UC2_{ij}}$$

CSC n°4: EMBV

The expected value of the best progeny among D_{ij} allocated to a cross is

$$\boldsymbol{EMBV_{ij} = PM_{ij} + INT^{1/D_{ij}} * \sigma_{ij}}$$

With $INT^{1/D_{ij}}$ is the expected value of the highest order statistic among a sample of D_{ij} statistics drawn from a N(0,1). An approximation of $INT^{1/D_{ij}}$ is given by (Burrows 1972):

$$INT^{N/M} = i^{\mathrm{q=\,N/M}} - \frac{(M-N)*q}{2N(M+1)*f(y_q)}$$

Where f is the density function of a Gaussian law N(0,1) and y_q the truncation threshold so that P($y \geq y_q$) = q = N/M. In our situation, N=1 and M= D_{ij} and $i^{\mathrm{q\,=N/M}} = f(y_q)/q$, so the formula of Burrows yields

$$INT^{1/D_{ij}} = i^{q_{ij}=1/D_{ij}} - \frac{D_{ij}-1}{2*(1+D_{ij})*i^{q_{ij}=1/D_{ij}}}$$

The objective function to maximize is

$$\boldsymbol{F_5 = \sum_{i,j} D_{ij} * EMBV_{ij}(D_{ij})}$$

CSC n°5: PROBA

This criterion ranks crosses based on the expected proportion of progeny exceeding a threshold λ, as suggested by Wellmann (2019) and Bijma et al. (2020). We defined λ as the genetic value (either TBV in TRUE scenarios or GEBV in ESTIMATED scenarios) of the best parental line. The probability for a cross to produce a progeny with genetic value superior to λ is $\omega_{i,j}$ = 1- F_{ij}(x≤ λ) with F being the cumulative distribution function for the cross i*j of the Gaussian law N($PM_{ij}, {\sigma_{ij}}^2$). The probability that no progeny of the cross P_i*P_j exceeds λ is $\omega_{i,j}^{D_{ij}}$. The probability that no progeny from all crosses exceeds λ is $\prod_{i,j} \omega_{i,j}^{D_{ij}}$, and so the log probability is $\sum_{i,j} D_{ij} * \log(\omega_{i,j})$. To maximize the expected number of progenies superior to λ, it is equivalent to minimize the objective function

$$\boldsymbol{F_6 = \sum_{i,j} D_{ij} * \log(\omega_{i,j})}$$

CSC n°6: OHV

Daetwyler et al. (2015) defined the Optimal Haploid Value as the value of the best inbred progeny that could be theoretically derived from a cross. For each genomic segment b, the effect of haplotypes carried by parent P_i and P_j are respectively called β_{bi} and β_{bj}. The OHV of a cross is computed as

$$\boldsymbol{OHV_{ij} = 2 * \sum_b \max(\beta_{bi}, \beta_{bj})}$$

Daetwyler et al. (2015) showed that there is an advantage of OHV compared to PM at later generations, not at the first generation. For bread wheat, they showed that one to three blocks per chromosome allowed higher genetic gain in simulations than smaller blocks. We defined three haplotypic blocks per chromosome, one block per chromosome arm plus one block for the centromere (frontiers of centromeric regions defined in Choulet et al. 2014). Distal regions of chromosomes' arms show high levels of recombination in bread wheat, while more proximal regions show lower recombination rates, and recombination is almost suppressed in centromeres (Choulet et al, 2014). Thus, recombination mostly occurs within chromosome arms and distal regions and centromeres can be seen as independently segregating parts. These three genomic segments per chromosome are thus good approximation of segregating blocks.

The objective function to maximize is

$$\boldsymbol{F_7 = \sum_{i,j} D_{ij} * OHV_{ij}}$$

Note that all these CSC (PM, UC1, UC2, PROBA, EMBV, OHV) are very accurate predictors of progeny distribution in TRUE scenarios (**supplementary Figure S2, appendices**).

- ***Optimization of mating plans***

For all CSC but EMBV, the objective function and constraints constitute a system of linear equations. We used an integer linear programming algorithm implemented in IBM ILOG CPLEX software (CPLEX Python API, IBM 2017) to maximize (or minimize) objective functions while respecting constraints. For criterion 4 (EMBV), the objective function and constraints do not form a system of linear equations, as the usefulness of a cross actually depends on the number of progenies allocated to the cross. To optimize mating plans for EMBV criterion, we used a Genetic Algorithm (GA). GAs are population-based Metaheuristics inspired by Darwinism (Goldberg 1989). The description of GA used in this study and parameters tunning are given in **supplementary Protocol S2**.

GAs are difficult to tune and get often stuck into local minima. To avoid premature convergence, a sharing process can be added before the selection (Xiaodong Yin et Germay 1993) in order to give more chance to candidates that are isolated in the search space. The sharing process requires the definition of a distance between candidate solutions. Candidate solutions were considered different if at least one D_{ij} was different.

We compared outputs of GA with or without sharing process on a subset of our data (supplementary Protocol S3). The sharing process did not improve significantly fitness of solutions, but was much more computation time-consuming. The mating plans were thus optimized without sharing.

We also tested whether a pre-selection of candidate crosses with highest PM had an influence on the value of the objective function to be maximized (**supplementary Protocol S4**). Most of the time and for all criteria, a pre-selection of the 10% highest PM crosses provided an objective function value after optimization that were 99% similar to the objective function value of the same population without pre-selection. Thus, to save computation time, we ran optimization of mating plans with the 10% highest PM crosses. Note that a pre-selection of crosses based on parental genetic values was also used in Zhong and Jannink (2007) and Bijma et al. (2020).

NO CONSTRAINTs scenarios did not need optimization for all CSC, and were thus very fast to compute. In that case, we wrote an algorithm that ranks crosses according to each CSC and allocate Dmax=60 progeny to the 55 best crosses (total = 3300 progeny).

- ***Similarities of mating plans***

We calculated the proportion of parents (or crosses) in common between mating plans obtained from each pair of CSC, divided by the total number of parents (respectively crosses) recruited. This proportion was averaged over the n=20 “selected populations” or n=20 “unselected populations”.

We also measured the genetic similarity between every pair of mating plan by computing the quantity

$$C = \frac{c_{CSC\,1} * G^{(2)} * c'_{CSC\,2}}{\sqrt{c_{CSC\,1} * G^{(2)} * c'_{CSC\,1}}\ \sqrt{c_{CSC\,1} * G^{(2)} * c'_{CSC\,2}}}$$

where $G^{(2)}$ is the genetic variance-covariance matrix of parents obtained from LDAK software (Speed et al. 2012) and $c_{CSC\,i}$ is the vector of progenies allocated to each parent in the mating plan optimized according to the CSC i. The quantity $c_{CSC\,1} * G^{(2)} * c'_{CSC\,2}$ is thus the average variances-covariance of parents recruited for two different CSC, and can be seen as a proxy of genetic similarities of parents recruited using different CSC. The quantities $c_{CSC\,1} * G^{(2)} * c'_{CSC\,1}$ and $c_{CSC\,2} * G^{(2)} * c'_{CSC\,2}$ give the genetic similarity of parents recruited within a mating plan, and are used to scale the numerator to provide a C quantity ranging between -1 and 1, of the same way on could calculate a correlation coefficient. The genetic similarity of mating plans was also averaged over the n = 20 “selected populations” and n = 20 “unselected populations”.

- ***Simulation of progenies***

The F5 RILs progenies of each mating plan were simulated using the R package MOBPS (Pook, et al. 2020). Each mating plan was simulated M=20 times, to take into account that progeny genotypes might vary due to mendelian sampling in parents and in RILs. Then, the TBV of progenies were computed as the cross product between QTLs effects and allelic dosage at QTLs loci.

- ***Ranking of CSC***

The mating plan’s optimization in this study focuses on two objectives: derive highly performing genotypes for commercial purpose, but also improve the breeding population with a limited decrease of genetic diversity.

To rank CSC’s ability to provide putative commercial varieties, we computed the mean proportion of progeny derived from the mating plan showing a TBV superior to a percentage ranging from 100% to 120% of the TBV of the best parental line of the breeding program.

To rank CSC's ability to improve the breeding population, we computed the relative increase (RI) of the progenies' mean TBV using a CSC compared to PM in a selected fraction K/D of progeny:

$$RI^{(K/D)} = \frac{\frac{1}{M}\sum_m TBV_{CSC}^{(K/D)} - \frac{1}{M}\sum_m TBV_{PM}^{(K/D)}}{\frac{1}{M}\sum_m TBV_{PM}^{(K/D)} - TBV^{parents}}$$

Where $\frac{1}{M}\sum_m TBV_{CSC}^{(K/D)}$ is the mean TBV of the K best progenies among D simulated progenies, averaged over M=20 mendelian sampling simulations. The relative increase $RI^{(K/D)}$ is then averaged over the n=20 simulated populations (either "selected populations" or "unselected populations"). The selection rate of progeny (K/D) ranged between 1/3300 and 10% (330/3300).

We especially focused on a selection rate of 7% which is a realistic selection rate a F5 stage in a bread wheat breeding program. These 7% best progenies are considered as the new breeding population derived from the mating plan. The genetic gain of this new breeding population was computed as

$$G = \frac{\frac{1}{M}\sum_m TBV_{CSC}^{(7\%)} - TBV^{parents}}{\sigma_0}$$

where σ_0 is the genetic standard deviation of the initial breeding population and is equal to $\sigma_0 = \sqrt{14}$. Then G was averaged over the n=20 simulated populations (either "selected populations" or "unselected populations").

Finally, we measured genetic diversity using genic variance computed as $\sigma_{gv}^{2} = \sum_l 4 * \beta_l^2 * p_l * (1 - p_l)$ with p_l the allelic frequency and β_l the true allelic effect of QTL at locus l. We measured the loss of genic variance in the new breeding population compared to the genic variance of parents using the metric:

$$V = \frac{\frac{1}{M}\sum_m \sigma_{gv}^{2}{}_{CSC}^{(7\%)} - \sigma_{gv}^{2}{}^{parents}}{\sigma_{gv}^{2}{}^{parents}}$$

Where $\frac{1}{M}\sum_m \sigma_{gv}^{2}{}_{CSC}^{(7\%)}$ is the average genic variance in the selected set of progenies among the M=20 mendelian simulations and $\sigma_a^{2}{}^{parents}$ is the genic variance of the parental population. Then V was averaged over the n=20 simulated populations (either "selected populations" or "unselected populations").

❖ Results

• *Rank correlation of CSC*

To better understand why CSC recruited different parents and crosses, we compared CSC values for all candidate crosses in TRUE scenarios (**Figure 1**) (see **supplementary Figure S3** for ESTIMATED scenarios and **supplementary File** to look at scatterplot relationships between CSC, in appendices of this thesis). CSC were tightly correlated to each other in both “selected” and “unselected populations”. The average Spearman rank correlation for 20 genetic architectures between PM and UC1 was ≥ 0.95 (0.95 in “selected populations”, 0.98 in “unselected populations”). The correlation between PM and UC2 was lower (≥ 0.83). This was expected as UC1 and UC2 are linear function of PM and progeny standard deviation σ, but the relative weights given to σ is higher in UC2 than in UC1 (UC1 = PM + 1.91* σ while UC2 = PM + 4* σ). UC2 would theoretically favor crosses with higher progeny variance compared to UC1. The EMBV criteria (using D_{ij} = 60 for all crosses i*j) highly correlated to UC1 (~1) and PM (>= 0.93), because the analytical formula of UC1 and EMBV are highly similar when the number of progenies per cross is fixed (EMBV = PM + 2.29 * σ). The PROBA criterion is almost null (< 1e-4), with 17% (“selected” populations) and 26% (“unselected” populations) of crosses showing a small value of PM (**supplementary File**). Taking into account only non-null values, the PROBA criterion is also tightly correlated to PM (≥ 0.77). The OHV criterion was the least correlated to PM (≥0.65) likely because OHV is the only criterion that consider only parental haplotypic complementarity and not consider the mean breeding value of parents. Overall, criteria based on the expected variance of progeny (UC1, UC2, EMBV and PROBA) were highly correlated to each other, and less correlated with criteria PM and OHV.

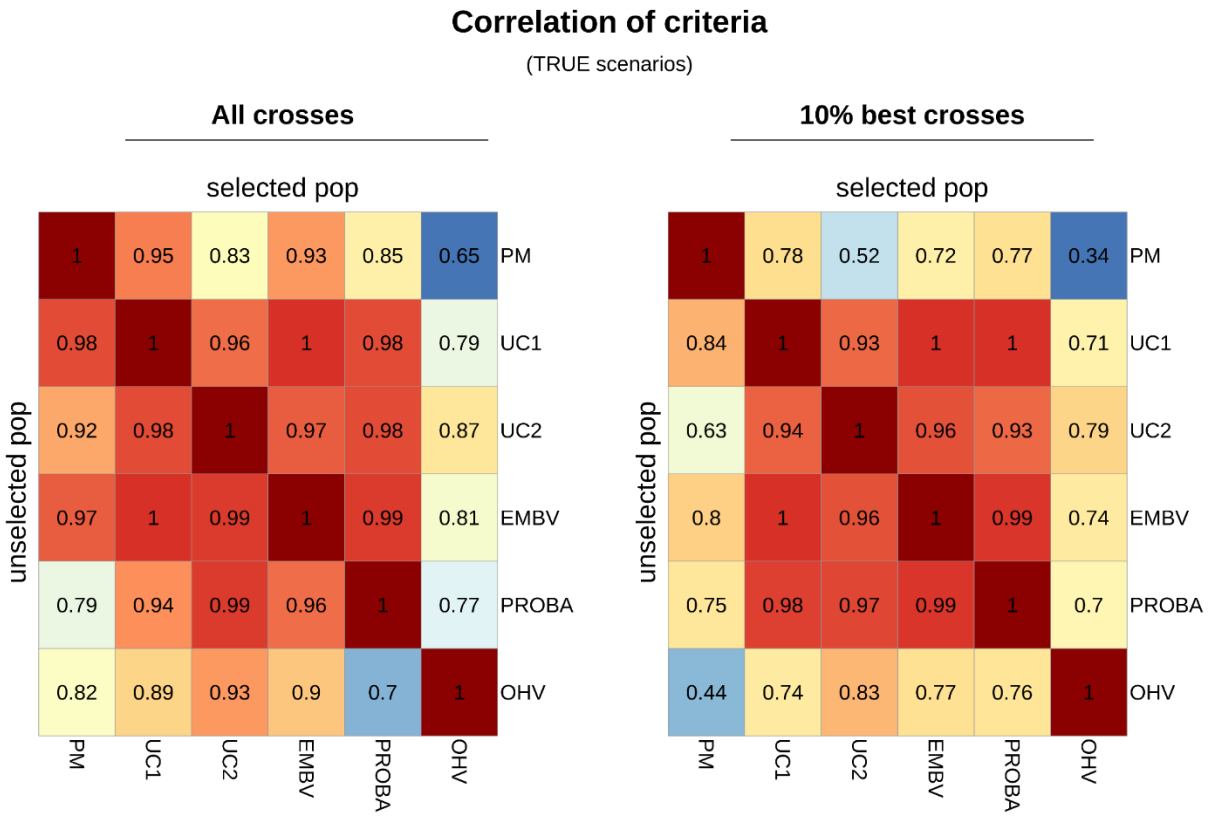


Figure 1: Spearman rank correlations between CSC.

Criteria were computed from known (simulated) QTLs effects (TRUE scenarios). The graph on the left shows correlations coefficients computed from all the candidate crosses (~400k crosses), while the graph on the right shows correlations coefficients computed for the subset of 10% crosses with highest PM (~40k crosses). Mating plans were optimized on this 10% subset. The upper diagonals give the average correlation coefficients for the "selected populations" (average on n=20 populations) while the lower diagonals give average correlation coefficients for the "unselected populations" (n=20).

To spare computation time, as the rank correlation between CSC and PM was very high, we tested the influence of pre-selection of crosses based on PM on mating plan. When pre-selecting 10% of highest PM crosses before mating plan optimization, the optimum objective functions' value were hardly not affected. The recruited parents and crosses were highly similar for all CSC but OHV and EMBV (**supplementary protocol S2**). This means that in our data, parental mean genetic value is a good proxy to roughly rank inferior and superior crosses for all CSC. Note that the rank correlations computed with the 10% highest PM crosses were slightly lower than correlations computed with all the crosses, especially for pairs of criteria involving PM or OHV (Figure 1). This means that PM and other CSC discriminate unanimously good from bad crosses but disagree about the ranking within the good crosses.

- ***Similarity of mating plans***

Mating plans were optimized using the 10% highest PM crosses. Optimization consisted in allocating the D = 3 300 progenies to crosses by maximizing one CSC while respecting constraints on parental contributions in CONSTRAINTS scenarios. Whereas the number of crosses can vary between 200 and 300, all mating plans except those using PM involved 200 crosses. Mating plans using PM recruited up to 226 crosses. The number of progenies per cross ranged from 5 to 60 in respect to constraints. Around 16% (± 2%) of crosses were attributed Dmax = 60 progenies. Whereas the algorithm allowed the number of parents to vary between 100 and 132, all mating plans but those using PM or OHV recruited 100 parents. Mating plans using PM or OHV recruited up to 103 (± 8) or 109 (± 10) parents respectively. The number of parents that reached the maximal number of progenies (Cmax = 250) varied between CSC. PM criterion allocated Cmax progenies to 24 parents (± 3), UC1 and PROBA to 20 parents (± 2), UC2 to 17 parents (± 4), EMBV to 16 parents (± 3) and OHV to 13 parents (± 4). Around 80% of progenies were allocated to 22 (± 3) parents and 69 (± 4) crosses. These CONSTRAINTS actually help at diversifying the mating plan. In the NO CONSTRAINTS scenarios, only 55 crosses were recruited (against 200 in CONSTRAINTS scenarios), and the number of parents was 33 (± 11) compared to 101 (± 5) in CONSTRAINTS scenarios.

To check wether different CSC recruited similar parents, we computed the proportion of parents and crosses shared by mating plans obtained by two different CSC. We also computed the "genetic similarity" of mating plans as the genetic variances-covariances of selected parents, weighted by their number of progenies (**Figure 2** for "selected populations" and **supplementary Figure S4** for "unselected populations"). Overall, many parents were shared between CSC (upper diagonal of **Figure 2A**), but were crossed differently (lower diagonal of **Figure 2A**). Indeed, the proportion of parents (crosses) in common across the 6 CSC ranged from 36% to 82% (3-43%). The criteria OHV recruited the most original parents, with 36% to 48% of parents in common with other criteria, while other criteria shared from 47% to 92% of parents. The genetic similarity of selected parents (off-diagonal of **Figure 2B**), increased when the genetic covariance of recruited parents increased. The CSC EMBV, PROBA UC1 and UC2 showed the highest genetic similarities (0.94-0.97). PM recruited slightly more different genetic background (0.89-0.96) compared to all other pairs of CSC expected pairs involving OHV, which recruited even more distinct genetic backgrounds (0.81-0.89). The similarity of parents that were attributed a minimum number of progenies (**supplementary Figures S5 and S6** was much lower than the similarity of parents that were attributed a high number of progenies. This highlights that superior parents (*i.e.,* parents that receive a lot of progenies) are more related, within and between mating plans.

We also checked whether some CSC recruited more diverse parents than others. The diagonal of **Figure 2B** gives the genetic variances-covariances of parents weighted by their number of progenies. The genetic similarity of parents was the highest for PM (PM > UC1 > EMBV > UC2 >

PROBA > OHV), meaning that PM criterion tends to recruit more similar parents than other criteria within a mating plan.

To conclude, all CSC but OHV recruited similar genetic backgrounds, likely because the parental mean genetic value is highly determinant in CSC' calculation (**Figure 1**). However, CSC based on progeny variance (UC1, EMBV, PROVA, UC2) and OHV involve more distinct genetic backgrounds than PM, which would be an advantage to maintain genetic diversity.

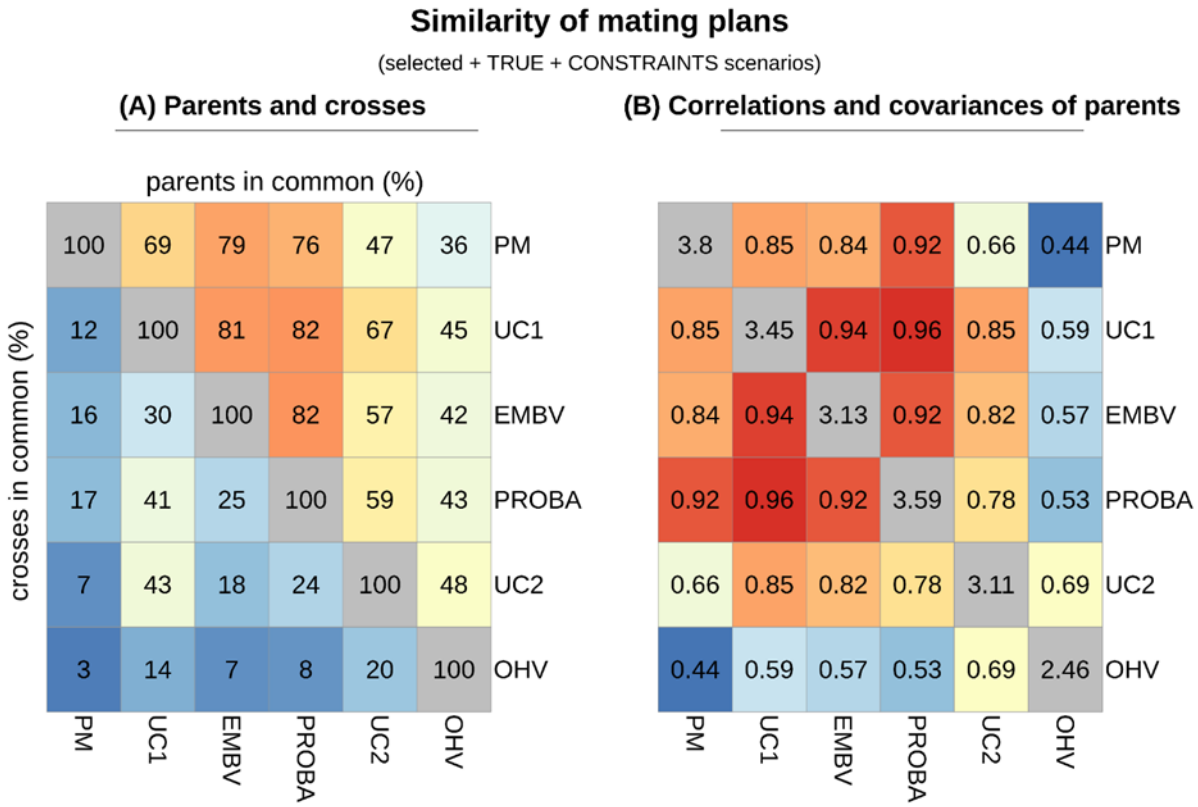


Figure 2: Similarity of mating plans

*Results are given for "selected populations", in TRUE + CONSTRAINTS scenarios (criteria are computed from true QTLs effects and parental contributions are constrained**). A: Parents and crosses shared by mating plans using different CSC.** For pairwise criteria, the number of crosses in common (and respectively the number of parents in common) divided by the total number of crosses (respectively the total number of parents) was computed and averaged over repetitions (n=20). **B: Genetic similarity of mating plans.** Genetic similarity of mating plans. On the diagonal, weighted variances of recruited parents computed as $c_{criteria1} * G^{(2)} * c'_{criteria1}$, averaged over repetitions (n=20). Off-diagonals, weighted correlations of recruited parents for each pair of criterion computed as $c_{criteria1} * G^{(2)} * c'_{criteria2}$, divided by the square root of $(c_{criteria1} * G^{(2)} * c'_{criteria1}) * (c_{criteria2} * G^{(2)} * c'_{criteria2})$. The $G^{(2)}$ relationship matrix was estimated from genotypes with LDAK software to decrease the bias due to LD between markers. The $c_{criteria}$ vector of parental*

contribution gives the proportion of progeny allocated to each parent in the mating plan optimized according to the CSC.

- ***Genetic gain after one generation of selection in "TRUE" scenarios***

The progenies of different mating plans were simulated *in silico* to estimate the proportion of transgressive lines, the genetic gain and the genetic diversity associated with each criterion and scenario.

Figure 3A gives the relative increase in the mean TBV of top progenies obtained by mating plans using the different criteria compared to the mean TBV of top progenies obtained with the mating plan using PM for "selected populations" (see **supplementary Figure S7** for results on "unselected populations"). For a selection rate of progeny comprised between 1/3300 (the very best progeny) and 10%, the relative benefits of all criteria compared to PM were always positive, except for the OHV criterion at a selection rate superior to ~ 1 %. At a selection rate of 7%, PROBA provided the highest gain (4.34% ± 0.82), followed by UC1 (4.33% ± 0.91), UC2 (2.87%± 0.86), EMBV (1.43% ± 0.84) and OHV (-3.32% ± 1.88). In "unselected populations", at a selection rate of 7%, PROBA provided also the highest gain (PROBA 3.85% ± 0.72; UC1 = 3.15% ± 0.79; UC2: 2.22% ± 0.94; EMBV = 0.86% ± 0.72.; OHV= -1.74% ±1.35).

We also computed the proportion of transgressive progenies, *as* the proportion of progenies superior to X% of best parental line, X varying from 100% to 120% (**Figure 3B**). The criteria PROBA and UC1 show the highest proportion of transgressive progenies, followed by the PM, UC2 and EMBV and finally by OHV.

For recurrent selection, we applied a selection rate of 7% among the progenies. The selected progenies constitues the new sets of parents for the next cycle. We estimated the genetic diversity in this set as an estimator of genetic diversity available for long-term genetic gain. We computed the genic variance (at QTLs) of the selected progenies. **Figure 3B** shows that the genic variance decreased in the selected progenies compared to the parents, because of a shift in allelic frequency due to selection and genetic drift. A black line joins criteria representing best trade-off between genetic gain and genetic diversity. The best trade-off criteria included OHV, UC2 and UC1, with OHV providing the highest genic variance and UC1 providing the highest genetic gain. The PM criterion did not belong to the best trade-off criteria, meaning than criteria taking into account recombination lead to a better compromise between genetic gain and genic variance.

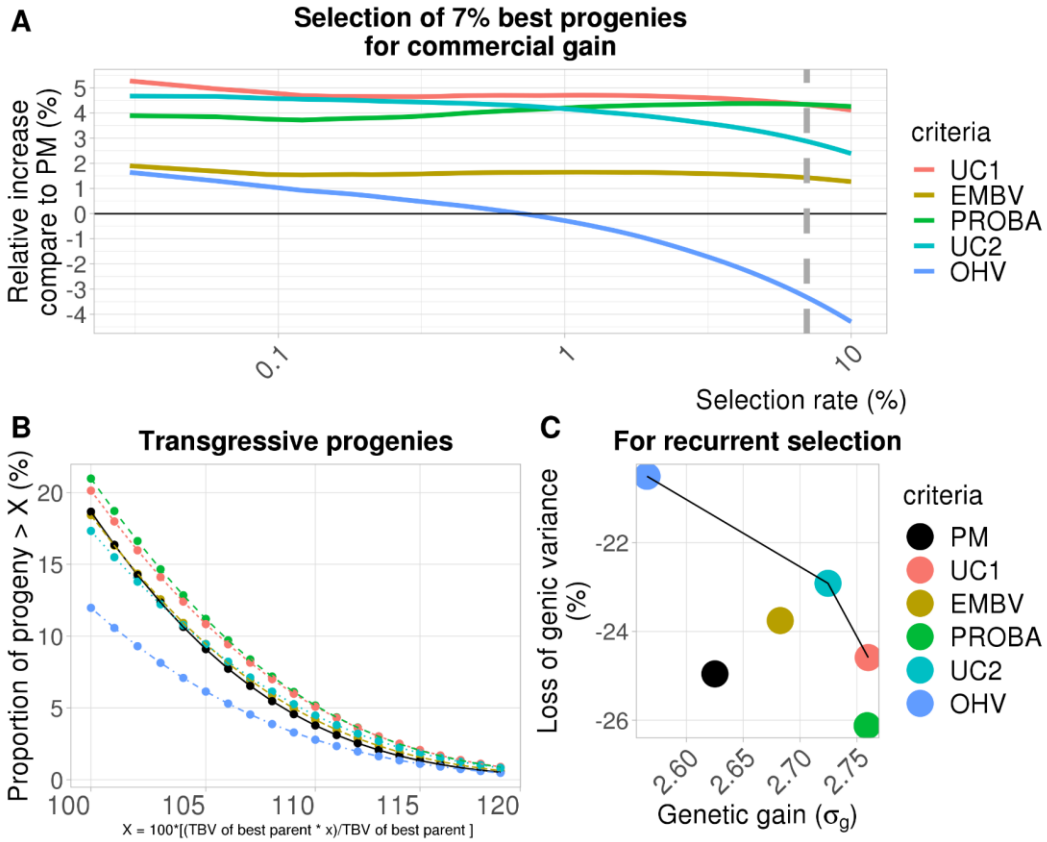


Figure 3: Benefits of each CSC in selected populations + TRUE + CONSTRAINTS scenarios

The results are presented for the "selected populations", in TRUE + CONSTRAINTS scenarios (QTLs effects are known and parental contributions are constrained). ***A. Relative increase in the mean TBV of selected progenies compare to PM criterion*** *The y-axis gives the relative increase in mean TBV of selected progeny for each CSC compare to PM criterion for a selection rate < 10%. The grey dashed line indicates a selection rate of 7%, which is the selected fraction of progeny supposed to form the new breeding population at the next cycle.* ***B. Proportion of transgressive progenies.*** *A transgressive progeny has a genetic value superior to X% of the best parental line, X ranging from 100 to 120%.* ***C. Genetic gain and loss of genic variance associated with each criterion in the 7% best progenies.*** *The 7% best progenies (grey dashed line in graph A) are considered as the new parents for the next selection cycle. The loss of genic variance is computed as the difference between the genic variance of the selected progeny and the genic variance of the former breeding population (n=900 parents in "selected populations"), divided by the genic variance of the former breeding population and multiply by 100. The genetic gain is computed as the difference between the average TBV of the selected progeny and the average TBV of the former breeding population, divided by the standard deviation of the TBV in the initial breeding population, multiply by 100. A black line joins criteria associated with the best trade-off between genetic gain and genic variance.*

Constraints on parental contributions are necessary to avoid producing progenies from highly-related parents. The diversity constraints force the algorithm to attribute progenies to more parents and not only the couple with highest CSC. The loss of genetic gain in CONSTRAINTS scenarios compared to NO CONSTRAINTS scenarios was 4% (± 3) (**supplementary Figures S8 and S9**), but the loss of genic variance in the new breeding population was almost divided by two (0.42 ± 0.12) in the CONSTRAINTS scenarios compared to the NO CONSTRAINTS scenarios. Thus, constraints had low impact on genetic gain but strongly reduce the loss of genic variance.

To sum up, we showed that mating plans were partly conserved across criteria because mean parental genetic value contribute a lot to the value of CSC in elite material. This explains the limited range of genetic gain in our data, but also highlight the even higher necessity of constraints on parental contributions for PM criteria in order to maintain putative genetic gain for the following generations.

- ***Genetic gain after one generation of selection in "ESTIMATED" scenarios***

When CSC are computed from estimated markers effects by a GBLUP model, the relative increase in TBVs in the selected progenies using alternative criteria compared to PM is close to 0% (**Figure 4** for "selected populations" and **supplementary Figures S10** for "unselected populations" optimized under CONSTRAINTS scenarios, **supplementary Figures S11 and S12** for NO CONSTRAINTS scenarios). In "selected populations", at a selection rate of 7%, all criteria provided a negative or null benefits (PROBA: ±0.03% ± 2.27; UC1: -0.17% ± 1.72; EMBV: -0.89 ± 1.96, UC2: -0.97% ± 2.44; OHV: -1.05 ± 3.47). The relative benefits of PROBA was positive at selection rate inferior to 7% but did not exceed 0.9%. In the "unselected" scenarios, the relative increase was positive for all criteria but EMBV and OHV (PROBA: 1.19% ± 1.34; UC1: 0.73% ±1.94; UC2: 0.04% ± 2.31; EMBV: -0.18% ± 2.11; OHV: -0.97 ±2.86). Note that for PROBA and UC1, there is a limited risk to obtain a genetic gain inferior to PM.

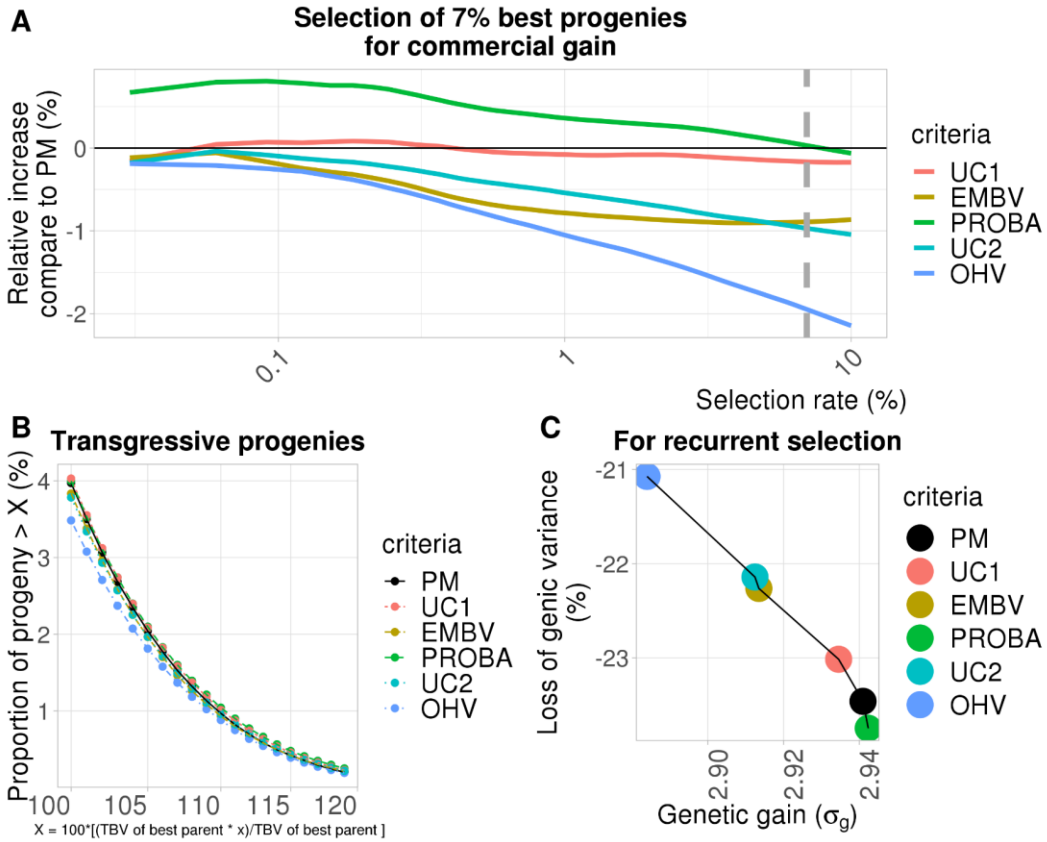


Figure 4: Benefits of each CSC in selected populations + ESTIMATED + CONSTRAINTS scenarios

CSC were computed from estimated marker effects derived from a GBLUP using the 900 parental genotypes and phenotypes as input.

The lower benefits of criteria based on progeny variance arise from the strong error in estimates of progeny variance. Notably, the progeny variance was strongly underestimated. The ratio between ESTIMATED variance of progeny standard deviation divided by the TRUE variance of progeny standard deviation was 0.27 ± 0.06 in "unselected" scenarios and 0.15± 0.03 in "selected scenarios". In comparison, the variance of progeny expected mean was less underestimated, with a ratio of 0.56 ± 0.06 (0.53 ± 0.1) between ESTIMATED and TRUE scenarios in "unselected" ("selected") scenarios.

As σ is more underestimated than PM, the criteria based on progeny variance (UC1, UC2, PROBA, EMBV) were even more influenced by PM than by σ, thus displaying a higher correlation in ESTIMATED scenarios than in TRUE scenarios (**supplementary Figure S3**). Consequently, the proportion of parents in common were also much higher (**supplementary Figures S13 and S14**), which explains the very limited range of supplementary genetic gain.

To sum up, we show that the robustness of CSC decreases when the estimation of marker effects is less accurate. Note that estimates are more robust in “unselected” scenarios than in “selected” scenarios. This likely comes from both a higher heritability, a higher genetic diversity and a lower LD between QTLs in “unselected populations”. This highlights the necessity to still improve the estimation of marker effects for cross value prediction.

❖ Discussion

• *Added-value of progeny variance estimation compared to progeny mean*

Without any genomic information, the mean value of parental phenotypes for several traits of interest is essential to design mating plans in a breeding program. Decreasing cost of genotyping and sequencing, and publication of high-density genetic and physical maps, along with the development of genomic prediction methodology, allow to predict genetic value of lines and interest of crosses without phenotyping the progeny. Those predictions are based on marker effect estimations. We showed that, in our material, parental mean genetic value was highly correlated to the different CSC we tested, especially when the estimation of marker effects was not accurate **(Figure 1** and supplementary **Figure S3**). Even with unperfect marker estimation, CSC and especially those that estimate progeny variance, along with diversity constraints in the mating optimization process, ensure to maintain a higher level of diversity, required for long term genetic gain, without significantly decrease short term genetic gain.

However, the increase of genetic gain using CSC compared to PM, especially those considering genetic variance estimation, was rather limited in this elite material. The benefits in terms of genetic gain of CSC based on progeny variance estimation compared to PM was actually shown by Zhong and Jannink (2007) and Bijma et al. (2020) to be function of the ratio between the variance of progeny standard deviation and the variance of the expected progeny mean value: var(σ)/var(PM). Authors that worked with CSC usually refer to this ratio to analyze the expected benefits of CSC based on progeny variance over PM (Lado et al. 2017; Lehermeier et al. 2017). A higher ratio means that the value of superior progenies are less determined by PM and more determined by σ, and so criteria based on progeny variance eventually lead to superior genetic gain if progeny variance estimates are accurate enough. Our data also illustrate the impact of a superior ratio var(σ)/var(PM) on the relative increase of the genetic value of selected progenies when mating plans are optimized with CSC based on progeny variance compared to PM (**Table 1**). This relative increase is even higher in “TRUE” and “selected” scenarios (4.34%) than in “unselected” scenarios (3.85%). This is consistent with the fact that we find a higher ratio var(σ)/var(PM) in “selected” populations (2.4% ± 3e-3) than in “unselected” populations (1.1% ± 2e-3). In “ESTIMATED” scenarios, the ratio var(σ)/var(PM) strongly decreased to 0.3% (0.5%) in “selected” (“unselected”)

populations leading to lower relative increase of TBVs in the selected progenies using CSC compared to PM.

Table 1: Relationship between the ratio var(σ)/var(PM) and the relative increase in the value of selected progeny.

Scenarios	Populations	Constraints	var(σ)/var(PM)	Relative increase of selected progeny TBV compare to PM
TRUE	Selected	CONSTRAINTS	2%	4.34% (PROBA)
TRUE	Unselected	CONSTRAINTS	1%	3.85% (PROBA)
ESTIMATED	Unselected	CONSTRAINTS	0.5%	1.19% (PROBA)
ESTIMATED	Selected	CONSTRAINTS	0.3%	0.03% (PROBA)
TRUE	Selected	NO CONSTRAINTS	2%	1.8% (UC1)
TRUE	Unselected	NO CONSTRAINTS	1%	1.4% (EMBV)
ESTIMATED	Unselected	NO CONSTRAINTS	0.5%	0.8% (UC1)
ESTIMATED	Selected	NO CONSTRAINTS	0.3%	-0.2% (EMBV)

Ratio var(σ)/var(PM) is given for the complete dataset of crosses. Relative increases of the mean TBV of selected progenies are given for the best CSC at a selection rate of 7%.

As the expected genetic gain increases with the ratio var(σ)/var(PM), the perspectives are to find which factors impact this ratio, thereby in which situation CSC based on progeny variance are interesting for breeders.

The ratio var(σ)/var(PM) probably depends on the genetic composition of the parental population. We observed that the magnitude of this ratio increased in “selected” scenarios compared to populations that have never been selected for the trait. As explained by Bijma et al. (2020), directional selection reduces the genetic variance $\sigma^2{}_a$ by producing negative covariance between QTLs (Bulmer 1971). The reduction of genetic variance σ_a also decreases var(PM) because var(PM) is equal to half the genetic variance [var(PM) = var($(P_i + P_j)/2$) = ½ $\sigma^2{}_a$]. But directional selection is not expected to affect mendelian sampling variance σ² and thus var(σ) in an infinitesimal model. In our data, which does not perfectly fit the infinitesimal model as there were “only” 300 QTLs that controlled the trait, we observed that the genetic variance and var(σ) were reduced by directional selection when comparing “selected” and “unselected” scenarios The genetic variance was reduced by a factor 1.5, var(PM) was reduced by a factor 3 and var(σ) was reduced by a factor 1.4. The reduction of progeny variance with selection was expected. Indeed, according to the progeny variance formula published by Lehermeier et al. (2017), progeny variance

depends on the polymorphism between parents and linkage phase between QTLs in parents, *i.e.*, coupling or repulsion phase. The decrease of var(σ) in "selected populations" can actually be caused by these two effects: 1) the lower polymorphism in parents is caused by a shift in allelic frequency toward 0 or 1 due to combined action of selection and genetic drift 2) the selection causes a higher repulsion phase between alleles described as the Bulmer effect.

In case of highly structured populations, the magnitude of var(σ) may increase. Structuration is actually associated with a higher polymorphism between groups (Wahlund effect) and more efficient recombination, leading to higher genetic variances in progenies when parents originate from different groups. This could explain why Lehermeier et al. (2017) obtained higher genetic gain compared to us starting from a Maize NAM population built with dent European landraces (Bauer et al. 2013), and Yao et al. (2018) on bread wheat crosses involving Chinese and Australian lines likely to be very differentiated. To compare with Lehermeier et al. (2017), we used scenarios with comparative methodology: UC1 criterion, "unselected parental populations" (no preliminary selection cycles), estimated marker effects, NO CONSTRAINTS for mating plans (**supplementary Figure S12**). The ratio var(σ) / var(PM) was 14% in average in Lehermeier et al. (2017) according to Table 2 with $h^2 = 0.2$ or $h^2 = 0.6$, without diversity constraints. The ratio in our study, using elite material, was 1.1% for TRUE "unselected populations" and 0.05% for ESTIMATED "unselected populations". In Lehermeier et al. (2017), the genetic gain provided by UC compared to PM was superior to 0.2 genetic standard deviation (σ_a) at a selection rate < 10% (Figure 4 of their article), without applying diversity constraints, and using Posterior Mean Variance model for marker effect estimation, with $h^2 = 0.2$ and $h^2 = 0.6$. This is four times (ten) higher than our results in "unselected" and TRUE (ESTIMATED) scenarios using UC1 criterion (respectively 0.05 and 0.02 σ_a). In Yao et al. (2018), using estimated marker effects, genetic gain provided by UC was 0.06 σ_a at $h^2 = 0.3$, 0.08 σ_a at $h^2 = 0.5$ σ_a and 0.13 at $h^2 = 0.8$, for a selection rate varying between 1 and 10%. This is of the same magnitude of what we observed in TRUE scenarios (0.05 σ_a, **supplementary Figure S9**) and much more to what we observed in ESTIMATED scenarios (0.02 σ_a, **supplementary Figure S12**).

If the ratio var(σ) / var(PM) is high enough, another element that might increase differentiation between mating plan based on PM and mating plans based on estimated progeny variance would be a negative covariance between PM and σ, which was reported in several publications in maize, (Bernardo 2014; Mohammadi et al. 2015), bread wheat (Lado et al. 2017) and barley (Abed et Belzile 2019; Neyhart et Smith 2019). We actually observed a triangular relationship between PM and σ in INRAE/AO data and in simulations. The negative relationship was stronger in ESTIMATED scenarios (**supplementary File**). A negative covariance indicates that crosses with low mean show a higher variance, and thus ranking crosses according to CSC based on progeny variance estimation may be useful in terms of genetic gain. However, a negative covariance between PM and σ is not sufficient in itself to ensure that criteria based on progeny variance will have any interest, because the most important factor is still a high var(σ)/var(PM) ratio (Lado et al. 2017).

The benefits of CSC based on progeny variance estimation also depends on the accuracy of marker effects estimates. Indeed, in our data, the benefits were close to 0 in ESTIMATED scenarios. The accuracy of progeny variance prediction has been shown to increase with heritability and size of training population (Lehermeier et al. 2017; Yao et al. 2018; Santos et al. 2019). Factors influencing accuracy of GEBV, such as quality of phenotyping, experimental design, statistical model used to take into account environmental effects, relationship between the candidate and the training population, will probably increase the estimation accuracy of marker effects, and consequently progeny variance and cross value. However, it is also currently reported in literature that prediction models underestimate progeny variance (Lian et al. 2015; Tiede et al. 2015; Lehermeier et al. 2017; Adeyemo et Bernardo 2019; Santos et al. 2019). If progeny variance are more underestimated than parental genetic values, thus the ratio var(σ)/var(PM) decreases, and benefits of criteria based on progeny variance estimates decrease. This is exactly what happened in our ESTIMATED scenarios. For example, because of progeny variance underestimation, the ratio var(σ)/var(PM) drops to 0.05%±1e-3 in ESTIMATED "unselected" scenarios when markers effects are estimated by a GBLUP model while it is 1.1% ± 2e-3 when QTLs effects are known. Some authors suggested to use adapted genomic prediction model to better estimate progeny variance (Zhong et Jannink 2007; Lehermeier et al. 2017; Santos et al. 2019). Different prediction models provide moderate improvement in accuracy of GEBV for quantitative traits, except when the variation of traits are controlled by a few and heterogenous QTLs effects (Daetwyler et al. 2008; Heslot et al. 2012). However, the prediction of progeny variance from estimated markers effects seems to be more sensitive to the choice of genomic prediction models. In our study, as in most of published papers on this subject (*e.g.* Bernardo et al. 2014 or Mohammadi et al., 2015), the expected variance of progeny ${\sigma^2}_{ij}$ derived from the cross $P_i * P_j$, was obtained replacing the QTLs effects by estimated marker effects (for the ESTIMATED scenarios). In matrix notations, this estimation is $\hat{\beta}' V_{ij} \hat{\beta}$ with $\hat{\beta}$ the vector of estimated effect and V_{ij} the genotypes (co)variance matrix in the progenies from $P_i * P_j$ cross. Lehermeier *et al*. (2017) proposed a better estimation of $\sigma_{ij}^2 = \beta' V_{ij} \beta$, they called PMV for Posterior Mean Variance, which is the expectation of $\beta' V_{ij} \beta$ on the space of β given the calibration population information. This expectation was obtained averaging samples ${\beta^{[s]}}' V_{ij} \beta^{[s]}$ of from the MCMC used to estimate the SNP effects ($\beta^{[s]}$ the effect vector in the s^{th} sample). This methodology takes into account uncertainty in marker effects estimates. It efficiently decreases the bias in progeny variance estimates and increases the correlation between true and estimated progeny variance, in comparison to our results. For example, for a h² = 0.4 and a training population size ranging from 100 to 600, bias on progeny variance was comprised between 0.21 to 0.06 in Lehermeier et al. (2017) versus -0.83 ± 0.03 in our data. The correlation ranged between 0.58 and 0.65 in their article versus 0.4 ± 0.08 in our data. Another type of model that seems to provide more accurate estimates of progeny variance than GBLUP model is the Bayesian Lasso model (Santos et al. 2019). This suggests that variable selection models may be of interest for prediction of progeny variance. Using haplotypic blocks

instead of markers might be of interest too (Cole and VanRaden (2011), Bonk et al. (2016)). The idea is that combinations of alleles in haplotypic blocks may be better estimated (if present in the training population) than individual SNPs. For bread wheat, recombination hotspots described in Danguy des Déserts et al. (2021) could be used as haplotype block separators.

- ***Compromise between genetic gain and genetic diversity***

The breeder's equation shows that genetic gain is proportional to selection intensity and genetic variance. However, the theory predicts that in a closed system, without extrinsic germplasm introduction, each selection step is associated with a reduction of genetic variance. Loss of genetic variance increases as selection intensity increases (Woolliams et al. 1993; Woolliams et al. 2015). Thus, genetic gain in successive generations is expected to decrease, and finally converge to 0 when there is no longer genetic variance in the breeding population (Jannink et al. 2010). This phenomenon is faster with genomic selection, that decreases generation interval, increases selection intensity if accuracy if high, and increase the probability to select related individuals (Clark et al. 2011, Pszczola et al. 2012).

The total genetic gain is defined at each selection step as the genetic gain cumulated since the start of the breeding program. In a long-term perspective, the best selection strategy would consist in selecting the mating plan at each selection step to maximize the total genetic gain at the breeding program's scale, with eventually an emphasis on short-term genetic gain depending on the breeder strategy. However, long term consequences of a mating plan are difficult to predict.

A practical strategy would rely on inbreeding rate management in parental population. Indeed, the loss of genetic variance due to selection is proportional to the increase of inbreeding rate ΔF (Facolner and Mackay 1996). This long-term inbreeding rate increases with the sum of squared long-term contributions of genitors to the progeny (Wray and Thompson, 1990), where the contributions of genitors correspond to the proportions of genes in the progeny originating from each genitor. Thus, authors suggested to constrain parental contribution to limit the increase of ΔF (for a review, see Woolliams et al. (2015)). By considering pairwise co-ancestry of parents, this leads to control the inbreeding coefficient of selected genitors weighted by their contribution to progeny: c'Φc where c is the parental contribution (either a binary value indicating if each parent is selected or not (Brisbane et Gibson 1995), or the proportion of progeny allocated to each parent (Meuwissen 1997) and Φ is the co-ancestry matrix. These methods are called Optimal Contribution Selection (Meuwissen 1997) or Optimal Cross Selections (Kinghorn et al. 2009; Allier al. 2019), OCS. In autogamous crops, such as bread wheat, where lines can be fully inbred, the inbreeding concept is not pertinent and Allier et al. (2019) suggested to replace ΔF by 1-ΔHe where He is the mean expected proportion of heterozygous loci in progeny derived from panmictic crosses of selected parents.

However, the simultaneous maximization of genetic gain (ΔG) and the minimization of the loss of genetic variance in progeny (approximated with the minimization of ΔF) form a double objective problem. There are many possible mating plans to test, each of them associated to a specific ΔG and a specific ΔF. As illustrated in Akdemir et al. (2019), the mating plans showing the best combination of ΔG and ΔF form a Pareto Front, a useful concept when exploring multi-objective problems. However, in complex situations, as the one considered here, exploring the space of possible solutions and finding the Pareto Front is highly demanding in terms of computation time and requires an efficient algorithmic strategy. Suppose that the Pareto Front is identified, the choice of the best mating plan on the Pareto Front is still not trivial. A strategy would be to define a unique objective function to maximize ΔG – α*ΔF, where α is a weight. The value of the weight must be chosen according to the economic consequences of inbreeding for animals (Wray et Goddard 1994) or economic importance of genetic diversity in crops (Allier et al. 2019a). A convenient strategy suggested initially by Meuwissen (1997) is to set a maximal threshold T for F and find the mating plan that maximize ΔG under F ~ cΦc' < T. This is called the ε-constraint method. The complexity of the problem is reduced to a single objective problem which can be efficiently solved by linear programming methods. However, the threshold for ΔF is difficult to define. One strategy is to test by simulation several trajectories for ΔF and check which one provides the highest commercial genetic gain over many selection cycles (Allier et al. 2019). Another strategy is to set a threshold to ensure maintaining a sufficient effective size in a breeding program. For example, the empirical view for FAO is that the rate of inbreeding should not exceed 1% per selection cycle in cattle, which corresponds to an effective population size of Ne = 50 (FAO 1998).

In our study, we implicitly defined a maximal threshold for F by setting constraints on parental contributions and avoiding crossing parents that were highly related. Our single objective problem was to maximize the objective function (one per criteria) under these constraints. This is in line with the propositions made by Toro et al. (1988) and Toro and Perez-Enciso (1990). This allows to use linear programming, except for criteria EMBV which does not fit requirements for linear programming. Our constraints on parental contributions were inspired from the dimensions of a real breeding programs and are thus very easy to set, at the opposite of an explicit threshold on F. These simple constraints have the advantage to be easy to set, but are likely suboptimal to limit the loss of genetic variance because they do not explicitly include co-ancestry of parents. This strategy might be adapted for a French bread wheat program but not to for other types of breeding programs. Indeed, French bread wheat breeding programs use a high proportion of genitors from other companies at each mating plan which reduces the need to manage internal genetic diversity. A supplementary step would be to add a supplementary constraint, which would ensure that the weighted inbreeding coefficient of selected genitors c'Φc does not exceed a threshold, which is still to be defined. Another supplementary step would be to vary these constraints and check which set of constraints is associated with the best Best trade-off of genetic gain and genetic diversity.

Note however that constraints on parental contributions can be seen as undesirable if the objective is limited to short-term genetic gain. In ESTIMATED and NO CONSTRAINTS scenarios, a very few number of parents and crosses actually contribute to the mating plan. In CONSTRAINTS scenarios, much more parents and crosses were used in mating plans. Diversify the sets of parents and crosses with empirical constraints provides the advantage to limit the risk to discard high performing crosses whose CSC are underestimated because of inaccuracy of marker estimates.

- ***Select your criteria***

For each CSC, we obtained a different mating plan. We supposed a selection rate of 7% to select the genitors of the next cycle among progenies of selected parents and to compute the corresponding genetic gain and genic variance. We identified the criteria associated with the best trade-off between genetic gain and genic variance for several scenarios: TRUE or ESTIMATED, “selected” or “unselected populations”. The PM criterion never belonged to the best trade-off, except in ESTIMATED + CONSTRAINTS or NO CONSTRAINTS scenarios, which were scenarios were progeny variance estimates were the less accurate. This shows that criteria considering reshuffling of alleles by meiotic recombination are superior to PM. The OHV criterion systematically belong to the best trade-off and was associated with a minor loss of genic variance but also a lower genetic gain. The property of OHV to maintain genetic diversity was already demonstrated by its authors (Daetwyler et al. 2015). The criteria PROBA, UC1 and EMBV were belong to best trade-off and were associated with the highest genetic gain and the highest loss of genic variance in most scenarios, but no more than PM. PROBA was the most efficient in CONSTRAINTS scenarios while UC1 and EMBV were the most efficient in NO CONSTRAINTS scenarios. PROBA was advantageous in CONSTRAINTS scenarios but show no interest in NO CONSTRAINTS scenarios, which suggests that PROBA superiority in CONSTRAINTS scenarios come from a better handle on the constraints. However, note that PROBA criterion requires to set a threshold. In this study, we choose to set the threshold as the genetic value (either TBV in TRUE scenarios, either GEBV in ESTIMATED scenarios) of the best parental line of the current breeding population. However, authors suggested to use different thresholds (Wellmann 2019; Bijma et al. 2020). Note that if the threshold is too high (or too low) compared to expected distributions of progeny (all crosses combined), most crosses will have a 0 value for PROBA (or 1), making this criterion difficult to optimize.

❖ Conclusion

Optimization of mating plans using recombination rate information and controlling the number of parents, crosses and progenies per cross allows to increase the rate of genetic gain with a lower loss of genetic diversity compared to classical mating plans based on parental genetic value only. The deployment of such optimization methods will be facilitated by the advancement and generalization of genomic prediction in breeding programs. The implementation of such methods does not represent a supplementary cost as it only requires genotyping of parents, which is already carried out for genomic prediction. However, the benefits of these methods depends on the genetic composition of the breeding population, and also efforts to improve accuracy of genomic prediction models, in particular for the estimation of gametic variances.

III.3 Conclusions sur la comparaison des bénéfices de plusieurs critères de sélection de croisements dans une population de blé tendre d'hiver

L'objectif du sélectionneur est de produire des variétés plus performantes à chaque génération tout en maintenant une variance génétique pour pérenniser le gain génétique sur le long terme. L'objectif de ce travail a été de comparer différents Critères de Sélection de Croisements (CSC), en termes de gain et de variance génétique après optimisation d'un plan de croisement dans un programme de sélection de blé tendre d'hiver élite. L'outil d'optimisation des plans de croisements (open-source) ainsi que les réflexions sur le choix du CSC le plus adapté aux objectifs court terme et long terme seront utiles aux sélectionneurs.

Pour comparer les critères dans le cadre d'un programme de sélection, nous avons utilisé comme génotypes parentaux de départ ceux du programme de sélection Agri-Obtentions-INRAE. Pour simuler des descendances, nous avons simulé une architecture de caractère avec une variance génétique ($\sigma^2_{TBV} = 14$) et une héritabilité ($h^2 = 0.4$) égale à celle du rendement dans le matériel de départ, contrôlée par 300 QTLs à effets additifs, une situation proche de celle du modèle infinitésimal.

III.3.1 Précision des effets aux marqueurs

Les CSC ont été comparés lorsque l'architecture du caractère (position et effets des QTLs) sont connus (scénario TRUE) ou lorsque les effets des marqueurs sont estimés par GBLUP (scénario ESTIMATED). Ce deuxième scénario suppose que les marqueurs sont en déséquilibre de liaison avec les QTLs. Le scénario TRUE, qui suppose une détection exhaustive et précise des QTLs est irréaliste pour des caractères quantitatifs comme le rendement où les QTLs sont nombreux, les effets des allèles faibles et les interactions entre génotypes et environnement fortes (Heffner et al. 2009; Millet et al. 2016). En pratique, il y a donc toujours des erreurs d'estimation dans les effets des marqueurs. Notre objectif ici était d'estimer le gain génétique maximum espérée si l'architecture du caractère était parfaitement estimée.

L'estimation de ces effets peut être améliorée en augmentant la taille de la population d'entrainement. Obtenir de très grandes populations d'entrainement est possible avec l'accumulation de données historiques et le développement du phénotypage haut-débit. Le phénotypage haut débit utilise notamment les méthodes d'imagerie numérique, par exemple les spectres de réflectance acquise avec des drones, corrélés à des caractères d'intérêt plus chers à phénotyper, tels que le rendement (Lozada et al. 2020; Kanke et al. 2016; Rutkoski et al. 2016; Sun et al. 2017; Crain et al. 2018; Juliana et al. 2018). La prédiction génomique « trait-assisted » permet d'augmenter la taille de la population d'entraînement à budget fixé en optimisant le phénotypage entre les deux caractères corrélés (Ben Sadoun et al. 2020).

La précision de la prédiction de la variance de la descendance σ² est aussi très dépendante du modèle de prédiction génomique utilisé. Santos et al. 2019 montrent que le modèle Bayésien Lasso fournit un estimateur de variance de la descendance moins biaisé, et mieux corrélé au vrai paramètre. Lehermeier et al. 2017 arrivent à obtenir un biais très proche de 0 et une corrélation très proche de 1 en utilisant le modèle Bayésien Posterior Mean Variance (PMV).

A noter que la prise en compte de QTLs majeurs dans la variance de la descendance d'un croisement est tout à fait possible. La connaissance du génotype au QTL chez les parents est alors indispensable, et il faut pouvoir lui attribuer un effet (estimé par prédiction génomique ou détection de QTLs).

III.3.2 Proposition de nouveaux critères de sélection de croisements à implémenter dans l'algorithme d'optimisation des plans de croisements

Sur les six critères testés dans l'article, PM, UC1, UC2 et OHV sont couramment comparés dans la littérature (Lehermeier et al. 2017; Allier et al. 2020; Goiffon et al. 2017; Yao et al. 2018; Daetwyler et al. 2015). Par contre, nous avons proposé une nouvelle approche pour le critère EMBV (Expected Maximum Breeding Value). Initialement proposé par Müller et al. (2018), il s'agit dans la publication d'origine de calculer l'EMBV d'un individu en supposant un nombre de gamètes constant, puis sélectionner les individus avec le meilleur EMBV. Nous calculons l'EMBV de la même manière mais pour une F1 issue d'un croisement biparental et pour un nombre de descendants qui varie entre couples. Nous maximisons $EMBV_{ij}(D_{ij})$ x D_{ij} sachant les contraintes sur les D_{ij} et sachant que $EMBV_{ij}$ se calcule comme une fonction non continue de D_{ij}, ${\sigma^2}_{ij}$ et PM_{ij}. Pour maximiser cette équation complexe, nous avons développé un algorithme génétique en collaboration avec Jean-Marc Alliot (Institut de Recherche Informatique de Toulouse, IRIT), Nicolas Durand (Ecole Nationale de l'Aviation Civile, ENAC), et avec l'appui de Daniel Ruiz (Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications, INP ENSEEIHT).

L'avantage de l'algorithme génétique, par rapport à d'autres méthodes heuristiques comme le recuit simulé, est qu'il permet d'identifier simultanément plusieurs optimums (Durand et al. 2004). Cet algorithme inclut notamment une possibilité de « sharing » qui limite les risques de converger vers des optimums locaux en donnant un bonus aux solutions isolées dans l'espace des possibles. Cet algorithme propose plusieurs plans de croisements aux sélectionneurs, de qualité éventuellement décroissante, de manière à ce que le sélectionneur ait le choix parmi plusieurs plans de croisements de bonne qualité. L'exploration de différents optimums est un avantage lorsque le problème d'optimisation a plusieurs objectifs négativement corrélés. Cette situation est courante en sélection, lorsqu'on souhaite maximiser le gain génétique pour le rendement par exemple, tout en maintenant un taux protéique élevé alors que celui-ci est négativement corrélé

au rendement, mais aussi en minimisant la perte de diversité génétique (**Figure 8** du manuscrit). L'algorithme génétique permet de construire le front de Pareto entre ces objectifs au sein duquel le sélectionneur pourra choisir.

Par ailleurs, nous avons proposé un nouveau CSC appelé PROBA, qui trie les croisements sur la proportion de descendants supérieurs à la valeur génétique de la meilleure lignée parentale du programme de sélection. Ce critère est pertinent dans le cadre du développement de nouvelles variétés élites, puisque l'exigence minimum est que la nouvelle variété soit meilleure que les variétés actuelles. Ce critère PROBA apporte généralement une plus forte proportion de descendants transgressifs et un meilleur gain génétique dans un plan de croisements sous contraintes. Les bénéfices du critère PROBA diminuent dans un plan de croisements sans contraintes par rapport aux bénéfices des autres CSC basés sur la variance de la descendance (**supplementary Figures S8, S9, S10**), ce qui signifie que ce critère est surtout pertinent lors d'un contrôle des contributions parentales.

III.3.4 Optimiser les contraintes sur les contributions parentales

Nous avons proposé de prendre en compte des contraintes empiriques simples sur les contributions parentales lors de l'optimisation du plan de croisements. Le nombre total de parents, de croisements et de descendants alloués à un croisement ou à un parent sont bornés. De plus, les parents trop proches génétiquement ne peuvent pas être croisés.

Nous avons utilisé deux méthodes pour identifier les couples trop similaires :

- D'après le pedigree : les couples dont les parents ont des probabilités d'identité (IBD) trop élevées. Le seuil d'IBD maximal autorisé est fixé par l'utilisateur. Dans notre article, nous avons rejeté les croisements dont les parents présentaient un IBD > 0.25, c'est-à-dire impliquant par exemple des demi-frères, grands-parents/petits-enfants, oncles/neveux.
- D'après la similarité génétique calculée à partir des génotypes : cette méthode présente l'avantage d'être plus précise que l'apparentement basé sur le pedigree (Goudet et al. 2018). Nous avons utilisé une matrice de variance-covariance entre génotypes construite avec le logiciel LDAK (Speed et al. 2012). Le premier avantage est que cette matrice prend en compte les fréquences alléliques lors du calcul de la similarité génétique, tel que décrit par Yang et al. 2010, Hayes et al. 2009 et Meuwissen et al. 2009. Les variants rares ont plus de poids, ce qui permet d'identifier des structures de populations plus fines (O'Connor et al. 2015; Baye et al. 2011; Shetty et al. 2020). Le deuxième avantage est qu'elle prend en compte l'hétérogénéité du DL le long du génome, phénomène très marqué chez le blé. Pour rappel, dans le **Chapitre 2**, nous avons montré que les télomères ont un taux de recombinaison historique élevé et un plus grand nombre de points chauds de recombinaison, donc un DL très faible, alors que les centromères montrent un taux de recombinaison historique très faible, donc un DL très fort.

Ainsi, la prise en compte de l'hétérogénéité du DL lors du calcul de l'apparentement permet d'éviter que certaines régions génomiques, tels les centromères et péricentromères (33% des marqueurs), contribuent plus que d'autres à l'apparentement.

Les contraintes de diversité que nous avons imposées à l'algorithme et les intensités de sélection appliquées sur les descendances correspondent aux pratiques des sélectionneurs partenaires. Nous n'avons pas testé si ces paramètres étaient optimaux en termes de gains et variances génétiques sur le long terme. En effet, ces contraintes ne prennent pas en compte la similarité génétique des parents, comme le font les méthodes OCS ou l'UCPC-He, et donc la perte de diversité est peu contrôlée. D'ailleurs, la diversité génétique dans les descendants issus des plans de croisements est très variable selon les CSC au sein d'un même scénario (exemple **sur Figure 3** de l'article). Nous avons développé une métrique pour mesurer la similarité génétique des parents recrutés par les différents plans de croisements optimisés grâce à différents CSC (diagonale de la **Figure 2B** de l'article). Les parents sélectionnés par le critère PM montrent une plus forte covariance génétique que les parents sélectionnés par les autres CSC, et à l'inverse la diversité génétique des parents sélectionnés par le critère OHV est nettement plus élevée que les autres CSC. Ceci illustre que le critère PM recrute des parents plus similaires génétiquement que les autres CSC qui sont donc intéressants pour contrôler la diversité génétique dans un programme de sélection sur le long terme. Cependant les contraintes simples sur les contributions parentales que nous avons utilisées ne permettent pas un contrôle fin sur la perte de diversité génétique, et les méthodes types OCS et UCPC-He pourraient s'avérer nettement plus optimales pour limiter la perte de diversité à long terme.

Ces contraintes simples permettent tout de même de diviser par deux la perte de variance génique par rapport à l'application du CSC sans contrainte. Nous avons calculé la variance génique comme la somme des variances des effets additifs à chaque QTL. La variance génique mesure la diversité génétique encore disponible pour un progrès génétique futur. En comparaison de la variance génique, la variance génétique tient compte du déséquilibre de liaison entre QTLs générés par l'effet Bulmer, et donc sous-estime la diversité génétique disponible pour l'amélioration génétique.

III.3.5 Optimisation des temps de calcul de l'algorithme d'optimisation des plans de croisements

Le calcul de la variance de la descendance est un premier défi calculatoire. Pour rappel, la valeur génétique d'un descendant se calcule comme le produit matriciel Xβ où β est le vecteur des effets marqueurs et X est le vecteur génotypique du descendant. Ce vecteur X est un vecteur aléatoire dont il faut déterminer la matrice de variance-covariance conditionnelle aux génotypes parentaux. Cette variance-covariance dépend 1) du polymorphisme chez les parents et 2) des phases et du taux de recombinaison entre les allèles favorables et défavorables chez les parents.

La variance de la descendance se calcule comme $\sigma^2 = V(X\beta) = \beta Var(X)\beta'$. Si l'on souhaite calculer la variance de la descendance pour chaque couple de N parents candidats pour une population génotypée sur M marqueurs, une approche frontale nécessiterait de calculer M*(M+1)/2 termes pour remplir la matrice Var(X) (matrice symétrique), et ce pour chacun des N(N-1)/2 couples, donc un nombre de termes à calculer de l'ordre de N^2M^2, soit de l'ordre de 10^{14} termes à calculer pour 900 lignées parentales génotypées sur 16K marqueurs. Pour limiter le coût calculatoire lié à la variance de la descendance, de nombreux auteurs pré-sélectionnent les parents sur leur valeur génétique (Bijma et al. 2020; Lehermeier et al. 2017; Zhong et Jannink 2007).
Dans le cadre de cette thèse, nous avons accéléré le calcul de l'estimation de la variance génétique dans la descendance en décomposant cette variance σ^2 de manière à identifier les quantités redondantes ou égales à 0 par construction (**supplementary protocol S1**). Avec cette méthode, nous avons pu calculer la variance des descendants de 400k croisements génotypés sur 16k marqueurs en moins de 2 heures en utilisant 2 CPU. En comparaison, Neyhart et Smith (2019) prédisent la variance de 330k croisements génotypés sur 6k marqueurs avec le logiciel PopVar (Mohammadi et al. 2015) en 32 heures de calcul avec 24 CPU.

Pour la majorité des CSC (PM, UC, PROBA, OHV), il est possible d'utiliser la programmation linéaire pour allouer les descendants de manière à maximiser les fonctions Objectif. Un plan de croisements comportant 40k croisements candidats est optimisé en moins de 5 min par le logiciel CPLEX et 2 CPU et un plan de croisements comportant 400k croisements est optimisé en moins de 5 heures par CPLEX, avec 4 CPU. Pour un programme de sélection qui dispose d'un jeu de données historiques de cet ordre de grandeur (900 lignées, 16k marqueurs), il y a donc peu d'intérêt à présélectionner un sous-ensemble de parents pour optimiser ces CSC.

Dans l'article, nous avons fait une pré-sélection des croisements car nous avons simulé environ une centaine de populations de sélection. Les coûts calculatoires sont donc bien plus importants que dans les programmes de sélection réels et justifient une pré-sélection des croisements. La pré-sélection est beaucoup plus avantageuse pour l'optimisation sur le critère EMBV. Ce critère nécessite d'être optimisé par un algorithme génétique, et nous avons observé que la fonction Objectif atteint un plateau à environ 400k itérations. Cela représente en moyenne 13 heures de calcul pour 40k croisements et 109 h de calcul (plus de 4 jours) pour 400k croisements avec 4CPU.

III.3.6 Résultats supplémentaires : Impact du profil de recombinaison sur le bénéfice des critères basés sur la variance de la descendance

Stapley et al. 2017 expliquent que le profil de recombinaison présente une multitude de niveaux de variabilité. D'une part, il y a la variabilité « topologique » du profil de recombinaison (variabilité de sa forme), qui varie selon les espèces. Un exemple est donné en **Figure 24**. La carte de Marey donne la distance génétique cumulée en fonction de la distance physique cumulée. Chez le blé (**Chapitre 2**) et le maïs (Bauer et al. 2013), les télomères recombinent beaucoup, la recombinaison est plus rare aux péricentromères et quasiment absente au niveau du centromère. La carte de Marey a donc la forme d'une fonction logit. A l'inverse, chez *Arabidopsis* et le peuplier (Salomé et al. 2011; Apuli et al. 2020), le taux de recombinaison est quasiment constant le long des chromosomes, et il en résulte une carte de Marey linéaire. L'hypothèse principale pour expliquer cette variabilité « topologique » du profil de recombinaison est que l'appariement des chromosomes homologues via le complexe synaptonémal est initié aux régions distales et se poursuit ensuite progressivement vers les régions proximales (Higgins et al. 2012). Des CO interférents (forme majoritaire de CO) se forment précocement aux régions distales, et limitent la formation d'autres CO dans les régions plus proximales. Chez les espèces à grands chromosomes (blé, maïs), cela produit un taux de recombinaison élevé aux télomères et faible aux péricentromères. Chez es espèces à petits chromosomes (A*rabidopsis*, le peuplier, le riz), les bras de chromosomes sont vraisemblablement trop courts pour observer ce phénomène d'extinction de la recombinaison aux péri-centromères. En conséquence, ces espèces montrent un profil de recombinaison plus homogène, sauf au niveau du centromère (Choi et al. 2013; Drouaud et al. 2013; Marand et al. 2019).

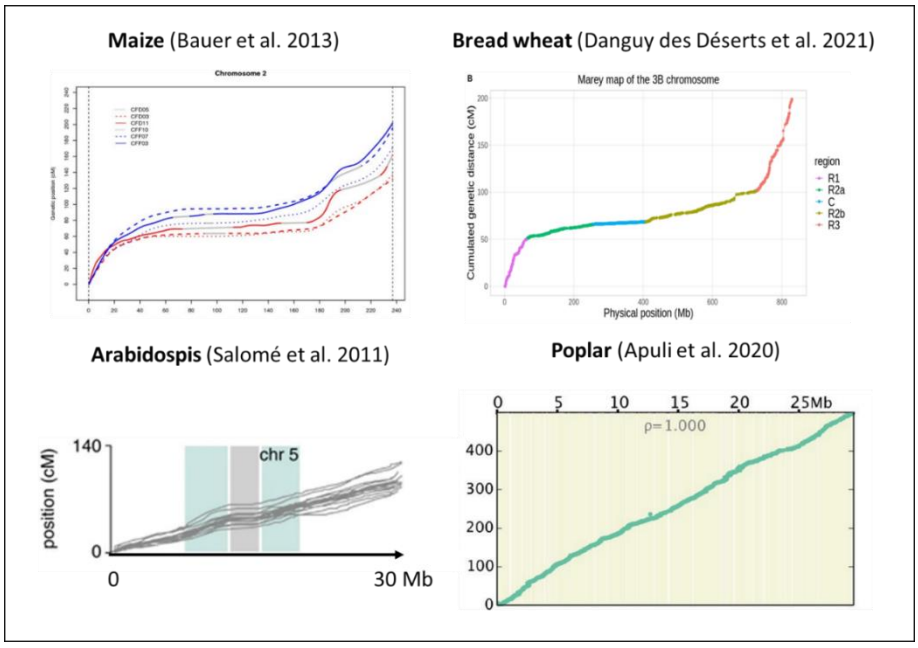


Figure 24 : Variabilité « topologique » du profil de recombinaison selon différentes espèces

D’autre part, il y a une variabilité de la recombinaison intra-espèce. Le profil de recombinaison varie entre individus en terme de positionnement des CO le long du génome (Schwarzkopf et al. 2020; Marand et al. 2019), mais aussi en termes de nombre de CO moyen par méiose (Jordan et al. 2018, Gardiner et al. 2019). Des mutations dans les voies de répression des CO, impliquant notamment l’hélicase atFANCM (Crismani et al. 2012; Girard et al. 2014), les hélicases du groupe RECQ4 (Séguéla-Arnaud et al. 2015) et la protéine FIDGETIN-LIKE-1 (Girard et al. 2015), augmentent aussi le taux de recombinaison. Les mutants *recq4* ont des cartes génétiques environ 3 fois plus longues que les cartes génétiques des témoins non mutés (Mieulet et al. 2018).

Ainsi, on peut se demander quel est l’apport des CSC basés sur la variance de la descendance selon ces trois caractéristiques de la recombinaison : l’aspect « topologique » du profil de recombinaison, le taux de recombinaison moyen, où la variabilité du profil de recombinaison entre individus.

Nous avons calculé la variance de la descendance des populations décrites au **Chapitre 3** (n=20 populations sélectionnées et n=20 populations non sélectionnées) avec plusieurs cartes génétiques. Les effets QTLs sont supposés connus (scénarios TRUE). La carte génétique de la population West Europe (WE) publiée par (Danguy des Déserts et al. 2021) est utilisée comme référence pour calculer la variance de la descendance (${\sigma^2}_{WE}$). Alternativement à la carte WE, nous avons calculé la variance de la descendance pour quatre autres cartes génétiques :

- la carte génétique de la population East Asia (EA), population dont le profil de recombinaison est le plus différencié de celui de WE lorsqu’on analyse la diversité génétique naturelle du profil de recombinaison (**Chapitre 2**). La carte de Marey de EA montre cependant de faibles différences avec WE (**Figure 25**), mais ces deux cartes montrent la variabilité maximale du profil de recombinaison que l’on peut observer chez le blé tendre.
- une carte génétique où le taux de recombinaison est constant le long du génome (nommée « constant »),
- une carte génétique 10 fois plus longue que WE (nommée « WE*10 »), dont les taux de recombinaison par intervalle ont simplement été multipliés par 10.
- Une carte génétique où chaque fréquence de recombinants vaut r = 0.5 pour toutes les paires de loci (nommée « free rec »).

Pour chacune des populations, la variance des écarts-types des descendants var (σ_{map}) pour chacune des quatre cartes génétiques alternatives (map = EA, constant, WE*10 ou free rec) a été calculée, puis divisée par la variance des écarts-types dans la population WE var (σ_{WE}).

Le graphe de gauche sur la **Figure 25** donne la carte de Marey pour trois types de cartes génétiques (WE, EA et constant). Le graphe de droite de la **Figure 25** donne l’augmentation ou la diminution du ratio entre var (σ_{map}) et var (σ_{WE}) selon la carte génétique « map » utilisée.

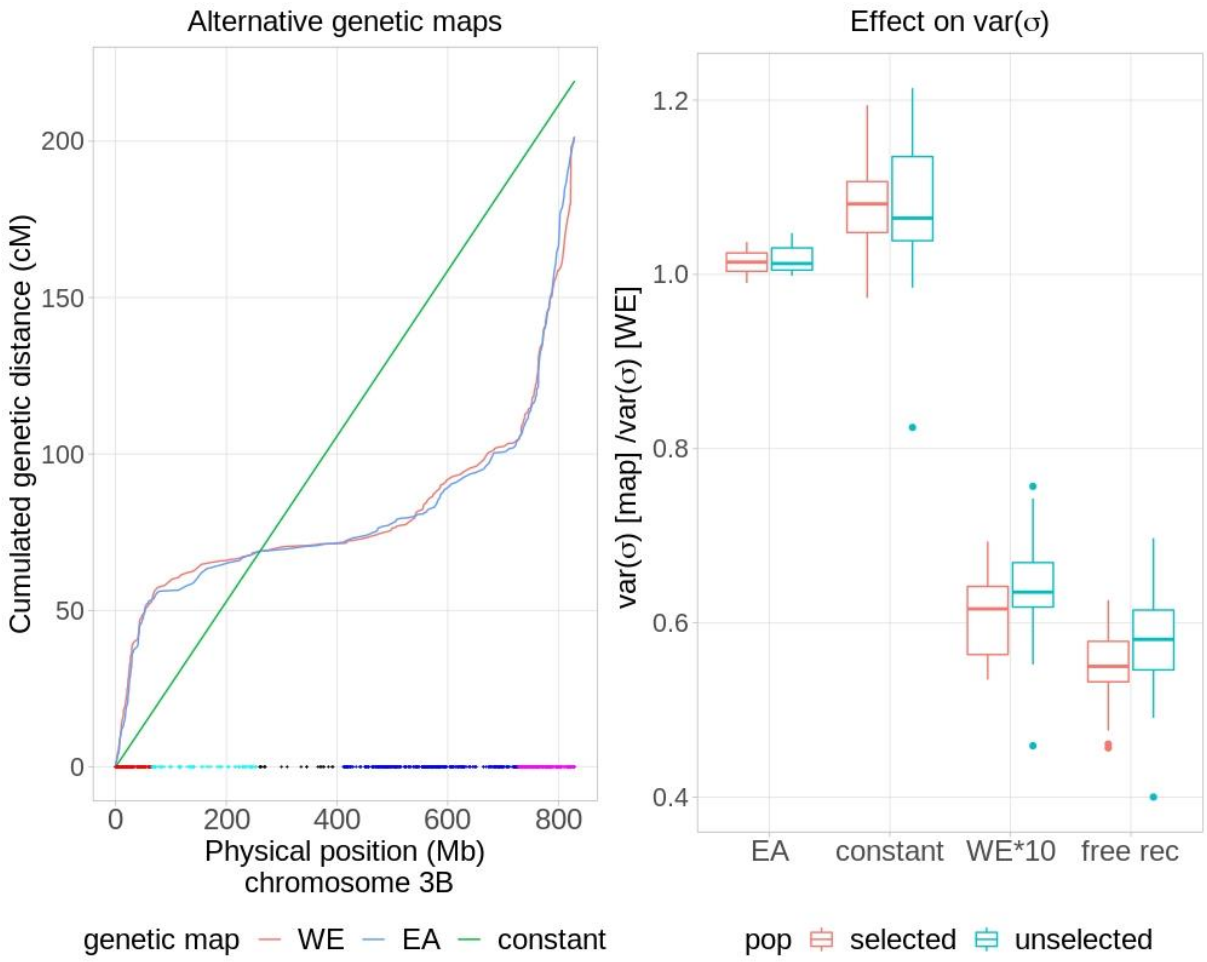


Figure 25 : Impact du profil de recombinaison sur la variabilité des écart-types des descendants

Gauche : *Carte génétiques alternatives dans le calcul de la variance de la descendance d'un croisement. WE : population West Europe ; EA : population East Asia ; constant : taux de recombinaison constant le long du génome. Exemple sur le chromosome 3B du blé tendre. Les points colorés dans la partie inférieure du graphique indiquent la position des marqueurs génotypés dans la population INRAE/AO utilisée au chapitre 2. Points rouges : marqueurs de la région télomérique R1 ; points cyan : péricentromère R2a ; points noirs : centromère C ; points bleus : péricentromère R2b ; points violets : télomère R3. Les bornes de chaque région chromosomique (R1, R2a, C, R2b, R3) étant définies par Choulet et al. (2014).*

Droite : *Impact de ces cartes génétiques sur la variabilité des écart-types des descendants. La variance des écarts-types des descendants dans une population ayant un profil de recombinaison alternatif est divisé par la variance des écarts-types des descendants dans une population ayant un profil de recombinaison WE. Un point donne le ratio pour une population. En rouge, les populations sélectionnées (n=20), en bleu les populations non sélectionnées (n=20). Les effets et les positions des QTLs sont connus.*

Les variances des écarts-types des descendants ne sont pas différentes dans la population WE ou dans la population EA (var(σ_{WE}) ~ var(σ_{EA})), ce qui est attendu au vu de la faible différenciation des profils de recombinaison. Par contre, la variance des écarts-types augmente de 20% dans une population où le taux de recombinaison est constant (var(σ_{WE}) < var($\sigma_{constant}$), et diminue d'environ de 30 à 50% dans une population où le taux de recombinaison est plus élevé

(var(σ_{WE}) > var($\sigma_{WE*10\ ou\ free\ rec}$).

Ce dernier résultat est cohérent avec une diminution de la variance des écarts-types des descendants lorsque les descendants d'un croisement sont produits par autofécondations successives (RILs F5) par rapport à une haplo-diploïdisation (**Figure 26**). Les génomes parentaux sont plus recombinés dans les descendants RILs que dans les descendants HDs. La variance des écart-types des descendants HDs est environ 20% supérieur par rapport à une population de RILs F5 (var(σ_{HDs}) > var($\sigma_{RILs\ F5}$), **Figure 26**). Limiter la recombinaison revient donc à augmenter la variance des écart-types var(σ).

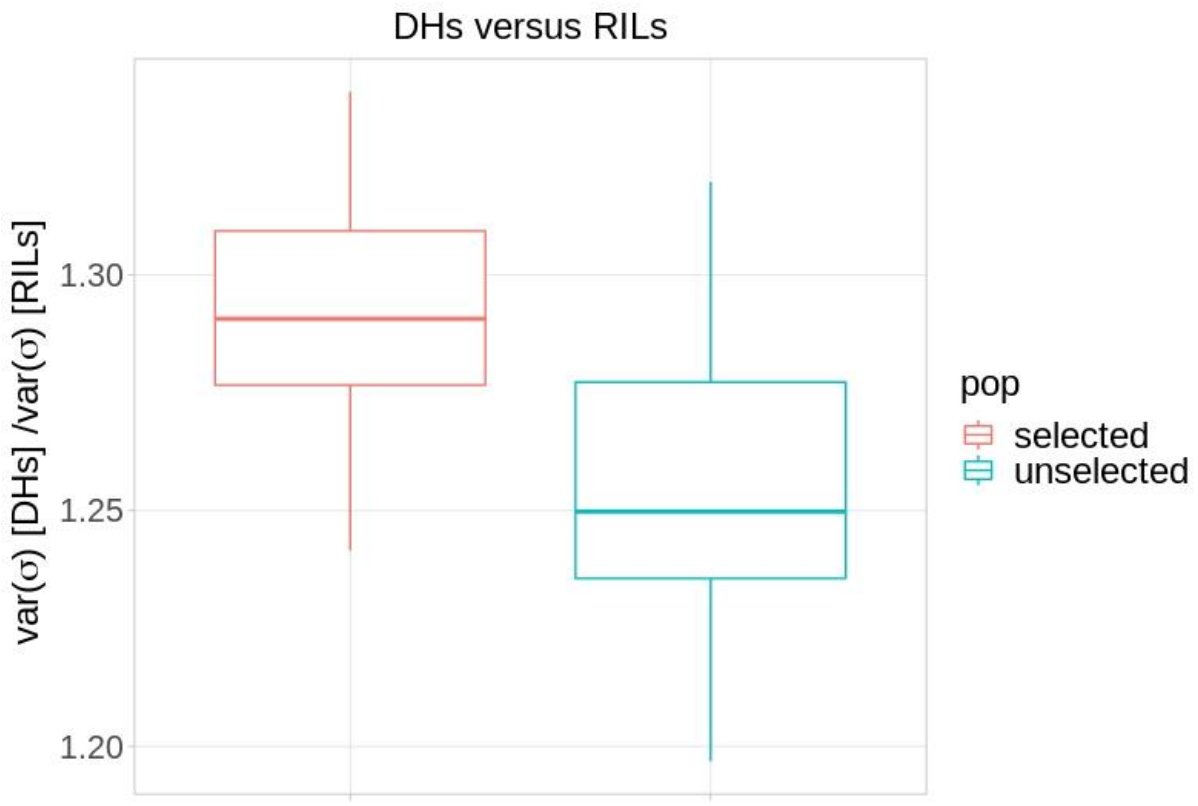


Figure 26 : Impact du type de descendants (RILs ou DHs) sur la variabilité des écart-types des descendants.

La variance des écart-types de chaque croisement est calculée avec le profil de recombinaison WE.

En conclusion, la variance des variances des descendants dépend de la « topologie » du profil de recombinaison, ainsi que du taux de recombinaison moyen. La variance des variances des descendants est un paramètre déterminant dans le bénéfice supplémentaire des critères de choix de croisements basés sur la variance de la descendance par rapport au choix des croisements sur la valeur génétique des parents. Ainsi, l'usage de critères de choix de croisements basés sur la variance de la descendance est plus pertinent chez les espèces (riz, peuplier) avec un profil de recombinaison homogène pour les marqueurs génotypés (les péricentromères recombinent autant que les télomères) et/ou avec une faible fréquence de recombinants dans la descendance. Pour comprendre l'impact de la recombinaison sur la variance des écart-types des descendants, nous sommes en train de chercher à exprimer la variance des variances des écart-types en fonction de la fréquence de recombinants r.

A noter que ces résultats ne permettent pas de dire quel profil de recombinaison est associé au plus fort gain génétique. Une autre question porte donc sur l'impact du profil de recombinaison sur le gain génétique.

Pour chaque carte génétique (EA, constant, WE*10 et free rec), nous avons donc l'écart-type de la descendance σ_{map} de chaque croisement, ce qui nous a permis de calculer l'UC1 de chaque croisement : $UC1_{map}$. Pour chaque croisement et dans chaque population, nous avons comparé les $UC1_{WE}$ et les σ_{WE} avec les $UC1_{map}$ et les σ_{map}. Pour les comparer nous avons calculé la moyenne des « bais » comme la moyenne de la quantité $\frac{parametre_{map} - parametre_{WE}}{parametre_{WE}}$ (avec paramètre = σ ou UC1) sur l'ensemble des croisements, mais aussi la corrélation entre les σ et UC1 au sein de chaque population. La **Figure 27** donne les biais et les corrélations pour les 10% meilleurs croisements (10% meilleurs PM) de la population.

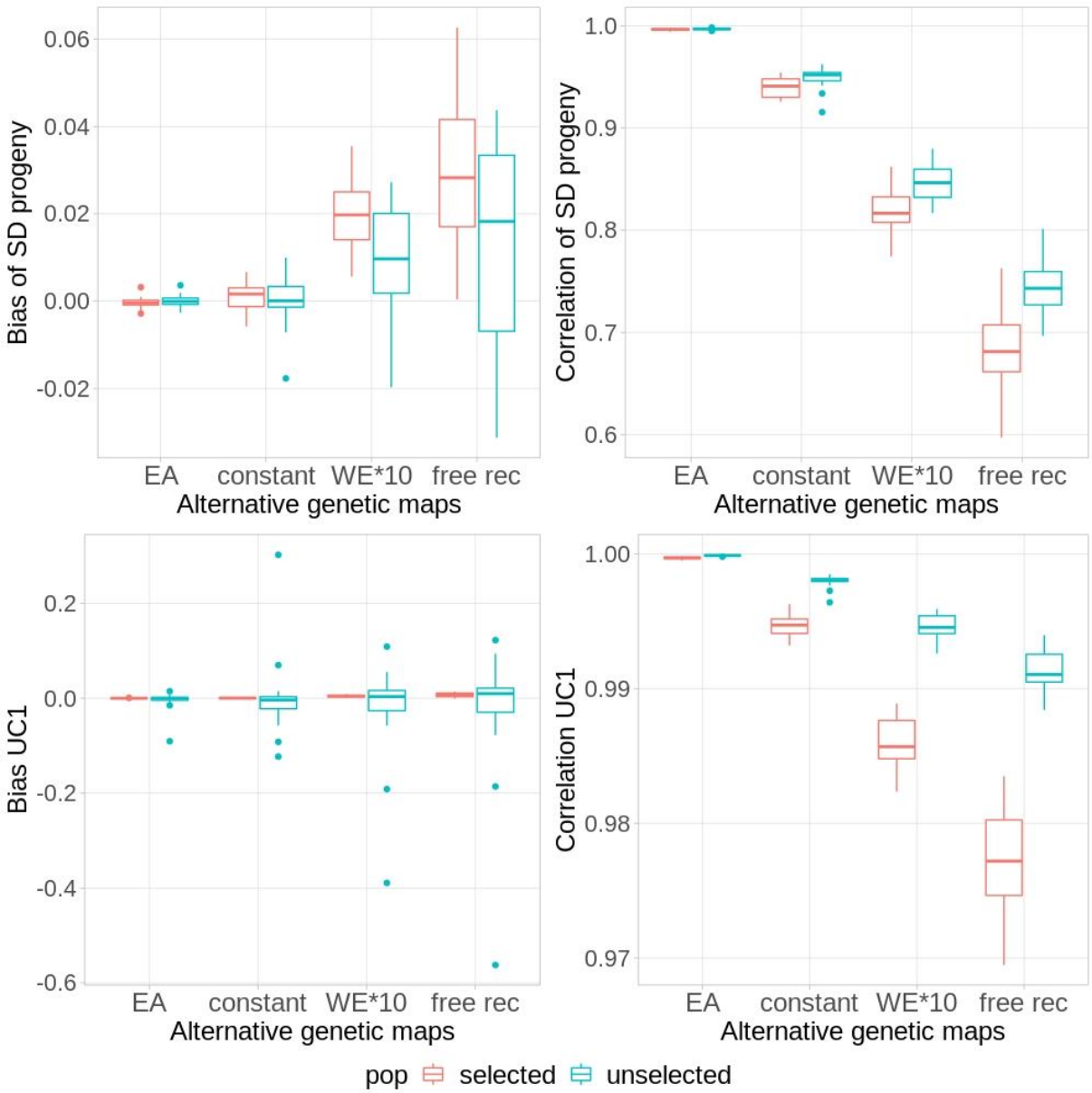


Figure 27 : Impact du profil de recombinaison sur le classement des meilleurs croisements

***Panneaux supérieurs** : Biais et corrélation de l'écart-type de la descendance calculés 1) avec une carte génétique alternative 2) avec la carte WE.*

***Panneaux inférieurs** : Biais et corrélation pour l'UC1. Un point = une valeur pour l'une des populations. En rouge, les populations sélectionnées, en bleu les populations non sélectionnés. Chaque point donne une moyenne sur les 10% meilleurs croisements par population (n=20 points par boxplot), présélectionnés sur la valeur génétique des parents.*

Les biais et corrélations des écarts-types des descendants et de l'UC1 calculés sur les profils WE et EA sont respectivement centrés sur 0 et proches de 1, avec peu de variabilité. Ce résultat indique que la variabilité « naturelle » du profil de recombinaison chez le blé ne suffit pas à générer des différences importantes dans le classement des meilleurs croisements.

Les profils de recombinaison « fortement recombinants » (WE*10 et free rec) sont associés à une plus grande variance de la descendance. Le biais atteint 6% sur l'écart-type de la descendance par rapport à la population WE, et la corrélation entre les écarts-types « alternatifs » et les écarts-types WE chute jusqu'à 0.6. Cependant, ces différences significatives n'auraient à priori qu'un impact limité sur le gain génétique car le biais de l'UC1 pour ces cartes alternatives est centré sur 0. Le « bonus » sur la variance de descendance permis par une augmentation de la recombinaison n'est vraisemblablement pas suffisant pour faire augmenter l'UC1 chez les 10% meilleurs croisements. De plus, la corrélation entre les UC1 est de l'ordre de 0.97, ce qui indique que le classement des meilleurs croisements est stable, même si la recombinaison augmente. Seule une optimisation du plan de croisements permettrait de pouvoir véritablement trancher la question du gain génétique.

Les résultats d'autres études montrent que l'augmentation de la recombinaison n'est pas déterminante sur le gain génétique à court terme pour un caractère polygénique. Par exemple, les résultats de Tourrette et al. (2019) montrent qu'effectivement augmenter la recombinaison a un impact limité sur le gain à court terme et ceux de McClosky et Tanksley (2013) montrent que le gain génétique court terme dépend surtout de la ségrégation des chromosomes. Gonen et al. (2017) montrent qu'une augmentation de la recombinaison augmente légèrement le gain à court terme. Les bénéfices semblent surtout concerner le gain génétique à long terme. Plusieurs études (Battagin et al. 2016; Tourrette et al. 2019) rapportent notamment que la diversité génétique dans les populations fortement recombinantes est plus élevée, ce qui augmente les possibilités de gain génétique à long terme.

En conclusion pour ces résultats supplémentaires, la variabilité des caractéristiques de la recombinaison (topologie, taux de recombinaison moyen, variabilité du profil de recombinaison) n'a vraisemblablement que peu de conséquences sur un gain génétique maximal à une génération. Une alternative à la modification globale du profil de recombinaison de la population de sélection serait de mieux comprendre le déterminisme de la position des crossing-over pour réaliser des recombinaisons ciblées.

Chapitre IV : Discussion générale

Chapitre IV : Discussion générale

IV.1 Apport de l'information de recombinaison dans l'optimisation des plans de croisements

IV.1.1 Les limitations dues au manque de précision dans l'estimation des effets des marqueurs

L'amélioration variétale cherche à maximiser la valeur génétique des meilleurs descendants. Produire des descendants extrêmes nécessite de croiser des parents génétiquement complémentaires, avec une probabilité de cumuler les allèles favorables par recombinaison méiotique suffisamment élevée. Une manière d'estimer la complémentarité génétique des parents est de calculer leur distance génotypique ou phénotypique. Cependant, ces méthodes ne garantissent pas une grande variance de la descendance des croisements retenus.

Les informations des cartes génétiques denses et les modèles de prédiction génomique permettent désormais de prédire en amont la variance de la descendance de tous les croisements possibles. La prédiction de la variance de la descendance n'entraine aucun coût supplémentaire si le programme de sélection utilise déjà la prédiction génomique en routine. Cependant, comme illustré au **Chapitre 3**, une mauvaise estimation de la variance limite le bénéfice de la prise en compte de la variance de la descendance dans le choix des meilleurs croisements. A noter que dans certains scénarios où les effets des marqueurs sont très mal estimés et où il n'y a pas de contraintes sur les contributions parentales, les critères de Sélection de croisements (CSC) basés sur la variance de la descendance sont même légèrement sous-performants par rapport au simple choix des croisements sur la moyenne des valeurs génétiques des parents (PM) (**supplementary Figure 10**).

La qualité d'estimation des variances est donc un aspect essentiel du déploiement des CSC en sélection. La qualité des estimations des variances passe nécessairement par la qualité des estimations des effets des marqueurs. Certains modèles de prédiction génomique sont plus adaptés à cette prédiction de la variance de la descendance par un meilleur traitement du bruit statistique associé aux erreurs d'estimation des effets des marqueurs. En particulier les modèles Bayésiens, tels que le modèle Bayesian Lasso, qui sélectionne les marqueurs expliquant une part importante de la variance génétique, ou le Posterior Mean Variance model (PMV) qui prend en compte l'incertitude sur les effets des marqueurs (Lehermeier et al. 2017). D'autres modèles pertinents sont proposés par Hofheinz et Frisch (2014). Les auteurs comparent la précision de plusieurs modèles de prédiction génomique pour estimer des effets marqueurs. Dans cette étude, les modèles les plus précis sont des modèles Ridge Regression où le shrinkage est ajusté pour chaque marqueur indépendamment grâce à des analyses préalables (par exemple, en estimant

préalablement la variance expliquée par chaque marqueur par une ANOVA). Alternativement, l'estimation des effets des marqueurs semble meilleure et plus portable dans des populations très recombinantes (Tourrette et al. 2019). Ils observent que les prédictions génomiques dans une population fortement recombinante (nombre plus élevé de CO par méiose) sont relativement plus précises pour des populations génétiquement distantes comparé à des populations moins recombinantes, probablement car les blocs de DL sont plus petits, et donc que les marqueurs en DL avec des QTLs estiment mieux leurs effets. Cependant, cela nécessite d'augmenter fortement la recombinaison dans la population d'entrainement, en utilisant par exemple des mutants comme géniteurs (lignées Hyper rec de (Mieulet et al. 2018). Dans la même perspective, l'estimation des effets des marqueurs dans des panels de diversité ou dans de populations très recombinées (populations MAGIC, Mackay et al. 2014) présentant une grande diversité haplotypique permettrait une meilleure portabilité des modèles

La comparaison de modèles de prédiction génomique pour prédire la variance d'une descendance réelle (phénotypique non simulée) n'a jamais été testée dans la littérature à ma connaissance. Pour les modèles les plus prometteurs (Bayesian Lasso et PMV), seules deux études ont utilisé ces modèles pour comparer variance prédite et variance réelle. Dans l'étude de Tiede et al. (2015), la corrélation entre la variance prédite par une modèle Bayesian Lasso et la variance observée sur une descendance réelle de croisement biparentaux de maïs est de l'ordre de 0.6. Wolfe et al. (2021) utilisent le modèle Posterior Mean Variance pour prédire la variance de la descendance de croisements de manioc (*cassava*), et la corrélation entre variance réelle et variance prédite n'excède pas 0.2. A noter que les variances phénotypiques et les variances génétiques estimées par prédiction génomique sont deux estimations de la vraie variance de la descendance. Aucune des deux ne peut être considérée comme une référence. Pour les variances estimées par prédiction génomiques, l'héritabilité du caractère, mais aussi la qualité de la population d'entrainement, influence la qualité des estimations des effets des marqueurs et donc l'estimation des variances des descendants. Chez les plantes, les forts effets de l'environnement ou d'interactions entre génotypes et environnements limitent probablement les valeurs de corrélation. Une meilleure prise en compte de ces effets sera vraisemblablement un levier pour améliorer la précision de la prédiction des variances.

Mesurer la précision des estimateurs de la variance de la descendance est une entreprise coûteuse. Par exemple, en blé tendre, cela implique de tester une certaine diversité de croisements (des croisements avec un faible variance, des croisements avec une forte variance), mais aussi de conserver de mauvais descendants (afin de capter toute la variance du couple) jusqu'à un stade avancé, puis obtenir une estimation précise de la valeur génétique de ces bons et mauvais descendants. Obtenir une estimation de la valeur génétique pour le rendement nécessite de multiplier suffisamment de semences pour faire plusieurs parcelles d'essai, et éventuellement de mesurer le phénotype dans plusieurs environnements. Alternativement, il est possible de mener une analyse rétrospective sur un programme de sélection pour connaître les caractéristiques des

couples ayant produit des variétés élites par le passé (moyenne, variance de la descendance, (Abed et Belzile 2019; Jean et al. 2021). Il s'avère que dans leur matériel de sélection, les couples qui ont produit des variété élites, ou au moins des lignées qui ont été sélectionnées jusqu'à un stade avancé du processus de sélection, montrent surtout une forte moyenne des valeurs génétiques et une variance de la descendance d'une amplitude faible ou commune (exemple **Figure 28**). Il est possible que les estimateurs de variance soient trop imprécis pour permettre de conclure que les croisements avec une forte variance ne donnent jamais de descendants transgressifs. Par ailleurs, il ne serait pas surprenant que les croisements aient surtout été choisis de manière classique, c'est-à-dire sur la moyenne des valeurs génétiques des parents, en imposant une certaine distance génétique entre les parents. Ce choix de croisements ne permet pas a *priori* d'explorer les croisements avec une descendance très variable.

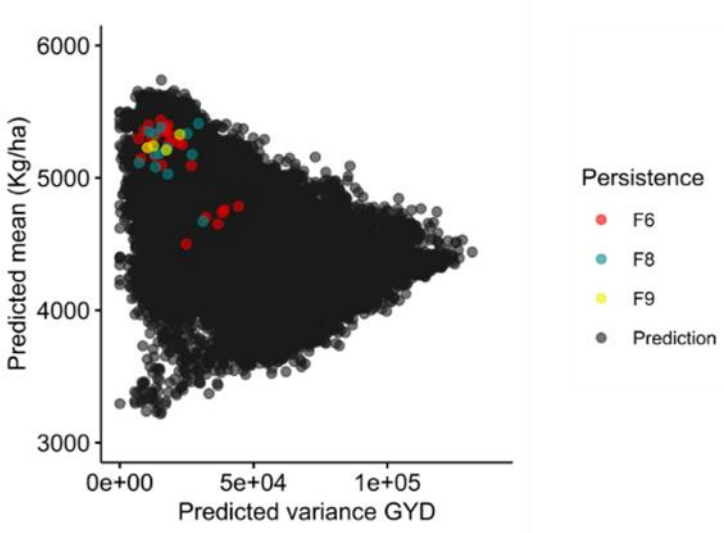


Figure 28 : Analyse rétrospective des croisements qui ont produit des lignées élites dans un programme de sélection orge

En abscisse, la variance de la descendance, en ordonnées l'espérance de la descendance. Les points colorés indiquent le stade d'avancement des descendants sélectionnés.

Abed et Belzile (2019)

La plupart des études qui ont cherché à prédire la variance d'une descendance, réelle ou simulée, rapportent que les variances de la descendance estimées avec la prédiction génomique sous estiment les estimateurs phénotypiques des variances de la descendance. Or, une sous-estimation de la variance s'accompagne d'une réduction du ratio entre variabilité des écart-types des descendances et variabilité des espérances des descendance var(σ)/var(PM). A l'instar de nombreux auteurs (Zhong et Jannink, 2007, Lehermeier et al. 2017, Lado et al. 2017, Bijma et al.

2020), le **Chapitre 3** met en évidence que l'intérêt de sélectionner un croisement avec des critères qui prennent en compte la variance de sa descendance dépend de ce ratio. Un faible ratio signifie que la supériorité des descendants d'un croisement s'explique surtout par la supériorité des valeurs génétiques des parents, plutôt que par la variance de la descendance. En conséquence, la variance de la descendance a moins d'impact sur la valeur du CSC, et donc le CSC favorise surtout les croisements avec une forte espérance, d'où une réduction des différences dans le plan de croisement entre un CSC basé sur la variance de la descendance et PM. Ce ratio est spécifique de chaque caractère d'intérêt, de chaque population parentale et dépend du profil de recombinaison. Le ratio augmente chez populations sélectionnées et les populations structurées. Entre les populations non sélectionnées et sélectionnées, ce ratio est doublé dans notre **Chapitre 3**, passant de 1% à 2%. Les populations structurées de Lehermeier et al. (2017) affichent un ratio moyen de 14%. Un profil de recombinaison homogène le long du génome permet d'augmenter le ratio par un facteur 0.2, et le fait de produire des descendants HDs au lieu de descendants RILs (donc en recombinant moins les génomes parentaux) augmente le ratio d'un facteur 0.3. Ces résultats illustrent bien que la pertinence des CSC en sélection est spécifique de chaque matériel de sélection.

On peut imaginer qu'en-dessous d'un certain ratio seuil, cela ne vaut plus la peine de calculer la variance des descendants et calculer l'espérance de la descendance PM devient suffisant. Dans notre étude par simulation, ce ratio seuil est compris entre 0.4% et 0.5%. Ce ratio de 0.4% a été observé dans les populations sélectionnées, dans le scénario où les effets des marqueurs sont les plus mal estimés. Dans ces scénarios, les CSC basés sur la variance de la descendance n'apportent quasiment aucun gain génétique par rapport à PM. Par contre, dans les populations non sélectionnées, le ratio était de 0.5% et les CSC apportent un bénéfice par rapport à PM. Bijma et al. (2020) montrent que le bénéfice d'un CSC qu'ils proposent, basé sur la variance de la descendance, dépend de l'intensité de sélection. Plus l'intensité de sélection est élevée, plus le ratio seuil est petit.

IV.1.2 Impact de la variabilité du profil de recombinaison pour un usage en sélection

La variabilité génétique du profil de recombinaison a été évalué au **Chapitre 2**. Les cartes génétiques des deux groupes génétiques les plus différenciés chez le blé tendre, le groupe génétique européen et le groupe génétique asiatique, montrent des différences significatives. Cependant, cette variabilité génétique n'est pas assez forte pour modifier le classement des meilleurs croisements. L'importance de la variabilité génétique du profil de recombinaison, ou au contraire l'aspect négligeable de cette variation, pourrait être confirmé en essayant de prédire les variances de descendance réelle de blé tendre en utilisant différents vecteurs de recombinaison.

Dans la perspective où la précision du vecteur de recombinaison n'impacte peu ou pas le classement des meilleurs croisements par un CSC basé sur la variance de la descendance, on peut se demander s'il est encore nécessaire d'estimer des cartes génétiques spécifiques de chaque population chez le blé tendre pour des applications directes en sélection.

En dehors de la prédiction de la variance de la descendance, une autre application courante des cartes génétique en sélection est la détection de QTL ou la recherche de marqueurs en DL avec des QTLs pour la sélection assistée par marqueur. Or, le développement du génotypage haut-débit ou du séquençage ciblé de grand fragments d'ADN, associé à la publication croissante d'assemblage de génomes, sur lesquels les marqueurs peuvent être positionnés, vont probablement diminuer l'intérêt des cartes génétiques pour la détection de QTLs ou la sélection assistée par marqueur dans les années à venir.

Cependant, les cartes génétiques resteront néanmoins toujours utiles chez le blé tendre. En effet, les variations structurales sont fréquentes (Cheng et al. 2019; Zhou et al. 2020; He et al. 2019), ce qui limite peut-être la portabilité des assemblages publiés. Par ailleurs, le génome du blé tendre est polyploïde, avec des séquences fortement conservées entre les trois génomes (Juery et al. 2021) ce qui complexifie beaucoup les assemblages de carte physique. Les cartes génétiques peuvent servir à assembler des génomes, en permettant d'ordonner et d'orienter les scaffolds de manière à limiter la distance génétique cumulée le long des chromosomes (Choulet, et al. 2014; Deokar et al. 2014; Mitros et al. 2019).

IV.1.3 Proposition de nouveaux critères de sélection de croisements

Les meilleurs croisements sont ceux qui ont à la fois une forte espérance et une forte variance dans la descendance. Il existe une certaine diversité de critères de choix de croisements, qui exploitent différemment la variance de la descendance. A un extrême, le critère PM trie les croisements sur la valeur génétique des parents, en ignorant la variance de la descendance. A l'autre extrême, le critère Optimal Haploid Value (OHV) calcule la valeur du meilleur individu qui puisse être dérivé d'un croisement, sans considération sur la probabilité d'obtenir cet individu. Entre ces deux extrêmes, une variété de critères a été proposée : espérance des q meilleurs descendants (UC), probabilité de produire un descendant supérieur à un seuil (PROBA), espérance du meilleur descendant parmi D descendants (EMBV).

L'inconvénient de ces critères est qu'ils s'appliquent individuellement à chaque couple, et qu'ils ne prennent pas en compte le plan de croisements dans son ensemble. L'étape suivante consisterait donc à maximiser la valeur génétique d'une fraction supérieure des descendants issus de l'ensemble du plan de croisement. Par exemple, Bijma et al (2020) proposent une CSC qui est proportionnel à la probabilité pour la descendance d'un taureau de dépasser le seuil de troncature

correspondant à la fraction des q meilleurs descendants de la génération suivante. Identifier ce seuil de troncature n'est pas trivial car la distribution des valeurs des descendants de la prochaine génération est un mélange de distributions (une distribution par taureau) pondérées par les effectifs relatifs des différentes descendances. Bijma et al. (2020) calculent ce seuil analytiquement avec plusieurs hypothèses. Ils supposent notamment que les mères ont une variance gamétique égale à l'espérance de la variance gamétique de la population ($0.25*\sigma_a$, Mrode 2005), et la variance intra-famille se calcule comme $\sigma^{2taureau}_{gametique}$ + $0.25*\sigma_a$. Par ailleurs, il est supposé que les taureaux soient accouplés aux 10% meilleurs mères.

Ce CSC permet d'optimiser le choix des reproducteurs pour maximiser la fraction supérieure de l'ensemble de la descendance qui va être sélectionnée. Le critère PROBA que nous avons proposé repose sur la même idée d'une sélection par troncature globale, contrairement aux autres CSC (UC, EMBV). En nous inspirant du critère proposé par Bijma et al. (2020), nous sommes en train de développer un critère, nommé UC3, qui permet de maximiser l'espérance des q meilleurs descendants du plan de croisements, en prenant en compte la variance des descendants de chaque couple (à la différence de Bijma et al. (2020) qui simplifient la variance gamétique des mères). La difficulté dans ce critère sera de définir conjointement le vecteur des descendants et le seuil de troncature correspondant au q meilleurs descendants du plan de croisements.

Nous envisageons aussi de généraliser le critère EMBV qui se calcule pour un couple comme l'espérance des K meilleurs descendants parmi D descendants alloués au couple, à un « EMBV du plan de croisements » (EMBV-MP, avec MP pour Mating Plan), qui donnerait l'espérance des K meilleurs descendants du plan de croisements parmi D descendants.

Les critères UC3 et EMBV-MP présentent l'avantage d'être potentiellement pertinents pour la sélection récurrente. Cependant, ces critères sont numériquement complexes à calculer et il faudra vraisemblablement un algorithme génétique pour optimiser ces plans de croisements. Ainsi, des critères plus simples peuvent leur être préférés (UC, PROBA). La simplicité d'un critère est particulièrement pertinente si de nombreuses simulations doivent être lancées comme lorsqu'il s'agit de mesurer les conséquences d'un CSC sur le gain long terme.

Nous avons montré que la pré-sélection des parents chez les plantes ne pose pas de problème dans du matériel élite, probablement car le ratio var(σ)/var(PM) est faible. Mais d'autres publications ont montré que la covariance entre la valeur génétique moyenne des parents et la variance de leurs descendants est souvent négative chez les plantes. Sous condition que le ratio var(σ)/var(PM) soit suffisamment élevé, il y aurait donc un intérêt potentiel à croiser des individus de valeur génétique intermédiaire mais complémentaires. Il serait donc intéressant de tester l'intérêt des différents CSC dans des populations plus diversifiées. Par exemple, en pré-breeding, les lignées ressources (L) sont croisées avec des lignées élites (H). Il est possible que la variance des descendants soit très élevée et très variable dans cette population H*L (**Figure 17**, gauche), ce qui permettrait d'obtenir un ratio var(σ)/var(PM) intéressant.

IV.1.4 Optimiser le plan de croisements pour plusieurs caractères

Le **Chapitre 3** consistait à maximiser le progrès génétique d'un caractère polygénique tel que le rendement sur une génération en optimisant les plans de croisements. Dans la réalité, les nouvelles variétés doivent présenter des performances minimales pour d'autres traits d'intérêt. Par exemple, chez le blé, le rendement est le premier critère plébiscité par les agriculteurs (France Agri Mer 2015), mais les caractéristiques agronomiques, la tolérance aux stress biotiques et abiotiques, ainsi que les caractéristiques technologiques de la farine sont aussi recherchées. Cependant, les différents caractères d'intérêt peuvent montrer des corrélations négatives, par exemple entre le taux de protéine et le rendement, ce qui complexifie le choix du plan de croisements.
Une première façon de construire un plan de croisements pour une amélioration multi-caractères est de faire une pré-sélection des croisements (Jean et al. 2021). Par exemple, éliminer d'avance tous les croisements impliquant deux parents sensibles à une maladie, ou dont la descendance présente une distribution prédite centrée sur des valeurs trop basses du taux de protéine. Cette technique est illustrée dans l'analyse rétrospective de Jean et al. (2021) chez le soja. Les auteurs montrent que les croisements ayant abouti à des descendants sélectionnés sont ceux qui présentent une moyenne des valeurs génétiques des parents supérieure aux autres croisements pour le rendement, et ce à l'intérieur d'une certaine fenêtre de précocité. Si éliminer un croisement (par exemple un croisement qui va produire un F1 hétérozygote pour un gène très important) peut paraître trop sévère, l'alternative est de ne calculer l'utilité du croisement que sur la fraction de descendants (RILs ou HDs) qui portent les gènes désirés. La distribution des valeurs génétiques pour un caractère (par exemple le rendement) en ne considérant que des descendants porteurs des gènes particuliers (par exemple, tolérance à une maladie) est réalisable facilement avec la simulation in-silico de descendants.

Une deuxième façon d'améliorer plusieurs caractères est d'utiliser des index pour sélectionner les croisements, c'est-à-dire calculer un CSC qui donne un poids économique différent à chaque caractère d'intérêt. Par exemple, (Wolfe et al. 2021) prédisent la distribution de la descendance pour deux index chez le manioc. Cette méthode requiert cependant de définir des pondérations entre les caractères d'intérêt ou bien de choisir le plan de croisements parmi la frontière de Pareto (**Figure 8**). Il serait intéressant de compléter notre algorithme d'optimisation des plans de croisement avec des UC multi-traits, grâce à une matrice de variance gamétique qui prenne en compte les effets joints des SNPs sur les caractères, et une optimisation qui considère des pondérations économiques.

Une troisième option suggérée par Yao et al. (2018), Wartha et Lorenz (2021); Neyhart et Smith (2019) pourrait être de sélectionner les croisements qui montrent la meilleure covariance entre les

deux caractères corrélés négativement. Par exemple, en blé tendre, la relation entre taux de protéines et rendement est négative. Mais certaines variétés présentent une déviation à cette relation négative moyenne, appelée « Grain Protein Deviation » (GPD, Monaghan et al. 2001; Oury et al. 2003). La corrélation négative entre caractères est théoriquement due à des facteurs pléiotropiques ou à la répulsion entre allèles favorables (Falconer et Mackay 1966). L'existence de variétés à déviation positive (par exemple, une teneur en protéine supérieure au taux moyen attendu pour ce niveau de rendement), suggère qu'il est possible de sélectionner un fond génétique en augmentant le niveau d'un premier caractère tout en limitant la perte sur le deuxième caractère. (Neyhart et Smith 2019) ont montré par simulation que la prise en compte de la corrélation négative entre la tolérance à la fusariose et la hauteur de plante lors du choix des croisements permet d'augmenter le gain génétique de l'ordre de 11-27%, en comparaison du simple choix des croisements sur la moyenne des caractères. La covariance entre deux caractères peut donc être considérée comme un caractère d'intérêt à maximiser dans le plan de croisement. Comme n'importe quelle sélection par troncation, il faudrait veiller à maintenir une diversité génétique sur l'un et l'autre des caractères d'intérêt.

La sélection directionnelle massale n'est pas la seule option pour améliorer un caractère. L'écophysiologie nous enseigne que les phénotypes complexes tels que le rendement sont la somme de plusieurs phénotypes et de leurs interactions. Ainsi, plutôt que d'utiliser les critères de choix de croisements pour maximiser le rendement dans une population de sélection, il est possible d'exploiter la variation génétique liée à ses différentes composantes. Par exemple, (Reynolds et al. 2021) proposent d'augmenter le rendement en utilisant la variation génétique associée aux composantes du rendement (assimilation, photosynthèse, harvest index).

Les ressources génétiques constituent une grande source de variabilité génétique pour améliorer différents caractères comme la tolérance aux stress biotiques et abiotiques. Mais le défi est de pouvoir phénotyper des ressources génétiques non adaptées à l'environnement cible. En effet, ces accessions présentent des allèles récessifs pour certains gènes majeurs impliqués dans la phénologie (hauteur, précocité) qui peuvent empêcher d'autres allèles favorables de s'exprimer. Pour évaluer leurs performances, Longin et Reif (2014) proposent de les croiser avec des lignées élites et de phénotyper les F1 avec les allèles dominants pour les gènes majeurs impliqués dans la phénologie. Pour certaines espèces comme le sorgho ou le maïs, la précision de prédiction de la valeur génétique de lignées exotiques (Yu et al 2016) ou de croisements avec l'UC ou l'OHV (Allier et al. 2019c), semble correcte sans passer par le phénotypage de F1.

Cependant, dissocier les allèles favorables des allèles délétères présents dans les lignées ressources nécessitent de nombreux rétrocroisements, ce qui encourage à développer une compréhension et une maîtrise du processus de recombinaison pour accélérer le cumul des allèles favorables.

IV.2 Exploitation de la recombinaison pour cumuler des allèles favorables et éliminer les allèles délétères

IV.2.1 Augmenter ou cibler la recombinaison

Deux approches sont couramment proposées pour accélérer le cumul des allèles favorables (synthèse dans (Taagen et al. 2020; Blary et Jenczewski 2019, **Figure 29**).

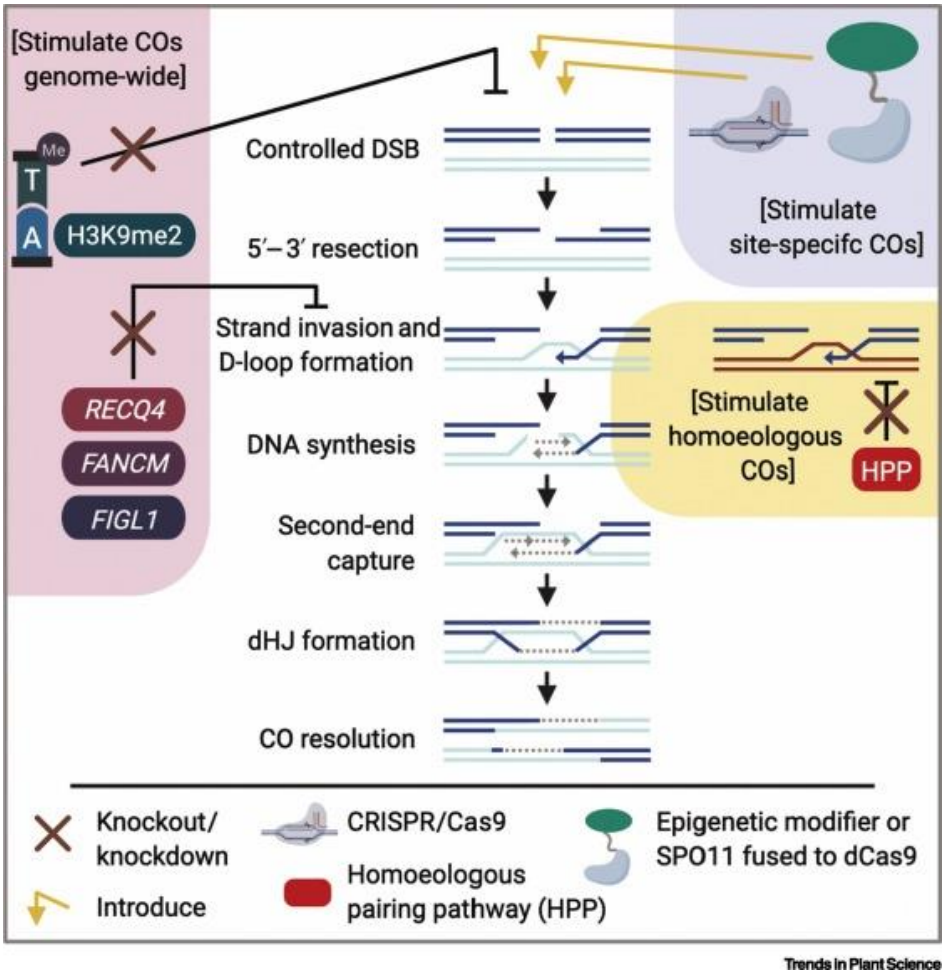


Figure 29 : Outils biotechnologiques pour augmenter ou cibler la recombinaison

Trois méthodologies peuvent modifier la recombinaison : 1) augmenter le nombre de crossing-over par méiose via des mutations sur les protéines impliquées dans la répression des crossing-over 2) augmenter la recombinaison entre génome homéologues via la mutation dans la voie métabolique HPP 3) cibler la recombinaison à des sites spécifiques de l'ADN en dirigeant les cassures double brins ou en modifiant le profil épigénétique.

Taagen et al. (2020)

La première approche consiste à augmenter significativement le nombre moyen de crossing-over (CO) par méiose en mutant des protéines impliquées dans des processus métaboliques répresseurs des CO. Les bénéfices de ces méthodes ont été testés par simulation (Battagin et al. 2016; McClosky et Tanksley 2013; Tourrette et al. 2019) et s'accompagnent d'un plus grand gain génétique (de l'ordre de 10-30%) à long terme. De nombreux gènes inhibiteurs de la formation des CO à forts effets ont été identifiés et sont fonctionnels chez plusieurs espèces de plantes (Mieulet et al. 2018). Cependant, obtenir ces mutants par des processus chimiques ou d'édition du génome est loin d'être facile techniquement et est encadré par la loi en Europe. Les mutations dans les gènes qui contrôlent la recombinaison peuvent s'accompagner des phénotypes indésirables, tel que la stérilité. Cependant, toutes les recombinaisons ne sont pas désirables. La recombinaison pour dissocier des allèles en répulsion est recherchée alors que la recombinaison pour dissocier des allèles en couplage ne l'est pas.

Une deuxième approche consiste à cibler la recombinaison à des endroits précis du génome. La formation induite de CO a déjà été réalisée expérimentalement. Par exemple, l'induction de CO a déjà été réalisée chez la levure à partir de protéines pouvant se lier à des séquences cibles de l'ADN (Sarno et al. 2017; Peciña et al. 2002). Le taux de recombinaison au niveau des séquences cibles a effectivement augmenté, sauf dans les régions où la chromatine est condensée. (Hayut, et al. 2017) ont effectivement pu induire des CO à des endroits précis chez la tomate, en utilisant la technologie d'édition génétique CRISPR/Cas9 pour induire la formation de cassures double brin DSB. Par contre, l'induction de DSB par CRISPR/Cas9 ne modifie pas le profil de recombinaison chez *Arabidopsis* (Yelina et al. 2021).

D'une manière générale, les DSB ne sont pas le facteur limitant de la formation des CO, et seulement 3 à 10% d'entre eux sont résolus en CO. Un autre enjeu est de pouvoir induire la formation de CO dans les régions où la chromatine est compactée (synthèse dans Taagen et al. 2020). Cela encourage à mieux comprendre le déterminisme de la position des CO, pour éventuellement identifier des leviers pour diriger la recombinaison. Le paysage chromatinien semble être une variable explicative importante du profil de recombinaison chez les plantes. (Underwood et al. 2018; Yelina et al. 2015) ont par exemple modifié le paysage chromatinien de *Arabidospis* en mutant des protéines impliquées dans la méthylation ou en supprimant certaines marques de méthylation. Cependant, les marques épigénétiques associés aux points chauds de recombinaison varient entre les plantes (Demirci et al. 2018), et une modification du paysage chromatinien peut engendrer des anomalies du développement.

Maitriser la recombinaison peut se révéler être une stratégie plus efficace pour cumuler les allèles favorables. Tourette et al. (2019) ont montré par simulation qu'une augmentation du taux de recombinaison en faveur des régions péri-centromériques permet un gain génétique plus élevé qu'une augmentation globale du taux de recombinaison. Par ailleurs, Gonen et al. (2017) ont montré par simulation qu'un déplacement des points chauds de recombinaison au cours de la vie

du programme de sélection s'accompagne d'un gain génétique et d'une meilleure sauvegarde de la diversité génétique. Bernardo (2017) et Ru et Bernardo (2019) proposent d'estimer les effets des marqueurs par prédiction génomique pour identifier les recombinaisons désirables dans les nouveaux descendants. Choisir la position des CO à chaque méiose permet au minimum de doubler le progrès génétique (augmentation d'un facteur 2 à un facteur 6 selon les espèces et le nombre de recombinaisons ciblées) en comparaison d'une recombinaison non contrôlée. Cependant, cela suppose que les estimations des effets des marqueurs soient fiables, et qu'il existe une technologie pour cibler la formation des CO.

IV.2.2 Evolution des profils de recombinaison chez les plantes

Les analyses du **Chapitre 2** ont permis de mettre en évidence une divergence du profil de recombinaison corrélée à la différenciation génétique des populations. Cependant, les profils de recombinaison du **Chapitre 2** ont été estimés à partir des patrons de déséquilibre de liaison, potentiellement biaisés par les forces évolutives. L'une des questions est de savoir si ces résultats sont cohérents avec ceux d'autres études chez les plantes. Nos résultats sont cohérents avec ceux de (Gardiner et al. 2019) sur le blé tendre. Ces auteurs ont utilisé une population NAM (pour Nested Association Mapping) de lignées recombinantes issues du croisement entre une lignée parentale commune et plusieurs lignées parentales. Les descendances issues des lignées parentales proches génétiquement ont des profils de recombinaison plus similaires que des lignées parentales génétiquement distantes (**Figure 30**).

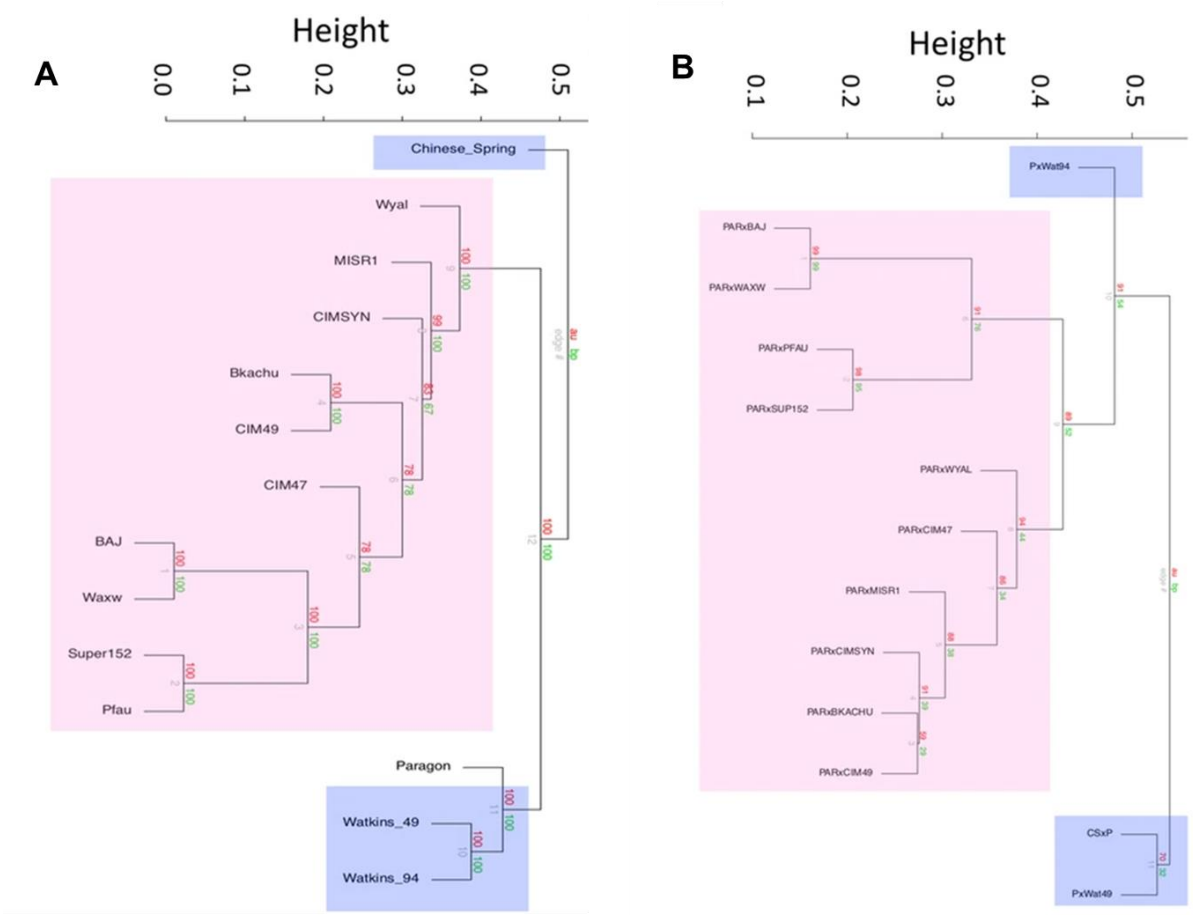


Figure 30 : Similarité du profil de recombinaison dans une population NAM de blé tendre

(A) *Similarité génétique des lignées parentales*

(B) *Similarité du profil de recombinaison des RIL issues du croisement entre chaque lignée parentale et la variété Parangon*

Gardiner et al. (2019)

La différenciation du profil de recombinaison entre populations génétiquement éloignées a aussi été mise en évidence chez d'autres espèces de plantes. Chez le riz, (Marand et al. 2019) ont comparé les profils de recombinaisons historiques des deux sous-espèces *Oryza sativa indica* et *O. s. japonica*. Ces deux populations ont un indice de différenciation F_{st} de l'ordre de 0.08 dans les régions génomiques neutres (d'après le Figure 4B de l'article de (Marand et al. 2019)) et seulement 20% de leurs points chauds sont partagés. Pour un niveau de différenciation équivalent chez le blé tendre, les populations WE et EA (F_{st} = 0.08) montrent seulement 11% de points chauds en commun. Les populations les moins différenciées WE et EE (F_{st} = 0.015) ont 20% de points chauds en commun, ce qui suggère que le profil de recombinaison évolue plus vite chez le blé que chez le riz. Chez le cacao, Schwarzkopf et al. (2020) ont estimé le profil de recombinaison historique de

huit populations divergentes (F_{st} 0.16-0.65). Environ 45% des points chauds de recombinaison historiques sont partagés entre populations. Des populations similaires partagent davantage de points chauds que des populations éloignées.

Chez le maïs, Rodgers-Melnick et al. (2015) ont estimé le profil de recombinaison de quatre populations : une population expérimentale (composée des lignées US-NAM et CN-NAM), une population de lignées modernes, une population de lignées plus anciennes (landraces), ainsi qu'une population de téosinte (*Zea mays ssp parviglumis*), l'espèce sauvage à partir de laquelle le maïs a été domestiqué il y a 9 000 ans (Beadle 1939). Le taux de recombinaison historique chez les populations de lignées modernes, de landraces et de téosintes a été mesuré au niveau des points chauds de recombinaison identifiés dans la population expérimentale. Il s'avère que le taux de recombinaison historique sous les points chauds de la population expérimentale augmente de 35% par rapport au reste du génome chez la population moderne, de 28% chez les landraces, mais n'augmente que très peu chez la téosinte (**Figure 31**). Cela suggère une différenciation progressive du profil de recombinaison au cours de la domestication du maïs.

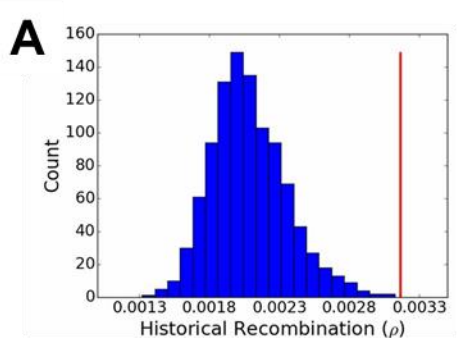


Figure 2. Mean historical recombination rate within improved maize lines over 1,000 permutations of controls (blue histogram) compared with the mean historical recombination rate within hotspots (red line).

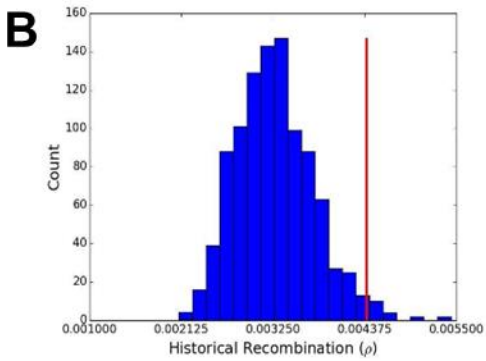


Figure S19. Mean historical recombination rate within maize landraces in hotspots compared to controls.

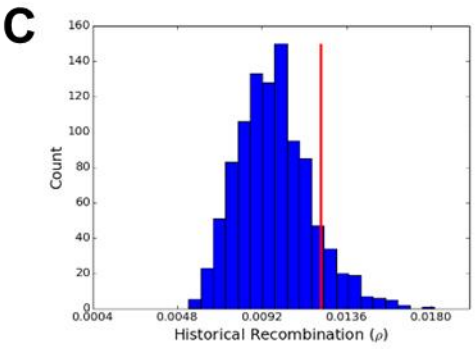


Figure S20. Mean historical recombination rate within teosintes in hotspots compared to controls.

Figure 31 : Similarité du profil de recombinaison chez le maïs

Distributions des taux de recombinaisons historiques (bleu) d'intervalles comprenant des points chauds chez US-NAM et CN-NAM (rouge).

***(A)** Le taux de recombinaison au niveau des points chauds de US et CN-NAM est significativement supérieur de 35% chez les lignées élites.*

***(B)** Supérieur de 28% chez les landraces.*

***(C)** Non significatif mais légèrement supérieur chez la teosinte.*

Rodgers-Melnick et al. (2015)

Chez le coton, Shen et al 2019 ont estimé le profil de recombinaison historique de 4 populations originaires de différentes régions de Chine et de 4 populations spécifiques d'une période de l'agriculture (1920 à 1950, 1950 à 1980, 1980 à 2000 et 2001 à 2010). A l'intérieur de chacun des deux groupes différenciés dans l'espace ou dans le temps, Shen et al. (2019) ont estimé les F_{st} et les corrélations des profils de recombinaison le long du génome pour chaque paire de populations (**Figure 32**). La similarité des profils de recombinaisons diminuait avec le F_{st} pour les groupes différenciés dans le temps (corrélation négative mais non significative) mais pas dans l'espace.

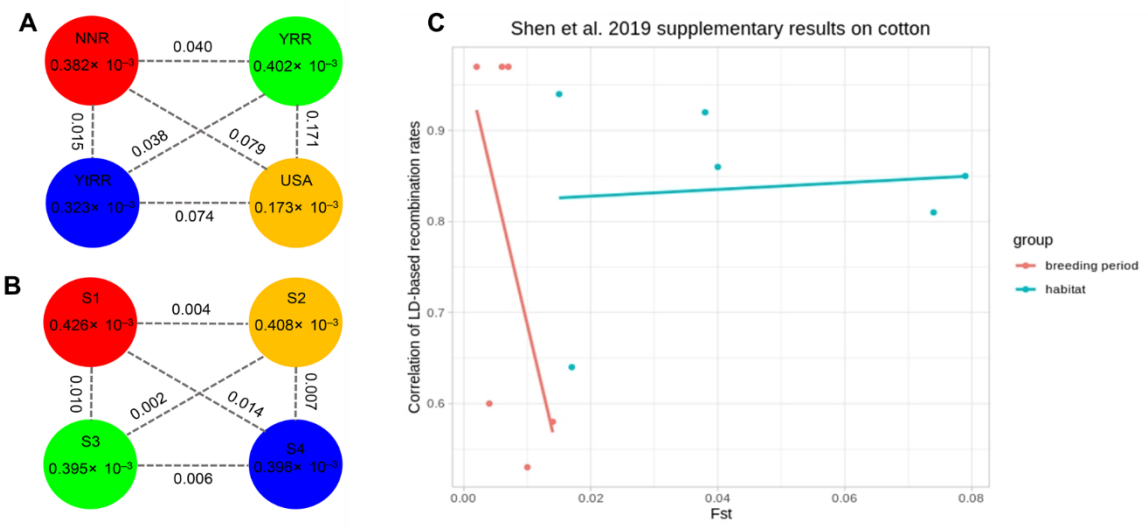


Figure 32 : Similarité du profil de recombinaison chez le coton

Indice F_{st} entre paires de populations de coton et indice de diversité nucléotidique intra-population pour :

***(A)** Les populations sont spécifiques d'un territoire : China northwestern inland region (NIR), China Yellow River region (YRR), China Yangtze River region (YtRR), and United-States territory (USA)*

***(B)** Les populations sont spécifiques d'une époque : des années 1920 à 1950 (S1), 1950 à 1980 (S2), 1980 à 2000 (S3) et de 2001 à 2010 (S4).*

***(C)** Relation entre le F_{st} et la corrélation du profil de recombinaison historique (figure personnelle).*

Shen et al. 2019

Chez le peuplier, les profils de recombinaison historiques ont été comparés entre trois espèces : *Populus. tremula*, *P tremuloides, P trichocarpa* (Wang et al. 2016). *P tremula* et *P. tremuloides* sont les espèces les plus apparentées d'après l'arbre phylogénétique de (Wang et al. 2014) (**Figure 33**). Ces populations montrent une corrélation des taux de recombinaison historiques de 0.51. En comparaison, la corrélation des taux de recombinaison entre *P tremula* et *P. trichocarpa* est de 0.31, et celle entre *P tremula* et *P. tremuloides* est de 0.32.

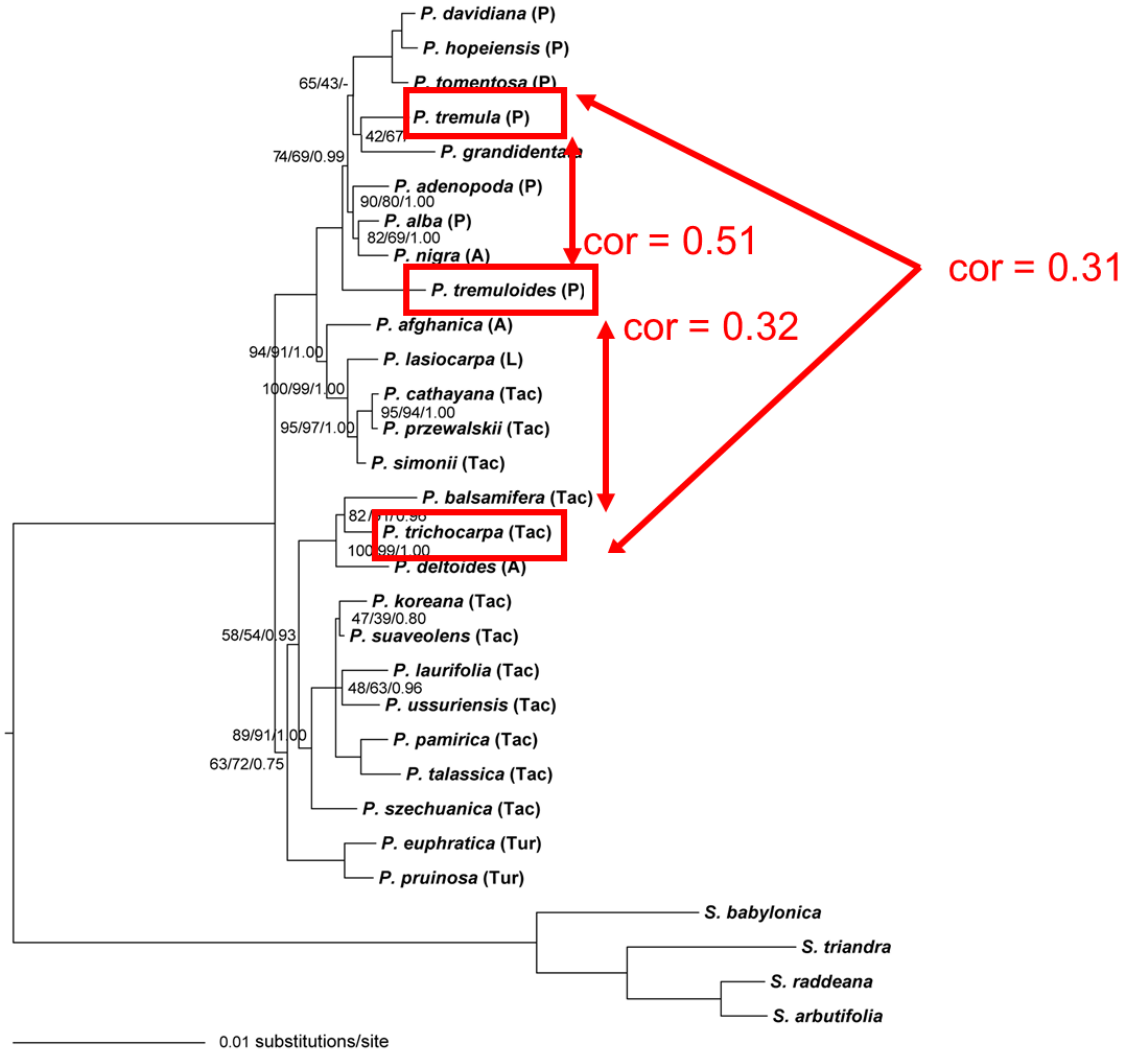


Figure 33 : Similarité du profil de recombinaison chez le peuplier

D'après les résultats de Wang et al. 2014 (arbre phylogénétique) et Wang et al. 2016 (corrélation des profils de recombinaison, en rouge)

En résumé, le profil de recombinaison chez les plantes montre de la variabilité génétique au sein des espèces. Plusieurs études précédemment décrites rapportent que deux populations proches ont des profils de recombinaison plus similaires, ce qui suggère que les facteurs qui déterminent la position des CO chez les plantes sont sujets à évolution. Une analyse fine de l'ensemble des cartes génétiques entre espèces permettrait de mieux caractériser les éléments de séquences qui corrèlent avec la position des points chauds de recombinaison, comme des motifs de séquence ADN ou bien l'enrichissement ou l'appauvrissement en éléments transposables et en marques de méthylation.

IV.2.3 Hypothèse d'un gène majeur dans le déterminisme génétique de la position des crossing-over chez les plantes

La conservation du profil de recombinaison à l'échelle des chromosomes, la concentration des CO en quelques points chauds et enfin la variabilité entre groupes génétiques laisse penser qu'il existe un déterminisme génétique de la recombinaison qu'il serait intéressant de comprendre pour pratiquer des recombinaisons ciblées.

Une hypothèse privilégiée est la variation du profil épigénétique entre groupes génétiques, qui conditionne le statut compacté/décompacté de la chromatine et donc l'accessibilité de l'ADN. Le profil épigénétique peut évoluer par dérive, ou bien il peut être le résultat de pressions de sélection, conduisant à un contrôle de l'expression de gènes. Plusieurs études rapportent une augmentation de la recombinaison à proximité de gènes d'immunité chez différentes espèces (Fulton et al. 2016; Tock et al. 2021; Choi et al. 2016). Chez le blé, Tock et al. (2021) rapportent effectivement que les régions riches en CO sont associées à des signatures de sélection. Les gènes codants pour les protéines d'immunité NLR sont en particulier surreprésentés dans les régions riches en CO. Par ailleurs, les gènes NLR associés à un fort taux de recombinaison présentent aussi une plus grande diversité structurale. Ces résultats sont cohérents avec le modèle de la course aux armements (Red Queen model, ; Howard et Lively 1994; Brockhurst et al. 2014), dans lequel la reproduction sexuée permet aux populations de générer de nouvelles séquences, par recombinaison ou mutation induite par la recombinaison, qui fournissent un avantage évolutif face aux parasites en co-évolution avec eux.

Il n'est pas exclu que la position des points chauds soit déterminée par le profil épigénétique, qui pourrait être lui-même sous contrôle génétique par un gène majeur. Plusieurs études chez les plantes rapportent un enrichissement en marques épigénétiques particulières au niveau des points chauds de recombinaison (enrichissement en H3K4me3 dans les régions riches en CO chez *Arabidopsis*, Choi et al. 2013; Aliyeva-Schnorr et al. 2015; Shilo et al. 2015), enrichissement en H3K27me3 chez le blé, Tock et al. 2021). Chez certains mammifères (humains, chimpanzés, souris, vaches, chevaux) (Ptak et al. 2005; Beeson et al. 2019; Sandor et al. 2012; Brunschwig et

al. 2012), la recombinaison est effectivement contrôlée via le dépôt d'une marque épigénétique par la protéine PRDM9 (Borde et De Massy, 2013). La protéine PRDM9 possède un domaine ayant une activité méthyl-transférase (SET) qui lui permet de tri-méthyler la lysine 4 de l'histone H3 (H3K4me3) à proximité d'un motif cible d'ADN. Cette marque épigénétique attire le complexe protéique SPO11, qui catalyse une cassure double brin, et favorise la formation d'un crossing-over. Les motifs d'ADN ciblés varient selon les allèles de PRDM9 (Baudat et al. 2010). Les points chauds de PRDM9 ont la particularité d'être situés en dehors des régions riches en gènes. Chez les espèces dépourvues de PRDM9 (plantes, oiseaux, levure, chiens et la majorité des espèces en général, (Zhang et Ma 2012), les points chauds de recombinaison sont concentrés en amont ou en aval des séquences codantes (Stapley et al. 2017). PRDM9 semble donc être une innovation évolutive dont le rôle supposé est d'éloigner la recombinaison des gènes, afin de limiter la dégradation des séquences codantes par les conséquences mutagènes de la recombinaison (Brick et al. 2012).

Dans la mesure où les recombinaisons sont concentrées en quelques endroits du génome et qu'elles ont des conséquences mutagènes sur la séquence ADN, l'affinité entre PRDM9 et ses séquences cibles devrait être théoriquement réduite au cours de l'histoire de la population. De cela il résulterait une disparition des points chauds. Cette situation appelée « le paradoxe des points chauds » consiste en une inadéquation entre l'observation des points chauds de recombinaison et la disparition prévisible des séquences cibles. Cependant, PRDM9 montre des traces de sélection diversifiante (Oliver et al. 2009). La diversification des allèles de PRDM9 entraine une diversification des séquences cibles au cours de l'évolution, ce qui résout le paradoxe des points chauds. Cette diversification des allèles de PRDM9 entraine une forte variabilité du profil de recombinaison au sein d'une même espèce ou entre espèces apparentées (Baudat et al. 2010; Hinch et al. 2011; Auton et al. 2012).

En résumé, un gène majeur qui contrôlerait la recombinaison par l'intermédiaire du dépôt de marques épigénétiques nécessite deux éléments : une séquence cible ; un renouvellement des séquences cibles pour résoudre le paradoxe des points chauds.

Les éléments transposables pourraient jouer ce double rôle chez les plantes. Chez le riz, la pomme de terre, le maïs et le blé, plusieurs études rapportent que les points chauds sont enrichis en éléments transposables (Fuentes et al. 2021; Marand et al. 2019; 2017; Schwarzkopf et al. 2020). Les motifs ciblés par PRDM9 sont aussi dérivés d'éléments transposables (Myers et al. 2008). Cette association entre éléments transposables et points chauds chez des espèces très diverses suggère un mécanisme universel dans le contrôle de la recombinaison (Darrier et al. 2017), avec un avantage évolutif à produire des crossing-over en dehors des gènes (promoteurs ou régions terminales des gènes chez les plantes, loin des gènes chez les espèces avec PRDM9). De plus, chez le maïs, les éléments transposables semblent se renouveler assez régulièrement dans les régions où la chromatine est accessible (Zhao et al. 2018), ce qui permettrait de résoudre le

paradoxe des points chauds. Cependant, il est aussi possible que la co-localisation des points chauds de recombinaison et de certaines séquences d'éléments transposables résultent d'une préférence commune pour la chromatine décompactée.

Si l'on souhaite réaliser une étude d'association pour identifier des gènes impliqués dans le contrôle de la recombinaison, le profil de recombinaison est un caractère particulièrement difficile à phénotyper. La génétique d'association consiste à identifier les associations entre une variation phénotypique à une variation génétique. Or le profil de recombinaison est un phénotype qui est difficilement réductible à une valeur numérique. Une solution serait d'utiliser les phénotypes décrits dans Coop et al. (2008). La population d'étude de Coop et al. (2008) est formée d'un ensemble de 1650 individus (humains) apparentés. Les relations d'apparentement sont connues (pedigree sur 13 générations) et 725 individus sont génotypés avec 500k marqueurs. Les auteurs ont pu extraire 50 couples génotypés ayant eu plus de deux descendants génotypés. La comparaison du génome phasé des parents et du génome phasé de leurs descendants a permis d'inférer le profil de recombinaison spécifique de chacun des 100 parents. Ils ont ensuite pu calculer pour chaque parent la proportion de CO qui co-localisent avec des points chauds de recombinaison inférés à partir des patrons de DL dans les données du projet Phase II HapMap.

En utilisant ce phénotype (proportion de CO qui tombe dans des points chauds précédemment estimés dans une autre population), Baudat et al. (2010) ont mis en évidence que les variants de PRDM9 étaient associés avec la variation du profil de recombinaison. La **Figure 34** montre que chez les individus porteurs des génotype AA et AB au locus de *Prdm9*, environ 60% des CO co-localisent avec les points chauds historiques, alors que cette proportion tombe à 20% chez les individus porteurs des allèles AI. Ces chiffres montrent aussi que PRDM9 n'explique pas à lui seul le déterminisme de la recombinaison.

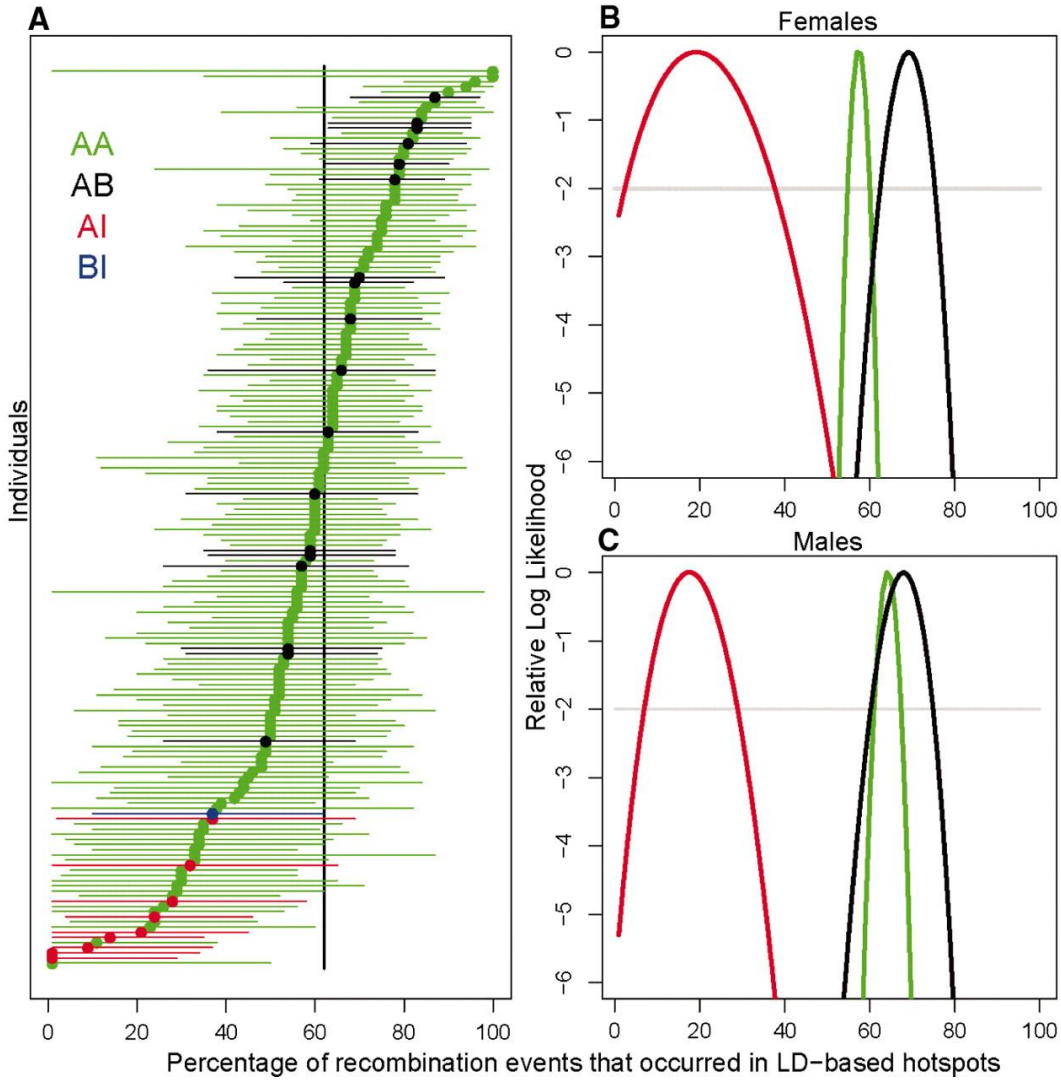


Figure 34 : Association entre les allèles de Prdm9 avec la proportion de crossing-over qui tombent dans les points chauds de recombinaison historiques

Les différentes couleurs indiquent le génotype au locus Prdm9. A : phénotype mesuré pour chaque individu. B et C : variation de la vraisemblance du paramètre (proportion de co-localisation entre points chauds historique et crossing-over) calculée par génotype. La barre horizontale grise donne l'intervalle de confiance à 95%.

Baudat et al. (2010)

La qualité de ce phénotype (proportion de CO qui tombent dans les points chauds d'une autre population) dépend de la résolution sur l'estimation de la localisation des points chauds historiques de recombinaison ou sur la position des CO. Dans les 4 groupes génétiques de blé tendre, nous avons montré que 80% des crossing-over sont concentrés dans 36% de la distance physique totale, au niveau des télomères (régions R1 et R3) où la densité en gènes est très forte. Si la résolution est mauvaise, il y a un risque de surestimer les co-localisations entre points chauds

historiques et CO, et de ne pas avoir assez de puissance pour détecter des différences fines entre populations.

Il existe plusieurs alternatives pour gagner en résolution sur la position des CO. Par exemple, les méthodes de sperm-typing permettent d'amplifier des régions génomiques ciblés au niveau des gamètes (le plus souvent des grains de pollen). Des sondes spécifiques de la région génomique d'intérêt doivent être développées afin de permettre leur liaison à l'ADN à proximité de points chauds préalablement identifiés (Tiemann-Boege et al. 2017; Baudat et de Massy 2007; Cole, Keeney, et Jasin 2010; Berg et al. 2011; Choi et al. 2013; Drouaud et al. 2013). Le dosage relatif entre séquences recombinées et non recombinées par rapport aux parents permet de déduire le taux de recombinaison dans ces régions.

Une méthode alternative est la technique du Chip-Seq. Il s'agit d'immuno-précipitation de la chromatine au niveau des sites de liaison entre des protéines spécifiquement choisies par l'utilisateur et la séquence ADN, puis d'amplifier les séquences ADN associées à ces protéines. Par exemple, Tock et al. (2021) ont utilisé cette technique chez le blé pour identifier les séquences d'ADN ciblées par les protéines DMC1 et ASY1, protéines impliquées dans la formation des CO.

Enfin, une autre alternative pour être exhaustif dans l'analyse du déterminisme génétique serait de séquencer le génome entier des descendants, le jour où les coûts de séquençage deviendront négligeables.

En parallèle, il est aussi nécessaire d'affiner la position des points chauds historiques. Dans notre étude, nous avons détecté environ 8 000 régions hautement recombinantes d'une taille médiane de 20kb. Or les points chauds de recombinaison font environ quelques kb (par exemple 5kb chez le riz (Marand et al. 2019), et plus généralement entre 1 et 10kb chez les plantes (Choi et Henderson 2015). Le séquençage de fragments de grandes tailles dans ces régions hautement recombinantes pourrait permettre de mieux les localiser.

Une difficulté supplémentaire chez le blé tendre est son statut d'allo-hexaploïde et la présence de nombreuse régions paralogues et homologues. Environ 37% des gènes sont présents en trois copies, une pour chacun des trois génomes homéologues (A, B et D, Juery et al. 2021). Ceci rend plus difficile toute approche de séquençage de régions spécifiques, mais aussi le génotypage, les alignements de séquences, le positionnement des marqueurs sur la séquence, le clonage de gènes et la puissance des études d'associations.

Une analyse de génétique d'association sur les RILs Chinese Spring*Renan nécessiterait des manipulations expérimentales coûteuses (sperm-typing ou Chip-Seq). Une alternative moins coûteuse à une analyse de génétique d'association serait de comparer les taux de recombinaisons entre séquences synthéniques des génomes homéologues du blé tendre (A, B et D), voire même avec le génome d'espèces apparentées comme l'orge (génome H), pour estimer l'influence de la séquence ADN ou d'autres éléments (marques de méthylation, transposons, distance au centromère) sur le taux de recombinaison locale.

Conclusion générale

Conclusion générale : Apport de la recombinaison dans l'optimisation des plans de croisements de blé tendre

La recombinaison permet de convertir la diversité génétique en progrès variétal. Grâce aux prédictions génomiques et à l'estimation des taux de recombinaison, il est désormais possible de trier les croisements les plus prometteurs en comparant la distribution prédite de leur descendance.

Si la prédiction génomique est déjà utilisée en routine par le programme de sélection, la prédiction de la descendance permet d'accélérer d'autant plus le progrès variétal à budget équivalent. De plus, choisir les croisements sur la base de la distribution de leur descendance permet aussi de limiter la perte de diversité génétique dans la population de lignées parentales, carburant essentiel pour le progrès variétal à long terme.

Cependant, prédire la distribution de la descendance nécessite des modèles de prédiction génomiques adaptés et une bonne estimation des taux de recombinaison le long du génome. En outre, la variabilité génétique des profils de recombinaison dans différentes populations de blé tendre a été peu explorée jusqu'à présent, ce qui laisse une incertitude sur la portabilité des profils de recombinaison estimés dans des populations distantes de celle du programme de sélection.

Le premier article de cette thèse consiste à estimer et comparer les profils de recombinaison des quatre principaux groupes génétiques du blé tendre. Afin d'estimer les profils de recombinaison, nous avons utilisé une méthode qui exploite la relation entre déséquilibre de liaison et taux de recombinaison. Cette méthode a permis d'obtenir des cartes génétiques sur l'ensemble du génome, les plus fines de la littérature à ce jour. Cependant, les patrons de déséquilibre de liaison peuvent aussi être influencés en théorie par les forces évolutives et les évènements démographiques rencontrés par la population au cours de son histoire évolutive. Nous n'avons pas pu mettre en évidence ces biais en comparant nos cartes basées sur le déséquilibre de liaison avec une carte préalablement obtenue en comptant les crossing-over dans une population biparentale, ces dernières étant moins fines mais dépourvues de ces biais.

Les variations fines des cartes basées sur le déséquilibre de liaison nous ont ensuite permis de faire une comparaison fine des taux de recombinaison entre les quatre groupes génétiques. Cette comparaison a permis de révéler que les facteurs qui déterminent la position des crossing-over chez le blé tendre sont vraisemblablement sujets à évolution. L'évolution du profil de recombinaison semble un processus commun à plusieurs autres espèces de plantes (riz, maïs, coton, cacao, peuplier), et présente des similarités intéressantes avec l'évolution du profil de recombinaison observée chez certaines mammifères (humains, chimpanzés, souris), ce qui suggère un mécanisme universel dans le déterminisme de la position des crossing-over. Une analyse plus

détaillée des cartes génétiques permettrait de mieux modéliser ce mécanisme, par exemple en caractérisant les facteurs génomiques qui corrèlent avec la position des points chauds de recombinaison ou bien en réalisant une analyse de génétique d'association pour identifier un gène majeur responsable de la position des crossing-over.

Le deuxième article de cette thèse utilise les profils de recombinaison précédemment estimés pour prédire la distribution de la descendance d'un grand nombre de croisements candidats à la sélection et optimiser le plan de croisements d'un programme de sélection simulé, en termes de gain et de variance génétique. Nous avons développé un pipeline pour optimiser les plans de croisements qui implémente plusieurs critères de sélection de croisements (CSC) déjà proposés dans la littérature. Ces critères classent différemment les croisements en exploitant différentes propriétés des distributions des descendants (par exemple, espérance des q (%) meilleurs descendants d'un croisement ou meilleur descendant réalisable).

Nous avons aussi proposé un nouveau critère PROBA, qui permet de classer les croisements en fonction de la proportion de leurs descendants dont la valeur génétique serait supérieure à un seuil donné, ce seuil pouvant être par exemple la valeur génétique de la meilleure lignée du programme de sélection. Ce critère PROBA est associé au plus fort progrès variétal lorsque les contributions parentales sont limitées afin de freiner la perte de diversité génétique (par exemple, contrôle du nombre de descendants attribués à chaque parent, parents génétiquement trop proches ne pouvant être accouplés…).

D'une manière générale, les critères de choix de croisements basés sur la distribution de la descendance sont supérieurs aux critères de choix de croisements basés sur la valeur génétique moyenne des parents, critère couramment employé dans les programmes de sélection. Ces critères sont bénéfiques à la fois pour le progrès variétal mais aussi pour la gestion de la diversité génétique.

Nous avons aussi remarqué que la variabilité des profils de recombinaison précédemment estimés n'est vraisemblablement pas suffisante pour influencer le classement des meilleurs croisements. Cependant, les bénéfices de ces critères dépendent de la qualité du modèle de prédiction génomique, ainsi que de la composition de la population parentale, notamment si ces lignées parentales sont issues d'un processus de sélection ou si elles sont structurées. Une prochaine étape consisterait à tester les bénéfices de ces critères dans le cadre de croisements entre lignées parentales élites et ressources génétiques (par exemple, lignées anciennes ou exotiques), qui appartiennent à des groupes génétiques très différenciés. Il serait aussi important de développer des outils pour la sélection des croisements afin d'optimiser le progrès variétal sur plusieurs caractères simultanément. Enfin, les critères majoritairement testés jusqu'à présent optimisent le

choix des croisements indépendamment (ou à la marge) des autres croisements. Or, il faudrait pouvoir prédire la distribution de l'ensemble des descendants issus d'un plan de croisements pour proposer des méthodes d'optimisation plus pertinentes en sélection. Cela encourage à développer de nouveaux critères d'optimisation de plans de croisements.

Ces deux aspects originaux (variation fine du profil de recombinaison et critères de choix de croisements) contribuent à fournir des outils aux sélectionneurs, outils leur permettant de développer de nouvelles variétés plus performantes et répondant bien aux enjeux actuels d'une amélioration de la production dans un contexte d'agriculture durable.

Chapitre V : Annexes

Chapitre V : Annexes

V.1 Supplementary: "Evolution of recombination landscapes in diverging populations of bread wheat"

Alice Danguy des Déserts [1], Sophie Bouchet [1], Pierre Sourdille [1*], Bertrand Servin [2*]

[1] INRAE-Université Clermont-Auvergne, UMR1095, Génétique Diversité Ecophysiologie des Céréales, 5 chemin de Beaulieu, 63000, Clermont-Ferrand, France

[2] INRAE, UMR 1388, Génétique, Physiologie et Systèmes d'Elevages, F-31326, Castanet-Tolosan, France

* Corresponding authors

- Bertrand Servin, PhD, email: bertrand.servin@inrae.fr
- Pierre Sourdille, PhD, email: pierre.sourdille@inrae.fr

V.1.1 Supplementary protocols

Protocol S1: Population-specific meiotic recombination profiles from LD-based recombination profiles

Assuming the LD-based recombination rate ρ is proportional to the meiotic recombination rate c (*e.g.* in a Wright Fisher model, $\rho = 4N_e c$, where N_e is the effective diploid size of the population), LD-based recombination profiles can be scaled based on the CsRe Bayesian meiotic recombination map: the ratio between the average CsRe Bayesian meiotic recombination rate and average LD-based recombination rate of a genomic region yields an estimate of the coefficient of proportionality. This scaling implies two hypotheses: 1) the average recombination rate in a genomic region is the same across the five populations CsRe, WE, EE, WA and EA. For example, CsRe has an average Bayesian meiotic recombination rate of 0.8 cM/Mb in 3BR1, so this would be the average recombination rate of this genomic region in the four populations of landraces. 2) There is no variation in the proportionality coefficient within a genomic region. Landraces and CsRe meiotic genetic maps can be found in supplementary file S4.

Protocol S2: Sensibility of LD-based recombination rate intensity profile (λ) to population sample size and prior parameter distribution

- ***S2A. Sensibility to population sample size***

To downsample WE population (127 landraces) and reach EE sample size (70 landraces), we build a hierarchical clustering of the 127 WE landraces using a simple-matching distance matrix (Balfourier & al. 2019) and Ward distance. We cut the hierarchical clustering into 70 groups and randomly sampled one landrace per group (supplementary Figure S13). This yielded a new population, named WEds, made of 70 landraces. This procedure allows WEds population to mimic at best, WE genetic composition. We then estimated WEds LD-based recombination profile with PHASE software using the WE SNPs dataset and the same PHASE settings as described in the article. Thus, for each interval of two successive SNPs, we had one posterior distribution of λ for WE and one posterior distribution of λ for WEds. For each interval and each population, we took the median of these distributions as estimates of local recombination intensity (supplementary Figure S14, left). To compare such correlation with the variability of PHASE inferences due to random start of the Markov Chain Monte Carlo (MCMC, "run effect"), we re-estimated WE LD-based recombination profile (same landraces set, same SNPs set) (supplementary Figure S14, right).

- ***S2B. Sensibility of LD-based recombination rate intensity (λ) to PHASE software prior distribution parameters***

According documentation for PHASE v2.1 (Stephens et al. 2004), several prior distribution parameters can be modified by user: "*The form for the prior on the background recombination parameter, $\underline{\rho}$, is that log(ρ) is normal with mean log(μ) and standard deviation σ (truncated so that $\underline{\rho}$ is forced to lie between 10^{-8} and 10^{3}). The value of σ is 0.5∗log(f), where f is chosen so that you would typically (95% of the time) expect $\underline{\rho}$ to be within a factor f of μ.*"

We estimated LD-based recombination rates of WE population using either the default prior for the background recombination rate ρ_w (μ = 0.0004 and f = 1e6), a prior centered on high values for ρ_w (μ = 1 and f = 1e6) and a prior centered on low values for ρ_w (μ = 1e-7 and f = 1e6). For each prior distribution and each interval of two successive SNPs, we extracted the median of the posterior distribution of $\boldsymbol{\lambda_i}$and of $\rho_i = \boldsymbol{\lambda_i} * \boldsymbol{\rho_{w(i)}}$.

Protocol S3: Differentiation at meiosis genes

From a list of 296 genes known to be involved in the meiosis process in bread wheat (Pierre Sourdille, personal communication), we kept 54 genes overlapped by at least 5 SNPs of the common SNPs dataset (including a 10 kilobases extra margin on each side of the gene). We also extracted a subset of 9,826 genes from RefSeq V1.0 annotation belonging to the same genomic region than meiosis genes and respecting similar SNPs coverage rules than meiosis genes. These genes are supposed to be neutral toward the meiosis process and will be considered as control genes to measure the background differentiation level across populations of landraces. F_{ST} values for each gene were computed with HAPFLK software using their overlapping SNPs. The differentiation level of each gene is given by the linear relationship (slope) between the F_{ST} and correlation of genome-wide recombination profile.

In a first place, to test whether meiosis genes were in average more or less differentiated than control genes of their genomic region, we used the linear model:

$$Y_{pg(rc)} = \mu + \alpha_r + \beta_c + (\omega + \delta_r + \gamma_c) * X_p + E_{pg(rc)} \text{ with } E_{pg(rc)} \sim N(0, \sigma^2) \qquad (1)$$

where p is the pair of populations (6 levels, WE-EE; WE-WA; WE-EA; EE-WA; EE-EA; WA-EA), g is the gene (9,880 levels), r is the genomic region (32 levels) and c is the gene category (2 levels, control or meiosis). $Y_{pg(rc)}$ is the pairwise F_{ST} value per pair of population p, for gene g belonging to genomic region r and gene category c. The term $(\mu + \alpha_r + \beta_c)$ gives the average pairwise F_{ST} per genomic region and gene category. The co-variables X_p are the median of correlation of log10(λ) across all genomic regions except centromeres (one measure per pair of population p, identical for every gene and every genomic region). The term ω is the linear relationship between F_{ST} and correlation, *i.e.* what we call the differentiation level. The term δ_r indicates that differentiation levels might vary across genomic regions. The term γ_c and its significance level (evaluated by t-test) indicates whether meiosis genes were in average more or less differentiated than control genes of their genomic region.

For each gene, we also estimated the deviation from the background differentiation level by adjusting a linear model for each genomic region independently:

$$Y_{pg} = \mu + \alpha_g + (\omega + \gamma_g) * X_{pg} + E_{pg} \text{ with } E_{pg)} \sim N(0, \sigma^2) \qquad (2)$$

The indices p gives the pair of populations (6 levels), the term g gives the gene name (from 9 levels in 7AC to 740 levels in 5AR3). The variables Y_{pg} and X_{pg} still represent pairwise F_{ST} and correlations respectively. The term $\mu + \alpha_g + \beta_c$ gives the average pairwise F_{ST} in the genomic region for the gene g; the term ω is the average relationship between F_{ST} and correlation of recombination profile and the term γ_g is the gene-specific deviation to average relationship between F_{ST} and similarity of recombination profile. Note that we set sum-to-0 constraints when estimating the γ_g terms.

The estimates of deviations and their standard errors were used to compute False Discoveries Rates using the ashr R package (Stephens et al. 2020). Genes showing a FDR lower than 0.01% were considered as significantly more differentiated than the genomic background.

Protocol S4: Identification of four diverging populations of landraces

Balfourier et al. (2019) analysed the genetic structure of the 632 landraces dataset and could pinpoint four main groups corresponding to the geographic origins of lines. Despite this structuration, the general pattern of differentiation in these data is somewhat continuous, a lot of individuals exhibiting admixed origins. Here, we subsampled the dataset in order to constitute populations of individuals that were both homogeneous within groups and clearly differentiated between groups. This was achieved in three steps. i) From the Balfourier et al. (2019) admixture analysis with K=4 groups, landraces exhibiting an admixture coefficient smaller than 50% of their dominant group were removed, yielding 534 low admixed landraces (supplementary Figure S1). ii) These 534 landraces were grouped into four populations by hierarchical clustering on the pairwise distance matrix estimated in Balfourier & al (2019) and using the Ward's grouping criterion. The four populations were named as West Europe (WE), East Europe (EE), West Asia (WA) and East Asia (EA) from the geographical origin of their members. The genetic difference (distance) between two landraces was the proportion of mismatched haplotypic alleles along the genome, computed using 8,741 haplotypic blocks containing up to 20 alleles per block (Figure 1 of Balfourier et al. 2019). iii) The last step aimed at discarding closely related individuals within each population to avoid over representing family specific recombination events. To spot closely related landraces, outliers were called from the distribution of genetic distance as follows. Within each population, we fitted a Normal distribution to the observed distribution of distances using robust estimators for the mean and variance (R MASS package, function rlm, (Venables and Ripley 2002). Based on this distribution we tested whether the distance of a particular pair of individuals was consistent with this normal distribution (note that the test is one sided as we only tested for outlying low values). We corrected for multiple testing by applying a False Discovery Rate correction based on the normal p-values (R qvalue package, function qvalue, Storey et al. 2015). Pairs of landraces showing a q-value lower than 0.001 were considered as related. An iterative algorithm was designed to suppress related landraces: at each step, the algorithm first computes the number of relatives per landrace, and then removes the landrace exhibiting the highest number of relationships. The algorithm stops when no related pair remains in the set of individuals. Relationships of individuals within populations are represented in supplementary Figure S19.

Protocol S5: Definition of PHASE windows

PHASE windows were defined of successive SNPs spanning 2 cM (centre 1 cM and borders 0.5 cM) according to the CsRe Bayesian genetic map. The genetic positions of SNPs that were not

polymorphic in CsRe were estimated based on their physical position and the physical and genetic positions of mapped flanking markers. A linear interpolation was fitted on the physical and the genetic positions of the 79,564 SNPs of the CsRe Bayesian map and used to predict genetic position of the other SNPs (R base package, function approxfun). Estimates of first SNPs in R1 and last SNPs in R3 in many chromosomes were often not possible, because no CsRe marker was mapped that far in chromosome extremities. The genetic positions of those extreme SNPs were estimated using the average recombination rate of CsRe intervals situated in R1 region (respectively situated in the R3 region) and their physical distance from the closest mapped marker. The final interpolated genetic map was set to start at 0 cM for each chromosome (example in supplementary Figure S23).

The total number of SNPs within PHASE windows had to be controlled: a minimum of 50 SNPs to ensure reliability of inferences, no more than 160 SNPs to reduce computational time. This required adjusting the number of SNPs in central and flanking parts for some windows (in supplementary Figure S24).

Protocol S6: Remove of 20% HRIs based on their physical length

We discarded the 10% smallest and 10% largest HRIs, assuming that very small intervals exhibiting high recombination rates likely reflect problems in genome assembly or that very wide HRIs do not allow to properly study colocalization of HRIs. After filtering, the size of the widest HRIs was around 100-times higher than the size of the smallest one (around 100k-times higher if filtering on HRIs size is not performed, supplementary Figures S25, S26).

V.1.2 Supplementary Figures

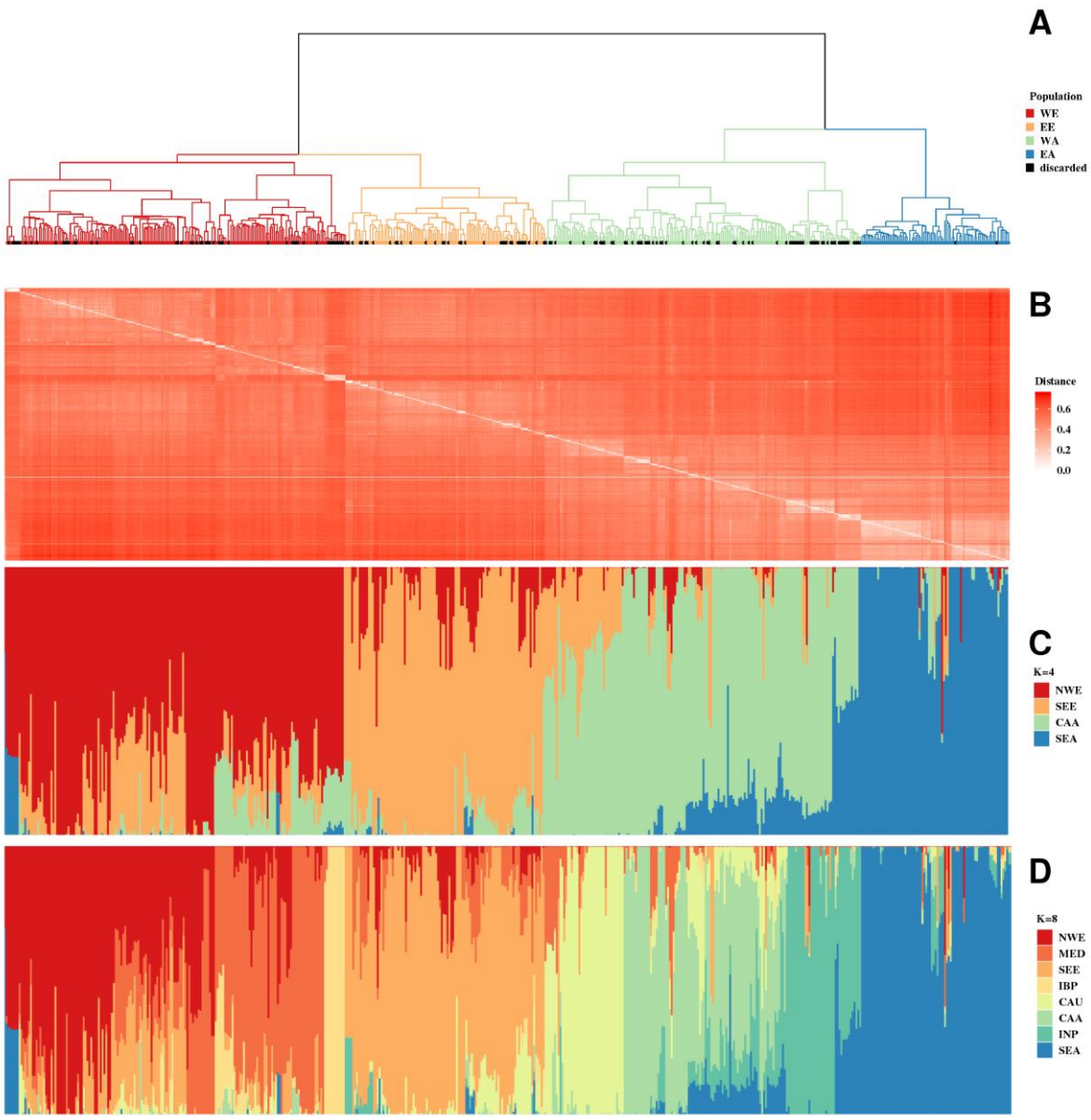
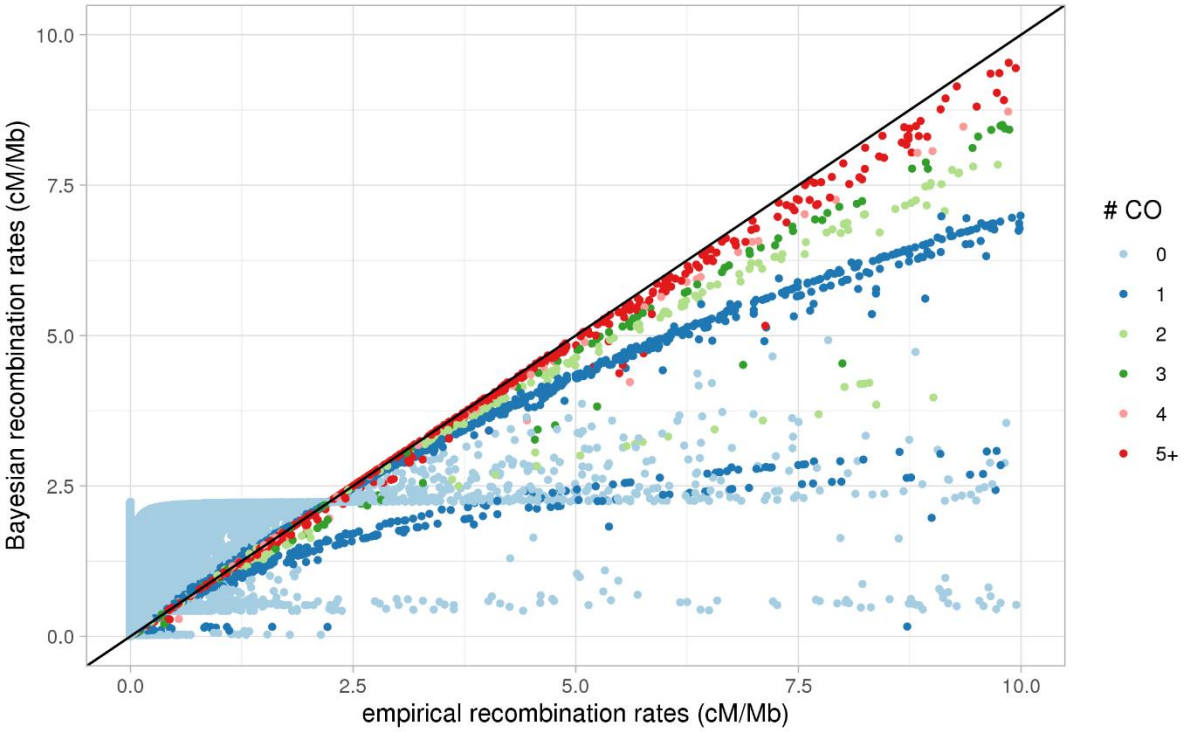


Figure S1: Bread wheat landraces genetic diversity and structuration at K=4 and K=8

A Hierarchical clustering to identify four bread wheat landrace populations. WA = West Asian population; EA = East Asian; WE = West European; EE = East European; discarded = too closely related landraces removed from the analysis. **B** Pairwise simple matching distance matrix from Balfourier et al. 2019 **C and D** STRUCTURE results for K=4 and K=8 groups, from Balfourier et al. (2019). NWE = North West European, MED = Mediterranean, SEE= South East European, IBP =



Iberian Peninsula, CAU = Caucasian, CAA = Central Asian and African group; INP = Indian Peninsula, SEA = South Est Asian.

Figure S2: Comparison between Bayesian and empirical (frequentist) estimates of CsRe meiotic recombination rates

Estimates of per-meiosis recombination rate are function of RILs recombination rate estimates (resulting from more than 1 meiosis). In the Bayesian model, RILs recombination rates is $C_{i(r)}^{bay} = \frac{y_i + \alpha_r}{M\ L_i + \beta_r}$ while empirical estimates is function of $C_i^{emp} = \frac{y_i}{M\ L_i}$, where i is the interval, y_i is the number of recombination events in the interval, M is the number of RILs, L_i is the physical size of the interval and α_r and β_r the parameters of the prior Gamma (supplementary Figure S18). Both estimates are also function of uncertainty in crossover locations. The number of crossovers per interval (#CO) is the average number of crossovers assigned to each interval over 1,000 iterations. Note that Bayesian model attributes similar estimates to intervals of similar length receiving the same number of crossovers in a region. This results in vertical lines in Figure 3 of the article.

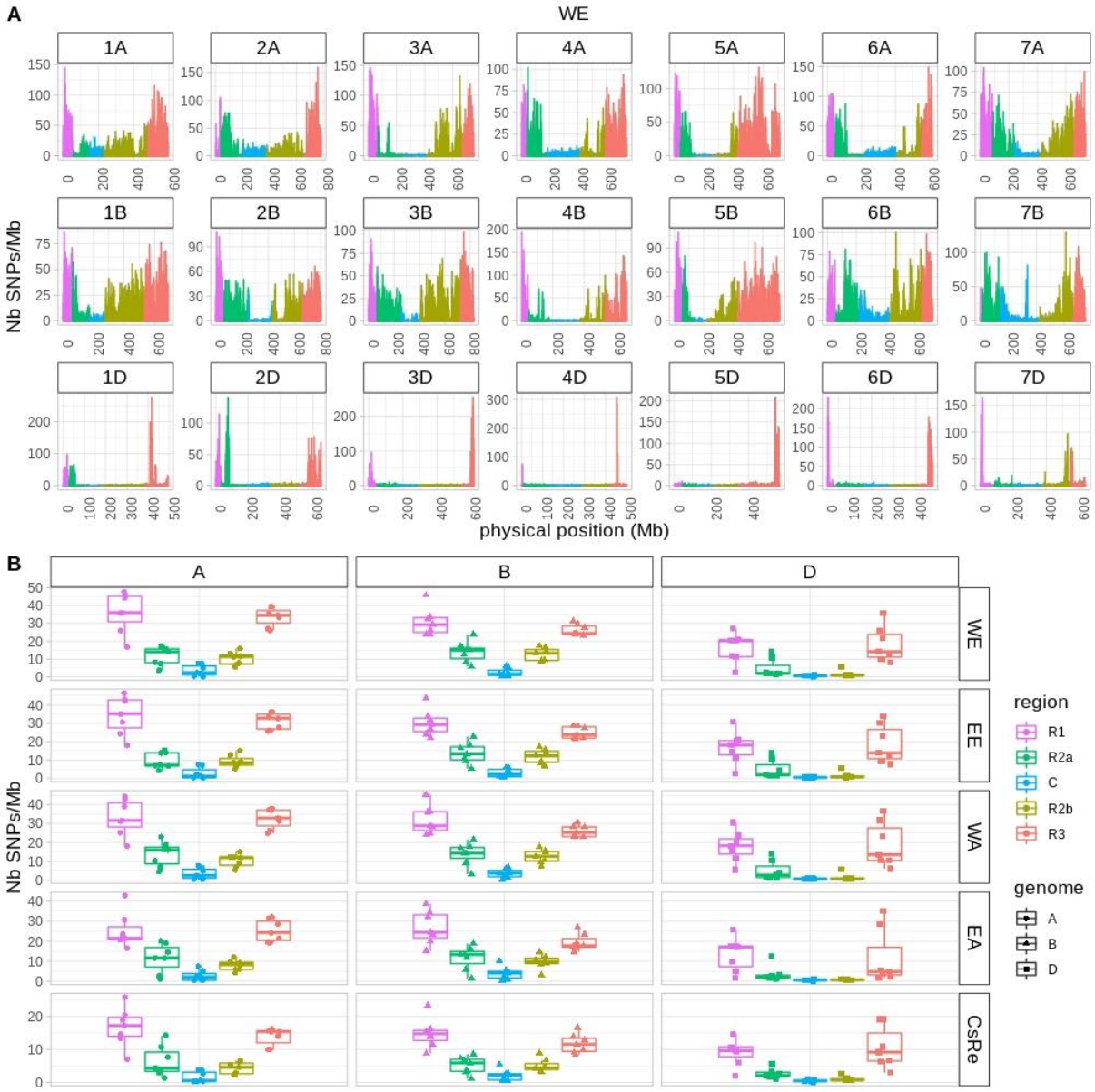


Figure S3: A. SNPs density along each chromosome for a Western European (WE) population of landraces. B. SNPs density per population, genomic region and biological origin of chromosomes

Regions R1 and R3 are telomeric regions, R2a and R2b pericentromeric regions and C are centromeric regions. Chromosomes of genome A (1A to 7A) are derived from *T. monococcum ssp. urartu* genome; chromosomes of genome B (1B to 7B) are derived from the genome of a yet-unknown species related to the *Sitopsis* section ; chromosomes of genome D (1D to 7D) are derived from *Aegilops tauschii* genome (D genome). Populations WE, EE, WA and EA are populations of unrelated landraces. The CsRe population is a biparental population of RILs.

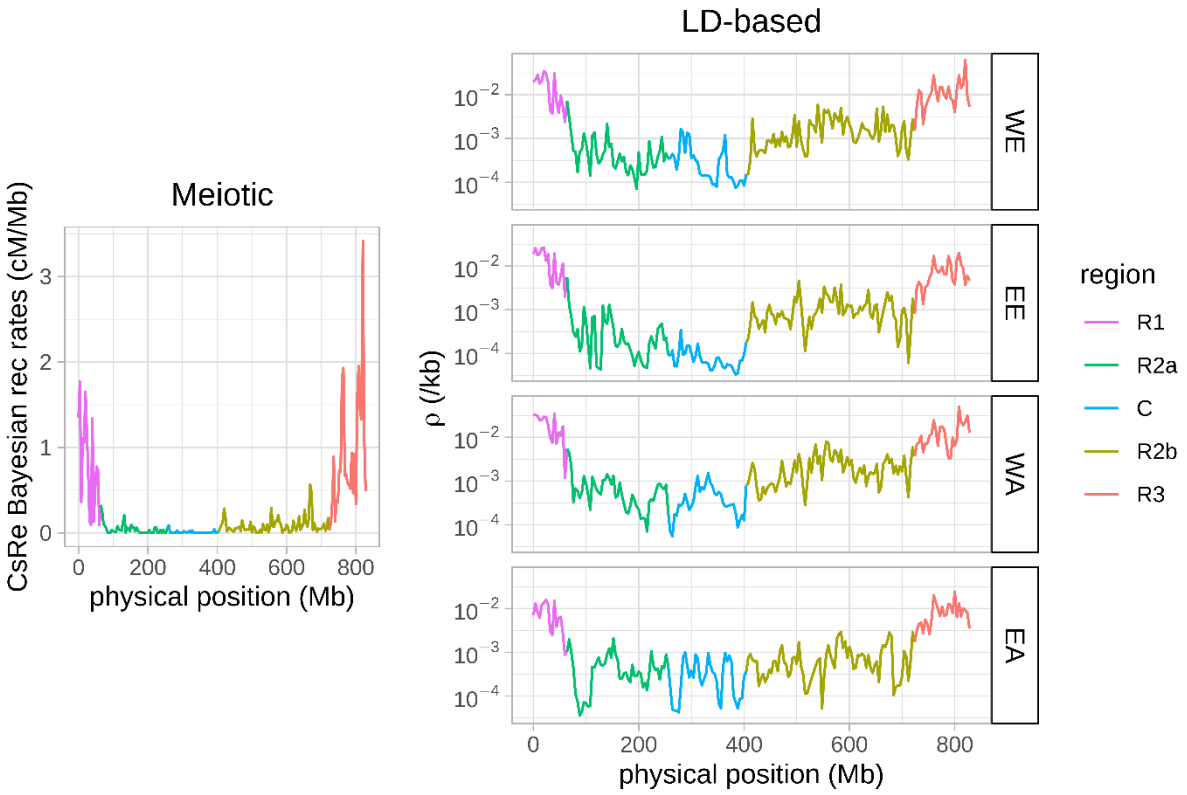


Figure S4: Meiotic and LD-based recombination profiles (log_{10} scale) in 4 Mb windows along chromosome 3B in the CsRe segregating population (left) and in the four West European (WE), East European (EE), West Asia (WA) and East Asia (EA) collections (right)

Each colour corresponds to genomic regions defined by Choulet et al. (2014): highly recombining telomeres R1 (magenta) & R3 (red); low recombining pericentromeres R2a (dark green) & R2b (light green); and centromere C (blue) where recombination rates are close to 0.

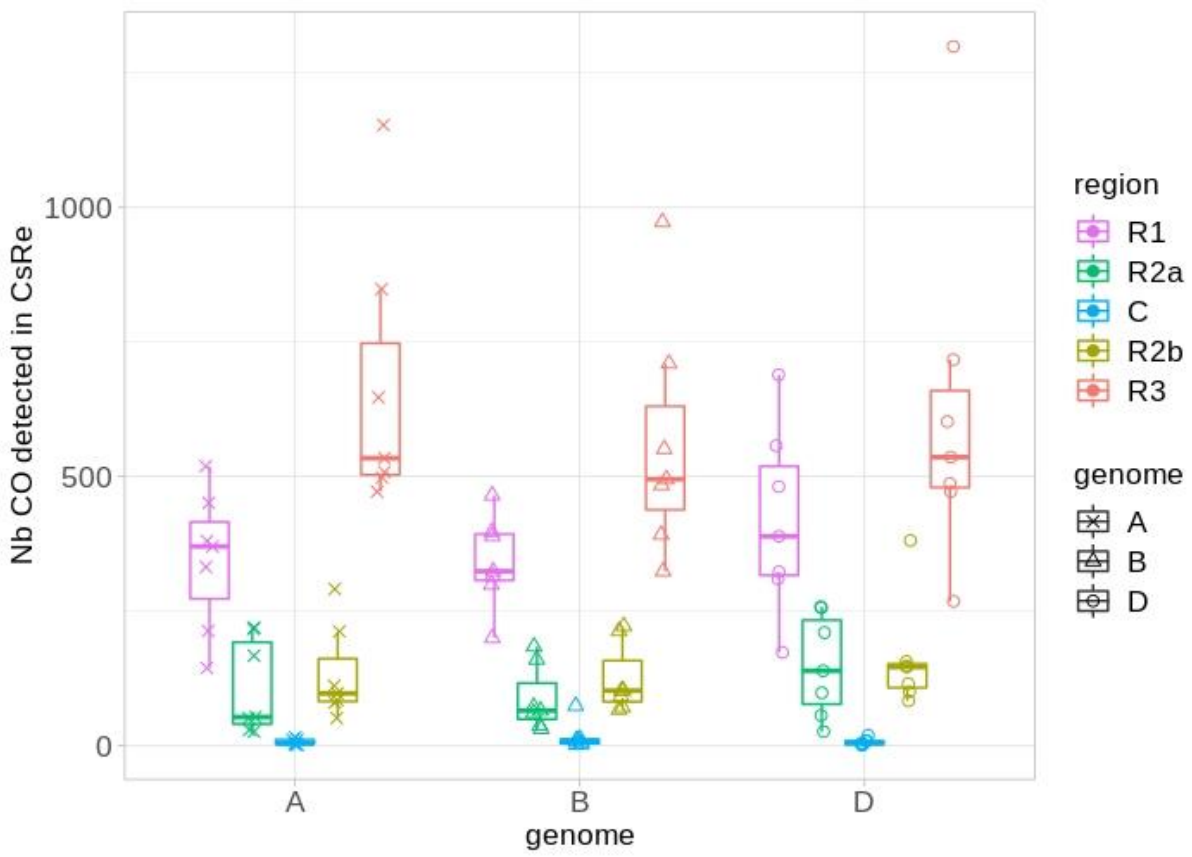


Figure S5: Number of crossovers (CO) detected per genomic region and biological origin of chromosomes in the CsRe population

Regions R1 and R3 are telomeric regions, R2a and R2b pericentromeric regions and C are centromeric regions. Chromosomes of genome A (1A to 7A) are derived from *T. monococcum ssp. urartu* genome; chromosomes of genome B (1B to 7B) are derived from the genome of a yet-unknown species related to the *Sitopsis* section ; chromosomes of genome D (1D to 7D) are derived from *Aegilops tauschii* genome (D genome). The CsRe population is a biparental population of RILs, while populations WE, EE, WA and EA are populations of unrelated landraces. One CO was counted at each parental allele switch along chromosomes in CsRe progeny.

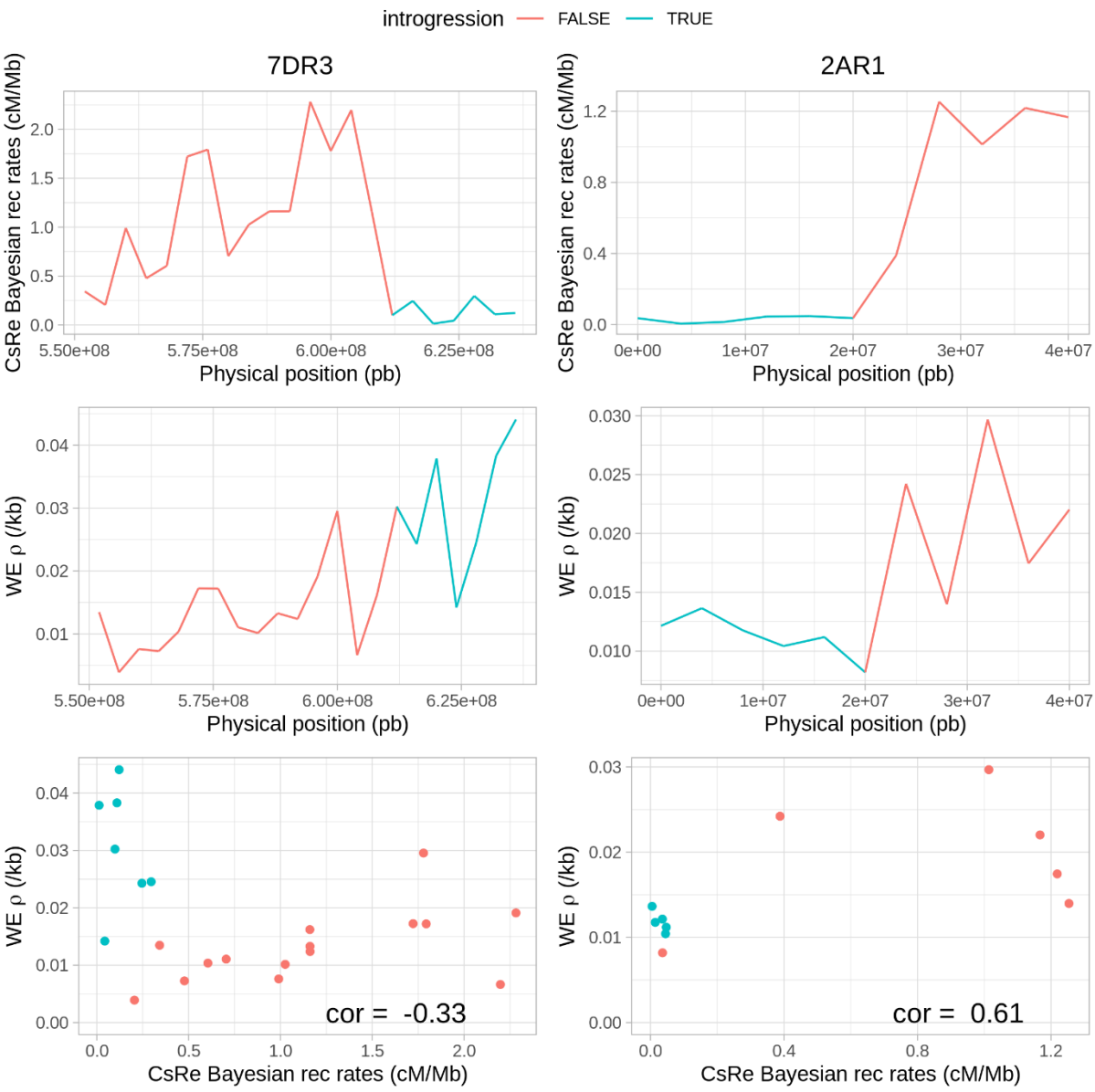


Figure S6: Recombination rates in the region R1 of chromosome 2A (left) and the region R3 of the chromosome 7D (right) where two genomic introgressions segregates in the CsRe biparental population (top panels) but not in WE landraces population (middle panels)

In the 2AR1 genomic region, Renan line (one parent of the CsRe biparental population) carries an 20 Mb introgression from *Aegilops ventricosa* (*Lr37/Sr38/Yr17* resistance gene cluster). In the 7DR3 genomic region, the Renan line carries a 28 Mb introgression from *Aegilops ventricosa* (eyespot resistance gene *Pch1*). Recombination profile overlapping locations of introgressed segments in the Renan line is colored in blue while the rest of the recombination profile is colored

in red. Recombination profiles are averaged within 4 Mb windows. The bottom panels give the relationship between CsRe and WE recombination profiles within each 4 Mb interval.

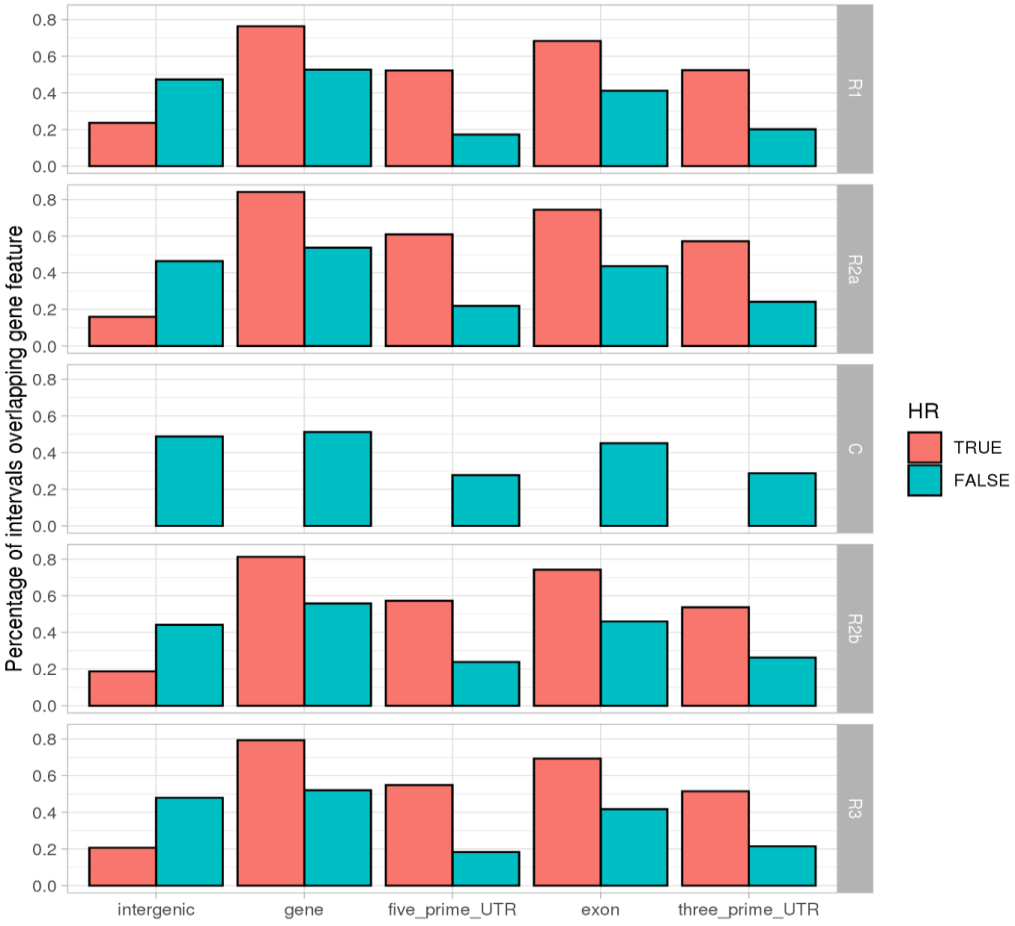


Figure S7: Proportion of Highly Recombining Intervals (HRI, red) overlapping genes features or intergenic segments compared to the proportion of non HRIs intervals (blue) overlapping genes features or intergenic segments

Position of genes features were extracted from the annotation of the RefSeq V1.0 assembly genome (IWGSC, 2018). The comparison has to be made independently within each of the 5 chromosomic regions because of decreasing density of genes and HRIs from telomeres (R1, R3) to pericentromeres (R2a, R2b) and centromeres (C) (Choulet et al. 2014). Note that the overlap was computed within each genomic region (1AR1...7DR3) and then averaged per chromosomic region (R1, R2a, c, R2b, R3). intergenic = genomic segments where no gene were annotated ; gene = genomic segment where genes were annotated, including sometimes 5'UTR, exon and 3'UTR segments within the gene segment. The proportion of HRIs co-localizing with genes and intergenic features is put in regard with the proportion of non HRIs overlapping such features, this second proportion representing here the expected overlap in a model where HRI's were randomly placed along the genome. On average, in all genomic regions, the 8,713 HRIs tend to highly co-localize with genes features compared to non HRIs intervals. For example, the proportion of HRIs

overlapping genes was 80%, but dramatically decreased to 53% when considering non-HRIs intervals.

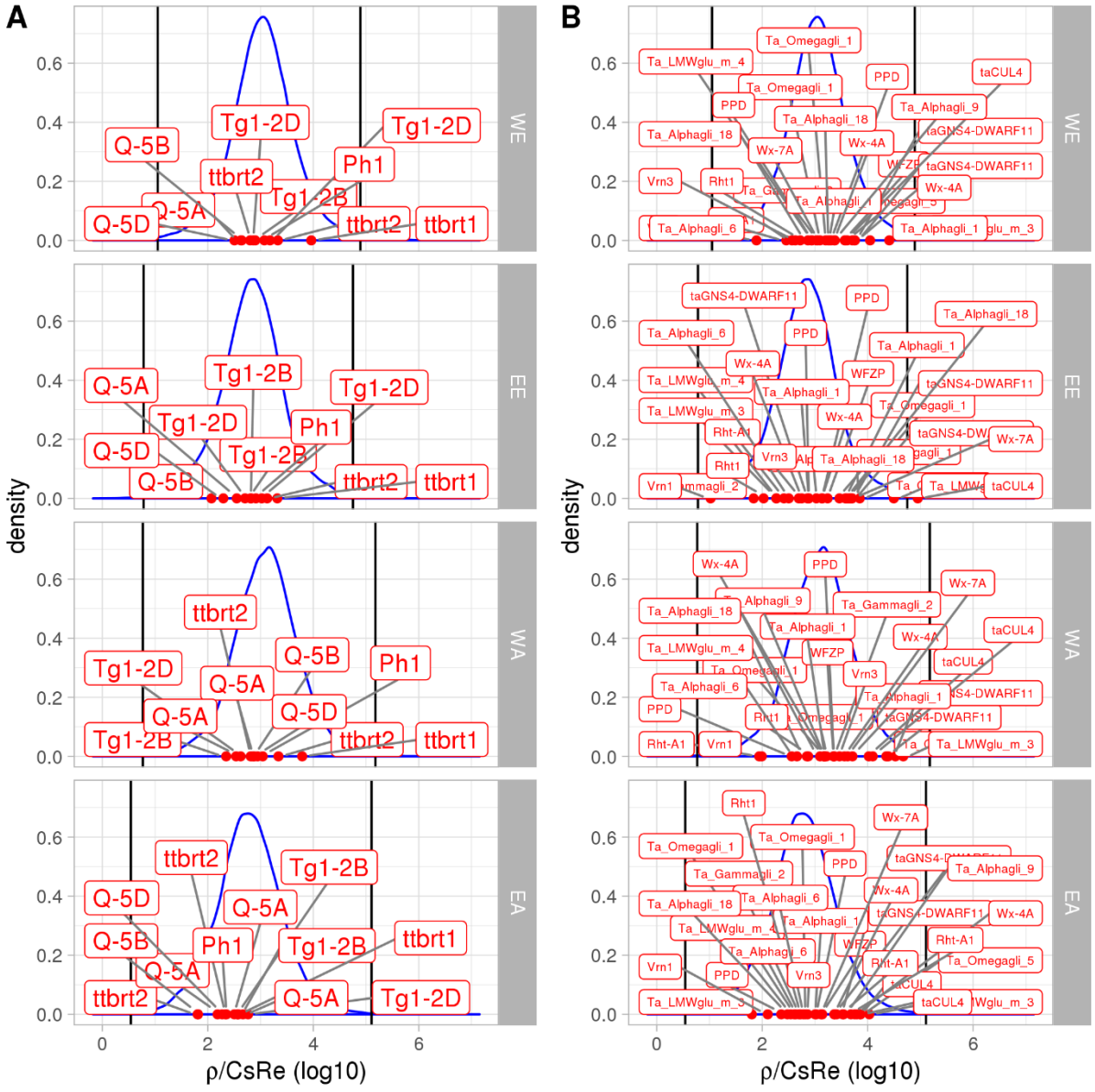


Figure S8: Comparison of LD-based and CsRe meiotic recombination rates nearby genes under selection

Blue curves represent the distribution of the ρ/CsRe ratio of each 4 Mb interval. **A** domestication genes. ttbrt = brittle rachis, Tg = tenacious glume, Ph = homoeologous pairing, Q = non-free-threshing character. **B** Improvement genes. Ppd = photoperiod sensitivity, VRN = vernalization, Rht = reduced height, Glu = Glutenin and Gli = Gliadin genes involved in seed storage protein accumulation, FZP = Frizzy panicle, GNS = grain number, Wx = waxy, as well as uniCULm = the

CUL gene driving plant architecture (names and position according Pont et al. (2019)). Vertical black lines indicate threshold of FDR ≤ 5% or FDR ≥ 5%.

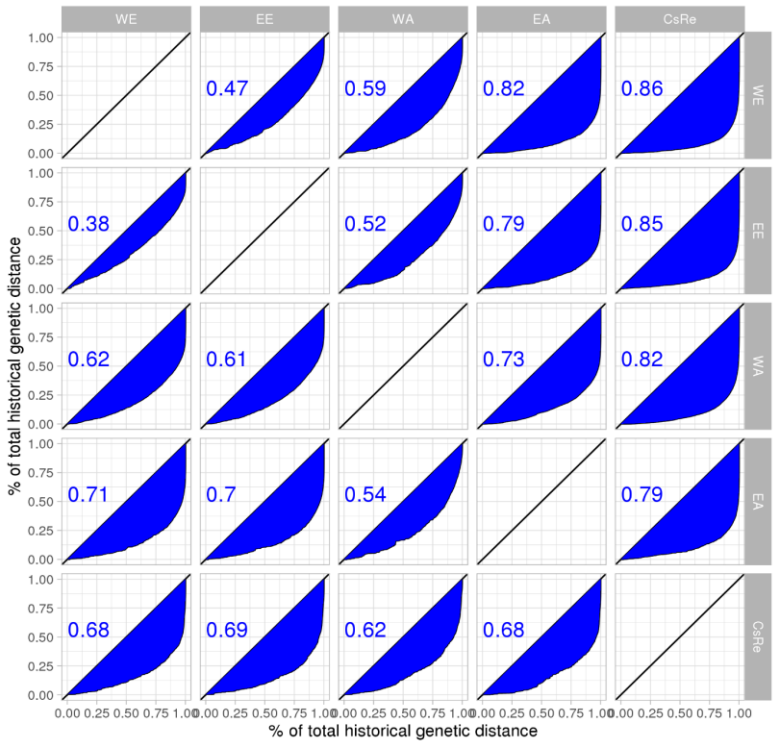


Figure S9: Comparison of distribution of genetic distance along genome in different populations

We computed the genetic distance of each interval i (defined at supplementary Figure S20) of two successive SNPs of each population this way : for WE, EE, WA and EA populations: genetic distance = LD-based recombination rates in interval i (λ_i^ *ρ_(w(i))^) * size of the interval; for CsRe population: genetic distance = c_(i(r))^bay* size of the interval). We also computed the total genetic distance of each population as the sum of genetic distances over all intervals. In each interval and each population, we computed the proportion of genetic distance by the interval as the genetic distance of the interval/total genetic distance. This allows us to compare distribution of genetic distances in intervals across populations. One possible representation of the unveness of the distribution of genetic distances across populations is given by the black curve on the graph. It gives the relationship between the proportion of genetic distance in one reference population (x-axis) and the corresponding proportion of genetic distance in another population (y-axis) considering the same intervals. On the x-axis, intervals are sorted from the interval catching the most genetic distance to the interval catching the lowest genetic distance in the reference population. In consequence, one pair of populations (example: WE-EE) yields different curves depending on which population is considered as the reference one. Another measure of

unevenness of a distribution is the Gini coefficient I. Note that Gini coefficients are technically computed as the area of the blue shape, i.e. area between y = x straight line and black curve.

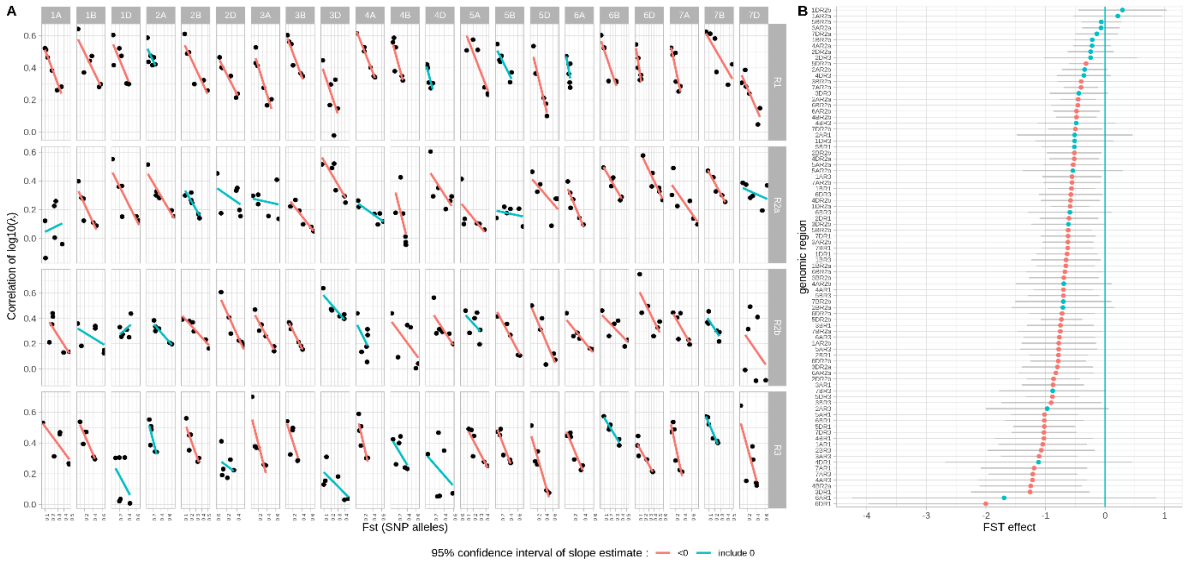


Figure S10: Relationships between correlation of local recombination intensity and F_{ST} per genomic region computed on SNPs alleles. Recombination rates are estimated from a population specific set of SNPs

A Relationship per genomic region. **B** Ranked slope estimates (coloured points) and their 95% confidence interval (grey bar). Blue colour represents slopes with a confidence interval overlapping 0 and red colour confidence interval not overlapping 0.

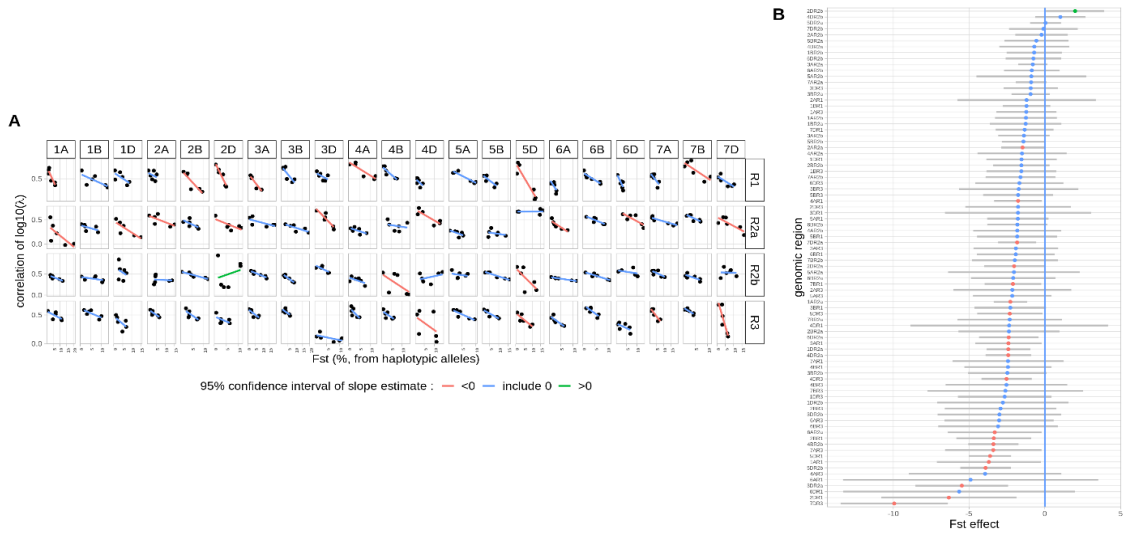


Figure S11: Relationships between correlation of local recombination intensity and F_{ST} per genomic region computed on SNPs alleles. Recombination rates are estimated using a set of SNPs which are polymorphic in all landraces populations (common SNPs dataset)

A Relationship per genomic region. **B** Ranked slope estimates (coloured points) and their 95% confidence interval (grey bar). Blue colour represents slopes with a confidence interval overlapping 0 and red colour confidence interval not overlapping 0.

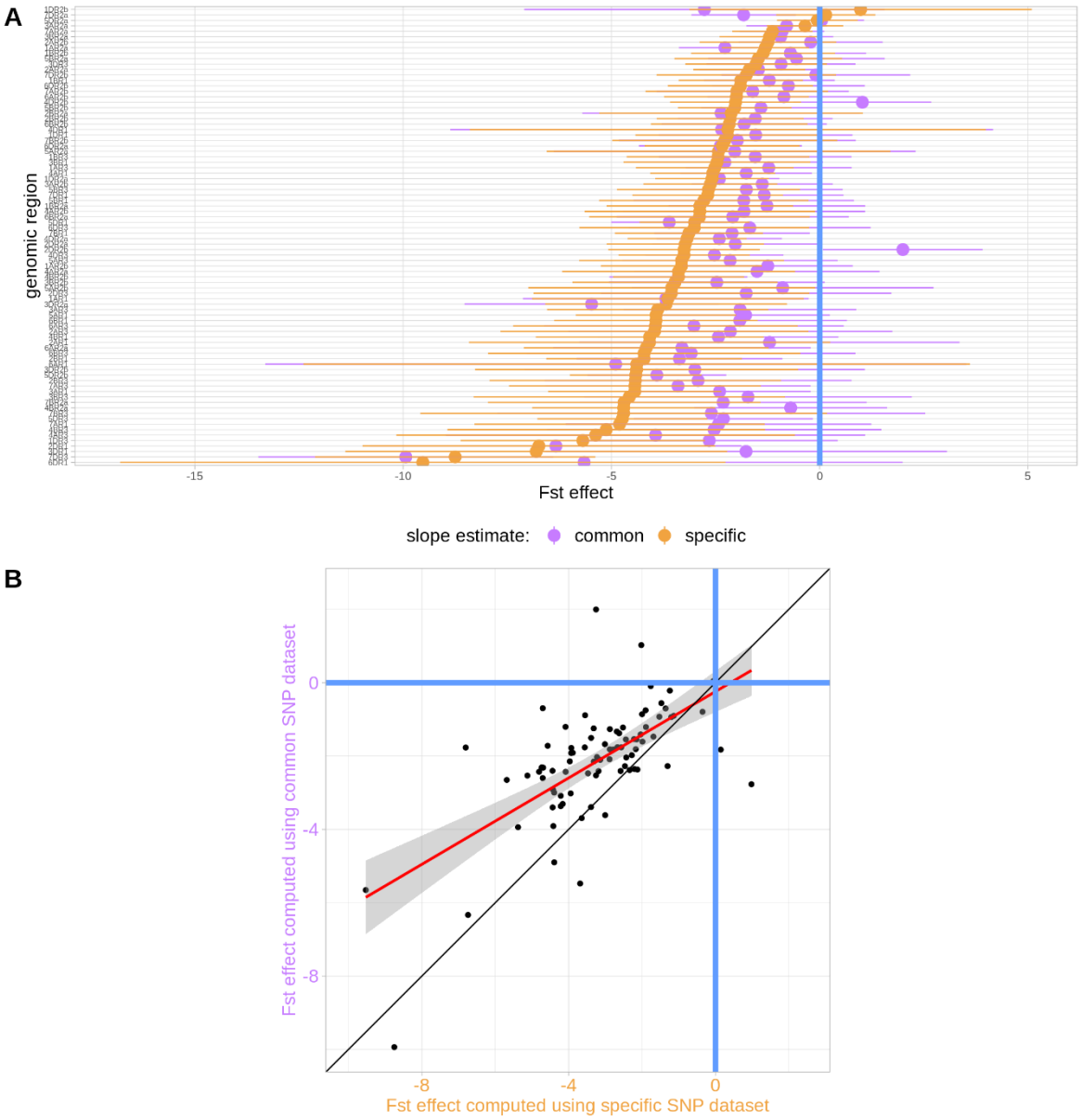


Figure S12: Comparison of slope estimates using specific (gold, Figure S10) and common (purple, Figure S11) SNPs datasets to estimate LD-based recombination rates. A. Ranking and confidence intervals of slopes estimates for each genomic region derived from either specific or common SNPs dataset. B Relationship between slopes estimates using either SNPs dataset

Most slopes are negative in both dataset and there is a positive significant relationship between estimates. Slopes from the common SNPs dataset tends to be higher (*i.e.* less negativ) than slopes from specific SNPs dataset.

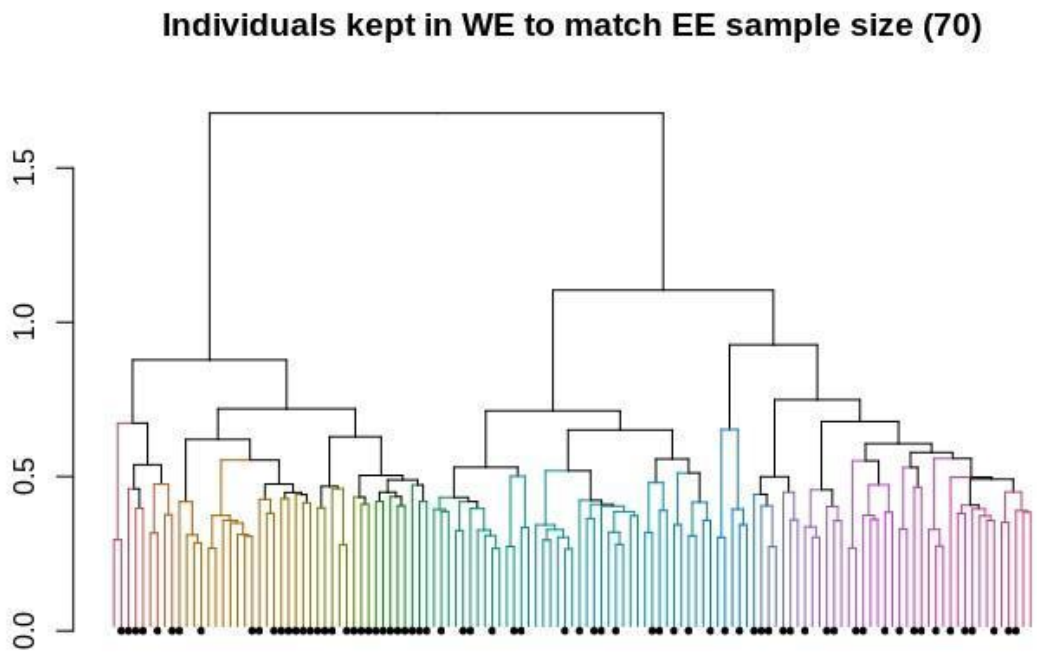


Figure S13: Hierarchical clustering of the 127 landraces of the WE population and subsampling of 70 landraces to mimic WE genetic diversity

Colored branches indicate the 70 clusters of this hierarchical clustering. Black leaves indicate the 70 sampled landraces (one per cluster) to form the WEds population.

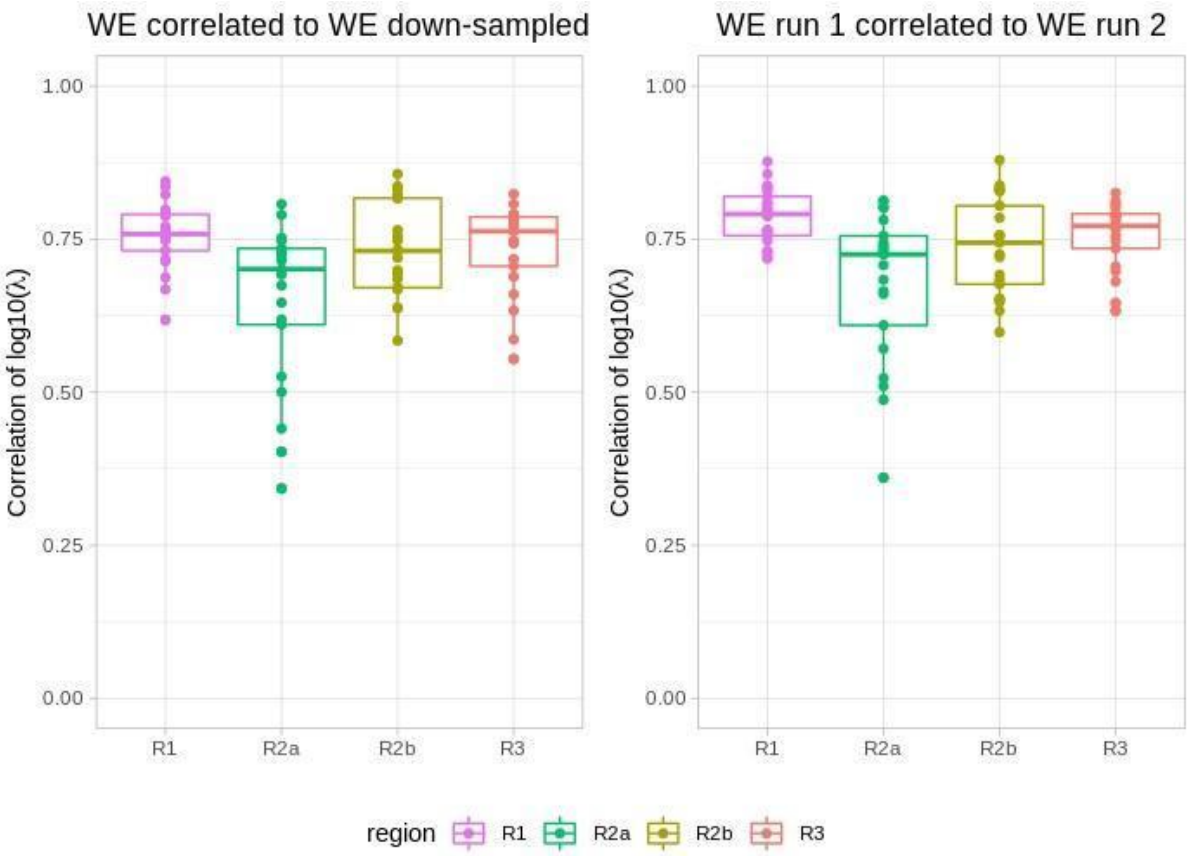


Figure S14: Average correlation of recombination profile intensity (log10(λ)) per genomic region between the WE and WEds population (left) and two independent analyses of the same population WE (right)

SNPs dataset and PHASE settings are the same across the analyses presented here. **Left.** The average correlation of recombination intensity profile of the WE and WEds populations across all genomic regions was 0.72 (∓0.1) (centromeric regions not included). **Right.** The average correlation of recombination intensity profile of WE population across two independent runs of PHASE is 0.74 (∓0.1) (centromeric regions not included). As the average correlation between WE and WEds is similar to the correlation obtained between two PHASE runs, we conclude that downsampling has little effect on LD-based recombination profile.

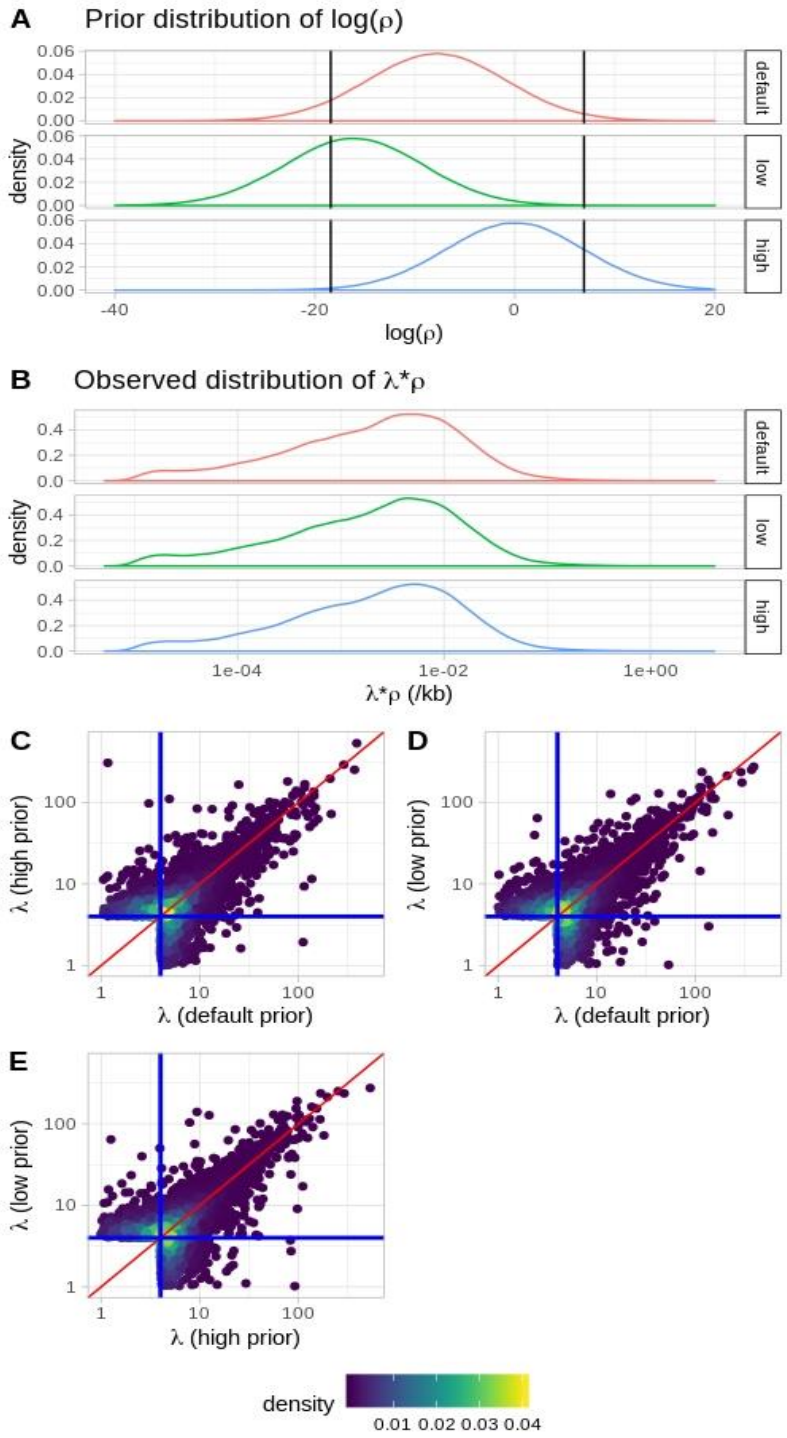


Figure S15: Robustness of PHASE inference to the prior distribution

A. The three tested prior distributions for background LD-based recombination rates: default prior distribution (mean of ρ_w = 4e-4); low prior distribution (1e-7) ; high prior distribution (1). **B**. Distributions of LD-based recombination rates estimates ($\rho_i = \lambda_i * \rho_{w(i)}$) using each prior distribution, and LD patterns of WE population. **C, D and E.** Relationships between local recombination rates intensities (λ_i) using two different prior distributions, in intervals claimed HRIs in one or both inferences in WE population.

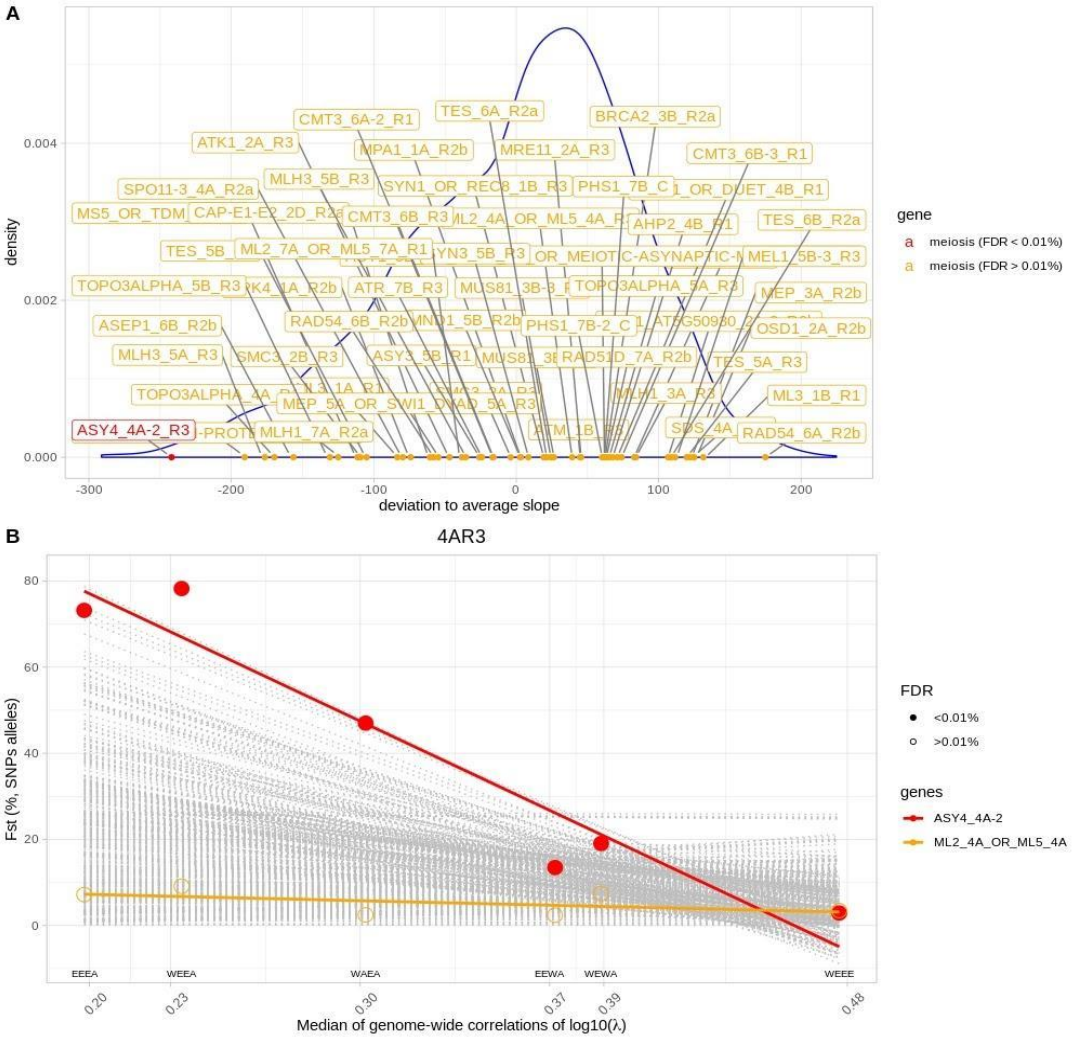


Figure S16: Differentiation of meiotic genes

Top: Distribution of deviations to average slope estimated with linear model (2) of supplementary protocol S3 for all control and meiotic genes. The average slope between F_{ST} and correlation of recombination profile is significantly negative, meaning that on average, the differentiation decreases with similarity in recombination profile. The deviation to average slope is interpreted here as the deviation to the background differentiation level. Very negative values indicate highly differentiated genes while very positive values indicate low differentiated genes. The 54 studied meiotic genes are represented by labels. Orange labels indicate meiotic genes whose deviation to background differentiation show a FDR ≥ 0.01%. Only one gene “ASY4_4A-2_R3” has a red label because its FDR is lower than 0.01%. This gene is located in the 4AR3 genomic region.

Bottom: Representation of relationship between F_{ST} for each gene of the 4AR3 genomic region with correlation of recombination intensity profile (measured by $\log_{10}(\lambda)$). Grey slopes represent linear relationship for each of the 475 control genes sampled in the genomic region. Coloured slopes represent linear relationship for the two genes of this genomic region possibly involved in meiotic process. Empty dots indicate whether theses meiotic genes show a FDR lower than 0.01%.

As we can see, the slope “ASY4_4A-2_R3” gene is very different from all others genes and very differentiated. At the opposite, the “ML2_4A” gene shows a low differentiation.

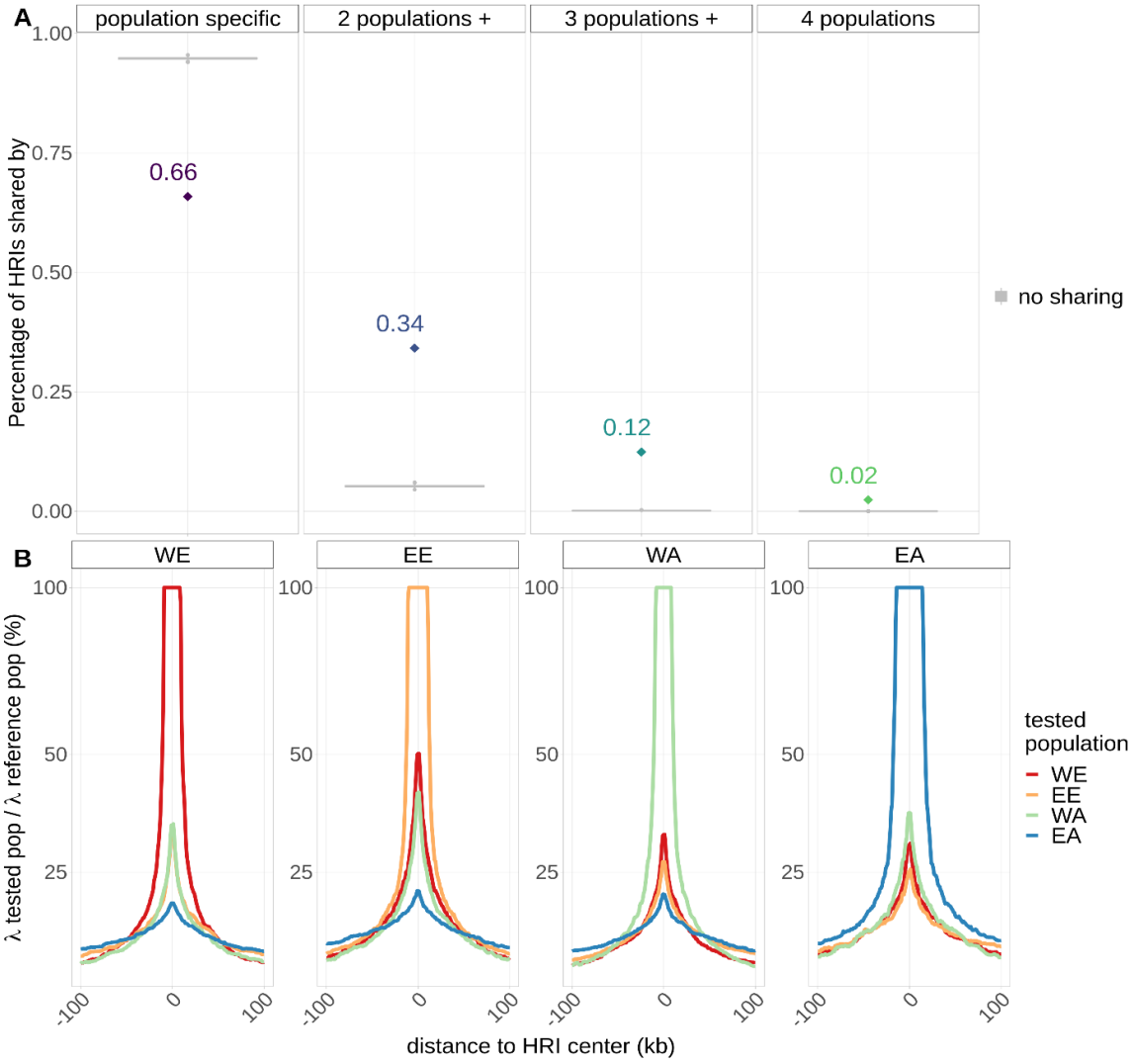


Figure S17: Proportion of shared HRIs when estimating LD-based recombination profiles on a common SNPs dataset for all populations

LD-based recombination rates were estimated using the same SNPs dataset for all populations (*i.e.* SNPs which are polymorphic in all populations). **A.** Proportion of co-localizing HR (coloured points) and simulated co-localizing values under random assignment of HRIs (grey boxplots) **B.** LD-based recombination intensity in each of the four populations WE, EE, WA and EA around HRIs specific to one population.

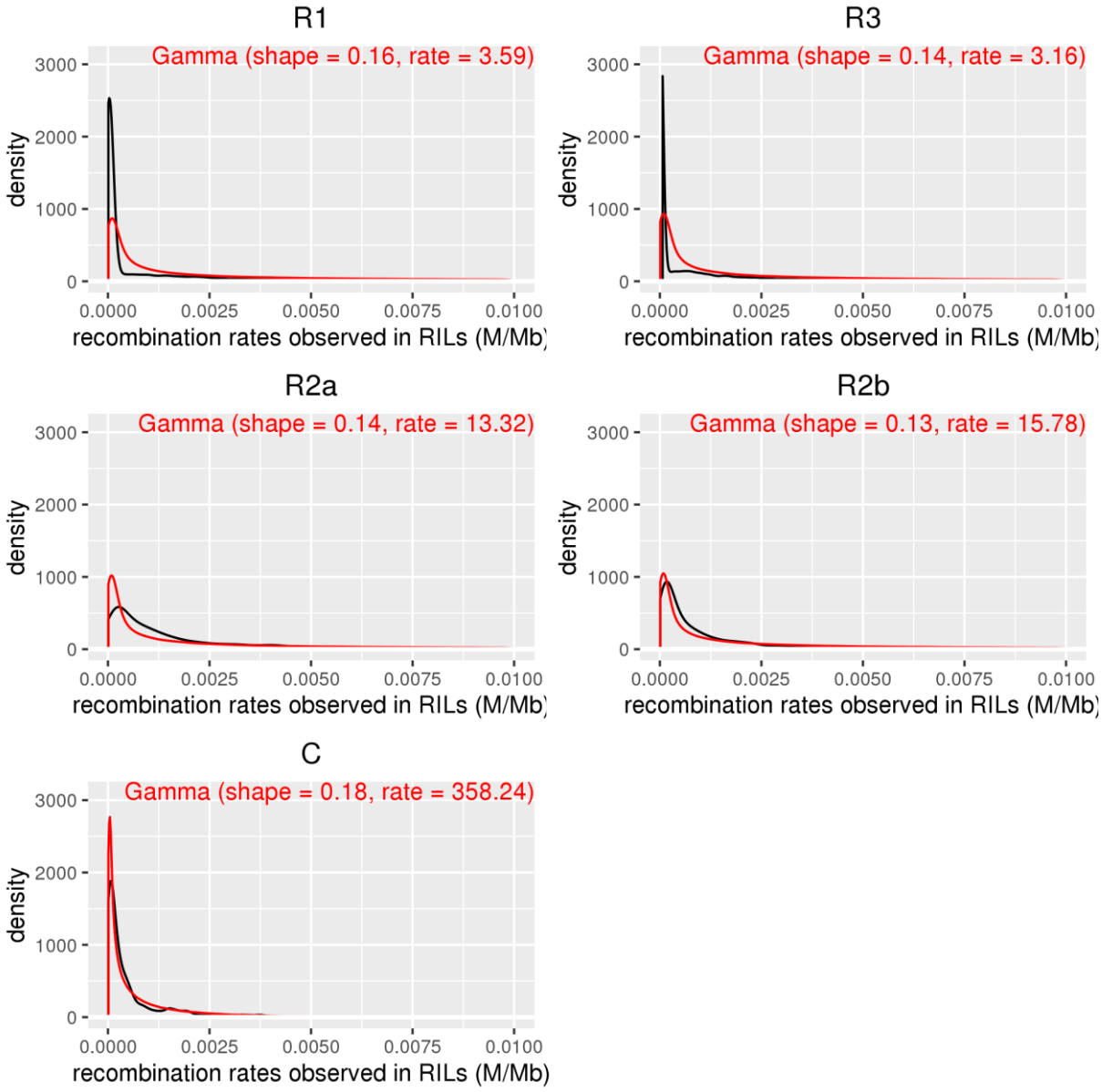


Figure S18: Prior distributions for RILs Bayesian recombination rates.

The black curve represents the distribution of frequentist recombination rates in RILs and the red curve the fitted Gamma distribution. The number of data points per black curve ranged from 4,799 in centromeric regions C to 28,756 in R3 regions, so that a single interval had a very low contribution when adjusting the Gamma distribution. The difference between the black and red distributions illustrates how prior inflates or shrinks RILs recombination rates in Bayesian model. While 60% of intervals had null frequentist estimates of recombination rates, they were replaced by the minimum recombination rates observed in the region to allow fitting a Gamma distribution (1.8e-3 M/Mb for R1, 2.5e-4 for R2a, 1e-6 for C, 1.7e-4 for R2b and 7.4e-4 for R3). However, those null intervals created a mass in empirical recombination rate distribution, decreasing the Gamma fitting quality. Note that the ratio between shape and rate for each region gives the average value of

recombination rate in the region. Bayesian estimates would converge toward those average values in case of very low informativity of intervals (no recombinant, small size of intervals).

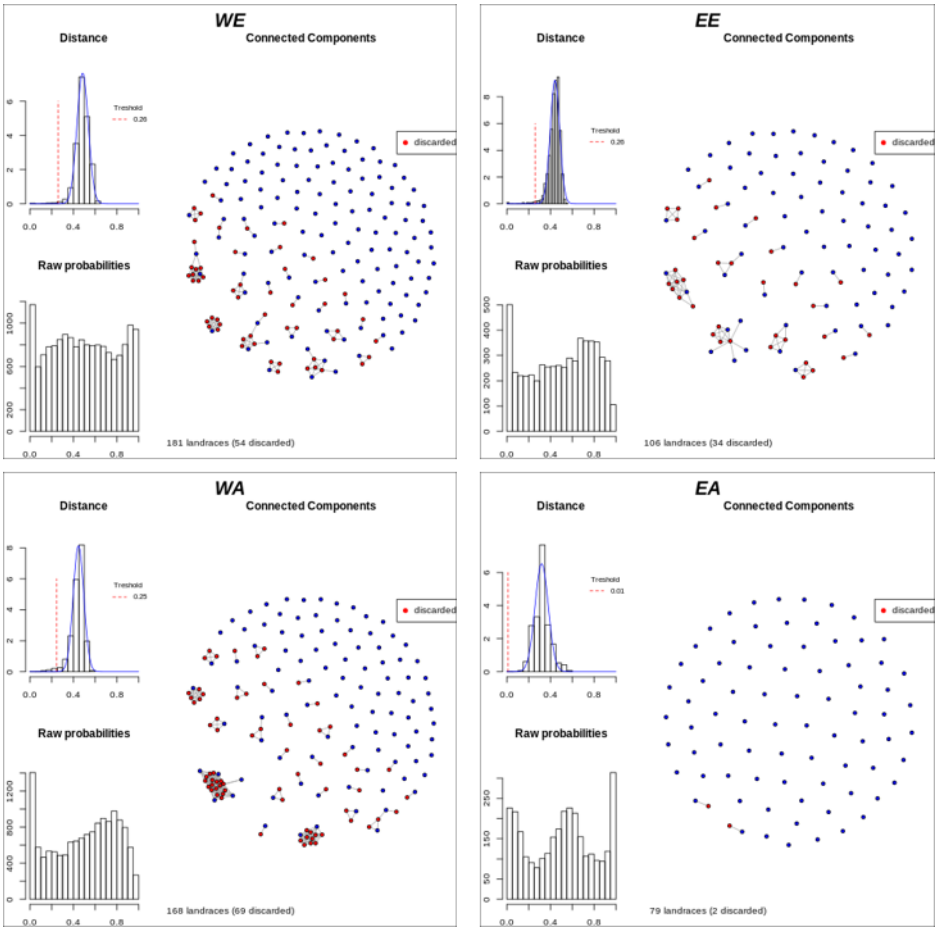


Figure S19: Identification and suppression of closely related landraces

Top left Distribution of simple matching distances, computed between each pair of landraces. Blue curve: Normal distribution to model similarity in the population, whose parameters were estimated using a Robust Fitting of Linear Models function in R, which is robust to outliers. Red dashed line: significant threshold to identify closely related landraces. **Bottom left** Distribution of P-values after, expected to follow a Uniform distribution if the simple matching distances distribution would match exactly the modelled Normal distribution. This P-values distribution was used as input in qvalue function to compute False Discovery Rate (FDR). A FDR of 1e-3 was used to set the significant threshold.

Right Representation of landraces relationship. Each point represents one landrace. Each strait line represents a close relationship. Red points correspond to landraces that were discarded to eliminate close relationships.

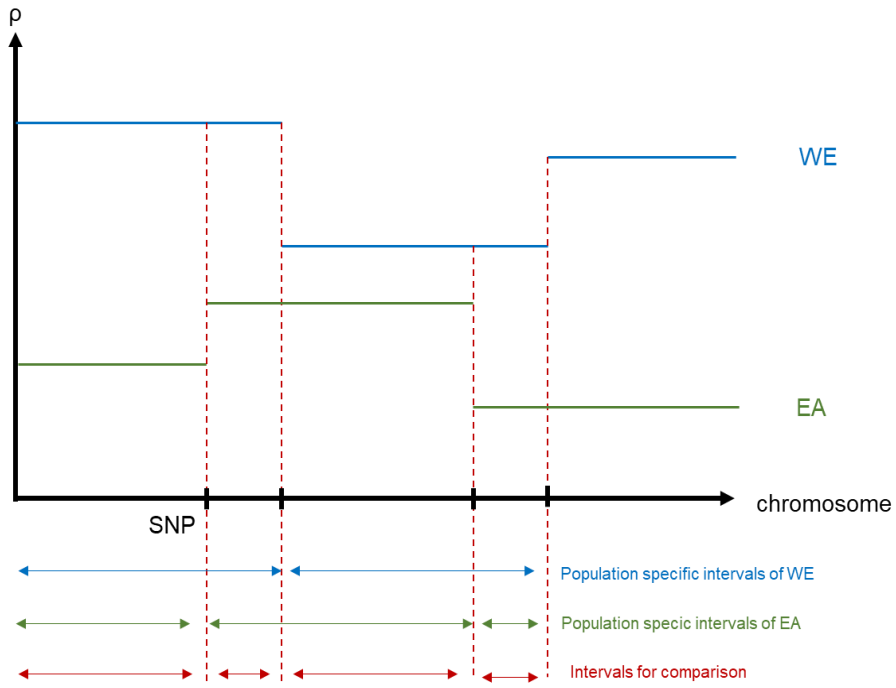


Figure S20: Defining intervals to compare fine-scale LD-based recombination rates using mixed models

Example with two populations WE and EA. LD-based recombination rates are not estimated in the same intervals in WE, EE, WA and EA because of the MAF filtering of SNPs. To compare LD-based recombination rates of the four populations, we defined intervals constituted of polymorphic SNPs in at least one of the four populations.

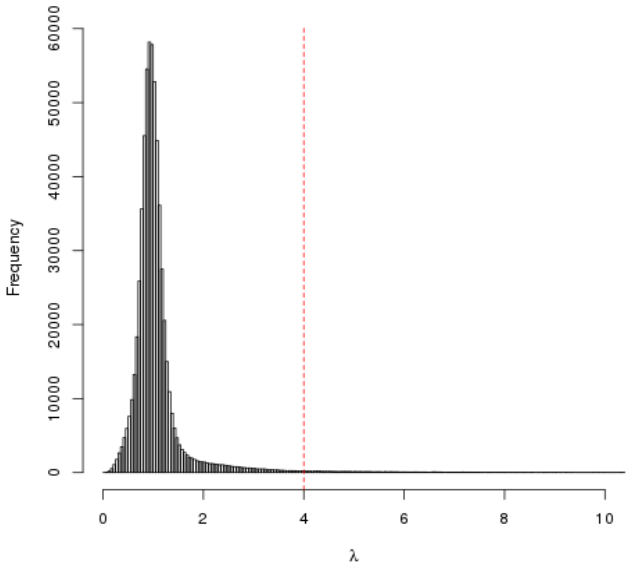


Figure S21: Distribution of LD-based recombination intensities λ for all intervals in all populations : Intervals with λ ≥ 4 are claimed HRIs.

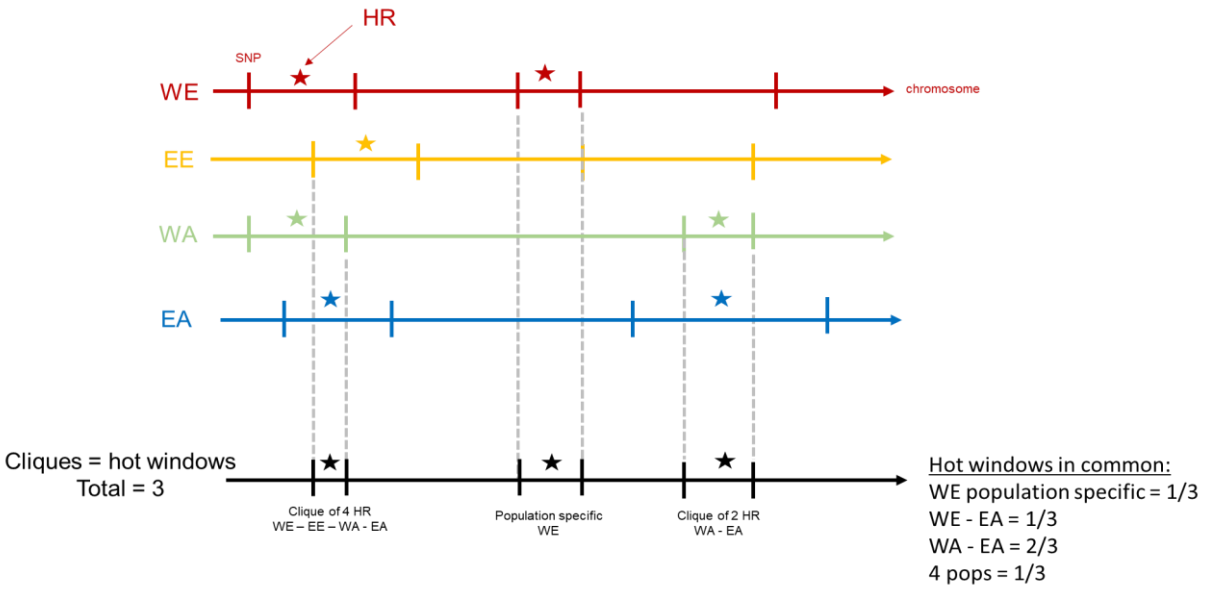


Figure S22: Examples of shared hot windows

A clique is defined as the systematic mutual overlap of HRIs of different populations. When a HRI doesn't overlap any HRI in other populations, it is called population specific. Upper and lower limits of hot windows are the inner join of HRIs borders.

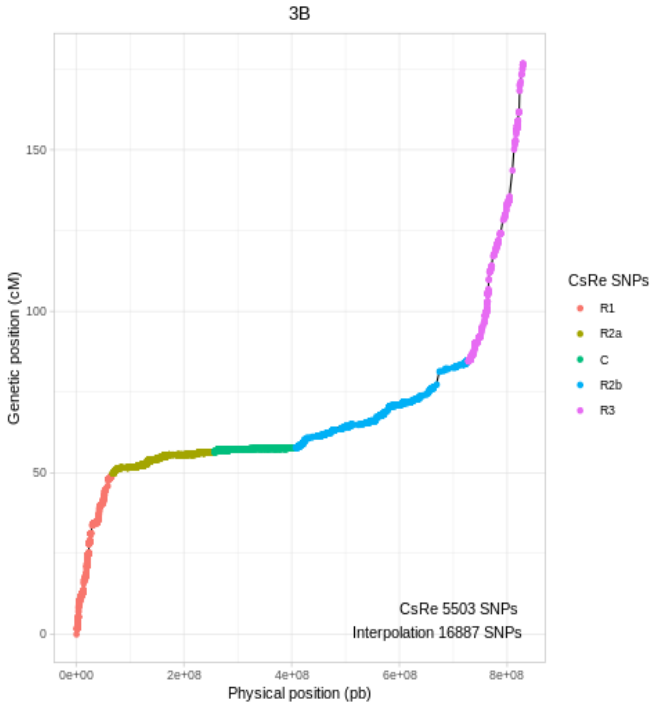


Figure S23: Interpolation of genetic position of SNPs located on 3B chromosome using CsRe Bayesian map

From mapped markers in CsRe population, we estimated the genetic position of umapped markers based on their physical position, by hypothesizing constant recombination rates within intervals of two successive mapped markers (supplementary protocol S5).

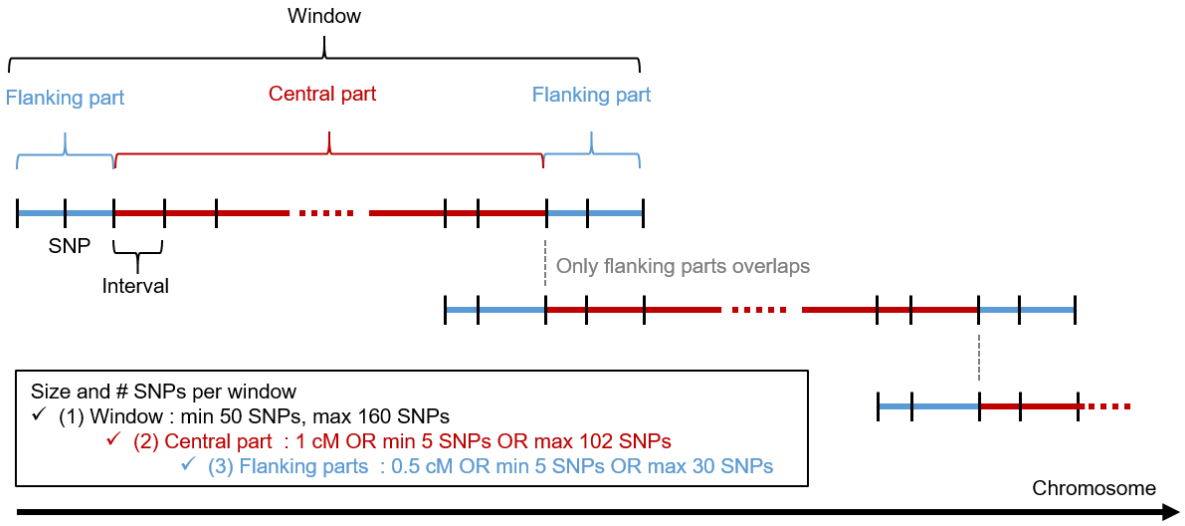


Figure S24: Definition of PHASE windows

To control both the genetic size of the window and the number of SNPs per window, we had to define constraints to form adequate windows, described in the black square (more details in supplementary protocol S5).

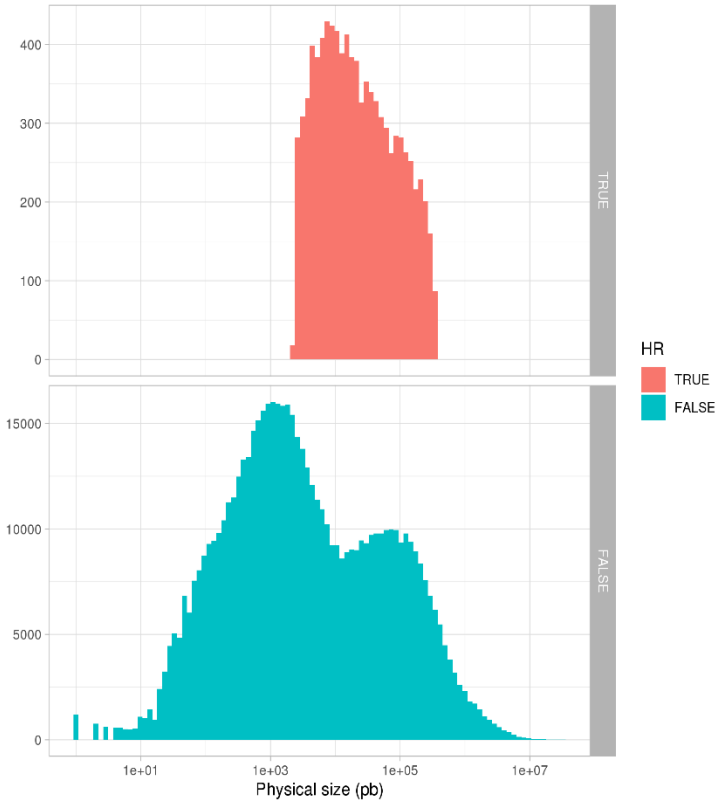


Figure S25: Distribution of physical size of Highly Recombing Intervals (HRIs, red) and other intervals (blue)

The filtering procedure yielded 8,713 HRIs, with size ranging from 2,369 to 344,607 pb.

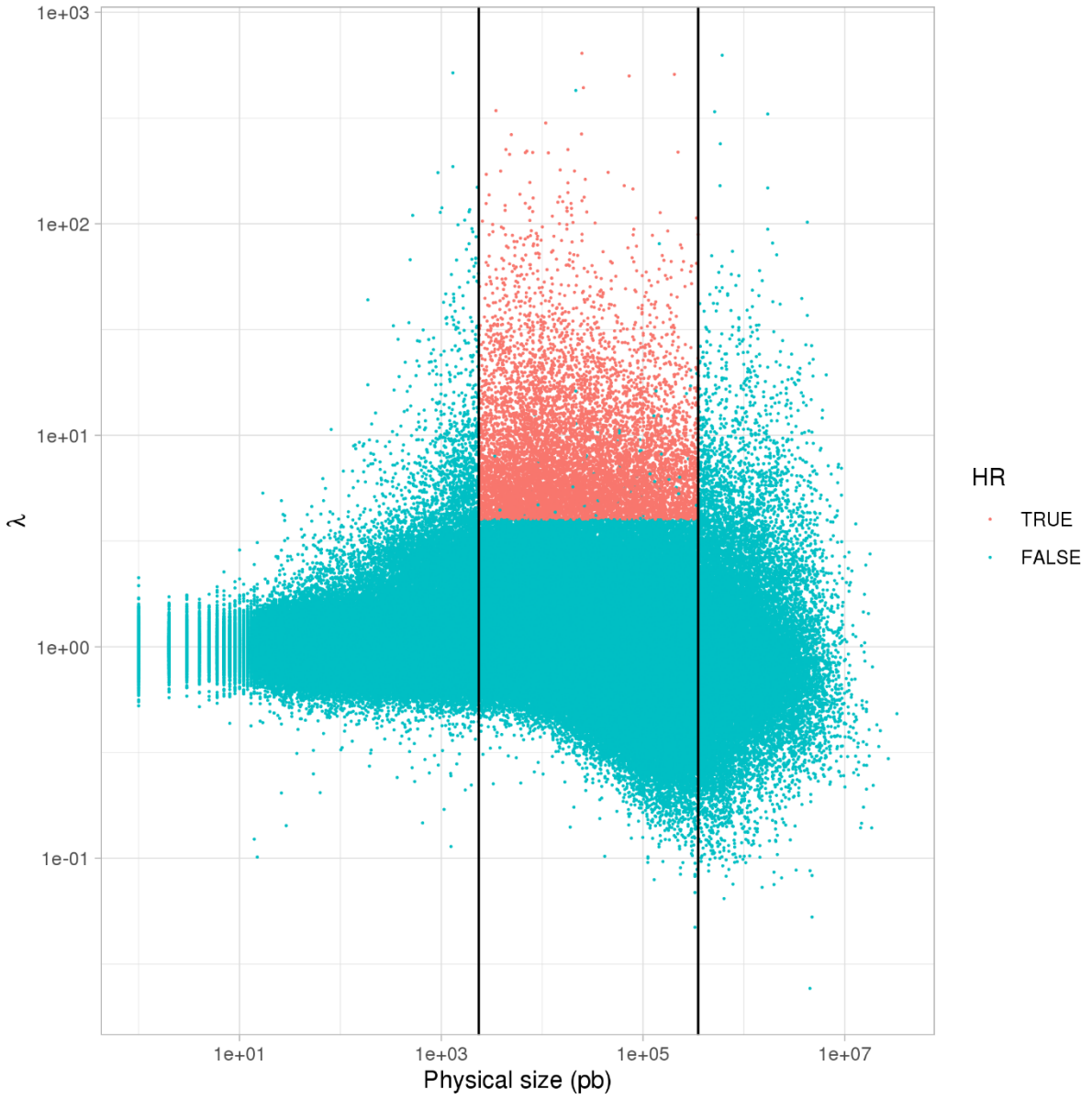


Figure S26: Relationship between historical recombination intensities λ related to the physical size of intervals

Note that intensity of Highly Recombining Intervals (HRIs) was quite independent of their physical size.

References

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. Sci Adv. 5(5):eaav0536.

Gini C. 1936. On the measure of concentration with special reference to income and statistics. Colo Coll Publ Gen Ser. 208:73–79.

IWGSC. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 361(6403):eaar7191.

Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W, Armisen D, Lang D, Bustos-Korts D, Goué N, Balfourier F, et al. 2019. Tracing the ancestry of modern bread wheats. Nat Genet. 51(5):905–911.

Stephens M, Smith NJ and Donnelly P. 2004 Documentation for PHASE, version 2.1. Chapter 6.2, p16. url: http://stephenslab.uchicago.edu/assets/software/phase/instruct2.1.pdf.

Stephens M, Carbonetto P, Gerard D, Lu M, Sun L, Willwerscheid J and Xiao N. 2020. ashr: Methods for Adaptive Shrinkage, using Empirical Bayes. R package version 2.2-47.

Storey JD, Andrew JB, Dabney A, David R. 2015. qvalue: Q-value estimation for false discovery rate control. R Package Version 2100.

Venables WN, Ripley BD. 2002. Modern applied statistics (Fourth S., editor) New York. Springer.

V.2 Supplementary: "Comparison of cross-selection criteria to optimize mating plans in a winter bread wheat breeding program"

V.2.1 Supplementary Protocols

- ***Supplementary Protocol S1: Fast computation of progeny variance***

Let's define N the number of candidate parents and M the number of markers. There are N(N-1)/2 candidate crosses. Progeny variance derived from the cross between parent i and j can be written as

$$\sigma_{ij}^2 = \beta' D^{ij} \beta \quad \textbf{[1]}$$

Where β is the vector of marker effects (size M*1) and D^{ij} is the variance-covariance matrix of progeny genotype specific of the cross i*j (size M*M).

Alternatively, progeny variance can be written as:

$$\sigma_{ij}^2 = (\beta^i - \beta^j)' \, \mathrm{D}^{(1)} \, (\beta^i - \beta^j) \quad \textbf{[2]}$$

Where β^i is the vector of marker effects for parent i (size M*1). Genotypes are coded 0, 1, 2. If the parent i is homozygous for the alternative allele (genotype coded with 2), the k-th element of the vector β^i is equal to 2*β_k, if parent i is heterozygous (genotype 1), the k-th element is equal to β_k and when parent i is homozygous for the dominant allele (genotype 0), the k-th element is equal to 0.

The variance-covariance matrix of progeny genotypes is named $\mathrm{D}^{(1)}$ (size M*M), and is common to every cross i*j. Diagonal elements of $\mathrm{D}^{(1)}$ are named $\mathrm{D}^{(1)}{}_{kk}$ for locus k, and are equal to 0.25. Off-diagonal elements of $\mathrm{D}^{(1)}$ are called $\mathrm{D}^{(1)}{}_{kl}$ for locus k and l. For Doubled Haploids (DHs) progenies, $\mathrm{D}^{(1)}{}_{kl}$= 0.25 (1−2r_{kl}) while for RILs F5 progenies $\mathrm{D}^{(1)}{}_{kl} = 0.25 * (1 - 2r_{kl}{}^5 - (0.5(1-2r_{kl}))^5)))$

Let's develop formula **[1]**:

$$\begin{aligned}\sigma_{ij}^2 &= (\beta^i - \beta^j)' \, \mathrm{D}^{(1)} \, (\beta^i - \beta^j) \\ &= \beta^{i\prime}\mathrm{D}^{(1)} \, \beta^i + \beta^{j\prime}\mathrm{D}^{(1)} \, \beta^j - 2 \, \beta^{i\prime}\mathrm{D}^{(1)} \, \beta^j \end{aligned} \quad \textbf{[3]}$$

Let's $\beta^{i\prime} = \gamma^{i\prime} \, \beta^{(2)}$ where γ^i is a vector (size 1*M) giving genotype of parent i. If the parent i is homozygous for the alternative allele (genotype 2), the k-th element of the vector γ^i is equal to 2, if parent i is heterozygous (genotype 1), the k-th element is equal to 1 and when parent i is homozygous for the dominant allele (genotype 0), the k-th element is equal to 0. The matrix $\beta^{(2)}$ is a diagonal matrix (size M*M) whose diagonal elements $\beta_{kk}^{(2)}$ is equal to marker effect β_k.

The formula **[3]** becomes:

$$\sigma_{ij}^2 = \gamma^{i\prime} \, \beta^{(2)} \, \mathrm{D}^{(1)} \, \beta^{(2)} \, \gamma^i + \gamma^{j\prime} \, \beta^{(2)} \, \mathrm{D}^{(1)} \, \beta^{(2)} \, \gamma^j - 2 \, \gamma^{i\prime} \, \beta^{(2)} \, \mathrm{D}^{(1)} \, \beta^{(2)} \, \gamma^j \quad \textbf{[4]}$$

Let's call $D^{(2)} = \beta^{(2)} D^{(1)} \beta^{(2)}$ the weighted variance-covariance matrix, common to every cross i*j. The formula **[4]** becomes:

$$\sigma_{ij}^2 = \gamma^{i\prime} D^{(2)} \gamma^i + \gamma^{j\prime} D^{(2)} \gamma^j - 2 \gamma^{i\prime} D^{(2)} \gamma^j \quad \textbf{[5]}$$

Let's define $\mu^i = \gamma^{i\prime} D^{(2)} \gamma^i$ (size 1*1). Let's also define the vector $\tau_i' = -2 \gamma^{i\prime} D^{(2)}$ (size 1*M) and κ_1^j the index of elements of γ^j equal to 1 and κ_2^j the index of elements of γ^j equal to 2. The quantity $\tau_i'\gamma^j$ (size 1*1) can be simply computed as $\tau_i'\gamma^j = \sum_{\{l|l\epsilon\kappa_1^j\}} \tau_{il} + 2*\sum_{\{l|l\epsilon\kappa_2^j\}} \tau_{il}$.

Example:

Take $\gamma^{j\prime}$= (2 2 0 1 0 0 2), so κ_1^j= (4) and κ_2^j = (1, 2, 7), so $\tau_i'\gamma^j = 2 * \tau_i'_1 + 2* \tau_i'_2 + \tau_i'_4 + 2 * \tau_i'_7$

Finally, the formula **[5]** can be converted in

$$\sigma_{ij}^2 = \mu^i + \mu^j + \tau_i'\gamma^j \quad \textbf{[6]}$$

Using formula **[1]**, one need to first compute one variance-covariance matrix D^{ij} (size M*M) for each of the N(N-1)/2 candidate crosses, which is already a huge task when the number of candidate parents N and/or the number of markers M are high.

To avoid this, one could use formula **[2]** to compute a unique variance-covariance matrix D (size M*M) common to every candidate crosses. Let's say that D has been computed. Computing progeny variance σ_{ij}^2 with the formula **[2]** would require to do $N(N-1)/2$ times M subtractions to compute $(\beta^i - \beta^j)$ for each candidate cross, and then $M \times (M+1)$ additions or products to compute $(\beta^i - \beta^j)'D$ for one cross, and finally M+1 supplementary additions or products to obtain $(\beta^i - \beta^j)'D(\beta^i - \beta^j)$ for one cross.

Let's define an addition (or subtraction) computation time T_a and a product computation time T_p. Computing progeny variance of each candidate cross with the formula **[2]** requires a total computation time of:

$$(N(N-1)/2) \times M \times (M+1) \times T_a + (N(N-1)/2) \times M \times (M+1) \times T_p \quad \textbf{[7]}$$

Formula **[3]** requires to compute the weighted variance-covariance matrix $D^{(2)} = \beta^{(2)} D \beta^{(2)}$ common to every candidate cross. This matrix $D^{(2)}$ requires some additional computation compare to D, but it should be computed only once. Let's say $D^{(2)}$ has been computed. Formula **[3]** also requires to compute

- N terms $\mu^i = \gamma^{i\prime} D^{(2)} \gamma^i$ (one per candidate parent). Each term μ requires $(M+1) T_a + M(M+1) T_p$ for a total of $N[M+1) T_a + M(M+1) T_p]$ computation time.
- N(N-1)/2 terms $\tau_i'\gamma^j$ (one per candidate cross). Each $\tau_i'\gamma^j$ term requires to compute $\tau_i' = -2 \gamma^{i\prime} D^{(2)}$ associated to a computation time of $MT_p + MT_a$ and then $\tau_i'\gamma^j$ associated to a computation time of around MT_a if all locus of second parent j are homozygous for alternative

allele or heterozygous. Thus, these N(N-1)/2 terms $\tau_i{}'\gamma^j$ are associated to a total of N(N-1)/2[MT_p+ $2MT_a$]. To sum up, formula **[6]** require a computation time of

$$N[M+1)\ T_a + M(M+1)\ T_p\ ] + N(N-1)/2[T_p + 2MT_a] \quad \textbf{[8]}$$

Directly computing progeny variance for every candidate cross with formula **[1]** or **[2]** yield a computation time of order N^2M^2 while using the piece-wise formula **[6]** yields a computation time of order $NM^2 + N^2M$. For any value of M and N, the second method **[6]** is always faster than the first methods **[1]** or **[2]** because $N^2M + NM^2 \ll N^2M^2$.

To gain even more computation time, one should compute progeny variance σ_{ij}^2 for each chromosome independently [$\sigma_{ij}^2{}_c$where c is the chromosome] and then sum the variance of all chromosomes [$\sigma_{ij}^2 = \sum_c \sigma_{ij}^2{}_c$]. Indeed, segregation of chromosomes during meiosis is random and thus the expected recombination frequency for locus located on different chromosomes is 0.5. The off-diagonal elements of variance-covariance of progeny genotypes are thus equal to 0 when locus are not located on the same chromosome.

- ***Supplementary protocol S2: Genetic algorithm***

The mating plans based on EMBV require a heuristic algorithm to be optimized. We used a Genetic Algorithm (GA) inspired from Darwinism. GA starts from a fixed-sized population of individuals representing candidate solutions. A selection process is applied to reproduce the most promising solutions. Some pairs of candidates are crossed to create children, other individuals are muted. The new solutions built replace their parents in the next generation.

In our case, a candidate solution is a vector of D_{ij}, initially randomly generated. In order to respect the constraints, this vector is corrected after every modification (crossover, or mutation), and before evaluation.

- First the number of parents is checked (Constraint C5): if too many parents are represented some D_{ij} are randomly set to zero, if too few parents are selected new Dijs are moved from zero to none zero values.

- Second, the number of progenies derived from one parent is checked (Constraint C4). If it is too big, some Dijs are reduced to respect this constraint without violating constraint C5.

- Third, if the number of different progenies (number of non-zero D_{ij}) is too small, new ones are added, if it is too big it is reduced (Constraint C3). This is done without violating constraints C5. Respecting constraint C4 can be difficult in some cases: consequently, if the number of attempts n tried to meet constraint C3-C5-C4 exceeds the size of the parent population, constraint C4 is

released and instead of accepting 250 progenies issued from the same parent we increase the boundary to 251, and so on until the constraint can be respected.

- Fourth, D_{ij} values are modified when necessary to respect constraint C2 without violating constraints C3, C4 and C5. Again, C4 can be released after a number of failed attempts exceeding the parent population size.

- Finally, the sum of D_{ij} is checked to meet the number of progenies chosen (Constraint C1). At this step, it can be difficult to randomly modify D_{ij} values and respect constraints C5, C4, C3 and C2, and exactly reach the fixed value. C4 can once again be released if necessary.

If C4 has been released the evaluation function in the GA is penalized: if x is the max number of progenies issued by a parent and x>250, the criteria presented in the previous section is multiplied by 250/x. With such a penalization, the best GA population element observed at the end of the algorithm always meets constraint C4 in our simulations.

Two crossover operators are defined and are equally randomly chosen. One randomly creates two children by randomly picking the $D_{ij}$from a solution candidate or the other. The other is inspired by the classic arithmetic crossover. In both cases the children are corrected to respect constraints C1 to C5.

The mutation operator is simple: two D_{ij} are chosen in a candidate solution and they exchange randomly between 1 and 5 progenies, the mutated candidate solution is corrected to respect constraints. The process is repeated randomly between 1 to 10 times and the best solution found is kept.

The sharing process requires the definition of a distance between candidate solutions. There are many ways to define such a distance. In the present case, we defined a simple distance which only sums non zero D_{ij} in one candidate solution that are zeros in the other one and reciprocally.

We used the following values for the application:

-Population size: 100

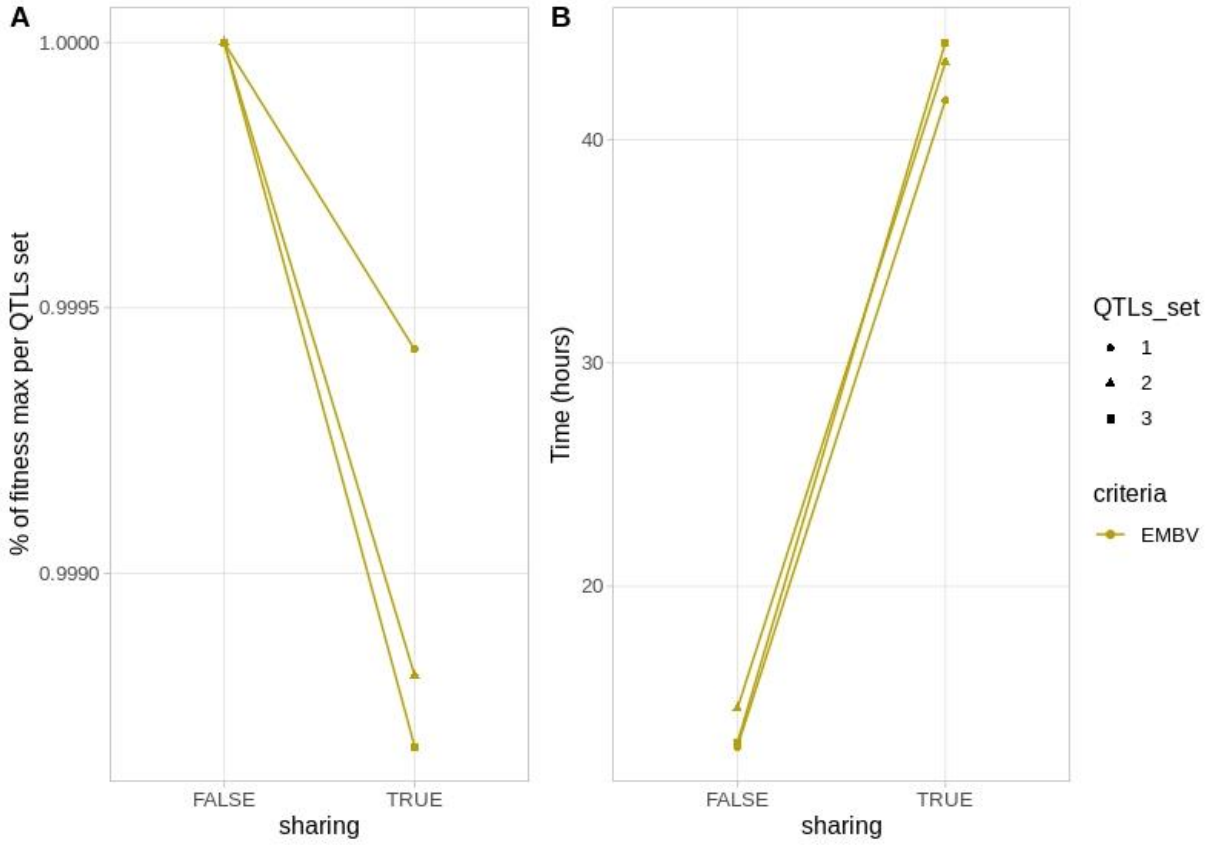
-Number of generations: 400 000

-Crossover rate: 30%

-Mutation rate: 50%

- ***Supplementary protocol S3: Effect of sharing in Genetic Algorithm convergence***

GAs are difficult to tune and get often stuck into local minima. To avoid premature convergence, a sharing process can be added before the selection (Yin and Germay 1993) in order to give more chance to candidates that are isolated in the search space. We tested the influence of sharing process on the final value of objective function when optimizing mating plan for criteria EMBV. We randomly choose 3 “selected populations” among the 20 populations available. We pre-selected 10% of crosses with highest PM value to limit computation time. For each of these 3 populations, we optimized mating plan with or without sharing. Each optimization was run two times independently (with different initialization of pseudorandom number generator). For each population, we kept the output with highest value of objective function among the 2 optimizations using sharing. We did the same for the 2 optimizations without sharing. Among these 2 remaining optimizations, the one associated with the highest value of the objective function was defined as control.



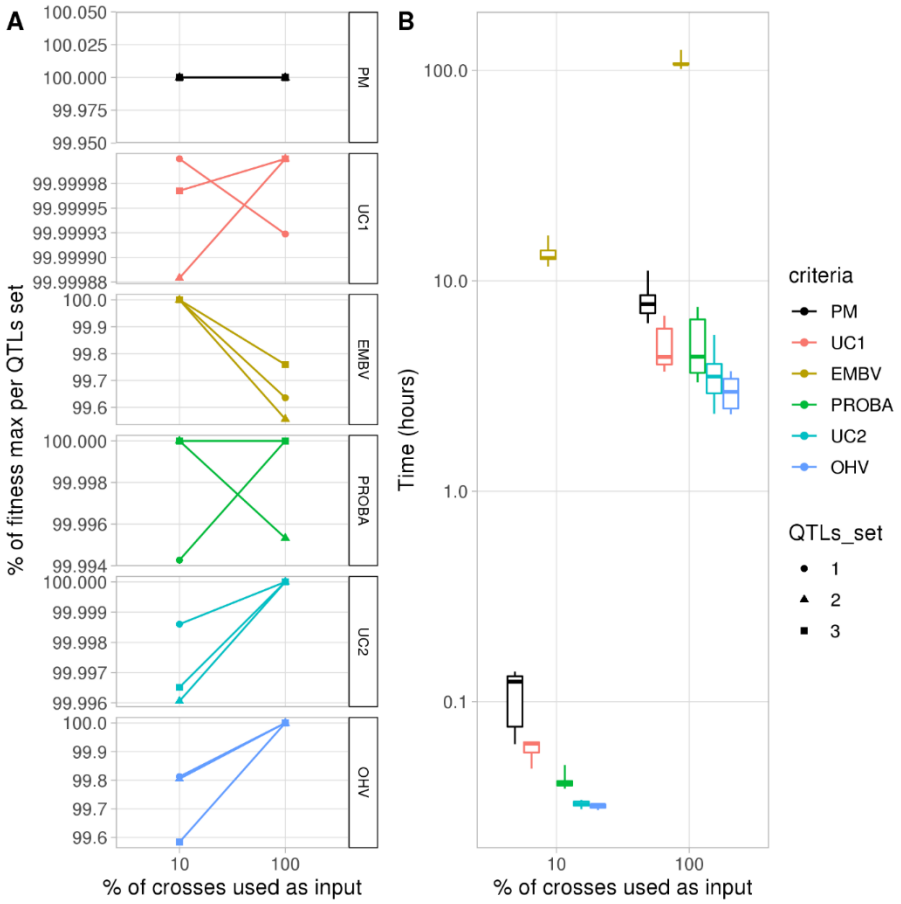
A. Influence of sharing on final value of objective function. *Each point represents one of the three randomly chosen population. On the y-axis is represented the ratio between the objective function value divided by the highest objective function value in optimizations with or without sharing. Each point represents one of the three randomly chosen population.*

B. Computation time associated with sharing.

The sharing did not provide higher optimal value of objective function, but the computing time was more than doubled.

- ***Supplementary protocol S4: Effect of pre-selecting the 10% best crosses in value of the objective function***

To limit computation time, we tested whether pre-selecting the 10% crosses with highest PM had an influence on the final value of objective function. We choose the same populations used in supplementary protocol S1. We optimized mating plans for each of the 6 usefulness criteria with or without a pre-selection of the 10% of crosses with highest PM value. Each mating plan was optimized two times independently (with different initialization of pseudorandom number generator). For each population and each dataset (either 10% crosses, either 100% crosses), we kept the output with highest value of objective function among the two optimizations. We also computed the standard deviation of the two objective function values to evaluate variability between two optimizations using the same dataset. Among these 2 best mating plans per population (one for 10% dataset, one for 100% dataset), the one associated with the highest value of the objective function was defined as control. Each point of the next figure is the ratio between the best value of the objective function obtained from a subset of crosses (either 10%, either 100%), divided by the control. Note that standard deviations (variability between two independent optimizations using the same dataset) for each point does not exceed 0.004 (from 0 to 0.00349).



A. Influence of pre-selection of crosses on final value of objective function. *Each point represents one of the three randomly chosen population. On the y-axis is represented the ratio between the objective function value divided by the highest objective function value obtained for each population.* ***B. Computation time associated with each data.*** The objective function value of pre-selected data represented minimum 99% of the objective function of complete data.

We also computed the proportion of crosses and parents in common within the two independent optimizations at 10% crosses, within the two independent optimizations at 100% crosses and between the two optimizations at 10% crosses and the two optimizations at 100% crosses.

In the next Table, proportion of parents and crosses in common are computed as the proportion of parents recruited twice between the two optimizations at 10%, twice between the two optimizations at 100%, and four times when considering the two optimizations at 10% and the two optimizations at 100%. Means and standard deviations are computed on the 3 populations.

CSC	Parents in common	Crosses in common	Parents in common	Crosses in common	Parents in common	Crosses in common
crosses	40k		400k		40k and 400k	
PM	100 ± 0	39 ± 3	100 ± 0	27 ± 1	100 ± 0	9 ± 4
UC1	98 ± 0	97 ± 1	98 ± 2	97 ± 2	97 ± 1	95 ± 3
EMBV	89 ± 5	35 ± 5	74 ± 2	22 ± 1	67 ± 3	7 ± 1
PROBA	97 ± 3	97 ± 1	98 ± 2	96 ± 4	93 ± 2	90 ± 4
UC2	99 ± 1	99 ± 1	100 ± 0	99 ± 2	95 ± 2	94 ± 3
OHV	100 ± 0	1 ± 0	100± 0	100 ± 0	80 ± 6	74 ± 8

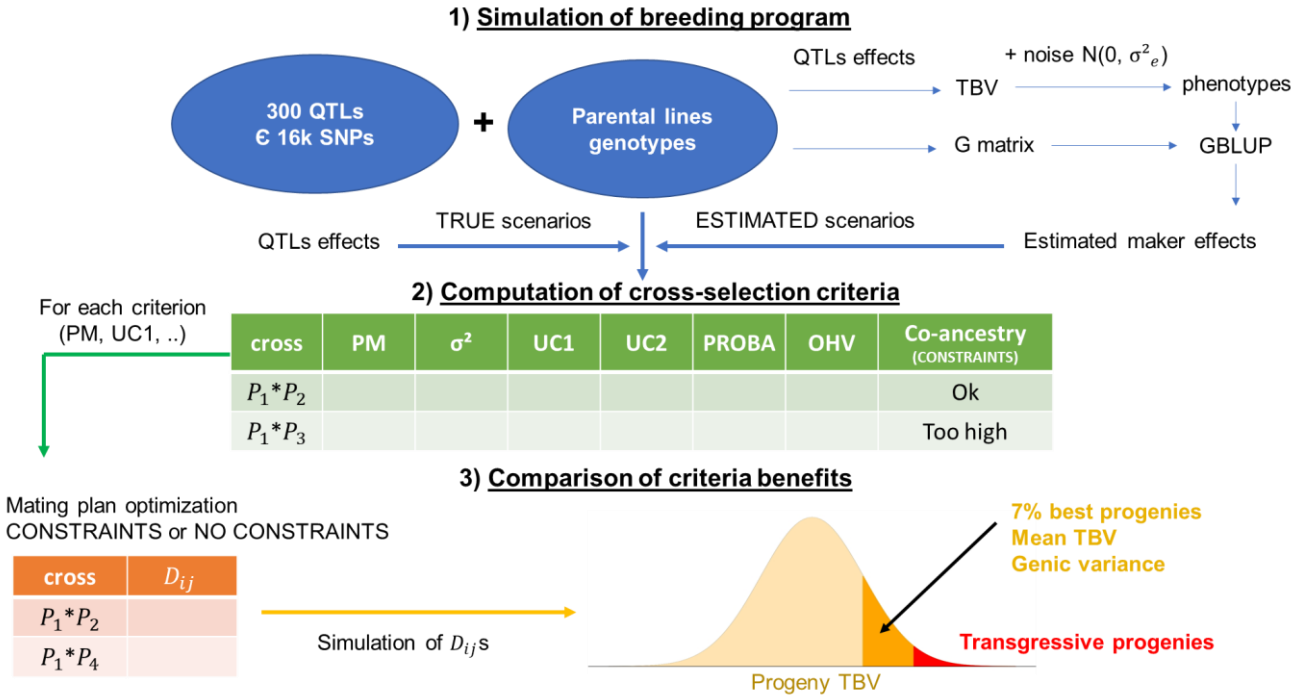
The proportion of parents in common in the PM optimization is always equal to 1. However, the proportion of crosses in common is quite low (≤ 39%). This suggest that PM criteria recruits the same parents but the way they are crossed does seems to matter much. In other words, PM likely seeks to maximize contributions of parent with highest genetic values, but there is a random part in mating decisions.

The EMBV criterion displays the lowest proportion of parents and crosses in common in all comparisons. This suggest that we didn't wait enough for the genetic algorithm to fully converge. However, despite big differences in mating plans, the objective function value of EMBV only slightly varies across two independent optimizations on the same dataset (mean > 99% of maximal objective function value with standard deviation < 0.004) and across optimizations using different datasets (from 100% of maximal objective function to > 99%, previous figure). This shows that this lack of convergence likely addresses the selection of crosses that does not contribute much to the objective value function, so crosses with low EMBV.

The CSC UC1, PROBA and UC2 show quite stable mating plans between optimizations, with more than 90% of crosses in common and more than 93% of parents in common between optimization at 40K crosses and optimization at 400k crosses. A part of these differences might be explained by a random process in the linear programming optimization. Indeed, on previous figure, some optimization for UC1 and PROBA show a lowest objective value at 400k crosses compare to 40k crosses, which would not be possible if the linear programming solver was fully deterministic. However, proportion of crosses and parents in common are slightly decreasing when comparing the 10% and the 100% crosses optimization. This suggest that pre-selection slightly affect mating plans. But the objective function is not much impacted (previous figure), suggesting that differences in mating plans addresses crosses with low usefulness (as in EMBV).

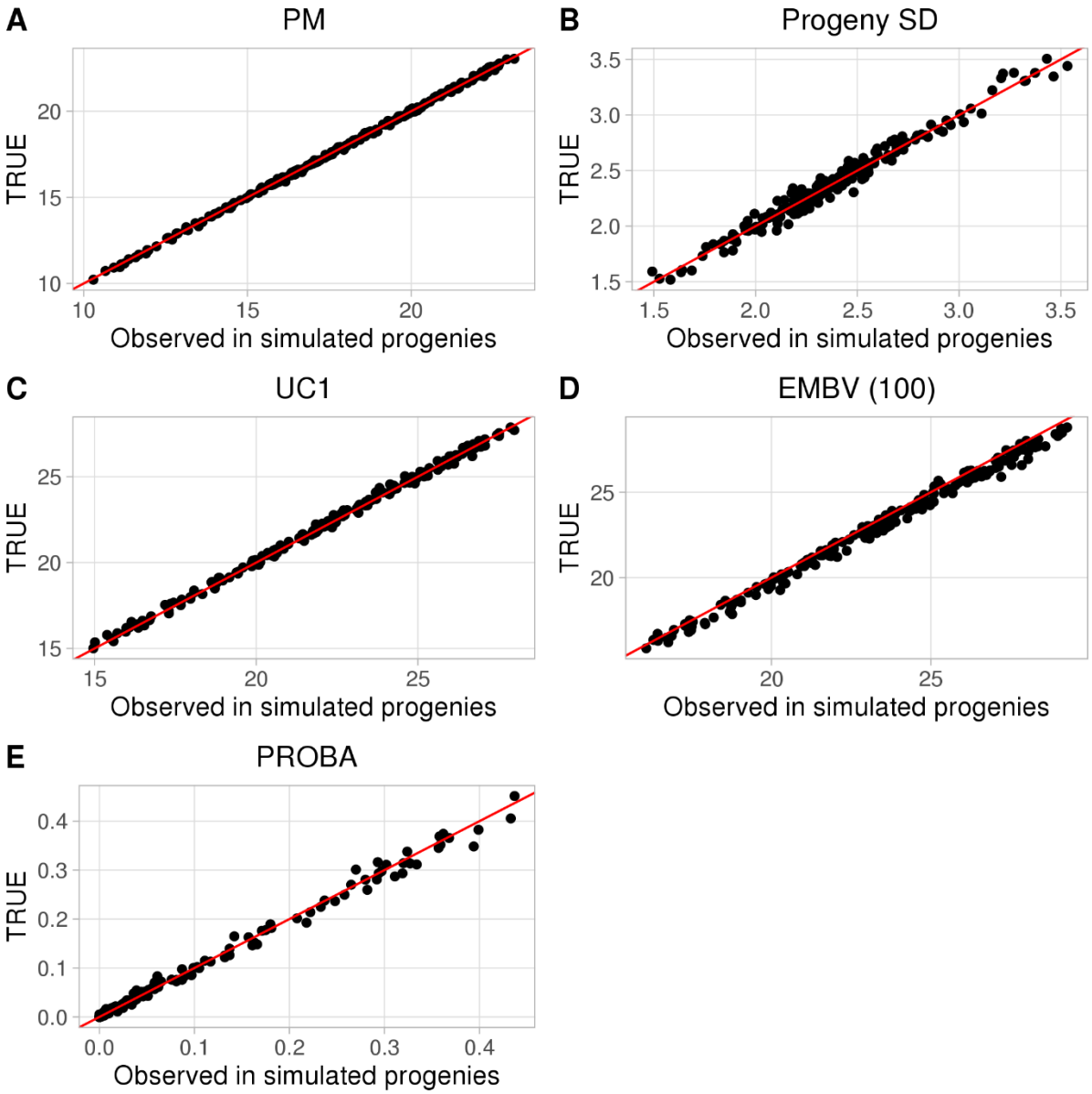
The criterion OHV seems the most sensitive criteria to a pre-selection of crosses. For OHV, the optimization at 40k crosses and at 400k crosses only show only 80% of parents and 74% of crosses in common. This is consistent with Figure 1 of the article showing that OHV is the less correlated criteria to PM.

V.2.2 Supplementary Figures



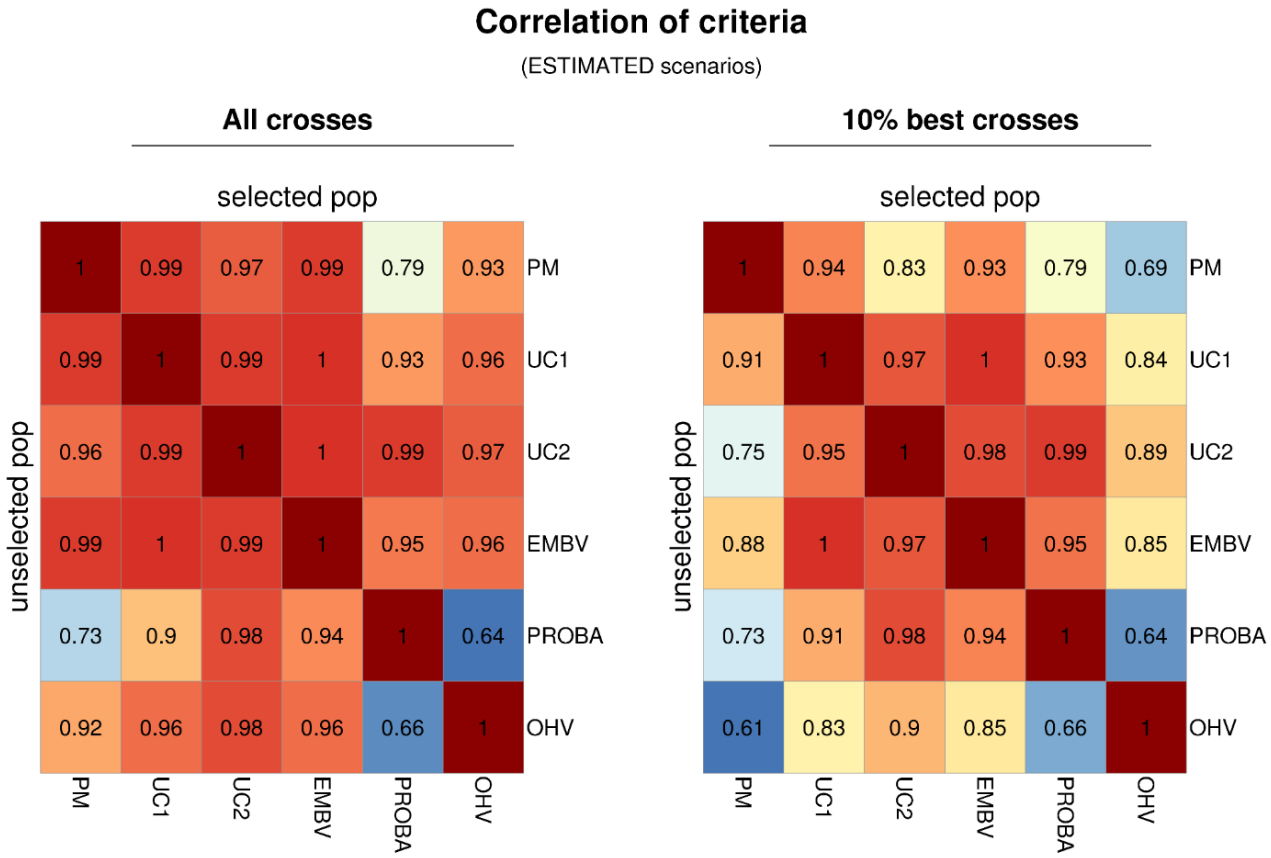
Supplementary Figure S1: Representation of simulated experiments

Parental lines genotypes can be either 835 INRAE/AO bread wheat lines in the “unselected populations” scenarios, or parental lines genotypes can be 900 lines derived from these 835 real lines by three selection-mating cycles.

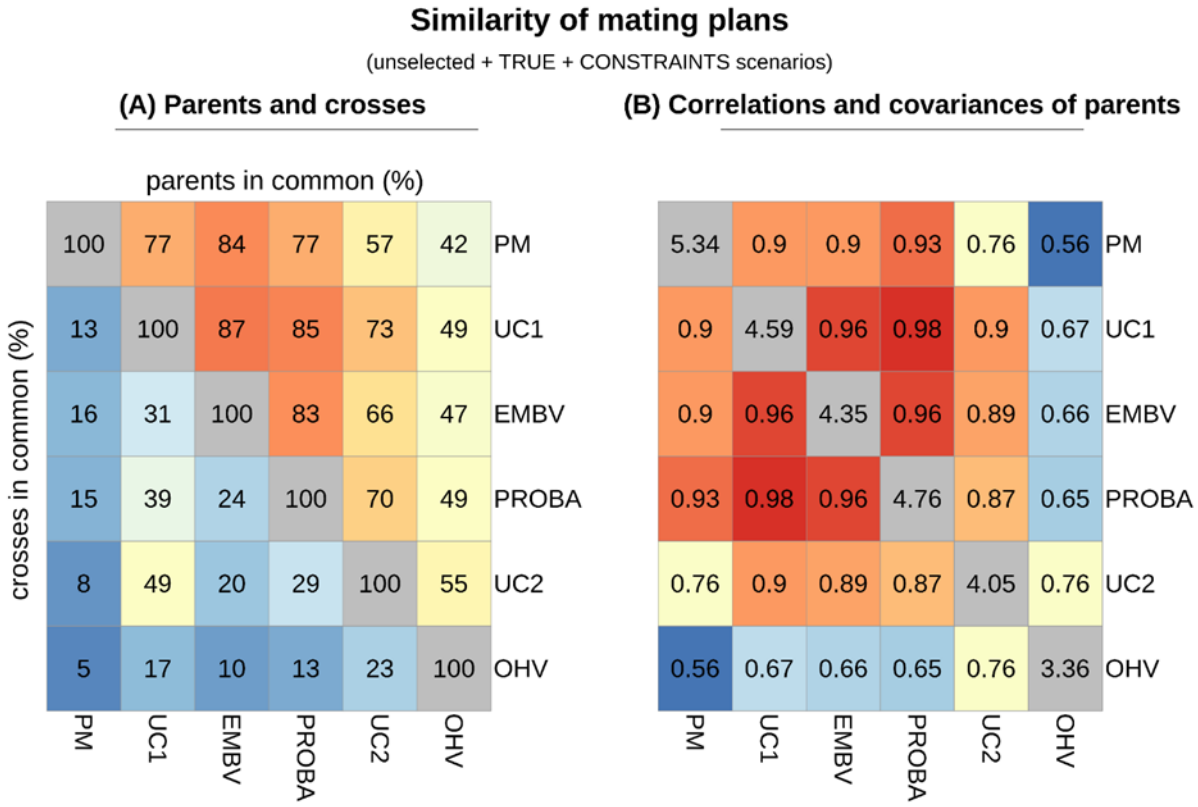


Supplementary Figure S2: Accuracy of usefulness criteria in TRUE scenarios

From QTLs effects and genotypes of 200 randomly chosen crosses, we computed the expected mean of progeny (PM), the expected standard deviation of progeny (Progeny SD), the expected mean of the 7% best progeny of each cross (UC1), the expected value of the best progeny when 100 progenies are allocated to each cross (EMBV(100)) and the expected proportion of progenies superior to the best parental line of the breeding program (PROBA). We also simulated 1 000 progenies per cross and computed these quantities from the distribution of progeny True Breeding Values. The red line is is y=x curve.

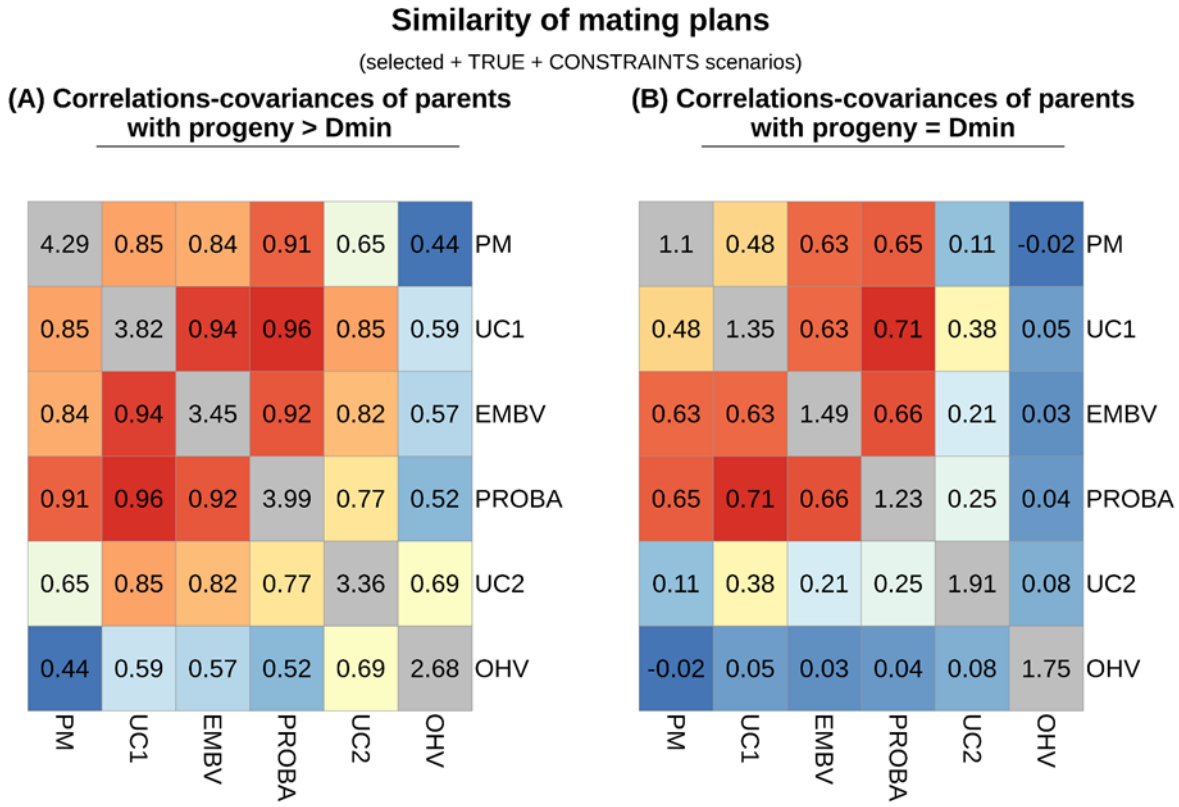


Supplementary Figure S3: Correlation of criteria in ESTIMATED scenarios



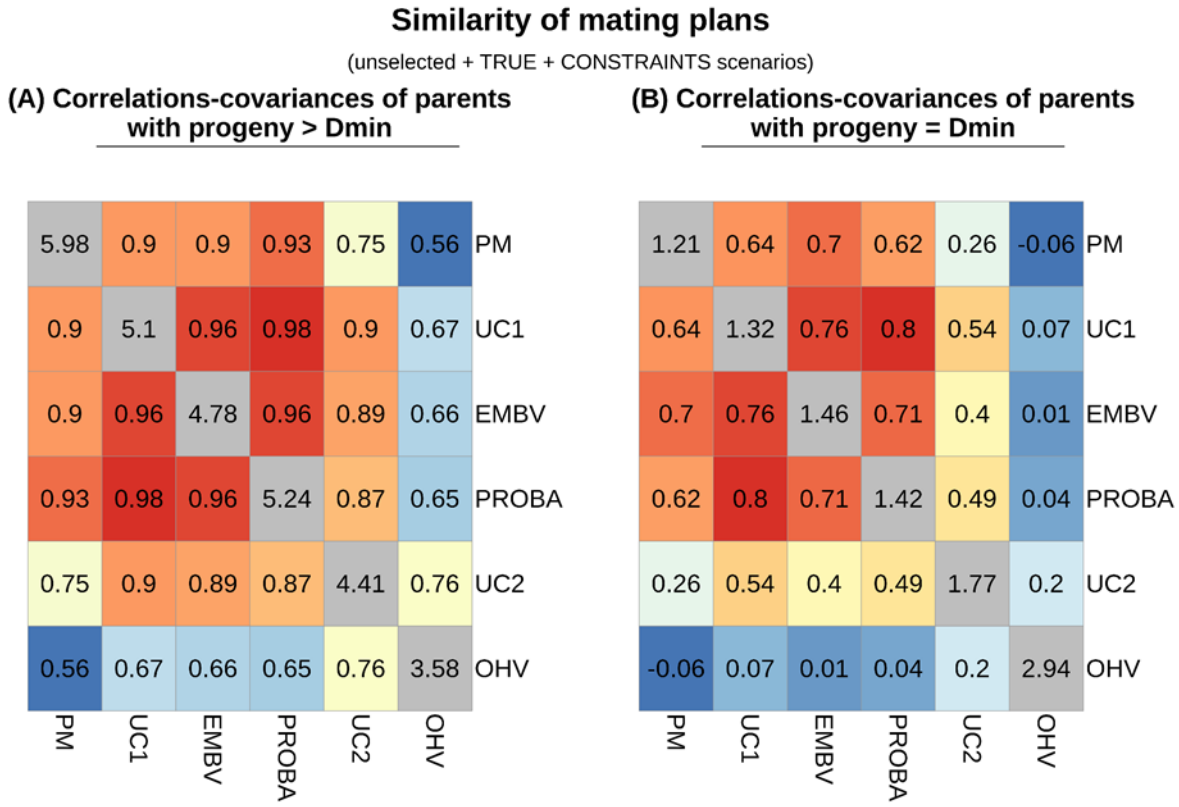
Supplementary Figure S4: similarity of mating plans

Results are given for unselected populations + TRUE+ CONSTRAINTS scenarios (criteria are computed from QTLs effects and parental contributions are limited). ***A: Parents and crosses shared by mating plans of each CSC.*** *For pairwise criteria, the number of crosses in common (and respectively the number of parents in common) divided by the total number of crosses (respectively the total number of parents) was computed and averaged over repetitions (n=20).* ***B: Genetic similarity of mating plans.*** *On the diagonal, weighted variances of recruited parents computed as* $c_{criteria1} * K * c'_{criteria1}$*, averaged over repetitions (n=20). Off-diagonals, weighted correlations of recruited parents for each pair of criterion computed as* $c_{criteria1} * K * c'_{criteria2}$*, divided by the square root of* $(c_{criteria1} * K * c'_{criteria1}) * (c_{criteria2} * K * c'_{criteria2})$*. The K relationship matrix was estimated from genotypes with LDAK software to decrease the bias due to LD between markers. The* $c_{criteria}$ *vector of parental contribution gives the proportion of progeny allocated to each parent in the mating plan optimized according to the CSC.*



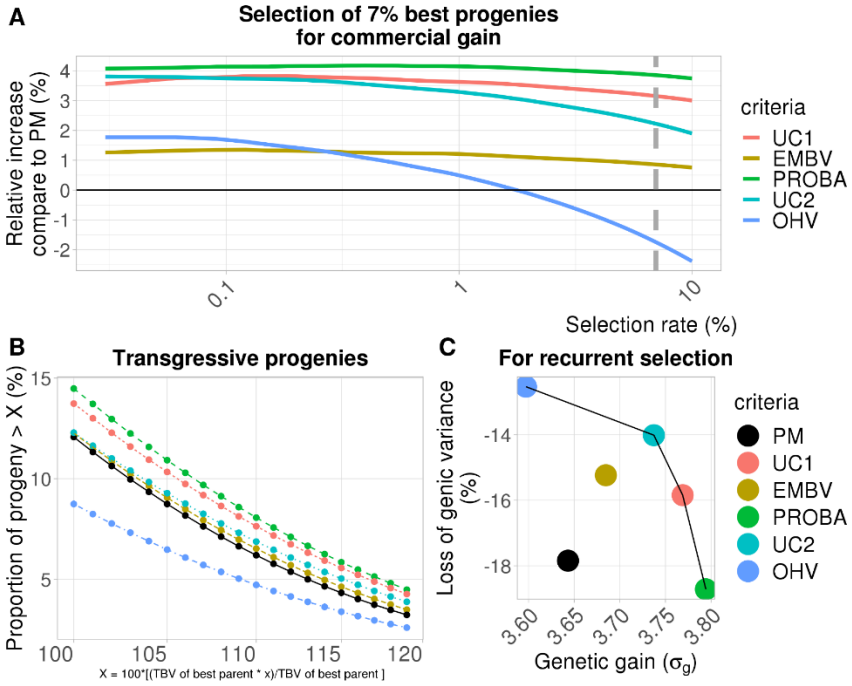
Supplementary Figure S5: Genetic similarity of selected parents with varying number of progenies in selected populations + TRUE + CONSTRAINTS scenarios

Genetic similarities were computed as in supplementary Figure S3. ***A****: Parents with more than Dmin progeny were kept.* ***B****: Parents with Dmin progeny were kept.*



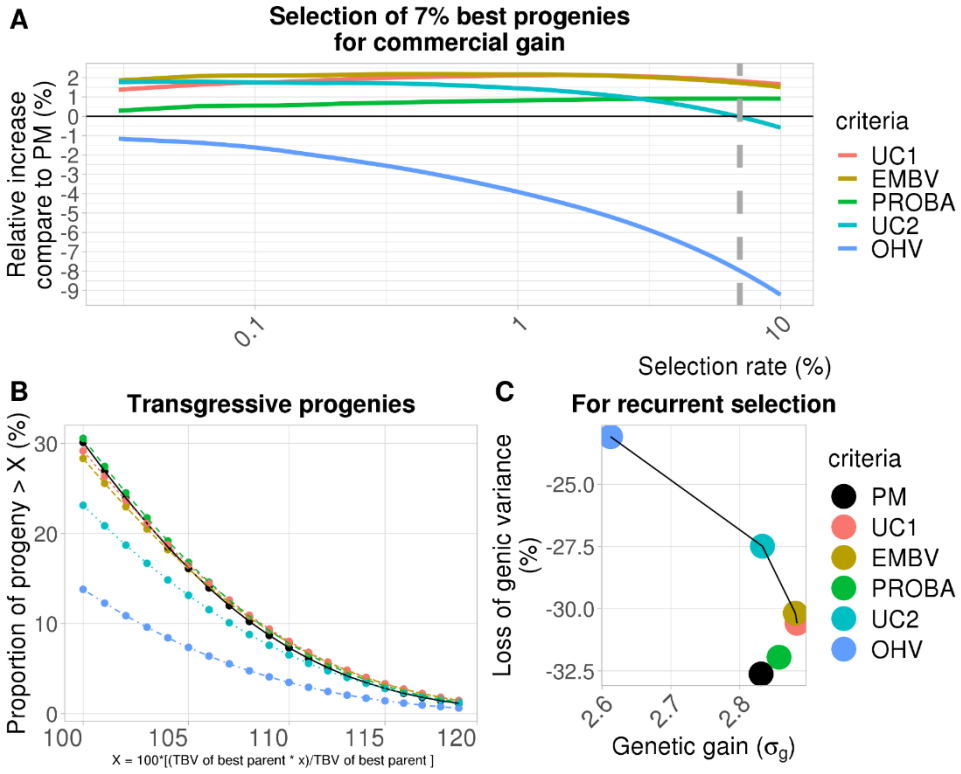
Supplementary Figure S6: Genetic similarity of selected parents with varying number of progenies in unselected populations + TRUE + CONSTRAINTS scenarios

Genetic similarities were computed as in supplementary Figure S3. ***A****: Parents with more than Dmin progeny were kept.* ***B****: Parents with Dmin progeny were kept.*

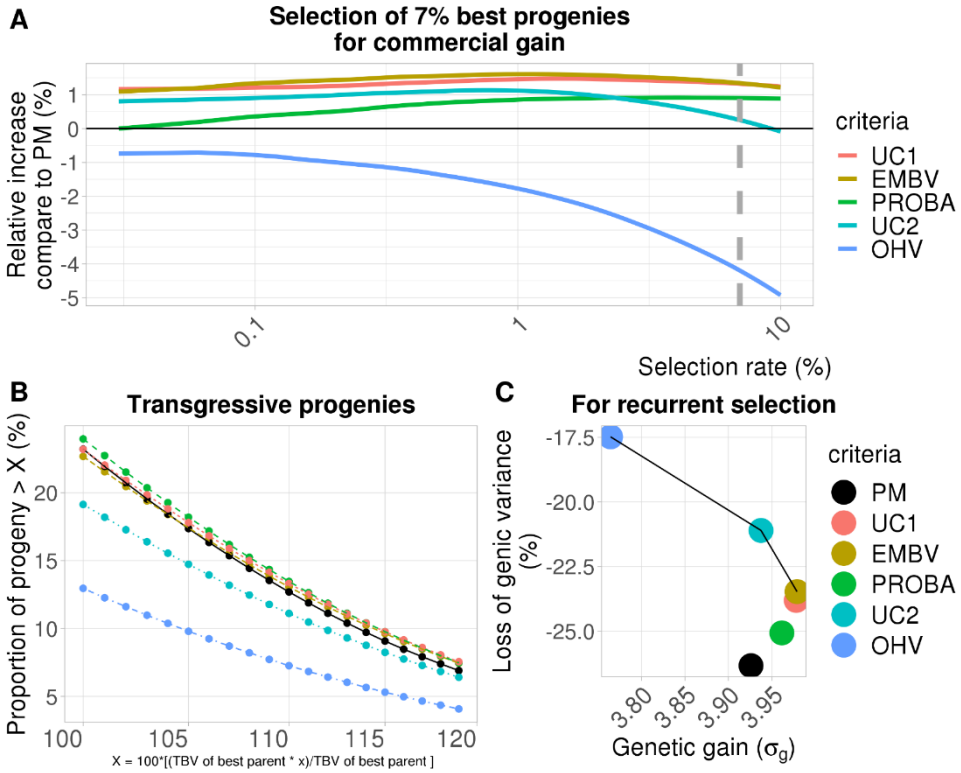


Supplementary Figure S7: Benefits of CSC in unselected populations + TRUE + CONSTRAINTS scenarios

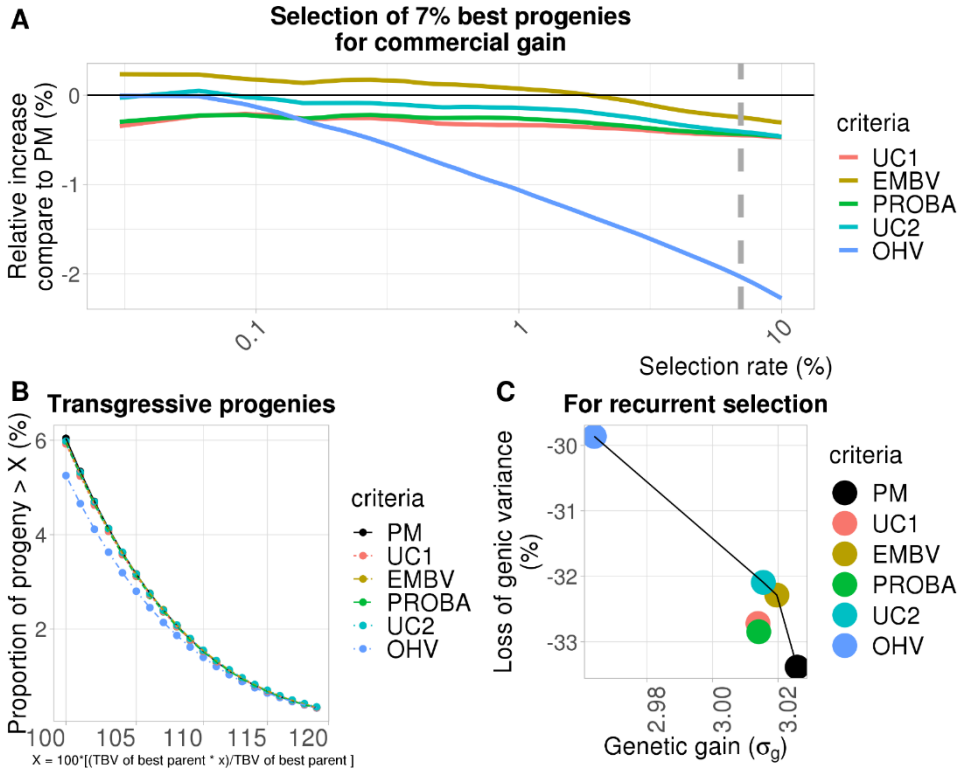
The results are presented for the “selected populations”, in TRUE + CONSTRAINTS scenarios (QTLs effects are known and parental contributions are constrained). **A. Relative increase in the mean TBV of selected progeny compare to PM criterion** The y-axis gives the relative increase in mean TBV of selected progeny for each CSC compare to PM criterion for a selection rate < 10%. The grey dashed line indicates a selection rate of 7%, which is the selected fraction of progeny supposed to form the new breeding population at the next cycle. **B. Proportion of transgressive progenies.** By transgressive, we mean superior to the best parental line of the breeding population. The proportion is reported for varying level of the value of the best parental line (from 100% of the value of the best parental line to 120%). **C. Genetic gain and loss of genic variance associated with each criterion in the 7% best progeny.** The 7% best progeny (grey dashed line in graph A) are considered as the new parents for the next selection cycle. Loss of genic variance is computed as the difference between genic variance of the selected progeny and genic variance of the former breeding population (n=900 parents in “selected populations” and n=835 parents in “unselected populations”), divided by genic variance of the former breeding population and multiply by 100. Genetic gain is computed as the difference between average TBV of selected progeny and average TBV of the former breeding population, divided by the standard deviation of TBV in the initial breeding population, multiply. The black line joining points show the criteria associated with the best trade-off between genetic gain and genic variance.



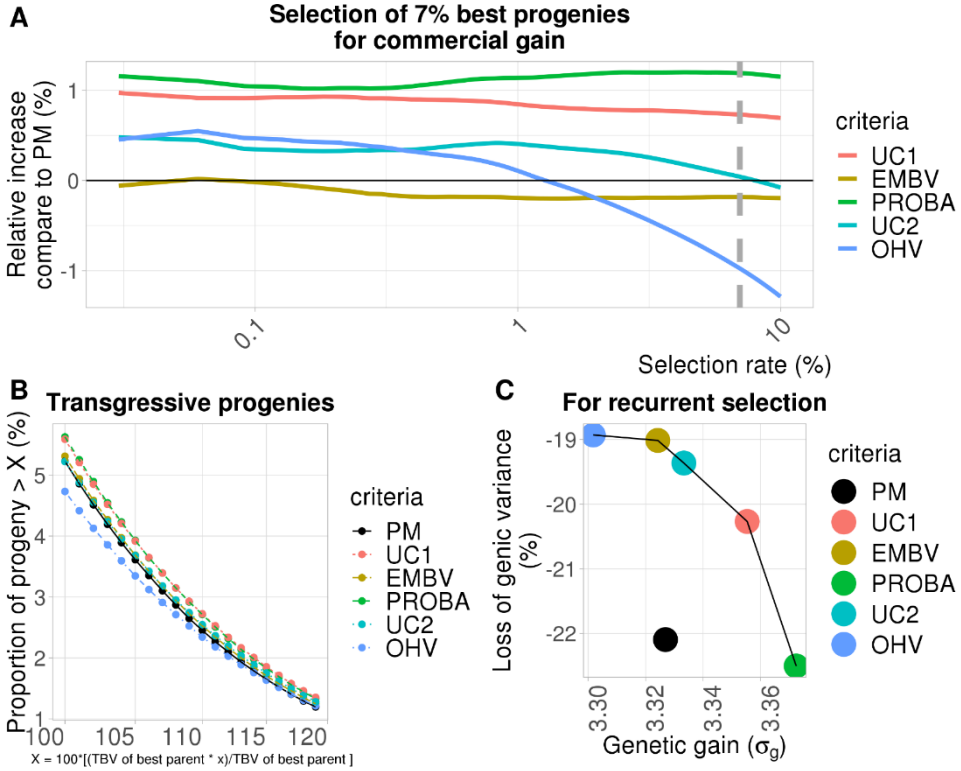
Supplementary Figure S8: Benefits of CSC in selected populations + TRUE + NO CONSTRAINTS



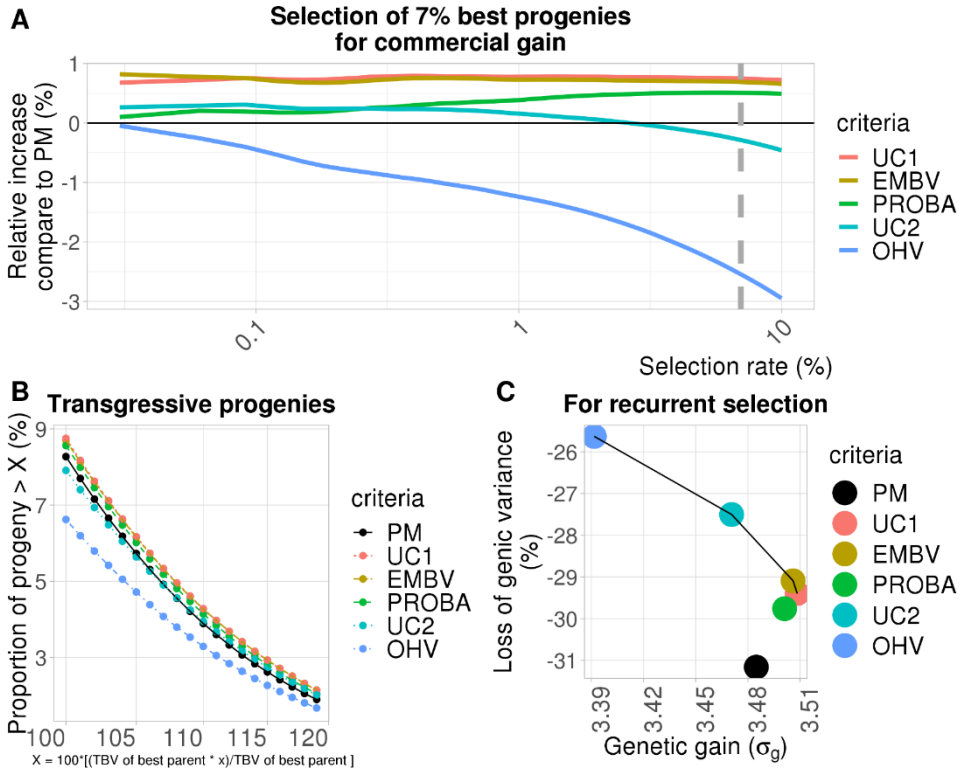
Supplementary Figure S9: Benefits of CSC in unselected populations + TRUE + NO CONSTRAINTS



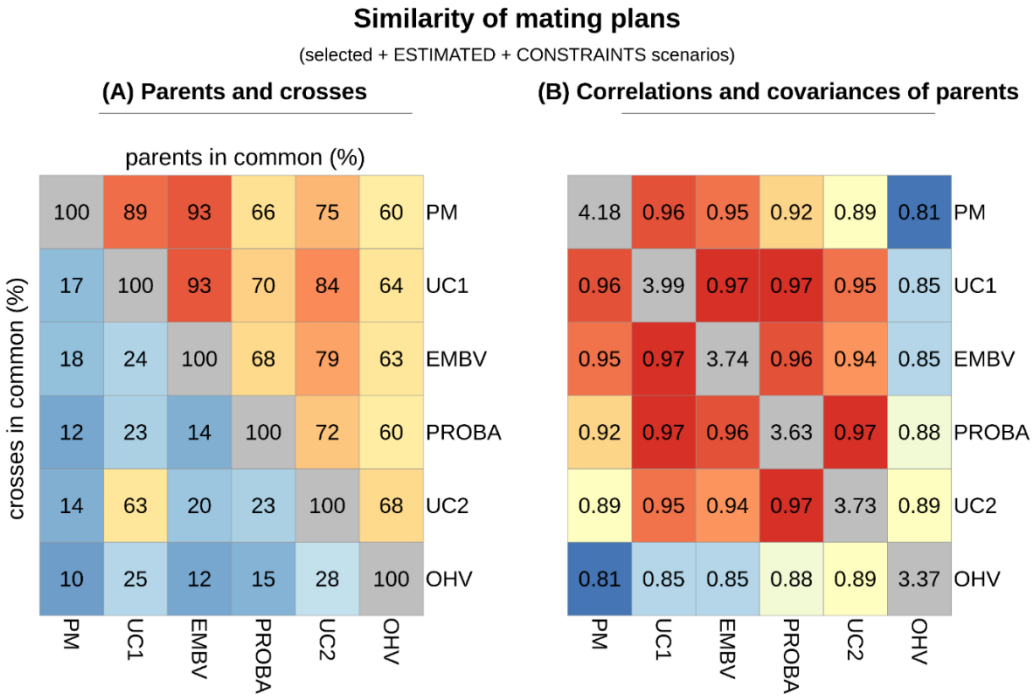
Supplementary Figure S10: Benefits of CSC in selected populations + ESTIMATED + NO CONSTRAINTS



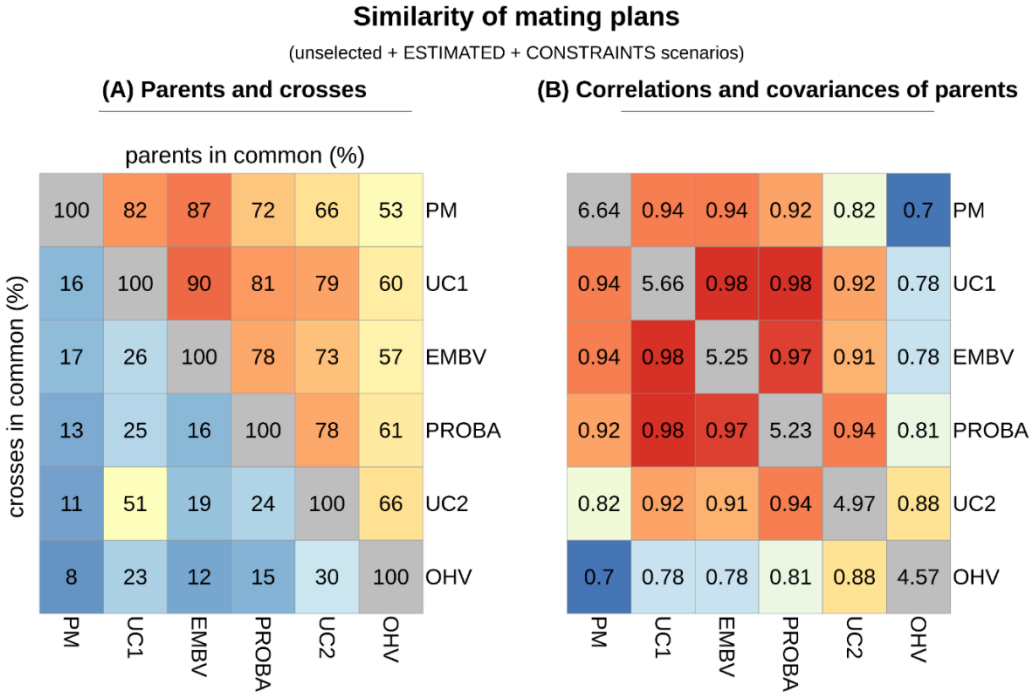
Supplementary Figure S11: Benefits of CSC in unselected populations + ESTIMATED + CONSTRAINTS



Supplementary Figure S12: Benefits of CSC in unselected + ESTIMATED + NO CONSTRAINTS

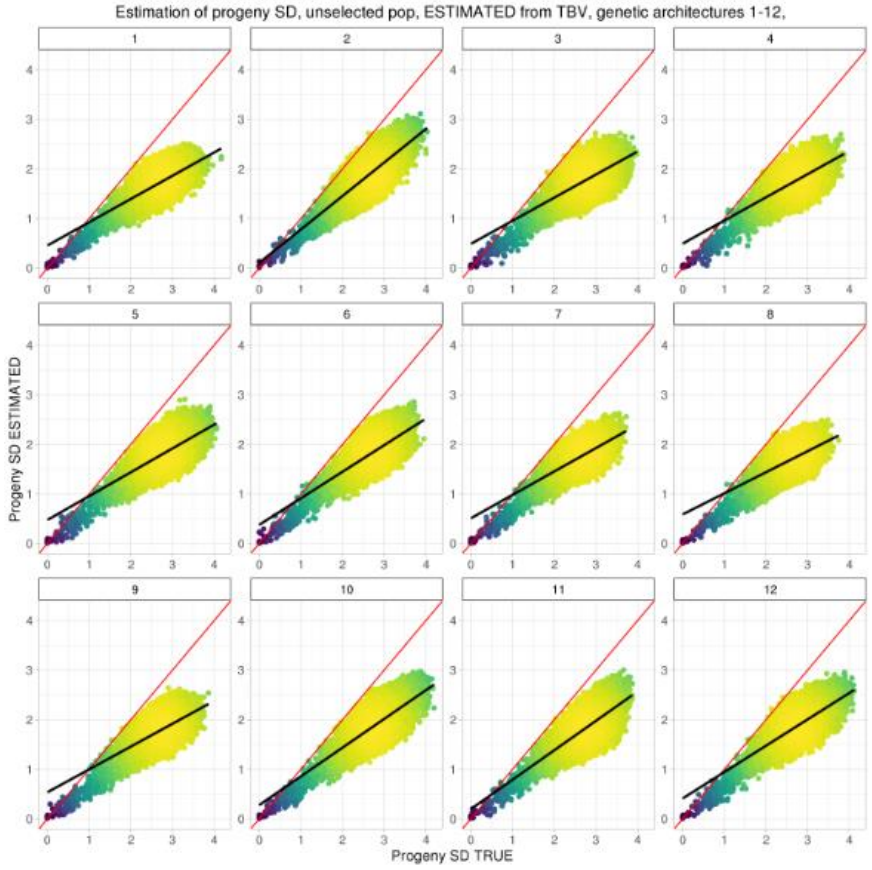


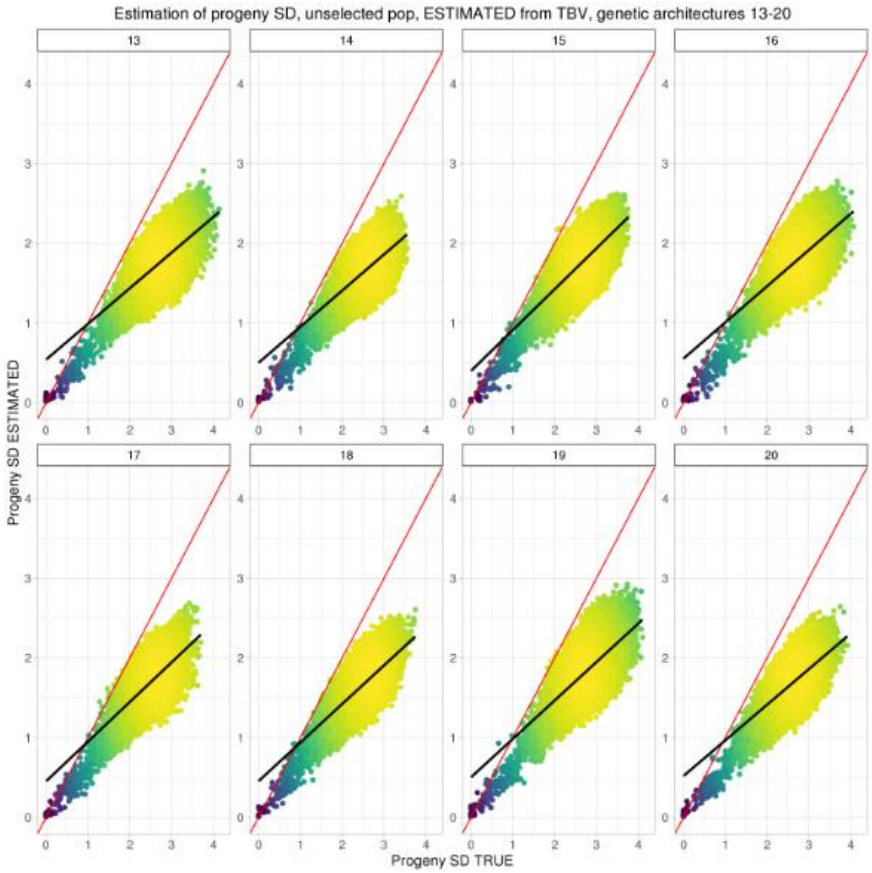
Supplementary Figure S13: Similarity of mating plan in selected populations + ESTIMATED

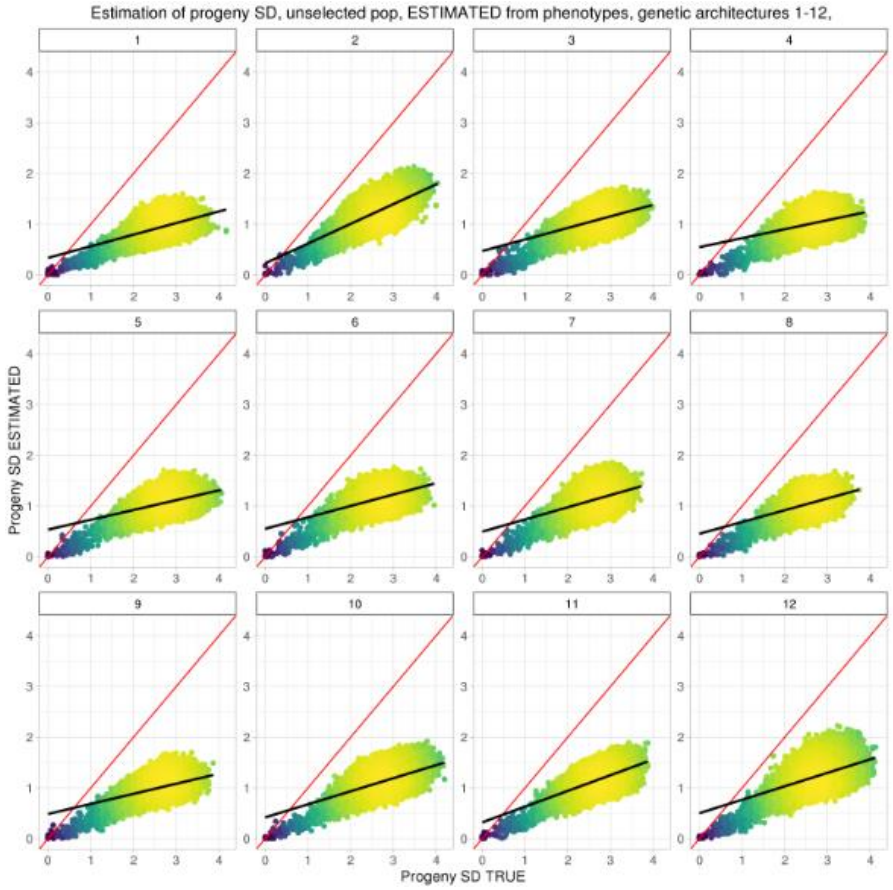


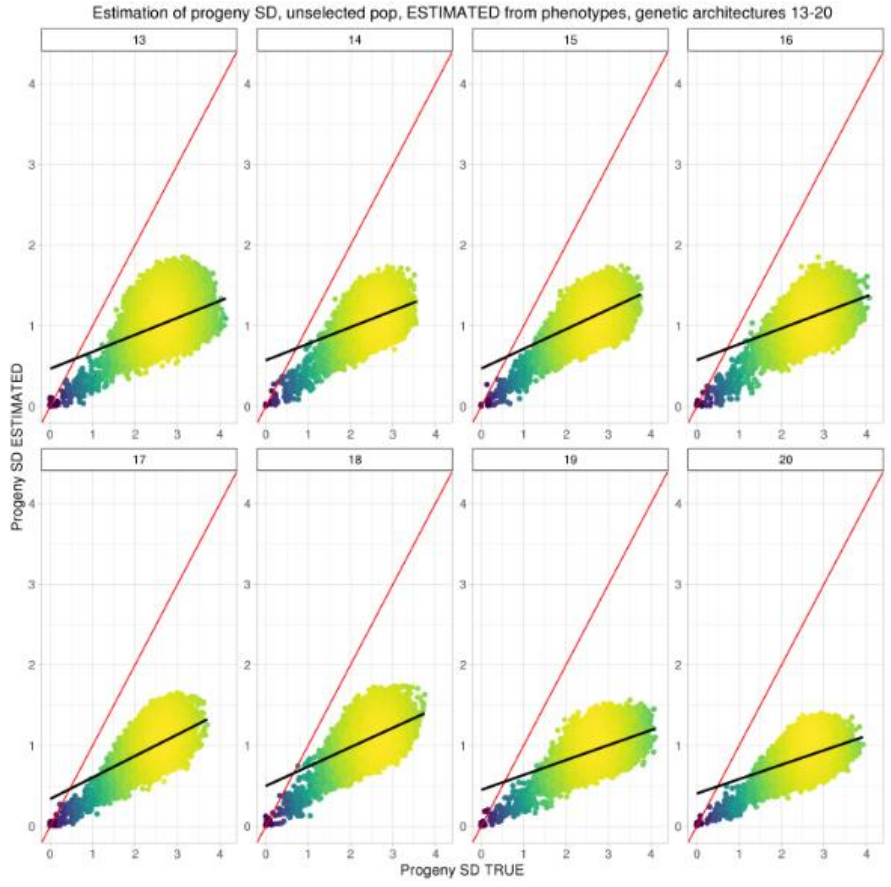
Supplementary Figure S14: Similarity of mating plan in unselected populations + ESTIMATED

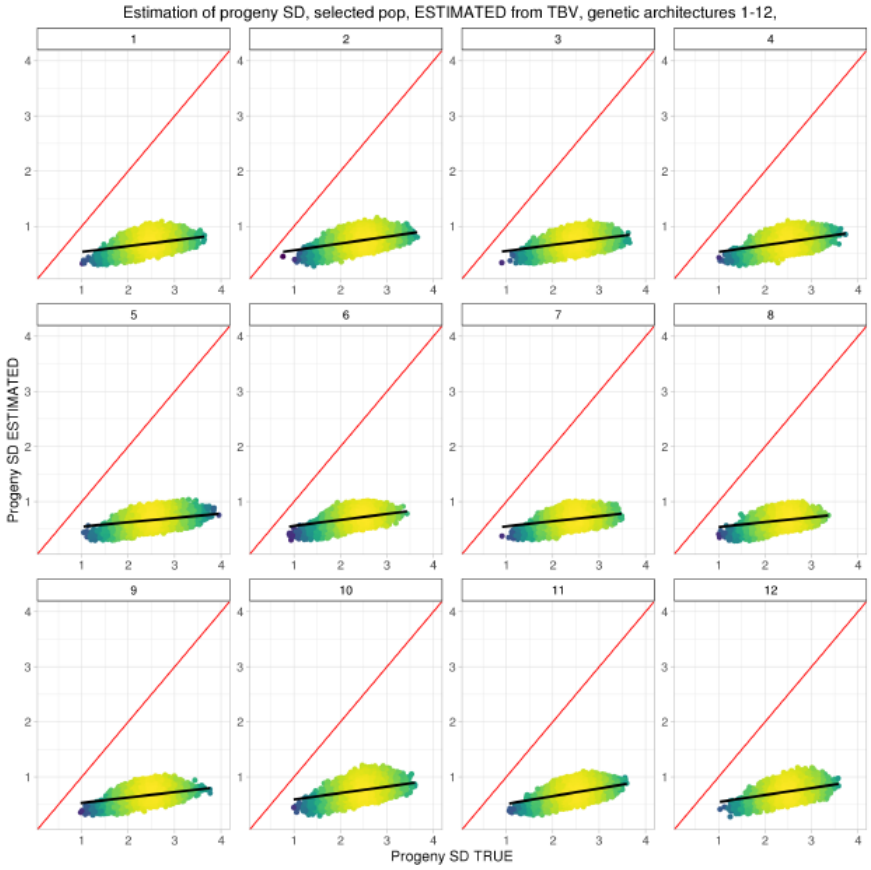
V.2.3 Supplementary File

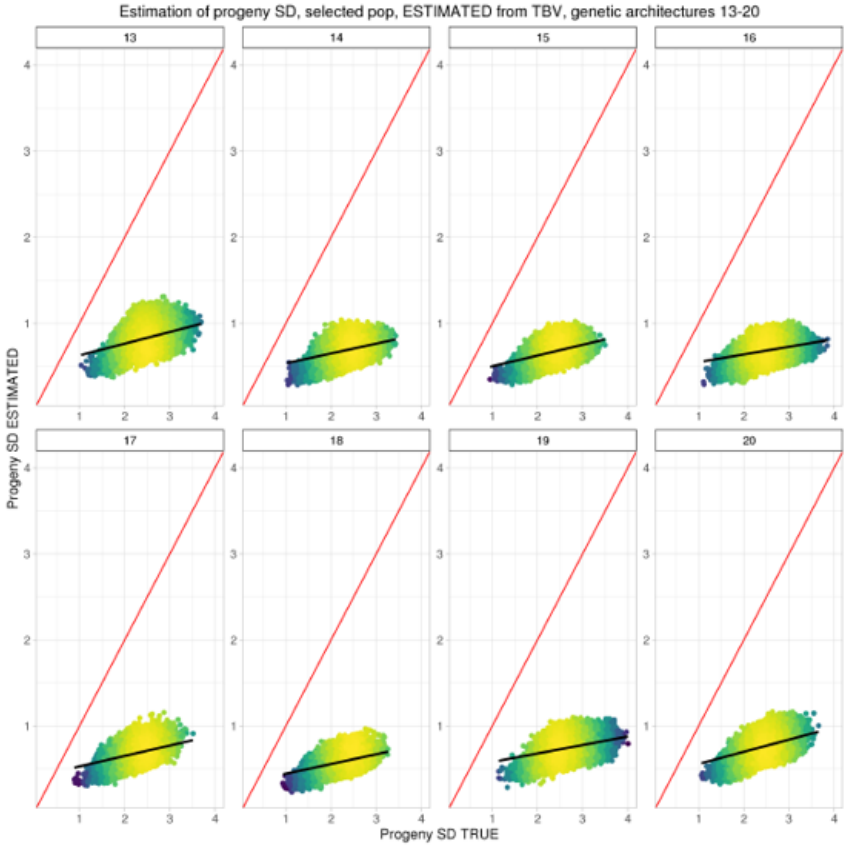


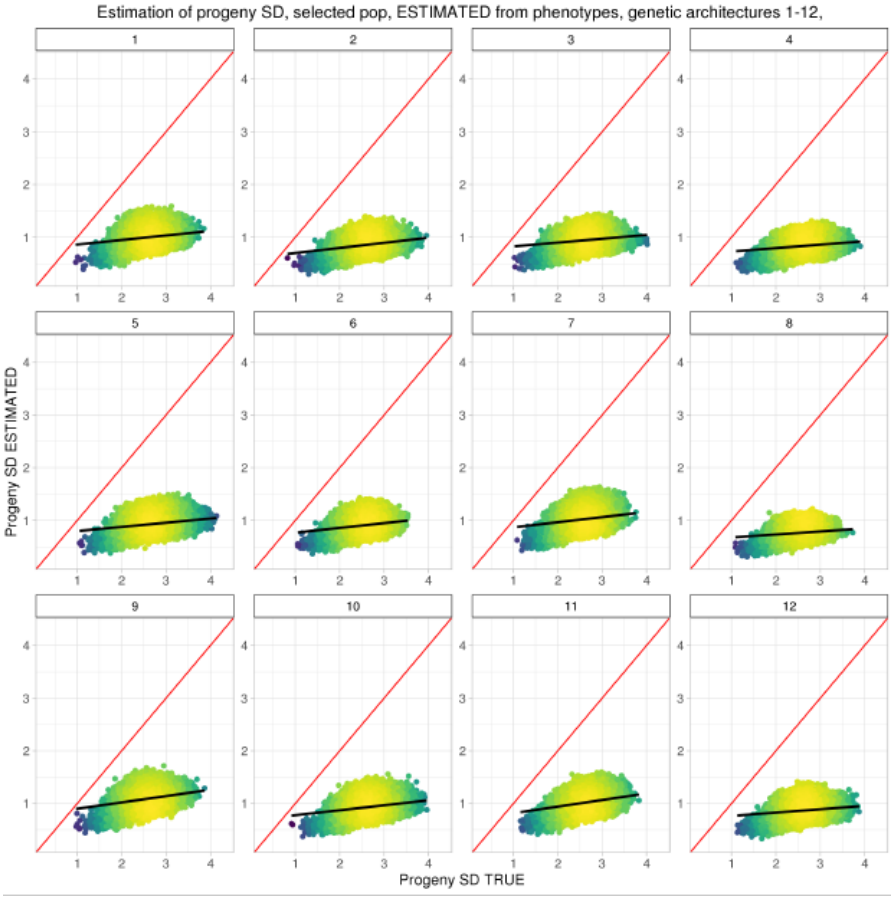


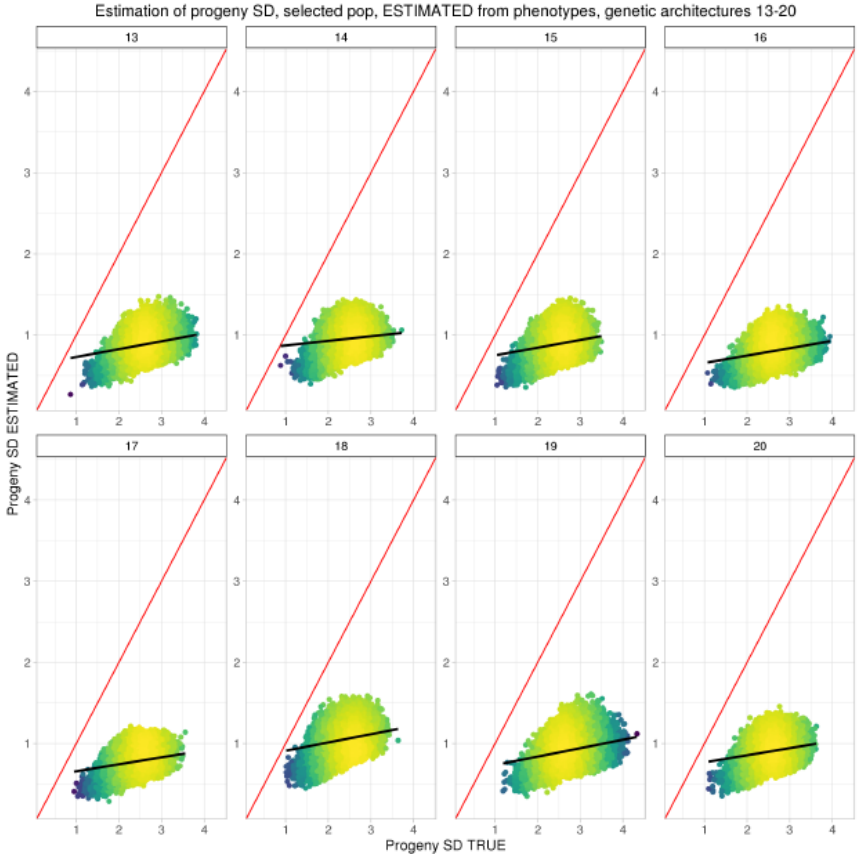

Estimation of progeny SD, unselected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD ESTIMATED
Progeny SD TRUE

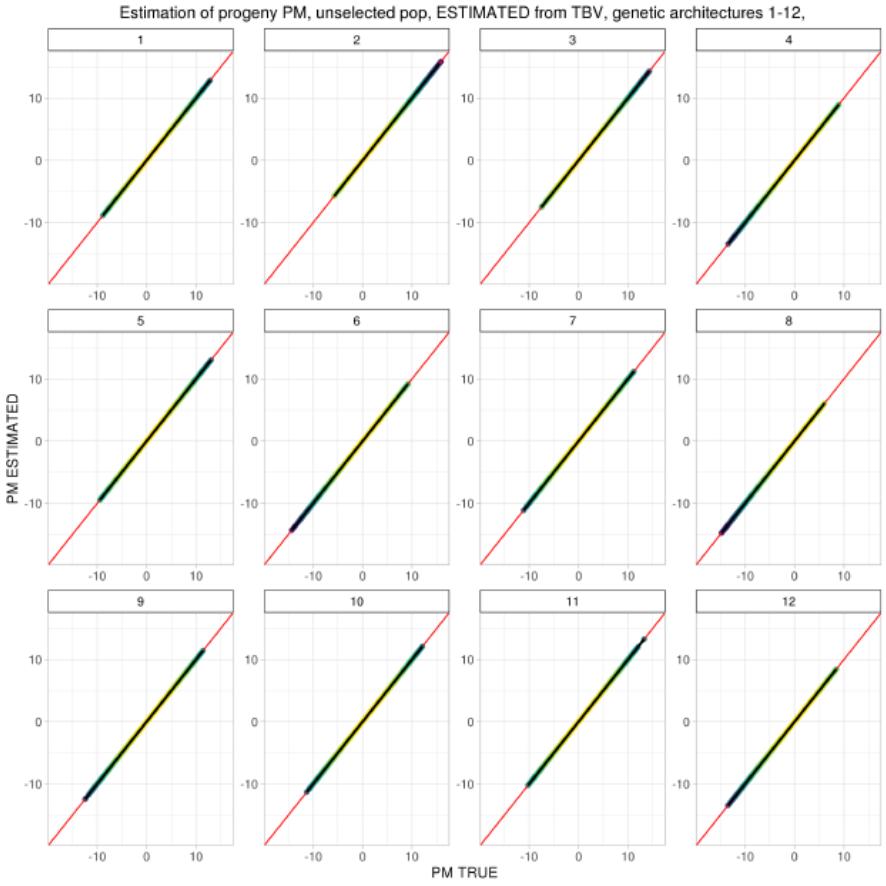

Estimation of progeny SD, unselected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD ESTIMATED
Progeny SD TRUE

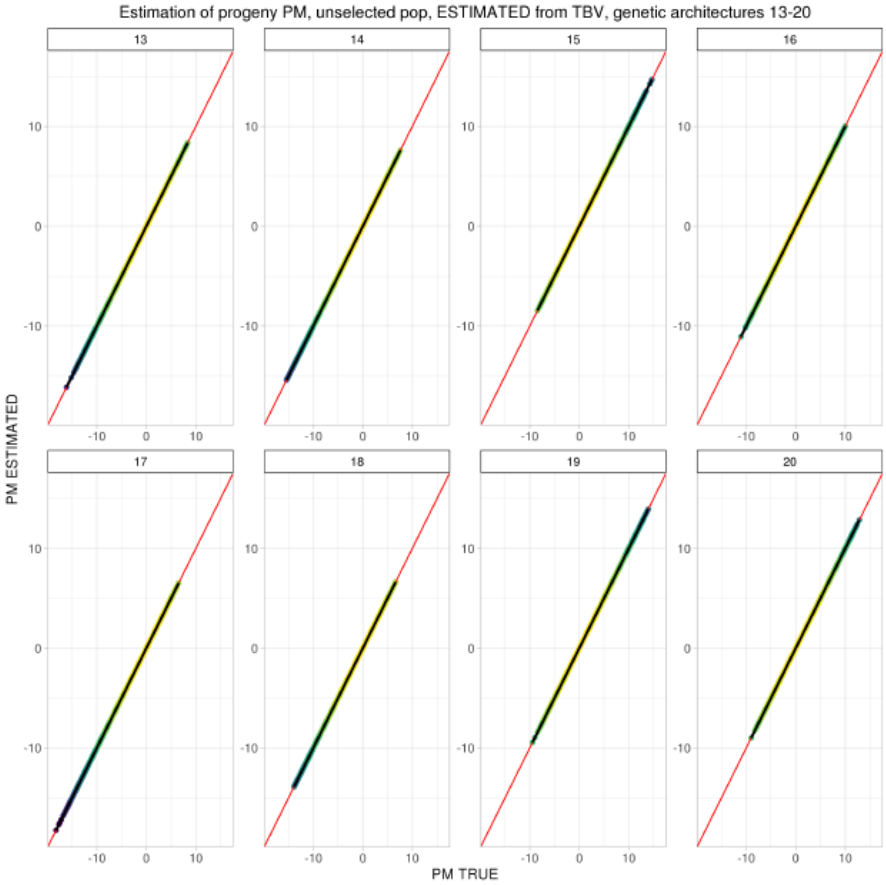

Estimation of progeny SD, selected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD ESTIMATED
Progeny SD TRUE

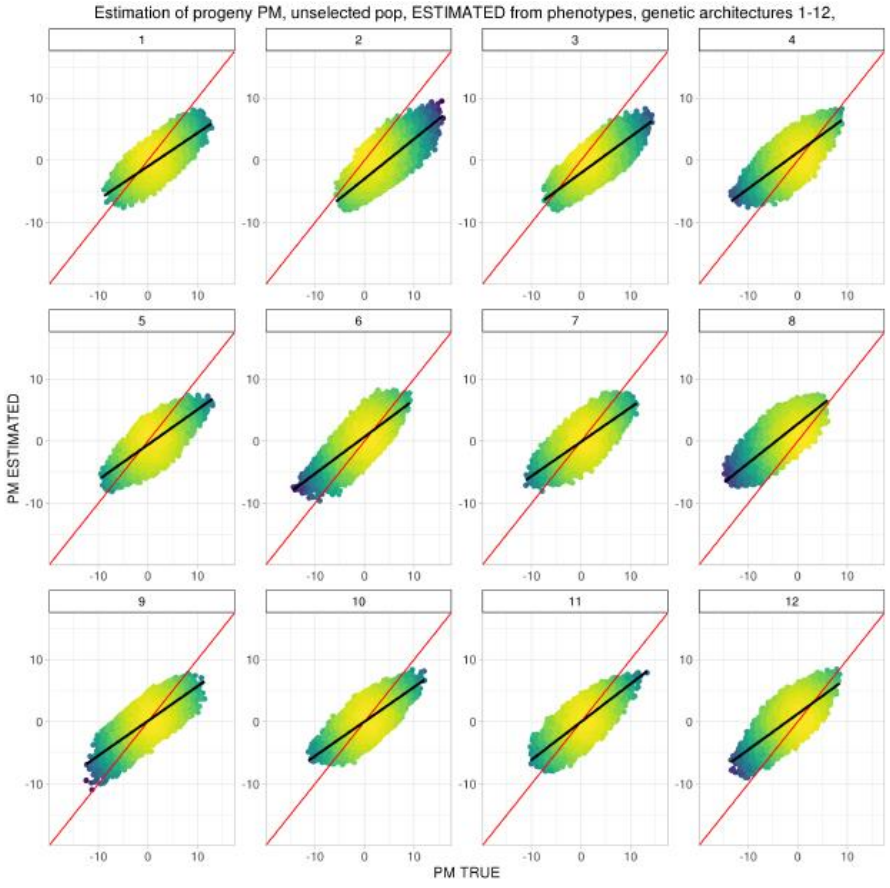

Estimation of progeny SD, selected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD ESTIMATED
Progeny SD TRUE

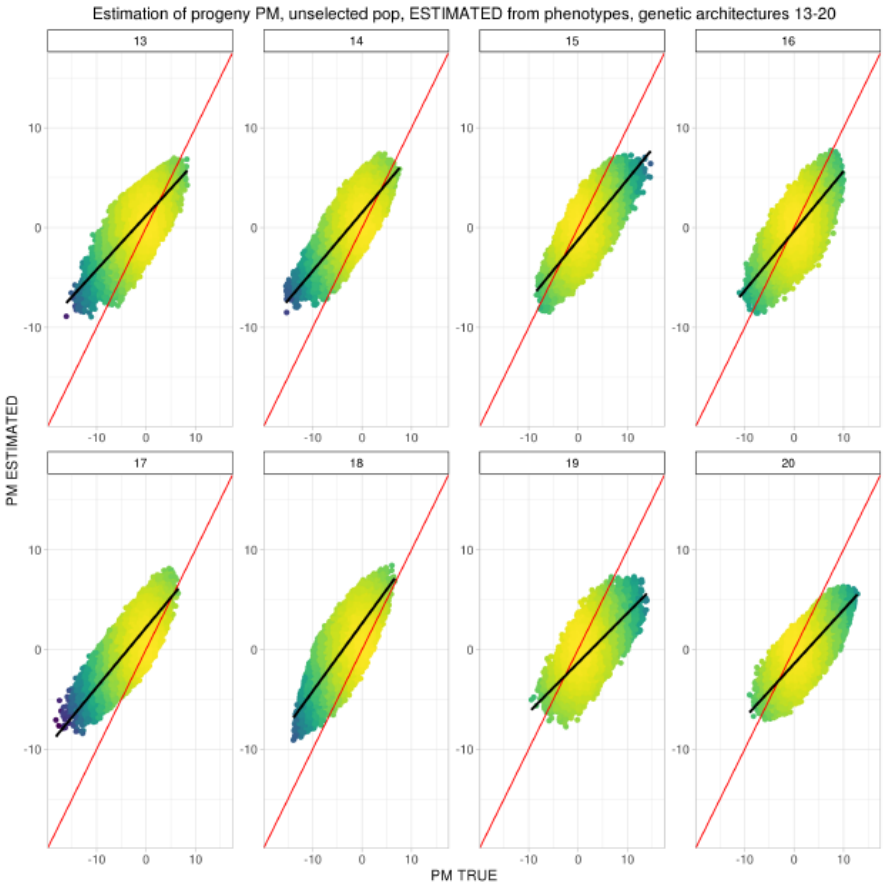

Estimation of progeny SD, selected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD ESTIMATED
Progeny SD TRUE

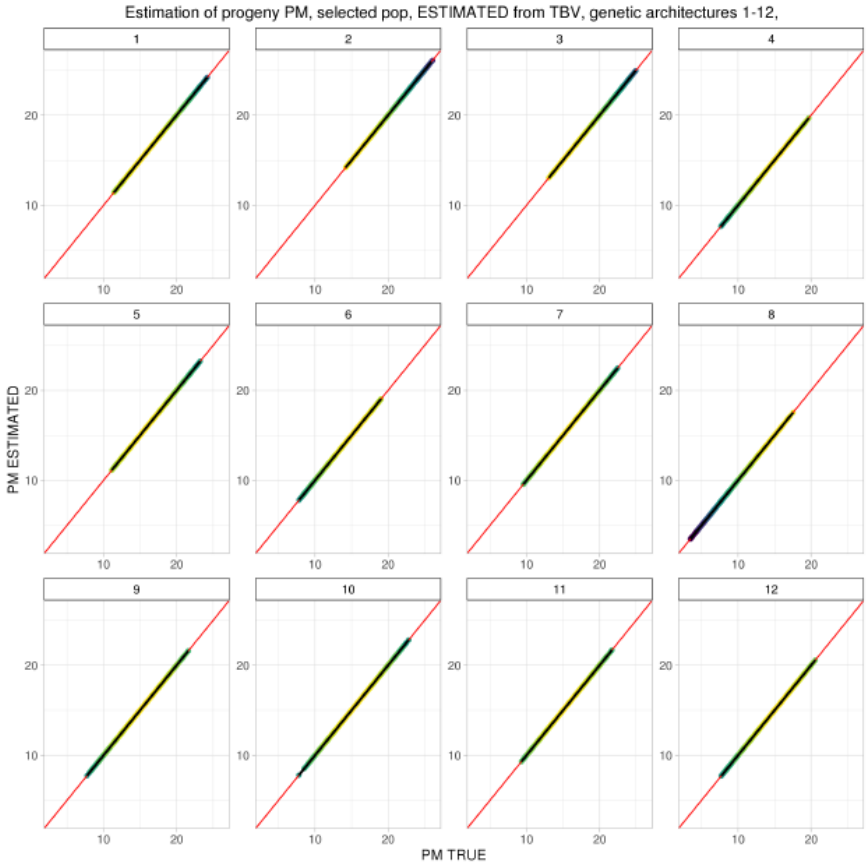

Estimation of progeny SD, selected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD ESTIMATED
Progeny SD TRUE

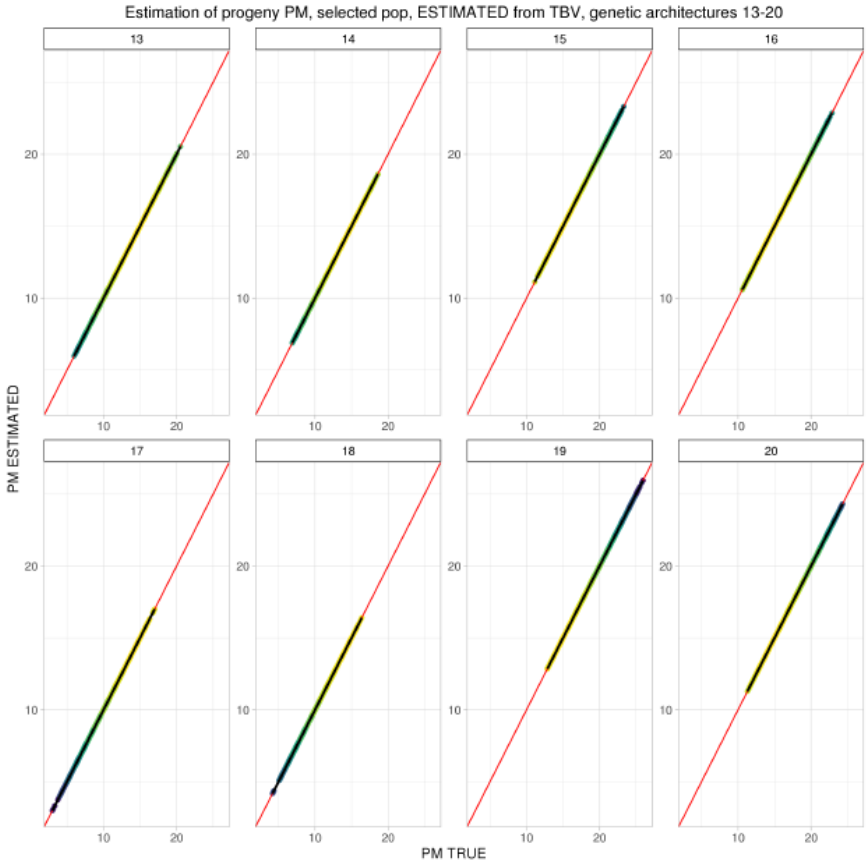

Estimation of progeny PM, unselected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PM ESTIMATED
PM TRUE

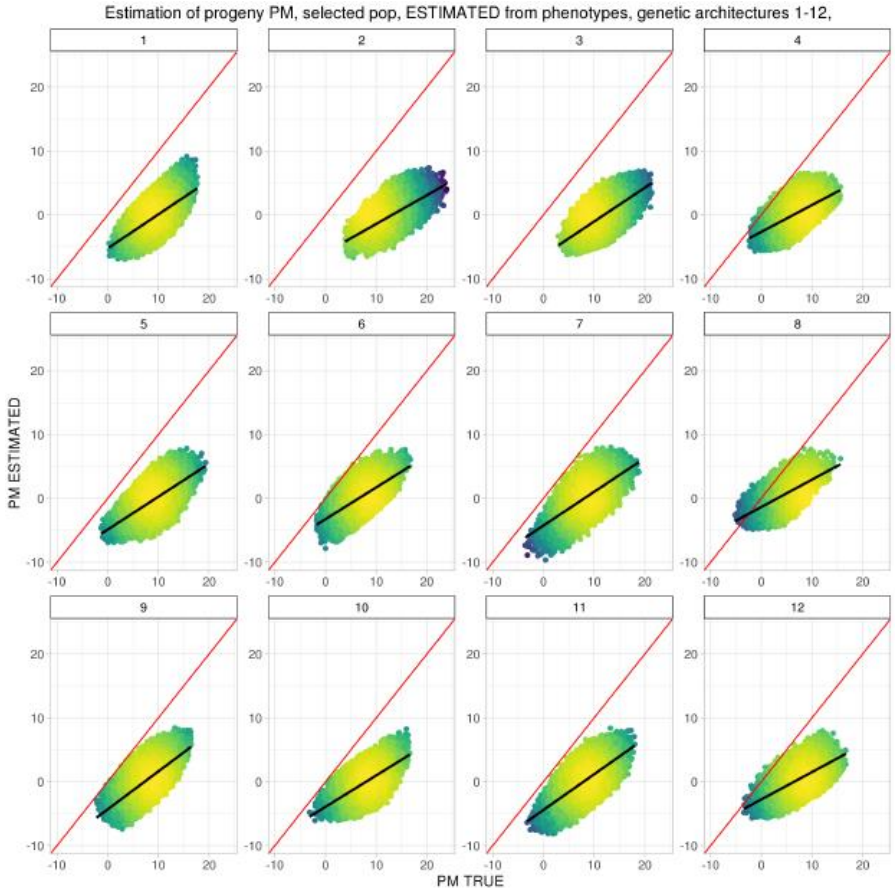

Estimation of progeny PM, unselected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
PM ESTIMATED
PM TRUE

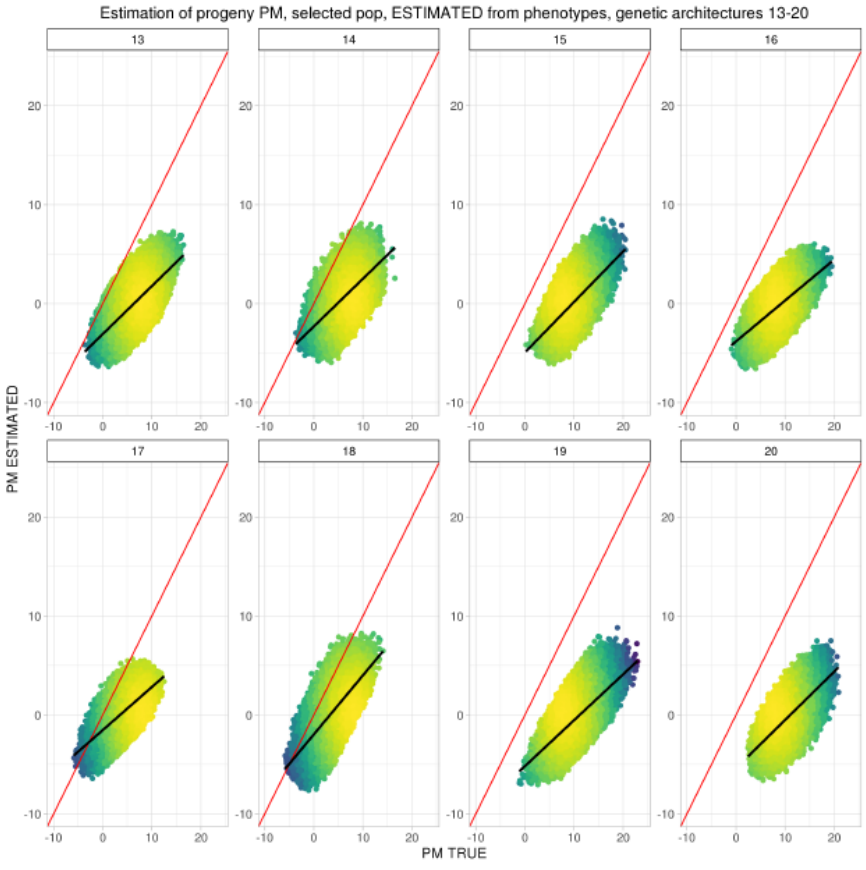

Estimation of progeny PM, unselected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PM ESTIMATED
PM TRUE

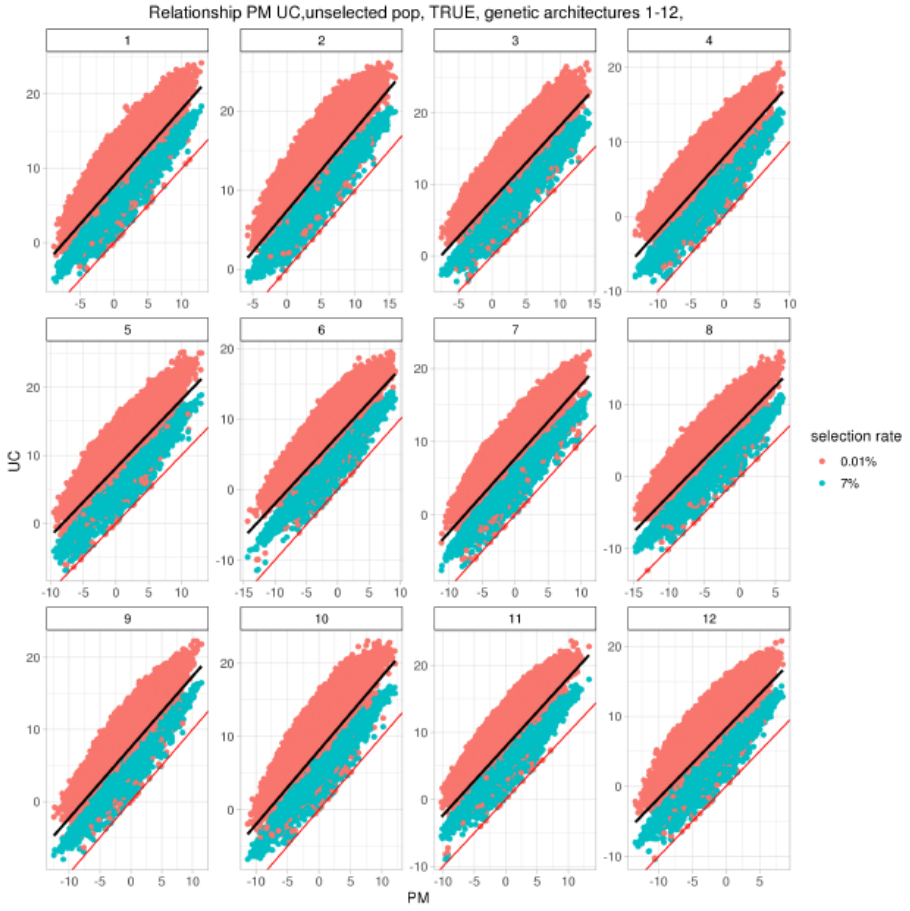

Estimation of progeny PM, unselected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
PM ESTIMATED
PM TRUE

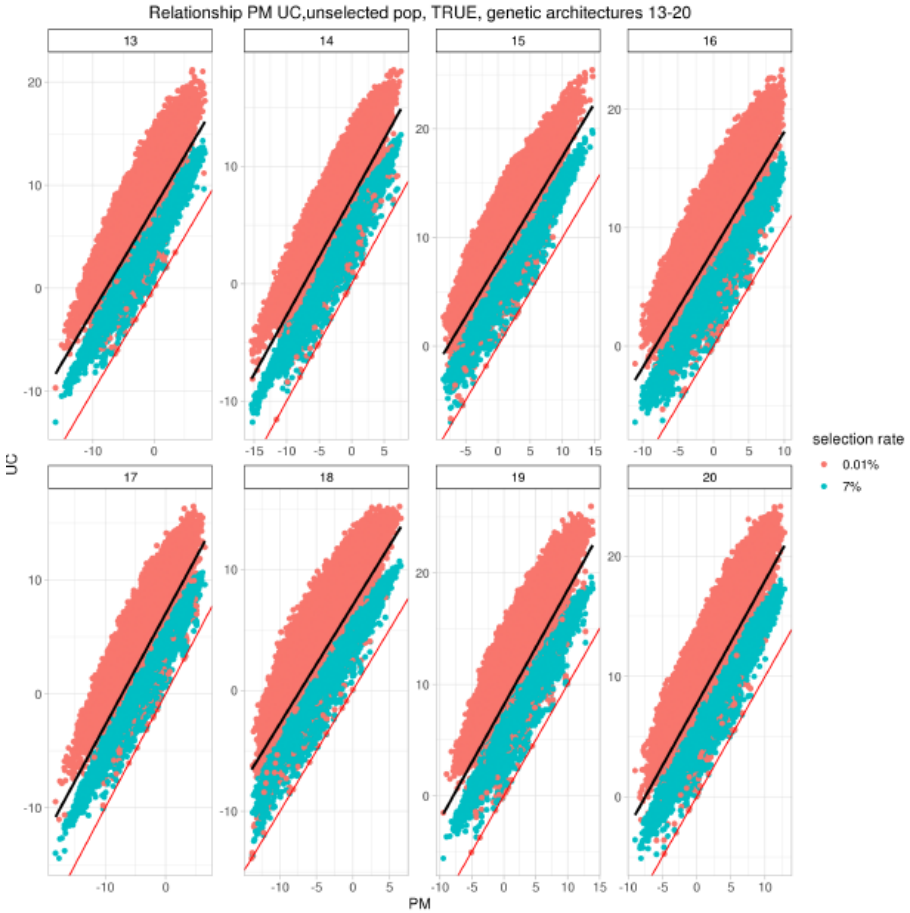

Estimation of progeny PM, selected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PM ESTIMATED
PM TRUE

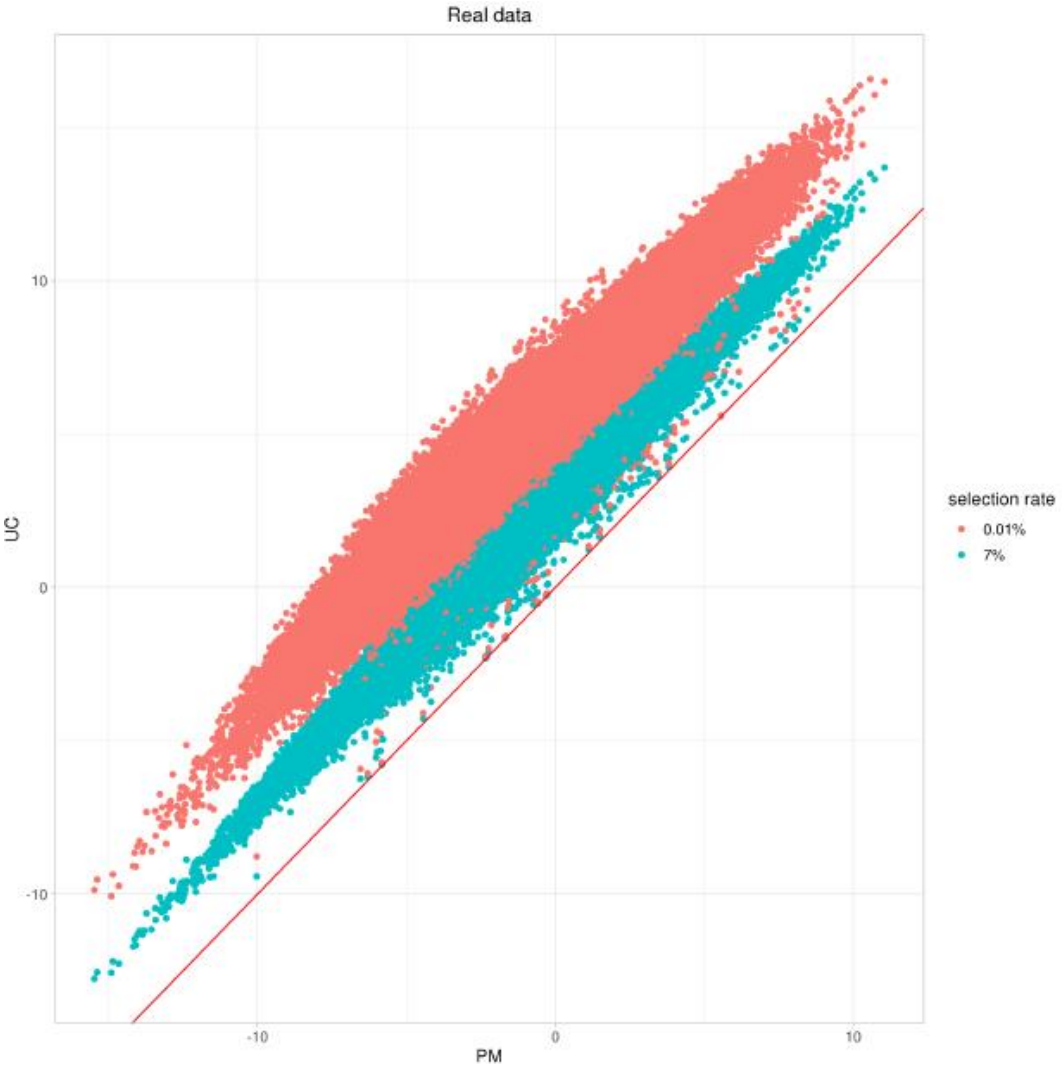

Estimation of progeny PM, selected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
PM ESTIMATED
PM TRUE

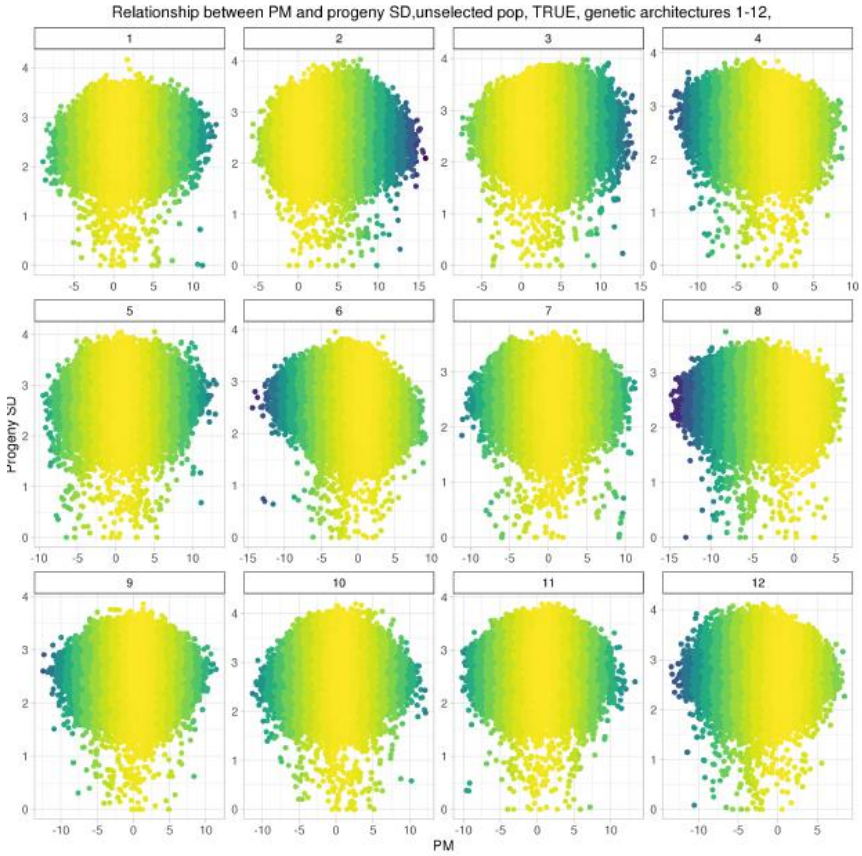

Estimation of progeny PM, selected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PM ESTIMATED
PM TRUE

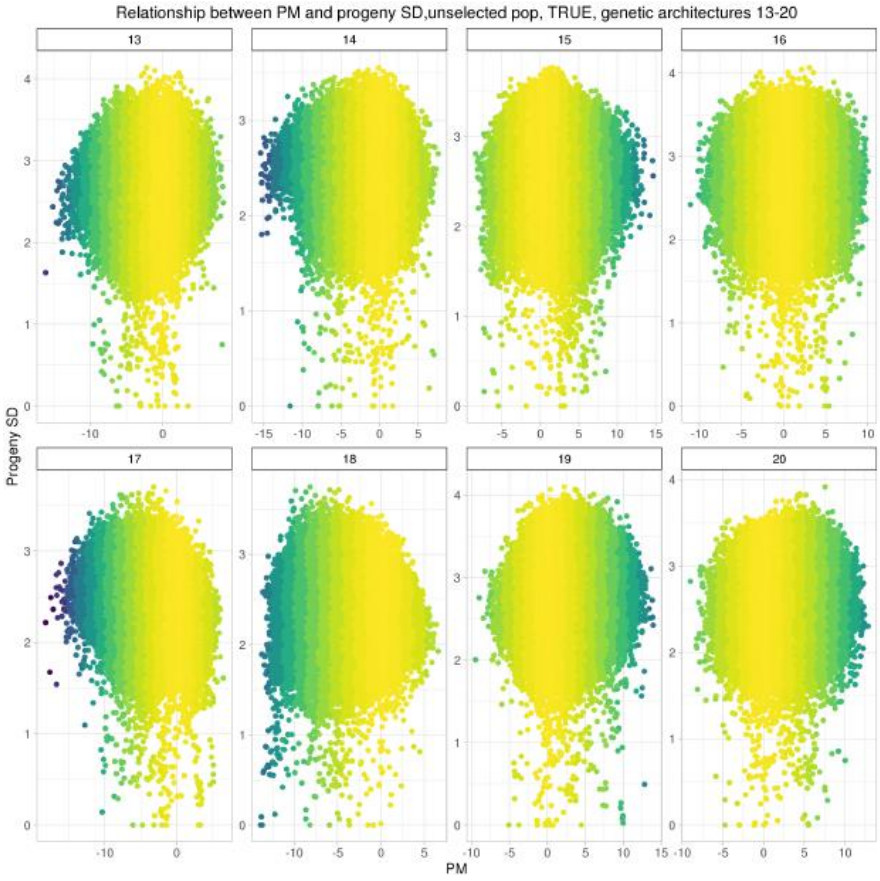

Estimation of progeny PM, selected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
PM ESTIMATED
PM TRUE

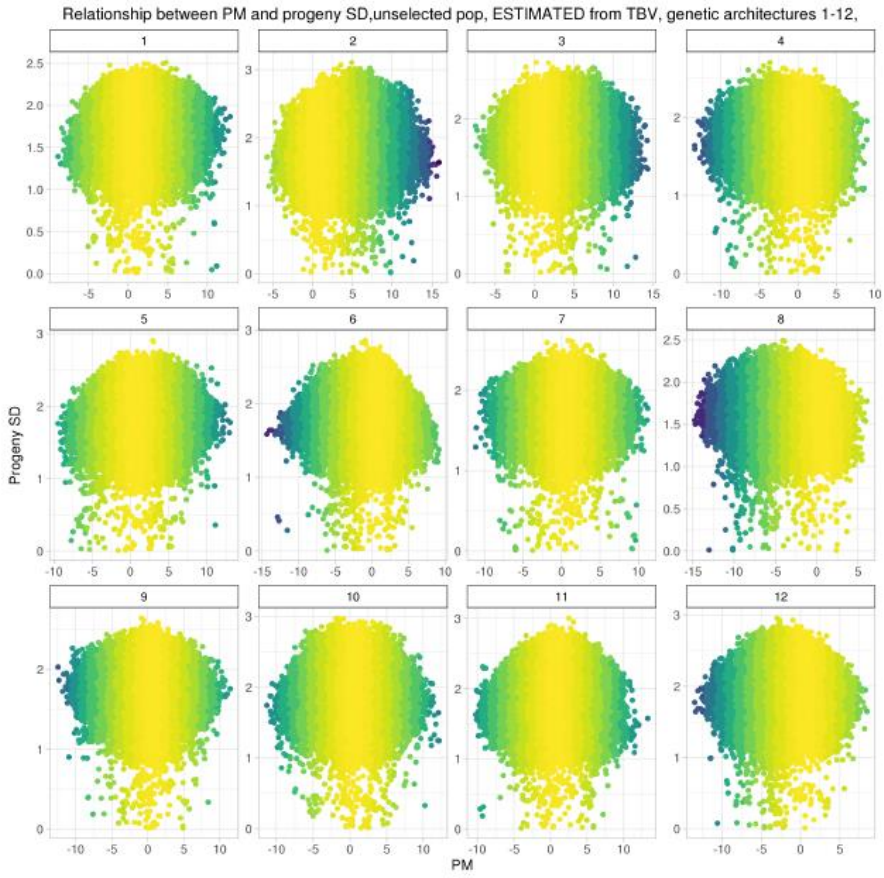

Relationship PM UC,unselected pop, TRUE, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
UC
PM
selection rate
0.01%
7%

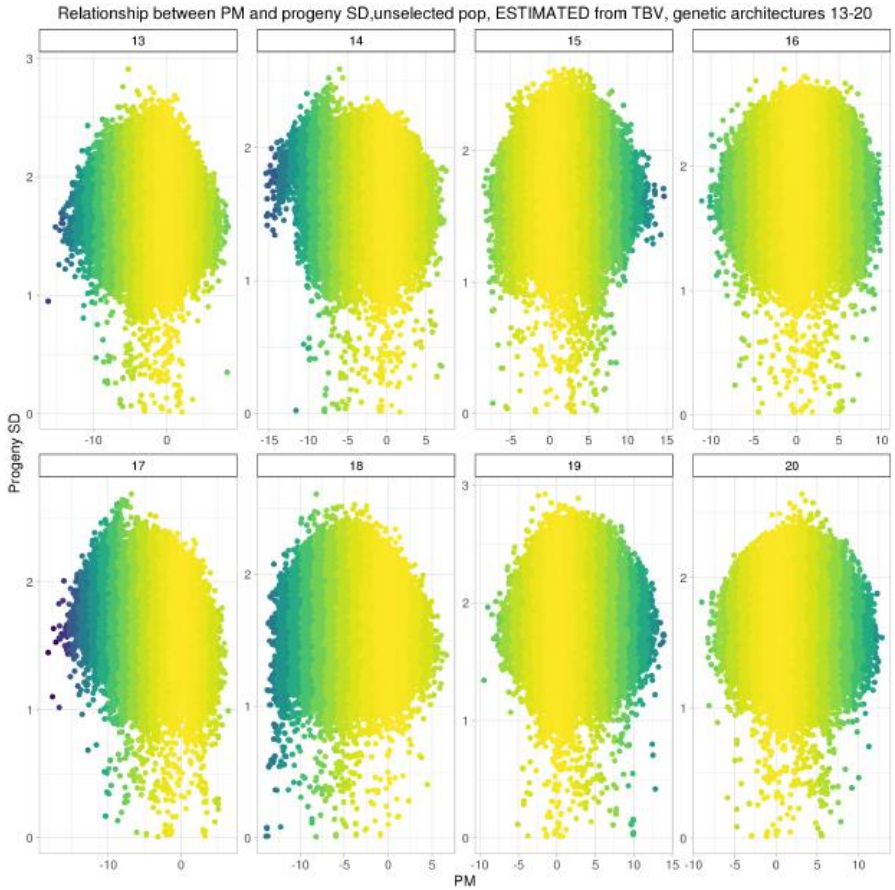

Relationship PM UC,unselected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
UC
PM
selection rate
0.01%
7%

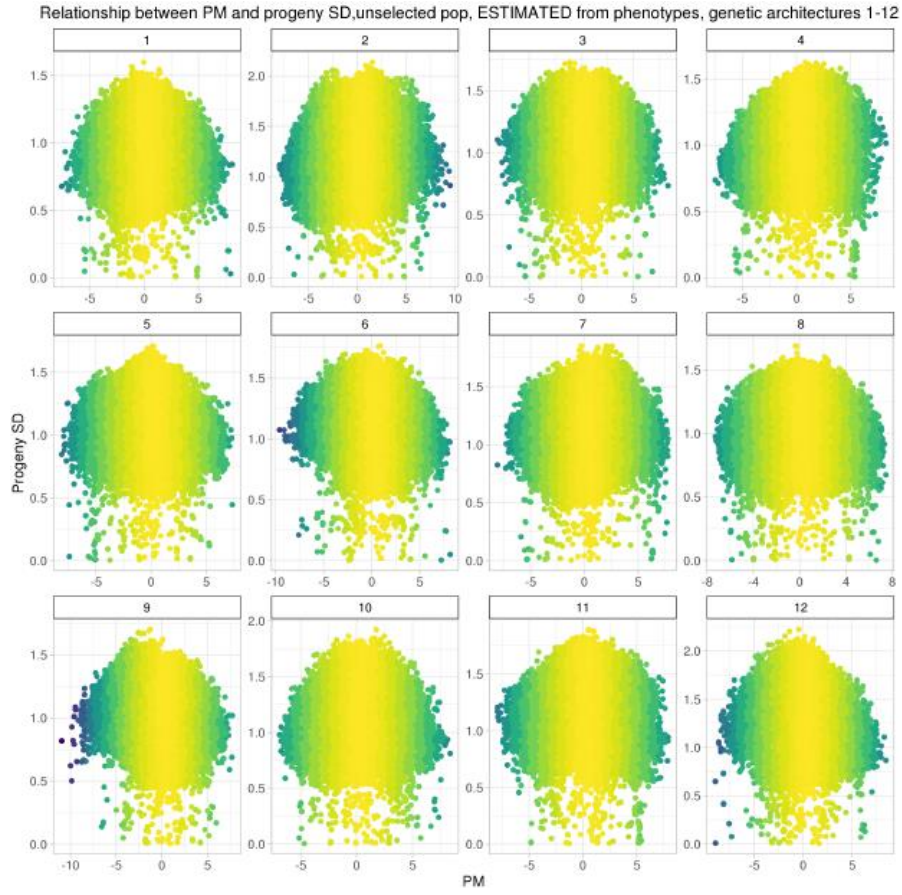

Real data
UC
PM
10
0
-10
-10
0
10
selection rate
0.01%
7%

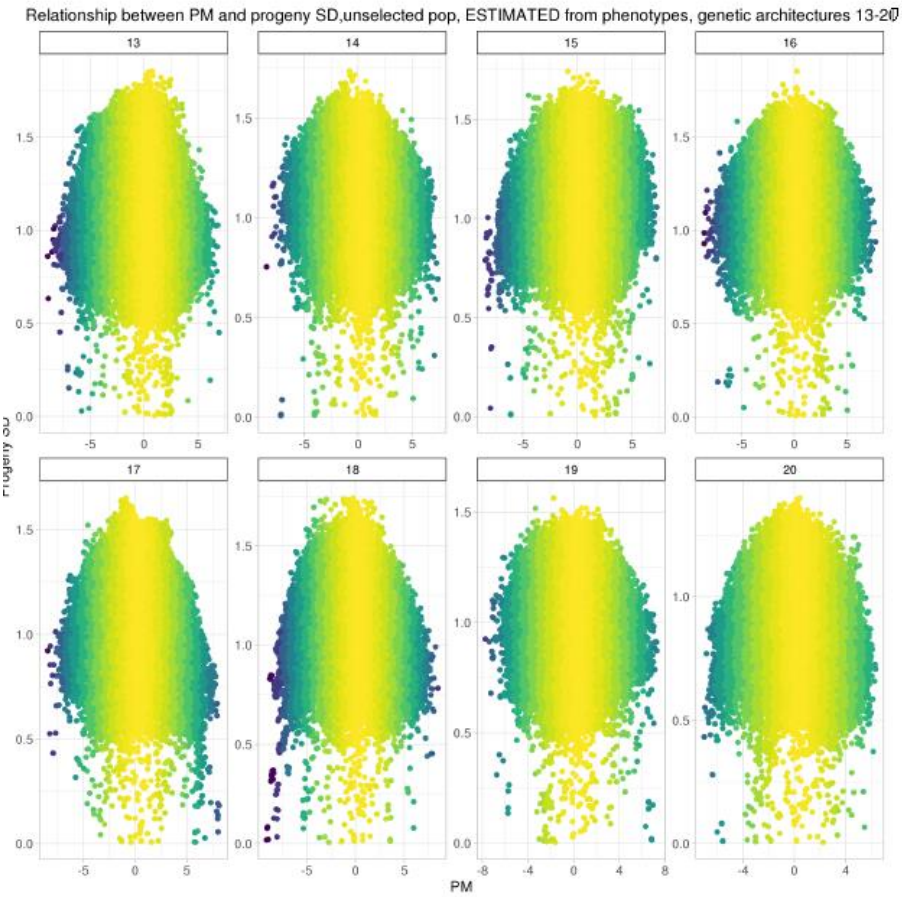

Relationship between PM and progeny SD,unselected pop, TRUE, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD
PM

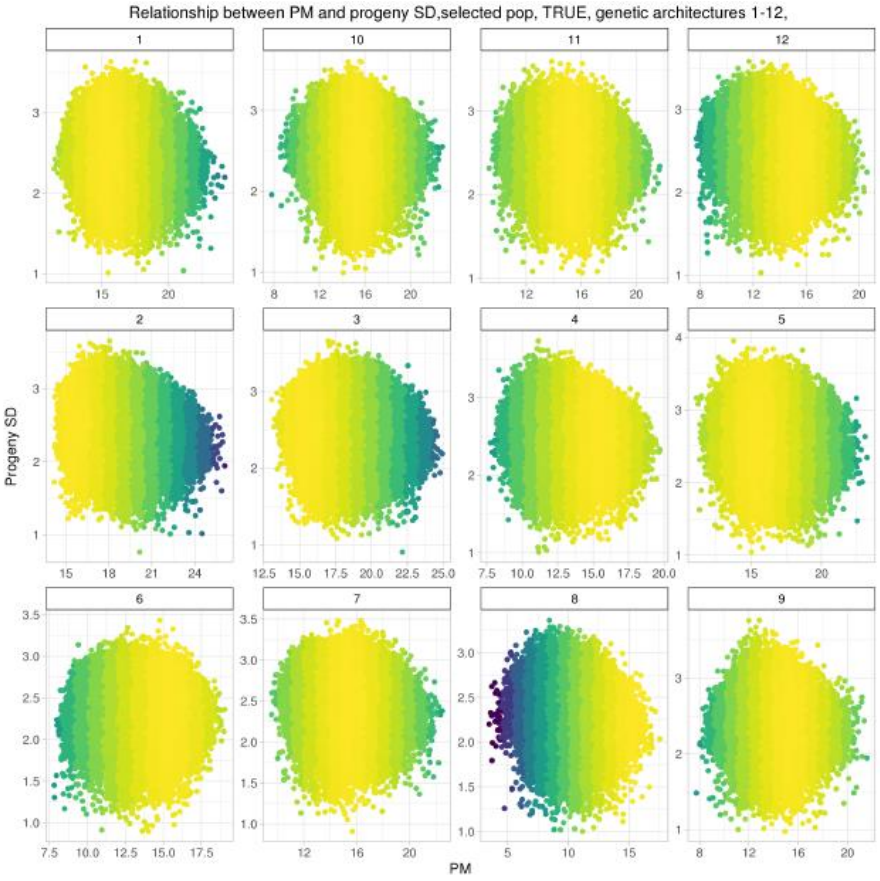

Relationship between PM and progeny SD,unselected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD
PM

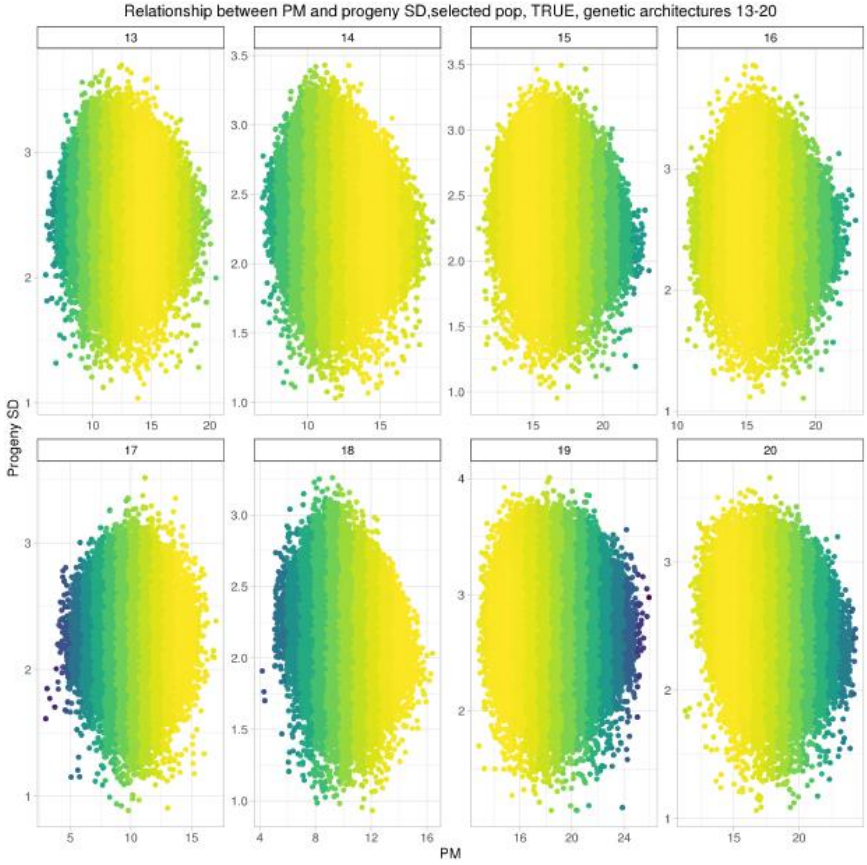

Relationship between PM and progeny SD,unselected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD
PM

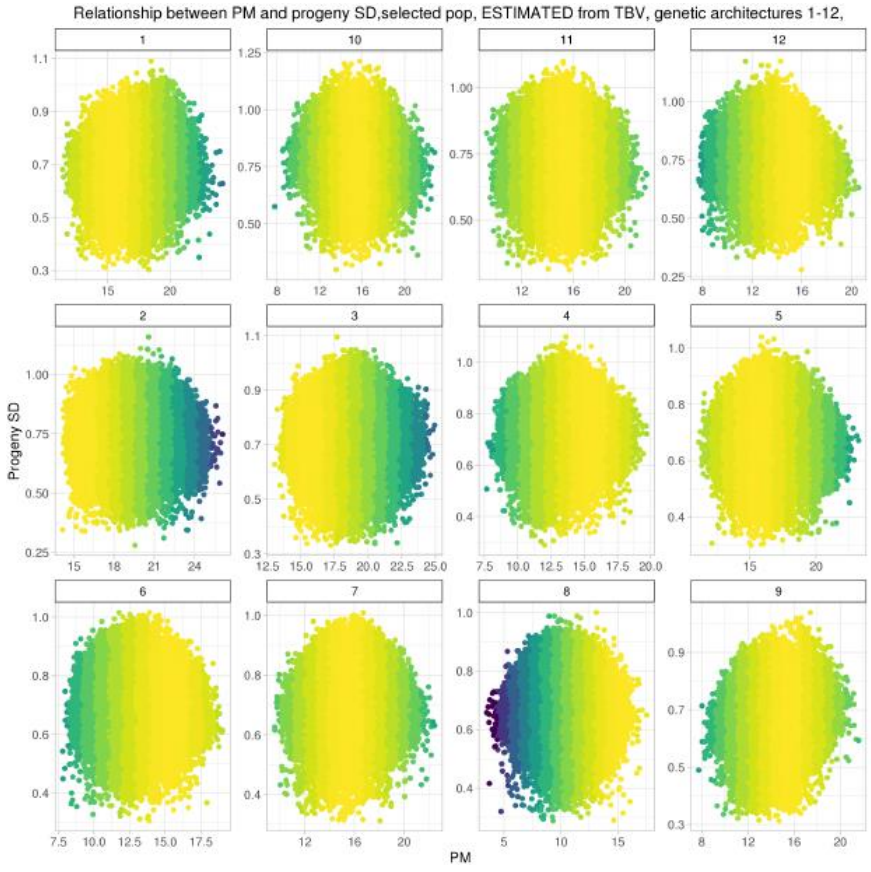

Relationship between PM and progeny SD,unselected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD
PM

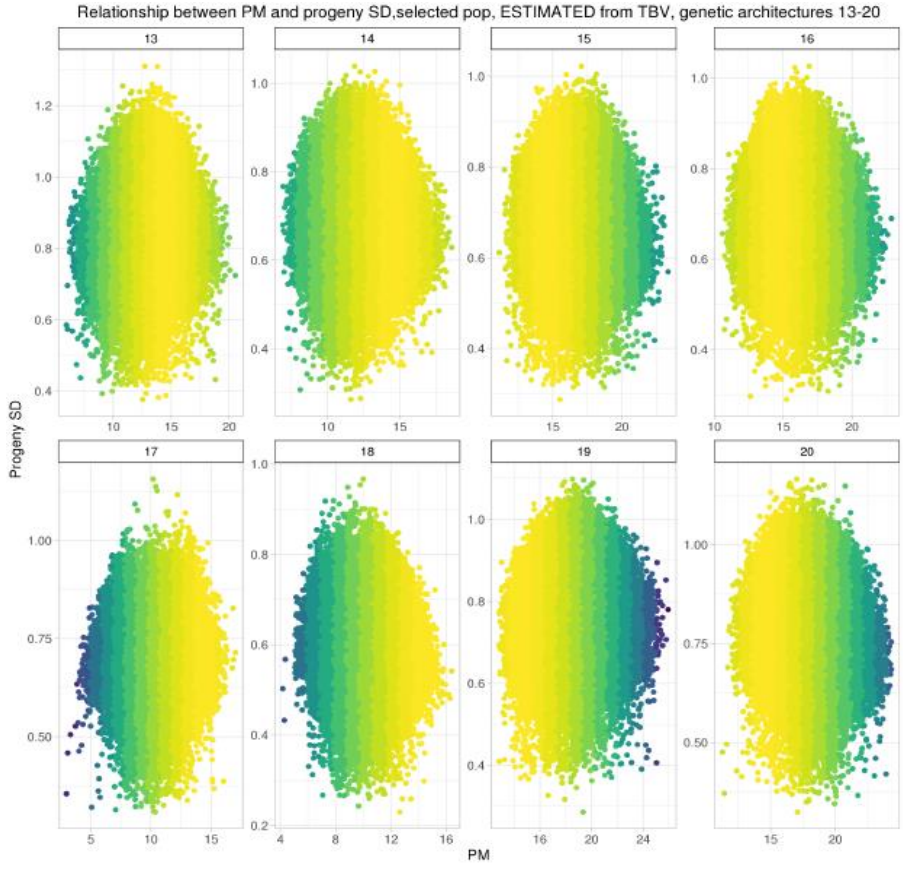

Relationship between PM and progeny SD,unselected pop, ESTIMATED from phenotypes, genetic architectures 1-12
1
2
3
4
5
6
7
8
9
10
11
12
Progeny SD
PM

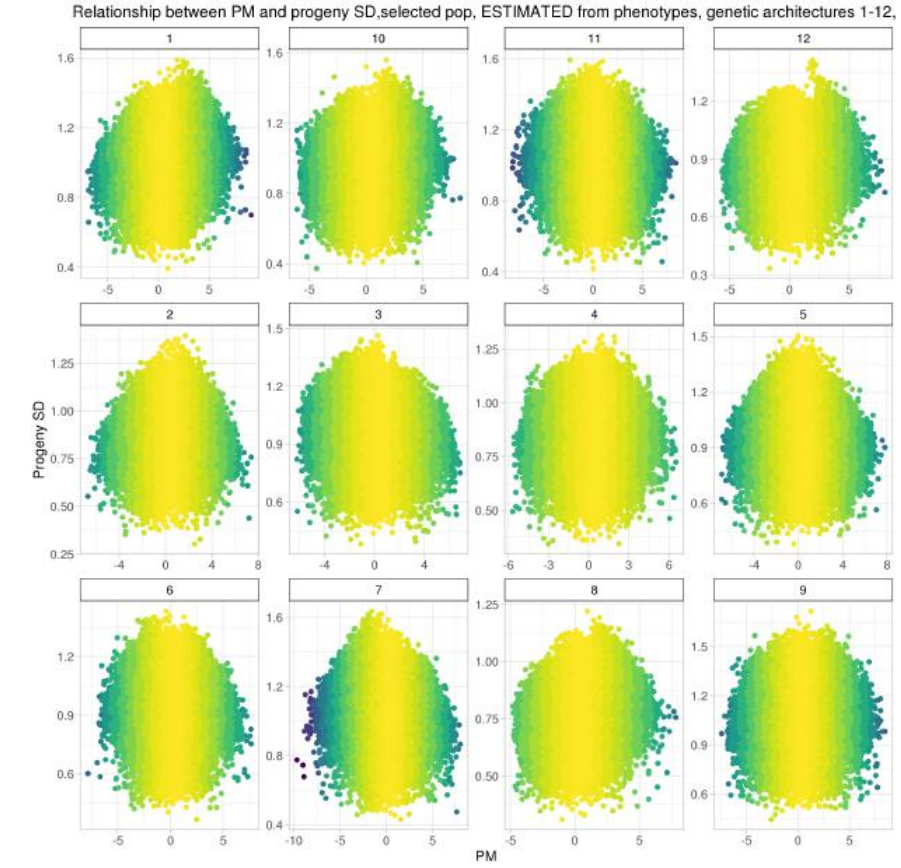

Relationship between PM and progeny SD,unselected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
PM

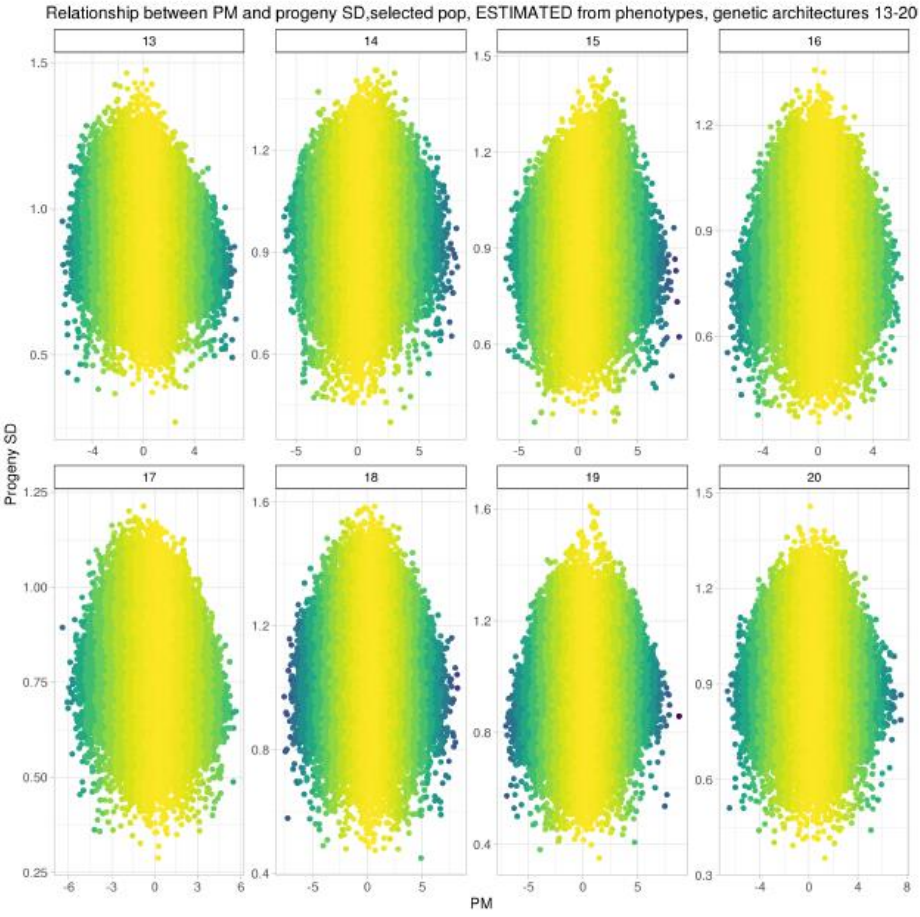

Relationship between PM and progeny SD,selected pop, TRUE, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
Progeny SD
PM

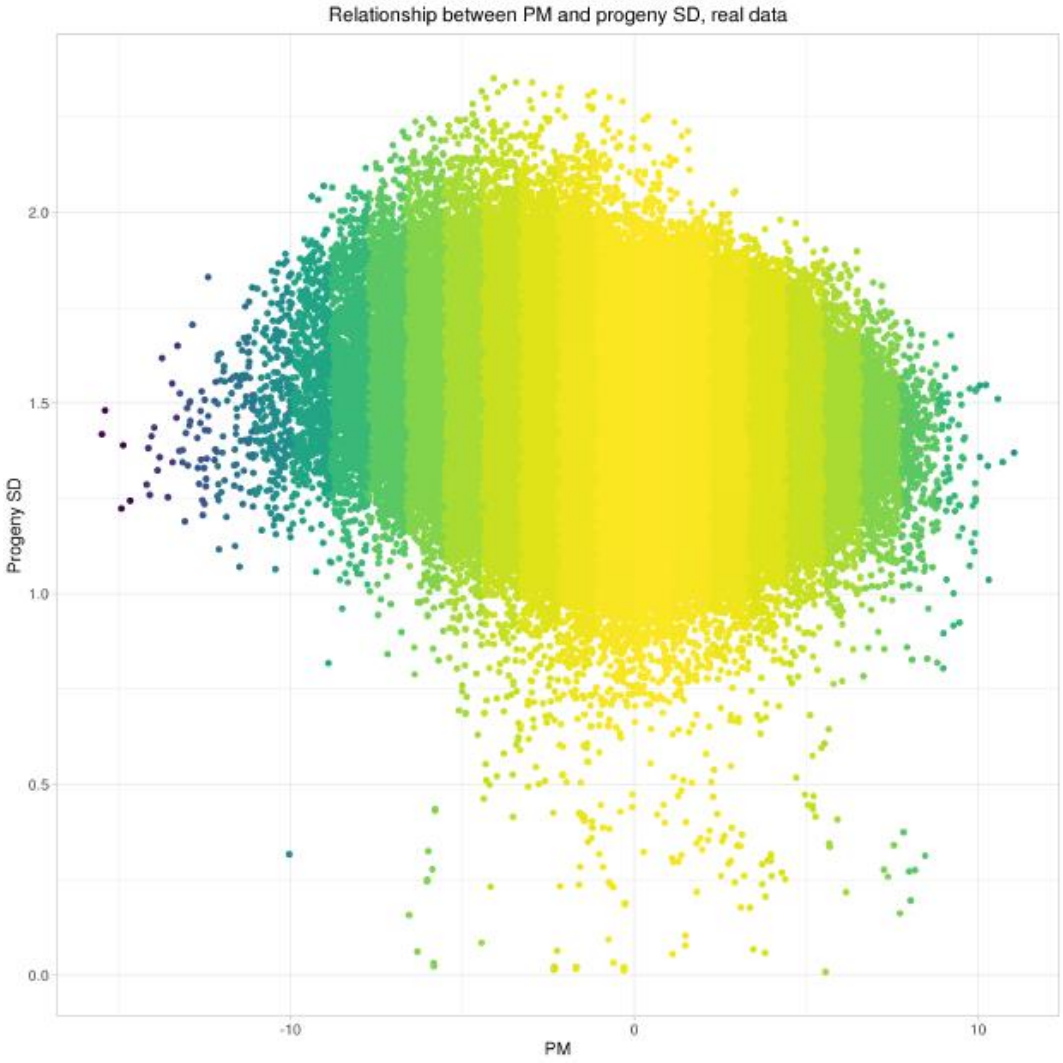

Relationship between PM and progeny SD,selected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD
PM

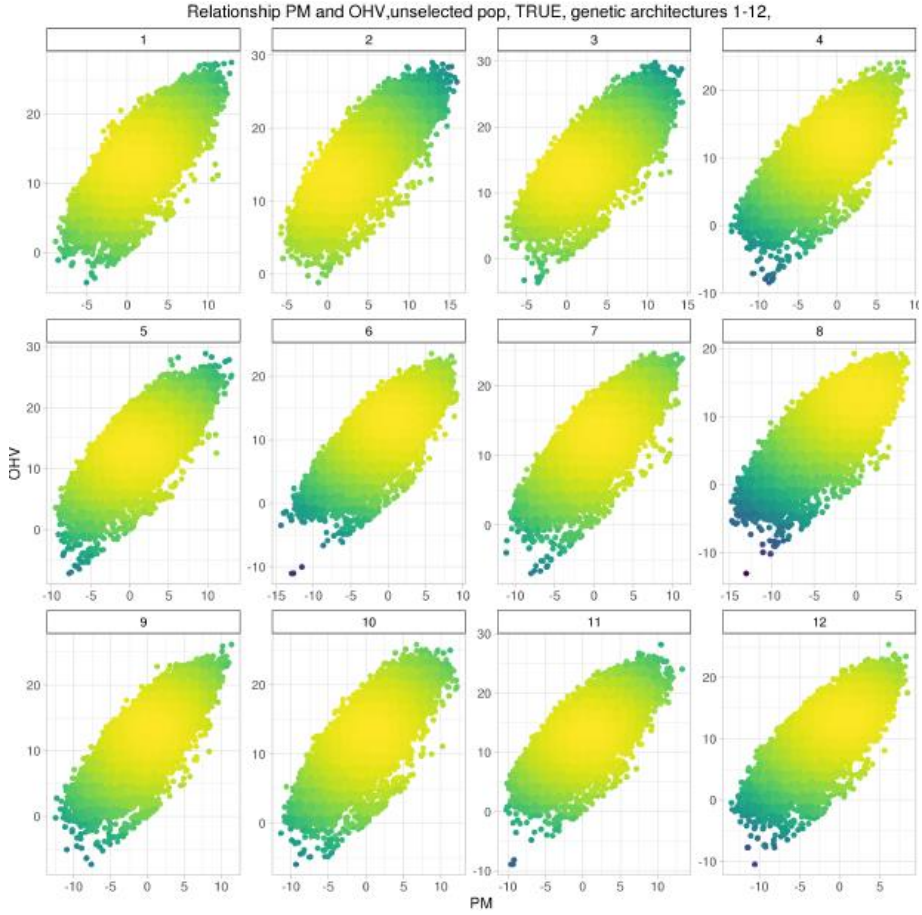

Relationship between PM and progeny SD,selected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
Progeny SD
PM

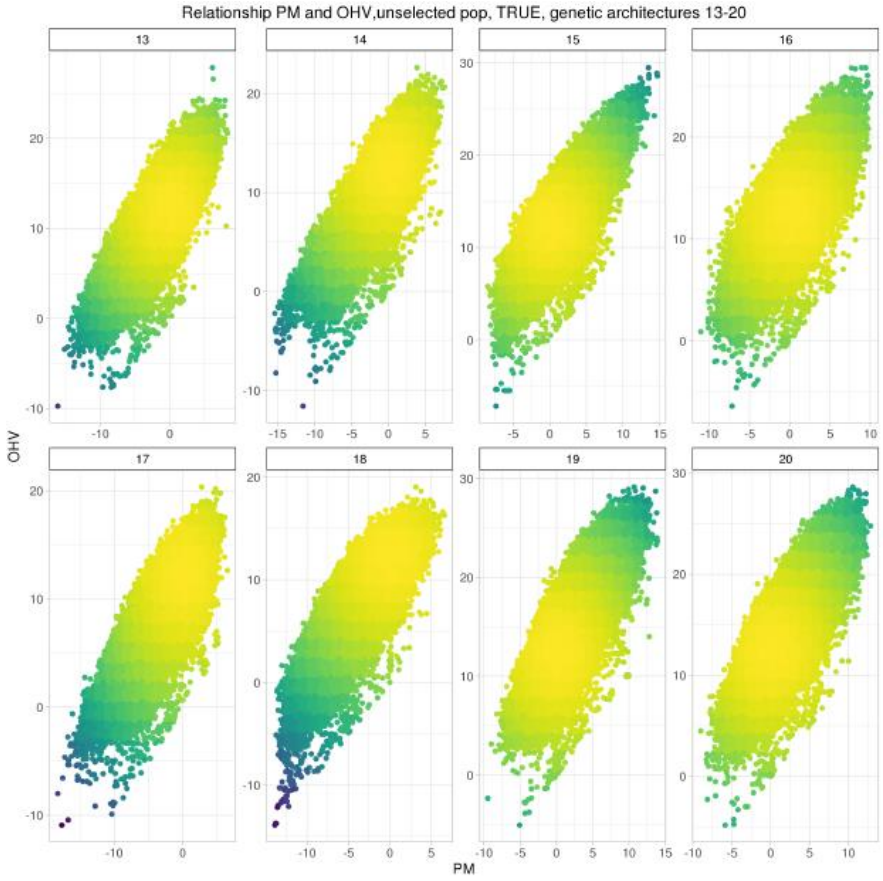

Relationship between PM and progeny SD,selected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD
PM

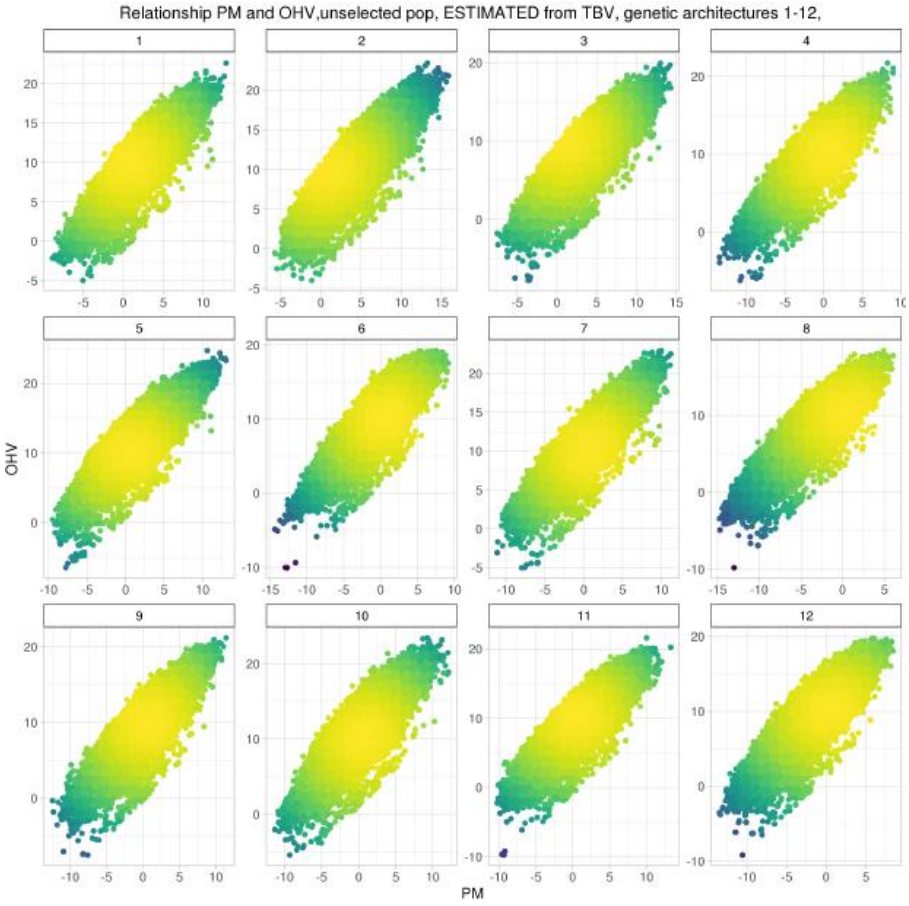

Relationship between PM and progeny SD,selected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
Progeny SD
PM

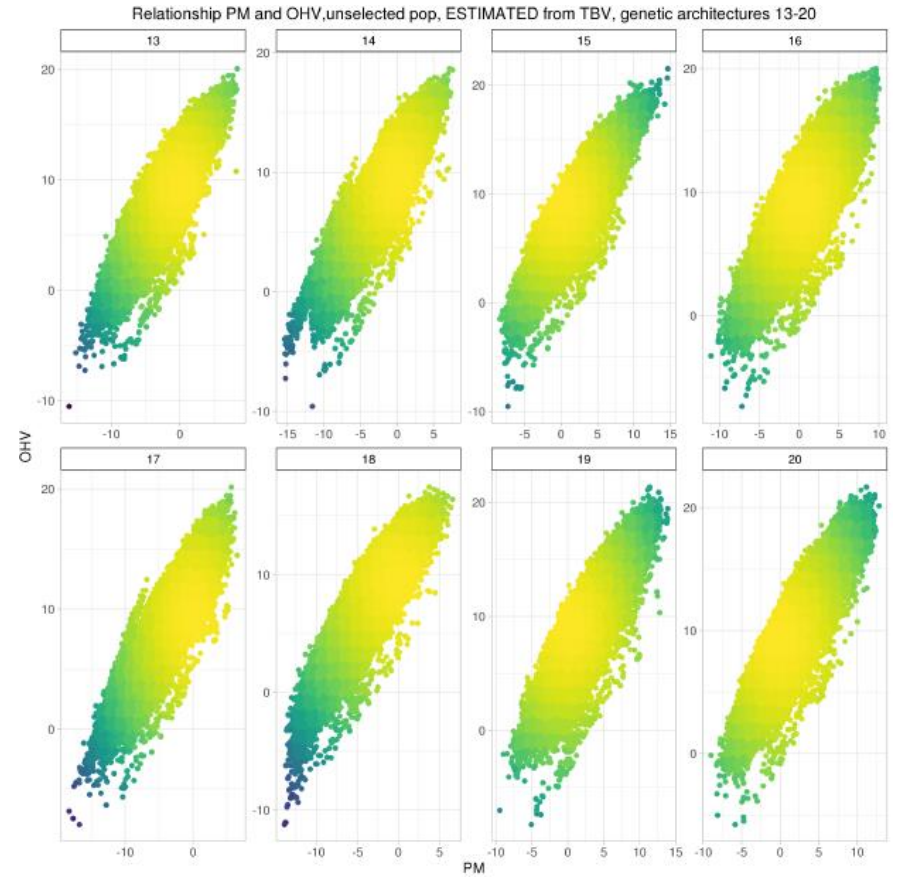

Relationship between PM and progeny SD,selected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
Progeny SD
PM

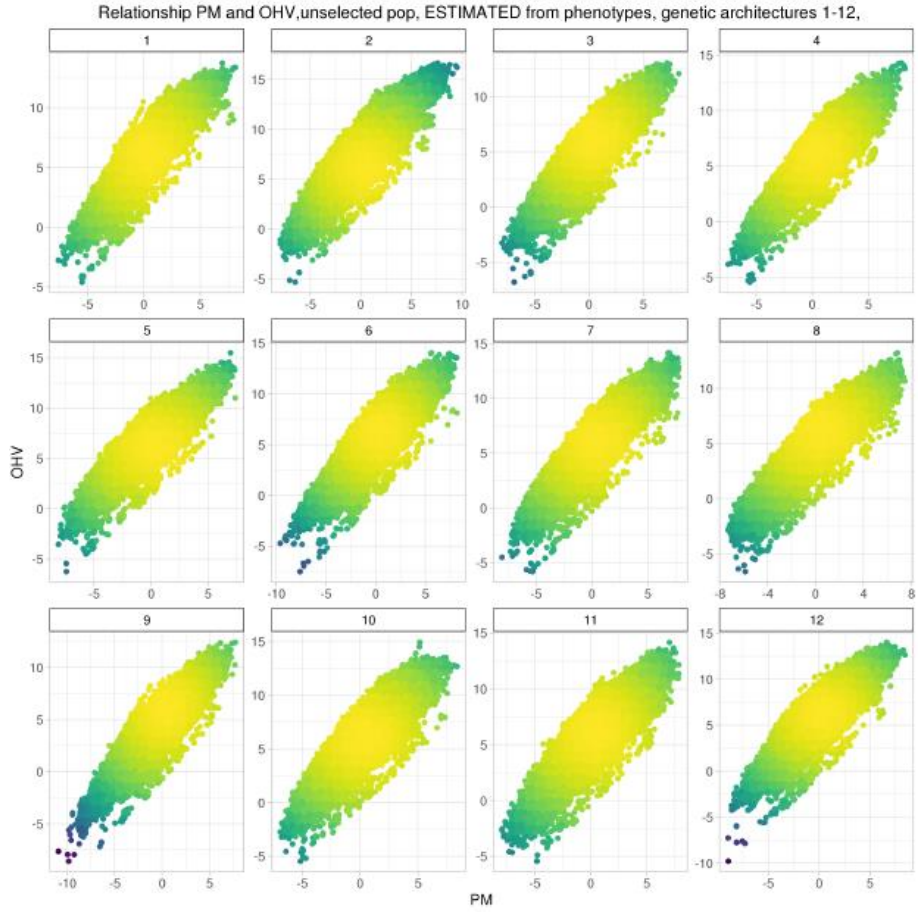

Relationship between PM and progeny SD, real data
Progeny SD
2.0
1.5
1.0
0.5
0.0
-10
0
10
PM

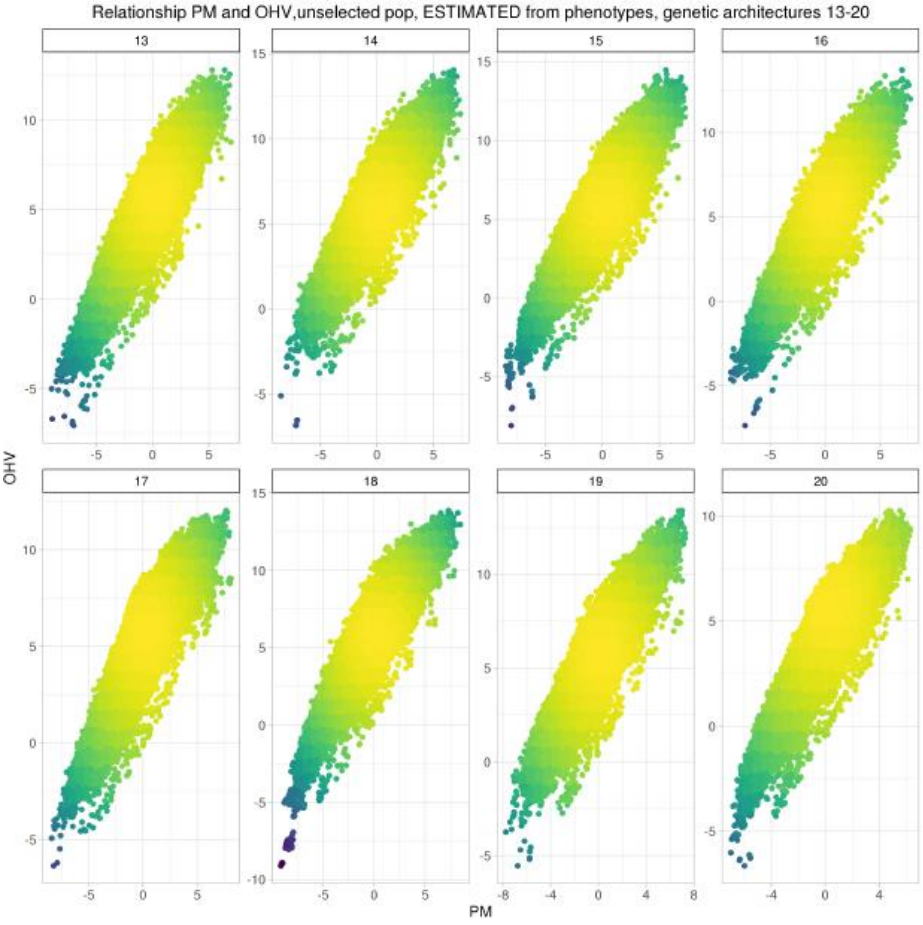

Relationship PM and OHV,unselected pop, TRUE, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
OHV
PM

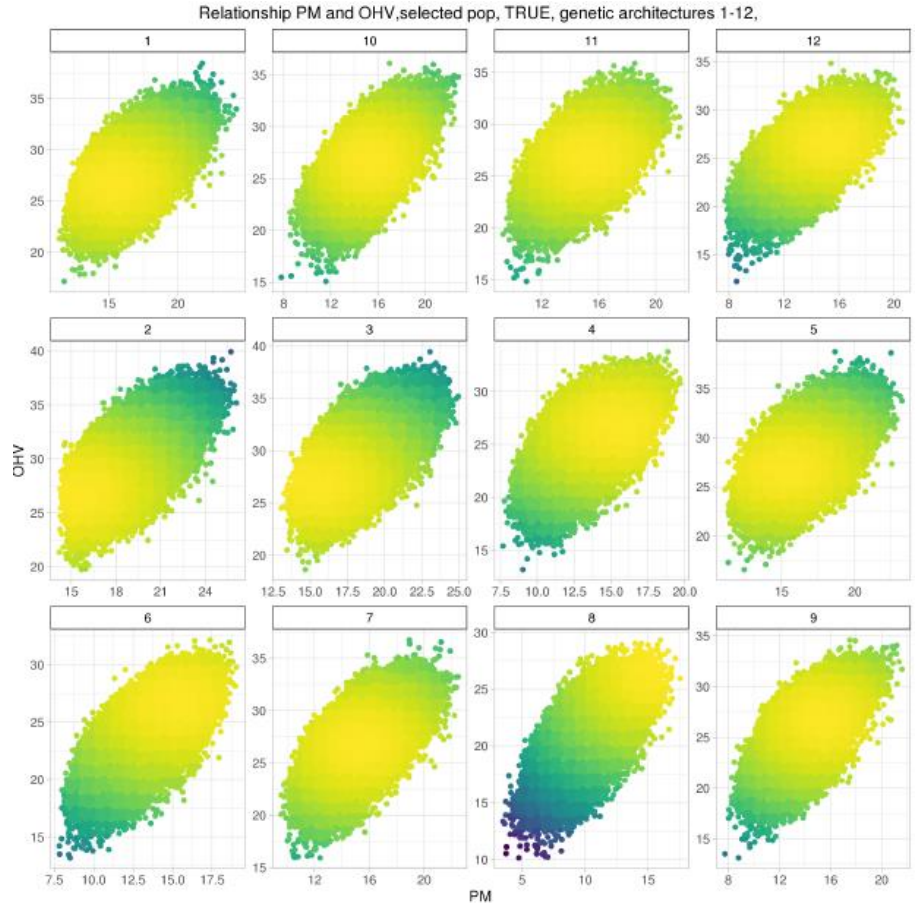

Relationship PM and OHV,unselected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

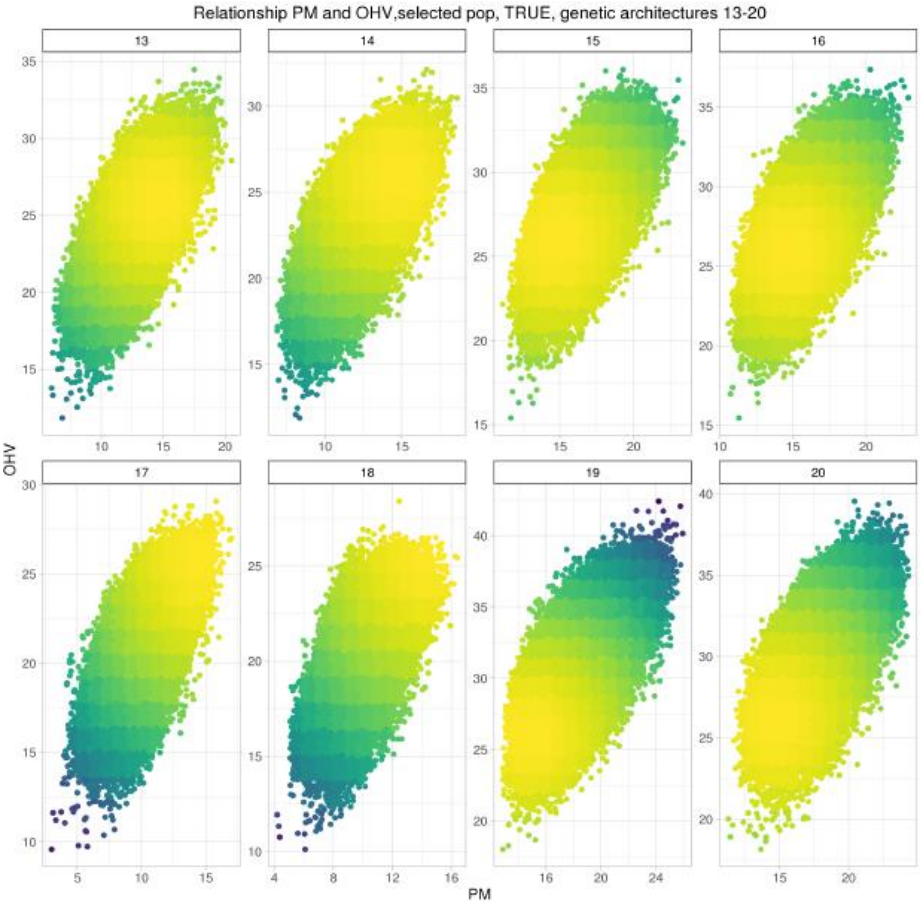

Relationship PM and OHV,unselected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
OHV
PM

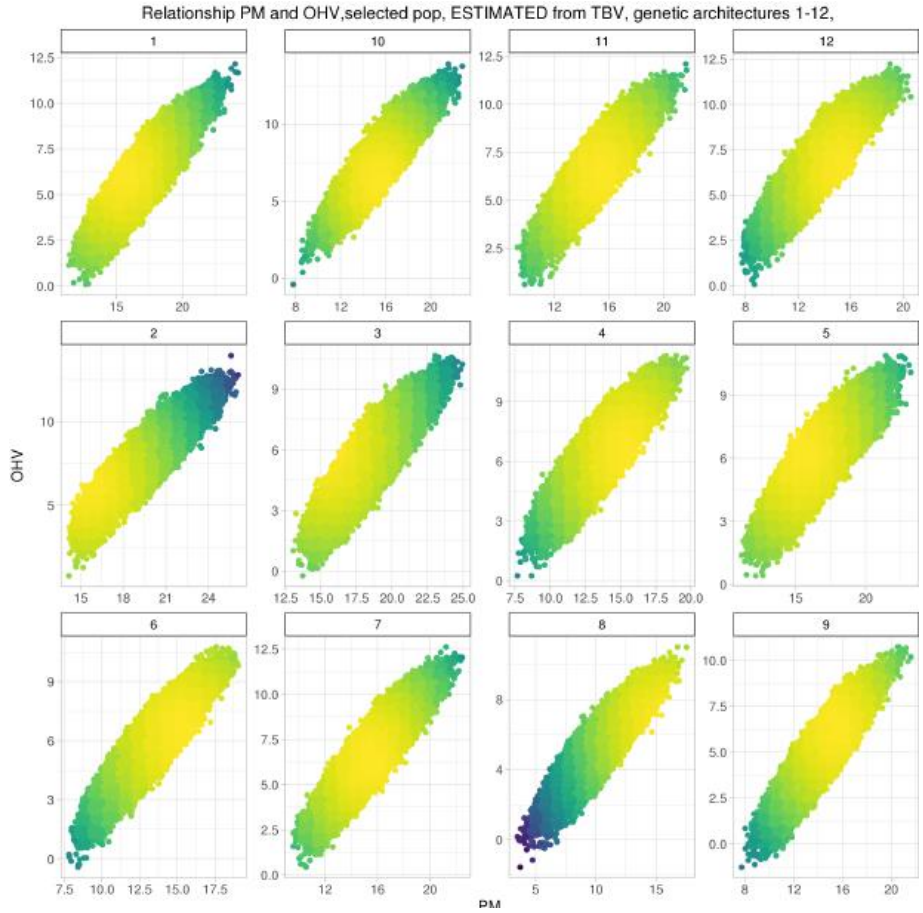

Relationship PM and OHV,unselected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

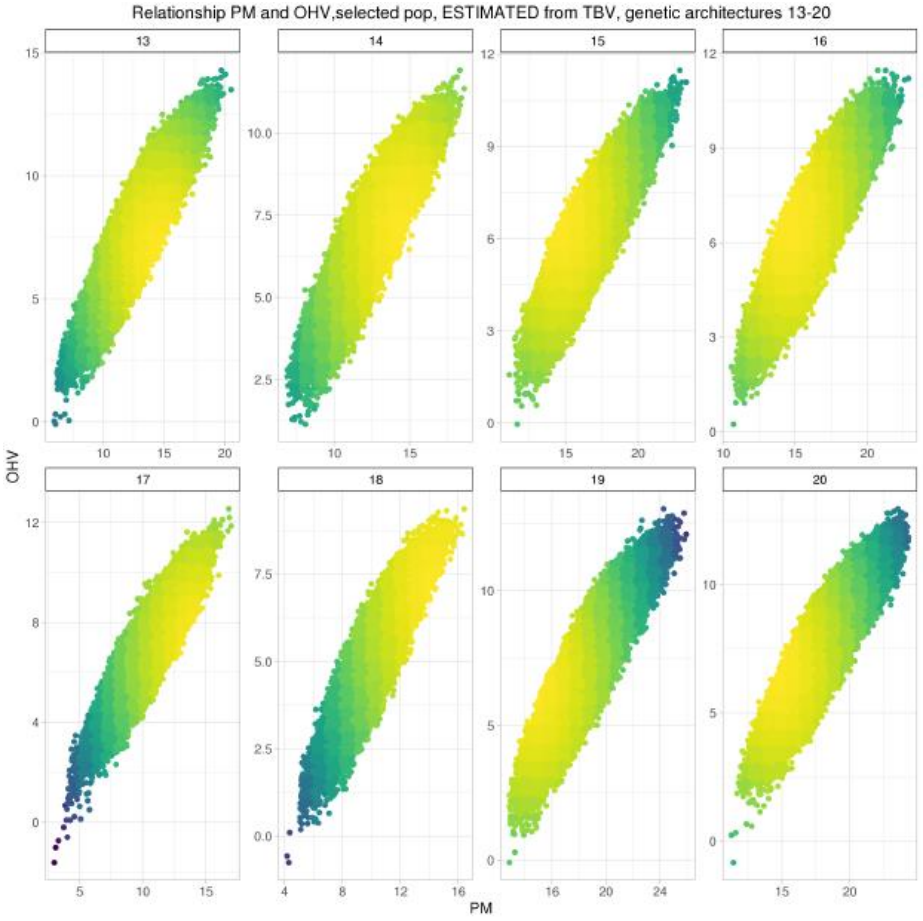

Relationship PM and OHV,unselected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
OHV
PM

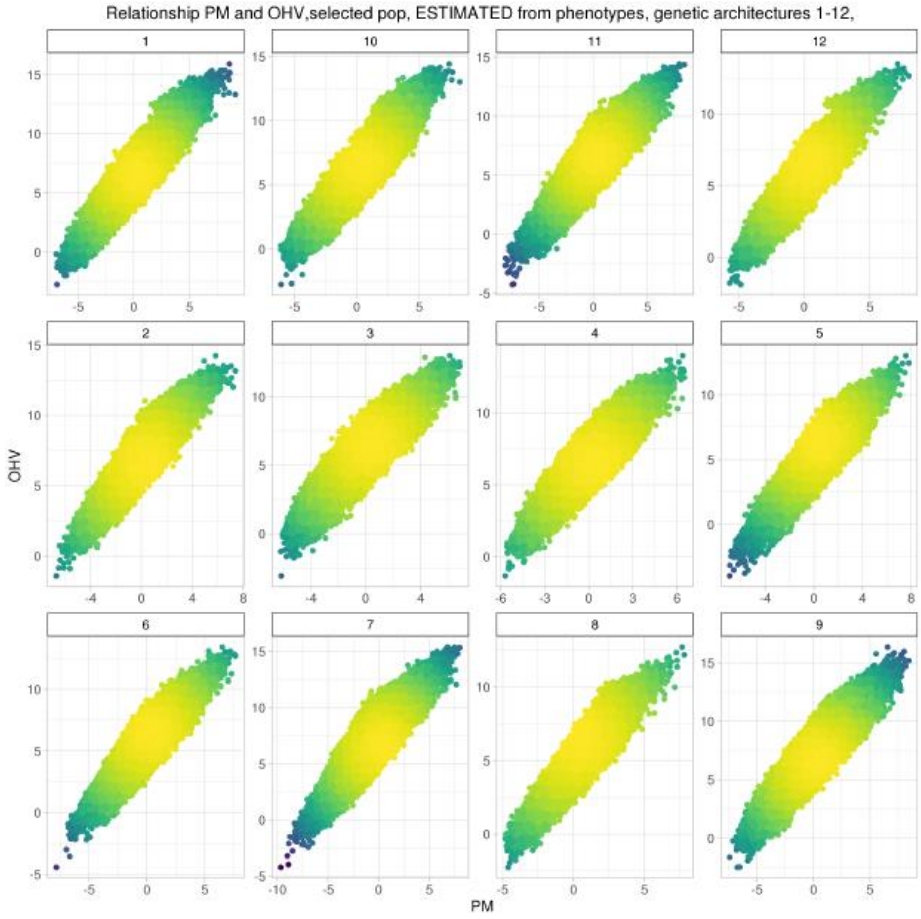

Relationship PM and OHV,unselected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

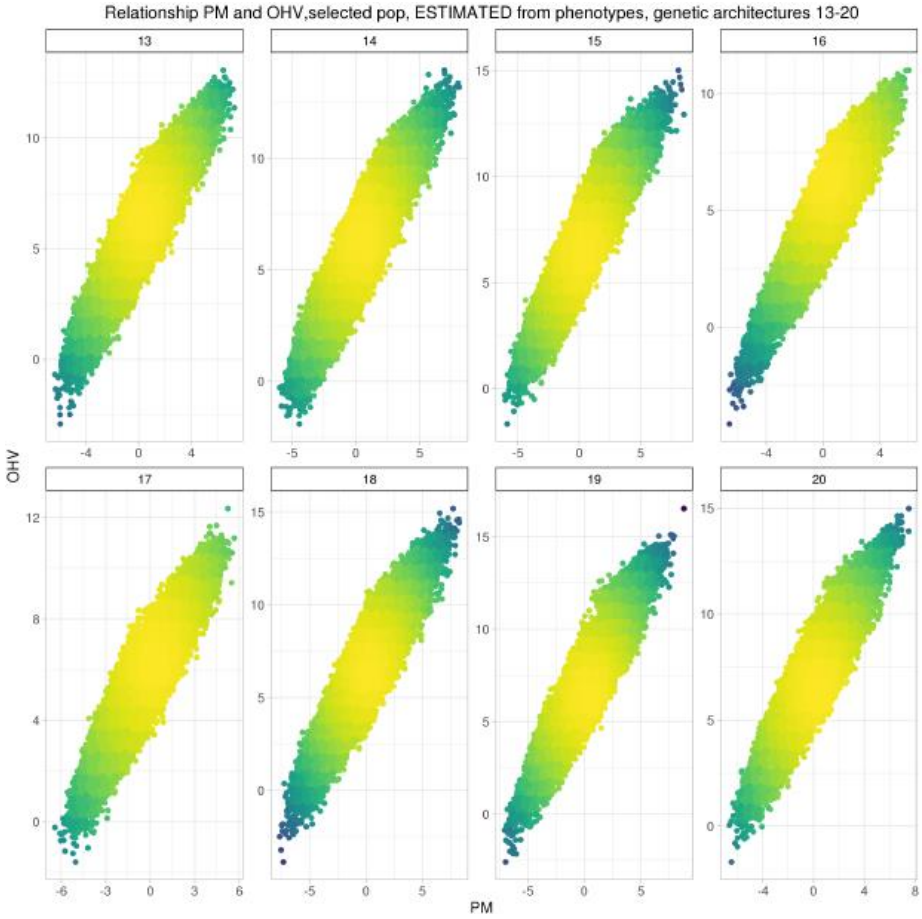

Relationship PM and OHV,selected pop, TRUE, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
OHV
PM

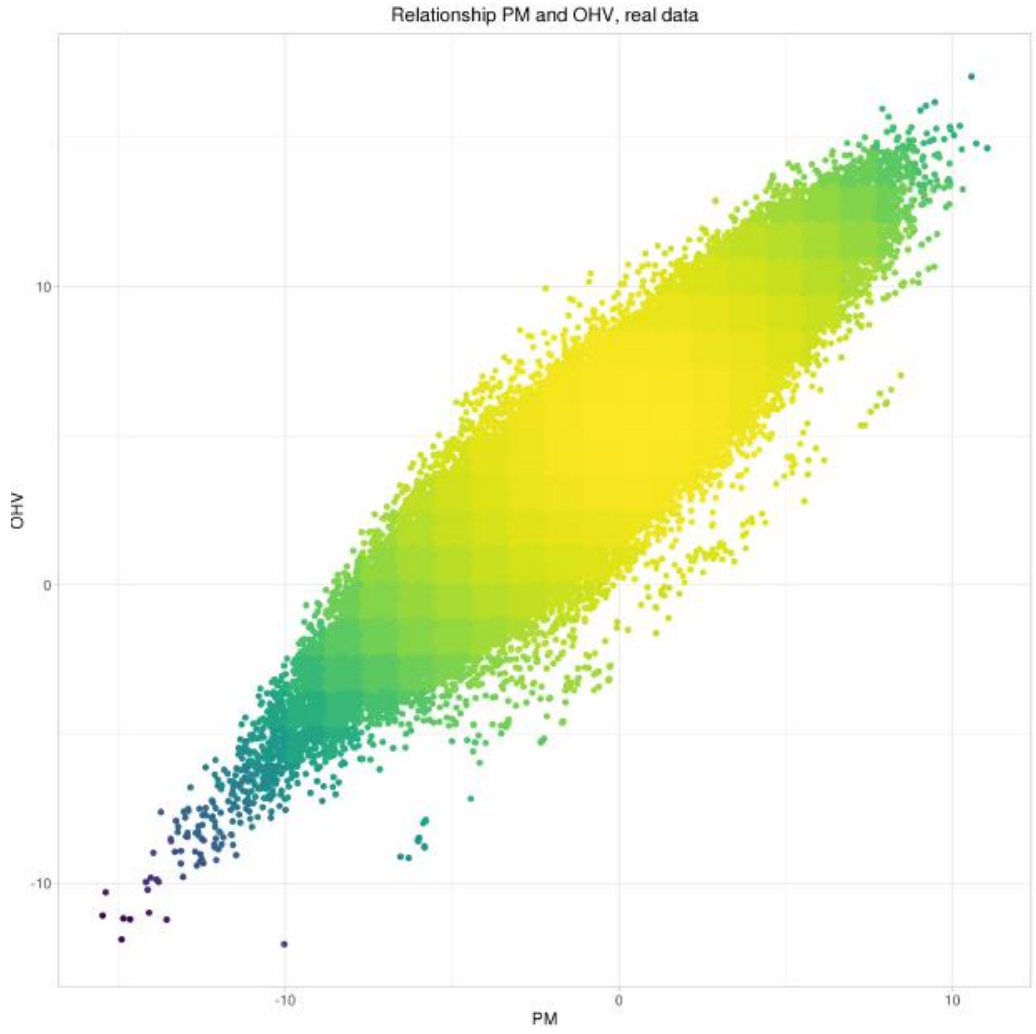

Relationship PM and OHV,selected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

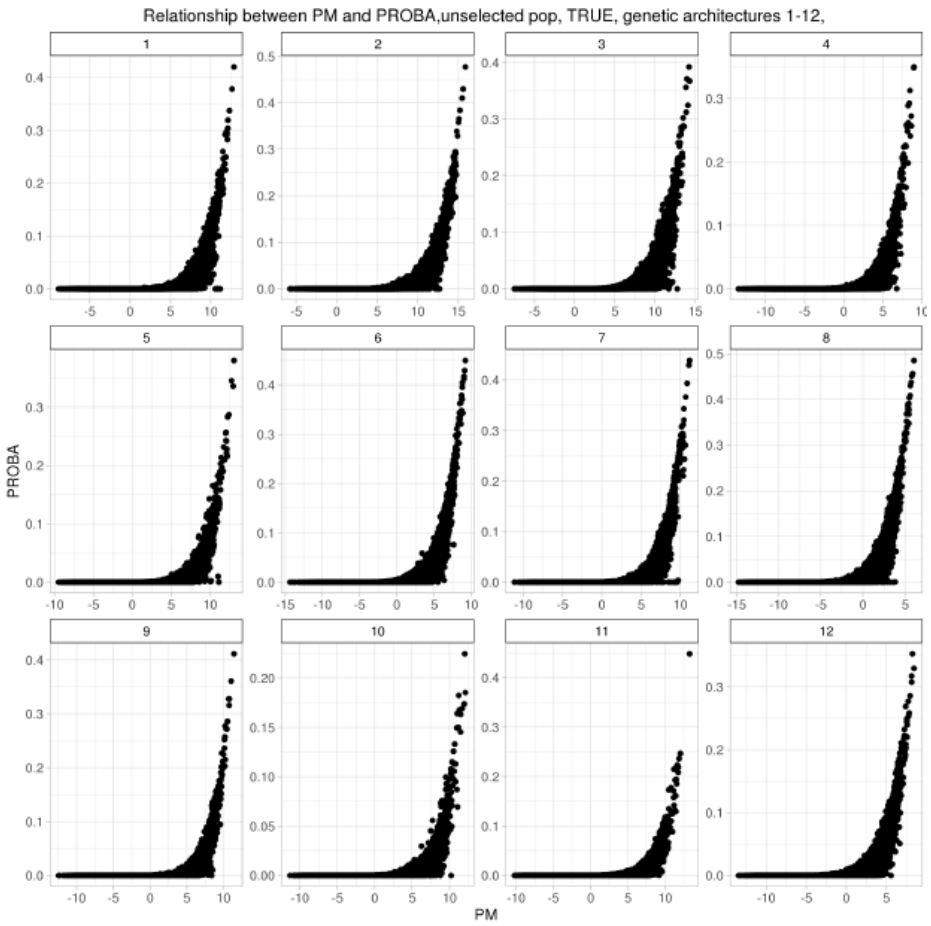

Relationship PM and OHV,selected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
OHV
PM

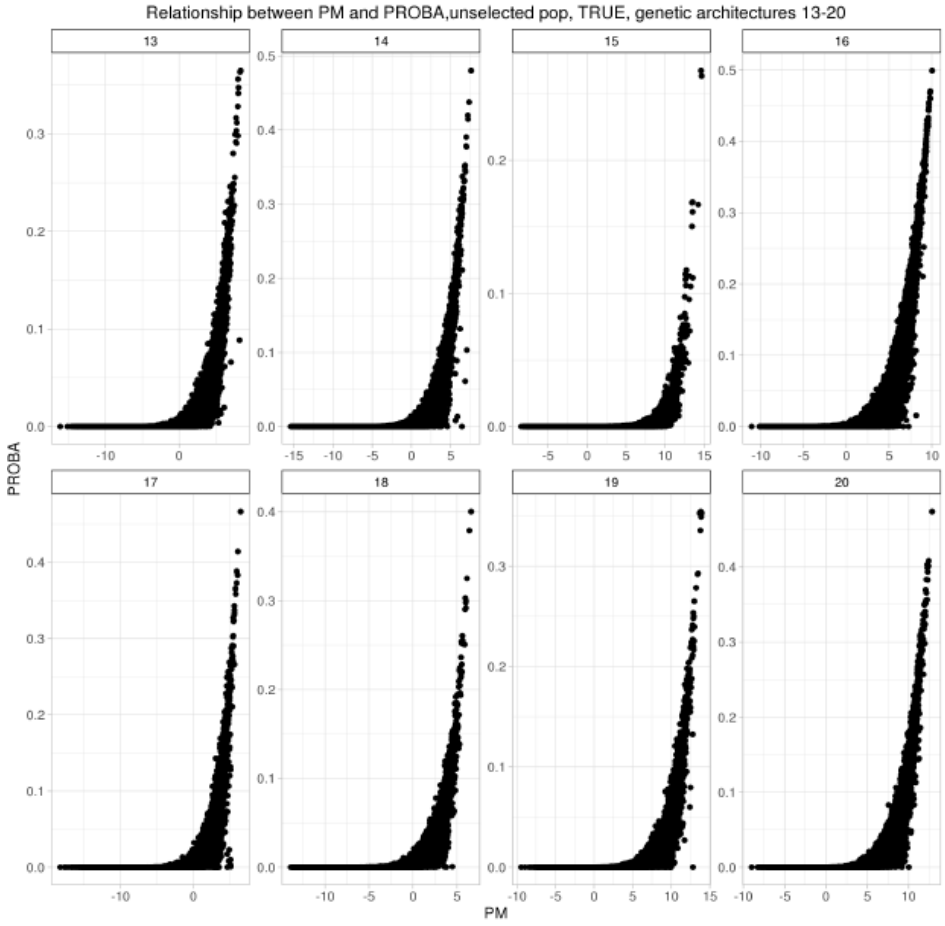

Relationship PM and OHV,selected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

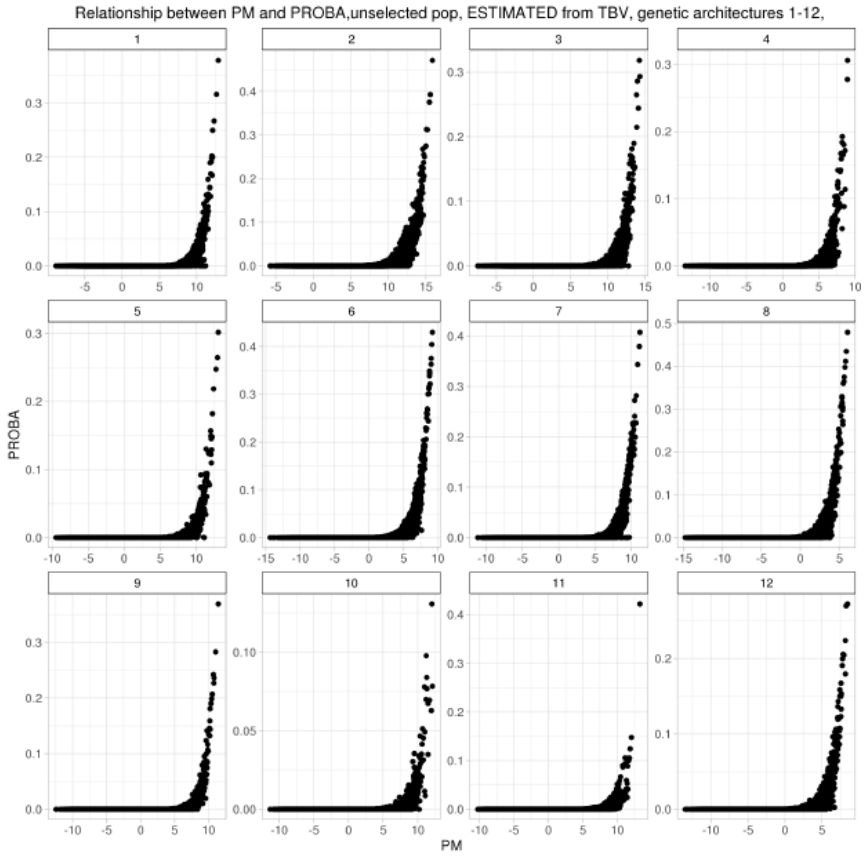

Relationship PM and OHV,selected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
OHV
PM

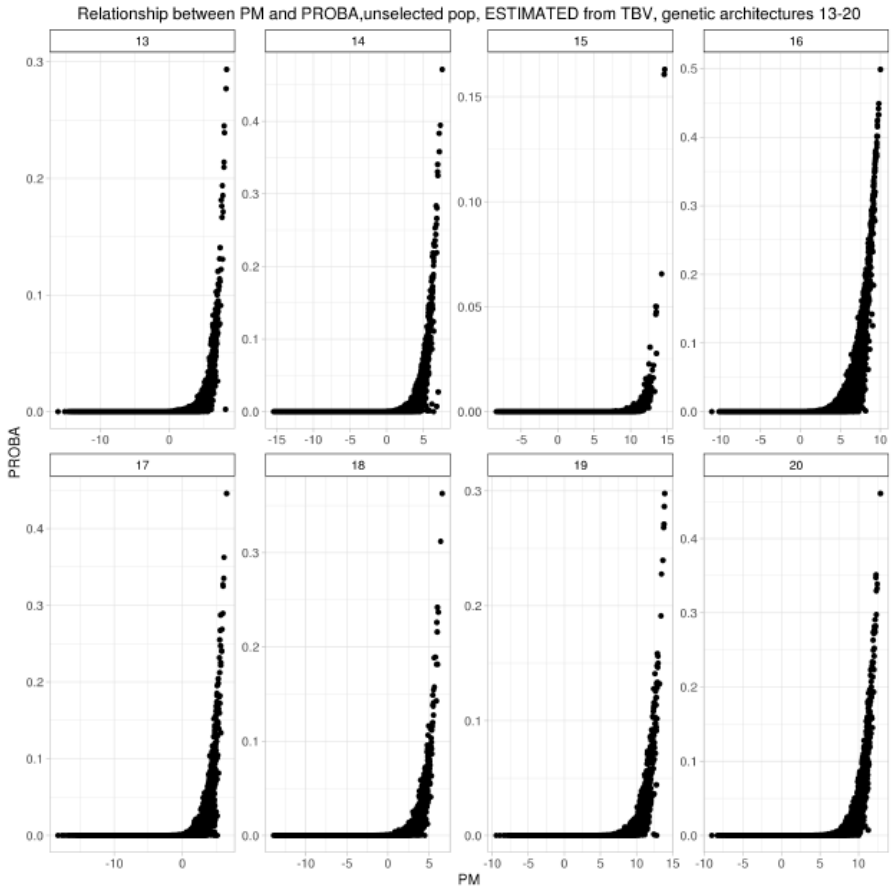

Relationship PM and OHV,selected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
OHV
PM

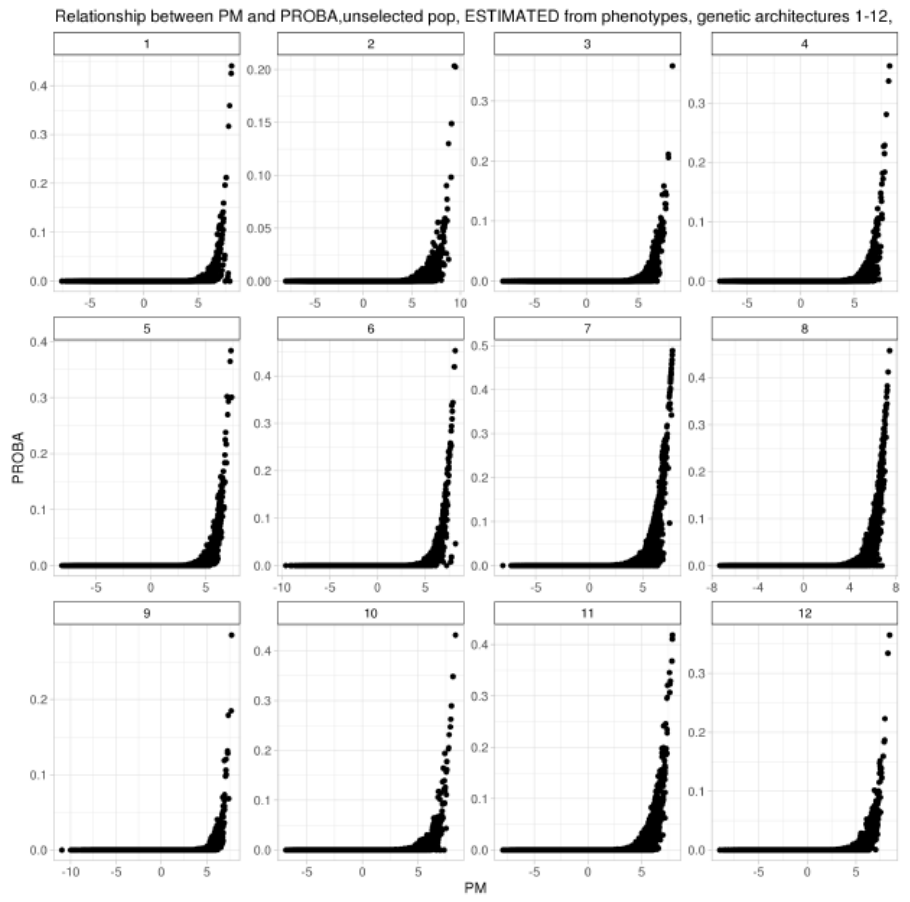

Relationship PM and OHV, real data
OHV
10
0
-10
-10
0
10
PM

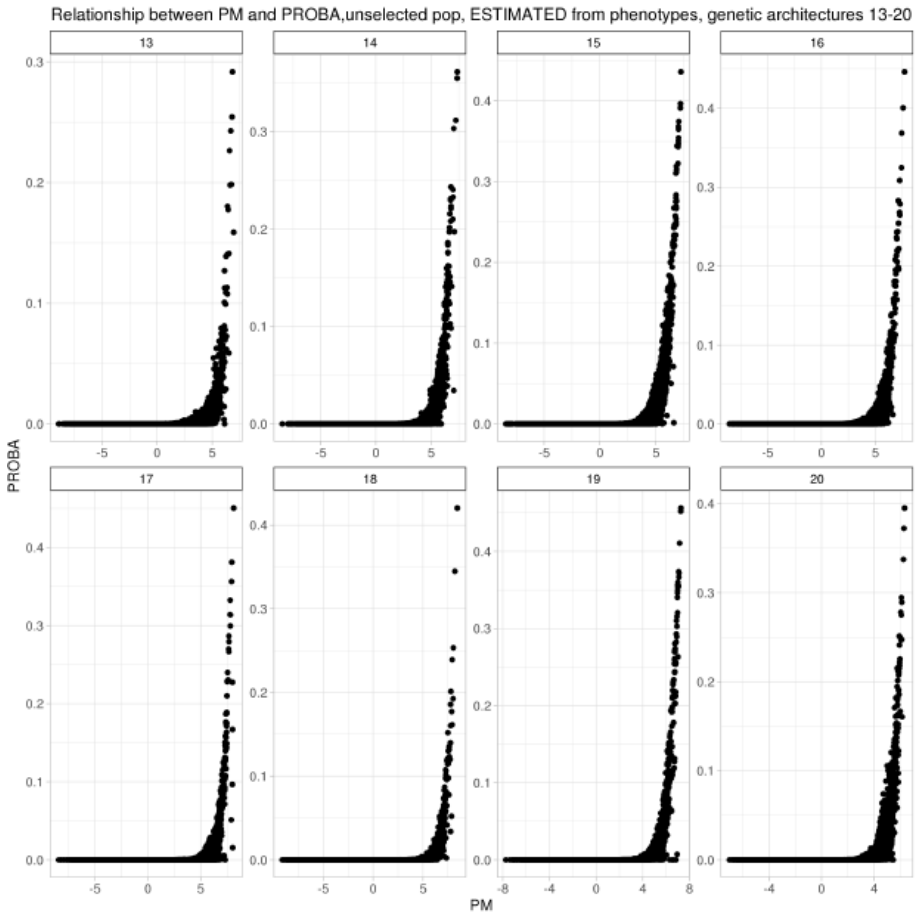

Relationship between PM and PROBA,unselected pop, TRUE, genetic architectures 1-12,
PROBA
PM

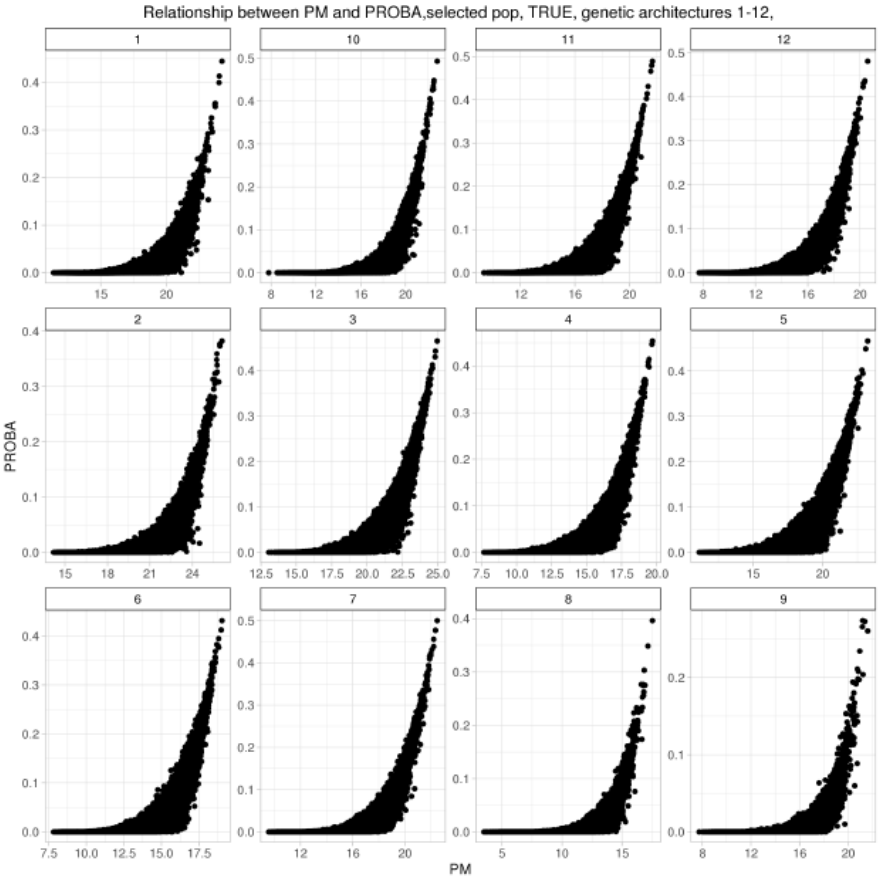

Relationship between PM and PROBA,unselected pop, TRUE, genetic architectures 13-20
PROBA
PM

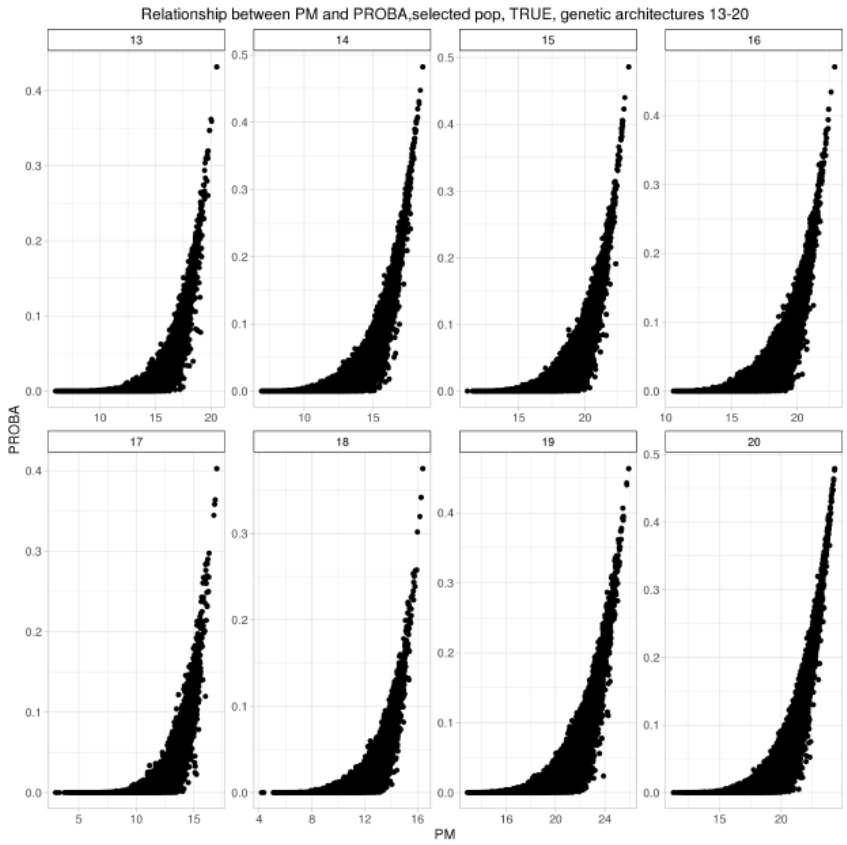

Relationship between PM and PROBA,unselected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PROBA
PM

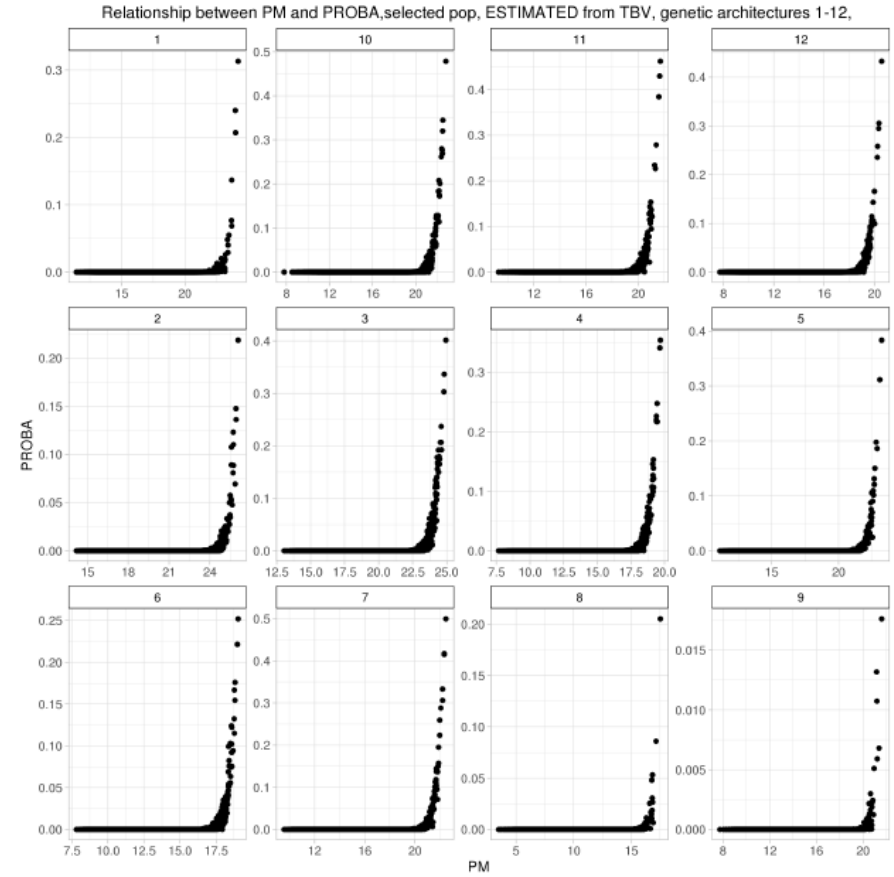

Relationship between PM and PROBA,unselected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
PROBA
PM

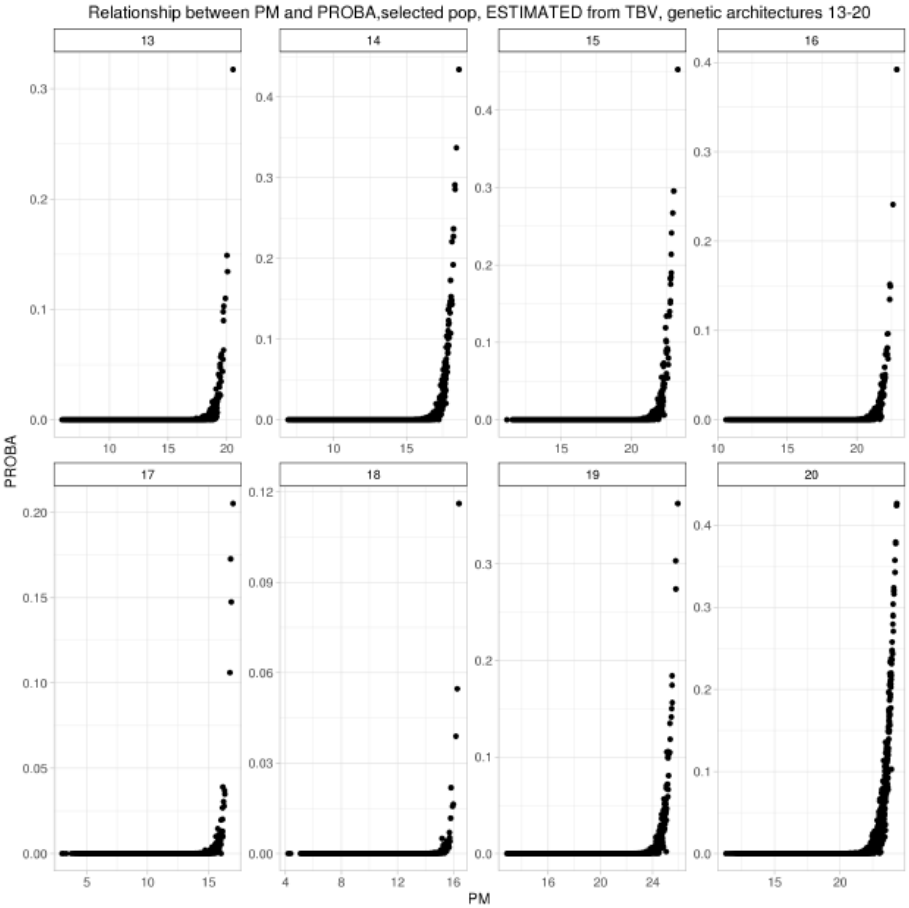

Relationship between PM and PROBA,unselected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
2
3
4
5
6
7
8
9
10
11
12
PROBA
PM

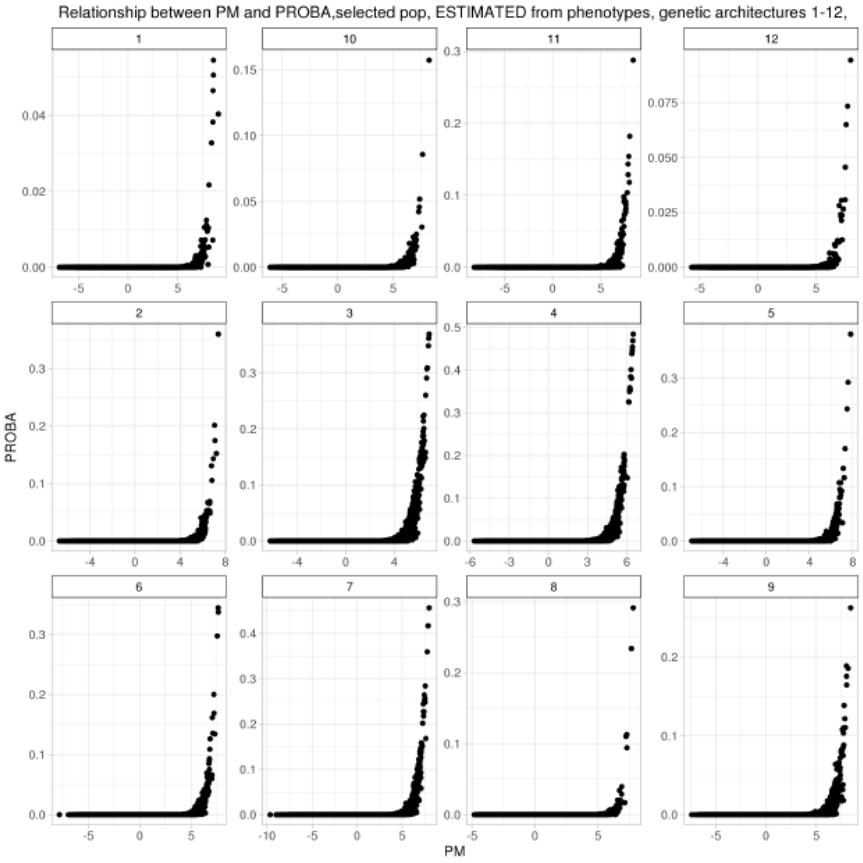

Relationship between PM and PROBA,unselected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
PROBA
PM

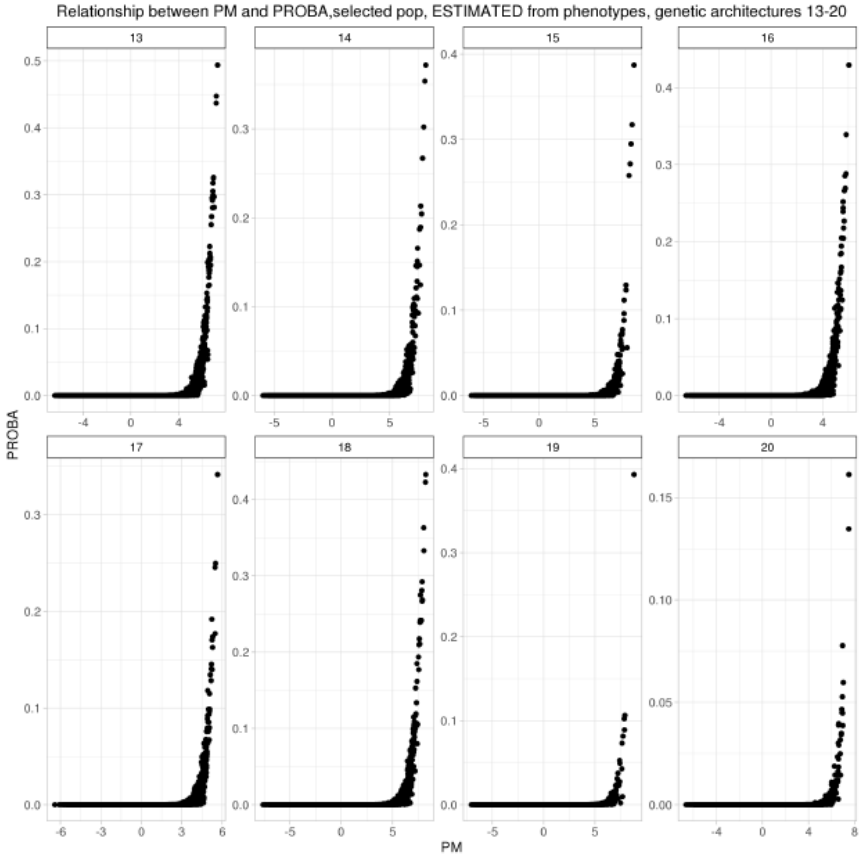

Relationship between PM and PROBA,selected pop, TRUE, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
PROBA
PM

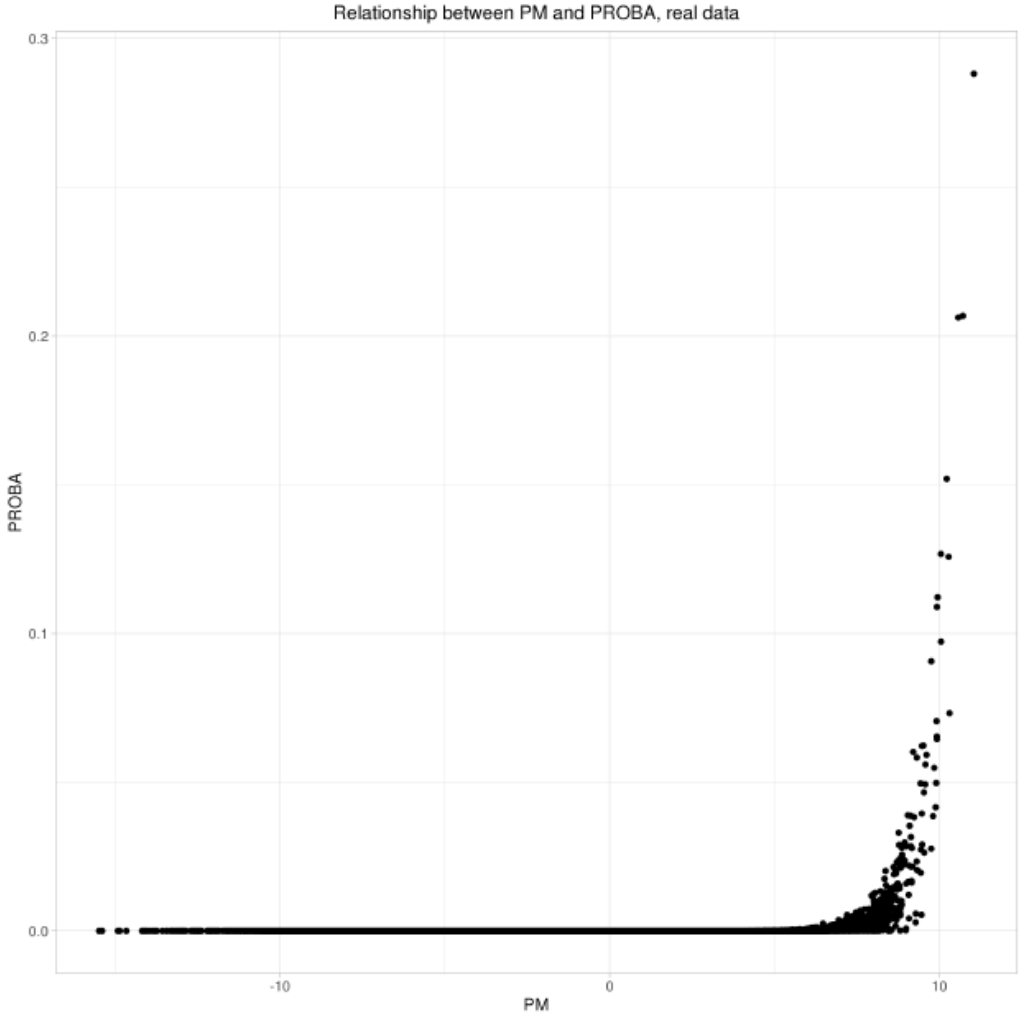

Relationship between PM and PROBA,selected pop, TRUE, genetic architectures 13-20
13
14
15
16
17
18
19
20
PROBA
PM


Relationship between PM and PROBA,selected pop, ESTIMATED from TBV, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
PROBA
PM


Relationship between PM and PROBA,selected pop, ESTIMATED from TBV, genetic architectures 13-20
13
14
15
16
17
18
19
20
PROBA
PM


Relationship between PM and PROBA,selected pop, ESTIMATED from phenotypes, genetic architectures 1-12,
1
10
11
12
2
3
4
5
6
7
8
9
PROBA
PM


Relationship between PM and PROBA,selected pop, ESTIMATED from phenotypes, genetic architectures 13-20
13
14
15
16
17
18
19
20
PROBA
PM


Relationship between PM and PROBA, real data
PROBA
0.3
0.2
0.1
0.0
-10
0
10
PM

Références bibliographiques

Abed, A., and Belzile, F. (2019). Exploring the realm of possibilities: trying to predict promising crosses and successful offspring through genomic mating in barley. Crop Breed. Genet. Genomics 1.

Adeyemo, E., and Bernardo, R. (2019). Predicting Genetic Variance from Genomewide Marker Effects Estimated from a Diverse Panel of Maize Inbreds. Crop Sci. 59, 583–590.

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A., and Wang, H. (2014). PREGSF90–POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. In 10. World Congress on Genetics Applied to Livestock Production (WCGALP), (American Society of Animal Science), p.

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A.K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity 122, 672–683.

Aliyeva-Schnorr, L., Beier, S., and Karafiátová, M. (2015). Cytogenetic mapping with centromeric BAC contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. Plant J 84, 385–394.

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019a). Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10, 1006.

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Melkior, S., Moreau, L., and Charcosset, A. (2019b). Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. Theor. Appl. Genet. 132, 1321–1334.

Allier, A., Teyssèdre, S., Lehermeier, C., Charcosset, A., and Moreau, L. (2020). Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Theor. Appl. Genet. 133, 201–215.

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., and Lehermeier, C. (2019c). Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression. G3 GenesGenomesGenetics 9, 1469.

Apuli, R.-P., Bernhardsson, C., Schiffthaler, B., Robinson, K.M., Jansson, S., Street, N.R., and Ingvarsson, P.K. (2020). Inferring the Genomic Landscape of Recombination Rate Variation in European Aspen (Populus tremula). G3 GenesGenomesGenetics 10, 299–309.

Auton, A., and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Res. 17, 1219–1227.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., et al. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science 336, 193.

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J.K., Hayward, J.J., Cohen, P.E., Greally, J.M., Wang, J., et al. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLOS Genet. 9, e1003984.

Baird, S.J.E. (2015). Exploring linkage disequilibrium. Mol. Ecol. Resour. 15, 1017–1019.

Balfourier, F., Bouchet, S., Robert, S., De Oliveira, R., Rimbert, H., Kitt, J., Choulet, F., and Paux, E. (2019). Worldwide phylogeography and history of wheat genetic diversity. Sci. Adv. 5, eaav0536.

Bassi, F.M., Bentley, A.R., Charmet, G., Ortiz, R., and Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci. 242, 23–36.

Battagin, M., Gorjanc, G., Faux, A.-M., Johnston, S.E., and Hickey, J.M. (2016). Effect of manipulating recombination rates on response to selection in livestock breeding programs. Genet. Sel. Evol. 48, 44.

Baudat, F., and de Massy, B. (2007). Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. Chromosome Res. 15, 565–577.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science 327, 836.

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Rincent, R., Schipprack, W., et al. (2013). Intraspecific variation of recombination rate in maize. Genome Biol. 14, 1–17.

Baye, T.M., He, H., Ding, L., Kurowski, B.G., Zhang, X., and Martin, L.J. (2011). Population structure analysis using rare and common functional variants. BMC Proc. 5, S8.

Beadle, G.W. (1939). Teosinte and the origin of maize. J. Hered. 30, 245–247.

Beckett, T.J., Rocheford, T.R., and Mohammadi, M. (2019). Reimagining Maize Inbred Potential: Identifying Breeding Crosses Using Genetic Variance of Simulated Progeny. Crop Sci. 59, 1457–1468.

Beeson, S.K., Mickelson, J.R., and McCue, M.E. (2019). Exploration of fine-scale recombination rate variation in the domestic horse. Genome Res.

Ben Sadoun, S., Rincent, R., Auzanneau, J., Oury, F.X., Rolland, B., Heumez, E., Ravel, C., Charmet, G., and Bouchet, S. (2020). Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. Theor. Appl. Genet. 133, 2197–2212.

Ben Sadoun, S. (2020). Optimisation du schéma de sélection chez le blé tendre : apport des prédictions génomiques et des caractères corrélés. Université de Clermont Auvergne.

Berchowitz, L.E., and Copenhaver, G.P. (2010). Genetic Interference: Dont Stand So Close to Me. Curr. Genomics 11, 91–102.

Berg, I.L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N.J., and Jeffreys, A.J. (2011). Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc. Natl. Acad. Sci. 108, 12378–12383.

Bernardo, R. (2014). Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations. Crop Sci. 54, 2586–2595.

Bernardo, R. (2017). Prospective Targeted Recombination and Genetic Gains for Quantitative Traits in Maize. Plant Genome 10, plantgenome2016.11.0118.

Bernardo, R., and Charcosset, A. (2006). Usefulness of Gene Information in Marker-Assisted Recurrent Selection: A Simulation Appraisal. Crop Sci. 46, 614–621.

Bijma, P., Wientjes, Y.C.J., and Calus, M.P.L. (2020). Breeding Top Genotypes and Accelerating Response to Recurrent Selection by Selecting Parents with Greater Gametic Variance. Genetics 214, 91.

Blary, A., and Jenczewski, E. (2019). Manipulation of crossover frequency and distribution for plant breeding. Theor. Appl. Genet. 132, 575–592.

Bohn, M., Utz, H.F., and Melchinger, A.E. (1999). Genetic Similarities among Winter Wheat Cultivars Determined on the Basis of RFLPs, AFLPs, and SSRs and Their Use for Predicting Progeny Variance. Crop Sci. 39, cropsci1999.0011183X003900010035x.

Boichard, D., Ducrocq, V., Croiseau, P., and Fritz, S. (2016). Genomic selection in domestic animals: Principles, applications and perspectives. C. R. Biol. 339, 274–277.

Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. Genet. Sel. Evol. 48, 36.

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature 485, 642–645.

Brisbane, J.R., and Gibson, J.P. (1995). Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. Theor. Appl. Genet. 91, 421–431.

Brockhurst, M.A., Chapman, T., King, K.C., Mank, J.E., Paterson, S., and Hurst, G.D.D. (2014). Running with the Red Queen: the role of biotic conflicts in evolution. Proc. R. Soc. B Biol. Sci. 281, 20141382.

Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. 98, 116–126.

Browning, S.R., and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81, 1084–1097.

Brunschwig, H., Levi, L., Ben-David, E., Williams, R.W., Yakir, B., and Shifman, S. (2012). Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. Genetics 191, 757–764.

Bulmer, M.G. (1971). The Effect of Selection on Genetic Variability. Am. Nat. 105, 201–211.

Burrows, P.M. (1972). Expected Selection Differentials for Directional Selection. Biometrics 28, 1091–1100.

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., and van Eeuwijk, F. (2016a). Improvement of Predictive Ability by Uniform Coverage of the Target Genetic Space. G3 GenesGenomesGenetics 6, 3733.

Calderini, D.F., and Slafer, G.A. (1998). Changes in yield and yield stability in wheat during the 20th century. Field Crops Res. 57, 335–347.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J.M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182, 375–385.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193, 327–345.

de los Campos, G., Sorensen, D.A., and Toro, M.A. (2019). Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). G3 GenesGenomesGenetics 9, 1429–1436.

Carlborg, Ö., Jacobsson, L., Åhgren, P., Siegel, P., and Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. Nat. Genet. 38, 418–420.

Chakraborty, R., Moreau, L., and Dekkers, J.C.M. (2002). A method to optimize selection on multiple identified quantitative trait loci. Genet. Sel. Evol. 34, 145–170.

Chambon, A., West, A., Vezon, D., Horlow, C., De Muyt, A., Chelysheva, L., Ronceret, A., Darbyshire, A., Osman, K., Heckmann, S., et al. (2018). Identification of ASYNAPTIC4, a Component of the Meiotic Chromosome Axis. Plant Physiol. 178, 233.

Chan, A.H., Jenkins, P.A., and Song, Y.S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLoS Genet. 8, e1003090.

Charlesworth, B., and Charlesworth, D. (2010). Elements of evolutionary genetics (United States of America: Roberts and Company Publishers Greenwood Village, CO).

Charmet, G. (2011). Wheat domestication: Lessons for the future. C. R. Biol. 334, 212–220.

Cheng, H., Liu, J., Wen, J., Nie, X., Xu, L., Chen, N., Li, Z., Wang, Q., Zheng, Z., Li, M., et al. (2019). Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol. 20, 136.

Choi, K., and Henderson, I.R. (2015). Meiotic recombination hotspots – a comparative view. Plant J. 83, 52–61.

Choi, K., Zhao, X., Kelly, K.A., Venn, O., Higgins, J.D., Yelina, N.E., Hardcastle, T.J., Ziolkowski, P.A., Copenhaver, G.P., Franklin, F.C.H., et al. (2013). Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat. Genet. 45, 1327–1336.

Choi, K., Reinhard, C., Serra, H., Ziolkowski, P.A., Underwood, C.J., Zhao, X., Hardcastle, T.J., Yelina, N.E., Griffin, C., and Jackson, M. (2016). Recombination rate heterogeneity within Arabidopsis disease resistance genes. PLoS Genet. 12, e1006179.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., and Paux, E. (2014). Structural and functional partitioning of bread wheat chromosome 3B. Science 345, 1249721.

Clark, S.A., Hickey, J.M., and Van der Werf, J.H. (2011). Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43, 1–9.

Cole, F., Keeney, S., and Jasin, M. (2010). Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. Mol. Cell 39, 700–710.

Cole, J.B., and VanRaden, P.M. (2011). Use of haplotypes to estimate Mendelian sampling effects and selection limits. J. Anim. Breed. Genet. 128, 446–455.

Coop, G., Wen, X., Ober, C., Pritchard, J.K., and Przeworski, M. (2008). High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science 319, 1395.

Cooper, M., Stucker, R.E., DeLacy, I.H., and Harch, B.D. (1997). Wheat Breeding Nurseries, Target Environments, and Indirect Selection for Grain Yield. Crop Sci. 37, cropsci1997.0011183X003700040024x.

Crain, J., Mondal, S., Rutkoski, J., Singh, R.P., and Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. Plant Genome 11, 170043.

Crismani, W., Girard, C., Froger, N., Pradillo, M., Santos, J.L., Chelysheva, L., Copenhaver, G.P., Horlow, C., and Mercier, R. (2012). FANCM limits meiotic crossovers. Science 336, 1588–1590.

Crossa, J., Campos, G. de los, Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., et al. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186, 713–724.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22, 961–975.

Crow, J.F. (1990). Mapping functions. Genetics 125, 669.

Crow, J.F., and Kimura, M. (1969). Evolution in Sexual and Asexual Populations: A Reply. Am. Nat. 103, 89–91.

Cullis, B.R., Smith, A.B., Cocks, N.A., and Butler, D.G. (2020). The Design of Early-Stage Plant Breeding Trials Using Genetic Relatedness. J. Agric. Biol. Environ. Stat. 25, 553–578.

Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLOS ONE 3, e3395.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and Woolliams, J.A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185, 1021–1031.

Daetwyler, H.D., Hayden, M.J., Spangenberg, G.C., and Hayes, B.J. (2015). Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity Relative to Genomic Selection. Genetics 200, 1341–1348.

Danguy des Déserts, A., Bouchet, S., Sourdille, P., and Servin, B. (2021). Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat. Genome Biol. Evol. 13.

Dapper, A.L., and Payseur, B.A. (2018). Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium. Mol. Biol. Evol. 35, 335–353.

Darrier, B., Rimbert, H., Balfourier, F., Pingault, L., Josselin, A.-A., Servin, B., Navarro, J., Choulet, F., Paux, E., and Sourdille, P. (2017). High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. Genetics 206, 1373–1388.

De Beukelaer, H., Badke, Y., Fack, V., and De Meyer, G. (2017). Moving Beyond Managing Realized Genomic Relationship in Long-Term Genomic Selection. Genetics 206, 1127–1138.

Dekkers, J.C.M., and Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3, 22–32.

Dekkers, J.C.M., Birke, P.V., and Gibson, J.P. (1995). Optimum linear selection indexes for multiple generation objectives with non-linear profit functions. Anim. Sci. 61, 165–175.

Demirci, S., Peters, S.A., de Ridder, D., and van Dijk, A.D.J. (2018). DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. Plant J. 95, 686–699.

Deokar, A.A., Ramsay, L., Sharpe, A.G., Diapari, M., Sindhu, A., Bett, K., Warkentin, T.D., and Tar'an, B. (2014). Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. BMC Genomics 15, 708.

Desta, Z.A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19, 592–601.

Doré, C., and Varoquaux, F. (2006). Histoire et amélioration de cinquante plantes cultivées (Editions Quae).

Dreissig, S., Mascher, M., and Heckmann, S. (2019). Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. Mol. Biol. Evol. 36, 2029–2039.

Drouaud, J., Khademian, H., Giraut, L., Zanni, V., Bellalou, S., Henderson, I.R., Falque, M., and Mézard, C. (2013). Contrasted Patterns of Crossover and Non-crossover at Arabidopsis thaliana Meiotic Recombination Hotspots. PLOS Genet. 9, e1003922.

Dudley, J.W., and Lambert, R.J. (2004). 100 generations of selection for oil and protein in corn. Plant Breed. Rev. 24, 79–110.

Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4, e1000071.

Duvick, D.N. (2005). Genetic progress in yield of United States maize (Zea mays L.). Maydica 50, 193.

Ellis, J.G., Lagudah, E.S., Spielmeyer, W., and Dodds, P.N. (2014). The past, present and future of breeding rust resistant wheat. Front. Plant Sci. 5, 641.

Esch, E., Szymaniak, J.M., Yates, H., Pawlowski, W.P., and Buckler, E.S. (2007). Using Crossover Breakpoints in Recombinant Inbred Lines to Identify Quantitative Trait Loci Controlling the Global Recombination Frequency. Genetics 177, 1851–1858.

Facolner, D.S., Mackay, T.F., and Fankham, R. (1996). Introduction to quantitative genetics (4th edn). Trends Genet. 12, 280.

Falconer, D.S., and Mackay, T.F.C. (1966). Introduction to Quantitative Genetics (Harlow, Royaume-Uni).

FAO (1998). Secondary guidelines for development of national farm animal genetic resources management plans: management of small populations at risk.

Fernandes, J.B., Wlodzimierz, P., and Henderson, I.R. (2019). Meiotic recombination within plant centromeres. Curr. Opin. Plant Biol. 48, 26–35.

Fisher, R.A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 10, 507–521.

Fisher, R.A., Sir, (1930). The genetical theory of natural selection. (Oxford: Clarendon Press).

Fradgley, N., Gardner, K.A., Cockram, J., Elderfield, J., Hickey, J.M., Howell, P., Jackson, R., and Mackay, I.J. (2019). A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLOS Biol. 17, e3000071.

France AgriMer (2015). Variétés de blé tendre, Récoltes 2015.

Fuentes, R.R., de Ridder, D., van Dijk, A.D.J., and Peters, S.A. (2021). Domestication shapes recombination patterns in tomato. Mol. Biol. Evol.

Fulton, J.E., McCarron, A.M., Lund, A.R., Pinegar, K.N., Wolc, A., Chazara, O., Bed'Hom, B., Berres, M., and Miller, M.M. (2016). A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex B region between BG2 and CD1A1. Genet. Sel. Evol. 48, 1.

Gallais, A. (1989). Optimization of recurrent selection on the phenotypic value of doubled haploid lines. Theor. Appl. Genet. 77, 501–504.

Gallais, A. (2011). Méthodes de création de variétés en amélioration des plantes (Quae).

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159, 907–911.

Gardiner, L.-J., Wingen, L.U., Bailey, P., Joynson, R., Brabbs, T., Wright, J., Higgins, J.D., Hall, N., Griffiths, S., Clavijo, B.J., et al. (2019). Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency. Genome Biol. 20, 69.

Gerton, J.L., and Hawley, R.S. (2005). Homologous chromosome interactions in meiosis: diversity amidst conservation. Nat. Rev. Genet. 6, 477–487.

Gill, B.S., Appels, R., Botha-Oberholster, A.-M., Buell, C.R., Bennetzen, J.L., Chalhoub, B., Chumley, F., Dvorák, J., Iwanaga, M., and Keller, B. (2004). A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics 168, 1087–1096.

Girard, C., Crismani, W., Froger, N., Mazel, J., Lemhemdi, A., Horlow, C., and Mercier, R. (2014). FANCM-associated proteins MHF1 and MHF2, but not the other Fanconi anemia factors, limit meiotic crossovers. Nucleic Acids Res. 42, 9087–9095.

Girard, C., Chelysheva, L., Choinard, S., Froger, N., Macaisne, N., Lehmemdi, A., Mazel, J., Crismani, W., and Mercier, R. (2015). AAA-ATPase FIDGETIN-LIKE 1 and helicase FANCM antagonize meiotic crossovers by distinct mechanisms. PLoS Genet. 11, e1005369.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. Sci. Adv. 5, eaav9188.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

Goddard, M. e., Hayes, B. j., and Meuwissen, T. h. e. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128, 409–421.

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P.S. (2017). Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection. Genetics 206, 1675.

Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning (New-York).

Gonen, S., Battagin, M., Johnston, S.E., Gorjanc, G., and Hickey, J.M. (2017). The potential of shifting recombination hotspots to increase genetic gain in livestock breeding. Genet. Sel. Evol. 49, 55.

Gorjanc, G., Gaynor, R.C., and Hickey, J.M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131, 1953–1966.

Goudet, J., Kay, T., and Weir, B.S. (2018). How to estimate kinship. Mol. Ecol. 27, 4121–4135.

Guo, Y., Zhang, G., Guo, B., Qu, C., Zhang, M., Kong, F., Zhao, Y., and Li, S. (2020). QTL mapping for quality traits using a high-density genetic map of wheat. PLOS ONE 15, e0230601.

Habier, D., Götz, K.-U., and Dempfle, L. (2007). Estimation of genetic parameters on test stations using purebred and crossbred progeny of sires of the Bavarian Piétrain. Livest. Sci. 107, 142–151.

Habier, D., Fernando, R.L., and Garrick, D.J. (2013). Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194, 597–607.

Haldane, J.B.S. (1919). THE CALCULATION OF DISTANCES BETWEEN THE LOCI OF LINKED FACTORS. J. Genet. 8, 299.

Haldane, J.B., and Waddington, C.H. (1931). Inbreeding and Linkage. Genetics 16, 357–374.

Halldorsson, B.V., Hardarson, M.T., Kehr, B., Styrkarsdottir, U., Gylfason, A., Thorleifsson, G., Zink, F., Jonasdottir, A., Jonasdottir, A., Sulem, P., et al. (2016). The rate of meiotic gene conversion varies by sex and age. Nat. Genet. 48, 1377–1384.

Hassold, T., and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. Nat. Rev. Genet. 2, 280–291.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92, 433–443.

Hayut, F.S., Melamed Bessudo, C., and Levy, A.A. (2017). Targeted recombination between homologous chromosomes for precise breeding in tomato. Nat. Commun. 8, 15605.

Hazel, L.N. (1943). The Genetic Basis for Constructing Selection Indexes. Genetics 28, 476–490.

Hazel, L.N., and Lush, J.L. (1942). THE EFFICIENCY OF THREE METHODS OF SELECTION*. J. Hered. 33, 393–399.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat. Genet. 51, 896–904.

He, S., Schulthess, A.W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J.C., and Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. Theor. Appl. Genet. 129, 641–651.

Heffner, E.L., Sorrells, M.E., and Jannink, J.-L. (2009). Genomic Selection for Crop Improvement. Crop Sci. 49, 1–12.

Heffner, E.L., Lorenz, A.J., Jannink, J.-L., and Sorrells, M.E. (2010). Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. 50, 1681–1690.

Hellsten, U., Wright, K.M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S.R., Schmutz, J., Willis, J.H., and Rokhsar, D.S. (2013). Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc. Natl. Acad. Sci. 110, 19478–19482.

Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31, 423–447.

Heslot, N., Yang, H.-P., Sorrells, M.E., and Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52, 146–160.

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., and Salamini, F. (1997). Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. Science 278, 1312–1314.

Higgins, J.D., Perry, R.M., Barakate, A., Ramsay, L., Waugh, R., Halpin, C., Armstrong, S.J., and Franklin, F.C.H. (2012). Spatiotemporal Asymmetry of the Meiotic Program Underlies the Predominantly Distal Distribution of Meiotic Crossovers in Barley. Plant Cell Online 24, 4096–4109.

Hill, W. g. (2017). “Conversion” of epistatic into additive genetic variance in finite populations and possible impact on long-term selection response. J. Anim. Breed. Genet. 134, 196–201.

Hill, W.G. (1982a). Rates of change in quantitative traits from fixation of new mutations. Proc. Natl. Acad. Sci. 79, 142–145.

Hill, W.G. (1982b). Predictions of response to artificial selection from new mutations. Genet. Res. 40, 255–278.

Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genet. Res. 8, 269–294.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature 476, 170–175.

Hofheinz, N., and Frisch, M. (2014). Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation. G3 GenesGenomesGenetics 4, 539–546.

Howard, R.S., and Lively, C.M. (1994). Parasitism, mutation accumulation and the maintenance of sex. Nature 367, 554–557.

Howe, F.S., Fischl, H., Murray, S.C., and Mellor, J. (2017). Is H3K4me3 instructive for transcription activation? BioEssays 39, e201600095.

Huguet-Robert, V., Dedryver, F., Röder, M.S., Korzun, V., Abélard, P., Tanguy, A.M., Jaudeau, B., and Jahier, J. (2001). Isolation of a chromosomally engineered durum wheat line carrying the Aegilops ventricosa Pch1 gene for resistance to eyespot. Genome 44, 345–349.

IBM (2017). IBM ILOG CPLEX 12.7 User's Manual. Int. Bus. Mach. Corp.

IWGSC (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.

IWGSC (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191.

Jahier, J., Chain, F., Barloy, D., Tanguy, A.-M., Lemoine, J., Riault, G., Margalé, E., Trottet, M., and Jacquot, E. (2009). Effect of combining two genes for partial resistance to Barley yellow dwarf virus-PAV (BYDV-PAV) derived from Thinopyrum intermedium in wheat. Plant Pathol. 58, 807–814.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9, 166–177.

Jansen, G.B., and Wilton, J.W. (1985). Selecting Mating Pairs with Linear Programming Techniques. J. Dairy Sci. 68, 1302–1305.

Jean, M., Cober, E., O'Donoughue, L., Rajcan, I., and Belzile, F. (2021). Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. Crop Sci.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D., and Jacob, H.J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Res. 14, 528–538.

Jiang, Y., and Reif, J.C. (2015). Modeling Epistasis in Genomic Selection. Genetics 201, 759–768.

Jinks, J.L., and Pooni, H.S. (1976). Predicting the properties of recombinant inbred lines derived by single seed descent. Heredity 36, 253–266.

Jordan, K.W., Wang, S., He, F., Chao, S., Lun, Y., Paux, E., Sourdille, P., Sherman, J., Akhunova, A., Blake, N.K., et al. (2018). The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. Plant J.

Juery, C., Concia, L., De Oliveira, R., Papon, N., Ramírez-González, R., Benhamed, M., Uauy, C., Choulet, F., and Paux, E. (2021). New insights into homoeologous copy number variations in the hexaploid wheat genome. Plant Genome 14, e20069.

Juliana, P., Singh, R.P., Poland, J., Mondal, S., Crossa, J., Montesinos-López, O.A., Dreisigacker, S., Pérez-Rodríguez, P., Huerta-Espino, J., Crespo-Herrera, L., et al. (2018). Prospects and Challenges of Applied Genomic Selection—A New Paradigm in Breeding for Grain Yield in Bread Wheat. Plant Genome 11, 180017.

Kanke, Y., Tubaña, B., Dalen, M., and Harrell, D. (2016). Evaluation of red and red-edge reflectance-based vegetation indices for rice biomass and grain yield prediction models in paddy fields. Precis. Agric. 17, 507–530.

Kemper, K.E., Bowman, P.J., Pryce, J.E., Hayes, B.J., and Goddard, M.E. (2012). Long-term selection strategies for complex traits using high-density genetic markers. J. Dairy Sci. 95, 4646–4656.

Kent, T.V., Uzunović, J., and Wright, S.I. (2017). Coevolution between transposable elements and recombination. Philos. Trans. R. Soc. B Biol. Sci. 372.

Kim, Y., and Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. Genetics 167, 1513.

Kinghorn, B.P., Banks, R., Gondro, C., Kremer, V.D., Meszaros, S.A., Newman, S., Shepherd, R.K., Vagg, R.D., and van der Werf, J.H.J. (2009). Strategies to Exploit Genetic Variation While Maintaining Diversity. In Adaptation and Fitness in Animal Populations: Evolutionary and Breeding Perspectives on Genetic Resource Management, J. van der Werf, H.-U. Graser, R. Frankham, and C. Gondro, eds. (Dordrecht: Springer Netherlands), pp. 191–200.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.Th., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467, 1099–1103.

Kotzamanidis, S.T., Lithourgidis, A.S., Mavromatis, A.G., Chasioti, D.I., and Roupakias, D.G. (2008). Prediction criteria of promising F3 populations in durum wheat: A comparative study. Field Crops Res. 107, 257–264.

Kuraparthy, V., Sood, S., Chhuneja, P., Dhaliwal, H.S., Kaur, S., Bowden, R.L., and Gill, B.S. (2007). A cryptic wheat–Aegilops triuncialis translocation with leaf rust resistance gene Lr58. Crop Sci. 47, 1995–2003.

Lado, B., Battenfield, S., Guzmán, C., Quincke, M., Singh, R.P., Dreisigacker, S., Peña, R.J., Fritz, A., Silva, P., Poland, J., et al. (2017). Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. Plant Genome 10.

Lam, I., and Keeney, S. (2015). Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. Science 350, 932–937.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4656–4663.

Legarra, A., Baloche, G., Barillet, F., Astruc, J.M., Soulas, C., Aguerre, X., Arrese, F., Mintegi, L., Lasarte, M., Maeztu, F., et al. (2014). Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97, 3200–3212.

Lehermeier, C., Teyssèdre, S., and Schön, C.-C. (2017). Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. Genetics 207, 1651.

Li, N., and Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics 165, 2213–2233.

Lian, L., Jacobson, A., Zhong, S., and Bernardo, R. (2015). Prediction of Genetic Variance in Biparental Maize Populations: Genomewide Marker Effects versus Mean Genetic Variance in Prior Populations. Crop Sci. 55, 1181–1188.

Lin, M., Corsi, B., Ficke, A., Tan, K.-C., Cockram, J., and Lillemo, M. (2020). Genetic mapping using a wheat multi-founder population reveals a locus on chromosome 2A controlling resistance to both leaf and glume blotch caused by the necrotrophic fungal pathogen Parastagonospora nodorum. Theor. Appl. Genet. 133, 785–808.

Lin, Z., Cogan, N.O., Pembleton, L.W., Spangenberg, G.C., Forster, J.W., Hayes, B.J., and Daetwyler, H.D. (2016). Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. Plant Genome 9, plantgenome2015.06.0046.

Liu, J., Tang, H., Qu, X., Liu, H., Li, C., Tu, Y., Li, S., Habib, A., Mu, Y., Dai, S., et al. (2020). A novel, major, and validated QTL for the effective tiller number located on chromosome arm 1BL in bread wheat. Plant Mol. Biol. 104, 173–185.

Lloyd, A., Morgan, C., H. Franklin, F.C., and Bomblies, K. (2018). Plasticity of Meiotic Recombination Rates in Response to Temperature in Arabidopsis. Genetics 208, 1409–1420.

Loidl, J. (1989). Effects of elevated temperature on meiotic chromosome synapsis in Allium ursinum. Chromosoma 97, 449–458.

Longin, C.F.H., and Reif, J.C. (2014). Redesigning the exploitation of wheat genetic resources. Trends Plant Sci. 19, 631–636.

Longin, C.F.H., Gowda, M., Mühleisen, J., Ebmeyer, E., Kazman, E., Schachschneider, R., Schacht, J., Kirchhoff, M., Zhao, Y., and Reif, J.C. (2013). Hybrid wheat: quantitative genetic parameters and consequences for the design of breeding programs. Theor. Appl. Genet. 126, 2791–2801.

Lozada, D.N., Godoy, J.V., Ward, B.P., and Carter, A.H. (2020). Genomic Prediction and Indirect Selection for Grain Yield in US Pacific Northwest Winter Wheat Using Spectral Reflectance Indices from High-Throughput Phenotyping. Int. J. Mol. Sci. 21, 165.

Lukaszewski, A.J., Kopecky, D., and Linc, G. (2012). Inversions of chromosome arms 4AL and 2BS in wheat invert the patterns of chiasma distribution. Chromosoma 121, 201–208.

Lupton, F.G.H. (1961). Studies in the breeding of self-pollinating cereals. Euphytica 10, 209–224.

Lynch, M., and Walsh, B. (1998). Genetics and analysis of quantitative traits.

Mackay, T.F.C. (2014). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat. Rev. Genet. 15, 22–33.

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of EthAcc. PLOS ONE 14, e0205629.

Marand, A.P., Jansky, S.H., Zhao, H., Leisner, C.P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A.J., Veilleux, R.E., et al. (2017). Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. Genome Biol. 18, 203.

Marand, A.P., Zhao, H., Zhang, W., Zeng, Z., Fang, C., and Jiang, J. (2019). Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. Plant Cell 31, 645.

Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., THE INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM, Jakobsen, K.S., Wulff, B.B.H., Steuernagel, B., Mayer, K.F.X., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science 345, 1250092.

McClosky, B., and Tanksley, S.D. (2013). The impact of recombination on short-term selection gain in plant breeding experiments. Theor. Appl. Genet. 126, 2299–2312.

McCouch, S., Baute, G.J., Bradeen, J., Bramel, P., Bretting, P.K., Buckler, E., Burke, J.M., Charest, D., Cloutier, S., Cole, G., et al. (2013). Feeding the future. Nature 499, 23–24.

McGaugh, S.E., Lorenz, A.J., and Flagel, L.E. (2021). The utility of genomic prediction models in evolutionary genetics. Proc. R. Soc. B Biol. Sci. 288, 20210693.

Melchinger, A.E., Gumber, R.K., Leipert, R.B., Vuylsteke, M., and Kuiper, M. (1998). Prediction of testcross means and variances among F3 progenies of F1 crosses from testcross means and genetic distances of their parents in maize. Theor. Appl. Genet. 96, 503–512.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., and Grelon, M. (2015). The Molecular Biology of Meiosis in Plants. Annu. Rev. Plant Biol. 66, 297–327.

Meunier, J., and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. 21, 984–990.

Meuwissen, T.H.E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75, 934–940.

Meuwissen, T.H., Solberg, T.R., Shepherd, R., and Woolliams, J.A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet. Sel. Evol. 41, 2.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819.

Michel, S., Ametz, C., Gungor, H., Epure, D., Grausgruber, H., Löschenberger, F., and Buerstmayr, H. (2016). Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor. Appl. Genet. 129, 1179–1189.

Mieulet, D., Aubert, G., Bres, C., Klein, A., Droc, G., Vieille, E., Rond-Coissieux, C., Sanchez, M., Dalmais, M., Mauxion, J.-P., et al. (2018). Unleashing meiotic crossovers in crops. Nat. Plants 4, 1010–1016.

Millet, E.J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S.D., Laborde, J., Bauland, C., Praud, S., Ranc, N., et al. (2016). Genome-Wide Analysis of Yield in Europe: Allelic Effects Vary with Drought and Heat Scenarios. Plant Physiol. 172, 749–764.

Misztal, I. (2008). Reliable computing in estimation of variance components. J. Anim. Breed. Genet. 125, 363–370.

Misztal, I., Aguilar, I., Lourenco, D., Ma, L., Steibel, J.P., and Toro, M. (2021). Emerging issues in genomic selection. J. Anim. Sci. 99.

Mitros, T., Lyons, J.B., Session, A.M., Jenkins, J., Shu, S., Kwon, T., Lane, M., Ng, C., Grammer, T.C., Khokha, M.K., et al. (2019). A chromosome-scale genome assembly and dense genetic map for Xenopus tropicalis. Dev. Biol. 452, 8–20.

Mohammadi, M., Tiede, T., and Smith, K.P. (2015). PopVar: A Genome-Wide Procedure for Predicting Genetic Variance and Correlated Response in Biparental Breeding Populations. Crop Sci. 55, 2068–2077.

Monaghan, J.M., Snape, J.W., Chojecki, A.J.S., and Kettlewell, P.S. (2001). The use of grain protein deviation for identifying wheat cultivars with high grain protein concentration and yield. Euphytica 122, 309–317.

Moreau, L., Charcosset, A., and Gallais, A. (2004). Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137, 111–118.

Mrode, M.R. (2005). Best linear unbiased prediction of breeding value: multivariate models. Linear Models Predict. Anim. Breed. Values 83–119.

Müller, D., Schopp, P., and Melchinger, A.E. (2018). Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection. G3 GenesGenomesGenetics 8, 1173–1181.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. 40, 1124–1129.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science 327, 876.

Nanda, G.S., Singh, P., and Gill, K.S. (1982). Epistatic, additive and dominance variation in a triple test cross of bread wheat (Triticum aestivum L.). Theor. Appl. Genet. 62, 49–52.

Neyhart, J.L., and Smith, K.P. (2019). Validating Genomewide Predictions of Genetic Variance in a Contemporary Breeding Program. Crop Sci. 59, 1062–1072.

Nordborg, M. (2000). Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. Genetics 154, 923–929.

O’Connor, T.D., Fu, W., NHLBI GO Exome Sequencing Project, ESP Population Genetics and Statistical Analysis Working Group, E.T., Mychaleckyj, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal, S.M., Smith, J.D., et al. (2015). Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans. Mol. Biol. Evol. 32, 653–660.

Olesen, J.E., Trnka, M., Kersebaum, K.C., Skjelvåg, A.O., Seguin, B., Peltonen-Sainio, P., Rossi, F., Kozyra, J., and Micale, F. (2011). Impacts and adaptation of European crop production systems to climate change. Eur. J. Agron. 34, 96–112.

Oliver, P.L., Goodstadt, L., Bayes, J.J., Birtle, Z., Roach, K.C., Phadnis, N., Beatson, S.A., Lunter, G., Malik, H.S., and Ponting, C.P. (2009). Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLOS Genet. 5, e1000753.

Osthushenrich, T. (2019). Genomic Prediction of Crossing Partners on Basis of the Expected Mean and Variance of their Derived Lines. Université Justus-Liebig de Giessen.

Otto, S.P. (2009). The Evolutionary Enigma of Sex. Am. Nat. 174, S1–S14.

Oury, F.X., Bérard, P., Brancourt-Hulmel, M., Heumez, E., Pluchard, P., Rousset, M., Doussinault, G., Rolland, B., Trottet, M., Giraud, A., et al. (2003). Yield and grain protein concentration in bread wheat: a review and a study of multi-annual data from a French breeding program [Triticum aestivum L.]. J. Genet. Breed. Italy.

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H.G., Tischfield, S.E., Zhu, X., Neale, M.J., Jasin, M., Socci, N.D., et al. (2011). A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. Cell 144, 719–731.

Peciña, A., Smith, K.N., Mézard, C., Murakami, H., Ohta, K., and Nicolas, A. (2002). Targeted Stimulation of Meiotic Recombination. Cell 111, 173–184.

Peñalba, J.V., and Wolf, J.B.W. (2020). From molecules to populations: appreciating and estimating recombination rate variation. Nat. Rev. Genet. 21, 476–492.

Perronne, R., Makowski, D., Goffaux, R., Montalent, P., and Goldringer, I. (2017). Temporal evolution of varietal, spatial and genetic diversity of bread wheat between 1980 and 2006 strongly depends upon agricultural regions in France. Agric. Ecosyst. Environ. 236, 12–20.

Petit, M., Astruc, J.-M., Sarry, J., Drouilhet, L., Fabre, S., Moreno, C.R., and Servin, B. (2017). Variation in Recombination Rate and Its Genetic Determinism in Sheep Populations. Genetics 207, 767–784.

Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., Colas, I., Oakey, H., Waugh, R., and Ramsay, L. (2015). The effect of temperature on the male and female recombination landscape of barley. New Phytol. 208, 421–429.

Poland, J.A., and Rife, T.W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome 5.

Pook, T., Schlather, M., and Simianer, H. (2020). MoBPS - Modular Breeding Program Simulator. G3 GenesGenomesGenetics 10, 1915–1918.

Pszczola, M., Strabel, T., Mulder, H.A., and Calus, M.P.L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95, 389–400.

Ptak, S.E., Hinds, D.A., Koehler, K., Nickel, B., Patil, N., Ballinger, D.G., Przeworski, M., Frazer, K.A., and Pääbo, S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. 37, 429–434.

R2D2 Consortium. (2021). Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience. Front. Genet. 12, 1185.

Ray, D.K., Ramankutty, N., Mueller, N.D., West, P.C., and Foley, J.A. (2012). Recent patterns of crop yield growth and stagnation. Nat. Commun. 3, 1293.

Reed, F.A., and Tishkoff, S.A. (2006). Positive Selection Can Create False Hotspots of Recombination. Genetics 172, 2011–2014.

Renaut, S., and Rieseberg, L.H. (2015). The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. Mol. Biol. Evol. 32, 2273–2283.

Reynolds, M., Atkin, O.K., Bennett, M., Cooper, M., Dodd, I.C., Foulkes, M.J., Frohberg, C., Hammer, G., Henderson, I.R., Huang, B., et al. (2021). Addressing Research Bottlenecks to Crop Productivity. Trends Plant Sci. 26, 607–630.

Riley, R., and Chapman, V. (1958). The production and phenotypes of wheat-rye chromosome addition lines. Heredity 12, 301–315.

Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., Duarte, J., Rivière, N., Eversole, K., on behalf of The International Wheat Genome Sequencing Consortium, et al. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. PLOS ONE 13, e0186329.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V.M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., et al. (2012). Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics 192, 715–728.

Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., and Segura, V. (2018). Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. G3 GenesGenomesGenetics 8, 3961–3972.

Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y., and Buckler, E.S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. 112, 3823–3828.

Roth, S.Y., Denu, J.M., and Allis, C.D. (2001). Histone acetyltransferases. Annu. Rev. Biochem. 70, 81–120.

Rowan, B.A., Heavens, D., Feuerborn, T.R., Tock, A.J., Henderson, I.R., and Weigel, D. (2019). An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. Genetics 213, 771.

Ru, S., and Bernardo, R. (2019). Targeted recombination to increase genetic gain in self-pollinated species. Theor. Appl. Genet. 132, 289–300.

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L.G., Crossa, J., Reynolds, M., and Singh, R. (2016). Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. G3 GenesGenomesGenetics 6, 2799–2808.

Saintenac, C., Faure, S., Remay, A., Choulet, F., Ravel, C., Paux, E., Balfourier, F., Feuillet, C., and Sourdille, P. (2011). Variation in crossover rates across a 3-Mb contig of bread wheat (Triticum aestivum) reveals the presence of a meiotic recombination hotspot. Chromosoma 120, 185–198.

Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin 8, 3.

Salomé, P.A., Bomblies, K., Fitz, J., Laitinen, R.A.E., Warthmann, N., Yant, L., and Weigel, D. (2011). The recombination landscape in Arabidopsis thaliana F2 populations. Heredity 108, 447.

Sánchez, L., Caballero, A., and Santiago, E. (2006). Palliating the impact of fixation of a major gene on the genetic variation of artificially selected polygenes. Genet. Res. 88, 105–118.

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., and Georges, M. (2012). Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. PLOS Genet. 8, e1002854.

Santiago, E., and Caballero, A. (1998). Effective Size and Polymorphism of Linked Neutral Loci in Populations Under Directional Selection. Genetics 149, 2105–2117.

Santos, D.J.A., Cole, J.B., Lawlor, T.J., Jr., VanRaden, P.M., Tonhati, H., and Ma, L. (2019). Variance of gametic diversity and its application in selection programs. J. Dairy Sci. 102, 5279–5294.

Sarno, R., Vicq, Y., Uematsu, N., Luka, M., Lapierre, C., Carroll, D., Bastianelli, G., Serero, A., and Nicolas, A. (2017). Programming sites of meiotic crossovers using Spo11 fusion proteins. Nucleic Acids Res. 45, e164–e164.

Schnell, F.W., and Utz, H.F. (1975). F1 Leistung und Elternwahl in der Zuchtung von Selbstbefruchtern. Ber Arbeitstag Arbeitsgem Saatzuchtleiter.

Schwarzkopf, E.J., Motamayor, J.C., and Cornejo, O.E. (2020). Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. BMC Genomics 21, 1–16.

Sears, E.R. (1958). Intergenomic chromosome relationships in hexaploid wheat. In Proc. Int. Congress Genet., pp. 258–259.

Segelke, D., Reinhardt, F., Liu, Z., and Thaller, G. (2014). Prediction of expected genetic variation within groups of offspring for innovative mating schemes. Genet. Sel. Evol. 46, 42.

Séguéla-Arnaud, M., Crismani, W., Larchevêque, C., Mazel, J., Froger, N., Choinard, S., Lemhemdi, A., Macaisne, N., Van Leene, J., and Gevaert, K. (2015). Multiple mechanisms limit meiotic crossovers: TOP3α and two BLM homologs antagonize crossovers in parallel to FANCM. Proc. Natl. Acad. Sci. 112, 4713–4718.

Serra, H., Choi, K., Zhao, X., Blackwell, A.R., Kim, J., and Henderson, I.R. (2018). Interhomolog polymorphism shapes meiotic crossover within the Arabidopsis RAC1 and RPP13 disease resistance genes. PLOS Genet. 14, e1007843.

Shen, C., Wang, N., Huang, C., Wang, M., Zhang, X., and Lin, Z. (2019). Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. Plant J. 0.

Shetty, A.C., Consortium, N.T.-O. for P.M. (TOPMed), Group, Topm.P.G.W., O'Connell, J., Mitchell, B.D., and O'Connor, T.D. (2020). Rare variant enriched identity-by-descent enables the detection of distant relatedness and older divergence between populations.

Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N., and Levy, A.A. (2015). DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. Plant Cell 27, 2427–2436.

Simmonds, N.W. (1995). The relation between yield and protein in cereal grain. J. Sci. Food Agric. 67, 309–315.

Singh, S. (1978). Intermating in early segregating generation and characterisation of genetic parameters in self pollinated crops. J Indian Soc Agric Stat 30, 159.

Singh, G., Bhullar, G.S., and Gill, K.S. (1986). Genetic control of grain yield and its related traits in bread wheat. Theor. Appl. Genet. 72, 536–540.

Singhal, S., Leffler, E.M., Sannareddy, K., Turner, I., Venn, O., Hooper, D.M., Strand, A.I., Li, Q., Raney, B., Balakrishnan, C.N., et al. (2015). Stable recombination hotspots in birds. Science 350, 928–932.

Smith, J.M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23–35.

Sonesson, A.K., Woolliams, J.A., and Meuwissen, T.H. (2012). Genomic selection requires genomic control of inbreeding. Genet. Sel. Evol. 44, 1–10.

Souza, E., and Sorrells, M.E. (1991). Prediction of progeny variation in oat from parental genetic relationship. Predict. Progeny Var. Oat Parent. Genet. Relatsh. 82, 233–241.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved Heritability Estimation from Genome-wide SNPs. Am. J. Hum. Genet. 91, 1011–1021.

Stapley, J., Feulner, P.G.D., Johnston, S.E., Santure, A.W., and Smadja, C.M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos. Trans. R. Soc. B Biol. Sci. 372.

Steffenson, B.J., Olivera, P., Roy, J.K., Jin, Y., Smith, K.P., and Muehlbauer, G.J. (2007). A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. Aust. J. Agric. Res. 58, 532–544.

Stumpf, M.P.H., and McVean, G.A.T. (2003). Estimating recombination rates from population-genetic data. Nat. Rev. Genet. 4, 959–968.

Sun, J., Rutkoski, J.E., Poland, J.A., Crossa, J., Jannink, J.-L., and Sorrells, M.E. (2017). Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. Plant Genome 10, plantgenome2016.11.0111.

Taagen, E., Bogdanove, A.J., and Sorrells, M.E. (2020). Counting on Crossovers: Controlled Recombination for Plant Breeding. Trends Plant Sci. 25, 455–465.

Tadesse, W., Sanchez-Garcia, M., Assefa, S.G., Amri, A., Bishaw, Z., Ogbonnaya, F.C., and Baum, M. (2019). Genetic gains in wheat breeding and its role in feeding the world. Crop Breed Genet Genom 1, e190005.

Tiede, T., Kumar, L., Mohammadi, M., and Smith, K.P. (2015). Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. Mol. Breed. 35, 199.

Tiemann-Boege, I., Schwarz, T., Striedner, Y., and Heissl, A. (2017). The consequences of sequence erosion in the evolution of recombination hotspots. Philos. Trans. R. Soc. B Biol. Sci. 372.

Tiret, M., and Hospital, F. (2017). Blocks of chromosomes identical by descent in a population: Models and predictions. PLOS ONE 12, e0187416.

Tock, A.J., Holland, D.M., Jiang, W., Osman, K., Sanchez-Moran, E., Higgins, J.D., Edwards, K.J., Uauy, C., Franklin, F.C.H., and Henderson, I.R. (2021). Crossover-active regions of the wheat genome are distinguished by DMC1, the chromosome axis, H3K27me3, and signatures of adaptation. Genome Res. 31, 1614–1628.

Toro, M., and Perez-Enciso, M. (1990). Optimization of selection response under restricted inbreeding. Genet. Sel. Evol. 22, 93–107.

Toro, M.A., Nieto, B., and Salgado, C. (1988). A note on minimization of inbreeding in small-scale selection programmes. Livest. Prod. Sci. 20, 317–323.

Tourrette, E., Bernardo, R., Falque, M., and Martin, O.C. (2019). Assessing by Modeling the Consequences of Increased Recombination in Recurrent Selection of Oryza sativa and Brassica rapa. G3 GenesGenomesGenetics 9, 4169–4181.

Uauy, C., Brevis, J.C., and Dubcovsky, J. (2006). The high grain protein content gene Gpc-B1 accelerates senescence and has pleiotropic effects on protein content in wheat. J. Exp. Bot. 57, 2785–2794.

Underwood, C.J., Choi, K., Lambing, C., Zhao, X., Serra, H., Borges, F., Simorowski, J., Ernst, E., Jacob, Y., Henderson, I.R., et al. (2018). Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. Genome Res. 28, 519–531.

Utz, H.F., Bohn, M., and Melchinger, A.E. (2001). Predicting Progeny Means and Variances of Winter Wheat Crosses from Phenotypic Values of Their Parents. Crop Sci. 41, 1470–1478.

Vaissayre, L., Ardisson, M., Borries, C., Santoni, S., David, J., and Roumet, P. (2012). Elite durum wheat genetic map and recombination rate variation in a multiparental connected design. Euphytica 185, 61–75.

Van Raden, P.M., Freeman, A.E., and Rothschild, M.F. (1984). Maximizing Genetic Gain under Multiple-Stage Selection1. J. Dairy Sci. 67, 1761–1766.

VanRaden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414–4423.

Varona, L., Legarra, A., Toro, M.A., and Vitezica, Z.G. (2018). Non-additive Effects in Genomic Selection. Front. Genet. 9, 78.

Wang, C., Hu, S., Gardner, C., and Lübberstedt, T. (2017). Emerging Avenues for Utilization of Exotic Germplasm. Trends Plant Sci. 22, 624–637.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W.M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94, 73–83.

Wang, J., Street, N.R., Scofield, D.G., and Ingvarsson, P.K. (2016). Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species. Genetics 202, 1185.

Wang, W., Vinocur, B., and Altman, A. (2003). Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta 218, 1–14.

Wang, Z., Du, S., Dayanandan, S., Wang, D., Zeng, Y., and Zhang, J. (2014). Phylogeny Reconstruction and Hybrid Analysis of Populus (Salicaceae) Based on Nucleotide Sequences of Multiple Single-Copy Nuclear Genes and Plastid Fragments. PLOS ONE 9, e103645.

Wartha, C.A., and Lorenz, A.J. (2021). Implementation of genomic selection in public sector plant breeding programs: Current status and opportunities. Crop Breed. Appl. Biotechnol. 21, e394621S15.

Weigel, K.A., and Lin, S.W. (2000). Use of Computerized Mate Selection Programs to Control Inbreeding of Holstein and Jersey Cattle in the Next Generation. J. Dairy Sci. 83, 822–828.

Wellmann, R. (2019). Optimum contribution selection for animal breeding and conservation: the R package optiSel. BMC Bioinformatics 20, 25.

Whittaker, J.C., Thompson, R., and Denham, M.C. (2000). Marker-assisted selection using ridge regression. Genet. Res. 75, 249–252.

Williams, J.S. (1962). The Evaluation of a Selection Index. Biometrics 18, 375–393.

Wimmer, V., Albrecht, T., Auinger, H.-J., and Schön, C.-C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28, 2086–2087.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., and Schön, C.-C. (2013). Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics 195, 573–587.

Wingen, L.U., West, C., Leverington-Waite, M., Collier, S., Orford, S., Goram, R., Yang, C.-Y., King, J., Allen, A.M., Burridge, A., et al. (2017). Wheat Landrace Genome Diversity. Genetics 205, 1657–1676.

Wolfe, M.D., Chan, A.W., Kulakow, P., Rabbi, I., and Jannink, J.-L. (2021). Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices. 2021.01.05.425443.

Woolliams, J.A., and Meuwissen, T.H.E. (1993). Decision rules and variance of response in breeding schemes. Anim. Sci. 56, 179–186.

Woolliams, J. a., Berg, P., Dagnachew, B. s., and Meuwissen, T. h. e. (2015). Genetic contributions and their optimization. J. Anim. Breed. Genet. 132, 89–99.

Woolliams, J.A., Wray, N.R., and Thompson, R. (1993). Prediction of long-term contributions and inbreeding in populations undergoing mass selection. Genet. Res. 62, 231–242.

Wray, N.R., and Goddard, M.E. (1994). MOET breeding schemes for wool sheep 1. Design alternatives. Anim. Sci. 59, 71–86.

Wray, N.R., and Thompson, R. (1990). Prediction of rates of inbreeding in selected populations. Genet. Res. 55, 41–54.

Yang, C.J., Sharma, R., Gorjanc, G., Hearne, S., Powell, W., and Mackay, I. (2020). Origin Specific Genomic Selection: A Simple Process To Optimize the Favorable Contribution of Parents to Progeny. G3 GenesGenomesGenetics 10, 2445–2455.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569.

Yao, J., Zhao, D., Chen, X., Zhang, Y., and Wang, J. (2018). Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (Triticum aestivum L.). Crop J. 6, 353–365.

Yelina, N.E., Lambing, C., Hardcastle, T.J., Zhao, X., Santos, B., and Henderson, I.R. (2015). DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. Genes Dev. 29, 2183–2202.

Yelina, N.E., Gonzalez-Jorge, S., Hirsz, D., Yang, Z., and Henderson, I.R. (2021). CRISPR targeting of MEIOTIC-TOPOISOMERASE VIB-dCas9 to a recombination hotspot is insufficient to increase crossover frequency in Arabidopsis.

Yin, X., and Germay, Noël. (1993). A Fast Genetic Algorithm with Sharing Scheme Using Cluster Analysis Methods in Multimodal Function Optimization. In Artificial Neural Nets and Genetic Algorithms, R.F. Albrecht, C.R. Reeves, and N.C. Steele, eds. (Vienna: Springer), pp. 450–457.

Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants. 2(10):1–7.

Zhang, L., and Ma, H. (2012). Complex evolutionary history and diverse domain organization of SET proteins suggest divergent regulatory interactions. New Phytol. 195, 248–263.

Zhao, H., and Speed, T.P. (1996). On genetic map functions. Genetics 142, 1369–1377.

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A.P., Wu, Y., and Jiang, J. (2018). Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. Plant Physiol. 176, 2789–2803.

Zhong, S., and Jannink, J.-L. (2007). Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. Genetics 177, 567.

Zhong, S., Dekkers, J.C.M., Fernando, R.L., and Jannink, J.-L. (2009). Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182, 355–364.

Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y., et al. (2020). Triticum population sequencing provides insights into wheat adaptation. Nat. Genet. 52, 1412–1422.

Résumé

Titre : Apport de la recombinaison dans l'optimisation des plans de croisements de blé tendre

Le choix du plan de croisements représente un enjeu crucial en sélection afin d'assurer le progrès variétal à court et long terme. La distribution des valeurs génétiques des descendants d'un croisement tient compte de la complémentarité allélique des parents et de la probabilité de cumuler ces allèles dans les descendants par recombinaison. L'objectif de cette thèse est double : d'une part, étudier la variabilité du profil de recombinaison entre groupes génétiques divergents de blé tendre, et d'autre part, évaluer l'apport d'une prise en compte de la recombinaison dans l'optimisation des plans de croisements dans le cadre d'un programme de sélection de blé tendre d'hiver. Dans un premier temps, cette thèse a permis d'estimer le profil de recombinaison des quatre principaux groupes génétiques du blé tendre, à partir de leurs patrons de déséquilibre de liaison. La corrélation des taux de recombinaison diminue avec la différenciation génétique des groupes génétiques. Ce résultat suggère que les facteurs qui déterminent la position des crossing-over chez le blé tendre sont soumis à des forces évolutives. Dans un deuxième temps, les taux de recombinaison précédemment estimés et un modèle de prédiction génomique ont été utilisés pour prédire la valeur des croisements de blé tendre dans le cadre d'un programme de sélection simulé. Plusieurs critères de sélection des croisements décrits dans la littérature ont été comparés. Un nouveau critère a été développé qui trie les croisements d'après la proportion de descendants dont la valeur génétique est supérieure à un seuil. L'optimisation des plans de croisements a consisté à répartir le nombre de descendants total entre les meilleurs croisements, en prenant en compte des contraintes empiriques sur les contributions parentales, pour limiter la perte de diversité génétique. Par rapport à une simple sélection des croisements sur la valeur génétique des parents, les critères basés la recombinaison permettent un gain génétique plus important tout en limitant la perte de diversité génétique. La principale limite est le manque de précision de l'estimation des effets des marqueurs par la prédiction génomique. D'un point de vue plus appliqué, cette thèse a permis le développement d'un outil d'optimisation des plans de croisements. Cet outil performant en termes de rapidité de calcul sera enrichi et pourra aider les sélectionneurs à améliorer la conversion de la diversité génétique en progrès génétique.

Mots clés : blé tendre, recombinaison, croisements, prédiction génomique

Title: Optimization of mating plans in bread wheat using recombination rate information

The selection of the mating plan design is crucial decision in plant breeding programs to ensure short and long-term genetic gain. The distribution of progenies breeding values of a cross depends on the allelic complementarity of parents, the distribution of favorable alleles in parents and recombination rate between QTLs. The objectives of the thesis were: first to estimate recombination rates along the genome and analyze its variability in divergent genetic groups of bread wheat, then to evaluate the interest of recombination information in mating design optimization in a winter bread wheat breeding program. The first objective was to estimate the recombination profile of the four main genetic groups of bread wheat, from linkage disequilibrium patterns. The correlation of recombination rates decreases with genetic differentiation among groups. This result suggests that factors that control crossing-over positions are under evolution pressures. The second objective was to estimate cross values and optimize mating plans in a simulated breeding programs using those recombination profiles and a genomic prediction model to estimate marker effects. Several cross selection criteria described in literature were compared. A new criterion was proposed that allow to rank crosses according to the proportion of progeny superior to a fixed threshold. The optimization of mating plans consisted in attributing the total number of progenies to best crosses, considering empirical constraints on parental contributions to limit loss of genetic diversity. Compared to the selection of mating plan based on parental genetic values only, the criteria based on recombination information allow higher genetic gain while limiting the loss of genetic diversity. The main limitation is the lack of accuracy in marker effect estimation. In a practical point of view, this thesis allowed to develop a tool that optimize mating designs. This tool was optimized in terms of computation time and will be extended. It could help breeders to improve the conversion of genetic diversity into genetic gain.

Key words: bread wheat, recombination, crosses, genomic prediction