Analyse de la différenciation génétique à l'ère des nouvelles technologies de séquençage

Valentin Hivert

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Génétique et Génomique

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Unité de recherche : Centre de Biologie pour la Gestion des Populations (CBGP)

Analyse de la différenciation génétique à l'ère des nouvelles technologies de séquençage

Présentée par Valentin HIVERT
Le 14 Décembre 2018

Sous la direction de Renaud VITALIS
et Mathieu GAUTIER

Devant le jury composé de

Joëlle RONFORT, Directrice de recherche, INRA Montpellier	Présidente du jury
Christine DILLMANN, Professeure, Université Paris-Sud	Rapportrice
Anna-Sapfo MALASPINAS, Professeure assistante, Université de Lausanne	Rapportrice
Miguel PÉREZ-ENCISO, Professeur associé, Université Autonome de Barcelone	Examinateur
Renaud VITALIS, Directeur de recherche, INRA Montpellier	Directeur de thèse
Mathieu GAUTIER, Chargé de recherche, INRA Montpellier	Co-Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

REMERCIEMENTS

Je tiens en premier lieu à remercier évidemment mes deux directeurs de thèse, Renaud et Mathieu. Vous avez été de véritables piliers tout au long de ces trois années. Merci pour votre confiance, votre soutien et pour la liberté que vous m'avez accordé tout en m'apportant un cadre. Vous avez toujours fait preuve d'une grande disponibilité à mon égard, de beaucoup de patience (et il en a fallu) et j'ai énormément appris à vos côtés. Cela n'a pas été facile tout les jours (les petits coups de pression étaient nécessaires) mais votre rigueur scientifique m'inspire et je vais tout faire pour cultiver ce sens du détail. Encore merci à vous deux. J'ai évidemment une pensée pour les différentes personnes qui ont participé à cette thèse et avec qui j'ai collaboré. Raph, tu as toujours été là pour répondre à mes questions, sans oublier Arnaud, Miguel, Nico R. et Eric Petit, merci à vous. J'aimerais remercier Anna-Sapfo Malaspinas et Christine Dillmann qui ont accepté d'être mes deux rapportrices, ainsi que Joëlle Ronfort et Miguel Pérez-Enciso pour avoir accepté de faire partie de mon jury. Merci à vous pour l'évaluation de ce travail. Je remercie aussi les membres de mes comités de thèse, Stéphanie Manel, Michael Blum, Simon Boitard et Bertrand Servin pour les différents échanges qui m'ont été d'une grande aide.

Plus q'une aventure scientifique, la thèse est aussi une aventure humaine faite de rencontres, toutes plus belles les unes que les autres. J'ai une pensée spéciale pour mes deux co-bureaux. Laure, tu as su m'écouter et me soutenir tout du long, et cela malgré mon sale caractère et mes blagues si recherchées (mais je sais que ça va te manquer). Pour le simple fait de nous avoir supporté avec Florian, tu mérites vraiment une médaille. Merci aussi à toi Florian pour cette belle amitié faite de science, de sport, de barbecues préhistoriques, de "3 brasseurs" et d'Oktoberfest au grand dam de mon foie. Et comment ne pas parler de toi Alex. En tant que collègue tu as toujours pris du temps pour me dépêtrer de mes soucis de codes, mais tu es surtout un véritable ami. Tu as été essentiel à ma bonne humeur avec ton humour si fin, pour mon plus grand plaisir ainsi que celui de Laure (quoique...). Nos discussions d'une

si grande profondeur d'esprit vont vraiment me manquer. Merci aussi à toi Louise pour ton amitié et ton soutien tout au long de ma thèse. Mélo pour tous ces supers moments passés ensemble. Ghais (le Syrien presque Creusois), Yurena, Isidora, Lili, Vitor, Fernando pour tous les moments partagés avec vous que ce soit au CBGP ou à l'extérieur. Merci aussi à Vincent, Maeva, Orianne, Leyli, Julie, Nath, Gwen, Antoine, Robin, Pierre, Allan, Nico L., Hossein, Eric Pierre et de façon générale tous les collègues et amis du CBGP pour votre accueil et votre gentillesse, vous avez tous contribué à ce que ces trois années de thèse soient pour moi un véritable bonheur.

Je tiens ensuite à remercier tout spécialement deux personnes. Hannah, tu as toujours le sourire sur le tatami et en dehors. Tu sais toujours me redonner la pêche avec cette énergie inépuisable qui te caractérise. Et Gaëlle ("Miss Häagen-Dazs") qui supporte ma bêtise depuis quelques années déjà (comme si tu n'avais pas assez de Fab). Un énorme merci à vous deux qui avez pris le temps de la relecture de ce travail. Je me dois aussi de remercier les personnes qui n'ont pas directement participé à cette thèse mais qui ont fait partie de mon quotidien et de mon équilibre. En commençant par la famille judo, merci à tous les membres du Jita Kyoei, et plus particulièrement à Lulu, Jeannot et Dylan. Vous allez tous me manquer. Merci à l'ensemble des ami(e)s creusois qui ont toujours été là. Avec mes principaux locataires, Florent et Rachel (on se souviendra de la baignade sous l'orage) ainsi que Pierro. Marie, que dire de toi qui est toujours là pour m'écouter, me conseiller et me remotiver dans les moments les plus difficiles. Sans parler de tous nos fous rires, je n'ai qu'une chose à dire : "Youpi rintintin". Et toi Fab, mon plus vieil et fidèle ami, déjà plus de 20 ans, je n'aurais pas de mots pour ça. Un grand merci aussi aux amis Bretons et plus spécialement à la promo MODE. Simon, tu sais à quel point tu comptes pour moi mon pote, ta philosophie du "t'inquiète" et ta technique "du Ratel" m'ont beaucoup aidé ;)

Enfin, je ne remercierai jamais assez ceux à qui je dédie cette thèse, mes parents pour m'avoir toujours laissé faire ce qui me plaît et soutenu dans mes choix. Et toi frangin, qui est toujours présent même si je suis parfois "une

huître" comme tu dis. Si tu daignes lire cette thèse, tu remarqueras très vite à ton grand désarroi qu'elle ne parle pas de vers de terre.

Table des matières

I Document de synthèse **1**

Introduction Générale **3**

1 Estimation du F_{ST} à partir de données Pool-seq **29**

1.1 Introduction 29

1.2 Modèle 31

1.3 Matériels et Méthodes 35

1.3.1 Étude par simulations 35

1.3.2 Autres estimateurs 38

1.3.3 Analyses des données Ind-seq 40

1.3.4 Exemple d'application : *Cottus asper* 40

1.4 Résultats 41

1.4.1 Comparaison d'estimations de F_{ST} Ind-seq et Pool-seq 41

1.4.2 Comparaison de différents estimateurs de F_{ST} pour données Pool-seq 43

1.4.3 Robustesse à des tailles de pool et à des couvertures variables 50

1.4.4 Robustesse aux erreurs de séquençage et d'expérimentation 50

1.4.5 Exemple d'application 55

1.5 Discussion 56

1.5.1 Analyse de variance et probabilités d'identité 57

1.5.2 Comparaison avec des estimateurs alternatifs 59

1.5.3 Application pour les études d'écologie évolutive 59

1.5.4 Limites du modèle et perspectives 60

2 Un modèle hiérarchique bayésien pour détecter l'adaptation locale à partir de données haplotypiques 63

2.1 Introduction 63
2.2 Description du modèle SelEstim multiallélique 71
2.2.1 Le critère de décision KLD pour l'identification des marqueurs potentiellement sous sélection. 76
2.3 Matériels et Méthodes 80
2.3.1 Échantillonneur par MCMC 80
2.3.2 Simulations 81
2.3.3 Comparaisons aux modèles FLK et HapFLK 83
2.3.4 Évaluation des performances de SelEstim sur la base des simulations 84
2.4 Résultats 86
2.4.1 Évaluation des performances de $\text{SelEstim}_{\text{HAP}}$ 86
2.4.2 Evaluation de la robustesse aux écarts des hypothèses du modèle 89
2.5 Discussion 93
2.5.1 Apport des données haplotypiques au modèle SelEstim 96
2.5.2 Les stratégies de groupement local des haplotypes . . . 96
2.5.3 Les balayages sélectifs faibles, le talon d'Achille du *scan* génomique 98
2.5.4 Conclusion 99

3 Intégration d'une variable bayésienne auxiliaire et d'un modèle de lissage intégré 101

3.1 Introduction 101
3.2 Description du modèle SelEstim Auxiliaire 102
3.2.1 Ajout de la variable bayésienne auxiliaire 102
3.2.2 Intégration d'un modèle d'auto-corrélation spatiale de Ising & Potts 107
3.3 Matériels et Méthodes 109

3.3.1 Échantillonneur par MCMC . . . 109
3.3.2 Simulations . . . 109
3.3.3 Comparaison des critères de décision KLD et BF, et évaluation des performances de SelEstim$^{\text{Ising}}_{\text{SNP}}$ sur la base des simulations . . . 110
3.4 Résultats . . . 111
3.4.1 Comparaison des critères de décision BF et KLD . . . 111
3.4.2 Évaluation des performances de SelEstim$^{\text{Ising}}_{\text{SNP}}$ . . . 114
3.5 Discussion . . . 115
3.5.1 Le choix du critère de décision . . . 115
3.5.2 Exploiter l'information de liaison en absence de données haplotypiques . . . 119

Conclusion **121**

Bibliographie **125**

II Annexes **147**

A Article 1 : Measuring Genetic Differentiation from Pool-seq Data **149**

B Un modèle hiérarchique bayésien pour détecter l'adaptation locale à partir de données haplotypiques **191**
B.1 Parameters Update . . . 192
B.1.1 $\mathbf{p_{ijk}}$ . . . 192
B.1.2 M_i . . . 194
B.1.3 $\pi_{\mathbf{j}}$ . . . 195
B.1.4 $\kappa_{\mathbf{ij}}$ . . . 196
B.1.5 σ_{ij} . . . 197
B.1.6 δ_j . . . 198
B.1.7 λ . . . 199
B.2 PODs algorithm . . . 200

C Intégration d'une variable bayésienne auxiliaire et d'un modèle de lissage intégré **203**

C.1 Parameters Update 204

C.1.1 Update of $\delta_j^\star$ 204

C.1.2 Update of σ_{ij} 204

C.1.3 Update of κ_{ij} 204

C.1.4 Update of ζ_j (when $b_{Is} = 0$) 204

C.1.5 Update of ζ_j (when $b_{Is} > 0$) 205

C.1.6 Update of P 205

Résumé/Abstract **206**

Liste des Figures

1 *Single Nucleotide Polymorphism* (SNP) 7
2 Représentation graphique de la mesure de différenciation génétique entre populations 9
3 Représentation des différents modèles de génétique des populations. 10
4 Distribution stationnaire des fréquences alléliques dans un modèle en îles (Wright 1931) selon un processus de diffusion . . . 12
5 Illustration de l'estimation de paramètre par maximum de vraisemblance et de distribution *a posteriori* dans un modèle bayésien 14
6 Représentation graphique d'individus haploïdes issus de différentes populations (dèmes) avec et sans sélection dans des environnements contrastés 17
7 Distribution stationnaire des fréquences alléliques dans un modèle en îles (Wright 1931) selon un processus de diffusion avec sélection 20
8 Comparaison entre séquençage individuel et Pool-seq 22
9 Haplotypes 24
10 Représentation de la signature d'un balayage sélectif sur le génome 26
11 Effet de l'auto-stop génétique 27

1.1 Représentation graphique des probabilités d'identité par état pour des données individuelles et Pool-seq. 30
1.2 Estimations de F_{ST} locus-spécifiques 42

1.3 Estimateurs de F_{ST} par paires 46
1.4 Surface de densité RMSE des F_{ST} Pool-seq sans erreurs de séquençage 48
1.5 Estimations de F_{ST} multilocus en fonction du nombre de dèmes échantillonnés. 49
1.6 Estimateur de F_{ST} global 52
1.7 Biais et précision des estimations de F_{ST} avec erreurs de séquençage et expérimentale. 53
1.8 Biais et précision des estimations de F_{ST} avec et sans filtre MRC. 54
1.9 Analyse du jeu de données Pool-seq de chabots piquants (*Cottus asper*). 55

2.1 Distributions de F_{ST} locus-spécifiques et mesurés le long de fenêtres de 5Mb dans le génome humain. 66
2.2 Comparaison de *scans* génomiques réalisés sur des données de moutons Nord européens avec FLK et HapFLK. 69
2.3 Exemple de groupements locaux d'haplotypes réalisés par `fastphase`. 70
2.4 Illustration de l'algorithme de regroupement local des haplotypes par blocs. 71
2.5 DAG de SelEstim$_{HAP}$. 75
2.6 Illustration de la KLD 77
2.7 Distribution de centrage et distributions *a posteriori* de δ_j pour des marqueurs neutres et sous sélection 78
2.8 Exemple de *scans* pour un jeu de données simulé, résultant d'une analyse avec SelEstim$_{SNP}$ et d'une analyse avec SelEstim$_{HAP}$ 80
2.9 Représentations graphiques du protocole de simulation pour les différents scénarios démographiques considérés. 82
2.10 Représentation graphique du biais de position et de la précision pour un signal. 85
2.11 Puissance en fonction de l'erreur de type I dans un modèle en îles 87

2.12 Distributions des estimations de σ_{ij} obtenus dans un modèle en îles. 88
2.13 Distribution du biais de position dans un modèle en îles. . . . 90
2.14 Distributions de la précision du signal dans un modèle en îles. 91
2.15 Puissance en fonction de l'erreur de type I dans un modèle d'arbre en étoile et d'arbre structuré hiérarchiquement. 92
2.16 Distribution du biais de position dans un modèle de dérive pure selon un arbre en étoile et avec populations structurées. . 94
2.17 Distribution de la précision dans un modèle de dérive pure selon un arbre en étoile et avec populations structurées. 95

3.1 Graphe orienté acyclique (DAG) du modèle SelEstim AUX . 103
3.2 Distribution de la densité de probabilité d'une loi Beta(0.2, 1.8).105
3.3 DAG de SelEstim avec modèle de lissage intégré de Ising . . 108
3.4 Exemple de *scan* KLD et BF 112
3.6 Graphique de puissance pour $\text{SelEstim}_{\text{SNP}}$, $\text{SelEstim}_{\text{HAP}}$ et $\text{SelEstim}^{\text{Ising}}_{\text{SNP}}$ 115
3.7 Distribution du biais de position et de la précision pour $\text{SelEstim}_{\text{HAP}}$, $\text{SelEstim}_{\text{SNP}}$, et $\text{SelEstim}^{\text{Ising}}_{\text{SNP}}$ 116
3.8 Illustration du phénomène de saturation des BF. 118

Liste des Tableaux

1.1 Résumé des notations principales du chapitre 1 32
1.2 Description des estimateurs de F_{ST} utilisés dans le texte. . . . 38
1.3 Liste bibliographiques d'articles Pool-seq mesurant des F_{ST} en 2016 et 2017 43
1.4 F_{ST} global estimé à partir de plusieurs pools 51

2.1 Résumé des notations principales du chapitre 2 73

3.1 Règle de Jeffreys et interprétation des différentes valeurs de BF 107
3.2 Résumé des notations liées aux mesures de performance . . . 110
3.3 Temps de calculs moyen des différents modèles SelEstim . . 114

Première Partie

Document de synthèse

Introduction Générale

Principes généraux de génétique des populations

Si Darwin, à travers des observations morphologiques, a mis en évidence le principe d'adaptation par la sélection naturelle dans son livre l'*Origine des Espèces* (Darwin 1859), il ne connaissait pas les mécanismes sous-jacents à la transmission des caractères ni leur "architecture génétique", c'est-à-dire l'ensemble des gènes impliqués dans leur expression. Il faudra attendre la redécouverte au début du 20ème siècle des travaux de Gregor Mendel (1822-1884) sur les lignées de pois pour poser les bases de ce que l'on connaît aujourd'hui sous le nom de la transmission mendélienne, selon laquelle un individu reçoit deux gènes provenant de chacun de ses parents.

La génétique des populations est née, à la suite de cette redécouverte au début du 20ème siècle, des travaux théoriques de son trio fondateur composé de S. Wright, R. Fisher et J.B.S Haldane. Cette discipline vise à expliquer l'origine et le maintien de la diversité génétique dans les populations, en étudiant l'évolution des fréquences alléliques sous l'influence des différentes forces évolutives : la mutation génératrice de polymorphisme génétique, les flux de gènes (la migration) qui peuvent aider au maintien de ce polymorphisme en introduisant des variants polymorphes au sein d'une population, la dérive génétique qui tend à éroder la diversité génétique et la sélection naturelle qui peut réduire (sélection directionnelle) ou bien maintenir le polymorphisme (sélection balancée).

La synthèse du principe de sélection naturelle de Darwin, de la trans-

mission mendélienne des caractères ainsi que des travaux théoriques en génétique des populations s'est traduite par la mise en place de la synthèse néo-Darwinienne, ou théorie synthétique de l'évolution.

Longtemps deux courants de pensée se sont affrontés. D'un côté l'école sélectionniste admettant que la sélection naturelle est la force évolutive majeure expliquant la diversité au sein et entre les espèces, négligeant ainsi l'effet des autres forces évolutives. Cette école était elle-même scindée en deux groupes : ceux supportant l"hypothèse classique" qui suppose que la diversité génétique est expliquée par la sélection directionnelle et ceux supportant l'"hypothèse de sélection balancée" qui suppose que la sélection équilibrante avantageant les individus polymorphes est majoritaire. À l'inverse, l'école neutraliste menée par Motoo Kimura (Kimura et Crow 1964) considère que la variation génétique peut être principalement expliquée par la dérive génétique (et donc le hasard en introduisant un aléa dans le système) tout en n'écartant pas l'importance ponctuelle de la sélection naturelle qui devient ainsi un facteur évolutif parmi d'autres.

Il y avait donc à l'époque une vraie nécessité pour les empiristes de pouvoir caractériser le polymorphisme génétique afin de comparer la théorie à l'observation.

Les marqueurs génétiques

Le génome d'un individu correspond à l'ensemble du matériel génétique porté par son ADN. Nous avons deux types de génomes chez les métazoaires, le génome nucléaire et le génome mitochondrial. Les plantes ont quant à elles un génome supplémentaire avec le génome chloroplastique. Le génome d'un individu est composé d'un nombre n de chromosomes représentant le support de l'hérédité. Les chromosomes peuvent être présents en plusieurs copies en fonction des espèces, c'est ce que l'on appelle le degré de ploïdie. Ainsi un individu diploïde possède deux jeux complets de chromosomes, soit $2n$ chromosomes. Si le génome contient ainsi nombre d'informations issues de nos ancêtres, il est cependant fréquemment remanié par les différents événements de mutation, de recombinaison ou encore par des ré-arrangements

chromosomiques qui modifient les séquences d'ADN. Haasl et Payseur (2016) comparent ainsi le génome à un palimpseste, dont les chromosomes sont le parchemin et les séquences d'ADN les écrits constamment remaniés au cours de l'évolution des espèces et qui remplacent partiellement voire totalement le texte ancien.

Les marqueurs génétiques peuvent être définis comme des variations d'une séquence d'ADN qui peuvent parfois s'exprimer phénotypiquement (p.e., marqueurs de coloration). Dans l'idéal, un marqueur génétique doit être polymorphe, co-dominant, transmis par un modèle simple (p.e., transmission mendélienne) et distribué de manière homogène dans le génome.

Historiquement, le développement des allozymes (des protéines dont la mobilité électrophorétique diffère en fonction de la séquence protéique et donc de la séquence d'ADN qui la code) avec les deux études pionnières de 1966 réalisées chez l'homme (Harris 1966) et chez *D. pseudoobscura* (Lewontin et Hubby 1966) ont permis de mettre en évidence une forte variabilité génétique au sein des populations (Lewontin 1985). Ces résultats faisant état d'une diversité génétique bien supérieure à ce qui avait été prédit, favorisaient l"hypothèse de sélection balancée" plutôt que l'"hypothèse classique" ou l'hypothèse neutraliste. Les allozymes ont cependant rapidement montré plusieurs inconvénients. Tout d'abord, ils ne permettaient qu'une estimation partielle de la variabilité génétique et n'étaient pas en mesure de révéler les variants n'affectant pas la séquence protéique ou ceux n'affectant pas la mobilité de la protéine sur gel. De plus, une autre limitation des allozymes concernaient la difficulté de les caractériser, avec peu de marqueurs utilisables, et une distribution hétérogène le long du génome qui ne permettait pas une étude de la variabilité génétique à l'échelle génomique. Il apparut donc comme évident que l'avenir de la génétique des populations se trouverait dans l'étude directe des variations d'ADN.

Le développement des technologies de séquençage de l'ADN dans les années 1980, suivi des techniques d'amplification par PCR à la fin des années 1980 ont vu l'apparition de deux nouveaux types de marqueurs : les microsatellites et les SNPs. Les microsatellites sont caractérisés par des motifs de quelques nucléotides (2 à 4 généralement) répétés en tandem. Ils sont sou-

vent hautement polymorphes car soumis à des taux de mutation élevés (de l'ordre de $10^{-3)}$ à 10^{-4} par méiose). Ce type de marqueurs est donc très vite devenu populaire et utilisé dans les études de génétique des populations, principalement dans l'étude des processus neutres (inférence de paramètre de migration, caractérisation structuration de la diversité génétique entre différentes populations, etc.). Bien que les microsatellites soient présents en très grande quantité dans le génome et de manière relativement homogène, il a été montré qu'ils sont présents en plus grande proportion dans les régions non codantes (Lawson et Zhang 2006; Metzgar et al. 2000)

À l'heure actuelle, le marqueur le plus populaire est le SNP, pour "Single Nucleotide Polymorphism", qui correspond à un polymorphisme ponctuel dans le génome avec une base montrant une ou plusieurs variations (quatre états possibles au maximum) à l'échelle de l'individu, de la population ou encore de l'espèce (voir Figure 1). La première étude de génétique des populations impliquant des SNPs a été menée par Kreitman (1983), qui utilisa la technique de Maxam et Gilbert (1977) pour reséquencer 11 copies indépendantes de la région du gène *adh* issues de 5 populations de *D. melanogaster*. Le reséquençage d'une région de 2721 paires de bases relevait de l'exploit à l'époque et se traduisit par l'identification de 43 SNPs, dont un seul était responsable des deux variants connus pour l'allozyme, les autres étant des mutations silencieuses.

En 1991, le "Human Genome Project" (HGP) vit le jour. Après 13 ans de travail et un budget de près de 3 milliard de dollars, le premier génome humain fut séquencé à partir de la technologie Sanger (Lander et al. 2001). S'en suivit le développement de nouvelles technologies de génotypage puis de séquençage haut débit (NGS pour "Next Generation Sequencing") qui ont fait l'effet d'une véritable révolution, nous faisant entrer dans l'ère de la génomique des populations (voir la section "Tirer profit des données à l'ère des NGS") . Nous disposons aujourd'hui de 84.7 millions de SNPs caractérisés dans le génome humain (Auton et al. 2015). Le SNP, en étant le marqueur majoritaire dans le génome, distribué à la fois en région codante ou non, est de fait aussi bien adapté à l'étude de phénomènes neutres qu'à l'étude de la sélection naturelle.

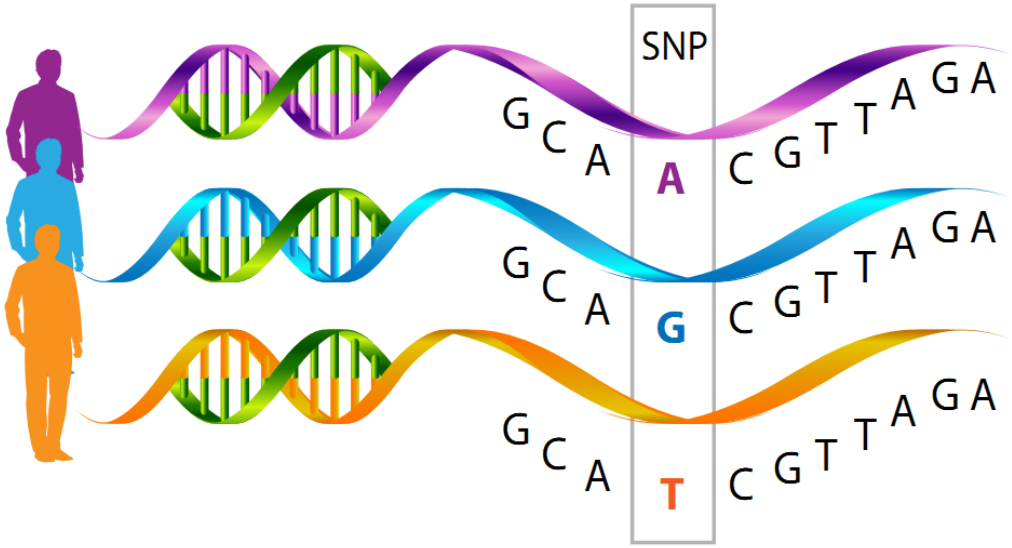


Figure 1 Représentation d'un marqueur "*Single Nucleotide Polymorphism*" (SNP), une paire de bases polymorphes, ici entre trois individus avec trois allèles A, G et T. La figure est issue du site https ://neuroendoimmune.wordpress.com/2014/03/27/dna-rna-snp-alphabet-soup-or-an-introduction-to-genetics/

Caractérisation de la différenciation génétique

L'un des objectifs majeurs de la génétique des populations consiste à caractériser la variabilité génétique entre individus au sein des populations mais aussi entre celles-ci. On s'attend à ce qu'une structuration génétique émerge lorsque les individus forment des populations, c'est-à-dire des groupes distincts organisés spatialement ou temporellement, au sein desquels ils se reproduisent préférentiellement ou exclusivement.

Nous nous placerons dans le cas particulier où une population est subdivisée en sous-populations spatialement distinctes ou dèmes.

Afin de caractériser la variabilité génétique au niveau des populations, G. Malécot et S. Wright ont introduit, à la fin des années 1940 et au début des années 1950, les statistiques F (Malécot 1948; Wright 1951). Les statistiques F ont pour but de décomposer la variabilité génétique et d'exprimer la corrélation entre des paires de gamètes ou de gènes échantillonnés aléatoirement à un niveau de subdivision de la population, relativement à la corrélation ob-

servée à un niveau supérieur. Ainsi, pour une population divisée en dèmes et composée d'individus diploïdes, la corrélation entre deux gènes échantillonnés au sein des individus (I) relativement à la sous-population (S) est exprimée par le F_{IS}. La corrélation entre deux gènes échantillonnés au sein des individus (I) relativement à la population totale (T) est exprimée par le F_{IT}. Enfin, la corrélation entre deux gènes échantillonnés au sein d'une sous-population (S) relativement à la population totale (T) est exprimée par le F_{ST}. Le F_{ST} peut aussi s'interpréter comme la proportion de variation génétique totale expliquée par la variation génétique entre les sous-populations, et donc à l'étendue de la différenciation entre celles-ci (Bhatia et al. 2013; Holsinger et Weir 2009). Nous nous consacreront particulièrement au F_{ST}. Le F_{ST} peut être traduit de plusieurs façons d'un point de vue paramétrique. Par exemple en terme de corrélation intra-classe de probabilités d'identité par état (IIS) entre paires de gènes pris au sein (Q_1) ou entre (Q_2) les sous-populations (Rousset 2007) (voir figure 2).

$$F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2} \tag{1}$$

Ainsi pour un locus, si l'ensemble des individus sont identiques dans l'ensemble des dèmes, aucune structure de population n'existe ($Q_1 = 1$ et $Q_2 = 1$) et le F_{ST} sera indéfini. À l'inverse, dans un cas extrême où tous les individus issus d'un dème sont identiques mais différents entre dèmes ($Q_1 = 1$ et $Q_2 = 0$), une structuration génétique forte existe entre les sous-populations et $F_{\mathrm{ST}} = 1$.

Un intérêt du F_{ST} est qu'il peut s'exprimer sous forme de paramètres démographique dans des modèles de génétique des populations.

Ainsi, sous l'hypothèse où deux populations de taille efficace N_{e} ont divergé depuis t générations pour évoluer selon un modèle de pure dérive génétique (sans mutation, ni migration, ni sélection, voir la Figure 3-B), il a alors été montré que l'ont peut relier le F_{ST} au temps de divergence t par la formule $F_{\mathrm{ST}} \approx 1 - \exp(-t/2N_{\mathrm{e}})$ (Reynolds et al. 1983).

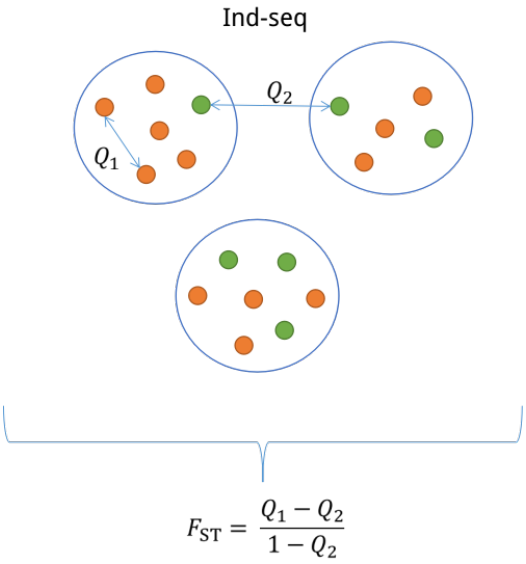


Figure 2 Représentation graphique de la mesure de différenciation génétique entre populations. On considère 3 dèmes avec des individus haploïdes dont les génotypes individuels à un locus biallélique sont représentés en orange et vert. Le F_{ST} est défini comme un ratio de probabilités d'identité par état entre paires de gènes pris au sein (Q_1) ou entre les populations (Q_2).

Le F_{ST} est également un paramètre central en inférence de la dispersion. Wright a montré que sous l'hypothèse d'une structuration des populations selon un modèle en îles (voir la Figure 3-A) alors $F_{\mathrm{ST}} \approx 1/\left(1 + 4N_{\mathrm{e}}m\right)$ (où N_{e} et m désignent respectivement l'effectif diploïde efficace de chaque île et le taux d'échange des migrants entre les îles), si le nombre d'îles tend vers l'infini et que l'on néglige la mutation (Wright 1931).

Il est cependant reconnu que cette relation est très simpliste et peu réaliste au vu des hypothèses sous-jacentes (Whitlock et McCauley 1999), notamment dans le cas où la migration n'est pas symétrique entre populations et où leur isolement géographique se traduit par une augmentation de la différenciation génétique. L'isolement par la distance (IBD, voir la Figure 3-C) (Rousset 1997) fut décrit pour la première fois par Wright (1943) puis le modèle fut dérivé par Malécot (1967). Wright (1943) voit l'IBD comme un phénomène écologique à partir duquel une structuration génétique peut apparaître. Il suppose ainsi qu'un individu a plus de chance de se reproduire localement que s'il doit disperser à grande distance. Contrairement à Wright (1943),

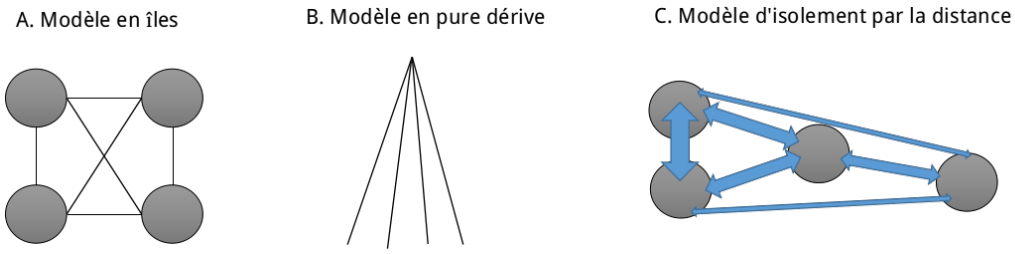


Figure 3 Représentation des différents modèles de génétique des populations. A) Le modèle en îles de Wright (1931) dans lequel les différents dèmes échangent des migrants entre eux avec le même taux de migration. B) Le modèle en dérive pure dans lequel des populations (quatre ici) divergent et évoluent de manière indépendante sans migration. C) Le modèle d'isolement par la distance dans lequel les dèmes échangent des migrants entre eux avec une intensité de migration (épaisseur de la flèche) d'autant plus importante qu'ils sont géographiquement proches.

Malécot (1948) voit l'IBD comme un patron de génétique des populations à partir duquel deux individus voient leur distance génétique augmenter avec la distance géographique qui les sépare (Ishida 2009). Ainsi le modèle IBD montre une différence majeure avec le modèle en îles de Wright (1931), qui considère que les populations échangent des migrants entre elles avec le même taux de migration, tout en négligeant leur positionnement géographique. Avec le modèle IBD, la distance génétique entre les populations devient ainsi fonction de la distance géographique qui les séparent. Il en résulte une relation entre les rapports $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$ pour l'ensemble des paires de populations et la distance géographique (Rousset 1997; Slatkin 1993) (le logarithme de cette distance) pour des populations évoluant dans un système linéaire en une dimension (respectivement pour un système en deux dimensions). On peut exprimer la pente de la régression comme $b = 1/(4D\pi\sigma^2)$ avec D la densité effective d'individus et σ^2 la distance quadratique axiale de dispersion ou plus simplement la distance moyenne séparant les parents de leurs descendants.

On perçoit donc l'intérêt de pouvoir estimer les statistiques F, et dans notre cas le F_{ST}, que ce soit dans un but de description de la structure génétique des populations ou d'inférence de paramètres démographiques. Dans les faits, l'estimation du F_{ST} à partir de données de comptages alléliques est

devenue une pratique courante et a été envisagée selon plusieurs méthodes.

Méthodes d'estimation du F_{ST}

Estimation par la méthode des moments

Un estimateur populaire du ratio de probabilités d'identités par état (IIS) Q_1 et Q_2 a été développé par la méthode des moments dans un cadre d'analyse de variance (Cockerham et Weir 1987; Rousset 2007; Weir et Goudet 2017). La variance est dans ce cas décomposée selon différents niveaux hiérarchiques : au sein des individus dans une sous-population, entre individus au sein d'une sous-population et entre individus provenant de sous-populations différentes.

L'estimation du F_{ST} par la méthode des moments a plusieurs avantages. Tout d'abord, elle montre un biais très faible, c'est-à-dire que si l'on réechantillonne un certain nombre de fois les mêmes populations, le $\hat{F}_{\mathrm{ST}}$ multilocus estimé va être très proche de la véritable valeur du paramètre F_{ST}. Enfin, cette méthode est computationnellement très efficace et ne repose sur aucun modèle de génétique des populations, ne faisant ainsi aucune hypothèse concernant la distribution des fréquences alléliques échantillonnées tant qu'elle dispose d'une moyenne et d'une variance (Holsinger et Weir 2009).

Si les F_{ST} sont classiquement estimés par la méthode des moments, il a aussi été proposé de les estimer par maximisation de la vraisemblance ou dans un cadre bayésien, en reposant dans les deux cas sur une modélisation de la distribution des fréquences alléliques.

Modélisation des fréquences alléliques et estimation du F_{ST}.

Sous l'hypothèse d'un système à l'équilibre entre les différentes forces évolutives, la distribution des fréquences alléliques peut être modélisée par une distribution stationnaire de densité de probabilité (Rannala 2013).

Par exemple, sous l'hypothèse d'une population de taille efficace diploïde N_{e}, issue d'un modèle en îles (Wright 1931) qui reçoit des migrants à un taux m. Si on néglige la sélection et la mutation, alors on peut modéliser

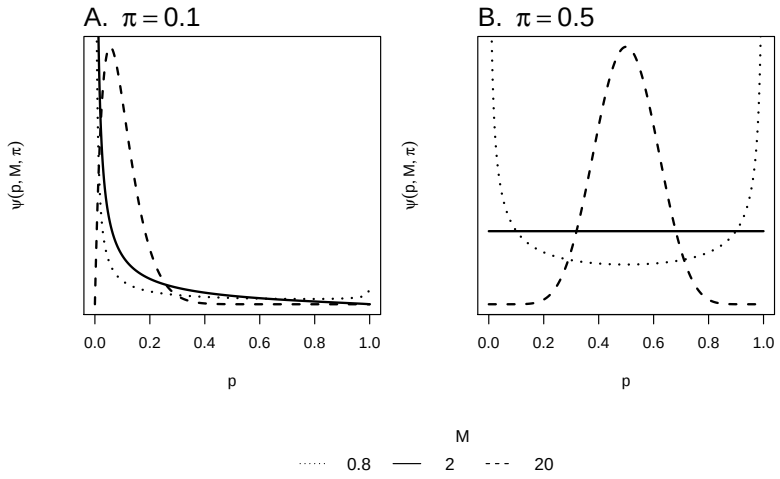


Figure 4 Distribution stationnaire des fréquences alléliques dans un modèle en îles (Wright 1931) selon un processus de diffusion (voir Equation 2). Les résultats sont présentés pour des fréquences allélique dans le pool de migrant π de 10% (A) et de 50% (B), ainsi que trois intensités de migration M=(0.8,2,20)

l'effet conjoint de la migration (qui va tendre à homogénéiser les fréquences alléliques) et de la dérive par la distribution stationnaire d'un processus de diffusion (Wright 1940) pour aboutir à la distribution stationnaire des fréquences d'un allèle donné (où $M = 4N_e m$ représente le taux de migration mis à l'échelle et π la fréquence de l'allèle de référence dans le pool de migrant) :

$$\psi(p, M, \pi) \quad = \quad \frac{\Gamma(M)}{\Gamma(M\pi)\Gamma(M(1-\pi))} p^{M\pi-1}(1-p)^{M(1-\pi)-1} \tag{2}$$

Ainsi, sous l'hypothèse d'un modèle en îles (Wright 1931), la distribution des fréquences alléliques p à un locus pour un allèle A, dans une population à l'équilibre sans sélection, est exprimée sous la forme d'une distribution beta de paramètre M et π (voir Figure 4). Rappelons que sous l'hypothèse du modèle de génétique des populations sous-jacent, $F_{\mathrm{ST}} = 1/(1+M)$.

L'estimation du F_{ST}, que ce soit par maximum de vraisemblance ou dans

un contexte bayésien, requiert également la spécification d'une distribution d'échantillonnage à partir de laquelle les comptages alléliques sont observés. Par exemple, dans le cas de comptages à un locus biallélique, de type SNP, issus d'un échantillonnage aléatoire au sein d'un dème dans un modèle en îles (Wright 1931), si l'allèle de référence est en fréquence p dans la population et qu'un échantillon de taille haploïde n est utilisé, on peut simplement définir la probabilité d'observer x comptages de l'allèle de référence par une distribution binomiale (ce qui revient à faire l'hypothèse que le dème est à l'équilibre de Hardy-Weinberg) :

$$\mathrm{P}(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3}$$

En intégrant sur l'ensemble de la distribution des fréquences alléliques, la probabilité d'observer x suit alors une distribution dite beta-binomiale :

$$\begin{aligned} \mathrm{P}(x) &= \int \mathrm{P}(x|p)\psi(p, M, \pi) \\ &= \frac{\Gamma(M)}{\Gamma(M+n)} \binom{n}{x} \frac{\Gamma(M\pi + x)\Gamma(M(1-\pi) + (n-x))}{\Gamma(M\pi)\Gamma(M(1-\pi))} \end{aligned} \tag{4}$$

La vraisemblance des données est ici exprimée pour un seul dème et un locus. Dans le cas de plusieurs dèmes et plusieurs locus, on multiplie simplement les vraisemblances pour chaque dème à chaque locus, ce qui revient à supposer une indépendance conditionnelle (échangeabilité) des marqueurs.

En se basant sur le modèle beta-binomial décrit ci-dessus, il est possible d'estimer le F_{ST} par maximum de vraisemblance. Nous pouvons en effet explorer les différentes valeurs possibles des paramètres (M et π) de manière à maximiser la vraisemblance du modèle, et donc estimer $\hat{M}$ et par conséquent $\hat{F}_{\mathrm{ST}} = 1/(1+\hat{M})$ sous l'hypothèse d'un modèle en îles (Wright 1931) (voir Figure 5-A). Différents modèles, reposant sur d'autres hypothèses ont également été développés. Il a ainsi été proposé d'estimer le F_{ST} par maximum de vraisemblance en supposant que les fréquences alléliques sont distribuées

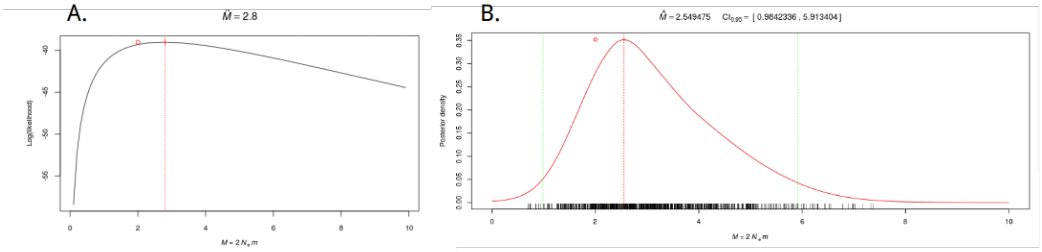


Figure 5 A) Estimation du paramètre $M \equiv 1/F_{\mathrm{ST}} - 1$ par maximum de vraisemblance et B) Estimation de la distribution *a posteriori* du paramètre par MCMC dans le F-model, en supposant une distribution *a priori* uniforme de $M \sim \mathrm{U}(0, 10)$, et une distribution uniforme de $\pi \sim \mathrm{U}(0, 1)$. La chaîne MCMC est composée d'un burn-in de 5000 itérations suivie de 5000 autres itérations. Dans les deux cas, les comptages alléliques ont été simulé pour un locus sous le modèle d'inférence dans 10 dèmes avec $M = 2$, $\pi = 0.5$ et une taille haploïde d'échantillons de 50 individus par dème. Notons qu'on simule ici des données haploïdes, d'où $M \equiv 2N_{\mathrm{e}}m$. Sur les différents graphes, la valeur estimée du paramètre est représentée par une croix rouge tandis que la valeur simulée est indiquée par un cercle rouge.

selon une gaussienne (Weir et Hill 2002).

Les approches par maximum de vraisemblance souffrent cependant d'une limitation majeure. Il est en effet parfois très difficile, voire impossible, de maximiser la vraisemblance du modèle. Que ce soit à cause d'un trop grand nombre de paramètres ou encore de l'impossibilité d'obtenir une expression algébrique de la vraisemblance. Pour pallier à ce problème, d'autres approches ont été proposées dans un contexte bayésien.

Avec les modèles bayésiens, nous ne cherchons pas à maximiser la probabilité d'observer les données sachant les paramètres du modèle, comme c'est le cas avec l'approche par maximum de vraisemblance, mais nous cherchons à estimer la distribution de probabilité des paramètres connaissant les données. L'approche bayésienne représente donc une philosophie bien différente. Plus précisément, dans un modèle bayésien, nous associons une distribution de probabilité *a priori* à chaque paramètre du modèle et nous cherchons à échantillonner des valeurs de paramètres dans leur distribution *a posteriori* (c'est à dire conditionnellement aux données et aux autres paramètres, selon le modèle considéré). Pour un paramètre Θ auquel on associe une distribution *a priori* $\mathrm{P}(\Theta)$, sa distribution *a posteriori* sachant les données D est donnée

par le théorème de Bayes :

$$\begin{aligned} \mathrm{P}(\Theta|D) &= \frac{\mathrm{P}(D|\Theta)\mathrm{P}(\Theta)}{\mathrm{P}(D)} \\ &\propto L(\Theta; D)P(\Theta) \end{aligned} \tag{5}$$

La distribution *a posteriori* du paramètre est donc proportionnelle au produit de la vraisemblance $L(\Theta; D)$ du modèle et à la distribution *a priori* du paramètre. L'idée va donc être d'échantillonner directement dans la distribution *a posteriori* des paramètres au moyen, par exemple, de procédures MCMC. Parmi les modèles qui ont été développés, le modèle F (Falush et al. 2003; Gaggiotti et Foll 2010) se base sur une distribution beta-binomiale des comptages alléliques (voir Équation 4). Les données de comptages alléliques x_{ij} de l'allèle de référence au locus j dans le dème i dépendent des fréquences alléliques p_{ij} inconnues, elles-mêmes fonction des hyper-paramètres π_j et F_{ST_j}. Un exemple d'estimation de la distribution *a posteriori* du paramètre M est donné dans la Figure 5-B.

Détecter la sélection à partir de la différenciation génétique

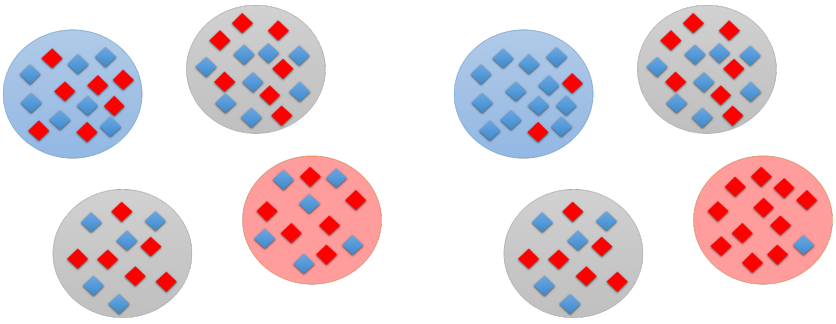
J'ai jusqu'à présent volontairement mis l'accent sur le lien entre le F_{ST} et des processus neutres, tels que la migration, en négligeant l'étude de la sélection naturelle. Or, celle-ci est au centre des études de génétique des populations et de biologie évolutive en général. Il a aussi été proposé de caractériser l'hétérogénéité des F_{ST} estimés entre marqueurs afin d'identifier les locus qui sont soumis à la sélection (Akey et al. 2002; Beaumont 2005; Beaumont et Nichols 1996; Cavalli-Sforza 1966; Lewontin et Krakauer 1973; Vitalis et al. 2001; Weir et al. 2005). Si la pression de sélection s'exerce directement au niveau du caractère, on peut s'attendre à ce que les loci impliqués dans sa variabilité génétique voient leur fréquence allélique affectée (voir Figure 6).

Nous utiliserons dès à présent le terme d'adaptation locale pour désigner le processus par lequel un caractère est soumis à la sélection naturelle, et dont la valeur sélective (c'est-à-dire le nombre moyen de descendants viables et fertiles fournis par les individus porteurs de ce caractère à la génération suivante) associée à son environnement local est plus élevée que dans n'importe quel autre environnement (Hoban et al. 2016; Kawecki et Ebert 2004; Savolainen et al. 2013). Nous devons différencier l'adaptation locale qui peut-être due à une plasticité phénotypique (un même génotype pouvant exprimer plusieurs phénotypes en fonction de l'environnement) de l'adaptation locale au sens génétique du terme (lorsque le phénotype avantageux va être déterminé majoritairement par une base génétique). On peut en effet exprimer la variance d'un phénotype comme $V(P) = V(G)+V(E)+V(G\times E)$, avec $V(G)$ la part de déterminisme génétique du caractère, V(E) l'effet de l'environnement sur le phénotype et $V(G \times E)$ l'interaction génotype-environnement (Fusco et Minelli 2010). Pour la suite, nous nous intéresserons plus spécifiquement au cas où la variance phénotypique est majoritairement expliquée par la variance génétique et où l'effet de l'environnement est négligeable.

Méthodes empiriques

Nous avons vu que les technologies NGS permettent de caractériser le polymorphisme génétique à une échelle pan-génomique. Akey et al. (2002) ont été les premiers à proposer un *scan génomique* de différenciation chez l'homme pour la recherche de signatures de sélection. Ils ont utilisé 26,530 SNPs issus de trois populations et ont identifié pour chaque chromosome les marqueurs sur-différenciés, c'est-à-dire dans le quantile à 97.5% de la distribution empirique des F_{ST} chromosomes-spécifiques.

Dans la même idée, Weir et al. (2005) ont proposé de mesurer les F_{ST} dans des fenêtres de 5Mb afin de réduire le bruit occasionné par l'utilisation d'une statistique locus-spécifique. Ils ont ensuite identifié pour chaque chromosome, les régions présentant un degré de différenciation supérieur (ou inférieur) au F_{ST} moyen plus ou moins trois fois l'écart-type de la distribution empirique de F_{ST} par chromosome.



a: Neutralité **b:** Adaptation locale

Figure 6 Représentation graphique d'individus haploïdes issus de différentes populations (dèmes) dont les environnements contrastés sont représentés par des couleurs différentes. Les individus sont génotypés à un locus à 2 allèles (rouges ou bleus). A) En l'absence de sélection les génotypes sont distribués selon un équilibre migration/dérive dans les populations. B) Adaptation locale des génotypes rouges (bleus) dans l'environnement rouge (bleu), les deux génotypes sont neutres dans les dèmes gris. On observe que les allèles sous sélection sont en fréquence plus élevée dans les environnements pour lesquels ils sont sélectionnés, augmentant ainsi le degré de différenciation génétique entre les populations, par rapport à l'absence de sélection.

Si ces approches empiriques sont computationnellement très efficaces et ne reposent sur aucun modèle de génétique des populations, elles ne permettent cependant pas de contrôler le degré de fausses découvertes.

D'autres approches ont ainsi été proposées pour détecter la sélection, notamment dans un cadre bayésien.

Méthodes basées sur des modèles de génétique des populations

Afin de contrôler le taux de fausses découvertes, plusieurs approches ont été proposées, à la fois dans des contextes fréquentiste et bayésien, en se basant sur des modèles de génétique des populations.

Historiquement, Lewontin et Krakauer (1973) ont proposé le premier test de détection de la sélection qui mesure à un locus, l'écart de son F_{ST} par rapport au F_{ST} moyen. Pour J locus échantillonnés dans n_d dèmes, ils définissent

le test LK au locus j par :

$$T_{\mathrm{LK}} \quad = \quad F_{\mathrm{st}}^{j} \frac{n_d - 1}{\bar{F}_{\mathrm{st}}} \tag{6}$$

avec $\bar{F}_{\mathrm{st}}$ le F_{ST} moyen sur les J locus. Sous l'hypothèse d'un modèle de génétique des populations en dérive pure (voir Figure 3-B), ils ont montré que le test LK suit une loi du Chi-2 avec $(n_d - 1)$ degrés de liberté. On peut donc utiliser les quantiles de cette distribution sous l'hypothèse de neutralité comme critère de décision afin d'identifier les locus sous sélection. Ce modèle est cependant sensible à la structure hiérarchique des populations (Robertson 1975), ce qui a motivé le développement du modèle FLK (Bonhomme et al. 2010). FLK est une extension du test LK de Lewontin et Krakauer (1973), qui introduit la matrice $\mathcal{F}$ de variance-covariance des fréquences alléliques, ce qui permet de corriger le test pour la structure des populations.

Au delà de ces approches fréquentistes, des modèles bayésiens de *scans génomiques* ont aussi été proposé pour identifier les locus sous sélection. L'idée commune aux différents modèles est de décomposer le F_{ST} en fonction d'effets population-spécifique et locus-spécifique afin d'identifier les locus soumis à la sélection, c'est-à-dire ceux qui montrent un effet locus-spécifique non nul.

L'ensemble de ces modèles se base sur des approximations de diffusion des fréquences alléliques et nombre d'entre eux n'incluent pas explicitement la sélection et supposent que la majorité des marqueurs sont neutres. C'est le cas de Nicholson et al. (2002) qui considèrent un modèle démographique neutre de dérive pure sans mutation, mais aussi de Beaumont et Balding (2004); Foll et Gaggiotti (2008); Gautier et al. (2010); Guo et al. (2009); Riebler et al. (2008), qui représentent des extensions du modèle F (et suppose donc un modèle en îles).

En 2014, Vitalis et al. (2014) ont quant à eux développé SelEstim, un modèle bayésien hiérarchique qui tient explicitement compte de la sélection. En effet, SelEstim utilise comme distribution *a priori* des fréquences alléliques la distribution stationnaire d'un processus de diffusion dans un modèle en îles avec de la sélection.

Considérons un locus j ayant deux allèles, un allèle de référence en fré-

quence p et l'allèle alternatif en fréquence $(1-p)$, dans une population de taille efficace diploïde N_e évoluant sous un modèle en îles (Wright 1931). Appliquons maintenant de la sélection à ce locus avec une intensité s de sorte que les valeurs sélectives des différents génotypes soient :

$$\begin{aligned} w_{AA} &= 1+s \\ w_{Aa} &= 1+s/2 \\ w_{aa} &= 1 \end{aligned} \tag{7}$$

Notons $\sigma = 2Ns$, l'intensité de sélection mise à l'échelle de la population. Si on néglige la mutation ($4N\mu << 1$, avec μ le taux de mutation par génération), on peut modéliser la distribution des fréquences alléliques dans un dème à un locus soumis à la sélection, à la dérive et à la migration par la densité stationnaire d'un processus de diffusion (voir Figure 7) (Barbour et al. 2000; Barton et Rouhani 1987; Donnelly et al. 2001; Wright 1949) tel que :

$$\begin{aligned} \psi(p,\pi,M,\sigma) &= \frac{\Gamma(M)}{\Gamma(M\pi)\Gamma(M(1-\pi))_1F_1(M\pi,M,\sigma)} \\ &\times \exp(\sigma p)p^{M\pi-1}(1-p)^{M(1-\pi)-1} \end{aligned} \tag{8}$$

où ${}_1F_1(a,b,z)$ est la fonction hypergéométrique confluente (Abramowitz et Stegun 1964).

De fait, on considère donc dans le cas de SelEstim, et contrairement aux modèles cités précédemment que l'ensemble des marqueurs est soumis dans une certaine mesure à de la sélection.

Les méthodes bayésiennes contrairement aux approches empiriques permettent de contrôler le taux de fausses découvertes en reposant sur des modèles précis de génétique des populations. L'approche bayésienne permet aussi de s'accommoder des petites tailles d'échantillons en intégrant sur l'incertitude des fréquences alléliques estimées à partir des données de comp-

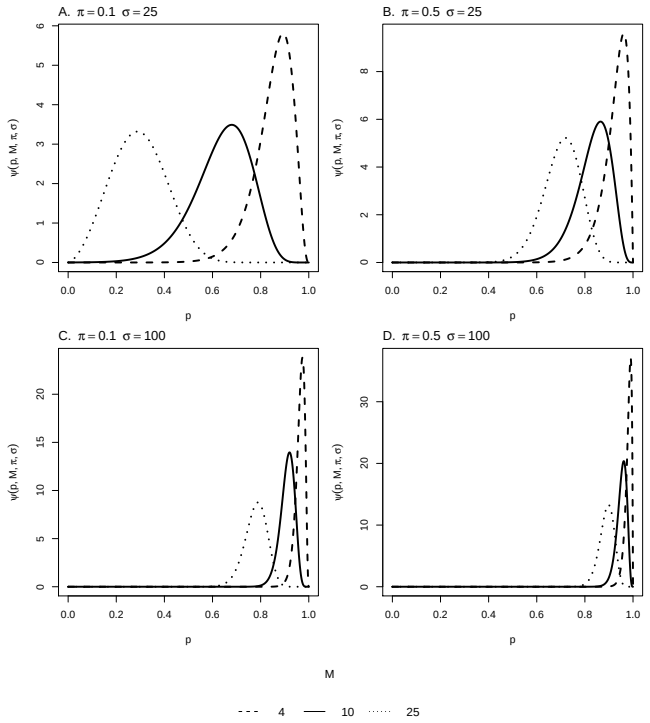


Figure 7 Distribution stationnaire des fréquences alléliques dans un modèle en îles (Wright 1931) selon un processus de diffusion. Les résultats sont présentés pour des fréquences alléliques dans le pool de migrants π de 10% (A) et de 50% (B), ainsi que trois intensités de migration M=(4, 10, 25)

tages. En contre partie elles ont aussi des inconvénient inhérents aux modèles sous-jacents qui peuvent être très restrictifs.

La génétique des populations a donc fait l'objet d'un grand nombre de développements théoriques pour caractériser la structure génétique des populations et étudier les différentes forces évolutives neutres ainsi que la sélection. La dernière décennie a vu l'émergence de nouveaux défis pour les généticiens des populations avec le développement des nouvelles technologies de séquençage NGS. La nature des données issues de ces technologies étant différente de celles citées jusqu'alors, il convient d'en tenir compte dans le modèle afin d'exploiter pleinement les informations qu'elles contiennent.

Tirer profit des données à l'ère des NGS

La révolution NGS

Suite au séquençage du génome humain, au début des années 2000, ce sont commercialisées les nouvelles technologies de séquençage (NGS), ou séquençage haut débit. Ces techniques font en effet preuve d'un débit bien supérieur aux générations précédentes et tout cela à coût réduit, ce qui a rendu beaucoup plus accessibles les études en génétique et à ouvert la porte à l'étude d'espèces non-modèles pour lesquelles les ressources génétiques étaient inexistantes ou lacunaires. Contrairement aux technologies utilisées jusqu'alors, elles permettent de séquencer en parallèle des milliers de fragments d'ADN sous la forme de génomes complets (en fragmentant aléatoirement les copies d'ADN) ou sous forme réduite (en fragmentant l'ADN à des régions bien précises, par exemple flanquant des sites de reconnaissance d'enzymes de restriction), comme c'est le cas des données issues du protocole "Restriction site associated DNA" (Rad-Seq) par exemple (Baird et al. 2008; Davey et al. 2010). Ces technologies ont représenté un tournant dans la nature des données utilisées, qui se présentent dorénavant sous la forme de milliers de courtes séquences d'ADN avec un certain nombre d'erreurs de séquençage. Ces propriétés particulières ont amené à repenser les formats de données utilisées. Elles ont aussi stimulé l'essor de la bio-informatique, avec le développement d'outils permettant de filtrer les données sur la base de leur qualité ainsi que d'aligner les séquences sur un génome de référence lorsqu'il est disponible, ou encore d'assembler les séquences sous forme de séquences continues d'ADN (contigs) dans le cas d'assemblages de-novo. Ces technologies permettent ainsi aujourd'hui de caractériser le polymorphisme génétique à une échelle pan-génomique chez de nombreuses espèces (Andrews et Luikart 2014; Belcaid et Toonen 2014), y compris chez les espèces "non-modèles".

Historiquement, le séquençage impliquait des échantillons marqués individuellement (voir The International HapMap Consortium 2005 pour l'homme par exemple). Dans ce cas précis, chaque lecture de séquençage peut être attribuée à un individu ce qui nous permet de déterminer les génotypes in-

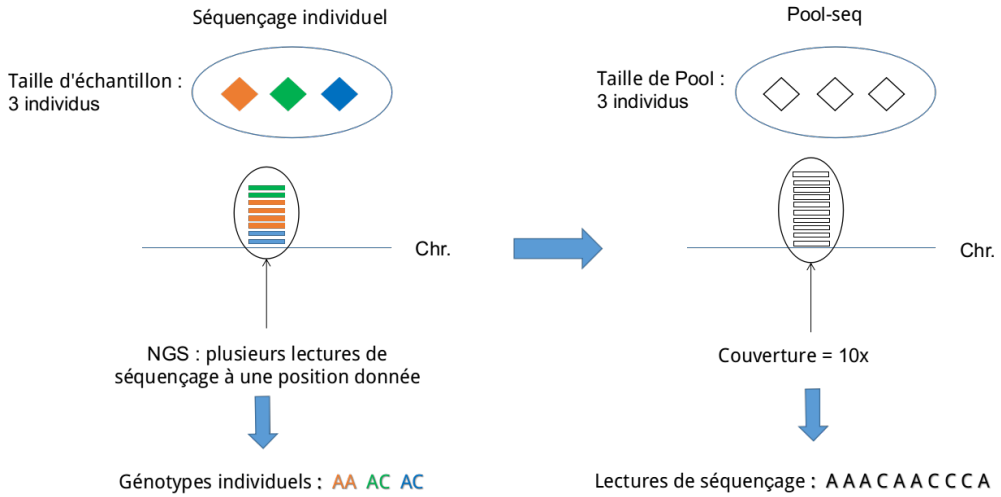


Figure 8 Comparaison entre séquençage individuel et Pool-seq

dividuels à chaque locus (voir figure 8-A). Cette approche, usuelle chez les espèces modèles tel que l'Homme ou encore l'abeille (Wragg et al. 2016) reste cependant trop coûteuse pour de nombreuses espèces non-modèles.

Le séquençage en Pool

Une alternative efficace et plus économique au séquençage individuel consiste à séquencer des mélanges (ou pools) d'ADN d'individus représentatifs de populations données (approches dites *Pool-seq* : voir pour revue, Schlötterer et al. (2014)). Puisque cela consiste à séquencer des librairies d'échantillons d'ADN en pool sans recourir au marquage individuel des séquences, le Pool-seq permet de produire des données génomiques de polymorphisme à moindre coût comparé au séquençage individuel (Schlötterer et al. 2014). C'est par exemple ce type de données qui est majoritairement utilisé au CBGP (Gautier et al. 2018; Leblois et al. 2018; Rohfritsch et al. 2018).

Il existe cependant une différence majeure dans les données issues de séquençage individuel et en Pool-seq. La première nous permet d'obtenir des génotypes individuels et donc des comptages alléliques. En revanche, les données issues de Pool-seq sont des lectures de séquençage que nous ne pouvons pas attribuer aux différents individus dans le pool. Nous travaillons

dans ce cas sur des comptages de lectures (voir Figure 8-B) tout en faisant l'hypothèse d'une homogénéité de la population ce qui revient à négliger la structuration intra-populationnelle.

D'un point de vue technique, des quantités non-équimolaires d'ADN provenant des individus en mélange, ainsi que la variation stochastique de l'efficacité de l'amplification des ADNs individuels ont d'ores et déjà été le sujet de questionnements sur la précision des estimations des fréquences alléliques, particulièrement dans le cas d'une faible profondeur de séquençage et de petites tailles de pools (Anderson et al. 2014; Cutler et Jensen 2010; Ellegren 2014). Cependant, il a aussi été montré qu'à effort de séquençage équivalent, le Pool-seq permet une estimation similaire, si ce n'est plus précise, des fréquences alléliques comparé au séquençage d'individus (Futschik et Schlötterer 2010; Gautier et al. 2013).

À cause des caractéristiques particulières de ces données (issues d'un double processus d'échantillonnage des individus dans les populations puis des individus dans les pools), le Pool-seq a fait l'objet de développements statistiques spécifiques (Ferretti et al. 2013; Futschik et Schlötterer 2010; Gautier et al. 2013; Rode et al. 2017). Néanmoins aucun intérêt particulier n'a été porté à l'estimation de F_{ST} à partir de ces données.

Haplotypes et exploitation de l'information apportée par le déséquilibre de liaison

Une autre caractéristique commune aux données NGS concerne la très forte densité de marqueurs qu'elles génèrent. Bien que classiquement considérés par les modèles mentionnés précédemment comme indépendants, les marqueurs génétiques portés sur un même chromosome ne le sont pas. On définit le déséquilibre de liaison (DL) comme l'association non aléatoire d'allèles à des locus différents (voir Sved et Hill (2018) pour revue). Le DL intervient à l'échelle gamétique lorsqu'il n'y a pas d'échange d'allèles à un locus par recombinaison réciproque entre chromosomes homologues durant la méiose.

Si la sélection affecte directement la mutation causale, elle va avoir un

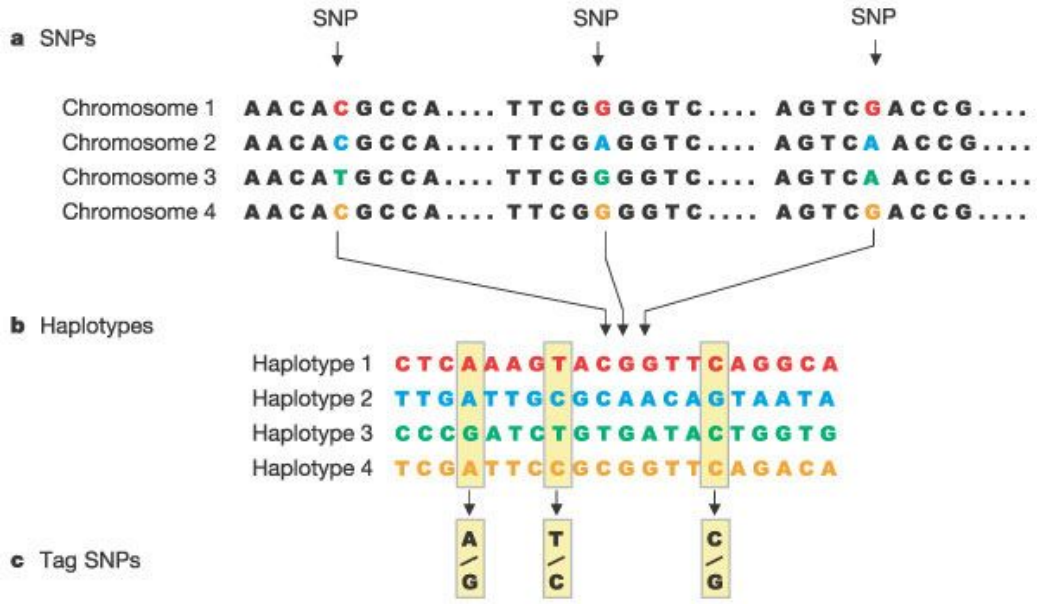


Figure 9 (A) Représentation de marqueurs SNPs identifiés sur quatre chromosomes et (B) des haplotypes correspondant aux associations des SNPs issus de chaque chromosome. Les allèles présents à chaque SNP sont ensuite décrits (C). La figure est tirée de Gibbs et al. (2003).

effet indirect sur les variants neutres, voisins de cette mutation et qui y sont physiquement liés (Maynard Smith et Haigh 1974; Stephan et al. 2006). Ainsi lorsqu'une mutation va rapidement augmenter en fréquence, elle va emmener avec elle toutes les mutations voisines selon un principe appelé auto-stop génétique (Maynard Smith et Haigh 1974; Storz 2005). Cela va avoir pour effet d'augmenter le DL autour de la mutation, avec une intensité de DL d'autant plus élevée que les marqueurs voisins sont physiquement proches et n'ont donc pas le temps, dans leur histoire commune, de recombiner (Maynard Smith et Haigh 1974).

Maynard Smith et Haigh (1974) ont décrit le phénomène de "balayage sélectif", processus par lequel une mutation avantageuse à un locus sous sélection va augmenter en fréquence, réduisant ainsi la diversité génétique neutre des locus adjacents. Un balayage sélectif peut être dit "fort" lorsque la sélection intervient sur une mutation avantageuse rare ou à l'extrême présente sous la forme d'une simple copie dans la population ; ou "faible" (Pennings

et Hermisson 2006) lorsqu'elle agit sur une mutation déjà présente dans la population et qui partage donc différents environnements génétiques (la mutation est portée par différents haplotypes). L'effet d'auto-stop génétique lors d'un balayage sélectif fort est représenté dans la Figure 11.

Les balayages sélectifs créent des patrons caractéristiques de DL au voisinage des régions sous sélection (Maynard Smith et Haigh 1974; Stephan et al. 2006) (voir Figure 10) qui peuvent être exploités dans la recherche de signatures de sélection dans les génomes. Utiliser les fréquences haplotypiques est dans ce cas plus intéressant qu'utiliser les fréquences marginales des différents SNPs : un haplotype est un fragment d'ADN, plus précisément une séquence provenant d'un même chromosome et qui est constituée de plusieurs SNPs. Si l'on considère chaque variant haplotypique (constitué par plusieurs allèles à des SNPs consécutifs) lui-même comme un allèle, on peut aboutir à un marqueur multi-locus hautement polymorphe. Si on note L la taille de l'haplotype en nombre de SNP, que l'on va supposer biallélique, le nombre d'allèles possibles est donc de 2^L, soit 8 allèles maximum pour un haplotype constitué de 3 SNPs. Contrairement à l'utilisation de SNPs que l'on considère classiquement "indépendants", l'utilisation de marqueurs haplotypiques nécessite de connaître la phase de chaque chromosome individuel et donc d'avoir accès à un minimum de ressources génétiques (assemblage du génome, carte génétique). Les "marqueurs" haplotypiques contiennent donc directement l'information de liaison génétique entre les SNPs qui les composent et ouvre ainsi la possibilité de prendre en compte l'information de déséquilibre de liaison dans la recherche de signatures de sélection.

Par exemple, si l'on considère deux locus à deux allèles A, a et B, b tous deux issus de deux populations où les haplotypes AB et ab sont présents en fréquence de 50% dans la population 1 et les haplotypes Ab et aB sont présents dans les mêmes proportions dans la populations 2. L'utilisation des fréquences marginales des allèles, en considérant les SNP comme indépendants, indique que nous avons 50% de A et 50% de B dans chaque population et qu'elles ne sont donc absolument pas différenciées. À l'inverse, considérer les fréquences haplotypiques nous indique que nous avons une forte différenciation génétique entre les deux populations.

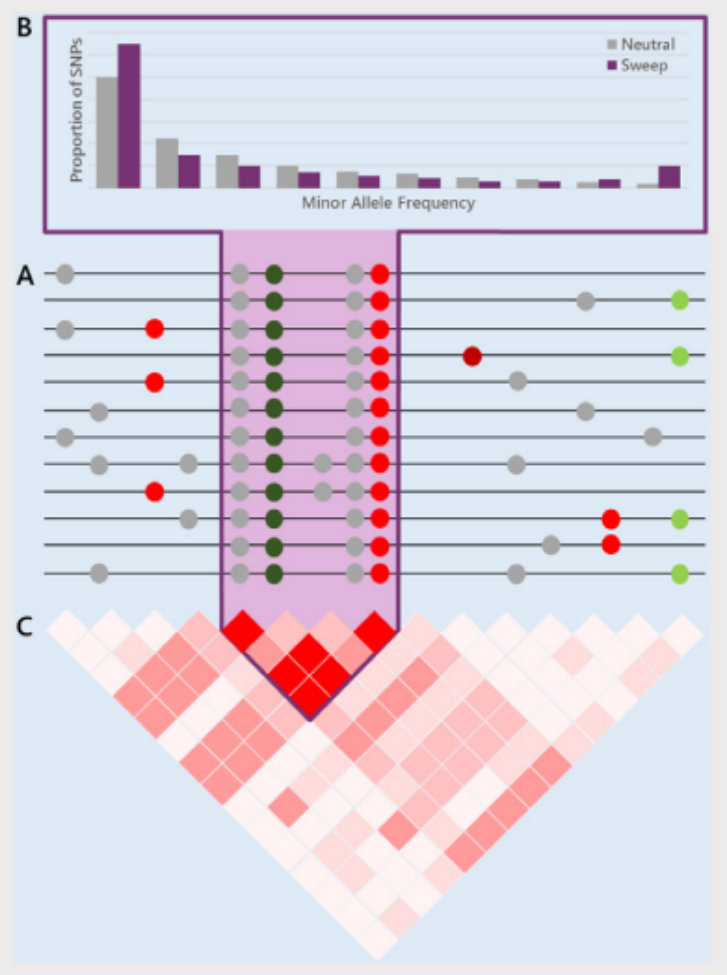


Figure 10 Représentation de la signature d'un balayage sélectif sur le génome (A) par une réduction de la diversité génétique dans la région sous sélection (violette), (B) un excès d'allèles rares et (C) une augmentation du DL dans la région sous sélection. Les différentes pastilles colorées représentent différentes mutations par rapport à leurs valeurs sélectives : en marron, fortement délétères (éliminées par la sélection naturelle) ; en rouge, faiblement délétères ; en gris, des mutations neutres ; en vert clair, des mutations faiblement avantageuses ; et en vert foncé, des mutations très avantageuses. Dans la région du balayage sélectif (violet), une mutation avantageuse a atteint la fixation en emmenant avec elle les marqueurs voisins liés, neutres ou approchant la neutralité. Dans cette région, la diversité génétique est réduite et le polymorphisme est partagé par les différents chromosomes créant un patron de fort DL, tandis que des mutations neutres récentes sont présentes sur deux chromosomes seulement. Cette figure est tirée de Casillas et Barbadilla (2017)

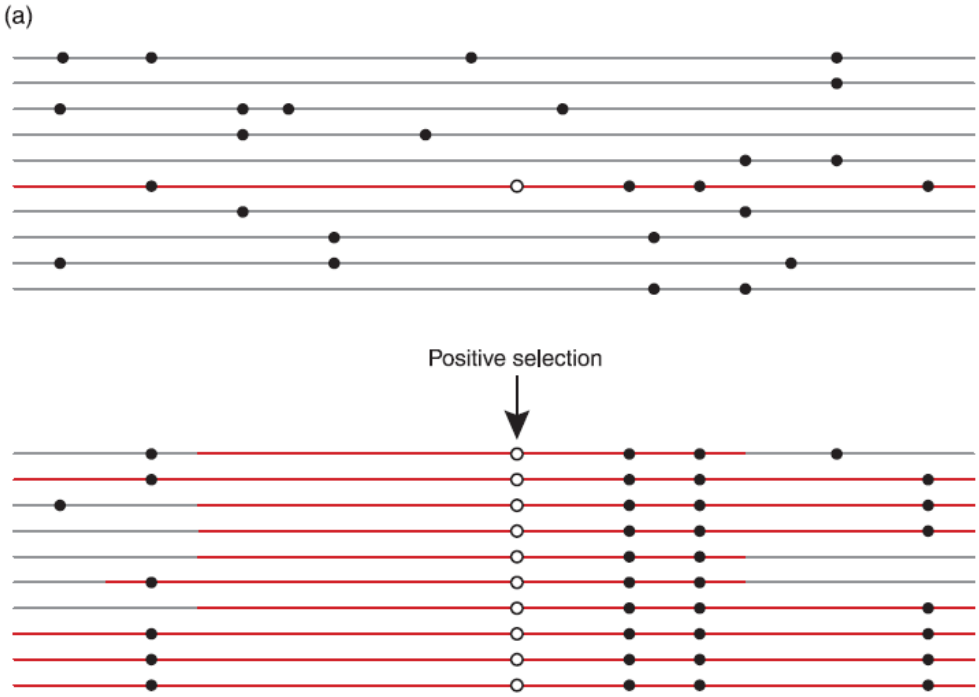


Figure 11 Figure montrant l'effet d'auto-stop génétique. Une mutation avantageuse apparaît dans la population et augmente en fréquence, emmenant avec elle les mutations neutres voisines et créant un déséquilibre de liaison important dans son voisinage. Figure issue de Storz (2005)

Des méthodes, empiriques (Browning et Weir 2010) ou encore basées sur un modèle de génétique des populations (Fariello et al. 2013), ont déjà été développé afin de tirer parti de l'information de DL contenue dans les données haplotypiques, pour l'identification de signatures de sélection. Je vais donc dédier le chapitre 3 de ma thèse à l'extension du modèle hiérarchique bayésien SelEstim (Vitalis et al. 2014) pour l'utilisation de données haplotypiques.

Objectifs de la thèse

Les objectifs principaux de ma thèse s'articulent donc autour de l'étude de la différenciation génétique à partir de données issues de technologies NGS. Le premier axe repose sur le développement d'un estimateur non biaisé de F_{ST} pour l'étude de la différenciation génétique entre populations à partir de

données Pool-seq. De ce fait, son applicabilité concerne à la fois les aspects neutres et de sélection en génétique des populations. Le second axe consiste à étendre le modèle hiérarchique bayésien SelEstim, développé au laboratoire (Vitalis et al. 2014), pour caractériser les signatures de sélection dans les génomes. Contrairement à une approche plus classique de *scan* F_{ST} (Akey et al. 2002; Weir et al. 2005), le modèle décompose les effets globaux (migration et dérive) des effets locaux (sélection). Ces développements vont se décliner en deux points :

(i) la prise en compte de l'information apportée par le DL dans les données. Pour cela nous allons étendre le modèle à l'utilisation de données multi-alléliques issues de données haplotypiques. Ceci représente un pari vis à vis des données Pool-seq majoritairement utilisées au laboratoire, à partir desquelles nous ne pouvons pas facilement obtenir d'haplotypes. Nous espérons que les données individuelles nécessaires à l'obtention d'haplotypes vont devenir de plus en plus accessibles dans le futur. De plus, des méthodes sont développées dans le but de reconstruire des haplotypes à partir de données Pool-seq (Franssen et al. 2017; Long et al. 2011).

(ii) l'exploration d'une modélisation alternative faisant intervenir une variable bayésienne auxiliaire, permettant l'utilisation d'un critère de décision alternatif pour discriminer les marqueurs sous sélection des marqueurs neutres. Cette modélisation nous permettra aussi d'intégrer un modèle de lissage éventuellement applicable aux données de type Pool-seq, et donc de comparer l'efficacité des différentes approches dans la mise en évidence de régions génomiques sous sélection.

Chapitre 1

Estimation du F_{ST} à partir de données Pool-seq

1.1 Introduction

Comme nous avons pu le voir dans l'introduction générale, les nouvelles technologies de séquençage (NGS) fournissent une quantité sans précédent de données de polymorphisme, à la fois pour des espèces modèles et non-modèles (Ellegren 2014). Bien qu'initialement le séquençage impliquait un marquage individuel, notamment chez l'homme (The International HapMap Consortium 2005), le séquençage de génome d'individus en mélange (Pool-seq) est désormais de plus en plus utilisé en génomique des populations (Schlötterer et al. 2014). S'il a été montré qu'à effort de séquençage équivalent, le Pool-seq permettait une estimation similaire, si ce n'est plus précise, des fréquences alléliques par rapport au séquençage d'individus (Ind-seq) (Futschik et Schlötterer 2010; Gautier et al. 2013), le problème est différent lorsqu'on s'intéresse aux paramètres de diversité et de différenciation, qui dépendent de moments d'ordre deux des fréquences alléliques ou, de manière équivalente, des mesures d'identité génétique par paires (voir Introduction Générale et Figure 1.1-A). Il est en effet impossible, à partir de données Pool-seq de distinguer des paires de lectures qui sont identiques parce qu'elles ont été séquencées à partir d'un seul gène, de celles qui sont identiques parce qu'elles

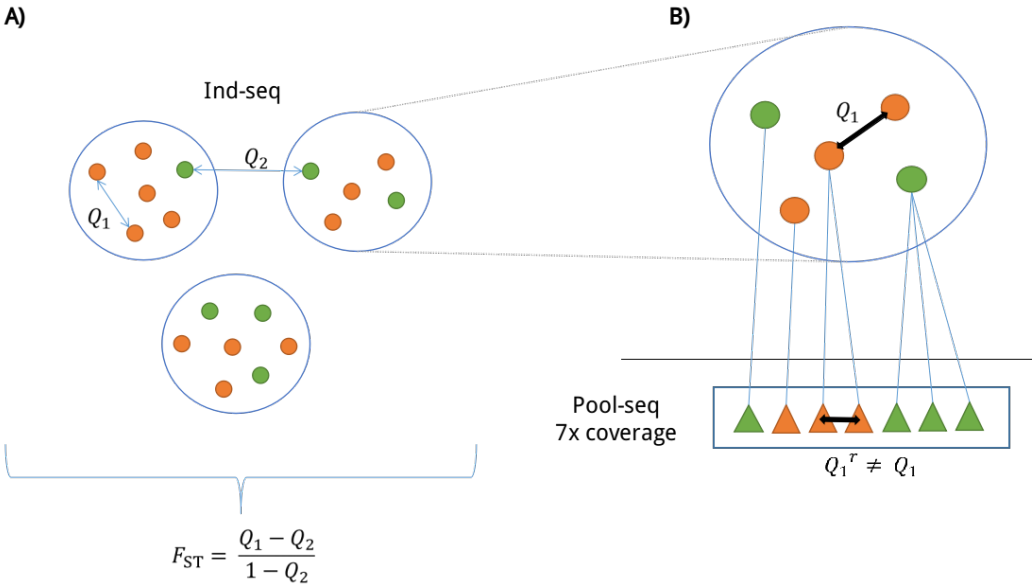


Figure 1.1 Représentation graphique des probabilités d'identité par état A) entre paires de gènes (orange et vert) au sein (Q_1) et entre les dèmes (Q_2) ainsi que B) les données de lectures issues du processus de séquençage Pool-seq qui ne permet plus de distinguer les gènes IIS parce qu'ils sont issus d'un même gène ou bien parce qu'ils sont issus de deux gènes distincts mais IIS.

ont été séquencées à partir de deux gènes distincts mais identiques par état (IIS) (Ferretti et al. 2013) (voir Figure 1.1-B). Notons, que nous utiliserons à partir de maintenant le terme "gène" pour désigner une unité génétique ségrégante (au sens mendélien du terme : voir Orgogozo et al. 2016).

Des estimateurs appropriés de diversité et de différenciation sont donc requis pour tenir compte à la fois de l'échantillonnage des gènes dans les mélanges d'individus, et de l'échantillonnage des lectures à partir de ces gènes. Plusieurs études définissant des estimateurs de F_{ST} pour données Pool-seq ont déjà été réalisées (Ferretti et al. 2013; Kofler et al. 2011), à partir de ratios d'hétérozygotie (ou de probabilités d'identité entre paires de lectures) au sein et entre les pools. Dans ce qui suit, nous allons montrer que ces estimateurs ainsi que d'autres approches déjà utilisées sont biaisés (c'est-à-dire qu'ils ne convergent pas vers la valeur attendue du paramètre) et que certains ont même des propriétés statistiques indésirables (par exemple le biais dépend de la taille de l'échantillon et de la couverture). Ici, en suivant les travaux

de Cockerham (1969), Cockerham (1973), Weir et Cockerham (1984), Weir (1996) et Rousset (2007), nous définissons un nouvel estimateur de F_{ST} basé sur la méthode des moments, et plus précisément sur une décomposition de la variance. Par la suite, nous évaluons la précision de notre estimateur sur la base d'analyses de jeux de données simulés pour le comparer aux estimations du logiciel PoPoolation2 (Kofler et al. 2011), et de Ferretti et al. (2013). Par ailleurs, nous testons aussi la robustesse de notre estimateur à différents écarts au modèle sous-jacent (incluant la contribution inégale des individus au pool et les erreurs de séquençage). Enfin, nous ré-analysons le jeu de données Pool-seq du chabot piquant (*Cottus asper*) (publié par Dennenmoser et al. 2017) et montrons comment l'utilisation d'estimateurs biaisés de F_{ST} peut conduire à une mauvaise interprétation de la structure des populations.

1.2 Modèle

Pour des raisons de clarté, les notations utilisées dans ce chapitre sont indiquées dans le Tableau 1. Nous dérivons dans un premier temps notre modèle pour un seul locus, les estimateurs qui en résultent pouvant ensuite être combinés pour plusieurs locus. Considérons un échantillon de n_{d} sous-populations, chacune constituée de n_{i} gènes ($i = 1, \ldots, n_{\mathrm{d}}$) et séquencée en mélange (n_{i} est donc la taille haploïde du pool i). On définit c_{ij} le nombre de lectures séquencées à partir du gène j ($j = 1, \ldots, n_{\mathrm{i}}$) dans la sous-population i au locus considéré. Notons que c_{ij} est une variable latente qui ne peut être directement observée à partir des données. Soit $X_{ijr:k}$ une variable indicatrice pour la lecture telle que $X_{ijr:k} = 1$ si la $r^{\text{ème}}$ lecture du $j^{\text{ème}}$ gène dans le dème i est de type k, et $X_{ijr:k} = 0$ autrement. Dans ce qui suit, nous utilisons les notations classiques pour les moyennes d'échantillons, c'est-à-dire : $X_{ij\cdot:k} \equiv \sum_r X_{ijr:k}/c_{ij}$, $X_{i\cdots:k} \equiv \sum_j \sum_r X_{ijr:k}/\sum_j c_{ij}$ et $X_{\cdots:k} \equiv \sum_i \sum_j \sum_r X_{ijr:k}/\sum_i \sum_j c_{ij}$. L'analyse de variance est basée sur le calcul de sommes de carrés, d'où :

Tableau 1.1 Résumé des notations principales

Notation	Description du paramètre
$X_{ijr:k}$	Variable indicatrice : $X_{ijr:k} = 1$ si la $r$ème lecture du $j$ème individu dans le pool i est de type k, et $X_{ijr:k} = 0$ autrement
$r_{i:k} = \sum_j \sum_r X_{ijr:k}$	Nombre de lectures du type k dans le $i$ème pool
c_{ij}	Nombre de lectures séquencées à partir de l'individu j dans la sous-population i (couverture individuelle non observée)
$C_{1i} \equiv \sum_j c_{ij}$	Nombre total de lectures dans le $i$ème pool (couverture du pool i)
$C_1 \equiv \sum_i C_{1i}$	Nombre total de lectures dans l'échantillon complet (couverture totale)
$C_2 \equiv \sum_i C_{1i}^2$	Carré du nombre de lectures dans l'échantillon complet
n_i	Nombre total de gènes dans le $i$ème pool (taille de pool haploïde)
$y_{i:k}$	Nombre de gènes (non observé) de type k dans le $i$ème pool
$\pi_k \equiv \mathbb{E}(X_{ijr:k})$	Fréquence attendue des lectures de type k dans l'échantillon complet
$\hat{\pi}_{ij:k} \equiv X_{ij\cdot:k}$	Fréquence moyenne (non observée) des lectures de type k pour l'individu j dans le $i$ème pool
$\hat{\pi}_{i:k} \equiv X_{i\cdots:k}$	Fréquence moyenne des lectures de type k dans le $i$ème pool
$\hat{\pi}_k \equiv X_{\cdots:k}$	Fréquence moyenne des lectures de type k dans l'échantillon complet
Q_1 (resp. Q_2)	Probabilité IIS pour deux gènes échantillonnés au sein des (resp. entre les) pools
Q_1^r (resp. Q_2^r)	Probabilité IIS pour deux lectures échantillonnées au sein des (resp. entre les) pools
$\hat{Q}_1^{\text{pool}}$ (resp. $\hat{Q}_2^{\text{pool}}$)	Estimateur non biaisé des probabilités IIS pour des gènes échantillonnés au sein des (resp. entre les) populations

$$\begin{aligned}
\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ijr:k}-X_{\cdots:k}\right)^2 &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ijr:k}-X_{ij\cdot:k}\right)^2 \\
&+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ij\cdot:k}-X_{i\cdot\cdot:k}\right)^2 \\
&+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{i\cdot\cdot:k}-X_{\cdots:k}\right)^2 \\
&\equiv SSR_{:k}+SSI_{:k}+SSP_{:k} \qquad (1.1)
\end{aligned}$$

Comme indiqué en annexe de l'article de Hivert et al. (2018), l'espérance de la somme des carrés dépend de l'espérance des fréquences alléliques π_k sur des réplicats indépendants du même processus évolutif, des probabilités d'identité par état (IIS) $Q_{1:k}$ que deux gènes dans le même pool soient tous deux de type k, et de la probabilité IIS $Q_{2:k}$ que deux gènes issus de pools différents soient tous deux de type k. Considérant les espérances, on obtient (voir le détail des calculs en annexe de l'article de Hivert et al. 2018) :

$$\mathbb{E}(SSR_{:k}) = 0 \qquad (1.2)$$

pour les lectures au sein des gènes individuels, en supposant qu'il n'y a pas d'erreur de séquençage, c'est-à-dire que l'ensemble des lectures issues d'un unique gène sont identiques (d'où $X_{ijr:k} = X_{ij\cdot:k}$ pour tout r). Pour les lectures au sein des pools, nous avons :

$$\mathbb{E}(SSI_{:k}) = (C_1 - D_2)\,(\pi_k - Q_{1:k}) \qquad (1.3)$$

où $C_1 \equiv \sum_i \sum_j c_{ij}$ est la couverture totale, soit le nombre total de lectures dans l'ensemble des échantillons, et $D_2 \equiv \mathbb{E}\left(\sum_i \sum_j c_{ij}^2 / C_{1i}\right)$. Parce qu'on ne peut pas observer les lectures c_{ij}, D_2 est approché en supposant une distribution multinomiale des c_{ij}'s entre les gènes : $D_2 \equiv \sum_i \left(C_{1i} + n_i - 1\right)/n_i$ (voir l'Équation A15 en annexe de l'article de Hivert et al. 2018). Pour les

lectures entre gènes provenant de pools différents, on a :

$$\mathbb{E}(SSP_{:k}) \;=\; \left(C_1 - \frac{C_2}{C_1}\right)(Q_{1:k} - Q_{2:k}) + (D_2 - D_2^\star)(\pi_k - Q_{1:k}) \quad (1.4)$$

où $C_2 \equiv \sum_i \left(\sum_j c_{ij}\right)^2$ et $D_2^\star \equiv \mathbb{E}\left(\sum_i \sum_j c_{ij}^2 / C_1\right)$. De même que D_2, $D_2^\star$ est aussi approché en supposant une distribution multinomiale des lectures c_{ij} entre les gènes : $D_2^\star \equiv \left[\sum_i C_{1i}\left(C_{1i} + n_i - 1\right)/n_i\right]/C_1$ (voir l'Équation A16 en annexe de l'article de Hivert et al. 2018). En réarrangeant les Équations 1.3–1.4 et en sommant sur les allèles, on obtient :

$$Q_1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) - (D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)\,(C_1 - C_2/C_1)} \quad (1.5)$$

et :

$$1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) + (n_\mathrm{c} - 1)\,(D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)\,(C_1 - C_2/C_1)} \quad (1.6)$$

où $n_\mathrm{c} \equiv (C_1 - C_2/C_1)\,/\,(D_2 - D_2^\star)$. Posons $MSI \equiv SSI/\,(C_1 - D_2)$ et $MSP \equiv SSP/\,(D_2 - D_2^\star)$. Alors :

$$F_\mathrm{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{\mathbb{E}(MSP) - \mathbb{E}(MSI)}{\mathbb{E}(MSP) + (n_\mathrm{c} - 1)\,\mathbb{E}(MSI)} \quad (1.7)$$

ce qui nous donne l'estimateur basé sur la méthode des moments :

$$\hat{F}_\mathrm{ST}^\mathrm{pool} = \frac{MSP - MSI}{MSP + (n_\mathrm{c} - 1)\,MSI} \quad (1.8)$$

Notons que l'Équation A27 de l'annexe de l'article de Hivert et al. (2018) donne l'expression de $\hat{F}_\mathrm{ST}^\mathrm{pool}$ en fonction des fréquences des lectures des échantillons. Si on considère le cas limite où chaque gène est séquencé exactement une fois, on retrouve le modèle Ind-seq : supposant $c_{ij} = 1$ pour tout (i, j), alors $C_1 = \sum_i^{n_\mathrm{d}} n_i \equiv S_1$, $C_2 = \sum_i^{n_\mathrm{d}} n_i^2 \equiv S_2$, $D_2 = n_\mathrm{d}$ et $D_2^\star = 1$. Par conséquent, $n_\mathrm{c} = (S_1 - S_2/S_1)\,/\,(n_\mathrm{d} - 1)$, et l'Équation 1.8 se réduit exactement à l'estimateur de F_ST pour des données haploïdes de type Ind-seq : voir Weir (1996), p. 182, et Rousset (2007), p. 977.

Comme dans Reynolds et al. (1983), Weir et Cockerham (1984), Weir

(1996) and Rousset (2007), un estimateur multi-locus est dérivé comme la somme des numérateurs locus spécifiques sur la somme des dénominateurs locus spécifiques :

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_l MSP_l - MSI_l}{\sum_l MSP_l + (n_{\mathrm{c}} - 1)\, MSI_l} \tag{1.9}$$

où MSI et MSP sont indicés avec l pour indiquer le $l$ème locus. Pour les données Ind-seq, Bhatia et al. (2013) se réfèrent à cet estimateur multi-locus comme un "ratio des moyennes" par opposition à une "moyenne des ratios", qui consisterait à faire la moyenne des F_{ST} locus spécifiques. Notre approche est justifiée dans l'annexe de Weir et Cockerham (1984) ainsi que dans Bhatia et al. (2013), qui ont analysé les deux estimateurs à l'aide de simulations basées sur le coalescent. Notons que l'Équation 1.9 suppose que la taille des pools est constante sur l'ensemble des locus. Ajoutons que la construction de l'estimateur dans l'Équation 1.9 est différente de celle de Weir et Cockerham (1984). Ces auteurs ont défini leur estimateur multi-locus comme un ratio de sommes de composantes de la variance (a, b et c dans leurs notations) sur les locus, ce qui donne le même poids à l'ensemble des locus, sans considération du nombre de gènes échantillonnés à chaque locus. L'Équation 1.9 suit l'esprit de ce qui est fait dans Genepop (Rousset 2008), qui donne plus de poids aux locus qui sont couverts plus intensément.

1.3 Matériels et Méthodes

1.3.1 Étude par simulations

Génération des génotypes individuels

Nous générons dans un premier temps des génotypes individuels à l'aide du logiciel `ms` (Hudson 2002), en supposant que la structure de nos populations suit un modèle en îles (Wright 1931). Pour chaque scénario simulé, nous considérons 8 dèmes, chacun constitué de $N = 5000$ individus haploïdes. Le taux de migration (m) a été fixé de manière à obtenir une valeur de F_{ST} (0.05 ou 0.2) en utilisant l'Équation 6 de Rousset (1996), soit : $M \equiv 2Nm = 16.569$

pour $F_{\mathrm{ST}} = 0.05$ et $M = 3.489$ pour $F_{\mathrm{ST}} = 0.20$. Le taux de mutation a été fixé à $\mu = 10^{-6}$, soit $\theta \equiv 2N\mu = 0.01$. Nous avons ensuite considéré à la fois des tailles fixes d'échantillons (n =10 ou 100), et des tailles variables entre dèmes. Dans ce cas, la taille haploïde n de l'échantillon a été déterminée pour chaque dème de façon indépendante à partir d'une distribution gaussienne de moyenne 100 et de déviation standard 30, le résultat étant ensuite arrondi à l'entier le plus proche, avec un minimum de 20 et un maximum de 300 gènes par dème. Nous avons généré 400 000 séquences pour chaque scénario, dans le but de retenir 50 000 SNPs polymorphes à partir des séquences contenant un seul et unique site polymorphe. Ce nombre relativement élevé de marqueurs a été simulé dans le but de retenir au moins 5 000 SNPs une fois les données pools générées. Chaque scénario a été répliqué 50 fois. Notons que nous utiliserons à partir de maintenant la dénomination Ind-seq pour des données issues de séquençage individuel et pour lesquelles le génotypage est supposé sans erreur.

Séquençage en pool

Pour chaque jeu de données simulé à partir de `ms`, nous avons généré un jeu de données Pool-seq par tirage dans une distribution binomiale (Gautier et al. 2013). Plus précisément, nous supposons pour chaque SNP, que le nombre $r_{i:k}$ de lectures de type allélique k dans le pool i suit :

$$r_{i:k} \sim \mathrm{Bin}(\frac{y_{i:k}}{n_i}, \delta_i) \tag{1.10}$$

où $y_{i:k}$ est le nombre de gènes portant l'allèle k dans le pool i, n_i est le nombre total de gènes dans le pool i (taille haploïde de pool), et δ_i est la couverture totale simulée pour le pool i. Dans ce qui suit, nous considérons à la fois une couverture fixe, avec $\delta_i = \Delta$ pour l'ensemble des pools et locus, ou bien une couverture variable entre pools et locus, avec $\delta_i \sim \mathrm{Pois}(\Delta)$.

Erreurs de séquençage

Nous avons simulé les erreurs de séquençage avec un taux $\mu_e = 0.001$, typique des séquenceurs Illumina (Glenn 2011; Ross et al. 2013). Chaque

erreur de séquençage modifie l'état allélique de la lecture vers un des trois états possibles avec une probabilité égale (il y a donc quatre états alléliques possibles au total, correspondant aux quatre nucléotides). Notons que seul les marqueurs biallélliques sont retenus dans les jeux de données finaux, ce qui constitue un premier filtre. Aussi, puisque nous débutons cette procédure seulement avec des marqueurs polymorphes, nous négligeons les erreurs de séquençage qui créeraient de faux SNPs. Cependant, de tels SNPs sont relativement rares dans les jeux de données réels, les marqueurs avec un nombre faible de lectures (c'est-à-dire en deçà d'un MRC, pour nombre minimal de lectures, donné) étant généralement supprimés.

Erreur expérimentale

Les quantités non-équimolaires d'ADN des différents individus au sein d'un pool ainsi que la variation stochastique de l'efficacité de l'amplification des ADNs individuels sont autant de sources d'erreur expérimentale dans le séquençage en pool. Afin de les simuler, nous avons utilisé le modèle dérivé par Gautier et al. (2013). Il suppose que la contribution $\eta_{ij} = c_{ij}/C_{1i}$ de chaque gène j à la couverture totale du pool i (C_{1i}) suit une distribution de Dirichlet :

$$\{\eta_{ij}\}_{1\leq j\leq n_j} \sim \mathrm{Dir}(\frac{\rho}{n_i}) \tag{1.11}$$

où le paramètre ρ contrôle la dispersion des contributions des gènes autour de la valeur $\eta_{ij} = 1/n_i$, attendue si tous les gènes contribuaient de manière égale au pool de lectures. Par souci de clarté, nous définissons l'erreur expérimentale ϵ comme le coefficient de variation de η_{ij}, c'est-à-dire : $\epsilon \equiv \sqrt{\mathbb{V}(\eta_{ij})}/\mathbb{E}(\eta_{ij}) = \sqrt{(n_i - 1)/(\rho + 1)}$ (voir Gautier et al. (2013)). Lorsque ϵ tend vers 0 (ou de manière équivalente, lorsque ρ tend vers l'infini), tous les individus contribuent de manière égale au pool et il n'y a pas d'erreur expérimentale. Nous avons testé la robustesse de notre estimateur à $\epsilon = 0.5$, ce qui correspond à une situation où (pour $n_i = 10$) 5 individus contribuent 2.8x plus au pool de lectures que les 5 autres.

Tableau 1.2 Description des estimateurs de F_{ST} utilisés dans le texte.

Notation	Definition
$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	Équation 1.8
FRP_{13}	Ferretti et al. (2013) et Équations 1.12, 1.16–1.17
NC_{83}	Nei et Chesser (1983)
$\mathrm{PP2}_{\mathrm{d}}$	Kofler et al. (2011) et Équations 1.12–1.14
$\mathrm{PP2}_{\mathrm{a}}$	Kofler et al. (2011) et Équation 1.15
WC_{84}	Weir et Cockerham (1984)

1.3.2 Autres estimateurs

Un résumé des notations pour les différents estimateurs de F_{ST} utilisés dans ce chapitre est donné dans le Tableau 1.2.

$\mathrm{PP2}_{\mathrm{d}}$ **:**

Cet estimateur de F_{ST} est celui implémenté par défaut dans le logiciel PoPoolation2 (Kofler et al. 2011). Il se base sur une définition du paramètre F_{ST} correspondant à la réduction globale de l'hétérozygotie moyenne relativement à la métapopulation :

$$\mathrm{PP2}_{\mathrm{d}} \equiv \frac{\hat{H}_{\mathrm{T}} - \hat{H}_{\mathrm{S}}}{\hat{H}_{\mathrm{T}}} \tag{1.12}$$

où $\hat{H}_{\mathrm{S}}$ est l'hétérozygotie moyenne au sein d'une sous-population et $\hat{H}_{\mathrm{T}}$ l'hétérozygotie moyenne dans la méta-population (obtenue en mélangeant toutes les sous-populations pour ne former qu'une seule entité virtuelle). Dans PoPoolation2, $\hat{H}_{\mathrm{S}}$ est l'hétérozygotie moyenne non pondérée au sein des sous-populations :

$$\hat{H}_{\mathrm{S}} = \frac{1}{n_{\mathrm{d}}} \sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(\frac{C_{1i}}{C_{1i} - 1} \right) \sum_{k} \left(1 - \hat{\pi}_{i:k}^2 \right) \tag{1.13}$$

(en utilisant les notations du Tableau 1) :

$$\hat{H}_{\mathrm{T}} = \left(\frac{\min_i(n_i)}{\min_i(n_i)-1}\right)\left(\frac{\min_i(C_{1i})}{\min_i(C_{1i})-1}\right)\sum_k\left(1-\hat{\pi}_k^2\right) \tag{1.14}$$

$\mathrm{PP2_a}$ **:**

Cet estimateur est l'estimateur de F_{ST} alternatif implémenté dans le logiciel PoPoolation2. Il est basé sur une interprétation de Kofler et al. (2011) de l'estimateur de F_{ST} de Karlsson et al. (2007), tel que :

$$\mathrm{PP2_a} \equiv \frac{\hat{Q}_1^{\mathrm{r}} - \hat{Q}_2^{\mathrm{r}}}{1-\hat{Q}_2^{\mathrm{r}}} \tag{1.15}$$

où $\hat{Q}_1^{\mathrm{r}}$ et $\hat{Q}_2^{\mathrm{r}}$ sont respectivement les fréquences de paires de lectures identiques au sein et entre les pools, simplement calculées en comptant les paires de gènes IIS. Ce sont des estimateurs de Q_1^{r}, la probabilité IIS pour deux lectures du même pool (peu importe qu'elles soient séquencées à partir du même gène ou non) et Q_2^{r}, la probabilité IIS pour deux lectures de pools différents. Notons que la probabilité IIS Q_1^{r} est différente de Q_1 dans l'Équation 1.7, qui telle qu'on l'a définie, représente la probabilité IIS entre deux gènes distincts du même pool. Cette approche confond ainsi des paires de lectures au sein d'un pool qui sont identiques parce que séquencées à partir d'un même gène, de celles qui sont identiques parce qu'elles sont séquencées à partir de gènes distincts mais IIS (voir Figure 1.1).

$\mathrm{FRP_{13}}$ **:**

Cet estimateur de F_{ST} a été développé par Ferretti et al. (2013) (voir leurs Équations 3 et 10–13). Ferretti et al. (2013) utilisent la même définition de F_{ST} que dans l'Équation 1.12, bien qu'ils estiment des hétérozygoties au sein et entre les pools comme des "diversités nucléotidiques moyennes par paires", qui sont formellement équivalentes aux probabilités IIS. Plus particulièrement, ils estiment l'hétérozygotie intra-pool comme (en utilisant les

notations du Tableau 1) :

$$\hat{H}_{\mathrm{S}} = \frac{1}{n_{\mathrm{d}}} \sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(1 - \hat{Q}^{\mathrm{r}}_{1i} \right) \tag{1.16}$$

et l'hétérozygotie totale entre les n_{d} sous-populations comme :

$$\hat{H}_{\mathrm{T}} = \frac{1}{n_{\mathrm{d}}^2} \left[\sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(1 - \hat{Q}^{\mathrm{r}}_{1i} \right) + \sum_{i \neq i'}^{n_{\mathrm{d}}} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'} \right) \right] \tag{1.17}$$

1.3.3 Analyses des données Ind-seq

L'estimation des F_{ST} sur données Ind-seq est effectuée en considérant à la fois l'estimateur de Nei et Chesser (1983) basé sur un ratio d'hétérozygotie (voir l'Équation 1.12), noté NC_{83}, ou l'estimateur basé sur une analyse de variance développé par Weir et Cockerham (1984), noté WC_{84}.

Afin de comparer équitablement les estimations sur jeux de données Ind-seq et Pool-seq, les F_{ST} sont estimés sur des sous-échantillons de 5 000 SNPs constitués par les mêmes marqueurs (c'est-à-dire remplissant les critères de sélection utilisés pour les différentes couvertures considérées en Pool-seq)

Tous les estimateurs ont été implémentés dans des fonctions R version 3.3.1 (R Core Team 2017). L'ensemble des fonctions a été précautionneusement vérifié à l'aune des logiciels existants.

1.3.4 Exemple d'application : *Cottus asper*

Dennenmoser et al. (2017) ont étudié les bases génomiques de l'adaptation aux conditions osmotiques chez le chabot piquant (*Cottus asper*), un poisson euryhalin abondant du Nord-Ouest de l'Amérique du Nord. Pour cela, ils ont séquencé l'ensemble du génome à partir de pools d'individus provenant de deux populations d'estuaire (CR, Capilano River Estuary ; FE, Fraser River Estuary) et deux populations d'eau douce (PI, Pitt Lake et HZ, Hatzic Lake) dans le sud de la Colombie Britanique (Canada). Nous avons téléchargé les quatre fichiers BAM correspondant sur le répertoire digital Dryad (doi : 10.5061/dryad.2qg01) que l'on a combinés en un seul fichier mpileup en uti-

lisant `SAMtools` version 0.1.19 (Li et al. 2009) avec les options par défaut, à l'exception de la couverture maximale par BAM fixée à 5 000 lectures. Le fichier résultant a ensuite été formaté à l'aide d'un script `awk`, pour détecter les SNPs et calculer les comptages de lectures, après suppression des bases dont le score de qualité de la base (BAQ) était inférieur à 25. Une position était ensuite considérée comme un SNP si : (*i*) seuls deux nucléotides avec plus d'une lecture étaient observés (les nucléotides avec $\leq$1 lecture étaient considérés comme des erreurs de séquençage) ; (*ii*) la couverture était comprise entre 10 et 300 dans chacun des quatre fichiers d'alignement ; (*iii*) la fréquence de l'allèle minoritaire (MAF), calculée à partir des comptages de lectures, était ≥ 0.01 dans les quatre populations. Le jeu de données final était constitué de 608 879 SNPs.

Notre but était de comparer la structure de population inférée à partir d'estimateurs de F_{ST} par paires, en utilisant d'un côté l'estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ et d'un autre côté l'estimateur $\mathrm{PP2_d}$. Nous avons également comparé dans les deux cas la structure de populations obtenue à celle inférée à partir du modèle hiérarchique bayésien implémenté dans le logiciel BayPass (Gautier 2015). BayPass permet en effet une estimation robuste de la matrice de covariance des fréquences alléliques entre populations pour des données Pool-seq, qui est reconnue pour être informative sur l'histoire démographique des populations (Pickrell et Pritchard 2012). Les éléments de la matrice estimée peuvent être interprétés comme des estimations de niveaux de différencation par paires et population-spécifiques (Coop et al. 2010), fournissant ainsi une description de la structure des populations utilisant l'ensemble des données disponibles.

1.4 Résultats

1.4.1 Comparaison d'estimations de F_{ST} Ind-seq et Pool-seq

Les estimations $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ locus spécifiques sont très corrélées avec l'estimateur classique WC_{84} calculé sur les données individuelles utilisées dans la simulation des pools (voir Figure 1.2). La variance de $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ sur des réplicats

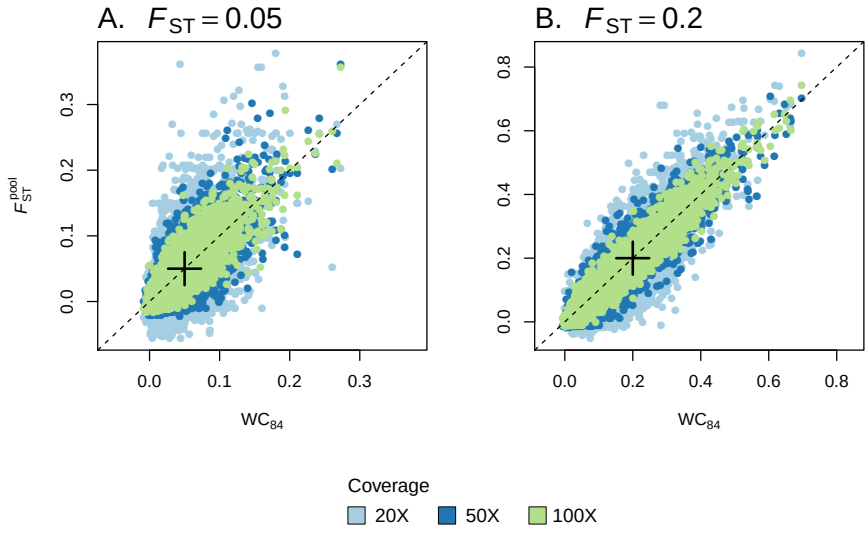


Figure 1.2 Estimations de F_{ST} locus-spécifiques. Nous avons comparé les estimations de F_{ST} locus-spécifiques basées sur des données de comptages alléliques obtenues à partir de génotypes individuels (Ind-seq), en utilisant l'estimateur WC_{84}, aux estimations issues de l'estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ à partir de données Pool-seq. Nous avons simulé 5 000 SNPs avec `ms`dans un modèle en îles avec $nd = 8$ dèmes. Nous avons utilisé deux taux de migration, correspondant à $F_{\mathrm{ST}} = 0.05$ (A) et $F_{\mathrm{ST}} = 0.20$ (B). La taille de chaque pool a été fixé à 100. Nous montrons les résultats pour différentes couvertures (20X, 50X and 100X). Dans chaque graphique, la croix indique la valeur de F_{ST} théorique.

indépendants diminue avec l'augmentation de la couverture. Ajoutons que la corrélation entre $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ et WC_{84} est encore plus forte pour des estimations multi-locus (voir Figure S1A de l'annexe de l'article de Hivert et al. (2018)).

1.4.2 Comparaison de différents estimateurs de F_{ST} pour données Pool-seq

Tableau 1.3 Liste bibliographiques non exhaustive d'articles référencés en 2016 et 2017 où des données Pool-seq ont été utilisées pour mesurer des F_{ST}.

Référence	Modèle biologique	Estimateur
Machado et al. (2016)	*Drosophila sp.*	WC_{84} [1]
Kang et al. (2016a)	*Drosophila melanogaster*	$\mathrm{PP2}_{\mathrm{d}}$
Collet et al. (2016)	*Drosophila melanogaster*	$\mathrm{PP2}_{\mathrm{d}}$
Yassin et al. (2016)	*Drosophila yakuba*	$\mathrm{PP2}_{\mathrm{d}}$
Kang et al. (2016b)	*Drosophila sp.*	$\mathrm{PP2}_{\mathrm{d}}$
Franchini et al. (2016)	Midas Cichlid group	$\mathrm{PP2}_{\mathrm{d}}$
Tyagi et al. (2016)	*Arabidopsis thaliana*	$\mathrm{PP2}_{\mathrm{d}}$
Sangwan et al. (2016)	Polynucleobacter	$\mathrm{PP2}_{\mathrm{d}}$
Burke (2016)	*Microplitis demolitor*	$\mathrm{PP2}_{\mathrm{d}}$
Deitz et al. (2016)	*Anopheles melas*	$\mathrm{PP2}_{\mathrm{d}}$
Kjærner-Semb et al. (2016)	*Salmo salar L.*	$\mathrm{PP2}_{\mathrm{d}}$
Fruciano et al. (2016)	*A.astorquii* et *A.zaliosus*	$\mathrm{PP2}_{\mathrm{d}}$
Bastide et al. (2016)	*Drosophila melanogaster*	$\mathrm{PP2}_{\mathrm{d}}$
Gammerdinger et al. (2016)	*Sarotherodon melanotheron*	$\mathrm{PP2}_{\mathrm{d}}$ [2]
Fleming et al. (2016)	Leghorn et Fayoumi chicken	$\mathrm{PP2}_{\mathrm{a}}$
Fu et al. (2016)	Cochon noir de Chine	$\mathrm{PP2}_{\mathrm{d}}$
Kadri et al. (2016)	*Apis mellifera*	$\mathrm{PP2}_{\mathrm{d}}$
Kaiser et al. (2016)	*Clunio marinus*	$\mathrm{PP2}_{\mathrm{d}}$
Lai et al. (2016)	*Capra hircus*	$\mathrm{PP2}_{\mathrm{d}}$
Wang et al. (2016)	*Capra hircus*	Akey et al. (2010)
Phillips et al. (2016)	*Drosophila melanogaster*	$1 - H_S/H_T$
Hendrick et Mathiasson (2016)	*Mimulus guttatus*	$\mathrm{PP2}_{\mathrm{d}}$

Suite sur la page suivante

Tableau 1.3 – (suite)

Référence	Modèle biologique	Estimateur
Griffin et al. (2017)	*Drosophila melanogaster*	$PP2_d$
Endler et al. (2016)	*Drosophila melanogaster*	$PP2_d$
Love et al. (2016)	*Anopheles gambiae*	$PP2_d$
Rellstab et al. (2016)	*Quercus spp.*	$PP2_d$
Konczal et al. (2016)	*Myodes glareolus*	$PP2_d$
Guo et al. (2016)	*Bufo andrewsi*	$PP2_d$
Kang et al. (2016a)	*Drosophila melanogaster*	$PP2_d$
Bertelsen et al. (2016)	Vaches Holstein Danoise	$PP2_d$
Günther et al. (2016)	*Arabidopsis thaliana*	$PP2_d$
Song et al. (2016)	*Capra hircus*	Nei (1987)
Guo et al. (2016)	*Clupea harengus*	$PP2_d$
Fischer et al. (2017)	*Arabidopsis halleri*	$PP2_d$
Graves et al. (2017)	*Drosophila melanogaster*	$1 - H_S/H_T$
Kaiser et al. (2016)	*Drosophila melanogaster*	$PP2_d$
Chen et al. (2016)	*Picea abies*	Hudson et al. (1992)
Bankers et al. (2017)	*Microphallus sp.*	$PP2_d$
Grande et al. (2017)	*Lasallia pustulata*	$PP2_d$
Carvajal-Rodríguez (2017)	HapMap data	Ferretti
Toomey et al. (2017)	*Serinus canaria*	$PP2_d$
Conte et al. (2017)	*Oreochromis niloticus*	Sex_SNP_finder (Gammerdinger et al. 2014)
Choi et al. (2017)	*Teladorsagia circumcincta*	$PP2_d$
Doyle et al. (2017)	*Onchocerca volvulus*	$PP2_d$
Wang et al. (2017)	Hongshan chicken	$PP2_d$
Zhao et Begun (2017)	*DrosFophila melanogaster*	Svetec et al. 2016
Oppold et al. (2017)	*Chironomus riparius*	$PP2_d$

Suite sur la page suivante

Tableau 1.3 – (suite)

Référence	Modèle biologique	Estimateur
Dennenmoser et al. (2017)	*Cottus asper*	BayScan 2.5
Neethiraj et al. (2017)	*Drosophila sp., corvus crows et papilio butterflies*	$PP2_d$
Gould et al. (2017)	*Mimulus guttatus*	$1 - H_S/H_T$
Li et al. (2017)	*Oreochromis niloticus*	$PP2_d$
Fustier et al. (2017)	*Zea mays ssp.*	Weir et Hill (2002)
Bourguinat et al. (2017)	*Dirofilaria immitis*	$PP2_d$
Hardy et al. (2018)	*Drosophila melanogaster*	$1 - H_S/H_T$
Kang et al. (2017)	*Mimulus guttatus*	$PP2_d$
Fontaine et al. (2017)	*Aedes aegypti*	WC_{84}
Kozak et al. (2017)	*Ostrinia nubilalis*	$PP2_d$

Nous avons trouvé que notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ avait un biais très faible : la moyenne sur tous les locus et réplicats converge vers la valeur simulée du paramètre, qui est basée sur le calcul de probabilités IIS pour des populations structurées selon un modèle en îles (voir l'Équation 6 dans Rousset 1996). Dans l'ensemble des situations étudiées, le biais ne dépend ni de la taille d'échantillon (c'est-à-dire la taille de chaque pool), ni de la couverture (voir Figure 1.3). Seule la variance de l'estimateur à travers les réplicats indépendants diminue avec l'augmentation de la taille d'échantillon et/ou de la couverture. À forte couverture, la moyenne ainsi que la racine carrée de l'erreur quadratique moyenne (RMSE) de $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ sur des réplicats indépendants sont quasiment indistinguables des estimations WC_{84} (voir Tableau S1 de l'annexe de l'article de Hivert et al. (2018)).

La Figure 1.4 montre le RMSE des F_{ST} estimés pour un large panel de

1. fait intervenir une correction par le calcul d'une taille efficace de pool.
2. script personnalisé

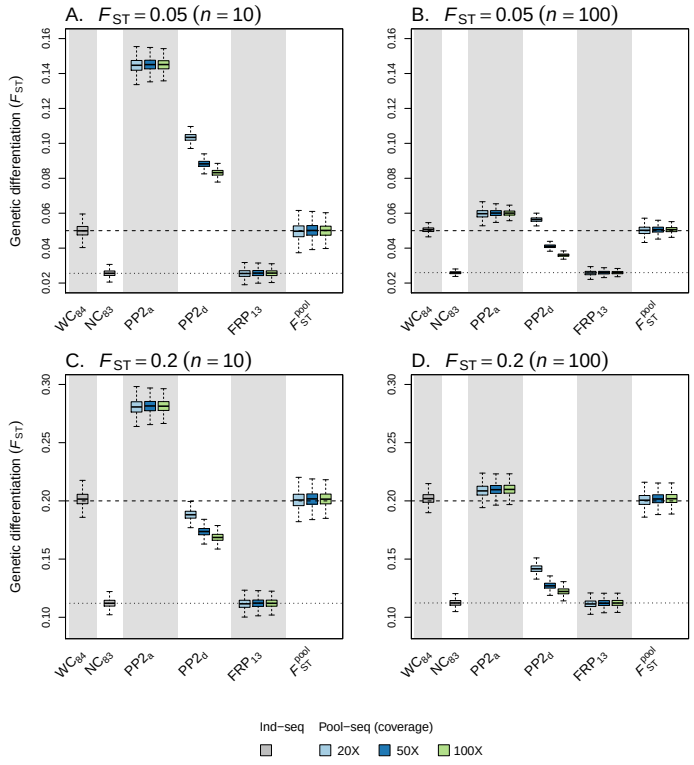


Figure 1.3 Estimateurs de F_{ST} par paires. Nous considérons deux estimateurs basés sur des données de comptages alléliques obtenues à partir de génotypes individuels (Ind-seq) : WC_{84} et NC_{83}. Pour les données issues de pools, nous calculons les deux estimateurs implémentés dans le logiciel PoPoolation2 ($\mathrm{PP2}_{\mathrm{d}}$ et $\mathrm{PP2}_{\mathrm{a}}$) ainsi que l'estimateur FRP_{13} et notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Équation 1.9). Chaque boite à moustaches représente la distribution des F_{ST} multilocus estimés pour toutes les comparaisons par paires dans un modèle en îles avec $n_{\mathrm{d}} = 8$ dèmes et entre 50 réplicats indépendants de simulations avec `ms`. Nous avons utilisé deux taux de migration, correspondant à $F_{\mathrm{ST}} = 0.05$ (A–B) et $F_{\mathrm{ST}} = 0.20$ (C–D). La taille de chaque pool était fixée soit à 10 (A et C) soit à 100 (B et D). Pour les données Pool-seq, nous montrons les résultats correspondant à différentes couvertures (20X, 50X et 100X). Dans chaque graphique, les tirets indiquent la valeur théorique de F_{ST} et la ligne en pointillé indique la médiane de la distribution des estimations NC_{83}.

tailles de pool et de couvertures. On voit que le RMSE décroît avec l'augmentation de la taille de pool et/ou de la couverture. On obtient une estimation plus précise du F_{ST} à faible niveau de différenciation. De plus, la Figure 1.4 peut être utilisée pour connaître les tailles de pool et de couverture nécessaire

à l'obtention d'un RMSE similaire à des données Ind-seq. Par exemple si on considère un échantillon de taille haploïde $n = 20$, pour $F_{ST} \leq 0.05$, alors (dans les conditions de nos simulations), le RMSE des F_{ST} estimés à partir des données Pool-seq tend vers le RMSE observé sur les données Ind-seq pour une taille de pool d'environ 200 individus haploïdes avec une couverture de 20X, ou pour une taille de pool d'environ 20 individus haploïdes séquencé à 200X.

À l'inverse, nous avons trouvé que $PP2_d$, l'estimateur par défaut du logiciel PoPoolation2, qui est l'estimateur le plus utilisé dans les études Pool-seq réalisées en 2016 et 2017 (voir Tableau 1.3), est biaisé lorsqu'on le compare à la valeur attendue du paramètre. Nous observons que le biais dépend à la fois de la taille d'échantillon et de la couverture (voir Figure 1.3). Nous pouvons aussi noter qu'avec l'augmentation de la couverture et de la taille d'échantillon, $PP2_d$ converge vers l'estimateur NC_{83} (Nei et Chesser 1983) calculé à partir des données individuelles (voir Figure S1 B de l'annexe de l'article de Hivert et al. 2018). Cet argument a été celui avancé par Kofler et al. (2011) pour valider leur approche, bien que l'estimation $PP2_d$ s'écarte de la valeur attendue du paramètre (voir Figure S1 B-C de l'annexe de l'article de Hivert et al. 2018).

Le second estimateur de F_{ST} implémenté dans PoPoolation2 que nous avons appelé $PP2_a$, est aussi biaisé (voir Figure 1.3). On note que le biais diminue avec l'augmentation de la taille d'échantillon. Ajoutons aussi que le biais ne dépend pas de la couverture (contrairement à la variance sur les réplicats indépendants). L'estimateur développé par Ferretti et al. (2013) que nous avons appelé FRP_{13}, est aussi biaisé (voir Figure 1.3). Cependant, le biais ne dépend ni de la taille de pool, ni de la couverture (contrairement à la variance sur des réplicats indépendants). FRP_{13} converge vers l'estimateur NC_{83} calculé sur les données individuelles (voir Figure 1.3). À forte couverture, la moyenne ainsi que le RMSE sur des réplicats indépendants sont quasiment indistinguables de ceux de l'estimateur NC_{83}. Ajoutons que ces deux estimateurs dépendent de la même manière du nombre de dèmes échantillonnés, le biais diminuant avec l'augmentation de ces derniers (voir Figure 1.5)

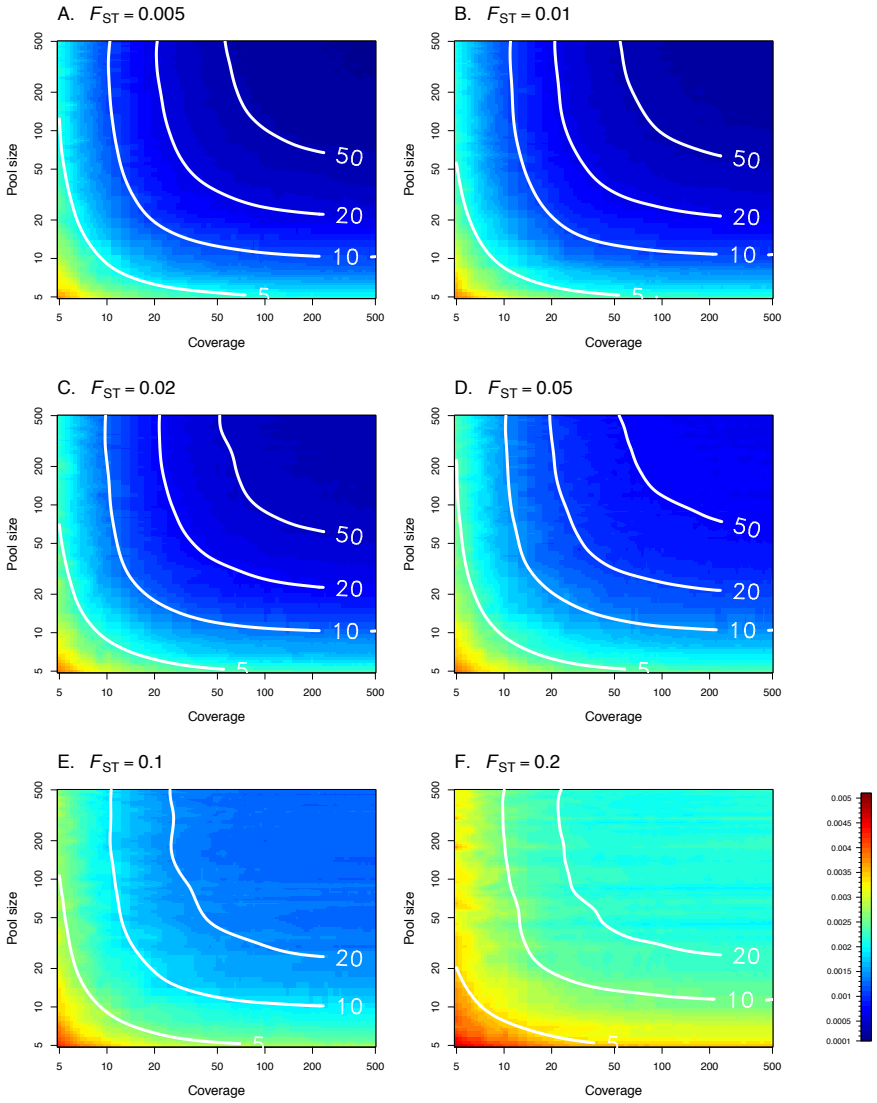


Figure 1.4 Erreur quadratique moyenne (RMSE) des F_{ST} estimés en fonction de la taille de pool et de la couverture, pour un F_{ST} qui varie de 0.005 à 0.2 (A–F). Chaque surface de densité donne le RMSE de notre estimateur $\hat{F}_{ST}^{pool}$, en utilisant un modèle linéaire d'interpolation à partir d'un jeu de 44×44 paires de valeurs de taille de pool et de couverture. Pour chaque taille de pool et couverture, 500 réplicats de 5 000 marqueurs ont été simulés. Les isolignes blanches représentent le RMSE de l'estimateur WC_{84} calculé à partir de données Ind-seq, pour différentes tailles d'échantillon (n = 5, 10, 20, et 50). Les isolignes ont été inférées par une régression par spline avec un paramètre de lissage $\lambda = 0.005$, implémenté dans le package R `field` (Douglas Nychka et al. 2017).

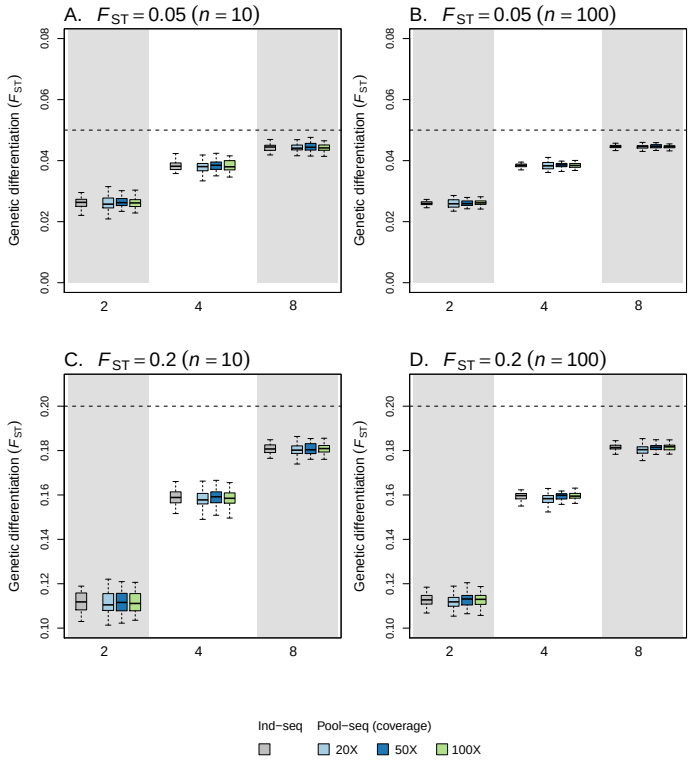


Figure 1.5 Estimations de F_{ST} multilocus en fonction du nombre de dèmes échantillonnés. Nous considérons deux estimateurs, l'un basé sur des données de comptages alléliques obtenues à partir de génotypes individuels (Ind-seq), NC_{83}, l'autre sur des données issues de pools (Pool-seq), FRP_{13}. Chaque boîte à moustaches représente la distribution des F_{ST} multilocus estimés pour toutes les comparaisons entre 2, 4 et 8 dèmes dans un modèle en îles avec $n_{\mathrm{d}} = 8$ dèmes et entre 50 réplicats indépendants de simulations `ms`. Nous avons utilisé deux taux de migration, correspondant à $F_{\mathrm{ST}} = 0.05$ (A–B) et $F_{\mathrm{ST}} = 0.20$ (C–D). La taille de chaque pool était fixée soit à 10 (A et C) soit à 100 (B et D). Pour les données Pool-seq, nous montrons les résultats correspondants à différentes couvertures (20X, 50X et 100X). Dans chaque graphique, les tirets indiquent la valeur théorique de F_{ST}.

Finalement, soulignons que notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ fournit aussi des estimations pour de multiples populations et n'est de ce fait pas restreint à l'analyse de paires de populations, contrairement aux estimateurs de PoPoolation2. Nous montrons que, même pour de faibles tailles d'échantillons ainsi que de faibles couvertures, les estimations Pool-seq sont quasiment indistin-

guables des estimations classiques sur données Ind-seq (voir Tableau 1.4).

1.4.3 Robustesse à des tailles de pool et à des couvertures variables

Nous avons évalué la précision de l'estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ lorsque les tailles d'échantillon varient entre pools, ainsi que lorsque les couvertures varient entre pools et locus (voir Figure 1.6). Nous avons trouvé qu'à faible couverture, des échantillons non-équilibrés ou avec une couverture variable, causent un écart négligeable de la médiane des estimations WC_{84} (Weir et Cockerham 1984) calculées sur des données individuelles. Pour une couverture de 100X, la distribution des estimations $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ est presque indistinguable de celle de WC_{84} (voir Figure 1.6 et Tableaux S2-S3 de l'annexe de l'article de Hivert et al. (2018)).

1.4.4 Robustesse aux erreurs de séquençage et d'expérimentation

La Figure 1.7 montre que les erreurs de séquençage causent un biais négatif négligeable pour les estimations de $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$. L'utilisation d'un filtre (en utilisant un nombre minimum de lectures MRC = 4), améliore légèrement l'estimation, mais cela à forte couverture (voir Figure 1.8 B). Il doit être noté, cependant, que le filtrage augmente le biais en l'absence d'erreurs de séquençage, et ce spécialement à faible couverture (voir Figure 1.8 A). Avec les erreurs expérimentales, c'est-à-dire lorsque les individus ne contribuent pas de manière équivalente au pool final, nous observons un biais positif pour les estimations $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (voir Figure 1.7). On note que le biais décroît avec l'augmentation de la taille des pools. La Figure S2 de l'annexe de l'article de Hivert et al. (2018) montre les RMSE des F_{ST} estimés pour un large panel de taille de pool, de couverture et de taux d'erreur expérimentale (ϵ). Pour $\epsilon \geq 0.25$, augmenter la couverture n'améliore pas la qualité de l'inférence si la taille de pool est trop petite. Lorsque l'étude Pool-seq est soumise à de fort taux d'erreur expérimentale, seule l'augmentation de la taille de pool peut

Tableau 1.4 F_{ST} global estimé à partir de plusieurs pools

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	WC_{84}
0.05	10	20×	0.050 (0.002)	
0.05	10	50×	0.051 (0.002)	0.050 (0.002)
0.05	10	100×	0.050 (0.002)	
0.05	100	20×	0.050 (0.001)	
0.05	100	50×	0.050 (0.001)	0.051 (0.001)
0.05	100	100×	0.050 (0.001)	
0.20	10	20×	0.200 (0.002)	
0.20	10	50×	0.201 (0.002)	0.201 (0.002)
0.20	10	100×	0.201 (0.002)	
0.20	100	20×	0.201 (0.003)	
0.20	100	50×	0.202 (0.003)	0.203 (0.003)
0.20	100	100×	0.203 (0.003)	

Le F_{ST} global a été estimé pour différentes conditions de F_{ST} théorique, de taille de pool (n) et de couverture (Cov.). Pour les données Pool-seq, nous avons calculé notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (voir Équation 1.9). La moyenne (RMSE) sur 50 réplicats indépendants de simulations `ms`sont renseigné, pour l'ensemble des populations ($n_{\mathrm{d}} = 8$). Pour comparaison, nous avons calculé WC_{84} à partir de données de comptages alléliques estimés à partir de génotypes individuels (Ind-seq).

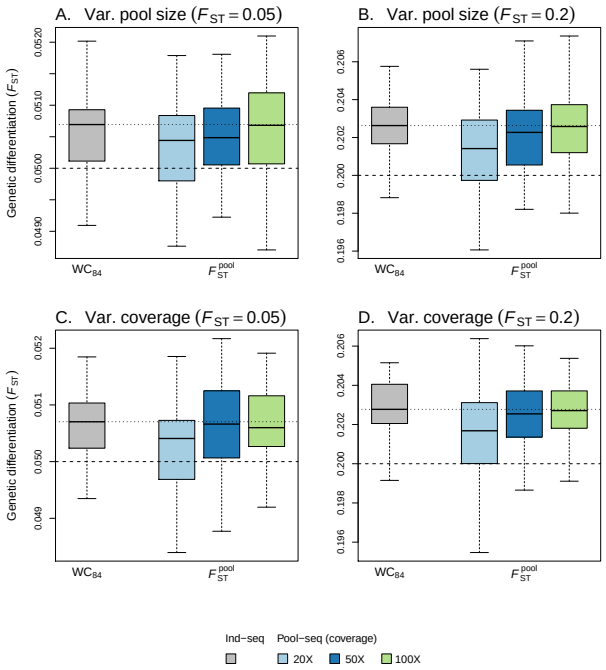


Figure 1.6 Estimateur de F_{ST} global. Notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Équation 1.9) a été calculé à partir de données Pool-seq sur l'ensemble des locus et dèmes et comparé à l'estimateur WC_{84}, appliqué sur données de comptages alléliques obtenues à partir de génotypes individuels (Ind-seq). Chaque boîte à moustaches représente la distribution de F_{ST} multilocus estimés sur 50 réplicats indépendants de simulation avec `ms`. Nous avons utilisé deux taux de migration, correspondant à $F_{\mathrm{ST}} = 0.05$ (A et C) ou $F_{\mathrm{ST}} = 0.20$ (B et D). Dans A–B la taille de pool est variable entre dèmes, avec une taille haploïde d'échantillon n tirée de manière indépendante pour chaque dème dans une distribution gaussienne de moyenne 100 et d'écart-type 30 (n a été arrondi à l'entier le plus proche, avec un minimum de 20 et un maximum de 300 gènes par dème). Dans C–D, la couverture (δ_i) était variable entre les dèmes et locus, avec $\delta_i \sim \mathrm{Pois}(\Delta)$ où $\Delta \in \{20, 50, 100\}$. Pour les données Pool-seq, nous montrons les résultats pour différentes couvertures (20X, 50X and 100X). Dans chaque graphique, les tirets indiquent la valeur théorique de F_{ST} et la ligne en pointillés la médiane de la distribution des estimations issues de WC_{84}.

améliorer la qualité de l'inférence. Filtrer (en utilisant un nombre minimum de lectures MRC = 4) n'améliorant pas l'estimation (voir Figure 1.8 C).

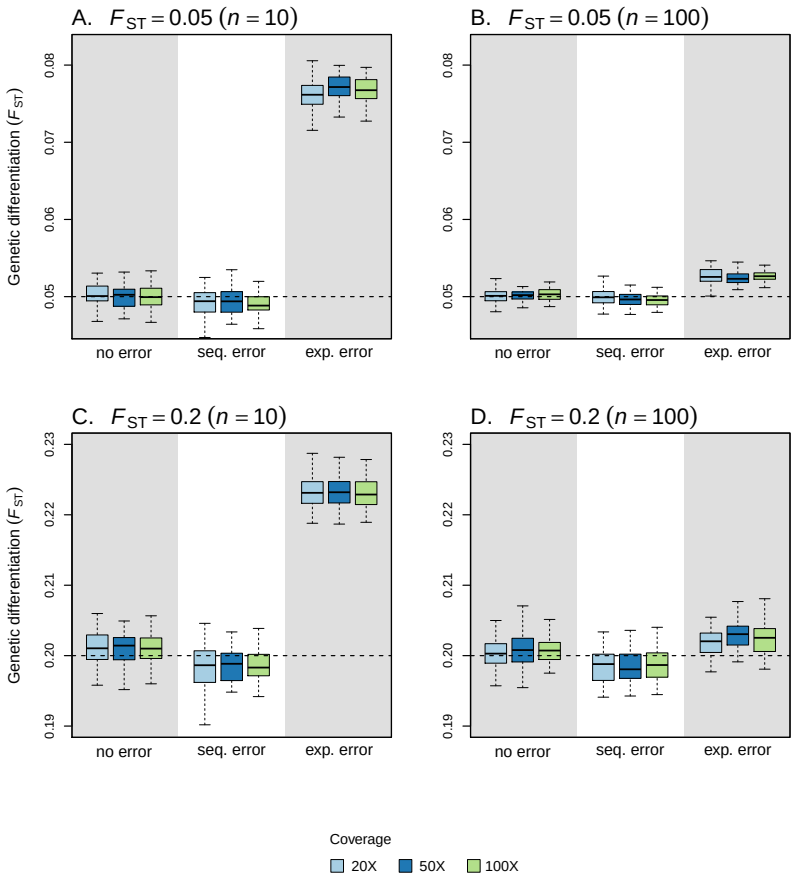


Figure 1.7 Biais et précision des estimations de F_{ST} avec erreurs de séquençage et expérimentale. Notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Équation 1.9) a été calculé à partir de données Pool-seq sur l'ensemble des locus et dèmes sans erreur ou avec erreur de séquençage (apparaissant à un taux $\mu_{\mathrm{e}} = 0.001$) ou et avec erreur expérimentale ($\epsilon = 0.5$). Chaque boîte à moustaches représente la distribution des estimations de F_{ST} multilocus sur 50 réplicats indépendants de simulation avec `ms`. Nous avons utilisé deux taux de migration, correspondant à $F_{\mathrm{ST}} = 0.05$ (A–B) ou $F_{\mathrm{ST}} = 0.20$ (C–D). La taille de chaque pool était fixée soit à 10 (A et C), soit à 100 (B et D). Pour les données Pool-seq, nous montrons les résultats pour différentes couvertures (20X, 50X and 100X). Dans chaque graphique, les tirets indiquent la valeur théorique de F_{ST}.

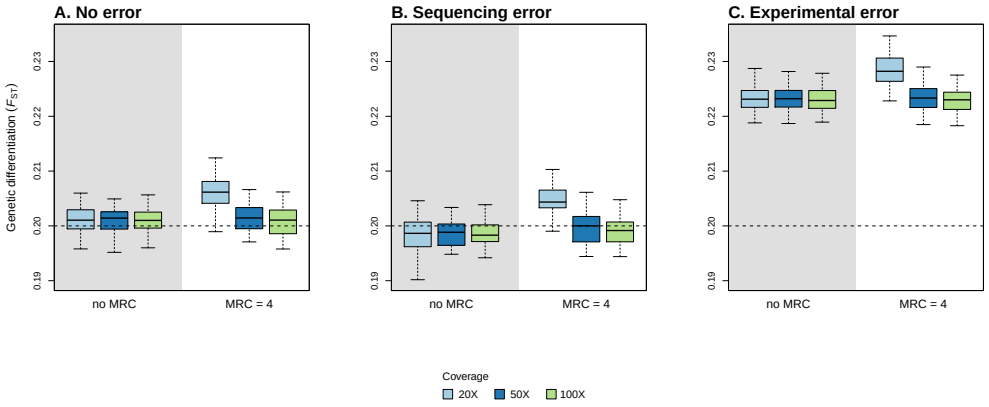


Figure 1.8 Biais et précision des estimations de F_{ST} avec et sans filtre MRC. Notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Équation 1.9) a été calculé à partir de données Pool-seq sur l'ensemble des locus et dèmes sans erreur (A), avec des erreurs de séquençage (B) et avec erreur expérimentale (C) (voir la légende de la Figure 7 pour plus de détails). Pour chaque cas, nous avons calculé le F_{ST} sans filtre (pas de MRC) et avec filtre (en utilisant un nombre minimum de lectures MRC = 4). Chaque boîte à moustaches représente la distribution des F_{ST} multilocus sur 50 réplicats indépendants de simulations `ms`. Nous avons utilisé un taux de migration correspondant à $F_{\mathrm{ST}} = 0.20$ et une taille de pool $n = 10$. Nous montrons les résultats pour différentes couvertures (20X, 50X and 100X). Dans chaque graphique, les tirets indiquent la valeur théorique de F_{ST}.

1.4.5 Exemple d'application

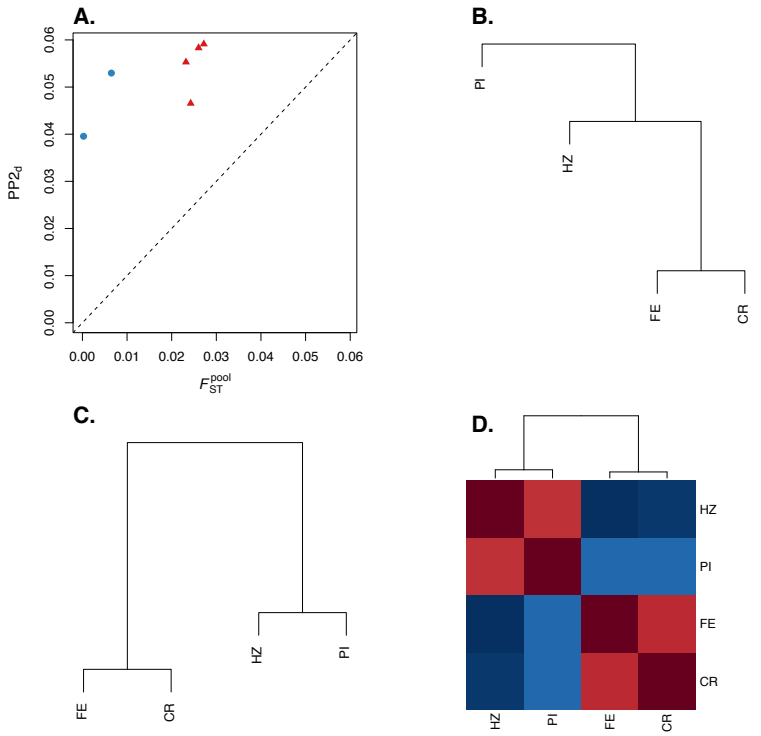


Figure 1.9 Analyse du jeu de données Pool-seq de chabots piquants (*Cottus asper*). Dans (A) nous comparons les estimations de F_{ST} par paires de populations issues des estimateurs $\mathrm{PP2_d}$, et $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Équation 1.9) pour chaque paire de populations échantillonnées en estuaire (CR et FE) et en eau douce (PI et HZ). Les comparaisons intra-écotypes sont représentées avec des ronds bleus tandis que les comparaisons inter-écotypes sont représentées avec des triangles rouges. Dans (B–C) nous montrons des arbres UPGMA construits à partir des estimations par paires $\mathrm{PP2_d}$ (B) et $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (C). Dans (D), nous montrons une représentation de la matrice de covariance entre les quatre populations de *C. asper*, estimée à partir du modèle bayésien hiérarchique implémenté dans le logiciel BayPass.

Nous avons obtenu des estimations de F_{ST} multi-locus par paires plus élevées en utilisant l'estimateur $\mathrm{PP2_d}$ (Kofler et al. 2011) par rapport à l'estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (voir Figure 1.9 A). De plus, les estimations $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ sont plus faibles pour les comparaisons intra-écotype que pour les comparaisons inter-écotypes. De ce fait, les relations inférées basées sur les estimations

$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ montrent une structure des populations marquée, séparant les échantillons d'eau saumâtre des échantillons d'eau douce (voir Figure 1.9 C). Nous n'avons pas retrouvé la même structure en utilisant les estimations de $\mathrm{PP2_d}$ (voir Figure 1.9 B). En appui, la matrice mise à l'échelle de covariance des fréquences alléliques entre échantillons est en accord avec la structure inférée à partir des estimations $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (voir Figure 1.9 D)

1.5 Discussion

Le séquençage des génomes d'individus en pools est de plus en plus utilisé pour la recherche en génomique des populations à la fois sur des espèces modèles et non-modèles (Schlötterer et al. 2014). Le développement de logiciels dédiés (listés dans Schlötterer et al. 2014) a sans aucun doute joué un rôle dans la diversité des questions de recherche abordées en utilisant le séquençage en pool. Cependant, l'analyse de la structure des populations à partir de données Pool-seq est compliquée compte tenu du double processus d'échantillonnage des gènes à partir des pools puis des lectures à partir de ces gènes (Ferretti et al. 2013).

L'approche naïve qui consisterait à calculer le F_{ST} à partir des lectures comme si elles étaient des comptages alléliques (comme dans Chen et al. 2016) ignore l'erreur supplémentaire apportée par l'échantillonnage aléatoire des lectures à partir des pools de gènes durant le séquençage, ce qui peut amener à de sévères biais d'estimations de la différenciation lorsque la taille de pool est faible (voir la Figure S2 de l'annexe de l'article de Hivert et al. 2018). D'autre part, l'approche alternative qui viserait à imputer les comptages alléliques à partir des comptages de lectures, sachant la taille haploïde des pools (comme dans Leblois et al. 2018; Smadja et al. 2012) ne peut être efficace que lorsque la couverture est bien plus grande que la taille de pool (voir la Figure S2 de l'annexe de l'article de Hivert et al. 2018).

Ici, nous avons développé un nouvel estimateur du paramètre F_{ST} pour données Pool-seq, dans un cadre d'analyse de la variance (Cockerham 1969, 1973). La précision de l'estimateur est à peine différentiable de celui de Weir et Cockerham (1984) pour données individuelles. De plus, nous avons montré

que cette précision ne dépendait ni de la taille de pool ni de la couverture, et qu'elle était robuste à des tailles de pools inégales ainsi qu'à des couvertures variables entre dèmes et loci.

1.5.1 Analyse de variance et probabilités d'identité

Dans un cadre d'analyse de variance, le F_{ST} est défini dans l'Équation 1.7 comme la corrélation intra-classe des probabilités d'identité par état (Cockerham et Weir 1987; Rousset 1996). En dehors des estimateurs d'analyse de variance introduits en génétique des populations par Cockerham (1969, 1973), des estimateurs basés sur le calcul de probabilités d'identités intra et inter groupes ont été proposés (voir, par exemple, Fleiss et Cuzick 1979; Mak 1988; Ridout et al. 1999). Ces estimateurs ont ensuite été utilisés en génétique des populations, où la "probabilité de réponse identique" a été interprétée comme la fréquence à laquelle les gènes sont IIS (Cockerham et Weir 1987; Rousset 2007; Weir 1996; Weir et Goudet 2017).

Cela suggère qu'avec les données Pool-seq, une autre stratégie pourrait consister à calculer les F_{ST} à partir des probabilités IIS entre paires de gènes (non observées), ce qui nécessite que des estimations non biaisées de ces quantités soit obtenues à partir des données de comptages des lectures. C'est ce que nous avons fait dans l'annexe de l'article de Hivert et al. (2018), où nous fournissons un estimateur non biaisé de F_{ST} (voir les Équations A44 et A48 dans l'annexe de l'article de Hivert et al. 2018). Ces estimateurs, notés $\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$ et $\tilde{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$), ont strictement la même forme que l'estimateur basé sur l'analyse de variance si les pools ont la même taille et que le nombre de lectures par pool est constant (voir l'Équation A33 dans l'annexe de l'article de Hivert et al. 2018). Cela fait écho aux dérivations faites par Rousset (2007) pour des données Ind-seq, qui montrent que l'approche par analyse de variance (Weir et Cockerham 1984) et la stratégie simple qui constitue à estimer les probabilités IIS en dénombrant les paires de gènes identiques donnent des estimations identiques lorsque les tailles d'échantillons sont égales (voir également Cockerham et Weir 1987; Karlsson et al. 2007).

Avec des échantillons non-équilibrés, nous avons trouvé que l'approche

d'analyse de variance montre un biais et une variance plus faible que celle basée sur les probabilités IIS (voir Figure S4). De manière intéressante, nous avons trouvé que les estimations de F_{ST} pour données Pool-seq basées sur les probabilités IIS montraient généralement un biais et une variance plus faibles lorsque l'ensemble des probabilités IIS intra- et inter-pools étaient calculées comme des moyennes non-pondérées (voir les Équation A39 et A43 dans l'annexe de l'article de Hivert et al. 2018) (comme dans Fleiss et Cuzick 1979). L'Équation A28 montre que notre estimateur peut être réécrit comme une fonction proche de $\left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$, qui dépendrait néanmoins de la somme $\sum_i \left(\hat{Q}_{1i} - \hat{Q}_1\right)$ au numérateur et au dénominateur. Cela suggère que si les Q_{1i} diffèrent entre populations, alors notre estimateur est une fonction de F_{ST} population-spécifiques (Weir et Goudet 2017; Weir et Hill 2002).

Les dérivations de l'annexe de l'article de Hivert et al. (2018) montrent que l'estimateur $\mathrm{PP2_a}$ (Équation 1.15) est biaisé car l'estimateur de probabilités IIS entre paires de lectures au sein d'un pool $\left(\hat{Q}_1^{\mathrm{r}}\right)$ est un estimateur biaisé des probabilités IIS entre paires de gènes distincts au sein du pool (voir l'Équation A34 dans l'annexe de l'article de Hivert et al. 2018). La quantité $\hat{Q}_1^{\mathrm{r}}$ confond en effet les paires de lectures qui sont identiques parce qu'elles ont été séquencées à partir d'une seule copie de gène et celles qui sont identiques parce qu'elles ont été séquencées à partir de deux gènes distincts mais identiques par état (IIS).

Un estimateur de F_{ST} mieux justifié a été proposé par Ferretti et al. (2013), d'après les développements de Futschik et Schlötterer (2010). Notons que, bien qu'ils définissent le F_{ST} comme un ratio de fonctions d'hétérozygotie, ils ont dans les faits utilisés des probabilités IIS (voir Équations 1.16 et 1.17). Cependant, bien que leur Équation 1.16 soit strictement équivalente à notre Équation A39 dans l'annexe de l'article de Hivert et al. (2018), l'hétérozygotie totale est calculée en intégrant sur l'ensemble des paires intra- et inter-pools (voir Équation 1.17), ce qui peut expliquer le biais observé dans la Figure 2 et la dépendance au nombre de populations échantillonnées observée dans la Figure 3.

1.5.2 Comparaison avec des estimateurs alternatifs

Un cadre de travail alternatif à l'analyse de variance de Weir et Cockerham (1984) a été développé par Masatoshi Nei et collaborateurs pour estimer le F_{ST} à partir de diversité génique (Nei 1973, 1977, 1986; Nei et Chesser 1983). L'estimateur $PP2_d$ (voir Équations 1.12–1.14) implémenté dans le logiciel PoPoolation2 (Kofler et al. 2011) suit cette logique. Cependant, il est reconnu depuis longtemps que les deux cadres de travail sont fondamentalement différents. L'approche par analyse de variance considère à la fois l'échantillonnage statistique et génétique (ou évolutif), tandis que Nei et collaborateurs ne le font pas (Excoffier 2007; Holsinger et Weir 2009; Weir et Cockerham 1984). De plus, l'attendu de l'estimateur de Nei et collaborateurs dépend du nombre de populations échantillonnées, avec un biais important pour un faible nombre de populations échantillonnées (Excoffier 2007; Goudet 1993; Weir et Goudet 2017). Par conséquent, nous ne recommandons pas l'utilisation de l'estimateur $PP2_d$ implémenté dans le logiciel PoPoolation2 (Kofler et al. 2011). Enfin, comme nous l'avons discuté précédemment, l'estimateur de Ferretti et al. (2013) bien que pouvant être appliqué sur plusieurs populations, a le même défaut que son pendant Ind-seq.

1.5.3 Application pour les études d'écologie évolutive

Le séquençage en pool est de plus en plus utilisé dans différents domaines d'application (Schlötterer et al. 2014), tels que la génétique de la conservation (voir, par exemple, Fuentes-Pardo 2017), la biologie de l'invasion (voir, par exemple , Dexter et al. 2017) et la biologie évolutive dans un sens plus large (voir, par exemple , Collet et al. 2016). Ces études utilisent un vaste panel d'outils, qui visent à caractériser la structure de populations à une échelle très fine (voir, par exemple , Fischer et al. 2017), la reconstruction d'histoires démographiques (voir, par exemple , Chen et al. 2016; Leblois et al. 2018), ou encore la recherche de signatures de sélection naturelle ou artificielle (voir, par exemple , Chen et al. 2016; Fariello et al. 2017; Leblois et al. 2018). Ici, nous avons réanalysé le jeux de données Pool-seq produit par Dennenmoser et al. (2017), qui ont étudié la divergence génomique adaptative entre des

écotypes d'eau douce et d'eau saumâtre chez le chabot piquant *C. asper*, un poisson euryhalin abondant du Nord-Ouest de l'Amérique du Nord. En mesurant la différenciation génétique par paires entre les échantillons à l'aide de l'estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$, nous avons trouvé une structure claire séparant les deux écotypes, eau douce et eau saumâtre. Une telle structure génétique est supportée par l'hypothèse que les populations sont localement adaptées à leurs conditions osmotiques dans ces deux habitats contrastés, comme cela a été discuté dans Dennenmoser et al. (2017). Cette structure qui est en désaccord avec celle inférée par les estimations issues de $\mathrm{PP2_d}$, n'est pas seulement supportée par la matrice de covariance des fréquences alléliques, mais aussi par des analyses préalables sur des marqueurs microsatellites, qui ont montré que les populations étaient génétiquement plus différenciées entre écotypes qu'au sein des écotypes (Dennenmoser et al. 2015, 2014).

1.5.4 Limites du modèle et perspectives

Nous avons montré que la source de biais la plus importante pour les estimations de $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ est la contribution inégale des individus aux pools. Cela s'explique par le fait que nous supposons que les lectures sont distribuées selon une loi multinomiale, ce qui suppose que l'ensemble des gènes contribue de manière égale au pool de lectures (Gautier et al. 2013), c'est-à-dire qu'il n'y a pas de variation de quantités d'ADN entre les individus et que la couverture de séquençage est égale pour l'ensemble des gènes (Rode et al. 2017). Puisque l'effet des contributions inégales est supposé plus important pour de faibles tailles de pool, il a été recommandé d'utiliser au moins 50 individus diploïdes par pool (Lynch et al. 2014; Schlötterer et al. 2014). Cependant, cette limite peut-être trop conservative pour l'estimation des fréquences alléliques (Rode et al. 2017), et nous avons montré ici que nous pouvons obtenir une excellente précision d'estimation du F_{ST} avec des tailles de pools allant jusqu'à $n = 10$ diploïdes. De plus, puisque les informations génotypiques sont perdues durant le processus d'expérimentation Pool-seq, nous supposons dans nos développement que les pools sont constitués d'haploïdes (et donc que le F_{IS} est nul). L'analyse de populations non-panmictiques (e.g, chez des espèces

autogames) est donc problématique.

Notre modèle, comme dans Weir et Cockerham (1984), suppose formellement que l'ensemble des populations sont des réplicats indépendants du même processus évolutif (Holsinger et Weir 2009). Ceci est plutôt irréaliste dans nombre de populations naturelles, ce qui a motivé Weir et Hill (2002) à développer un estimateur de F_{ST} population-spécifique pour données individuelles (voir ausi Vitalis et al. 2001). Bien que l'utilisation de l'estimateur de Weir et Hill (2002) reste rare dans la littérature (mais voir Vitalis 2012; Weir et al. 2005), Weir et Goudet (2017) ont récemment proposé une réinterprétation des estimations populations-spécifiques de F_{ST} en terme de proportions de correspondances alléliques, ce qui est strictement équivalent aux probabilités IIS entre paires de gènes. Il serait ainsi possible d'étendre l'estimateur de F_{ST} population-spécifique de Weir et Goudet (2017) pour l'analyse de données Pool-seq, en utilisant un estimateur non biaisé de probabilité IIS tel que celui fourni dans l'annexe de l'article de Hivert et al. (2018).

Finalement, si nous n'avons pas évoqué le calcul d'intervalles de confiance pour les mesures de F_{ST} pour lesquels Leviyang et Hamilton (2011) ont étudié les propriétés de plusieurs techniques de bootstrapping. Weir et Cockerham (1984) conseillent en effet de construire ces intervalles en rééchantillonnant les loci. Cependant, l'étude réalisée se place dans un contexte de faible nombre de marqueurs (< 20) non liés, bien loin de la quantité de marqueurs NGS qui constituent les jeux de données actuels. Ainsi, il pourrait être intéressant d'étudier les propriétés de différentes méthodes de construction d'intervalles de confiance sur des données génomiques, et plus particulièrement en présence de déséquilibre de liaison, afin de pouvoir identifier la méthode la plus adaptée à ce type de données.

Chapitre 2

Un modèle hiérarchique bayésien pour détecter l'adaptation locale à partir de données haplotypiques

2.1 Introduction

L'identification des "signatures de sélection" dans les génomes est un sujet central en biologie évolutive puisqu'elle est un élément clé dans la compréhension de la réponse des individus aux conditions environnementales, variables dans le temps et l'espace (Bernatchez 2016). L'adaptation locale peut aussi être un moteur de la spéciation (Kirkpatrick et Barton 2006; Schluter 2001), et est étudiée dans de nombreux domaines, que ce soit en sélection animale (Bernatchez 2016; Druet et al. 2014), en gestion forestière (Rellstab et al. 2016) ou en épidémiologie et recherche médicale, à travers l'étude de l'évolution des pathogènes et l'identification de gènes sous sélection en vue d'une lutte contre ceux-ci (Montano et al. 2015).

En 1966, Luigi Luca Cavalli-Sforza (Cavalli-Sforza 1966) notait : "if we analyse a species the demographic structure of which can be accurately studied [...] then we can predict the exact amount of variation due to drift and

can compare this with the observed one. The likelihood that geographical or temporal variation of selective coefficients simulate exactly this expectation will become smaller the more independent gene systems we examine, as the expectation of drift, unlike selective variation, will be the same for all genes". Détecter des signatures génomiques de sélection revient *in fine* à distinguer, parmi les forces agissant sur l'évolution des fréquences alléliques, celles qui ont un effet global sur l'ensemble du génome (migration et dérive), de celles qui ont un effet local (sélection et mutation). Pour cela, nous devons être capables de caractériser la variabilité génétique à l'échelle du génome. Si ces méthodes ont souvent été limitées par des aspects techniques, le développement des technologies de génotypage et de séquençage à haut débit (NGS pour *Next Generation Sequencing*) permet désormais de caractériser la variabilité génétique à une échelle pan-génomique chez de nombreuses espèces (Andrews et Luikart 2014; Belcaid et Toonen 2014), y compris chez les espèces "non-modèles" pour lesquelles les ressources génomiques sont encore limitées. Ainsi, les données de type NGS ont motivé le développement de nombreuses méthodes de détection de signatures de sélection au cours de ces dernières années (Hoban et al. 2016; Savolainen et al. 2013).

Nombreuses sont celles basées sur des mesures de F_{ST} visant à identifier des locus atypiques par rapport à l'hypothèse de neutralité, en supposant que la majorité des locus sont neutres. Comme mentionné en Introduction Générale, si les approches qui utilisent la distribution empirique des F_{ST} pour identifier les marqueurs sous sélection (Akey et al. 2002; Weir et al. 2005) sont computationnellement très efficaces, elles ne permettent cependant pas de contrôler le degré de fausses découvertes. Lewontin et Krakauer (1973) avaient dans ce but, proposé un test en modélisant la distribution neutre des F_{ST}. Cependant la modélisation trop simpliste des F_{ST} rendait leur test sensible aux effets confondants de la démographie et de la structure des populations (Hoban et al. 2016; Savolainen et al. 2013) l'exposant très tôt à de vives critiques (Beaumont et Nichols 1996; Nei et Maruyama 1975; Robertson 1975), tout en motivant le développement de nouvelles méthodes (voir Introduction Générale 0.3.2).

Différentes approches ont été proposées, afin de prendre en compte la dé-

mographie des populations en supposant un modèle démographique neutre de dérive pure (Gautier et al. 2010; Nicholson et al. 2002) ou encore un modèle en îles à l'équilibre migration-dérive (Beaumont et Balding 2004; Foll et Gaggiotti 2008; Guo et al. 2009; Riebler et al. 2008), dont la distribution *a priori* des fréquences alléliques est issue d'une approximation de diffusion (voir Équation 2 dans l'Introdution Générale). C'est dans ce contexte que s'est inscrit le développement de SelEstim (Vitalis et al. 2014). Ce modèle hiérarchique bayésien distingue les effets globaux (migration et dérive) des effets locaux (sélection) sur les fréquences alléliques, en modélisant explicitement l'effet de la sélection. De fait, on considère dans ce modèle que l'ensemble des marqueurs sont soumis, dans une certaine mesure à la sélection, modélisée avec des coefficients de sélection locus- et population-spécifiques.

Cependant la plupart de ces modèles négligent encore souvent l'information potentiellement apportée par le déséquilibre de liaison (DL) entre les marqueurs génétiques, en les considérant comme "indépendants" alors qu'il existe une liaison physique entre ceux-ci (voir Introduction Générale : "Tirer profit des données à l'ère des NGS").

Or, les balayages sélectifs, par l'effet d'auto-stop génétique, créent des patrons caractéristiques de DL au voisinage des régions sous sélection (Maynard Smith et Haigh 1974; Stephan et al. 2006), qui vont ainsi montrer un écart à l'hypothèse d'évolution neutre (Pritchard et al. 2010). Afin d'exploiter cette information de DL dans la recherche de régions génomiques sous sélection, plusieurs approches ont été proposées, que l'on peut décliner en trois catégories dans le contexte des *scans* génomiques de différenciation.

(i) Les méthodes de lissages en post-traitement

Les statistiques locus-spécifiques montrent souvent une grande variance d'un locus à l'autre. Weir et al. (2005) ont proposé de réaliser des mesures de F_{ST} multi-locus le long de fenêtres glissantes (de 5 Mb dans leur cas) afin de réduire la variance des estimations et donc le bruit de fond (voir Figure 2.1). Ils ont considéré comme extrêmement différenciée, toute fenêtre dont le F_{ST} est supérieur à trois fois l'écart-type mesuré à partir de la distribution empirique de F_{ST} pour chaque chromosome. Myles et al. (2008) ont quant à

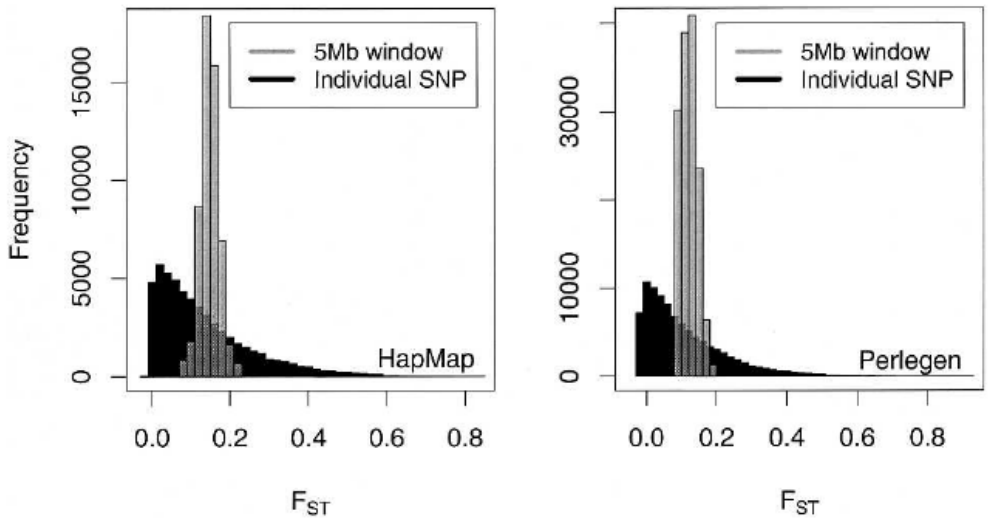


Figure 2.1 Distributions de F_{ST} locus-spécifiques (en noir) et mesurés le long de fenêtres de 5Mb (en gris) dans le génome humain. On voit distinctement une réduction de la variance pour ces derniers contrairement à l'approche locus-spécifique. Figure tirée de Weir et al. (2005)

eux choisis de considérer comme très différenciées les fenêtres dont la valeur se trouve dans les 10% maximum de la distribution empirique de F_{ST}. L'objectif est d'identifier des régions de différenciation extrême, en espérant qu'elles soient causées par la sélection et non par l'histoire démographique des populations. Si d'autres auteurs ont proposé, sur le même principe, des variantes, que ce soit en terme de statistiques utilisées ou de définition des fenêtres, nous sommes toujours confronté à la même limite. La définition des paramètres liés à la construction des fenêtres (comme leur taille) et le choix des critères de décision sont très arbitraires.

Plus récemment, Fariello et al. (2017) ont proposé une approche qui combine, au sein de fenêtres, les signaux locus-spécifiques d'une statistique choisie sous forme de score local, fonction de la somme des statistiques observées. Les auteurs appuient que le choix du critère de décision utilisé repose sur une base statistique solide au regard de l'utilisation de p.valeurs et de la définition du modèle neutre. La méthode repose cependant sur le choix du paramètre dit "de tuning", dont le choix peut fortement influencer la position des régions identifiées ce qui peut s'avérer problématique d'un point de vue opérationnel.

(ii) les méthodes de lissage intégrées

Une approche alternative consiste à incorporer directement les méthodes de lissage au sein des modèles. Le principe est de considérer que deux marqueurs voisins ont une probabilité plus élevée de montrer des patrons de différenciation similaires entre populations.

L'idée est donc de tenir compte de la corrélation statistique qui existe entre marqueurs voisins dans le calcul d'un ou plusieurs paramètres locus-spécifiques et d'encourager l'agrégation des marqueurs atypiques le long du génome. Dans le contexte des modèles bayésiens, cela s'est par exemple traduit par l'utilisation de modèles spatiaux auto-régressifs. À titre d'exemple, Guo et al. (2009) ont introduit un modèle auto-régressif conditionnel (CAR) au sein du "F-model" de sorte que pour chaque marqueur, son effet locus-spécifique sur le F_{ST} tient compte des effets des marqueurs voisins. Dans le même esprit, Duforet-Frebourg et al. (2014) ont quant à eux utilisé un modèle 1D de Ising et Potts afin d'encourager le modèle à discriminer comme étant sous sélection des marqueurs voisins.

Contrairement aux approches par fenêtres, les méthodes intégrées de lissage prennent explicitement en compte la corrélation entre les marqueurs dans le calcul des statistiques locus-spécifiques et ne nécessitent pas à proprement parler de définir une taille de fenêtre. Cependant, elles sont également soumises à la définition préalable, et parfois subjective, de l'intensité de la corrélation spatiale qui existe entre les marqueurs. De plus, ces approches représentent une modélisation "simpliste" du DL en considérant uniquement les corrélations des fréquences alléliques et non des fréquences haplotypiques sous-jacentes.

(iii) les méthodes exploitant l'information haplotypiques

L'ensemble des méthodes vues jusqu'à présent exploite l'information de déséquilibre de liaison de différentes manières à travers une mesure des relations des fréquences alléliques à différents marqueurs. Néanmoins, comme nous avons pu le voir aucune ne s'intéresse directement aux informations portées par les données génétique et plus particulièrement à la structure des haplotypes. Or, celle-ci représente le résultat le plus direct de l'effet conjoint

des différentes forces évolutives et du déséquilibre de liaison sur le génome, ce qui représente donc un atout pour la détection de la sélection. En effet, dans le cas où une nouvelle mutation se retrouve sous sélection, elle va rapidement augmenter en fréquence (voir Introduction Générale : "Détecter la sélection à partir de la différenciation génétique"), entraînant avec elle les mutations neutres adjacentes, par l'effet d'auto-stop génétique, qui n'auront pas eu le temps de recombiner. Cela va résulter en un haplotype d'autant plus long que l'intensité de la sélection sera importante et le taux de recombinaison faible. Partant de ce constat, plusieurs modèles ont été développés afin d'exploiter directement la structure des haplotypes au sein d'une population (Sabeti et al. 2002; Voight et al. 2006), ou d'une paire de populations (Sabeti et al. 2007; Tang et al. 2007).

En 2010, Browning et Weir (2010) ont proposé de réaliser un *scan* génomique de F_{ST} en regroupant localement des haplotypes similaires autour de chaque SNP et en considérant chaque groupe comme autant d'allèles pour former des locus multialléliques. Leur approche a ouvert la voie à l'analyse de l'information haplotypique dans un contexte de populations structurées. Sur le même principe, Fariello et al. (2013) ont développé HapFLK, une extension haplotypique du modèle FLK de Bonhomme et al. (2010), qui prend en compte la structure hiérarchique des populations. Ils ont ainsi montré que l'utilisation de données haplotypiques dans un modèle de mesure de différenciation entre populations sous l'hypothèse de dérive pure, améliore sensiblement la puissance de détection des régions sous sélection par rapport à l'analyse de SNPs considérés indépendants (voir la Figure 2.2 pour une comparaison des différents *scans* génomiques FLK et HapFLK).

Parmi les différentes approches évoquées pour prendre en compte l'information de DL dans les méthodes de détection de la sélection, l'utilisation de données haplotypiques paraît être la plus puissante. La création de marqueurs haplotypique peut se faire de différentes manières. L'approche par blocs qui semble être la plus intuitive consiste à définir des haplotypes en mettant en place une fenêtre autour des SNPs, dont la taille est fixée par l'utilisateur. Cela peut-être un certain nombre de SNPs, une taille physique, ou encore un nombre d'allèles minimal à obtenir. Si elles sont très simples à mettre en place,

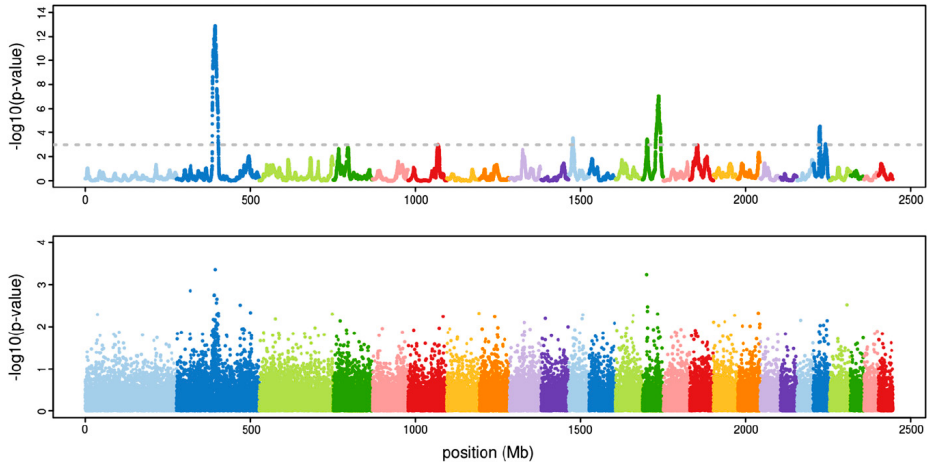


Figure 2.2 Adaptée de Fariello et al. (2013). La figure représente deux *scans* génomiques réalisés sur des données de moutons Nord européens avec FLK et HapFLK. Le *scan* HapFLK (en haut) montre des signaux de sélection beaucoup plus clairs que FLK (en bas), principalement sur les chromosomes 2, 14 et 22. Le signal est quant à lui aussi plus lissé avec HapFLK que FLK.

ces approches ne font pas preuve d'une très grande flexibilité et ne sont pas très appropriées pour prendre en compte les variations locales de DL. Ainsi, des approches basées sur des chaînes de Markov cachées (HMM) ou encore des chaînes de Markov de tailles variables (VLMC) ont été développées respectivement par Scheet et Stephens (2006) et Browning et Browning (2007) afin de phaser les données de génotypages et réaliser un groupement local des haplotypes. Ces méthodes, plus flexibles que les précédentes, prennent mieux en compte les variations locales de DL et regroupent des haplotypes similaires dans un même groupe en tenant compte de la corrélation qui existe entre marqueurs voisins. Cela représente une différence majeure par rapport aux approches par blocs, pour lesquelles une seule mutation dans un bloc haplotypique crée un nouvel allèle (voir Figure 2.3, pour un exemple de groupement local des haplotypes réalisé par `fastphase` (Scheet et Stephens 2006). La Figure 2.4 représente un exemple de groupement local des haplotypes par blocs.

Bien que prenant en compte l'information de structure des haplotypes, ainsi que la structure des populations, HapFLK ne permet pas d'estimer l'intensité de la sélection à chaque locus et dans chaque population (même si une

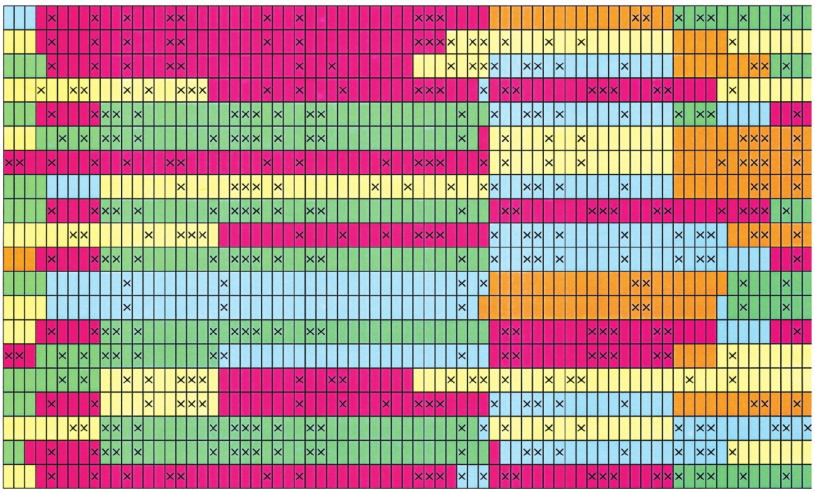


Figure 2.3 Exemple de groupements locaux d'haplotypes réalisés par `fastphase`. La figure est issue de l'article de Scheet et Stephens (2006) et de l'analyse des données humaines sur 60 individus indépendants. Ici, 10 individus diploïdes sont représentés, chaque ligne correspondant à un chromosome. Chaque colonne correspond à un SNP dont les deux allèles sont représentés par une case vide ou cochée. Enfin, les couleurs correspondent aux différents groupes haplotypiques inférés par `fastphase` parmis les K groupes possibles.

étape de post-traitement permet d'identifier les populations sous sélection à un marqueur donné), à l'instar du modèle SelEstim (Vitalis et al. 2014). Les deux modèles reposent sur des hypothèses démographiques différentes : des populations ayant divergé et qui évoluent en dérive pure pour HapFLK ; un modèle à nombre infini d'îles qui échangent des migrants pour SelEstim.

Ici, nous proposons donc d'étendre le modèle hiérarchique bayésien SelEstim (Vitalis et al. 2014) à l'utilisation des marqueurs multialléliques. Notre but est d'exploiter l'information de DL potentiellement présente dans les données, en définissant des marqueurs multialléliques par une méthode naïve de regroupement local d'haplotypes. La stratégie employée par HapFLK n'est pas réalisable avec SelEstim d'un point de vue computationnel puisqu'elle demanderait plusieurs analyses indépendantes d'un même jeu de données. Nous avons donc opté pour une autre stratégie de regroupement local par construction de blocs haplotypiques : pour un nombre K d'allèles minimum à considérer, une fenêtre est créée autour de chaque SNP, la taille

de la fenêtre étant adaptée de part et d'autre du marqueur jusqu'à obtenir au moins K allèles (voir Figure 2.4). Les blocs haplotypiques sont ensuite considérés comme autant de marqueurs multialléliques pour les analyses SelEstim.

Si cette approche est moins flexible que `fastphase` pour capturer l'information de DL local, dont l'étendue peut varier le long du génome ; elle est cependant plus stable et nous permet d'éviter les problèmes de maximums locaux observés avec les méthodes HMM telles que `fastphase` (Scheet et Stephens 2006) ou encore Beagle (Browning et Weir 2010).

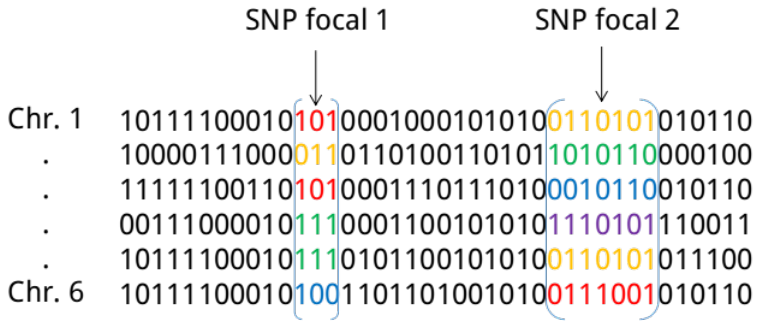


Figure 2.4 Illustration de l'algorithme de regroupement local des haplotypes par blocs. Des haplotypes locaux autour de deux marqueurs focaux (indiqués par les flèches) issus de 6 haplotypes (Chr.) sont construits dans le but d'obtenir au moins $K = 4$ allèles. Dans le cas du premier SNP focal, une fenêtre de 3 SNPs (1 de chaque côté du SNP focal) permet d'obtenir exactement 4 haplotypes, considérés comme autant d'allèles. À l'inverse, pour le deuxième SNP focal, une fenêtre de 7 SNPs résulte en 5 allèles, une taille de fenêtre inférieure ne permettant pas d'obtenir le nombre minimum de 4 allèles.

2.2 Description du modèle SelEstim multiallélique

Nous allons maintenant décrire le modèle bayésien hiérarchique SelEstim permettant d'utiliser des données multialléliques. Par soucis de clarté, nous nommerons dorénavant ce modèle $\text{SelEstim}_{\text{HAP}}$, par opposition au modèle développé uniquement pour des données bialléliques indépendantes de

type SNP par Vitalis et al. (2014), que nous nommerons SelEstim$_{\text{SNP}}$.

On considère des données génomiques sur n_d dèmes pour J locus ayant chacun un nombre d'allèles K_j. On peut noter x_{ijk} le nombre d'allèles k échantillonnés au locus j dans le dème i. On obtient ainsi le vecteur de comptages alléliques pour le dème i au locus j $\boldsymbol{n}_{ij} \equiv (x_{ij1}, x_{ij2}, ..., x_{ijK_j})$ et le nombre total de comptages $n_{ij} = \sum_{k=1}^{K} x_{ijk}$.

SelEstim considère un modèle démo-génétique à nombre infini d'îles. Ainsi, pour le dème i, on définit le paramètre de migration mis à l'échelle $M_i \equiv 4N_i m_i$, avec N_i la taille diploïde totale du dème i et m_i le taux d'immigration total vers le dème i.

On note p_{ijk} la fréquence de l'allèle k dans le dème i et π_{jk} sa fréquence dans la métapopulation (i.e., dans le pool de migrants sous l'hypothèse d'un nombre infini d'îles). De même, $\boldsymbol{p}_{ij} \equiv (p_{ij1}, p_{ij2}, ..., p_{ijK_j})$ est le vecteur des fréquences alléliques au locus j dans le dème i et $\boldsymbol{\pi}_j \equiv (\pi_{j1}, \pi_{j2}, ..., \pi_{jK_j})$ est le vecteur des fréquences alléliques dans le pool de migrants.

Un point-clé du modèle SelEstim est que nous considérons que l'ensemble des locus j est soumis, dans une certaine mesure, à la sélection. À chaque locus et dans chaque dème, on considère qu'un seul et unique allèle est sélectionné, référencé par la variable indicatrice $\kappa_{ij} \in \lceil 1, K_j \rceil$.

Ainsi pour chaque dème i au locus j on peut définir $\tilde{p}_{ij} = p_{ij\kappa_{ij}}$ la fréquence de l'allèle sous sélection au locus j dans le dème i, et $\tilde{\pi}_j = \pi_{j\kappa_{ij}}$ sa fréquence dans le pool de migrants. On définit le coefficient de sélection mis à l'échelle dans le dème i au locus j comme $\sigma_{ij} \equiv 2N_i s_{ij}$, avec s_{ij} l'intensité de la sélection au locus j dans le dème i. Dans le dème i au locus j, un homozygote pour l'allèle sous sélection aura une valeur sélective de $1 + s_{ij}$, un hétérozygote portant un allèle sous sélection aura une valeur sélective de $1 + s_{ij}/2$, et l'ensemble des autres génotypes possibles auront une valeur sélective égale à 1.

Conditionnellement au vecteur de fréquences alléliques $\boldsymbol{p}_{ij}$, les comptages $\boldsymbol{n}_{ij}$ au locus j dans le dème i sont supposés suivre une distribution multinomiale :

Tableau 2.1 Résumé des notations principales

Notation	Description
n_{ij}	Taille haploïde (nombre de gènes) de l'échantillon dans le dème i au locus j
$\boldsymbol{n}_{ij}$	Vecteur des comptages alléliques du dème i échantillonné au locus j
N_i	Effectif diploïde total du dème i
$M_i \equiv 4N_i m_i$	Paramètre de migration mis à l'échelle du dème i, avec m_i le taux d'immigration total dans le dème i
p_{ijk}	Fréquence de l'allèle k au locus j dans le dème i
$\tilde{p}_{ij}$	Fréquence de l'allèle sous sélection au locus j dans le dème i
$\boldsymbol{p}_{ij}$	Vecteur des fréquences alléliques au locus j dans le dème i
π_{jk}	Fréquence de l'allèle k au locus j dans le pool de migrants
$\tilde{\pi}_j$	Fréquence de l'allèle sous sélection au locus j dans le pool de migrants
$\boldsymbol{\pi}_j$	Vecteur des fréquences alléliques au locus j dans le pool de migrants
κ_{ij}	Variable indicatrice de l'allèle sous sélection au locus j dans le dème i
$\sigma_{ij} \equiv 2N_i s_{ij}$	Coefficient de sélection mis à l'échelle du locus j dans le dème i, avec s_{ij} le coefficient de sélection
δ_j	Coefficient de sélection au locus j dans la métapopulation
λ	Effet de la sélection sur l'ensemble du génome pour tout les dèmes et loci échantillonnés. Écart à l'hypothèse de modèle en îles.

$$\mathcal{L}(\boldsymbol{p}_{ijk};\boldsymbol{n}_{ij}) = \frac{n_{ij}!}{\prod_{k=1}^{K_j} x_{ijk}!}\prod_{k=1}^{K_j} p_{ijk}^{x_{ijk}} \tag{2.1}$$

Sous l'hypothèse du modèle à nombre infini d'îles ainsi que de notre modèle de sélection et à l'équilibre entre les différentes forces évolutives, la distribution des fréquences alléliques $\boldsymbol{p}_{ij}$ peut être approchée par la distribution stationnaire d'un processus de diffusion multiallélique, fonction des différents paramètres décrits jusqu'alors et dont la densité s'écrit :

$$\psi(\boldsymbol{p}_{ij},\boldsymbol{\pi}_j,M_i,\sigma_{ij},\kappa_{ij}) = C^{-1}\exp(\sigma_{ij}\tilde{p}_{ij})\prod_{k=1}^{K_j} p_{ijk}^{M_i\pi_{jk}-1} \tag{2.2}$$

(Barbour et al. 2000; Barton et Turelli 1987; Donnelly et al. 2001; Wright 1949), où C est la constante d'intégration de la densité :

$$\begin{aligned} C &= \int_0^1 \exp(\sigma_{ij}\tilde{p}_{ij})\prod_{k=1}^{K_j} p_{ijk}^{M_i\pi_{jk}-1}\,\mathrm{d}\mathbf{p}_{ij} \\ &= \frac{\prod_{k=1}^{K_j}\Gamma(M_i\pi_{jk})}{\Gamma(M_i)}{}_1F_1(M_i\tilde{\pi}_{ij},M_i,\sigma_{ij}) \end{aligned} \tag{2.3}$$

avec ${}_1F_1(a,b,z)$ la fonction hypergéométrique confluente (Abramowitz et Stegun 1964). Ajoutons que la dérivation de cette constante est possible grâce à notre hypothèse selon laquelle un seul et unique allèle est sous sélection pour chaque locus dans chaque dème.

Selon le modèle spécifié par les Équations 2.1 et 2.2, nous sommes intéressés par l'évaluation des paramètres $\boldsymbol{\kappa} \equiv (\kappa_{11},...,\kappa_{ij},...,\kappa_{n_dL})$, $\boldsymbol{\sigma} \equiv (\sigma_{11},...,\sigma_{ij},...,\sigma_{n_dL})$, $\boldsymbol{\pi} \equiv (\pi_{11},...,\pi_{jK_j},...,\pi_{LK_L})$ et $\boldsymbol{M} \equiv (M_1,...,M_i,...,M_{n_d})$ à partir des données de comptages alléliques $\boldsymbol{n}$ sur l'ensemble des dèmes et loci échantillonnés.

Parce que nous supposons qu'un seul allèle se trouve sous sélection pour chaque locus dans chaque dème (celui référencé par κ_{ij}), ce modèle est une

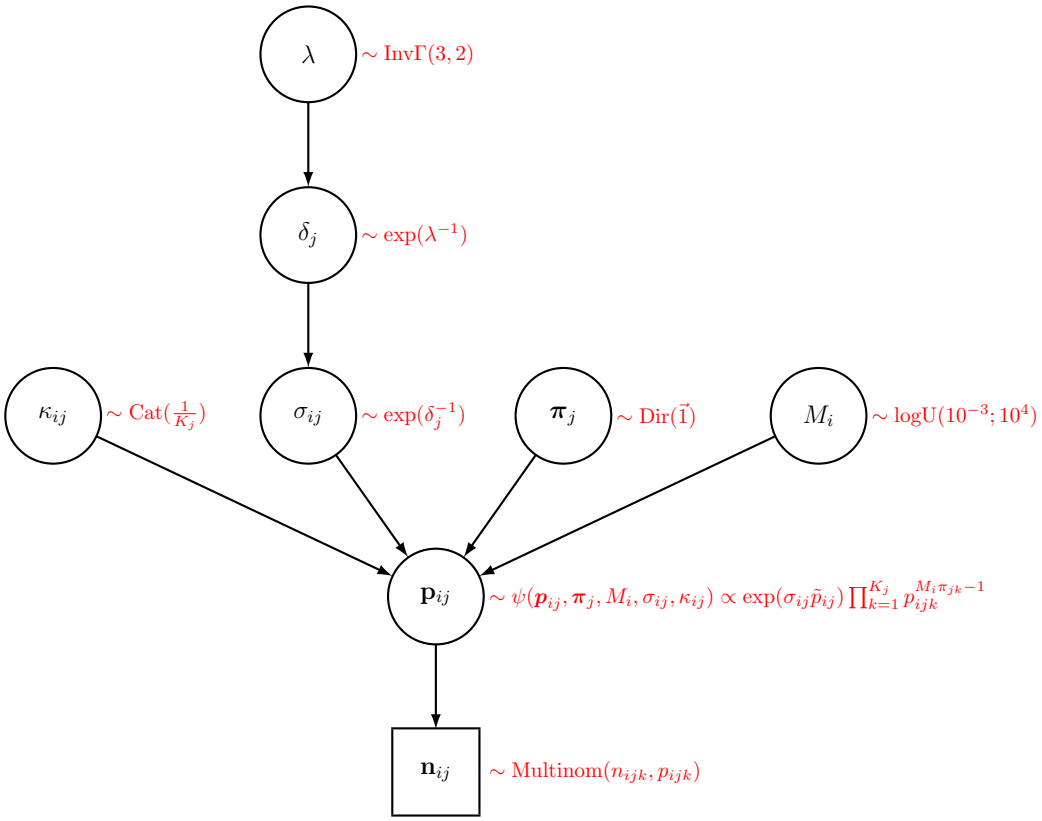


Figure 2.5 Graphe orienté acyclique (DAG) de SELESTIM$_{\text{HAP}}$. Les données de comptages alléliques sont représentées dans le carré par les $\mathbf{n}_{ij}$, tandis que les hyperparamètres sont représentés dans les cercles. La dépendance entre les paramètres du modèle est symbolisée par les flèches et les distributions *a priori* des paramètres sont renseignées à côté de chacun d'entre eux.

généralisation multivariée du modèle développé par Vitalis et al. (2014). Dans le cas de marqueurs bialléliques ($K_j = 2$ pour tout j), ce modèle est strictement identique au modèle développé par Vitalis et al. (2014).

Le graphe orienté acyclique (DAG) de l'extension multialléliques de SelEstim est représenté sur la Figure 2.5.

On suppose que la distribution *a priori* des paramètres κ_{ij} est une distribution généralisée de Bernoulli (Cat) de paramètre $1/K_j$, *i.e.*, $\kappa_{ij} \sim \text{Cat}(1/K_j)$. Pour les paramètres $\boldsymbol{\pi}_j$, on suppose une distribution *a priori* de type Dirichlet uniforme, avec $\boldsymbol{\pi}_j \sim \text{Dir}(\frac{\vec{1}}{K_j})$. Comme dans Vitalis et al. (2014), les paramètres M_i ont une distribution *a priori* log-uniforme sur un support allant de 0.001 à 1000, d'où $\log(M_i) \sim U(\log(10^{-3}), \log(10^4))$. De même, comme dans Vitalis et al. (2014), les distributions *a priori* des coefficients de sélection σ_{ij} sont modélisées hiérarchiquement (voir Figure 2.5). On suppose ainsi que σ_{ij} suit une distribution *a priori* exponentielle $f(\sigma_{ij}|\delta_j) \sim \exp(\delta_j^{-1})$, avec δ_j le coefficient de sélection moyen locus spécifique sur l'ensemble des dèmes. La distribution *a priori* de δ_j est une loi exponentielle $f(\delta_j|\lambda) \sim \exp(\lambda^{-1})$, qui dépend à son tour de l'hyper-paramètre λ, qui peut s'interpréter comme l'effet de la sélection sur l'ensemble des marqueurs, ou encore comme l'écart à l'hypothèse du modèle en îles. Finalement, on suppose ici que la distribution *a priori* de λ est $f(\lambda) \sim \text{Inv}\Gamma(3, 2)$), d'espérance et de variance égales à 1.

2.2.1 Le critère de décision KLD pour l'identification des marqueurs potentiellement sous sélection.

Étant donné que nous disposons des coefficients de sélection locus-spécifique δ_j, nous allons naturellement nous intéresser à leurs distributions *a posteriori*. Puisque nous supposons que l'ensemble des marqueurs est soumis à la sélection, nous nous attendons à ce que les δ_j s'écartent vers des valeurs positives, d'autant plus que la pression de sélection est forte, et tendent vers zéro pour des marqueurs neutres. Ce raisonnement néglige néanmoins un point important, à savoir l'effet moyen de la sélection sur l'ensemble des marqueurs. Nous allons donc calculer, pour chaque locus, la divergence entre la distribution *a posteriori* du coefficient de sélection δ_j, et une distribution de "centrage",

paramétrée par la moyenne *a posteriori* de l'hyper-paramètre λ (voir Vitalis et al. 2014). Pour cela, nous allons utiliser la divergence de Kullback-Leibler (KLD), qui mesure la divergence (ou entropie relative) entre deux distributions (voir Figure 2.6). Notons que si la KLD est souvent considérée comme une mesure de distance entre deux distributions, elle est asymétrique et ne respecte pas le critère d'inégalité triangulaire caractéristique d'une distance.

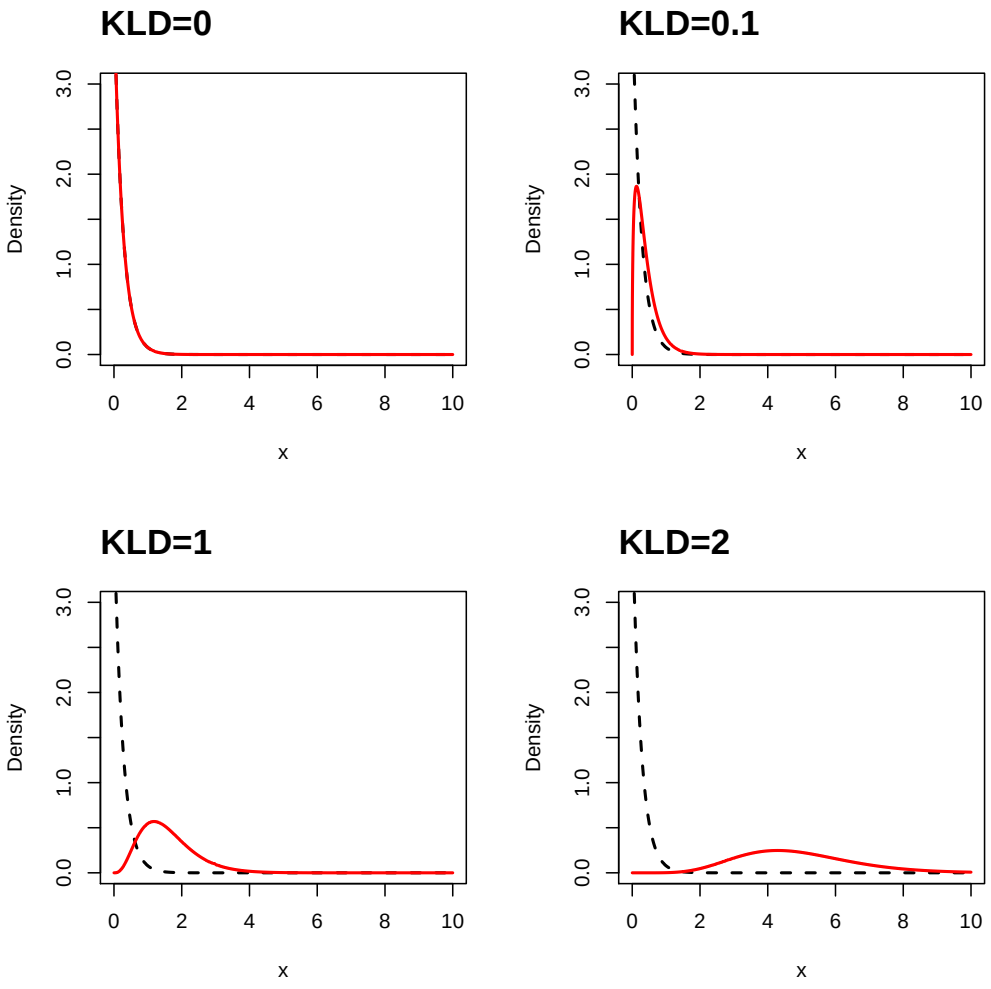


Figure 2.6 Exemples de valeurs prise par la KLD mesurée entre la distribution $\gamma(A)$ (en rouge) et la distribution $\gamma(B)$ (en noir). La figure illustre l'augmentation de la KLD au fur et à mesure de la divergence entre les deux distributions.

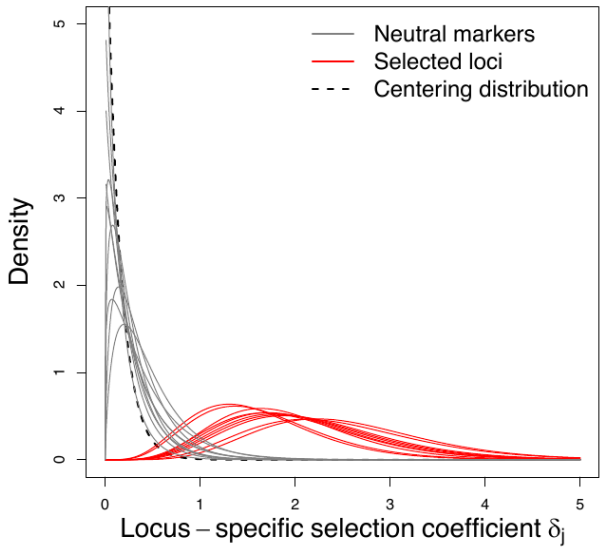


Figure 2.7 Distribution de centrage (tirets noirs) et distributions *a posteriori* de δ_j pour des marqueurs neutres (en gris) et sous sélection (en rouge). Tiré de Vitalis et al. (2014)

Dans notre cas, la distribution *a posteriori* du paramètre δ_j d'un locus neutre est censée être très proche de la distribution de centrage, et donc avoir une KLD faible, tandis qu'un marqueur sous sélection verra la sienne s'en éloigner et avoir une KLD élevée (voir Figure 2.7).

Afin de pouvoir calculer analytiquement la KLD, nous approximons la distribution *a posteriori* des paramètres δ_j par une distribution $\Gamma(k_0 = \bar{x}^2_{\delta_j}/s^2_{\delta_j}, \theta_0 = s^2_{\delta_j}/\bar{x}_{\delta_j})$, avec $\bar{x}_{\delta_j}$ la moyenne et $s^2_{\delta_j}$ la variance de la distribution *a posteriori* de δ_j échantillonnée par MCMC.

On approxime la distribution de centrage par une distribution $\Gamma(1, \theta_1 = \bar{x}_\lambda)$ où $\bar{x}_\lambda$ est la moyenne de la distribution *a posteriori* de l'hyper-paramètre λ issue des échantillons MCMC. Nous utilisons ensuite la mesure de divergence de Kullback-Leibler (KLD) entre les distributions postérieures approchées des δ_j et la distribution de centrage. Pour ce faire, nous utilisons l'estimateur de KLD entre une distribution $\Gamma(k_0, \theta_0)$ et une distribution $\Gamma(1, \theta_1)$ décrite dans l'Équation 6 de Vitalis et al. (2014) :

$$\begin{aligned} \mathrm{KLD}[\Gamma(k_0,\theta_0)||\Gamma(1,\theta_1)] &= log[\frac{\theta_1}{\Gamma(k_0)\theta_0^{k_0}}] + k_0\frac{\theta_0-\theta_1}{\theta_1} \\ &+ (k_0-1)[log(\theta_0)+\digamma(k_0)] \end{aligned} \tag{2.4}$$

où $\digamma(x) \equiv \Gamma'(x)/\Gamma(x)$ est la fonction digamma. On espère ainsi, que plus un marqueur sera soumis à une forte intensité de sélection, plus l'écart de sa distribution *a posteriori* δ_j à la distribution de centrage sera important et donc que la KLD sera élevée.

À ce stade nous n'avons toujours pas de critère de décision et il nous faut donc calibrer la KLD afin de définir une valeur limite au-delà de laquelle un marqueur sera considéré sous sélection.

Nous avons donc choisi de calibrer la KLD à l'aide d'une distribution empirique obtenue sous le modèle nul à partir de données pseudo-observées (PODs). Ces données PODs sont générées à partir du modèle d'inférence avec les valeurs des hyperparamètres fixées à leurs moyennes *a posteriori* des paramètres. Les δ_j sont générés à partir de la distribution de centrage (le modèle nul est donc différent d'un modèle Dirichlet-multinomial classique). L'idée dans la génération de données PODs est de simuler un jeu de données dont les locus vont montrer un niveau de différenciation similaire au jeu de données original. Dans notre procédure, nous considérons des marqueurs échangeables (i.e., conditionnellement indépendants). L'algorithme de génération des données POD est décrit dans l'Annexe B. Les quantiles de la distribution des KLD calculées sur les données PODs sont ensuite utilisés pour calibrer la KLD. Ainsi, le quantile à 95% de la distribution empirique de KLD sur les PODs donne un seuil de KLD à 5%, utilisé comme critère de décision pour discriminer un locus neutre d'un locus sous sélection dans le jeu de données original.

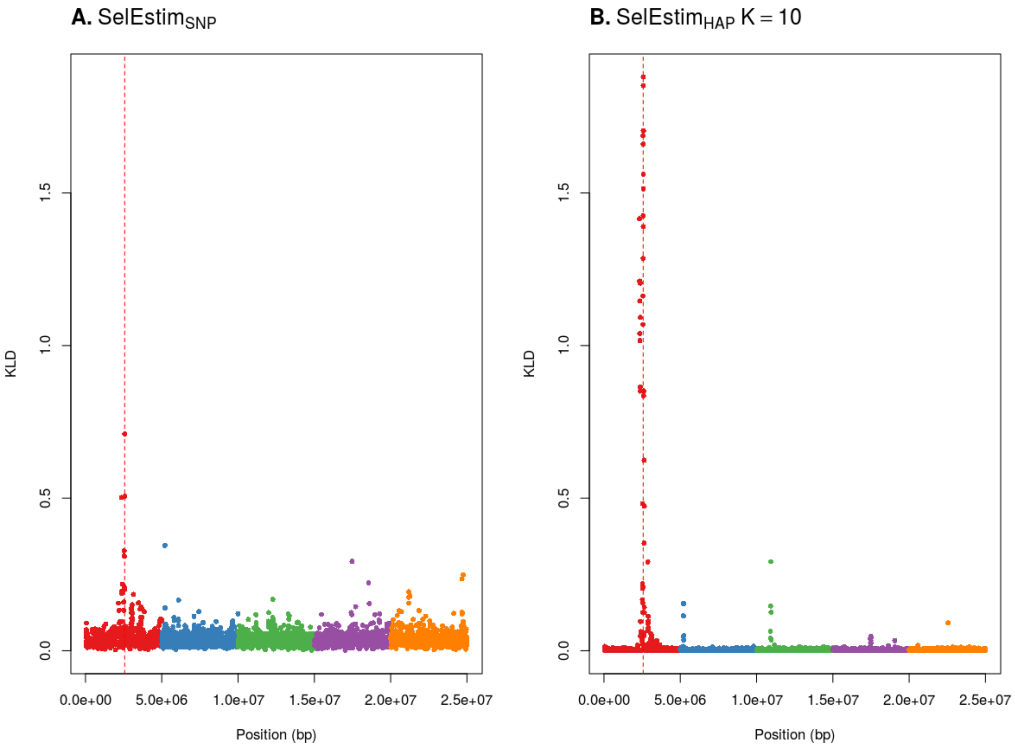


Figure 2.8 Exemple de *scans* génomiques pour un jeu de données simulé, résultant d'une analyse avec SELESTIM$_{SNP}$ (A) et d'une analyse avec SELESTIM$_{HAP}$ (B) dans un modèle en dérive pure d'arbre en étoile. La position simulée sous sélection est indiquée par la ligne pointillée rouge.

2.3 Matériels et Méthodes

2.3.1 Échantillonneur par MCMC

Tout comme le modèle SELESTIM original développé par Vitalis et al. (2014), les distributions *a posteriori* des différents paramètres des deux modèles sont estimées conjointement par MCMC avec algorithme de Metropolis-Hasting (détaillé dans l'Annexe B). En pratique, les différents paramètres sont actualisés les uns après les autres à chaque itération de la chaîne MCMC. Dans cette étude, une analyse SELESTIM est constituée de 30 chaînes "pilotes" de 500 itérations chacune, de manière à ajuster les distributions instrumentales des paramètres $\boldsymbol{\kappa}$,$\boldsymbol{\pi}$,$\boldsymbol{M}$,$\boldsymbol{\sigma}$,$\boldsymbol{\delta}$ et λ, et à obtenir des taux d'acceptation

compris entre 25% et 40% afin de s'assurer de bonnes propriétés de convergence (Gilks et al. 1996). S'ensuit une chaîne de 50 000 itérations faisant suite à une période de "*burn-in*", elle aussi de 50 000 itérations. Afin de réduire l'auto-corrélation entre les valeurs échantillonnées successivement pour les différents paramètres, nous utilisons une procédure de "*thinning*" avec un échantillonnage des paramètres toute les 20 itérations après "*burn-in*".

2.3.2 Simulations

Afin d'évaluer les performances de notre modèle, nous avons réalisé une étude par simulation. Le protocole est inspiré de celui de Fariello et al. (2013). Deux types de scénarios démographiques ont été simulés : un modèle en îles et un modèle en pure dérive. Ce dernier est représenté par un arbre "en étoile" typique des dispositifs d'évolution expérimentale et par un second modèle en pure dérive représenté par un arbre structuré hiérarchiquement avec deux événements de divergence successifs. Le protocole de simulation est illustré dans la Figure 2.9.

Dans un premier temps, nous générons du polymorphisme génétique en simulant des séquences haplotypiques individuelles (5 chromosomes de 5Mb par individu) avec le logiciel de simulations de coalescence `ms` (Hudson 2002). Les données sont ensuite importées dans le logiciel de simulation individu-centré `SimuPop` (Peng et Kimmel 2005) qui nous permet de faire évoluer nos populations selon le modèle démographique souhaité et de faire intervenir la sélection. Notons que le taux de recombinaison a été fixé à $r = 10^{-8}$ (1cM/Mb) et que la recombinaison agit à la fois dans la phase de simulation avec `ms` (Hudson 2002) et avec `SimuPop` (Peng et Kimmel 2005) contrairement à la mutation qui n'agit que dans la première phase de simulation avec un taux $\mu = 10^{-8}$.

La sélection agit sur un seul locus, choisi au plus près du centre du chromosome 1, parmi les loci dont la fréquence f_0 de l'allèle minoritaire dans le dème sous sélection est égale à une fréquence prédéfinie représentant un cas de fort ($f_0 = 0.01$) ou de faible ($f_0 = 0.2$) balayage sélectif. L'allèle minoritaire se trouve ainsi sous sélection positive dans le dème rouge (voir Figure 2.9)

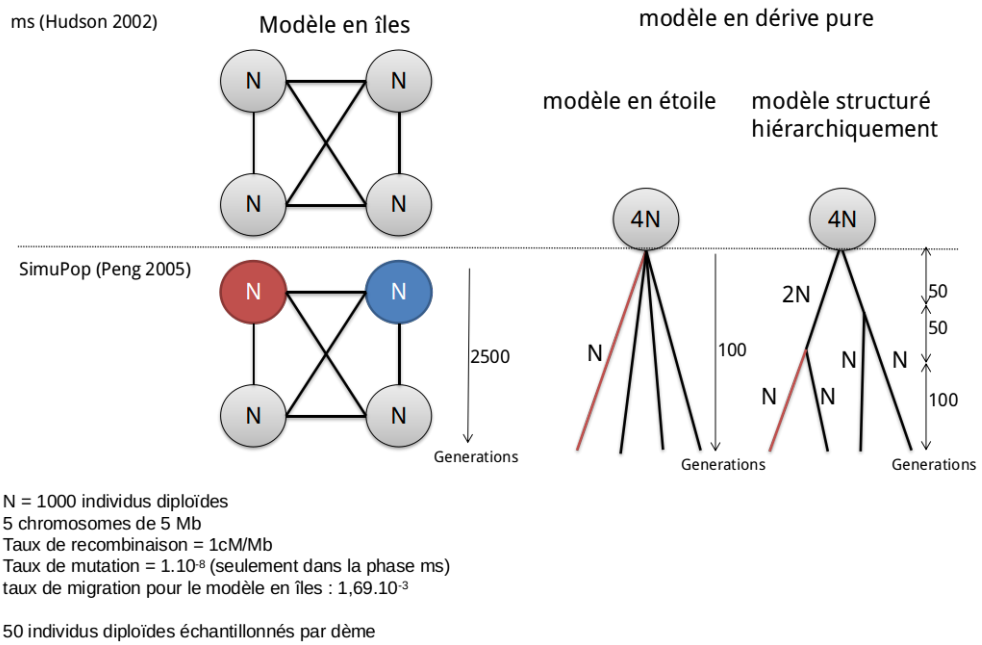


Figure 2.9 Représentations graphiques du protocole de simulation pour les différents scénarios démographiques considérés. Les différentes valeurs de paramètres utilisés pour les simulations sont renseignées pour chaque scénario. La sélection apparaît dans le dème rouge après la génération de polymorphismes génétiques avec `ms` (Hudson 2002). Dans le cas du modèle en îles, la sélection agit aussi dans le dème bleu au même locus que dans le dème rouge mais pour l'allèle alternatif. Lorsque l'on considère les modèle en dérive pure, une population ancestrale est simulée avec `ms` (Hudson 2002) avant d'être séparée en deux ou quatre dèmes pour la phase de simulation avec `SimuPop` (Peng et Kimmel 2005). On notera dans le cas du modèle structuré hiérarchiquement que la sélection apparaît uniquement après le second événement de divergence, au bout de 100 générations.

avec une valeur sélective relative de $1 + s$ pour l'homozygote sous sélection, $1 + s/2$ pour l'hétérozygote, et 1 pour l'homozygote de l'allèle majoritaire qui n'est pas sous sélection, avec $s = 0.05$ et 0.0125 (seulement pour le modèle en îles) soit un coefficient de sélection mis à l'échelle $\sigma \equiv 2Ns = 100$ et 25, respectivement. Dans le cas du modèle en îles, l'allèle majoritaire au locus choisi est quant à lui aussi sous sélection dans le dème bleu, ceci afin de créer un contraste entre les différents dèmes, nécessaire à l'obtention d'une signature de sélection malgré l'homogénéisation due à la migration (Savolainen et al. 2013). Le locus est neutre dans les autres dèmes. Un nombre élevé de générations pour le modèle en îles a été choisi afin d'atteindre un état d'équilibre migration-dérive-sélection. Enfin, seules les simulations pour lesquelles la fréquence de l'allèle sous sélection est supérieure ou égale à 0.6 sont conservées. Nous échantillonnons ensuite aléatoirement 50 individus par dème et ne gardons que les SNPs avec une MAF supérieure à 5%. Ces marqueurs sont ensuite échantillonnés aléatoirement afin d'obtenir une densité de 125 SNPs/Mb. Nous nous attendons donc à ce que, comme dans la plupart des jeux de données réels, le marqueur causal ne soit pas fréquemment présent. Chaque scénario a été répliqué 500 fois de manière indépendante.

2.3.3 Comparaisons aux modèles FLK et HapFLK

Les performances de SelEstim ont été comparées à deux autres modèles, FLK (Bonhomme et al. 2010) et son extension pour données haplotypiques, HapFLK (Fariello et al. 2013).

Afin d'offrir un traitement équitable des données avec les différentes méthodes, la matrice $\mathcal{F}$ de variance-covariance des fréquences alléliques utilisée par FLK et HapFLK n'a pas été renseignée et a été estimée pour chaque jeu de données par les programmes. Chaque jeu de données étant composé d'un chromosome avec sélection et de quatre chromosomes neutres, l'estimation de la matrice de variance-covariance des fréquences alléliques est assez précise dans le cas d'un modèle en pure dérive, sans migration. De plus, HapFLK implémente directement l'algorithme de Markov caché (HMM) de phasage et de groupement haplotypique `fastphase` (Scheet et Stephens 2006). Nous

avons ici utilisé l'option `-phased` du logiciel afin de renseigner directement les haplotypes simulés et ne pas effectuer le phasage, ce qui aurait pu ajouter des erreurs. Le nombre K de groupes haplotypiques choisi dépend du scénario considéré. Nous avons choisi $K = 10$ et $K = 20$ pour l'arbre en étoile et l'arbre structuré hiérarchiquement, respectivement, ceci afin de maximiser la puissance de détection des modèles FLK et HapFLK (Fariello et al. 2013). $K = 20$ a aussi été choisi pour le modèle en îles. Étant donné que le modèle de groupement haplotypique `fasphase` (Scheet et Stephens 2006) converge vers des maximums locaux, HapFLK moyenne les statistiques obtenues après différentes itérations d'algorithmes de maximisation (EM). Dans cette étude, en accord avec Fariello et al. (2013), 5 itérations ont été utilisées.

2.3.4 Évaluation des performances de SelEstim sur la base des simulations

Afin d'évaluer les performances des différents modèles sur la base des simulations réalisées, nous avons mesuré la puissance et la "qualité de signal" en distinguant le biais de position (c'est-à-dire l'écart entre la position réellement simulée sous sélection et la position des marqueurs détectés sous sélection) et la précision du signal (l'écart-type des positions détectées sous sélection). Ces deux concepts sont représentés dans la Figure 2.10

Pour cela nous devons être capables, pour une statistique S donnée (telle que la KLD ou les statistiques FLK ou HapFLK), de définir une valeur-seuil au delà de laquelle un marqueur sera considéré sous sélection. Puisque nous nous trouvons dans le cadre de simulations pour lesquelles nous connaissons la vérité, il apparaît plus judicieux de définir ce seuil en fonction de la distribution de la statistique S observée sur les données neutres que nous avons simulées. En pratique, pour un scénario démographique donné, nous relevons, pour chaque simulation, la valeur maximale de la statistique S (S_{max}) mesurée sur chacun des 4 chromosomes neutres. À partir de 500 jeux de données simulés, soit 2000 valeurs (4 chromosomes neutres par simulation), nous pouvons construire la distribution de S^{max} observée sous le modèle neutre. Pour une erreur de type I (α) choisie, la valeur-seuil sera définie comme le

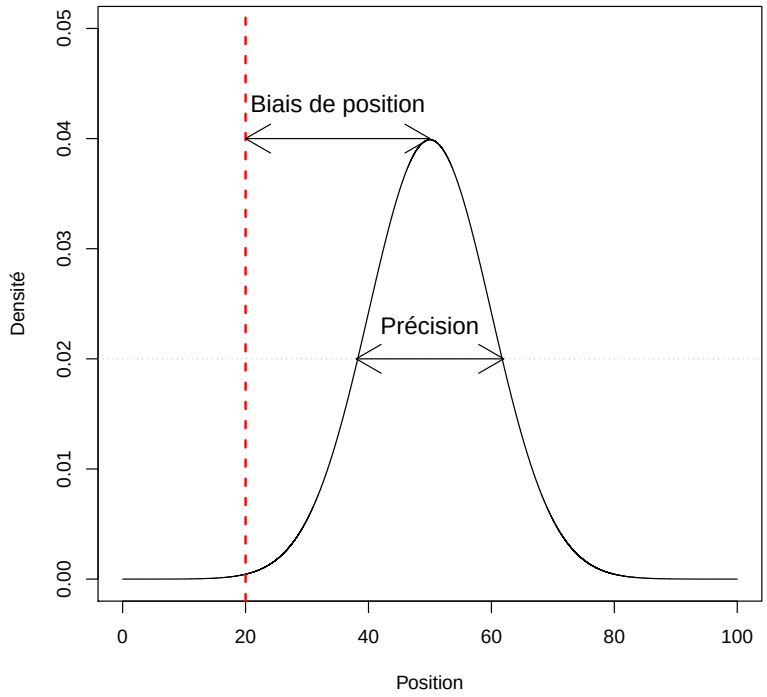


Figure 2.10 Représentation graphique du biais de position et de la précision pour un signal représenté par la distribution en noir tandis que la position de la mutation sous sélection est représentée par la ligne rouge en pointillés. La valeur-seuil au delà de laquelle les valeurs sont considérées comme étant détectées positivement est représentée par la ligne grise en pointillés.

$(1-\alpha)$ème quantile de la distribution empirique des S^{max}.

La puissance sera donc définie comme la proportion de simulations pour lesquelles la valeur maximale de la statistique observée sur le chromosome sous sélection est supérieure à la valeur-seuil. Ce choix est justifié par le fait qu'en présence de marqueurs haplotypiques, et en l'absence du marqueur causal, nous n'avons pas d'attendu concernant les marqueurs qui doivent être détectés sous sélection (car cet attendu dépend de la définition des marqueurs haplotypiques ainsi que de l'étendue du DL). Par défaut, et sauf mention contraire, nous avons considéré une erreur de type I de 5%.

L'ensemble des analyses réalisées sur les résultats des différents modèles

a été réalisé sous R version 3.3.1 (R Core Team 2017)

2.4 Résultats

2.4.1 Évaluation des performances de SelEstim$_{\text{HAP}}$

.

Afin d'évaluer les performances de SelEstim$_{\text{HAP}}$, nous avons réalisé des simulations selon différents scénarios démographiques (voir Matériels et Méthodes). Nous allons dans un premier temps évaluer les performances des modèles sous l'hypothèse démographique sous-jacente à SelEstim$_{\text{HAP}}$: le modèle en îles (Wright 1931). Les performances du modèles seront ensuite comparées à celles de SelEstim$_{\text{SNP}}$, FLK et HapFLK.

Dans un contexte de fort balayage sélectif, lorsque la sélection agit sur une mutation initialement présente en faible fréquence (1%) dans la population cible, l'utilisation de données haplotypiques par SelEstim$_{\text{HAP}}$ augmente considérablement la puissance de détection du modèle (voir Figures 2.11 A-C). Le gain le plus important apparaît lorsque l'intensité de la sélection est la plus élevée ($\sigma = 100$, voir Figure 2.11 A). En effet, pour une erreur de type I de 5%, l'utilisation d'haplotypes par SelEstim se traduit par un gain de puissance de 31% par rapport à l'utilisation de SNPs considérés comme indépendants par SelEstim$_{\text{SNP}}$. Comme nous pouvions nous y attendre, une intensité de sélection plus modérée (c'est-à-dire $\sigma = 25$) entraîne une diminution drastique de la puissance pour l'ensemble des modèles testés, avec une chute de SelEstim$_{\text{HAP}}$ et SelEstim$_{\text{SNP}}$ à 42% et 32% (pour une erreur de type I de 5%). FLK et HapFLK (pour lesquels le modèle en îles représente un écart important au modèle démographique sous-jacent) sont les moins performants. On notera cependant que pour une forte intensité de sélection et dans un contexte de fort balayage sélectif, l'utilisation de données haplotypiques apporte aussi un gain de puissance à HapFLK par rapport à FLK (Fariello et al. 2013).

Dans un contexte de faible balayage sélectif, la différence de performance entre SelEstim$_{\text{HAP}}$ et SelEstim$_{\text{SNP}}$ est fortement réduite, quelle que soit

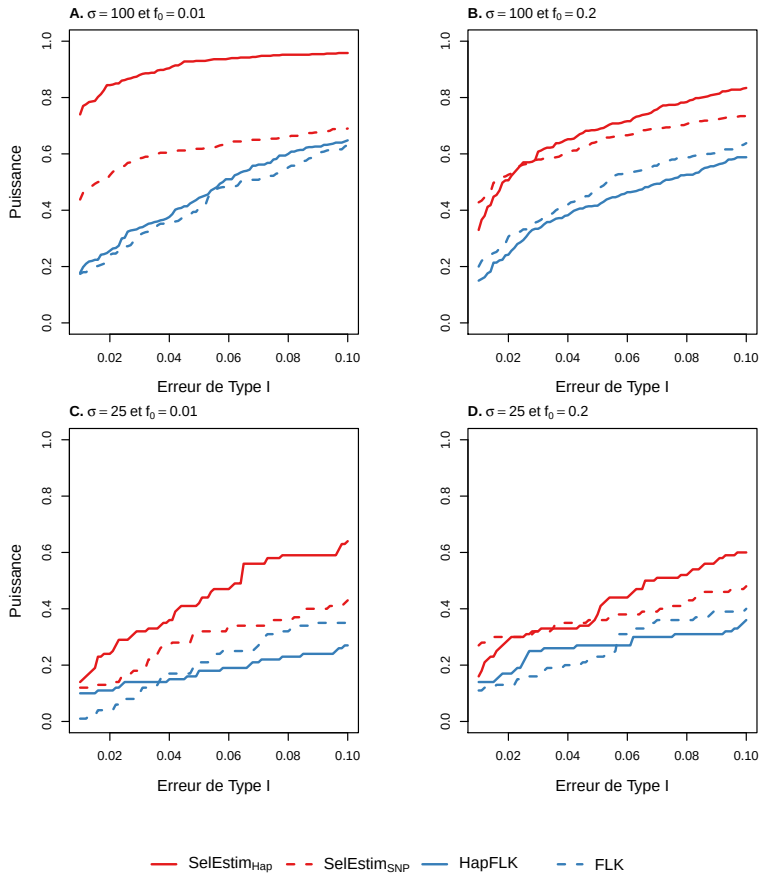


Figure 2.11 Puissance en fonction de l'erreur de type I dans un modèle en îles avec $n_d = 4$ dèmes. Les résultats sont représentés selon deux intensités de sélection, $2Ns = 100$ (A-B) et $2Ns = 25$ (C-D) ainsi que deux fréquences initiales (f_0) de l'allèle sous sélection, 1% (A-C) et 20% (B-D).

l'intensité de la sélection. Ce résultat est plutôt dû à une diminution de la puissance de SELESTIM$_{\text{HAP}}$, la puissance de SELESTIM$_{\text{SNP}}$ étant proche du cas de balayage sélectif fort.

La Figure 2.12 nous montre que les coefficients de sélection σ estimés à partir de leurs distributions *a posteriori* nous permettent d'identifier dans tous les cas le dème rouge comme étant celui où la sélection agit, et cela même pour un balayage sélectif faible et une faible intensité de sélection (voir la Figure 2.12 D). Il ne nous est cependant pas possible de comparer

directement les valeurs obtenues par rapport aux valeurs simulées. En effet, nous observons ici un σ estimé à partir de données haplotypiques tandis que la valeurs simulée est uniquement valable pour la mutation causale.

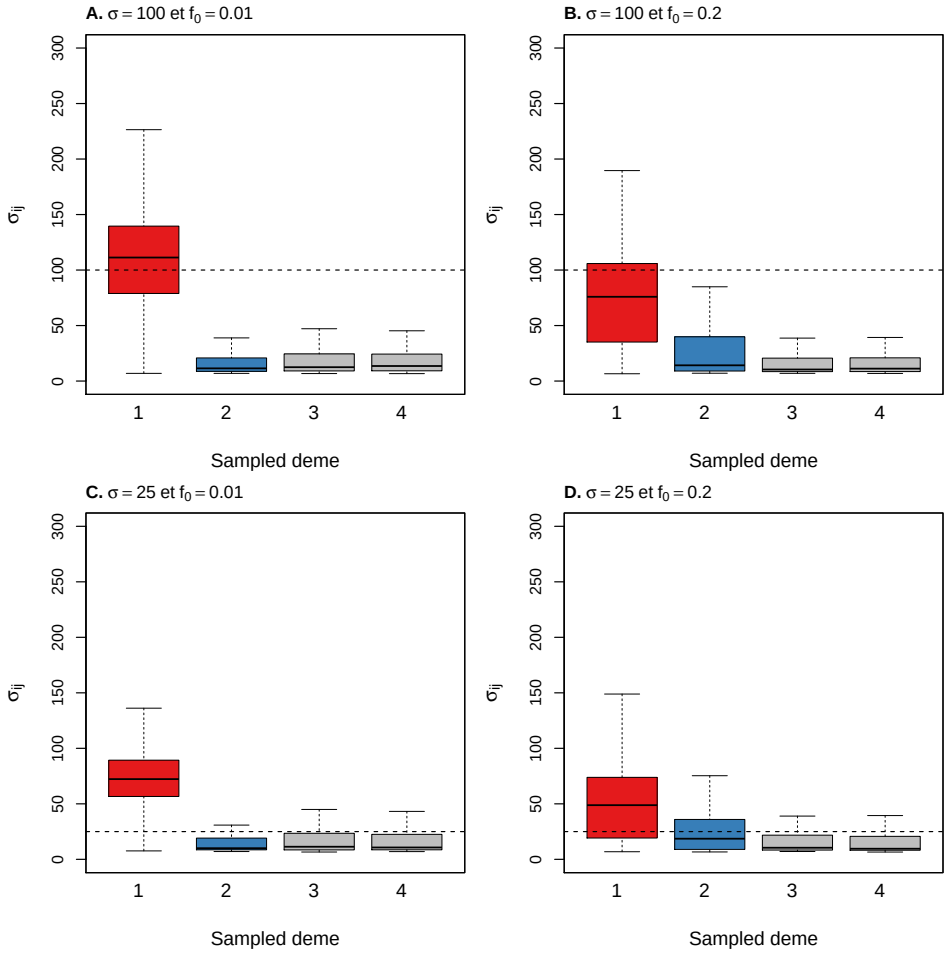


Figure 2.12 Distributions des estimations de σ_{ij} (moyennes *a posteriori* obtenus dans un modèle en îles avec $n_d = 4$ dèmes. Les résultats sont représentés selon deux intensités de sélection, $\sigma = 100$ (A-B) et $\sigma = 25$ (C-D) ainsi que deux fréquences initiales (f_0) de l'allèle sous sélection, 1% (A-C) et 20% (B-D). Le dème rouge représente le dème sous sélection pour l'allèle de référence tandis que le dème bleu est sous sélection pour l'allèle alternatif.

Bien qu'informatives, nos mesures de puissance ne mesurent que la capacité des modèles à obtenir un signal sur le chromosome d'intérêt et n'évaluent donc pas la précision des signaux obtenus. À cette fin, nous avons mesuré les

biais de position (distance médiane entre la valeur maximale de la statistique d'intérêt le long du chromosome et la position simulée sous sélection). De la même manière, nous avons aussi mesuré la précision (écart-type des distances entre les positions détectées sous sélection et la position simulée sous sélection). Pour des simulations sous le modèle en îles, SELESTIM$_{HAP}$ fait preuve d'un biais de position bien plus faible que FLK et HapFLK (voir Figure 2.13). Ceci traduit donc le fait que les signaux perçus par FLK et HapFLK s'écartent beaucoup plus de la position réellement simulée sous sélection, contrairement à SELESTIM$_{HAP}$ et SELESTIM$_{SNP}$. Notons que de manière générale, l'utilisation de SNPs considérés indépendants avec SELESTIM$_{SNP}$ résulte en un biais de position plus faible. Les mesures de précision montrent quant à elles que l'utilisation de données haplotypiques, que ce soit pour SELESTIM$_{HAP}$ et HapFLK résultent en un signal plus précis que leurs pendants utilisant des SNPs indépendants (voir Figure 2.14). Pris deux à deux, SELESTIM$_{HAP}$ et HapFLK montrent des précisions similaires, de même que SELESTIM$_{SNP}$ et FLK. On notera tout de même une meilleure précision pour SELESTIM$_{SNP}$ lorsque l'intensité de la sélection est modérée (voir Figures 2.14 C-D). Pour résumer ces premiers résultats dans un contexte de modèle en îles, favorable à SELESTIM, l'utilisation de données haplotypiques se traduit par une augmentation de la puissance de détection de marqueurs sous sélection, associée à un signal de bonne qualité, dont la précision se trouve accrue par rapport à l'utilisation de SNPs indépendants. Si le biais de position de SELESTIM$_{HAP}$ reste en deçà de celui de SELESTIM$_{SNP}$, cela peut s'expliquer par le fait que de nombreux haplotypes, autour de la position sous sélection, peuvent contenir plusieurs SNPs fortement différenciés.

2.4.2 Evaluation de la robustesse aux écarts des hypothèses du modèle

Afin d'évaluer la robustesse de SELESTIM$_{HAP}$ lorsque l'histoire des populations s'écarte de celle d'un modèle en îles, j'ai évalué ses performances pour deux modèles de dérive pure : l'un selon un arbre en étoile et le second selon un arbre structuré hiérarchiquement avec deux événements indépendants

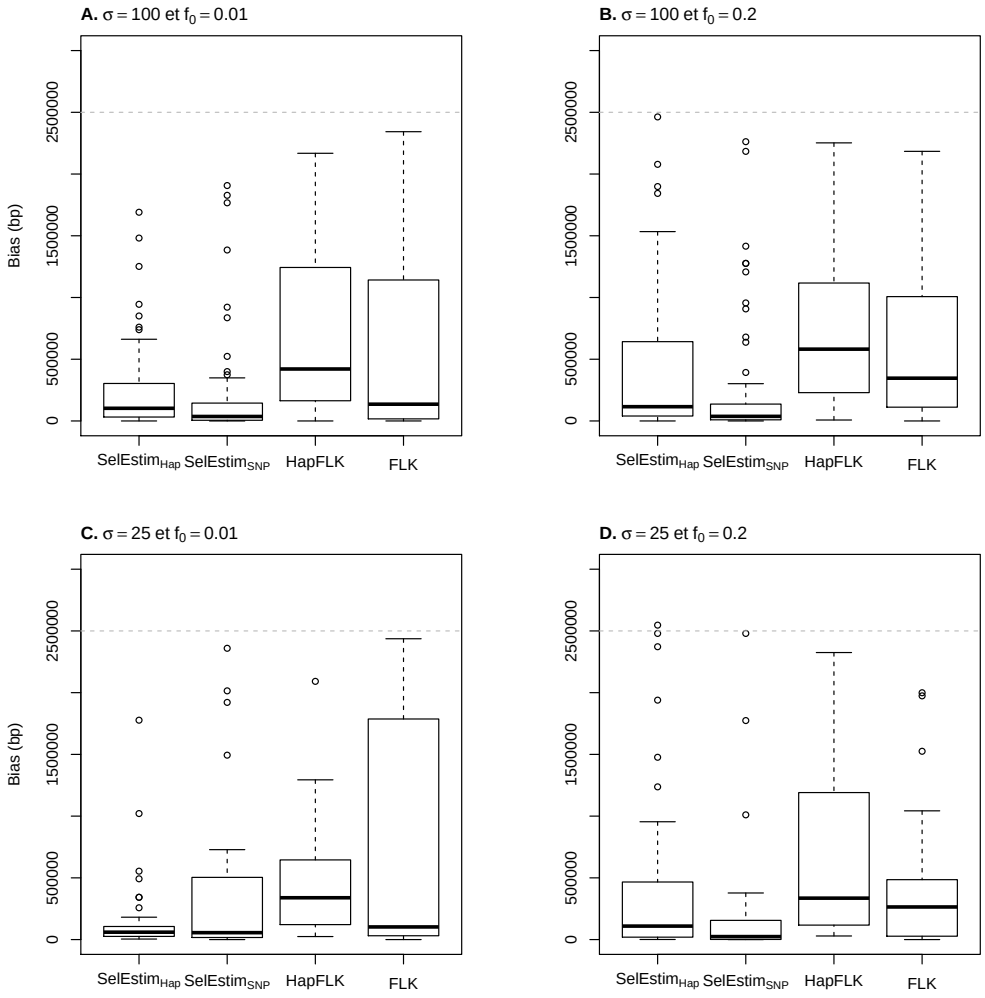


Figure 2.13 Distribution du biais de position (distance médiane de la/les valeur(s) maximale(s) à la position sous sélection) dans un modèle en îles avec $n_d = 4$ dèmes. Les résultats sont représentés selon deux intensités de sélection, $\sigma = 100$ (A-B) et $\sigma = 25$ (C-D) ainsi que deux fréquences initiales (f_0) de l'allèle sous sélection, 1% (A-C) et 20% (B-D). Les comparaisons sont effectuées entre les différents modèles, SelEstim$_{\text{HAP}}$ (KLD_Hap), SelEstim$_{\text{SNP}}$ (KLD), HapFLK et FLK.

de divergence, qui vont nous permettre d'évaluer la robustesse du modèle à l'histoire des populations. Contrairement au modèle en îles vu précédemment, ces deux scénarios sont particulièrement adaptés aux modèles FLK et HapFLK. En effet ces derniers supposent que les populations évoluent en

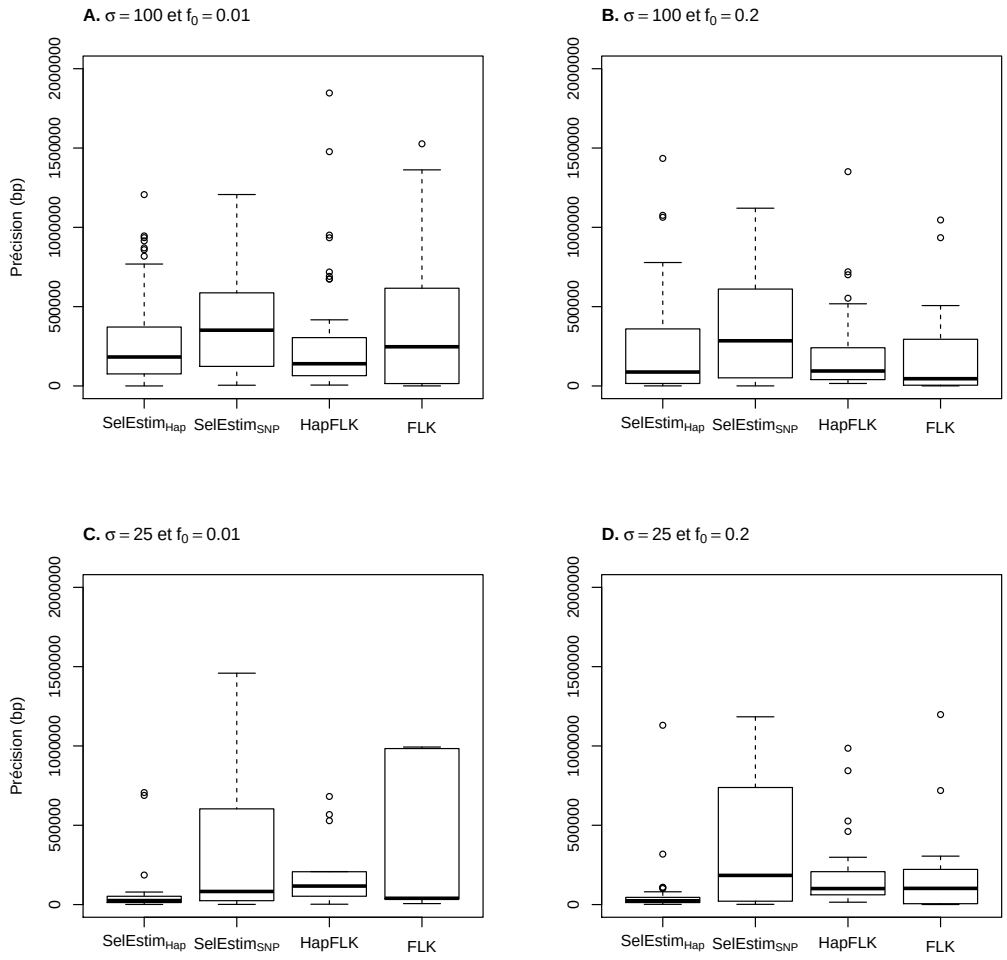


Figure 2.14 Distributions de la précision du signal(écart-type de la distance à la position sous sélection des différents marqueurs détectés positivement) dans un modèle en îles avec $n_d = 4$ dèmes. Les résultats sont représentés selon deux intensités de sélection, $2Ns = 100$ (A-B) et $2Ns = 25$ (C-D) ainsi que deux fréquences initiales (f_0) de l'allèle sous sélection, 1% (A-C) et 20% (B-D).

dérive pure, avec une taille de population constante dans chaque branche et aucun événement de migration.

Lorsque nous considérons une démographie simple avec un arbre en étoile, on constate que HapFLK et FLK montrent les meilleures puissances (voir Figures 2.15 A-B). Notons tout de même que si la puissance de SelEstim$_{\text{SNP}}$

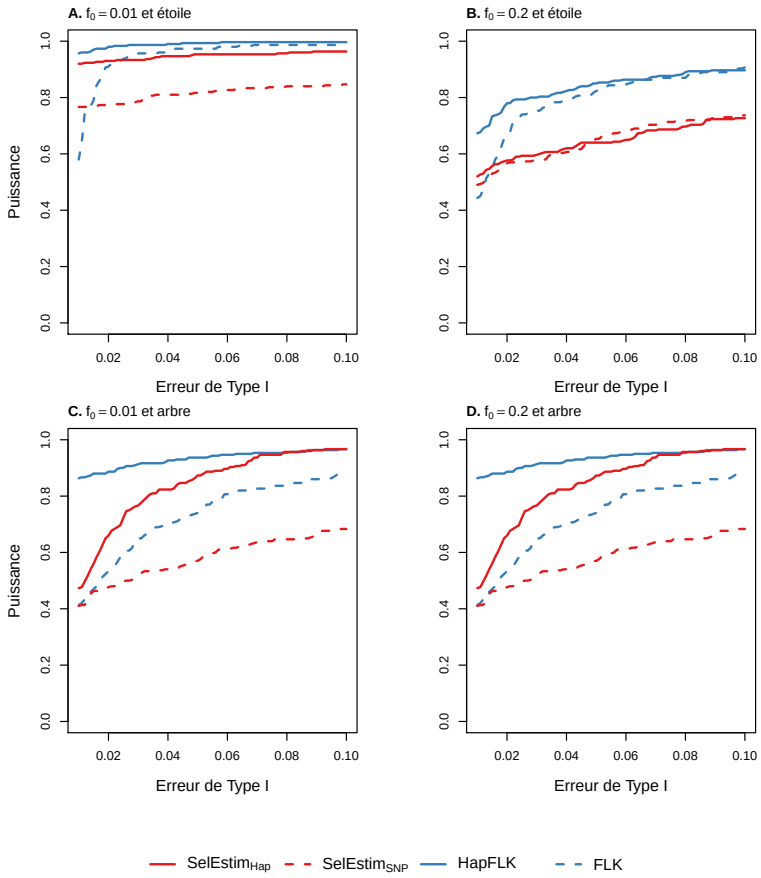


Figure 2.15 Puissance en fonction de l'erreur de type I dans un modèle d'arbre en étoile (A-B) et d'arbre structuré hiérarchiquement (C-D) pour un fort balayage sélectif avec une fréquence initiale (f_0) de l'allèle sous sélection de 1% (A et C) et un faible balayage sélectif pour lequel la fréquence initiale est de 20% (B et D).

est la plus faible dans le cas de forts balayage sélectifs (bien qu'avoisinant les 80%), les performances de SELESTIM$_{\text{HAP}}$ sont portées au niveau d'HapFLK et FLK lorsqu'on utilise des marqueurs haplotypiques. Contrairement à HapFLK, pour lequel l'utilisation de marqueurs haplotypiques se traduit aussi par une augmentation de puissance lors d'un faible balayage sélectif. SELESTIM$_{\text{HAP}}$ montre des performances quasiment indistinguables de celles de SELESTIM$_{\text{SNP}}$. Notons que ces performances restent inférieures à celles

de HapFLK et FLK. Le type de scénario simulé correspond parfaitement à l'hypothèse sous-jacente à FLK et HapFLK tout en ne représentant qu'un écart minimal à l'hypothèse du modèle en îles faite par SelEstim$_{HAP}$ et SelEstim$_{SNP}$ (car la distribution des fréquences alléliques dans un modèle de dérive pure peut aussi être approximée par une distribution beta).

Les résultats les plus intéressants apparaissent lorsque l'on considère des populations qui évoluent selon un arbre hiérarchiquement structuré. Ce dernier étant le scénario le moins propice à SelEstim, il est cependant celui où le gain de puissance relatif à l'utilisation de données haplotypiques est le plus important et montre une meilleure robustesse du modèle à la structure de population (pour le modèle démographique testé). En effet, pour un fort balayage sélectif, SelEstim$_{HAP}$ voit sa puissance augmentée de 30.3% et portée au niveau d'HapFLK pour une erreur de type I de 5% (voir Figure 2.15 C), contre 13.6% pour le scénario d'arbre en étoile vu précédemment. Finalement, l'utilisation d'haplotypes augmente aussi la puissance de SelEstim dans un contexte de balayage sélectif faible (voir Figure 2.15 D), le plaçant devant HapFLK et FLK qui montrent des résultats difficilement distinguables l'un de l'autre.

De la même manière que pour les simulations sous le modèle en îles, nous avons mesuré la qualité du signal obtenu à partir des différents modèles pour les deux scénarios en pure dérive. Une fois de plus, on observe un biais de position supérieur pour les méthodes utilisant les haplotypes par rapport aux SNPs (voir Figure 2.16), mais dans une mesure bien moindre à que ce que nous observions pour le modèle en îles. La précision des signaux est quant à elle très similaire entre les différentes méthodes avec une meilleure précision dans le cas de faible balayage sélectif (voir Figure 2.17).

2.5 Discussion

Les *scans* génomiques de différenciation sont devenus particulièrement populaires pour la recherche de signatures de sélection. S'ils ont longtemps été dédiés aux espèces modèles, le développement des technologies de séquençage NGS, et l'augmentation des ressources génomiques disponibles, a rendu

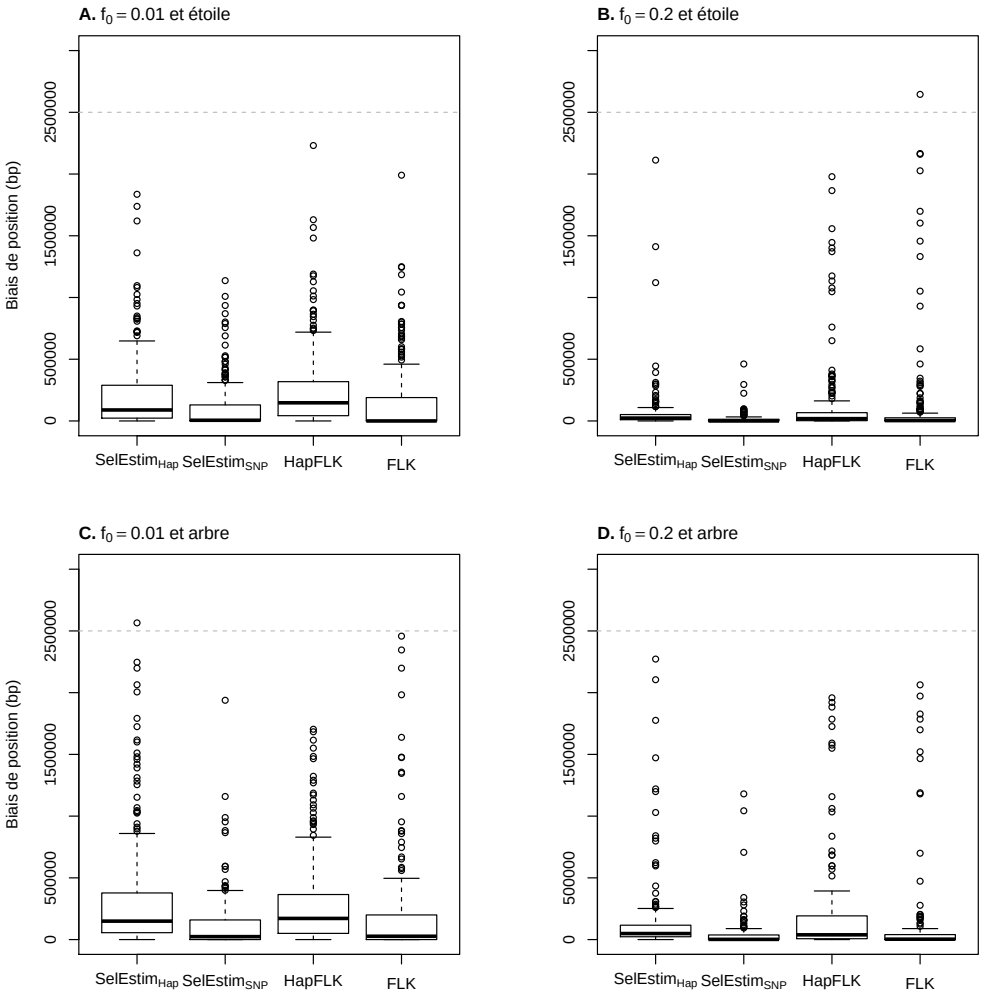


Figure 2.16 Distribution du biais de position (distance médiane de la/les valeur(s) maximale(s) à la position sous sélection) dans un modèle de dérive pure selon un arbre en étoile (A-B) et avec populations structurées (C-D) pour un fort balayage sélectif avec une fréquence initiale de l'allèle sous sélection (f_0) de 1% (A et C) et un faible balayage sélectif pour lequel la fréquence initiale est de 20% (B et D). Les comparaisons sont effectuées entre les différents modèles, SelEstim$_{HAP}$ (KLD_Hap), SelEstim$_{SNP}$ (KLD), HapFLK et FLK.

possible leur application chez de nombreuses espèces non-modèle.

La quantité de ressources n'étant plus un facteur limitant, de nouveaux défis ont émergé pour les génomiciens des populations. Parmi eux, la prise en compte de l'information apportée par lDL dans les données représente un

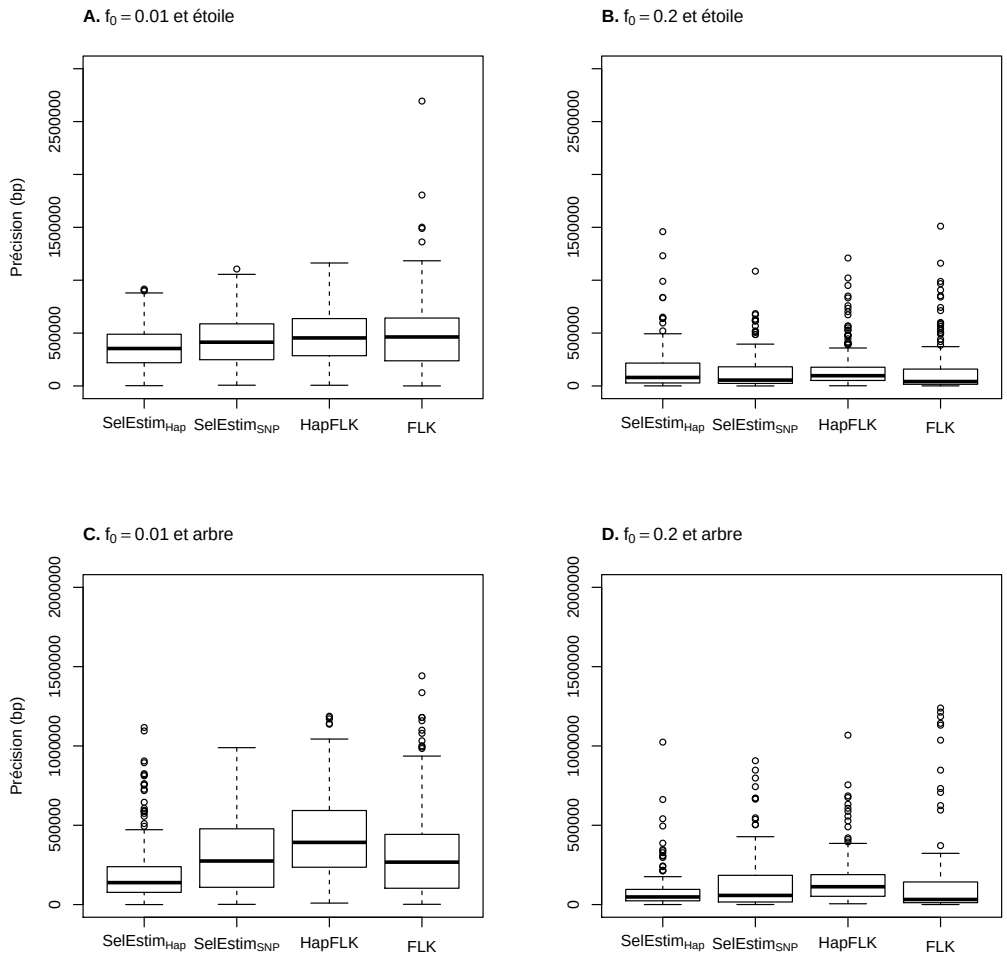


Figure 2.17 Distribution de la précision (écart-type de la distance à la position sous sélection des différents marqueurs détectés positivement) dans un modèle de dérive pure selon un arbre en étoile (A-B) et avec populations structurées (C-D) pour un fort balayage sélectif avec une fréquence initiale (f_0) de l'allèle sous sélection de 1% (A et C) et un faible balayage sélectif pour lequel la fréquence initiale est de 20% (B et D). Les comparaisons sont effectuées entre les différents modèles, SELESTIM$_{\text{HAP}}$ (KLD_Hap), SELESTIM$_{\text{SNP}}$ (KLD), HAPFLK et FLK.

enjeu majeur. En effet, le DL représente une source d'information intéressante au vu de son effet sur les marqueurs neutres au voisinage de la mutation causale selectionnée.

2.5.1 Apport des données haplotypiques au modèle SelEstim

Nous avons proposé ici une nouvelle version de SelEstim, un modèle hiérarchique bayésien de détection et d'estimation de la sélection (Vitalis et al. 2014) initialement conçu pour l'utilisation de données bialléliques de type SNP. Ce nouveau modèle, basé sur un processus de diffusion multiallélique des fréquences alléliques nous permet d'utiliser des données haplotypiques. Ces haplotypes caractérisés dans une fenêtre sont considéré dans ce modèle comme un ensemble d'allèles à un locus donné. Ces marqueurs haplotypiques locaux permettent ainsi d'exploiter l'information de déséquilibre de liaison qui existe entre les marqueurs. Nous avons pu voir, par le moyen de données simulées, que dans le cas de données de génotypage dense pour lesquelles le marqueur causal n'est pas toujours présent, l'application de cette stratégie résulte en une hausse de la puissance de détection des régions sous sélection, et ce pour les différents scénarios démographiques testés. Notons tout de même qu'au même titre que Innan et Kim (2008) et Fariello et al. (2013), nous avons observé que dans le cas de données de séquences complètes (avec une très forte densité de marqueurs et toujours en présence de la mutation causale), l'utilisation de données SNPs considérées indépendantes par $\text{SelEstim}_{\text{SNP}}$ montrait des performances au moins aussi bonnes que $\text{SelEstim}_{\text{HAP}}$ (résultats non présentés), tout en nécessitant un temps de calcul bien plus court. Cependant, Fariello et al. (2013) ont déjà argumenté l'intérêt d'utiliser une méthode haplotypique même en présence de données de séquences. En effet, la sélection agissant souvent sur des haplotypes multi-locus, que ce soit à travers une sélection polygénique ou encore sur des mutations récurrentes au sein d'un même gène.

2.5.2 Les stratégies de groupement local des haplotypes

Un point important à discuter concerne les différentes stratégies de regroupement local des haplotypes utilisées pour définir nos marqueurs mul-

tialléliques. Nous avons utilisé deux méthodes différentes, l'algorithme HMM implémenté dans `fastphase` (Scheet et Stephens 2006) et ré-implémenté dans le logiciel hapFLK d'une part, et une approche de regroupement local par blocs haplotypiques pour SelEstim$_{\text{HAP}}$ d'autre part. Si les méthodes HMM telles qu'implémentées dans `fastphase` (Scheet et Stephens 2006) sont réputées plus flexibles, avec notamment une meilleure prise en compte de la variation locale de DL, elles sont aussi soumises à des problèmes de convergence vers des maxima locaux. HapFLK contourne ce problème en réalisant les mesures de statistique HapFLK à partir de chaque procédure de phasage basée sur un algorithme EM, puis en les moyennant. Cette stratégie permet ainsi à HapFLK d'intégrer sur l'incertitude dû à l'algorithme HMM `fastphase` (Scheet et Stephens 2006) et de réaliser indirectement un lissage du signal. Nous ne pouvons cependant pas utiliser cette approche avec SelEstim$_{\text{HAP}}$, puisqu'elle demanderait de réaliser plusieurs analyses indépendantes et serait donc beaucoup trop coûteuse computationellement. Nous avons donc choisi une approche de construction par blocs haplotypiques, plus simple, mais plus stable. Bien que ce type d'approche demeure en principe plus limité dans la prise en compte du DL, nous avons tout de même observé une amélioration de la puissance de détection avec SelEstim$_{\text{HAP}}$. Notre approche, conditionnée sur un nombre d'allèles minimum à atteindre localement est finalement proche de la méthode développée par Utsunomiya et al. (2016). Cette dernière conditionne les blocs haplotypiques en fonction du nombre de SNPs présents dans chaque fenêtre. Leur méthode a d'ailleurs ensuite été appliquées à la détection de signatures de sélection chez les bovins (Utsunomiya et al. 2017).

L'inconvénient de notre procédure repose majoritairement sur le fait que nous comparons deux méthodes haplotypiques qui n'utilisent pas les mêmes données, ce qui rend l'interprétation de certains résultats plus complexe. C'est le cas par exemple lorsque SelEstim$_{\text{HAP}}$ montre une puissance supérieure aux autres modèles dans le cas d'un balayage sélectif de faible intensité pour une démographie en arbre structuré hiérarchiquement (voir Figure 2.15). La différence avec HapFLK est-elle réellement due au modèle SelEstim ou aux données utilisées, qui auraient capturé une information de DL différente de celle issue de `fastphase` ? Un moyen éventuel de répondre à cette question

serait de fournir à HapFLK les données utilisées par SelEstim$_{\text{HAP}}$. Celles-ci ayant cependant un nombre d'allèles variables entre locus, elles deviennent de fait incompatibles avec la version actuelle du modèle. Une perspective concernant SelEstim$_{\text{HAP}}$ serait d'intégrer un modèle de phasage et de regroupement locale des haplotypes. On peut par exemple penser au modèle Beagle (Browning et Weir 2010) plus rapide que `fastphase` (Scheet et Stephens 2006), et qui autorise aussi une variation locale du nombre de groupes haplotypiques (et donc potentiellement une meilleure prise en compte de la variation locale de DL), ce qui est compatible avec SelEstim$_{\text{HAP}}$. L'intégration d'un tel modèle pourrait se faire directement entre le niveau des données et l'estimation des p_{ijk} de manière à intégrer sur l'incertitude de phasage et de regroupement local des haplotypes. Bien que réalisable en théorie, cela demanderait cependant un vrai travail d'optimisation du modèle afin de le rendre applicable à des jeux de données réels en un temps limité.

2.5.3 Les balayages sélectifs faibles, le talon d'Achille du *scan* génomique

SelEstim n'échappe pas à la longue lignée de modèles spécifiquement adaptés à la détection de balayages sélectifs forts sur des mutations rares. Cependant, nous avons pu voir que l'utilisation d'haplotypes peut aussi augmenter la puissance de détection des balayages sélectifs faibles. Puisque SelEstim considère un seul et unique allèle sous sélection (à travers la variable κ_{ij}) pour chaque locus et dans chaque dème, il n'est de fait pas adapté aux balayages sélectifs faibles (qui s'exercent sur un variant porté par différents haplotypes, qui vont donc tous augmenter en fréquence). Sur la base d'études réalisées chez l'homme et la drosophile, un engouement tout particulier s'est fait sentir concernant les balayages sélectifs faibles, qui seraient légion dans la nature, contrairement aux balayages forts.

Jensen (2014) a cependant mis en avant que même dans le cas de balayages sélectifs forts, les méthodes de détection de la sélection pouvaient échouer face à des démographies complexes, ce qui a depuis fait l'objet de nombreux développements (Fariello et al. 2013; Foll et al. 2014; Gautier 2015). De plus,

Jensen (2014) porte aussi un regard intéressant sur le fait que nous considérons souvent des systèmes à l'équilibre dans nos modèles, bien que ce soit rarement le cas dans la nature, ce qui peut compliquer la détection d'événements de sélection en cours.

Finalement, il parait compliqué de pouvoir rendre SelEstim plus adapté à la détection d'épisodes de sélection portant sur une mutation présente en forte fréquence et portée par plusieurs haplotypes. Il faudrait pour cela considérer, pour chaque locus non pas un mais plusieurs allèles sous sélection. Si la théorie de la diffusion peut considérer l'ensemble des allèles avec une valeur sélective associée (Barbour et al. 2000), la simplification du modèle en ne considérant qu'un seul allèle sous sélection est finalement ce qui nous permet de dériver la constante d'intégration C, indispensable au modèle SelEstim.

2.5.4 Conclusion

Nous avons donc développé un nouveau modèle de *scan* génomique capable d'exploiter l'information de DL en utilisant des marqueurs haplotypiques. Nous avons vu que notre méthode augmente considérablement la puissance de détection de la sélection ainsi que la précision du signal dans le cas d'un balayage sélectif fort s'appliquant dans des populations organisées selon un modèle en îles (Wright 1931). Cependant nous avons aussi vu que les résultats sont dépendants du modèle démographique analysé. Ainsi il est important lors de l'analyse de données de choisir un modèle adapté à la démographie de l'espèce lorsque celle-ci est connue, ou le plus robuste possible dans le cas contraire. J'insisterai aussi sur le fait que nous considérons un modèle démographique à l'équilibre. Il faut donc être précautionneux lors de l'analyse de données chez des espèces dont on sait que la démographie n'est pas à l'équilibre. Cela peut par exemple être le cas lors d'une invasion biologique ou après une colonisation, comme ça a été le cas chez l'homme, où l'on peut retrouver le phénomène d'*allele surfing* sur le front d'expansion (Hofer et al. 2009). Dans ce cas, l'*allele surfing* peut augmenter la fréquence des allèles dans la zone nouvellement colonisée et donc montrer des patrons de différenciation génétique élevée par rapport à d'autre sous-populations. Ces

patrons de différenciation, bien qu'uniquement dû à la dérive, peuvent être confondus avec des signaux de sélection.

Je terminerai sur le potentiel d'application de SelEstim$_{HAP}$ dans un futur proche à partir de données Pool-seq. En effet des études ont été menées pour essayer d'obtenir des phases localement, à partir de ces données dans un contexte d'évolution expérimentale. Si elles sont encore difficilement applicables et nécessitent souvent de connaître les différents haplotypes fondateurs présents dans le mélange (Long et al. 2011), d'autres méthodes s'affranchissent de cette contrainte. Dans un contexte d'évolution expérimentale où les différentes lignées vont être séquencées à plusieurs générations (Franssen et al. 2017) ont proposé de reconstruire les haplotypes en tenant compte des trajectoires corrélées des fréquences alléliques au cours d'un balayage sélectif fort (Franssen et al. 2017). Il n'est donc pas impossible que nous soyons capables dans un futur proche d'exploiter l'information d'haplotypes locaux issus de données Pool-seq pour la recherche de signatures de sélection dans les génomes.

Chapitre 3

Intégration d'une variable bayésienne auxiliaire et d'un modèle de lissage intégré

3.1 Introduction

Dans le chapitre précédent, nous avons développé une version multiallélique de SelEstim, un modèle hiérarchique bayésien de détection de la sélection. Cette version multiallélique nous a permis d'utiliser des marqueurs haplotypiques et d'exploiter l'information apportée par le déséquilibre de liaison (DL). L'ensemble des résultats obtenus sont basés sur un seul et unique critère de décision utilisé par SelEstim, à savoir la mesure de divergence KLD. Celle-ci consiste à mesurer l'écart de la distribution *a posteriori* des coefficients de sélection locus-spécifiques par rapport à une distribution de centrage, qui représente l'effet moyen de la sélection sur l'ensemble des marqueurs en supposant qu'ils sont tous soumis à un certain point à la sélection. Nous avons vu que cette mesure de KLD peut être calibrée par la simulation et l'analyse de données pseudo-observées (PODs), générées à partir des distributions *a posteriori* estimées sur le jeu de données d'intérêt. L'inconvénient principal de cette procédure (décrite dans la section Matériels et Méthodes du chapitre 2) est d'ordre computationnel, car la génération et la réanalyse

de jeux de données PODs peuvent s'avérer coûteuses en temps de calcul, et ce, d'autant plus, dans le cas de données multialléliques. Une approche alternative consiste à intégrer au modèle le fait de tester pour chaque locus s'il est sous sélection (coefficient $\sigma \neq 0$) ou non (coefficient $\sigma = 0$). À cette fin, nous proposons d'introduire une modélisation s'appuyant sur une variable auxiliaire binaire (Duforet-Frebourg et al. 2014; Gautier 2015; Riebler et al. 2008) attachée à chaque locus. Ce modèle va nous permettre de dériver un nouveau critère de décision, le facteur de Bayes (BF). Nous allons donc commencer par comparer, au moyen de données simulées, les performances des critères de décision basés sur la KLD ou le BF avec SelEstim.

De plus, si les données haplotypiques prises en compte par SelEstim$_{\text{HAP}}$ augmentent la puissance de détection de régions génomiques sous sélection, elles nécessitent une information préalable de phasage. Nous pouvons donc nous demander dans quelle mesure il est possible d'exploiter, dans le contexte des modèles SelEstim, l'information de DL lorsque nous n'avons pas accès aux phases (voir l'introduction du chapitre 2). Si nous disposons d'information sur la position relative des marqueurs les uns par rapport aux autres, la modélisation avec variable auxiliaire va en effet nous permettre d'introduire un modèle de lissage intégré (Duforet-Frebourg et al. 2014; Gautier 2015). Dans un second temps, nous comparerons les performances de SelEstim en appliquant ce type de modèle de lissage à des données SNPs considérées indépendantes, par rapport à l'utilisation de données haplotypiques.

3.2 Description du modèle SelEstim Auxiliaire

3.2.1 Ajout de la variable bayésienne auxiliaire

En considérant le modèle multiallélique saturé (au sens où tous les loci sont supposés sous sélection, c'est-à-dire $\delta_j > 0$ pour tout j) développé dans le chapitre 2, nous allons maintenant décrire le modèle SelEstim qui intègre une variable auxiliaire bayésienne (modèle AUX).

Nous avons jusqu'ici fait l'hypothèse que l'ensemble des marqueurs sont soumis à la sélection ; or il pourrait être intéressant de comparer à chaque

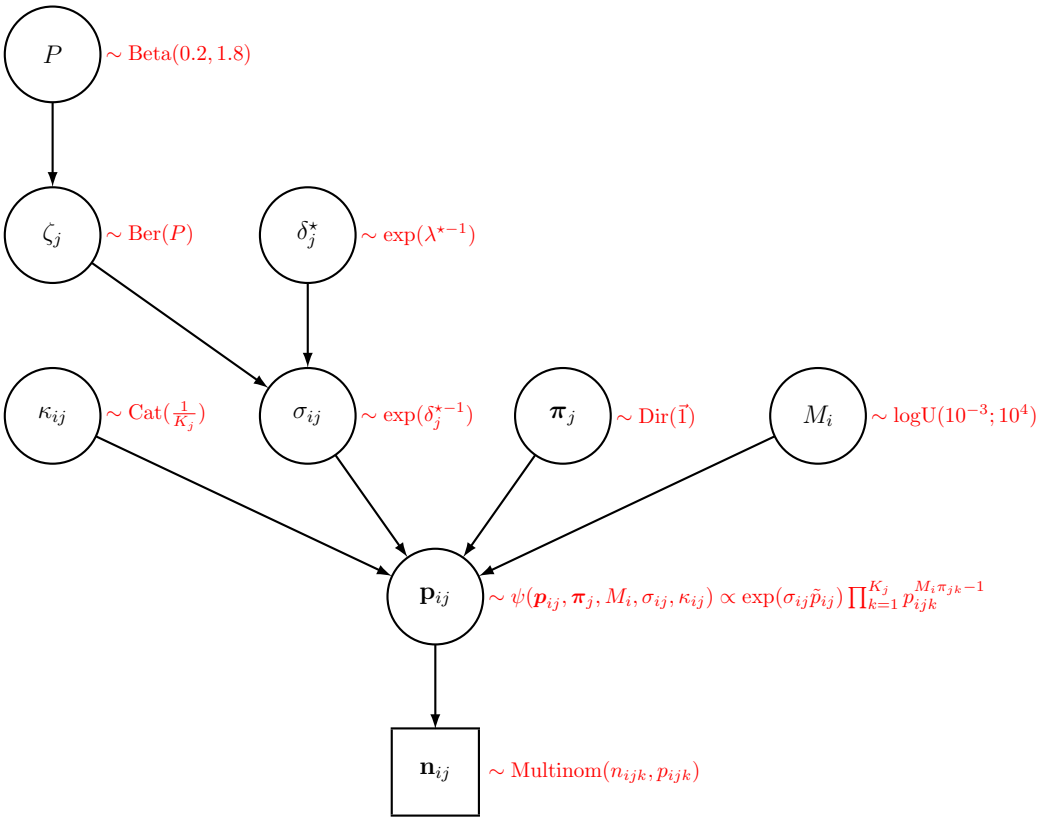


Figure 3.1 Graphe orienté acyclique (DAG) du modèle SelEstim AUX. Le modèle AUX est très similaire au modèle saturé décrit dans le chapitre 2 (voir Figure 2.5), mais contrairement à ce dernier, une variable auxiliaire binaire ζ_j est introduite afin d'indiquer si le locus j est sous sélection ($\zeta_j = 1$) ou non ($\zeta_j = 0$). La variable auxiliaire dépend d'un hyper-paramètre P qui peut s'interpréter comme la proportion de locus sous sélection.

locus un modèle nul considérant le locus comme neutre à un modèle pour lequel le locus est sous sélection (avec un effet locus-spécifique non nul). Par souci de clarté, nous nommerons dorénavant ce modèle $\text{SelEstim}^{\text{AUX}}_{\text{HAP}}$ lorsqu'il est utilisé avec des données haplotypiques et $\text{SelEstim}^{\text{AUX}}_{\text{SNP}}$ lorsqu'on utilise des données SNPs considérées comme indépendantes. Le modèle AUX présenté ici est une extension du modèle saturé décrit dans le chapitre 2. Le modèle AUX s'inspire des travaux de Riebler et al. (2008) et Gautier (2015) et consiste à utiliser une variable auxiliaire binaire et locus-spécifique, ζ_j. La valeur prise à un locus par cette variable indique si celui-ci est sous sélection ($\zeta_j = 1$), ou s'il est neutre ($\zeta_j = 0$, ce qui entraîne $\sigma_{ij} = 0$ pour tout dème i). On note $\boldsymbol{\zeta} \equiv (\zeta_1, \zeta_2, ..., \zeta_J)$ le vecteur des ζ_j sur l'ensemble des J locus. On assigne à chaque variable auxiliaire ζ_j la distribution *a priori* suivante :

$$f(\zeta_j|P) \propto Ber(P) \tag{3.1}$$

où P est un hyper-paramètre pouvant s'interpréter comme la proportion de marqueurs sous sélection dans le génome. On suppose une distribution *a priori* de type beta $f(P|a, b) \sim \text{Beta}(a, b)$, avec $a = 0.2$ et $b = 1.8$, ce qui correspond à une proportion moyenne de marqueurs sous sélection $\mathbb{E}(P) = 0.1$. Cette distribution *a priori* nous permet de supposer que seule une fraction du génome est sous sélection, tout en disposant d'une densité de probabilité non négligeable sur l'ensemble du domaine de définition [0 ;1] (voir Figure 3.2). L'estimation de P, contrairement à l'utilisation d'une valeur fixée, nous permet d'effectuer une correction pour les tests multiples. En effet, chaque locus peut être un candidat potentiel sous sélection, et nous disposons donc de 2^J modèles possibles, c'est à dire de combinaisons différentes de locus considérés sous sélection. Or, si nous avions fixé P nous aurions contraint la gamme de combinaisons à explorer. Ici, en estimant P il est possible de tester l'ensemble des modèles possibles avec des valeurs de P allant potentiellement de 0 à 1.

Ainsi, la distribution *a posteriori* $P[\zeta_j = 1|\mathcal{D}]$ est par définition la probabilité d'inclusion *a posteriori* (PIP) du locus j au modèle de sélection, en intégrant sur l'ensemble des modèles possibles. Nous allons donc chercher

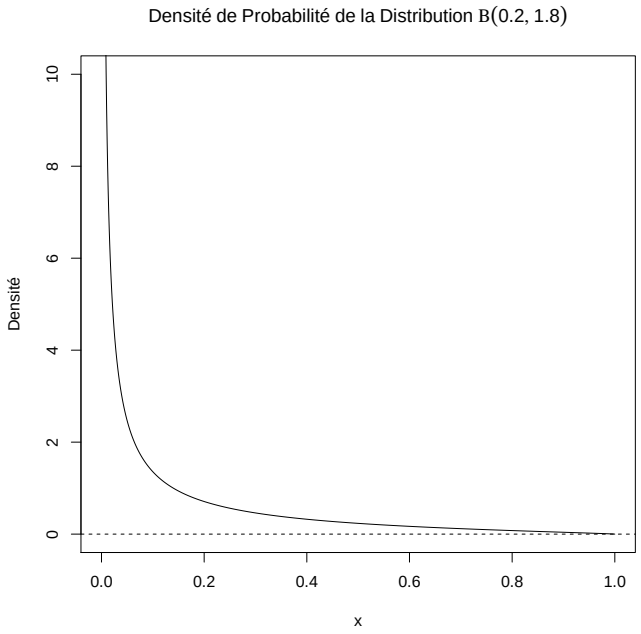


Figure 3.2 Distribution de la densité de probabilité d'une loi Beta(0.2, 1.8).

à estimer cette distribution. Avec l'ajout de la variable auxiliaire, quelques ajustements par rapport au modèle saturé vont être nécessaires. La principale modification intervient sur la spécification de la distribution *a priori* du coefficient de sélection locus- et population-spécifique σ_{ij}. En effet, déclarer un locus comme non sélectionné (i.e., $\zeta_j = 0$) revient à dire que $\forall_i \in (1, I), \sigma_{ij} = 0$, d'où la distribution *a priori* :

$$f(\sigma_{ij}|\zeta_j, \delta_j) = \begin{cases} 0, & \text{si } \zeta_j = 0. \\ \Gamma(1, \frac{1}{\delta_j^\star}), & \text{si } \zeta_j = 1. \end{cases} \tag{3.2}$$

avec $\delta_j^\star \sim \exp(\lambda^{\star -1})$ les nouveaux coefficients de sélection locus-spécifiques dont la distribution *a priori* est donnée par $f(\delta_j^\star|\lambda^\star) \sim \Gamma(1, \frac{1}{\lambda^\star})$.

Nous cherchons le paramètre $\lambda^\star$ tel que la distribution *a priori* des σ_{ij} ait la même espérance que celle des σ_{ij} du modèle saturé. Dans le cas du modèle

saturé on a :

$$\mathbb{E}(\sigma_{ij}|\lambda) = \mathbb{E}_{\delta}(\mathbb{E}(\sigma_{ij}|\delta_j, \lambda)) = \mathbb{E}_{\delta}(\mathbb{E}(\sigma_{ij}|\delta_j)) = \mathbb{E}_{\delta}(\delta_j) = \lambda \tag{3.3}$$

Pour le modèle AUX, l'espérance des σ_{ij} est :

$$\mathbb{E}(\sigma_{ij}|\lambda^{\star}) = \mathbb{E}_{\delta,\zeta}(\mathbb{E}(\sigma_{ij}|\zeta_j, \delta_j^{\star}, \lambda^{\star})) = \mathbb{E}_{\delta}(\delta_j^{\star})\mathbb{E}(\zeta_j) = P\lambda^{\star} \tag{3.4}$$

d'où $\lambda^{\star} = \frac{\lambda}{P}$. En pratique, et puisque nous ne connaissons pas λ, nous avons choisi de l'estimer à partir du modèle saturé durant une première phase de *burn-in* tel que décrit dans la partie Matériels et Méthodes.

Afin de discriminer les marqueurs sous sélection des marqueurs neutres, nous allons nous intéresser directement aux PIP ($P[\zeta_j = 1|\mathcal{D}]$) des différents marqueurs, estimées comme les moyennes des distributions *a posteriori* des variables auxiliaires correspondantes. Si on peut penser dans un premier temps à s'intéresser directement à la valeur des PIP pour définir un seuil critique (en utilisant une calibration par simulation par exemple, voir Riebler et al. 2008) au delà duquel un marqueur serait considéré sous sélection, cette approche négligerait la probabilité d'inclusion *a priori* du marqueur. Un moyen de mesurer à quel point la PIP d'un marqueur s'écarte de sa probabilité d'inclusion *a priori* est de calculer un facteur de Bayes (BF) qui, pour un locus donné, compare le modèle neutre ($\zeta_j = 0$) au modèle avec sélection ($\zeta_j = 1$). Les BF sont directement calculés à partir des moyennes des PIP des différents marqueurs et on définit le BF en déciban (dB) :

$BF_{dB} = 10 \log_{10}(\frac{cote\ \ a\ posteriori}{cote\ \ a\ priori})$
avec :
— cote *a priori* $= \mathbb{P}[\zeta_j = 1]/[1 - \mathbb{P}[\zeta_j = 1] = \frac{\mathbb{E}(P)}{1-\mathbb{E}(P)}$
— cote *a posteriori* $= \frac{\text{PIP}}{1-\text{PIP}}$

où, pour rappel, $P[\zeta_j = 1|\mathcal{D}]$ est la PIP du marqueur j parmi les 2^J modèles de sélection possibles, estimée comme la moyenne de la distribution

Tableau 3.1 Règle de Jeffreys et interprétation des différentes valeurs de BF

Valeurs de BF (en dB)	Interprétation
10<BF<15	Forte preuve
15<BF<20	Très forte preuve
BF>20	Preuve probante

a posteriori de la variable auxiliaire.

Lorsqu'on utilise le BF sur des jeux de données réels, nous appliquons la règle de décision de Jeffreys (Jeffreys 1961) définie selon l'échelle décrite dans le Tableau 3.1 pour interpréter les BF.

3.2.2 Intégration d'un modèle d'auto-corrélation spatiale de Ising & Potts

Le modèle AUX que nous avons décrit précédemment traite les marqueurs comme indépendants. Supposons que l'on dispose de l'information de la position relative des marqueurs les uns par rapport aux autres, il devient possible d'introduire une dépendance spatiale entre ceux-ci à l'aide d'un modèle de Ising sur une dimension. Le but étant d'encourager la proposition de modèles où des loci voisins sont déclarés sous sélection. Le modèle de Ising permet d'intégrer une dépendance spatiale entre les marqueurs dans la distribution *a priori* des ζ_j. La paramétrisation du modèle a été inspirée des travaux de Duforet-Frebourg et al. (2014) et Gautier (2015). Dans ce modèle, la distribution conditionnelle de $\boldsymbol{\zeta} \equiv (\zeta_1, ..., \zeta_J)$ est donnée par :

$$f(\boldsymbol{\zeta}|P, b_{Is}, \boldsymbol{\zeta}) \propto P^{S_1}(1-P)^{S_0} \exp^{\eta b_{Is}} \tag{3.5}$$

avec respectivement $S_1 = \sum_{j=1}^{J} 1_{\zeta_j=1}$ le nombre de marqueurs potentiellement sous sélection, et $S_0 = J - S_1$ le nombre de marqueurs neutres. $\eta = \mathbb{I}_{\zeta_j=\zeta_j-1} + \mathbb{I}_{\zeta_j=\zeta_j+1}$ (avec $\mathbb{I}$ la fonction identité) correspond au nombre

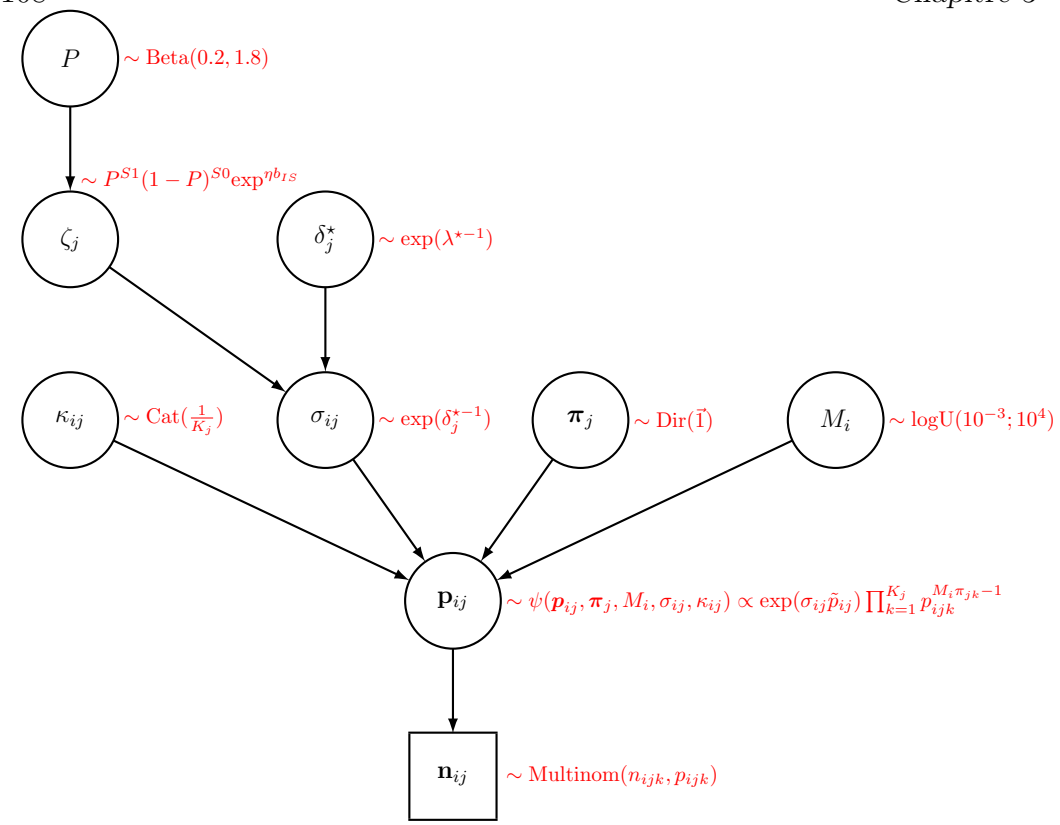


Figure 3.3 DAG de SelEstim avec modèle de lissage intégré de Ising . Ce modèle est strictement identique au modèle AUX à l'exception de la distribution *a priori* des variables auxiliaires ζ_j dans laquelle nous intégrons le modèle de dépendance spatiale.

de marqueurs voisins dont l'état de la variable auxiliaire est identique. b_{Is} est le paramètre de température inverse de Ising (si égal à 0, le modèle est identique au modèle AUX). Notons que puisque nous ne prenons en compte que la position relative des marqueurs et non leur position physique, nous supposons ici leur densité homogène le long du génome. Nous appellerons dorénavant ce modèle SelEstim$^{\text{Ising}}_{\text{SNP}}$ lorsqu'il est utilisé avec des données SNPs considérées comme indépendantes et lorsque $b_{Is} > 0$. Le DAG du modèle SelEstim AUX avec modèle de Ising est représenté dans la Figure 3.3.

Notons enfin qu'il est difficile de dériver simplement un BF avec SelEstim$^{\text{Ising}}_{\text{SNP}}$,

car le calcul de la cote *a priori* devient très compliqué lorsque $b_{Is} > 0$.

3.3 Matériels et Méthodes

3.3.1 Échantillonneur par MCMC

L'échantillonneur par MCMC du modèle AUX est similaire à celui du modèle saturé décrit dans le chapitre 2. En pratique, dans le cas du modèle AUX, on utilise le modèle saturé durant les chaînes pilotes (afin d'ajuster les différentes distributions instrumentales) et durant la première moitié de la phase de *burn-in* pour ajuster la distribution *a priori* sur les $\boldsymbol{\delta}$. Ce dernier point est nécessaire afin de régler les problèmes d'identifiabilité qui surviennent lorsque les paramètres $\boldsymbol{\zeta}$ et λ sont tous les deux inclus. Finalement, nous fixons λ à sa valeur moyenne obtenue à l'issue de la première phase de *burn-in*. Les paramètres σ_{ij} $\delta_j^\star$ et κ_{ij} sont mis à jour en échantillonnant dans leurs distributions *a priori* lorsque $\zeta_j = 0$ ou comme dans le modèle saturé lorsque $\zeta_j = 1$. Les paramètres ζ_j sont estimées conjointement par MCMC avec algorithme de Metropolis-Hasting tandis que l'hyper-paramètre P est estimé par un algorithme de Gibbs (les mises à jour sont décrites dans l'Annexe C).

3.3.2 Simulations

Afin d'évaluer les performances des modèles, nous avons réalisé une étude par simulation. Nous avons simulé des données de polymorphismes génétiques selon un modèle en îles (Wright 1931) en utilisant le même protocole que dans le chapitre 2, qui est illustré dans la Figure 2.9. De la même manière, nous avons utilisé 500 réplicats indépendants par jeu de paramètres. Cependant, nous avons uniquement simulé des données avec une intensité de sélection $\sigma = 2Ns = 100$ et un balayage sélectif fort (fréquence initiale f_0 de l'allèle minoritaire à 1%) et faible (fréquence initiale f_0 de l'allèle minoritaire à 20%).

Tableau 3.2 Résumé des notations liées aux mesures de performance

Notation	Description
Rappel	Taux de vrais positifs estimé comme la proportion de simulations pour lesquelles au moins un marqueur a été détecté sur le chromosome où la sélection agit
FDR	Taux de fausses découvertes estimé comme la proportion de faux positifs parmi les marqueurs détectés sous sélection
Précision	1-FDR

3.3.3 Comparaison des critères de décision KLD et BF, et évaluation des performances de SelEstim$^{\text{Ising}}_{\text{SNP}}$ sur la base des simulations

Nous allons comparer les critères KLD au critère BF en réalisant des courbes de Précision-Rappel (courbes PR). Ces courbes représentent le rappel (équivalent à la puissance ou taux de vrais positifs : voir Table 3.2) en fonction de la précision (définie comme 1-FDR, où FDR est le taux de fausse découverte : voir Table 3.2). Les courbes PR prennent mieux en compte que les courbes ROC (Davis et Goadrich 2006) le grand nombre de faux positifs potentiels (dans notre cas, par exemple, nous simulons 4 chromosomes neutres et un seul sous sélection dans chaque jeux de données). Lorsque nous comparons plusieurs courbes PR, celle ayant une aire sous la courbe maximale est celle montrant les meilleures performances (l'idée étant d'avoir pour une puissance donnée le plus faible FDR possible).

Nous évaluerons ensuite les performances du modèle SelEstim$^{\text{Ising}}_{\text{SNP}}$ en le comparant à SelEstim$_{\text{SNP}}$ et SelEstim$_{\text{HAP}}$, dont les marqueurs haplotypiques sont définis au moyen de l'algorithme par blocs décrit dans le chapitre 2 (voir Figure 2.4). Pour les comparaisons nous utilisons les mesures de puissance, de biais de position et de précision selon le protocole décrit dans le chapitre 2 (voir chapitre 2 – Matériels et Méthodes) avec une erreur de type I de 5% pour les mesures de qualité du signal.

L'ensemble des analyses a été réalisé sous R version 3.3.1 (R Core Team 2017), et les courbes PR ont été obtenues à l'aide du package R PRROC (Grau et al. 2015; Keilwagen et al. 2014)

3.4 Résultats

Un exemple de *scans* génomiques obtenus avec le modèle SelEstim$_{\text{SNP}}$, SelEstim$^{\text{AUX}}_{\text{SNP}}$ ainsi que SelEstim$^{\text{Ising}}_{\text{SNP}}$ (pour $b_{Is} = 0.5$ et 1.0), sur des données simulées sous le modèle en îles (Wright 1931) est présenté Figure 3.4. On peut voir que pour l'ensemble des modèles testés, nous obtenons un signal de sélection dans la région de la mutation causale simulée sous sélection. On observe un bruit supérieur pour SelEstim$^{\text{AUX}}_{\text{SNP}}$ par rapport aux autres modèles (voir Figure 3.4). Ce bruit semble dû à une forte variabilité pour les faibles BF (voir Figure 3.4-A). Enfin, on observe que l'utilisation d'un modèle de lissage avec SelEstim$^{\text{Ising}}_{\text{SNP}}$ améliore visuellement la précision du signal et réduit le bruit sur les chromosomes neutres de manière d'autant plus importante que la valeur de b_{Is} est élevée (voir Figure 3.4-C et D)

3.4.1 Comparaison des critères de décision BF et KLD

Pour chaque scénario, les courbes de Précision-Rappel (PR) moyennes issues des 500 jeux de données simulées pour les deux critères de décision de SelEstim, déclinés pour les données de type SNPs et haplotypiques, sont tracées dans la Figures 3.5 [1]. On note de manière générale que pour un type de marqueur donné, les critères KLD et BF montrent des performances similaires. On note que comme dans le chapitre 2, l'utilisation de marqueurs haplotypiques avec SelEstim$_{\text{HAP}}$ surpasse de loin les performances réalisées par SelEstim$_{\text{SNP}}$ pour un fort balayage sélectif (voir Figure 3.5 A). De même, dans ce cas, le critère BF de SelEstim$^{\text{AUX}}_{\text{SNP}}$ montre un très léger avantage de performance par rapport au critère KLD de SelEstim$_{\text{SNP}}$,

1. Nous ne présentons pas les résultats de SelEstim$^{\text{AUX}}_{\text{HAP}}$ pour lesquels on observe des puissances supérieures à 0.8 qui chutent à 0 tout en gardant une précision avoisinant les 70% ce qui indique une valeur de BF maximale plus élevée chez les marqueurs neutres que chez ceux appartenant au chromosome contenant le variant sous sélection.

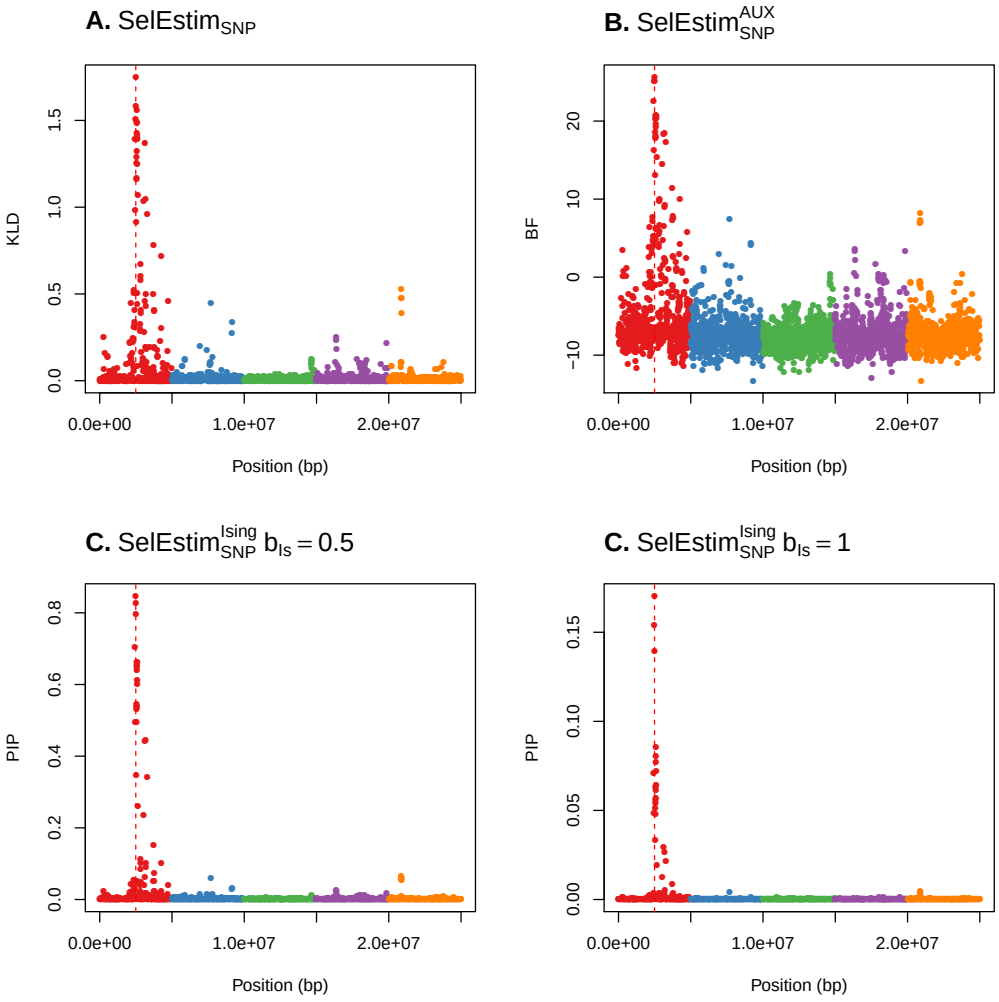


Figure 3.4 Exemple de *scans* génomiques obtenus pour un jeu de données simulé avec (A) SelEstim$_{SNP}$, et (B) SelEstim$^{AUX}_{SNP}$ ainsi que SelEstim$^{Ising}_{SNP}$ pour (C) $b_{Is} = 0.5$ et (D) $b_{Is} = 1$. Les différentes couleurs représentent les cinq chromosomes simulés et la position simulée sous sélection est indiquée par la ligne pointillée rouge.

avec respectivement une aire sous la courbe de 0.80 et de 0.78. Ces premiers résultats ne permettent donc pas de discriminer clairement les critères de décisions KLD et BF sur la base de leurs performances.

On observe cependant un avantage certain du BF par rapport à la KLD lorsqu'on s'intéresse au temps d'analyse moyen sur les données simulées avec,

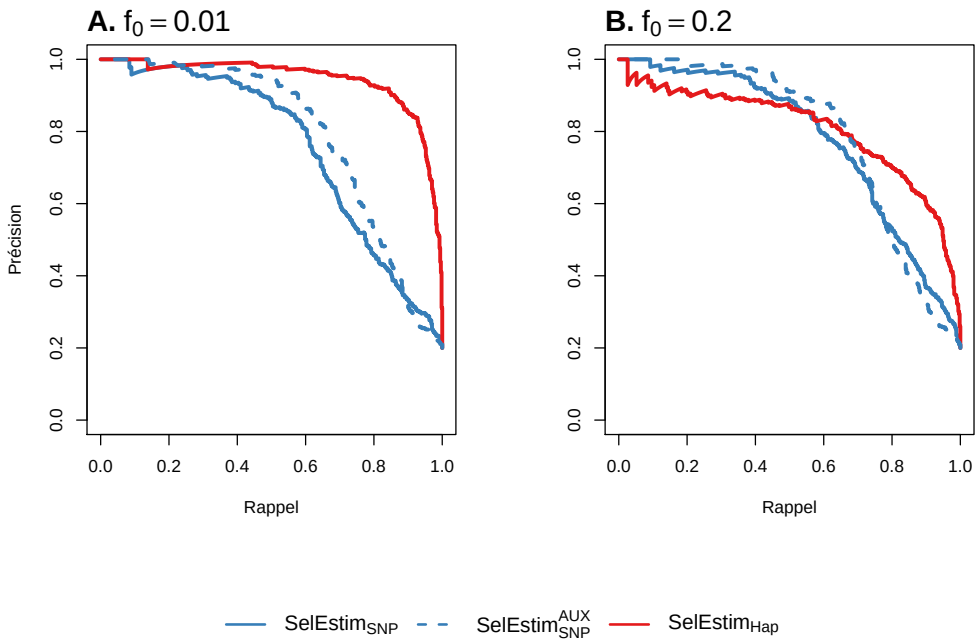


Figure 3.5 Courbes de Précision-Rappel pour les simulations sous le modèle en îles. Les résultats sont présentés (A) pour un balayage sélectif fort et (B) faible avec respectivement des fréquences initiales de l'allèle sous sélection f_0 de 1% (A) et 20% (B).

pour un type de données, un temps de calcul deux fois plus faible pour les modèles AUX (voir Tableau 3.3).

Tableau 3.3 Temps de calculs moyen des différents modèles sur 500 réplicats avec un balayage sélectif fort ($f_0 = 1\%$) et faible ($f_0 = 20\%$) dans un modèle en îles. Les temps de calcul des modèles utilisant la KLD sont estimés en doublant les temps de calcul relevés après l'analyse des jeux de données, ceci afin d'éviter la génération et l'analyse des jeux de données PODs.

fo	Modèle (critère de décision)	Temps de calcul moyen (en sec.)
1%	SelEstim$_{\text{SNP}}$ (KLD)	5728.5
1%	SelEstim$^{\text{AUX}}_{\text{SNP}}$ (BF)	2570.5
1%	SelEstim$_{\text{HAP}}$ (KLD)	30421.2
1%	SelEstim$^{\text{AUX}}_{\text{HAP}}$ (BF)	15213.4
20%	SelEstim$_{\text{SNP}}$ (KLD)	5628.0
20%	SelEstim$^{\text{AUX}}_{\text{SNP}}$ (BF)	2461.4
20%	SelEstim$_{\text{HAP}}$ (KLD)	29272.2
20%	SelEstim$^{\text{AUX}}_{\text{HAP}}$ (BF)	14223.9

3.4.2 Évaluation des performances de SelEstim$^{\text{Ising}}_{\text{SNP}}$

Afin de voir dans quelle mesure nous pouvons exploiter l'information de DL des données à partir du modèle d'auto-corrélation spatiale implémenté dans SelEstim$^{\text{Ising}}_{\text{SNP}}$, nous avons comparé ses performances aux modèles SelEstim$_{\text{SNP}}$ et SelEstim$_{\text{HAP}}$, en faisant varier l'intensité d'homogénéisation des valeurs de la variable auxiliaire entre marqueurs voisins ($b_{Is} =$ 0.5, 0.8 et 1.0). Si l'on observe, en accord avec le chapitre 2, une puissance maximale pour SelEstim$_{\text{HAP}}$, on a une diminution de puissance avec SelEstim$^{\text{Ising}}_{\text{SNP}}$ par rapport à SelEstim$_{\text{SNP}}$ (voir Figure 3.6). La diminution de puissance est d'autant plus importante que le paramètre b_{Is} est faible et lorsqu'on se trouve dans un cas d'un balayage sélectif faible (voir Figure 3.6-B).

Les mesures liées à la qualité du signal montrent que SelEstim$^{\text{Ising}}_{\text{SNP}}$ fait preuve du biais de position le plus faible, et cela pour les différentes valeurs

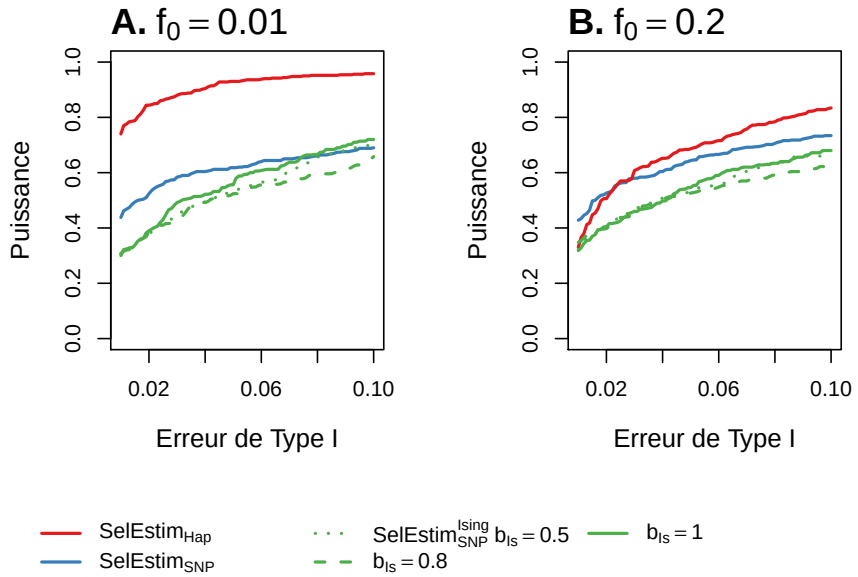


Figure 3.6 Puissance en fonction de l'erreur de type I dans un modèle en îles avec $n_d = 4$ dèmes. Les résultats sont représentés pour un balayage sélectif fort avec une fréquence initiale de l'allèle sous sélection $f_0 = 1\%$ (A), et faible avec une fréquence initiale de l'allèle sous sélection $f_0 = 20\%$.

de paramètre b_{Is} sans que la valeur prise par ce dernier résulte en une différence flagrante (voir Figure 3.7-A-B). La précision du signal, SELESTIM$^{\text{Ising}}_{\text{SNP}}$ est similaire à celle de SELESTIM$_{\text{SNP}}$, qui reste moins bonne que celle obtenue en utilisant des marqueurs haplotypiques avec SELESTIM$_{\text{HAP}}$. Là encore, les différentes valeurs testées pour le paramètre b_{Is} ne montrent pas d'effet important sur le résultat obtenu.

3.5 Discussion

3.5.1 Le choix du critère de décision

Vitalis et al. (2014) ont présenté le modèle SELESTIM$_{\text{SNP}}$ en utilisant un unique critère de décision, la divergence KLD. Ils émettaient déjà l'idée de pouvoir calculer un BF, que ce soit par une stratégie d'algorithme *Reversible*

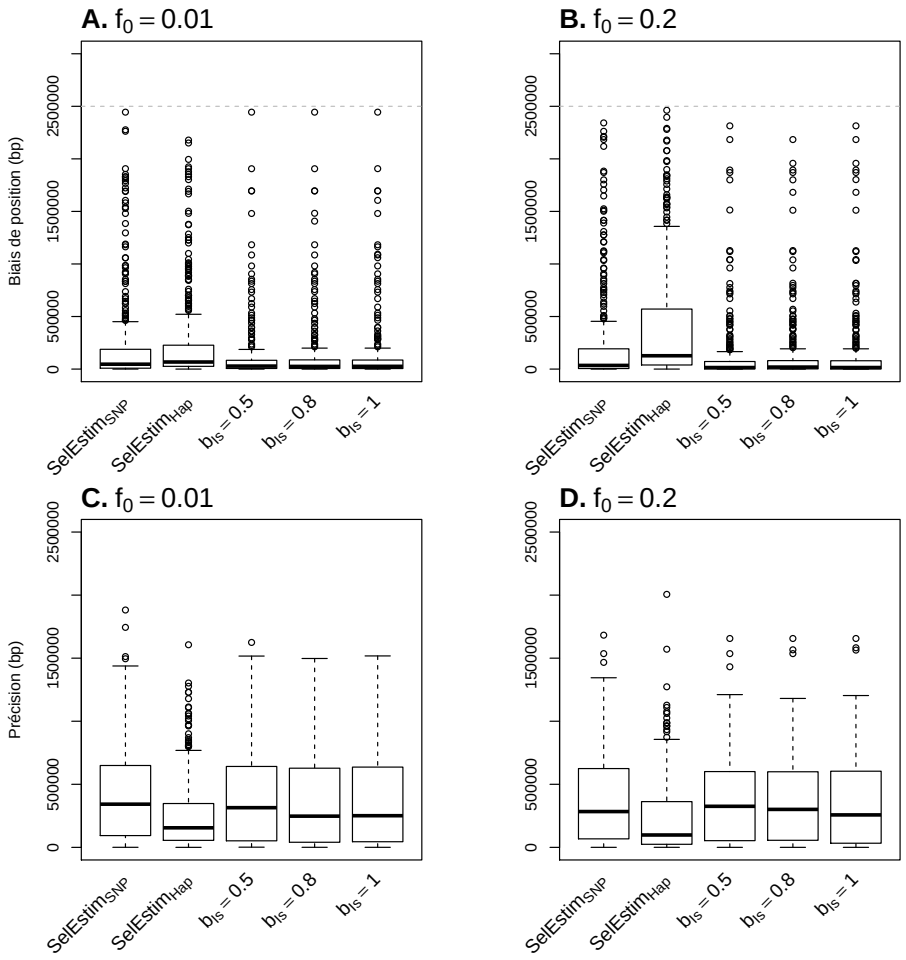


Figure 3.7 (A-B) Distribution du biais de position (distance médiane de la/les valeur(s) maximale(s) à la position sous sélection) et de (C-D) la précision (écart-type de la distance à la position sous sélection des différents marqueurs détectés positivement) pour une erreur de type I de 5% dans un modèle en île avec $n_d = 4$ dèmes. Les résultats sont représentés selon deux fréquences initiales de l'allèle sous sélection, 1% (A-C) et 20% (B-D). Les comparaisons sont effectuées entre les différents modèles, SELESTIM$_{\text{HAP}}$, SELESTIM$_{\text{SNP}}$, et SELESTIM$^{\text{Ising}}_{\text{SNP}}$ avec trois valeurs de paramètre b_{Is} : 0.5, 0.8 et 1.0.

Jump MCMC comme c'est le cas dans Foll et Gaggiotti (2008), ou bien encore en réalisant une reparamétrisation du modèle, inspirée de Riebler et al. (2008). Cette dernière approche a été retenue pour construire notre version

AUX de SelEstim. Nous avons pu voir que les deux critères, KLD pour le modèle saturé et BF pour le modèle AUX, montrent des performances similaires. Bien que le modèle AUX soit moins coûteux computationellement que le modèle saturé qui demande la simulation et l'analyse d'un jeux de données PODs, il montre aussi des limites. En effet, comme cela à déjà été discuté par Vitalis et al. (2014), il faut noter que, contrairement à la KLD, l'estimation du BF dépend en pratique de la longueur de la chaîne de Markov utilisée pour l'analyse. Cette dernière va d'ailleurs déterminer la valeur maximale qui peut être prise par le BF. Il peut en résulter un phénomène de saturation des BF (voir Figure 3.8), pour lequel de nombreux marqueurs montrent le même signal de sélection (Vitalis et al. 2014). Dans le cas où nous voulons uniquement utiliser le BF pour discriminer les régions génomiques identifiées sous sélection, sans classification des différents signaux, cela ne pose pas de problème. Dans le cas contraire, la KLD semble être plus appropriée en apportant un support visuel intéressant permettant de discriminer plus facilement les signaux en fonction de leur intensité.

Finalement, le problème computationnel lié à la KLD se résume simplement à un problème de calibration pour le choix du critère de décision. Une approche alternative à été testée par Vitalis et al. (2014). Elle repose sur l'argument de Peng et Dey (1995), qui définissent une valeur limite de KLD sur la base d'une expérience de pensée : imaginons un tirage entre deux pièces, l'une truquée (avec une probabilité ν pour pile) et l'autre non. Une KLD peut alors être calculée entre une distribution de Bernoulli de probabilité ν (tirage truqué) et une distribution de Bernoulli de probabilité 0.5 (tirage non truqué). Une valeur-seuil de KLD peut alors être définie comme la valeur de KLD obtenue par le calcul de la KLD entre deux distributions de Bernoulli, Ber(0.5) et Ber(ν). Cette valeur-seuil de KLD peut alors être utilisée comme critère de décision au seuil ν. Si cette approche est extrêmement rapide à mettre en place, elle avait été jugée trop conservative (Vitalis et al. 2014). Une autre calibration, quasi-immédiate à mettre en œuvre pourrait être envisagée si l'on trouvait une transformation de la KLD permettant d'approcher une distribution connue, et donc d'utiliser les quantiles de cette dernière pour dériver un critère de décision. Malheureusement, nous n'avons pas trouvé à

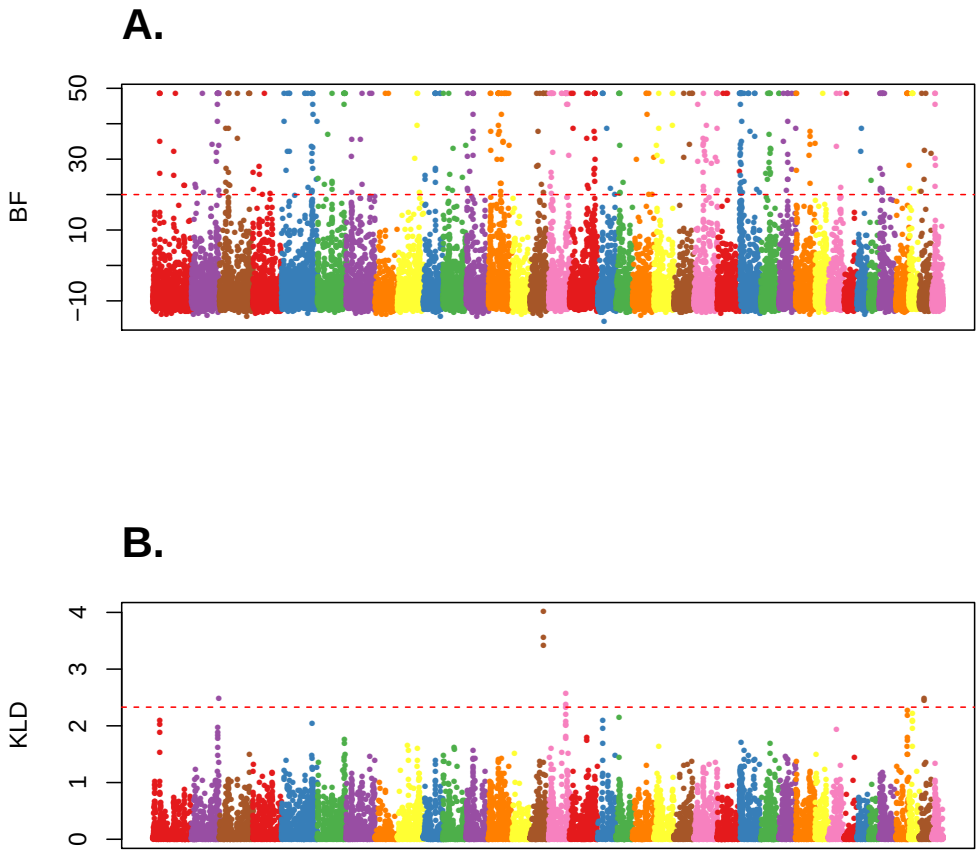


Figure 3.8 Illustration du phénomène de saturation des BF par rapport au critère de type KLD. Les deux *scans* génomiques ont été réalisés avec SelEstim$^{\text{AUX}}_{\text{HAP}}$ (A) SelEstim$_{\text{HAP}}$ (B) sur un jeux de données haplotypiques issu de Schweizer et al. (2016). Le jeux de données contient 6 populations de loups et 42 587 SNPs répartis sur 38 chromosomes (les différentes couleurs représentent les différents chromosomes). Les données haplotypiques ont été obtenues en phasant les génotypes avec le logiciel `fastphase` (Scheet et Stephens 2006), puis en utilisant un algorithme de regroupement local par blocs (voir Figure 2.4) avec $K = 20$. Les lignes rouges en tirés représentent le seuil de BF=20 (A) et le seuil de KLD à 99% estimé par la réanalyse d'un jeu de données PODs (B).

l'heure actuelle de transformation satisfaisante de la KLD et devons donc nous contenter de l'approche par PODs.

3.5.2 Exploiter l'information de liaison en absence de données haplotypiques

Dans ce chapitre, le développement d'une version AUX de SelEstim nous a permis d'intégrer un modèle d'auto-corrélation spatiale des marqueurs. L'avantage de cette modélisation est son potentiel d'application sur des données non-phasées. Nous avons donc pu évaluer l'effet d'une méthode de lissage intégrée au modèle en utilisant des marqueurs SNPs considérés comme indépendants, afin d'évaluer les possibilités d'exploitation du DL en absence de données haplotypiques. En terme de puissance de détection des régions génomiques sous sélection, l'utilisation de données haplotypiques montre de bien meilleures performances que les données SNPs malgré l'utilisation de la méthode de lissage. Nous avons même pu observer une baisse de puissance avec $\text{SelEstim}^{\text{Ising}}_{\text{SNP}}$ par rapport au modèle saturé de SelEstim. Cette diminution de puissance peut être liée à la correction pour les tests multiples intégrée au modèle AUX. De plus, contrairement à $\text{SelEstim}_{\text{HAP}}$ qui s'intéresse directement aux fréquences haplotypiques, $\text{SelEstim}^{\text{Ising}}_{\text{SNP}}$ s'intéresse uniquement aux fréquences alléliques marginales qui ne contiennent pas la même information de liaison (voir Introduction Générale). Notons tout de même que nous avons observé un gain important sur la qualité du signal avec un biais de position plus faible lorsque nous utilisons le modèle $\text{SelEstim}^{\text{Ising}}_{\text{SNP}}$. De plus, si la valeur de l'intensité d'auto-corrélation spatiale (b_{Is}) ne semble pas avoir d'effet sur les résultats liés à l'évaluation du modèle, on distingue un effet clair lorsqu'on s'intéresse directement aux résultats graphiques des *scans*. En effet, la Figure 3.4 montre une nette amélioration du lissage avec l'augmentation de l'intensité d'auto-corrélation spatiale.

Ainsi, dans la limite d'un compromis entre puissance du modèle et qualité du signal espérée, le modèle AUX de SelEstim avec modèle de Ising semble être une bonne alternative lorsque nous ne pouvons pas avoir accès à des données haplotypiques. C'est le cas par exemple des données issues de

séquençage d'individus en mélange (Pool-seq). Ce dernier type de données peut-être assez facilement intégré au modèle SelEstim : on considère alors que les lectures de l'allèle de référence à un locus j et dans un dème i sont issues d'un tirage binomial de paramètre c_{ij} (la couverture), et (x_{ij}/n_{ij}) où x_{ij} est le nombre de comptages (non-observés) de l'allèle de référence et n_{ij} la taille du pool. Ce type d'implémentation est déjà réalisé dans la version actuelle de SelEstim (Vitalis et al. 2014) ou encore dans le modèle BayPass de Gautier (2015).

Conclusion

Dans cette thèse, nous avons abordé la question de l'analyse de la différenciation génétique sous plusieurs angles complémentaires. Dans la première partie, nous avons développé un estimateur non biaisé de F_{ST} pour des données de type Pool-seq par la méthode des moments, reposant sur une approche de décomposition de la variance. Nous avons montré, à travers une application sur un jeu de données réelles (Dennenmoser et al. 2017), comment l'utilisation d'un estimateur non-biaisé pouvait remettre en cause l'interprétation de la structure génétique d'une espèce.

L'analyse de la distribution de mesures de différenciation génétique le long du génome est souvent réalisée dans un contexte de recherche de signatures de sélection. Or nous avons montré que certains estimateurs du paramètre F_{ST} pour des données de type Pool-seq pouvaient être biaisés, avec un biais qui dépend en partie de la couverture (comme c'est le cas par exemple pour l'estimateur par défaut du logiciel Popoolation2 : voir Kofler et al. 2011). L'utilisation de tels estimateurs dans une approche de type *genome scan* pourrait donc potentiellement produire des variations de F_{ST} le long du génome qui seraient uniquement dues à des variations de couverture et non à la sélection. Il serait donc intéressant de comparer les performances de ces estimateurs et de notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ dans le cadre de *scans* génomiques de différenciation. Dans ce contexte, nous pourrions également réaliser des estimations multilocus de F_{ST} le long de fenêtres glissantes, afin de réduire l'effet de la variabilité des mesures de F_{ST} locus-specifiques (voir, par exemple, Weir et al. 2005).

De tels *scans* génomiques de différenciation montrent toutefois un certain nombre de limites, notamment dans le cas où il existe une structure hié-

rarchique des populations (Robertson 1975). Pour y pallier, nous pourrions étendre le modèle FLK (Bonhomme et al. 2010) aux données Pool-seq. FLK suppose un modèle de populations divergeant en pure dérive et repose sur l'estimation de la matrice $\mathcal{F}$ de variance-covariance des fréquences alléliques par le calcul des distances génétiques de Reynolds (Reynolds et al. 1983). Ce point demanderait donc d'étendre l'estimateur de distance de Reynolds (Reynolds et al. 1983) aux données Pool-seq. Cela parait réalisable, d'autant plus que Reynolds et al. (1983) proposent un estimateur basé sur une approche de décomposition de la variance. Rappelons que FLK, tout comme le test LK original (Lewontin et Krakauer 1973) suppose que la distribution neutre du test suit une distribution du Chi-2 avec $n_{\mathrm{d}} - 1$ degrés de liberté, où n_{d} est le nombre de dèmes échantillonnés. Il faudrait donc s'assurer que la statistique qui résulterait de cette extension suit toujours cette distribution.

Une autre extension pourrait être envisagée avec OutFlank (Whitlock et Lotterhos 2015). Comme pour le test FLK (Bonhomme et al. 2010), OutFlank (Whitlock et Lotterhos 2015) vise à corriger le test LK de Lewontin et Krakauer (1973) dans le cas de populations structurées hiérarchiquement, et donc à diminuer le taux de faux positifs. OutFlank repose sur l'estimation d'une distribution nulle de F_{ST}, dans le cas neutre, sans supposer un quelconque modèle démographique *a priori*. En supposant que la majorité des marqueurs est neutre, la distribution des F_{ST} observés est ajustée à une distribution du Chi-2, qui est ensuite utilisée pour identifier les marqueurs atypiques. OutFlank pourrait donc être facilement étendu à l'utilisation de données Pool-seq, en substituant l'estimateur de F_{ST} par défaut par notre estimateur $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$.

Dans la seconde partie de ma thèse, nous avons mis en œuvre plusieurs stratégies pour prendre en compte l'information de déséquilibre de liaison (DL) pour la recherche de signatures de sélection. Nous avons proposé différents développements au modèle bayésien SelEstim. Dans un premier temps, nous avons fait le choix de nous éloigner du contexte de données Pool-seq pour proposer un modèle qui exploite l'information de DL à partir de données haplotypiques. Dans un second temps, nous avons introduit un modèle de lissage à SelEstim qui prend en compte la dépendance spatiale

entre marqueurs adjacents. Cette approche est applicable à des données bialléliques, issues de génotypes individuels ou en mélange (Pool-seq). Au vu de nos résultats, et bien que la stratégie alternative de lissage améliore la qualité du signal de sélection, l'utilisation d'haplotypes semble être la meilleure stratégie pour exploiter l'information de DL contenue dans les données.

Dans le contexte de données Pool-seq il paraît donc difficile de pouvoir pleinement exploiter cette information de DL, car la caractérisation de blocs haplotypiques à partir de données de lectures n'est pas un problème trivial. Des méthodes existent actuellement pour inférer des fréquences haplotypiques à partir de données Pool-seq (Cao et Sun 2015; Kessner et al. 2013; Long et al. 2011), mais elles nécessitent de connaître l'ensemble des haplotypes ancestraux, ce qui représente une limitation importante. Franssen et al. (2017) ont quant à eux développé une méthode qui permet d'inférer les haplotypes sous sélection à partir de données Pool-seq dans un contexte d'évolution expérimentale, en séquençant les populations à différentes générations. Leur méthode suppose que le variant sous sélection au cours de l'expérimentation est initialement présent en faible fréquence, et qu'un fort balayage sélectif a lieu pendant l'expérience. Dans ces conditions, ils supposent que peu d'événements de recombinaison ont lieu et ils caractérisent l'évolution des fréquences alléliques corrélées au cours du temps pour reconstruire les haplotypes au voisinage des mutations sous sélection. Les différentes méthodes proposées offrent donc des perspectives prometteuses pour inférer des données haplotypiques à partir de données Pool-seq, mais elles restent néanmoins peu adaptées à l'étude des populations naturelles. Cependant, les avancées technologiques nous permettront très certainement de séquencer des fragments de plus en plus longs et d'avoir ainsi accès à des données phasées localement, même dans un contexte de Pool-seq.

Les différentes méthodes proposées dans cette thèse pour détecter la sélection consistent à rechercher des locus fortement différenciés, sans considération de la variable causale qui crée la pression de sélection. Dans le cas où l'on dispose de données environnementales, pouvoir identifier des locus sous sélection spécifiquement associés à ces covariables environnementales serait particulièrement intéressant. Ce type d'approche a déjà été développée par

Coop et al. (2010), De Villemereuil et Gaggiotti (2015) et Gautier (2015). Puisque SelEstim inclut des paramètres de sélection population-spécifiques (σ), l'association d'une covariable environnementale à ce niveau semble assez naturelle. Une telle modélisation n'est cependant pas triviale puisqu'en l'état, la sélection est modélisée hiérarchiquement dans SelEstim. Inclure une dépendance à une covariable environnementale demanderait de gommer cette structure hiérarchique.

Enfin, dans une mise en perspective plus globale de nos travaux, nous nous devons d'évoquer certaines limites intrinsèques aux méthodes développées pour la recherche de signatures de sélection. Pritchard et al. (2010) ont par exemple montré que l'analyse d'un même jeu de données par différentes méthodes pouvait conduire à la caractérisation de différentes signatures de sélection. Chaque méthode étant développée selon différentes hypothèses (modèles démo-génétiques, types de sélection), il est difficile de pouvoir interpréter ces résultats en l'absence d'une validation fonctionnelle des gènes identifiés comme étant sous sélection. Il faut donc considérer les *scans* génomiques comme une étape préliminaire à l'identification de gènes "candidats", qu'il faudra valider par la suite par des approches fonctionnelles pour pouvoir confirmer ou infirmer des hypothèse biologiques. Enfin, notons que même en présence d'un signal de sélection potentiellement associé à une variable environnementale, des effets confondants autres que la sélection (par exemple des mécanismes d'auto-incompatibilité pré- et/ou post-zygotiques) peuvent expliquer des patrons de sur-différenciation (Bierne et al. 2011).

Bibliographie

Abramowitz, M. et Stegun, I. A. (1964). *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

Akey, J. M., Ruhe, A. L., Akey, D. T., Wong, A. K., Connelly, C. F., Madeoy, J., Nicholas, T. J., et Neff, M. W. (2010). Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci. USA*, 107(3) :1160–1165.

Akey, J. M., Zhang, G., Jin, L., et Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.*, 12 :1805–1814.

Anderson, E. C., Skaug, H. J., et Barshis, D. J. (2014). Next-generation sequencing for molecular ecology : a caveat regarding pooled samples. *Mol. Ecol.*, 23 :502–512.

Andrews, K. R. et Luikart, G. (2014). Recent novel approaches for population genomics data analysis. *Mol. Ecol.*, 23(7) :1661–7.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571) :68–74.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, 3(10) :e3376.

Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., et Neiman, M. (2017). Genomic evidence for population-specific responses to co-evolving parasites in a new zealand freshwater snail. *Mol. Ecol.*, 26(14) :3663–3675.

Barbour, A. D., Ethier, S. N., et Griffiths, R. C. (2000). A transition function expansion for a diffusion model with selection. *Ann. Appl. Probab.*, 10(1) :123–162.

Barton, N. et Rouhani, S. (1987). The frequency of shifts between alternative equilibria. *J. Theor. Biol.*, 125 :397–418.

Barton, N. H. et Turelli, M. (1987). Adaptive landscapes, genetic distance and the evolution of quantitative characters. *Genet. Res.*, 49(02) :157–173.

Bastide, H., Lange, J. D., Lack, J. B., Yassin, A., et Pool, J. E. (2016). A variable genetic architecture of melanic evolution in *Drosophila melanogaster*. *Genetics*, 204(3) :1307–1319.

Beaumont, M. A. (2005). Adaptation and speciation : What can F_{ST} tell us ? *Trends Ecol. Evol.*, 20(8) :435–440.

Beaumont, M. A. et Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, 13(4) :969–980.

Beaumont, M. A. et Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263 :1619–1626.

Belcaid, M. et Toonen, R. J. (2014). Demystifying computer science for molecular ecologists. *Mol. Ecol.*, 24(11) :2619–2640.

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change : considerations from population genomics in fishes. *J. Fish Biol.*, 89(6) :2519–2556.

Bertelsen, H. P., Gregersen, V. R., Poulsen, N., Nielsen, R. O., Das, A., et al. (2016). Detection of genetic variation affecting milk coagulation properties in danish holstein dairy cattle by analyses of pooled whole-genome sequences from phenotypically extreme samples (pool-seq). *Journal of Animal Science*, 94(4) :1365–1376.

Bhatia, G., Patterson, N., Sankararaman, S., et Price, A. L. (2013). Estimating and interpreting F_{ST} : the impact of rare variants. *Genome Res.*, 23 :1514–1521.

Bierne, N., Welch, J., Loire, E., Bonhomme, F., et David, P. (2011). The coupling hypothesis : Why genome scans may fail to map local adaptation genes. *Miol. Ecol.*, 20(10) :2044–2072.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et SanCristobal, M. (2010). Detecting selection in population trees : The Lewontin and Krakauer test extended. *Genetics*, 186(1) :241–262.

Bourguinat, C., Lefebvre, F., Sandoval, J., Bondesen, B., Moreno, Y., et Prichard, R. K. (2017). *Dirofilaria immitis* jyd-34 isolate : whole genome analysis. *Parasites Vectors*, 10(2) :494.

Browning, S. R. et Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-Data inference for whole-Genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81 :1084–1097.

Browning, S. R. et Weir, B. S. (2010). Population structure with localized haplotype clusters. *Genetics*, 185(4) :1337–1344.

Burke, G. R. (2016). Analysis of genetic variation across the encapsidated genome of *Microplitis demolitor* bracovirus in parasitoid wasps. *PLOS ONE*, 11(7) :e0158846.

Cao, C. C. et Sun, X. (2015). Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplo-

type carriers by overlapping pool sequencing. *Bioinformatics*, 31(4) :515–522.

Carvajal-Rodríguez, A. (2017). Hacdivsel : Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations. *PLOS ONE*, 12(4) :e0175944.

Casillas, S. et Barbadilla, A. (2017). Molecular population genetics. *Genetics*, 205 :1003–1035.

Cavalli-Sforza, L. L. (1966). Population structure and human evolution. *Proc. R. Soc. Lond., B, Biol. Sci*, 164(995) :362–379.

Chen, J., Källman, T., Ma, X.-F., Zaina, G., Morgante, M., et Lascoux, M. (2016). Identifying genetic signatures of natural selection using pooled populations sequencing in *Picea abies*. *G3*, 6 :1979–1989.

Choi, Y.-j., Bisset, S. A., Doyle, S. R., Hallsworth-pepin, K., Martin, J., Grant, W. N., et Mitreva, M. (2017). Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLOS Genet.*, 13(1) :e1006857.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, 23 :72–84.

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, 74 :679–700.

Cockerham, C. C. et Weir, B. S. (1987). Analyses of gene frequencies. *Proc. Natl. Acad. Sci. USA*, 84 :8512–8514.

Collet, J. M., Fuentes, S., Hesketh, J., Hill, M. S., Innocenti, P., Morrow, E. H., Fowler, K., et Reuter, M. (2016). Rapid evolution of the inter-sexual genetic correlation for fitness in *Drosophila melanogaster*. *Evolution*, 70 :781–795.

Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J., et Kocher, T. D. (2017). A high quality assembly of the Nile Tilapia (*Oreochromis*

niloticus) genome reveals the structure of two sex determination regions. *BMC Genomics*, 18(1) :341.

Coop, G., Witonsky, D., Di Rienzo, A., et K., P. J. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185 :1411–1423.

Cutler, D. J. et Jensen, J. D. (2010). To pool, or not to pool ? *Genetics*, 186 :41–43.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. Murray, London. or the Preservation of Favored Races in the Struggle for Life.

Davey, J. L. W., Davey, J. L. W., Blaxter, M. W. L., et Blaxter, M. W. L. (2010). RADSeq : next-generation population genetics. *Brief. Funct. Genomics*, 9(5-6) :416–23.

Davis, J. et Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

De Villemereuil, P. et Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11) :1248–1258.

Deitz, K. C., Athrey, G. A., Jawara, M., Overgaard, H. J., Matias, A., et Slotman, M. A. (2016). Genome-wide divergence in the west-african malaria vector *Anopheles melas*. *G3 : Genes, Genomes, Genetics*, 6(9) :2867–2879.

Dennenmoser, S., Nolte, A. W., Vamosi, S. M., et Rogers S, M. (2015). Phylogeography of the prickly sculpin (*Cottus asper*) in north-western North America reveals parallel phenotypic evolution across multiple coastal-inland colonizations. *J. Biogeogr.*, 42 :1626–1638.

Dennenmoser, S., Rogers, S. M., et Vamosi, S. M. (2014). Genetic population structure in prickly sculpin (*Cottus asper*) reflects isolation-by-

environment between two life-history ecotypes. *Biol. J. Linnean Soc.*, 113 :943–957.

Dennenmoser, S., Vamosi, S. M., Nolte, S. W., et Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Mol. Ecol.*, 26 :25–42.

Dexter, E., Bollens, S. M., Cordell, J., Soh, H. Y., Rollwagen-Bollens, G., Pfeifer, S. P., Goudet, J., et Vuilleumier, S. (2017). A genetic reconstruction of the invasion of the calanoid copepod *Pseudodiaptomus inopinus* across the North American Pacific Coast. *Biol. Invasions*, 20(6) :1577–1595.

Donnelly, P., Nordborg, M., et Joyce, P. (2001). Likelihoods and simulation methods for a class of nonneutral population genetics models. *Genetics*, 159 :853–867.

Douglas Nychka, Reinhard Furrer, John Paige, et Stephan Sain (2017). fields : Tools for spatial data. R package version 9.6.

Doyle, S. R., Bourguinat, C., Nana-Djeunga, H. C., Kengne-Ouafo, J. A., Pion, S. D. S., et al. (2017). Genome-wide analysis of ivermectin response by *Onchocerca volvulus* reveals that genetic drift and soft selective sweeps contribute to loss of drug sensitivity. *PLOS Neglected Tropical Diseases*, 11(7) :1–31.

Druet, T., Ahariz, N., Cambisano, N., Tamma, N., Michaux, C., Coppieters, W., Charlier, C., et Georges, M. (2014). Selection in action : dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC Genomics*, 15 :796.

Duforet-Frebourg, N., Bazin, E., et Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using a bayesian factor model. *Mol. Biol. Evol.*, 31(9) :2483–2495.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.*, 29 :51–63.

Endler, L., Betancourt, A. J., et Nolte, V. (2016). Reconciling differences in Pool-GWAS between populations : a case study of female abdominal Pigmentation in *Drosophila melanogaster*. *Genetics*, 202 :843–855.

Excoffier, L. (2007). Analysis of population subdivision. In Balding, D. J., Bishop, M., et Cannings, C., editors, *Handbook of Statistical Genetics*, pages 980–1020, Chichester. John Wiley & Sons, Ltd.

Falush, D., Stephens, M., et Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data : Linked loci and correlated allele frequencies. *Genetics*, 164(4) :1567–1587.

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., et al. (2017). Accounting for Linkage Disequilibrium in genome scans for selection without individual genotypes : the local score approach. *Mol. Ecol.*, 26 :3700–3714.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., et Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3) :929–941.

Ferretti, L., Ramos-Onsins, S. E., et Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Mol. Ecol.*, 22(22) :5561–5576.

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., Holderegger, R., et Widmer, A. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, 18 :69.

Fleiss, J. L. et Cuzick, J. (1979). The reliability of dichotomous judgements : unequal numbers of judges per subject. *Appl. Psychol. Meas.*, 3 :537–542.

Fleming, D. S., Koltes, J. E., Rothschild, M. F., Schmidt, C. J., Ashwell, C. M., Persia, M. E., Reecy, J. M., et Lamont, S. J. (2016). Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genomics*, 17 :1–11.

Foll, M. et Gaggiotti, O. (2008). A genome-Scan method to identify selected loci appropriate for both dominant and codominant markers : a bayesian perspective. *Genetics*, 180(2) :977–993.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., et Excoffier, L. (2014). Widespread signals of convergent adaptation to high altitude in Asia and america. *Am. J. Hum. Genet.*, 95(4) :394–407.

Fontaine, A., Filipovi, I., Fansiri, T., Hoffmann, A. A., Cheng, C., Kirkpatrick, M., Ra, G., et Lambrechts, L. (2017). Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, *Aedes aegypti*. *Genome Biol. Evol.*, 9 :2322–2335.

Franchini, P., Xiong, P., Fruciano, C., et Meyer, A. (2016). The Role of microRNAs in the repeated parallel diversification of lineages of midas cichlid fish from Nicaragua. *Genome Biol. Evol.*, 8(5) :1543–1555.

Franssen, S. U., Barton, N. H., et Schlotterer, C. (2017). Reconstruction of haplotype-blocks selected during experimental evolution. *Mol. Biol. Evol.*, 34(1) :174–184.

Fruciano, C., Franchini, P., Kovacova, V., Elmer, K. R., Henning, F., et Meyer, A. (2016). Genetic linkage of distinct adaptive traits in sympatrically speciating crater lake cichlid fish. *Nat. Commun.*, 7 :1–8.

Fu, Y., Li, C., Tang, Q., Tian, S., Jin, L., Chen, J., Li, M., et Li, C. (2016). Genomic analysis reveals selection in Chinese native black pig. *Nature Publishing Group*, 6 :1–9.

Fuentes-Pardo, A. P.and Ruzzente, D. E. (2017). Whole-genome sequencing approaches for conservation biology : Advantages, limitations and practical recommendations. *Mol. Ecol.*, 26 :5369–5406.

Fusco, G. et Minelli, A. (2010). Phenotypic plasticity in development and evolution : facts and concepts. *Philos. Trans. R. Soc. B-Biol. Sci.*, 365(1540) :547–556.

Fustier, M.-A., Brandenburg, J.-T., Boitard, S., Lapeyronnie, J., Eguiarte, L. E., Vigouroux, Y., Manicacci, D., et Tenaillon, M. I. (2017). Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. *Mol. Ecol.*, 26(10) :2738–2756.

Futschik, A. et Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, 186 :207–218.

Gaggiotti, O. E. et Foll, M. (2010). Quantifying population structure using the F-model. *Mol. Ecol. Res.*, 10(5) :821–830.

Gammerdinger, W. J., Conte, M. A., Acquah, E. A., Roberts, R. B., et Kocher, T. D. (2014). Structure and decay of a proto-y region in tilapia, *Oreochromis niloticus*. *BMC Genomics*, 15(1) :975.

Gammerdinger, W. J., Conte, M. A., Baroiller, J.-f., Cotta, H. D., et Kocher, T. D. (2016). Comparative analysis of a sex chromosome from the blackchin tilapia, *Sarotherodon melanotheron*. *BMC Genomics*, 17 :1–10.

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

Gautier, M., Gharbi, K., Cezaerd, T., Galan, M., Loiseau, A., et al. (2013). Estimation of population allele frequencies from next-generation sequencing data : pool-versus individual-based genotyping. *Mol. Ecol.*, 22 :3766–3779.

Gautier, M., Hocking, T. D., et Foulley, J. L. (2010). A bayesian outlier criterion to detect snps under selection in large data sets. *PLOS ONE*, 5(8) :e11913.

Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., et al. (2018). The genomic basis of colour pattern polymorphism in the harlequin ladybird. *Curr. Biol.*, 28(1) :1–7.

Gelman, A., Carlin, J. B., Stern, H. S., et Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., et al. (2003). The International HapMap Project. *Nature*, 426 :789–796.

Gilks, W. R., Richardson, S., et Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Res.*, 11 :759–769.

Goudet, J. (1993). *The genetics of geographically structured populations*. PhD thesis, University of Wales, Bangor.

Gould, B. A., Chen, Y., et Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Mol. Ecol.*, 26(1) :163–177.

Grande, F. D., Sharma, R., Meiser, A., Rolshausen, G., Büdel, B., et al. (2017). Adaptive differentiation coincides with local bioclimatic conditions along an elevational cline in populations of a lichen-forming fungus. *BMC Evo. Biol.*, 17(1) :93.

Grau, J., Grosse, I., et Keilwagen, J. (2015). Prroc : computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, 31(15) :2595–2597.

Graves, Jr, J., Hertweck, K., Phillips, M., Han, M., Cabral, L., et al. (2017). Genomics of parallel experimental evolution in *Drosophila*. *Mol. Biol. Evol.*, 34(4) :831–842.

Griffin, P. C., Hangartner, S. B., Fournier-Level, A., et Hoffmann, A. A. (2017). Genomic trajectories to desiccation resistance : convergence and divergence among replicate selected *Drosophila* lines. *Genetics*, 205(2) :871–890.

Guo, B., Lu, D., Liao, W. B., et Merilä, J. (2016). Genomewide scan for adaptive differentiation along altitudinal gradient in the andrew's toad *Bufo andrewsi*. *Mol. Ecol.*, 25(16) :3884–3900.

Guo, F., Dey, D. K., et Holsinger, K. E. (2009). A bayesian hierarchical model for analysis of single-Nucleotide polymorphisms diversity in multilocus, multipopulation samples. *J. Am. Stat. Assoc.*, 104(485) :142–154.

Günther, T., Lampei, C., Barilar, I., et Schmid, K. J. (2016). Genomic and phenotypic differentiation of *Arabidopsis thaliana* along altitudinal gradients in the north italian alps. *Mol. Ecol.*, 25(15) :3574–3592.

Haasl, R. J. et Payseur, B. A. (2016). Fifteen years of genomewide scans for selection : trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.*, 25 :5–23.

Hardy, C. M., Burke, M. K., Everett, L. J., Han, M. V., Lantz, K. M., et Gibbs, A. G. (2018). Genome-wide analysis of starvation-selected *Drosophila melanogaster*—a genetic model of obesity. *Mol. Biol. Evol.*, 35(1) :50–65.

Harris, H. (1966). Enzyme polymorphisms in man. *Proc. R. Soc. Lond. B Biol. Sci.*, 164(995) :298–310.

Hendrick, M. F. et Mathiasson, M. E. (2016). The genetics of extreme microgeographic adaptation : an integrated approach identifies a major gene underlying leaf trichome divergence in Yellowstone *Mimulus guttatus*. *Mol. Ecol.*, 25(22) :5647–5662.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., et Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, 210(1) :315–330.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., et al. (2016). Finding the genomic basis of local adaptation : pitfalls, practical solutions, and future directions. *Am. Nat.*, 188(4) :379–397.

Hofer, T., Ray, N., Wegmann, D., et Excoffier, L. (2009). Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann. Hum. Genet.*, 73(1) :95–108.

Holsinger, K. S. et Weir, B. S. (2009). Genetics in geographically structured populations : defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.*, 10 :639–650.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18 :337–338.

Hudson, R. R., Slatkin, M., et Maddison, W. P. (1992). Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2) :583–589.

Innan, H. et Kim, Y. (2008). Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*, 179(3) :1713–1720.

Ishida, Y. (2009). Sewall Wright and Gustave Malécot on isolation by distance. *Philos. Sci.*, 76 :784–796.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, London/New York/Oxford, 3 edition.

Jensen, J. D. (2014). On the unfounded enthusiasm for soft selective sweeps. *Nat. Commun.*, 5 :1–10.

Kadri, S. M., Harpur, B. A., Orsi, R. O., et Zayed, A. (2016). A variant reference data set for the Africanized honeybee, *Apis mellifera*. 3 :1–6.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., et al. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540 :69–73.

Kang, L., Aggarwal, D. D., Rashkovetsky, E., Korol, A. B., et Michalak, P. (2016a). Rapid genomic changes in *Drosophila melanogaster* adapting to desiccation stress in an experimental evolution system. *BMC Genomics*, 17(1) :233.

Kang, L., Garner, H. R., Price, D. K., et Michalak, P. (2017). A test for gene flow among sympatric and allopatric hawaiian picture-winged *Drosophila*. *Journal of Molecular Evolution*, 84(5) :259–266.

Kang, L., Settlage, R., McMahon, W., Michalak, K., Tae, H., et al. (2016b). Genomic signatures of speciation in sympatric and allopatric hawaiian picture-winged *Drosophila*. *Genome Biology and Evolution*, 8(5) :1482–1488.

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H. C., Zody, M. C., et al. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.*, 39 :1321–1328.

Kawecki, T. J. et Ebert, D. (2004). Conceptual issues in local adaptation. *Ecol. Lett.*, 7 :1225–1241.

Keilwagen, J., Grosse, I., et Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PLOS ONE*, 9(3) :e92209.

Kessner, D., Turner, T. L., et Novembre, J. (2013). Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Mol. Biol. Evol.*, 30(5) :1145–1158.

Kimura, M. et Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49 :725–738.

Kirkpatrick, M. et Barton, N. (2006). Chromosome Inversions, local adaptation and speciation. *Genetics*, 173 :419–434.

Kjærner-Semb, E., Ayllon, F., Furmanek, T., Wennevik, V., Dahle, G., et al. (2016). Atlantic salmon populations reveal adaptive divergence of immune related genes - a duplicated genome under selection. *BMC Genomics*, 17(1) :610.

Kofler, R., Pandey, R. V., et Schlötterer, C. (2011). PoPoolation2 : identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27 :3435–3436.

Konczal, M., Koteja, P., Orlowska-Feuer, P., Radwan, J., Sadowska, E. T., et Babik, W. (2016). Genomic response to selection for predatory behavior in a mammalian model of adaptive radiation. *Mol. Biol. Evol.*, 33(9) :2429–2440.

Kozak, G. M., Wadsworth, C. B., Kahne, S. C., Bogdanowicz, S. M., Harrison, R. G., Coates, B. S., et Dopman, E. B. (2017). A combination of sexual and ecological divergence contributes to rearrangement spread during initial stages of speciation. *Mol. Ecol.*, 26(8) :2331–2347.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, 304(5925) :412–417.

Lai, F.-n., Zhai, H.-l., Cheng, M., Ma, J.-y., Cheng, S.-f., et al. (2016). Whole-genome scanning for the litter size trait associated genes and SNPs under selection in dairy goat (*Capra hircus*). *Sci. Rep.*, 6 :38096.

Lander, E. S. et al., E. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.

Lawson, M. J. et Zhang, L. (2006). Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol.*, 7(2) :1–11.

Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., et al. (2018). Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*. *Mol. Ecol.*, 27 :264–278.

Leviyang, S. et Hamilton, M. B. (2011). Properties of weir and cockerham's F_{ST} estimators and associated bootstrap confidence intervals. *Theor. Popul. Biol.*, 79 :39–52.

Lewontin, R. C. (1985). Population Genetics. *Annu. Rev. Genet.*, 19 :81–102.

Lewontin, R. C. et Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54(2) :595–609.

Lewontin, R. C. et Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics*, 74 :175–195.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 :2078–2079.

Li, H. L., Gu, X. H., Li, B. J., Chen, C. H., Lin, H. R., et Xia, J. H. (2017). Genome-wide qtl analysis identified significant associations between hypoxia tolerance and mutations in the gpr132 and abcg4 genes in nile tilapia. *Mar. Biotechnol.*, 19(5) :441–453.

Long, Q., Jeffares, D. C., Zhang, Q., Ye, K., Nizhynska, V., Ning, Z., Tyler-Smith, C., et Nordborg, M. (2011). PoolHap : Inferring haplotype frequencies from pooled samples by next generation sequencing. *PLoS ONE*, 6(1) :e15292.

Love, R. R., Dame, N., Life, G., Dame, N., Dame, N., Centre, T., et Pharmacy, M. (2016). Chromosomal inversions and ecotypic differentiation in *Anopheles gambiae* : the perspective from whole-genome sequencing. *Mol. Ecol.*, 25(23) :5889–5906.

Lynch, M., Bost, D., Wilson, S., Maruki, T., et Harrison, S. (2014). Population-genetic inference from pooled-sequencing data. *Genome Biol. Evol.*, 6 :1210–1218.

Machado, H. E., Bergland, A. O., Emily, L., Schmidt, P. S., et Petrov, D. A. (2016). HHS Public Access. *Mol. Ecol.*, 25(3) :723–740.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 37 :344–352.

Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson, Paris.

Malécot, G. (1967). Identical loci and relationship. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 4 :317–332.

Maxam, a. M. et Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, 74(2) :560–4.

Maynard Smith, J. et Haigh, J. (1974). The hitch-hikink effect of a favorable gene. *Genet. Res.*, 23 :23–35.

Metzgar, D., Bytof, J., et Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.*, 10(1) :72–80.

Montano, V., Didelot, X., Foll, M., Linz, B., Reinhardt, R., Suerbaum, S., Moodley, Y., et Jensen, J. D. (2015). Worldwide population structure, long-Term demography, and local adaptation of *Helicobacter pylori*. *Genetics*, 200 :947–963.

Myles, S., Davison, D., Barrett, J., Stoneking, M., et Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Med. Genomics*, 1(1) :22.

Neethiraj, R., Hornett, E. A., Hill, J. A., et Wheat, C. W. (2017). Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach : New insights into the genetics underlying colour variation in diverse taxa. *Mol. Ecol.*, 26(19) :4990–5002.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA*, 70 :3321–3323.

Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.*, 41 :225–233.

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution*, 40 :643–645.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press.

Nei, M. et Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.*, 47 :253–259.

Nei, M. et Maruyama, T. (1975). Lewontin-Krakauer test for neutral genes. *Genetics*, 80 :395.

Nicholson, G., Smith, A. V., et Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Series B Stat. Methodol.*, 64(4) :695–715.

Oppold, A.-M., Schmidt, H., Rose, M., Hellmann, S. L., Dolze, F., et al. (2017). *Chironomus riparius* (diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence. *Mol. Ecol.*, 26(12) :3256–3275.

Orgogozo, V., Peluffo, A. E., et Morizot, B. (2016). The "mendelian gene" and the "molecular gene" : Two relevant concepts of genetic units. In Orgogozo, V., editor, *Genes and Evolution*, volume 119 of *Current Topics in Developmental Biology*, pages 1–26. Academic Press.

Peng, B. et Kimmel, M. (2005). simuPOP : A forward-time population genetics simulation environment. *Bioinformatics*, 21(18) :3686–3687.

Peng, F. et Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *Can. J. Stat.*, 23(2) :199–213.

Pennings, P. S. et Hermisson, J. (2006). Soft Sweeps III : The signature of positive selection from recurrent mutation. *PLOS Genet.*, 2(12) :e186.

Phillips, M. A., Long, A. D., Greenspan, Z. S., Greer, L. F., Burke, M. K., et al. (2016). Genome-wide analysis of long-term evolutionary domestication in *Drosophila melanogaster*. *Nature Publishing Group*, 6 :1–12.

Pickrell, J. K. et Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.*, 8(11) :e1002967.

Pritchard, J. K., Pickrell, J. K., et Coop, G. (2010). The genetics of human adaptation : hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.*, 20(4) :208–215.

R Core Team (2017). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. (2013). *Stationary Allele Frequency Distributions*. American Cancer Society.

Rellstab, C., Zoller, S., Lesur, I., et Pluess, A. R. (2016). Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions. *Mol. Ecol*, 25(23) :5907–5924.

Reynolds, J., Weir, B. S., et Cockerham, C. C. (1983). Estimation of the coancestry coefficient : basis for a short-term genetic distance. *Genetics*, 105 :767–779.

Ridout, M. S., Demktrio, C. G. B., et Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, 55 :137–148.

Riebler, A., Held, L., et Stephan, W. (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178(3) :1817–1829.

Robertson, A. (1975). Remarks on the Lewontin-Krakauer test. *Genetics*, 80 :396.

Rode, N. O., Holtz, Y., Loridon, K., Santoni, S., Ronfort, J., et Gay, J. (2017). How to optimize the precision of allele and haplotype frequency estimates using pooled-sequencing data. *Mol. Ecol. Res.*, 18(2) :194–203.

Rohfritsch, A., Galan, M., Gautier, M., Gharbi, K., Olsson, G., et al. (2018). Preliminary insights into the genetics of bank vole tolerance to Puumala hantavirus in Sweden. *Ecol. Evol.* Accepted.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., et Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, 14(5) :R51.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142 :1357–1362.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145 :1219–1228.

Rousset, F. (2007). Inferences from spatial population genetics. In Balding, D. J., Bishop, M., et Cannings, C., editors, *Handbook of Statistical Genetics*, pages 945–979, Chichester. John Wiley & Sons, Ltd.

Rousset, F. (2008). genepop'007 : a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Res.*, 8 :103–106.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419 :832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449 :913–919.

Sangwan, N., Zarraonaindia, I., Hampton-Marcell, J. T., Ssegane, H., Eshoo, T. W., Rijal, G., Negri, M. C., et Gilbert, J. A. (2016). Differential functional constraints cause strain-level endemism in polynucleobacter populations. *mSystems*, 1(3).

Savolainen, O., Lascoux, M., et Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.*, 14(11) :807–820.

Scheet, P. et Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data : applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78(4) :629–644.

Schlötterer, C., Tobler, R., Kofler, R., et Nolte, V. (2014). Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.*, 15 :749–763.

Schluter, D. (2001). Ecology and the origin of species. *Ecol. Lett.*, 16(7) :372–380.

Schweizer, R. M., VonHoldt, B. M., Harrigan, R., Knowles, J. C., Musiani, M., Coltman, D., Novembre, J., et Wayne, R. K. (2016). Genetic subdivision and candidate genes under selection in North American grey wolves. *Mol. Ecol*, 25(1) :380–402.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, 47 :264–279.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J., et Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. *Evolution*, 66 :2723–2738.

Song, S., Yao, N., Yang, M., Liu, X., Dong, K., et al. (2016). Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the tibetan cashmere goat (*Capra hircus*). *BMC Genomics*, 17(1) :122.

Stephan, W., Song, Y. S., et Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4) :2647–2663.

Storz, J. F. (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.*, 14(3) :671–688.

Sved, J. A. et Hill, W. G. (2018). One hundred years of Linkage Disequilibrium. *Genetics*, 209 :629–636.

Tang, K., Thornton, K. R., et Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLOS Biology*, 5(7) :e171.

Toomey, M. B., Lopes, R. J., Johnson, J. D., Afonso, S., Mota, P. G., et al. (2017). High-density lipoprotein receptor SCARB1 is required for carotenoid coloration in birds. *PNAS*, 114(20) :5219–5224.

Tyagi, A., Yadav, A., Tripathi, A. M., et Roy, S. (2016). High light intensity plays a major role in emergence of population level variation in *Arabidopsis thaliana* along an altitudinal gradient. *Sci. Rep.*, 6 :26160.

Utsunomiya, Y. T., Milanesi, M., Utsunomiya, A. T. H., et Ajmone-marsan, P. (2016). GHap : an R package for genome-wide haplotyping. *Bioinformatics*, 32 :2861–2862.

Utsunomiya, Y. T., Taiti, A., Utsunomiya, H., Torrecilha, B. P., Kim, E.-s., et al. (2017). Mutation contributed to stature recovery in modern cattle. *Sci. Rep.*, 7 :17140.

Vitalis, R. (2012). DetSel : An R-Package to detect marker loci responding to selection. In Pompanon, F. et Bonin, A., editors, *Data Production and Analysis in Population Genomics : Methods and Protocols*, volume 888 of *Methods in Molecular Biology*, pages 277–293, New York. Humana Press.

Vitalis, R., Dawson, K., et Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics*, 158 :1811–1823.

Vitalis, R., Gautier, M., Dawson, K. J., et Beaumont, M. a. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, 196 :799–817.

Voight, B. F., Kudaravalli, S., Wen, X., et Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLOS Biology*, 4(3) :e72.

Wang, Q., Pi, J., Pan, A., Shen, J., et Qu, L. (2017). A novel sex-linked mutant affecting tail formation in Hongshan chicken. *Sci. Rep.*, 7 :10079.

Wang, X., Liu, J., Zhou, G., Guo, J., Yan, H., et al. (2016). Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Sci. Rep.*, 8 :38932.

Weir, B. S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Inc., Sunderland, MA.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., et Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Res.*, 15 :1468–1476.

Weir, B. S. et Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38 :1358–1370.

Weir, B. S. et Goudet, J. (2017). An unified characterization of population structure and relatedness. *Genetics*, 206 :2085–2103.

Weir, B. S. et Hill, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.*, 36 :721–750.

Whitlock, M. C. et Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation : Inference of a null model through trimming the distribution of f_{st}. *Am. Nat.*, 186(S1) :S24–S36. PMID : 26656214.

Whitlock, M. C. et McCauley, D. E. (1999). Indirect measures of gene flow and migration : F_{ST} not equal to 1/(4Nm + 1). *Heredity*, 82 (Pt 2) :117–125.

Wragg, D., Marti-marimon, M., Basso, B., Bidanel, J.-p., Labarthe, E., Bouchez, O., Conte, Y. L., et Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Sci. Rep.*, 6 :27168.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16 :97–159.

Wright, S. (1940). Breeding structure of populations in relation to speciation. *Am. Nat.*, 74(752) :232–248.

Wright, S. (1943). Isolation by distance. *Genetics*, 28 :114–138.

Wright, S. (1949). *Adaptation and selection*. Genenetics, Paleontology, and Evolution, Princeton, university edition.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*, 15 :323–354.

Yassin, A., Debat, V., Gidaszewski, N., David, J. R., et Pool, J. E. (2016). Recurrent specialization on a toxic fruit in an island *Drosophila* population. *PNAS*, 113(17) :4771–4776.

Zhao, L. et Begun, D. J. (2017). Genomics of parallel adaptation at two timescales in *Drosophila*. *PLOS Genet.*, 13(10) :e1007016.

Deuxième Partie

Annexes

Annexe A

Article 1 : Measuring Genetic Differentiation from Pool-seq Data

Valentin Hivert, Raphaël Leblois, Eric J. Petit, Mathieu Gautier and Renaud Vitalis

Measuring Genetic Differentiation from Pool-seq Data

Valentin Hivert,*,† Raphaël Leblois,*,† Eric J. Petit,‡ Mathieu Gautier,*,†,1 and Renaud Vitalis*,†,1,2

*CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, 34988 Montferrier-sur-Lez Cedex, France, †Institut de Biologie Computationnelle, Univ Montpellier, 34095 Montpellier Cedex, France, and ‡ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, 35042 Rennes, Cedex, France

ORCID IDs: 0000-0002-5144-6956 (V.H.); 0000-0002-3051-4497 (R.L.); 0000-0001-5058-5826 (E.J.P.); 0000-0001-7257-5880 (M.G.); 0000-0001-7096-3089 (R.V.)

ABSTRACT The advent of high throughput sequencing and genotyping technologies enables the comparison of patterns of polymorphisms at a very large number of markers. While the characterization of genetic structure from individual sequencing data remains expensive for many nonmodel species, it has been shown that sequencing pools of individual DNAs (Pool-seq) represents an attractive and cost-effective alternative. However, analyzing sequence read counts from a DNA pool instead of individual genotypes raises statistical challenges in deriving correct estimates of genetic differentiation. In this article, we provide a method-of-moments estimator of F_{ST} for Pool-seq data, based on an analysis-of-variance framework. We show, by means of simulations, that this new estimator is unbiased and outperforms previously proposed estimators. We evaluate the robustness of our estimator to model misspecification, such as sequencing errors and uneven contributions of individual DNAs to the pools. Finally, by reanalyzing published Pool-seq data of different ecotypes of the prickly sculpin *Cottus asper*, we show how the use of an unbiased F_{ST} estimator may question the interpretation of population structure inferred from previous analyses.

KEYWORDS F_{ST}; genetic differentiation; pool sequencing; population genomics

IT has long been recognized that the subdivision of species into subpopulations, social groups, and families fosters genetic differentiation (Wahlund 1928; Wright 1931). Characterizing genetic differentiation as a means to infer unknown population structure is therefore fundamental to population genetics and finds applications in multiple domains, including conservation biology, invasion biology, association mapping, and forensics, among many others. In the late 1940s and early 1950s, Malécot (1948) and Wright (1951) introduced *F*-statistics to partition genetic variation within and between groups of individuals (Holsinger and Weir 2009; Bhatia *et al.* 2013). Since then, the estimation of *F*-statistics has become standard practice (see, *e.g.*, Weir 1996, 2012; Weir and Hill 2002) and the most commonly used estimators of F_{ST} have been developed in an analysis-of-variance framework (Cockerham 1969, 1973; Weir and Cockerham 1984), which can be recast in terms of probabilities of identity of pairs of homologous genes (Cockerham and Weir 1987; Rousset 2007; Weir and Goudet 2017).

Assuming that molecular markers are neutral, estimates of F_{ST} are typically used to quantify genetic structure in natural populations, which is then interpreted as the result of demographic history (Holsinger and Weir 2009): large F_{ST} values are expected for small populations among which dispersal is limited (Wright 1951), or between populations that have long diverged in isolation from each other (Reynolds *et al.* 1983). When dispersal is spatially restricted, a positive relationship between F_{ST} and the geographical distance for pairs of populations generally holds (Slatkin 1993; Rousset 1997). It has also been proposed to characterize the heterogeneity of F_{ST} estimates across markers for identifying loci that are targeted by selection (Cavalli-Sforza 1966; Lewontin and Krakauer 1973; Beaumont and Nichols 1996; Vitalis *et al.* 2001; Akey *et al.* 2002; Beaumont 2005; Weir *et al.* 2005; Lotterhos and Whitlock 2014, 2015; Whitlock and Lotterhos 2015).

Next-generation sequencing (NGS) technologies provide unprecedented amounts of polymorphism data in both model

doi: https://doi.org/10.1534/genetics.118.300900
Manuscript received March 9, 2018; accepted for publication July 21, 2018; published Early Online July 25, 2018.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6856781.
[1]These authors are joint senior authors on this work.
[2]Corresponding author: Centre de Biologie pour la Gestion des Populations, Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez Cedex, France. E-mail: renaud.vitalis@inra.fr

and nonmodel species (Ellegren 2014). Although the sequencing strategy initially involved individually tagged samples in humans (The International HapMap Consortium 2005), whole-genome sequencing of pools of individuals (Pool-seq) is being increasingly used for population genomic studies (Schlötterer *et al.* 2014). Because it consists of sequencing libraries of pooled DNA samples and does not require individual tagging of sequences, Pool-seq provides genome-wide polymorphism data at considerably lower cost than sequencing of individuals (Schlötterer *et al.* 2014). However, non-equimolar amounts of DNA from all individuals in a pool and stochastic variation in the amplification efficiency of individual DNAs have raised concerns with respect to the accuracy of the so-obtained allele frequency estimates, particularly at low sequencing depth and with small pool sizes (Cutler and Jensen 2010; Anderson *et al.* 2014; Ellegren 2014). Nonetheless, it has been shown that, at equal sequencing efforts, Pool-seq provides similar, if not more accurate, allele frequency estimates than individual-based analyses (Futschik and Schlötterer 2010; Gautier *et al.* 2013). The problem is different for diversity and differentiation parameters, which depend on second moments of allele frequencies or, equivalently, on pairwise measures of genetic identity: with Pool-seq data, it is indeed impossible to distinguish pairs of reads that are identical because they were sequenced from a single gene from pairs of reads that are identical because they were sequenced from two distinct genes that are identical in state (IIS) (Ferretti *et al.* 2013).

Appropriate estimators of diversity and differentiation parameters must therefore be sought to account for both the sampling of individual genes from the pool and the sampling of reads from these genes. There has been several attempts to define estimators for the parameter F_{ST} for Pool-seq data (Kofler *et al.* 2011; Ferretti *et al.* 2013), from ratios of heterozygosities (or from probabilities of genetic identity between pairs of reads) within and between pools. In the following, we will argue that these estimators are biased (*i.e.*, they do not converge toward the expected value of the parameter) and that some of them have undesired statistical properties (*i.e.*, the bias depends on sample size and coverage). Here, following Cockerham (1969, 1973), Weir and Cockerham (1984), Weir (1996), Weir and Hill (2002), and Rousset (2007), we define a method-of-moments estimator of the parameter F_{ST} using an analysis-of-variance framework. We then evaluate the accuracy and precision of this estimator, based on the analysis of simulated data sets, and compare it to estimates defined in the software package PoPoolation2 (Kofler *et al.* 2011) and in Ferretti *et al.* (2013). Furthermore, we test the robustness of our estimators to model misspecifications (including unequal contributions of individuals in pools and sequencing errors). Finally, we reanalyze the prickly sculpin (*Cottus asper*) Pool-seq data (published by Dennenmoser *et al.* 2017), and show how the use of biased F_{ST} estimators in previous analyses may challenge the interpretation of population structure.

Note that throughout this article, we use the term "gene" to designate a segregating genetic unit (in the sense of the "Mendelian gene" from Orgogozo *et al.* 2016). We further use the term "read" in a narrow sense, as a sequenced copy of a gene. For the sake of simplicity, we will use the term "Ind-seq" to refer to analyses based on individual data, for which we further assume that individual genotypes are called without error.

Model

F-statistics may be described as intraclass correlations for the IIS probability of pairs of genes (Cockerham and Weir 1987; Rousset 1996, 2007). F_{ST} is best defined as:

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2}, \tag{1}$$

where Q_1 is the IIS probability for genes sampled within subpopulations, and Q_2 is the IIS probability for genes sampled between subpopulations. In the following, we develop an estimator of F_{ST} for Pool-seq data by decomposing the total variance of read frequencies in an analysis-of-variance framework. A complete derivation of the model is provided in the Supplemental Material, File S1.

For the sake of clarity, the notation used throughout this article is given in Table 1. We first derive our model for a single locus and eventually provide a multilocus estimator of F_{ST}. Consider a sample of n_d subpopulations, each of which is made of n_i genes ($i = 1, \ldots, n_d$) sequenced in pools (hence n_i is the haploid sample size of the ith pool). We define c_{ij} as the number of reads sequenced from gene j ($j = 1, \ldots, n_i$) in subpopulation i at the locus considered. Note that c_{ij} is a latent variable that cannot be directly observed from the data. Let $X_{ijr:k}$ be an indicator variable for read r ($r = 1, \ldots, c_{ij}$) from gene j in subpopulation i, such that $X_{ijr:k} = 1$ if the rth read from the jth gene in the ith deme is of type k, and $X_{ijr:k} = 0$ otherwise. In the following, we use standard dot notation for sample averages, *i.e.*: $X_{ij\cdot:k} \equiv \sum_r X_{ijr:k}/c_{ij}$, $X_{i\cdots:k} \equiv \sum_j \sum_r X_{ijr:k} / \sum_j c_{ij}$, and $X_{\cdots:k} \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij}$. The analysis-of-variance is based on the computation of sums of squares, as follows:

$$\begin{aligned}\sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{ijr:k} - X_{\cdots:k}\right)^2 &= \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{ijr:k} - X_{ij\cdot:k}\right)^2 \\ &+ \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{ij\cdot:k} - X_{i\cdots:k}\right)^2 \\ &+ \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{i\cdot\cdot:k} - X_{\cdots:k}\right)^2 \\ &\equiv SSR_{:k} + SSI_{:k} + SSP_{:k}.\end{aligned} \tag{2}$$

Table 1 Summary of main notations used

Notation	Parameter definition
$X_{ijr:k}$	Indicator variable: $X_{ijr:k} = 1$ if the rth read from the jth individual in the ith pool is of type k, and $X_{ijr:k} = 0$ otherwise
$r_{i:k} = \sum_j \sum_r X_{ijr:k}$	Number of reads of type k in the ith pool
c_{ij}	Number of reads sequenced from individual j in subpopulation i (unobserved individual coverage)
$C_{1i} \equiv \sum_j c_{ij}$	Total number of reads in the ith pool (pool coverage)
$C_1 \equiv \sum_i C_{1i}$	Total number of reads in the full sample (total coverage)
$C_2 \equiv \sum_i C_{1i}^2$	Squared number of reads in the full sample
n_i	Total number of genes the ith pool (haploid pool size)
$y_{i:k}$	(Unobserved) number of genes of type k in the ith pool
$\pi_k \equiv \mathbb{E}(X_{ijr:k})$	Expected frequency of reads of type k in the full sample
$\tilde{\pi}_{ij:k} \equiv X_{ij:k}$	(Unobserved) average frequency of reads of type k for individual j in the ith pool
$\hat{\pi}_{i:k} \equiv X_{i\cdots k}$	Average frequency of reads of type k in the ith pool
$\hat{\pi}_k \equiv X_{\cdots k}$	Average frequency of reads of type k in the full sample
Q_1 (respectively Q_2)	IIS probability for two genes sampled within (respectively between) pools
Q_1^r (respectively Q_2^r)	IIS probability for two reads sampled within (respectively between) pools
$\hat{Q}_1^{\text{pool}}$ (respectively $\hat{Q}_2^{\text{pool}}$)	Unbiased estimator of the IIS probability for genes sampled within (respectively between) pools

As is shown in File S1, the expected sums of squares depend on the expectation of the allele frequency π_k over all replicate populations sharing the same evolutionary history, as well as on the IIS probability $Q_{1:k}$ that two genes in the same pool are both of type k, and the IIS probability $Q_{2:k}$ that two genes from different pools are both of type k. Taking expectations (see the detailed computations in File S1), one has:

$$\mathbb{E}(SSR_{:k}) = 0 \tag{3}$$

for reads within individual genes, since we assume that there is no sequencing error, *i.e.*, all the reads sequenced from a single gene are identical and $X_{ijr:k} = X_{ij:k}$ for all r. For reads between genes within pools, we get:

$$\mathbb{E}(SSI_{:k}) = (C_1 - D_2)(\pi_k - Q_{1:k}), \tag{4}$$

where $C_1 \equiv \sum_i \sum_j c_{ij} = \sum_i C_{1i}$ is the total number of reads in the full sample (total coverage), C_{1i} is the coverage of the ith pool, and $D_2 \equiv \sum_i (C_{1i} + n_i - 1)/n_i$. D_2 arises from the assumption that the distribution of the read counts c_{ij} is multinomial (*i.e.*, that all genes contribute equally to the pool of reads; see Equation A15 in File S1). For reads between genes from different pools, we have:

$$\mathbb{E}(SSP_{:k}) = \left(C_1 - \frac{C_2}{C_1}\right)(Q_{1:k} - Q_{2:k}) + (D_2 - D_2^\star)(\pi_k - Q_{1:k}), \tag{5}$$

where $C_2 \equiv \sum_i C_{1i}^2$ and $D_2^\star \equiv \left[\sum_i C_{1i}(C_{1i} + n_i - 1)/n_i\right] \Big/ C_1$ (see Equation A16 in File S1). Rearranging Equation 4 and Equation 5 and summing over alleles, we get:

$$Q_1 - Q_2 = \frac{(C_1 - D_2)\mathbb{E}(SSP) - (D_2 - D_2^\star)\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \tag{6}$$

and

$$1 - Q_2 = \frac{(C_1 - D_2)\mathbb{E}(SSP) + (n_c - 1)(D_2 - D_2^\star)\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)}, \tag{7}$$

where $n_c \equiv (C_1 - C_2/C_1)/(D_2 - D_2^\star)$. Let $MSI \equiv SSI/(C_1 - D_2)$ and $MSP \equiv SSP/(D_2 - D_2^\star)$. Then, using the definition of F_{ST} from Equation 1, we have:

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{\mathbb{E}(MSP) - \mathbb{E}(MSI)}{\mathbb{E}(MSP) + (n_c - 1)\mathbb{E}(MSI)}, \tag{8}$$

which yields the method-of-moments estimator

$$\hat{F}_{ST}^{\text{pool}} = \frac{MSP - MSI}{MSP + (n_c - 1)MSI}, \tag{9}$$

where

$$MSI = \frac{1}{C_1 - D_2} \sum_k \sum_i^{n_d} C_{1i} \hat{\pi}_{i:k}(1 - \hat{\pi}_{i:k}) \tag{10}$$

and

$$MSP = \frac{1}{D_2 - D_2^\star} \sum_k \sum_i^{n_d} C_{1i}(\hat{\pi}_{i:k} - \hat{\pi}_k)^2 \tag{11}$$

(see Equations A25 and A26 in File S1). In Equation 10 and Equation 11, $\hat{\pi}_{i:k} \equiv X_{i\cdots k}$ is the average frequency of reads of type k within the ith pool, and $\hat{\pi}_k \equiv X_{\cdots k}$ is the average frequency of reads of type k in the full sample. Note that from the definition of $X_{\cdots k}$, $\hat{\pi}_k \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij} = \sum_i C_{1i} \hat{\pi}_{i:k} / \sum_i C_{1i}$ is the weighted average of the sample frequencies with weights equal to the pool coverage. This is equivalent to the weighted analysis-of-variance in Cockerham (1973) (see also Weir and Cockerham 1984; Weir 1996; Weir and Hill 2002; Rousset 2007; Weir and

Goudet 2017). Finally, the full expression of $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ in terms of sample frequencies develops as:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\sum_k \left[(C_1 - D_2) \sum_i^{n_\mathrm{d}} C_{1i} (\hat{\pi}_{i:k} - \hat{\pi}_k)^2 - (D_2 - D_2^\star) \sum_i^{n_\mathrm{d}} C_{1i} \hat{\pi}_{i:k} (1 - \hat{\pi}_{i:k}) \right]}{\sum_k \left[(C_1 - D_2) \sum_i^{n_\mathrm{d}} C_{1i} (\hat{\pi}_{i:k} - \hat{\pi}_k)^2 + (n_\mathrm{c} - 1)(D_2 - D_2^\star) \sum_i^{n_\mathrm{d}} C_{1i} \hat{\pi}_{i:k} (1 - \hat{\pi}_{i:k}) \right]}.$$

If we take the limit case where each gene is sequenced exactly once, we recover the Ind-seq model: assuming $c_{ij} = 1$ for all (i,j), then $C_1 = \sum_i^{n_\mathrm{d}} n_i$, $C_2 = \sum_i^{n_\mathrm{d}} n_i^2$, $D_2 = n_\mathrm{d}$, and $D_2^\star = 1$. Therefore, $n_\mathrm{c} = (C_1 - C_2/C_1)/(n_\mathrm{d} - 1)$, and Equation 9 reduces exactly to the estimator of F_{ST} for haploids: see Weir (1996), p. 182, and Rousset (2007), p. 977.

As in Reynolds *et al.* (1983), Weir and Cockerham (1984), Weir (1996), and Rousset (2007), a multilocus estimate is derived as the sum of locus-specific numerators over the sum of locus-specific denominators:

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_l MSP_l - MSI_l}{\sum_l MSP_l + (n_\mathrm{c} - 1) MSI_l}, \tag{13}$$

where *MSI* and *MSP* are subscripted with l to denote the lth locus. For Ind-seq data, Bhatia *et al.* (2013) refer to this multilocus estimate as a "ratio of averages" as opposed to an "average of ratios," which would consist of averaging single-locus F_{ST} over loci. This approach is justified in the appendix of Weir and Cockerham (1984) and in Bhatia *et al.* (2013), who analyzed both estimates by means of coalescent simulations. Note that Equation 13 assumes that the pool size is equal across loci. Also note that the construction of the estimator in Equation 13 is different from Weir and Cockerham's (1984). These authors defined their multilocus estimator as a ratio of sums of components of variance (a, b, and c in their notation) over loci, which give the same weight to all loci whatever the number of sampled genes at each locus. Equation 13 follows GENEPOP's rationale (Rousset 2008) instead, which gives more weight to loci that are more intensively covered.

Materials and Methods

Simulation study

Generating individual genotypes: We first generated individual genotypes using ms (Hudson 2002), assuming an island model of population structure (Wright 1931). For each simulated scenario, we considered eight demes, each made of $N = 5000$ haploid individuals. The migration rate (m) was fixed to achieve the desired value of F_{ST} (0.05 or 0.2), using equation 6 in Rousset (1996) leading to, *e.g.*, $M \equiv 2Nm = 16.569$ for $F_{\mathrm{ST}} = 0.05$ and $M = 3.489$ for $F_{\mathrm{ST}} = 0.20$. The mutation rate was set at $\mu = 10^{-6}$, giving $\theta \equiv 2N\mu = 0.01$. We considered either fixed or variable sample sizes across demes. In the latter case, the haploid sample size n was drawn independently for each deme from a Gaussian distribution with mean 100 and SD 30; this number was rounded up to the nearest integer, with a minimum of 20 and maximum of 300 haploids per deme. We generated a very large number of sequences for each scenario and sampled independent single nucleotide polymorphisms (SNPs) from sequences with a single segregating site. Each scenario was replicated 50 times (500 times for Figure 3 and Figure S2).

Pool sequencing: For each ms simulated data set, we generated Pool-seq data by drawing reads from a binomial distribution (Gautier *et al.* 2013). More precisely, we assume that for each SNP, the number $r_{i:k}$ of reads of allelic type k in pool i follows:

$$r_{i:k} \sim \mathrm{Bin}\left(\frac{y_{i:k}}{n_i}, \delta_i \right), \tag{14}$$

where $y_{i:k}$ is the number of genes of type k in the ith pool, n_i is the total number of genes in pool i (haploid pool size), and δ_i is the simulated total coverage for pool i. In the following, we either consider a fixed coverage, with $\delta_i = \Delta$ for all pools and loci, or a varying coverage across pools and loci, with $\delta_i \sim \mathrm{Pois}(\Delta)$.

Sequencing error: We simulated sequencing errors occurring at rate $\mu_\mathrm{e} = 0.001$, which is typical of Illumina sequencers (Glenn 2011; Ross *et al.* 2013). We assumed that each sequencing error modifies the allelic type of a read to one of three other possible states with equal probability (there are therefore four allelic types in total, corresponding to four nucleotides). Note that only biallelic markers are retained in the final data sets. Also note that, since we initiated this procedure with polymorphic markers only, we neglect sequencing errors that would create spurious SNPs from monomorphic sites. However, such SNPs should be rare in real data sets, since markers with a low minimum read count (MRC) are generally filtered out.

Experimental error: Nonequimolar amounts of DNA from all individuals in a pool and stochastic variation in the amplification efficiency of individual DNAs are sources of experimental errors in Pool-seq. To simulate experimental errors, we used the model derived by Gautier *et al.* (2013). In this model, it is assumed that the contribution $\eta_{ij} = c_{ij}/C_{1i}$ of each gene j

to the total coverage of the ith pool (C_{1i}) follows a Dirichlet distribution:

$$\{\eta_{ij}\}_{1\le j\le n_i} \sim \mathrm{Dir}\left(\frac{\rho}{n_i}\right), \tag{15}$$

where the parameter ρ controls the dispersion of gene contributions around the value $\eta_{ij} = 1/n_i$, which is expected if all genes contributed equally to the pool of reads. For convenience, we define the experimental error ϵ as the coefficient of variation of η_{ij}, *i.e.*, $\epsilon \equiv \sqrt{\mathbb{V}(\eta_{ij})}\Big/\mathbb{E}(\eta_{ij}) = \sqrt{(n_i-1)/(\rho+1)}$ (see Gautier *et al.* 2013). When ϵ tends toward 0 (or equivalently, when ρ tends to infinity), all individuals contribute equally to the pool and there is no experimental error. We tested the robustness of our estimates to values of ϵ between 0.05 and 0.5. The case $\epsilon = 0.5$ could correspond, for example, to a situation where (for $n_i = 10$) five individuals contribute 2.8× more reads than the other five individuals.

Other estimators

For the sake of clarity, a summary of the notation of the F_{ST} estimators used throughout this article is given in Table 2.

PP2$_d$: This estimator of F_{ST} is implemented by default in the software package PoPoolation2 (Kofler *et al.* 2011). It is based on a definition of the parameter F_{ST} as the overall reduction in average heterozygosity relative to the total combined population (see, *e.g.*, Nei and Chesser 1983):

$$\mathrm{PP2_d} \equiv \frac{\hat{H}_T - \hat{H}_S}{\hat{H}_T}, \tag{16}$$

where $\hat{H}_S$ is the average heterozygosity within subpopulations, and $\hat{H}_T$ is the average heterozygosity in the total population (obtained by pooling together all subpopulations to form a single virtual unit). In PoPoolation2, $\hat{H}_S$ is the unweighted average of within-subpopulation heterozygosities:

$$\hat{H}_S = \frac{1}{n_d}\sum_i^{n_d}\left(\frac{n_i}{n_i-1}\right)\left(\frac{C_{1i}}{C_{1i}-1}\right)\left(1-\sum\nolimits_k \hat{\pi}^2_{i:k}\right) \tag{17}$$

(using the notation from Table 1). Note that in PoPoolation2, $\mathrm{PP2_d}$ is restricted to the case of two subpopulations only ($n_d = 2$). The two ratios in the right-hand side of Equation 17 are presumably borrowed from Nei (1978) to provide an unbiased estimate, although we found no formal justification for the expression in Equation 17 for Pool-seq data. The total heterozygosity is computed as (using the notation from Table 1):

$$\hat{H}_T = \left(\frac{\min_i(n_i)}{\min_i(n_i)-1}\right)\left(\frac{\min_i(C_{1i})}{\min_i(C_{1i})-1}\right)\left(1-\sum_k \hat{\pi}^2_k\right). \tag{18}$$

Table 2 Definition of the F_{ST} estimators used in the text

Notation	Definition
$\hat{F}_{ST}^{pool}$	Equation 12
$\mathrm{FRP_{13}}$	Ferretti *et al.* (2013) and Equation 16, Equation 20, and Equation 21
$\mathrm{NC_{83}}$	Nei and Chesser (1983)
$\mathrm{PP2_d}$	Kofler *et al.* (2011) and Equation 16, Equation 17, and Equation 18
$\mathrm{PP2_a}$	Kofler *et al.* (2011) and Equation 19
$\mathrm{WC_{84}}$	Weir and Cockerham (1984)

PP2$_a$: This is the alternative estimator of F_{ST} provided in the software package PoPoolation2. It is based on an interpretation by Kofler *et al.* (2011) of Karlsson *et al.*'s (2007) estimator of F_{ST}, as:

$$\mathrm{PP2_a} \equiv \frac{\hat{Q}_1^r - \hat{Q}_2^r}{1-\hat{Q}_2^r}, \tag{19}$$

where $\hat{Q}_1^r$ and $\hat{Q}_2^r$ are the frequencies of identical pairs of reads within and between pools, respectively, computed by simple counting of IIS pairs. These are estimates of Q_1^r, the IIS probability for two reads in the same pool (whether they are sequenced from the same gene or not), and Q_2^r, the IIS probability for two reads in different pools. Note that the IIS probability Q_1^r is different from Q_1 in Equation 1, which, from our definition, represents the IIS probability between distinct genes in the same pool. This approach therefore confounds pairs of reads within pools that are identical because they were sequenced from a single gene from pairs of reads that are identical because they were sequenced from distinct, yet IIS genes.

FRP$_{13}$: This estimator of F_{ST} was developed by Ferretti *et al.* (2013) (see their equations 3, 10, 11, 12, and 13). Ferretti *et al.* (2013) use the same definition of F_{ST} as in Equation 16 above, although they estimate heterozygosities within and between pools as "average pairwise nucleotide diversities," which, from their definitions, are formally equivalent to IIS probabilities. In particular, they estimate the average heterozygosity within pools as (using the notation from Table 1):

$$\hat{H}_S = \frac{1}{n_d}\sum_i^{n_d}\left(\frac{n_i}{n_i-1}\right)\left(1-\hat{Q}_{1i}^r\right) \tag{20}$$

and the total heterozygosity among the n_d populations as:

$$\hat{H}_T = \frac{1}{n_d^2}\left[\sum_i^{n_d}\left(\frac{n_i}{n_i-1}\right)\left(1-\hat{Q}_{1i}^r\right) + \sum_{i\neq i'}^{n_d}\left(1-\hat{Q}_{2ii'}^r\right)\right]. \tag{21}$$

Analyses of Ind-seq data

For the comparison of Ind-seq and Pool-seq data sets, we computed F_{ST} on subsamples of 5000 loci. These subsamples were defined so that only those loci that were polymorphic in all coverage conditions were retained, and the same loci were

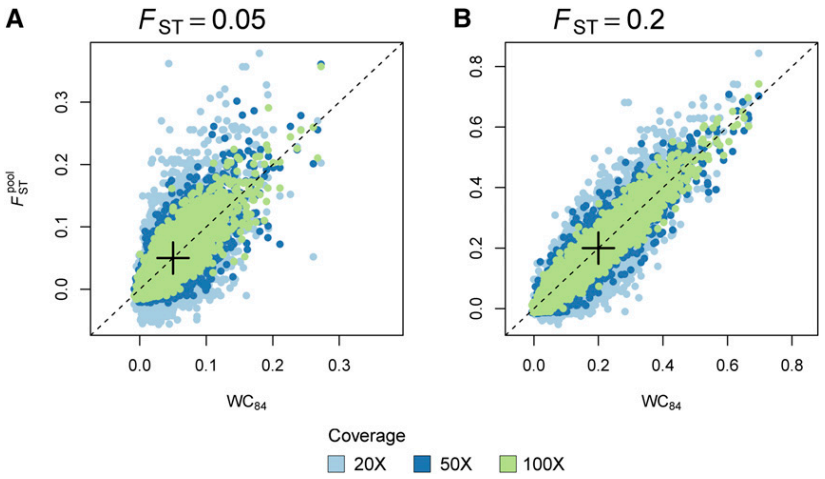


Figure 1 Single-locus estimates of F_{ST}. We compared single-locus estimates of F_{ST} based on allele count data inferred from individual genotypes (Ind-seq), using the WC_{84} estimator, to $\hat{F}_{ST}^{pool}$ estimates from Pool-seq data. We simulated 5000 SNPs using ms in an island model with $n_d = 8$ demes. We used two migration rates corresponding to (A) $F_{ST} = 0.05$ and (B) $F_{ST} = 0.20$. The size of each pool was fixed to 100. We show the results for different coverages (20×, 50×, and 100×). In each graph, the cross indicates the simulated value of F_{ST}.

used for the analysis of the corresponding Ind-seq data. For the latter, we used either the Nei and Chesser's (1983) estimator based on a ratio of heterozygosity (see Equation 16 above), hereafter denoted by NC_{83}, or the analysis-of-variance estimator developed by Weir and Cockerham (1984), hereafter denoted by WC_{84}.

All the estimators were computed using custom functions in the R software environment for statistical computing, version 3.3.1 (R Core Team 2017). All of these functions were carefully checked against available software packages to ensure that they provided strictly identical estimates.

Application example: C. asper

Dennenmoser *et al.* (2017) investigated the genomic basis of adaption to osmotic conditions in the prickly sculpin (*C. asper*), an abundant euryhaline fish in northwestern North America. To do so, they sequenced the whole genome of pools of individuals from two estuarine populations (Capilano River Estuary, CR; Fraser River Estuary, FE) and two freshwater populations (Pitt Lake, PI; Hatzic Lake, HZ) in southern British Columbia (Canada). We downloaded the four corresponding BAM files from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.2qg01) and combined them into a single mpileup file using SAMtools version 0.1.19 (Li *et al.* 2009) with default options, except the maximum depth per BAM that was set to 5000 reads. The resulting file was further processed using a custom awk script to call SNPs and compute read counts, after discarding bases with a base alignment quality (BAQ) score <25. A position was then considered a SNP if: (1) only two different nucleotides with a read count >1 were observed (nucleotides with ≤1 read being considered as a sequencing error); (2) the coverage was between 10 and 300 in each of the four alignment files; (3) the minor allele frequency, as computed from read counts, was ≥0.01 in the four populations. The final data set consisted of 608,879 SNPs.

Our aim here was to compare the population structure inferred from pairwise estimates of F_{ST} using the estimator $\hat{F}_{ST}^{pool}$ (Equation 12) with that of $PP2_d$. To determine which of the two estimators performs better, we then compared the population structure inferred from $\hat{F}_{ST}^{pool}$ and $PP2_d$ to that inferred from the Bayesian hierarchical model implemented in the software package BayPass (Gautier 2015). BayPass allows the robust estimation of the scaled covariance matrix of allele frequencies across populations for Pool-seq data, which is known to be informative about population history (Pickrell and Pritchard 2012). The elements of the estimated matrix can be interpreted as pairwise and population-specific estimates of differentiation (Coop *et al.* 2010) and therefore provide a comprehensive description of population structure that makes full use of the available data.

Data availability

An R package called poolfstat, which implements F_{ST} estimates for Pool-seq data, is available at the Comprehensive R Archive Network (CRAN): https://cran.r-project.org/web/packages/poolfstat/index.html.

The authors state that all data necessary for confirming the conclusions presented in this article are fully represented within the article, figures, and tables. Supplemental material (including Figures S1–S4, Tables S1–S3, and a complete derivation of the model in File S1) available at Figshare: https://doi.org/10.25386/genetics.6856781.

Results

Comparing Ind-seq and Pool-seq estimates of F_{ST}

Single-locus estimates of $\hat{F}_{ST}^{pool}$ are highly correlated with the classical estimates of WC_{84} (Weir and Cockerham 1984) computed on the individual data that were used to generate the pools in our simulations (see Figure 1). The variance of $\hat{F}_{ST}^{pool}$ across independent replicates decreases as the coverage increases. The correlation between $\hat{F}_{ST}^{pool}$ and WC_{84} is stronger for multilocus estimates (see Figure S1A).

Comparing Pool-seq estimators of F_{ST}

We found that our estimator $\hat{F}_{ST}^{pool}$ has extremely low bias (<0.5% over all scenarios tested: see Table 3 and Tables S1–S3). In other words, the average estimates across multiple

Table 3 Overall F_{ST} estimates from multiple pools

		Pool-seq		Ind-seq
F_{ST}	n	Coverage	$\hat{F}_{ST}^{pool}$	WC_{84}
0.05	10	20×	0.050 (0.002)	
0.05	10	50×	0.051 (0.002)	0.050 (0.002)
0.05	10	100×	0.050 (0.002)	
0.05	100	20×	0.050 (0.001)	
0.05	100	50×	0.050 (0.001)	0.051 (0.001)
0.05	100	100×	0.050 (0.001)	
0.20	10	20×	0.200 (0.002)	
0.20	10	50×	0.201 (0.002)	0.201 (0.002)
0.20	10	100×	0.201 (0.002)	
0.20	100	20×	0.201 (0.003)	
0.20	100	50×	0.202 (0.003)	0.203 (0.003)
0.20	100	100×	0.203 (0.003)	

Multilocus $\hat{F}_{ST}^{pool}$ estimates were computed for various conditions of expected F_{ST}, pool size (n), and coverage in an island model with $n_d = 8$ subpopulations (pools). The mean (RMSE) is over 50 independent simulated data sets, each made of 5000 loci. For comparison, we computed multilocus WC_{84} estimates from individual genotypes (Ind-seq).

loci and replicates closely equal the expected value of the F_{ST} parameter, as given by equation 6 in Rousset (1996), which is based on the computation of IIS probabilities in an island model of population structure. In all the situations examined, the bias does not depend on the sample size (*i.e.*, the size of each pool) or on the coverage (see Figure 2). Only the variance of the estimator across independent replicates decreases as the sample size increases and/or as the coverage increases. At high coverage, the mean and root mean squared error (RMSE) of $\hat{F}_{ST}^{pool}$ over independent replicates are virtually indistinguishable from that of the WC_{84} estimator (see Table S1).

Figure 3 shows the RMSE of F_{ST} estimates for a wide range of pool sizes and coverages. The RMSE decreases as the pool size and/or the coverage increases. The F_{ST} estimates are more precise and accurate when differentiation is low. Figure 3 provides some clues to evaluate the pool size and the coverage that is necessary to achieve the same RMSE as for Ind-seq data. Consider, for example, the case of samples of $n = 20$ haploids. For $F_{ST} \leq 0.05$ (in the conditions of our simulations), the RMSE of F_{ST} estimates based on Pool-seq data tends to the RMSE of F_{ST} estimates based on Ind-seq data either by sequencing pools of ~200 haploids at 20×, or by sequencing pools of 20 haploids at ~200×. However, the same precision and accuracy are achieved by sequencing ~50 haploids at ~50×.

Conversely, we found that $PP2_d$ (the default estimator of F_{ST} implemented in the software package PoPoolation2) is biased when compared to the expected value of the parameter. We observed that the bias depends on both the sample size and the coverage (see Figure 2). We note that, as the coverage and the sample size increase, $PP2_d$ converges to the estimator NC_{83} (Nei and Chesser 1983) computed from individual data (see Figure S1B). This argument was used by Kofler *et al.* (2011) to validate their approach, even though the estimates of $PP2_d$ depart from the true value of the parameter (Figure S1, B and C).

The second of the two estimators of F_{ST} implemented in PoPoolation2, which we refer to as $PP2_a$, is also biased (see Figure 2). We note that the bias decreases as the sample size increases. However, the bias does not depend on the coverage (only the variance over independent replicates depends on coverage). The estimator developed by Ferretti *et al.* (2013), which we refer to as FRP_{13}, is also biased (see Figure 2). However, the bias does not depend on the pool size or on the coverage (only the variance over independent replicates depends on coverage). FRP_{13} converges to the estimator NC_{83}, computed from individual data (see Figure 2). At high coverage, the mean and RMSE over independent replicates are virtually indistinguishable from that of the NC_{83} estimator.

Lastly, we stress that our estimator $\hat{F}_{ST}^{pool}$ provides estimates for multiple populations and is therefore not restricted to pairwise analyses, contrary to PoPoolation2's estimators. We show that, even at low sample size and low coverage, Pool-seq estimates of differentiation are virtually indistinguishable from classical estimates for Ind-seq data (see Table 3).

Robustness to unbalanced pool sizes and variable sequencing coverage

We evaluated the accuracy and the precision of the estimator $\hat{F}_{ST}^{pool}$ when sample sizes differ across pools and when the coverage varies across pools and loci (see Figure 4). We found that, at low coverage, unequal sampling or variable coverage causes a negligible departure from the median of WC_{84} estimates computed on individual data, which vanishes as the coverage increases. At 100× coverage, the distribution of $\hat{F}_{ST}^{pool}$ estimates is almost indistinguishable from that of WC_{84} (see Figure 4 and Tables S2 and S3).

Robustness to sequencing and experimental errors

Figure 5 shows that sequencing errors cause a negligible negative bias for $\hat{F}_{ST}^{pool}$ estimates. Filtering (using an MRC of 4) improves estimation slightly, but only at high coverage (Figure 6B). It must be noted, however, that filtering increases the bias in the absence of sequencing error, especially at low coverage (Figure 6A). With experimental error, *i.e.*, when individuals do not contribute evenly to the final set of reads, we observed a positive bias for $\hat{F}_{ST}^{pool}$ estimates (Figure 5). We note that the bias decreases as the size of the pools increases. Figure S2 shows the RMSE of F_{ST} estimates for a wider range of pool sizes, coverage, and experimental error rate (ϵ). For $\epsilon \geq 0.25$, increasing the coverage cannot improve the quality of the inference if the pool size is too small. When Pool-seq experiments are prone to large experimental error rates, increasing the size of pools is the only way to improve the estimation of F_{ST}. Filtering (using an MRC of 4) does not improve estimation (Figure 6C).

Application example

The reanalysis of the prickly sculpin data revealed larger pairwise estimates of multilocus F_{ST} using the $PP2_d$ estimator,

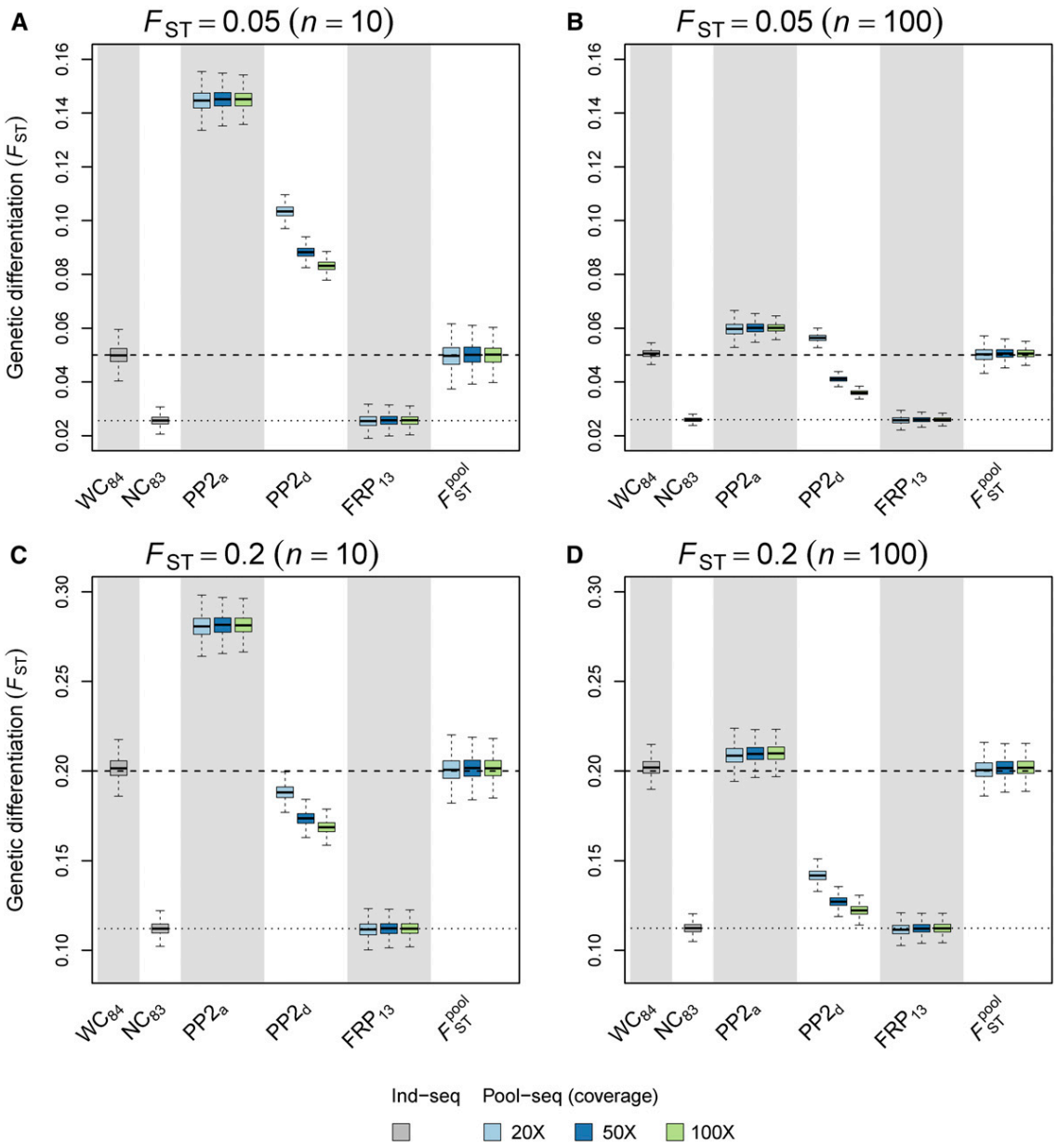


Figure 2 Precision and accuracy of pairwise estimators of F_{ST}. We considered two estimators based on allele count data inferred from individual genotypes (Ind-seq): WC_{84} and NC_{83}. For Pool-seq data, we computed the two estimators implemented in the software package PoPoolation2, which we refer to as $PP2_d$ and $PP2_a$, as well as the FRP_{13} estimator and our estimator $\hat{F}_{ST}^{pool}$. Each boxplot represents the distribution of multilocus F_{ST} estimates across all pairwise comparisons in an island model with $n_d = 8$ demes and across 50 independent replicates of the ms simulations. We used two migration rates, corresponding to (A and B) $F_{ST} = 0.05$ and (C and D) $F_{ST} = 0.20$. The size of each pool was either fixed to (A and C) 10 or to (B and D) 100. For Pool-seq data, we show the results for different coverages (20×, 50×, and 100×). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of NC_{83} estimates.

as compared to $\hat{F}_{ST}^{pool}$ (see Figure 7A). Furthermore, we found that $\hat{F}_{ST}^{pool}$ estimates are smaller for within-ecotype pairwise comparisons as compared to between-ecotype comparisons. Therefore, the inferred relationships between samples based on pairwise $\hat{F}_{ST}^{pool}$ estimates show a clear-cut structure, separating the two estuarine samples from the freshwater ones (see Figure 7C). We did not recover the same structure using $PP2_d$ estimates (see Figure 7B). Additionally, the scaled covariance matrix of allele frequencies across samples is consistent with the structure inferred from $\hat{F}_{ST}^{pool}$ estimates (see Figure 7D).

Discussion

Whole-genome sequencing of pools of individuals is increasingly popular for population genomic research on both model and nonmodel species (Schlötterer *et al.* 2014). The development of dedicated software packages (reviewed in

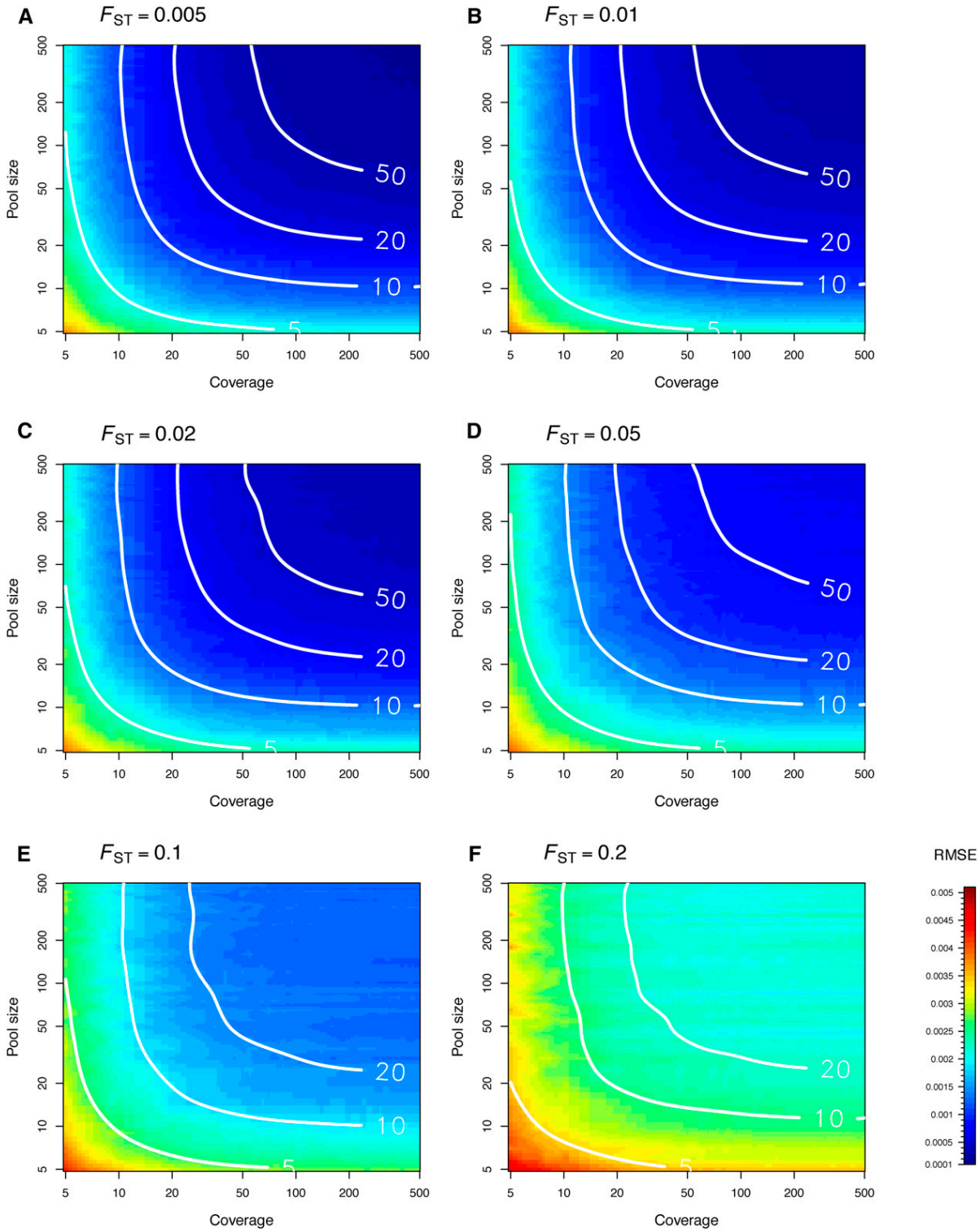


Figure 3 (A–F) Precision and accuracy of our estimator $\hat{F}_{ST}^{pool}$ as a function of pool size and coverage for simulated F_{ST} values ranging from 0.005 to 0.2. Each density plot, which represents the RMSE of the estimator $\hat{F}_{ST}^{pool}$, was obtained using simple linear interpolation from a set of 44×44 pairs of pool size and coverage values. For each pool size and coverage, 500 replicates of 5000 markers were simulated from an island model with $n_d = 8$ demes. White isolines represent the RMSE of the WC_{84} estimator computed from Ind-seq data for various sample sizes (n = 5, 10, 20, and 50). Each isoline was fitted using a thin plate spline regression with smoothing parameter $\lambda = 0.005$, implemented in the fields package for R (Nychka *et al.* 2017).

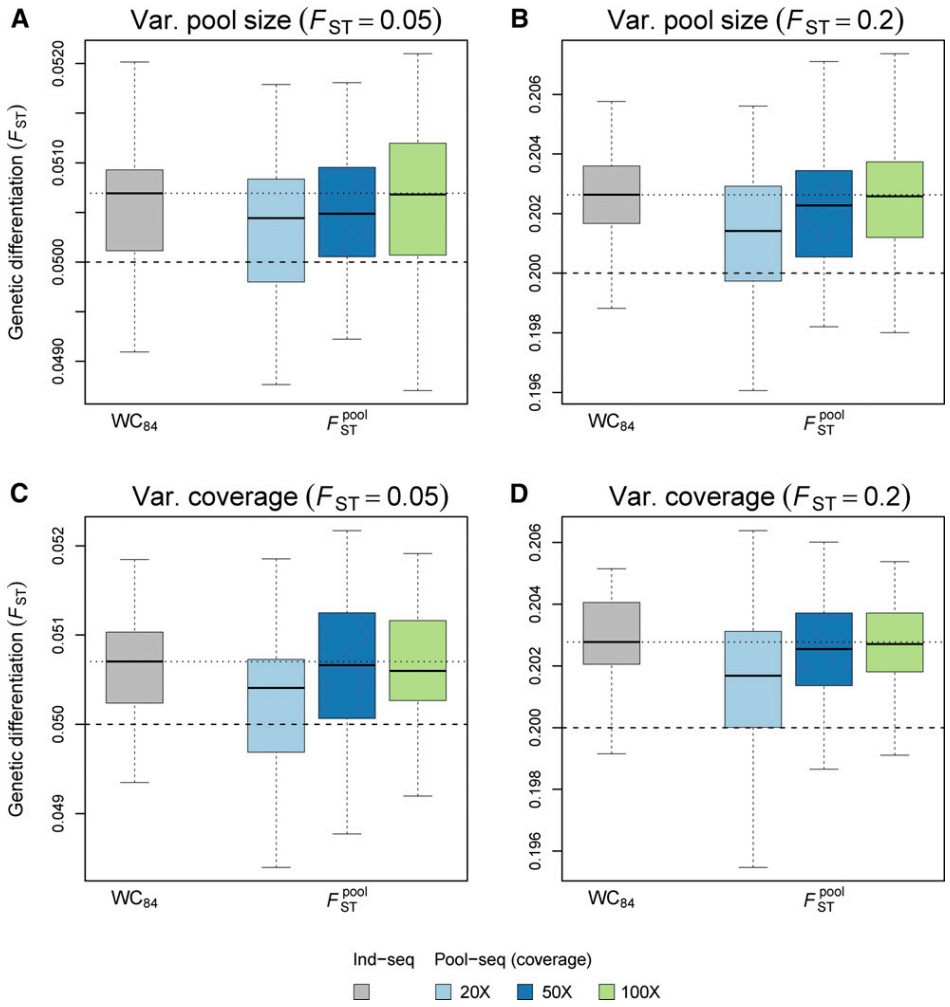


Figure 4 Precision and accuracy of F_{ST} estimates with varying pool size or varying coverage. Our estimator $\hat{F}_{ST}^{pool}$ was calculated from Pool-seq data over all demes and loci and compared to the estimator WC_{84}, computed from Ind-seq data. Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the ms simulations. We used two migration rates, corresponding to (A and C) $F_{ST} = 0.05$ and (B and D) $F_{ST} = 0.20$. (A and B) The pool size was variable across demes, with haploid sample size n drawn independently for each deme from a Gaussian distribution with mean 100 and SD 30; n was rounded up to the nearest integer, with a minimum of 20 and a maximum of 300 haploids per deme. (C and D) The pool size was fixed ($n = 100$) and the coverage (δ_i) was varying across demes and loci, with $\delta_i \sim \text{Pois}(\Delta)$ where $\Delta \in \{20, 50, 100\}$. For Pool-seq data, we show the results for different coverages (20×, 50×, and 100×). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of WC_{84} estimates. Var., variable.

Schlötterer *et al.* 2014) undoubtedly has something to do with the breadth of research questions that have been tackled using Pool-seq. However, the analysis of population structure from Pool-seq data are complicated by the double sampling process of genes from the pool and sequence reads from those genes (Ferretti *et al.* 2013).

The naive approach that consists of computing F_{ST} from read counts as if they were allele counts (*e.g.*, as in Chen *et al.* 2016) ignores the extra variance brought by the random sampling of reads from the gene pool during Pool-seq experiments. Furthermore, such computation fails to consider the actual number of lineages in the pool (haploid pool size). Altogether, these limits may result in severely biased estimates of differentiation when the pool size is low (see Figure S3). A possible alternative is to compute F_{ST} from allele counts imputed from read counts using a maximum-likelihood approach conditional on the haploid size of the pools (*e.g.*, as in Smadja *et al.* 2012; Leblois *et al.* 2018), or from allele frequencies estimated using a model-based method which accounts for the sampling effects and the sequencing error probabilities inherent to pooled NGS experiments (see Fariello *et al.* 2017). However, these latter approaches may only be accurate in situations where the coverage is much larger than pool size, allowing for a reduction of the sampling variance of reads (see Figure S3). We therefore developed a new estimator of the parameter F_{ST} for Pool-seq data in an analysis-of-variance framework (Cockerham 1969, 1973). The accuracy of this estimator is barely distinguishable from that of the Weir and Cockerham's (1984) estimator for individual data. Furthermore, it does not depend on the pool size or on the coverage, and it is robust to unequal pool sizes and varying coverage across demes and loci.

In our analysis, the frequency of reads within pools is a weighted average of the sample frequencies, with weights equal to the pool coverage. Therefore, our approach follows Cockerham's (1973) one, which he referred to as a weighted analysis-of-variance (see also Weir and Cockerham 1984; Weir 1996; Weir and Hill 2002; Weir and Goudet 2017). With unequal pool sizes, weighted and unweighted analyses differ. As discussed recently in Weir and Goudet (2017), the unweighted approach seems appropriate when the between component exceeds the within component, *i.e.*, when F_{ST} is large (Tukey 1957). It turns out that optimal weighting depends upon the parameter to be estimated (Cockerham 1973) and is only efficient at lower levels of differentiation (Robertson 1962). In a likelihood analysis of the island

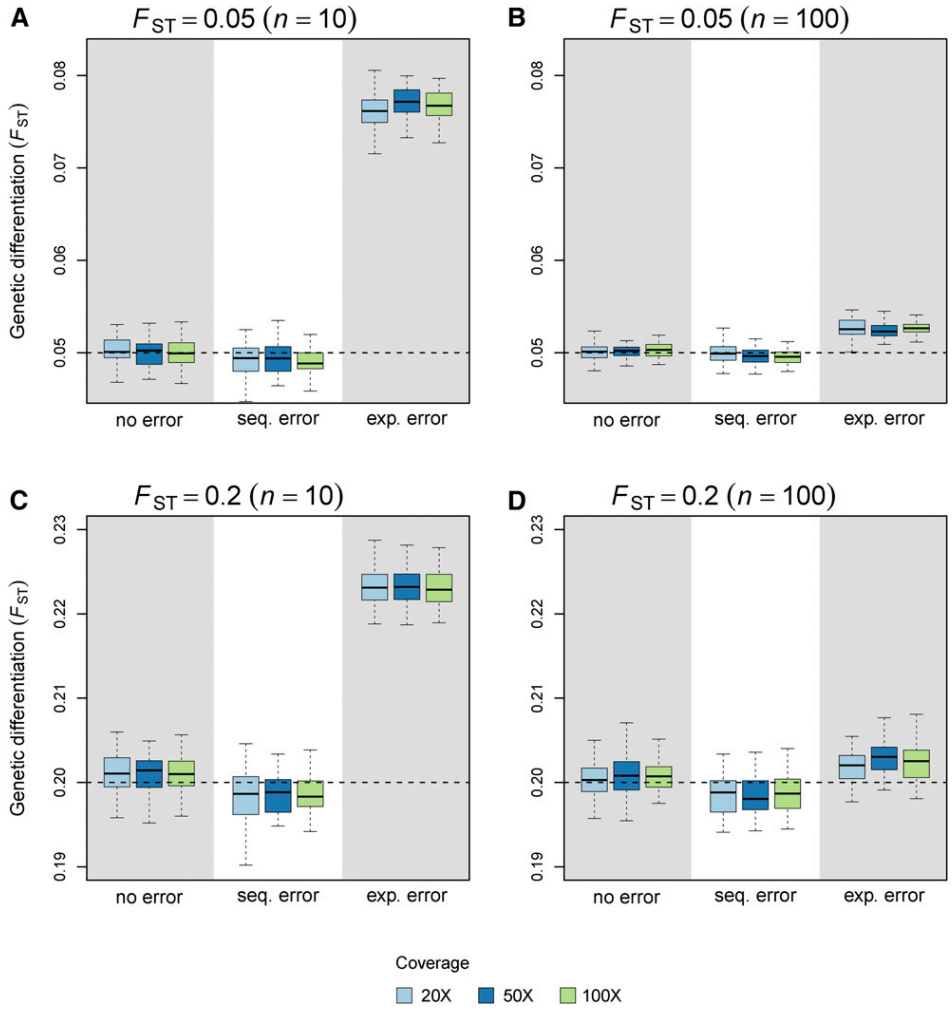


Figure 5 Precision and accuracy of F_{ST} estimates with sequencing and experimental errors. Our estimator $\hat{F}_{ST}^{pool}$ was computed from Pool-seq data over all demes and loci without error, with sequencing error (occurring at rate $\mu_e = 0.001$), and with experimental error ($\epsilon = 0.5$). Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the ms simulations. We used two migration rates, corresponding to (A and B) $F_{ST} = 0.05$ or (C and D) $F_{ST} = 0.20$. The size of each pool was either fixed to (A and C) 10 or to (B and D) 100. For Pool-seq data, we show the results for different coverages (20×, 50×, and 100×). In each graph, the dashed line indicates the simulated value of F_{ST}. Exp., experimental; Seq., sequencing.

model, Rousset (2007) derived asymptotically efficient weights that are proportional to n_i^2 for the sum of squares of different samples (see also Robertson 1962). To the best of our knowledge, such optimal weighting has never been considered in the literature.

Analysis-of-variance and probabilities of identity

In the analysis-of-variance framework, F_{ST} is defined in Equation 1 as an intraclass correlation for the probability of IIS (Cockerham and Weir 1987; Rousset 1996). Extensive statistical literature is available on estimators of intraclass correlations. Beside analysis-of-variance estimators, introduced in population genetics by Cockerham (1969, 1973), estimators based on the computation of probabilities of identical response within and between groups have been proposed (see, *e.g.*, Fleiss 1971; Fleiss and Cuzick 1979; Mak 1988; Ridout *et al.* 1999; Wu *et al.* 2012), which were originally referred to as kappa-type statistics (Fleiss 1971; Landis and Koch 1977). These estimators have later been endorsed in population genetics, where the "probability of identical response" was then interpreted as the frequency with which the genes are alike (Cockerham 1973; Cockerham and Weir 1987; Weir 1996; Rousset 2007; Weir and Goudet 2017).

This suggests that, with Pool-seq data, another strategy could consist of computing F_{ST} from IIS probabilities between (unobserved) pairs of genes, which requires that unbiased estimates of such quantities are derived from read count data. We have done this in the second section of File S1 and we provide alternative estimators of F_{ST} for Pool-seq data (see Equations A44 and A48 in File S1). These estimators (denoted by $\hat{F}_{ST}^{pool-PID}$ and $\tilde{F}_{ST}^{pool-PID}$) have exactly the same form as the analysis-of-variance estimator if the pools all have the same size and if the number of reads per pool is constant (Equation A33 in File S1). This echoes the derivations by Rousset (2007) for Ind-seq data, who showed that the analysis-of-variance approach (Weir and Cockerham 1984) and the simple strategy of estimating IIS probabilities by counting identical pairs of genes provide identical estimates when sample sizes are equal (see Equation A28 in File S1 and also Cockerham and Weir 1987; Weir 1996; Karlsson *et al.* 2007). With unbalanced samples, we found that analysis-of-variance estimates have better precision and accuracy than IIS-based estimates, particularly for low levels of differentiation (see

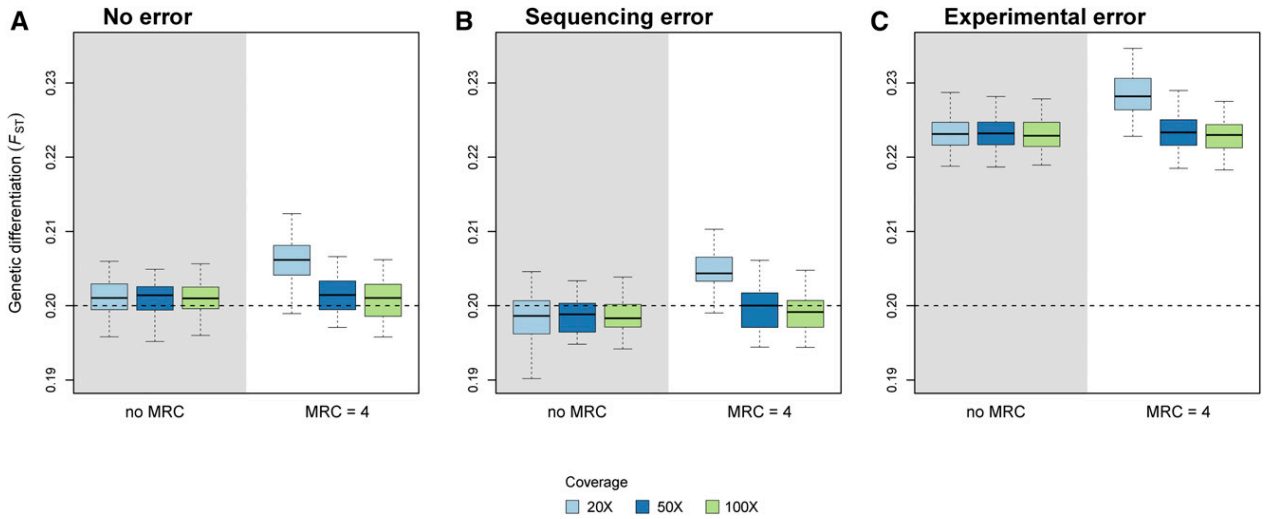


Figure 6 Precision and accuracy of F_{ST} estimates with and without filtering. Our estimator $\hat{F}_{ST}^{pool}$ was computed from Pool-seq data over all demes and loci (A) without error, (B) with sequencing error, and (C) with experimental error (see the legend of Figure 5 for further details). For each case, we computed F_{ST} without filtering (no MRC) and with filtering (using a MRC = 4). Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the ms simulations. We used a migration rate corresponding to $F_{ST} = 0.20$ and pool size $n = 10$. We show the results for different coverages (20×, 50×, and 100×). In each graph, the dashed line indicates the simulated value of F_{ST}.

Figure S4). Interestingly, we found that IIS-based estimates of F_{ST} for Pool-seq data have generally lower bias and variance if the overall estimates of IIS probabilities within and between pools are computed as unweighted averages of population-specific or pairwise estimates (see Equations A39 and A43 in File S1), as compared to weighted averages (Equations A46 and A47 in File S1). Equation A28 in File S1 further shows that our estimator may be rewritten as a function close to $(\hat{Q}_1 - \hat{Q}_2)/(1 - \hat{Q}_2)$, except that it also depends on the sum $\sum_i(\hat{Q}_{1i} - \hat{Q}_1)$ in both the numerator and the denominator. This suggests that if the Q_{1i}'s differ among subpopulations, then our estimator provides an estimate of an average of population-specific F_{ST} (Weir and Hill 2002; Weir and Goudet 2017).

It follows from the derivations in File S1 that the estimator $PP2_a$ (Equation 19) is biased because the IIS probability between pairs of reads within a pool $(\hat{Q}_1^r)$ is a biased estimator of the IIS probability between pairs of distinct genes in that pool (see Equations A34–A36 in File S1). This is the case because the former confounds pairs of reads that are identical because they were sequenced from a single gene from pairs of reads that are identical because they were sequenced from distinct, yet IIS genes.

A more justified estimator of F_{ST} has been proposed by Ferretti *et al.* (2013), based on previous developments by Futschik and Schlötterer (2010). Note that, although they defined F_{ST} as a ratio of functions of heterozygosities, they actually worked with IIS probabilities (see Equation 20 and Equation 21). However, although Equation 20 is strictly identical to Equation A39 in File S1, we note that they computed the total heterozygosity by integrating over pairs of genes sampled both within and between subpopulations (compare Equation 21 with Equation A43 in File S1), which may explain the observed bias (see Figure 2).

Comparison with alternative estimators

An alternative framework to Weir and Cockerham's (1984) analysis-of-variance has been developed by Masatoshi Nei and coworkers to estimate F_{ST} from gene diversities (Nei 1973, 1977, 1986; Nei and Chesser 1983). The estimator $PP2_d$ (see Equation 16, Equation 17, and Equation 18) implemented in the software package PoPoolation2 (Kofler *et al.* 2011) follows this logic. However, it has long been recognized that both frameworks are fundamentally different in that the analysis-of-variance approach considers both statistical and genetic (or evolutionary) sampling, whereas Nei and coworkers' approach do not (Weir and Cockerham 1984; Excoffier 2007; Holsinger and Weir 2009). Furthermore, the expectation of Nei and coworkers' estimators depend on the number of sampled populations, with a larger bias for lower numbers of sampled populations (Goudet 1993; Excoffier 2007; Weir and Goudet 2017). This is the case because the computation of the total diversity in Equation 18 and Equation 21 includes the comparison of pairs of genes from the same subpopulation, whereas the computation of IIS probabilities between subpopulations do not (see, *e.g.*, Excoffier 2007). Therefore, we do not recommend using the estimator $PP2_d$ implemented in the software package PoPoolation2 (Kofler *et al.* 2011).

Applications in evolutionary ecology studies

Pool-seq is being increasingly used in many application domains (Schlötterer *et al.* 2014), such as conservation genetics (see, *e.g.*, Fuentes-Pardo and Ruzzente 2017), invasion biology (see, *e.g.*, Dexter *et al.* 2018), and evolutionary biology in a broader sense (see, *e.g.*, Collet *et al.* 2016). These studies use a large range of methods, which aim at characterizing fine-scaled population structure (see, *e.g.*, Fischer *et al.*

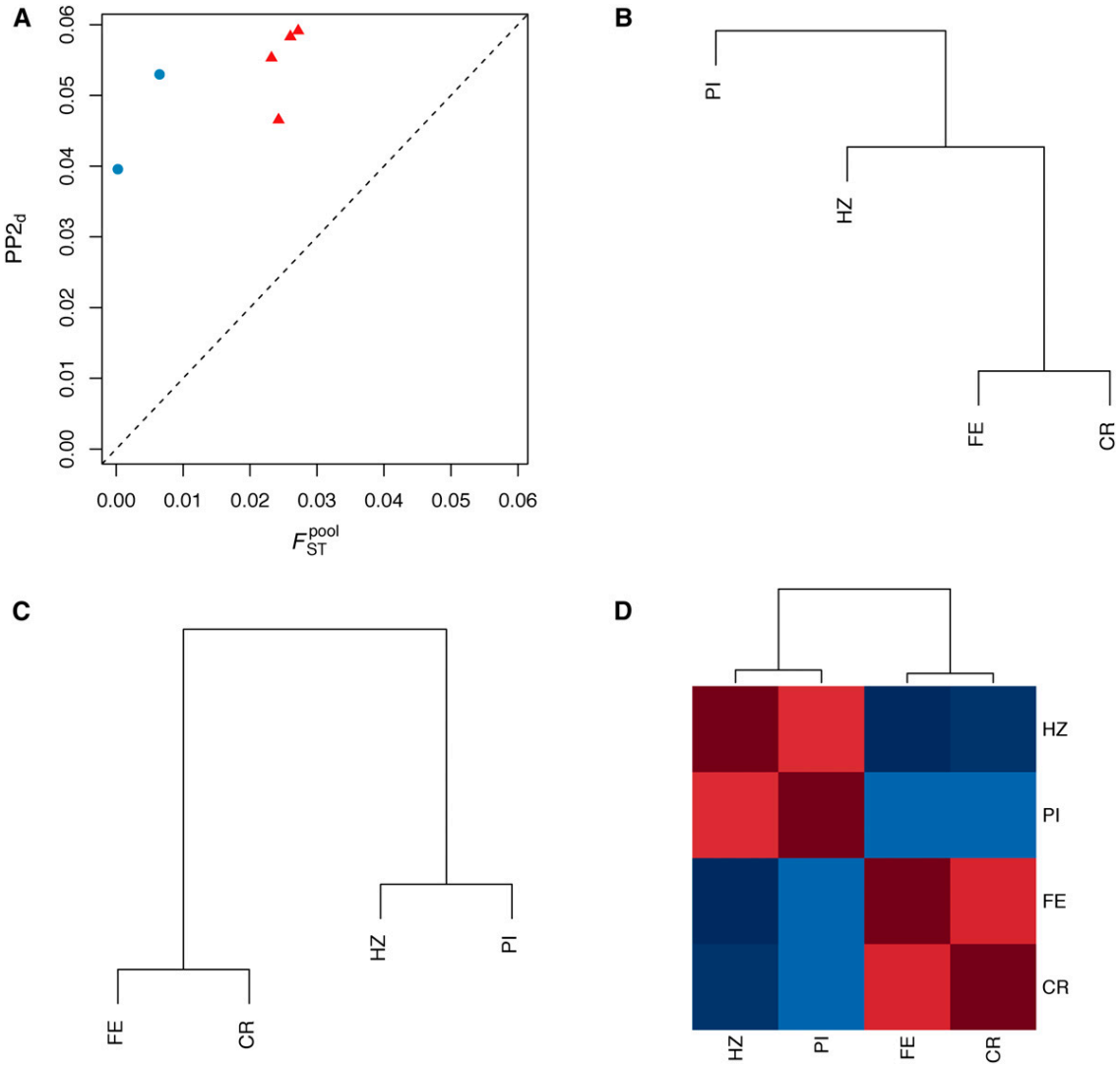


Figure 7 Reanalysis of the prickly sculpin (*C. asper*) Pool-seq data. (A) We compare the pairwise F_{ST} estimates $PP2_d$ and $\hat{F}_{ST}^{pool}$ for all pairs of populations from the estuarine (CR and FE) and freshwater samples (PI and HZ). Within-ecotype comparisons are depicted as ● and between-ecotype comparisons as ▲. (B and C) We show hierarchical cluster analyses based on (B) $PP2_d$ and (C) $\hat{F}_{ST}^{pool}$ pairwise estimates using unweighted pair group method with arithmetic mean (UPGMA). (D) We show a heatmap representation of the scaled covariance matrix among the four *C. asper* populations, inferred from the Bayesian hierarchical model implemented in the software package BayPass.

2017), reconstructing past demography (see, *e.g.*, Chen *et al.* 2016; Leblois *et al.* 2018), or identifying footprints of natural or artificial selection (see, *e.g.*, Chen *et al.* 2016; Fariello *et al.* 2017; Leblois *et al.* 2018).

Here, we reanalyzed the Pool-seq data produced by Dennenmoser *et al.* (2017), who investigated the adaptive genomic divergence between freshwater and brackish-water ecotypes of the prickly sculpin *C. asper*, an abundant euryhaline fish in northwestern North America. Measuring pairwise genetic differentiation between samples using $\hat{F}_{ST}^{pool}$, we found a clear-cut structure separating the freshwater from the brackish-water ecotypes. Such genetic structure supports the hypothesis that populations are locally adapted to osmotic conditions in these two contrasted habitats, as discussed in Dennenmoser *et al.* (2017). This structure, which is at odds with that inferred from $PP2_d$ estimates, is not only supported by the scaled covariance matrix of allele frequencies, but also by previous microsatellite-based studies, which showed that populations were genetically more differentiated between ecotypes than within ecotypes (Dennenmoser *et al.* 2014, 2015).

Limits of the model and perspectives

We have shown that the stronger source of bias for the $\hat{F}_{ST}^{pool}$ estimate is unequal contributions of individuals in pools. This is because we assume in our model that the read counts are multinomially distributed, which supposes that all genes contribute equally to the pool of reads (Gautier *et al.* 2013), *i.e.*, that there is no variation in DNA yield across individuals and that all genes have equal sequencing coverage (Rode *et al.* 2018). Because the effect of unequal contribution is expected

to be stronger with small pool sizes, it has been recommended to use Pool-seq with at least 50 diploid individuals per pool (Lynch *et al.* 2014; Schlötterer *et al.* 2014). However, this limit may be overly conservative for allele frequency estimates (Rode *et al.* 2018) and we have shown here that we can achieve very good precision and accuracy of F_{ST} estimates with smaller pool sizes. Furthermore, because genotypic information is lost during Pool-seq experiments, we assume in our derivations that pools are haploid (and therefore that F_{IS} is nil). Analyzing nonrandom mating populations (*e.g.*, in selfing species) is therefore problematic.

Finally, our model, as in Weir and Cockerham (1984), formally assumes that all populations provide independent replicates of some evolutionary process (Excoffier 2007; Holsinger and Weir 2009). This may be unrealistic in many natural populations, which motivated Weir and Hill (2002) to derive a population-specific estimator of F_{ST} for Ind-seq data (see also Vitalis *et al.* 2001). Even though the use of Weir and Hill's (2002) estimator is still scarce in the literature (but see Weir *et al.* 2005; Vitalis 2012), Weir and Goudet (2017) recently proposed a reinterpretation of population-specific estimates of F_{ST} in terms of allelic matching proportions, which are strictly equivalent to IIS probabilities between pairs of genes. It is therefore straightforward to extend Weir and Goudet's (2017) estimator of population-specific F_{ST} for the analysis of Pool-seq data, using the unbiased estimates of IIS probabilities provided in File S1.

Acknowledgments

We thank Alexandre Dehne-Garcia for his assistance in using computer farms. We thank two anonymous reviewers for their positive comments and suggestions. Analyses were performed on the GenoToul bioinformatics platform Toulouse Midi-Pyrénées (http://bioinfo.genotoul.fr) and the High Performance Computational platform of the Centre de Biologie pour la Gestion des Populations. This work is part of V.H.'s Ph.D.; V.H. was supported by a grant from the Institut National de la Recherche Agronomique's Plant Health and Environment (SPE) Division and by the BiodivERsA project EXOTIC (ANR-13-EBID-0001). Part of this work was supported by the project SWING (ANR-16-CE02-0015) of the French National Research Agency, and by the CORBAM project of the French region Hauts-de-France.

Literature Cited

Akey, J. M., G. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12: 1805–1814. https://doi.org/10.1101/gr.631202

Anderson, E. C., H. J. Skaug, and D. J. Barshis, 2014 Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. Mol. Ecol. 23: 502–512. https://doi.org/10.1111/mec.12609

Beaumont, M. A., 2005 Adaptation and speciation: what can F_{ST} tell us? Trends Ecol. Evol. 20: 435–440. https://doi.org/10.1016/j.tree.2005.05.017

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. Biol. Sci. 263: 1619–1626. https://doi.org/10.1098/rspb.1996.0237

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. 23: 1514–1521. https://doi.org/10.1101/gr.154831.113

Cavalli-Sforza, L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. B Biol. Sci. 164: 362–379. https://doi.org/10.1098/rspb.1966.0038

Chen, J., T. Källman, X.-F. Ma, G. Zaina, M. Morgante *et al.*, 2016 Identifying genetic signatures of natural selection using pooled populations sequencing in *Picea abies*. G3 (Bethesda) 6: 1979–1989. https://doi.org/10.1534/g3.116.028753

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution 23: 72–84. https://doi.org/10.1111/j.1558-5646.1969.tb03496.x

Cockerham, C. C., 1973 Analyses of gene frequencies. Genetics 74: 679–700.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA 84: 8512–8514. https://doi.org/10.1073/pnas.84.23.8512

Collet, J. M., S. Fuentes, J. Hesketh, M. S. Hill, P. Innocenti *et al.*, 2016 Rapid evolution of the intersexual genetic correlation for fitness in *Drosophila melanogaster*. Evolution 70: 781–795. https://doi.org/10.1111/evo.12892

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423. https://doi.org/10.1534/genetics.110.114819

Cutler, D. J., and J. D. Jensen, 2010 To pool, or not to pool? Genetics 186: 41–43. https://doi.org/10.1534/genetics.110.121012

Dennenmoser, S., S. M. Rogers, and S. M. Vamosi, 2014 Genetic population structure in prickly sculpin (*Cottus asper*) reflects isolation-by-environment between two life-history ecotypes. Biol. J. Linn. Soc. Lond. 113: 943–957. https://doi.org/10.1111/bij.12384

Dennenmoser, S., A. W. Nolte, S. M. Vamosi, and S. M. Rogers, 2015 Phylogeography of the prickly sculpin (*Cottus asper*) in north-western North America reveals parallel phenotypic evolution across multiple coastal-inland colonizations. J. Biogeogr. 42: 1626–1638. https://doi.org/10.1111/jbi.12527

Dennenmoser, S., S. M. Vamosi, S. W. Nolte, and S. M. Rogers, 2017 Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. Mol. Ecol. 26: 25–42. https://doi.org/10.1111/mec.13805

Dexter, E., S. M. Bollens, J. Cordell, H. Y. Soh, G. Rollwagen-Bollens *et al.*, 2018 A genetic reconstruction of the invasion of the calanoid copepod *Pseudodiaptomus inopinus* across the North American Pacific Coast. Biol. Invasions 20: 1577–1595. https://doi.org/10.1007/s10530-017-1649-0

Ellegren, H., 2014 Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29: 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Excoffier, L., 2007 Analysis of population subdivision, pp. 980–1020 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop, and C. Cannings. John Wiley & Sons, Chichester, United Kingdom.

Fariello, M. I., S. Boitard, S. Mercier, D. Robelin, T. Faraut *et al.*, 2017 Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. Mol. Ecol. 26: 3700–3714. https://doi.org/10.1111/mec.14141

Ferretti, L., S. Ramos Onsins, and M. Pérez-Enciso, 2013 Population genomics from pool sequencing. Mol. Ecol. 22: 5561–5576. https://doi.org/10.1111/mec.12522

Fischer, M. C., C. Rellstab, M. Leuzinger, M. Roumet, F. Gugerli *et al.*, 2017 Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. BMC Genomics 18: 69. https://doi.org/10.1186/s12864-016-3459-7

Fleiss, J. L., 1971 Measuring nominal scale agreement among many raters. Psychol. Bull. 76: 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L., and J. Cuzick, 1979 The reliability of dichotomous judgements: unequal numbers of judges per subject. Appl. Psychol. Meas. 3: 537–542. https://doi.org/10.1177/014662167900300410

Fuentes-Pardo, A. P., and D. E. Ruzzente, 2017 Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. Mol. Ecol. 26: 5369–5406. https://doi.org/10.1111/mec.14264

Futschik, A., and C. Schlötterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186: 207–218. https://doi.org/10.1534/genetics.110.114397

Gautier, M., 2015 Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics 201: 1555–1579. https://doi.org/10.1534/genetics.115.181453

Gautier, M., K. Gharbi, T. Cezaerd, M. Galan, A. Loiseau *et al.*, 2013 Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol. Ecol. 22: 3766–3779. https://doi.org/10.1111/mec.12360

Glenn, T. C., 2011 Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11: 759–769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Goudet, J., 1993 The genetics of geographically structured populations. Ph.D. Thesis, University of Wales, Bangor, Wales.

Holsinger, K. S., and B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat. Rev. Genet. 10: 639–650. https://doi.org/10.1038/nrg2611

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. C. Salmon Hillbertz, M. C. Zody *et al.*, 2007 Efficient mapping of Mendelian traits in dogs through genome-wide association. Nat. Genet. 39: 1321–1328. https://doi.org/10.1038/ng.2007.10

Kofler, R., R. V. Pandey, and C. Schlötterer, 2011 PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27: 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Landis, J. R., and G. G. Koch, 1977 A one-way components of variance model for categorical data. Biometrics 33: 671–679. https://doi.org/10.2307/2529465

Leblois, R., M. Gautier, A. Rohfritsch, J. Foucaud, C. Burban *et al.*, 2018 Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*. Mol. Ecol. 27: 264–278. https://doi.org/10.1111/mec.14411

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. Genetics 74: 175–195.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lotterhos, K. E., and M. C. Whitlock, 2014 Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. Mol. Ecol. 23: 2178–2192. https://doi.org/10.1111/mec.12725

Lotterhos, K. E., and M. C. Whitlock, 2015 The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol. Ecol. 24: 1031–1046. https://doi.org/10.1111/mec.13100

Lynch, M., D. Bost, S. Wilson, T. Maruki, and S. Harrison, 2014 Population-genetic inference from pooled-sequencing data. Genome Biol. Evol. 6: 1210–1218. https://doi.org/10.1093/gbe/evu085

Mak, T. K., 1988 Analysing intraclass correlation for dichotomous variables. J. R. Stat. Soc. Ser. C Appl. Stat. 37: 344–352.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson, Paris.

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70: 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Nei, M., 1977 *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41: 225–233. https://doi.org/10.1111/j.1469-1809.1977.tb01918.x

Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583–590.

Nei, M., 1986 Definition and estimation of fixation indices. Evolution 40: 643–645. https://doi.org/10.1111/j.1558-5646.1986.tb00516.x

Nei, M., and R. K. Chesser, 1983 Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47: 253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x

Nychka, D., R. Furrer, J. Paige, and S. Sain, 2017 fields: tools for spatial data. R package version 9.6. University Corporation for Atmospheric Research, Boulder, CO. DOI: 10.5065/D6W957CT

Orgogozo, V., A. E. Peluffo, and B. Morizot, 2016 The "mendelian gene" and the "molecular gene": two relevant concepts of genetic units, pp. 1–26 in *Genes and Evolution. Current Topics in Developmental Biology*, Vol. 119, edited by V. Orgogozo. Academic Press, New York.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967. https://doi.org/10.1371/journal.pgen.1002967

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767–779.

Ridout, M. S., C. G. B. Demktrio, and D. Firth, 1999 Estimating intra-class correlation for binary data. Biometrics 55: 137–148. https://doi.org/10.1111/j.0006-341X.1999.00137.x

Robertson, A., 1962 Weighting in the estimation of variance components in the unbalanced single classification. Biometrics 18: 413–417. https://doi.org/10.2307/2527485

Rode, N. O., Y. Holtz, K. Loridon, S. Santoni, J. Ronfort *et al.*, 2018 How to optimize the precision of allele and haplotype frequency estimates using pooled-sequencing data. Mol. Ecol. Resour. 18: 194–203. https://doi.org/10.1111/1755-0998.12723

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon *et al.*, 2013 Characterizing and measuring bias in sequence data. Genome Biol. 14: R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142: 1357–1362.

Rousset, F., 1997 Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics 145: 1219–1228.

Rousset, F., 2007 Inferences from spatial population genetics, pp. 945–979 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop, and C. Cannings. John Wiley & Sons, Ltd., Chichester, England.

Rousset, F., 2008 genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol. Ecol. Resour. 8: 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Schlötterer, C., R. Tobler, R. Kofler, and V. Nolte, 2014 Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15: 749–763. https://doi.org/10.1038/nrg3803

Slatkin, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47: 264–279. https://doi.org/10.1111/j.1558-5646.1993.tb01215.x

Smadja, C. M., B. Canbäck, R. Vitalis, M. Gautier, J. Ferrari *et al.*, 2012 Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution 66: 2723–2738. https://doi.org/10.1111/j.1558-5646.2012.01612.x

The International HapMap Consortium, 2005 A haplotype map of the human genome. Nature 437: 1299–1320. https://doi.org/10.1038/nature04226

Tukey, J. W., 1957 Variances of variance components: II. The unbalanced single classification. Ann. Math. Stat. 28: 43–56. https://doi.org/10.1214/aoms/1177707036

Vitalis, R., 2012 DetSel: an R-Package to detect marker loci responding to selection, pp. 277–293 in *Data Production and Analysis in Population Genomics: Methods and Protocols*. *Methods in Molecular Biology*, Vol. 888, edited by F. Pompanon, and A. Bonin. Humana Press, New York.

Vitalis, R., P. Boursot, and K. Dawson, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics 158: 1811–1823.

Wahlund, S., 1928 Zusammens etzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. Hereditas 11: 65–106. https://doi.org/10.1111/j.1601-5223.1928.tb02483.x

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Inc., Sunderland, MA.

Weir, B. S., 2012 Estimating *F*-statistics: a historical view. Philos. Sci. 79: 637–643. https://doi.org/10.1086/667904

Weir, B. S., and C. C. Cockerham, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weir, B. S., and J. Goudet, 2017 A unified characterization of population structure and relatedness. Genetics 206: 2085–2103. https://doi.org/10.1534/genetics.116.198424

Weir, B. S., and W. G. Hill, 2002 Estimating *F*-statistics. Annu. Rev. Genet. 36: 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. Genome Res. 15: 1468–1476. https://doi.org/10.1101/gr.4398405

Whitlock, M. C., and K. E. Lotterhos, 2015 Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. Am. Nat. 186: S24–S36. https://doi.org/10.1086/682949

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Wu, S., C. M. Crespi, and W. K. Wong, 2012 Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp. Clin. Trials 33: 869–880. https://doi.org/10.1016/j.cct.2012.05.004

Communicating editor: M. Beaumont

Measuring genetic differentiation from Pool-seq data

Valentin Hivert, Raphaël Leblois, Eric J. Petit, Mathieu Gautier and Renaud Vitalis

SUPPLEMENTAL FILE S1: DETAILED MATHEMATICAL DERIVATIONS

Analysis of variance for Pool-seq data

In the following, we first derive our model for a single locus. Consider a sample of n_d subpopulations, each of which is made of n_i genes ($i = 1, \ldots, n_\mathrm{d}$) sequenced in pools (hence n_i is the haploid sample size of the ith pool). We define c_{ij} as the number of reads sequenced from gene j ($j = 1, \ldots, n_\mathrm{i}$) in subpopulation i at the locus considered. Note that c_{ij} is a latent variable, that cannot be directly observed from the data. Let $X_{ijr:k}$ be an indicator variable for read r ($r = 1, \ldots, c_{ij}$) from gene j in subpopulation i, such that $X_{ijr:k} = 1$ if the rth read from the jth gene in the ith deme is of type k, and $X_{ijr:k} = 0$ otherwise. In the following, we use standard dot notations for sample averages, i.e.: $X_{ij\cdot:k} \equiv \sum_r X_{ijr:k}/c_{ij}$, $X_{i\cdots:k} \equiv \sum_j \sum_r X_{ijr:k} / \sum_j c_{ij}$ and $X_{\cdots:k} \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij}$. The analysis of variance is based on the computation of sums of squares, as follows:

$$\begin{aligned}
\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{\cdots:k})^2 &= \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{ij\cdot:k})^2 \\
&+ \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ij\cdot:k} - X_{i\cdots:k})^2 \\
&+ \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{i\cdots:k} - X_{\cdots:k})^2 \\
&\equiv SSR_{:k} + SSI_{:k} + SSP_{:k} \qquad \text{(A1)}
\end{aligned}$$

We express the sum of squares for reads within individuals as:

$$\begin{aligned} SSR_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ijr:k} - X_{ij\cdot:k}\right)^2 \\ &= 0 \end{aligned} \tag{A2}$$

since we assume that there is no sequencing error, i.e. all the reads sequenced from a single gene are identical (therefore $X_{ijr:k} = X_{ij\cdot:k}$, for all r). The sum of squares for genes within pools reads:

$$\begin{aligned} SSI_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ij\cdot:k} - X_{i\cdots:k}\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ij\cdot:k} - \pi_k\right)^2 - \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{i\cdots:k} - \pi_k\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} c_{ij} \left(X_{ij\cdot:k} - \pi_k\right)^2 - \sum_{i}^{n_{\mathrm{d}}} C_{1i} \left(X_{i\cdots:k} - \pi_k\right)^2 \end{aligned} \tag{A3}$$

where π_k is the expectation of the frequency of allele k over independent replicates of the evolutionary process, and $C_{1i} \equiv \sum_j c_{ij}$ is the total number of observed reads in the ith pool. Likewise, the sum of squares for genes between pools reads:

$$\begin{aligned} SSP_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{i\cdots:k} - X_{\cdots:k}\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} C_{1i} \left(X_{i\cdots:k} - \pi_k\right)^2 - C_1 \left(X_{\cdots:k} - \pi_k\right)^2 \end{aligned} \tag{A4}$$

where $C_1 \equiv \sum_i \sum_j c_{ij} = \sum_i C_{1i}$ is the total number of observed reads in the full sample. These sums can be expressed as functions of the average frequency of reads of type k for individual j: $\hat{\pi}_{ij:k} \equiv X_{ij\cdot:k}$, of the average fre-

quency of reads of type k within the ith pool: $\hat{\pi}_{i:k} \equiv X_{i\cdot:k}$, and of the average frequency of reads of type k in the full sample: $\hat{\pi}_k \equiv X_{\cdots:k}$. Note that from the definition of $X_{\cdots:k}$, $\hat{\pi}_k \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij} = \sum_i C_{1i} \hat{\pi}_{i:k} / \sum_i C_{1i}$ is the weighted average of the sample frequencies with weights equal to the pool coverage. Our approach is therefore equivalent to the weighted analysis-of-variance in Cockerham (1973) (see also Weir and Cockerham 1984; Weir 1996; Weir and Hill 2002; Rousset 2007; Weir and Goudet 2017). Then, developing the square in the first term in the right-hand side of Equation A3, we get:

$$\begin{aligned}
(X_{ij\cdot:k} - \pi_k)^2 &= \left(\frac{\sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{c_{ij}} \right)^2 \\
&= \frac{1}{c_{ij}^2} \left(\sum_r^{c_{ij}} X_{ijr:k} - c_{ij}\pi_k \right)^2 \\
&= \frac{1}{c_{ij}^2} \left(\sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} - 2c_{ij}^2 X_{ij\cdot:k}\pi_k + c_{ij}^2 \pi_k^2 \right) \\
&= \frac{1}{c_{ij}^2} \left(c_{ij} X_{ij\cdot:k} + c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&- \left. 2c_{ij}^2 X_{ij\cdot:k}\pi_k + c_{ij}^2 \pi_k^2 \right) \\
&= \hat{\pi}_{ij:k} - 2\pi_k \hat{\pi}_{ij:k} + \pi_k^2 \qquad \text{(A5)}
\end{aligned}$$

The sums of squares also depend on the unobserved frequency of pairs of genes sampled in the ith pool that are both of type k, i.e. the probability of identity in state (IIS) for allele k, for two distinct genes in the ith pool: $\hat{Q}_{1i:k} \equiv \left(\sum_{j \neq j'} \sum_{r,r'} X_{ijr:k} X_{ij'r':k} \right) / \left(C_{1i}^2 - \sum_j c_{ij}^2 \right)$. Then, developing the

square in the second term in the right-hand side of Equation A3, we get:

$$
\begin{aligned}
(X_{i\cdot\cdot:k} - \pi_k)^2 &= \left(\frac{\sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{C_{1i}} \right)^2 \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k} - C_{1i}\pi_k \right)^2 \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_j^{n_i} \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} \right. \\
&+ \left. \sum_{j \neq j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{ij'r':k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} c_{ij} X_{ij\cdot:k} + \sum_j^{n_i} c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&+ \left. \left(C_{1i}^2 - \sum_j^{n_i} c_{ij}^2 \right) \hat{Q}_{1i:k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} c_{ij}^2 \left(X_{ij\cdot:k} - X_{i\cdot\cdot:k} \right) + \left(C_{1i}^2 - \sum_j^{n_i} c_{ij}^2 \right) \left(\hat{Q}_{1i:k} - X_{i\cdot\cdot:k} \right) \right. \\
&+ \left. C_{1i}^2 X_{i\cdot\cdot:k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \hat{\pi}_{i:k} - 2\pi_k \hat{\pi}_{i:k} + \pi_k^2 + \sum_j^{n_i} \frac{c_{ij}^2}{C_{1i}^2} \left(\hat{\pi}_{ij:k} - \hat{\pi}_{i:k} \right) \\
&+ \left(1 - \sum_j^{n_i} \frac{c_{ij}^2}{C_{1i}^2} \right) \left(\hat{Q}_{1i:k} - \hat{\pi}_{i:k} \right) \qquad \text{(A6)}
\end{aligned}
$$

Last, the sums of squares depend on the unobserved frequency of pairs of genes sampled in the same pool that are both of type k, i.e. the IIS probability for allele k for two distinct genes in the same pool: $\hat{Q}_{1:k} \equiv \left(\sum_i \sum_{j \neq j'} \sum_{r,r'} X_{ijr:k} X_{ij'r':k} \right) / \left(C_2 - \sum_i \sum_j c_{ij}^2 \right)$, and of the unobserved frequency of pairs of genes sampled in different pools that are both of type k: $\hat{Q}_{2:k} \equiv \left(\sum_{i \neq i'} \sum_{j,j'} \sum_{r,r'} X_{ijr:k} X_{i'j'r':k} \right) / \left(C_1^2 - C_2 \right)$, where $C_2 \equiv \sum_i C_{1i}^2$.

Developing the second term in the right-hand side of Equation A4, we get:

$$
\begin{aligned}
(X_{\cdots:k} - \pi_k)^2 &= \left(\frac{\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{C_1} \right)^2 \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k} - C_1 \pi_k \right)^2 \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} \right. \\
&+ \sum_i^{n_\mathrm{d}} \sum_{j \neq j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{ij'r':k} + \sum_{i \neq i'}^{n_\mathrm{d}} \sum_{j,j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{i'j'r':k} \\
&- \left. 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij} X_{ij\cdot:k} + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&+ \left. \left(C_2 - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \right) \hat{Q}_{1:k} + \left(C_1^2 - C_2 \right) \hat{Q}_{2:k} - 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \left(X_{ij\cdot:k} - X_{\cdots:k} \right) + \left(C_2 - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \right) \left(\hat{Q}_{1:k} - X_{\cdots:k} \right) \right. \\
&+ \left. \left(C_1^2 - C_2 \right) \left(\hat{Q}_{2:k} - X_{\cdots:k} \right) + C_1^2 X_{\cdots:k} - 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \hat{\pi}_k - 2\pi_k \hat{\pi}_k + \pi_k^2 + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \frac{c_{ij}^2}{C_1^2} \left(\hat{\pi}_{ij:k} - \hat{\pi}_k \right) \\
&+ \left(\frac{C_2}{C_1^2} - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \frac{c_{ij}^2}{C_1^2} \right) \left(\hat{Q}_{1:k} - \hat{\pi}_k \right) + \left(1 - \frac{C_2}{C_1^2} \right) \left(\hat{Q}_{2:k} - \hat{\pi}_k \right) \qquad \text{(A7)}
\end{aligned}
$$

Hence, developing the first term in the right-hand side of Equation A3 using Equation A5, we have:

$$
\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij} \left(X_{ij\cdot:k} - \pi_k \right)^2 = C_1 \left(\hat{\pi}_k - 2\pi_k \hat{\pi}_k + \pi_k^2 \right) \qquad \text{(A8)}
$$

Likewise, developing the second term in the right-hand side of Equation A3 using Equation A6, we get:

$$\begin{aligned}\sum_{i}^{n_{\mathrm{d}}} C_{1i}\left(X_{i\cdot:k}-\pi_k\right)^2 &= C_1\left(\hat{\pi}_k-2\pi_k\hat{\pi}_k+\pi_k^2\right)+\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{ij:k}-\hat{\pi}_{i:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{Q}_{1i:k}-\hat{\pi}_{i:k}\right) \qquad \text{(A9)}\end{aligned}$$

Last, developing the second term in the right-hand side of Equation A4 using Equation A7, we get:

$$\begin{aligned}C_1\left(X_{\cdots:k}-\pi_k\right)^2 &= C_1\left(\hat{\pi}_k-2\pi_k\hat{\pi}_k+\pi_k^2\right)+\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\left(\hat{\pi}_{ij:k}-\hat{\pi}_k\right)\\ &+ \left(\frac{C_2}{C_1}-\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\right)\left(\hat{Q}_{1:k}-\hat{\pi}_k\right)\\ &+ \left(C_1-\frac{C_2}{C_1}\right)\left(\hat{Q}_{2:k}-\hat{\pi}_k\right) \qquad \text{(A10)}\end{aligned}$$

Then, from Equations A3, A8 and A9:

$$\begin{aligned}SSI_{:k} &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{i:k}-\hat{\pi}_{ij:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{\pi}_{i:k}-\hat{Q}_{1i:k}\right) \qquad \text{(A11)}\end{aligned}$$

and from Equations A4, A9 and A10:

$$\begin{aligned}SSP_{:k} &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{ij:k}-\hat{\pi}_{i:k}\right)-\sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{\pi}_{i:k}-\hat{Q}_{1i:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\left(\hat{\pi}_k-\hat{\pi}_{ij:k}\right)+\left(\frac{C_2}{C_1}-\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\right)\left(\hat{\pi}_k-\hat{Q}_{1:k}\right)\\ &+ \left(C_1-\frac{C_2}{C_1}\right)\left(\hat{\pi}_k-\hat{Q}_{2:k}\right) \qquad \text{(A12)}\end{aligned}$$

Taking expectation over all possible samples from all replicate populations sharing the same evolutionary history, we get from Equation A11:

$$\begin{aligned}
\mathbb{E}(SSI_{:k}) &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1i}}\right) \\
&+ \sum_{i}^{n_{\mathrm{d}}} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{Q}_{1i:k}\right) \mathbb{E}\left(C_{1i} - \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right) \\
&= (\pi_k - Q_{1:k}) \left(C_1 - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right)\right) \qquad \text{(A13)}
\end{aligned}$$

where $Q_{1:k}$ is the expected IIS probability that two genes in the same pool are both of type k. Likewise, from Equation A12:

$$\begin{aligned}
\mathbb{E}(SSP_{:k}) &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1i}}\right) + \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1}}\right) \\
&- \sum_{i}^{n_{\mathrm{d}}} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{Q}_{1i:k}\right) \mathbb{E}\left(C_{1i} - \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right) \\
&+ \mathbb{E}\left(\hat{\pi}_{k} - \hat{Q}_{1:k}\right) \mathbb{E}\left(\frac{C_2}{C_1} - \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_1}\right) \\
&+ \left(C_1 - \frac{C_2}{C_1}\right) \mathbb{E}\left(\hat{\pi}_{k} - \hat{Q}_{2:k}\right) \\
&= (\pi_k - Q_{1:k}) \left(\frac{C_2}{C_1} - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_1}\right)\right) \\
&- (\pi_k - Q_{1:k}) \left(C_1 - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right)\right) \\
&+ \left(C_1 - \frac{C_2}{C_1}\right) (\pi_k - Q_{2:k}) \qquad \text{(A14)}
\end{aligned}$$

where $Q_{2:k}$ is the expected IIS probability that two genes from different pools are both of type k. Note that the expected sums $\mathbb{E}\left(\sum_i \sum_j c_{ij}^2\right)/C_{1i}$ and $\mathbb{E}\left(\sum_i \sum_j c_{ij}^2\right)/C_1$ in Equations A13 and A14 depend on the latent variable

c_{ij}, that cannot be directly observed from the data. Therefore, we must make an assumption on the distribution of the c_{ij}'s to proceed. In the following, we assume that for each pool i, c_{ij} follows a multinomial distribution with parameter C_{1i} (the number of trials, i.e. the total number of reads in the ith pool) and probabilities $(1/n_i, \ldots, 1/n_i)$ for the n_i individuals in the pool. Then:

$$\begin{aligned}
\mathbb{E}\left(\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\frac{c_{ij}^2}{C_{1i}}\right) &= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\mathbb{E}\left(c_{ij}^2\right) \\
&= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\left(\mathbb{E}\left(c_{ij}\right)^2+\mathbb{V}\left(c_{ij}\right)\right) \\
&= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\left(\left(\frac{C_{1i}}{n_i}\right)^2+\frac{C_{1i}}{n_i}\left(\frac{n_i-1}{n_i}\right)\right) \\
&= \sum_i^{n_\mathrm{d}}\left(\frac{C_{1i}}{n_i}+\left(\frac{n_i-1}{n_i}\right)\right) \equiv D_2 \qquad \text{(A15)}
\end{aligned}$$

and:

$$\begin{aligned}
\mathbb{E}\left(\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\frac{c_{ij}^2}{C_1}\right) &= \frac{1}{C_1}\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\mathbb{E}\left(c_{ij}^2\right) \\
&= \frac{1}{C_1}\sum_i^{n_\mathrm{d}}C_{1i}\left[\frac{C_{1i}}{n_i}+\left(\frac{n_i-1}{n_i}\right)\right] \equiv D_2^\star \qquad \text{(A16)}
\end{aligned}$$

Hence, from Equations A13 and A15, we have:

$$\mathbb{E}(SSI_{:k}) = (C_1 - D_2)\left(\pi_k - Q_{1:k}\right) \qquad \text{(A17)}$$

and from Equations A14 and A16:

$$\begin{aligned}\mathbb{E}(SSP_{:k}) &= \left(\frac{C_2}{C_1} - D_2^\star\right)(\pi_k - Q_{1:k}) - (C_1 - D_2)(\pi_k - Q_{1:k}) \\ &+ \left(C_1 - \frac{C_2}{C_1}\right)(\pi_k - Q_{2:k}) \\ &= \left(C_1 - \frac{C_2}{C_1}\right)(Q_{1:k} - Q_{2:k}) \\ &+ (D_2 - D_2^\star)(\pi_k - Q_{1:k}) \end{aligned} \tag{A18}$$

Summing over alleles, we get the following expressions for the expected sums of squares for genes between individuals within pools:

$$\mathbb{E}(SSI) = \sum_k \mathbb{E}(SSI_{:k}) = (C_1 - D_2)(1 - Q_1) \tag{A19}$$

and for genes between individuals from different pools:

$$\begin{aligned}\mathbb{E}(SSP) &= \sum_k \mathbb{E}(SSP_{:k}) \\ &= \left(C_1 - \frac{C_2}{C_1}\right)(Q_1 - Q_2) + (D_2 - D_2^\star)(1 - Q_1) \end{aligned} \tag{A20}$$

Rearranging Equations A19–A20, we get:

$$Q_1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) - (D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \tag{A21}$$

and:

$$1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) + (n_\mathrm{c} - 1)(D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \tag{A22}$$

where $n_\mathrm{c} \equiv (C_1 - C_2/C_1)/(D_2 - D_2^\star)$. Let $MSI \equiv SSI/(C_1 - D_2)$ and $MSP \equiv SSP/(D_2 - D_2^\star)$. Then, using the definition of F_{ST} from Equation 1

in the main text, and rearranging Equations A21–A22, we get:

$$F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{\mathbb{E}(MSP) - \mathbb{E}(MSI)}{\mathbb{E}(MSP) + (n_{\mathrm{c}} - 1)\,\mathbb{E}(MSI)} \tag{A23}$$

which yields the method-of-moments estimator:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{MSP - MSI}{MSP + (n_{\mathrm{c}} - 1)\,MSI} \tag{A24}$$

Since SSI (Equation A3) and SSP (Equation A4) may be rewritten in terms of sample frequencies as:

$$\begin{aligned} SSI &= \sum_k SSI_{:k} = \sum_k^{n_{\mathrm{d}}} \sum_i^{n_i} \sum_j^{c_{ij}} \sum_r (X_{ij:k} - X_{i\cdot:k})^2 \\ &= \sum_k^{n_{\mathrm{d}}} \sum_i C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right) \end{aligned} \tag{A25}$$

and:

$$\begin{aligned} SSP &= \sum_k SSP_k = \sum_k^{n_{\mathrm{d}}} \sum_i^{n_i} \sum_j^{c_{ij}} \sum_r (X_{i\cdot:k} - X_{\cdot\cdot:k})^2 \\ &= \sum_k^{n_{\mathrm{d}}} \sum_i C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 \end{aligned} \tag{A26}$$

our estimator then takes the form:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\sum_k \left[(C_1 - D_2) \sum_i^{n_{\mathrm{d}}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 - (D_2 - D_2^{\star}) \sum_i^{n_{\mathrm{d}}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right)\right]}{\sum_k \left[(C_1 - D_2) \sum_i^{n_{\mathrm{d}}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 + (n_{\mathrm{c}} - 1)\,(D_2 - D_2^{\star}) \sum_i^{n_{\mathrm{d}}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right)\right]} \tag{A27}$$

The estimator in Equation A24 can also be expressed as a function of the

frequencies of identical pairs of genes $\hat{Q}_1 = \sum_k \hat{Q}_{1:k}$ and $\hat{Q}_2 = \sum_k \hat{Q}_{2:k}$, as:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\left(\hat{Q}_1 - \hat{Q}_2\right)\alpha + \left(C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_1}\right)\beta}{\left(1 - \hat{Q}_2\right)\alpha + \left(C_2/C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_1}\right)\beta} \tag{A28}$$

where:

$$\alpha \equiv \left(C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_{1i}}\right)\left(C_1 - \frac{C_2}{C_1}\right) \tag{A29}$$

and:

$$\beta \equiv \sum_i \left(C_{1i} - \sum_j \frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{Q}_{1i} - \hat{Q}_1\right) \tag{A30}$$

If we take the limit case where the number of sequenced reads per gene is constant, i.e. if $C_{1i} = C$, for all $i \in (1, \ldots, n_{\mathrm{d}})$, then it can be shown that Equation A28 reduces exactly to Equations 28A29–28A30 in Rousset (2007), p. 977. Furthermore, if the pools have all the same size, i.e. if $n_i = n$ for all $i \in (1, \ldots, n_{\mathrm{d}})$, then $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$.

If the pools have all the same size and if the number of reads per pool is constant, then one can also show that Equations A25–A26 reduce to:

$$SSI = n_{\mathrm{d}}(C - 1)\left(1 - \hat{Q}_1^{\mathrm{r}}\right) \tag{A31}$$

and:

$$SSP = C(n_{\mathrm{d}} - 1)\left(1 - \hat{Q}_2^{\mathrm{r}}\right) - (n_{\mathrm{d}} - 1)(C - 1)\left(1 - \hat{Q}_1^{\mathrm{r}}\right) \tag{A32}$$

where $\hat{Q}_1^{\mathrm{r}}$ and $\hat{Q}_2^{\mathrm{r}}$ are the frequencies of identical pairs of reads within and between pools, respectively, computed by simple counting of IIS pairs. These

are (unweighted) averages of the population-specific estimates $\hat{Q}^{\mathrm{r}}_{1i}$ (Equation A34) and the pairwise estimates $\hat{Q}^{\mathrm{r}}_{2ii'}$ (Equation A40), respectively. Then, from Equation A24, we get:

$$\hat{F}^{\mathrm{pool}}_{\mathrm{ST}} = 1 - \left(\frac{1-\hat{Q}^{\mathrm{r}}_{1}}{1-\hat{Q}^{\mathrm{r}}_{2}}\right)\left(\frac{n}{n-1}\right) \tag{A33}$$

IIS probabilities for Pool-seq data

In this Appendix, we provide unbiased estimates of IIS probabilies between pairs of genes, computed from read count data. Let $r_{i:k} = \sum_j \sum_r X_{ijr:k}$ be the number of reads of type k in the ith pool. A straightforward estimate of the IIS probability between pairs of reads in the ith pool is given by:

$$\hat{Q}^{\mathrm{r}}_{1i} \equiv \frac{\sum_k r_{i:k}\left(r_{i:k}-1\right)}{C_{1i}\left(C_{1i}-1\right)} \tag{A34}$$

where $C_{1i} = \sum_k r_{i:k}$. As above (see Equations A15 and A16), we assume that in each pool, the conditional distribution of the read counts $r_{i:k}$, given the (unobserved) allele counts $y_{i:k}$, is binomial, i.e.: $r_{i:k} \mid y_{i:k} \sim \mathrm{Bin}\left(y_{i:k}/n_i, C_{1i}\right)$. The conditional expectation of the number of reads is therefore given by: $\mathbb{E}(r_{i:k} \mid y_{i:k}) = C_{1i}\left(y_{i:k}/n_i\right)$, and the conditional expectation of the squared number of reads by: $\mathbb{E}(r^2_{i:k} \mid y_{i:k}) = C_{1i}(C_{1i}-1)\left(y_{i:k}/n_i\right)^2 + C_{1i}\left(y_{i:k}/n_i\right)$. Therefore, the conditional expectation of the IIS probability between pairs of reads in the ith pool reads:

$$\mathbb{E}\left(\hat{Q}^{\mathrm{r}}_{1i} \mid y_{i:k}\right) = \frac{\sum_k \mathbb{E}\left(r^2_{i:k} - r_{i:k}\right)}{C_{1i}\left(C_{1i}-1\right)} = \sum_k \left(\frac{y_{i:k}}{n_i}\right)^2 \tag{A35}$$

Since

$$\hat{Q}_{1i} \equiv \frac{\sum_k y_{i:k}\left(y_{i:k}-1\right)}{n_i\left(n_i-1\right)} \tag{A36}$$

is an unbiased estimate of the IIS probability between pairs of distinct genes in the ith pool, Equation A35 implies that $\hat{Q}^{\mathrm{r}}_{1i}$ (Equation A34) is a biased estimate of that quantity (i.e., the IIS probability between pairs of reads within a pool is a biased estimate of the IIS probability between pairs of

distinct genes in that pool). This is so, because the former confounds pairs of reads that are identical because they were sequenced from a single gene copy, from pairs of reads (from distinct gene copies) that are identical because they share a common ancestor. However, inspection of Equation A35 suggests that an unbiased estimate of $\hat{Q}_{1i}$ may be given by:

$$\hat{Q}_{1i}^{\text{pool}} \equiv 1 - \frac{n_i}{n_i - 1}\left(1 - \hat{Q}_{1i}^{\text{r}}\right) \tag{A37}$$

Taking expectation of Equation A37, we get indeed:

$$\begin{aligned}
\mathbb{E}\left(\hat{Q}_{1i}^{\text{pool}} \mid y_{i:k}\right) &= \frac{n_i}{n_i - 1}\mathbb{E}\left(\hat{Q}_{1i}^{\text{r}}\right) - \frac{1}{n_i - 1} \\
&= \frac{n_i}{n_i - 1}\sum_k \left(\frac{y_{i:k}}{n_i}\right)^2 - \frac{n_i}{n_i(n_i - 1)} \\
&= \frac{\sum_k y_{i:k}^2}{n_i(n_i - 1)} - \frac{\sum_k y_{i:k}}{n_i(n_i - 1)} \\
&= \frac{\sum_k y_{i:k}(y_{i:k} - 1)}{n_i(n_i - 1)} \equiv \hat{Q}_{1i}
\end{aligned} \tag{A38}$$

Following Weir and Goudet (2017), we define the overall IIS probability between pairs of genes within pools as the unweighted average of population-specific estimates, leading to:

$$\hat{Q}_1^{\text{pool}} \equiv \frac{\sum_i \hat{Q}_{1i}^{\text{pool}}}{n_{\text{d}}} = 1 - \frac{1}{n_{\text{d}}}\sum_i \frac{n_i}{n_i - 1}\left(1 - \hat{Q}_{1i}^{\text{r}}\right) \tag{A39}$$

A straightforward estimate of the IIS probability between pairs of reads taken in different pools i and i' is given by:

$$\hat{Q}_{2ii'}^{\text{r}} \equiv \frac{\sum_k r_{i:k} r_{i':k}}{C_{1i} C_{1i'}} \tag{A40}$$

Since we assume that pools are conditionally independent, taking expectation gives:

$$\begin{aligned}\mathbb{E}\left(\hat{Q}^{\mathrm{r}}_{2ii'} \mid y_{i:k}, y_{i':k}\right) &= \frac{\sum_k \mathbb{E}(r_{i:k})\mathbb{E}(r_{i':k})}{C_{1i}C_{1i'}} \\ &= \sum_k \left(\frac{y_{i:k}y_{i':k}}{n_i n_{i'}}\right) \equiv \hat{Q}_{2ii'} \end{aligned} \tag{A41}$$

Therefore, the IIS probability between pairs of reads sampled in different pools is an unbiased estimate of the IIS probability between pairs of genes in these pools, and an unbiased estimate of the IIS probabilitiy of genes sampled from different pools is given by:

$$\hat{Q}^{\mathrm{pool}}_{2ii'} \equiv \hat{Q}^{\mathrm{r}}_{2ii'} \tag{A42}$$

As above, we define the overall IIS probability between pairs of genes sampled from different pools as the unweighted average of pairwise estimates, i.e.:

$$\hat{Q}^{\mathrm{pool}}_{2} \equiv \frac{\sum_{i\neq i'} \hat{Q}^{\mathrm{pool}}_{2ii'}}{n_{\mathrm{d}}(n_{\mathrm{d}}-1)} = 1 - \frac{1}{n_{\mathrm{d}}(n_{\mathrm{d}}-1)} \sum_{i\neq i'} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'}\right) \tag{A43}$$

We can then derive an IIS-based estimator of F_{ST}, as:

$$\begin{aligned}\hat{F}^{\mathrm{pool-PID}}_{\mathrm{ST}} &\equiv \frac{\hat{Q}^{\mathrm{pool}}_{1} - \hat{Q}^{\mathrm{pool}}_{2}}{1 - \hat{Q}^{\mathrm{pool}}_{2}} = 1 - \frac{1 - \hat{Q}^{\mathrm{pool}}_{1}}{1 - \hat{Q}^{\mathrm{pool}}_{2}} \\ &= 1 - \frac{\sum_i \left[\left(1 - \hat{Q}^{\mathrm{r}}_{1i}\right) n_i / \left(n_i - 1\right)\right]}{\sum_{i\neq i'} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'}\right) / \left(n_{\mathrm{d}} - 1\right)} \end{aligned} \tag{A44}$$

which, to the extent that we may take the expectation of a ratio to be the ratio of expectations, is unbiased. If the pools have all the same size (i.e., if

$n_i = n$ for all i), then Equation A44 reduces to:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}} = 1 - \left(\frac{1-\hat{Q}_1^{\mathrm{r}}}{1-\hat{Q}_2^{\mathrm{r}}}\right)\left(\frac{n}{n-1}\right) \tag{A45}$$

where $\hat{Q}_1^{\mathrm{r}} \equiv \sum_i \hat{Q}_{1i}^{\mathrm{r}}/n_{\mathrm{d}}$ and $\hat{Q}_2^{\mathrm{r}} \equiv \sum_{i\neq i'} \hat{Q}_{2ii'}^{\mathrm{r}}/\left[n_{\mathrm{d}}(n_{\mathrm{d}}-1)\right]$. Note that Equation A45 is strictly identical to Equation A33. Therefore, if the pools have all the same size and if the number of reads per pool is constant, the analysis-of-variance estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ is strictly equivalent to the estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$ based on the computation of IIS probabilities between pairs of reads, with appropriate bias correction (see Equation A37). This echoes the derivations by Rousset (2007) for Ind-seq data, who showed that the analysis-of-variance approach (Weir and Cockerham 1984) and the simple strategy of estimating IIS probabilities by counting identical pairs of genes provides identical estimates when sample sizes are equal (see also Cockerham and Weir 1987; Karlsson et al. 2007).

Alternatively, the overall IIS probability between pairs of genes within pools may be defined as the weighted average of population-specific estimates, with weights equal to the number of pairs of genes in each pool (see Rousset 2007), i.e.:

$$\tilde{Q}_1^{\mathrm{pool}} \equiv \frac{\sum_i n_i(n_i-1)\hat{Q}_{1i}^{\mathrm{pool}}}{\sum_i n_i(n_i-1)} \tag{A46}$$

Likewise, the overall IIS probability between pairs of genes sampled from different pools may be defined as the weighted average of pairwise estimates, with weights equal to the number of pairs of genes sampled between pools,

i.e.:

$$\tilde{Q}_2^{\text{pool}} \equiv \frac{\sum_{i \neq i'} n_i n_{i'} \hat{Q}_{2ii'}^{\text{pool}}}{\sum_{i \neq i'} n_i n_{i'}} \tag{A47}$$

We can then derive an IIS-based estimator of F_{ST}, using weighted IIS probabilities, as:

$$\begin{aligned} \tilde{F}_{\text{ST}}^{\text{pool-PID}} &\equiv \frac{\tilde{Q}_1^{\text{pool}} - \tilde{Q}_2^{\text{pool}}}{1 - \tilde{Q}_2^{\text{pool}}} = 1 - \frac{1 - \tilde{Q}_1^{\text{pool}}}{1 - \tilde{Q}_2^{\text{pool}}} \\ &= 1 - \frac{\sum_i \left[n_i^2 \left(1 - \hat{Q}_{1i}^{\text{r}} \right) \right] / \sum_i n_i (n_i - 1)}{\sum_{i \neq i'} n_i n_{i'} \left(1 - \hat{Q}_{2ii'}^{\text{r}} \right) / \sum_{i \neq i'} n_i n_{i'}} \end{aligned} \tag{A48}$$

If the pools have all the same size (i.e., if $n_i = n$ for all i), then Equation A48 reduces to Equation A45, and $\tilde{F}_{\text{ST}}^{\text{pool-PID}} = \hat{F}_{\text{ST}}^{\text{pool-PID}}$. With unbalanced samples, simulation analyses show that $\tilde{F}_{\text{ST}}^{\text{pool-PID}}$ has larger bias and variance than $\hat{F}_{\text{ST}}^{\text{pool-PID}}$, in particular for low levels of differentiation (see Figure S4).

LITERATURE CITED

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, 74:679–700.

Cockerham, C. C. and Weir, B. S. (1987). Analyses of gene frequencies. *Proc. Natl. Acad. Sci. USA*, 84:8512–8514.

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H. C., Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R., Kulbokas, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler, H., Kämpe, O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson, L., and Lindblad-Toh, K. (2007). Efficient mapping of Mendelian traits in dogs through genome-wide association. *Nat. Genet.*, 39:1321–1328.

Rousset, F. (2007). Inferences from spatial population genetics. In Balding, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 945–979, Chichester. John Wiley & Sons, Ltd.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc., Sunderland, MA.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370.

Weir, B. S. and Goudet, J. (2017). An unified characterization of population structure and relatedness. *Genetics*, 206:2085–2103.

Weir, B. S. and Hill, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.*, 36:721–750.

Table S1 Comparison of pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{ST}^{pool}$	WC_{84}
0.05	10	20×	0.051 (0.004)	
0.05	10	50×	0.051 (0.004)	0.051 (0.003)
0.05	10	100×	0.051 (0.003)	
0.05	100	20×	0.051 (0.003)	
0.05	100	50×	0.051 (0.003)	0.051 (0.002)
0.05	100	100×	0.051 (0.002)	
0.20	10	20×	0.203 (0.007)	
0.20	10	50×	0.202 (0.006)	0.202 (0.007)
0.20	10	100×	0.201 (0.006)	
0.20	100	20×	0.201 (0.006)	
0.20	100	50×	0.201 (0.006)	0.201 (0.005)
0.20	100	100×	0.202 (0.005)	

Pairwise multilocus $\hat{F}_{ST}^{pool}$ estimates were computed for various conditions of expected F_{ST}, pool size (n) and coverage (Cov.) in an island model with $n_d = 8$ subpopulations (pools). The mean (RMSE) is computed for a single pair of subpopulations, over 50 independent simulated datasets, each made of 5,000 loci. For comparison, we computed multilocus WC_{84} estimates from allele count data inferred from individual genotypes (Ind-seq).

Table S2 Effect of unequal sampling on pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	WC_{84}
0.05	$\mathcal{N}(100, 30)$	20×	0.051 (0.003)	
0.05	$\mathcal{N}(100, 30)$	50×	0.052 (0.003)	0.051 (0.002)
0.05	$\mathcal{N}(100, 30)$	100×	0.051 (0.002)	
0.20	$\mathcal{N}(100, 30)$	20×	0.202 (0.007)	
0.20	$\mathcal{N}(100, 30)$	50×	0.202 (0.006)	0.202 (0.006)
0.20	$\mathcal{N}(100, 30)$	100×	0.202 (0.006)	

Pairwise multilocus $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates were computed for various conditions of expected F_{ST} and coverage (Cov.) in an island model with $n_{\mathrm{d}} = 8$ subpopulations (pools). The pool size (n) was variable across demes, with haploid sample size n drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; n was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. The mean (RMSE) is computed for a single pair of subpopulations, over 50 independent simulated datasets, each made of 5,000 loci. For comparison, we computed multilocus WC_{84} (Weir and Cockerham 1984) estimates from allele count data inferred from individual genotypes (Ind-seq).

Table S3 Effect of variable coverage on pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Δ	$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	WC_{84}
0.05	10	20	0.050 (0.006)	
0.05	10	50	0.050 (0.004)	0.050 (0.004)
0.05	10	100	0.050 (0.004)	
0.05	100	20	0.051 (0.003)	
0.05	100	50	0.051 (0.002)	0.051 (0.002)
0.05	100	100	0.051 (0.002)	
0.20	10	20	0.200 (0.007)	
0.20	10	50	0.200 (0.007)	0.200 (0.007)
0.20	10	100	0.200 (0.007)	
0.20	100	20	0.202 (0.006)	
0.20	100	50	0.203 (0.006)	0.203 (0.005)
0.20	100	100	0.203 (0.005)	

Pairwise multilocus $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates were computed for various conditions of expected F_{ST} and pool size (n) in an island model with $n_{\mathrm{d}} = 8$ subpopulations (pools). The coverage (δ_i) was varying across demes and loci, with $\delta_i \sim \mathrm{Pois}(\Delta)$. The mean (RMSE) is computed for a single pair of subpopulations, over 50 independent simulated datasets, each made of 5,000 loci. For comparison, we computed multilocus WC_{84} estimates from allele count data inferred from individual genotypes (Ind-seq).

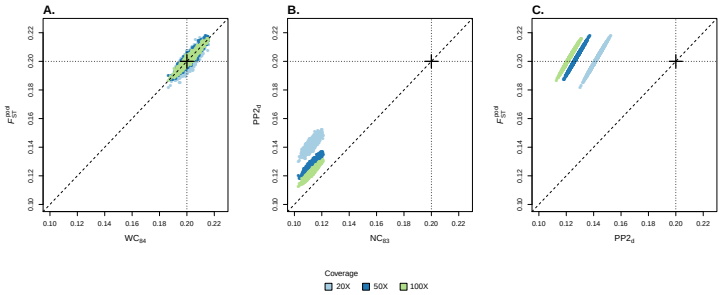


Figure S1 Pairwise estimators of F_{ST}. A. Multilocus estimates $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ computed from read counts, as a function of WC_{84} estimates computed from individual genotypes. B. Multilocus estimates $\mathrm{PP2_d}$ computed from read counts, as a function of NC_{83} estimates computed from individual genotypes. C. Multilocus estimates $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ as a function of multilocus $\mathrm{PP2_d}$ estimates. In each graph, the dots represent multilocus estimates of F_{ST} across all pairs of subpopulations from an 8-island model, and over 50 replicate `ms` simulations. We specified the migration rate corresponding to $F_{\mathrm{ST}} = 0.20$. The size of each pool was fixed to 100. The results are shown for different coverages (20X, 50X and 100X). The cross indicates the simulated value of the parameter F_{ST}.

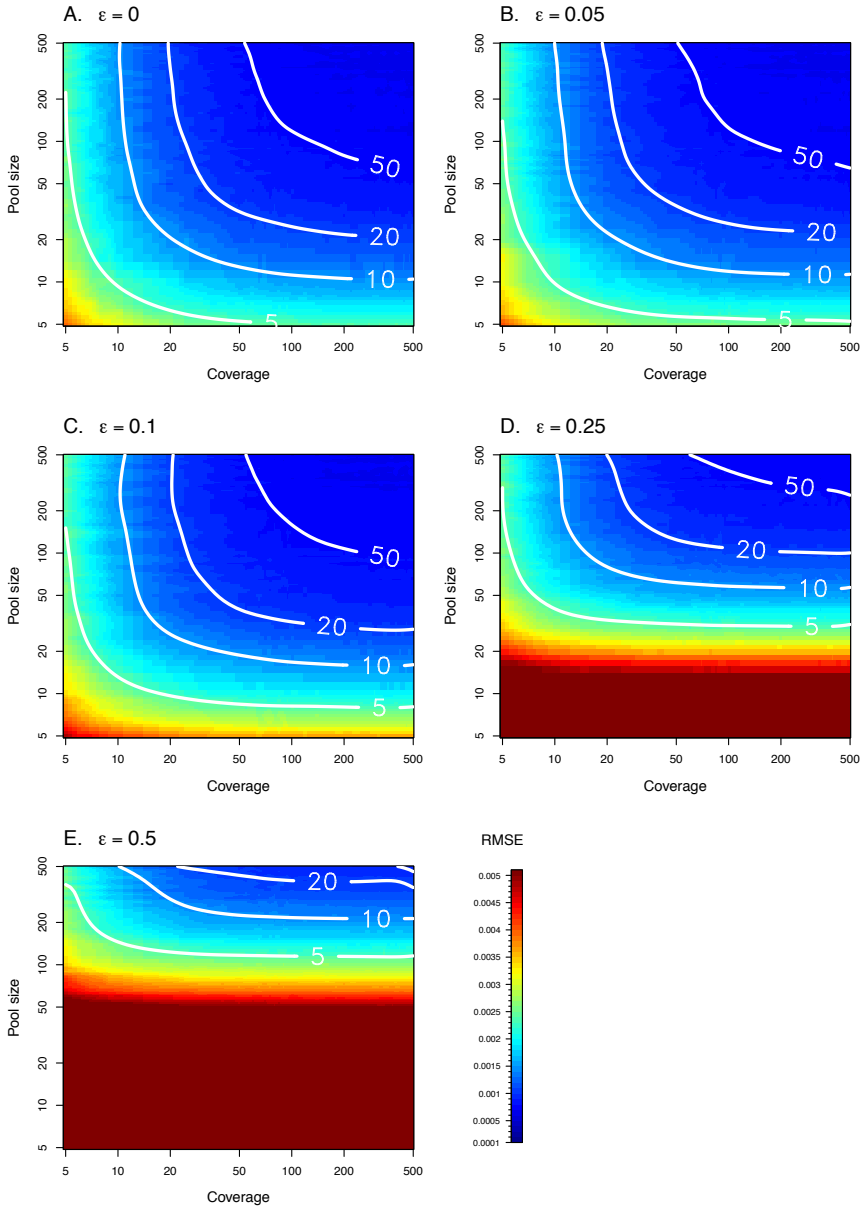


Figure S2 Precision and accuracy of our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ as a function of pool size and coverage, with experimental error rate ϵ varying from 0 to 0.5 (A–E). Each density plot, which represents the root mean squared error (RMSE) of the estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$, was obtained using simple linear interpolation from a set of 44×44 pairs of pool size and coverage values. For each pool size and coverage, 500 replicates of 5,000 markers were simulated from an island model with $n_{\mathrm{d}} = 8$ demes. Plain white isolines represent the RMSE of the WC_{84} estimator computed from Ind-seq data, for various sample sizes ($n =$ 5, 10, 20, and 50). Each isoline was fitted using a thin plate spline regression with smoothing parameter $\lambda = 0.005$, implemented in the `fields` package for R.

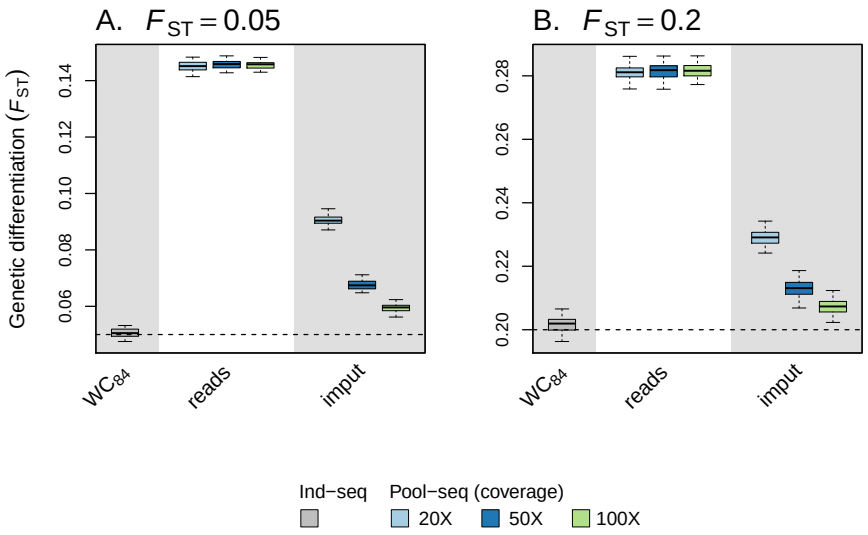


Figure S3 Precision and accuracy of naive estimators of F_{ST} for Pool-seq data. We compare WC_{84} estimates computed from allele count data inferred from individual genotypes (Ind-seq) to WC_{84} estimates computed: (*i*) directly from read counts, as if they were allele counts ("reads"); (*ii*) from allele counts imputed by maximum-likelihood ("imput"). Each boxplot represents the distribution of multilocus F_{ST} estimates across all demes in an 8-island model, and over 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to $F_{\mathrm{ST}} = 0.05$ (A) and $F_{\mathrm{ST}} = 0.20$ (B). The size of each pool was fixed to 10. For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST}.

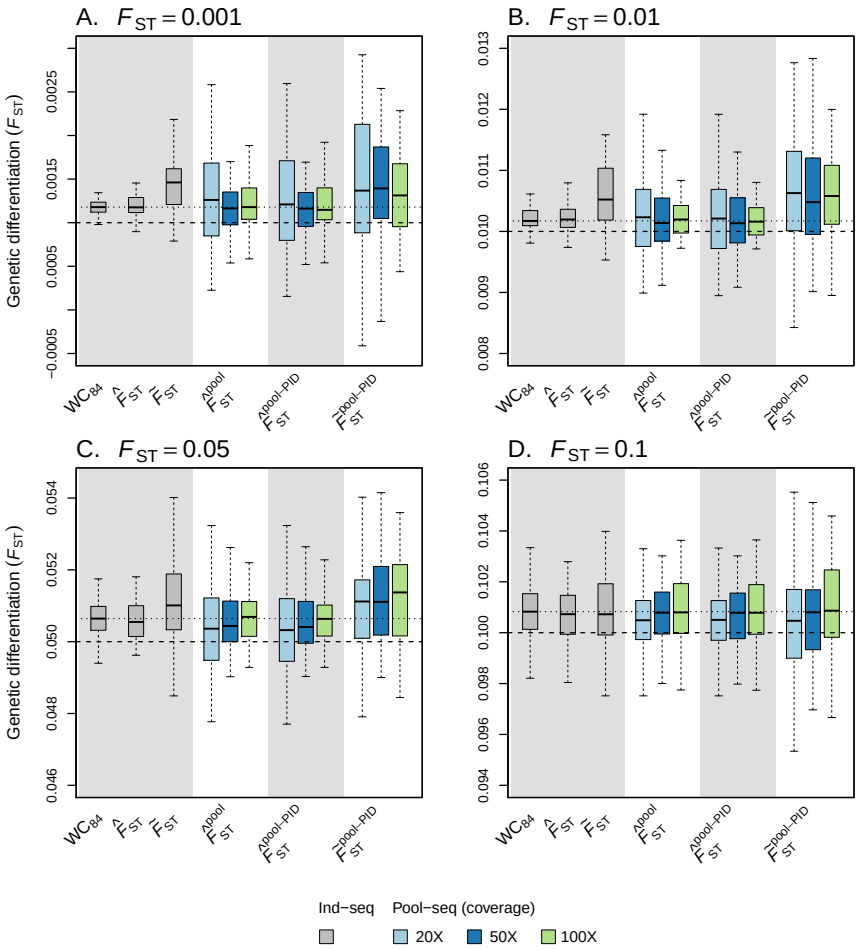


Figure S4 Precision and accuracy of alternative estimators of F_{ST} with varying pool size, for various levels of differentiation (A–D). The haploid pool size n drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; n was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. We considered three estimators based on allele count data inferred from individual genotypes (Ind-seq): WC_{84}, $\hat{F}_{\mathrm{ST}} \equiv \left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$ (where $\hat{Q}_1$ and $\hat{Q}_2$ are the weighted frequencies of identical pairs of genes within and between subpopulations, respectively, with weights equal to the number of pairs of genes) and $\tilde{F}_{\mathrm{ST}} \equiv \left(\tilde{Q}_1 - \tilde{Q}_2\right) / \left(1 - \tilde{Q}_2\right)$ (where $\tilde{Q}_1$ and $\tilde{Q}_2$ are the unweighted frequencies of identical pairs of genes within and between subpopulations, respectively). For Pool-seq data, we considered the estimators $\hat{F}^{\mathrm{pool}}_{\mathrm{ST}}$ (Equation 12), $\hat{F}^{\mathrm{pool-PID}}_{\mathrm{ST}}$ (Equation A44) and $\tilde{F}^{\mathrm{pool-PID}}_{\mathrm{ST}}$ (Equation A45). Each boxplot represents the distribution of multilocus F_{ST} over 50 independent replicates of the `ms` simulations. For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of WC_{84} estimates.

Annexe B

Un modèle hiérarchique bayésien pour détecter l'adaptation locale à partir de données haplotypiques

We provide here the computational details for the componentwise Markov chain Monte Carlo updates. Our aim is to sample from the joint posterior distribution of $f(\lambda, \delta, \sigma, \kappa, \pi, M|n)$, which is specified by the directed acyclic graph (DAG) in Figure 2.5. To do so, we use a combination of Metropolis–Hastings algorithm and Gibbs sampler to generate observations from $f(\lambda, \delta, \sigma, \kappa, \pi, M|n)$ using outputs from a Markov chain. Each Markov chain is initialized with random values of the parameters drawn from their prior densities, except for the parameters p_{ijk}, (use of the observed frequencies) and the parameters π_{jk}, (Laplace values are calculated from the dataset frequencies). The updating sequence is as follow :

— all $J \times I \times \sum_{j=1}^{J} K_j$ parameters p_{ijk}
— all I parameters M_i
— all $J \times \sum_{j=1}^{J} K_j$ parameters π_{jk}
— the hyperparameter λ
— all J hyperparameters δ_j
— all $J \times I$ parameters σ_{ij}
— all $J \times I$ parameters κ_{ij}

Since the full posterior distribution of the model can be decomposed as a product over loci and over populations, each update only requires the recomputation of the relevant terms of the distribution $f(\lambda, \delta, \sigma, \kappa, \pi, M|n)$. This considerably improves the computational efficiency of the algorithm.

B.1 Parameters Update

B.1.1 $\mathbf{p_{ijk}}$

The parameters $\mathbf{p_{ijk}}$ are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, the new allele frequencies $\mathbf{p'_{ijk}}$ are chosen as random variables drawn from a dirichlet distribution around the current values p_{ijk} with k the allele :

$$q(\mathbf{p}_{ij} \rightarrow \mathbf{p}'_{ij}) = \mathrm{Dir}(\mathbf{p}_{ij} \times \alpha)$$

$q(\mathbf{p}_{ij} \to \mathbf{p}'_{ij}) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K_j} \Gamma(\mathbf{p}_{ijk} \times \alpha)} \prod_{k=1}^{K_j} p'^{\mathbf{p}_{ijk} \times \alpha - 1}_{ijk}$

Log Scale :
$q(\mathbf{p}_{ij} \to \mathbf{p}'_{ij}) = \log \Gamma(\alpha) + \sum_{k=1}^{K_j} [(\mathbf{p}_{ijk} \times \alpha - 1) \log(\mathbf{p}'_{ijk}) - \log \Gamma(\mathbf{p}_{ijk} \times \alpha)]$

We calculate the forward (pfwd) and backward (pbwd) probabilities to compute the Hasting terms.

The variance scale parameter α is a constant adjusted during 30 short pilot runs of 500 iterations, in order to get an acceptance rate between 0.25 and 0.40. As the Dirichlet jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (due to the assymetry of the proposal distribution), their probability of acceptance proportionately decreases. In order to avoid computational problems with excessively small or large p_{ijk}, all moves outside the interval $[\epsilon; 1 - \epsilon]$ with $\epsilon = 10^{-8}$ are fixed to the closest boundary, a normalization is thus require after the proposal step. The proposed values $\mathbf{p}'_{\mathbf{ij}.}$ are accepted according to the appropriate Metropolis-Hastings probability, which is :

$1 \wedge \frac{\mathcal{L}(p'_{ij}, \mathbf{n_{ij}}) \psi(\mathbf{p}'_{\mathbf{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij})}{\mathcal{L}(p_{ij}, \mathbf{n_{ij}}) \psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij})} \times \frac{q(\mathbf{p}'_{ij} \to \mathbf{p}_{ij})}{q(\mathbf{p}_{ij} \to \mathbf{p}'_{ij})}$

It can be rewritten as :
$1 \wedge \exp(\sigma_{ij}(\tilde{p'_{ij}} - \tilde{p_{ij}})) \prod_{k=1}^{K_j} [(\frac{p'_{ijk}}{p_{ijk}})^{x_{ijk} + M_i \pi_{jk} - 1}]$

Log scale :

$\log(1) \wedge \sigma_{ij}(\tilde{p'_{ij}} - \tilde{p_{ij}}) + \sum_{k=1}^{K_j} [(x_{ijk} + M_i \pi_{jk} - 1)(\log(p'_{ijk}) - \log(p_{ijk}))] + (\log(pbwd) - \log(pfwd))$

Simulation Step :
1) Generate proposal from dirichlet sampling

2) Compute pfwd and pbwd
3) Correct proposal for boundaries
4) Compute the Metropolis-Hasting ratio

B.1.2 M_i

The parameters M_i are updated iteratively, one deme at a time. The proposed value M_i' is drawn from a lognormal distribution with a median equal to the current value M_i, i.e :

$$q(M_i \rightarrow M_i') = \frac{1}{M_i' \nu_M \sqrt{2\pi}} \exp(\frac{-ln(M_i'/Mi)^2}{2\nu_M^2})$$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant adjusted during 25 short pilot runs of 500 iterations, in order to get acceptance rates between 0.25 and 0.40. As the lognormal jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (due to the assymetry of the proposal distribution), their probability of acceptance proportionately decreases. Here, the ratio $\frac{q(M_i' \rightarrow M_i)}{q(M_i \rightarrow M_i')} = \frac{M_i'}{M_i}$ reduces to $\frac{M_i'}{M_i}$. In order to avoid computational problems with excessively small or large M_i values, all moves falling outside the interval $[10^{-3}; 10^3]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value M_i' is accepted according to the appropriate Metropolis-Hastings probability, which is :

$$1 \wedge \frac{\prod_{j=1}^{L} \psi(\mathbf{p_{ij}}, M_i', \pi_j, \sigma_{ij}, \kappa_{ij}) f(M_i')}{\prod_{j=1}^{L} \psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij}) f(M_i)} \times \frac{q(M_i' \rightarrow M_i)}{q(M_i \rightarrow M_i')}$$

It can be rewritten as :

$$1 \wedge [\frac{\Gamma(M_i')}{\Gamma(M_i)}]^L \prod_{j=1}^{L} [\frac{{}_1F_1(M_i \pi_{ij}^{\cdot}, M_i, \sigma_{ij})}{{}_1F_1(M_i' \pi_{ij}^{\cdot}, M_i', \sigma_{ij})} \prod_{k=1}^{K_j} \frac{\Gamma(M_i \pi_{jk})}{\Gamma(M_i' \pi_{jk})} p_{ijk}^{\pi_{jk}(M_i' - M_i)}]]$$

Log scale :

$\log(1) \wedge J(log\Gamma(M_i')-log\Gamma(M_i))+\sum_{j=1}^{L}[log({}_1F_1(M_i'\tilde{\pi}_j, M_i', \sigma_{ij}))-log({}_1F_1(M_i\tilde{\pi}_j, M_i, \sigma_{ij}))+ \sum_{k=1}^{K}(log\Gamma(M_i\pi_{jk}) - log\Gamma(M_i'\pi_{jk}) + \pi_{jk}(M_i' - M_i)log(p_{ijk}))]$

B.1.3 $\pi_{\mathbf{j}}$

The parameters $\pi_{\mathbf{jk}}$ are updated iteratively, one locus at a time. In the ith deme, at locus j, the news allelic frequencies $\pi'_{\mathbf{jk}}$ are chosen as a random variable drawn from a dirichlet distribution around the currents values π_{jk} with k the allele :

$q(\pi_{\mathbf{j}} \rightarrow \pi'_{\mathbf{j}}) = \mathrm{Dir}(\pi_{\mathbf{j}} \times \alpha)$

The variance scale parameter α is a constant adjusted during 25 short pilot runs of 500 iterations, in order to get acceptance rates between 0.25 and 0.40. As the Dirichlet jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (due to the assymetry of the proposal distribution), their probability of acceptance proportionately decreases. In order to avoid computational problems with excessively small or large π_{jk}, all proposal outside the interval $[\epsilon; 1-\epsilon]$ with $\epsilon = 10^{-8}$ are fixed to the closest boundary, a normalization is thus require after the proposal step. The proposed value $\pi'_{\mathbf{j}}$ is accepted according to the appropriate Metropolis-Hastings probability, which is :

$1 \wedge \frac{\prod_{i=1}^{n_d} \psi(\mathbf{p_{ij}}, M_i, \pi'_j, \sigma_{ij}, \kappa_{ij}) f(\pi'_j)}{\prod_{i=1}^{n_d} \psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij}) f(\pi_j)} \times \frac{q(\pi'_{\mathbf{j}} \rightarrow \pi_{\mathbf{j}})}{q(\pi_{\mathbf{j}} \rightarrow \pi'_{\mathbf{j}})}$

It can be rewritten as :

$1 \wedge \prod_{i=1}^{nd} [\frac{{}_1F_1(M_i\tilde{\pi_j}, M_i, \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi'_j}, M_i, \sigma_{ij})} \prod_{k=1}^{K_j} [\frac{\Gamma(M_i\pi_{jk})}{\Gamma(M_i\pi'_{jk})} p_{ijk}^{M_i(\pi'_{jk}-\pi_{jk})}]] \times \frac{pbwd}{pfwd}$

Log scale :

$\log(1) \wedge \sum_{i=1}^{nd} [\log({}_1F_1(M_i\tilde{\pi_j}, M_i, \sigma_{ij})) - \log({}_1F_1(M_i\tilde{\pi'_j}, M_i, \sigma_{ij})) + \sum_{k=1}^{K} [\log\Gamma(M_i\pi_{jk}) - \log\Gamma(M_i\pi'_{jk}) + M_i(\pi'_{jk} - \pi_{jk})\log(p_{ijk})]] + (\log(pbwd) - \log(pfwd))$

Simulation Step :
1) Generate proposal from dirichlet sampling
2) Compute pfwd and pbwd
3) Correct proposal for boundaries
4) Compute the Metropolis-Hasting ratio

B.1.4 κ_{ij}

The parameters κ_{ij} are updated iteratively in each deme, one locus at a time. in the ith deme, at locus j, the variable κ_{ij}, which indicates which of the two alleles is selected for, is updated using Gibbs sampling based on the conditional posterior distribution :

$f(\kappa_{ij}|\theta_{[-\kappa_{ij}]}) \propto \psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij}) f(\kappa_{ij})$

κ_{ij} can take K_j values, we have :

$Pr(\kappa_{ij} = k|\theta_{[-\kappa_{ij}]}) \propto \frac{1}{K} \times \psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij}) f(\kappa_{ij})$

It can be rewritten as (ρ) where ρ is the Pr=k vector.

Metropolis-Hasting sampling :

The parameters κ_{ij} are updated iteratively in each deme, one locus at a time. The proposed value κ_{ij} is sample from a discret uniform distribution with interval $[1; K_j]$ with step equal to 1 and K_j the number of allele at loci

j. This proposal is symetric. The proposed value κ'_{ij} is accepted according to the appropriate Metropolis-Hastings probability, which is :

$$1 \wedge \frac{\psi(\mathbf{p_{ij}}, M_i, \sigma_{\mathbf{ij}}, \kappa'_{\mathbf{ij}}, \pi_{\mathbf{j}}) f(\kappa'_{\mathbf{ij}})}{\psi(\mathbf{p_{ij}}, M_i, \sigma_{\mathbf{ij}}, \kappa_{\mathbf{ij}}, \pi_{\mathbf{j}}) f(\kappa_{\mathbf{ij}})}$$

can be rewritten as :

$$1 \wedge \frac{{}_1F_1(M_i\tilde{\pi'_{\mathbf{ij}}}, M_i, \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi_{\mathbf{ij}}}, M_i, \sigma_{ij})} \exp(\sigma_{ij}(\tilde{\mathbf{p'_{ij}}} - \tilde{\mathbf{p_{ij}}}))$$

Log scale :

$$\log(1) \wedge \log({}_1F_1(M_i\tilde{\pi'_{ij}}, M_i, \sigma_{ij})) - \log({}_1F_1(M_i\tilde{\pi_{ij}}, M_i, \sigma_{ij})) + \sigma_{ij}(\tilde{p'_{ij}} - \tilde{p_{ij}})$$

B.1.5 σ_{ij}

The parameters σ_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme at locus j, the proposed valueof the parameters σ'_{ij} is drawn from a lognormal distribution with median equal to the current value σ_{ij},i.e. :

$q(\sigma_{ij} \to \sigma'_{ij}) = \frac{1}{\sigma'_{ij}\nu_{\sigma_{ij}}\sqrt{2\pi}} exp(\frac{-ln(\sigma'_{ij}/\sigma_{ij})^2}{2\nu^2_{\sigma_{ij}}})$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. As the Lognormal jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (due to the assymetry of the proposal distribution), their probability of acceptance proportionately decreases . Here, the ratio $\frac{q(\sigma'_{ij} \to \sigma_{ij})}{q(\sigma_{ij} \to \sigma'_{ij})} = \frac{\sigma'_{ij}}{\sigma_{ij}}$ reduces to $\frac{\sigma'_{ij}}{\sigma_{ij}}$. In order to avoid computational problems with excessively small or large σ_{ij} values, all moves falling outside the interval $[0; 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value

σ'_{ij} is accepted according to the appropriate Metropolis-Hastings probability, which is :

$$1 \wedge \frac{\psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma'_{ij}, \kappa_{ij}) f(\kappa_{ij}) f(\sigma'_{ij}|\delta_j)}{\psi(\mathbf{p_{ij}}, M_i, \pi_j, \sigma_{ij}, \kappa_{ij}) f(\kappa_{ij}) f(\sigma_{ij}|\delta_j)} \times \frac{q(\sigma'_{ij} \to \sigma_{ij})}{q(\sigma_{ij} \to \sigma'_{ij})}$$

It can be rewritten as :

$$\frac{\sigma'_{ij}}{\sigma_{ij}} \exp[(\sigma'_{ij} - \sigma_{ij})(\tilde{p_{ij}} - \frac{1}{\delta_j})] \frac{{}_1F_1(M_i\tilde{\pi}_i, M_i, \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_i, M_i, \sigma'_{ij})}$$

Log scale :

$$\log(1) \wedge log(\sigma'_{ij}) - \log(\sigma_{ij}) + \log({}_1F_1(M_i\tilde{\pi_{ij}}, M_i, \sigma_{ij})) - \log({}_1F_1(M_i\tilde{\pi_{ij}}, M_i, \sigma'_{ij})) + (\sigma'_{ij} - \sigma_{ij})(\tilde{p_{ij}} - \frac{1}{\delta_j})$$

B.1.6 δ_j

The parameters δ_j are updated iteratively, one locus at a time. The proposed value of the hyperparameters δ'_j is drawn from a lognormal distribution with median equal to the current value δ_j,i.e. :

$$q(\delta_j \to \delta'_j) = \frac{1}{\delta'_j \nu_\delta \sqrt{2\pi}} exp(\frac{-ln(\delta'_j/\delta_j)^2}{2\nu_\delta^2})$$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. As the Lognormal jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (due to the assymetry of the proposal distribution), their probability of acceptance proportionately decreases. Here, the ratio $\frac{q(\delta'_j \to \delta_j)}{q(\delta_j \to \delta'_j)} = \frac{\delta'_j}{\delta_j}$ reduces to $\frac{\delta'_j}{\delta_j}$. In order to avoid computational problems with excessively small or large δ_j values, all moves falling outside the interval $[0; 500]$ are discarded

(i.e., the chain is kept unchanged). Otherwise, the proposed value δ'_j is accepted according to the appropriate Metropolis-Hastings probability, which is :

$$1 \wedge \frac{[\prod_{i=1}^{nd} f(\sigma_{ij}|\delta'_j)]f(\delta'_j|\lambda)}{[\prod_{i=1}^{nd} f(\sigma_{ij}|\delta_j)]f(\delta_j|\lambda)} \times \frac{q(\delta'_j \to \delta_j)}{q(\delta_j \to \delta'_j)}$$

It can be rewritten as :

$$1 \wedge (\frac{\delta_j}{\delta'_j})^{n_d - 1} \exp[(\delta'_j - \delta_j)(\frac{\sum_{i=1}^{n_d} \sigma_{ij}}{\delta_j \delta'_j} - \frac{1}{\lambda})]$$

Log scale :

$$\log(1) \wedge (nd-1)(\log(\delta_j) - \log(\delta'_j) + (\delta'_j - \delta_j)(\frac{\sum_{i=1}^{nd} \sigma_{ij}}{\delta_j \delta'_j} - \frac{1}{\lambda})$$

B.1.7 λ

The proposed value of the hyperparameter λ' is drawn from a lognormal distribution with median equal to the current value λ,i.e. :

$$q(\lambda \to \lambda') = \frac{1}{\lambda' \nu_\lambda \sqrt{2\pi}} exp(\frac{-ln(\lambda'/\lambda)^2}{2\nu_\lambda^2})$$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. As the Lognormal jumping rule is asymmetric, a Metropolis-Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes reffered to as the "Hasting term" : see, e.g., Gelman et al. 2004). It means that when some moves are more likely to happen (because of the assymetry of the proposal distribution), their probability of acceptance proportionately decreases. Here, the ratio $\frac{q(\lambda' \to \lambda)}{q(\lambda \to \lambda')} = \frac{\lambda'}{\lambda}$ reduces to $\frac{\lambda'}{\lambda}$. In order to avoid computational problems with excessively small or large λ values, all moves falling outside the interval $[0; 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value

λ' is accepted according to the appropriate Metropolis-Hastings probability, which is :

$$1 \wedge \frac{[\prod_{j=1}^{L} f(\delta_j|\lambda')]f(\lambda'|\Lambda)}{[\prod_{j=1}^{L} f(\delta_j|\lambda)]f(\lambda|\Lambda)} \times \frac{q(\lambda' \to \lambda)}{q(\lambda \to \lambda')}$$

can be rewritten as :

$$1 \wedge (\frac{\lambda}{\lambda'})^{L-1} exp[(\lambda' - \lambda)(\frac{\sum_{j=1}^{L} \delta_j}{\lambda\lambda'} - \frac{1}{\Lambda})]$$

log scale :

$$log(1) \wedge (L-1)(log(\lambda) - log(\lambda') + (\lambda' - \lambda)(\frac{\sum_{j=1}^{L} \delta_j}{\lambda\lambda'} - \frac{1}{\Lambda})$$

B.2 PODs algorithm

In order to provide a decision criterion to discriminate between neutral and selected markers, we calibrate the Kullback-Leibler divergence (KLD) using simulations from a predictive distribution based on the observed data set. To that end, we generate pseudo-observed data as follows.

We set the hyperparameters M_i,π_{jk} and λ to their respective posterior means $\overline{M_i}$, $\overline{\pi_{jk}}$, $\overline{\lambda}$ as estimated from the MCMC. Then we draw δ_j from an exponential distribution $exp(\Lambda^{-1})$ and we draw σ_{ij} from an exponential distribution $\exp(\delta_j^{-1})$. Last, the parameter κ_{ij} is drawn from a Multinomial distribution (one sample with parameter the posterior frequencies of κ_{ijk}).

We aim at sampling the allele frequencies p_{ij} from the distribution with density $f(p_{ij})$ defined by equations 2 and 3 in the main text. As the cumulative distribution function of the distribution with density $f(p_{ij})$ is not tractable, we use a rejection-sampling algorithm. To that end, we define an instrumental distribution $g(p_{ijk})$ $\mathrm{Dir}(M_i\pi_{jk})$, with density :

$$g(p_{ijk}) = \frac{\Gamma(\sum(M_i\pi_{jk}))}{\prod\Gamma(M_i\pi_{jk})} p_{ijk}^{M_i\pi_{jk}-1}$$

We need to define a constant u, with $f(p_{ijk}) < [ug(p_{ijk})]$ over the support

[0,1]. Noting that :

$\frac{f(p_{ijk})}{g(p_{ijk})} = \frac{\exp(\sigma_{ij}\tilde{p}_{ijk})}{{}_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})}$

Then, if we define $u \equiv \exp(\sigma_{ij})/{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})$ we get :

$\frac{f(p_{ijk})}{ug(p_{ijk})} = \exp(\sigma_{ij}(\tilde{p}_{ij} - 1))$

Since $0 \leq \tilde{p}_{ij} \leq 1$ and $\sigma_{ij} \geq 0$, by definition, we have $\exp(\sigma_{ij}(\tilde{p}_{ij} - 1)) \leq 1$ and therefore $f(p_{ij}) \leq [ug(p_{ij})]$. A straightforward algorithm to sample from the distribution with density $f(p_{ij})$ is :

(i) Sample x from a Dirichlet distribution Dir($M_i\pi_{jk}$) and y from $U(0, 1)$ (the uniform distribution over the unit interval).

(ii) Check whether or not y<f(x)/[ug(x)] or equivalently if $log(y) < \sigma_{ij}(\tilde{p}_{ij} - 1)$:

— If this hold, accept x.

— If not, reject the values of x and repeat the sampling step (1).

Finally we draw the allele counts n_{ijk} in the ith deme at the jth locus and kth allele by a random draw from the multinomiale distribution $M(n_{ij}, p_{ijk})$. We repeat this procedure for each locus j in each deme i.

If this algorithm is computationnaly efficient for low and medium value of σ, it could become stuck for high values of σ due to a huge number of trials required until acceptance of a x. This number of trials is exactly the bounding constant $u \equiv \exp(\sigma_{ij})/{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})$. Therefore, we adopt an alternative algorithm in order to avoid this problem when $u > 10^4$:

we draw $x[\kappa]$ from a $Beta(\alpha, \beta)$ with the same first two moments as the target distribution.

$\alpha = m_1(m_2 - m_1)/(m_1^2 - m_2)$ and $\beta = \alpha(1/m_1 - 1)$, where
$m_1 = \tilde{\pi}_{ij}(\frac{{}_1F_1(M_i\tilde{\pi}_{ij}+1;M_i+1;\sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})})$

$m_2 = \tilde{\pi}_{ij}\left(\frac{{}_1F_1(M_i\tilde{\pi}_{ij}+2;M_i+2;\sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})}\right)$

Then, we draw the others $x[k_{\text{for}k\neq\kappa}]$ from a flat dirichlet Dir($\overrightarrow{1}$). It is possible as $u > 10^4$ when there is only one allele at high frequencies and the others very rare.

Annexe C

Intégration d'une variable bayésienne auxiliaire et d'un modèle de lissage intégré

C.1 Parameters Update

C.1.1 Update of $\delta_j^\star$

If $\zeta_j = 1$, we perform as in the full model. If $\zeta_j = 0$, then we sample the new δ_j in the prior $\Gamma(1, \frac{1}{\lambda^\star})$.

C.1.2 Update of σ_{ij}

If $\zeta_j = 1$, we perform as in the full model. If $\zeta_j = 0$, then we sample the new values of σ_{ij} in the prior $(\Gamma(1, \frac{1}{\delta_j^\star}))$ (we also shift the hypergeomteric terms to 0 in the log-likelihood)

C.1.3 Update of κ_{ij}

If $\zeta_j = 1$, we perform as in the full model. If $\zeta_j = 0$, then we sample the new values of κ_{ij} in a $Categorical(\frac{1}{K_j})$, with K_j the number of allele at locus j

C.1.4 Update of ζ_j (when $b_{Is} = 0$)

The full-conditional has an usual form $f(\zeta_j|.) \propto f(\zeta_j|P)(\prod_{i=1}^{I} f(\boldsymbol{\alpha}_{ij}|\sigma_{ij}, \kappa_{ij}))$. Moreover, because the variable is binary, it follows a a Bernouilli distribution. Indeed, we have :

(i) $f(\zeta_j = 0|.) = (1 - P)(\prod_{i=1}^{I} f(\alpha_{ij}|\sigma_{ij} = 0))$

(ii) $f(\zeta_j = 1|.) = P(\prod_{i=1}^{I} f(\alpha_{ij}|\sigma_{ij}, \kappa_{ij}))$

Hence :

$$(\zeta_j|.) \sim Ber(\pi = \frac{f(\zeta_j=1|.)}{f(\zeta_j=1|.)+f(\zeta_j=0|.)})$$

We denote that $\frac{1}{\pi} = 1 + \frac{f(\zeta_j=0|.)}{f(\zeta_j=1|.)} = 1 + \frac{(1-P)}{P} \prod_{i=1}^{I} \frac{{}_1F_1(M_i \tilde{\gamma}_j, M_i, \sigma_{ij})}{exp(\sigma_{ij} \tilde{p_{ij}})}$

C.1.5 Update of ζ_j (when $b_{Is} > 0$)

The full-conditional has an usual form $f(\zeta_j|.) \propto f(\zeta_j|P, bIs, \delta_{-j})(\prod_{i=1}^{I} f(\alpha_{ij}|\sigma_{ij}, \kappa_{ij}))$. Moreover, because the variable is binary, it goes to be a Bernouilli distribution. Indeed, we have :

(i) $f(\zeta_j = 0|.) = (1-P)e^{b_{Is}\eta_0}(\prod_{i=1}^{I} f(\boldsymbol{\alpha}_{ij}|\sigma_{ij} = 0))$

(ii) $f(\zeta_j = 1|.) = P \exp^{b_{Is}\eta_1}(\prod_{i=1}^{I} f(\boldsymbol{\alpha}_{ij}|\sigma_{ij}, \kappa_{ij}))$

with b_{Is} the inverse Ising temperature (see Ising and Potts litterature), and η_1 (respectively η_0) the number of neighboors pairs in the same state ($\mathbb{I}_{\zeta_j=\zeta_j-1} + \mathbb{I}_{\zeta_j=\zeta_j+1}$, with $\mathbb{I}$ the identity function), so associated with selection state (respectively neutral)

Hence :

$$(\zeta_j|.) \sim Ber(\pi = \tfrac{f(\zeta_j=1|.)}{f(\zeta_j=1|.)+f(\zeta_j=0|.)})$$

We denote that $\frac{1}{\pi} = 1 + \frac{f(\zeta_j=0|.)}{f(\zeta_j=1|.)} = 1 + \frac{(1-P)}{P} \exp^{b_{Is}\eta_0-\eta_1} \prod_{i=1}^{I} \frac{{}_1F_1(M_i\tilde{\pi}_j, M_i, \sigma_{ij})}{exp(\sigma_{ij}\tilde{p_{ij}})}$

C.1.6 Update of P

The full conditional distribution of P is a Beta distribution allowing a simple Gibbs update. Indeed we have :

$f(P|.) \propto f(P)f(\zeta|P)$ $f(P|.) \propto P^{ap-1}(1-P)^{bp-1}P^{\sum_{j=1}^{J}\zeta_j}(1-P)^{J-\sum_{j=1}^{J}\zeta_j}$
$f(P|.) \propto P^{ap-1+\sum_{j=1}^{J}\zeta_j}(1-P)^{bp-1+J-\sum_{j=1}^{J}\zeta_j}$

Hence :

$P|. \propto Beta(ap + \sum_{j=1}^{J}\zeta_j, bp + J - \sum_{j=1}^{J}\zeta_j)$

with $ap = 0.2$ and $bp = 1.8$.

Résumé :

L'avancée des technologies de séquençage et de génotypage à haut-débit permet la comparaison de patrons de polymorphisme à un très grand nombre de marqueurs génétiques. L'analyse de la différenciation des populations à une échelle génomique rend ainsi possible la recherche de régions génomiques impliquées dans l'adaptation locale des organismes à leur environnement. Dans cette thèse, nous avons suivi deux approches complémentaires pour caractériser la différenciation génétique à partir de données de génotypage à haut-débit. Dans un premier temps, nous avons développé un estimateur non-biaisé du paramètre F_{ST} pour des données de génotypage d'individus en mélange (*Pool-seq*). La construction de cet estimateur, dans un contexte d'analyse de variance, a nécessité de bien prendre en compte les différentes étapes de l'échantillonnage : des gènes dans le mélange d'individus et des lectures de séquençage parmi les gènes. Nous montrons qu'il surpasse les estimateurs utilisés jusqu'à présent. Dans un deuxième temps, nous avons développé une méthode d'analyse de la différenciation génétique à l'échelle du génome, dans le cadre d'un modèle bayésien hiérarchique, pour distinguer l'effet de la démographie de celui de la sélection. Pour cela, nous avons implémenté plusieurs extensions au modèle SelEstim, pour exploiter l'information de déséquilibre de liaison entre les marqueurs. Une première stratégie a consisté à analyser des données multialléliques, obtenues par le regroupement local de marqueurs SNPs en blocs d'haplotypes. Une stratégie alternative a consisté à intégrer un modèle de lissage prenant en compte la dépendance spatiale entre marqueurs adjacents. Cette approche repose sur l'analyse de données bialléliques, ce qui la rend applicable à la fois à des données de génotypage individuel et à des données *Pool-seq*. Nous discutons, sur la base de l'analyse de jeux de données simulées, des mérites relatifs de ces différentes approches.

Mots-clés : Génomique des populations, *Pool-seq*, inférence statistique, génétique de l'adaptation, modèles bayésiens hiérarchiques, analyse de variance

---

Abstract :

The advent of high throughput sequencing and genotyping technologies allows the comparison of patterns of polymorphisms at a very large number of genetic markers. The analysis of genetic differentiation between populations at a whole-genome scale makes it possible to characterize genomic regions involved in the local adaptation of organisms to their environment. In this thesis, we followed two complementary approaches to characterize differentiation from high-throughput genotyping data. First, we developed an unbiased estimator of the parameter F_{ST} for individuals sequenced in pools (*Pool-seq*). Deriving this estimator, in an analysis-of-variance framework, required to properly account for the different sampling steps : individual genes from the pool, and sequence reads from these genes. We show that it outperforms previously proposed estimators. Second, we developed a method to analyze genetic differentiation at a whole-genome scale in a hierarchical bayesian framework, in order to untangle the effect of demography from that of selection. To this end, we implemented different extensions to the SelEstim model, aimed at leveraging the information from linkage disequilibrium between markers. A first approach consisted in analyzing multiallelic data derived from the local clustering of SNPs into haplotype blocks. An alternative strategy consisted in including a smoothing model, which accounts for the spatial dependency between neighboring markers. This strategy relies on the analysis of biallelic data, and can be used both with individual genotype data or *Pool-seq* data. We discuss the relative benefits of these different approaches, based on the analysis of simulated data sets.

Keywords : Population genomics, Pool-seq, statistical inference, adaptation genetics, hierarchical bayesian modeling, analysis-of-variance