Décalage génétique de population face aux changements environnementaux

Clément Gain

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : Translational Innovation in Medicine and Complexity

Décalage génétique de population face aux changements environnementaux

Population genetic offset in the face of climate change.

Présentée par :

Clément GAIN

Direction de thèse :

Olivier FRANCOIS Professeur des Universités, Grenoble INP	Directeur de thèse
Flora JAY Chargé de recherche, CNRS Délégation Ile-de-France Sud	Co-encadrante de thèse

Rapporteurs :

MATHIEU GAUTIER
Directeur de recherche, INRAE CENTRE OCCITANIE-MONTPELLIER
IVAN SCOTTI
Directeur de recherche, INRAE CENTRE PROVENCE-ALPES-COTE D'AZUR

Thèse soutenue publiquement le **10 mai 2023**, devant le jury composé de :

OLIVIER FRANÇOIS Professeur des Universités, GRENOBLE INP	Directeur de thèse
LAURENCE DESPRES Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Présidente
MATHIEU GAUTIER Directeur de recherche, INRAE CENTRE OCCITANIE-MONTPELLIER	Rapporteur
IVAN SCOTTI Directeur de recherche, INRAE CENTRE PROVENCE-ALPES-COTE D'AZUR	Rapporteur
SIMON BOITARD Chargé de recherche HDR, INRAE CENTRE OCCITANIE-MONTPELLIER	Examinateur

Invités :

THIBAUT CAPBLANCQ
Docteur en Sciences, UNIVERSITE GRENOBLE ALPES
FLORA JAY
Chargé de recherche, CNRS DELEGATION ILE-DE-FRANCE SUD

Remerciements

Je tiens tout d'abord à remercier sincèrement mon directeur de thèse Olivier François. Merci pour ta volonté de transmettre ton savoir et de partager tes idées. Merci pour la patience et la bienveillance dont tu as fait preuve durant ces trois années. Merci pour ta rigueur et ton exigence, indispensable dans ce milieu et merci d'avoir été cette porte d'entrée dans le monde de la recherche. Je veux aussi remercier ma co-directrice de thèse, Flora Jay. Merci d'avoir su occuper une place complémentaire dans l'encadrement de ma thèse et ceux, malgré les modifications par rapport au sujet initial.

Je tiens également à remercier toutes les personnes qui ont croisé ma route durant cette thèse et qui m'ont aidé, d'une manière ou d'une autre, à mener à bien mon travail. Je pense bien sûr en premier lieu à mon équipe, MAGE, avec laquelle j'ai pris plaisir à partager des moments tant conviviaux que cérébraux. Je veux dire merci à tous les personnels du laboratoire dont le travail est absolument indispensable au bon fonctionnement de notre recherche et dont on oublie souvent de souligner l'importance. Je pense par exemple à Olivier Pedano du service informatique et à notre gestionnaire financière Stéphanie Imbert mais je pourrai en citer tant d'autres. Je veux aussi remercier toutes les personnes avec qui j'ai pu échanger autour de sujets scientifiques divers et variés durant des évènements et au cours de mes réunions avec mon comité de suivi individuel. Merci pour la passion et la curiosité que vous mettez dans votre travail.

Je tiens pour finir à remercier tout mon entourage, ma famille et mes amis, qui me permettent de vivre, en parallèle de cette aventure scientifique, une belle aventure humaine.

Résumé

Le changement climatique mondial modifie les habitats à un rythme sans précédent. Ces changements environnementaux ont un impact important sur la biodiversité et il y a un intérêt grandissant dans la compréhension de la réponse des populations à ces changements environnementaux. Cette thèse porte sur l'utilisation de données génomiques intraspécifiques afin d'informer sur la prédiction de ces réponses. Plus précisément, nous apportons notre contribution à la notion de décalage génétique. Le décalage génétique cherche à quantifier la maladaptation génétique des populations. On parle de maladaptation génétique lorsque la composition génétique d'une population ne correspond pas à celle requise pour l'habitat dans lequel elle évolue. Notre travail de thèse se concentre sur différents axes. On présentera tout d'abord une nouvelle mesure de décalage génétique, appelée fossé génétique, visant à résoudre certaines limites des méthodes existantes, tel que la prise en compte des facteurs de confusion et de l'aspect polygénique de l'adaptation. On établira également un cadre théorique permettant la mise en place d'une relation entre le décalage génétique et la valeur sélective d'un individu dans un environnement modifié. Plus précisément, nous montrerons que le fossé génétique est proportionnel au logarithme de la valeur sélective dans l'environnement modifié. Nous validerons ce résultat théorique sur des données simulées à l'aide du logiciel SLiM, et sur des données réelles à l'aide d'une expérience de jardin commun pour des populations de mil (*Pennisetum glaucum*). En parallèle de ces travaux sur le décalage génétique, nous établirons une relation théorique entre l'analyse en composantes principales (ACP) et l'indice de fixation de Wright, deux approches essentielles dans la compréhension de la structure de population existant chez des individus échantillonnés. Cette relation nous dit que dans un modèle à K populations discrètes, la valeur de F_{ST} moyenne le long du génome est approchée par les $(K-1)$ valeurs propres de l'ACP standardisée. Notre thèse contribue donc à une meilleure interprétation du décalage génétique par la mise en place d'un cadre théorique autour de cette notion et facilite également son utilisation par l'implémentation d'une fonction de calcul du fossé génétique dans la librairie R `LEA 3`. Elle contribue également à justifier l'utilisation de l'ACP pour décrire la structure génétique des populations en précisant le lien existant entre cette méthode et l'indice de fixation de Wright.

Recommandations de lecture

Ce manuscrit se divise en six chapitres. Le premier chapitre est un chapitre d'introduction qui vise à présenter les questions auxquelles nous cherchons à répondre dans cette thèse. Il donne le contexte minimal pour la compréhension de ces questions et passe volontairement rapidement sur certaines notions qui seront approfondies dans le chapitre 2 d'état de l'art. Le chapitre 2 donne le contexte historique dans lequel s'inscrivent les questions de cette thèse et présente les notions essentielles à la compréhension de nos travaux. Les trois chapitres suivants forment le contenu de la thèse. Les chapitres 3 et 4 constituent le coeur de notre travail sur le décalage génétique et il est recommandé de les lire dans l'ordre car le chapitre 4 s'appuie sur des notions présentées dans le chapitre 3. Le chapitre 5 présente les travaux autour de la relation entre les valeurs propres de l'analyse en composantes principales (ACP) et la F_{ST} et peut être lu de manière indépendante. Ces trois chapitres commencent par une introduction qui reprend certains des éléments de l'état de l'art afin de situer les apports du chapitre par rapport à la littérature scientifique existante. Il n'est donc pas indispensable de connaître en détail le contenu de l'état de l'art pour lire l'un des chapitre de contenu mais l'état de l'art fournit toutefois une introduction plus complète. Enfin, le manuscrit se termine par un chapitre de conclusion qui rappelle les grands résultats de nos travaux et offre des perspectives de travaux futurs pour prolonger le contenu présenté dans cette thèse. Bonne lecture !

Table des matières

Chapitre 1

Introduction

Le changement climatique mondial modifie les habitats à un rythme sans précédent (Rellstab, 2021). Ces changements environnementaux ont un impact important sur la biodiversité et il y a un intérêt grandissant dans la compréhension de la réponse des populations à ces changements environnementaux (Foden et al., 2019). Cette thèse porte sur l'utilisation de données génomiques intraspécifiques afin d'informer sur la prédiction de ces réponses. Plus précisément, nous apportons notre contribution à la notion de décalage génétique. Le décalage génétique cherche à quantifier la maladaptation génétique des populations (Rellstab, Dauphin et al., 2021) (Capblancq, Fitzpatrick et al., 2020). On parle de maladaptation génétique lorsque la composition génétique d'une population ne correspond pas à celle requise pour l'habitat dans lequel elle évolue. Dans ce court chapitre introductif, nous commençons par introduire la problématique des impacts du changement climatique sur la biodiversité en définissant certaines notions de base. Nous définissons ensuite le concept du décalage génétique et établissons les problématiques auxquelles cette thèse cherche à répondre. Cette introduction vise à donner le contexte minimum pour énoncer les questions posées par notre thèse, le contexte historique de l'apparition de la notion de décalage génétique ainsi que l'approfondissement de certaines notions essentielles à la compréhension de nos travaux sont établis dans le chapitre 2 d'état de l'art.

1. Biodiversité et changements environnementaux

Le sujet de thèse s'inscrit dans le sujet plus large de la compréhension des impacts du changement climatique sur la biodiversité. Il convient tout d'abord de définir certaines de ces notions essentielles et de rendre compte des liens qui existent entre changement climatique et perte de biodiversité.

1. 1. Biodiversité

La biodiversité désigne la variété des formes de vie sur Terre. Le mot correspond à la contraction du terme "diversité biologique" et on situe son apparition à l'année 1985 (Z. X. Chen et al., 2004) On distingue en général trois niveaux différents de biodiversité :

- **Diversité génétique** : L'ensemble des gènes contenus dans le monde du vivant
- **Diversité d'espèce** : L'ensemble des différentes espèces, les différences au sein et entre les espèces.
- **Diversité d'écosystème** : L'ensemble des habitats, des communautés écologiques, et des processus écologiques.

La notion d'espèce On retrouve un grand nombre de manières différentes de définir la notion d'espèce dans la littérature scientifique (Mayden, 1997). Il est en effet très difficile de trouver une définition qui puisse s'appliquer à tous les organismes. Le débat autour de cette question a été surnommé le "problème des espèces" (Queiroz, 2005). Darwin lui même avait mentionné cette problématique dans son oeuvre "On the origin of species" :
"*No one definition has satisfied all naturalists ; yet every naturalist knows vaguely what he means when he speaks of a species. Generally the term includes the unknown element of a distinct act of creation*"
Dans le cadre de ce projet de thèse, nous pourrons nous servir de la définition d'Ernst Mayr (Mayr, 1942) qui se sert du concept d'espèce biologique. L'espèce est alors définie comme un groupe de populations naturelles capables de se reproduire et qui est isolé sur le plan de la reproduction des autres groupes.

La notion de population Tout comme la notion d'espèce, il n'existe pas de définition unique autour du concept de population en écologie. Dans ce projet de thèse, nous définissons la population comme étant le sous-ensemble de l'ensemble des individus d'une espèce qui occupe une certaine zone géographique dans le monde.

La notion de valeur sélective (fitness) Il convient également d'aborder le concept de valeur sélective d'un individu. Ce concept, appelé fitness en anglais, peut être défini de nombreuses façons. Dans son sens technique, on peut le voir comme une mesure relative ou absolue de l'efficacité ou du succès reproductif (Krimbas, 2004). Elle peut par exemple être mesurée par la proportion de ses descendants qui atteignent la maturité sexuelle.

1. 2. Climat et changements environnementaux

Climat Dans le cadre de ce projet de thèse, nous nous intéressons aux impact des variables climatiques sur la biodiversité. Il est donc important de commencer par définir ce qu'est le climat. Nous choisissons pour cela la définition donnée par le groupe d'experts intergouvernemental sur l'évolution du climat :

"*Climate in a narrow sense is usually defined as the "average weather," or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period ranging from months to thousands or millions of years. The classical period is 30 years, as defined by the World Meteorological Organization (WMO). These quantities are most often surface variables such as temperature, precipitation, and wind. Climate in a wider sense is the state, including a statistical description, of the climate system.*" (Giec, 2017)

Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) Le GIEC est un organisme intergouvernemental chargé d'évaluer la réalité, les causes et les conséquences du changement climatique en cours. Les évaluations du GIEC sont principalement fondées sur les publications scientifiques et techniques. Elles sont publiées chaque année sous la forme de rapports synthétiques ou dédiés à un aspect particulier du changement climatique.

Changement climatique Le terme de changement climatique fait référence à des variations du climat qui persistent sur des périodes de plusieurs décennies au moins. La planète Terre a connu plusieurs périodes de changements climatiques, mais elle traverse aujourd'hui une augmentation moyenne des températures qui est plus rapide et dont la cause principale est l'Homme. La période de réchauffement actuelle s'explique par des émissions importantes de gaz à effet de serre tel que le dioxyde de carbone et le méthane dû aux activités humaines. (Giec, 2017)

Modèles d'évolutions climatiques Afin d'évaluer comment pourrait évoluer le climat, on cherche à estimer l'évolution de la concentration en dioxyde de carbone dans l'atmosphère. Pour effectuer cette estimation, on se base sur des scénarios d'évolutions socio-économiques mondiales, Shared Socioeconomic Pathways (SSPs) en anglais, projetés jusqu'en 2100. Le GIEC a ainsi distingué 5 scénarios par ordre croissant d'émissions, de SSP1, dit de durabilité "*Taking the Green Road*", à SSP5, dit de développement alimenté par les combustibles fossibles "*Taking the highway*" (Tableau 1.1).
Dans un commentaire publié dans Nature (Hausfather et Peters, 2021), le scénario SSP5 est décrit comme hautement improbable, le SSP3 comme improbable et le SSP2 comme probable. Toutefois, un rapport de PNAS (Schwalm et al., 2012) montre que le SSP5 est la meilleure correspondance avec les émissions cumulées de 2005 à 2020.
Ces scénarios nous fournissent donc des prédictions sur l'évolution du climat. Ces changements importants ont un impact sur la biodiversité. Intéressons nous désormais à ces impacts.

1. 3. Biodiversité et changements environnementaux

Perte de biodiversité On parle d'une période d'extinction massive de la biodiversité lors d'une période où la Terre perd plus des trois quarts de ces espèces dans un intervalle de temps géologiquement court. On distingue cinq périodes d'extinctions massives au cours des 540 millions d'années passées. Au vu des pertes d'espèces connues au cours des derniers siècles, on parle aujourd'hui de la sixième extinction de masse. (Barnosky et al., 2011) Voici quelques chiffres pour prendre la mesure de cette extinction :

- Parmi un ensemble de 177 espèces de mammifères étudiées, 40% ont subi des pertes de population supérieures à 80% (Ceballos et al., 2017)
- Entre 1989 et 2016, on observe une diminution de 76% du nombre d'insectes volants (Hallmann et al., 2017)
- 41% des espèces d'insectes sont en déclin (Sánchez-Bayo et Wyckhuys, 2019)

	Near term, 2021–2040		Mid-term, 2041–2060		Long term, 2081–2100	
Scenario	Best estimate (°C)	*Very likely* range (°C)	Best estimate (°C)	*Very likely* range (°C)	Best estimate (°C)	*Very likely* range (°C)
SSP1-1.9	1.5	1.2 to 1.7	1.6	1.2 to 2.0	1.4	1.0 to 1.8
SSP1-2.6	1.5	1.2 to 1.8	1.7	1.3 to 2.2	1.8	1.3 to 2.4
SSP2-4.5	1.5	1.2 to 1.8	2.0	1.6 to 2.5	2.7	2.1 to 3.5
SSP3-7.0	1.5	1.2 to 1.8	2.1	1.7 to 2.6	3.6	2.8 to 4.6
SSP5-8.5	1.6	1.3 to 1.9	2.4	1.9 to 3.0	4.4	3.3 to 5.7

Table 1.1 – Changements de la température de surface de la Terre, pour des périodes de 20 ans et pour les 5 scénarios d'émisssions envisagés (Giec, 2017)

Causes de la perte de biodiversité La perte de biodiversité est dû à l'activité humaine et on la qualifie donc d'extinction de l'holocène (Dirzo et al., 2014). On distingue 5 causes majeures à cette extinction massive (SCDB, 2010) (Figure 1.1) :

- **Modifications des habitats** : Destruction, fragmentation, artificialisation, déforestation, pollution lumineuse ...
- **Surexploitation de la biodiversité** : Taux de prélèvement des ressources qui dépasse le taux de renouvellement. La surpêche est un exemple.
- **Pollutions** provoquées par la révolution industrielle.
- **Espèces exotiques envahissantes** : Les introductions d'espèces invasives sont causées notamment par le commerce international.
- **Changements climatiques** : provoqués par des émissions importantes de gaz à effet de serre liées à l'activité humaines.

Dans le cadre de ce projet de thèse, nous nous intéresserons plus particulièrement au lien entre la perte de biodiversité et le changement climatique.

Perte de biodiversité, changement climatique et variables d'intérêt Le changement climatique a un impact sur plusieurs variables environnementales parmi lesquelles on peut citer la température, les précipitations, la concentration en CO_2. Au delà des variations de la moyenne, c'est également les évènements climatiques dits extrêmes qui vont être de plus en plus fréquents. Du fait de ces changements, les risques d'extinction pourraient s'accélérer avec les futures températures mondiales, menaçant jusqu'à une espèce sur six dans le cadre des politiques actuelles (Urban, 2015).
La prise de conscience des impacts du changement climatique sur la biodiversité a fait naître tout un domaine de recherche dont le but est l'évaluation de la vulnérabilité des espèces au changement climatique (Foden et al., 2019). Cette évaluation est une condition préalable à l'élaboration de stratégies efficaces pour les conserver. Notre thèse porte sur l'une de ses méthodes, cherchant à prédire une mesure génomique de la maladaptation, appelée décalage génétique (Rellstab, Dauphin et al., 2021) (Fitzpatrick et Keller, 2015).

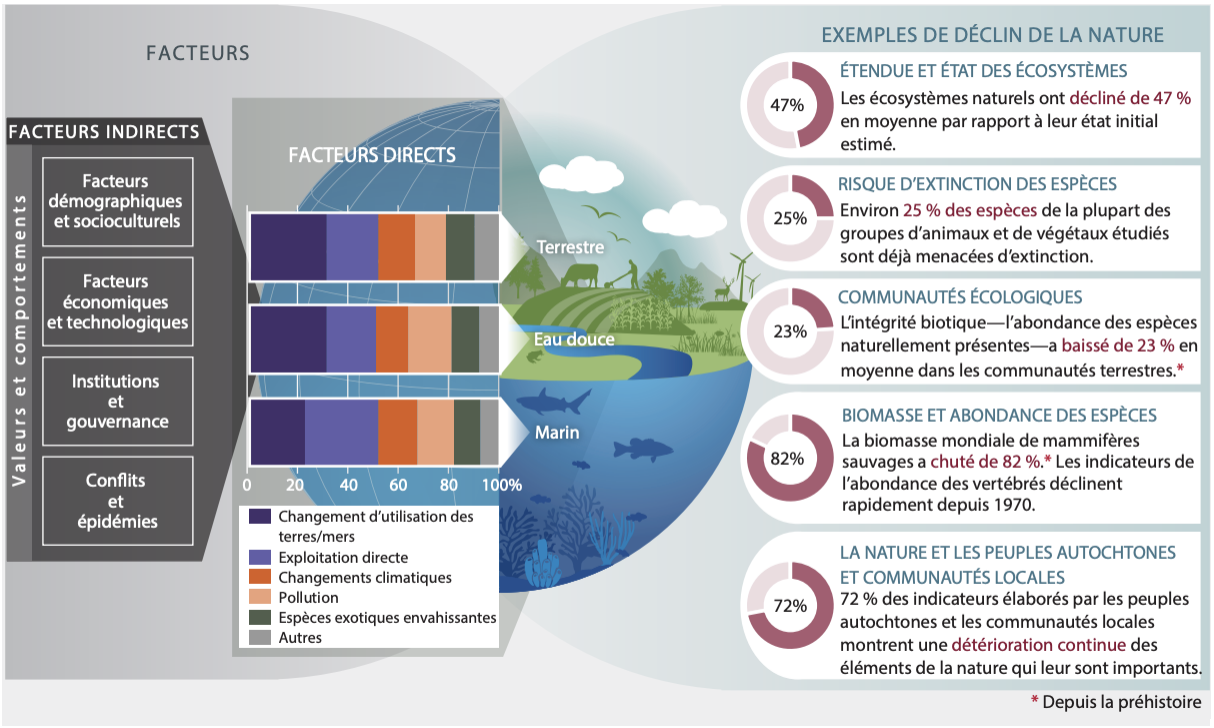

Figure 1.1 – Exemples de déclins observés dans la nature au niveau mondial, soulignant le recul de la biodiversité provoqué par des facteurs de changement directs et indirects. (Díaz et al., 2019)

2. Le décalage génétique

La mesure de décalage génétique se base sur une relation statistique entre des données génomiques et des données environnementales, on parle de gene environment association (GEA). On se base ensuite sur cette relation pour établir une distance entre la composition génétique d'une population adaptée à son environnement et la composition théoriquement requise pour que cette dernière soit adaptée à des conditions environnementales modifiées (Capblancq, Fitzpatrick et al., 2020) (Rellstab, Dauphin et al., 2021) (Figure 1.2).

Le concept de décalage génétique s'inscrit dans une compréhension des impacts du changement climatique sur la biodiversité. Cette mesure est construite pour prédire la valeur sélective des individus dans un environnement donné sur la base de leur génotype. Cette indication est précieuse dans bon nombre d'applications telles que l'agriculture et la conservation. On peut par exemple utiliser ces mesures pour prédire quels génotypes fourniront les meilleures récoltes pour un environnement donné ou encore pour prédire à quelle point une population naturelle aura besoin de migrer pour atteindre une région au climat adapté et pour informer des politiques de flux de gènes assistés (Aitken et Whitlock, 2013) (Aitken et Bemmels, 2016) (Keller et al., 2018) (Steane et al., 2014) (Supple et al., 2018).

3. Problématique de la thèse

Durant cette thèse, nous cherchons à prolonger le travail déjà effectué autour de la notion de décalage génétique. Notre travail se concentre sur différents axes qui sont résumés dans cette section.

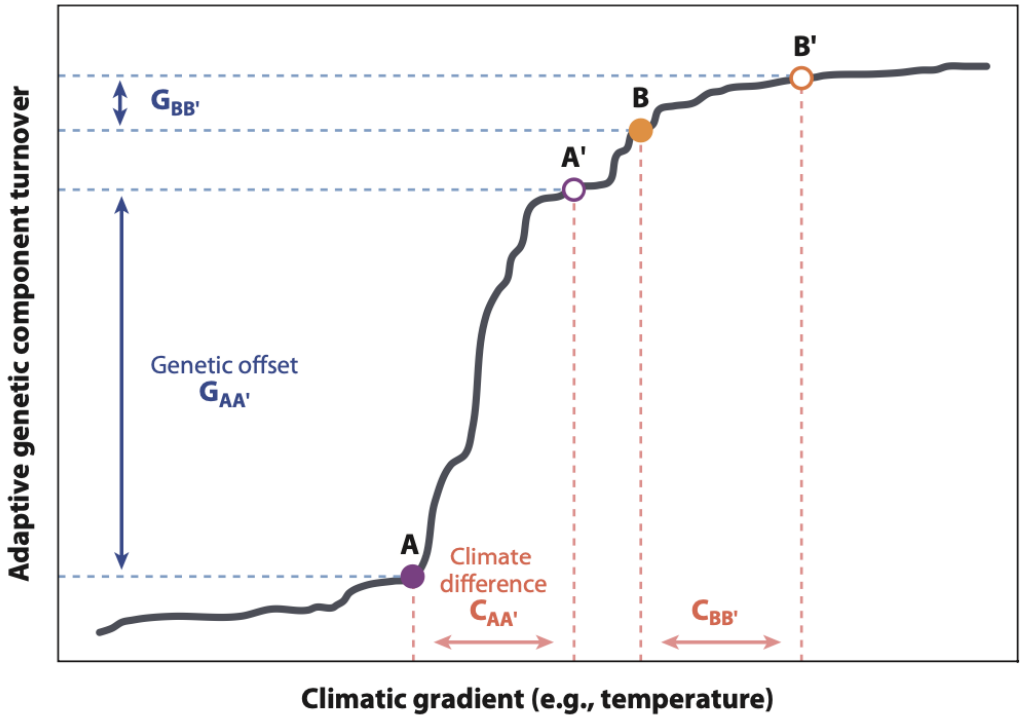

Figure 1.2 – Illustration du concept de décalage génétique issue de (Capblancq, Fitzpatrick et al., 2020)

Une variation le long du gradient climatique est associée à une modification de la composition génétique adaptative, quantifiée par le décalage génétique. On constate qu'une variation de A à A' donne lieu à une valeur de décalage génétique élevée alors qu'une variation équivalente de B vers B' donne lieu à une valeur plus faible. Cela illustre l'intérêt d'intégrer la composante adaptative pour quantifier la maladaptation plutôt qu'une simple distance euclidienne entre les environnements.

3. 1. Création d'une nouvelle mesure

Il existe plusieurs méthodes permettant le calcul du décalage génétique. Certaines limites ont été identifiées quant à l'utilisation de ces méthodes. La description des méthodes existantes et les limites associées sont détaillées dans le chapitre 2 d'état de l'art. Nous avons fait le choix de développer une nouvelle méthode intitulée fossé génétique, permettant de résoudre certaines de ces limites. Notre méthode permet notamment la correction pour les facteurs de confusion, la possibilité d'appliquer la méthode à un grand nombre de single nucleotide polymorphisms (SNPs cf chapitre 2 pour une définition plus précise) afin de prendre en compte l'aspect polygénique de l'adaptation. Cette méthode est détaillée dans le chapitre 3 de cette thèse. Nous donnons une double interprétation de notre mesure à la fois comme distance dans la niche écologique, correspondant à la somme des conditions d'habitat qui permettent aux individus d'une population de survivre et de se reproduire, mais également comme distance génétique. Ce chapitre détaille également un cadre théorique pour l'interprétation de notre mesure de décalage génétique en terme de valeur sélective dans l'environnement modifié.

3. 2. Interprétabilité des mesures

Comme expliqué précédemment, les mesures de décalage génétique ont initialement été pensée afin de renseigner la maladaptation d'une population lorsqu'elle fait face à un environnement modifié. Toutefois, il est important de souligner que le décalage génétique correspond à une distance entre deux compositions génétiques et qu'il n'existe aucune littérature qui nous renseigne sur le lien entre cette mesure et la valeur sélective dans un environnement modifié. Dans le chapitre 3, nous établissons un cadre théorique nous permettant d'établir cette relation. Nous utilisons également ce cadre pour unifier les différentes mesures de décalage génétique et pour quantifier l'importance des variables impliquées dans le processus d'adaptation locale. Une fois la relation théorique établie entre le décalage génétique et la valeur sélective, nous cherchons également à vérifier cette relation de manière empirique à l'aide de données réelles et de données simulées.

3. 3. Validation des mesures

Il existe quelques expériences visant à valider les mesures de décalage génétique mais ces dernières restent encore très rares (Rellstab, Dauphin et al., 2021). Dans le chapitre 4, nous proposons de valider les mesures de décalage génétique à l'aide de deux outils, la simulation et les données réelles. Nos expériences de simulations visent à reproduire une expérience d'adaptation locale suivie d'un changement brutal d'environnement. Notre outil de simulation nous permet alors de comparer les valeurs de décalage génétique aux valeurs sélectives dans l'environnement modifié. Il nous fournit donc un critère objectif pour comparer les performances des méthodes dans différents scénarios. Nous utilisons également des données de mil (*Pennisetum glaucum*) et une expérience de jardin commun pour vérifier la relation entre décalage génétique et valeur sélective dans le cas de données réelles.

3. 4. Mise en place d'une relation théorique entre l'indice de fixation F_{ST} et l'ACP

Dans le chapitre 3, nous établissons un lien entre le décalage génétique et des mesures classiques de génétique des populations, notamment les indices de Wright D_{ST} et la F_{ST}. Dans cette thèse, nous fournissons également une relation entre ces mesures et les valeurs propres de l'analyse en composante principale (ACP) de la matrice de génotype. L'ACP et les mesures de F_{ST} et D_{ST}

constituent deux approches essentielles dans la compréhension de la structure de population chez des individus échantillonnés et nous prolongeons le travail effectué dans la compréhension des liens existants entre ces deux approches. Cette théorie est présentée dans le chapitre 5 de cette thèse.

Chapitre 2

État de l'art

La fonction principale de ce chapitre est de donner toutes les clefs nécessaires à la compréhension de nos travaux de thèse. Il se concentre pour cela sur deux aspects. Il fournit le contexte scientifique dans lequel s'inscrit notre sujet de thèse en effectuant un historique des idées ayant mené à l'apparition du concept de décalage génétique et un historique du travail effectué autour du lien entre l'indice de fixation F_{ST} et l'analyse en composantes principales (ACP). Il présente plusieurs notions essentielles à la compréhension de notre thèse. On rappelle notamment certains fondamentaux de la génétique des populations en définissant certains termes utilisés en introduction. Nous présentons également les hypothèses nécessaires à l'élaboration de notre théorie quantitative du décalage génétique, le modèle infinitésimal de Fisher et la sélection stabilisatrice gaussienne. Enfin, nous présenterons le modèle mixte à facteurs latents (LFMM) qui sera fréquemment utilisé dans nos travaux de thèse. Ce chapitre fournit donc le socle de connaissance nécessaire à la compréhension des trois prochains chapitres.

1. Notions essentielles

Cette section vise à introduire quelques notions essentielles qui seront utilisées dans toute la thèse. Nous allons notamment introduire le type de données avec lesquelles nous allons travailler et le domaine de la génétique des populations.

1. 1. Format des données génétique

Durant cette thèse, nous avons travaillé avec les polymorphismes génétiques d'un seul nucléotide, single nucleotide polymorphisms (SNP), qui correspond à la substitution d'un seul nucléotide à une position spécifique du génome, appelé le locus. A certains locus, il existe donc plusieurs variants, appelés allèle. Durant cette thèse, nous avons fait l'approximation qu'il n'existe que 2 versions possibles pour un locus donné, l'allèle de référence (ou ancestral) et l'allèle alternatif (ou dérivé). Techniquement, l'allèle ancestral correspond à l'allèle présent avant l'apparition de la mutation et l'allèle dérivé serait l'allèle résultant de la mutation mais la distinction entre les deux n'est pas toujours possible. On parle donc dans ce cas plutôt d'allèle de référence et d'allèle alternatif. Ainsi, pour un individu, à un locus donné, trois valeurs sont possibles :

— 0 lorsque l'individu est homozygote pour l'allèle de référence

— 1 lorsque l'individu est hétérozygote

— 2 lorsque l'individu est homozygote pour l'allèle alternatif

Les données génétiques que nous avons manipulées sont donc des matrices Y de taille $n \times L$ où n correspond au nombre d'individus et L correspond au nombre de locus. Toutes les entrées de la matrice sont donc des valeurs parmi 0, 1 ou 2. Il est également possible de rassembler les individus par population. Nous manipulons alors des matrices de taille $p \times L$ où p correspond au nombre de populations. Les entrées correspondent à la fréquence de l'allèle dit dérivé dans la population et sont donc comprises entre 0 et 1.

1. 2. Génétique des populations

Dans le cadre de ce projet de thèse, nous faisons régulièrement appel à des connaissances appartenant au domaine de la génétique des populations. Le principe de base de la génétique des populations consiste à comparer l'ADN de différentes populations. Concrètement, on cherche à expliquer les différences génétiques au sein des populations et entre ces dernières. Cette discipline apporte une compréhension quant à l'histoire d'une espèce, les migrations, les variations de taille de population et les pressions évolutives qu'a pu subir cette espèce (Luikart et al., 2003). Parmi les pionniers de cette discipline, on peut citer Sewall Wright, J. B. S. Haldane et Ronald Fisher. Les avancées en terme de génotypage, le processus permettant de déterminer le génome des individus, permettent de travailler aujourd'hui à l'échelle de génome complet. Lorsque les études sont faites sur des génomes entiers, on peut alors parler de génomique des populations. Les différences de fréquences d'allèles entre populations sont expliqués par différentes processus évolutifs :

Dérive génétique Plaçons nous dans une population de taille finie où la reproduction est aléatoire. En dehors de toute autre force évolutionnaire, nous allons constater que les fréquences alléliques vont évoluer au cours des générations. Cela est dû à la dérive génétique. La dérive génétique correspond à la variation de fréquence d'allèle dû au hasard.

Flux de gènes Le flux de gènes correspond à l'échange de gènes ou de leurs allèles entre différentes populations apparentées en raison de la migration d'individus fertiles ou de leurs gamètes. Le flux de gènes vient donc évidemment modifier les fréquences d'allèles de la population "visitée".

De novo mutation Une mutation de novo est "une mutation du gène apparaissant chez un individu alors qu'aucun des parents ne la possède dans son patrimoine génétique" (*Wikipedia* s. d.). Si une mutation de novo apparait à un locus dans une population mais pas dans les autres, on observera nécessairement une différence de fréquence d'allèle à ce locus.

Sélection naturelle La sélection naturelle correspond à "un avantage ou un désavantage reproductif, procuré par la présence ou l'absence de variations génétiques propices ou défavorables, face à un environnement qui peut se modifier" (*Wikipedia* s. d.). La sélection naturelle est un concept central du décalage génétique et la section suivante le définit plus en détail.

2. Historique de l'adaptation locale

En biologie, on appelle évolution la transformation des êtres vivants dans le temps. Dans cette section, nous présenterons succinctement un concept central de l'évolution, présenté de manière rigoureuse par Charles Darwin, la sélection naturelle (Darwin, 1859). Nous montrerons que le phénomène de sélection naturelle donne lieu au phénomène d'adaptation locale. Enfin, nous nous expliquerons que l'adaptation locale peut donner lieu à un autre phénomène : l'inadéquation évolutive, concept étroitement lié à la notion de décalage génétique.

2. 1. La sélection naturelle

Afin de définir le concept de sélection naturelle, nous reprendrons les termes de Charles Darwin dans son oeuvre "L'origine des espèces" : "J'ai donné à ce principe, en vertu duquel une variation si insignifiante qu'elle soit se conserve et se perpétue, si elle est utile, le nom de sélection naturelle" (Darwin, 1859). La sélection naturelle repose sur trois principes : la variation, l'adaptation et l'hérédité (Mayr, 1982).

Variation Comme expliqué en introduction, un organisme est constitué d'un ensemble de traits appelés phénotype. Pour qu'il y ait sélection naturelle sur ces traits, il est nécessaire qu'il y ait une variation de ces traits au sein des organismes de la population.

Adaptation Les traits sous sélection doivent présenter un avantage adaptatif. On entend par là qu'être porteur d'un trait particulier puisse nous apporter soit une plus grande capacité de survie, on augmente alors ses chances d'arriver à l'âge adulte et donc de se reproduire, soit un avantage reproductif, en présentant des traits attirants pour le partenaire de sexe opposé par exemple.

Hérédité Enfin, les traits doivent être héréditaires. Autrement dit, il est nécessaire qu'il puisse y avoir transmission du dit trait à la génération suivante. On sait aujourd'hui que cela correspond au trait déterminé par la génétique.
Ainsi, la sélection naturelle donne lieu au phénomène d'adaptation : les organismes s'ajustent au milieu dans lequel ils évoluent.

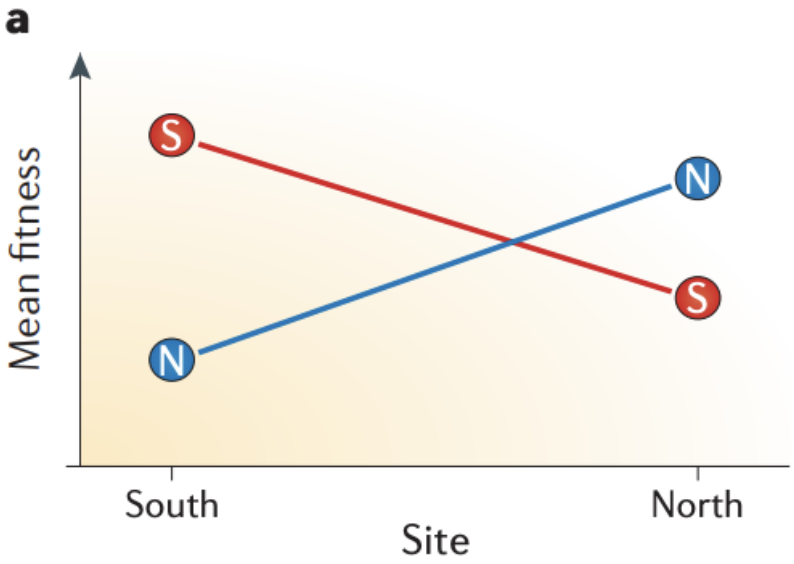

Figure 2.1 – Illustration du phénomène d'adaptation locale (Savolainen et al., 2012) Dans les expériences de transplantations réciproques, on constate que chaque population adaptée à son site d'origine a une meilleure valeur sélective que toute autre population dans le même site.

2. 2. L'adaptation locale

On dit qu'une population est adaptée localement lorsque les organismes de cette dernière présentent des traits phénotypiques différents de ceux des autres populations de la même espèce, et que ces traits offrent une valeur sélective plus élevée à cette population dans son environnement d'origine par rapport aux populations provenant d'autres endroits de l'aire de répartition de l'espèce (Williams, 2018) (Kawecki et Ebert, 2004). Afin de tester l'adaptation locale, on a historiquement eu recours aux expériences de transplantations réciproques. Les expériences de transplantations réciproques consistent à utiliser des jardins communs dans un lieu précis afin de comparer la valeur sélective de la population native de ce lieu à celle de populations introduites (Figure 2.1) (Savolainen et al., 2012). On retrouve des expériences de transplantations réciproques dès 1940 (Clausen et al., 1940). Bien que majoritairement effectué sur des plantes, on retrouve également des expériences réalisées sur des grenouilles des bois (Berven, 1982b) (Berven, 1982a) ou encore sur les saumons (Fraser et al., 2011). Ainsi, les organismes s'adaptent au cours du temps à l'environnement dans lequel ils évoluent. Toutefois, les traits qui apportent un avantage adaptatif peuvent finalement devenir maladaptés à cause de changement de l'environnement. On parle alors d'inadéquation évolutive (Schlaepfer et al., 2002).

2. 3. L'inadéquation évolutive

L'inadéquation évolutive est un état de déséquilibre entre un organisme et son environnement. L'inadéquation survient lorsque les adaptations qui ont contribué à la survie d'un organisme dans des environnements précédents deviennent inadaptées dans un environnement modifié. (Lloyd et al., 2011). On retrouve dans la littérature scientifique, plusieurs études faisant état d'inadéquations évolutives (Cook et Saccheri, 2013) (Knight, 2011). L'exemple emblématique est celui de la phalène du bouleau qui est également une mise en évidence du phénomène de sélection naturelle (Poulton, 1890). La forme ancestrale de la phalène du bouleau est la forme claire. Cette couleur claire lui aurait permis de se dissimuler lorsqu'elle se posait sur les

surfaces des troncs de bouleau. Toutefois, la pollution industrielle a rendu les bouleaux plus foncés et on a alors vu une augmentation rapide de la fréquence des individus sombres. En pratique, la mise en évidence d'inadéquation évolutive et la compréhension de l'adaptation locale se sont longtemps appuyées sur des observations directes dans l'environnement ou dans des expériences de jardins communs. Ces expériences sont coûteuses à la fois en temps et en ressources. Toutefois, les nouveaux outils de séquençage de l'ADN ont permis de réaliser des études sur l'ensemble du génome et ont offert un autre axe de compréhension de l'adaptation locale : la base génétique de l'adaptation locale (Savolainen et al., 2012).

2. 4. L'étude de la base génétique de l'adaptation locale

Une étape importante dans la compréhension de l'adaptation locale est l'identification des SNPs sous sélection. Il existe une littérature fournie concernant l'identification de SNPs causaux (Rellstab, Gugerli et al., 2015) (Hoban et al., 2016) et différentes manières pour identifier ces derniers. Dans cette sous-section, nous présenterons deux catégories de méthodes : les méthodes basées sur la différenciation des populations et notamment le coefficient de consanguinité de Wright F_{ST} (défini plus en détail dans la section **5.1**) et les méthodes basées sur la corrélation entre les fréquences d'allèles et les variables environnementales.

Méthodes basées sur la différenciation des populations Ces méthodes consistent à identifier des SNPs ayant des valeurs aberrantes de F_{ST} (De Mita et al., 2013) (Luu et al., 2017). On suppose alors que ces valeurs aberrantes sont le résultat d'une sélection sur ces SNPs. Toutefois, d'autres procédés démographiques peuvent donner lieu à des valeurs aberrantes de F_{ST} et il faut donc interpréter la liste comme une liste de candidats potentiels pour lesquels des investigations plus poussées peuvent être nécessaires.

Méthodes basées sur la corrélation entre fréquences d'allèles et variables environnementales Ces méthodes reposent sur l'estimation d'une relation statistique entre les gradients environnementaux et les fréquences d'allèles. Cela peut être réalisé à l'aide de méthodes bayesiennes avec notamment les programmes Bayenv (Coop et al., 2010) et BayPass (Gautier, 2015) ou avec des modèles mixtes à facteurs latents (Caye, Jumentier et al., 2019). Ces méthodes proposent des moyens pour corriger pour les facteurs de confusions, souvent présents dans les données de génétique des populations.
Les données génétiques peuvent donc être utilisée pour identifier les SNPs sous sélection. Ces données ont également été utilisé afin de prédire l'inadéquation évolutive : c'est l'apparition du concept de décalage génétique.

3. Le décalage génétique

Le décalage génétique est le concept central de notre thèse. Comme expliqué en introduction, cette mesure vise à prédire la maladaptation future des populations dans un environnement changeant.

3. 1. Historique de l'apparition du concept

Les modèles de distribution d'espèces Avant l'utilisation de modèle de décalage génétique, la prédiction de la réaction de la biodiversité au changement climatique s'est d'abord appuyé sur des modèles de distribution d'espèces (Guisan et Zimmermann, 2000) (Guisan

et Thuiller, 2005) (Elith et Leathwick, 2009). Les modèles de distribution d'espèces cherchent à établir comment les conditions influencent la présence d'une espèce. Ils peuvent s'appuyer pour cela sur des modèles corrélatifs, qui utilisent la corrélation entre les données d'occurrence d'espèce et les données environnementales afin d'établir les conditions environnementales nécessaires à l'existence de l'espèce. Ils peuvent alors être utilisé dans un but prédictif, notamment pour obtenir des indications sur la distribution future de l'espèce dans le cadre du changement climatique. L'utilisation des modèles de distribution d'espèces permet une meilleure compréhension du déplacement de l'aire de répartition des espèces en réponse au changement climatique. Toutefois, du fait de la diversité génétique au sein d'une espèce, différentes populations peuvent répondre de manière différente en raison de l'adaptation à leur environnement (Aitken, Yeaman et al., 2008) (F. Jay et al., 2012). On voit alors apparaître de nouvelles approches pour prévoir les effets du changement climatique en utilisant des données génétique, c'est l'apparition du concept de décalage génétique.

L'apparition du concept de décalage génétique Le terme de décalage génétique (genetic offset en anglais) apparaît pour la première fois dans la littérature scientifique dans le papier de Fitzpatrick et Keller, 2015. Ils proposent alors d'utiliser deux méthodes initialement utilisées pour la modélisation de la composition de communauté écologique à partir de gradients environnementaux, gradient dissimilarity modelling (GDM) (Ferrier et al., 2007) et gradient forests (GF) (Ellis et al., 2012). Plutôt que de modéliser des communautés écologiques, ils appliquent les méthodes pour modéliser des compositions génétiques. Ils se servent ensuite de ces modèles pour obtenir des différences dans les fréquences alléliques prédites à des paires de points dans la niche écologique, correspondant à la somme des conditions d'habitat qui permettent aux individus d'une population de survivre et de se reproduire, et appelle cette mesure le décalage génétique. D'autres méthodes sont ensuite utilisées pour obtenir des mesures similaires. Rellstab propose en 2016 une mesure appelée risk of non adaptedness (Rellstab, Zoller et al., 2016). Capblancq décrit dans deux papiers (Capblancq, Fitzpatrick et al., 2020) (Capblancq et Forester, 2021) comment la redundancy analysis (RDA) peut être utilisée pour obtenir des mesures de décalage génétique. La sous-section suivante vise à décrire plus en détail les différentes méthodes.

3. 2. Description des méthodes

Dans cette partie, nous allons présenter plusieurs méthodes de calcul de décalage génétique. Toutes les méthodes que nous présentons ici repose sur l'estimation d'une relation entre fréquences d'allèles et gradients environnementaux en utilisant des modèles d'association gène environnement (GEA). Dans toute cette partie définissons $\mathbf{Y}$ une matrice de génotype de taille $n \times L$ avec n correspondant au nombre d'individus et L au nombres de locus. Définissons $\mathbf{X}$ une matrice de l'environnement associé à chacun des individus échantillonnés de taille $n \times d$ où d représente le nombre de variables environnementales. Définissons $\mathbf{X}^\star$ la matrice de l'environnement modifié pour lequel on veut évaluer notre décalage génétique de taille $n \times d$. Définissons également $\mathbf{Y_{freq}}$ la matrice de fréquence de génotype de taille $F \times L$ où F représente le nombre de populations, on définit de manière similaire $\mathbf{X_{freq}}$, la matrice d'environnement moyen par population de taille $F \times d$ et $\mathbf{X^\star_{freq}}$ la matrice de l'environnement modifié.

L'approche space-for-time Avant de présenter en détail les méthodes, nous allons introduire une notion sur laquelle repose l'entraînement de ces dernières, l'approche space-for-time. Rellstab définit cette approche de la manière suivante : "L'utilisation de données spatiales pour déduire la dynamique temporelle en l'absence de données temporelles. Des sites séparés dans

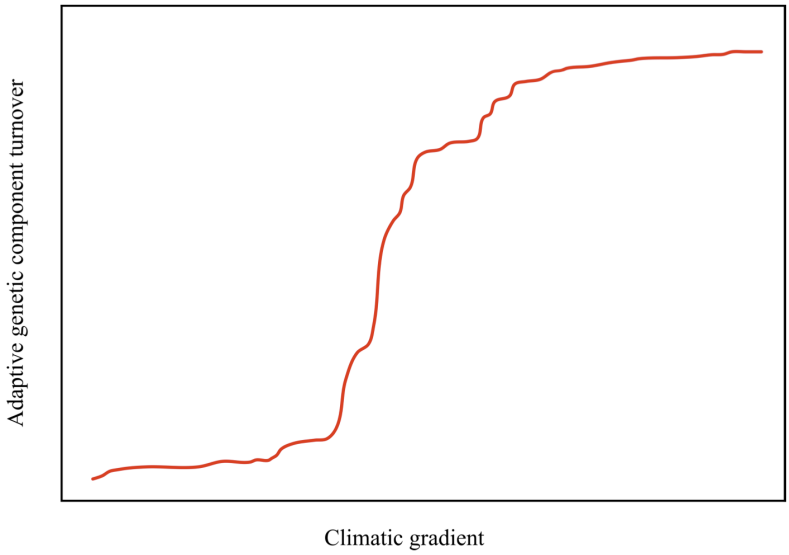

Figure 2.2 – Exemple de fonction F_{fp} de changement de fréquence d'allèle à un locus pour un prédicteur donnée
Pour chaque locus, on entraine une forêt aléatoire f à prédire la fréquence d'allèle à ce locus en fonction d'un prédicteur environnemental p, on obtient alors la fonction F_{fp}.

l'espace le long de gradients écologiques ou environnementaux servent de substituts pour prédire des séries temporelles." (Rellstab, Dauphin et al., 2021). L'estimation de la relation statistique entre fréquences d'allèles et gradients environnementaux se base en effet sur des données d'entraînements. En l'absence de données échantillonnées à différents moments pour une même population, on se sert de données de plusieurs populations échantillonnées dans différents lieux géographiques (et faisant donc face à différents environnements).

Gradient forests (GF) La méthode Gradient Forests (GF) a été présenté dans un papier de 2012 (Ellis et al., 2012). La méthode a été initialement appliquée pour modéliser la composition d'une communauté écologique à partir des données environnementales. Dans notre cas d'usage, GF sera utilisé pour modéliser la composition allélique d'une population. GF est basé sur l'algorithme de machine learning Random Forest (RF), lui même basé sur des arbres de décision dont l'entraînement est effectué sur des sous-ensembles de données sélectionnées selon la méthode du bootstrapping (tirage avec remise d'un nombre donné d'échantillons). Voici un descriptif plus précis du fonctionnement de GF : Pour chacun des locus, on entraîne une RF à prédire la fréquence des populations à partir des données environnementales. Il est également possible de chercher à prédire la valeur de chacun des SNPs des différents individus. On retrouve à nouveau le principe des GEA à cette étape. On entraîne donc une RF par locus. Pour chacune de ces RF on peut obtenir R^2_{fp} la part de variance expliquée par le prédicteur p dans la forêt f. A partir de cette RF entraînée, on peut construire la fonction F_{fp} la fonction de changement de fréquence d'allèle (Figure 2.2). Quelque soit x, $F_{fp}(x)$ correspond à la part de variance de la réponse expliquée par le prédicteur entre 0 et x.
On agrège alors ensuite les résultats des différentes forêts f, en pondérant l'importance de chacun des forêts par leur capacité de prédiction du SNPs résumé par R^2_f la part de variance expliquée par la forêt f. On obtient alors pour chacun des prédicteurs p la fonction F_p. On utilise alors $F_p(x_2) - F_p(x_1)$ comme un indicateur de la variation des fréquences d'allèles lorsque le prédicteur p passe de la valeur x_1 à x_2. La figure 2.3 résume comment obtenir la valeur de décalage génétique dans le cas d'un seul prédicteur environnemental.

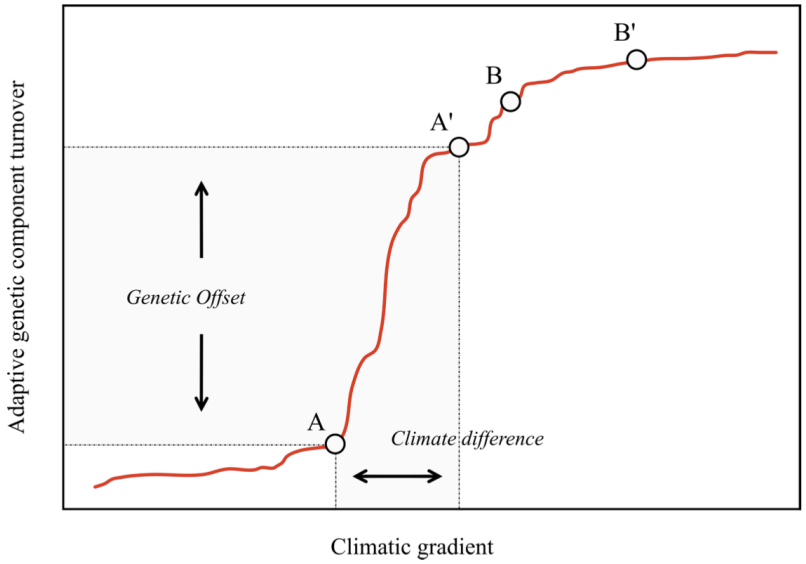

Figure 2.3 – Illustration du calcul de décalage génétique avec GF extrait d'une présentation de Stephen Keller
Obtention de la valeur de décalage génétique pour un prédicteur environnemental subissant une modification de A vers A'

Pour obtenir le décalage génétique avec plusieurs prédicteurs, on calcule le vecteur $CI(\mathbf{X})$ et $CI(\mathbf{X_{pred}})$ correspondant à la projection de chacun des prédicteurs p de $\mathbf{X}$ et $\mathbf{X_{pred}}$ sur l'axe des ordonnées de leur fonctions respectives F_p. Le décalage génétique correspond alors à la distance euclidienne entre $CI(\mathbf{X})$ et $CI(\mathbf{X_{pred}})$.

Risk of non-adaptedness (Rona) Le Rona est une méthode qui a été présentée par Rellstab en 2016 (Rellstab, Zoller et al., 2016). Dans ce papier, Rellstab travaille sur les fréquences d'allèles mais on pourrait facilement étendre le raisonnement pour des matrices de génotypes d'individus. La première étape du Rona consiste à effectuer la régression linéaire des SNPs en fonction de chacune des variables environnementales. Soit x_{freq_p}, le vecteur de taille F correspondant à la valeur moyenne du prédicteur environnemental p pour chacune des populations, et $x^{\star}_{freq_p}$ le vecteur de taille F correspondant à la valeur moyenne du prédicteur environnemental p **modifié** pour chacune des populations . On effectue alors pour chacun des prédicteurs p, la régression suivante :

$$\mathbf{Y_{freq}} \sim x_{freq_p}$$

On obtient alors pour chacun des SNPs ℓ, et pour chacun des prédicteurs environnementaux p, la taille d'effet $\beta_{p\ell}$ et l'ordonnée à l'origine $\alpha_{p\ell}$. Une fois ces valeurs obtenues, on calcul le Rona, pour un prédicteur p donné, de la manière suivante.

$$Rona = \sum_{\ell=1}^{L} \mid (x^{\star}_{freq_p}\beta_{p\ell} + \alpha_{p\ell}) - \mathbf{Y_{freq}}^{\ell} \mid$$

où $\mathbf{Y_{freq}}^{\ell}$ correspond à la fréquence d'allèle au locus ℓ. La figure 2.4 représente le Rona pour trois espèces de chênes (*Quercus petraea, Q. pubescens, Q. robur*) (Rellstab, Zoller et al., 2016).

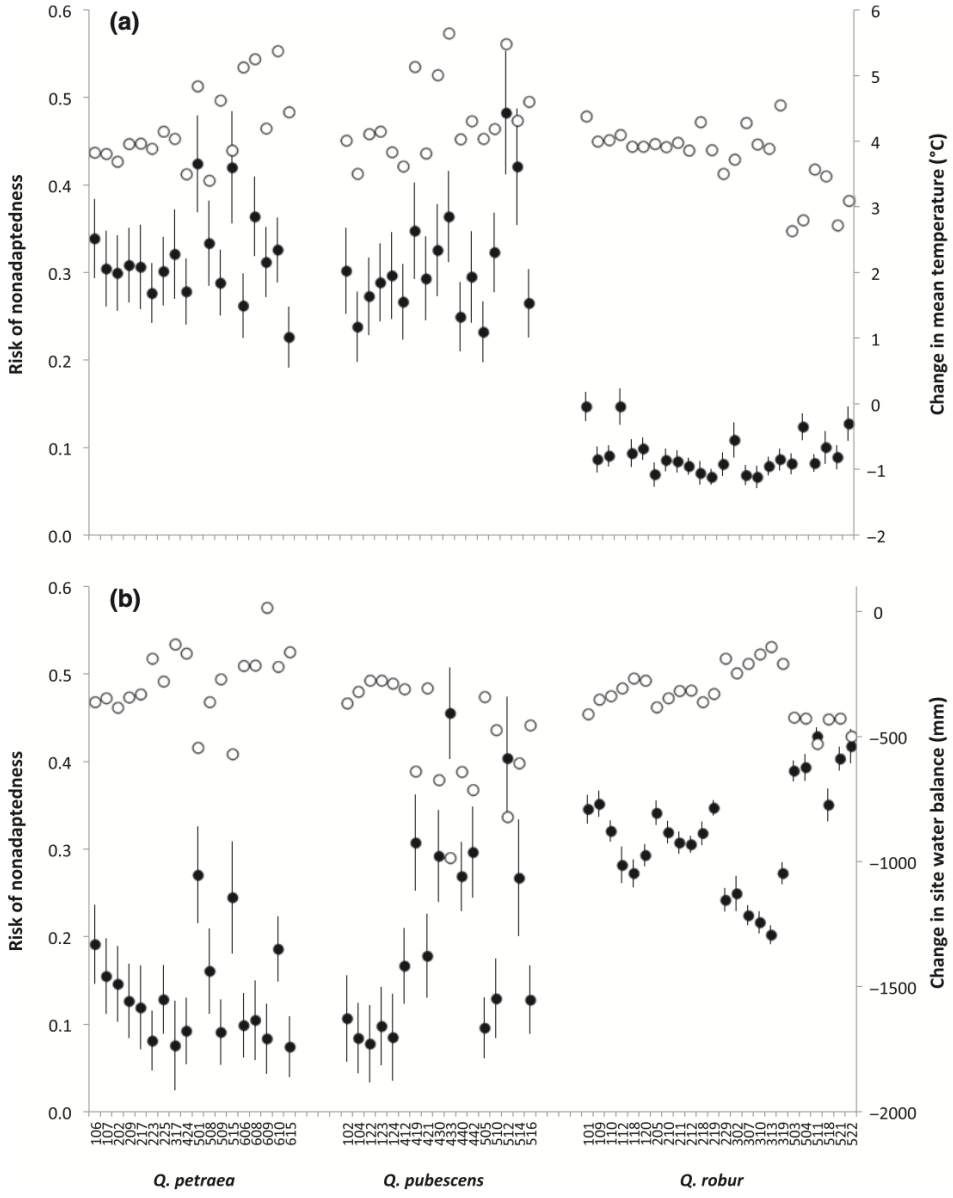

Figure 2.4 – **Valeurs de Rona pour des espèces de chênes et pour deux prédicteurs environnementaux (Rellstab, Zoller et al., 2016)**

Rona obtenu pour
(a) La température moyenne (moyenne sur 20 locus)
(b) modification de la quantité de précipitation (moyenne sur 17 locus)
Cercle blanc : Changement des conditions climatiques entre les périodes 1931-1960 et 2071-2100 dans les habitats des 71 populations étudiées.
Cercle noir : Rona moyen obtenu sur un ensemble de locus sélectionnés
Les barres d'erreur représentent l'écart type.

Redundancy analysis (RDA) Capblancq a utilisé la RDA afin de calculer des mesures de décalage génétique (Capblancq, Fitzpatrick et al., 2020) (Capblancq et Forester, 2021).
La première étape de la RDA consiste a réalisé une régression linéaire similaire à celle décrite dans la partie précédente. Une fois encore, il est possible d'effectuer cette régression sur tous les individus où par fréquence d'allèle dans chaque population. Dans cette partie, nous présenterons la méthode basée sur les individus. On effectue donc la régression suivante.

$$\mathbf{Y} \sim \mathbf{X}$$

On peut alors obtenir la matrice des valeurs ajustées de la régression :

$$\mathbf{Y_{fit}} = \mathbf{X}\beta$$

où β correspond aux tailles d'effets de la régression linéaire. De la même manière on obtient la matrice $\mathbf{Y_{pred}}$ correspondant au nouvel environnement

$$\mathbf{Y_{pred}} = \mathbf{X}^{\star}\beta$$

On effectue alors une ACP sur la matrice $\mathbf{Y_{fit}}$ puis on projette $\mathbf{Y_{fit}}$ et $\mathbf{Y_{pred}}$ sur les composantes principales. L'ACP est une méthode utilisée notamment pour réduire la dimension de nos données tout en conservant une quantité maximale d'information. Elle repose sur le fait de combiner linéairement nos variables initiales afin de créer un ensemble artificiel de variables, appelés composantes principales. Elle sera présentée plus en détail en section **5.2** de ce chapitre. Le décalage génétique de RDA correspond alors à la distance euclidienne entre les projections de $\mathbf{Y_{fit}}$ et $\mathbf{Y_{pred}}$ sur les composantes principales.

3. 3. Limites des méthodes existantes

Dans une review de 2021, Rellstab, Dauphin et al., 2021 identifient certaines limites liées à l'utilisation des décalages génétiques. Dans cette sous-section, nous allons présenter certaines d'entre elles.

Facteurs de confusion Comme nous l'avons vu en introduction, au sein des populations, les changements de fréquences d'allèles peuvent être expliqués par quatre composantes évolutives : le flux de gènes dû à la migration, la dérive génétique, la mutation et la sélection. Les méthodes présentées plus tôt reposent sur l'identification de signaux de sélection. Il faut cependant faire attention 'a ne pas identifier comme de la sélection certaines différences de fréquence qui incomberaient en fait à d'autres facteurs démographiques. Il est donc essentiel de corriger nos méthodes pour les effets dûs à la structure de population. Ce problème est bien connu et les méthodes présentées proposent presque toutes des solutions pour corriger pour les facteurs de confusion.

L'approche space-for-time Comme expliqué en introduction, le calcul du décalage génétique repose sur une approche space for time. On utilise les données spatiales pour faire de l'inférence sur les données temporelles. Une hypothèse forte relative à cette approche est de considérer que les populations sont actuellement adaptées à leur habitat local et il a déjà été montré que ce n'est pas systématiquement le cas. (Browne et al., 2019)

Sélection de SNPs et adaptation polygénique Nous avons vu en section **2.4** de ce chapitre qu'il est parfois préférable d'effectuer une sélection de SNPs causaux sur lesquels faire tourner les méthodes de décalage génétique. L'adaptation est en revanche souvent très polygénique et il est donc important que les méthodes soient en mesure de tourner sur des grands ensembles de SNPs et de détecter des effets très faibles localisés sur ces ensembles.

Quantifier l'incertitude des modèles Les méthodes de décalage génétique ne fournisse pas d'information quant à l'incertitude liée à leurs mesures. Cette information peut être importante pour renseigner la fiabilité de la mesure obtenue.

Interprétabilité des mesures L'interprétation biologique des mesures a été peu exploré jusqu'ici. On verra dans la partie suivante qu'on arrive à lier le décalage génétique avec des mesures expérimentales en lien avec la valeur sélective des populations mais il n'existe pas de littérature concernant des interprétations théoriques.

3. 4. Validation des méthodes

Le décalage génétique est une mesure qui a été développé initialement pour donner une indication sur la vulnérabilité des populations. Le décalage génétique a d'ailleurs déjà été appelé "vulnérabilité génomique" dans la littérature scientifique (Bay et al., 2018). On suppose donc que le décalage génétique reflète la perte de valeur sélective provoquée par un changement d'environnement. Différents protocoles expérimentaux ont été développé pour tester la véracité de cette supposition. Dans cette partie, nous présenterons deux de ces protocoles de validation par les données réelles.

Comparaison du décalage génétique avec la tendance démographique actuelle de la population Dans leur papier de 2018, Bay et al., 2018 compare les valeurs de décalage génétique pour le climat de 2050 avec le scénario RCP2.6 pour les populations de paruline jaune dans leur aire de reproduction située en Amérique du Nord avec les tendances démographiques actuelles de ces populations (Figure 2.5). L'idée de cette approche est de dire que si le changement climatique futur est corrélé aux changements récents, alors on s'attend à ce que le changement climatique ayant déjà eu lieu ait déjà impacté négativement les populations dont le décalage génétique est élevé. Comme le montre la figure 2.5, ils mettent en évidence un lien faible mais significatif entre le déclin des populations et leur décalage génétique. Leur approche a ensuite été critiqué dans un commentaire en 2018 (Fitzpatrick, Keller et Lotterhos, 2018), notamment parce que dans ce commentaire, ils ont comparé les tendances climatiques historiques et futures et ont trouvé peu de preuves pour soutenir l'hypothèse selon laquelle les changements climatiques historiques et futurs seraient corrélés.

Expérience de jardin commun Il est également possible d'utiliser des jardins communs afin de valider nos mesures de décalage génétique. La validation s'effectue de la façon suivante, on prédit les valeurs de décalage génétique de chaque population pour un environnement donné $\mathbf{X}^\star$ puis on effectue la transplantation de cette population dans cet environnement contrôlé. On effectue alors un suivi de certaines mesures d'intérêts pour rendre compte de la valeur sélective de la population dans ce nouvel environnement. Il nous est alors possible de comparer les valeurs de décalage génétique avec les traits observés rendant compte de la valeur sélective dans l'environnement modifié afin d'étudier s'il y a effectivement un lien entre les deux. Pour illustrer cette procédure, nous pouvons présenter les travaux de Fitzpatrick, Chhatre et al., 2021. Ils ont planté des individus recueillis dans toute l'aire de répartition du peuplier baumier

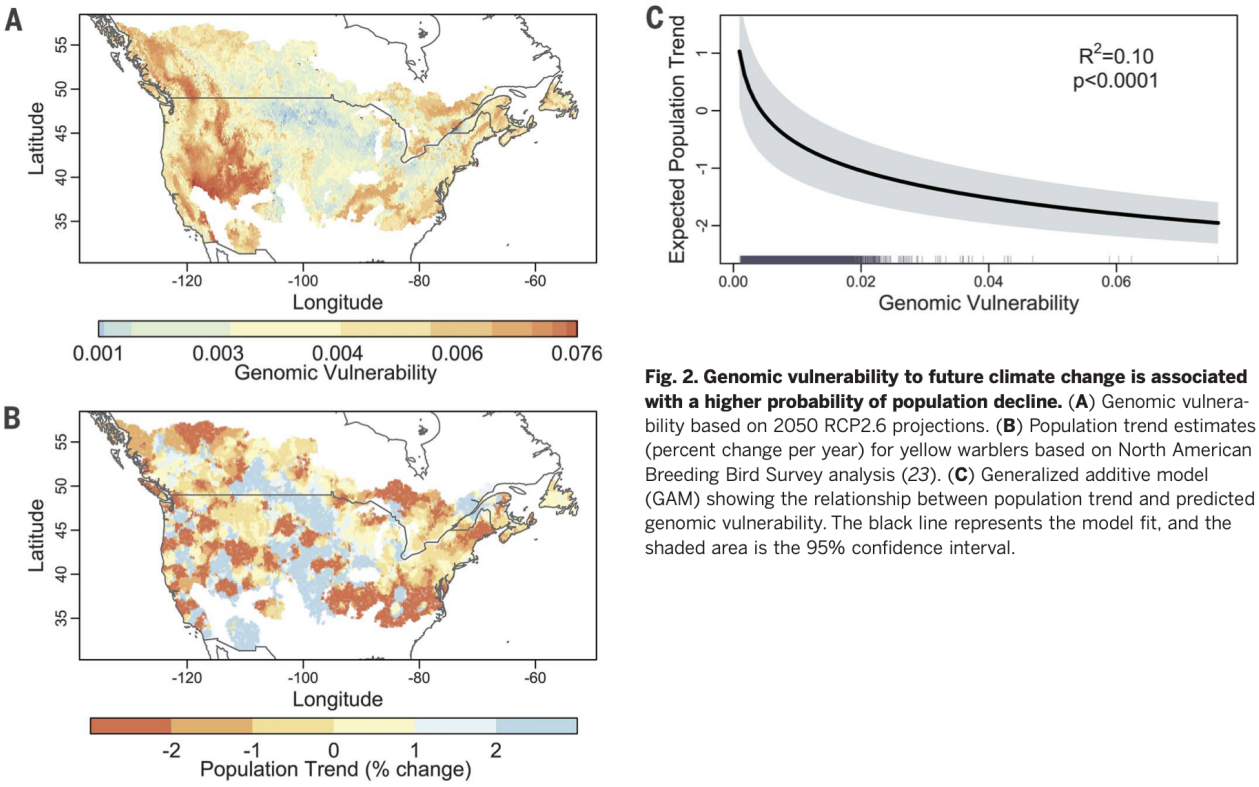

Fig. 2. Genomic vulnerability to future climate change is associated with a higher probability of population decline. (**A**) Genomic vulnerability based on 2050 RCP2.6 projections. (**B**) Population trend estimates (percent change per year) for yellow warblers based on North American Breeding Bird Survey analysis (*23*). (**C**) Generalized additive model (GAM) showing the relationship between population trend and predicted genomic vulnerability. The black line represents the model fit, and the shaded area is the 95% confidence interval.

Figure 2.5 – Comparaison de valeur de décalage génétique avec la tendance démographique actuelle de populations de paruline jaune (Bay et al., 2018)

(*Populus balsamifera*) dans deux jardins communs. Ils ont généré les décalages génétiques à l'aide de GF et ont comparé ces valeurs aux mesures de performance des populations dans le jardin commun. Ils ont choisi comme mesure de performance l'augmentation annuelle de la hauteur des peupliers. Ils retrouvent la relation inverse attendue entre le décalage génétique et la performance dans les jardins communs : les populations avec des décalages génétiques prédits plus importants ont eu de moins bonnes performances dans les jardins communs que les populations avec des décalages plus faibles. De plus, le décalage génétique prédit mieux la performance dans les jardins communs qu'une distance climatique naïve. Des résultats similaires ont été retrouvé dans une autre expérience de jardins communs pour des pins gris (*Pinus banksiana*) et le sapin de douglas (*Pseudotsuga menziesii*) (Lind et al., 2023).

3. 5. Application des méthodes

Ces différentes méthodes ont été appliquées à des données réelles pour prédire des décalages génétiques de population dans des climats futurs et ainsi, obtenir des informations complémentaires pour la mise en oeuvre de politiques de conservation. Par exemple, Rhoné et al., 2020 ont appliqué la méthode GF sur des données du mil (*Cenchrus americanus*). Plus précisément, ils ont calculé le décalage génétique dans toute la zone de culture du mil en Afrique de l'Ouest aux horizons 2050 et 2100 à partir des projections climatiques futures réalisées par un modèle climatique dédié. La projection de ce décalage en 2050 est résumé dans la figure 2.6.
Ils prédisent alors que les zones les plus vulnérables pourraient bénéficier de l'utilisation de variétés qui poussent déjà dans des conditions climatiques équivalentes actuellement. Il faudrait donc envisager des scénarios d'échange de semences nécessitant des migrations assistées sur de longues distances et au-delà des frontières. Leur conclusion est que l'exploitation de la

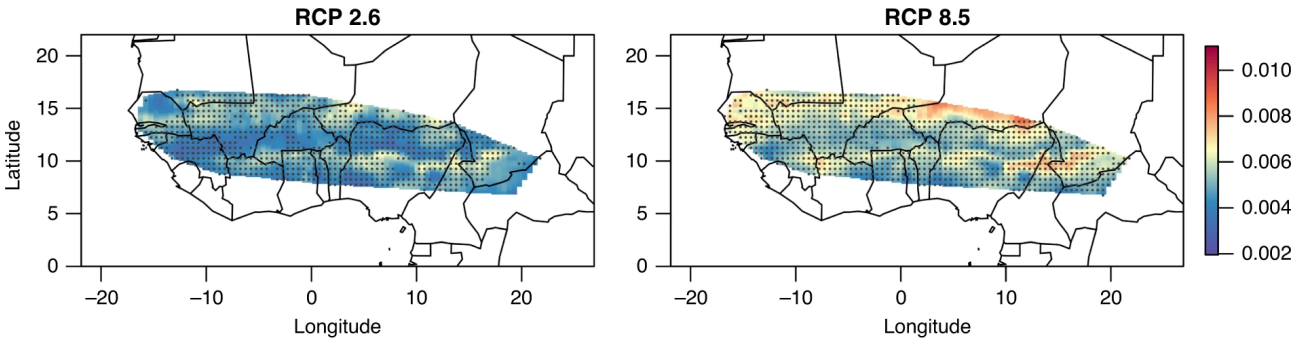

Figure 2.6 – **Décalage génétique du mil pour le climat en 2050 de 2 scénarios d'emissions (Rhoné et al., 2020)**

diversité génétique comme stratégie d'atténuation du changement climatique en Afrique de l'Ouest nécessitera une collaboration régionale.

3. 6. Conclusions

Dans cette partie, nous avons présenté un court historique de l'apparition du concept de décalage génétique. Nous avons ensuite résumé le fonctionnement de certaines des méthodes les plus utilisées pour le calcul de décalage génétique. Nous avons vu également les limites de ces méthodes, des procédures de validation des méthodes et des exemples d'applications. Dans cette thèse, nous présentons une nouvelle méthode que l'on appelle "fossé génétique" cherchant à résoudre certaines des limites présentées dans cette partie. Nous cherchons également à unifier les différentes approches en faisant des liens entre les différentes mesures de décalage génétique. Nous développons également notre propre procédure de validation à l'aide de simulations et de l'outil SLiM 3 (Haller et Messer, 2019). Enfin, nous développons l'interprétation théorique des mesures de décalage génétique. Plus précisément, nous établissons une relation entre le décalage génétique et la valeur sélective d'une population faisant face à un nouvel environnement. Afin d'établir cette relation, il est nécessaire d'utiliser certaines connaissances de génétique quantitative que nous résumons dans la section suivante.

4. Fondements théoriques pour l'interprétation du décalage génétique

Pour ce projet de thèse, nous avons travaillé sur l'interprétation du décalage génétique. L'objectif du décalage génétique est de prédire la maladaptation d'une population dans un environnement modifié. On s'attend donc à ce qu'il existe un lien entre le décalage génétique et la valeur sélective d'une population dans un environnement modifié. Tel que défini actuellement, le décalage génétique correspond à une mesure de différence de fréquences d'allèles entre deux points de la niche écologique. Afin de faire le lien entre cette mesure de différence de fréquences d'allèles et la valeur sélective, nous avons utilisé les hypothèses du modèle infinitésimal de Fisher (Fisher, 1919) et de la sélection stabilisatrice Gaussienne (Schmalhausen, 1941). Le modèle infinitésimal de Fisher nous est utile pour faire le lien du génotype vers les valeurs de traits adaptatifs et la sélection stabilisatrice Gaussienne nous permet d'établir un lien entre les traits adaptatifs et la valeur sélective.

4. 1. Le modèle infinitésimal de Fisher

L'hypothèse de base de ce modèle est que la variation d'un trait phénotypique est influencée par un nombre infiniment grand de gènes, chacun d'eux apportant une contribution infiniment petite. Il a été présenté pour la première fois par Fisher (Fisher, 1919) et est étudié en détail dans un papier de Barton (Barton et al., 2017). Dans le cadre de cette thèse, on peut résumer le modèle infinitésimal de la manière suivante : Considérons un trait adaptatif z, régi par L mutations chacune ayant un effet infinitésimal, $a_l \approx \pm\sqrt{a/L}$ sur ce trait. On définit la valeur du trait de la manière suivante :

$$z = \sum_{l=1}^{L} a_l y_l \tag{2.1}$$

Ainsi, les mesures de décalage génétique nous permettant de quantifier les différence de fréquences d'allèles provoquées par un changement environnemental, il est ensuite possible d'utiliser le modèle infinitésimal pour interpréter comment un changement environnemental impacte la valeur d'un trait adaptatif.

4. 2. La sélection stabilisatrice Gaussienne

On appelle sélection stabilisatrice gaussienne un type de sélection naturelle pour lequel la distribution de la valeur sélective en fonction du trait phénotypique forme une courbe gaussienne. La sélection stabilisatrice a tendance à supprimer les valeurs de phénotypes les plus extrêmes, donnant lieu au succès reproductif des phénotypes moyens. Cette hypothèse nous est utile pour établir une relation entre la différence des traits adaptifs et la valeur sélective. Ces deux hypothèses nous permettent donc d'établir une relation entre le décalage génétique et la valeur sélective. Ce sera l'objet du chapitre 3. En plus d'en donner une interprétation en terme de valeur sélective, nous cherchons également à lier notre mesure à des concepts de génétique des populations. En parallèle du travail pour établir une relation entre le décalage génétique et des concepts de génétique des populations, nous avons effectué un travail sur le lien entre les indices de fixation de Wright et l'analyse en composante principale. Plus précisément, nous avons cherché à établir une relation entre les valeurs propres de la matrice de données génétique et les indices de fixation. Le travail existant sur le sujet ainsi que certaines notions essentielles dans l'étude de la structure de population sont résumés dans la section suivante.

5. Historique autour de la notion de structure de population

Cette section vise à présenter le travail existant autour du lien entre l'ACP et l'indice de fixation de Wright (Wright, 1965). Dans cette section, nous introduisons rapidement les concepts d'ACP et d'indice de fixation puis nous présentons les travaux ayant été effectués sur le lien entre ces deux notions. Pour rappel, notre travail consiste à établir un lien entre les valeurs propres de l'ACP de la matrice de génotype et l'indice de fixation de Wright.

5. 1. Indice de fixation de Wright

Pour introduire, le concept d'indice de fixation de Wright, plaçons nous dans le cas simple d'une population répartie en 2 sous-populations et sur un gène comportant 2 allèles A et a. Appelons p la fréquence de l'allèle A et q la fréquence de l'allèle a à l'échelle de la population. Appelons

Genotype	AA	Aa	aa
Frequency	$p^2(1-F_{ST})+pF_{ST}$	$2pq(1-F_{ST})$	$q^2(1-F_{ST})+qF_{ST}$

Table 2.1 – Déviation des fréquences HWE
Fréquence des génotypes à un locus lorsque les hypothèses HWE ne sont pas respectées.

p_1 et q_1 les fréquences dans la populations 1 et p_2 et q_2 les fréquences dans la population 2. Appelons c_1 et c_2 les proportions d'individus dans la population 1 et 2. Commençons par définir 2 valeurs : H_T et H_S.

— $H_T = 2pq$, l'hétérozygotie totale

— $H_S = 2 \times (c_1 p_1 q_1 + c_2 p_2 q_2)$, l'hétérozygotie moyenne des sous-populations.

Définissons aussi rapidement le principe d'Hardy Weinberg (Hartl et al., 1997). Plaçons nous dans une population dite idéale (taille infinie, reproduction aléatoire, dérive aléatoire comme seule pression evolutive). Le principe d'Hardy-Weinberg nous dit que la relation mathématique entre les fréquences d'allèles et les fréquences de géntoype est la suivante :

$$AA : p^2 \qquad Aa : 2pq \qquad aa : q^2 \tag{2.2}$$

Cette relation reste inchangée au cours des générations. On dispose désormais de tous les éléments pour définir l'indice de fixation de Wright (Wright, 1965). Wright a mis en place ces mesures appelés statistiques F pour rendre compte des écarts par rapport aux fréquences théoriques de l'équilibre d'Hardy Weinberg. Elles sont parmi les statistiques descriptives les plus utilisées en génétique des populations et de l'évolution (Cockerham, 1969) (Nei, 1973) (Weir et Cockerham, 1984) (Slatkin, 1991) (Holsinger et Weir, 2009). Le F tient ici pour indice de fixation. Le terme fixation est utilisée ici dans le sens d'homozygotie accrue. Dans cette partie nous allons présenter une de ces statistiques, la F_{ST}. La F_{ST} est une mesure de la structure de population, elle permet de rendre compte de la différenciation génétique entre sous-populations. Il est possible de définir la F_{ST} de plusieurs manières, toutes équivalentes. On peut par exemple la définir de la manière suivante.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

On peut donner plusieurs interprétations à la F_{ST}.

— Une quantification de la réduction globale d'hétérozygotie.

— La déviation des fréquences des génotypes par rapport aux fréquences HWE. (Table 1)

— La part de variance expliquée par la structure de population.

Cette statistique a été beaucoup utilisée pour évaluer la structure de population. Elle nécessite toutefois d'avoir une subdivision prédéfinie des populations et permet d'étudier la structure de population à l'échelle d'un seul locus. Avec l'apparition de technique de séquençage à grande échelle, les approches non supervisées telles que l'analyse en composantes principales (ACP) ont toutefois pris une place importante dans les analyses récentes de la structure des populations.

5. 2. L'analyse en composante principale (ACP)

L'ACP est une méthode utilisée notamment pour réduire la dimension de nos données tout en conservant une quantité maximale d'information. Elle repose sur le fait de combiner linéairement nos variables initiales afin de créer un ensemble artificiel de variables, appelés composantes

principales. Le premier axe (ou première composante principale) est choisi de telle sorte que la projection des échantillons le long de celui ci explique la plus grande variance possible parmi tous les axes possibles. On construit ensuite le second axe de telle sorte à maximiser la variance résiduelle parmi tous les axes possibles orthogonaux au premier. On répète ce processus pour tous les axes suivants. Grâce à l'ACP, on peut alors projeter nos données originales sur ces nouveaux axes permettant un résumé efficace de la structure des données. Définissons formellement le fonctionnement de l'ACP. Soit $\mathbf{Y}$ une matrice centrée sur les colonnes de taille $n \times L$. L'ACP définit une transformation des données par rotation et étirement, résumée par la matrice $\mathbf{P}$, telle que l'application de $\mathbf{P}$ aux données originales ($\mathbf{T} = \mathbf{PY}$) conduit à des données ayant les propriétés suivantes

1. La matrice de données transformées $\mathbf{T}$ a la même dimension que les données originales
2. Chaque entrée de la matrice $\mathbf{T}$, t_{ij}, représente la projection de la i^{eme} donnée sur la j^{eme} composante principale, on appellera ces valeurs les loadings.
3. La corrélation entre 2 colonnes de $\mathbf{T}$ vaut 0
4. La somme des variances des colonnes est égale à la variance totale des données originales
5. La variance des colonnes est décroissante.

L'obtention des composantes principales se fait par calcul des vecteurs propres de la matrice de covariance. Ces vecteurs propres sont triés par ordre décroissant de leur valeur propre associée. (McVean, 2009). L'ACP a été envisagée très tôt en biologie humaine, et est devenue une méthode populaire pour étudier la structure génétique des populations (Menozzi et al., 1978). On s'intéresse notamment à la projection des individus sur les axes principaux, donnant des indications sur leur degré de mélange avec les populations sources lorsque ces dernières sont représentées dans l'échantillon (Patterson et al., 2006) (McVean, 2009) (Haak et al., 2015). Il existe donc une littérature fournie autour de l'interprétation des projections sur les axes de l'ACP. Toutefois, des difficultés résident dans l'interprétation des valeurs propres de l'ACP. Les sections suivantes résument les travaux ayant été réalisé dans cette direction.

5. 3. McVean, lien entre la leading eigenvalue et la F_{ST}

Dans son papier de 2009, McVean s'appuie sur la théorie de la coalescence pour démontrer un lien entre F_{ST} et ACP dans une situation à 2 sous-populations A et B (McVean, 2009).

Coalescence L'un des objectifs de la théorie de coalescence est de déterminer la durée écoulée depuis la vie de l'ancêtre commun le plus récent. (Slatkin, 1991)
Pour établir son lien, McVean s'appuie d'abord sur un résultat de Slatkin qui démontre l'expression suivante de la F_{ST} :

$$F_{ST} = 1 - \frac{\bar{t}_w}{\bar{t}} \tag{2.3}$$

où $\bar{t}_w$ est le temps de coalescence moyen pour une paire d'échantillons au sein d'une sous population et $\bar{t}$, celui pour une paire d'échantillons au sein de la population complète. McVean reformule cette expression de la façon suivante :

$$F_{ST} = \frac{\tau c(1-c)}{\bar{t}} \tag{2.4}$$

où c correspond à la fraction de la population dans la première sous-population et où $\tau = 2t_{AB} - t_{AA} - t_{BB}$ avec t_{AB} le temps de coalescence moyen entre un individu de la population

A et un individu de la population B, t_{AA}, celui entre 2 individus de la population A et t_{BB}, entre 2 individus de la population B. McVean montre alors que la part de variance expliquée par le premier axe de l'ACP vaut également $\frac{\tau c(1-c)}{t}$. Il établit ainsi une égalité entre la part de variance expliquée par le premier axe de l'ACP et la F_{ST} dans un cadre à 2 sous-populations. Dans le cadre de la thèse, nous cherchons à étendre ce résultat à un contexte plus général. Plus précisément, nous cherchons à créer une relation dans un cadre à K sous-populations ou K est un entier quelconque.

5. 4. Un test de détection de la structure de population

Dans leur papier de 2006, Patterson et al. cherche à répondre à la question suivante : Est ce que les individus échantillonnés proviennent d'une population homogène ou d'une population contenant des sous-groupes génétiquement distincts (Patterson et al., 2006). Autrement dit, ils cherchent à savoir si les données échantillonnées proviennent d'une population structurée. Ils développent dans ce papier un test statistique permettant de répondre à cette question. Ce test statistique est lié à l'ACP et s'intéressent au valeur propre de la matrice de corrélation des données génétiques. Dans notre thèse, nous proposons également un test pour l'existence de structure dans nos données. Le test statistique de Patterson repose sur une distribution de Tracy-Widom tandis que le notre repose sur une condition de séparation entre les valeurs propres de deux matrices obtenues à partir de la matrice de génotype.

5. 5. Théorie des matrices aléatoires.

La théorie spectrale que nous développons dans le chapitre 5 s'appuie sur la théorie des matrices aléatoires (Marčenko et L.A., 1967) (Johnstone, 2001) (Johnstone et Paul, 2018) (Bryson et al., 2019). Nous nous appuyons notamment sur un résultat concernant la distribution des valeurs propres de l'ACP de la matrice de génotype en l'absence de structure de population (Bryc et al., 2013) (Patterson et al., 2006). Ce résultat stipule que les valeurs propres suivent une distribution appelée distribution de Marchenko-Pastur. Notre théorie repose sur la séparation de la matrice de génotype en deux matrices, une matrice inter-population, et une matrice intra-population. On vérifie alors que les valeurs propres de l'ACP de la matrice intra-population sont en adéquation avec la distribution de Marchenko-Pastur.

6. Introduction aux modèles mixtes à facteurs latents (LFMMs)

Les mesures de décalage génétique que nous avons présentées reposent sur les GEA. La mesure que nous développons s'appuie également sur les GEA. Bien que notre théorie puisse s'appliquer à tout modèle de prédiction de fréquence d'allèle, nous proposons d'utiliser les modèles mixtes à facteurs latents (LFMM) pour l'intérêt qu'ils présentent dans l'estimation des facteurs de confusions liés à l'historique démographique des populations. Cette section vise à présenter le modèle et à présenter l'utilisation de ce modèle en génomique du paysage pour la détection de SNPs sous sélection.

6. 1. Présentation du modèle

Nous avons vu l'importance des GEA dans le calcul du décalage génétique. Nous avons également vu que les études d'associations gène-environnement peuvent faire face au problème des

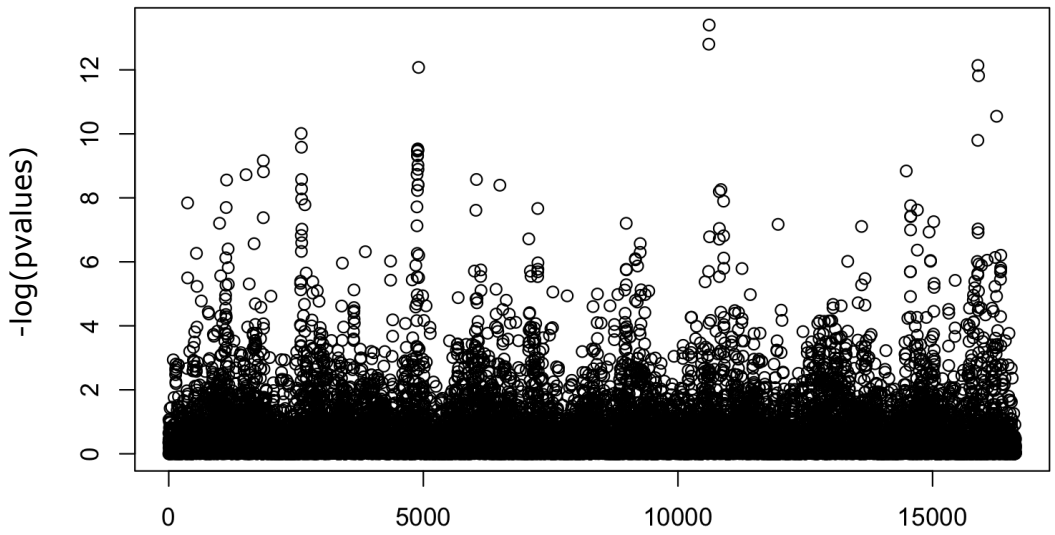

Figure 2.7 – **Manhattan plot des pvaleurs obtenus avec un test de significativité global pour des données de mil avec 30 prédicteurs environnementaux**

facteurs de confusion dus aux facteurs démographiques non observés. Dans l'optique de pallier à ce problème, LFMMs cherche à modéliser les fréquences alléliques à chaque locus comme la réponse mixte des effets fixes des variables environnementales observées et des effets latents de K variables non observées. Cela se traduit mathématiquement de la façon suivante :

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^{\mathbf{T}} + \mathbf{U}\mathbf{V}^{\mathbf{T}} + \epsilon \tag{2.5}$$

où $\mathbf{Y}$ représente la matrice de génotype de taille $n \times L$ avec n le nombre d'individus et L le nombre de locus, $\mathbf{X}$ représente la matrice des variables environnementales de taille $n \times d$ où d correspond au nombre de prédicteurs et $\mathbf{B}$ correspond aux tailles d'effets associés à $\mathbf{X}$, $\mathbf{U}$ de taille $n \times K$ correspond à la matrice des variables latentes et $\mathbf{V}$ correspond aux loadings associés à $\mathbf{U}$. Les applications de LFMMs sont nombreuses. En génomique du paysage, ils sont notamment utilsés pour la détection de SNPs adaptatifs.

6. 2. Utilisation du modèle pour la détection de SNPs

Une fois que nous avons estimé les tailles d'effets environnementaux en corrigeant pour les facteurs de confusion, il est ensuite possible de tester la significativité de ces tailles d'effets pour chacun des SNPs nous permettant donc d'identifier les SNPs considérés comme adaptatifs. Dans le cas où il y a plusieurs prédicteurs environnementaux, deux options s'offrent à nous : effectuer un test de significativité globale de la régression reposant sur un test de Fisher où l'hypothèse nulle correspond à "l'ensemble des coefficients de la régression sont nuls" ou effectuer un test sur chacun des coefficients reposant sur un test de Student. En effectuant un test de significativité globale, on obtient alors une p-valeur par SNPs que l'on peut résumer dans un Manhattan plot (Figure 2.7) et on peut ensuite sélectionner les SNPs en définissant un seuil de significativité.

6. 3. Conclusions et perspectives

Dans cette partie, nous avons présenté les modèles mixtes à facteur latents. En génétique du paysage, ils ont jusqu'ici été essentiellement utilisés pour la détection de SNPs adaptatifs. Dans le cadre de notre thèse, ils seront une composante essentielle de notre méthode fossé génétique. De plus, nous utiliserons les facteurs latents estimés par LFMM afin de corriger les différentes méthodes existantes pour le calcul de décalage génétique.

7. Conclusions

Ce chapitre remplit deux objectifs. Le premier objectif est d'inscrire le travail de thèse dans un contexte plus large. On y résume notamment l'historique des idées ayant donné lieu à l'apparition du concept de décalage génétique. Le décalage génétique s'inscrit dans la continuité du concept d'inadéquation évolutive qui correspond à un état de déséquilibre entre un organisme et son environnement. Le décalage génétique cherche à quantifier ce déséquilibre en s'appuyant sur des relations statistiques entre les fréquences d'allèles et les gradients environnementaux et en utilisant ces relations pour prédire les différences entre fréquences alléliques prédites à des paires de points dans la niche écologique. Il existe plusieurs méthodes de calcul du décalage génétique. Ces méthodes font face à certaines limites parmi lesquelles la gestion des facteurs de confusion et l'interprétabilité des mesures. Dans cette thèse, nous développons une nouvelle mesure de décalage génétique qui vise à répondre à ces limites. Nous travaillons notamment à établir une relation entre la mesure de décalage génétique d'une population lorsqu'elle fait face à un environnement modifié et la valeur sélective des individus de cette population dans cet environnement modifié. En parallèle du travail sur le décalage génétique, nous avons travaillé sur la relation existante entre les indices de fixation de Wright et l'ACP des données génétique. Ce travail étend notamment les résultats obtenus à partir de la théorie de la coalescence pour deux populations divergentes par McVean à tout modèle de population discrète. Le second objectif est d'introduire certaines notions essentielles pour la mise en place de nos résultats. Nous abordons notamment les hypothèses qui seront nécessaires pour établir la théorie quantitative du décalage génétique, le modèle infinitésimal et la sélection stabilisatrice gaussienne. Nous donnons également quelques références autour de la théorie des matrices aléatoires importantes pour la compréhension de la théorie spectrale des indices de fixation de Wright. Ce chapitre contient donc tous les éléments importants pour la compréhension de nos travaux de thèse présents dans les trois chapitres suivants.

Chapitre 3

Une nouvelle mesure de décalage génétique : le fossé génétique

Dans la partie sur l'état de l'art, nous avons présenté différentes méthodes existantes pour le calcul du décalage génétique. Nous avons identifié certaines limites liées à ces méthodes : les biais de prédictions que peuvent entraîner la structure de population, la difficulté pour certaines méthodes de prendre en compte l'aspect polygénique de l'adaptation. Nous avons aussi mis en évidence le manque d'interprétabilité des mesures. Dans ce chapitre, nous allons définir une nouvelle mesure de décalage génétique que nous appellerons fossé génétique visant à répondre aux limites citées précédemment. Nous allons unifier les méthodes présentées dans l'état de l'art en identifiant les liens statistiques existant entre ces dernières. Nous allons montrer qu'il existe une interprétation duale de notre nouvelle mesure, comme distance dans l'espace écologique et dans l'espace génétique. Nous allons également clarifier la relation théorique qui existe entre notre mesure et les traits adaptatifs en établissant une théorie quantitative du décalage génétique. Nos résultats offrent une vision unifiée des statistiques de décalage génétique, et soulignent l'importance que peuvent avoir ces dernières dans un contexte de conservation face aux changements environnementaux.

1. Introduction

La quantification du changement nécessaire pour qu'un trait phénotypique adapté à un environnement soit adapté à un environnement modifié est une question de longue date en écologie et évolution, on parle d'inadéquation évolutive (Schlaepfer et al., 2002) (Cook et Saccheri, 2013). Avec la disponibilité grandissante des données génomiques, on cherche désormais à déterminer ces changements à partir des effets environnementaux sur les locus du génome qui contrôlent les traits adaptatifs, contournant ainsi le besoin d'avoir accès à des mesures phénotypiques directes (Capblancq, Fitzpatrick et al., 2020) (Waldvogel et al., 2020). Cette approche permet d'informer sur la possibilité pour une population naturelle de résister à un changement soudain dans la niche écologique (Grinnell, 1917) (Sork et al., 2010) (F. Jay et al., 2012) (Aitken et Whitlock, 2013) (Schoville et al., 2012) (Foden et al., 2019). Les approches consistant à intégrer les données génomiques pour quantifier la maladaptation des populations face aux changements environnementaux sont appelées décalage génétique (Fitzpatrick et Keller, 2015). Les statistiques de décalage génétique estiment d'abord une relation statistique entre les gradients environnementaux et les fréquences alléliques à l'aide de modèles d'association génotype-environnement (GEA) (Forester et al., 2018). La relation déduite est ensuite utilisée pour évaluer les différences entre les fréquences alléliques prédites à des paires de points dans la niche écologique (Rellstab, Zoller et al., 2016) (Gougherty et al., 2021). Plusieurs méthodes ont été proposées pour le calcul du décalage génétique mais certaines limites ont été identifiées, les prédictions sont biaisées du fait de la structure de population, il est difficile de tenir compte de l'aspect polygénique de l'adaptation ou de la corrélation des prédicteurs environnementaux (Rellstab, Dauphin et al., 2021). De plus, les statistiques de décalage génétique manquent d'une théorie qui clarifierait la manière dont ces statistiques sont liées les unes aux autres, et qui préciserait la manière de les interpréter en terme biologique. Dans ce chapitre, nous proposons une nouvelle méthode de calcul du décalage génétique, appelée fossé génétique, et visant à répondre aux limites des méthodes précédentes. Nous proposons également une théorie permettant d'établir une relation entre le décalage génétique et la valeur sélective des individus subissant une modification brutale de leur environnement.

2. Une nouvelle mesure de décalage génétique : le fossé génétique

Cette partie vise à définir notre nouvelle mesure de décalage génétique et à en donner une interprétation à la fois comme distance dans la niche écologique et comme distance génétique.

2. 1. Définition mathématique du fossé génétique

Dans cette section, nous allons définir le calcul du fossé génétique. Nous donnons dans les sous-sections suivantes les différentes interprétations de cette nouvelle mesure. Considérons tout d'abord une matrice de génotype $\mathbf{Y}$ de taille $n \times L$ où n correspond au nombre d'individus et L correspond au nombre de locus. La matrice de génotype est une matrice centrée. Considérons également $\mathbf{X}$ une matrice de prédicteurs environnementaux de taille $n \times d$ où d correspond au nombre de prédicteurs. La matrice $\mathbf{X}$ est centrée et réduite de sorte que les prédicteurs sont sans unité. Cherchons ensuite à estimer un modèle de GEA, et plus précisément, un modèle mixte à facteurs latents tel que défini dans le chapitre 2 section 6 :

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^{\mathbf{T}} + \mathbf{U}\mathbf{V}^{\mathbf{T}} + \epsilon \tag{3.1}$$

On obtient alors une matrice de taille d'effets $\mathbf{B}$ de taille $L \times d$ où chaque entrée correspond aux effets de chaque prédicteur environnemental $j = 1, \ldots, d$ sur les fréquences d'allèles à chaque locus génétique $\ell = 1, \ldots, L$. On définit alors la matrice $\mathbf{C_b} = \mathbb{E}[\mathbf{b}\mathbf{b}^T]$. Dans tout ce chapitre, on supposera que la valeur moyenne des tailles d'effets le long du génome est nulle $\mathbb{E}[\mathbf{b}] \approx 0$. Du fait de la définition arbitraire de l'allèle de référence, cette approximation est valide dans la plupart des jeux de données empiriques. Ainsi, $\mathbf{C_b}$ peut être vu comme la matrice de covariance des tailles d'effets de taille $d \times d$. La notation mathématique $\mathbb{E}[.]$ correspond à la moyenne sur les L locus inclus dans l'analyse. Si $\mathbf{q}_l$ est une quantité évaluée au locus ℓ, alors $\mathbb{E}[\mathbf{q}] = \sum_{l=1}^{L} \mathbf{q}_\ell / L$. Soient $\mathbf{x}$ et $\mathbf{x}^\star$ deux vecteurs environnementaux de taille d. Nous allons définir le produit scalaire $\langle \mathbf{x}, \mathbf{x}^\star \rangle_\mathbf{b} = \mathbf{x}\mathbf{C_b}\mathbf{x}^{\star T}$. Nous disposons désormais de tous les éléments pour définir le fossé génétique :

$$G^2(\mathbf{x}, \mathbf{x}^\star) = \langle \mathbf{x} - \mathbf{x}^\star, \mathbf{x} - \mathbf{x}^\star \rangle_\mathbf{b} = (\mathbf{x} - \mathbf{x}^\star)\mathbf{C_b}(\mathbf{x} - \mathbf{x}^\star)^T = \|\mathbf{x} - \mathbf{x}^\star\|_\mathbf{b}^2 . \tag{3.2}$$

Nous disposons ainsi d'une première définition mathématique du décalage génétique. Voyons désormais la manière de l'interpréter.

2. 2. Interprétation comme distance dans la niche écologique pondérée par la génétique

Une manière naïve de quantifier la maladaptation d'une population à un nouvel environnement est le calcul de la distance euclidienne entre l'environnement actuel de la population, mesuré par $\mathbf{x}$ et son nouvel environnement, mesuré par $\mathbf{x}^\star$. Cette approche prend en compte le changement environnemental mais n'utilise pas l'information génétique. Elle peut alors donner lieu à des résultats biaisés lorsqu'on intègre dans $\mathbf{x}$ des prédicteurs qui n'ont pas d'impact sur l'adaptation locale, ou encore des prédicteurs corrélés entre eux. La racine carrée du fossé génétique définie également une distance mathématique entre les vecteurs $\mathbf{x}$ et $\mathbf{x}^\star$ mais cette distance est pondérée par les effets des prédicteurs environnementaux sur les fréquences alléliques permettant de pallier au problème que nous venons de citer.

2. 3. Interprétation comme distance génétique

En sous-section **2.1** de ce chapitre, nous avons établi la définition mathématique du fossé génétique. Nous allons voir dans cette section que cette formule peut être interprétée comme une mesure de distance génétique entre la population adaptée aux conditions environnementales $\mathbf{x}$ et une population similaire mais adaptée aux conditions environnementales $\mathbf{x}^\star$. Commençons par définir $f(\mathbf{x}) = \mathbf{x}\mathbf{b}^T + \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_k^T$ correspondant à la fréquence d'allèle prédite par le modèle LFMM pour l'environnement $\mathbf{x}$. On interprète alors $f(\mathbf{x})$ comme une fréquence allélique dans la population adaptée aux conditions environnementales $\mathbf{x}$. On définit de façon similaire $f(\mathbf{x}^\star) = \mathbf{x}^\star\mathbf{b}^T + \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_k^T$ pouvant être interprétée comme une fréquence allélique dans la population adaptée aux nouvelles conditions environnementales $\mathbf{x}^\star$. Le but de cette sous section est de montrer que :

$$G^2(\mathbf{x}, \mathbf{x}^\star) = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2] ,$$

Commençons par rappeler que $\mathbb{E}[\mathbf{b}] \approx 0$. Il est toutefois possible d'étendre les résultats lorsque cette condition n'est pas respectée. Sous cette condition, on peut définir la variance de $(\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T$ le long du génome de la manière suivante :

$$\mathbb{V}\mathrm{ar}[(\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T] = (\mathbf{x} - \mathbf{x}^\star)\mathbf{C_b}(\mathbf{x} - \mathbf{x}^\star)^T .$$

et on obtient également l'égalité suivante :

$$\mathbb{V}\mathrm{ar}[(\mathbf{x}-\mathbf{x}^\star)\mathbf{b}^T] = \mathbb{E}[((\mathbf{x}-\mathbf{x}^\star)\mathbf{b}^T)^2] = \mathbb{E}[(\mathbf{x}\mathbf{b}^T - \mathbf{x}^\star\mathbf{b}^T)^2]\,,$$

Or

$$\mathbb{E}[(\mathbf{x}\mathbf{b}^T - \mathbf{x}^\star\mathbf{b}^T)^2] = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]\,.$$

Ainsi, on obtient bien :

$$G^2(\mathbf{x},\mathbf{x}^\star) = \mathbb{V}\mathrm{ar}[(\mathbf{x}-\mathbf{x}^\star)\mathbf{b}^T] = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]\,.$$

Le fossé génétique peut donc être vu comme une distance génétique entre deux populations adaptées respectivement aux conditions environnementales $\mathbf{x}$ et $\mathbf{x}^\star$.

2. 4. Valeurs propres, vecteurs propres et importance des variables

Dans cette sous-section, nous proposons un moyen pour évaluer l'importance des différentes variables environnementales dans l'adaptation locale. Rappelons que la matrice de variables environnementales $\mathbf{X}$ est standardisée. On travaille ensuite une fois encore sur la matrice de covariance des tailles d'effets, et, plus précisément, sur les valeurs propres et vecteurs propres de cette dernière. On peut alors trier les vecteurs propres par la taille de leur valeur propre associée. On nommera λ_i la valeur propre associée au i ème vecteur propre. On récupère alors également le coefficient de la k ème variable environnementale, associé au i ème vecteur propre, μ_{ik}. Le coefficient μ_{ik} élevé au carré nous donne la contribution en pourcentage de la k ème variable environnementale au sein du i ème vecteur propre. Il est alors possible de s'intéresser à la contribution d'une variable environnementale pour chacun des vecteurs propres et de pondérer cette dernière par la valeur propre associée à chacun des vecteurs afin d'obtenir une mesure globale de l'importance d'une variable. On résume donc l'importance de la variable k, I_k, de la façon suivante :

$$I_k = \sum_{i=1}^{d} \lambda_i \times \mu_{ik}^2$$

Cette variable nous donne une estimation de l'importance globale d'une variable environnementale. On peut également obtenir le détail de cette estimation en regardant la contribution de la variable dans chacun des vecteurs propres.

2. 5. Conclusions

Dans cette section, nous avons donné la définition mathématique du fossé génétique. Cette mesure permet de résoudre le problème des biais causés par la structure de population. En effet, les effets $\mathbf{B}$ sont estimés à l'aide de LFMM permettant de corriger pour les facteurs de confusion liés à la structure de population. Nous avons montré que le fossé génétique pouvait être interprété comme une distance géométrique dans la niche écologique pondérée par les effets de l'adaptation locale. Nous avons également donné une définition équivalente du décalage génétique comme distance génétique entre deux matrices de génotypes adaptés à des environnements différents. Dans la section suivante, nous allons établir la relation théorique entre le fossé génétique et la valeur sélective de l'organisme qui fait face à un nouvel environnement.

3. Une théorie quantitative du fossé génétique

L'objectif de cette section est d'établir le lien qui existe, sous certaines hypothèses, entre le fossé génétique et la valeur sélective d'un trait contrôlé par les locus servant à le calculer. Pour ce faire, nous allons d'abord rappeler les hypothèses et les notations nécessaires à l'établissement de cette théorie. Nous utiliserons ensuite le modèle infinitésimal pour montrer le lien entre fossé génétique et trait adaptatif. Nous utiliserons la sélection stabilisatrice gaussienne pour étendre directement ce lien à la valeur sélective. Nous utiliserons enfin la théorie du fardeau génétique pour retrouver cette relation d'une manière différente. Nous obtiendrons ainsi une théorie définissant la relation entre le fossé génétique et la valeur sélective d'un trait adapté à l'environnement $\mathbf{x}$ et plongé dans l'environnement $\mathbf{x}^\star$.

3. 1. Hypothèses, rappels théoriques, et notations

Dans la section **4** du chapitre 2, nous avons présenté la sélection stabilisatrice gaussienne et le modèle infinitésimal de Fisher. Pour démontrer notre théorie quantitative du fossé génétique, nous utiliserons ces théories comme hypothèse de départ. Cette première sous-section vise à rappeler ces hypothèses, à introduire certaines notations et à fournir les éléments essentiels à la compréhension de notre théorie quantitative du décalage génétique.

Notations $\bar{z}$ et $\bar{z}^\star$ On utilisera dans cette section les notations $\bar{z}$ et $\bar{z}^\star$. Ces notations correspondent à la valeur du trait de valeur sélective conditionnée respectivement à $\mathbf{x}$ et $\mathbf{x}^\star$. Autrement dit, $\bar{z} = \mathrm{E}[z|\ \mathbf{x}]$ et $\bar{z}^\star = \mathrm{E}[z|\ \mathbf{x}^\star]$. Ici la notation E[.] désigne l'espérance sur la population.

Notation $\omega(\mathbf{x}, \mathbf{x}^\star)$ Dans ce chapitre nous cherchons à lier le fossé génétique à la valeur sélective d'un individu dont les traits phénotypiques sont adaptés à l'environnement $\mathbf{x}$ et qui est placé dans un environnement $\mathbf{x}^\star$. Nous noterons cette valeur $\omega(\mathbf{x}, \mathbf{x}^\star)$.

Modèle infinitésimal de Fisher Notre théorie repose sur le fait que les traits liés à la valeur sélective suivent le modèle infinitésimal de Fisher. Plus précisément, nous allons supposer que la composante génétique de la valeur des traits est la suivante :

$$z = \sum_{\ell=1}^{L} a_\ell y_\ell \tag{3.3}$$

avec $a_\ell \approx \pm\sqrt{a/L}$ et où y_ℓ est la fréquence d'allèle au locus ℓ. L'hypothèse selon laquelle les traits adaptatifs suivent le modèle infinitésimal de Fisher vont nous permettre d'exprimer l'espérance de la différence des traits de valeur sélective conditionnellement aux variables environnementales en fonction du fossé génétique.

Sélection stabilisatrice gaussienne Nous supposons ensuite que la valeur sélective est déterminée par la sélection stabilisatrice gaussienne. Autrement dit, le lien entre la valeur sélective dans le nouvel environnement $\omega(\mathbf{x}, \mathbf{x}^\star)$ et la différence des traits de valeur sélective est le suivant.

$$\omega(\mathbf{x}, \mathbf{x}^\star) = \exp -\frac{S(\bar{z}^\star - \bar{z})^2}{2}$$

Ces éléments permettront alors de déterminer une expression de $\omega(\mathbf{x}, \mathbf{x}^\star)$ en fonction du fossé génétique.

Fardeau génétique On peut voir le fardeau génétique comme un fardeau migratoire, qui survient lorsque la migration remplace les allèles locaux par des allèles étrangers censés être moins adaptés que les allèles locaux. Dans la section **4.3.4**, on interprétera un changement environnemental de $\mathbf{x}$ à $\mathbf{x}^\star$ comme une migration géographique d'un lieu d'environnement $\mathbf{x}$ à un lieu d'environnement $\mathbf{x}^\star$. On retrouvera alors à nouveau le lien entre fossé génétique et valeur sélective mais en s'appuyant cette fois sur des arguments liés au calcul du fardeau génétique, nous offrant ainsi une interprétation alternative du fossé génétique.

Epistasie multiplicative L'épistasie multiplicative signifie que la valeur sélective d'un individu est déterminée par le produit de tous les effets locus-spécifique. Autrement dit :

$$\omega(\mathbf{x}, \mathbf{x}^\star) = \prod_{\ell=1}^{L} \omega_\ell(\mathbf{x}, \mathbf{x}^\star)\,.$$

3. 2. Lien entre trait de valeur sélective et fossé génétique

Dans cette sous-section, nous allons chercher à exprimer $(\bar{z}^\star - \bar{z})^2$ en fonction du fossé génétique. Compte tenu du fait que le trait suit un modèle infinitésimal de Fisher, on peut l'écrire de la façon suivante :

$$z = \sum_{l=1}^{L} a_l y_l.$$

En estimant la fréquence allélique au locus ℓ pour les environnements $\mathbf{x}$ et $\mathbf{x}^\star$ de la manière suivante : $f_\ell(\mathbf{x}) = \mathbf{x}\mathbf{b}_\ell^T + \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_{k\ell}^T$ et $f_\ell(\mathbf{x}^\star) = \mathbf{x}^\star\mathbf{b}_\ell^T + \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_{k\ell}^T$, on obtient alors

$$\bar{z} = \sum_{\ell=1}^{L} a_\ell f_\ell(\mathbf{x})\,,$$

et

$$\bar{z}^\star = \sum_{\ell=1}^{L} a_\ell f_\ell(\mathbf{x}^\star)\,.$$

On peut donc calculer l'expression de la différence des traits de valeur sélective de la façon suivante :

$$\bar{z}^\star - \bar{z} = (\mathbf{x}^\star - \mathbf{x}) \sqrt{\frac{a}{L}} \sum_{\ell=1}^{L} \mathbf{b}_\ell^T\,.$$

On peut alors appliquer le théorème central limite et on obtient que la distribution de $\bar{z}^\star - \bar{z}$ est une gaussienne de moyenne 0 et de variance $a\mathbb{V}\mathrm{ar}[(\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T]$. Dans la section **2.3** de ce chapitre on a montré que $\mathbb{V}\mathrm{ar}[(\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T] = G^2(\mathbf{x}, \mathbf{x}^\star)$. On obtient donc finalement

$$\bar{z}^\star - \bar{z} |\ \mathbf{x}, \mathbf{x}^\star \sim N(0, aG^2(\mathbf{x}, \mathbf{x}^\star))\,,$$

On peut ensuite déterminer la distribution conditionnelle de la différence élevée au carrée. Il s'agit d'une loi du χ^2 avec un degré de liberté. On obtient finalement une expression de $(\bar{z}^\star - \bar{z})^2$ en fonction du fossé génétique :

$$(\bar{z}^\star - \bar{z})^2 |\ \mathbf{x}, \mathbf{x}^\star \sim aG^2(\mathbf{x}, \mathbf{x}^\star)\, \chi_1^2.$$

Nous avons donc établi une relation entre le fossé génétique et la différence de trait adaptatif. On cherche désormais à exploiter cette relation pour établir une relation entre le fossé génétique et la valeur sélective d'un trait optimal dans l'environnement $\mathbf{x}$ placé dans l'environnement $\mathbf{x}^\star$

3. 3. Lien entre valeur sélective et fossé génétique

Nous disposons désormais d'un lien entre trait phénotypique lié à la valeur selective et fossé génétique. En utilisant la théorie de la sélection stabilisatrice gaussienne, nous allons pouvoir faire le lien entre le trait et la valeur sélective et donc, finalement, un lien entre valeur sélective dans l'environnement modifié et décalage génétique. Dans la sous-section **3.2**, on a établi que :

$$\omega(\mathbf{x}, \mathbf{x}^\star) = \exp -\frac{S(\bar{z}^\star - \bar{z})^2}{2}$$

En transformant cette expression, on obtient :

$$-2\log\omega(\mathbf{x}, \mathbf{x}^\star) = S(\bar{z}^\star - \bar{z})^2,$$

A l'aide des résultats de la sous section précédente, et en remplaçant la distribution du χ^2 par son espérance (égale à 1), on obtient alors :

$$-2\log\omega(\mathbf{x}, \mathbf{x}^\star) = aS\, G^2(\mathbf{x}, \mathbf{x}^\star)\,,$$

On obtient bien un lien entre valeur sélective dans le nouvel environnement et fossé génétique. Pour la suite de ce manuscrit, on posera $s = aS$ de telle sorte qu'en moyenne :

$$-2\log\omega(\mathbf{x}, \mathbf{x}^\star) = s\, G^2(\mathbf{x}, \mathbf{x}^\star)\,,$$

On obtient alors l'un des résultats essentiels de ce chapitre, il existe une relation linéaire entre le fossé génétique et la valeur sélective d'un individu placé dans un environnement modifié :

$$G^2(\mathbf{x}, \mathbf{x}^\star) = -2\log(\omega(\mathbf{x}, \mathbf{x}^\star))/s\,.$$

Ce résultat est résumé dans la figure 3.1.

3. 4. Obtention du résultat en s'appuyant sur la théorie du fardeau génétique

Dans cette sous section, on cherche à retrouver le résultat précédent en s'appuyant sur la théorie du fardeau génétique. Sous sélection stabilisatrice gaussienne, la valeur sélective au locus ℓ est décrite par le gradient de sélection suivant :

$$\omega_\ell(\mathbf{x}, \mathbf{x}^\star) = \exp\left(-\frac{s}{2L}(f_\ell(\mathbf{x}) - f_\ell(\mathbf{x}^\star))^2\right)\,.$$

Sous l'hypothèse de l'épistatsie multiplicative, et par les propriétés de l'exponentielle :

$$\omega(\mathbf{x}, \mathbf{x}^\star) = \exp\left(-\frac{s}{2}\mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]\right)\,.$$

Or, on a montré dans la sous-section **2.3** de ce chapitre que :

$$G^2(\mathbf{x}, \mathbf{x}^\star) = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]\,.$$

On retrouve alors que le fossé génétique est relié linéairement au logarithme de la valeur sélective dans un environnement modifié :

$$sG^2(\mathbf{x}, \mathbf{x}^\star) = -2\log(\omega(\mathbf{x}, \mathbf{x}^\star))\,.$$

Cela permet une interprétation du fossé génétique comme fardeau génétique environnemental, où les mutations délétères sont remplacées par des changements des conditions environnementales ayant des effets délétères sur la viabilité.

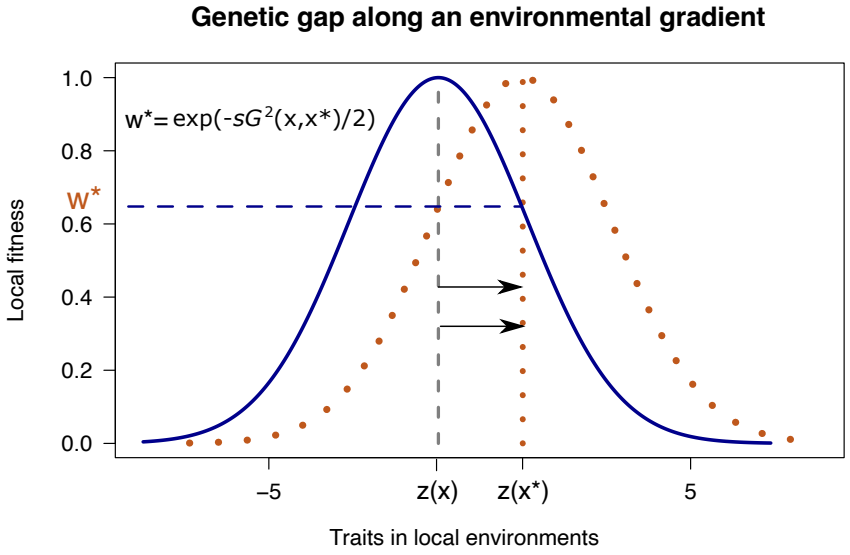

Figure 3.1 – **Variation de la valeur sélective en fonction du fossé génétique dans le cas de la sélection stabilisatrice gaussienne**

Les deux points, $z(\mathbf{x})$ et $z(\mathbf{x}^\star)$, représente les valeurs d'optimum local d'un trait adaptatif pour les environnements de valeurs respectives $\mathbf{x}$ and $\mathbf{x}^\star$. La courbe affiche les valeurs sélectives pour le trait dans chacun des environnements. Un organisme avec le trait $z(\mathbf{x})$, optimal pour l'environnement $\mathbf{x}$, étant placé dans un environnement modifié $\mathbf{x}^\star$, a alors une valeur sélective égale à $\omega^\star = \exp(-sG^2(\mathbf{x},\mathbf{x}^\star)/2)$ où $G^2(\mathbf{x},\mathbf{x}^\star)$ est le fossé génétique et s est l'intensité de sélection.

3. 5. Prolongement des résultats au cas des variables non causales

Nous avons supposé jusqu'ici que les prédicteurs environnementaux observés influencent directement les traits sous sélection. Le but de cette sous section est d'étendre ce résultat à des variables $\mathbf{x}$ définies comme combinaison linéaire ($\mathbf{A}$) des variables causales $\mathbf{x}'$, $\mathbf{x} = \mathbf{x}'\mathbf{A}^T$. Les coefficients de $\mathbf{A}$ sont inconnus. On peut alors montrer que le fossé génétique, calculé à partir des prédicteurs non causaux, est équivalent au fossé génétique calculé avec les variables causales.

$$G^2(\mathbf{x},\mathbf{x}^\star) = (\mathbf{x}' - \mathbf{x}'^\star)\mathbb{E}[\mathbf{c}^T\mathbf{c}](\mathbf{x}' - \mathbf{x}'^\star)^T\,,$$

où $\mathbf{c} = \mathbf{bA}$ correspond aux tailles d'effets pour les variables causales non observées. Cette formule signifie que notre définition du fossé génétique est robuste aux corrélations dans les effets causaux et fonctionne avec une combinaison inconnue de ces effets. Comme la sélection stabilisatrice agit sur les variables causales, le fossé génétique est proportionnel à la distance au carrée entre les environnements causaux dans la niche écologique.

$$G^2(\mathbf{x},\mathbf{x}^\star) \propto (\mathbf{x}' - \mathbf{x}'^\star)\mathbf{C}^{-1}(\mathbf{x}' - \mathbf{x}'^\star)^T\,,$$

où $\mathbf{C}$ est la matrice de covariance inconnue des effets causaux.

3. 6. Conclusions

En conclusion, nous avons développé, sous certaines hypothèses, une théorie quantitative du fossé génétique. Le grand résultat de cette partie est le lien établi entre fossé génétique et le logarithme de la valeur sélective dans l'environnement modifié :

$$G^2(\mathbf{x},\mathbf{x}^\star) = -2\log(\omega(\mathbf{x},\mathbf{x}^\star))/s\,. \tag{3.4}$$

Ce résultat a été obtenu sur la base du modèle infinitésimal sous sélection stabilisatrice gaussienne. Nous avons également obtenu ce résultat avec une approche basée sur le fardeau génétique. De plus, nous avons montré que ce résultat restait valide lorsque les prédicteurs environnementaux étaient reliés indirectement aux facteurs qui influencent les traits sous sélection. En conclusion, le fossé génétique est robuste aux corrélations des effets causaux, et les résultats sont valides pour des combinaisons linéaires de ces effets.

4. Unification de différentes mesures de décalage génétique

Dans cette section, nous allons nous servir de la théorie présentée dans la section précédente afin d'unifier les différentes mesures de décalage génétique existantes. Nous allons notamment comparer notre mesure de fossé génétique avec les méthodes présentées au chapitre 2 section **2.3**, le décalage génétique basé sur la redundancy analysis (RDA), Risk of non-adaptedness (Rona) et gradient forests (GF).

4. 1. Lien avec Rona

Il est possible de lier la définition du fossé génétique avec celle de Rona. Toutefois, pour ce faire, il est nécessaire d'apporter quelques modifications par rapport à la version présentée par Rellstab, Zoller et al., 2016. La définition originale de Rona était basée sur les prédictions d'un modèle linéaire simple pour chaque prédictions. Nous proposons ici une régression linéaire multiple sur la matrice de tous les prédicteurs environnementaux $\mathbf{X}$. De plus, afin de supprimer les biais dûs à la structure de population, Rona peut être calculé à partir des prédictions de LFMM à la place d'une régression linéaire classique. On retrouve une implémentation de Rona identique à celle décrite ici dans le papier de Aquino et al., 2022. Avec ces modifications, on peut réécrire Rona de la façon suivante :

$$\text{Rona}(\mathbf{x}, \mathbf{x}^\star) = \mathbb{E}[|f(\mathbf{x}) - f(\mathbf{x}^\star)|] = \mathbb{E}[|(\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T|]\,,$$

où les $\mathbf{b}$ correspondent aux tailles d'effets estimées avec LFMM. Contrairement au fossé génétique, Rona calcule la différence en valeur absolue des fréquences d'allèles plutôt que la différence au carré. Afin d'expliquer Rona dans un modèle de sélection stabilisatrice, supposons que la courbe de valeur sélective dans l'environnement modifié est décrite par la courbe exponentielle

$$\omega_\ell(\mathbf{x}, \mathbf{x}^\star) = \exp\left(-\frac{s}{L}|f_\ell(\mathbf{x}) - f_\ell(\mathbf{x}^\star)|\right)\,,$$

où les fréquences alléliques sont estimées à partir des prédicteurs environnementaux dans les environnements observés et modifiés à chaque locus. On peut à nouveau exprimer la valeur sélective dans l'environnement modifié comme le produit de tous les effets locus spécifique et on obtient alors :

$$\omega(\mathbf{x}, \mathbf{x}^\star) = \exp\left(-s\,\text{Rona}(\mathbf{x}, \mathbf{x}^\star)\right)\,.$$

Tout comme le fossé génétique, Rona est relié linéairement au logarithme de la valeur sélective dans l'environnement modifié. La différence entre les deux mesures réside dans les modèles de sélection stabilisatrice, gaussienne pour le fossé génétique, exponentielle pour Rona. De plus, les deux mesures sont reliés par l'inégalité de Cauchy-Schwartz, Rona est toujours plus petit que la racine carrée du fossé génétique,

$$\mathbb{E}[|f(\mathbf{x}) - f(\mathbf{x}^\star)|] \leq \sqrt{\mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]}\,.$$

4. 2. Lien avec RDA

La définition du fossé génétique comme distance dans la niche écologique est proche de celle de RDA que nous avons décrit dans la sous section **3.2** du chapitre 2. Dans cette partie nous allons supposer que 1) La régression, première étape de la méthode RDA, est réalisée en utilisant les d prédicteurs environnementaux et les K facteurs latents obtenus avec LFMM. 2) Les projections dans l'espace de la RDA incluent les $(d + K)$ composantes RDA, autrement dit, autant de composantes que la dimension de l'espace latent et environnemental. Sous ces conditions, on fournit dans cette sous section une preuve mathématique de l'équivalence entre le décalage génétique élevé au carré de la RDA et du fossé génétique. Le coeur de la démonstration repose sur une propriété générale de la projection de l'ACP qui préserve la distance euclidienne.
Sous la première hypothèse, les predicteurs linéaires des fréquences alléliques sont définis dans la matrice $\mathbf{XB}^T + \mathbf{UV}^T$. L'ACP décompose les valeurs prédites de la façon suivante :

$$(\mathbf{XB}^T + \mathbf{UV}^T)/\sqrt{n-1} = \mathbf{QP}^T\,,$$

où $\mathbf{Q}$ correspond aux scores des composantes principales (ou projections des échantillons) et $\mathbf{P}$ correspond aux loadings (de taille $L \times n$). Pour tout vecteur de dimension d de prédicteurs environnementaux, $\mathbf{x}$, et facteur de dimensions K, $\mathbf{u}$, la projection dans l'espace RDA est donnée par

$$\mathbf{proj}(\mathbf{x}) = (\mathbf{xB}^T + \mathbf{uV}^T)\mathbf{P}\,.$$

De manière analogue, pour un vecteur de prédicteurs environnementaux $\mathbf{x}^\star$, on obtient la projection suivante

$$\mathbf{proj}(\mathbf{x}^\star) = (\mathbf{x}^\star\mathbf{B}^T + \mathbf{uV}^T)\mathbf{P}\,.$$

$\mathbf{P}$ étant une transformation linéaire, le terme $\mathbf{uV}^T$ disparait lorsqu'on effectue la différence des projections, et on obtient

$$\mathbf{proj}(\mathbf{x}^\star) - \mathbf{proj}(\mathbf{x}) = (\mathbf{x}^\star - \mathbf{x})\mathbf{B}^T\mathbf{P}\,.$$

La distance euclidienne au carré dans l'espace RDA est alors obtenu comme suit,

$$\|\mathbf{proj}(\mathbf{x}^\star) - \mathbf{proj}(\mathbf{x})\|^2 = (\mathbf{x}^\star - \mathbf{x})\mathbf{B}^T\mathbf{P}\mathbf{P}^T\mathbf{B}(\mathbf{x}^\star - \mathbf{x})^T\,.$$

Par les propriétés de l'ACP, $\mathbf{P}$ est une matrice semi-orthogonale, de telle sorte que

$$\mathbf{PP}^T = \mathbf{I}\,,$$

où $\mathbf{I}$ est la matrice identité de taille $L \times L$. De plus, on a

$$\mathbb{E}[\mathbf{b}^T\mathbf{b}] = \mathbf{B}^T\mathbf{B}/L\,.$$

Ainsi, on obtient

$$\frac{1}{L} \times \|\mathbf{proj}(\mathbf{x}^\star) - \mathbf{proj}(\mathbf{x})\|^2 = (\mathbf{x}^\star - \mathbf{x})\mathbb{E}[\mathbf{b}^T\mathbf{b}](\mathbf{x}^\star - \mathbf{x})^T = G^2(\mathbf{x}, \mathbf{x}^\star)\,.$$

En d'autres termes, le fossé génétique est égal à la distance euclidienne au carré dans l'espace RDA divisé par le nombre de locus considérés dans l'analyse.

4. 3. Lien avec Gradient Forests

GF est une méthode non paramétrique non linéaire basée sur les forêts aléatoires. L'approche ne suppose aucun modèle de sélection à priori, mais infère le modèle des données observées. Dans la sous section **3.2** du chapitre 2, on explique que le décalage génétique GF correspond à la distance euclidienne entre les vecteurs des projections des prédicteurs environnementaux $\mathbf{x}$ et $\mathbf{x}^\star$. On avait appelé F_j la fonction qui, pour une valeur du prédicteur j, donnait sa projection. On a alors

$$\mathrm{GF}^2(\mathbf{x}, \mathbf{x}^\star) = \sum_{j=1}^{d} (F_j(x_j) - F_j(x_j^\star))^2$$

La construction du décalage génétique de GF partage des similarités avec le fossé génétique et le décalage génétique RDA. En effet, on peut voir le fossé génétique comme la projection pondérée par les valeurs propres de la matrice de covariance, représentant l'importance des prédicteurs environnementaux inférée des données.

$$G^2(\mathbf{x}, \mathbf{x}^\star) = \sum_{j=1}^{d} (\mathrm{proj}_j(\mathbf{x}) - \mathrm{proj}_j(\mathbf{x}^\star))^2 \,.$$

Une différence importante entre le fossé génétique et le décalage génétique GF est la capacité de GF à gérer une réponse non linéaire des fréquences alléliques aux gradients environnementaux. La non linéarité fait de GF un modèle plus flexible que les modèles linéaires. Toutefois, les degrés additionnels de liberté ne sont pas toujours souhaitables lorsqu'on fait face au compromis biais variance. Le choix de la méthode dépend donc de notre cas d'usage (quantité de données, type de réponse des fréquences alléliques..)

5. Conclusions

Dans ce chapitre, nous définissons une nouvelle statistique de décalage génétique reposant sur LFMM que nous avons nommé fossé génétique. Nous établissons de plus une théorie quantitative du fossé génétique. Cette théorie géométrique fournit un cadre unifié qui permet une meilleure compréhension des mesures de décalage génétique existantes. En nous basant sur les théories de la sélection stabilisatrice gaussienne et du modèle infinitésimal de Fisher, nous montrons que le fossé génétique croit linéairement avec le logarithme de la valeur sélective dans l'environnement modifié. Nous donnons une interprétation duale du fossé génétique à la fois comme distance au carrée dans l'espace environnemental et dans l'espace génétique. Le fossé génétique correspond à une mesure classique de génétique des populations, la valeur moyenne de la D_{ST} de Nei, évaluant la diversité génétique entre la population adaptée aux conditions initiales et celle adaptée aux conditions modifiées. En conclusion, dans ce chapitre, nous répondons à un certain nombre des limites identifiées par Rellstab dans sa review (Rellstab, Dauphin et al., 2021). Le fossé génétique résoud le problème des prédictions biaisées par la corrélation existant entre structure de population et prédicteurs environnementaux en utilisant les facteurs latents comme covariables dans le modèle de prédiction. Le problème des corrélations existants entre prédicteurs est adressé par la modélisation de la covariance des tailles d'effets. Les valeurs propres et les vecteurs propres de la matrice de covariance nous offre un moyen de quantifier l'importance des prédicteurs environnementaux impliqués dans le processus d'adaptation locale. Enfin, nous fournissons une théorie qui permet une unification et une interprétation des statistiques de décalage génétique.

Chapitre 4

Validation des méthodes par la simulation et les données réelles

Dans le chapitre précédent, nous avons établi plusieurs résultats théoriques importants des statistiques de décalage génétique. Dans cette partie, nous chercherons à valider et consolider la compréhension de ces résultats par l'expérience. Nous utiliserons pour cela deux types de données, des données réelles et des données simulées. Les données simulées ont été obtenues avec le logiciel SLiM 3 (Haller et Messer, 2019). Les données réelles sont des données de mil, une céréale nutritive cultivée sur des sols arides en Afrique subsaharienne (Rhoné et al., 2020). Dans les deux cas, la validation reposera sur la comparaison des valeurs de décalage génétique avec le logarithme de la valeur sélective, ou d'une approximation de cette dernière dans le cas des données réelles. Nous allons tout d'abord présenter les caractéristiques générales de nos simulations. Nous proposerons ensuite de valider certains des résultats théoriques présentés dans le chapitre précédent. Puis, nous comparerons la performance des méthodes sur différents scénarios de simulation. Enfin, nous chercherons à tester les différentes statistiques de décalage génétique sur les données réelles. Ce chapitre offre une validation empirique des résultats présentés au chapitre précédent et vient confirmer l'utilité du recours au décalage génétique pour la compréhension des impacts du changement climatique sur la biodiversité.

1. Introduction

De nombreuses études ont cherché à valider empiriquement les statistiques de décalage génétique. Pour ce faire, ces études cherchent à comparer des mesures de traits phénotypiques d'individus évoluant dans des jardins communs ou des mesures de recensement de populations naturelles avec les mesures de décalage génétique (Bay et al., 2018) (Ruegg et al., 2018) (Fitzpatrick, Chhatre et al., 2021) (Y. Chen et al., 2022) (Sang et al., 2022). Les études de jardins communs visent à reproduire expérimentalement un changement environnemental abrupt pour des populations en extrayant les populations de leur environnement d'origine et en les faisant évoluer dans les conditions environnementales du jardin commun. On s'intéresse alors aux valeurs des traits phénotypiques dans le jardin commun et on compare finalement les valeurs de décalage génétique à ces traits phénotypiques. Ces études montrent alors un lien de corrélation entre le décalage génétique et les changements phénotypiques. Par exemple, Fitzpatrick, Chhatre et al., 2021 montrent qu'il existe une corrélation négative entre le décalage génétique et la taille du peuplier baumier (*Populus balsamifera*). Une étude de Bay et al., 2018 montre une corrélation entre les valeurs de décalage génétique pour le climat de 2050 avec le scénario RCP2.6 (Giec, 2017) pour les populations de paruline jaune dans leur aire de reproduction située en Amérique du Nord avec les tendances démographiques actuelles de ces populations. Dans le prolongement de ces études, nous proposons dans ce chapitre d'utiliser une expérience de jardin commun de mil (*Pennisetum glaucum*). Nous nous servons de cette expérience pour comparer les valeurs de décalage génétique obtenues avec les différentes méthodes avec le poids total moyen de graines obtenues pour chacune des populations cultivées dans le jardin. En parallèle de validation empirique à l'aide de données réelles, nous proposons une validation par la simulation. L'utilisation de la simulation est intéressante puisque cette dernière est moins coûteuse en temps et en ressource que la récolte de données réelles et permet d'explorer une multitude de scénarios différents. Une étude de Làruson et al., 2022 propose d'étudier la relation entre le déclin de valeur sélective en réponse à un changement environnemental brutal et le décalage génétique calculé avec la méthode GF. Nous proposons dans ce chapitre d'étudier également cette relation pour l'ensemble des mesures de décalage génétique. Nous proposons d'étudier cette relation avec et sans correction des méthodes pour la structure de population. Nous étudions différents types de scénarios avec une adaptation faiblement et fortement polygénique et différents degrés de corrélation entre les variables environnementales et les composantes principales. Ces différentes expériences offrent des critères objectifs de comparaison des méthodes de décalage génétique et permettent de valider empiriquement la théorie établie dans le chapitre précédent.

2. Présentation de l'outil de simulation et détails des scénarios

Pour tester les mesures de décalage génétique, nous avons fait le choix d'utiliser la simulation. Plus précisément, nous utilisons le logiciel SLiM 3.7 (Haller et Messer, 2019) pour effectuer des simulations spatialement explicites, basées sur l'individu. Dans tout ce chapitre, le scénario général des simulations reste le même. Nous allons simuler l'adaptation locale des individus dans différents environnements, $\mathbf{X}$. Nous simulerons ensuite un changement brutal de $\mathbf{X}$ vers $\mathbf{X}^\star$ et nous relèverons les nouvelles valeurs sélectives des individus suite à ce changement. Nous pourrons alors comparer les valeurs de décalage génétique calculées pour la variation d'environnement de $\mathbf{X}$ vers $\mathbf{X}^\star$ avec le logarithme de la valeur sélective des individus dans l'environnement $\mathbf{X}^\star$. Nous résumons dans cette section toutes les caractéristiques générales de

nos simulations (Figure 4.2). Les différences spécifiques à chacun des scénarios seront présentées dans les sections suivantes.

Géographie et gradients environnementaux Toutes les simulations ont lieu dans un carré de côté 10 unités. Il y a systématiquement deux variables environnementales, x_1 dont le gradient varie d'Est en Ouest, et x_2 dont le gradient varie du Sud au Nord. On définit donc $\mathbf{x} = (x_1, x_2)$ Ces gradients environnementaux influencent la viabilité des génomes des individus, et aucune autre variable n'est liée à la sélection.

Cycle de vie et évolution dans la simulation L'entité temporelle de base de nos simulations est le cycle de vie. Le cycle de vie est un cycle durant lequel chaque individu passe par une phase de reproduction, de migration et de sélection. Les mutations neutres apparaissent avec un taux de 3.0×10^{-6} par paire de base par cycle de vie. Le taux de recombinaison est égal à 1.0×10^{-2} par paire de base par cycle de vie. La taille du génome est de 110 kb. Les individus migrent en suivant une marche aléatoire de loi uniforme (-0.1, 0.1) sur chacun des axes géographiques. A la fin d'un cycle de vie, un individu se reproduit avec un partenaire choisit aléatoirement dans un rayon de 0.5 unité autour de l'individu. Le nombre de descendants résultant de ce couple d'individus suit une loi de Poisson de paramètre 10%. Le lieu de naissance du descendant correspond au barycentre des localisation des parents avec l'ajout d'un bruit gaussien d'écart type $\sigma = 0.05$. A la fin d'un cycle de vie, un génome survit jusqu'au prochain cycle de vie avec une probabilité égale au produit des composantes de la densité de régulation et de l'adaptation locale (cf paragraphes ci-dessous).

Phénotypes Le génome d'un individu contient des mutations neutres et des QTLs qui sont controlés par des mutations non neutres. L'apparition des mutations neutres n'est pas stochastique, elle est dépendante du scénario, les détails d'apparition seront spécifiées dans les paragraphes ci dessous. Tous les scénarios sont composés de 2 QTLs, chacun déterminant un trait phénotypique spécifique sous sélection contrôlée par une variable environnementale. Dans chacun des QTLs, la valeur des traits est déterminée par un effet additif. Un allèle dérivé dans un QTL augmente la valeur du trait d'un montant fixé qui varie selon le degré de polygénicité, lui même dépendant du scénario développé. Le montant est déterminé de sorte que la valeur du trait phénotypique soit comprise entre 0 et 1. La figure 4.1 résume l'obtention de la valeur d'un trait lié à un QTL composé de 5 SNPs. Le montant est alors de 0.1.

Régulation de la densité La densité des individus est régulée par la compétition spatiale. La compétition spatiale est régie par la formule suivante selon Haller et Messer, 2019

$$\text{Competition spatiale} = \frac{11\pi R^2}{N_{\text{ind}} + 1} \tag{4.1}$$

où N_{ind} est le nombre d'individus présents dans un cercle de rayon $R = 0.8$ autour de l'individu pour lequel la composante de compétition spatiale est calculée. Combinée à la composante d'adaptation locale (que nous appellerons valeur sélective dans ce chapitre), la compétition spatiale fournit la probabilité de survie d'un génome dans un prochain cycle de vie.

Valeur sélective La valeur sélective est calculée sur la base de la formule de sélection stabilisatrice gaussienne (Bürger, 2000).

$$\text{Valeur selective} = \exp\left(-\frac{1}{2}(\mathbf{z} - \bar{\mathbf{z}})C^{-1}(\mathbf{z} - \bar{\mathbf{z}})^T\right) \tag{4.2}$$

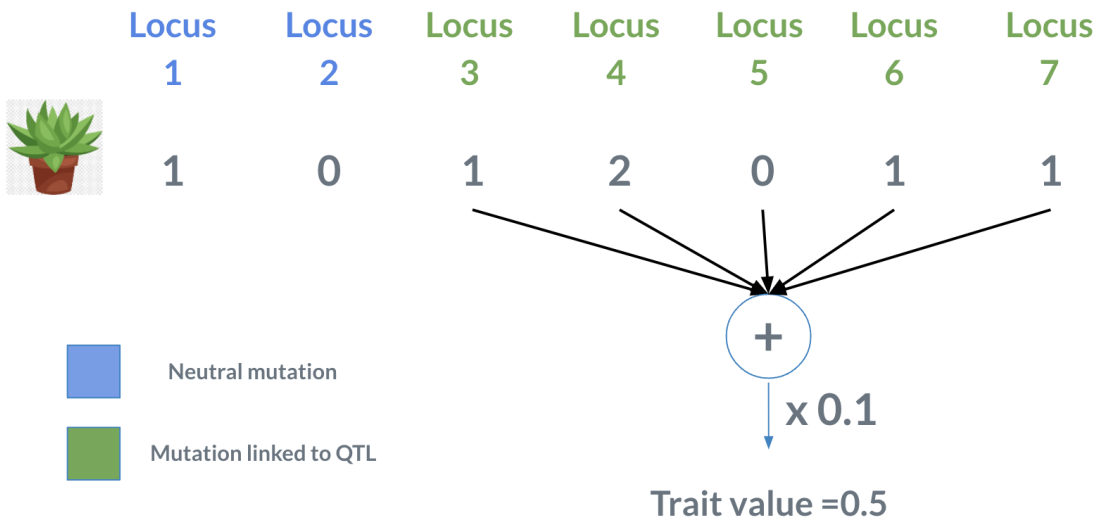

Figure 4.1 – Obtention de la valeur d'un trait dans notre simulation
La valeur des traits de phénotypes est basée sur un effet additif. Ici, la présence d'un allèle dérivé dans un locus du QTL augmente la valeur du phénotype de 0.1 permettant à ce trait de varier entre 0 et 1.

où $\bar{\mathbf{z}} \approx \mathbf{x}$ le vecteur de variables environnementales aux coordonnées géographiques de l'individu, $\mathbf{z}$ est le vecteur de trait phénotypique et C est une matrice de covariance pour les coefficients de sélection. Nous avons utilisé :

$$C = \begin{pmatrix} 0.04 & 0 \\ 0 & 0.04 \end{pmatrix}.$$

La valeur de 0.04 a été choisi de sorte que l'adaptation locale arrive dans un temps de simulation raisonnable, avant le changement environnemental.

Probabilité de survie La probabilité de survie d'un individu est dépendante de la composante de compétition spatiale et de la valeur sélective de la manière suivante :

$$\text{Probabilite de survie} = \text{Valeur selective} \times \text{Competition spatiale}$$

Dans certains scénarios, l'adaptation locale n'est pas mise en place en début de scénario. La probabilité de survie est alors égale à la compétition spatiale. Ces éléments nous permettent de simuler de l'adaptation locale chez les individus.

Changement environnemental A l'issue des simulations, on provoque un changement environnemental brutal. Pour ce faire, on ajoute un bruit uniforme de paramètres $(-0.3, 0.3)$ à chaque valeur de x_1 et x_2. Un exemple de perturbation environnemental est donné en figure 4.3a. Ce bruit uniforme donne lieu à une hétérogénéité de la perturbation, deux populations proches dans l'espace subissent des perturbations indépendantes et donc potentiellement très différentes.

Population Nous avons fait le choix de réunir les individus en population en fonction de leur position géographique. Pour ce faire, nous avons divisé l'aire d'étude de carré de côté de longueur 10 unités en 100 carrés de côté de longueur 1 unité. La population d'un individu dépend alors du carré dans lequel il se trouve. Le carré en haut à gauche de la carte correspond

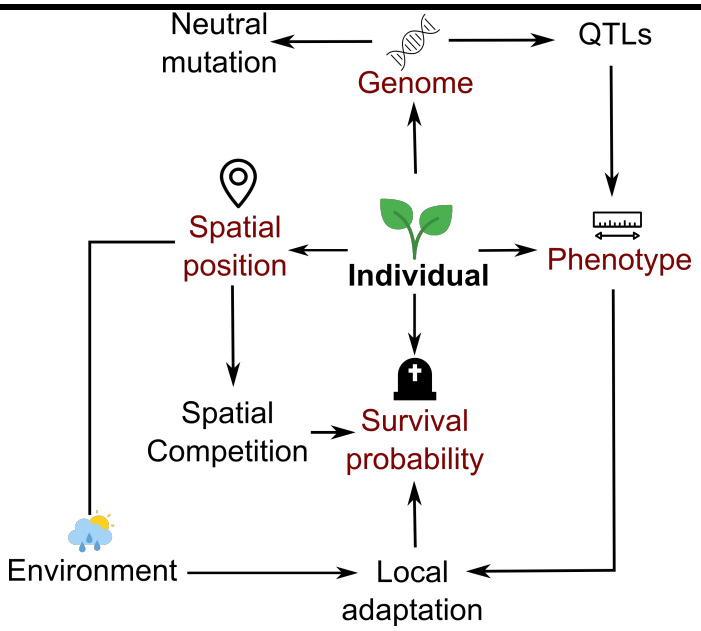

Figure 4.2 – Résumé des caractéristiques de nos simulations basées sur l'individu
Nos simulations sont basées sur l'individu. Cette individu a un génome, constitué de mutations neutres et de mutations codant pour un QTL. Le QTL détermine la valeur d'un trait phénotypique. La position d'un individu détermine les variables environnementales auxquelles il fait face. La combinaison d'un trait adaptatif et de la variable environnementale associée à ce trait donne la composante d'adaptation locale. Le produit des composante d'adaptation locale et de compétition spatiale nous donne la probabilité de survie du génome de l'individu.

à la population 1 puis on incrémente de 1 pour chaque décalage sur la droite (ou pour un passage à la ligne suivante dans le cadre d'une population multiple de 10).

Phases de la simulation Chaque simulation consiste en deux phases distinctes. Durant la première phase, appelée phase démographique, la probabilité de survie des individus se base uniquement sur la régulation de la densité. Cette phase vise à façonner la structure de population. Durant la seconde phase, appelée phase adaptative, la composante d'adaptation locale rentre en compte dans le calcul de la probabilité de survie. Cela permet aux populations d'évoluer afin d'atteindre leur optimum de valeur sélective. Au début de la phase adaptative, on fait apparaître les mutations liées au QTLs chez 300 individus. La fréquence initiale de 300 a été choisie afin que les allèles dérivés ne soient pas supprimés par la dérive génétique dans une population de taille fluctuante d'environ 2000 individus. A l'issue des simulations, les traits adaptatifs des individus correspondent aux valeurs des variables environnementales dans les populations auxquelles ils sont liés.

Exploitation des simulations Durant les simulations, les individus s'adaptent à un environnement X. A l'issue des simulations, nous allons simuler un changement environnemental brutal donnant lieu à un nouvel environnement $\mathbf{X}^\star$. Il nous est alors possible, pour chacun des individus d'obtenir la valeur $\omega(\mathbf{x}, \mathbf{x}^\star)$ correspondant à la nouvelle valeur sélective (i.e la composante d'adaptation locale) des individus. Nous calculons alors la valeur moyenne de cette composante pour chacune des 100 populations et nous comparons cette valeur au décalage génétique calculé avec les différentes méthodes.

3. Validation de la théorie par la simulation à travers un exemple

Rappel de la théorie Notre théorie quantitative du décalage génétique nous donne une expression du fossé génétique en fonction du logarithme de $\omega(\mathbf{x}, \mathbf{x}^\star)$. Elle nous permet aussi de quantifier l'importance des variables en se basant sur les valeurs propres et vecteurs propres de la matrice de covariance des tailles d'effet. Dans cette section, nous allons retrouver expérimentalement ces résultats à l'aide des simulations.

Comparaison des performances de régression avec le J-test Dans tout ce chapitre, nous allons effectuer des regressions du logarithme de la valeur sélective en fonction du décalage génétique. Cette régression sera effectuée pour les décalages génétiques calculés avec différentes méthodes et nous chercherons à comparer les R^2 de ces différentes régressions. Pour comparer des régressions effectuées avec des prédicteurs différents, nous avons fait le choix d'utiliser le J-test (Davidson et MacKinnon, 1981). L'idée du J-test est la suivante : Si un modèle contient le bon prédicteur (ici un décalage génétique prédisant mieux $\omega(\mathbf{x}, \mathbf{x}^\star)$), inclure les valeurs ajustées d'un second modèle ne devrait apporter aucune amélioration significative. Ainsi pour comparer deux modèles de régression, les valeurs ajustées du modèle 1 sont incluses comme prédicteurs dans le modèle 2 et vice versa.

Description du scénario Le scénario d'illustration respecte toutes les caractéristiques présentées en section **2**. En plus des variables x_1 et x_2, nous avons simulé deux variables supplémentaires x_3 et x_4 corrélées aux deux premières variables (Figure 4.3a). Les variables x_3 et x_4 ne sont pas impliquées dans le processus d'adaptation locale. Les perturbations de ces variables sont également représentées dans la Figure 4.3a.
Comme attendu par l'équation (3.4), les valeurs de fossé génétique calculées à l'aide de l'équation 3.2 varient linéairement avec le logarithme de la valeur sélective observée ($r^2 \approx 78\%, P < 0.001$, Figure 4.3b-c). Le pouvoir prédictif du fossé génétique est significativement plus élevé que le carré de la distance environnementale euclidienne entre les prédicteurs et leur valeur modifiée ($r^2 \approx 45\%$, $J = 11.3$, $P < 0.001$). Bien que calculé avec des prédicteurs causaux et non-causaux, le fossé génétique s'ajuste presque parfaitement à la fonction quadratique qui détermine l'intensité de la sélection stabilisatrice gaussienne ($r^2 = 97\%$, $P < 0.001$, Figure 4.4). Il est intéressant de remarquer qu'il n'est pas possible de mieux prédire la valeur sélective dans le nouvel environnement que par cette fonction quadratique. On constate toutefois une différence entre la valeur sélective observée et la valeur théorique calculée sur la base de l'équation 4.2. Cela s'explique par le bruit causé par l'aléa de la simulation qui implique que tous les individus ne sont pas parfaitement adaptés à leur environnement au moment de l'échantillonnage. Les deux premières valeurs propres de la matrice de covariance des tailles d'effets sont beaucoup plus grandes que les deux dernières (Figure 4.3b-c). Les coefficients des premiers axes donnent plus de poids aux predicteurs liés à la sélection naturelle. Ces résultats donnent la preuve que les plus grandes valeurs propres contiennent l'information utile à propos de l'adaptation locale. Ainsi, nous retrouvons bien les résultats théoriques présentés dans le chapitre précédent par le biais de la simulation.

4. Comparaison des méthodes par la simulation

On attend des mesures de décalage génétique qu'elles soient un indicateur de la vulnérabilité des populations. Autrement dit, le décalage génétique doit être prédictif de la valeur sélective

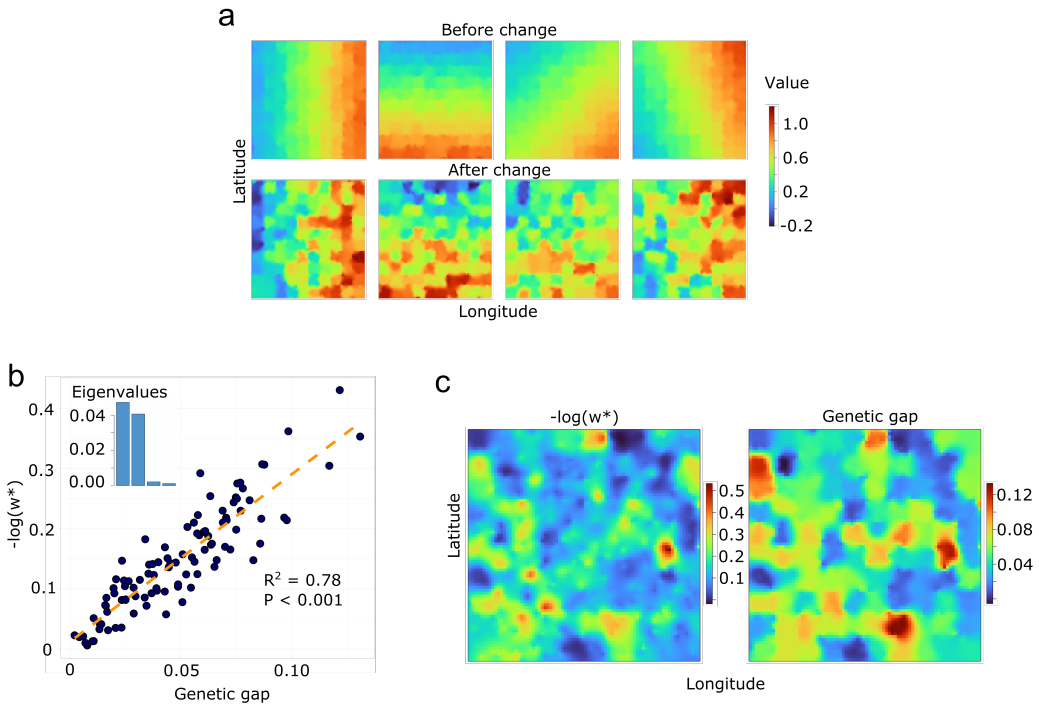

Figure 4.3 – **Comparaison de la valeur sélective avec le fossé génétique pour un scénario spécifique**
a) Carte géographique de 4 prédicteurs environnementaux avant et après changement environnemental **b**) Logarithme de la valeur sélective dans l'environnement modifiée en fonction du fossé génétique. Les valeurs propres de la matrice de covariance des tailles d'effets sont affichées en haut à gauche **c**) Carte géographique du logarithme de la valeur sélective dans l'environnement modifié (gauche) et fossé génétique (droite).

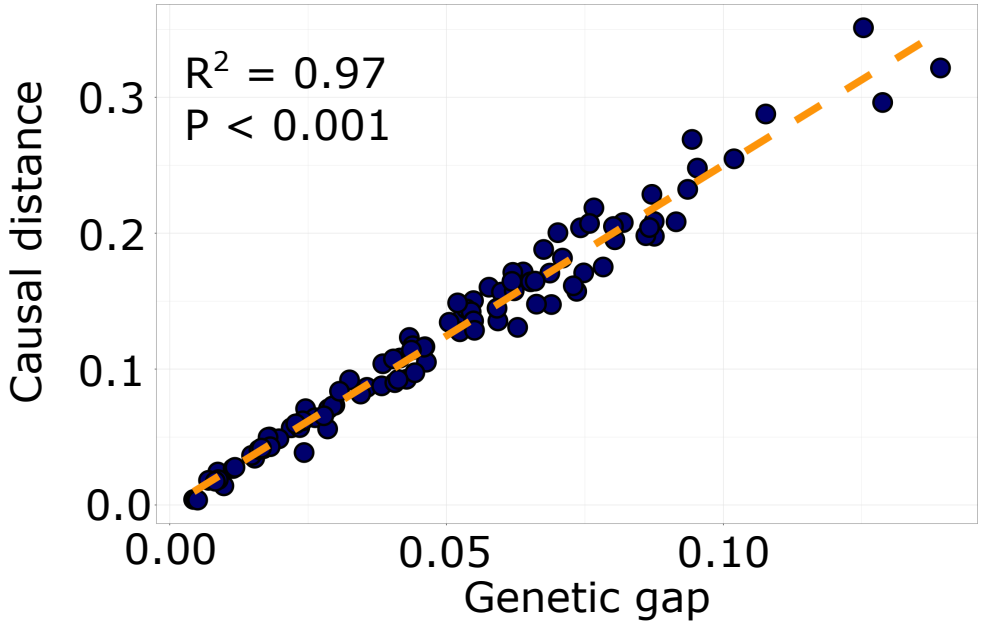

Figure 4.4 – **Fossé génétique et distance quadratique causale dans l'exemple**
Distance au carrée entre les environnements qui déterminent l'intensité de la sélection en fonction du fossé génétique

de la population dans un nouvel environnement. Dans le chapitre précédent, nous avons prouvé que cela était vrai en théorie. Nous proposons dans cette section une procédure de comparaison des méthodes par la simulation.

4. 1. Procédure de comparaison

Nous avons conçu 4 types de scénarios, décrits en détail dans la sous-section **4.3** de ce chapitre. Les types de scénarios sont basés sur 2 caractéristiques distinctes : le niveau de polygénicité des QTLs (faible ou élevé), et la corrélation des variables causales avec les premières composantes principales. Tous les scénarios respectent les caractéristiques présentées en section **2**. A l'issue des scénarios, nous récupérons les données de génotype de chacun des génotypes, la position des QTLs, les coordonnées géographiques des individus, les variables environnementales avant et après la variation brutale, et la valeur sélective des individus avant et après la variation brutale. Nous allons ensuite calculer le décalage génétique à l'aide de 4 méthodes différentes , fossé génétique, RDA, Rona, et GF pour chacune des populations de l'environnement pour lequel elles sont adaptées vers l'environnement modifié. Nous allons ensuite effectuer la régression linéaire de la valeur sélective de chacune des population dans l'environnement modifié en fonction des valeurs de décalage génétique. Nous utiliserons alors le R^2 de la régression linéaire comme mesure de performance de chacune des méthodes. Nous répliquerons chacun des scénarios 30 fois et comparerons la distribution des valeurs de R^2 pour les différentes méthodes.

4. 2. Implémentation des méthodes

Afin d'obtenir les valeurs de décalage génétique, nous utiliserons les méthodes de la manière suivante :

Fossé génétique Les calculs du fossé génétique ont été obtenu avec la fonction `genetic.gap` dans la librairie `R` LEA version 3.9.5 (Gain et François, 2021).

Décalage génétique RDA Nous avons implémenté deux versions de RDA, une sans correction et une avec correction pour les facteurs de confusion dûs aux effets démographiques. Pour la version sans correction, nous avons ajusté un modèle de régression linéaire à la matrice de génotype à l'aide de la fonction `lm`. Nous avons alors effectué une analyse en composante principale sur la matrice ajustée avec la fonction `prcomp`. Nous avons finalement projeté les prédicteurs environnementaux sur les composantes principales obtenues et avons calculé la distance au carré entre les projections de l'environnement actuel et de l'environnement modifié sur l'ensemble des axes de la RDA. Pour la version avec correction, nous avons ajusté un LFMM avec K facteurs latents à la place du modèle de régression linéaire simple puis avons appliqué les mêmes étapes que pour la version sans correction. Notre implémentation diffère de celle de Capblancq, Morin et al., 2020, qui eux utilisent la projection le long de 2 axes uniquement et incluent des poids liés à la part de variance expliquée par chacun des axes. Notre implémentation rend le décalage génétique RDA corrigé théoriquement équivalent au fossé génétique comme nous l'avons montré dans le chapitre précédent.

Décalage génétique GF Nous avons calculé le décalage génétique GF en obtenant les projections des prédicteurs environnementaux actuels et modifiés sur les courbes de changement de fréquences alléliques. Le décalage génétique GF correspond alors à la distance euclidienne au carré des vecteurs de ces projections. Les calculs ont été effectué à l'aide de la librairie `R`

`gradientForest` version 0.1. Afin de corriger pour les facteurs de confusion, nous avons inclus les K facteurs latents estimés dans LFMM.

Décalage génétique Rona Dans la version présentée en 2016, Rellstab propose un calcul du Rona pour chaque prédicteur environnemental (cf chapitre 2 section **3.2**). Ici, nous proposons une modification de cette méthode. En effet, plutôt que d'ajuster une régression linéaire simple pour chaque prédicteur, nous avons ajuster une régression linéaire multiple avec comme variables d'entrée l'ensemble des prédicteurs environnementaux. On effectue la régression linéaire des fréquences d'allèles en fonction de l'ensemble des prédicteurs environnementaux. Le décalage génétique Rona a été calculé comme la valeur moyenne de la distance en valeur absolue entre les valeurs ajustées et prédites des fréquences d'allèles le long du génome. La version sans correction se base sur un ajustement des fréquences d'allèles avec un modèle de régression linéaire multiple. La version avec correction se base sur un LFMM avec K facteurs latents. (Aquino et al., 2022). Notre version diffère de celle de Rellstab, Zoller et al., 2016 pour une autre raison. En effet pour le calcul du Rona, Rellstab, Zoller et al., 2016 utilise la distance entre la fréquence prédite pour l'environnement futur et la fréquence **observée** là où nous utilisons la distance entre la fréquence prédite pour l'environnement futur et la fréquence **prédite pour l'environnement actuel**.

4. 3. Description des différents scénarios

Comme expliqué dans la sous-section **4.1**, 4 types de scénarios ont été développés. Ils se distinguent par le degré de polygénicité des QTLs, et par la corrélation entre la structure de population et les variables environnementales. Dans cette partie nous résumons comment sont construits ces scénarios.

Trait faiblement et fortement polygénique Nous avons fait varier le nombre de mutations qui contrôlent les deux QTLs impliqués dans le processus d'adaptation locale. Dans les scénarios avec une polygénicité élevée, les QTLs sont contrôlés par 120 mutations avec des effets additifs. L'allèle dérivé d'une mutation augmente la valeur du trait de 0.005, permettant au trait de varier entre 0 et 1 afin de correspondre à la variable environnementale auquel il est lié qui varie dans le même intervalle. Il y a 20 mutations supplémentaires par rapport au 100 initialement requises pour faire varier le phénotype entre 0 et 1. Ce choix a été effectué car, l'effet des mutations étant relativement faible, certaines mutations seront fixées à une fréquence d'allèle de 0. Dans les scénarios faiblement polygéniques, les QTLs sont contrôlés par 10 mutations. Un allèle dérivé d'une mutation augmente la valeur du trait de 0.05, permettant au trait de varier entre 0 et 1.

Corrélation entre structure de population et gradient environnemental Nous avons conçu des scénarios avec différents degrés de corrélation entre le gradient environnemental et la structure de population. La différence entre les scénarios se trouve durant la phase démographique. Dans les scénarios avec une forte corrélation entre la structure de population et le gradient environnemental, nous simulons une migration vers le Nord. Pour ce faire, en début de simulation, la partie Nord de la carte est hostile aux individus (leur présence dans cette zone implique une probabilité de survie au prochain cycle de vie de 0), cette zone devient progressivement habitable permettant aux individus de graduellement coloniser cette zone. En simulant l'expansion de l'aire de répartition à partir du sud, la structure de la population a été orientée selon la deuxième variable environnementale.

4. 4. Ensemble de SNPs causaux

Nous avons testé les méthodes sur un ensemble de SNPs, dit SNPs causaux et sur la totalité des SNPs. En pratique, nos simulations nous permettent d'avoir accès aux vrais SNPs causaux. Toutefois, pour que la procédure soit la plus proche possible de ce qui est fait dans le cas des données réelles, les SNPs causaux ont été détecté à l'aide de LFMM et de la librairie `qvalue` en appliquant un taux de fausse découverte de 10%. Cette procédure nous permet également de tester les méthodes en présence de faux positifs.

4. 5. Résultats

Les valeurs de R^2 des régressions du logarithme de la valeur sélective en fonction de la statistique de décalage génétique sont très proches peu importe la méthode utilisée pour obtenir la statistique de décalage génétique (Figure 4.5 et Figure 4.6). Bien que la différence soit faible, les méthodes qui n'utilisent pas de correction ou qui corrigent à l'aide de variables de structure de population (et non de facteurs latents) fonctionnent en moyenne moins bien que les méthodes avec une correction par les facteurs latents (Figure 4.7 et Figure 4.8). Une fois corrigées, les statistiques ont des performances similaires sur tous les scénarios. La capacité du fossé génétique à prédire le logarithme de la valeur sélective est égale à celle de RDA corrigée. Elle est légèrement supérieure à Rona corrigé et à GF corrigé. Toutes les mesures sont fortement corrélées au fossé génétique (Figure 4.9). Le fossé génétique affiche une corrélation élevée avec la distance quadratique entre les prédicteurs causaux expliquant les traits sous sélection (Figure 4.10). Lorsque tous les locus de la matrice de génotype sont inclus pour le calcul du décalage génétique, les prédictions restent proches de celles basées sur le sous ensemble de locus identifié par LFMM, GF atteignant alors des performances similaires aux autres statistiques (Figure 4.11).

5. Comparaison du fossé génétique avec des méthodes contraintes

La méthode linéaire que nous utilisons est non contrainte. Cela signifie que nous n'avons pas de garantie que les valeurs prédites soient comprises entre 0 et 1 et donc nous n'avons pas l'assurance de pouvoir interpréter les valeurs obtenues comme de réelles fréquences alléliques. Nous proposons dans cette section de comparer les valeurs de fossé génétique obtenues avec des méthodes linéaires aux méthodes non linéaires contraintes, notamment une mesure basée sur un glm et une autre basée sur une méthode d'apprentissage profond.

5. 1. Décalage génétique avec modèle linéaire généralisé (GLM)

La mesure basée sur le GLM est similaire au fossé génétique mais le modèle linéaire est remplacée par une régression logistique (Cox, 1958). Nous utilisons la fonction logit, $\sigma(x) = 1/(1 + e^{-x})$, dans un modèle de régression logistique des fréquences d'allèle en fonction des prédicteurs environnementaux, $\mathbf{x}$, et des facteurs latents, $\mathbf{u}$. Après avoir ajusté le modèle aux données, les fréquences alléliques prédites sont de la forme suivante

$$f_c(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{a}^T + \mathbf{u}\mathbf{w}^T)\,,$$

où $\mathbf{a}$ et $\mathbf{w}$ sont les tailles d'effet et les loadings estimés dans le modèle de régression logistique. Ces prédicteurs contraints prennent des valeurs entre 0 et 1 et peuvent donc être interprété en terme de fréquences d'allèle.

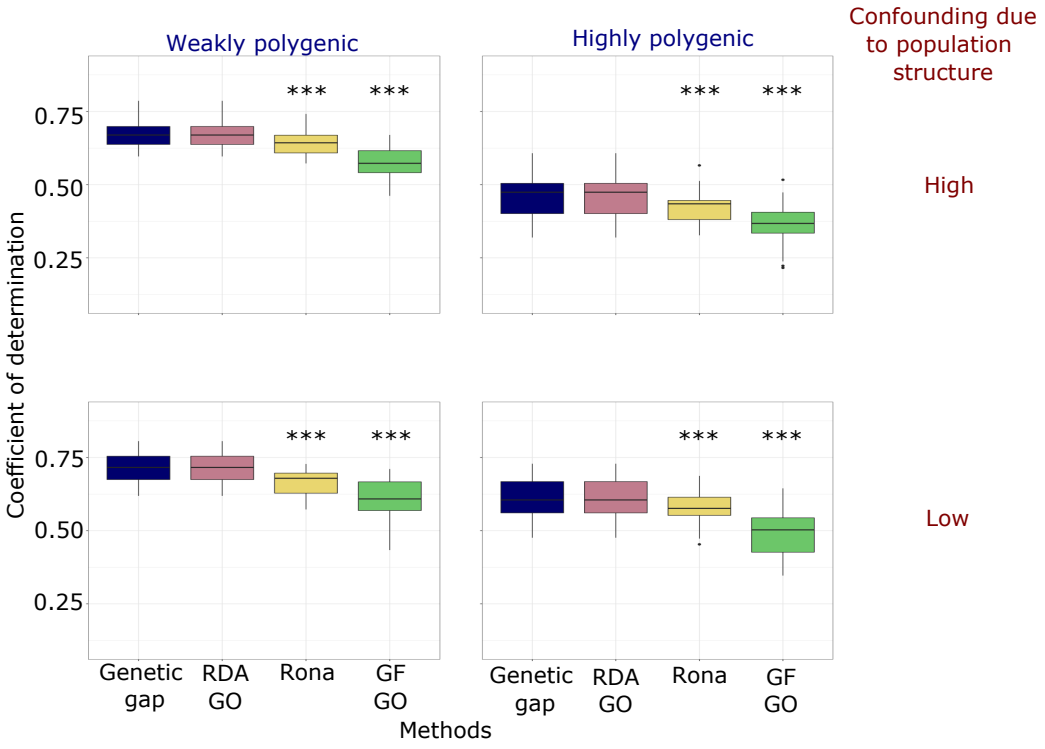

Figure 4.5 – Performance prédictive des statistiques de décalage génétique
Part de variance de la valeur sélective dans l'environnement modifié expliqué par les statistiques de décalage génétique (coefficient de détermination). 4 scénarios avec des niveaux distincts de polygénicité dans les traits adaptatifs et de corrélation entre les prédicteurs environnementaux et la structure de populations ont été implémenté. Les étoiles correspondent à la comparaison par paire entre le fossé génétique et les autres statistiques de décalage génétique (paired t-tests, *** : $P < 0.001$). Les boxplots affichent la médiane, le premier quartile et le troisième quartile. L'extrémité haute correspond à la plus grande valeur située au maximum à 1,5 intervalle interquartile (IQR) du troisième quartile. L'extrémité basse correspond à la plus petite valeur située au maximum à 1,5 intervalle interquartile (IQR) du premier quartile.

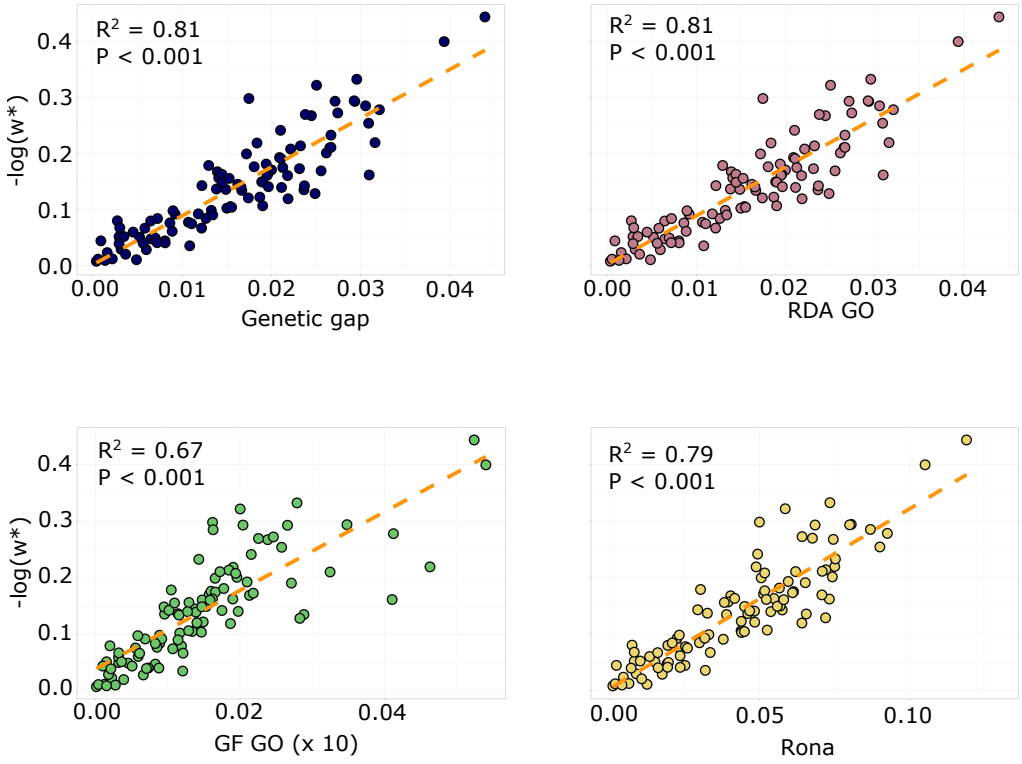

Figure 4.6 – **Graphique de régression de la valeur sélective en fonction de différentes statistiques de décalage génétique**

Graphiques de régression pour des données simulées avec des traits faiblement polygénique et un niveau élevé de corrélation entre la structure de population et l'environnement. Le fossé génétique, RDA, Rona et GF incluent des corrections basées sur les facteurs latents.

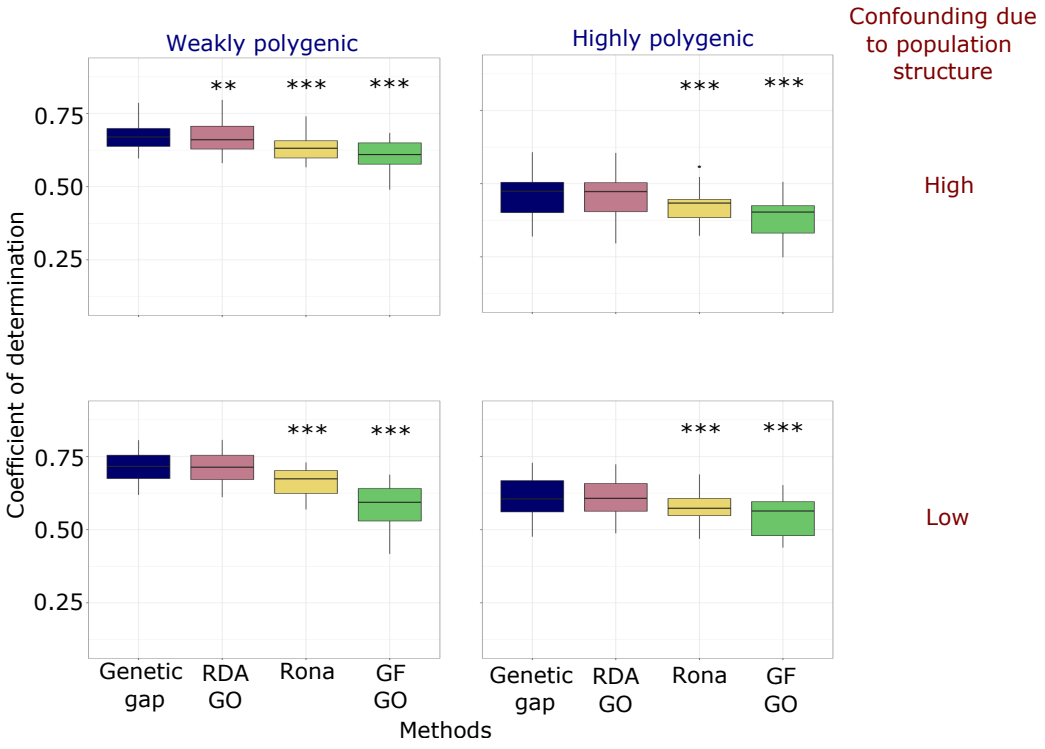

Figure 4.7 – Performance prédictive des statistiques de décalage génétique non corrigées

Part de variance de la valeur sélective dans l'environnement modifié expliqué par les statistiques de décalage génétique (coefficient de détermination) non corrigées.

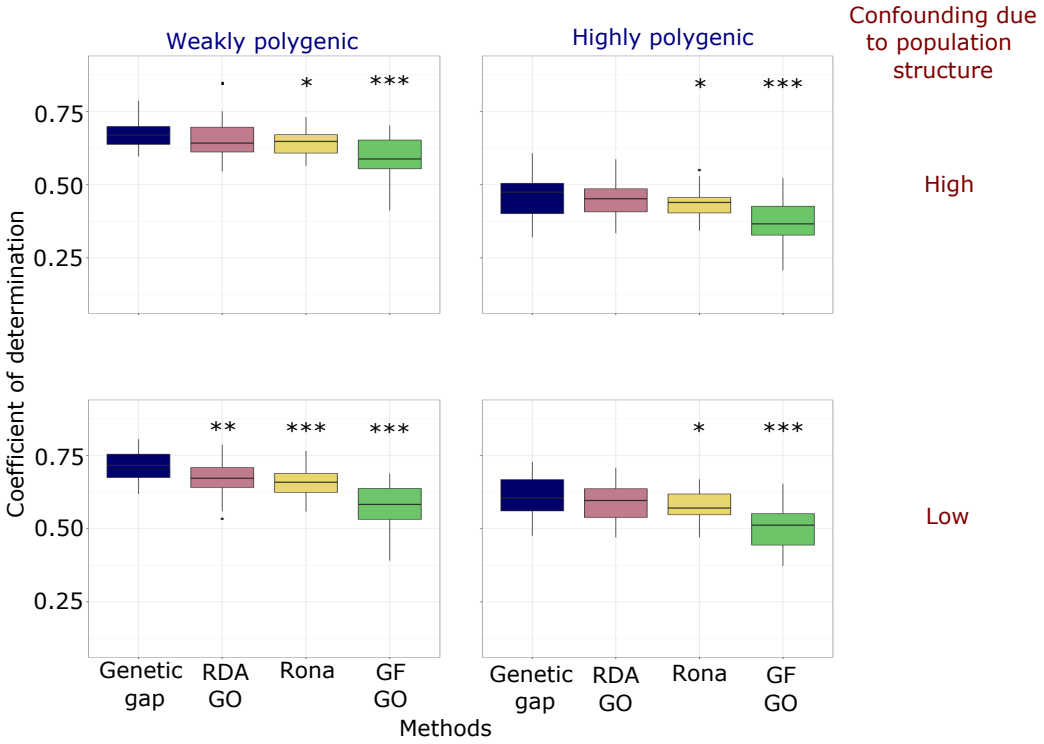

Figure 4.8 – Performance prédictive des statistiques de décalage génétique corrigées à l'aide des PCs

Part de variance de la valeur sélective dans l'environnement modifié expliqué par les statistiques de décalage génétique (coefficient de détermination) corrigées à l'aide des 10 composantes principales de la matrice de génotype.

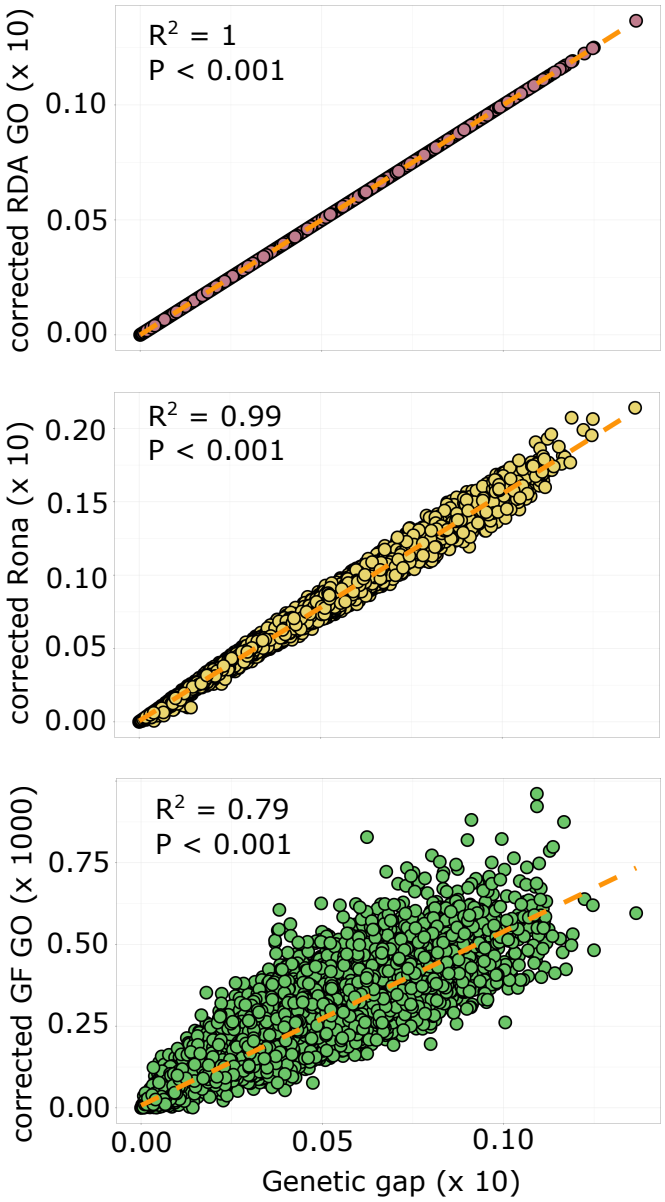

Figure 4.9 – Statistiques de décalage génétique en fonction du fossé génétique
Valeurs de décalage génétique de RDA, Rona et GF en fonction du fossé génétique. Les statistiques sont calculées sur tous les locus de la matrice de géntoype pour 30 réplications des 4 scénarios (tous les résultats sont représentés simultanément).

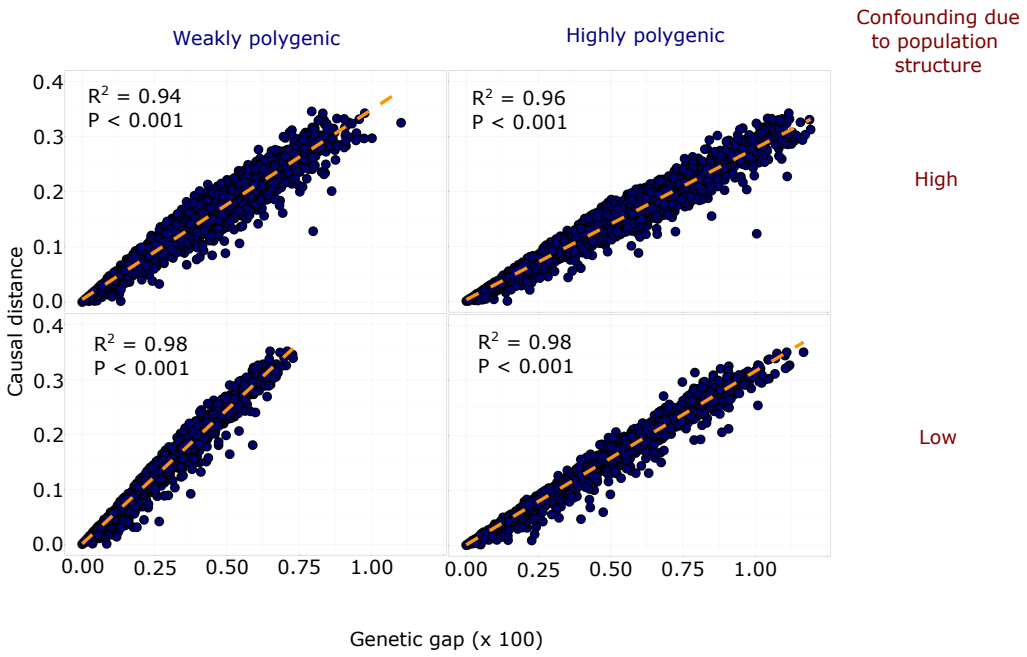

Figure 4.10 – **Corrélation du décalage génétique avec la distance quadratique causale** Distance au carrée entre les variables qui impactent la valeur sélective dans les simulations et leur valeur modifée en fonction du fossé génétique calculé sur tous les locus. Les niveaux élevés de corrélations indiquent que les prédictions faites avec le genetic sont proche de l'optimal.

Pour de petites variations environnementales, une série de Taylor de la différence des fréquences alléliques prédites nous donne

$$f_c(\mathbf{x}^\star) - f_c(\mathbf{x}) \approx f_c(\mathbf{x})(1 - f_c(\mathbf{x}))(\mathbf{x}^\star - \mathbf{x})\mathbf{a}^T \,.$$

En posant

$$\mathbf{b} = f_c(\mathbf{x})(1 - f_c(\mathbf{x})) \times \mathbf{a} = H_0 \times \mathbf{a}/2 \,,$$

où H_0 correspond à l'hétérozygotie au locus considéré, on a alors

$$f(\mathbf{x}^\star) - f(\mathbf{x}) = (\mathbf{x}^\star - \mathbf{x})\mathbf{b}^T \,.$$

Ainsi, pour de petites variations environnementales, le décalage génétique obtenu à l'aide de la régression logistique $\mathbb{E}[(f_c(\mathbf{x}^\star) - f_c(\mathbf{x}))^2]$ doit correctement approcher le fossé génétique.

5. 2. Décalage génétique avec un auto-encodeur variationnel (VAE)

Nous avons également utilisé les données environnementales et génomiques dans un modèle génératif d'apprentissage profond. Ce modèle génératif est construit de façon à générer une composition génétique conditionnée aux données environnementales. Cette approche nous permet d'obtenir des prédictions contraintes de fréquences alléliques, $f_c(\mathbf{x})$, contrairement à l'approche linéaire. Pour ce faire nous avons utilisé une architecture de type auto-encodeur variationnel conditionnel (Kingma et Welling, 2013). Plus précisément nous avons utilisé l'architecture présentée dans Liao et Lin, 2021, composée de deux encodeurs probabilistes avec une relation étudiant-professeur (Figure 4.15). Nous obtenons ensuite le décalage génétique à l'aide de la formule $\mathbb{E}[(f_c(\mathbf{x}) - f_c(\mathbf{x}^\star))^2]$, où $f_c(\mathbf{x})$ et $f_c(\mathbf{x}^\star)$ sont obtenus à partir du modèle d'apprentissage profond (Figure 4.13).

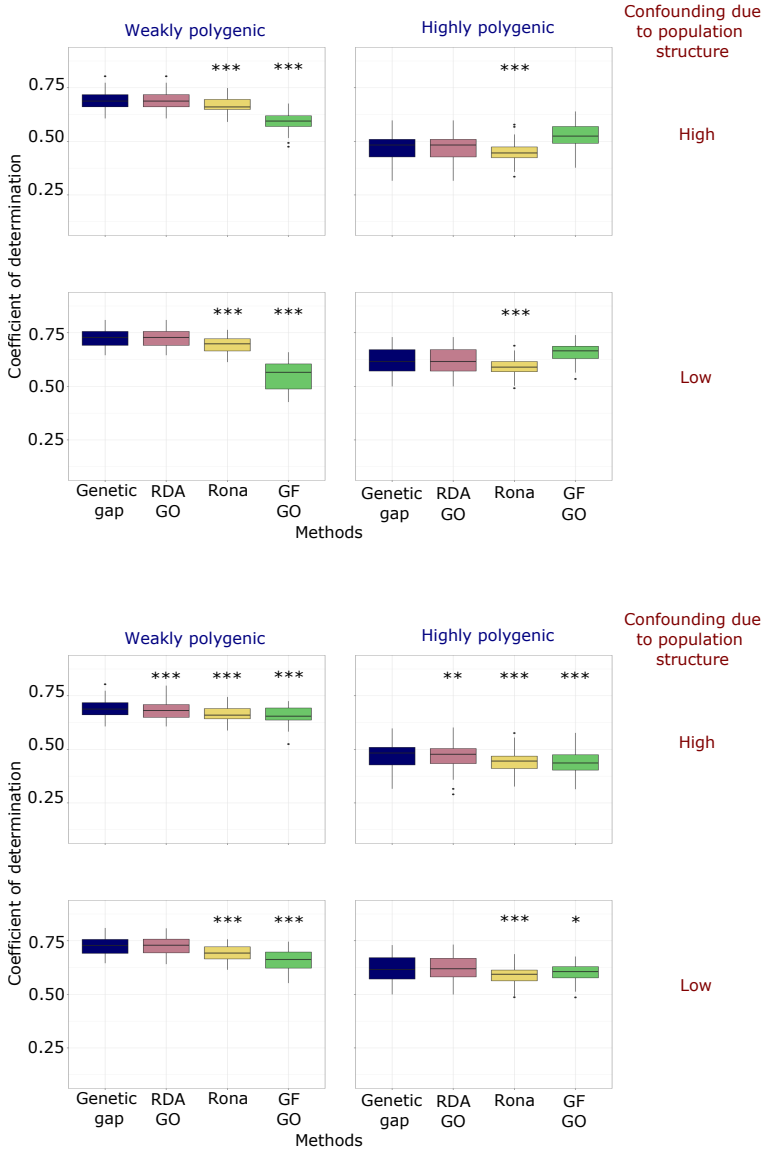

Figure 4.11 – Performance prédictive des statistiques de décalage génétique calculées sur tout le génome

Capacité des statistiques de décalage génétique à prédire la valeur sélective calculées sur l'ensemble du génome. Les statistiques sont calculées avec correction en utilisant 10 facteurs latents (haut) et sans correction (bas)

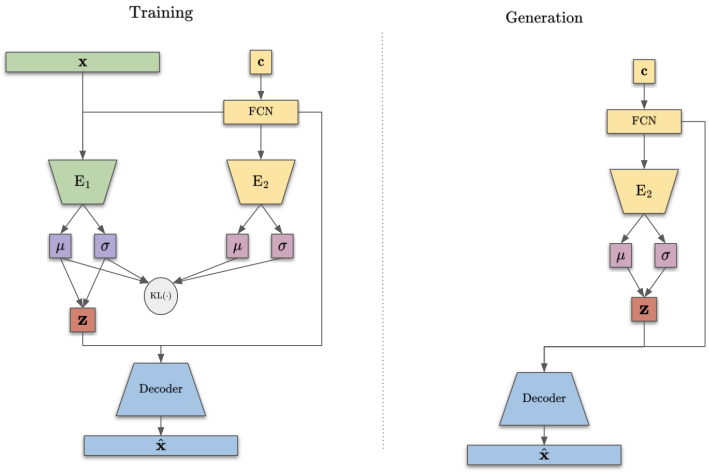

Figure 4.12 – Architecture basée sur (Liao et Lin, 2021), utilisant une relation étudiant-professeur

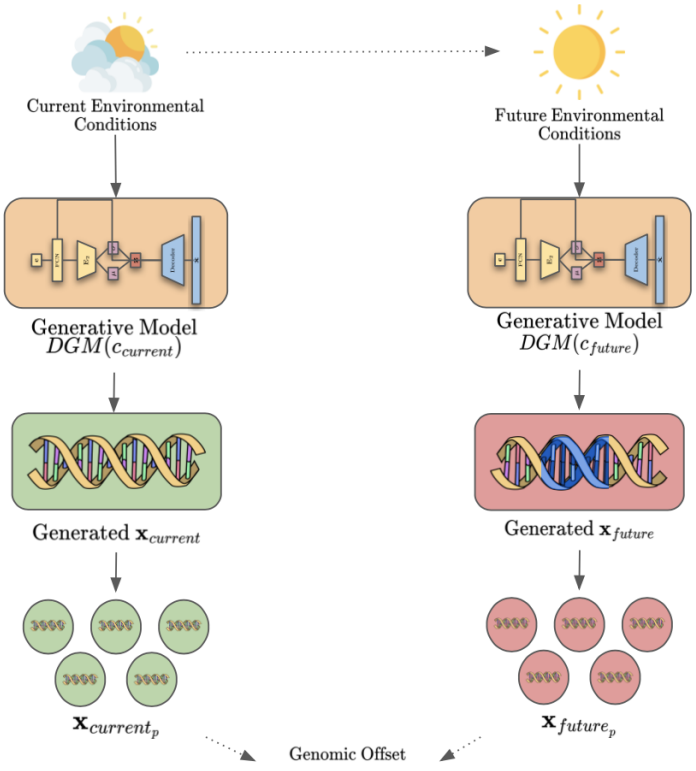

Figure 4.13 – Décalage génétique obtenu avec un modèle d'apprentissage profond

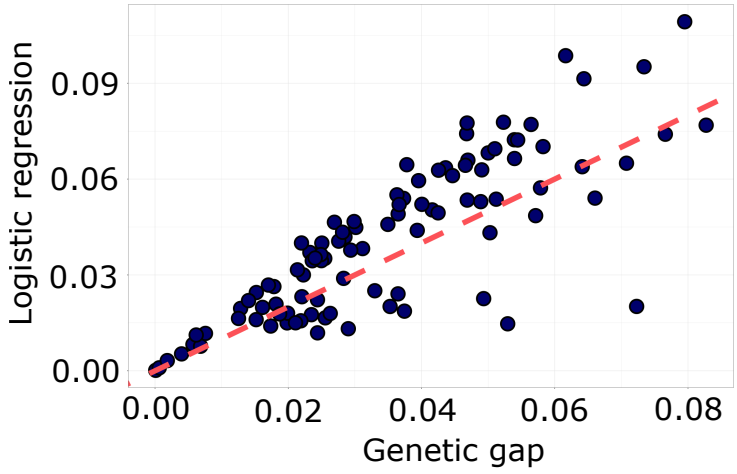

Figure 4.14 – Comparaison d'un décalage génétique basé sur le GLM avec le fossé génétique

Comparaison de la distance quadratique entre des prédicteurs contraints , $\mathbb{E}[(f_c(\mathbf{x}) - f_c(\mathbf{x}^\star))^2]$, basés sur la régression logistique avec le fossé génétique, basé sur des prédicteurs linéaires non contraints.

5. 3. Expériences et résultats

Nous allons désormais comparer les valeurs de fossé génétique aux valeurs de décalage génétique obtenues avec les méthodes contraintes. Le scénario utilisé pour la comparaison au glm est le même que celui dans l'exemple en section **5.2** mais sans les variables non causales x_3 et x_4. On constate que pour les petites variations, la correspondance entre le glm et fossé génétique est presque parfaite. Pour de plus fortes variations, on peut voir l'apparition d'un biais (Figure 4.14). Les comparaisons avec la méthode d'apprentissage profond ont été réalisés sur l'ensemble des scénarios de la sous section **4.3** de ce chapitre. On constate que la qualité de l'ajustement est très bonne pour les différents scénarios avec des valeurs de R^2 entre le décalage génétique basé sur l'apprentissage profond et le fossé génétique allant de 0.78 à 0.89 ($P < 0.001$, Figure 4.15).

6. Validation des statistiques de décalage génétique dans le cas des données réelles

En complément des validations par la simulation, nous avons cherché à valider les méthodes à l'aide de données réelles. Nous allons utilisé pour cela des données de mil échantillonnées dans la région du Sahel et qui ont été utilisées dans une expérience de jardin commun.

6. 1. Présentation des données et de l'expérience de jardin commun

Les données ont été échantillonnées pour 158 populations situées en Afrique sub-saharienne. Pour chacune des populations, 100 plantes ont été séquencées à 138,948 sites polymorphiques avec la méthode pool-seq. On réalise ensuite un filtre sur la fréquence d'allèle minimale et sur la qualité des SNPs et on conserve finalement 16,154 SNPs. On obtient donc une matrice de taille $n \times L$ avec $n = 158$ et $L = 16154$, où chaque entrée (i, j) de la matrice correspond à la fréquence de l'allèle pour la population i au locus j. Un jardin commun a été mis en place à Sadoré (13° 14' 0" N, 2° 17' 0" E, Niger, Africa) (Rhoné et al., 2020) dans lequel on a fait pousser

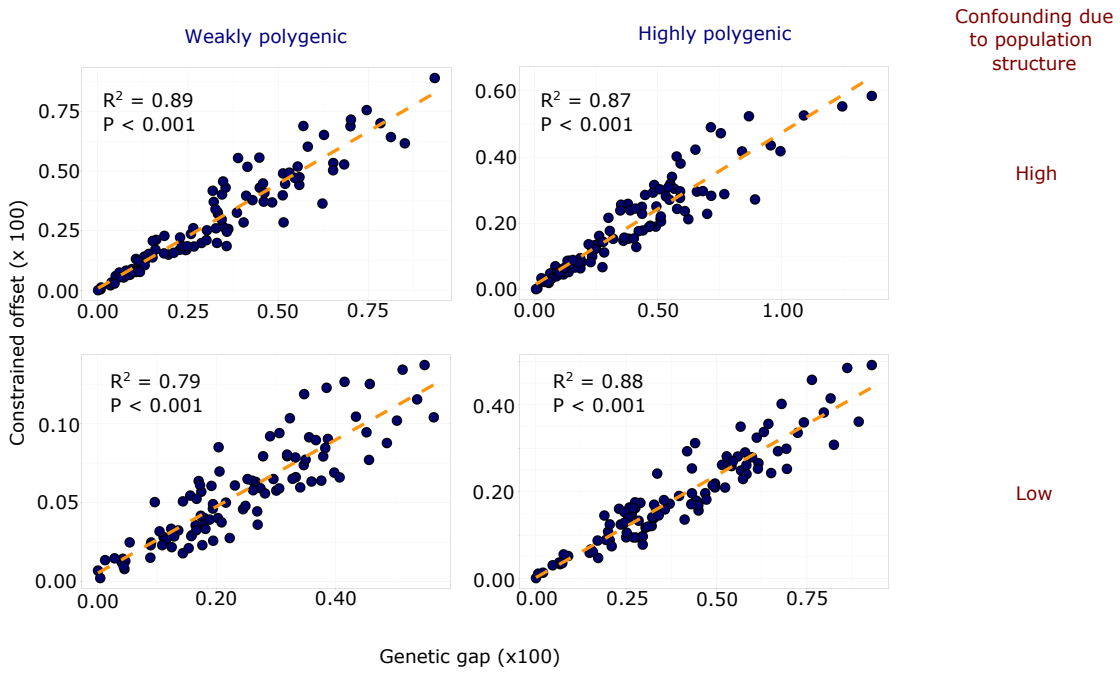

Figure 4.15 – Comparaison d'un décalage génétique basé sur un modèle d'apprentissage profond avec fossé génétique

Comparaison de la distance quadratique entre des prédicteurs contraints , $\mathbb{E}[(f_c(\mathbf{x}) - f_c(\mathbf{x}^\star))^2]$, basés sur un modèle d'apprentissage profond avec le fossé génétique, basé sur des prédicteurs linéaires non contraints.

chacune des 158 populations. Pour chacun des lieux de séquençage, les données climatiques ont été exploitées pour calculer 157 métriques divisées en trois catégories : précipitation, temperature et rayonnement. Ces prédicteurs étant fortement corrélés, nous avons utilisé des ACP sur chacune des 3 catégories afin de ne conserver que 27 variables : 15 liées à l'ACP des variables de température, 10 liées à l'ACP des variables de précipitation et 2 liées à l'ACP des variables de rayonnement. On appellera dans toute cette section $\mathbf{x}$, les prédicteurs environnementaux au lieu d'origine des populations et $\mathbf{x}^\star$ les prédicteurs environnementaux aux conditions locales de Sadoré. Pour chacune des populations cultivées dans le jardin commun, on a récolté le poids total des graines de l'épi principal de 10 plantes durant 2 années consécutives. On fait alors l'hypothèse que le poids total moyen de graine ainsi récolté est proportionnel à $\omega(\mathbf{x}, \mathbf{x}^\star)$ pour chacune des populations. Le lieu de séquençage des populations ainsi que l'approximation de $\omega(\mathbf{x}, \mathbf{x}^\star)$ sont résumés dans la figure 4.16. De manière analogue à l'expérience précédente avec les données simulées, nous allons alors comparer les valeurs $\omega(\mathbf{x}, \mathbf{x}^\star)$ aux mesures de décalage génétique des différentes méthodes calculées d'un environnement $\mathbf{x}$ vers $\mathbf{x}^\star$. A nouveau, les statistiques seront évaluées sur un ensemble de SNPs sélectionnées via LFMM et un taux de fausse découverte de 10%, appelés SNPs causaux, et sur l'ensemble des SNPs à notre disposition. Nous estimons finalement une relation linéaire entre les statistiques de décalage génétique et le logarithme de la valeur sélective et utilisons le coefficient de corrélation de Pearson au carré comme mesure de la qualité de l'ajustement. Le J-test a été utilisé pour comparer les performances prédictives des différentes méthodes (Davidson et MacKinnon, 1981).

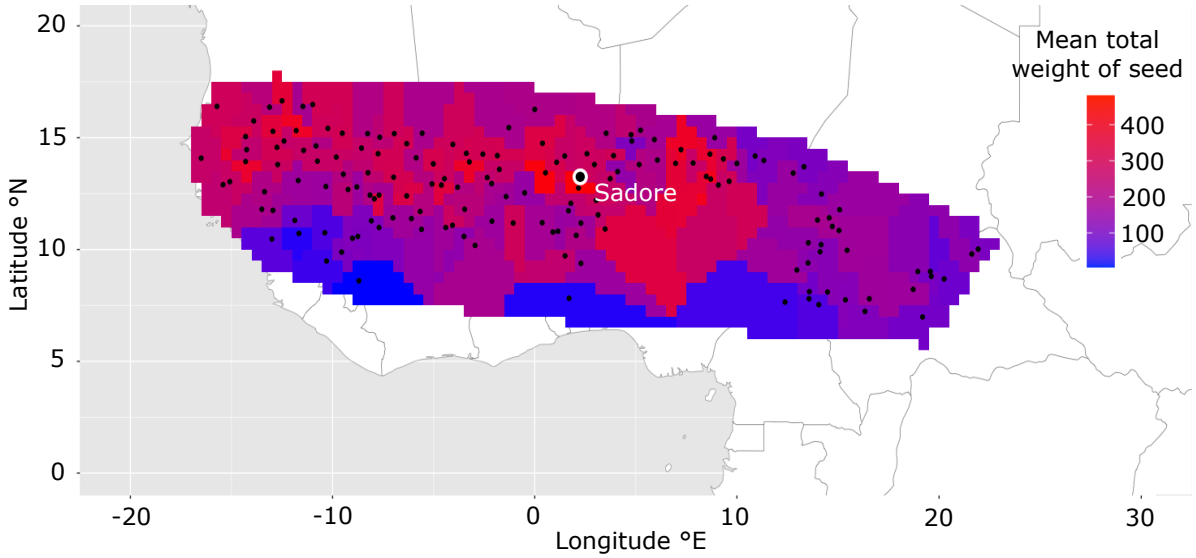

Figure 4.16 – Localisation des échantillons et gradient de la valeur sélective du mil Pour l'expérience du jardin commun situé à Sadoré (Niger), les couleurs représentent la valeur sélective mesurée à partir du poids total moyen de graine obtenu pour chacune des populations que l'on a fait pousser dans le jardin commun. La localisation de l'origine de chacune des populations est représentée par un point. Les valeurs sélectives pour les lieux non échantillonnés ont été interpolé à partir de la localisation la plus proche en utilisant la méthode de pondération par la distance inverse.

6. 2. Résultats

Toutes les statistiques de décalage génétique affichent une relation linéaire significative avec le logarithme du poids des graines. Les meilleures prédictions obtenues l'ont été avec le fossé génétique et la version corrigée de Rona ($r^2 = 61\%$, $P < 0.001$, Figure 4.17). Sur l'ensemble de SNPs causaux, les méthodes affichent des performances similaires. La correction pour les facteurs de confusions améliorent significativement la performance des méthodes RDA et Rona. Lorsqu'on applique les méthodes sur l'ensemble des SNPs, la correction par les facteurs latents offrent des performances significativement meilleures également. Les valeurs propres et les vecteurs propres de la matrice de covariance des tailles d'effet suggèrent que la température avait plus d'importance dans la variation de valeur sélective que les précipitations et le rayonnement.

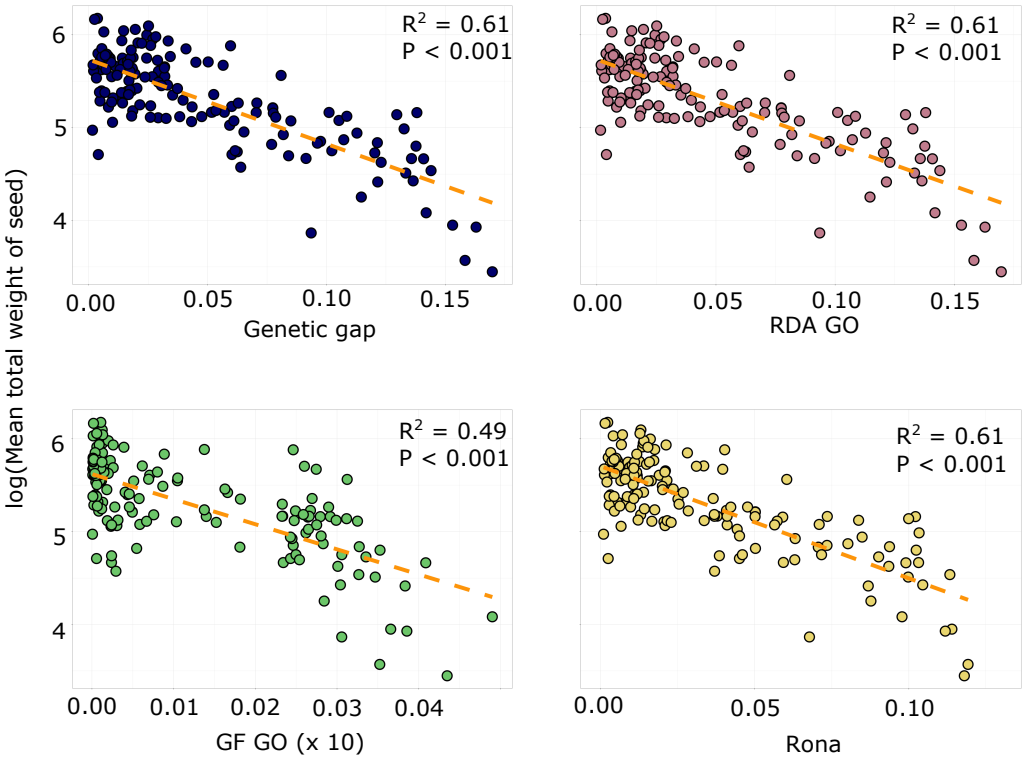

Figure 4.17 – **Performance des méthodes pour la prédiction de la valeur sélective des populations dans le jardin commun**

Méthodes évaluées sur l'ensemble de SNPs causaux et corrigés pour les facteurs de confusion à l'aide de 10 facteurs latents de LFMM

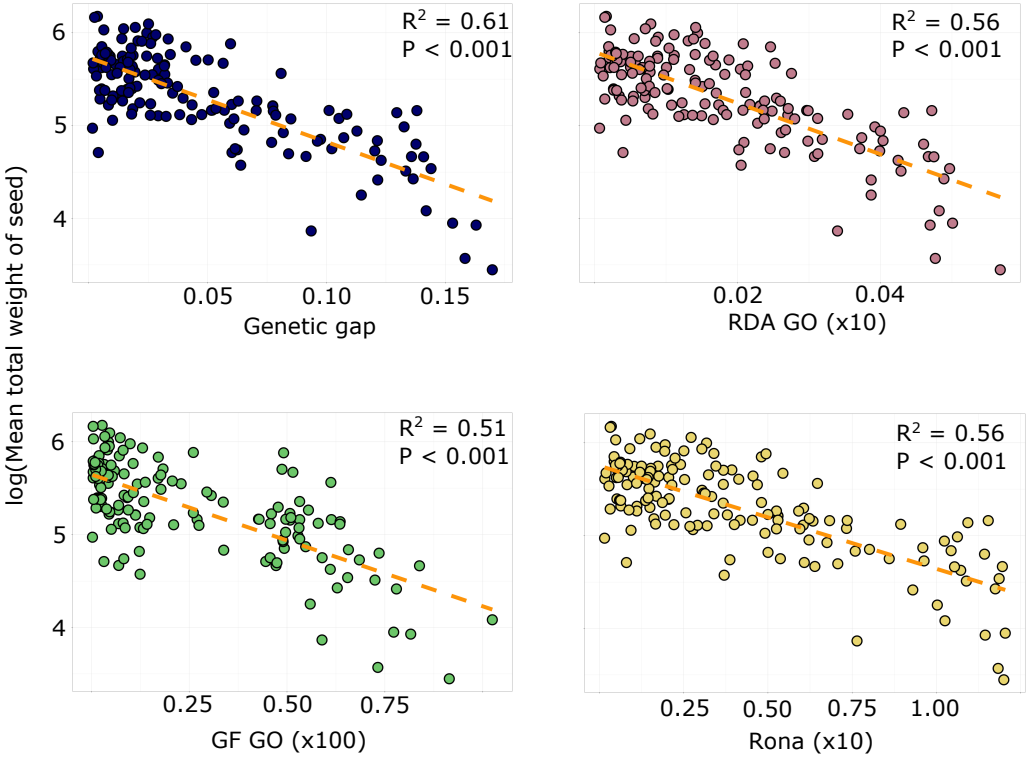

Figure 4.18 – **Performance des méthodes pour la prédiction de la valeur sélective des populations dans le jardin commun**
Méthodes évaluées sur l'ensemble de SNPs causaux sans correction pour les facteurs de confusion

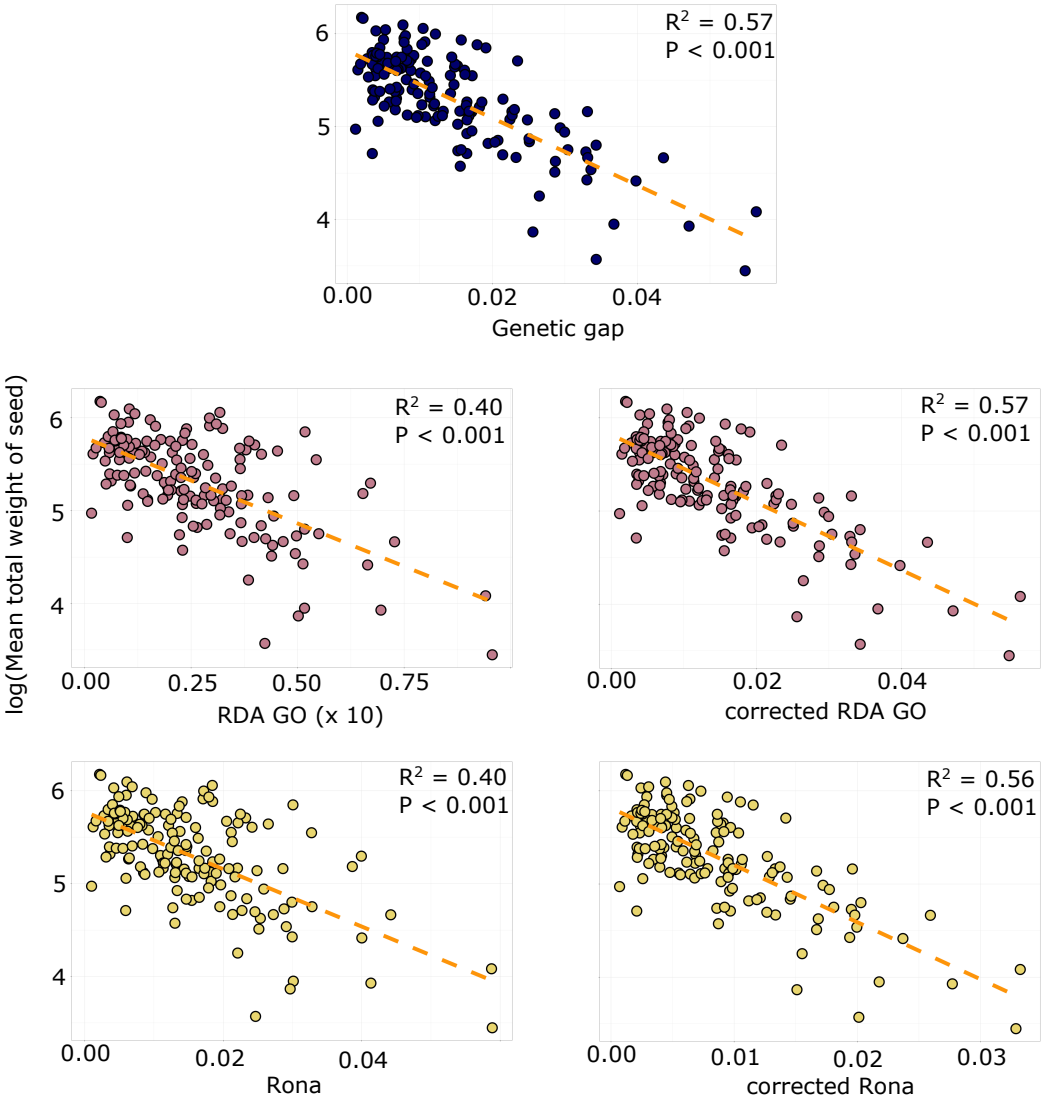

Figure 4.19 – **Performance des méthodes pour la prédiction de la valeur sélective des populations dans le jardin commun**

Méthodes évaluées sur tous les SNPs

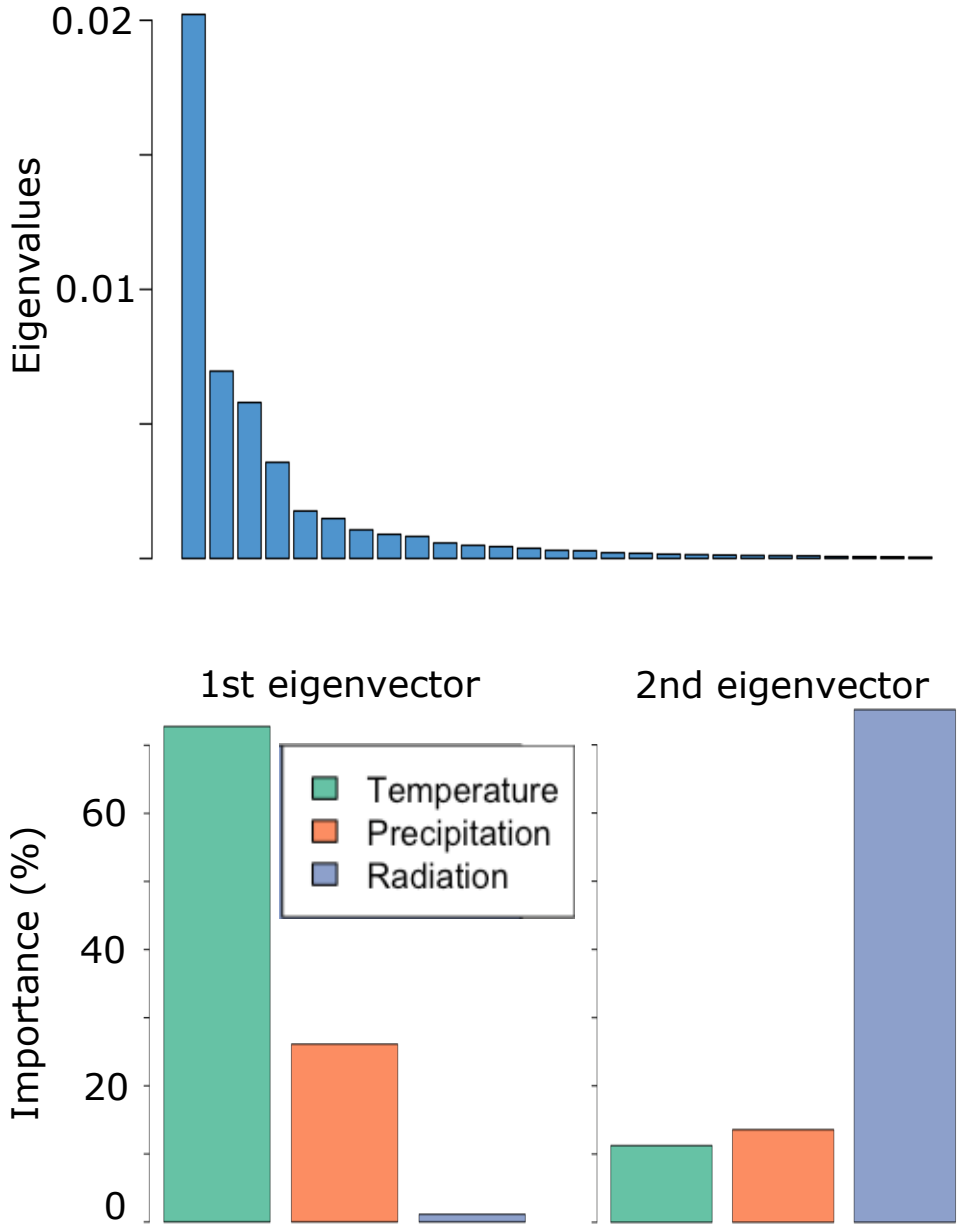

Figure 4.20 – **Importance relative des variables bioclimatiques**
Haut : Valeurs propres de la matrice de covariance des tailles d'effets. Bas : Importance relative des prédicteurs environnementaux dans les deux premiers axes. L'importance relative a été calculée sur la base du carré des coefficients sur les premier et deuxième vecteurs propres de la matrice de covariance des tailles d'effet.

7. Conclusions

Dans ce chapitre, nous avons validé les résultats théoriques présentés au chapitre précédent. Nous avons montré, à l'aide de données simulées et de données réelles, qu'il existait bien une relation linéaire entre les mesures de décalage génétique et le logarithme de la valeur sélective. Dans leur review, Rellstab, Dauphin et al., 2021 remarquent que la corrélation entre la structure de population et les prédicteurs environnementaux peut générer des prédictions biaisées de la part des statistiques de décalage génétique. Dans notre étude comparative, nous utilisons les facteurs latents de LFMM pour corriger les méthodes de décalage génétique. L'utilisation des facteurs latents comme variable de correction permet l'amélioration des performances prédictives de ces méthodes. Le fossé génétique résout la problématique des prédicteurs corrélés en modélisant la matrice de covariance des tailles d'effet des prédicteurs. Cette matrice de covariance permet également de mesurer l'importance des prédicteurs environnementaux à l'aide des valeurs propres et vecteurs propres de cette matrice. Lorsque plusieurs prédicteurs redondants sont présents lors de l'analyse statistique, le fossé génétique diminue l'importance de ces prédicteurs redondants.

Les statistiques de décalage génétique ont été estimées dans une expérience de jardin commun avec des populations de mil. Dans l'étude originale (Rhoné et al., 2020), le coefficient de détermination de la régression entre le décalage génétique (calculé avec GF) et la mesure de valeur sélective (approximée par le poids des graines) était de $r^2 \approx 9.5 - 17\%$. Dans notre étude, nous avons amélioré de manière significative les valeurs de coefficient de détermination atteignant une valeur de $r^2 \approx 61\%$ avec la méthode du fossé génétique. Ces résultats soutiennent les conclusions de Rhoné et al., 2020 et montrent la pertinence de l'utilisation du décalage génétique dans la prédiction des variations de valeur sélective face à des changements environnementaux. Cette étude nous a également permis de montrer que la température était la variable environnementale la plus impliquée dans le processus d'adaptation locale chez les populations de mil.

Toutefois il est important de souligner certaines limites de nos investigations empiriques. Nous n'avons exploré à l'aide de nos simulations qu'un nombre restreint de scénarios. Nous avons par exemple étudié des scénarios pour lesquels les traits phénotypiques sont déterminés à partir d'effets génétiques additifs. Il serait intéressant d'étudier d'autres types d'effets tel que la dominance ou l'épistasie. De plus, à l'exception de GF, les mesures de décalage génétique sont basées sur une relation linéaire entre le génotype et l'environnement. Cela génère des prédictions qui sont invariantes par translation dans la niche écologique. Cela rend les prédictions pertinentes au centre de l'aire de répartition mais peut être moins aux extrémités de cette dernière. Dans nos simulations, nous avons tenté d'utiliser des modèles non linéaires, mais cela donne des performances similaires aux modèles linéaires. Cela peut s'expliquer par le fait que bien qu'offrant des statistiques de décalage génétique plus flexibles que les modèles linéaires, les modèles non linéaires obtiennent un moins bon compromis biais-variance, peut être parce qu'un plus grand nombre de données seraient nécessaires pour leur application.

En conclusion de ces deux chapitres, nous avons développé un cadre théorique qui relie les statistiques de décalage génétique à une géométrie non euclidienne de la niche écologique. La théorie quantitative a permis de proposer une interprétation en terme de valeur sélective dans l'environnement modifié, unifiant plusieurs approches existantes et abordant certaines de leurs limites. Sur la base de simulations numériques approfondies et de données recueillies dans une expérience de jardin commun, notre étude a indiqué que les statistiques de décalage génétique sont des outils importants pour la gestion de la conservation face au changement climatique. Les résultats du chapitre 3 et 4 ont fait l'objet d'une publication pour l'instant disponible dans bioRxiv (Gain, Rhone et al., 2023)

Chapitre 5

Théorie spectrale des indices de fixation de Wright

L'indice de fixation de Wright, F_{ST}, est une mesure fondamentale de la génétique des populations. En supposant connue la population de chaque échantillon, cette statistique est utilisée pour évaluer la structure de la population à un locus génomique donné. Toutefois, lorsqu'on travaille sur des donnnées en grande dimension, les approches non supervisées telles que l'analyse en composantes principales (ACP) ont pris une place importante dans l'étude de la structure de population. Dans ce chapitre, nous allons mettre en avant les liens existants entre les indices de fixation de Wright et l'ACP dans un modèle à K populations. Notre théorie fournit une définition équivalente de l'indice de fixation F_{ST} basée sur la décomposition de la matrice de génotype en une matrice intra-population et une matrice inter-population. Nous montrerons que la valeur moyenne de F_{ST} le long du génome peut être obtenue à partir de l'ACP de la matrice inter-population. De plus, en supposant qu'une condition de séparation que nous définirons est vérifiée et que le jeu de données est suffisamment grand, on peut également approcher cette valeur par la part de variance expliquée par les $(K-1)$ composantes principales de la matrice de génotype. Ces résultats nous permettront notamment d'étendre la notion de F_{ST} à des matrices de génotype modifiées pour lesquelles nous n'avons pas accès aux fréquences d'allèles. La théorie développée nous permet donc d'interpréter les résultats de l'ACP à partir de concepts de génétique de population et permet d'étendre le concept d'indice de fixation à des études de génétique de population prenant en compte les effets temporels, géographiques et environnementaux.

1. Introduction

1. 1. Vers une interprétation des valeurs propres de l'ACP de la matrice de génotype en génétique des populations

Défini par Sewall Wright et Gustave Malécot, l'indice de fixation ou coefficient de consanguinité, F_{ST}, mesure la part de variance expliquée par la structure de population. Autrement dit, il mesure la quantité de variation génétique trouvée entre les populations par rapport à la quantité trouvée au sein des populations (Wright, 1965 ; Malécot, 1948). L'indice de fixation est notamment utilisé comme mesure de la structure de population et fait partie des statistiques les plus utilisées en génétique des populations (Cockerham, 1969) (Nei, 1973) (Slatkin, 1991) (Holsinger et Weir, 2009). La statistique a été initialement utilisée pour l'étude de cette structure à l'échelle d'un seul locus. Avec la quantité de données génétiques disponibles de nos jours, le besoin d'utiliser des méthodes permettant de rendre compte de la structure à l'échelle de plusieurs locus est apparu. Plusieurs méthodes sont apparues pour répondre à ce besoin tel que le logiciel STRUCTURE (Pritchard et al., 2000), TESS3 (Caye, F. Jay et al., 2018). Parmi elles, on retrouve la méthode de l'analyse en composante principale (ACP) (I. Jolliffe, 1986) (Menozzi et al., 1978).

A partir d'une matrice de génotype où les colonnes sont centrées ou standardisées, l'ACP calcule les valeurs propres et les vecteurs propres de la matrice de covariance de l'échantillon. Les valeurs propres et les vecteurs propres peuvent être calculés efficacement en utilisant la décomposition en valeurs singulières (SVD) de la matrice de données centrée (I. T. Jolliffe et Cadima, 2016). L'ACP permet l'obtention de différents résultats d'intérêts notamment liés au métissage des échantillons à partir des populations sources, la projection des échantillons sur les axes de l'ACP et les valeurs propres de l'ACP. Alors que le lien entre la projection des échantillons et le métissage est aujourd'hui bien compris, l'interprétation des valeurs propres de l'ACP est encore difficile. Les principales contributions allant dans ce sens se restreignent à des modèles de divergence à deux populations. Ces résultats se basent sur la théorie des matrices aléatoires (TMA) (Patterson et al., 2006) (Bryc et al., 2013) et sur la théorie de la coalescence (McVean, 2009). En se basant sur la TMA, (Patterson et al., 2006) propose un seuil pour la valeur de F_{ST} en dessous duquel on ne peut conclure quant la présence de structure de population. Pour un modèle de divergence à deux populations, (McVean, 2009) utilise des résultats théoriques de temps de coalescence pour démontrer une relation entre la plus grande valeur propre de l'ACP et la F_{ST}.

Ce chapitre vise à développer une théorie spectrale de la matrice de génotype afin de prolonger la compréhension de la relation entre l'ACP et les coefficients de Wright dans un modèle de populations discrètes. Cette théorie suppose que les génotypes observés correspondent à l'échantillonnage de K populations discrètes. Cette théorie s'appuie sur une décomposition de la matrice de génotype en une matrice inter-population et une matrice intra-population. Le résultat principal établit que la valeur moyenne de F_{ST} le long du génome est égale à la norme de Hilbert-Schmidt au carré de la matrice inter-population. Cette norme peut être obtenue par une analyse spectrale. Sous réserve de la validation d'une condition de séparation entre les matrices intra et inter population, nous montrerons que la somme des $(K-1)$ premières valeurs propres de l'ACP standardisée est un moyen d'approcher la valeur moyenne de la statistique de F_{ST} le long du génome. Afin de décrire la variance résiduelle qui n'est pas expliquée par le modèle de population discrète, on s'appuie sur une approximation des valeurs propres de la matrice intra-population via la TMA (Patterson et al., 2006 ; Johnstone et Paul, 2018). La théorie ainsi établie nous fournit une définition alternative des coefficients de fixation nous permettant d'étendre la notion de F_{ST} à des génotypes modifiés. Pour illustrer cette nouvelle

définition, on calculera l'indice de fixation généralisé pour des données d'ADN ancien d'humains après correction pour la couverture génomique et les modifications dues à la différence de date de l'échantillonnage (François et Jay, 2020). Dans une deuxième illustration, nous calculons la statistique de F_{ST} pour les échantillons scandinaves de l'espèce de plante *Arabidopsis thaliana* après avoir retiré la variation génétique associées à certaines variables environnementales extraites d'une base de données climatiques (I. Wang et al., 2013 ; Alonso-Blanco et al., 2016).

1. 2. Notations

Ce chapitre fera appel à plusieurs notations que nous résumons dans la boîte 1

Boîte 1. Notations

n	Taille d'échantillons
L	Nombre de locus
n_k	Le nombre d'échantillons présents dans la population k
c_k	La proportion d'échantillons présents dans la population k
F_{ST}	Indice de fixation de Wright, obtenu avec la formule de Nei avec correction pour les tailles d'échantillons inégales
H_{S}	Diversité génétique intra-population
H_{T}	Diversité génétique totale
D_{ST}	Diversité génétique inter-population
$\mathbf{X}$	Matrice de SNPs pour n individus à L locus
$\mathbf{P}$	Vecteur des fréquences de SNPs pour les L locus
$\mathbf{Z}$	Matrice de génotypes centrés, $\mathbf{X} - \mathbf{P}$
$\mathbf{Z}^{\mathrm{sc}}$	Matrice de génotypes standardisés, $\mathbf{Z}/\sqrt{\mathbf{P}(1-\mathbf{P})}$
$\mathbf{Z}_{\mathrm{ST}}$	Une matrice de taille $n \times L$ décrivant les données inter-populations répétés pour chaque individu de la même population
$\mathbf{Z}_{\mathrm{S}}$	Une matrice de taille $n \times L$ décrivant les données intra-population
$\sigma_k^2(\mathbf{Z})$	Valeur propre de la matrice de covariance empirique $\mathbf{ZZ^T/n}$ (PCA non standardisée)
$\rho_k^2(\mathbf{Z}^{\mathrm{sc}})$	Valeur propre de la matrice de corrélation empirique $\mathbf{Z^{sc}Z^{scT}}/n$ (PCA standardisée) , égal à L fois la part de variance expliquée par les composantes principales

2. Partition de la variation génétique en deux matrices distinctes, lien entre valeur propre et F_{ST}

Dans cette section, nous commencerons par décrire la manière dont nous partitionnons la matrice de génotype en deux matrices distinctes, une matrice intra-population et une matrice inter-population. Nous donnerons ensuite le théorème qui spécifie l'égalité entre la valeur moyenne de F_{ST} le long du génome et la part de variance expliquée par les $K-1$ premiers axes de l'ACP de la matrice inter-population. Dans toute cette section, on utilise la notation $\mathbb{E}[Q] = \sum_{\ell=1}^{L} Q_\ell / L$ pour exprimer la valeur moyenne de Q_ℓ le long du génome.

2. 1. Partition de la variation génétique

Considérons un échantillon de n individus pour lesquels un grand nombre de locus a été génotypé, donnant lieu à une matrice $\mathbf{X} = (x_{i\ell})$, avec n lignes et L colonnes. Pour les individus haploïdes, on pose $x_{i\ell} = 0, 1$, et pour les diploïdes $x_{i\ell} = 0, 1, 2$ pour compter le nombre d'allèles dérivés au locus ℓ pour l'individu i. Nous simplifions ce chapitre en considérant qu'un échantillon de diploïdes peut être représenté par un échantillon d'haploïdes de deux fois la taille de l'échantillon initial. Bien que cela soit une condition non nécessaire, les locus sont considérés comme non liés. En suivant l'approche de Wright dans la description de la structure de population, notre hypothèse principale est que les individus sont issus de K populations discrètes prédéfinies.
Pour analyser la structure de population, l'ACP peut être réalisée après avoir centré ou standardisé la matrice de génotype. La matrice standardisée est notée $\mathbf{Z}^{\text{sc}}$, la matrice centrée $\mathbf{Z}$. L'ACP standardisée calcule les valeurs propres, $\rho_k^2(\mathbf{Z}^{\text{sc}})$, de la matrice de corrélation empirique. L'ACP non standardisée calcule les valeurs propres, $\sigma_k^2(\mathbf{Z})$, de la matrice de covariance empirique (I. Jolliffe, 1986 ; Johnstone et Paul, 2018). Les valeurs propres sont rangées par ordre décroissant et $\rho_k^2(\mathbf{Z}^{\text{sc}})/L$ est souvent interprétée comme la part de variance expliquée par le $k$ème axe de l'ACP. L'ACP peut être réalisée via l'algorithme de décomposition en valeurs singulières (ou singular value decomposition SVD). Dans ce cas, les valeurs propres de l'ACP standardisée (ou non standardisée) correspondent au valeurs singulières au carré de la matrice standardisée (ou centrée) divisée par $\sqrt{n}$ (I. Jolliffe, 1986 ; Johnstone et Paul, 2018).
Afin d'établir la relation entre l'ACP et les coefficients de fixation, nous décomposons la matrice centrée en une somme de deux matrices, $\mathbf{Z} = \mathbf{Z}_{\text{ST}} + \mathbf{Z}_{\text{S}}$, correspondant aux composantes inter et intra population. La décomposition est réalisée de la manière suivante. Soit i un individu échantillonné de la population k. A un locus particulier, ℓ, le génotype, $x_{i\ell}$, est égal à 0 ou 1, et $p_{k\ell}$ correspond à la fréquence de l'allèle dérivé dans la population k à ce locus. Le coefficient de la matrice centrée, $z_{i\ell}$, est égal à

$$z_{i\ell} = \sum_{j \neq k} c_j (p_{k\ell} - p_{j\ell}) + (x_{i\ell} - p_{k\ell})$$

où $c_k = n_k/n$ représente la proportion d'individus échantillonnés de la population k. Avec la formulation suivante, la matrice inter-population, $\mathbf{Z}_{\text{ST}}$, a pour terme général

$$z_{i\ell}^{st} = \sum_{j \neq k} c_j (p_{k\ell} - p_{j\ell})$$

répété pour tous les individus de la population k. Par construction, le rang de $\mathbf{Z}_{\text{ST}}$ est égal à $(K-1)$. La matrice intra-population, $\mathbf{Z}_{\text{S}}$, a pour terme général

$$z_{i\ell}^{s} = x_{i\ell} - p_{k\ell}$$

La décomposition peut facilement être étendue à la matrice standardisée, définie par $z_{i\ell}^{sc} = z_{i\ell}/\sqrt{P_\ell(1-P_\ell)}$, où $P_\ell = \sum_{k=1}^{K} c_k p_{k\ell}$ est la fréquence de l'allèle dérivée dans l'échantillon total au locus ℓ. Nous disposons désormais de tous les éléments pour établir le théorème central de ce chapitre.

2. 2. Théorème spectral des indices de fixation

Commençons par rappeler la définition des deux mesures d'intérêts notées D_{ST} et F_{ST}. Considérons n échantillons issus de K populations discrètes et définissons D_{ST} et F_{ST} d'après Wright, 1965 et Nei, 1973 ; Nei et Chesser, 1983, avec des tailles de populations inégales. À un locus

particulier, on pose $H_{\mathrm{S}} = 2\sum_{k=1}^{K} c_k p_k (1-p_k)$ et $H_{\mathrm{T}} = 2P(1-P)$, on a alors $D_{\mathrm{ST}} = H_{\mathrm{T}} - H_{\mathrm{S}}$. Le coefficient de consanguinité de Wright est défini comme tel :

$$F_{\mathrm{ST}} = D_{\mathrm{ST}}/H_{\mathrm{T}}\,.$$

Le résultat principal établit que la valeur moyenne de l'indice de fixation le long du génome peut être calculé à partir des valeurs singulières de la matrice inter-population, $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}$. Une relation similaire est également établie pour D_{ST} et la matrice non standardisée $\mathbf{Z}_{\mathrm{ST}}$. Les valeurs singulières des matrices intra et inter-population peuvent être obtenues à l'aide de l'algorithme SVD. Le coût computationnel de ces opérations est d'ordre $O(n^2 L)$. On peut réduire ce coût à l'aide de méthodes variées, comme par exemple le fait de calculer uniquement les $K-1$ premières valeurs singulières uniquement. Les conclusions décrites ci dessous sont valides peu importe le modèle de génétique des populations utilisé, indépendamment du seuil de fréquence de l'allèle minoritaire choisi, et lorsque les locus sont liés physiquement.

Théorème 1. *Soit $K \geq 2$. Soit $\mathbf{Z}$ et $\mathbf{Z}^{\mathrm{sc}}$ telles que décrites dans la boîte 1. En définissant $\mathbf{Z}_{\mathrm{ST}}$ et $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}$ comme dans la section précédente, on a*

$$\mathbb{E}\left[F_{\mathrm{ST}}\right] = \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}})/L\,, \tag{5.1}$$

et

$$\mathbb{E}\left[D_{\mathrm{ST}}\right]/2 = \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L\,. \tag{5.2}$$

2. 3. Démonstration du théorème

La démonstration du théorème repose sur plusieurs éléments. Le premier point de cette démonstration consiste à donner une autre formulation de l'expression de la D_{ST}. Le second point consiste à partir de la norme de Hilbert-Schmidt de la matrice Z_{ST} et de développer l'expression afin de retrouver le résultat du théorème.

Reformulation de l'expression de la D_{ST} Reprenons les notations des parties précédentes et ajoutons $q_k = 1 - p_k$. Rappelons que $D_{\mathrm{ST}} = H_{\mathrm{T}} - H_{\mathrm{S}}$. On peut alors reformuler :

$$D_{\mathrm{ST}} = 2\left(\sum_{j=1}^{K}\sum_{k=1}^{K} c_j c_k p_j q_k - \sum_{j=1}^{K} c_j p_j q_j\right)\,.$$

qui peut se réécrire de la façon suivante

$$D_{\mathrm{ST}} = 2\left(\sum_{j=1}^{K} c_j p_j \sum_{k=1}^{K} c_k (p_j - p_k)\right) = 2\sum_{j<k} c_j c_k (p_j - p_k)^2\,.$$

En développant l'expression

$$A = \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j (p_j - p_k)\right)^2\,,$$

on trouve l'égalité suivante

$$D_{\mathrm{ST}} = 2\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j (p_j - p_k)\right)^2\,.$$

Rappel sur la norme de Hilbert-Schmidt Soit $\mathbf{X}$ une matrice quelconque de dimension $n \times L$, la norme de Hilbert-Schmidt, ou norme de Frobenius, de $\mathbf{X}$ est définie de la manière suivante :

$$\|\mathbf{X}\| = \left(\sum_{i=1}^{n} \sum_{\ell=1}^{L} x_{i\ell}^2 \right)^{1/2} .$$

Cette norme peut se réécrire de la manière suivante :

$$\|\mathbf{X}\|^2 = \mathbf{Tr}(\mathbf{X}^T \mathbf{X}).$$

Or, la trace d'une matrice correspond à la somme de ces valeurs propres. Ainsi :

$$\|\mathbf{X}\|^2 = n \sum_{i=1}^{\min(n,L)} \sigma_i^2(\mathbf{X}) \,,$$

Obtention des résultats du théorème Rappelons d'abord l'expression de $\mathbf{Z}_{\mathrm{ST}}$. Pour un individu i issu de la population k et un locus arbitraire ℓ, le terme général de $\mathbf{Z}_{\mathrm{ST}}$ est

$$z_{i\ell}^{st} = \sum_{j=1}^{K} c_j (p_{k\ell} - p_{j\ell}) \,,$$

Ce terme est répété n_k fois pour chaque locus, pour chaque individu i de la population k. Les fréquences d'allèles, $p_{k\ell}$, sont spécifiques à chaque locus. La norme de la matrice inter-population s'écrit :

$$\|\mathbf{Z}_{\mathrm{ST}}\|^2 = \sum_{i=1}^{n} \sum_{\ell=1}^{L} (z_{i\ell}^{st})^2 \,,$$

En développant, on obtient l'expression suivante :

$$\|\mathbf{Z}_{\mathrm{ST}}\|^2 = nL\, \mathbb{E}\left[\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j (p_j - p_k) \right)^2 \right] = nL\, \mathbb{E}[D_{\mathrm{ST}}]/2 \,.$$

Comme Z_{ST} est de rang $K-1$ et d'après l'expression de la norme donnée dans la partie précédente, on déduit alors :

$$nL\, \mathbb{E}[D_{\mathrm{ST}}]/2 \; = n \sum_{i=1}^{K-1} \sigma_i^2(\mathbf{Z_{ST}}),$$

en faisant passer n et L de l'autre côté de l'égalité on obtient finalement l'expression donnée dans le Théorème 1 :

$$\mathbb{E}\left[D_{\mathrm{ST}}\right]/2 = \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L \,.$$

La matrice standardisée, $\mathbf{Z}^{\mathrm{sc}}$, peut être obtenue à partir de $\mathbf{Z}$ en divisant chaque colonne par $\sqrt{P_\ell(1-P_\ell)}$. On obtient ainsi la seconde expression du théorème :

$$\mathbb{E}\left[F_{\mathrm{ST}}\right] = \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}})/L \,.$$

2. 4. Conclusions

Dans cette section, nous avons proposé un partitionnement de la matrice de génotype en une matrice intra-population, rendant compte de la variance au sein des populations, et une matrice inter-population, rendant compte de la variance entre les populations. Nous avons ensuite établi un théorème qui fournit une expression de la valeur moyenne des statistiques F_{ST} et D_{ST} en fonction des valeurs singulières de la matrice inter-population. Nous avons finalement donné la démonstration de ce théorème qui repose sur une reformulation de l'expression de la D_{ST} et sur un réagencement des termes de la norme de Hilbert-Schmidt de la matrice inter-population. Nous allons désormais montrer qu'il est possible d'étendre ce résultat aux valeurs propres de la matrice de génotype centrée, ou standardisée (et non sa composante inter-population).

3. Extension des résultats aux valeurs propres de l'ACP de la matrice de génotype

Dans cette section, nous allons montrer que la valeur moyenne de F_{ST} le long du génome peut être approchée par les $K-1$ plus grandes valeurs propres de l'ACP de la matrice de génotype. Le résultat est proche de celui de la section précédente mais s'applique directement à la matrice de génotype et non à la matrice inter-population.

3. 1. Hypothèse issue de la théorie des matrices aléatoires

Pour énoncer le résultat de cette section, nous allons nous appuyer sur une hypothèse issue de la théorie des matrices aléatoires (Patterson et al., 2006 ; Marčenko et L.A., 1967 ; Johnstone, 2001 ; Johnstone, 2008 ; Bryson et al., 2019). Nous allons, dans cette sous-section, préciser en quoi consiste cette hypothèse. Rappelons que nous partitionnons la matrice de génotype en deux matrices, une matrice inter-population et une matrice intra-population. Notre hypothèse consiste à dire que la distribution de la part de variance expliquée par chacun des axes principaux de la matrice Z_S peut être approchée par la densité de probabilité de Marchenko-Pastur :

$$p(x) = L\frac{\sqrt{(x_M - x)(x - x_m)}}{2x\pi}\,, \quad x_m = (1-\sqrt{\gamma})^2/L \leq x \leq x_M = (1+\sqrt{\gamma})^2/L\,,$$

la part de variance expliquée par la première composante principale est approchée par $(1/\sqrt{L} + 1/\sqrt{n-1})^2$. Nous n'avons pas démontré théoriquement que cette hypothèse est valide dans les modèles considérés par la suite (F-modèles, données empiriques) mais nous constaterons dans la section suivante qu'elle est vérifiée empiriquement dans le cadre de simulation et de données réelles.

3. 2. Énoncé du résultat

Nous avons établi dans la section précédente que la valeur moyenne de l'indice de fixation peut être obtenue à partir des valeurs singulières de la matrice inter-population. Nous étendons désormais ce résultat aux valeurs propres de l'ACP de la matrice de génotype standardisée.

$$\mathbb{E}\left[F_{\mathrm{ST}}\right] \approx \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}^{\mathrm{sc}})/L\,, \tag{5.3}$$

et pour l'ACP de la matrice centrée,

$$\mathbb{E}\left[D_{\mathrm{ST}}\right] \approx 2\sum_{k=1}^{K-1}\sigma_k^2(\mathbf{Z})/L\,. \tag{5.4}$$

Ce résultat est vérifié sous les conditions suivantes. Il requiert tout d'abord que les valeurs propres de $\mathbf{Z}/\sqrt{n}$ triées par ordre décroissant correspondent à une approximation des valeurs propres de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ suivies par une approximation des valeurs propres de $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$. Dit autrement, la plus petite valeur singulière non nulle de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ doit être plus grande que la plus grande valeur singulière de $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$.

$$\sigma_{K-1}^2(\mathbf{Z}_{\mathrm{ST}}) > \sigma_1^2(\mathbf{Z}_{\mathrm{S}})\,. \tag{5.5}$$

On suppose le rapport L/n constant pour des grandes valeurs de L et n, et effectuons les hypothèses suivantes : 1) La condition de séparation (5.4) est vérifiée, 2) La plus grande valeur propre de l'ACP de $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$ est d'ordre $(1/\sqrt{n}+1/\sqrt{L})^2$ (Hypothèse TMA). Alors, sous ces conditions, la précision de l'approximation dans les équations (5.3) et (5.4) est d'ordre $O(K/L)$. Dit autrement, pour toute valeur singulière, $\sigma_k(\mathbf{Z}_{\mathrm{ST}})$, de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, Il existe une valeur singulière, $\sigma_k(\mathbf{Z})$, de $\mathbf{Z}/\sqrt{n}$ tel que nous avons :

$$\left|\sigma_k^2(\mathbf{Z})/L - \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L\right| = O(1/L)\,,\quad k=1,\ldots,K-1\,.$$

Un résultat similaire existe pour les $K-1$ première valeurs propres de l'ACP de la matrice standardisée, $\rho_k^2(\mathbf{Z}^{\mathrm{sc}})$ et $\rho_k^2(\mathbf{Z}^{\mathrm{sc}}_{\mathrm{ST}})$. Cela prouve que la valeur moyenne de F_{ST} le long du génome peut être approchée par la somme des $(K-1)$ plus grandes valeurs propres de l'ACP avec une précision proportionnelle au nombre de populations et à l'inverse du nombre de locus dans la matrice de génotype. Procédons désormais à la démonstration de ce résultat.

3. 3. Démonstration du résultat

Nous cherchons donc à prouver dans cette partie que, pour toute valeur singulière non nulle, $\sigma_k(\mathbf{Z}_{\mathrm{ST}})$, de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, il existe une valeur singulière $\sigma_k(\mathbf{Z})$, de $\mathbf{Z}/\sqrt{n}$, de telle sorte que nous avons :

$$|\sigma_k^2(\mathbf{Z})/L - \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L| \leq C/L\,,\quad k=1,\ldots,K-1\,.$$

Commençons par remarquer que les matrices $\mathbf{Z}_{\mathrm{ST}}$ et $\mathbf{Z}_{\mathrm{S}}$ satisfont la condition d'orthogonalité : $\mathbf{Z}_{\mathrm{ST}}^T\mathbf{Z}_{\mathrm{S}} = \mathbf{0}$. Cela se vérifie en effectuant le produit de matrice pour un coefficient quelconque. Ces coefficients sont égaux à :

$$\sum_{i=1}^{n} z_{i\ell}^{st} z_{im}^{s} = n\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j(p_{k\ell}-p_{j\ell})\right) \sum_{i\in \mathrm{pop_k}} (x_{i\ell}-p_{k\ell})\,.$$

Or, quelque soit k, nous avons

$$\sum_{i\in \mathrm{pop_k}} (x_{i\ell}-p_{k\ell}) = 0\,.$$

La condition d'orthogonalité est donc vérifiée.
Pour $\sigma = \sigma(\mathbf{Z}_{\mathrm{ST}})$, une valeur singulière non nulle de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, et $\tilde{\sigma} = \tilde{\sigma}(\mathbf{Z}_{\mathrm{S}})$, une valeur singulière non-nulle de $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$, les vecteurs singuliers gauche $\mathbf{u}$ et $\tilde{\mathbf{u}}$, associés à σ et $\tilde{\sigma}$ sont des vecteurs orthogonaux. Pour le prouver, calculons

$$n^2\sigma^2\tilde{\sigma}^2\mathbf{u}^T\tilde{\mathbf{u}} = \mathbf{u}^T\mathbf{Z}_{\mathrm{ST}}(\mathbf{Z}_{\mathrm{ST}}^T\mathbf{Z}_{\mathrm{S}})\mathbf{Z}_{\mathrm{S}}^T\tilde{\mathbf{u}}.$$

Comme $\mathbf{Z}_{\mathrm{ST}}^T\mathbf{Z}_{\mathrm{S}} = \mathbf{0}$, le terme du milieu est nul et donc l'orthogonalité est vérifiée.
Montrons maintenant que les valeurs singulières de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ sont des approximations des valeurs singulières de $\mathbf{Z}/\sqrt{n}$. La démonstration est similaire pour les valeurs singulières de $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$. Supposons que $L/n = \gamma$ pour de grandes valeurs de L et n, soit $\mathbf{u}$ un vecteur singulier gauche de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, $\mathbf{v}$ le vecteur singulier droit associé et σ la valeur singulière correspondante. On a alors

$$\mathbf{Z}\mathbf{Z}^T\mathbf{u}/n = \mathbf{Z}_{\mathrm{ST}}\mathbf{Z}_{\mathrm{ST}}^T\mathbf{u}/n + \mathbf{Z}_{\mathrm{S}}\mathbf{Z}_{\mathrm{ST}}^T\mathbf{u}/n.$$

Car $\mathbf{Z} = \mathbf{Z_{ST}} + \mathbf{Z_S}$ et $\mathbf{u} = \mathbf{Z_{ST}}\mathbf{v}/\sigma$ donc $\mathbf{Z_S^T u} = 0$. Pour le premier terme, $\mathbf{Z}_{\mathrm{ST}}\mathbf{Z}_{\mathrm{ST}}^T\mathbf{u}/n = \sigma^2\mathbf{u}$. Pour le second terme, $\mathbf{Z}_{\mathrm{ST}}^T\mathbf{u}/\sqrt{n} = \sigma\mathbf{v}$, et alors

$$\mathbf{Z}\mathbf{Z}^T\mathbf{u}/n = \sigma^2\mathbf{u} + \sigma\mathbf{Z}_{\mathrm{S}}\mathbf{v}/\sqrt{n}\,.$$

Par définition du rayon spectral, on a

$$\|\mathbf{Z}_{\mathrm{S}}\mathbf{v}/\sqrt{n}\|^2 \leq \mathrm{radius}(\mathbf{Z}_{\mathrm{S}}\mathbf{Z}_{\mathrm{S}}^T/n)\,.$$

Or, la valeurs propre σ^2 est d'ordre $O(L)$ (On a $\sigma^2 < \|\mathbf{Z}_{\mathrm{ST}}\|^2/n = O(L)$ d'après le Théorème 1). D'après l'hypothèse 2) pour $\mathbf{Z}_{\mathrm{S}}$, on a

$$\sigma^2\|\mathbf{Z}_{\mathrm{S}}\mathbf{v}/\sqrt{n}\|^2 = (1+\sqrt{\gamma})^2 \times O\left(1\right)\,.$$

Mises ensemble, les dernières équations nous donnent

$$\mathbf{Z}\mathbf{Z}^T\mathbf{u}/n = \sigma^2\mathbf{u} + O(1)\,.$$

Ainsi pour des grandes valeurs de n et L, $\mathbf{Z}\mathbf{Z}^T\mathbf{u}/n \approx \sigma^2\mathbf{u}$ plus un terme négligeable, ce qui signifie que $\mathbf{u}$ est une approximation d'un vecteur singulier de $\mathbf{Z}/\sqrt{n}$, et σ^2 une approximation d'une valeur propre de $\mathbf{Z}/\sqrt{n}$. Maintenant, considérons l'espace vectoriel engendré par les $K-1$ premières valeurs propres de $\mathbf{C}_{\mathrm{ST}} = \mathbf{Z}_{\mathrm{ST}}\mathbf{Z}_{\mathrm{ST}}^T/n$. On a alors :

$$\mathbf{C}_{\mathrm{ST}}\mathbf{U} = \mathbf{U}\Sigma\,,$$

où $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{K-1}^2)$ et $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{K-1})$ sont les valeurs propres et les vecteurs singuliers gauches de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$. Pour $\mathbf{C} = \mathbf{Z}\mathbf{Z}^T/n$, on a :

$$\mathbf{U}^T\mathbf{C}\mathbf{U} = \Sigma + \mathbf{U}^T O(1)\,.$$

Le conditionnement de $\mathbf{U}$ étant égal à 1, on peut appliquer le théorème de Bauer-Fike (Bauer et Fike, 1960 ; Bhatia, 2013). Pour chaque valeur propre non nulle de $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, $\sigma^2(\mathbf{Z}_{\mathrm{ST}})$, il existe une valeur propre de $\mathbf{Z}/\sqrt{n}$, $\sigma^2(\mathbf{Z})$, telle que

$$|\sigma^2(\mathbf{Z}) - \sigma^2(\mathbf{Z}_{\mathrm{ST}})| = O(1)\,.$$

En reprenant les résultats de la section précédente :

$$\|\mathbf{Z}_{\mathrm{ST}}\|^2 = nL\mathbb{E}[D_{\mathrm{ST}}/2],$$

et

$$\|\mathbf{Z}_{\mathrm{ST}}\|^2 = \mathbf{Tr}(\mathbf{Z}_{\mathrm{ST}}\mathbf{Z}_{\mathrm{ST}}^T) = n\sum_{k=1}^{K-1}\sigma_k^2(\mathbf{Z}_{\mathrm{ST}})\,.$$

Alors, sous la condition de séparation et l'hypothèse TMA, on a :

$$\sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}) \approx \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})\,,$$

et

$$L \times \mathbb{E}[D_{\mathrm{ST}}]/2 = \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}_{\mathrm{ST}}) \approx \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z})\,.$$

Pour l'ACP centrée, on obtient que $L\mathbb{E}[D_{\mathrm{ST}}]/2$ est proche de la somme des $(K-1)$ plus grandes valeurs propres de $\mathbf{Z}/\sqrt{n}$. Les résultats s'étendent facilement à l'ACP standardisée. Comme la matrice de corrélation empirique est obtenue en standardisant les colonnes de $\mathbf{Z}$, une décomposition spectrale est valide pour la matrice standardisée également. Pour l'ACP standardisée, la somme des $K-1$ plus grandes valeurs propres est égale à

$$L \times \mathbb{E}[F_{\mathrm{ST}}] = \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}_{\mathrm{ST}}) \approx \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z})\,.$$

En d'autres termes, $\mathbb{E}[F_{\mathrm{ST}}]$ peut être approchée par la part de variance expliquée par les $(K-1)$ premières composantes principales de la matrice de génotype. L'espérance de la précision de l'approximation est d'ordre $O(K/L)$.

3. 4. Conclusions

On vient donc de montrer que, pour une matrice de génotypes constituées de K populations, la part de variance expliquée par les $K-1$ premières composantes principales de la matrice est une manière d'approcher $\mathbb{E}[F_{\mathrm{ST}}]$ le long du génome. Dans cette section et dans la précédente, nous avons démontré théoriquement la validité des résultats sous certaines hypothèses. Nous proposons désormais de valider ces résultats de manière empirique à l'aide de simulation et de données réelles.

4. Validation des résultats précédents à l'aide du F-modèle et de données réelles

Nous proposons dans cette section de vérifier les résultats énoncés dans la partie précédente à l'aide de simulation et de données réelles. Les simulations s'appuieront sur le F-modèle (Balding et Nichols, 1995) et les données réelles seront issues d'échantillons de populations humaines provenant du 1000 Genomes Project (Consortium et al., 2015).

4. 1. F-modèle

Les F-modèles sont des modèles de populations discrètes dans lesquels K populations divergent d'une diversité génétique ancestrale. Les modèles sont sans mutation, migration ou sélection. Dans la diversité génétique ancestrale, la fréquence de l'allèle dérivé est égale à p_{anc}. Les populations divergent les unes des autres avec des coefficients de dérive spécifique à chaque population, F_k, relativement à la population ancestrale. Conditionnellement à p_{anc}, la fréquence d'allèle à un locus particulier dans la population k suit une distribution beta de paramètres $p_{\mathrm{anc}}(1-F_k)/F_k$ et $(1-p_{\mathrm{anc}})(1-F_k)/F_k$. Pour créer une distribution le long des L locus, p_{anc} est tiré d'une densité de probabilité beta ayant pour paramètres a and b. Les paramètres du F-modèle sont choisis

de tel sorte que $\mathbb{E}[p_k|p_{\text{anc}}] = p_{\text{anc}}$ et $\text{Var}(p_k|p_{\text{anc}}) = p_{\text{anc}}(1-p_{\text{anc}})F_k$, pour tout k. L'espérance de l'hétérozygotie ancestrale est égale à

$$\mathbb{E}[H_{\text{A}}] = \int_0^1 2p_{\text{anc}}(1-p_{\text{anc}})f(p_{\text{anc}})dp_{\text{anc}} = \frac{2ab}{(a+b)(a+b+1)}\,.$$

Dans cette section, nous comparerons les premières valeurs propres de l'ACP avec l'espérance des valeurs de F_{ST} et D_{ST}. Les comparaisons se feront à partir des valeurs moyennes mais également avec les valeurs théoriques. La sous-section suivante calcule les valeurs théoriques de D_{ST}, H_S et H_T dans le F-modèle

4. 2. Valeur attendue dans le F-modèle

Dans cette sous-section, nous allons donner l'expression de la valeur théorique D_{ST}, H_S et H_T dans le F-modèle. Puis, nous donnerons la valeur attendue des valeurs propres de l'ACP dans un modèle à 3 populations.

D_{ST}, H_S et H_T Rappelons que $D_{\text{ST}} = 2\mathbb{E}[\sum_{j<k} c_j c_k (p_j - p_k)^2]$. Sachant p_{anc}, l'espérance conditionnelle de la différence des fréquences d'allèles au carré $(p_j - p_k)^2$ est égale à

$$\mathbb{E}[(p_j - p_k)^2|p_{\text{anc}}] = \mathbb{E}[(p_j - p_{\text{anc}})^2|p_{\text{anc}}] + \mathbb{E}[(p_k - p_{\text{anc}})^2|p_{\text{anc}}]\,.$$

La somme des carrés découle de l'indépendance conditionnelle des fréquences alléliques p_j et p_k étant donné p_{anc}. Cela nous donne

$$\mathbb{E}[(p_j - p_k)^2|p_{\text{anc}}] = p_{\text{anc}}(1-p_{\text{anc}})(F_j + F_k)\,.$$

Ainsi

$$\mathbb{E}[\sum_{j\neq k} c_j c_k (p_j - p_k)^2|p_{\text{anc}}] = p_{\text{anc}}(1-p_{\text{anc}}) \sum_{j\neq k} c_j c_k (F_j + F_k)\,.$$

En réorganisant la somme on obtient

$$\sum_{j\neq k} c_j c_k F_j = \sum_{j=1}^{K} c_j F_j \sum_{k\neq j}^{K} c_k = \sum_{j=1}^{K} c_j(1-c_j)F_j\,,$$

on peut alors conclure

$$\mathbb{E}[D_{\text{ST}}] = \left(\sum_{k=1}^{K} c_k(1-c_k)F_k\right)\mathbb{E}[H_{\text{A}}]\,.$$

A l'aide de calculs similaires, les résultats pour H_{T} et H_{S} sont donnés par

$$\mathbb{E}[H_{\text{T}}] = \left(1 - \sum_{k=1}^{K} c_k^2 F_k\right)\mathbb{E}[H_{\text{A}}]\,.$$

et

$$\mathbb{E}[H_{\text{S}}] = \left(1 - \sum_{k=1}^{K} c_k F_k\right)\mathbb{E}[H_{\text{A}}]\,.$$

Modèle à 3 populations Pour des modèles à 3 populations de taille égale et pour des fréquences d'allèles ancestrales distribuées selon une distribution uniforme beta($a = 1$, $b = 1$), la matrice Λ correspondant à la matrice de covariance inter-population pour les fréquences d'allèles est la suivante

$$\mathbf{\Lambda} = \frac{1}{162}\begin{pmatrix} 4F_1 + F_2 + F_3 & F_3 - 2(F_1 + F_2) & F_2 - 2(F_1 + F_3) \\ * & F_1 + 4F_2 + F_3 & F_1 - 2(F_2 + F_3) \\ * & * & F_1 + F_2 + 4F_3 \end{pmatrix},$$

où les étoiles correspondent à des coefficients symétriques. A l'aide d'algébre linéaire élémentaire, on obtient la première valeur propre de $\mathbf{\Lambda}$

$$\lambda_1 = \frac{1}{54}\left(F_1 + F_2 + F_3 + \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1}\right).$$

La seconde valeur propre

$$\lambda_2 = \frac{1}{54}\left(F_1 + F_2 + F_3 - \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1}\right),$$

et on a $\mathbf{Tr}(\mathbf{\Lambda}) = \lambda_1 + \lambda_2 = \mathbb{E}[D_{\mathrm{ST}}]/2$.

4. 3. Résultats liés au F-modèle

Exemple Pour illustrer l'approximaton de $\mathbb{E}[F_{\mathrm{ST}}]$ par les valeurs propres de l'ACP, nous présentons ici un exemple de simulation, dans lequel la matrice de génotype est générée par un F-modèle à trois populations. Dans ce premier exemple, la fréquence ancestrale moyenne est égale à 20%, et les paramètres de dérive sont $F_1 = 5\%$, $F_2 = 10\%$ et $F_3 = 30\%$. Les populations 1 et 2 sont plus proches génétiquement l'une de l'autre que de la population 3, la population la plus divergente. On a génotypé 300 échantillons ($n_k = 100$, pour $k = 1, 2, 3$) à 10,000 locus. L'ACP a été appliquée sur $L = 9740$ SNP après avoir retiré les sites monomorphiques. La valeur moyenne de F_{ST} le long du génome est égale à 9.52%, elle est correctement approchée par la somme des deux plus grandes valeurs propres de l'ACP (9.54%). Les premiers axes de l'ACP expliquent respectivement 6.78%, 2.76%, et 0.47% de la variation totale (Figure 5.1). Les valeurs singulières au carré non-nulles de $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}/\sqrt{n}$, 6.77% et 2.75%, sont proches de celles obtenues pour les deux premiers axes de l'ACP de la matrice de génotype. Leur somme est égale à $\mathbb{E}[F_{\mathrm{ST}}]$ comme spécifiée dans le théorème 1. La plus petite valeur singulière au carré de $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}/\sqrt{n}$ est clairement séparée de la plus grande valeur singulière au carré de $\mathbf{Z}_{\mathrm{S}}^{\mathrm{sc}}/\sqrt{n}$ (0.47%), laquelle est proche de la troisième valeur propre de l'ACP et de la prédiction de la TMA (0.31%). Pour montrer l'effet d'une numérotation incorrecte des population pour les échantillons, nous avons utilisé la même matrice de génotype, et nous avons répliqué l'analyse en groupant les individus des populations 2 et 3 au sein d'une même population (Figure 2, $n_1 = 200$ and $n_2 = 100$, $K = 2$). La valeur moyenne de F_{ST} est alors égale 3.46%, et ne correspond pas à la plus grande valeur propre de l'ACP (6.78%). La valeur singulière au carré de $\mathbf{Z}_{\mathrm{S}}^{\mathrm{sc}}/\sqrt{n}$ (6.06%) ne vérifie pas la condition de séparation, et diffère de sa prédiction TMA (0.32%). La projection des échantillons sur les axes principaux pour $\mathbf{Z}_{\mathrm{S}}^{\mathrm{sc}}$ montre la structure résiduelle de la population au sein de la population fictive composée des populations 2 et 3. Ces résultats soulignent l'utilité de visualiser la matrice résiduelle dans l'optique d'évaluer le nombre de populations de la matrice de génotype (Figure 5.2).

Modèle à une population Nous réalisons tout d'abord des simulations sans structure de population, afin de vérifier si les prédictions TMA sont valides pour le F-modèle. Pour les F-modèle à une population, les résultats montrent que la plus grande valeur propre de l'ACP est

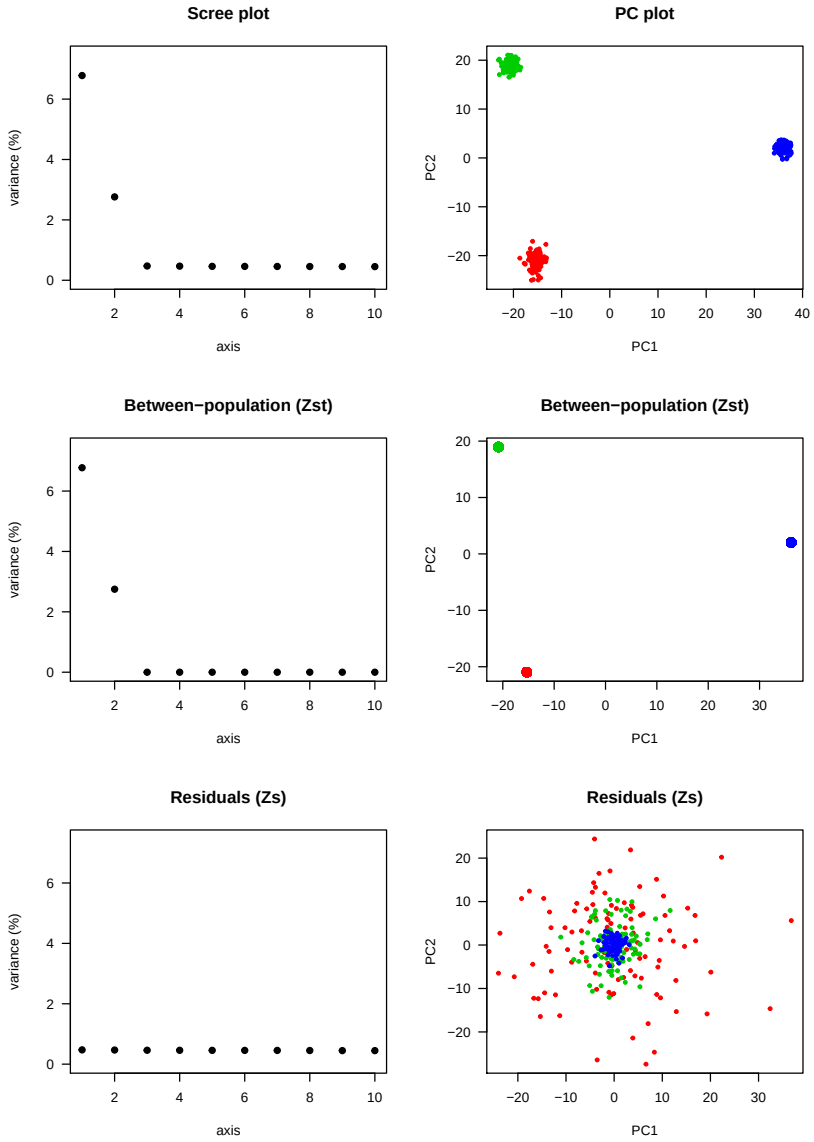

Figure 5.1 – **Analyse spectrale d'un modèle à 3 populations**
Part de variance expliquée par l'ACP (gauche) et projection des échantillons sur les composantes principales (droite) pour la matrice standardisée, $\mathbf{Z}^{\text{sc}}$, pour la matrice inter-population, $\mathbf{Z}^{\text{sc}}_{\text{ST}}$, et pour la matrice résiduelle, $\mathbf{Z}^{\text{sc}}_{\text{S}}$ de données simulées. Simulations réalisée avec $n = 300$ individus et un F-modèle ($F_1 = 5\%, F_2 = 10\%, F_3 = 30\%$) avec une fréquence ancestrale tirée d'une distribution beta(1,4).

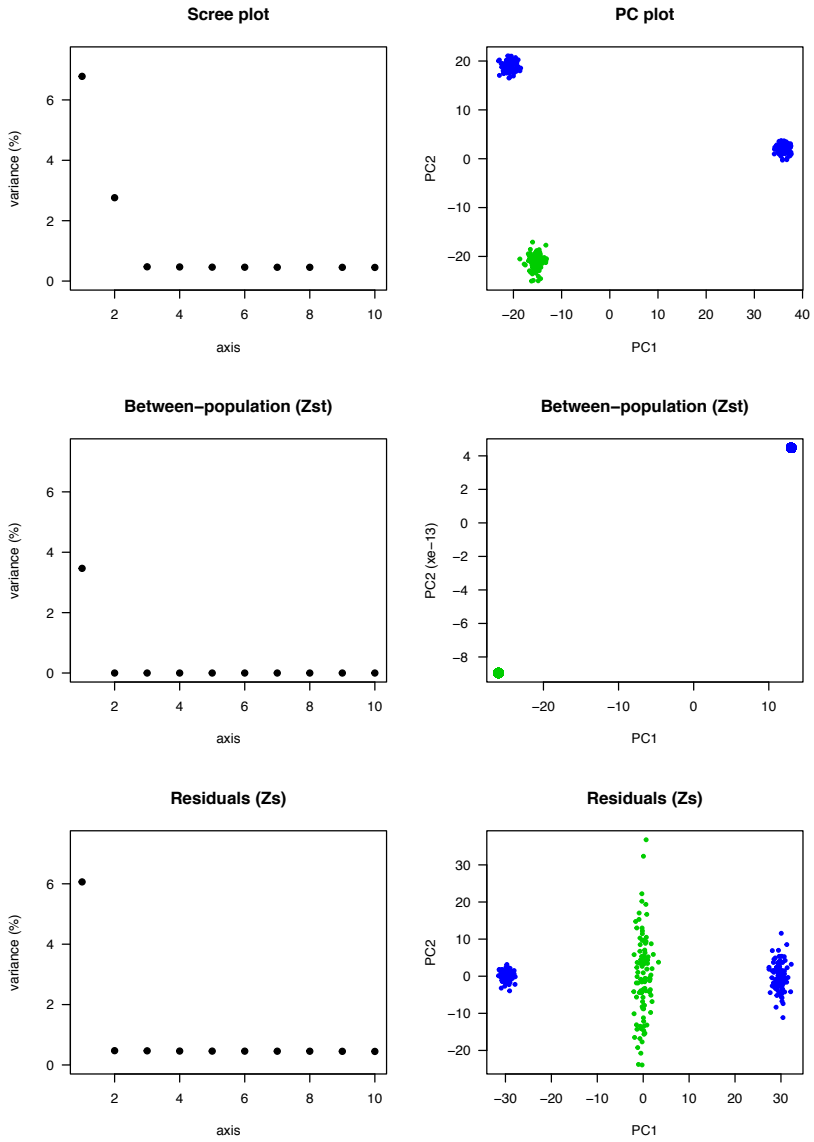

Figure 5.2 – **Analyse spectrale avec des populations incorrectement numérotées**
Pour la même matrice de genotype que dans la Figure 5.1, les échantillons des populations 2 et 3 (bleu) sont groupés face à la population 1 (vert). Le scree plot de l'ACP et le graphique des PC pour la matrice standardisée, $\mathbf{Z}^{\text{sc}}$, pour la matrice inter-population, $\mathbf{Z}^{\text{sc}}_{\text{ST}}$, et pour la matrice résiduelle, $\mathbf{Z}^{\text{sc}}_{\text{S}}$ des données simulées. F_{ST} est inférieur à la plus grande valeur singulière au carré de la matrice résiduelle, et est différente de la plus grande valeur propre de l'ACP.

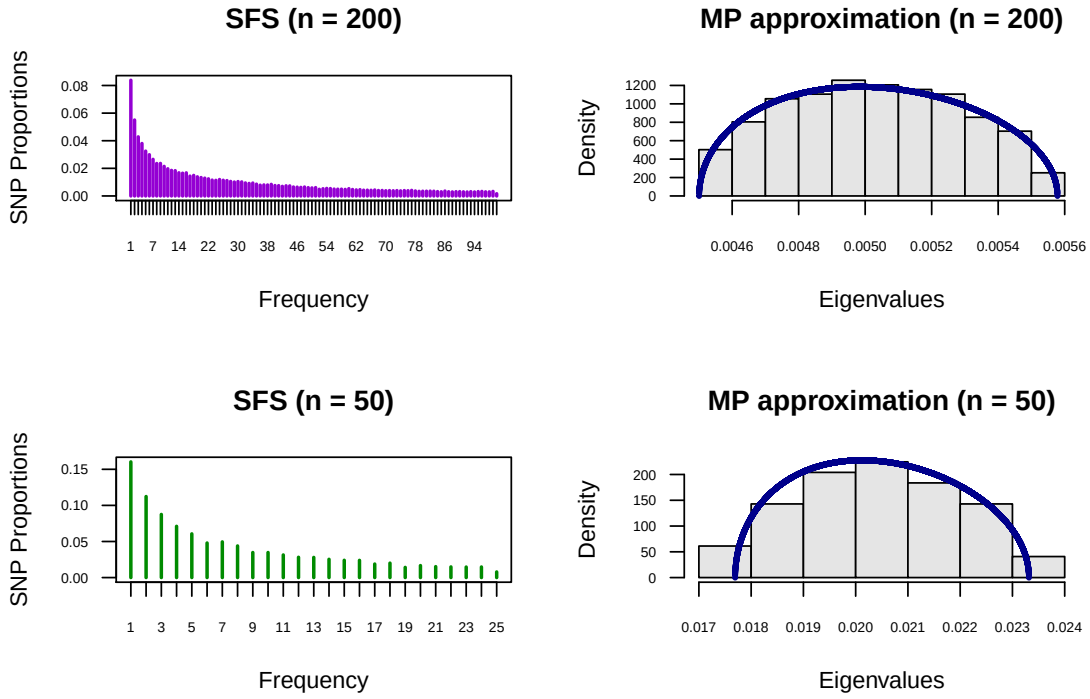

Figure 5.3 – **Approximation par la théorie des matrices aléatoires de la part de variance expliquée dans un F-modèle à 1 population**

2 simulations ont été réalisées avec p_{anc} tiré d'une distribution beta de paramètres $a = 1$ and $b = 9$, et $F = 15\%$. **Haut** : $n = 200$ individus et $L = 69,248$ SNPs. **Bas** : $n = 50$ individus et $L = 10,331$ SNPs. **SFS** : Site Frequency Spectrum, **MP approximation** : Approximation Marchenko-Pastur de la distribution des valeurs propres de l'ACP standardisée (courbe bleue). Les histogrammes des valeurs propres de l'ACP standardisée représentant la part de variance expliquée par les $(n-1)$ composantes principales sont affichés en gris.

correctement prédite par la distribution de Marchenko-Pastur (Figure 5.3). Nous avons ensuite cherché à savoir si la condition de séparation (5.5) pouvait être vérifiée dans des situations où il n'y avait pas de structure dans les données, et où deux populations avaient incorrectement été identifiées dans une phase préliminaire d'analyse de structure. Nous avons simulé 200 scénarios de modèle à une population ($n = 100$ et $L \approx 10,000$), et, pour chacun des jeux de données, nous avons séparé les échantillons en deux groupes, selon le signe de leur première composante principale. Cette procédure maximise la probabilité de détecter des groupes fictifs, et donne lieu à une valeur moyenne de $F_{\text{ST}} \approx 1.1\%$. Pour ces groupes fictifs, nous avons calculé les valeurs singulières non nulles de la matrice inter-population, $\mathbf{Z}^{\text{sc}}_{\text{ST}}/\sqrt{n-1}$, et la plus grande valeur singulière de la matrice intra-population, $\mathbf{Z}^{\text{sc}}_{\text{S}}/\sqrt{n-1}$. Pour les simulations, la condition de séparation n'a jamais été vérifiée, rejetant ainsi l'existence de structure de population dans tous les cas. Pour des échantillons de plus petite taille ($n = 10$ et $L \approx 1,000$), la condition de séparation était incorrectement vérifiée dans 21% des simulations, montrant qu'il est plus difficile de distinguer des groupes fictifs avec des petites tailles d'échantillons (Figure 5.4).

Modèle à deux populations Nous avons également réalisé des simulations de F-modèle avec deux populations. Pour ces simulations, la condition de séparation est vérifiée sur l'ensemble des jeux de données. On trouve une correspondance presque parfaite entre la plus grande valeur propre de l'ACP centrée, $\sigma^2_1(\mathbf{Z})$, et la valeur moyenne de $D_{\text{ST}}/2$ le long du génome (Figure

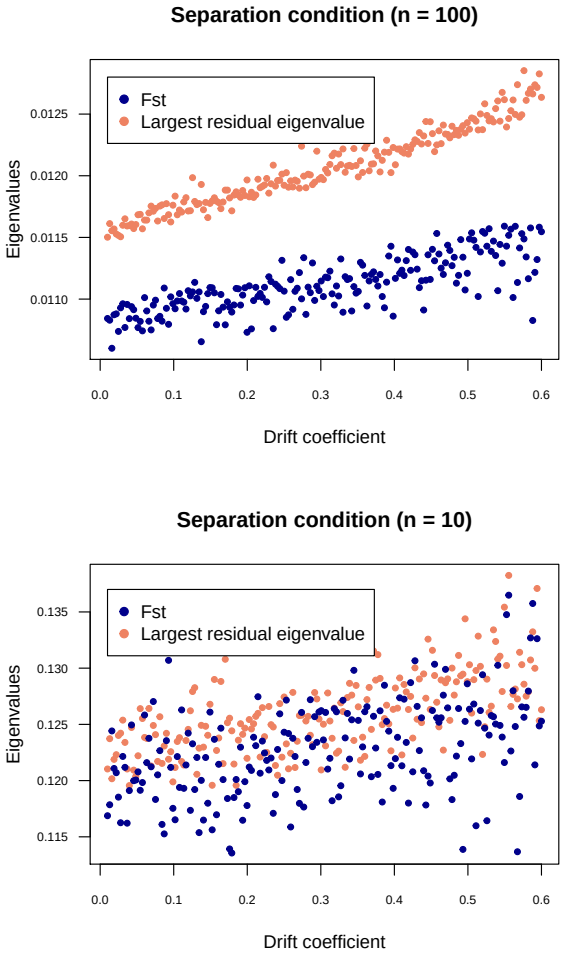

Figure 5.4 – Condition de séparation dans les échantillons de populations fictives construites à partir d'un F-modèle à 1 population

Pour chaque valeur de coefficient de dérive, un couple de points bleu et orange représente un jeu de données simulées. La F_{ST} (points bleus) correspond à la valeur singulière non nulle au carré de la matrice inter-population, $\mathbf{Z}_{\mathrm{ST}}$, et la "Largest residual eigenvalue" correspond à la plus grande valeur singulière au carré de $\mathbf{Z}_{\mathrm{S}}$. La structure de population est détectée lorsque le point bleu est au dessus du point orange. **Haut** : 200 simulations avec $n = 100$ individus et L environ 10,000 SNPs. **Bas** : 200 simulations avec $n = 10$ individus et L environ 1,000 SNPs. Les simulations ont été réalisés avec p_{anc} tiré d'une distribution beta de paramètres $a = 1$ et $b = 9$.

5.5A), ainsi qu'avec sa valeur théorique. La correspondance est également presque parfaite entre la plus grande valeur propre de l'ACP standardisée, $\rho_1^2(\mathbf{Z}^{sc})$, et la valeur moyenne de la F_{ST} le long du génome (Figure 5.5C). La seconde plus grande valeur propre est correctement prédite par la TMA pour la version non-standardisée et standardisée de l'ACP (Figure 5.5B-D). Nous avons réalisé des simulations supplémentaires avec $F_1 = F_2 = 7\%$, en s'intéressant à la distribution des valeurs singulières au carré de la matrice résiduelle. Dans un échantillon de $n = 200$ et $L \approx 85,500$ SNPs, la première composante principale explique 3.11% de la variance génétique total, correspondant à la valeur moyenne de F_{ST} (3.11%, Figure 5.7A). La condition de séparation est vérifiée, et la seconde valeur propre de l'ACP (0.536%) est très proche de la prédiction TMA, donnée par $(1 - \rho_1^2(\mathbf{Z}^{sc})) \times (1/\sqrt{L} + 1/\sqrt{n-2})^2 = 0.537\%$ (Figure 5.7A). La distribution des valeurs propres résiduelles, correspondant à la variation intra-population, est correctement représentée par la fonction de densité de probabilité de Marchenko-Pastur (Figure 5.7B). Avec un échantillon plus petit de $n = 20$ individus et $L \approx 12,500$ SNPs, l'axe principal explique 5.24% de la variance génétique total, correspondant encore à la valeur moyenne de F_{ST} le long du génome (5.23%, Figure 5.7C). La densité de Marchenko-Pastur est à nouveau une approximation précise des valeurs propres de l'ACP de la matrice résiduelle (Figure 5.7D).

Modèle à trois populations Nous avons effectué des simulations de F-modèle à 3 populations pour vérifier que les données sont en accord avec les prédictions théoriques des valeurs propres principales, λ_1 et λ_2, pour la D_{ST} et F_{ST}. Avec des coefficients de dérive aléatoires et $n = 100$, $L = 20000$, la condition de séparation était vérifiée dans tous les cas. On observe une correspondance presque parfaite entre $\lambda_1 + \lambda_2$ et la valeur moyenne de $D_{ST}/2$ (ACP non standardisée) ou F_{ST} (ACP standardisée) (Figure 5.8AC). La valeur propre principale de l'ACP non standardisée présente un biais petit mais visible par rapport à la valeur théorique prédite de λ_1 (Figure 5.8B). La troisième valeur propre de l'ACP standardisée est proche de l'approximation par la TMA (Figure 5.8D).
Afin d'étudier des cas pour lesquels la condition de séparation n'était pas vérifiée, nous avons considéré des plus petites valeurs de n et L, et des valeurs plus faibles de coefficients de dérives ($F_k \leq 10\%$). Pour des petites valeurs de n et L, une proportion significative des jeux de données ne vérifient pas la condition de séparation (Figure 5.9). Ces résultats fournissent une preuve supplémentaire des biais dans l'analyse de la structure de population dans le cas de petits jeux de données.

4. 4. Données réelles

Dans cette sous-section, nous allons confronter les prédictions de nos résultats théoriques à un jeu de données réelles.

Présentation des données Nous utilisons ici des jeux de données composés de paires et de triplets de populations humaines issues du 1000 Genomes Project. Dans ces comparaisons, le nombre de SNPs est de $L \approx 1.3M$ après un filtrage des fréquences d'allèles inférieures à 5%. Nous avons utilisé des échantillons issus des populations de Chinois Han de Pékin (CHB, $n = 100$), Yoruba (YRI, $n = 158$), résidents d'Utah d'ascendance européenne (CEU, $n = 104$), Ibérique (IBS, $n = 147$). Nous avons également utilisé des populations d'ascendance métissée, américains d'ascendance africaine dans le Sud-Est des Etats-Unis d'Amérique (ASW, $n = 97$), Colombiens de Medellin (CLM, $n = 102$), Puerto Ricains (PUR, $n = 94$), individus de Los Angeles d'ascendance mexicaine (MXL, $n = 100$), et africains caribéens de barbades (ACB, $n = 98$). Certaines paires et triplets contiennent des individus d'ascendance métissée et d'autres non.

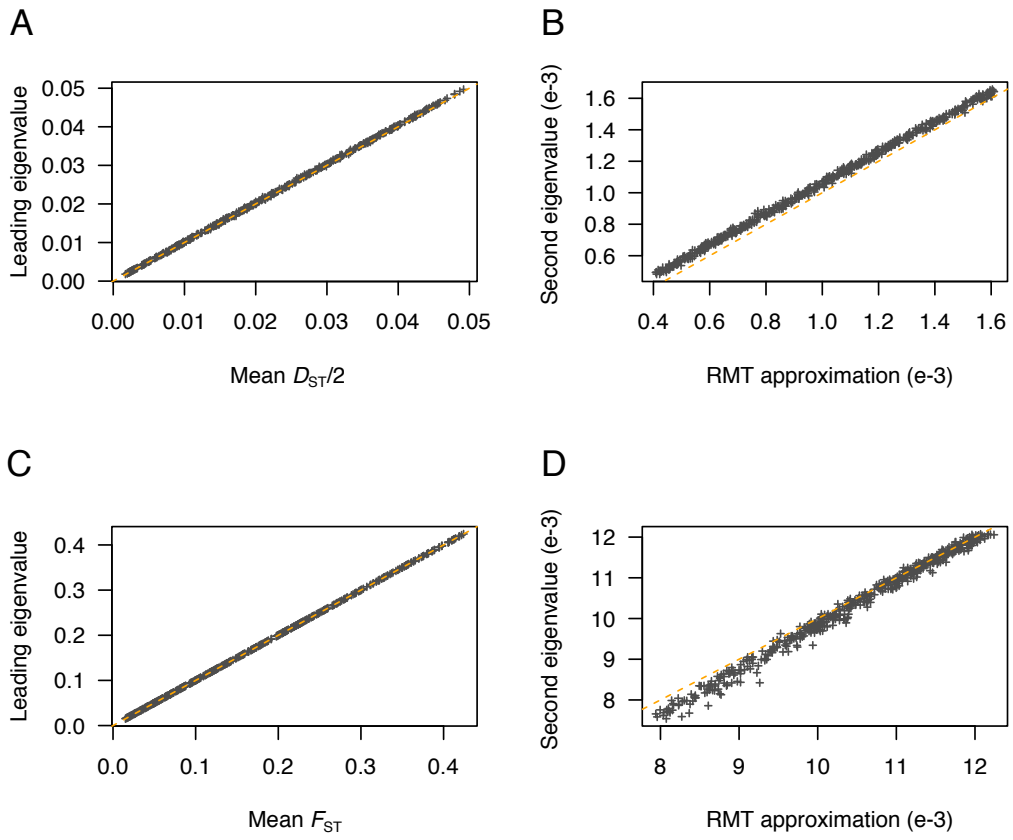

Figure 5.5 – Comparaison des estimations de D_{ST} et F_{ST} avec la plus grande valeur propre de l'ACP dans un modèle à 2 populations

(**A**) Plus grande valeur propre de l'ACP centrée en fonction de la moyenne de $D_{\mathrm{ST}}/2$ le long du génome.
(**B**) Seconde plus grande valeur propre de l'ACP centrée en fonction de son approximation par la TMA.
(**C**) Plus grande valeur propre de l'ACP standardisée en fonction de la valeur moyenne de F_{ST} le long du génome.
(**D**) Seconde plus grande valeur propre de l'ACP standardisée en fonction de son approximation par la TMA (approximation de la plus grande valeur propre de la matrice résiduelle).
La ligne en pointillée correspond à la droite $y = x$.Les simulations ont été réalisées pour $n = 100$ individus (coefficient de consanguinité entre 1% et 75%, proportion d'individus dans la population 1 entre 10% et 50%, fréquence ancestrale tirée d'une distribution beta(1,4)

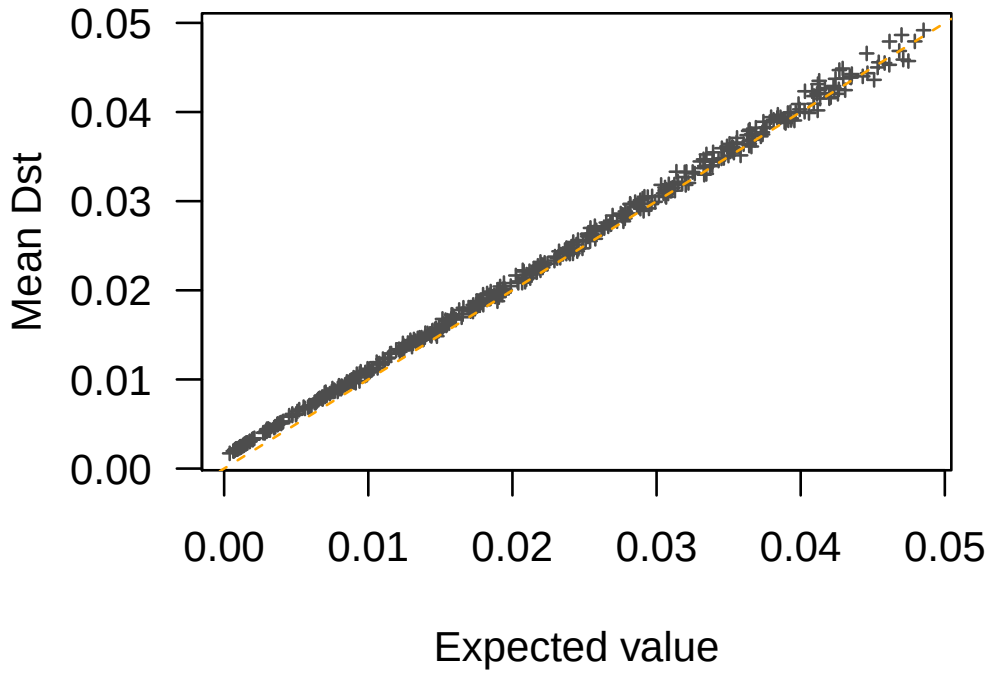

Figure 5.6 – **Précision de l'estimation de D_{ST} comparée à sa valeur théorique dans un F-modèle avec 2 populations**

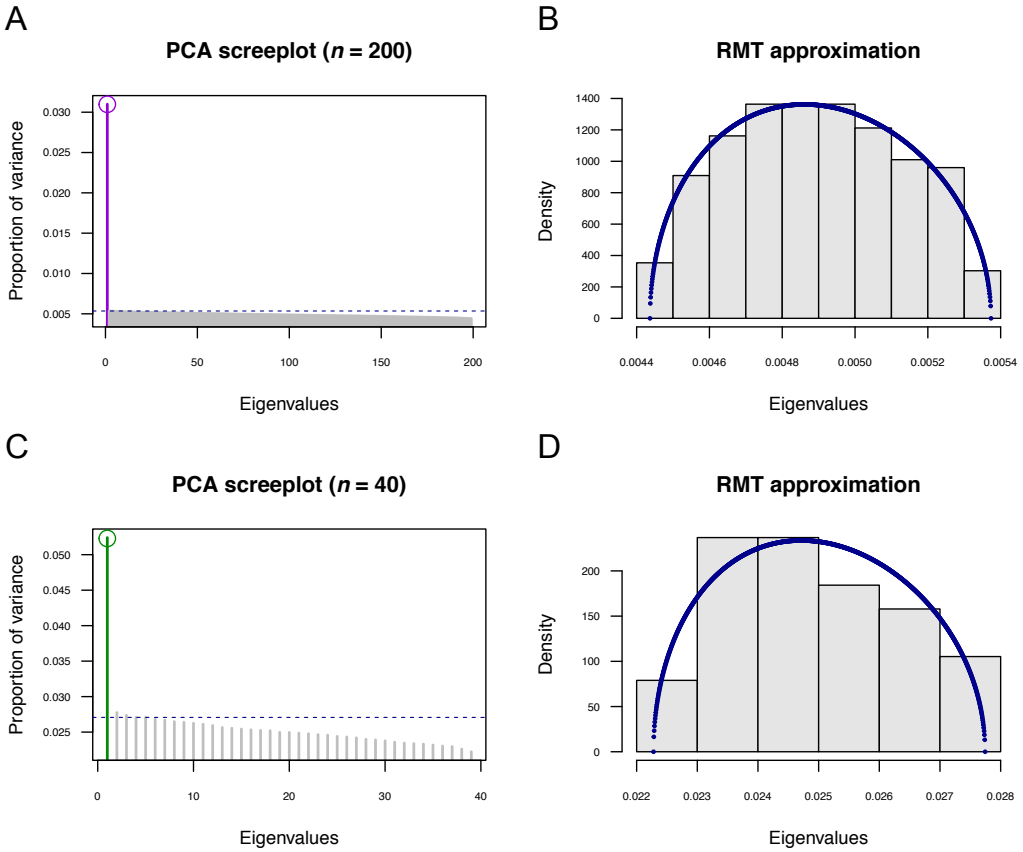

Figure 5.7 – **Screeplot et approximation par la TMA dans un modèle à 2 populations**
(**A**) Part de variance expliquée par les composantes principales, le cercle correspond la valeur moyenne de F_{ST} le long du génome. $n = 200$ individus et $L = 85,540$ SNPs.
(**B**) Histogramme des valeurs singulières au carré de la matrice résiduelle, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-2}$, pour les données simulées en (A).
(**C**) Part de variance pour $n = 40$ individus et $L = 12,650$ SNPs.
(**D**) Histogramme des valeurs singulières au carré de la matrice résiduelle, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-2}$, pour les données simulées en (C).
La ligne en pointillé représente l'approximation par la TMA de la plus grande valeur singulière au carré de la matrice résiduelle. La courbe bleue représente la densité de probabilité Marchenko-Pastur. p_{anc} tirée d'une distribution beta(1,9) et $F_1 = F_2 = 7\%$.

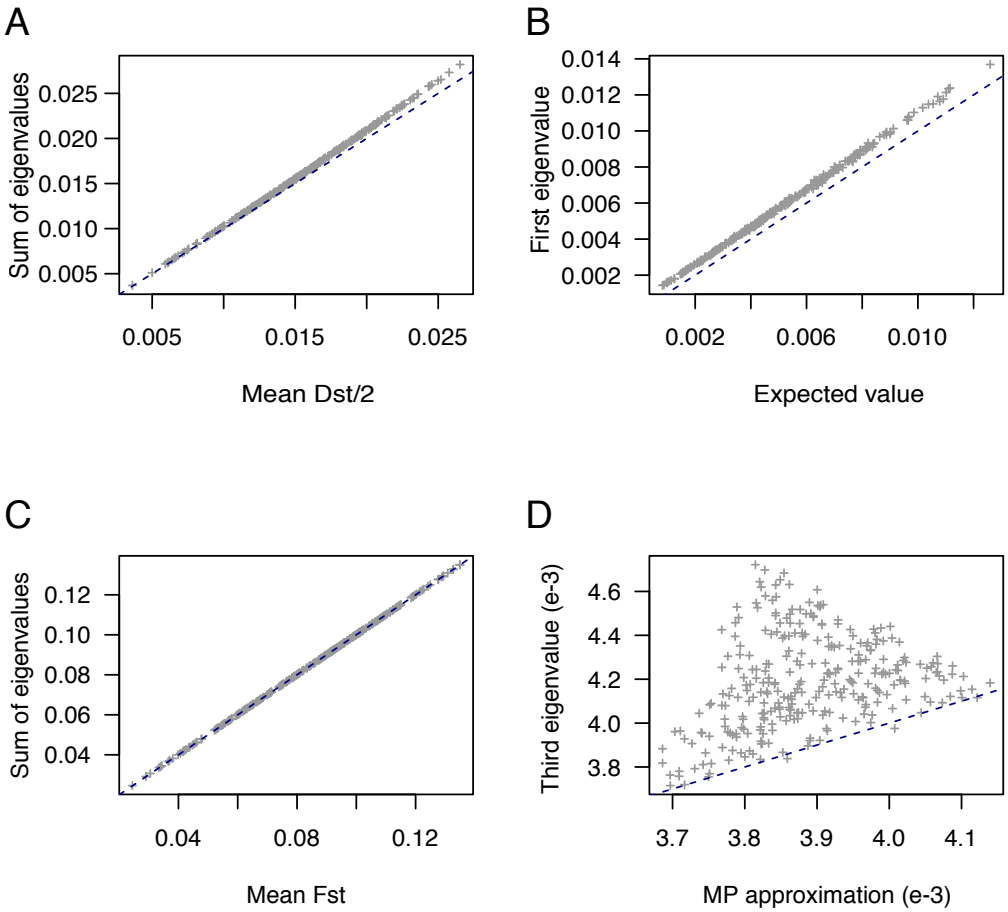

Figure 5.8 – **Valeurs propres principales dans un F-modèle à 3 populations**
(**A**) Somme des deux premières valeurs propres de l'ACP centrée en fonction de la valeur moyenne de $D_{\mathrm{ST}}/2$ le long du génome
(**B**) Première valeur propre de l'ACP centrée en fonction de sa valeur théorique $\lambda_1 = (F_1 + F_2 + F_3 + \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1})/54$.
(**C**) Somme des deux premières valeur propres de l'ACP standardisée en fonction de la valeur moyenne de F_{ST} le long du génome.
(**D**) Troisième valeur propre de l'ACP standardisée en fonction de son approximation par la TMA.
MP approximation : Approximation Marchenko-Pastur de la plus grande valeur singulière au carrée de la matrice résiduelle, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-3}$, égale à $(1-\rho_1^2-\rho_2^2)\times(1/\sqrt{L}+1/\sqrt{n-3})^2$. La ligne en pointillé correspond à la droite $y = x$. Les simulations F-modèle ont été réalisé pour $n = 100$ individus avec des coefficients de dérive F_1, F_2, F_3 entre 1% et 25%, un nombre d'individus égal dans chaque population, et des fréquences ancestrales tirées d'une distribution uniforme ($L = 20000$ locus).

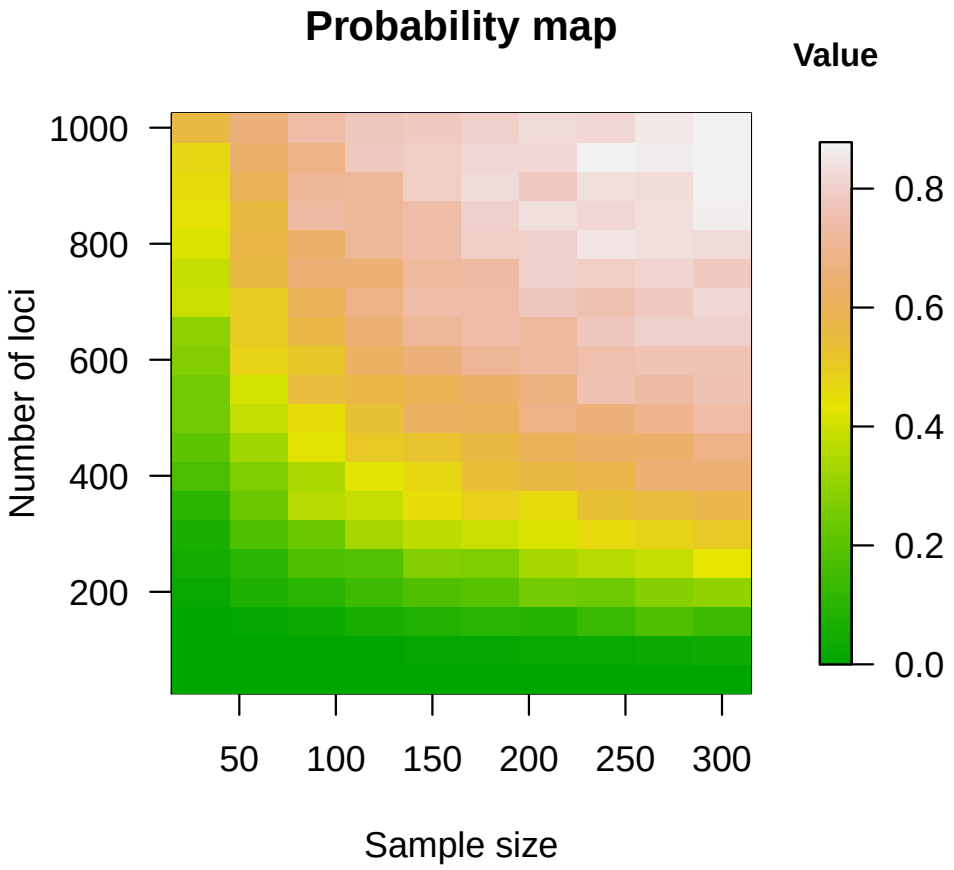

Figure 5.9 – **Condition de séparation dans un F-modèle à 3 populations**
Probabilité que la condition de séparation soit vérifiée pour des tailles d'échantillons allant de $n = 30$ à $n = 300$ individus, and un nombre de locus allant de $L = 100$ à $L = 1000$. Les simulations sont réalisées avec un coefficient de dérive aléatoire inférieur à 10%, and des fréquences ancestrales tirés d'une distribution uniforme. Le nombre de réplication pour chaque combinaison de n et L est de 500.

Résultats Afin de prouver la validité de notre théorie concernant la relation entre les valeurs propres de l'ACP et les valeurs de F_{ST}, nous avons calculé la valeur de F_{ST}, son approximation par l'ACP et la plus grande valeur propre au carré de la matrice résiduelle pour des paires et triplets de populations. (Tableau 5.1 et Tableau 5.2). Dans les analyses par paires excluant les échantillons métissés, la condition de séparation est toujours vérifiée à l'exception de la paire CEU-IBS, formée de deux populations européennes proches. La plus grande valeur propre de l'ACP standardisée est correctement approchée par $\mathbb{E}[F_{\mathrm{ST}}]$, et la plus grande valeur singulière au carré de la matrice résiduelle est correctement prédite par la TMA. Pour les analyses de triplets sans échantillon métissé, la condition de séparation est également vérifiée et la somme des deux premières valeurs propres de l'ACP standardisée est correctement approchée par $\mathbb{E}[F_{\mathrm{ST}}]$. La TMA prédit également correctement la plus grande valeur singulière au carré de la matrice résiduelle. Dans les analyses de paires et triplets incluant des échantillons métissés, la condition de séparation est vérifiée à l'exception de la paire ACB-ASW (Table ?). L'approximation de $\mathbb{E}[F_{\mathrm{ST}}]$ par la plus grande valeur propre de l'ACP standardisée est moins précise que pour les analyses sans échantillon métissé. Pour la paire CEU-ASW par exemple, $\mathbb{E}[F_{\mathrm{ST}}]$ ($= 4.55\%$) est inférieure à la plus grande valeur propre de l'ACP ($= 4.87\%$). Une explication pourrait être que F_{ST} nous informe de la proportion de métissage entre les populations métissées et leur population d'origine. Avec des échantillons métissés, on observe également des différences entre la plus grande valeur singulière au carré de la matrice résiduelle et sa prédiction par la TMA. Les résultats suggèrent que les données ne concordent pas avec un modèle à K populations discrètes, et une définition modifiée de la F_{ST} pourrait être plus appropriée pour décrire la structure de population en présence d'individus métissés (Martins et al., 2016 ; Ochoa et JD., 2021).

5. Utilisation des résultats pour le calcul de F_{ST} sur des matrices de génotype modifiées

Une application intéressante des résultats théoriques est que la définition alternative de F_{ST} et D_{ST} obtenue à l'aide du théorème 1 nous permet d'étendre ces notions à des matrices de génotypes modifiées. Par exemple, les matrice de génotypes modifiées, ou ajustées, apparaissent lorsqu'on souhaite corriger les biais liés aux artefacts techniques, comme la couverture génomique ou l'effet de lot (batch effect) pour les données de génomique de populations (DJ., 2006 ; Ross et al., 2013). Dans cette section, nous expliquerons le principe de la correction puis nous proposerons deux applications de ces résultats, à des données d'ADN ancien puis en génomique écologique.

5. 1. Principe

En général, la valeur de F_{ST} peut être ajustée pour n'importe quel effet spécifique en considérant les résidus des modèles de régressions à facteurs latents (Leek et al., 2012 ; J. Wang et al., 2017 ; Caye, Jumentier et al., 2019). Plus précisément, pour $\mathbf{Z}$ (ou $\mathbf{Z}^{\mathrm{sc}}$) et pour un ensemble de covariables $\mathbf{Y}$, les modèles de régression à facteurs latents estiment une matrice, $\mathbf{W}$, en ajustant un modèle de régression de la forme

$$\mathbf{Z} = \mathbf{Y}\mathbf{B}^T + \mathbf{W} + \epsilon\,.$$

Dans ce modèle, la matrice $\mathbf{B}$ contient les tailles d'effets pour chaque variable dans $\mathbf{Y}$, et ϵ est une matrice qui représente l'erreur centrée. La matrice latente $\mathbf{W}$ a un rang spécifique, k, inférieur à n moins le nombre de covariables. Le rang k correspond au nombre de facteurs latents

	Lead. eigen. of PCA*	F_{ST} **across locus**	**Lead. eigen. res. matrix****	**RMT approximation*****
CHB-CEU	5.65%	5.65%	0.42%	0.48%
CHB-YRI	8.35%	8.35%	0.36%	0.37%
CEU-YRI	7.21%	7.21%	0.35%	0.37%
CEU-IBS	0.41%	0.38%	0.37%	0.41%
CEU-YRI-CHB	9.99%	9.98%	0.24%	0.26%
CEU-ASW	4.87%	4.55%	0.75%	0.52%
CEU-YRI-ASW	6.12%	5.82%	0.44%	0.29%

Table 5.1 – Estimation de la F_{ST} pour des populations du projet 1,000 Genomes
* Sum of the leading eigenvalues of the PCA
** Sum of the leading eigenvalues of the within-population (residual) matrix
*** RMT approximation for the leading eigenvalue of the within-population matrix

IBS : Iberian ($n = 147$), **CHB** : Han Chinese in Beijing ($n = 100$), **YRI** : Yoruba ($n = 158$), **CEU** : Utah residents with European ancestry ($n = 104$). **ASW** : Americans of African Ancestry in SW USA ($n = 97$).

	Lead. eigen. of PCA*	F_{ST} **across locus**	**Lead. eigen. res. matrix****	**RMT approximation*****
YRI-IBS	7.27%	7.27%	0.31%	0.32%
YRI-IBS-CHB	9.75%	9.74%	0.25%	0.25%
ACB-ASW	1.26%	0.60%	1.05%	0.56%
PUR-ASW	3.53%	3.01%	0.95%	0.56%
CEU-CLM	1.40%	1.16%	0.75%	0.53%
CEU-MXL	2.45%	1.86%	1.06%	0.53%
CEU-CLM-CHB	4.77%	4.60%	0.51%	0.35%
CLM-IBS-ASW	4.65%	4.19%	0.52%	0.31%
ACB-CHB-CEU	9.00%	8.87%	0.36%	0.34%

Table 5.2 – Estimation de la F_{ST} pour des populations du projet 1,000 Genomes
* Leading eigenvalue of the PCA
** Leading eigenvalue of the within-population matrix
*** RMT approximation for the leading eigenvalue of the within-population matrix
IBS : Iberian ($n = 147$), **CHB** : Han Chinese in Beijing ($n = 100$), **YRI** : Yoruba ($n = 158$), **CEU** : Utah residents with European ancestry ($n = 104$). **CLM** : Colombians from Medellin Colombia ($n = 102$), **ASW** : Americans of African Ancestry in SW USA ($n = 97$), **PUR** : Puerto Ricans from Puerto Rico ($n = 94$), **MXL** : Individuals of Mexican Ancestry from Los Angeles USA ($n = 100$), **ACB** : African Caribbeans in Barbados ($n = 98$).

inclus dans le modèle. La matrice $\mathbf{Z}^{\text{adj}} = \mathbf{W} + \epsilon$, assimilable à une matrice de génotypes modifiés, donne lieu à une définition ajustée du coefficient de consanguinité, $F_{\text{ST}}^{\text{adj}}$. Le coefficient de consanguinité ajusté peut être calculé comme la norme au carré de la matrice inter-population, $\mathbf{Z}_{\text{ST}}^{\text{adj+sc}}$, après standardisation. Il est également possible de calculer $F_{\text{ST}}^{\text{adj}}$ à partir de la valeur moyenne du coefficient de détermination, R^2, obtenu à partir de la régression de chacune des valeurs du génotype standardisée en fonction de l'identifiant de la population. Les définitions sont équivalentes et on a :

$$\mathbb{E}[F_{\text{ST}}^{\text{adj}}] = \sum_{k=1}^{K-1} \rho^2(\mathbf{Z}_{\text{ST}}^{\text{adj+sc}})/L = \mathbb{E}[R^2]\,.$$

5. 2. Application aux données d'ADN ancien

Dans cette sous-section, nous obtenons nos valeurs de F_{ST} ajustées pour des données d'ADN ancien, pour lesquelles les problématiques principales sont les biais dus à la couverture génomique et les distortions temporelles créées par la dérive génétique.

Présentation des données Nous avons analysé 143,081 SNPs pseudo-haploïdes d'échantillons anciens des premiers agriculteurs d'Anatolie (EFA, $n = 23$), des pasteurs des steppes de la culture Yamnaya (Steppe, $n = 15$), des chasseurs-cueilleurs occidentaux de Serbie (WHG, $n = 31$), et des agriculteurs de culture campaniforme d'Angleterre et d'Allemagne (BKK, $n = 38$). Les données ont été obtenues à partir d'un jeu de données publiques disponible dans le laboratoire de David Reich (`reich.hms.harvard.edu`) (Allentoft et al., 2015 ; Mathieson, Lazaridis et al., 2015 ; Mathieson, Roodenberg et al., 2018). Les échantillons anciens ont une couverture minimale de 0.25x, une couverture médiane de 2.69x (moyenne de 2.98x) et une couverture maximale de 13.54x. Les génotypes ont été ajustés pour la couverture à l'aide d'un modèle de régression à facteurs latents avec un nombre de facteurs égal au nombre d'échantillons moins deux. La matrice a ensuite été ajustée pour la distortion causée par les différences d'âge des échantillons, donnant lieu à une matrice de génotypes modifiés encodés par des valeurs continues sans aucune interprétation directe possible en terme de fréquence d'allèle (François et Jay, 2020).

Résultats Après ajustement pour la couverture et la correction pour les distortions liées aux différences d'âge des échantillons, les estimations ajustées de F_{ST} sont égales à 4.7% pour *EFA – Steppe*, 5.8% pour *EFA – WHG*, 5.1% pour *Steppe – WHG* (Tableau 5.3). La condition de séparation est vérifiée dans toutes les comparaisons. Les scores de l'ACP des individus sont impactés par la couverture et la distortion temporelle (Figure 5.10), mais ces effets non désirés ne génèrent pas de biais substantiel pour les valeurs propres de l'ACP, donnant lieu à des estimations de la F_{ST} qui sont similaires avec ou sans ajustement. Pour des modèles à trois populations incluant des échantillons EFA, WHG et Steppe, l'estimation ajustée de F_{ST} est égale à 7.0%, légèrement inférieure à l'estimation non corrigée (7.2 %, Tableau 5.3). La plus petite valeur singulière au carré de la matrice inter-population est plus grande (2.6%) que la plus grande valeur singulière au carré de la matrice résiduelle (1.8%). Cette condition n'est plus vérifiée quand les Bell Beaker d'Angleterre et d'Allemagne sont inclus dans le jeu de données. Avec les échantillons Bell Beaker, la plus petite valeur singulière au carré de la matrice inter-population est plus petite (1.0%) que la plus grande valeur singulière au carré de la matrice résiduelle (1.0%, Tableau 5.3). Une explication de ce résultat est que l'ascendance partagée des individus Bell Beaker (Mathieson, Roodenberg et al., 2018) a rendu les résultats de l'ACP incompatibles avec un modèle à quatre populations.

	F_{ST} **without correction**	F_{ST} **with correction**	**Lead. eigen. res. matrix***	**RMT threshold****
EFA-Steppe	4.8%	4.7%	3.1%	2.8%
EFA-WHG	5.9%	5.8%	3.3%	2.0%
Steppe-WHG	5.2%	5.1 %	3.8%	2.3%
EFA-Steppe-WHG	7.2%	7.0%	1.8 % (2.6%)	1.5 %
EFA-Steppe-WHG-BBK	5.9 %	5.8%	1.8 % (1.0%)	1.0 %

Table 5.3 – Estimations de F_{ST} pour des échantillons anciens d'eurasiens avec correction pour la couverture génomique

EFA : Early Farmers from Anatolia, **WHG** : Western Hunter-Gatherers, **Steppe** : Yamnaya pastoralists, **BBK** : Bell Beakers from England and Germany.

* Plus grande valeur singulière au carré de la matrice intra-population (plus petite valeur singulière au carré de la matrice inter-population)

** Seuil de TMA pour la mise en évidence de structure de population pour des paires : $(1/\sqrt{L} + 1/\sqrt{n-1})^2$, Approximation TMA pour triplets et quadruplets : $(1 - \mathbb{E}[F_{\mathrm{ST}}])(1/\sqrt{L} + 1/\sqrt{n-K})^2$. L : nombre de locus, n : taille d'échantillon.

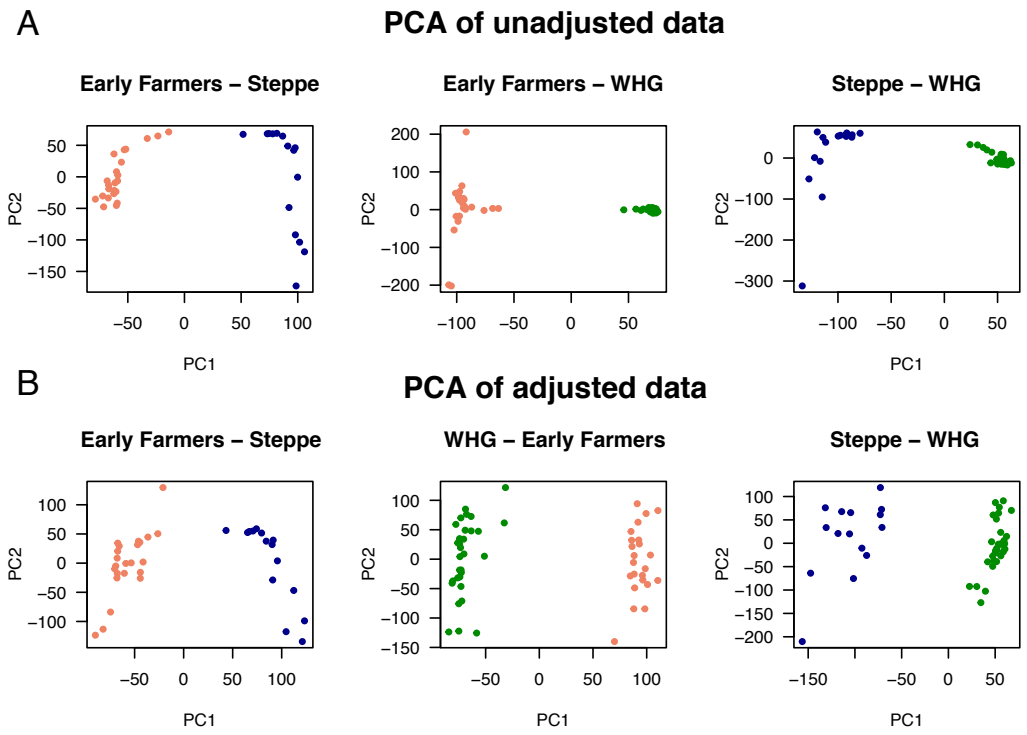

Figure 5.10 – **Correction pour la couverture dans les graphiques PC pour des paires d'échantillons de populations anciennes**

(**A**) ACP de génotypes non ajustés
(**B**) ACP de données de génotypes non binaires ajustées pour la couverture
Agriculteurs d'Anatolie (couleur saumon, $n_1 = 23$), Pasteurs des steppes (couleur bleue, $n_2 = 15$), Chasseurs cueilleurs occidentaux (couleur verte, $n_3 = 31$).

5. 3. Application aux données d'*Arabidopsis thaliana*

Nous avons ensuite étudié le rôle des facteurs bioclimatiques dans la création de structure génétique. Ici, l'objectif est d'évaluer la part de différenciation expliquée par la température et les précipitations chez *Arabidopsis thaliana*.

Présentation des données Nous avons étudié 241 échantillons de plantes suédoises de l'espèce *Arabidopsis thaliana* issues de la base de données 1,001 Genomes. La matrice de génotypes a été obtenue en considèrant les variants ayant une fréquence d'allèle minimale supérieure à 5% et une densité de variants d'environ un SNP tous les 1kb (167,475 SNPs). Les individus ont été rassemblés en deux groupes basés sur une analyse de la structure de population prenant en compte la proximité géographique (Caye, F. Jay et al., 2018). Les données bioclimatiques correspondant aux coordonnées géographiques des individus ont été téléchargées à partir de la base de données WorldClim (`https://worldclim.org`). La matrice de données environnementales est constituée de 18 variables bioclimatiques, obtenues à partir de la température mensuelle et des valeurs de précipitation. La correction pour les effets des variables environnementales à été réalisé avec le modèle de régression à facteur latents de la librairie R `lfmm` (Caye, Jumentier et al., 2019). Pour la matrice de génotypes centrés, $\mathbf{Z}$, et la matrice des 18 variables bioclimatiques, $\mathbf{Y}$, la fonction estime une matrice de génotypes modifiés , $\mathbf{W}$, à l'aide du modèle $\mathbf{Z} = \mathbf{Y}\mathbf{B}^T + \mathbf{W} + \epsilon$. Afin que l'estimation de la matrice latente ($\mathbf{W}$) reste le plus proche possible de $\mathbf{Z}$, on utilise $k = n - 19 = 222$ facteurs pour calculer $\mathbf{W}$.

Résultats Comme expliqué dans le paragraphe précédent, les individus ont été regroupés en deux populations localisées dans le Sud et le Nord de la Suède (Figure 5.11A). Pour ces groupes, la valeur moyenne de F_{ST} le long du génome est de 7.9%. Cette valeur est plus grande que la plus grande valeur singulière au carré de la matrice intra-population, égale à 4.9 %. La part de variance expliquée par le premier axe de l'ACP est égal à 8.5%, supérieure à la F_{ST} (Figure 5.11), suggérant que le modèle à deux populations pourrait ne pas convenir aux données. La structure de population a ensuite été évaluée en utilisant $K = 3$ populations ancestrales. Les individus du Sud ont été divisés en 2 groupes le long d'un axe Est-Ouest, mettant en évidence une ascendance métissée de ces groupes (Figure 5.12). Avec trois groupes, les plus grandes valeurs singulières au carré de la matrice inter-population sont égales à 7.8 % et 2.5%. La seconde valeur singulière au carré est plus petite que la plus grande valeur singulière au carré de la matrice intra-population, égale à 3.7 %. Suite à cette expérience, nous avons décidé de nous focaliser sur le modèle à 2 populations. Après ajustement pour la variation bioclimatique, la plus grande valeur propre de l'ACP est égale à 6.5% (Figure 5.11C). La plus grande valeur singulière au carré de la matrice inter-population, qui définit la valeur moyenne de F_{ST} pour les génotypes modifiés, est égale à $\mathbb{E}[F_{\text{ST}}^{\text{adj}}] = 5.3\%$. Les valeurs propres suivantes de l'ACP sont égales à 4.9%, 3.2%, 2.3%, et ces valeurs ne sont pas affectées par les variables bioclimatiques (Figure 5.11B). De plus, ces valeurs propres coïncident avec les plus grandes valeurs singulières au carré de la matrice résiduelle, $Z_{\text{S}}^{\text{adj}}$, égale à 5.1%, 3.3%, 2.6% (Figure 5.11B). En comparant $\mathbb{E}[F_{\text{ST}}^{\text{adj}}]$ à $\mathbb{E}[F_{\text{ST}}]$, ces résultats montrent que la part de variance relative expliquée par le climat le long du premier axe est d'environ 33%. Ces résultats fournissent une preuve que le climat a eu un impact sur la différenciation des populations le long de l'axe Sud-Nord, mais moins sur les autres axes de variation génétique. En résumé, ces résultats suggèrent que les conditions bioclimatiques ont joué un rôle majeur dans la divergence génétique des populations du Sud et du Nord d'*A. thaliana* en Scandinavie.

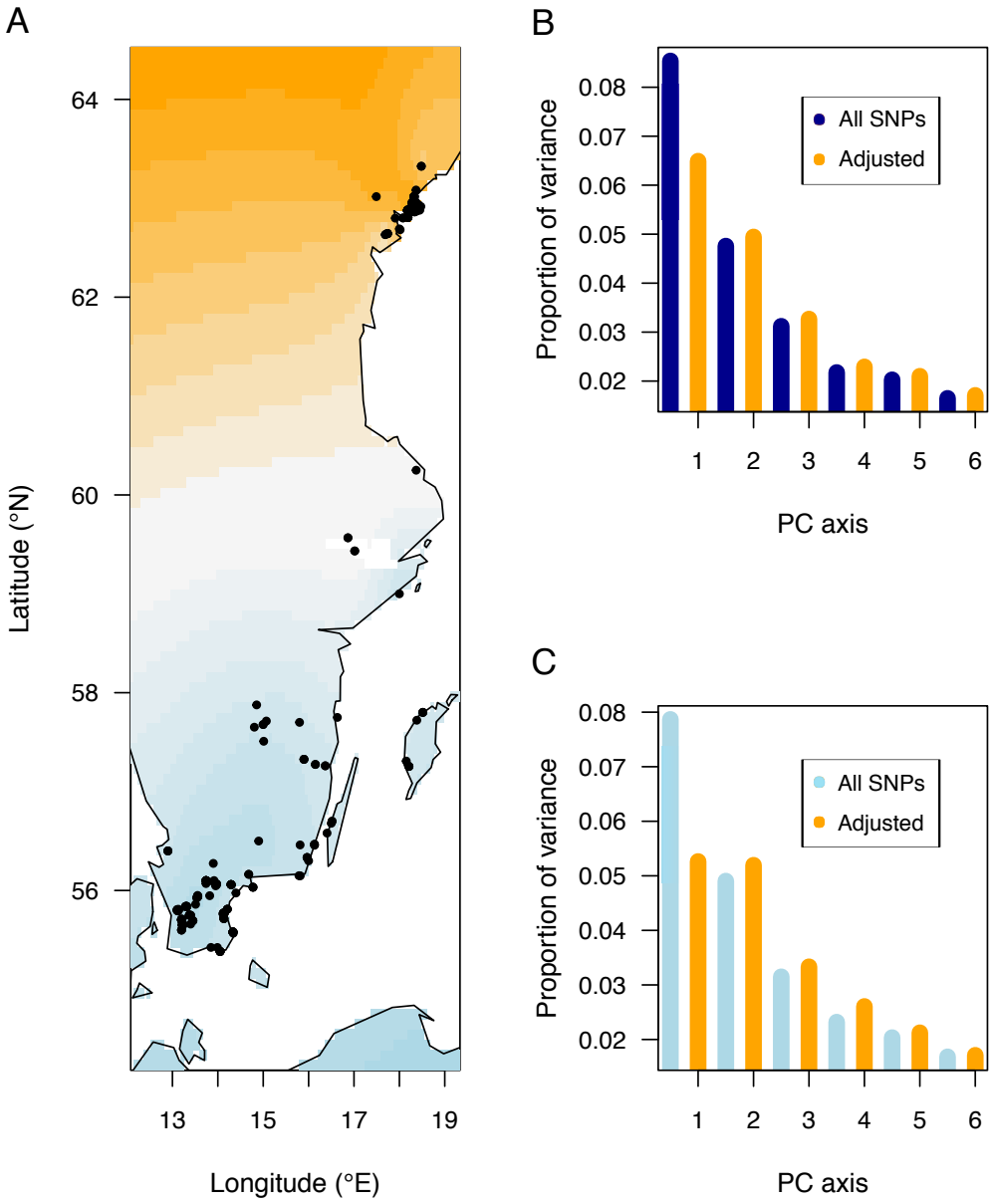

Figure 5.11 – F_{ST} **neutre pour** *Arabidopsis thaliana* **en Scandinavie.**
(**A**) Localisations géographiques de 241 échantillons et inférence de la structure de population à partir d'une méthode spatiale (Couleur bleue : groupe du Sud, couleur orange : groupe du Nord).
(B) Part de variance expliqués par les axes PC avant ajustement du génotypes pour les variables bioclimatiques (bleue) et après ajustement (couleur orange).
(C) Part de variance expliquée par le premier axe de la matrice inter-population, et par les premiers axes de la matrice résiduelle (5 composantes) avant ajustement (bleu) et après ajustement (orange). Le coefficient de Wright est représenté par les valeurs du premier axe

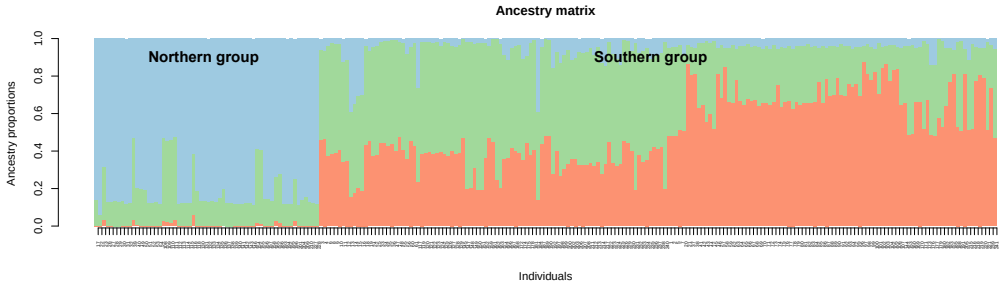

Figure 5.12 – **Estimations des coefficients d'ascendance pour 241 individus suédois d'Arabidopsis Thaliana**

Les coefficients d'ascendances ont été obtenu à l'aide d'un programme d'estimation de l'ascendance spatiallement explicite `tess3r` avec $K = 3$ populations. Le groupe du Sud affiche un haut niveau d'ascendance métissé.

6. Utilisation des résultats pour un décalage génétique interprétable comme une valeur de F_{ST}

Toute la théorie développée dans ce chapitre a généré des réflexions autour du développement d'un autre indice de décalage génétique interprétable comme une F_{ST}. Cet indice a été implémenté dans la librairie R `LEA` (Gain et François, 2021) et utilise le modèle LFMM.

6. 1. Définition du décalage génétique

Rappelons que pour une population donnée, les statistiques de décalage mesurent la divergence en fréquence d'allèle dans les conditions actuelles et dans une population fictive ayant des fréquences correspondant aux nouvelles conditions environnementales. Ici, nous proposons comme mesure de divergence l'indice F_{ST}, et nous proposons de le calculer de la manière suivante. Pour la population, on considère deux ensembles de variables environnementales, $\mathbf{Y}_{\mathrm{current}}$ et $\mathbf{Y}_{\mathrm{future}}$. La matrice de variables environnementales actuelles est d'abord utilisée pour ajuster un modèle LFMM, et le modèle ajusté est ensuite appliqué aux nouvelles données pour la prédiction des réponses. On contruit alors deux matrices $\mathbf{Z}_{\mathrm{fit}}$ et $\mathbf{Z}_{\mathrm{pred}}$

$$\mathbf{Z}_{\mathrm{fit}} = \mathbf{Y}_{\mathrm{current}}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T\,,$$

et

$$\mathbf{Z}_{\mathrm{pred}} = \mathbf{Y}_{\mathrm{future}}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T\,,$$

où $\mathbf{B}$, sont les tailles d'effet, $\mathbf{UV}$ la matrice des facteurs latents ajustés par un modèle LFMM à l'aide des données actuelles. Ensuite, on pose σ_{pred} et σ_{fit} les plus grandes valeurs singulières des matrices $\mathbf{Z}_{\mathrm{pred}}$ et $\mathbf{Z}_{\mathrm{fit}}$, et $\sigma_{\mathrm{pred+fit}}$ la plus grande valeur singulière de la matrice concaténée $(\mathbf{Z}_{\mathrm{pred}}, \mathbf{Z}_{\mathrm{fit}})^T$. Toutes les matrices sont standardisées. On calcule alors un décalage génétique, F_{offset}, de la manière suivante

$$1 - F_{\mathrm{offset}} = \frac{1 - \sigma^2_{\mathrm{pred+fit}}}{1 - (\sigma^2_{\mathrm{pred}} + \sigma^2_{\mathrm{fit}})/2}\,.$$

Ce décalage génétique mesure la quantité de dérive génétique séparant la population adaptée aux variables environnementales actuelles de la population fictive adaptée aux variables pro-

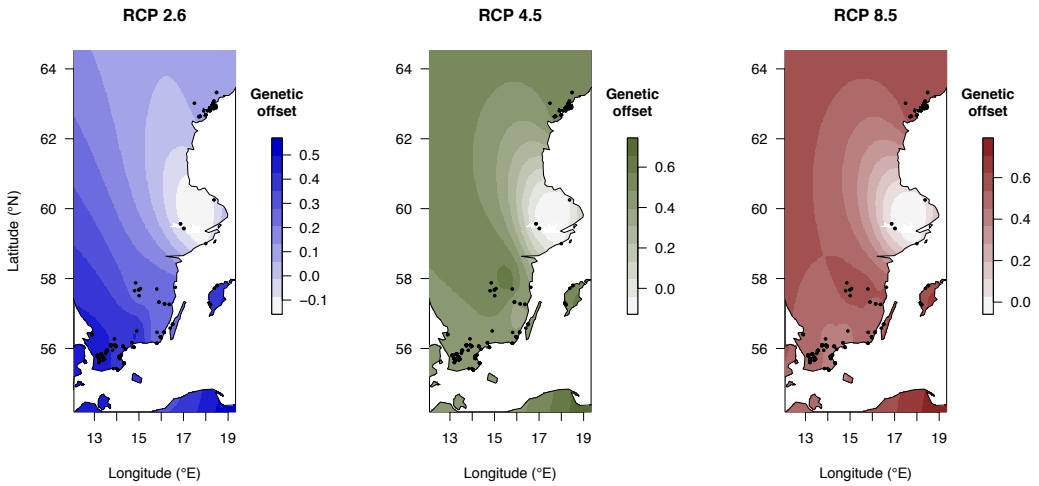

Figure 5.13 – **Décalage génétique pour les populations scandinaves d'***Arabidopsis Thaliana*
Décalage génétique calculé à partir de la projection de 18 variables bioclimatiques à 3 modèles de climats RCP 2.6, 4.5 et RCP 8.5 (70 ans) (IPCC, 2014). Les statistiques de décalage génétique ont été calculé pour 8 populations. Les valeurs ont été interpolé en utilisant l'algorithme de kriging implémenté dans `fields` 10.2 au 241 sites d'échantillonages représentés comme des points noirs.

jetées. Ce décalage est interprétable comme un indice de F_{ST} ce qui en fait un avantage par rapport à d'autres indices de décalage génétique.

6. 2. Application aux données *Arabidopsis thaliana*

Pour tester cette mesure, nous l'avons appliquée aux données d'*A. Thaliana* présentées dans la section précédente. Cette fois, les individus ont été rassemblés dans 8 groupes suite à une analyse préliminaire de la structure à l'aide du logiciel `snmf`, une méthode rapide et efficace d'estimation du métissage des individus basée sur des algorithmes de factorisation de matrices non négatives (Frichot et al., 2014), et sur la base de la proximité géographique. Nous utilisons $K = 4$ facteurs dans la prédiction LFMM. Les variables environnementales projetées ont été obtenues à partir de trajectoire RCP. Nous avons choisi les scénarios RCP 2.6, RCP 4.5 et RCP 8.5 à 70 ans.
Sous le scénario RCP 2.6, les statistiques de décalage génétique vont de 0% à 56%, avec une valeur moyenne de 29% (Figure 5.13). Pour le RCP 4.5 et 8.5, les valeurs vont respectivement jusque 74% et 79% avec des valeurs moyennes de 51% et 60%. Dans le RCP 2.6, la population la plus exposée est à une latitude d'environ 50°N dans le sud de la Scandinavie. Pour RCP 4.5 et 8.5, les populations les plus exposées sont dans le Nord aux latitudes 62°-64°N.

6. 3. Conclusions

En supposant un modèle à K populations discrètes, nous avons établi une relation entre le coefficient de consanguinité de Wright, F_{ST}, et les $(K-1)$ plus grandes valeurs singulières au carré de la matrice inter-population et de l'ACP standardisée de la matrice de génotype. Nous avons également établi une relation similaire entre la D_{ST} et les valeurs propres de l'ACP non standardisée. Ces relations justifie l'utilisation de l'ACP pour décrire la structure de population dans le cadre de grandes matrices de génotype. Elles étendent les résultats obtenus à partir de la théorie de la coalescence pour deux populations divergentes dans la Ref. (McVean, 2009)

à n'importe quel modèle de population discrète. En supposant que la TMA s'applique aux matrices résiduelles, elles augmentent la précision des résultats précédents, en clarifiant pour quelle taille d'échantillon et pour quelle quantité de locus ces relations peuvent être valides. Dans nos simulations, nous avons trouvé que la plus grande valeur singulière au carré de la matrice résiduelle était bien prédite par la TMA. La TMA nous fournit également une valeur seuil de F_{ST}, égale à $\theta = (1/\sqrt{L} + 1/\sqrt{n-1})^2$, en-dessous de laquelle il n'y pas de preuve de la structure de population pour deux populations ou plus. Ce seuil diffère de la valeur seuil de $1/\sqrt{nL}$ proposée par Ref. (Patterson et al., 2006), et il a été validé par des simulations de modèle à une population. En plus de faire le lien entre l'ACP de la matrice de génotype et les coefficients de consanguinité, nos résultats ont des implications pour l'analyse de matrices ajustées, en fournissant une statistique analogue à la F_{ST} pour ces données. Les génotypes ajustés sont présents dans diverses applications, tel que l'ADN ancien, pour corriger pour les biais dû aux artefacts techniques ou d'échantillonages, ou en génomique écologique où cela permet d'évaluer la part de différenciation des populations expliquée par les variations environnementales. L'estimation proposée des coefficients de consanguinité est donc de grande importance pour la compréhension de l'historique démographique des populations et leur adaptation aux variations environnementales. Enfin, nous avons pu appliquer cette théorie afin de développer une mesure de décalage génétique interprétable comme une F_{ST} entre une population adaptée aux conditions environnementales et sa population homologue adaptée à des conditions modifiées. L'ensemble des résultats de ce chapitre a fait l'objet d'une publication dans PLoS Genetics (François et Gain, 2021).

Chapitre 6

Conclusions et perspectives

Dans ce chapitre, nous rappelons les principales contributions de notre travail. Autour de la notion de décalage génétique, nous proposons une nouvelle mesure qui cherche à répondre à certaines limites identifiées dans la littérature scientifique. Nous proposons également une théorie quantitative du décalage génétique permettant de lier cette statistique à la valeur sélective dans l'environnement modifié et d'unifier les méthodes existantes. Nous validons cette théorie à l'aide de simulation et de données réelles, soulignant l'importance que peuvent avoir ces statistiques de décalage génétique pour la gestion de la conservation face au changement climatique. Nous établissons également une théorie spectrale des indices de fixation de Wright mettant en évidence une relation entre l'analyse en composantes principales et les indices de fixation. Nous proposons ensuite quelques pistes de réflexion pour prolonger le travail durant cette thèse. Ces pistes se concentrent sur les limites persistantes des mesures de décalage génétique que nous n'avons pas résolu durant cette thèse et sur un prolongement de la théorie lorsqu'on prend également en compte les effets directs de l'environnement sur le phénotype.

1. Contributions de la thèse

Notre thèse s'est principalement focalisé sur la notion de décalage génétique. Un travail parallèle a également été effectué sur la relation entre l'analyse en composante principales et les indices de fixation de Wright.

1. 1. Une nouvelle mesure de décalage génétique

Tout d'abord, nous proposons une nouvelle mesure de décalage génétique, que nous avons appelée fossé génétique. Nous proposons une double interprétation de cette mesure, à la fois comme distance dans la niche écologique, et comme distance génétique. Cette méthode prend en compte les facteurs de confusion en utilisant un modèle mixte à facteurs latents. Il s'agit d'une méthode multivariée qui modélise les effets environnementaux corrélés. On peut obtenir l'importance relative des variables environnementales et cette méthode prend en compte l'architecture polygénique des traits adaptatifs. Nous avons montré théoriquement que, sous les hypothèses du modèle infinitésimal de Fisher et de sélection stabilisatrice gaussienne, le fossé génétique est proportionnel au logarithme de la valeur sélective dans l'environnement modifié. Nous avons également prouvé une équivalence théorique entre la RDA et le fossé génétique et nous avons établi des relations avec GF et Rona.

1. 2. Validation empirique des mesures

Ensuite, nous avons voulu valider les résultats théoriques établis au chapitre 3 à l'aide de données simulées et de données réelles. Nous avons notamment utilisé le logiciel SLiM 3.7 pour effectuer des simulations spatiallement explicites basées sur l'individu. Ces simulations nous ont permis de simuler des processus d'adaptation locale suivi d'une variation brutale de l'environnement et de récupérer les valeurs sélectives des individus faisant face à ces modifications de l'environnement. Il nous a alors été possible de comparer les valeurs de décalage génétique avec les valeurs sélectives dans l'environnement modifié et de vérifier que la relation établie théoriquement était également valable dans la simulation. Nous avons également retrouvé cette relation avec des données réelles et une expérience de jardin commun pour du mil. Le poids total moyen des graines obtenu dans le jardin commun a alors été utilisé comme approximation de la valeur sélective dans l'environnement modifié.

1. 3. Relation entre l'ACP et les indices de fixation de Wright

Enfin, nous avons établi une relation théorique entre l'ACP et les indices de fixation de Wright. Dans un modèle à K populations discrètes la F_{ST} moyenne le long du génome est approchée par les $(K-1)$ valeurs propres de l'ACP standardisée. Ce résultat permet d'obtenir une valeur de F_{ST} pour des matrices de génotype modifiées et nous fournit une définition alternative de mesure de décalage génétique.

2. Perspectives

Nous avons résumé dans la section précédente, les contributions de notre thèse. Nous proposons maintenant quelques pistes de réflexions permettant d'approfondir la compréhension de la mesure de décalage génétique.

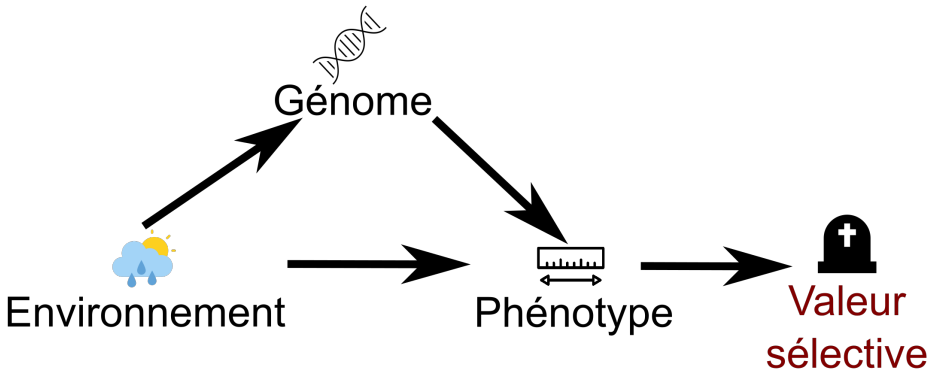

Figure 6.1 – Illustration des effets directs et indirects de l'environnement sur les traits adaptatifs

La variance des traits phénotypiques est dépendantes des effets médiés par le génotype, on parle alors d'effets indirects. Ce sont ces effets qui sont pris en compte dans le calcul du décalage génétique. La variance dépend également des effets directs.

2. 1. Valeur sélective non linéaire

Les scénarios de simulations dans cette thèse sont des scénarios pour lesquelles la valeur sélective varie généralement linéairement avec les variations d'environnements. Cela permet donc aux méthodes linéaires de ne pas être désavantagées par rapport à d'autres méthodes. Il pourrait toutefois être intéressant de comparer les méthodes dans des contextes non linéaires, notamment pour voir comment se comporte GF ainsi que la méthode basée sur des VAE présentée dans le chapitre 3. Pour ce faire, on pourrait modifier la fonction de valeur sélective de nos scénarios SLiM.

2. 2. Effets directs de l'environnement sur le phénotype

Notre théorie quantitative du décalage génétique s'appuie sur une hypothèse sous-jacente que nous n'avons pas abordée dans ce manuscrit. Les effets de l'environnement sur les traits adaptatifs sont systématiquement médiés par le génome. En pratique, une partie de la variance des traits adaptatifs doit être attribuée aux effets directs de l'environnement sur ces derniers (Figure 6.1). Une piste de réflexion serait donc d'étudier les implications de la négligence des effets directs, voire de développer le cadre théorique permettant de quantifier à la fois les effets directs et indirects. Toutefois, pour pouvoir ensuite appliquer ce nouveau cadre théorique à des données réelles, il sera alors nécessaire d'avoir accès à des données de phénotypes en plus des données de génotypes.

2. 3. Hypothèse d'adaptation préalable

Une des hypothèses fortes du calcul du décalage génétique et l'adaptation des populations à l'environnement dans lequel elles sont échantillonnées. Il pourrait être intéressant d'étudier théoriquement et à l'aide de simulations les implications lorsque cette hypothèse n'est pas vérifiée. Autrement dit, lorsqu'il existe déjà un décalage génétique de la population entre son environnement optimal et l'environnement dans lequel elle est prélevée. On voudrait notamment comprendre quelles sont les répercussions en terme d'interprétation des valeurs de décalage

génétique dans de nouveaux environnements, par exemple en comprenant comment cela impact la relation avec la valeur sélective dans l'environnement modifié.

Bibliographie

Aitken, S. N. et J. B. Bemmels. "Time to get moving : assisted gene flow of forest trees." In : *Evolutionary applications* 9(1) (2016), p. 271-290.

Aitken, S. N. et M. C. Whitlock. "Assisted gene flow to facilitate local adaptation to climate change." In : *Annual Review of Ecology, Evolution, and Systematics* 44 (2013), p. 367-388.

Aitken, S. N., S. Yeaman, J.A. Holliday, T. Wang et S. Curtis-McLane. "Adaptation, migration or extirpation : climate change outcomes for tree populations." In : *Evolutionary applications* 1 (2008), p. 95-111.

Allentoft, ME, M Sikora, Sjögren KG, Rasmussen S, Rasmussen M et Stenderup J. "Population genomics of Bronze Age Eurasia." In : *Nature* 522 (2015), p. 167-172.

Alonso-Blanco, C., J. Andrade, C. Becker, F. Bemm, J. Bergelson, K. M. Borgwardt et X. Zhou. "1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana". In : *Cell* 166 (2016), p. 481-491.

Aquino, S. O. de, C. Kiwuka, R. Tournebize, C. Gain, P. Marraccini et al. "Adaptive potential of Coffea canephora from Uganda in response to climate change." In : *Molecular Ecology* 31 (2022), p. 1800-1819.

Balding, DJ. et RA. Nichols. "A method for quantifying differentiation between populations at multi-allelic locus and its implications for investigating identity and paternity." In : *Genetica* 96 (1995), p. 3-12.

Barnosky, A., N. Matzke et S. Tomiya. "Has the Earth's sixth mass extinction already arrived ?" In : *Nature* 471 (2011), p. 51-57.

Barton, N. H., A. M. Etheridge et A. Véber. "The infinitesimal model : Definition, derivation, and implications." In : *Theor Popul Biol* 118 (2017), p. 50-73.

Bauer, F. L. et C. T. Fike. "Norms and exclusion theorems." In : *Numerische Mathematik* 2 (1960), p. 137-141.

Bay, R. A., R. J. Harrigan, V. Le Underwood, H. L. Gibbs, T. B. Smith et K. Ruegg. "Genomic signals of selection predict climate-driven population declines in a migratory bird." In : *Science* 359(6371) (2018), p. 83-86.

Berven, K. A. "The genetic basis of altitudinal variation in the wood frog Rana sylvatica II. An experimental analysis of larval development." In : *Oecologia* 52(3) (1982), p. 360-369.

— "The genetic basis of altitudinal variation in the wood frog Rana sylvatica. I. An experimental analysis of life history traits." In : *Evolution* (1982), p. 962-983.

Bhatia, R. *Matrix Analysis*. Springer Science et Business Media, 2013.

Browne, L., J. W. Wright, S. Fitz-Gibbon, P. F. Gugger et V. L. Sork. "Adaptational lag to temperature in valley oak (Quercus lobata) can be mitigated by genome-informed assisted gene flow." In : *Proceedings of the National Academy of Sciences of the United States of America* 116 (2019), p. 25179-25185.

Bryc, K., W. Bryc et J. W. Silverstein. "Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations". In : *Theoretical population biology* 89 (2013), p. 34-43.

Bryson, J., R. Vershynin et H. Zhao. "Marchenko-Pastur law with relaxed independence conditions." In : *arXiv* (2019).

Bürger, R. *The mathematical theory of selection, recombination, and mutation.* John Wiley & Sons., 2000.

Capblancq, T., M. C. Fitzpatrick, R. A. Bay, M. Exposito-Alonso et S. R. Keller. "Genomic prediction of (mal)adaptation across current and future climatic landscapes." In : *Annual Review of Ecology, Evolution, and Systematics* 51 (2020), p. 245-269.

Capblancq, T. et B. R. Forester. "Redundancy analysis : A Swiss Army Knife for landscape genomics." In : *Methods in Ecology and Evolution* 12(12) (2021), p. 2298-2309.

Capblancq, T., X. Morin, M. Gueguen, J. Renaud, S. Lobreaux et Bazin E. "Climate-associated genetic variation in Fagus sylvatica and potential responses to climate change in the French Alps." In : *Journal of Evolutionary Biology* 33(6) (2020), p. 783-796.

Caye, K., F. Jay, O. Michel et O. François. "Fast inference of individual admixture coefficients using geographic data." In : *Ann Appl Stat* 12 (2018), p. 586-608.

Caye, K., B. Jumentier, J. Lepeule et O. François. "LFMM 2 : fast and accurate inference of gene-environment associations in genome-wide studies." In : *Molecular Biology and Evolution* 36(4) (2019), p. 852-860.

Ceballos, G., P.R. Ehrlich et R. Dirzo. "Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines." In : *PNAS* 114(30) (2017), p. 6089-6096.

Chen, Y., Z. Jiang, P. Fan, P. G. Ericson, G. Song, X. Luo et Y. Qu. "The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability." In : *Nature Communications* 13(1) (2022), p. 4821.

Chen, Z. X., S. Y. Chen et D. W. Dickson. *Nematology : advances and perspectives.* 2004.

Clausen, J., D. D. Keck et W. M. Hiesey. "Experimental studies on the nature of species. I. Effect of varied environments on western North American plants." In : (1940).

Cockerham, C. C. "Variance of gene frequencies." In : *Evolution* (1969), p. 72-84.

Consortium, G. P., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison et H. M. Kang. "A global reference for human genetic variation." In : *Nature* 526 (2015), p. 68-74.

Cook, L. M. et I. J. Saccheri. "The peppered moth and industrial melanism : evolution of a natural selection case study." In : *Heredity* 110(3) (2013), p. 207-212.

Coop, G., D. Witonsky, A. Di Rienzo et J. K. Pritchard. "Using environmental correlations to identify locus underlying local adaptation." In : *Genetics* 185(4) (2010), p. 1411-1423.

Cox, D. R. "The regression analysis of binary sequences." In : *Journal of the Royal Statistical Society : Series B (Methodological)* 20(2) (1958), p. 215-232.

Darwin, C. *On the origin of species.* 1859.

Davidson, R. et J. MacKinnon. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." In : *Econometrica* 49 (1981), p. 781-793.

De Mita, S., A. C. Thuillet, L. Gay, N. Ahmadi, S. Manel, J. Ronfort et Y. Vigouroux. "Detecting selection along environmental gradients : analysis of eight methods and their effectiveness for outbreeding and selfing populations." In : *Molecular ecology* 22(5) (2013), p. 1383-1399.

Díaz, S. et al. "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services." In : (2019).

Dirzo, R., H. S. Young, M. Galetti, G. Ceballos, N. J. Isaac et B. Collen. "Defaunation in the Anthropocene." In : *Science* 345(6195) (2014), p. 401-406.

DJ., Balding. “A tutorial on statistical methods for population association studies.” In : *Nat Rev Genet.* 7(10) (2006), p. 781-791.

Elith, J. et J. R. Leathwick. “Species distribution models : ecological explanation and prediction across space and time.” In : *Annual Review of Ecology, Evolution and Systematics* 40(1) (2009), p. 677-697.

Ellis, N., S. Smith et C. Pitcher. “Gradient forests : calculating importance gradients on physical predictors.” In : *Ecology* 93 (2012), p. 156-168.

Ferrier, S., G. Manion, J. Elith et K. Richardson. “Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment.” In : *Diversity and distributions* 13(3) (2007), p. 252-264.

Fisher, R. A. “XV.—The correlation between relatives on the supposition of Mendelian inheritance.” In : *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52(2) (1919), p. 399-433.

Fitzpatrick, M. C., V. E. Chhatre, R. Y. Soolanayakanahally et S. R. Keller. “Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests.” In : *Molecular Ecology Resources* 21(8) (2021), p. 2749-2765.

Fitzpatrick, M. C. et S. R. Keller. “Ecological genomics meets community-level modelling of biodiversity : Mapping the genomic landscape of current and future environmental adaptation.” In : *Ecology Letters* 18(1) (2015), p. 1-16.

Fitzpatrick, M. C., S. R. Keller et K. E. Lotterhos. “Comment on “Genomic signals of selection predict climate-driven population declines in a migratory bird””. In : *Science* 361(6401) (2018), eaat7279.

Foden, W. B., B. E. Young, H. R. Akçakaya, R. A. Garcia, A. A. Hoffmann et B. A. Stein. “Climate change vulnerability assessment of species.” In : *Wiley interdisciplinary reviews : climate change* 10(1) (2019), e551.

Forester, B. R., J. R. Lasky, H. H. Wagner et D. L. Urban. “Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations.” In : *Molecular Ecology* 27(9) (2018), p. 2215-2233.

François, O. et C. Gain. “A spectral theory for Wright’s inbreeding coefficients and related quantities.” In : *PLoS Genetics* 17(7) (2021), e1009665.

François, O. et F Jay. “Factor analysis of ancient population genomic samples.” In : *Nat Commun.* 11 (2020), p. 4661.

Fraser, D. J., L. K. Weir, L. Bernatchez, M. M. Hansen et E. B. Taylor. “Extent and scale of local adaptation in salmonid fishes : review and meta-analysis.” In : *Heredity* 106(3) (2011), p. 404-420.

Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard et O. François. “Fast and efficient estimation of individual ancestry coefficients.” In : *Genetics* 196(4) (2014), p. 973-983.

Gain, C. et O. François. “LEA 3 : Factor models in population genetics and ecological genomics with R.” In : *Molecular Ecology Resources* 21(8) (2021), p. 2738-2748.

Gain, C., B. Rhone, P. Cubry, I. Salazar, F. Forbes, Y. Vigouroux et O. François. “A quantitative theory for genomic offset statistics.” In : *bioRxiv* (2023).

Gautier, M. “Genome-wide scan for adaptive divergence and association with population-specific covariates.” In : *Genetics* 201(4) (2015), p. 1555-1579.

Giec. “Intergovernmental Panel on Climate Change. Appendix I : Glossary.” In : (2017).

Gougherty, A. V., S. R. Keller et M. C. Fitzpatrick. “Maladaptation, migration and extirpation fuel climate change risk in a forest tree species.” In : *Nature Climate Change* 11 (2021), p. 166-171.

Grinnell, J. “The niche-relationships of the California Thrasher.” In : *The Auk* 34 (1917), p. 427-433.

Guisan, A. et W. Thuiller. "Predicting species distribution : offering more than simple habitat models." In : *Ecology letters* 8(9) (2005), p. 993-1009.

Guisan, A. et N. E. Zimmermann. "Predictive habitat distribution models in ecology." In : *Ecological modelling* 135(2-3) (2000), p. 147-186.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas et D. Reich. "Massive migration from the steppe was a source for Indo-European languages in Europe." In : *Nature* 522(7555) (2015), p. 207-211.

Haller, B. et P.W. Messer. "SLiM 3 : Forward Genetic Simulations Beyond the Wright–Fisher Model." In : *Molecular Biology and Evolution* 36(3) (2019), p. 632-637.

Hallmann, C. A., M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan et H. De Kroon. "More than 75 percent decline over 27 years in total flying insect biomass in protected areas." In : *PloS one* 12(10) (2017).

Hartl, D. L., A. G. Clark et A. G. Clark. *Principles of population genetics*. Sunderland : Sinauer associates., 1997.

Hausfather, Z. et G.P. Peters. "Emissions – the 'business as usual' story is misleading." In : *Nature* 577 (2021), p. 618-620.

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd, D. B. Lowry et M. C. Whitlock. "Finding the genomic basis of local adaptation : pitfalls, practical solutions, and future directions." In : *The American Naturalist* 188(4) (2016), p. 379-397.

Holsinger, K. E. et B. S. Weir. "Genetics in geographically structured populations : defining, estimating and interpreting FST." In : *Nature Reviews Genetics* 10(9) (2009), p. 639-650.

IPCC. "Climate Change 2014 : Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change." In : (2014).

Jay, F., S. Manel, N. Alvarez, E. Y. Durand, W. Thuiller et R. Holderegger. "Forecasting changes in population genetic structure of alpine plants in response to global warming." In : *Molecular Ecology* 21(10) (2012), p. 2354-2368.

Johnstone, I. M. "Multivariate analysis and Jacobi ensembles : largest eigenvalue, Tracy-Widom limits and rates of convergence." In : *Ann Stat.* 36 (2008), p. 2638-2716.

— "On the distribution of the largest eigenvalue in principal components analysis." In : *The Annals of statistics* 29(2) (2001), p. 295-327.

Johnstone, I. M. et D. Paul. "PCA in high dimensions : An orientation." In : *Proc IEEE* 106 (2018), p. 1277-1292.

Jolliffe, I. *Principal component analysis*. Springer, 1986.

Jolliffe, I. T. et J. Cadima. "Principal component analysis : a review and recent developments." In : *Philosophical transactions of the royal society A : Mathematical, Physical and Engineering Sciences* 374(2065) (2016), p. 20150202.

Kawecki, T. J. et D. Ebert. "Conceptual issues in local adaptation." In : *Ecology letters* 7(12) (2004), p. 1225-1241.

Keller, S. R., V. E. Chhatre et M. C. Fitzpatrick. "Influence of range position on locally adaptive gene–environment associations in Populus flowering time genes." In : *Journal of Heredity* 109(1) (2018), p. 47-58.

Kingma, D. P. et M. Welling. "Auto-encoding variational bayes." In : *arXiv* (2013).

Knight, C. ""Most people are simply not designed to eat pasta" : evolutionary explanations for obesity in the low-carbohydrate diet movement." In : *Public Understanding of Science* 20(5) (2011), p. 706-719.

Krimbas, C. B. "On fitness." In : *Biology and Philosophy* 19(2) (2004), p. 185-203.

Làruson, À. J., M. C. Fitzpatrick, S. R. Keller, B. C. Haller et K. E. Lotterhos. "Seeing the Forest for the trees : Assessing genetic offset predictions with Gradient Forest." In : *Evolutionary Applications* 15(3) (2022), p. 403-416.

Leek, J.T., W.E. Johnson, H.S. Parker, Jaffe A.E. et Storey J.D. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." In : *Bioinformatics* 28(6) (2012), p. 882-883.

Liao, C. H. et F. L. Lin. "Deep generative models of gravitational waveforms via conditional autoencoder." In : *Physical Review D* 103(12) (2021), p. 124051.

Lind, B. M., R. Candido-Ribeiro, P. Singh, M. Lu, D. Obreht Vidakovic, T. R. Booker et S. N. Aitken. "How useful is genomic data for predicting maladaptation to future climate ?" In : *bioRxiv* (2023).

Lloyd, E., D. S. Wilson et E. Sober. "Evolutionary mismatch and what to do about it : A basic tutorial." In : *Evolutionary Applications* (2011), p. 2-4.

Luikart, G., P. R. England, D. Tallmon, S. Jordan et P. Taberlet. "The power and promise of population genomics : from genotyping to genome typing." In : *Nature reviews genetics* 4(12) (2003), p. 981-994.

Luu, K., E. Bazin et M. G. Blum. "pcadapt : an R package to perform genome scans for selection based on principal component analysis." In : *Molecular ecology resources* 17(1) (2017), p. 67-77.

Malécot, G. *Les mathématiques de hérédité*. Masson, Paris, 1948.

Marčenko, V.A. et Pastur L.A. "Distribution of eigenvalues for some sets of random matrices." In : *Mat Sb* 1 (1967), p. 457.

Martins, H, K Caye, K Luu, M. Blum et O. François. "Identifying outlier locus in admixed and in continuous populations using ancestral population differentiation statistics." In : *Mol Ecol* 25 (2016), p. 5029-5042.

Mathieson, I, I Lazaridis, N Rohland, S Mallick, N Patterson et SA Roodenberg. "Genome-wide patterns of selection in 230 ancient Eurasians." In : *Nature* 528 (2015), p. 499.

Mathieson, I, S Roodenberg, C Posth, A Szécsényi-Nagy, N Rohland et S Mallick. "The genomic history of southeastern Europe." In : *Nature* 555 (2018), p. 197.

Mayden, R.L. "A hierarchy of species concepts : the denoument in the saga of the species." In : *The units of diversity* (1997), p. 381-423.

Mayr, E. *Systematics and the Origin of Species*. New York : Columbia University Press, 1942.

— *The growth of biological thought : Diversity, evolution, and inheritance*. Harvard University Press., 1982.

McVean, G. "A genealogical interpretation of principal components analysis." In : *PLoS genetics* 5(10) (2009).

Menozzi, P., A. Piazza et L. Cavalli-Sforza. "Synthetic Maps of Human Gene Frequencies in Europeans : These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic." In : *Science* 201(4358) (1978), p. 786-792.

Nei, M. "Analysis of gene diversity in subdivided populations." In : *Proc. Natl. Acad. Sci.* 70 (1973), p. 3321-3323.

Nei, M. et R. K. Chesser. "Estimation of fixation indices and gene diversities." In : *Annals of human genetics* 47(3) (1983), p. 253-259.

Ochoa, A et Storey JD. "Estimating F_{ST} and kinship for arbitrary population structures." In : *PLoS Genet* 17 (2021).

Patterson, N., A. L. Price et D. Reich. "Population structure and eigenanalysis." In : *PLoS genetics* 2(12) (2006).

Poulton, E. B. *The colours of animals : their meaning and use, especially considered in the case of insects*. D. Appleton., 1890.

Pritchard, JK, M Stephens et Donnelly P. "Inference of population structure using multilocus genotype data." In : *Genetics* 155 (2000), p. 945-959.

Queiroz, K. "Ernst Mayr and the modern concept of species." In : *PNAS* 102 (2005), p. 6600-6607.

Rellstab, C. "Genomics helps to predict maladaptation to climate change." In : *Nature Climate Change* 11(2) (2021), p. 85-86.

Rellstab, C., B. Dauphin et M. Exposito-Alonso. "Prospects and limitations of genomic offset in conservation management." In : *Evolutionary applications* 14(5) (2021), p. 1202-1212.

Rellstab, C., F. Gugerli, A. J. Eckert, A. M. Hancock et R. Holderegger. "A practical guide to environmental association analysis in landscape genomics." In : *Molecular Ecology* 24(17) (2015), p. 4348-4370.

Rellstab, C., S. Zoller, L. Walthert, C. Bodénès, I. Lesur et A. R. Pluess. "Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions." In : *Molecular Ecology* 25(23) (2016), p. 5907-5924.

Rhoné, B., D. Defrance, C. Berthouly-Salazar, C. Mariac, P. Cubry et M. Couderc. "Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration." In : *Nature communications* 11(1) (2020), p. 1-9.

Ross, MG, C Russ, M Costello, A Hollinger, NJ Lennon, R Hegarty, C Nusbaum et DB. Jaffe. "Characterizing and measuring bias in sequence data." In : *Genome Biol* 14(5) (2013), p. 1-20.

Ruegg, K, RA Bay, EC Anderson, JF Saracco et RJ Harrigan. "Ecological genomics predicts climate vulnerability in an endangered southwestern songbird." In : *Ecol. Lett.* 21(7) (2018), p. 1085-96.

Sánchez-Bayo, F. et A.G Wyckhuys. "Worldwide decline of the entomofauna : A review of its drivers." In : *Biological Conservation* 232 (2019), p. 8-27.

Sang, Y., Z. Long, X. Dan, J. Feng, T. Shi, C. Jia et J. Wang. "Genomic insights into local adaptation and future climate-induced vulnerability of a keystone forest tree in East Asia." In : *Nature Communications* 13(1) (2022), p. 6541.

Savolainen, O., M. Lascoux et J. Merilä. "Ecological genomics of local adaptation." In : *Nature Reviews Genetics* 14(11) (2012), p. 807-820.

SCDB. "3e édition des Perspectives mondiales de la diversité biologique". In : *Secrétariat de la Convention sur la diversité biologique 2010* (2010).

Schlaepfer, M. A., M. C. Runge et P. W. Sherman. "Ecological and evolutionary traps." In : *Trends in ecology & evolution* 17(10) (2002), p. 474-480.

Schmalhausen, I. I. "Stabilizing selection and its place among the factors of evolution." In : *Zh. Obshch. Biol* 2(3) (1941), p. 307-354.

Schoville, S. D., A. Bonin, O. François, S. Lobreaux, C. Melodelima et S. Manel. "Adaptive genetic variation on the landscape : methods and cases." In : *Annual Review of Ecology, Evolution, and Systematics* 43 (2012), p. 23-43.

Schwalm, C. R., Glendon et P.B. Duffy. "RCP8.5 tracks cumulative CO2 emissions." In : *PNAS* 117 (33) (2012), p. 19656-19657.

Slatkin, M. "Inbreeding coefficients and coalescence times." In : *Genetics Research* 58(2) (1991), p. 167-175.

Sork, V. L., F. W. Davis, R. Westfall, A. Flint, M. Ikegami, H. Wang et D. Grivet. "Gene movement and genetic association with regional climate gradients in California valley oak (Quercus lobata Née) in the face of climate change." In : *Molecular Ecology* 19(17) (2010), p. 3806-3823.

Steane, D. A., B. M. Potts, E. McLean, S. M. Prober, W. D. Stock, R. E. Vaillancourt et M. Byrne. "Genome-wide scans detect adaptation to aridity in a widespread forest tree species." In : *Molecular Ecology* 23(10) (2014), p. 2500-2513.

Supple, M. A., J. G. Bragg, L. M. Broadhurst, A. B. Nicotra, M. Byrne, R. L. Andrew et J. O. Borevitz. "Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change." In : *Elife* 7 (2018).

Urban, M. C. "Accelerating extinction risk from climate change." In : *Science* 348(6234) (2015), p. 571-573.

Waldvogel, A. M., B. Feldmeyer, G. Rolshausen, M. Exposito-Alonso, C. Rellstab et R. Kofler. "Evolutionary genomics can improve prediction of species' responses to climate change." In : *Evolution Letters* 4(1) (2020), p. 4-18.

Wang, IJ, RE Glor et Losos JB. "Quantifying the roles of ecology and geography in spatial genetic divergence." In : *Ecol Lett.* 16 (2013), p. 175-182.

Wang, J, Q Zhao, T Hastie et AB. Owen. "Confounder adjustment in multiple testing." In : *Ann Stat.* 45(5) (2017), p. 1863-1894.

Weir, B. S. et C. C. Cockerham. "Estimating F-statistics for the analysis of population structure." In : *Evolution* (1984), p. 1358-1370.

Wikipedia. URL : https://www.wikipedia.org/.

Williams, G. C. *Adaptation and natural selection : A critique of some current evolutionary thought (Vol. 75)*. Princeton university press., 2018.

Wright, S. "The interpretation of population structure by F-statistics with special regard to systems of mating." In : *Evolution* (1965), p. 395-420.