# Génomique des populations appliquée : détection de signatures de sélection au sein de populations expérimentales

Jean-Noël Hubert

université PARIS-SACLAY

NNT : 2018SACLS141

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

# Génomique des populations appliquée : détection de signatures de sélection au sein de populations expérimentales

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°581 Agriculture, alimentation, biologie, environnement et santé (ABIES)
Spécialité de doctorat: Sciences de la Vie et de la Santé

Thèse présentée et soutenue à Paris, le 21 juin 2018, par

**M. Jean-Noël Hubert**

Composition du Jury :

| | |
|---|---|
| M. Pierre Boursot<br>Directeur de Recherche, CNRS – ISEM, Montpellier | Président |
| Mme Hélène Gilbert<br>Directrice de Recherche, INRA – UMR GenPhySE , Toulouse | Rapporteur |
| M. Lluis Quintana-Murci<br>Directeur de Recherche, CNRS – Institut Pasteur , Paris | Rapporteur |
| Mme Laurène Gay<br>Chargée de Recherche, INRA – UMR AGAP, Montpellier | Examinateur |
| M. Thomas Heams<br>Maitre de Conférences, AgroParisTech – UMR GABI, Paris | Examinateur |
| M. Frédéric Hospital<br>Directeur de Recherche, INRA – UMR GABI , Jouy en Josas | Directeur de thèse |

# Remerciements

« *Traitez les gens comme s'ils étaient ce qu'ils devraient être, et vous les aiderez ainsi à devenir ce qu'ils peuvent être.* »

Johann Wolfgang Goethe

S'il adhère implicitement à un contrat méthodologique rigoureux, tout chercheur évolue dans plusieurs communautés. Evidemment, il est d'abord membre d'une petite communauté de spécialistes qui permet la validation et la diffusion du travail scientifique. Mais il fait aussi partie d'autres sphères – interdisciplinaires, sociales, amicales, familiales, etc. – qui lui permettent d'avoir un équilibre de vie et alimentent sa réflexion, influençant y compris ses développements scientifiques futurs. J'ai personnellement pu compter sur la bienveillance, la sagesse et le soutien d'un large entourage, qui m'a permis d'avoir un parcours plus riche que je n'aurais jamais pu l'imaginer (et dont je m'étonne encore aujourd'hui !). Je souhaite remercier toutes les personnes qui m'ont soutenu et qui m'ont aidé à accomplir ce long chemin.

J'aimerais en premier lieu remercier les membres de mon jury de thèse pour avoir consacré de leur temps à l'évaluation de mon travail. Merci de m'avoir accompagné dans cette ultime ligne droite de la thèse !

Je remercie Pierre Boursot d'avoir accepté de présider ce jury, et de l'avoir fait avec le sérieux, la justesse et la bienveillance qui le caractérisent.

Je remercie Hélène Gilbert et Lluis Quintana-Murci d'avoir accepté avec enthousiasme d'être mes rapporteurs. Je ne pouvais imaginer un duo plus stimulant et plus complémentaire pour juger de façon approfondie ce travail. Merci pour vos rapports écrits qui témoignent d'un intérêt sincère pour le travail accompli durant cette thèse.

Je remercie Laurène Gay qui a accepté d'examiner cette thèse et de mettre sa fine connaissance de la génomique des populations appliquée au service du jury, et ce sans pour autant mettre mal à l'aise le petit nouveau que je suis dans ce milieu.

J'adresse mes plus sincères remerciements à Thomas Heams. En plus de m'avoir fait l'honneur de participer à mon jury de thèse, Thomas a été un personnage-clé dans ma progression : il m'a accueilli à AgroParisTech et a grandement facilité mes premiers contacts avec les généticiens de ce qui devint finalement mon laboratoire d'accueil pour la thèse. Son engagement pour l'enseignement et la diffusion du savoir scientifique au-delà des frontières disciplinaires a été pour moi une inspiration, et son soutien m'a été précieux. Merci beaucoup Thomas !

J'aimerais exprimer à Fred, qui a encadré cette thèse, ma profonde reconnaissance. Je le remercie en premier lieu de m'avoir mis sur ce passionnant sujet et de m'avoir fait confiance malgré mon inexpérience. Lors de notre première rencontre, j'arrivais sur la pointe des pieds et il m'a très

chaleureusement accueilli dans le monde de la génétique des populations. Cette thèse a été faite de hauts et de bas ; je crois en tout cas que le manuscrit reflète un travail original – dans tous les sens de ce terme – et j'espère qu'il le satisfait. Je remercie Fred de m'avoir sincèrement soutenu, guidé, fait partager ses intuitions avec justesse et transmis ses conseils en toute franchise. Je pense tout simplement que je n'aurais pas pu avoir un autre directeur de thèse que lui ! J'associe évidemment Isabelle Goldringer à ces remerciements, car elle aussi m'a soutenu et accordé sa confiance. Je souhaite à Adèle (merci pour ta présence à la soutenance ; j'ai aussi pu apprécier ton super coup de crayon ! ☺) et Mayeul beaucoup de réussite et de moments de bonheur, depuis le Perche jusqu'en Amérique du sud !

Je remercie les membres de l'équipe de choc qui, aux côtés de Fred, a constitué mon comité de thèse : Joëlle Ronfort, Lounès Chikhi, Simon Boitard et Mathieu Gautier. Je suis tout d'abord honoré d'avoir pu compter sur le soutien d'une équipe de généticiens aussi éminents et je leur suis infiniment reconnaissant de m'avoir régulièrement consacré de leur temps pour me guider et me conseiller. Nos échanges ont toujours été riches en enseignements et m'ont permis de prendre confiance en moi. Merci !

Je souhaite aussi remercier tout particulièrement mon école doctorale de rattachement, ABIES, qui m'a accompagné avec sérieux et bienveillance depuis le jour de mon audition au concours. J'adresse notamment mes plus sincères remerciements à Irina Vassileva, Christine Duvaux-Ponter, Corinne Fiers, Sylvie Ponsonnet, Cyril Kao et Alexandre Péry. Je mesure la chance que j'ai eue de côtoyer une équipe si fiable et proche des doctorants. J'aimerais associer à ces remerciements Armel Guyonvarch, qui m'a accueilli à l'Université Paris-Sud, Bruno Bost, qui m'a soutenu sans réserve et m'a permis d'obtenir un complément de financement, et Valérie Le Port, qui m'a permis de déposer ce manuscrit dans de bonnes conditions.

J'ai beaucoup apprécié le site de Jouy-en-Josas, où s'est déroulé la quasi-totalité de mon travail, tant pour son cadre préservé que pour les personnes que j'y ai rencontrées. Je remercie l'ensemble des habitués du bâtiment 211, où j'ai rencontré des gens humbles, passionnés et accessibles.

A tout seigneur, tout honneur : commençons par Denis « *de Mesmaeker* » Laloë qui, en plus d'être un animateur d'équipe présent et bienveillant, a le bon goût d'aimer la BD franco-belge (et de le revendiquer !). Les réunions PSGen vont finalement me manquer (et je ne dirai rien sur la petite bière du mercredi, chuuut ! ☺).

Je remercie Tatiana Zerjal pour sa présence, ses grandes qualités scientifiques et humaines. Nous avons finalement partagé beaucoup de choses ; j'ai beaucoup appris à son contact et je lui souhaite de connaître une réussite totale dans ses projets professionnels (avec j'espère pour très bientôt un lauréat au concours ABIES !) et personnels. Je vais faire des jaloux en haut lieu à PSGen mais j'avoue avoir eu la chance de goûter la *pasta* préparée par Tatiana *di persona* et c'était évidemment un délice ! Je transmets en cette occasion mes meilleures pensées à Catarina et Nicolas.

Au sein de l'équipe PSGen, j'ai eu la chance de côtoyer des scientifiques expérimentés soucieux de transmettre aux plus jeunes malgré les multiples sollicitations dont ils font l'objet. Je pense en particulier à Michèle Tixier-Boichard, Xavier Rognon et Etienne Verrier, au contact desquels j'ai beaucoup appris. Malgré son implication à tous les étages de notre communauté de travail (et même au-delà : je suspecte là aussi un intérêt marqué pour le 9ème art ; à croire qu'il s'agit d'un prérequis pour l'exercice des plus hautes fonctions au sein de PSGen ☺), Etienne a trouvé le temps de me donner des conseils lorsque je le sollicitais et m'a toujours soutenu dans mon parcours. Merci beaucoup !

Je remercie Maria Bernard pour sa bonne humeur, sa réactivité, sa patience (que dis-je, son opiniâtreté !) dans le cadre du projet EFFICACE, et pour avoir accepté de partager avec moi quelques-unes de ses connaissances en bioinformatique (et aussi pour sa capacité à supporter Mathieu C. au quotidien… Quel mérite ! ☺).

Je remercie Andrea Rau de nous faire partager ses grandes qualités scientifiques et humaines. Comme elle a déjà tout ou presque (Américaine, elle s'exprime mieux en français que beaucoup de francophones natifs, est déjà HDR, etc.), je lui souhaite d'avoir un chien noir avec une liste blanche et des taches feu sur le museau et les pattes. ☺

Même si ça me coûte (☺), je remercie Gwendal Restoux… qui est tout simplement admirable de dévouement à son métier (malgré quelques passions inavouables, *the dark side of the* roux…). Je lui souhaite un succès professionnel à la mesure de son investissement.

Je remercie Mathieu Charles, que j'aurais aimé côtoyer davantage malgré des comportements parfois suspects ☺ [1]. Je crois que l'équipe a beaucoup de chance de compter dans ses rangs quelqu'un d'aussi agréable et compétent.

Je remercie Agathe Vieaud, Dominique Montagu (merci pour le livre !), Eléonore Charvolin-Lemaire, Florence Jaffrezic, Marco Moroldo, Wendy Brand-Williams pour tous les agréables moments passés en leur compagnie !

Je suis aussi redevable à des chercheurs du labo issus d'autres équipes, et en particulier de la très estimée équipe GenAqua ! Je remercie ainsi René Guyomard, qui a accepté que je travaille sur les carpes malgaches, Marc Vandeputte et Edwige Quillet, tous deux très investis dans la progression et la réussite des jeunes chercheurs, pour m'avoir aidé, fait partager leurs idées et une partie de leur immense expérience.

Je remercie Déborah Jardet et Bertrand Bed'hom qui m'ont très gentiment proposé leur aide pour l'impression de ce manuscrit.

Je remercie bien sûr Claire Rogel-Gaillard pour son investissement à la tête de l'UMR GABI, son soutien et son dynamisme contagieux !

Et GABI, c'est aussi et surtout des jeunes gens merveilleux qui font la vie scientifique quotidienne du labo : Adélie Tholance (j'espère que tu es sur un chemin qui te comble !), Alexis Michenet (promis, un jour je passerai mes vacances à *La Foulquetière* – et puis on ira voir Jeman avec Michel D. ☺), Anna-

Charlotte Doublet, Audrey Hulot, Belén Jiménez, Chris Hozé & Romain Saintilan, Clémentine Escouflaire, Dávid Jónás, Edin Hamzic, Gabriel Guillocheau & Thierry Heirman (qui ont tant aimé ce beau dîner en salle Apicula… ☺), Gilles Monneret, Iola Croué (qui sait concilier science et football, le tout à haut niveau !), Kenza Bazi-Kabbaj, Lenha Mobuchon, Marie Bérodier, Mélina Gallopin (#STGA2014 forever !), Parsaoran Silalahi, Pauline Michot, RooOOooxane Vallée ☺, Sébastien Taussat (La Grande-Motte, ça le botte !), Shizhi Wang, Sonia Eynard (je crois que l'affichette de ta « disparition » est toujours dans mon ex-bureau^^), Zih-Hua Fang et… *Mathieu Tiret* (s'il a autant d'amis c'est surtout grâce à son pouf rose qui est cool hein… ☺ Merci pour tout Mat' !). Je vous adresse à chacun un immense merci ! Ne changez pas, vous êtes déjà les meilleurs, j'espère qu'on aura l'occasion de se recroiser !

Je remercie Sonia Le Mentec et Emily Ruiz, qui ont été des étudiantes enthousiastes et patientes lors de leur stage sur nos thématiques.

Merci aux footeuses/-eux de l'INRA (j'ai beau avoir mis ma carrière entre parenthèses, je n'oublierai jamais ces petits matchs du jeudi qui nous offraient à tous une halte sportive méritée) : merci en particulier à Alexis, Aurélie Vinet & Sébastien Fritz, Elisandra Kern, Gabriel, Iola, Jean-Pierre Bidanel, Jehanne Mauxion, Marc Teissier, Mathieu T., Pascal Croiseau, Renaud Fleurot, Romain, Sylvain Marthey, Thierry pour ces moments d'efforts partagés (j'ai encore le souvenir d'oppositions épiques par -2°C dans la boue rouge du Val d'Enfer…).

Je remercie Hans Erhard, Nicolas Gilbert et Pierre Pudlo, que je n'ai pas eu l'occasion de recroiser durant mon doctorat, mais dont la sympathie et le soutien m'ont été précieux lors d'étapes importantes de mon parcours antérieur.

Je remercie mes proches et amis franciliens : Aude & Aurélien (et Madryn !), Christine & Manu, Françoise & Pascal (un grand merci pour votre appui dans nos recherches de logement !). Nous n'avons pas partagé assez de moments ensemble (et souvent par ma faute), mais ils furent toujours intenses et vrais. Merci !

Je remercie ma belle-famille, qui a été présente en nombre à ma soutenance. Merci Odile, merci Dominique, merci Amélie, et merci Guillaume ! Merci à Léna et à Daniel pour leurs messages de soutien. Votre présence à mes côtés me fait chaud au cœur !

Je remercie mon oncle Hubert et sa femme Anne-Marie pour leur soutien, leur gentillesse et pour m'avoir honoré de leur présence à ma soutenance. Merci aussi à Olivier, qui a eu la sagesse de prévoir son mariage pour après ma soutenance. ☺

J'ai une pensée émue en souvenir de mon oncle et parrain Yves qui nous a quittés avant le terme de ce parcours. Je remercie Gabrielle pour son soutien et lui transmets mes meilleures pensées.

Je remercie avec émotion ma marraine Jeanne. Avoir obtenu l'agrégation de sciences physiques à 21 ans est révélateur de capacités exceptionnelles que tu as, avec humanisme et rigueur, mises au service

de toutes les générations. Merci Jeanne pour ton soutien, pour m'avoir enseigné et fait partager tant de choses… Sans toi, je n'aurais tout simplement pas pu avoir ce parcours.

Je remercie mes parents, Régine et Marc, pour leur amour et pour l'indéfectible soutien qu'ils accordent à leurs enfants, quelque soient les choix de ces derniers. Merci pour votre présence toujours réconfortante, votre patience et votre empathie, vous êtes des personnes appréciées de tous ceux qui ont croisé votre chemin et des parents admirables ! J'associe à ces remerciements Isabelle & Antonin et Veronica & Guillaume, à qui je souhaite le bonheur de parcours personnels et professionnels riches et exaltants. Merci Guillaume d'être rentré à temps de Taïwan pour assister à ma soutenance ! ☺

Enfin, je remercie Anne-Laure. Comme un chapitre entier ne suffirait pas à rendre compte de l'ampleur de ton soutien, de ton amour et de ta confiance (sans compter que cela pourrait ennuyer les éventuels autres lecteurs ☺), j'aimerais simplement te rappeler ce moment symbolique du printemps 2011 où, près de Montpellier-Méditerranée, tu m'as donné ce qui est devenu à jamais « la photo du bonheur »…

A Baly…

[1] Robin, P. (1990). Getting to like the burn of chili pepper. Chemical Senses II. Irritation. 'New York'and Basel: Marcel Dekker, 23, l-269

# TABLE DES MATIERES

-

# Liste des tableaux



# Liste des figures

# Liste des annexes

# Liste des communications scientifiques

## Publications

Hubert JN, Zerjal T & Hospital F (submitted to *PLoS One*) Cancer- and behavior-related genes are targeted by selection in the Tasmanian devil (*Sarcophilus harrisii*).

Hubert JN & Hospital F (submitted to *Molecular Biology and Evolution*) The power to detect selection from short-term Evolve & Resequence experiments.

Hubert JN, Allal F, Hervet C et al. (2016) How could fully scaled carps appear in natural waters in Madagascar? Proc. R. Soc. B. 283:1837 DOI: 10.1098/rspb.2016.0945

## Communications lors de congrès internationaux

Hubert JN, Zerjal T & Hospital F (2018, accepted) DFTD-driven selection in the Tasmanian devil (Sarcophilus harrisii). Oral presentation at the second Joint Congress on Evolutionary Biology – Montpellier, France.

Hubert JN & Hospital F (2016) Detecting strong contemporary selection in small populations. Poster presentation at the Population Variation Genetics: Experimental Strategies and Analysis workshop, Earlham Institute – Norwich, UK.

Hubert JN & Hospital F (2015) Detecting selection within ten generations thanks to population genomics. Poster presentation at the Software and Statistical Methods for Population Genetics meeting – Aussois, France.

Hubert JN & Hospital F (2015) The power to detect selection from SNP allele frequencies variation. Poster presentation at the European Society for Evolutionary Biology meeting – University of Lausanne, Switzerland.

Hubert JN & Hospital F (2015) The power to detect selection from SNP allele frequencies variation. Poster presentation at the Society for Molecular Biology and Evolution, Biological Adaptation satellite meeting – Le Hameau de l'Etoile, France.

# Liste des abréviations

ACV : Analyse de Cycle de Vie

ADN : Acide DésoxyriboNucléique

ADO : *Allelic Dropout*

AFLP : *Amplified Fragment-Length Polymorphism*

aGPCR : *adhesion G Protein-Coupled Receptor*

ATP : Adénosine TriPhosphate

CDH : Cadhérines

CDK : *Cycline-Dependent Kinase*

CIPA : Comité Interprofessionnel des Produits de l'Aquaculture

CMH : Complexe Majeur d'Histocompatibilité

CPO : *Conditional Predictive Ordinate*

CTVT : *Canine Transmissible Venereal Tumor*

DFTD : *Devil Facial Tumor Disease*

DIC : *Deviance Information Criterion*

DL : Déséquilibre de Liaison

E&R : *Evolve and Resequence*

EFFICACE : EFFICacité Alimentaire : allocation des ressourCes Energétiques

EHH : *Extended Haplotype Homozygosity*

FAO : *Food and Agriculture Organization (of the United Nations)*

FDR : *False-Discovery Rate*

FGFR : *Fibroblast Growth Factor Receptors*

FLK : Test de Lewontin-Krakauer

FN : Freycinet

$F_{ST}$ : Indice de fixation de Wright

GAP : *GTPase Activating Protein*

GTPase : *G*uanosine TriPhosphatase

HAL : Hyper Articles en Ligne

HD : Haploïde Doublé

HLA : *Human Leucocyte Antigen*

HPC : *High Perfomance Computing*

IgSF-CAM : *Immunoglobulin SuperFamily - Cell Adhesion Molecule*

iHS : *integrated Haplotype Score*

INRA : Institut National de la Recherche Agronomique

INRA-PRN : souche de truite arc-en-ciel à ponte printanière

IPA : *Ingenuity Pathway Analysis*

kb : kilobases (*i.e.*, kilo- paire de bases)

LDD : *Linkage Disequilibrium Decay*

LNS : *Laboratory Natural Selection*

LPML : *Logarithm of the PseudoMarginal Likelihood*

LRH : *Long-Range Haplotype*

LRT : *Likelihood Ratio Test*

MAF : *Minor Allele Frequency*

MAP : *Mitogen-activated protein*

MCMC : *Monte-Carlo Markov Chain*

MMP : MétalloProtéinases

$N_e$ : Effectif efficace

NHGRI : *National Human Genome Research Institute*

NGS : *Next-Generation Sequencing*

NP : Narawntapu

pb : paire de bases (*i.e.*, paire de nucléotides)

PCR : *Polymerase Chain Reaction*

PE RAD-seq : *Paired-End Restriction-site Associated DNA sequencing*

PI3k/Akt : Phosphatidylinositol 3-Kinase

PPP-value : *Posterior Predictive P-value*

PRL : *Phosphatase Regenerating Liver*

QTL : *Quantitative Trait Loci*

RAD-seq : *Restriction site Associated DNA sequencing*

RAD tags : *Restriction site Associated DNA tags*

REHH : *Relative Extended Haplotype Homozygosity*

RIL : *Recombinant Inbred Lines*

ROH : *Runs Of Homozygosity*

RTK : Récepteur à activité Tyrosine Kinase

SE RAD-seq : *Single-End Restriction-site Associated DNA sequencing*

SFS : *Site Frequency Spectrum*

SGE : *Sun Grid Engine*

SJR : *Scimago Rank Journal*

SLURM : *Simple Linux Utility for Resource Management*

SMS : *Single Molecule Sequencing*

SNC : Système Nerveux Central

SNP : *Single Nucleotide Polymorphism*

SODD : *Silencer Of Death Domain*

ST5 : *Suppression of Tumorigenicity* 5

STAT3 : *Signal Transducers and Activators of Transcription* 3

TEAD : *Transcriptional Enhancer Factor Domain*

*TGF*β : *Transforming Growth Factor* β

TNF-R1 : *Tumor Necrosis Factor Receptor* 1

TRIM : *Tripartite Motif Family*

TSA : Troubles du Spectre Autistique

TSPAN : Tétraspanines

WP : West Pencil Pine

XP-EHH : *Cross Population-Extended Haplotype Homozygosity*

# Chapitre I

# TABLE DES MATIERES

-

# Chapitre I – Avant-propos

## — 1. Résumé

Grâce à la combinaison d'outils moléculaires modernes et de moyens de calcul puissants permettant d'analyser une grande masse de données, la génomique des populations rend possible la mise en évidence de traces, ou empreintes, de sélection dans le génome. Les travaux effectués dans ce domaine considèrent en général une échelle de temps longue (*i.e.*, plusieurs centaines ou milliers de générations). En comparaison, peu d'intérêt a été porté aux études expérimentales de court terme (*i.e.*, une dizaine de générations). De telles expériences sont pourtant susceptibles de nous renseigner sur la base génétique de caractères complexes. L'enjeu de la thèse est donc de déterminer s'il est possible de détecter des signatures de sélection contemporaines correspondant aux cas pratiques actuellement rencontrés chez certaines populations expérimentales. Nous nous intéressons plus spécifiquement aux petites populations soumises à une sélection directionnelle intense sur une dizaine de générations pour lesquelles des échantillons génétiques ont été acquis au cours du temps, formant une série génétique temporelle. De telles données peuvent être obtenues durant une expérience de sélection menée en laboratoire comme durant une « expérience de sélection naturelle » consécutive à une perturbation environnementale soudaine. Nous proposons une méthode de vraisemblance basée sur un modèle de Wright-Fisher pour exploiter au mieux les séries génétiques temporelles générées lors de ces expériences de sélection. Nous montrons par simulation que notre méthode permet de différencier les signaux dus à la combinaison de la sélection et de la dérive génétique de ceux dus à la seule dérive. Nous montrons également par simulation qu'il est possible d'estimer le coefficient de sélection appliqué à un locus testé. De plus, nous illustrons l'intérêt de notre méthode pour la détection de marqueurs candidats à la sélection au travers de deux études génomiques sur données réelles. Ces applications mettent en évidence des régions génomiques candidates pour des phénotypes complexes dans des contextes différents : en réponse à un cancer transmissible naturel chez le diable de Tasmanie (*Sarcophilus harrisii*) et en réponse à une sélection divergente artificielle sur l'efficacité alimentaire chez la truite arc-en-ciel (*Oncorhynchus mykiss*). Dans l'ensemble, nos résultats montrent qu'il est possible de détecter des gènes sujets à une sélection directionnelle intense à partir d'échantillons génétiques temporels, même si la sélection est de courte durée et si les populations examinées ont un faible effectif.

## — 2. Organisation du manuscrit

Ce manuscrit se compose de sept chapitres et quatre annexes. Chaque chapitre est précédé d'une table des matières spécifique et se termine par une section listant les travaux cités en référence (N.B. les noms d'auteurs homographes sont distingués dans le texte par l'ajout de l'initiale du prénom). Le Chapitre I, qui fait office d'avant-propos, comprend un résumé des travaux présentés ainsi que cette note. Le Chapitre II constitue l'introduction bibliographique du manuscrit. Nous y passons en revue les tests usuels de détection de signatures de sélection et y évoquons l'intérêt d'études couplant évolution expérimentale et reséquençage pour mieux comprendre la base génétique de caractères complexes. Le Chapitre III est essentiellement composé de la version de soumission d'un article (Article I) présentant une méthode de vraisemblance pour la détection de signatures de sélection à court terme (*i.e.*, une dizaine de générations). Ce travail s'appuie principalement sur des simulations. Nous proposons par la suite deux études d'application de cette méthode de vraisemblance. Le Chapitre IV correspond à la première de ces deux études. Nous nous y intéressons à la réponse évolutive à un cancer transmissible chez le diable de Tasmanie. Le Chapitre V présente le séquençage de marqueurs RAD (*Restriction site Associated DNA*), qui permet la découverte de polymorphisme *de novo*. Le Chapitre VI correspond à la seconde étude d'application, qui s'appuie sur la constitution d'un jeu de marqueurs RAD *de novo* pour rechercher des signatures de sélection liées à l'efficacité alimentaire au sein de lignées expérimentales de truite arc-en-ciel. Le Chapitre VII est consacré à la discussion générale, qui tire les conséquences globales de nos travaux et observations, et clôt le manuscrit. Les annexes comprennent deux articles (Article II, la version de soumission de l'article associé au Chapitre IV ; Article III, la version publiée d'un travail mentionné dans la discussion), une figure supplémentaire de l'Article II et une section d'aide à l'écriture de scripts de lancement de tâches.

# Chapitre II

# TABLE DES MATIERES

-

# Chapitre II – Inférer la présence de sélection au sein des génomes

## — 1. La génomique des populations

La combinaison d'outils moléculaires modernes fondés sur les techniques de séquençage haut débit et de moyens de calcul puissants a favorisé, ces dernières années, le développement d'une nouvelle branche de la génétique des populations : la génomique des populations. Celle-ci repose sur l'acquisition, pour un échantillon issu d'une population d'intérêt, de plusieurs centaines à plusieurs centaines de milliers de marqueurs génétiques (*e.g.,* des SNP – *Single Nucleotide Polymorphism*) dispersés le long du génome. Autrefois dominée par des approches théoriques, la génétique des populations devient un domaine envahi par les données (Pool *et al.*, 2010), permettant ainsi l'estimation de paramètres théoriques historiques, comme le taux de mutation ou l'effectif efficace des populations ($N_e$).

Un objectif central de la génomique des populations est de distinguer les effets locus-spécifiques (*e.g.,* mutation, sélection) de ceux qui affectent le génome dans son ensemble (*e.g.,* dérive génétique, flux de gènes). En s'appuyant sur la grande quantité de données désormais disponibles, les généticiens des populations cherchent à faire des inférences sur l'histoire des populations, non seulement dans le cas du polymorphisme neutre (*e.g.,* taille de population, diversité), mais aussi dans le cas du polymorphisme sélectionné. L'application des méthodes de génomique des populations permet en particulier de mettre en évidence des signatures de sélection dans le génome. Ce sont par exemple des régions où le profil de polymorphisme moléculaire observé n'est pas cohérent avec une évolution neutre, ce qui fait que l'action d'autres forces évolutives, comme la sélection, doit être envisagée. A l'aide de différentes statistiques, on cherche alors à distinguer un « signal » dû à la sélection au milieu d'un « bruit de fond » neutre (Fig. II-1). Une partie importante de la littérature en génomique des populations a été consacrée aux travaux visant à mettre en évidence ces signatures de sélection.

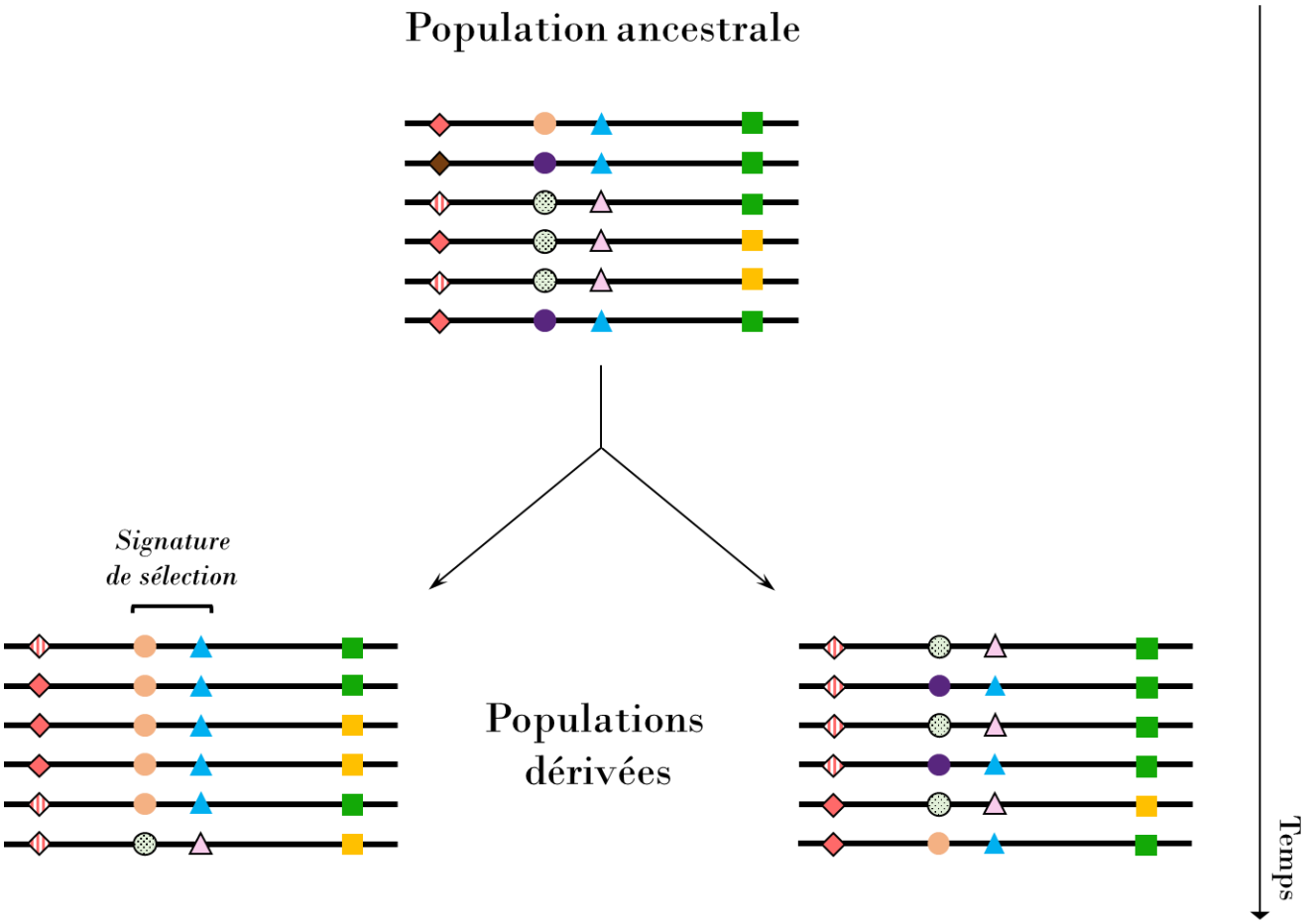



**Figure II-1. La sélection laisse des patrons de variation génétique localement identifiables dans le génome**. Des échantillons de six séquences sont tirés d'une population ancestrale et de deux populations dérivées. Les formes géométriques colorées représentent les différents allèles présents aux locus polymorphes. Un évènement sélectif a laissé une empreinte dans la population dérivée de gauche, ce qui se traduit par une augmentation de la fréquence de l'allèle bénéfique (disque orange), mais aussi de l'allèle neutre voisin (triangle bleu) du fait de sa liaison à l'allèle bénéfique. Ce phénomène est désigné sous le terme d'autostop génétique. La signature localement engendrée par cette sélection positive peut être repérée par différentes approches de génomique des populations, selon les données disponibles. La présence de sélection peut ainsi être inférée en observant (i) une structure haplotypique fortement conservée parmi la variation génétique disponible au sein de la population sélectionnée (*cf.* paragraphe 2.2 *Les méthodes de recherche de signatures de sélection intra-population*), (ii) une divergence localement importante au niveau du locus sélectionné, s'il est possible de comparer la variation génétique de la population sélectionnée à celle d'une population dérivée non sélectionnée (ou sélectionnée de façon divergente) (*cf.* paragraphe 2.1 *Lewontin, Krakauer et les mesures de différenciation*), (iii) une évolution temporelle compatible avec une action de la sélection, s'il est possible de comparer des échantillons génétiques temporels issus des populations ancestrale et sélectionnée (*cf.* section 4. *Les expériences d'Evolution et Reséquençage*).

# — 2. La recherche de signatures de sélection

## 2.1. Lewontin, Krakauer et les mesures de différenciation

La recherche de signatures de sélection a débouché sur plusieurs résultats marquants auprès d'une large audience, comme l'identification de régions génomiques hébergeant des mutations causales soumises à une sélection positive au sein de populations humaines (Tishkoff *et al.*, 2007 ; Barreiro *et al.*, 2008 ; Vasseur & Quintana-Murci, 2013) ou d'animaux domestiques (Bovine HapMap Consortium, 2009 ; Olsson *et al.*, 2011 ; Petersen *et al.*, 2013).

Historiquement, un principe fondateur dans cette branche de la génétique des populations est l'idée formalisée dès 1973 par Richard Lewontin et Jesse Krakauer (Cavalli-Sforza, 1966 ; Lewontin & Krakauer, 1973) que certaines différences entre populations sont le fruit d'une sélection naturelle positive susceptible d'avoir laissé des patrons de variation localement distincts dans le génome. La comparaison des fréquences alléliques de populations divergentes devrait par conséquent permettre d'identifier les locus qui ont effectivement été ciblés par la sélection. En pratique, Lewontin et Krakauer ont proposé d'utiliser un estimateur de l'indice de fixation de Wright ($F_{ST}$) (Wright, 1931 ; pour une revue récente, voir par exemple Holsinger & Weir, 2009) comme mesure de différenciation pour révéler de potentielles signatures de sélection. Des valeurs de $F_{ST}$ localement extrêmes (si les fréquences alléliques sont très différenciées d'une population à l'autre, le $F_{ST}$ tend vers 1 ; si en revanche elles sont homogènes, il tend vers 0) par rapport à la distribution supposée neutre du $F_{ST}$, obtenue grâce à l'ensemble des marqueurs génomiques disponibles, suggèrent la présence d'une sélection positive (*i.e.,* en cas de $F_{ST}$ suffisamment fort) ou équilibrante (*i.e.,* en cas de $F_{ST}$ suffisamment faible). Tester l'écart du $F_{ST}$ locus-spécifique au $F_{ST}$ génomique moyen semblait ainsi être un moyen intéressant de séparer les effets de la sélection de ceux de la démographie afin d'identifier des marqueurs génétiques candidats à la sélection.

La principale limite de ce type d'approche a très vite été soulignée : dès 1975, Robertson, Nei et Maruyama montreront qu'un $F_{ST}$ localement élevé du fait de la structure de la population peut conduire à considérer comme candidat à la sélection un marqueur pourtant neutre (Nei & Maruyama, 1975 ; Robertson, 1975). Cet effet prête d'autant plus à confusion que la structure des populations s'écarte de l'hypothèse faite par Lewontin et Krakauer d'un modèle en îles ne prenant pas en compte les migrations préférentielles entre populations voisines ni la possibilité d'une histoire démographique complexe. Les limites des tests de détection de signatures de sélection basés sur l'utilisation du $F_{ST}$ ont été bien caractérisées (Beaumont, 2005 ; Excoffier *et al.*, 2009 ; Hermisson, 2009 ; Bierne *et al*., 2013), ce qui n'a toutefois pas nui à leur succès. En effet, avec l'avènement des données génomiques à haut

débit et la baisse du coût du génotypage par marqueur, ce type de test a été employé à de multiples reprises, permettant d'identifier des gènes candidats à une sélection directionnelle au sein de nombreuses populations (*e.g.*, Akey *et al.*, 2002 ; Flori *et al.*, 2009 ; Nielsen, E. E., *et al.*, 2009 ; Chávez-Galarza *et al.*, 2013 ; Petersen *et al.*, 2013 ; McRae *et al.*, 2014).

L'idée de Lewontin et Krakauer a inspiré le développement de statistiques de test basées sur le $F_{ST}$ plus sophistiquées, reposant sur des modèles démographiques plus réalistes que celui initialement postulé (*e.g.,* Vitalis *et al.*, 2001 ; Foll & Gaggiotti, 2008). Des tests incorporant d'autres types d'information, comme la matrice de parenté entre populations (Bonhomme *et al.*, 2010 ; Fariello *et al.*, 2013), une variable de « différenciation environnementale » (de Villemereuil & Gaggiotti, 2015) ou bien une distribution neutre du $F_{ST}$ déduite des génotypes (Whitlock & Lotterhos, 2015) ont aussi été développés pour aider à différencier les effets liés à la démographie de ceux dus à la sélection.

D'autres stratégies de mesure de la différenciation entre populations n'utilisent pas le $F_{ST}$, mais sont conceptuellement liées à l'idée de Lewontin et Krakauer. C'est le cas de nombreuses méthodes d'association génotype-environnement, qui permettent d'identifier le polymorphisme affichant une forte corrélation avec des variables reflétant les pressions écologiques (Coop *et al.*, 2010). Ces méthodes prennent aujourd'hui en compte l'effet de la démographie, des effets population-spécifiques (via des variables environnementales mais aussi phénotypiques), voire les effets de la dépendance entre marqueurs au sein du génome (*i.e.,* le déséquilibre de liaison, DL) et sont très prisées pour étudier la base génétique d'une adaptation locale (Hoban *et al.*, 2016). Les développements récents implémentent par exemple des modèles bayésiens hiérarchiques incorporant une statistique de test apparentée au $F_{ST}$ (Günther & Coop 2013 ; Gautier *et al.*, 2015) ou encore des modèles de régression statistique (Frichot et *al.*, 2013 ; Frichot & François, 2015).

Les méthodes de recherche de signatures de sélection exploitant la différenciation entre populations forment un domaine de la recherche en génomique des populations arrivant à maturité. Beaucoup de méthodes flexibles et abouties sont aujourd'hui distribuées sous la forme de programmes documentés (*e.g.,* BayPass, Bayescenv, hapFLK) ou de packages R (*e.g.,* LEA, outFLANK, PCAdapt), si bien que l'utilisateur n'a que l'embarras du choix.

## 2.2. Les méthodes de recherche de signatures de sélection intra-population

Le principe des méthodes de recherche de signatures de sélection intra-population n'est pas très différent de celui des méthodes de différenciation : l'objectif est toujours l'identification de zones localement distinctes du reste du génome, mais sans avoir la possibilité de comparer des génomes ayant subi une différenciation. Souvent, les tests effectués sont empiriques, c'est-à-dire que la

distribution neutre de la statistique de test utilisée ne dépend pas explicitement d'un modèle, mais directement des données observées.

Une stratégie consiste à rechercher une déviation dans le spectre des fréquences alléliques (SFS) – c'est-à-dire la proportion du nombre d'allèles présents au sein de différentes classes de fréquence allélique – attendu sous l'hypothèse d'une évolution neutre. En effet, en augmentant localement la présence d'allèles rares par autostop génétique, la sélection tend à décaler le SFS vers de très faibles valeurs au voisinage du site sélectionné (Braverman *et al.*, 1995). A la suite de l'introduction du *D* de Tajima (1989), plusieurs statistiques basées sur l'estimation de la quantité de polymorphisme disponible dans la population (Watterson, 1975) ont été développées – *D** et *F** de Fu & Li (1993), *H* de Fay & Wu (2000) – pour repérer au travers d'une déviation du SFS les régions génomiques potentiellement sous sélection. Ces tests ont été très largement utilisés pour l'identification de signatures de sélection (Nielsen, R., *et al.*, 2005). D'autres tests de neutralité basés sur le SFS ont été développés plus récemment (*e.g.,* Boitard & Rocha, 2013 ; Ronen *et al.*, 2013).

Certaines approches visent à identifier les régions génomiques arborant une réduction locale extrême du polymorphisme par rapport à la moyenne observée sur l'ensemble du génome, ce qui peut constituer la signature d'une sélection passée récente (Maynard-Smith & Haigh, 1974). En pratique, les statistiques disponibles ciblent des segments de génome hébergeant un grand nombre de marqueurs consécutifs homozygotes, comme les *Runs Of Homozygosity* (ROH) (McQuillan *et al.*, 2008 ; Metzger *et al.*, 2015).

L'idée de cibler les régions génomiques comportant des portions homozygotes anormalement longues est commune à plusieurs indicateurs très utilisés pour détecter la sélection, consécutivement à la publication de Sabeti *et al.* (2002). Celle-ci définit l'*Extended Haplotype Homozygosity* (EHH), qui permet de révéler des structures haplotypiques inhabituelles suggérant une sélection passée, et qui est utilisée par plusieurs tests de détection de signatures de sélection. Dans le cas du test *Long-Range Haplotype* (LRH), la présence de sélection est testée en identifiant les haplotypes fréquents possédant une forte EHH relative (REHH), c'est-à-dire une EHH élevée par rapport aux haplotypes voisins, ce qui permet de prendre en compte des fluctuations du taux de recombinaison d'une région du génome à l'autre (Sabeti *et al.* 2002). En distinguant de la sorte les allèles probablement bénéfiques (*i.e.,* des allèles ayant subi une augmentation si rapide de leur fréquence allélique que la variation neutre environnante est aussi affectée), le LRH représente une approche puissante pour identifier la signature d'une sélection positive (Gautier & Vitalis, 2012). D'autres tests n'utilisant pas l'information haplotypique, comme le test *Linkage Disequilibrium Decay* (LDD), s'appuient sur la même idée (Wang *et al.*, 2006).

L'approche de Sabeti *et al.* (2002) a suscité le développement d'autres tests recherchant des haplotypes présents à fréquence élevée tout en étant associés à un fort DL, et s'appuyant pour cela sur l'EHH. Une extension de l'EHH est l'*integrated Haplotype Score* (iHS ; Voight *et al.*, 2006). En observant que l'intégrale de l'EHH par rapport à la distance physique est plus grande pour une région sujette à la sélection que pour une région neutre, Voight *et al.* (2006) ont proposé l'iHS pour comparer les intégrales obtenues pour un SNP à l'état ancestral et à l'état dérivé. Les valeurs extrêmes de l'iHS indiquent de possibles signatures de sélection. D'autres statistiques basées sur la mesure de l'EHH, mais plutôt destinées à des analyses comparatives, ont aussi été proposées. C'est le cas du *Rsb* (Tang *et al.*, 2007) et de la *Cross Population-EHH* (XP-EHH ; Sabeti *et al.*, 2007), qui comparent un même SNP entre deux populations afin d'identifier une EHH plus étendue dans une population que dans l'autre, ce qui suggère une signature de sélection.

Tous ces tests intra-population ont été largement utilisés pour la détection de signatures de sélection, en particulier chez l'Homme (*e.g.,* Bamshad & Wooding 2003 ; Akey *et al.*, 2004 ; Yu *et al.*, 2005 ; Walsh *et al.*, 2006 ; Magalon *et al.*, 2008 ; Yang *et al.*, 2011 ; pour une synthèse, voir Vasseur & Quintana-Murci, 2013). Comme pour les tests de différenciation, il existe des logiciels et packages simples d'usage implémentant les approches intra-population usuelles décrites dans cette section – par exemple, *DnaSP* (Rozas *et al.*, 2017), *selscan* (Szpiech & Hernandez, 2014) et *rehh* (Gautier & Vitalis, 2012).

## — 3. L'évolution expérimentale

« L'évolution expérimentale, c'est la biologie évolutive dans sa forme la plus empirique ». Dans l'introduction de leur synthèse sur l'évolution expérimentale, Rose & Garland (2009) suggèrent à la fois l'importance des études expérimentales en biologie évolutive, mais aussi la grande variété des approches employées dans ce domaine. L'évolution expérimentale concerne un spectre étendu de travaux, comme par exemple la recherche d'une preuve expérimentale de la pléiotropie antagoniste (Cooper & Lenski, 2000), l'observation d'un processus de domestication en temps réel (Belyaev, 1979 ; Simões *et al.*, 2007 ; Christie *et al.*, 2016) ou la modification des traits d'histoire de vie de populations naturelles du fait de leur transfert dans un nouvel environnement (Reznick *et al.*, 1990). Un principe commun à toutes ces expériences d'évolution est de suivre l'action des forces évolutives au cours du temps sur une population placée dans un environnement particulier.

Garland (2003) a proposé de classer les expériences d'évolution au sein de quatre grandes catégories selon le type d'altération environnementale considéré. La première catégorie comprend les expériences de sélection artificielle, où la fraction des individus conservés pour procréer chaque génération est définie selon un critère phénotypique (*e.g.,* un trait morphologique ou un comportement). C'est principalement à cette catégorie qu'appartiennent les lignées entretenues au sein des unités expérimentales de l'INRA. La deuxième catégorie concerne les expériences de sélection naturelle menées en laboratoire (LNS). Contrairement aux expériences de sélection artificielle, les individus ne font pas l'objet d'une mesure phénotypique, mais sont exposés à une altération de leurs conditions environnementales (*e.g.,* un changement de la photopériode ou de la température). Ces expériences sont en général menées chez des populations d'organismes modèles (*e.g.,* bactéries, levures, nématodes et drosophiles), aussi bien pour tester des prédictions théoriques de portée générale que pour évaluer la capacité de la population examinée à s'adapter à une altération environnementale particulière. La troisième catégorie est en réalité une variante de LNS, que l'on nomme « expériences d'abattage ». Ces expériences consistent à soumettre la population étudiée à un stress environnemental intense de façon à ne laisser survivre qu'un nombre d'individus déterminé à l'avance par l'expérimentateur. Pour des considérations éthiques évidentes, ces expériences ne sont généralement menées qu'avec des invertébrés – par exemple, chez les moustiques (Juliano & Gravel, 2002). Enfin, la quatrième catégorie proposée par Garland (2003) correspond aux expériences effectuées sur le terrain, c'est-à-dire dans un milieu naturel ou semblable à un milieu naturel. Cette catégorie inclut les introductions d'individus dans un nouvel environnement, que celles-ci consistent en des expériences intentionnelles (Reznick *et al.*, 1990 ; Bach *et al.*, 2018) ou fortuites (Hendry *et al.*, 2000 ; Liebl *et al.*, 2015).

Cette quatrième catégorie se distingue des expériences de sélection plus conventionnelles où les individus sont maintenus en captivité. Si elles ne peuvent pas toujours être aussi facilement contrôlées et répliquées que les expériences LNS, les études menées en milieu naturel permettent en revanche d'intégrer la complexité écologique des habitats naturels. La comparaison entre expériences de sélection naturelle incorporant cette complexité et expériences LNS montre par exemple que l'amélioration de la valeur sélective constatée en laboratoire n'est pas forcément visible dans des conditions plus réalistes au plan écologique (Bach *et al.*, 2018). Les expériences de sélection naturelle sont donc indispensables pour comprendre les bases de l'adaptation à un environnement nouveau. En particulier, les altérations d'habitat naturel peuvent donner lieu à des expériences de sélection naturelles « fortuites » nous renseignant sur la capacité des populations à s'adapter rapidement (Irshick & Reznick, 2009). L'étude approfondie de certains évènements naturels inattendus, comme le passage de populations d'organismes marins à un milieu dulcicole du fait d'un tremblement de terre (Lescak *et al.*, 2015) ou l'émergence d'un agent pathogène létal au sein d'une petite population (Epstein *et al.*, 2016), est susceptible d'améliorer notre compréhension des mécanismes sous-tendant l'adaptation rapide, notamment lorsque des données longitudinales ont pu être collectées.

## — 4. Les expériences d'Evolution et Reséquençage (E&R)

Une application récente de la génomique des populations consiste à coupler évolution expérimentale et séquençage de façon à pouvoir examiner l'impact de la sélection à l'échelle génomique. Ce type d'approche est aujourd'hui communément désignée par le terme d'expérience « d'Evolution et Reséquençage » (E&R) (Turner *et al.*, 2011), et concerne des expériences dont la durée peut s'étendre d'une seule génération (Christie *et al.*, 2016) à plusieurs dizaines de milliers (Good *et al.*, 2017). Les expériences E&R prévoient très fréquemment l'acquisition de séries génétiques temporelles (Fig. II-2), c'est-à-dire la collecte d'échantillons séparés d'une ou plusieurs générations en vue d'un génotypage. Travailler avec des séries génétiques temporelles permet de suivre la trajectoire de marqueurs génétiques au cours du temps. Les séries génétiques temporelles générées lors d'expériences E&R documentent ainsi en temps réel le processus sous-tendant l'adaptation à de nouvelles conditions environnementales.

Les expériences E&R offrent de nouvelles opportunités d'améliorer notre compréhension de l'évolution biologique sur des questions de longue date, comme par exemple la dynamique des mutations ou le rôle de la variation génétique préexistante (de l'anglais *standing genetic variation*, désignant la variation génétique déjà disponible dans la population) dans l'adaptation (Long *et al.*, 2015). A l'heure actuelle, les expériences E&R ont surtout été effectuées sur des populations de microorganismes ou d'invertébrés pour lesquelles le nombre de générations d'évolution et les tailles de population peuvent être élevés. Par exemple, l'effectif efficace ($N_e$) des populations de nématodes (*Caenorhabditis elegans*) et de drosophiles est compris entre quelques centaines et quelques milliers d'individus (Barrière & Félix, 2005 ; Mueller *et al.*, 2013), tandis que celui des populations d'*Escherichia coli* de la célèbre expérience de Richard Lenski est de l'ordre de $10^7$ (Good *et al.*, 2017). Sur une période d'un an, il est possible d'obtenir 25 générations d'évolution avec des drosophiles (Long *et al.*, 2015), 100 générations avec *C. elegans* (Hodgkin, 2002), et au moins 2000 générations avec *E. coli* (Good *et al.*, 2017).

En comparaison, les populations de plus gros organismes eucaryotes ayant un temps de génération plus important (plus d'un an par génération), présentant un $N_e$ de l'ordre de quelques dizaines d'individus et soumises à une forte sélection, ont été peu étudiées. Pourtant, les séries génétiques temporelles obtenues chez ces populations pourraient permettre de mieux comprendre la base génétique de phénotypes complexes, ainsi que la capacité de certaines petites populations à s'adapter rapidement.

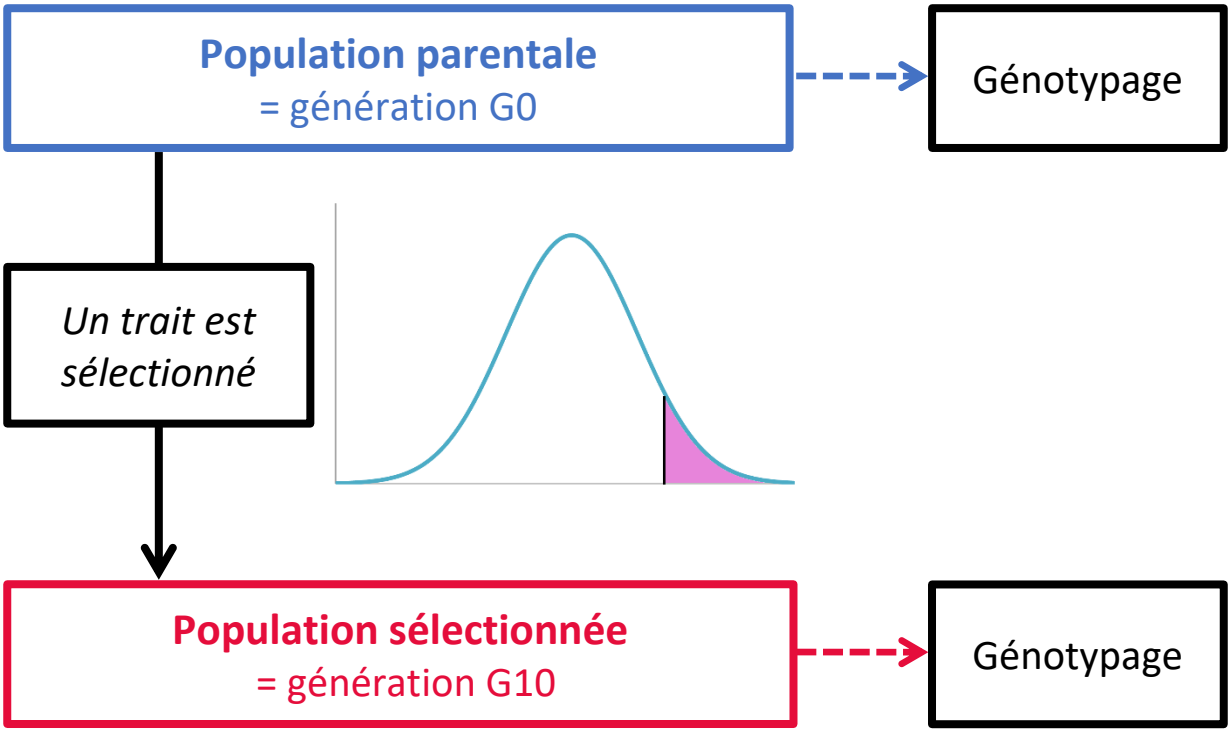



**Figure II-2. Exemple d'expérience d'Evolution et Reséquençage (E&R).** Dans cet exemple, une petite population ($N_e$ ≈ 50) est sélectionnée de façon artificielle sur une période de dix générations (*i.e.,* de G0 à G10). A chaque génération, une sélection par troncature est effectuée : les individus sont mesurés pour un trait et seule la fraction des individus affichant les meilleures performances pour ce trait est conservée pour procréer la génération suivante. Des échantillons sont prélevés en début (G0) et en fin (G10) d'expérience afin d'être génotypés. Notre objectif est d'identifier des locus potentiellement ciblés par la sélection à partir des échantillons génétiques temporels recueillis.

## — 5. Exploiter les expériences E&R courtes chez les petites populations

Nous avons vu qu'il existait déjà beaucoup de méthodes de détection de signatures de sélection, mais celles-ci ont surtout été utilisées pour révéler des empreintes laissées par un évènement de sélection passé, remontant à plusieurs centaines ou plusieurs milliers de générations (*e.g.,* Imhoff & Schlötterer, 2001 ; Bersaglieri *et al.*, 2004 ; Jarvis *et al.*, 2012 ; Rubin *et al.*, 2012 ; Carneiro *et al.*, 2014). Par contre, il n'existe pas de méthode véritablement adaptée au traitement des expériences E&R, en particulier celles qui sont menées sur de courtes périodes (*e.g.,* sur dix générations ; *cf.* Fig. II-2) au sein de petites populations ($N_e \approx 50$), alors que la baisse des coûts de génotypage rend possible l'acquisition de ce type de séries génétiques temporelles y compris chez les espèces non-modèles.

En l'absence de méthodologie dédiée, on ne trouve aujourd'hui dans la littérature que peu d'analyses de signatures de sélection utilisant des séries génétiques temporelles acquises dans ces conditions, et celles-ci sont principalement menées chez les drosophiles. Les inférences s'appuient alors sur des approches empiriques comparant les fréquences alléliques au début et au terme d'une expérience E&R, soit avec des tests classiquement utilisés pour l'analyse de tableaux de contingence (Burke *et al.*, 2010 ; Orozco-terWengel *et al.*, 2012 ; Martins *et al.*, 2014 ; Jalvingh *et al.*, 2016), soit en utilisant le $F_{ST}$ (Remolina *et al.*, 2012 ; Dubois *et al.*, 2017), soit, plus rarement, en adaptant une statistique basée sur l'EHH (Epstein *et al.*, 2016). Peu d'approches ont en revanche été fondées sur des modèles de génétique des populations rendant compte d'une évolution des fréquences alléliques au cours du temps, alors que ce type de méthodologie permettrait de calculer la probabilité d'observer un saut fréquentiel donné sous une hypothèse de neutralité réaliste.

Dans le cadre de cette thèse, nous proposons d'exploiter les séries génétiques temporelles générées lors d'expériences E&R en couplant un modèle de Wright-Fisher à une approche de maximisation de la vraisemblance. Williamson & Slatkin (1999) ont développé ce type d'approche avec un modèle de dérive génétique pure pour estimer $N_e$ à partir de la variation temporelle des fréquences alléliques. L'idée a ensuite été reprise en incorporant l'effet de la sélection, afin d'estimer non seulement $N_e$ mais aussi le coefficient de sélection associé aux potentiels SNP sous sélection (Bollback *et al.*, 2008). Pour des raisons d'efficacité computationnelle, le développement de ces approches s'est limité à l'utilisation d'approximations du modèle de Wright-Fisher (Bollback *et al.*, 2008 ; Malaspinas *et al.*, 2012 ; Mathieson & McVean, 2013 ; Feder *et al.*, 2014), ce qui les destine plutôt à traiter les cas où la sélection est modérée et où $N_e$ est élevé.

Il existe pourtant des données acquises au sein de petites populations sujettes à une sélection directionnelle intense sur quelques générations, comme par exemple au sein des lignées expérimentales de populations animales entretenues par sélection artificielle ou chez les petites populations naturelles soumises à un brusque changement environnemental. La recherche de signatures de sélection à partir des séries génétiques temporelles générées dans le cadre de ce régime de sélection (*i.e.,* forte sélection, faible nombre de générations de sélection et faible $N_e$) pourrait ainsi bénéficier d'une méthode calculant la vraisemblance exacte des trajectoires alléliques observées sous un modèle de Wright-Fisher incluant la sélection. Dans les chapitres suivants, nous présenterons cette méthode de vraisemblance, une étude par simulation de sa puissance statistique, ainsi que son application au cas de deux petites populations réelles soumises à une forte sélection pendant moins de dix générations.

# — 6. Références


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

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., & Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS biology*, *2*(10), e286.

Bach, L. T., Lohbeck, K. T., Reusch, T. B., & Riebesell, U. (2018). Rapid evolution of highly variable competitive abilities in a key phytoplankton species. *Nature ecology & evolution*, *2*(4), 611.

Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, *4*(2), 99.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature genetics*, *40*(3), 340.

Barrière, A., & Félix, M. A. (2005). High local genetic diversity and low outcrossing rate in Caenorhabditis elegans natural populations. *Current Biology*, *15*(13), 1176-1184.

Beaumont, M. A. (2005). Adaptation and speciation: what can Fst tell us?. *Trends in ecology & evolution*, *20*(8), 435-440.

Belyaev, D. K. (1979). Destabilizing selection as a factor in domestication. *Journal of Heredity*, *70*(5), 301-308.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., ... & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, *74*(6), 1111-1120.

Bierne, N., Roze, D., & Welch, J. J. (2013). Pervasive selection or is it…? why are FST outliers sometimes so frequent?. Molecular ecology, 22(8), 2061-2064.

Boitard, S., & Rocha, D. (2013). Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed. Animal genetics, 44(5), 579-583.

Bollback, J. P., York, T. L., & Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics*, *179*(1), 497-502.

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

Bovine HapMap Consortium. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, *324*(5926), 528-532.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, *140*(2), 783-796.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., & Long, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, *467*(7315), 587.

Carneiro, M., Rubin, C. J., Di Palma, F., Albert, F. W., Alföldi, J., Barrio, A. M., ... & Younis, S. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, *345*(6200), 1074-1079.

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

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Muñoz, I., ... & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (Apis mellifera iberiensis) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology*, *22*(23), 5890-5907.

Christie, M. R., Marine, M. L., Fox, S. E., French, R. A., & Blouin, M. S. (2016). A single generation of domestication heritably alters the expression of hundreds of genes. *Nature Communications*, *7*, 10676.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, *185*(4), 1411-1423.

Cooper, V. S., & Lenski, R. E. (2000). The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature*, *407*(6805), 736.

de Villemereuil, P., & 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.

Dubois, A., Galan, M., Cosson, J. F., Gauffre, B., Henttonen, H., Niemimaa, J., ... & Charbonnel, N. (2017). Microevolution of bank voles (Myodes glareolus) at neutral and immune-related genes during multiannual dynamic cycles: Consequences for Puumala hantavirus epidemiology. *Infection, Genetics and Evolution*, *49*, 318-329.

Duforet-Frebourg, N., Bazin, E., & Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular biology and evolution*, *31*(9), 2483-2495.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ... & Storfer, A. (2016). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nature communications*, *7*, 12684.

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, *103*(4), 285.

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

Feder, A. F., Kryazhimskiy, S., & Plotkin, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics*, *196*(2), 509-522.

Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S., ... & Gautier, M. (2009). The genome response to artificial selection: a case study in dairy cattle. *PloS one*, *4*(8), e6595.

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

Frichot, E., Schoville, S. D., de Villemereuil, P., Gaggiotti, O. E., & François, O. (2015). Detecting adaptive evolution based on association with ecological gradients: Orientation matters!. *Heredity*, *115*(1), 22.

Frichot, E., & François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925-929.

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*(3), 693-709.

Garland, T., & Rose, M. R. (2009). Darwin's other mistake. In *Experimental evolution: concepts, methods, and applications of selection experiments* (pp. 3-30). University of California Press.

Gautier, M., & Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176-1177.

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

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, *551*(7678), 45.

Gulcher, J., & Stefansson, K. (1998). Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clinical Chemistry and Laboratory Medicine*, *36*(8), 523-527.

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205-220.

Hendry, A. P., Wenburg, J. K., Bentzen, P., Volk, E. C., & Quinn, T. P. (2000). Rapid evolution of reproductive isolation in the wild: evidence from introduced salmon. *Science*, *290*(5491), 516-518.

Hermisson, J. (2009). Who believes in whole-genome scans for selection? *Heredity* 103, 283–284.

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

Hodgkin, J. (2002). Exploring the envelope: systematic alteration in the sex-determination system of the nematode Caenorhabditis elegans. *Genetics*, *162*(2), 767-780.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F ST. *Nature Reviews Genetics*, *10*(9), 639.

Imhof, M., & Schlötterer, C. (2001). Fitness effects of advantageous mutations in evolving Escherichia coli populations. *Proceedings of the National Academy of Sciences*, *98*(3), 1113-1117.

Irschick, D. J., & Reznick, D. (2009). Field experiments, introductions, and experimental evolution. In *Experimental evolution: concepts, methods, and applications of selection experiments* (pp. 173-194). University of California Press.

Jalvingh, K. M., Chang, P. L., Nuzhdin, S. V., & Wertheim, B. (2014). Genomic changes under rapid evolution: selection for parasitoid resistance. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1779), 20132303.

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., ... & Mezey, J. (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics*, *8*(4), e1002641.

Juliano, S. A., & Gravel, M. E. (2002). Predation and the evolution of prey behavior: an experiment with tree hole mosquitoes. *Behavioral Ecology*, *13*(3), 301-311.

Lescak, E. A., Bassham, S. L., Catchen, J., Gelmond, O., Sherbick, M. L., von Hippel, F. A., & Cresko, W. A. (2015). Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, *112*(52), E7204-E7212.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.

Liebl, A. L., Schrey, A. W., Andrew, S. C., Sheldon, E. L., & Griffith, S. C. (2015). Invasion genetics: Lessons from a ubiquitous bird, the house sparrow Passer domesticus. *Current Zoology*, *61*(3), 465-476.

Long, A., Liti, G., Luptak, A., & Tenaillon, O. (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, *16*(10), 567.

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

Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L., & Heyer, E. (2008). Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. *European Journal of Human Genetics*, *16*(2), 243.

Martins, N. E., Faria, V. G., Nolte, V., Schlötterer, C., Teixeira, L., Sucena, É., & Magalhães, S. (2014). Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences*, *111*(16), 5938-5943.

Mathieson, I., & McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, *193*(3), 973-984.

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., ... & MacLeod, A. K. (2008). Runs of homozygosity in European populations. *The American Journal of Human Genetics*, *83*(3), 359-372.

McRae, K. M., McEwan, J. C., Dodds, K. G., & Gemmell, N. J. (2014). Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. *BMC genomics*, *15*(1), 637.

Metzger, J., Karwath, M., Tonda, R., Beltran, S., Águeda, L., Gut, M., ... & Distl, O. (2015). Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC genomics*, *16*(1), 764.

Mueller, L. D., Joshi, A., Santos, M., & Rose, M. R. (2013). Effective population size and evolutionary dynamics in outbred laboratory populations of Drosophila. *Journal of genetics*, *92*(3), 349-361.

Nei, M., & Maruyama, T. (1975). Lewontin-Krakauer test for neutral genes. *Genetics*, *80*(2), 395-395.

Nielsen, E. E., Hemmer-Hansen, J., Poulsen, N. A., Loeschcke, V., Moen, T., Johansen, T., ... & Carvalho, G. R. (2009). Genomic signatures of local directional selection in a high gene flow marine organism ; the Atlantic cod (Gadus morhua). *BMC evolutionary biology*, *9*(1), 276.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, *39*, 197-218.

Olsson, M., Meadows, J. R., Truve, K., Pielberg, G. R., Puppo, F., Mauceli, E., ... & Bassols, A. (2011). A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS genetics*, *7*(3), e1001332.

Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., & Schlötterer, C. (2012). Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology*, *21*(20), 4931-4941.

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., ... & Brama, P. (2013). Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS genetics*, *9*(1), e1003211.

Pool, J. E., Hellmann, I., Jensen, J. D., & Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome research*, *20*(3), 291-300.

Remolina, S. C., Chang, P. L., Leips, J., Nuzhdin, S. V., & Hughes, K. A. (2012). Genomic basis of aging and life-history evolution in Drosophila melanogaster. *Evolution*, *66*(11), 3390-3403.

Reznick, D. A., Bryga, H., & Endler, J. A. (1990). Experimentally induced life-history evolution in a natural population. *Nature*, *346*(6282), 357.

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181-193.

Rose, M. R., Graves, J. L., & Hutchinson, E. W. (1990). The use of selection to probe patterns of pleiotropy in fitness characters. In *Insect life cycles* (pp. 29-42). Springer, London.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. *Mol. Biol. Evol.* 34**:** 3299-3302. DOI: 10.1093/molbev/msx248

Rubin, C. J., Megens, H. J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., ... & Archibald, A. L. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences*, *109*(48), 19529-19536.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., ... & Ackerman, H. C. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., ... & Schaffner, S. F. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913.

Simões, P., Rose, M. R., Duarte, A., Gonçalves, R., & Matos, M. (2007). Evolutionary domestication in Drosophila subobscura. *Journal of evolutionary biology*, *20*(2), 758-766.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, *23*(1), 23-35.

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*, *31*(10), 2824-2827.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585-595.

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

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., ... & Ibrahim, M. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, *39*(1), 31.

Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., & Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS genetics*, *7*(3), e1001336.

Vasseur, E., & Quintana-Murci, L. (2013). The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evolutionary applications*, *6*(4), 596-607.

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

Walsh, E. C., Sabeti, P., Hutcheson, H. B., Fry, B., Schaffner, S. F., de Bakker, P. I., ... & Winkler, C. (2006). Searching for signals of evolutionary selection in 168 genes related to immune function. *Human genetics*, *119*(1-2), 92-102.

Wang, E. T., Kodama, G., Baldi, P., & Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(1), 135-140.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, *7*(2), 256-276.

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F ST. *The American Naturalist*, *186*(S1), S24-S36.

Williamson, E. G., & Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, *152*(2), 755-761.

Wright, S. (1949). The genetical structure of populations. *Annals of Human Genetics*, *15*(1), 323-354.

Yang, K., Zheng, H., Qin, Z., Lu, Y., Farina, S. E., Li, S., ... & Li, H. (2011). Positive selection on mitochondrial M7 lineages among the Gelong people in Hainan. *Journal of human genetics*, *56*(3), 253.

Yu, F., Sabeti, P. C., Hardenbol, P., Fu, Q., Fry, B., Lu, X., ... & Leal, S. M. (2005). Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS genetics*, *1*(3), e41.

# Chapitre III

# TABLE DES MATIERES

-

# Chapitre III – Détecter la sélection à court terme avec une méthode de vraisemblance

## — 1. Présentation de l'Article I

Nous nous intéressons à l'exploitation de séries génétiques temporelles acquises au sein de petites populations ($N_e \sim 50$) sujettes à une sélection directionnelle, comme dans le cas de certaines expériences E&R. Pour exploiter ces séries temporelles, nous proposons une méthode de détection de signatures de sélection par maximisation de la vraisemblance s'appuyant sur un modèle de Wright-Fisher. Le principe de notre méthode est d'effectuer une comparaison de la vraisemblance des données (*i.e.,* les fréquences alléliques d'un allèle d'intérêt au cours du temps) selon deux hypothèses : l'une ($H_0$) considérant un modèle de Wright-Fisher incluant la dérive seule, la seconde ($H_1$) considérant un modèle de Wright-Fisher incluant la sélection en plus de la dérive. Un test de rapport de vraisemblance (LRT) permet de quantifier la plausibilité du rôle de la sélection en associant une *p-value* à chaque SNP testé.

Nos travaux basés sur des simulations individu-centrées ont montré l'intérêt de notre méthode afin d'inférer la sélection en temps réel à partir de données E&R :

*(i)* Il est possible d'identifier des empreintes laissées par la sélection compatibles avec la densité en marqueurs permise par les projets de génotypage actuels.

*(ii)* Les SNP sélectionnés sont détectés même si ceux-ci étaient déjà à fréquence élevée dans la population avant la sélection.

*(iii)* Les signatures de sélection peuvent être repérées à très court terme (une poignée de générations peut suffire).

*(iv)* La sélection peut même être inférée dans des populations de très faible effectif efficace ($N_e \leq 50$) si son intensité est suffisamment importante.

*(v)* Le taux de faux-positifs demeure faible dans toutes les configurations examinées pour le régime de sélection qui nous intéresse. En particulier, un biais dans l'estimation de $N_e$ semble n'avoir qu'un effet modéré (en cas de forte surestimation) sur le taux de faux-positifs.

*(vi)* Les coefficients de sélection sont correctement estimés si l'intervalle entre deux échantillons temporels est modéré (une dizaine de générations). Lorsque l'intervalle est plus important, il est encore possible d'obtenir une estimation correcte en ajoutant un échantillon dans la série génétique temporelle.

Ce travail a permis de préciser les conditions optimales d'utilisation de notre méthode selon les contraintes pesant sur l'analyse. Dans un manuscrit intitulé « *The power to detect selection from short-term Evolve and Resequence experiments* » soumis au journal *Molecular Biology & Evolution* (et ci-après désigné « Article I »), nous avons proposé des résultats synthétiques de nos simulations permettant à d'éventuels utilisateurs d'analyser des séries temporelles issues d'une expérience de sélection en toute connaissance de cause. Le texte de la version de soumission, suivi des tableaux et figures, est disponible dans la section suivante du présent chapitre.

En plus de la présentation de la méthode et des résultats de simulation, l'Article I inclut des résultats issus d'une application sur données réelles. Il s'agit de données déjà publiées (Epstein *et al.*, 2016a, 2016b) ayant permis la mise en évidence de deux régions génomiques candidates à une forte sélection contemporaine chez le diable de Tasmanie (*Sarcophilus harrisii*). Un objectif était de tester la capacité de notre méthode à identifier ces deux régions. Nos résultats ont effectivement permis de retrouver les deux régions candidates, mais aussi d'en identifier d'autres, suggérant une empreinte de la sélection contemporaine dans le génome du diable de Tasmanie plus étendue que ce l'on pensait jusqu'alors. Compte tenu du nombre important de régions candidates, leur analyse détaillée dépasse le cadre de l'Article I et fait partie du prochain chapitre (Chapitre IV) consacré à l'évolution rapide des populations de diable de Tasmanie sous l'effet d'un cancer transmissible.

## — 2. Article I: The power to detect selection from short-term Evolve and Resequence experiments

# The power to detect selection from short-term Evolve and Resequence experiments

Jean-Noël Hubert and Frédéric Hospital

GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

Corresponding author: Jean-Noël Hubert (jean-noel.hubert@inra.fr)

## Abstract

Evolve and Resequence (E&R) experiments associate experimental evolution with Next Generation Sequencing technologies to generate valuable temporal data for exploring biological adaptation. Exploiting genomic samples separated by as few as ten generations in small populations submitted to selection is still a challenge though such opportunities are becoming frequent. An important issue lies in the ability to discriminate between the effects of selection and those of genetic drift. Here we address this concern by coupling a diallelic Wright-Fisher model to a maximum-likelihood approach to follow the variation of allele frequencies. We study the performances of such a strategy through various forward simulation scenarios. We show that our approach can detect selected SNP on the very short-term in small populations undergoing strong directional selection. In particular, our approach detects signatures of selection from standing genetic variation, and provides accurate estimates of the selection coefficient. In addition, we performed a genome scan on E&R publicly available data from Tasmanian devil populations facing the spread of the Devil Facial Tumor Disease (DFTD), a transmissible cancer. Several candidate-genes putatively mediating cancer onset or progression (*CBL*, *DAPK2*, *KLF10*, *MCAM*, *THY1*, *USP2*…) could be identified through our approach, supporting the recent report of a rapid evolutionary response to DFTD. Our results highlight the importance of short-term E&R experiments for further characterizing the underlying architecture of biological functions. We also provide guidelines to design and optimize such experiments.

# Introduction

Experimental evolution ranks as a valuable field for testing evolutionary theory (Bennett and Lenski 1999), ranging today from analyzing the effects of a single generation of selection (Christie et al. 2016) to considering strategies for empirically investigating macroevolution (Bell 2016). A guiding principle consists in measuring the effects of real-time evolution on a population submitted to a specific environment. By offering different control levels – on the ancestral population, on the progeny or on the environment –, experimental evolution helps examine theoretical predictions. Important evolutionary issues have been discussed from suitable experiments, such as the potential adaptive impact of strong genetic drift (Keightley and Caballero 1997 ; Matute 2013), the processes associated to thermal adaptation (Zhao et al. 2015 ; Hangartner and Hoffman, 2016) or the evolution of senescence (Reznick et al. 2004 ; Zwoinska et al. 2017). This highlights the versatility of the field, which provides a wealth of adjustable experimental resources to surround and test hypotheses about evolution.

Selection experiments can be organized in four categories: artificial selection, laboratory culling, laboratory natural selection, and field experiments (Garland 2003). In addition, experimental setups can also be distinguished according to the timescale investigated. Long-term studies typically make use of model microorganisms, which offer large population sizes, short generation intervals and the potential for many replications (Matute 2013 ; Morgan et al. 2014 ; McGuigan et al. 2015 ; Boyle et al. 2017 ; Graves et al. 2017). Specifically, such experiments aim to assess and quantify the relative impact of tangled evolutionary factors on the processes affecting the long-term fate of populations (Travisano 2009). Long-term setups can cover several thousands of generations (Lenski and Travisano 1994), but there is no standard duration since some long-term studies have been carried out on various timescales (Keightley and Caballero 1997 ; Morgan et al. 2014 ; McGuigan et al. 2015 ; Scanlan et al. 2015 ; Smukowski Heil et al. 2017). As a contrast, short-term studies focus on the immediate evolutionary trajectory of a given phenotype and aim at assessing its potential to evolve (Travisano 2009). Non-model and larger organisms with smaller population sizes can be considered (Kraaijeveld-Smit et al. 2006 ; Cooke et al. 2007 ; Konczal et al. 2015), and the duration of the selection experiment is then comprised between a single (Christie et al. 2016) and less than one hundred generations for most experiments (Travisano 2009).

At a time when assessing the impacts of human activities on natural populations is being raised as a major priority, the analysis of short-term field experiments provides fundamental

indications about the evolution of local populations currently facing an alteration of habitat (Hendry et al. 2006 ; Epstein et al. 2016a ; Papetti et al. 2016). In parallel, short-term artificial selection experiments emancipate from ecological reality and represent a very powerful mean for investigating biological functions. Especially, research on livestock species has implemented valuable selection protocols for detecting phenotypic changes at different levels of biological organization (Hocquette et al. 2012). Some experiments directly look after highly integrated phenotypes, such as behavior (Minvielle et al. 2002 ; Lattorff and Moritz 2013), organismal performances (Cooke et al. 2007 ; Hiramatsu et al. 2017) or production traits (Le Boucher et al. 2012 ; Zan et al. 2017). Some others will rather track changes at much lower integration levels, using in some cases the concentration of a single target molecule or ion within a particular organ as the main selection criterion (Griffin et al. 1989 ; Alnahhas et al. 2014). Therefore, short-term studies in large multicellular organisms undergoing strong directional selection provide a large collection of scenarios to examine the evolution of complex forms and functions from a single molecule to populations in their ecological context.

The interest of short-term selection experiments is particularly strengthened by the current spread of Next-Generation Sequencing (NGS), making it possible to track the trajectories of alleles at some loci as they are being targeted by contemporary selection. Mapping such causative loci is the very purpose of the growing area of Evolve & Resequence (E&R) experiments, which associates experimental evolution with NGS (Turner et al. 2011). E&R experiments produce genomic time-series and therefore initiate the opportunity of looking for signatures of selection in real time, as populations are being evolved, by measuring the changes in allele frequencies (Long et al. 2015 ; Schlöterrer et al. 2015). First E&R studies were analyzed through the modification of existing approaches, mostly based on contingency tables (Orozco-terWengel et al. 2012 ; Schlöterrer et al. 2015) or direct comparisons of allele frequencies (Turner et al. 2011 ; Jah et al. 2015 ; Kessner and Novembre 2015 ; Konczal et al. 2015 ; Schlöterrer et al. 2015 ; Epstein et al. 2016a). However, there is a need for specific data analysis methods and guidelines for maximizing the ability to determine the genetic basis of the investigated traits, so that the enormous potential of experimental evolution for understanding complex phenotypes can be better exploited.

Recent studies focus on the existing strategies for detecting candidate-genes from E&R experiments and for improving experimental designs (Kofler and Schlöterrer 2014 ; Schlöterrer et al. 2015). Simulation studies have preferably investigated large population sizes ($N_e \sim 1000$) or long-term (over 100 generations) scenarios (Turner et al. 2011 ; Orozco-terWengel et al.

2012 ; Baldwin-Brown et al. 2014 ; Kofler and Schlöterrer 2014 ; Long et al. 2015 ; Kessner and Novembre 2015 ; Schlöterrer et al. 2015), as frequently found in laboratory evolution experiments. Yet, the case of short-term E&R experiments with strong selection in small populations of large organisms have remained very little explored. This is the topic of the present paper.

We aim to extend the current methodological effort by characterizing a suitable strategy for exploiting E&R experiments that implement short-term (from a single to some tens of generations) and strong ($s \geq 0.1$) directional selection on small populations ($N_e \sim 50$). Considering small effective population sizes ($N_e$) means the amount of genetic drift is expected to be large. Consequently, a particularly challenging issue lies in the ability to discriminate between drift alone and drift plus selection when measuring a shift in allele frequency. We address these concerns by coupling a diallelic Wright-Fisher model to a maximum-likelihood approach. Probabilities of observing a particular shift in allele frequency, knowing $N_e$ and the number of generations of selection, are compared with ($H_1$) and without ($H_0$) incorporating selection in the model through a Likelihood Ratio Test. In addition to providing an objective criterion for disentangling the effects of selection from those of drift, such an approach allows estimating the selection coefficient.

We present below performances obtained by implementing our approach for identifying selected loci and for estimating selection coefficients in forward simulation scenarios. In particular, we investigated the effects of the variation of relevant population parameters for experimental biologists that may wish to design and analyze E&R experiments. We also present the results of an application of our method to real data from a recently published short-term E&R study in Tasmanian devil (*Sarcophilus harrisii*) populations (Epstein et al. 2016a). In light of all these results, we provide guidelines to exploit E&R experiments through our approach.

## New approaches

### *Wright-Fisher models*

Our objective is to offer a method for detecting signatures of selection from genetic short-term time-series collected in small populations undergoing strong selection. The logic behind our approach is to compare two Wright-Fisher models, one including only drift and the other considering drift plus selection. We do not account for mutation since such an event is highly

unlikely to occur on the short-term. As well, investigated populations are considered to be closed and we do not account for migration. At the genomic level, we consider independent diallelic loci (SNP) in diploid populations with discrete generations.

Let $a_1$ and $a_2$ the two alleles at an investigated SNP before any evolution. Our null hypothesis ($H_0$) is that the only driving force for allele frequency change is the random sampling of reproducing individuals due to constant finite size, in other words genetic drift. Under a Wright-Fisher model including drift only, the probability of getting $X_{n+1}$ copies of $a_1$ after one generation of evolution is provided in Equation 1.

$$\Pr(X = X_{n+1} | X_n, N_e) = \binom{2N_e}{X_{n+1}} p^{X_{n+1}} (1-p)^{2N_e - X_{n+1}} \quad (1)$$

where $X_n$ is the number of copies for the allele of interest ($a_1$) at generation $n$, $N_e$ the effective population size, and $p$ ($= X_n/2N_e$) the frequency of $a_1$ at generation $n$.

If we consider selection, the model includes a new parameter, the selection coefficient ($s$). Here we consider the two alleles $a_1$ and $a_2$ as codominant, $a_1$ being the beneficial one. As a consequence of directional selection, genotypes $a_1a_1$, $a_1a_2$ and $a_2a_2$ will have relative fitnesses $\omega_{11} = 1 + 2s$, $\omega_{12} = 1 + s$, and $\omega_{22} = 1$, respectively. Frequency of $a_1$ at generation $n$ (after selection occurred) then becomes $p'$ (Equation 2).

$$p' = \frac{p(p\omega_{11} + q\omega_{12})}{\bar{\omega}} \quad (2)$$

where $\bar{\omega}$ denotes the average fitness (Equation 3).

$$\bar{\omega} = p^2\omega_{11} + 2pq\omega_{12} + q^2\omega_{22} \quad (3)$$

Thus, in case of selection ($H_1$), the probability of having $X_{n+1}$ copies of $a_1$ at the next generation is obtained by replacing $p$ by $p'$ in Equation 1.

$$\Pr(X = X_{n+1} | X_n, N_e, s) = \binom{2N_e}{X_{n+1}} p'^{(X_{n+1})} (1-p')^{2N_e - X_{n+1}} \quad (4)$$

The idea is to find the most suitable Wright-Fisher model, between the first that includes only drift ($H_0$, Equation 1) and the second that includes both drift and selection ($H_1$, Equation 4), to describe the evolution of $a_1$ over time.

### *Comparison of models through Likelihood Ratio Test*

We now consider as an example a SNP of interest after an E&R experiment implementing directional selection over 10 generations. A time-series of two points, before (at generation $G_0$) and after (at generation $G_{10}$) selection, gives the numbers of copies $X_0$ and $X_{10}$, respectively, for $a_1$. The likelihood function that describes the temporal change in $a_1$ according to our models is given in Equation 5 if we do not consider sampling.

$$L(s) = P(X_{10}|X_0, N_e, s) \quad (5)$$

with $s = 0$ in case of $H_0$ (absence of selection) and $s = \hat{s}_{ML}$ in case of $H_1$ (presence of selection). $\hat{s}_{ML}$ is the value of the parameter $s$ that maximizes the likelihood function given the data in case we postulate selection ($H_1$). In practice, the selection coefficient is maximized numerically for positive values ranging from 0 to 1, and negative selection is addressed by symmetry (*i.e.*, positive selection on the other allele). When the allele of interest is fixed in a time sample, it is never possible to determine exactly at which generation the fixation did occur (*e.g.*, if there are only two samples, $G_0$ and $G_{10}$, and the allele is fixed in $G_{10}$, one cannot decide whether fixation took place exactly at $G_{10}$ or at any previous generation). In such case of lack of information, the estimate $\hat{s}_{ML}$ can be very high, because stronger selection and earlier fixation is always more likely. To avoid that, we set a limit of $\hat{s}_{ML} = 1$, which gives more realistic results in case of fixation. Note however that those results were still incorporated in the means of the estimates. As a consequence, estimates are more prone to bias when the probability of fixation is higher (*e.g.*, in case of small $N_e$). In addition, we assume that $N_e$ is constant over time. Getting the likelihoods requires computing transition probabilities over 10 generations under each Wright-Fisher model (with and without selection).

In case of sampling, the likelihood function will display the probabilities of observing a given number of copies of $a_1$ due to a sampling event from the distribution of its allele frequency in the population (Equation 6).

$$L(s) = \sum_{p_0} \sum_{p_{10}} P(p_0|N_e, s) P(X_0|p_0, N_{s0})\, P(p_{10}|p_0, N_e, s) P(X_{10}|p_{10}, N_{S10}) \quad (6)$$

where $N_{s0}$ and $N_{S10}$ are the sample sizes, and $p_0$ and $p_{10}$ the allele frequency of $a_1$, at generations $G_0$ and $G_{10}$, respectively.

Williamson and Slatkin (1999) already used an equation very similar to Equation 6, except here we use the complete matrix of transition between all genotypes, and do not group them into "allele configurations" hence providing a gain in accuracy, but very small. As suggested by these authors, we also assumed that the distribution of initial allele frequencies $p_0$ is uniform, giving the probability $P(p_0|N_e, s)$. All others factors of Equation 6 consist of binomial probabilities that can be computed from Equations 1 and 4, by replacing the effective population size ($N_e$) by sampling sizes ($N_{s0}$ or $N_{s10}$) when necessary.

Furthermore, we may integrate data from intermediate sampling points if available. If we consider a third sample collected at generation $G_{int}$ in the interval $]G_0 ; G_{10}[$, the likelihood function may be written as in Equation 7.

$$L(s) = \sum_{p_0} \sum_{p_{int}} \sum_{p_{10}} P(p_0|N_e, s)P(X_0|p_0, N_{s0})\, P(p_{int}|p_0, N_e, s)P(X_{int}|p_{int}, N_{Sint})$$

$$. P(p_{10}|p_{int}, N_e, s)P(X_{10}|p_{10}, N_{S10}) \quad (7)$$

where all symbols ($X_{int}$, $p_{int}$, $N_{sint}$ ) related to the intermediate temporal point are subscripted with *int*.

The relative adjustment of the two Wright-Fisher models (with and without selection) to the data is quantified through a Likelihood Ratio Test (LRT). The LRT statistics is computed for each SNP (Equation 8) and provides an objective criterion to discriminate between drift alone ($H_0$) and drift plus selection ($H_1$).

$$LRT = -2ln\left[\frac{L(s=0)}{L(s=\hat{s}_{ML})}\right] \quad (8)$$

The significance of the test is assessed by assuming that the LRT follows a $\chi^2$ distribution with one degree of freedom (Wilks' theorem) under the hypothesis of absence of selection ($H_0$).

To assess the ability of this approach to infer selection from short-term E&R experiments in small populations, we performed forward simulations with varying key population parameters (see Materials and Methods). Unless stated otherwise, we applied a Type I error threshold ($\alpha$) of 0.01 to the LRT test, because such a value was shown enough to limit the expected false-discovery rate in the multilocus case for several simulation scenarios (data not shown). We suggest using a FDR method such as the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) for any application on real genomic data.

### *Availability of the method*

Our method is available in the form of an R implementation at https://github.com/hubert-pop/signasel.

## Results

### *Detecting selective sweeps*

When performing a genome scan looking for a causal variant, genotyping is likely to miss the truly selected locus. Even though not directly targeted by selection, the variation surrounding a causal variant may harbor signatures of selection due to linkage disequilibrium (LD) through the process of selective sweep. Finding candidate-loci for selection therefore requires a statistics able to identify patterns of selective sweep from genomic data.

We assessed the ability of our approach to reveal such a pattern by simulating the evolution of a 50 cM-long portion of chromosome over 10 generations in a population with an effective size ($N_e$) of 50. The portion of chromosome consisted in one selected SNP located at one end, and 50 neutral SNP. The whole SNP were regularly spaced (one locus every cM along the chromosome), and their allele frequency before evolution was 0.5. In addition, alleles were in total linkage disequilibrium before evolution, meaning that we had two haplotypes each composed of 51 polymorphisms in the starting population. We considered two scenarios: either selection started immediately for 10 generations, or the population was randomly reproduced for 3 generations prior to the 10 generations of selection, in order to reduce the initial LD. In any case allele frequencies at all loci were recorded twice over time: just before the first generation of selection and after the last ($10^{th}$) generation of selection. Our approach for detecting signatures of selection was applied to the *in silico* genomic time-series generated from these two samples.

Figure 1 shows the proportion of significant likelihood ratio tests, hereafter designated as power, obtained for each locus tracked along the simulated portion of chromosome, according to a false-discovery rate (FDR) of 0.1 (Benjamini and Hochberg 1995). Each time an allele had been targeted by selection (s = {0.2, 0.4, 0.7}), the power was maximal at the selected position, and then decreased as the genetic distance from the selected site increased, as expected. A reduction in power of 50 % occurred within 10 cM from the selected site. When all loci were neutral ($s$ = 0), the null hypothesis (absence of selection) was largely favored whatever the position. In other words, a neutral marker associated to a selected locus was almost always given a better power than a neutral marker associated to a neutral locus. In addition, the amount of LD existing between the selected locus and the neighboring neutral polymorphisms when selection starts also affected the power. Reducing LD via three preliminary events of recombination actually lowered the power of the test, this effect being proportionally more pronounced for larger selection coefficients.

We showed the ability of our approach to efficiently tell strong selection from strong drift and to reveal patterns of selective sweep from genomic samples separated by 10 generations. Maximizing the power to detect signatures of selection would require SNP in high LD with a strongly selected causal locus.

*Detecting standing genetic variation*

Directional selection targets mainly standing genetic variation on the short-term in small populations (Barrett and Schluter 2008). We investigated the effect of the variation of a preexisting beneficial allele by considering the selected SNP at an initial frequency ($f_0$) ranging from 0 to 1 in a population of $N_e = 50$. Then, we recorded the frequency of the beneficial allele after 10 generations of selection. We assessed the ability of our approach at identifying the SNP of interest and estimating the selection coefficient applied to this SNP.

Intermediate initial allele frequencies increased the probability of our test to identify a positively selected allele (Fig. 2A). The power represents the proportion of significant likelihood ratio tests at the investigated SNP with $\alpha = 0.01$ over 500 simulations. When selection is very strong ($s \geq 0.4$), our approach detected the selected allele in at least 75% of the simulations if $f_0$ was comprised between 0.1 and 0.6. For the same frequency range, the power remained over 25% for the smallest value of $s$ examined ($s = 0.2$). Our approach allows detecting with good power estimates the variants that were already common in the starting population, on condition that they are selected strongly enough during the experimental evolution.

A limit to the identification of selection exists for high values of $f_0$ ($f_{0\_lim} = 0.82$ here). Above this limit, no detection was possible, even if the investigated allele was strongly selected. In such a case, the preeminence of the allele is initially so high that even its fixation does not allow concluding that the effect of selection is significant.

The selection coefficient $s$ was accurately estimated in any case if $f_0$ was comprised between 0.1 and 0.6 (Figs. 2B and 3). Beyond this interval, the accuracy of estimation was lower as $f_0$ was close to fixation. The test still provided correct estimates for weaker selection coefficients ($s < 0.1$), preferably when $f_0$ was close to 0.5 (data not shown). Interestingly, our approach allows good performances in estimating a wide range of selection coefficients when the beneficial allele is at intermediate frequency before selection.

We showed here that our approach allows jointly detecting variation favored in the course of short-term selection and estimating the selection coefficient associated to the identified SNP. Importantly, the test performs well even if the selected allele was already common in the population before selection. In particular, the best conditions for maximizing both the power of detection and the accuracy of the estimation of $s$ are obtained when the allele frequency is initially comprised between 0.1 and 0.6.

*Impact of the effective population size*

Because our approach relies on a Wright-Fisher model, an estimate of the effective size ($N_e$) of the investigated population is required as an input parameter. We therefore examined the effect of variations in population size on the performances of our approach by simulating a beneficial allele submitted to selection over 10 generations in small populations with $N_e$ ranging from 35 to 200. Figure 4 shows the power to detect a beneficial allele as a function of $N_es$, integrating the impacts of both the effective size ($N_e$) and the selection coefficient ($s$). For the 4 values of $N_e$ assayed, the power was over 90% when $N_es \geq 30$. Our method seems therefore well adapted for detecting strong $N_es$ (effective selection, see for example Lawrie and Petrov 2014) from short-term (10 generations) E&R studies. Interestingly, at constant $N_es$, power was shown to be higher for smaller $N_e$. A possible strategy to maximize the power to detect candidate-SNP from E&R experiments could therefore theoretically be to reduce the fraction of reproducing individuals, even if the investigated population has a small $N_e$.

The performances of the estimators of $N_e$ from genetic data are currently a matter of discussion, including in the conditions of controlled demography (Gilbert and Whitlock 2015), as it is often the case in experimental evolution. We measured how the effects of errors on the estimate of $N_e$ could impact the performances of the test. We considered that the true $N_e$ was the size of the population generated by simulation. Then, the genetic time-series generated through simulation (true $N_e = 50$) were analyzed with our approach by inputing a $N_e$ from 10 to 100. Results indicate that an underestimation of $N_e$ would cause a decrease in power, whereas an overestimation would translate into a gain in power of the test (Fig. 5). The loss of power can be important, depending on the extent of the underestimation of $N_e$ and on the strength of the selection. In particular, we see that the power was completely lost when $N_e$ was estimated to be 15 whereas the true value for the investigated population was 50. The false-positive rate ($s = 0$) also increased as $N_e$ was overestimated, but remained below 4% even when $N_e$ was considered equal to 100 instead of 50. The tendency to outperform when $N_e$ is overestimated is due to the fact that larger $N_e$ are favored by the test (Fig. 6). For a given initial allele frequency ($f_0$), the required shift in frequency to identify a SNP as selected after 10 generations of directional selection ($\Delta f_{10}^*$) was actually smaller as $N_e$ increased. For instance if $f_0 = 0.5$, a shift in frequency $\Delta f_{10}^* = 0.35$ was required to reach significance if $N_e = 50$, whereas $\Delta f_{10}^* = 0.275$ for $N_e = 100$.

Thus, a bias in $N_e$ can affect the performances of the test, particularly in the case of an underestimation, which can strongly reduce the power to detect a selected SNP. Overestimating $N_e$ has the opposite effect but seems less problematic, as long as the false-discovery rate is efficiently controlled.

### *Impact of sampling*

Keeping costs at a minimum in analyses of signatures of selection often implies genotyping samples rather than full populations. We investigated the impact of various sample sizes ($N_s$ from 10 to 50) on the performances of our method in detecting a selected SNP from a 10-generation-long experiment in a population of $N_e = 50$.

Figure 7 shows how much sampling is likely to reduce power compared to the ideal case where the whole population is genotyped ($N_s = N_e = 50$). The impact of sampling on power is very limited ($\leq 6\%$) when $N_s$ is 30 or greater. At maximum, the power dropout due to sampling reached 22% for the smallest sample size investigated ($N_s = 10$, $s = 0.4$). In addition, sampling did not inflate the too much false positive rate, which reached a maximum of 3% for the smallest sample size ($N_s = 10$).

The effect of sampling on estimates of the selection coefficient is shown on figure 8. A tendency to overestimate the highest selection coefficients ($s = 0.4$ and $s = 0.7$) is visible for the smallest sample size investigated ($N_s = 10$). However, for $N_s = 20$ and higher, there is no major difference in estimating $s$ between the cases of sampling and the ideal situation ($N_s = N_e = 50$).

Thus, simulations indicate that inferring selection is here still possible with a small sample size ($N_s = 10$, that is to say 20% of the $N_e$ value). A good compromise between performances and genotyping costs would be to get a sample of 20 individuals (40% of $N_e$), what seems to guarantee that power and accuracy of estimates of the selection coefficient are close to those of the ideal case of computing statistics with the whole effective population.

### *Impact of the number of generations of selection*

Because it affects the duration and the cost of an E&R experiment, the number of generations of selection is a critical parameter. A well-reasoned choice is here of great importance, in particular in species for which the generation interval is long or the maintenance of breeding lines expensive. We therefore investigated the effect of the number of generations of selection on the performances of our method, under pure genetic drift ($s = 0$) or drift plus strong selection ($s = 0.4$), for 4 different levels of drift – $N_e = 35$, 50, 100 and 200.

Simulations show that the selection coefficient is overestimated when there were more than 10 generations of selection between the two temporal samples (Fig. 9A). This bias is due to an increased probability of fixation and was consequently more pronounced in populations with small $N_e$. An interval of 10 generations of selection was enough to get power over 80%, except

for the smallest population ($N_e$ = 35) that required 20 generations (Fig. 9B). Note that a very small number of generations (less than 5) can be enough to detect selected SNP with a very high power when $N_e \geq 100$. In addition, the false-positive rate remained very limited even after 20 generations whatever $N_e$.

We focused more specifically on a period of 20 generations of selection because such an interval was shown adapted to get enough power to detect selection in a wide range of situations. Of course, analyzing E&R experiments of more than 20 generations is possible (Table 1). However, the difficulty of disentangling the effects of drift alone and drift plus selection on allele frequency becomes more acute as the number of generations of selection increases. It is here necessary to take into account the existence of an upper limit in the number of generations that can be analyzed. In fact, the maximal number of generations between two consecutive samples in a time-series depends on the initial frequency of the examined SNP and on the $N_e$ of the population (Fig. 10). Beyond this limit, selection, whatever its strength, is no longer detectable with $\alpha = 0.01$ and two temporal samples. This point must be taken into account before analyzing E&R data, or even better, when choosing the time interval between two samples in the experimental design.

### *Impact of intermediate sampling*

So far, we considered the analysis of E&R experiments from two samples only, at $G_0$ (*i.e.*, before selection) and at the last generation of the experiment. However, there may be an advantage in considering intermediate sampling points, because it should give additional information about the trajectory of the SNP frequency in the course of the experiment. We therefore assessed to what extent adding extra temporal samples could improve the performances of our method.

We considered a population of $N_e$ = 50 subjected to 20 generations of strong selection ($s \geq 0.1$). We collected genotypes at each of the 21 generations (from $G_0$ to $G_{20}$ included) of selection. We could detect no clear improvement in the power of our method to detect selection, whatever the simulation settings, from adding extra temporal samples (data not shown). However, using more than 2 samples permitted to correct the bias in the estimation of *s* that may occur when the number of generations explored is greater than 10. We investigated more in detail the benefit of extra temporal samples on the estimation of *s* in the same conditions as for Fig. 9 ($s$ = 0.4). Figure 11 displays a comparison between three sampling strategies. As previously mentioned (Fig. 9A), the 2-sample strategy overestimated s ($\hat{s}$ = 0.81). Sampling all 21 generations

substantially reduced the bias ($\hat{s}$ = 0.44). The 3-sample strategy gave intermediate results. Interestingly, results with three samples were very similar to those with 21 samples when the third sample was taken before $G_{10}$. In such a case, the best estimates ($\hat{s}$ = 0.45) were actually provided when picking the third sample at generations $G_4$, $G_5$ or $G_6$. Thus, considering a single extra sample may be optimal to estimate the selection coefficient. This strategy was here tested for a 20-generation-long E&R experiment, but is still valid for longer experiments (Table 1).

### *Application to the Tasmanian devil dataset*

In addition to our simulation work, we applied our method to genomic time-series collected from Tasmanian devil populations (Epstein et al. 2016b) during the spread of Devil Facial Tumor Disease (DFTD), an infectious cancer associated to high mortality rates (Loh et al. 2006). Results issued from our method are consistent with the original findings of Epstein et al. (2016a). Briefly, our analysis allowed identifying 84 scaffolds harboring putative selection signals, including the two signatures of contemporary selection previously identified in Tasmanian devils (Epstein et al. 2016a). Here we retained only regions hosting at least one SNP with $P < 0.0001$ or several co-localizing SNP with $P < 0.01$, which corresponds to a FDR of about 13%. Most identified selection signals seem population-specific, suggesting that the three populations might have evolved responses to DFTD relying on different genes.

Only one candidate-region was shared among the three populations according to our analysis, on chromosome 2. Detected SNP stand at less than 100 kb of an ortholog of the *CRBN* gene (Fig. 12), which exactly matches the first candidate-locus identified in the study of Epstein et al. (2016a). CRBN is able to bind immunomodularity drugs (IMiD) and is seen as a predictive biomarker for myeloma therapy outcome (Schuster et al. 2014). Polymorphism in this region, which also includes an ortholog of *TRNT1*, has otherwise been shown to be associated to moderate mental disabilities in children (Papuc et al. 2015). In addition, we could retrieve the second signature of selection revealed by Epstein et al. (2016a) in an ortholog of the *MFRP* gene on chromosome 3 (Fig. 13). The *MFRP* gene and its close neighbor *C1QTNF5* are mainly related to ophthalmic disorders in mammals (Kameya et al. 2002). However, as highlighted in Epstein et al. (2016a), the four other genes located in the immediate vicinity of the selection signal are probably better candidates here, since they are orthologous to *THY1*, *USP2*, *MCAM* and *CBL*, all of them being involved in cancer progression (Schlagbauer-Wadl et al. 1999 ; Lung et al. 2005 ; Kales et al. 2010 ; Metzig et al. 2011). The identification of these two loci

illustrates the ability of our method to locate known candidate-genes in small real populations submitted to short-term directional selection, as suggested by simulations.

Additionally, we could identify other interesting selection signals. Our investigation of the complete set of candidate regions is still under progress, but we have already characterized relevant signatures of selection in each Tasmanian devil population examined. In particular, a strong selection signal was detected in a close ortholog to the *DAPK2* gene on chromosome 1 in the FN population, and to a lesser extent in the WP population (Fig. 14). This gene belongs to the DAP-kinase (Death Associated Protein kinase) family that encodes serine-threonine kinases involved in the transmission of pro-apoptotic signals (Shohat et al. 2002 ; Shiloh et al. 2014). *DAPK2* is the smallest member of the family and was first identified as an inducer of membrane blebbing (Inbal et al. 2002 ; Shoval et al. 2011), an early key event for cells undergoing apoptosis. Other studies confirmed that *DAPK2* was a regulator of apoptosis, but its precise roles in the apoptotic pathways remain unclear and probably cell context-dependent (Schlegel et al. 2014 ; Geering et al. 2015). For instance, *DAPK2* plays a dual role in death signaling pathways, by displaying both pro-apoptotic (Britschgi et al. 2008 ; Britschgi et al. 2010) and anti-apoptotic behaviors (Schlegel et al. 2014). *DAPK2* has also been shown to promote other functions related to the control of cell fate, such as autophagy or differentiation (Bialik and Kimchi 2006 ; Britschgi et al. 2010 ; Geering et al. 2015). Three other genes (*HERC1*, *MRPL46* and *FAM96A*) stand in the immediate vicinity of the selection signal located within *DAPK2*. *HERC1* plays a role in intracellular membrane trafficking (Garcia-Gonzalo et al. 2003) and is involved in several phenotypes, including nervous system disorders (Bachiller et al. 2015 ; Nguyen et al. 2016) and cancer progression (Garcia-Gonzalo et al. 2003 ; Goto et al. 2011). *MRPL46* encodes a mitochondria ribosomal protein that may be upregulated in cancer (Sotgia et al. 2012). *FAM96A* is known for its key role in the regulation of iron metabolism (Stehling et al. 2013), but has also recently been suggested as a tumor-suppressor (Schwamb et al. 2015 ; Zhang et al. 2017).

Another example of selection signal most probably related to cancer was found in the NP population on chromosome 2 (Fig. 15). Detected SNP are located at 4 kb of an ortholog of the *KLF10* gene, in the middle of a large region without any other known protein coding gene. Krüppel-Like Factors (KLFs) encode DNA-binding transcriptional regulators of the cell cycle. They also play a role in other signaling pathways with well-known oncogenic alterations. In total, alterations of KLFs are characterized in at least 23 tumor types (Tetreault et al. 2013). Antitumor effects of *KLF10* are well documented. For instance, the inactivation of *KLF10*

highlighted its tumor suppressor activity in normal cells through the control of the G1/S-phase transition (Song et al. 2012). Actually, KLF10 activates CDKN1A/p21$^{CIP1}$, which is a pivotal inhibitor of the cyclin E-cdk2 complex that promotes DNA replication. *KLF10* has also been shown to play a key role in estrogen-induced apoptosis in breast cancer (Hsu et al. 2011), to be part of the antiproliferative action of TGF-β signaling in hepatocellular carcinoma (Jiang et al. 2012) and to limit the proliferative effects of H-RAS mutations (Song et al. 2012), found in many cancers. Although the critical role of the KLF family in cancer is recognized, the extent of their involvement is not yet fully understood. Individual KLFs may have a dual role in cancer progression, depending on the cell context (Tetreault et al. 2013 ; Weng et al. 2017). In particular, recent works suggested potential tumorigenic effects for KLF10 (Heo et al. 2015, 2017). It should also be noted here that *KLF10* affects immunity through the modulation of regulatory T cells (Venuprasad et al. 2008), which themselves have a dual role towards cancer progression (Tetreault et al. 2013). Given its demonstrated multiple links with cancer, *KLF10* is a very interesting candidate for selection in Tasmanian devils.

Finally, each population harbored a strong signal of selection in a locus reported to encode known regulators of cancer progression. The identification of these candidate-genes demonstrates the ability of our method to retrieve recently published candidates and furthermore to extend the list of genes potentially under contemporary selection in Tasmanian devils. This also highlights the potential of short-term field E&R experiments to look for footprints of selection.

# Discussion

## *Identification of short-term selection signals in small experimental populations*

In the present study, we suggested a method for detecting traces of selection from short-term (from a single to some tens of generations) E&R experiments involving small populations (from $N_e$ = 30 to 200) subjected to strong selection ($s \geq 0.1$). We reviewed through simulation the impact of the variation of key parameters on the ability of our method to identify reliable candidate-genes. By successively controlling these parameters, we identified conditions that promote the inference of selection and some particular features of our method.

The first important feature of the method is its ability to identify selection from standing genetic variation, which may be decisive in the short-term evolutionary response to a novel challenge

(Barrett and Schluter 2008). This concern is particularly relevant in the case of field experiments performed in vertebrates, for which the demographic events that shaped contemporary populations are difficult to assess and the genetic composition of the starting populations is generally not under control (Irschick and Reznick 2009 ; Bradic et al. 2012 ; Lillie et al. 2014 ; Norén et al. 2014 ; Konczal et al. 2015). The good news is that our method offers best and relatively stable performances when dealing with common variants. The vast majority of the simulation results presented here considered common variants, highlighting the propensity of our method to target selection from standing genetic variation in small populations. Importantly, our method works best on a timescale of 10 generations, which is optimal to catch standing genetic variation (*i.e.*, the exclusive raw material for selection at this timescale).

Another important point is the ability of our method to target selection in very small populations. This helps address a major issue in the inference of directional selection: to what extent can we discriminate between the effects of drift and drift plus selection? Our work illustrated the potentiality of identifying patterns caused by selection when $N_e$ reaches as low values as 30. Successfully detecting SNP subjected to selection despite very small population sizes is actually possible here, if the selection is strong enough and the number of generations of selection is adapted (see Table 1). Such combinations of population parameters match very well to available breeding lines from livestock species submitted to artificial selection setups (Griffin et al. 1989 ; Minvielle et al. 2002 ; Toro et al. 2011 ; Le Boucher et al. 2012 ; Alnahhas et al. 2014). Our work therefore offers a way to exploit genomic time-series from livestock selection experiments, what should help better understand the genetic architecture of complex phenotypes such as behavior or production traits.

A third key feature of our method lies in its high power to detect selection on extremely short timescales. Our results showed in particular that less than 5 generations of selection may be enough to detect strongly selected SNP when $N_e$ is at least 100. Classical laboratory animal models, like Drosophila, allow maintaining such levels of genetic diversity in the course of selection experiments (Promislow et al. 1998 ; Orozco-terWengel et al. 2012), due to the greater ease of keeping relatively large numbers of individuals at each generation in these species. Simulations indicated that genomic time-series gained from such laboratory populations over a very limited amount of time (from a couple to no more than 20 generations, what would corresponds to several weeks to less than 1 year of experiment in Drosophila) can provide reliable gene discovery.

Finally, simulations showed that our method was well adapted to detect selection from short-term E&R experiments whatever the category of selection experiment implemented (Garland 2003). When a particular constraint cannot be directly circumvented (*e.g.*, a very small $N_e$ due to prior demographic events), experimenters can still gain flexibility through other population variables (*e.g.*, the fraction of reproducing individuals or the number of generations of selection) to improve the power in detecting footprints of selection (see Table 1). In addition, our method provides accurate estimates of the selection coefficients associated to the analyzed SNP from either a 2-sample or a 3-sample time-series. Using more than three temporal samples does not provide further decisive information, given the third sample is appropriately chosen. We provided in Table 1 guidelines for planning short-term E&R experiments in small populations, including recommendations for the choice between 2-sample- and 3-sample-strategies for estimating the selection coefficient.

### *Genes associated to cancer are being targeted by selection in Tasmanian devil populations*

"Unplanned natural experiments" are part of the large scope of experimental evolution and are complementary to more standardized studies conducted at the lab (Irschick and Reznick 2009 ; Hubert et al. 2016). Furthermore, generating E&R data from the field offers a unique opportunity to observe real-time evolution subsequent to an environmental alteration, while taking into account the ecological complexity of natural environments. E&R data collected in such a context provide a valuable material that need to be well exploited. The case of wild Tasmanian devil populations that are currently facing the spread of a new infectious disease – a rare transmissible cancer called DFTD – is a highly symbolic example of the necessity of investigating cases of unplanned natural experiments. Despite the low levels of genetic diversity displayed in the species (Epstein et al. 2016a) and the high prevalence and mortality associated to DTFD, populations still persist in low numbers in the areas affected by the disease (Loh et al. 2006), suggesting the potential for a resistance to DFTD. Looking for signatures of selection in such a context may nevertheless seem tricky. In addition to the very low effective population sizes ($N_e \sim 30$), available data were collected over a very short temporal interval of ~ 10 years, what corresponds to ~ 5 generations of selection in Tasmanian devils. However, the study of Epstein et al. (2016a) showed the possibility of identifying candidate-genes for an evolutionary response to DFTD. By targeting genomic regions that accumulated extreme values both in allele frequency change and in *Rsb* statistic (Tang et al. 2007) over time, this study introduced a way for exploiting contemporary genomic time-series acquired from nature despite a priori

unfavorable conditions to detect selection. Our work confirms the possibility of performing successful genome scans for selection under such tricky conditions. In particular, our method allowed not only confirming the two signatures of selection previously revealed from the investigated dataset (Epstein et al. 2016a) but also adding new genes to the list of candidates for contemporary selection in response to DFTD. We could identify at least one cancer-associated locus in each of the three populations examined. The response to cancer may therefore have followed independent evolutionary routes in Tasmanian devils, according to the genetic variation specifically hosted by each population. One way to confirm or exclude this hypothesis would be to further genotype or sequence candidate-regions in the three populations examined. In addition, looking for potential functional polymorphisms in genes located in signatures of selection and related to cancer would be highly interesting. These genes are supposed to be involved in different pathways well-known to mediate cancer progression. A current issue in cancer cell biology is the difficulty to explore the precise role of many regulators and effectors that operate in multilayer intricate networks. Especially, more and more attention is being paid to the dual role of several key cellular factors towards tumor progression (Garcia-Gonzalo et al. 2003 ; Bialik et al. 2006 ; Britschgi et al. 2008 ; Britschgi et al. 2010 ; Goto et al. 2011 ; Schlegel et al. 2014 ; Bachiller et al. 2015 ; Geering et al. 2015 ; Evangelou et al. 2016 ; Nguyen et al. 2016 ; Wu et al. 2017). Revealing which causal polymorphisms might have been targeted by natural selection in putative two-faced candidates like *DAPK2*, *KLF10* or *MCAM* could therefore help improve our general understanding of key pathways mediating cancer progression.

## *Conclusion*

E&R experiments represent promising tools for investigating the capacity of small populations to evolve on the short-term. We provide here new resources to properly design and investigate such setups. Briefly, we presented a method that implements a model-based maximum-likelihood inference to target selection and estimate its strength from E&R experiments. Such a strategy was shown able to efficiently discriminate between drift alone and drift plus selection from both simulated and real E&R data. Real data analysis allowed the detection of consistent candidate-genes involved in the evolutionary response to DFTD in Tasmanian devils.

# Materials and methods

## *Simulations*

We tested the performances of our method by simulating the forward evolution of SNP under different scenarios. The standard scenario included SNP starting at an allele frequency of 0.5 in a population of $N_e$ = 50 undergoing strong directional selection ($s \geq 0.1$) over 10 generations. From this standard, we varied classical population genetics parameters within a range relevant to explore the scope of the method. Each parameter of interest was considered in turn. For instance, the first set of simulations implemented a portion of chromosome to specifically take into account the effect of linkage disequilibrium and therefore assess the ability of our method to report neutral SNP linked to a selected locus. Then, all subsequent simulations considered only a single SNP submitted to selection. All simulations consisted of Python scripts using the simuPOP library (Peng and Kimmel 2005).

## *Effective population size*

As our method involves a Wright-Fisher model, genetic drift is taken into account through the effective population size ($N_e$). Users may therefore specify an estimate of the $N_e$ of the examined population as an input parameter. Strategies for inferring $N_e$ represent a subfield of population genetics of its own and there are currently several types of methods for getting $N_e$ from genetic data (Waples 2016). Our preliminary assays for estimating $N_e$ (data not shown) worked well with the linkage disequilibrium (LD) and sibship frequency estimators, two single-sample methods reviewed in Wang (2016). Also, $N_e$ can be simply estimated by maximizing the joint likelihood of all SNP with no selection ($s$ = 0) using Equation 1, as done by Williamson and Slatkin (1999).

In conservation biology, a $N_e$ of 50 is highly symbolic. It has long been considered as a low acceptable limit in natural small populations to circumvent the risk of inbreeding depression on the short-term (Lehmkuhl 1984 ; Franklin 2014). Actually, a $N_e$ between 50 and 100 is also a convenient rough estimate for many animal populations undergoing artificial selection (Konczal et al. 2015 ; Howard et al. 2017). To cover this range, we investigated values of $N_e$ from 30 to 200. Using our method outside these limits is in theory possible, but only very strong selection may be detected when $N_e < 30$, and the amount of computing time may be substantial when $N_e > 200$.

## *Selection coefficient*

Current reviews about E&R experiments generally consider long-term studies and an upper limit of $s = 0.1$ for the selection coefficient (Kofler and Schlötterer 2014). Selection coefficients over 0.1 were examined here, as we aimed to investigate strong directional selection on the short-term.

Rather than the selection coefficient ($s$), biologists that implement such setups use the selection intensity ($i$) to quantify the strength of selection, because the latter is easier to connect to the selection pressure they may apply and to the response to selection they may expect. There is no precise definition of what is strong selection pressure, but a selected proportion of 20% individuals or below (*i.e.,* a selection intensity $i = 1.4$ or above) is generally considered as matching to the highest range of selection (Falconer 1981, San Cristobal et al. 1998 ; Walsh 2004 ; Fernandez et al. 2014, Hangartner and Hoffmann 2016). For instance, in a recent study of artificial selection aiming at providing ecologically relevant insights about the potential for rapid climate change adaptation in vertebrates, Hangartner and Hoffmann (2016) applied a selection intensity of $i = 1.7$ to Drosophila lines over 10 generations.

As presented in Falconer (1981, Equation 11.8), a QTL undergoing directional selection can be approximated as a locus with a selection coefficient that is function of the selection intensity (Equation 9).

$$s = \frac{2a}{\sigma_P} i \tag{9}$$

where $2a$ represents the difference in mean phenotypic values between the two homozygous genotypes $a_1a_1$ and $a_2a_2$, and $\sigma_P$ is the standard deviation of the phenotype before selection.

The ratio $2a/\sigma_P$ is called the *standardized effect* of the locus and quantifies its contribution to the phenotype of interest. Building on this link between quantitative and population genetic models, a strong selection intensity $i = 1.4$ applied to a SNP with a standardized effect of 10% would translate into a selection coefficient of $s = 0.14$. We therefore considered the value of $s = 0.1$ as a good starting point to explore strong selection.

### *Application to real data*

A real data analysis was performed on three genetic time-series previously investigated in a recent paper (Epstein et al. 2016a). Authors showed a genomic evidence for short-term selection in natural populations of Tasmanian devils (*Sarcophilus harrisii*) in response to Devil Facial Tumor Disease (DFTD), a serious transmissible cancer. Data analyzed in this study were made

publicly available on *Dryad* (Epstein et al. 2016b). We considered here the same samples as those analyzed in the originating paper. Samples were collected at different times from three distinct populations, Freycinet (FN), Narawntapu (NP) and West Pencil Pine (WP). We looked for candidates for selection by applying our method to 2 temporal samples in the FN (the sample from 1999 and the combination of those from 2012 and 2013) and WP (the sample from 2006 and the combination of those from 2013 and 2014) populations, and to three temporal samples (samples from 1999, 2004 and 2009) in the NP population (Table 2).

We reproduced the SNP filtering strategy reported by Epstein et al. (2016a) in their Methods section. In brief, we performed filtering according to (i) MAF computed over the whole dataset (SNP with MAF less than 0.01 were discarded), (ii) observed heterozygosities computed over the whole dataset (SNP with heterozygosity over 0.5 were discarded), (iii) the proportion of missing genotypes (SNP with less than one-third of genotypes either in the whole dataset or in a sample were discarded, except for the two smallest samples in which SNP with less than half genotypes were removed), (iv) the linkage disequilibrium between neighboring SNP (we removed using PLINK (Purcell et al. 2007) one SNP from pairs of SNP harboring $R^2 > 0.99$ over 20 successive SNP and 50 kb of distance in any sample).

In addition, we discarded SNP without polymorphism in the pre-DFTD samples (corresponding to samples from 1999 for the FN and the NP populations, and from 2006 for the WP population), since such SNP are meaningless for our method. Concerning the NP population, SNP fixed in 2004 were also removed. Finally, we applied our method to filtered datasets of respectively 16978, 27173 and 5401 SNP for the FN, NP and WP populations.

Our method requires specifying a value of $N_e$ for each investigated population. We considered here the effective population sizes assumed in Epstein et al. (2016a), that is 34, 37 and 26 for the FN, NP and WP populations, respectively.

*Related works*

In this paper, we suggested to explore genetic time-series from experimental populations undergoing strong directional selection by coupling Wright-Fisher models to a maximum-likelihood framework. Williamson and Slatkin (1999) already used such a strategy to estimate population size and growth rate from neutral polymorphisms. Building on the same theoretical idea, a first implementation of our method to detect selection from an E&R experiment in wheat was used in Thépot et al. (2015). Here, we performed an extensive simulation work to provide

guidelines for maximizing the performances of our approach. A recent work considered a similar maximum-likelihood approach to investigate Pool-seq data (Iranmehr et al. 2017).

## References

Alnahhas, N., Berri, C., Boulay, M., Baéza, E., Jégo, Y., Baumard, Y., ... & LeBihan-Duval, E. (2014). Selecting broiler chickens for ultimate pH of breast muscle: Analysis of divergent selection experiment and phenotypic consequences on meat quality, growth, and body composition traits. Journal of animal science, 92(9), 3816-3824.

Bachiller, S., Rybkina, T., Porras-García, E., Pérez-Villegas, E., Tabares, L., Armengol, J. A., ... & Ruiz, R. (2015). The HERC1 E3 ubiquitin ligase is essential for normal development and for neurotransmission at the mouse neuromuscular junction. *Cellular and molecular life sciences*, *72*(15), 2961-2971.

Baldwin-Brown, J. G., Long, A. D., & Thornton, K. R. (2014). The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. Molecular biology and evolution, msu048.

Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in ecology & evolution*, *23*(1), 38-44.

Bell, G. (2016). Experimental macroevolution. In Proc. R. Soc. B (Vol. 283, No. 1822, p. 20152547). The Royal Society.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

Bennett, A. F., & Lenski, R. E. (1999). Experimental evolution and its role in evolutionary physiology. American Zoologist, 39(2), 346-362.

Bialik, S., & Kimchi, A. (2006). The death-associated protein kinases: structure, function, and beyond. Annu. Rev. Biochem., 75, 189-210.

Boyle, K. E., Monaco, H. T., Deforet, M., Yan, J., Wang, Z., Rhee, K., & Xavier, J. (2017). Metabolism and the evolution of social behavior. *Molecular Biology and Evolution*.

Bradic, M., Beerli, P., García-de León, F. J., Esquivel-Bobadilla, S., & Borowsky, R. L. (2012). Gene flow and population structure in the Mexican blind cavefish complex (Astyanax mexicanus). BMC evolutionary biology, 12(1), 9.

Britschgi, A., Trinh, E., Rizzi, M., Jenal, M., Ress, A., Tobler, A., ... & Tschan, M. P. (2008). DAPK2 is a novel E2F1/KLF6 target gene involved in their proapoptotic function. Oncogene, 27(43), 5706-5716.

Britschgi, A., Simon, H. U., Tobler, A., Fey, M. F., & Tschan, M. P. (2010). Epigallocatechin-3-gallate induces cell death in acute myeloid leukaemia cells and supports all-trans retinoic acid-induced neutrophil differentiation via death-associated protein kinase 2. British journal of haematology, 149(1), 55-64.

Christie, M. R., Marine, M. L., Fox, S. E., French, R. A., & Blouin, M. S. (2016). A single generation of domestication heritably alters the expression of hundreds of genes. Nature communications, 7.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101.

Cooke, S. J., Suski, C. D., Ostrand, K. G., Wahl, D. H., & Philipp, D. P. (2007). Physiological and behavioral consequences of long-term artificial selection for vulnerability to recreational angling in a teleost fish. Physiological and Biochemical Zoology, 80(5), 480-490.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ... & Storfer, A. (2016a). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nature communications*, *7*, 12684.

Epstein B, Jones M, Hamede R, Hendricks S, McCallum H, Murchison EP, Schönfeld B, Wiench C, Hohenlohe P, Storfer A (2016b) Data from: Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.r60sv

Evangelou, K., Galanos, P., & Gorgoulis, V. G. (2016). The Janus face of p21. Molecular & cellular oncology, 3(5), e1215776.

Falconer, D.S. (1981). Selection: I. The response and its prediction. Introduction to quantitative genetics. Ed. 2. Longmans Green, London/New York.

Fernández, J., Toro, M. Á., Sonesson, A. K., & Villanueva, B. (2014). Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. Frontiers in genetics, 5.

Foll, M., Shim, H., & Jensen, J. D. (2015). WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular ecology resources, 15(1), 87-98.

Franklin, I.R. (2014). The 50/500 rules is still valid – Reply to Frankham et al. Biological Conservation, 176, 284-285.

Garcia-Gonzalo, F. R., Cruz, C., Muñoz, P., Mazurek, S., Eigenbrodt, E., Ventura, F., ... & Rosa, J. L. (2003). Interaction between HERC1 and M2-type pyruvate kinase. FEBS letters, 539(1-3), 78-84.

Garland Jr, T. (2003). Selection experiments: an under-utilized tool in biomechanics and organismal biology. *Vertebrate biomechanics and evolution*, 23-56.

Geering, B. (2015). Death-associated protein kinase 2: Regulator of apoptosis, autophagy and inflammation. The international journal of biochemistry & cell biology, 65, 151-154.

Gilbert, K. J., & Whitlock, M. C. (2015). Evaluating methods for estimating local effective population size with and without migration. Evolution, 69(8), 2154-2166.

Goto, N., Hiyoshi, H., Ito, I., Tsuchiya, M., Nakajima, Y., & Yanagisawa, J. (2011). Estrogen and antiestrogens alter breast cancer invasiveness by modulating the transforming growth factor-β signaling pathway. Cancer science, 102(8), 1501-1508.

Graves Jr, J. L., Hertweck, K. L., Phillips, M. A., Han, M. V., Cabral, L. G., Barter, T. T., ... & Rose, M. R. (2017). Genomics of Parallel Experimental Evolution in Drosophila. *Molecular biology and evolution*, *34*(4), 831-842.

Griffin, H., Acamovic, F., Guo, K., & Peddie, J. (1989). Plasma lipoprotein metabolism in lean and in fat chickens produced by divergent selection for plasma very low density lipoprotein concentration. Journal of lipid research, 30(8), 1243-1250.

Hangartner, S., & Hoffmann, A. A. (2016). Evolutionary potential of multiple measures of upper thermal tolerance in Drosophila melanogaster. *Functional Ecology*, *30*(3), 442-452.

Hendry, A. P., Grant, P. R., Grant, B. R., Ford, H. A., Brewer, M. J., & Podos, J. (2006). Possible human impacts on adaptive radiation: beak size bimodality in Darwin's finches. Proceedings of the Royal Society of London B: Biological Sciences, 273(1596), 1887-1894.

Heo, S. H., Jeong, E. S., Lee, K. S., Seo, J. H., Lee, W. K., & Choi, Y. K. (2015). Krüppel-like factor 10 null mice exhibit lower tumor incidence and suppressed cellular proliferation activity following chemically induced liver tumorigenesis. Oncology reports, 33(4), 2037-2044.

Heo, S. H., Jeong, E. S., Lee, K. S., Seo, J. H., Lee, W. K., & Choi, Y. K. (2017). Knockout of krüppel-like factor 10 suppresses hepatic cell proliferation in a partially hepatectomized mouse model. Oncology Letters, 13(6), 4843-4848.

Hiramatsu, L., Kay, J., Thompson, Z., Singleton, J., Claghorn, G., de Albuquerque, R. L., ... & Garland, T. (2017). Maternal exposure to Western diet affects adult body composition and voluntary wheel running in a genotype-specific manner in mice. *Physiology & Behavior*.

Hocquette, J. F., Capel, C., David, V., Guemene, D., Bidanel, J., Ponsart, C., ... & Barbezant, M. (2012). Objectives and applications of phenotyping network set-up for livestock. Animal Science Journal, 83(7), 517-528.

Howard, J. T., Pryce, J. E., Baes, C., & Maltecca, C. (2017). Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. Journal of Dairy Science.

Hsu, C. F., Sui, C. L., Wu, W. C., Wang, J. J., Yang, D. H., Chen, Y. C., ... & Chang, H. S. (2011). Klf10 induces cell apoptosis through modulation of BI-1 expression and Ca 2+ homeostasis in estrogen-responding adenocarcinoma cells. The international journal of biochemistry & cell biology, 43(4), 666-673.

Hubert, J. N., Allal, F., Hervet, C., Ravakarivelo, M., Jeney, Z., Vergnet, A., ... & Vandeputte, M. (2016). How could fully scaled carps appear in natural waters in Madagascar?. In Proc. R. Soc. B (Vol. 283, No. 1837, p. 20160945). The Royal Society.

Inbal, B., Bialik, S., Sabanay, I., Shani, G., & Kimchi, A. (2002). DAP kinase and DRP-1 mediate membrane blebbing and the formation of autophagic vesicles during programmed cell death. J Cell Biol, 157(3), 455-468.

Iranmehr, A., Akbari, A., Schlötterer, C., & Bafna, V. (2017). CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments. Genetics, 206(2), 1011-1023.

Irschick, D.J., & Reznick, D. (2009). Field experiments, introductions, and experimental evolution: a review and practical guide. Experimental evolution: concepts, methods, and applications of selection experiments. University of California Press, California.

Jha, A. R., Miles, C. M., Lippert, N. R., Brown, C. D., White, K. P., & Kreitman, M. (2015). Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in Drosophila melanogaster. *Molecular biology and evolution*, *32*(10), 2616-2632.

Jiang, L., Lai, Y. K., Zhang, J. F., Chan, C. Y., Lu, G., Lin, M. C., ... & Kung, H. F. (2012). Transactivation of the TIEG1 confers growth inhibition of transforming growth factor-β-susceptible hepatocellular carcinoma cells. World Journal of Gastroenterology: WJG, 18(17), 2035.

Kales, S. C., Ryan, P. E., Nau, M. M., & Lipkowitz, S. (2010). Cbl and human myeloid neoplasms: the Cbl oncogene comes of age. Cancer research, 70(12), 4789-4794.

Kameya, S., Hawes, N. L., Chang, B., Heckenlively, J. R., Naggert, J. K., & Nishina, P. M. (2002). Mfrp, a gene encoding a frizzled related protein, is mutated in the mouse retinal degeneration 6. Human molecular genetics, 11(16), 1879-1886.

Keightley PD, Caballero A. (1997) Genomic mutation rates for lifetime reproductive output and lifespan in Caenorhabditis elegans. Proc Natl Acad Sci USA 94: 3823-3827

Kessner, D., & Novembre, J. (2015). Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. Genetics, 199(4), 991-1005.

Kofler, R., & Schlötterer, C. (2014). A guide for the design of evolve and resequencing studies. Molecular biology and evolution, 31(2), 474-483.

Konczal, M., Babik, W., Radwan, J., Sadowska, E. T., & Koteja, P. (2015). Initial molecular-level response to artificial selection for increased aerobic metabolism occurs primarily through changes in gene expression. *Molecular biology and evolution*, 32(6), 1461-1473.

Kraaijeveld-Smit, F. J., Griffiths, R. A., Moore, R. D., & Beebee, T. J. (2006). Captive breeding and the fitness of reintroduced species: a test of the responses to predators in a threatened amphibian. *Journal of Applied Ecology*, *43*(2), 360-365.

Lattorff, H. M. G., & Moritz, R. F. (2013). Genetic underpinnings of division of labor in the honeybee (Apis mellifera). *Trends in Genetics*, *29*(11), 641-648.

Lawrie, D. S., & Petrov, D. A. (2014). Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, *30*(4), 133-139.

Le Boucher, R., Dupont-Nivet, M., Vandeputte, M., Kerneïs, T., Goardon, L., Labbe, L., ... & Quillet, E. (2012). Selection for adaptation to dietary shifts: towards sustainable breeding of carnivorous fish. PloS one, 7(9), e44898.

Lehmkuhl, J. F. (1984). Determining size and dispersion of minimum viable populations for land management planning and species conservation. Environmental Management, 8(2), 167-176.

Lenski, R. E., & Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proceedings of the National Academy of Sciences, 91(15), 6808-6814.

Lillie, M., Shine, R., & Belov, K. (2014). Characterisation of major histocompatibility complex class I in the Australian cane toad, Rhinella marina. PloS one, 9(8), e102824.

Loh, R., Bergfeld, J., Hayes, D., O'hara, A., Pyecroft, S., Raidal, S., & Sharpe, R. (2006). The pathology of devil facial tumor disease (DFTD) in Tasmanian devils (Sarcophilus harrisii). Veterinary Pathology, 43(6), 890-895.

Long, A., Liti, G., Luptak, A., & Tenaillon, O. (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nature Reviews Genetics, 16(10), 567-582.

Lung, H. L., Bangarusamy, D. K., Xie, D., Cheung, A. K. L., Cheng, Y., Kumaran, M. K., ... & Fang, Y. (2005). THY1 is a candidate tumour suppressor gene with decreased expression in metastatic nasopharyngeal carcinoma. Oncogene, 24(43), 6525-6532.

Matute, D. R. (2013). The role of founder effects on the evolution of reproductive isolation. *Journal of evolutionary biology*, *26*(11), 2299-2311.

McGuigan, K., Collet, J. M., McGraw, E. A., Yixin, H. Y., Allen, S. L., Chenoweth, S. F., & Blows, M. W. (2014). The nature and extent of mutational pleiotropy in gene expression of male Drosophila serrata. Genetics, 196(3), 911-921.

Metzig, M., Nickles, D., Falschlehner, C., Lehmann-Koch, J., Straub, B. K., Roth, W., & Boutros, M. (2011). An RNAi screen identifies USP2 as a factor required for TNF-α-induced NF-κB signaling. International journal of cancer, 129(3), 607-618.

Minvielle, F., Mills, A. D., Faure, J. M., Monvoisin, J. L., & Gourichon, D. (2002). Fearfulness and performance related traits in selected lines of Japanese quail (Coturnix japonica). Poultry Science, 81(3), 321-326.

Morgan, A. D., Ness, R. W., Keightley, P. D., & Colegrave, N. (2014). Spontaneous mutation accumulation in multiple strains of the green alga, Chlamydomonas reinhardtii. Evolution, 68(9), 2589-2602.

Nguyen, L. S., Schneider, T., Rio, M., Moutton, S., Siquier-Pernet, K., Verny, F., ... & Cormier-Daire, V. (2016). A nonsense variant in HERC1 is associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy. *European Journal of Human Genetics*, *24*(3), 455-458.

Norén, K., & Angerbjörn, A. (2014). Genetic perspectives on northern population cycles: bridging the gap between theory and empirical studies. Biological Reviews, 89(2), 493-510.

Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., & Schloetterer, C. (2012). Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Molecular ecology, 21(20), 4931-4941.

Papetti, C., Lucassen, M., & Pörtner, H. O. (2016). Integrated studies of organismal plasticity through physiological and transcriptomic approaches: examples from marine polar regions. *Briefings in functional genomics*, *15*(5), 365-372.

Papuc, S. M., Hackmann, K., Andrieux, J., Vincent-Delorme, C., Budişteanu, M., Arghir, A., ... & Di Donato, N. (2015). Microduplications of 3p26. 3p26. 2 containing CRBN gene in patients with intellectual disability and behavior abnormalities. European journal of medical genetics, 58(5), 319-323.

Peng, B., & Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics, 21(18), 3686-3687.

Promislow, D. E., Smith, E. A., & Pearse, L. (1998). Adult fitness consequences of sexual selection in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *95*(18), 10687-10692.

Reznick, D. N., Bryant, M. J., Roff, D., Ghalambor, C. K., & Ghalambor, D. E. (2004). Effect of extrinsic mortality on the evolution of senescence in guppies. Nature, 431(7012), 1095-1099.

SanCristobal-Gaudy, M., Elsen, J. M., Bodin, L., & Chevalet, C. (1998). Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. Genetics Selection Evolution, 30(5), 423.

Scanlan, P. D., Hall, A. R., Blackshields, G., Friman, V. P., Davis Jr, M. R., Goldberg, J. B., & Buckling, A. (2015). Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. Molecular biology and evolution, 32(6), 1425-1435.

Schlagbauer-Wadl, H., Jansen, B., Müller, M., Polterauer, P., Wolff, K., Eichler, H. G., ... & Johnson, J. P. (1999). Influence of MUC18/MCAM/CD146 expression on human melanoma growth and metastasis in SCID mice. International journal of cancer, 81(6), 951-955.

Schlegel, C. R., Fonseca, A. V., Stöcker, S., Georgiou, M. L., Misterek, M. B., Munro, C. E., ... & Costa-Pereira, A. P. (2014). DAPK2 is a novel modulator of TRAIL-induced apoptosis. Cell Death & Differentiation, 21(11), 1780-1791.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., & Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity, 114(5), 431-440.

Schuster, S. R., Kortuem, K. M., Zhu, Y. X., Braggio, E., Shi, C. X., Bruins, L. A., ... & Mikhael, J. (2014). The clinical significance of cereblon expression in multiple myeloma. Leukemia research, 38(1), 23-28.

Schwamb, B., Pick, R., Fernández, S. B. M., Völp, K., Heering, J., Dötsch, V., ... & Zörnig, I. (2015). FAM96A is a novel pro-apoptotic tumor suppressor in gastrointestinal stromal tumors. International journal of cancer, 137(6), 1318-1329.

Shiloh, R., Bialik, S., & Kimchi, A. (2014). The DAPK family: a structure-function analysis. Apoptosis: an international journal on programmed cell death, 19(2), 286.

Shohat, G., Shani, G., Eisenstein, M., & Kimchi, A. (2002). The DAP-kinase family of proteins: study of a novel group of calcium-regulated death-promoting kinases. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1600(1), 45-50.

Shoval, Y., Berissi, H., Kimchi, A., & Pietrokovski, S. (2011). New modularity of DAP-kinases: alternative splicing of the DRP-1 gene produces a ZIPk-like isoform. PloS one, 6(2), e17344.

Smukowski Heil, C. S., DeSevo, C. G., Pai, D. A., Tucker, C. M., Hoang, M. L., & Dunham, M. J. (2017). Loss of heterozygosity drives adaptation in hybrid yeast. Molecular Biology and Evolution, 34(7), 1596-1612.

Song, K. D., Kim, D. J., Lee, J. E., Yun, C. H., & Lee, W. K. (2012). KLF10, transforming growth factor-β-inducible early gene 1, acts as a tumor suppressor. Biochemical and biophysical research communications, 419(2), 388-394.

Sotgia, F., Whitaker-Menezes, D., Martinez-Outschoorn, U. E., Salem, A. F., Tsirigos, A., Lamb, R., ... & Lisanti, M. P. (2012). Mitochondria "fuel" breast cancer metabolism: fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells. Cell cycle, 11(23), 4390-4401.

Stehling, O., Mascarenhas, J., Vashisht, A. A., Sheftel, A. D., Niggemeyer, B., Rösser, R., ... & Lill, R. (2013). Human CIA2A-FAM96A and CIA2B-FAM96B integrate iron homeostasis and maturation of different subsets of cytosolic-nuclear iron-sulfur proteins. Cell metabolism, 18(2), 187-198.

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS biology, 5(7), e171.

Tetreault, M. P., Yang, Y., & Katz, J. P. (2013). Krüppel-like factors in cancer. Nature Reviews Cancer, 13(10), 701-713.

Thépot, S., Restoux, G., Goldringer, I., Gouache, D., Hospital, F., Mackay, I., & Enjalbert, J. (2015). Efficiently tracking selection in a multiparental population: the case of earliness in wheat. Genetics, 199(2), 609-623.

Toro, M. A., Fernández, J., Shaat, I., & Mäki-Tanila, A. (2011). Assessing the genetic diversity in small farm animal populations. animal, 5(11), 1669-1683.

Travisano, M. (2009). Long-term experimental evolution and adaptive radiation. Experimental evolution: concepts, methods, and applications of selection experiments. University of California Press, California.

Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., & Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. PLoS Genet, 7(3), e1001336.

Venuprasad, K., Huang, H., Harada, Y., Elly, C., Subramaniam, M., Spelsberg, T., ... & Liu, Y. C. (2008). The E3 ubiquitin ligase Itch regulates expression of transcription factor Foxp3 and airway inflammation by enhancing the function of transcription factor TIEG1. Nature immunology, 9(3), 245-253.

Walsh, B. (2004). Population-and quantitative-genetic models of selection limits. Plant breeding reviews, 24(1), 177-226.

Wang, J. (2009). A new method for estimating effective population sizes from a single sample of multilocus genotypes. Molecular Ecology, 18(10), 2148-2164.

Wang, J. (2016). A comparison of single-sample estimators of effective population sizes from genetic marker data. Molecular ecology, 25(19), 4692-4711.

Waples, R. S. (2016). Making sense of genetic estimates of effective population size. Molecular ecology, 25(19), 4689-4691.

Weng, C. C., Hawse, J. R., Subramaniam, M., Chang, V. H. S., Yu, W. C. Y., Hung, W. C., ... & Cheng, K. H. (2017). KLF10 loss in the pancreas provokes activation of SDF-1 and induces distant metastases of pancreatic ductal adenocarcinoma in the KrasG12D p53flox/flox model. Oncogene.

Williamson, E. G., & Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics, 152(2), 755-761.

Wu, W. K., Li, X., Wang, X., Dai, R. Z., Cheng, A. S., Wang, M. H., ... & Wong, S. H. (2017). Oncogenes without a neighboring tumor-suppressor gene are more prone to amplification. Molecular biology and evolution, 34(4), 903-907.

Zan, Y., Sheng, Z., Lillie, M., Ronnegard, L., Honaker, C. F., Siegel, P. B., & Carlborg, O. (2017). Artificial selection response due to polygenic adaptation from a multi-locus, multi-allelic genetic architecture. *Molecular Biology and Evolution* doi: 10.1093/molbev/msx194

Zhang, M. Y., & Wang, J. P. (2017). A multi-target protein of hTERTR-FAM96A presents significant anticancer potent in the treatment of hepatocellular carcinoma. Tumor Biology, 39(4), 1010428317698341.

948 Zhao, X., Bergland, A. O., Behrman, E. L., Gregory, B. D., Petrov, D. A., & Schmidt, P. S.
949 (2015). Global transcriptional profiling of diapause and climatic adaptation in Drosophila
950 melanogaster. *Molecular biology and evolution*, *33*(3), 707-720.

951 Zwoinska, M. K., Maklakov, A. A., Kawecki, T. J., & Hollis, B. (2017). Experimental
952 evolution of slowed cognitive aging in Drosophila melanogaster. *Evolution*, *71*(3), 662-670.

953

954 # Figure legends

955 **Figure 1.** A SNP associated to 4 different selection coefficients *s* [0 – 0.7] was positioned at
956 the beginning of a 50-cM long genomic region from a single chromosome. This region was
957 genotyped through a set of 51 SNP regularly spaced 1 centiMorgan (cM) apart, including the
958 selected SNP at position 0 and 50 neutral SNP. The whole SNP were at a frequency of 0.5 and
959 in total linkage disequilibrium before evolution. Evolution consisted in simulating either 10
960 generations of selection (solid curves) or 3 generations of random mating prior to 10 generations
961 of selection (dashed curves) in a population of $N_e$ = 50. Our method was applied to look for
962 selection from the simulated data recorded at the first generation ($G_0$) and the last generation
963 ($G_{10}$). Power shows the proportion of SNP declared as selected by the test over 500 replications
964 with a FDR of 10% and is represented here with regard to the genetic distance from the selected
965 position.

966 **Figure 2.** A SNP associated to 4 different selection coefficients *s* [0 – 0.7] was followed over
967 10 generations with constant $N_e$ = 50. The initial allele frequency $f_0$ (*i.e.*, before selection
968 occurred) ranged from 0.01 to 0.99. Our method was applied to look for selection from the
969 simulated data recorded at the first generation ($G_0$) and the last generation ($G_{10}$). (A) Power
970 shows the proportion of SNP declared as selected by our method with α = 0.01 over 500
971 replications. Power is null above $f_{0\_lim}$ = 0.82. (B) Estimates for the selection coefficient ($\hat{s}$) as
972 provided by our method from averaging 500 replications, are plotted with regard to $f_0$. Dashed
973 lines indicate the expected value according to simulation settings. The hatched area corresponds
974 to frequencies above the detection limit ($f_{0\_lim}$ = 0.82), beyond which power was lost.

975 **Figure 3.** A SNP associated to 4 different selection coefficients *s* [0 – 0.7] was followed over
976 10 generations with constant $N_e$ = 50. Our method was applied to look for selection from the
977 simulated data recorded at the first generation ($G_0$) and the last generation ($G_{10}$). Box plots show
978 the selection coefficients ($\hat{s}$) as estimated by our method for the 500 replicated simulations, with
979 regard to the expected values (*s*), also specified through the horizontal dashed lines. The initial
980 allele frequencies (*i.e.*, at $G_0$, before selection occurred) were 0.25 (A) and 0.5 (B).

981 **Figure 4.** We considered the product $N_e s$ associated to a SNP in the range [0 - 50], with constant
982 $N_e$ in the range [35 – 200] over a forward-in-time evolution of 10 generations. The initial allele
983 frequency of the SNP was 0.5 and simulations were repeated 500 times for each condition. Our
984 method was applied to look for selection from the simulated data recorded at the first generation

($G_0$) and the last generation ($G_{10}$). Power shows the proportion of SNP declared as selected by our method with $\alpha = 0.01$ over 500 replications, and is represented here as a function of $N_e s$.

**Figure 5.** A SNP starting at frequency $f_0 = 0.5$ was associated to 4 different selection coefficients $s$ [0 – 0.7], and followed over 10 generations in a population of size $N_e = 50$. Our method was applied to look for selection from the simulated data recorded at the first generation ($G_0$) and the last generation ($G_{10}$). Analysis was here carried out with wrong estimates of the effective population size ($N_e$) ranging from 10 to 100. Power shows the proportion of simulated alleles declared as selected by our method with $\alpha = 0.01$ over 500 replications, and is represented here with regard to the estimates of $N_e$. The dashed vertical line indicates the true effective population size ($N_e = 50$).

**Figure 6.** The minimal shift in frequency required to detect selection after 10 generations of selection ($\Delta f_{10}^*$) with $\alpha = 0.01$ is represented as a function of the initial allele frequency ($f_0$) for effective populations sizes ($N_e$) ranging from 30 to 200. Larger populations are more likely to display selection, since declaring a SNP as selected requires a smaller frequency change as $N_e$ increases.

**Figure 7.** A SNP starting at frequency $f_0 = 0.5$ was associated to 4 different selection coefficients $s$ [0 – 0.7], and followed over 10 generations in a population of size $N_e = 50$. Our method was applied to look for selection from the simulated data recorded at the first generation ($G_0$) and the last generation ($G_{10}$). Unlike previous cases where allele frequency was computed from the whole population, analysis was here performed from samples with sizes ranging from $N_s = 10$ to $N_s = N_e = 50$. For each examined condition, bar plots show the proportion of SNP declared as selected by our method with $\alpha = 0.01$ over 500 replications.

**Figure 8.** A SNP starting at frequency $f_0 = 0.5$ was associated to 4 different selection coefficients $s$ [0 – 0.7], and followed over 10 generations in a population of size $N_e = 50$. Our method was applied to detect selection from the simulated data recorded at the first generation ($G_0$) and the last generation ($G_{10}$). Unlike previous cases where allele frequency was computed from the whole population, analysis was here performed from samples with sizes ranging from $N_s = 10$ to $N_s = N_e = 50$. For each examined condition, box plots show the selection coefficients ($\hat{s}$) as estimated by our method over the 500 replicated simulations. The true values of $s$ are given by the horizontal dashed lines. Red points indicate the mean.

**Figure 9.** A SNP starting at frequency $f_0 = 0.5$ was followed over a temporal window ranging from 1 to 20 generations of forward-in-time evolution. Our method was applied to look for

selection from the simulated data recorded at the first generation ($G_0$) and the last generation (from $G_1$ to $G_{20}$). Two cases were considered: strong selection ($s$ = 0.4, red curves) and absence of selection ($s$ = 0, blue curves). For each case, 4 different levels of strong drift were investigated, from $N_e$ = 35 to $N_e$ = 200. (A) Estimates for the selection coefficient ($\hat{s}$) as provided by our method from averaging 500 replications, are plotted with regard to the number of generations investigated ($n_g$). Dashed lines indicate the true value of $s$. (B) Power shows the proportion of simulated SNP declared as selected by our method with $\alpha = 0.01$.

**Figure 10.** Between two temporal samples, there is a maximal time interval beyond which selection cannot be detected. This maximal number of evolved generations ($n_{g\_max}$) between two consecutive samples in the time-series depends on the initial frequency ($f_0$) of the examined SNP and on the effective population size ($N_e$). Three different situations regarding the initial frequency ($f_0$ = 0.25, 0.5 or 0.75) of the examined SNP are represented.

**Figure 11.** A SNP starting at frequency $f_0$ = 0.5 was followed over 20 generations of selection ($s$ = 0.4) in a population of $N_e$ = 50. Our method was applied to estimate the selection coefficient from simulated data (i) at $G_0$ and $G_{20}$ (red dashed line), (ii) at all 21 generations from $G_0$ to $G_{20}$ (blue dashed line), and (iii) at $G_0$, $G_{20}$, and an intermediate generation $G_{int}$ taking values from $G_1$ to $G_{19}$ (black solid line). The green dashed line indicates the true value of $s$.

**Figure 12.** Application of our method to the Tasmanian devil dataset (Epstein et al. 2016b) allowed retrieving the first selection signal identified by Epstein et al. (2016a) on chromosome 2 in the vicinity of an ortholog of CRBN. The genomic window represented here corresponds to scaffold GL841593 (≈ 5.04 Mb) of the Tasmanian devil genome. Given physical positions, expressed in base pairs (bp), are relative to the investigated scaffold. Colored dots indicate each examined SNP within the scaffold according to the sampled population (FN = Freycinet, NP = Narawntapu and WP = West Pencil Pine). (A) $-log_{10}(p)$ is plotted against the physical position, where $p$ is the p-value obtained at the likelihood ratio test from our method. The candidate-region, which includes several neighboring SNP that pass the significance threshold in the three populations, is depicted in blue. Protein coding genes that stand at less than 100 kb of the candidate-region according to Ensembl *Sarcophilus harrisii* version 89.7 are depicted in red. (B) Maximum-likelihood estimation of the selection coefficient ($\hat{s}$) is plotted against physical position for each non-fixed SNP analyzed. Fixed SNP were excluded here because of the risk of overestimation of $s$ in case of fixation.

**Figure 13.** Application of our method to the Tasmanian devil dataset (Epstein et al. 2016b) allowed retrieving the second selection signal identified by Epstein et al. (2016a) on chromosome 3 in an ortholog of MFRP. The genomic window represented here corresponds to scaffold GL849657 (≈ 1.63 Mb) of the Tasmanian devil genome. Given physical positions, expressed in base pairs (bp), are relative to the investigated scaffold. Colored dots indicate each examined SNP within the scaffold according to the sampled population (FN = Freycinet, NP = Narawntapu and WP = West Pencil Pine). (A) $-log_{10}(p)$ is plotted against the physical position, where $p$ is the p-value obtained at the likelihood ratio test from our method. The candidate-region, which includes 2 neighboring SNP that clearly pass the significance threshold ($p < 10^{-5}$) in the WP population, is depicted in blue. Protein coding genes that stand at less than 100 kb of the candidate-region according to Ensembl *Sarcophilus harrisii* version 89.7 are depicted in red. (B) Maximum-likelihood estimation of the selection coefficient ($\hat{s}$) is plotted against physical position for each non-fixed SNP analyzed. Fixed SNP were excluded here because of the risk of overestimation of $s$ in case of fixation.

**Figure 14.** Application of our method to the Tasmanian devil dataset (Epstein et al. 2016b) allowed identifying a new signature of selection in an ortholog of DAPK2 on chromosome 1**.** The genomic window represented here corresponds to scaffold GL834484 (≈ 3.84 Mb) of the Tasmanian devil genome. Given physical positions, expressed in base pairs (bp), are relative to the investigated scaffold. Colored dots indicate each examined SNP within the scaffold according to the sampled population (FN = Freycinet, NP = Narawntapu and WP = West Pencil Pine). (A) $-log_{10}(p)$ is plotted against the physical position, where $p$ is the p-value obtained at the likelihood ratio test from our method. The candidate-region, which includes several neighboring SNP that pass the significance threshold, is depicted in blue. Protein coding genes that stand at less than 100 kb of the candidate-region according to Ensembl *Sarcophilus harrisii* version 89.7 are depicted in red. Another signature of selection is located in a weakly conserved region at around 2 Mb, which does not host any ortholog of mammal species. (B) Maximum-likelihood estimation of the selection coefficient ($\hat{s}$) is plotted against physical position for each non-fixed SNP analyzed. Fixed SNP were excluded here because of the risk of overestimation of $s$ in case of fixation.

**Figure 15.** Application of our method to the Tasmanian devil dataset (Epstein et al. 2016b) allowed identifying a new signature of selection in an ortholog of KLF10 on chromosome 2. The genomic window represented here corresponds to scaffold GL841374 (≈ 4.02 Mb) of the Tasmanian devil genome. Given physical positions, expressed in base pairs (bp), are relative to

the investigated scaffold. Colored dots indicate each examined SNP within the scaffold according to the sampled population (FN = Freycinet, NP = Narawntapu and WP = West Pencil Pine). (A) $-log_{10}(p)$ is plotted against the physical position, where $p$ is the p-value obtained at the likelihood ratio test from our method. The candidate-region, which includes 2 neighboring SNP that pass the significance threshold in the NP population, is depicted in blue. Protein coding genes that stand at less than 100 kb of the candidate-region according to Ensembl *Sarcophilus harrisii* version 89.7 are depicted in red. (B) Maximum-likelihood estimation of the selection coefficient ($\hat{s}$) is plotted against physical position for each non-fixed SNP analyzed. Fixed SNP were excluded here because of the risk of overestimation of $s$ in case of fixation.

## Tables

Table 1. Guidelines for inferring selection from short-term E&R experiments through our method.

| $N_e$ (Effective population size) | $n_g$ (Number of generations of selection)[a] | $s^*$ (Detectable selection coefficient)[b] | $\hat{s}^*$ (Estimate of $s^*$)[c] | $[G_{inf} ; G_{sup}]$ (Best interval for a 3rd temporal sampling point)[d] | *FPR* (False-positive rate)[c] |
|---|---|---|---|---|---|
| 30 | 5 | 0.95 | 0.81 | Useless | 0.02 |
| 30 | 10 | 0.75 | 0.72 | Useless | 0.01 |
| 30 | 20 | 0.4 | *0.45* | $[G_4 ; G_7]$ | *0.02* |
| 30 | **100** | 0.45 | *0.5* | $G_5$ | *0.01* |
| 50 | 5 | 0.55 | 0.58 | Useless | < 0.01 |
| 50 | 10 | 0.4 | 0.45 | Useless | 0.01 |
| 50 | 20 | 0.3 | *0.34* | $[G_4 ; G_7]$ | 0.03 |
| 50 | 30 | 0.25 | *0.3* | $[G_4 ; G_{10}]$ | *0.03* |
| 50 | **100** | 0.25 | *0.27* | $[G_3 ; G_4]$ | *0.01* |
| 100 | 5 | 0.3 | 0.32 | Useless | 0.01 |
| 100 | 10 | 0.25 | 0.27 | Useless | 0.01 |
| 100 | 20 | 0.2 | 0.22 | Useless | 0.01 |
| 100 | 30 | 0.15 | *0.16* | $[G_9 ; G_{11}]$ | 0.01 |
| 100 | 60 | 0.1 | *0.11* | $[G_{13} ; G_{16}]$ | 0.03 |
| 100 | **100** | 0.1 | *0.11* | $[G_8 ; G_{17}]$ | *0.02* |
| 200 | 5 | 0.2 | 0.21 | Useless | 0.02 |
| 200 | 10 | 0.15 | 0.15 | Useless | 0.01 |
| 200 | 20 | 0.1 | 0.11 | Useless | 0.01 |
| 200 | 120 | 0.05 | *0.06* | $[G_7 ; G_{40}]$ | 0.03 |
| 200 | **200** | 0.05 | *0.06* | $[G_5 ; G_{40}]$ | *0.01* |

Table 1. (continued)

[a]There is an upper limit in the acceptable number of generations between two consecutive samples (see Fig. 10). Beyond this limit, it is necessary to increase the number of samples in the time-series to detect SNP. Bold values indicate when the number of generations investigated passes the limit.

[b]The detectable selection coefficient $s^*$ is the required selection coefficient to get at least a power of $(1 - \beta) = 0.8$, where $\beta$ is the Type II error, in accordance with the convention suggested by Cohen (1992) for power analyses. We explored the range of possible $s^*$ by steps of 0.05.

[c]Estimates for $s^*$ and *FPR* were obtained from two-sample time-series, except for values in italics that required an extra sample.

[d]When a third sample may improve the estimation of $s^*$ (see Fig. 11), we give the lower ($G_{inf}$) and upper ($G_{sup}$) bounds of the interval [$G_{int}$ ; $G_{sup}$] that provides the best accuracy.

N.B. All results displayed here derive from simulations of a SNP with an initial frequency of 0.5. Our method was applied under different experimental scenarios by varying $N_e$, $n_g$ and the number of available genetic time-samples. Generations are numbered with regards to the first generation ($G_0$) submitted to selection. Estimates of $s^*$ and *FPR* were computed over 500 replications and rounded to the nearest 0.01.

Table 2. General information about the genomic time-series from the Tasmanian devil dataset (Epstein et al. 2016a, 2016b).

| Population symbol | Sampling location | $N_e$ | $t_0$ | $t_1$ | $t_2$ | Distance between $t_0$ and $t_2$ | Nb. of SNP analyzed |
|---|---|---|---|---|---|---|---|
| FN | Freycinet | 34 | 1999 | - | 2012-2013 | 6 generations | 16978 |
| NP | Narawntapu | 37 | 1999 | 2004 | 2009 | 5 generations | 27173 |
| WP | West Pencil Pine | 26 | 2006 | - | 2013-2014 | 3 generations | 5401 |

## Figures

### Figure 1

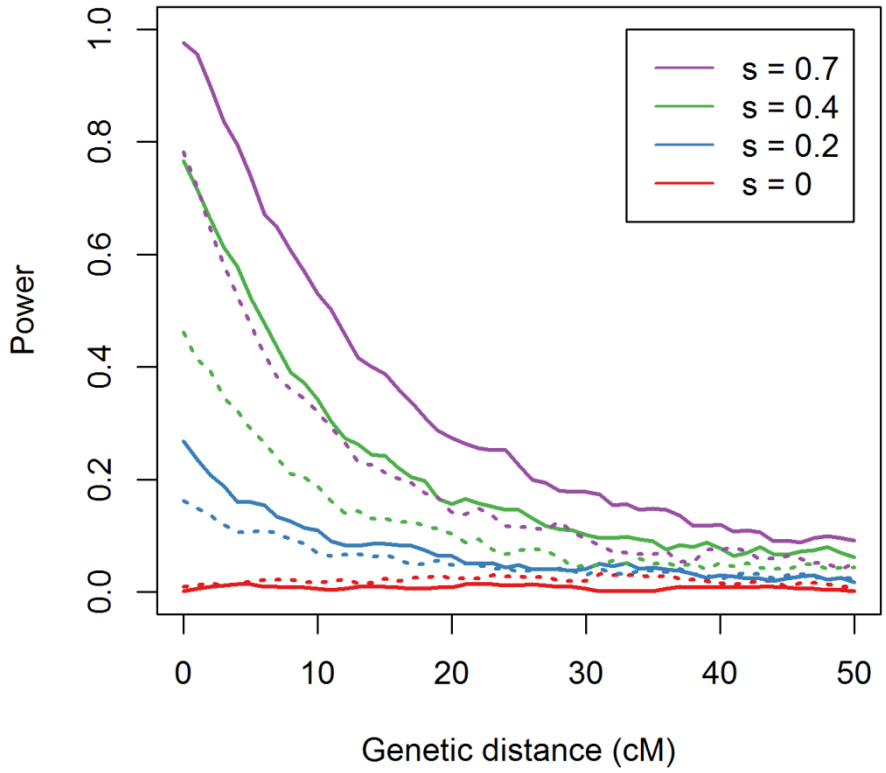


**Figure 2**

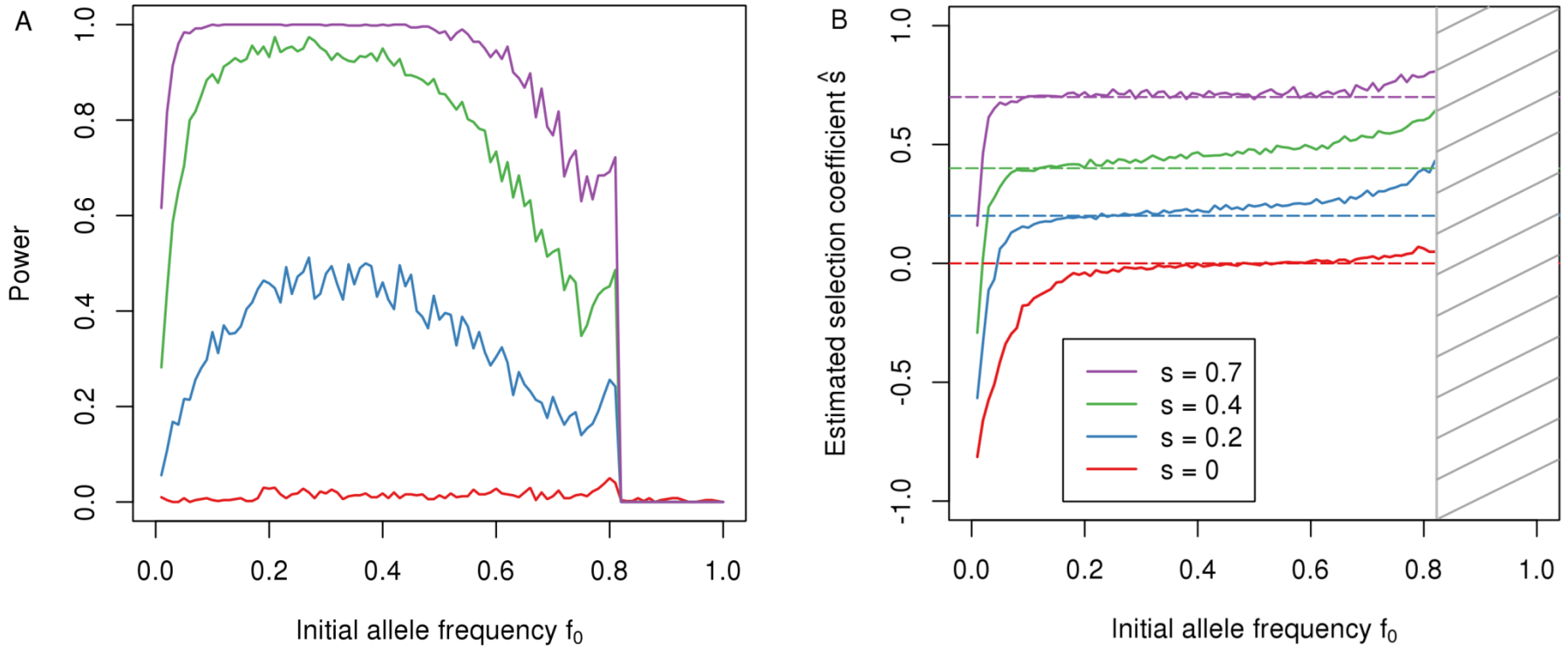

A
Power
0.0
0.2
0.4
0.6
0.8
1.0
Initial allele frequency $f_0$
B
Estimated selection coefficient $\hat{s}$
-1.0
-0.5
0.0
0.5
1.0
s = 0.7
s = 0.4
s = 0.2
s = 0
Initial allele frequency $f_0$

**Figure 3**

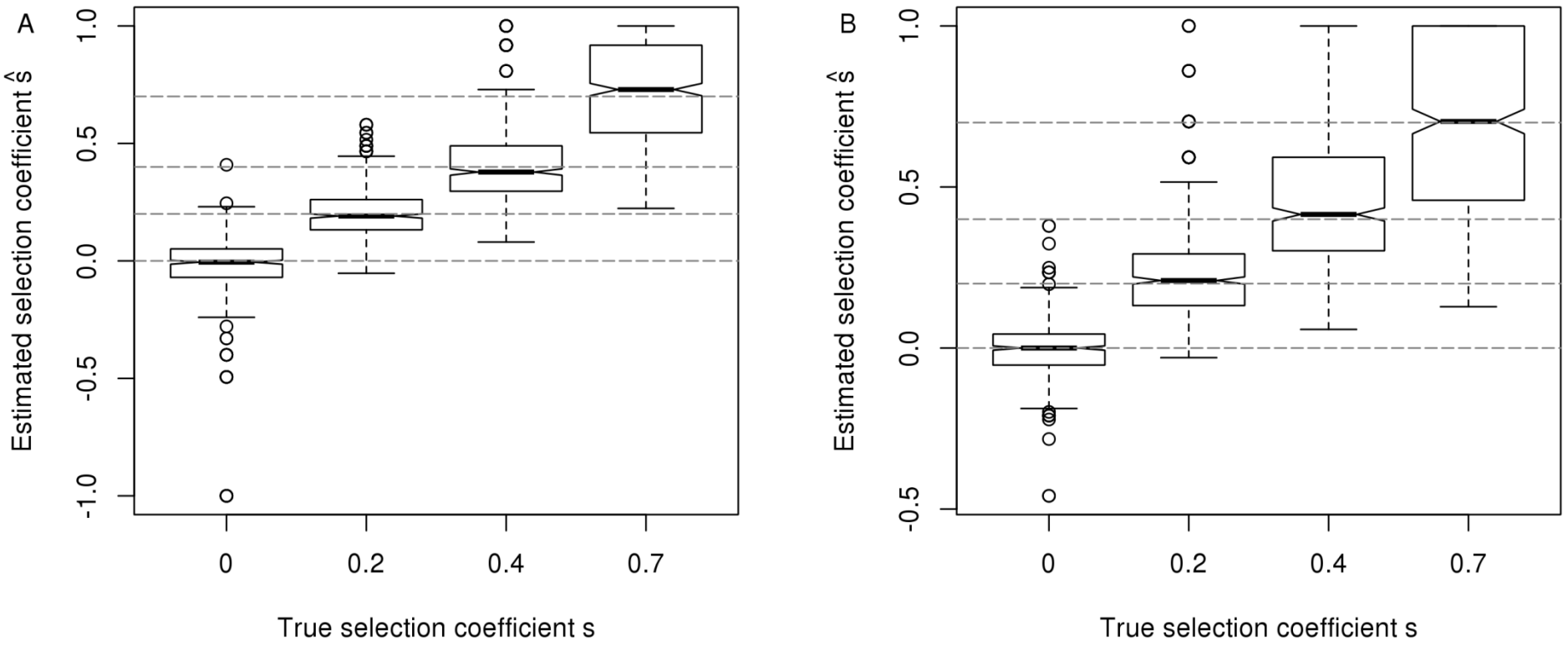

A
B
Estimated selection coefficient ŝ
True selection coefficient s
1.0
0.5
0.0
-0.5
-1.0
0
0.2
0.4
0.7

**Figure 4**

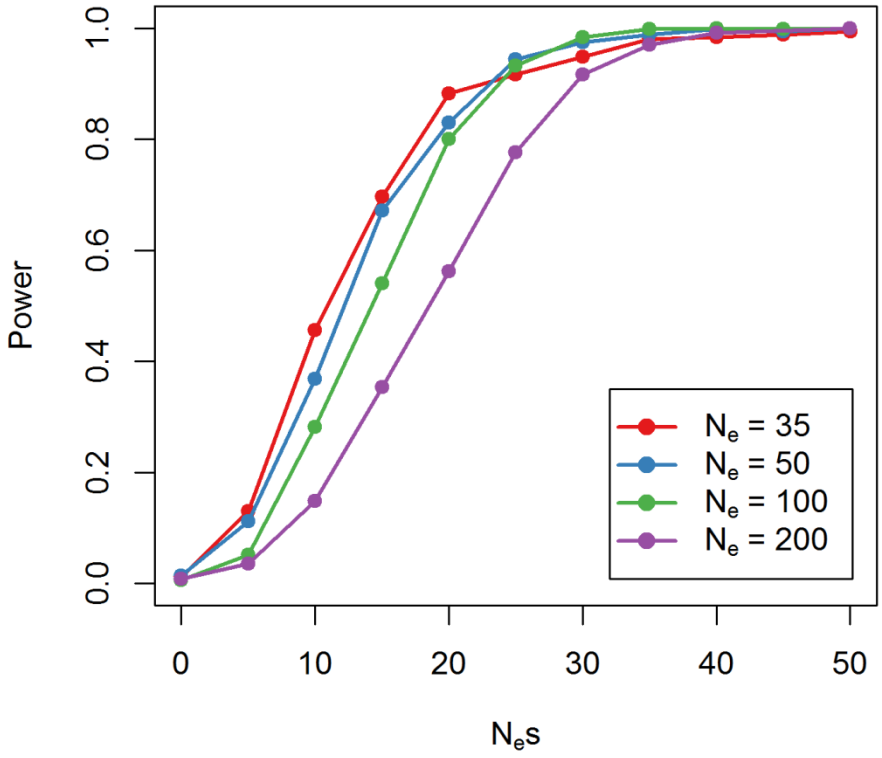



**Figure 5**

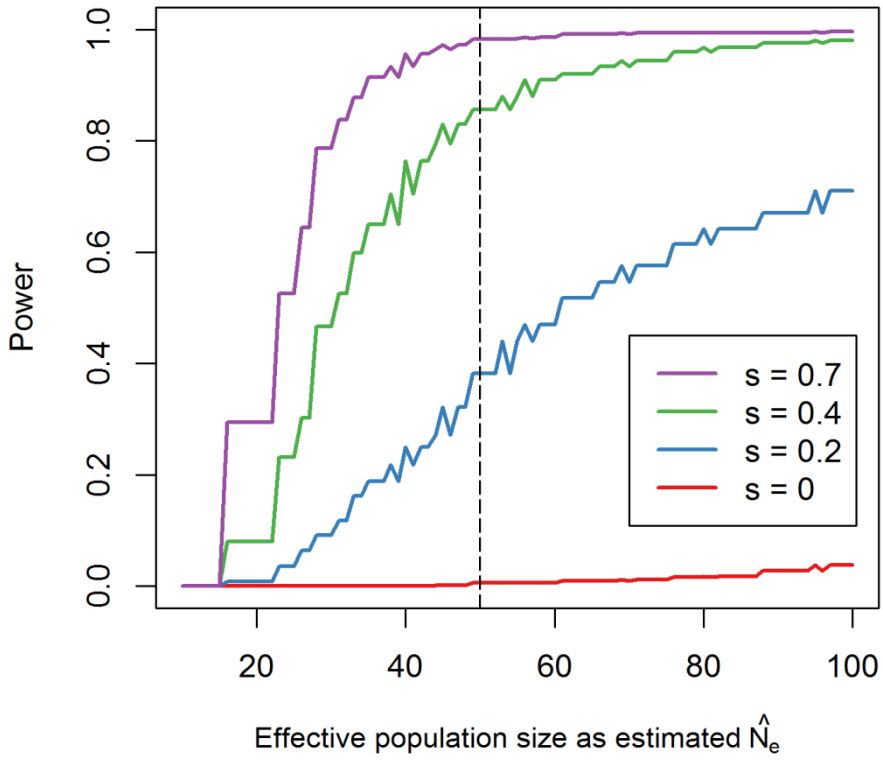


**Figure 6**

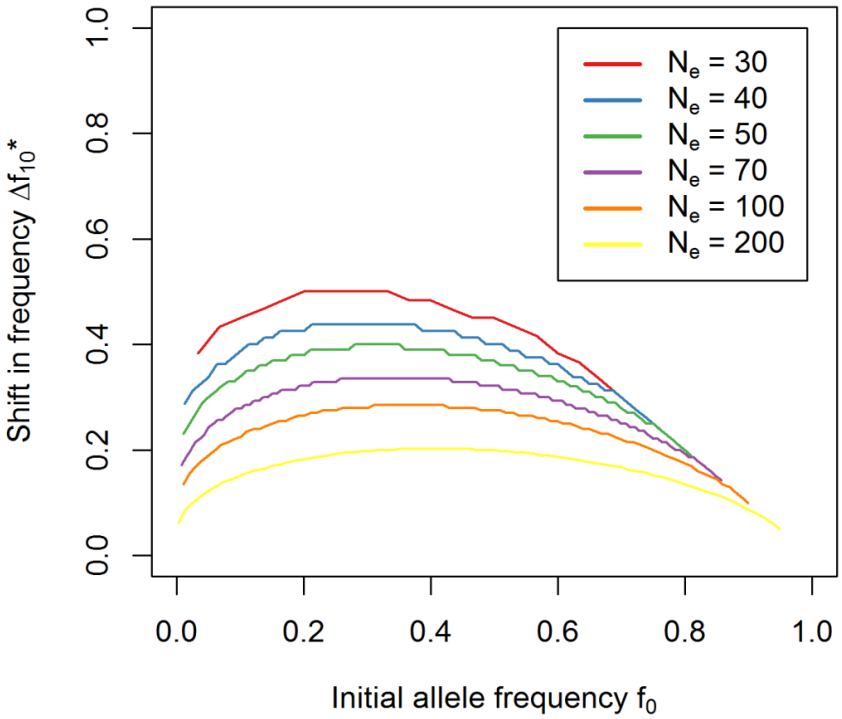
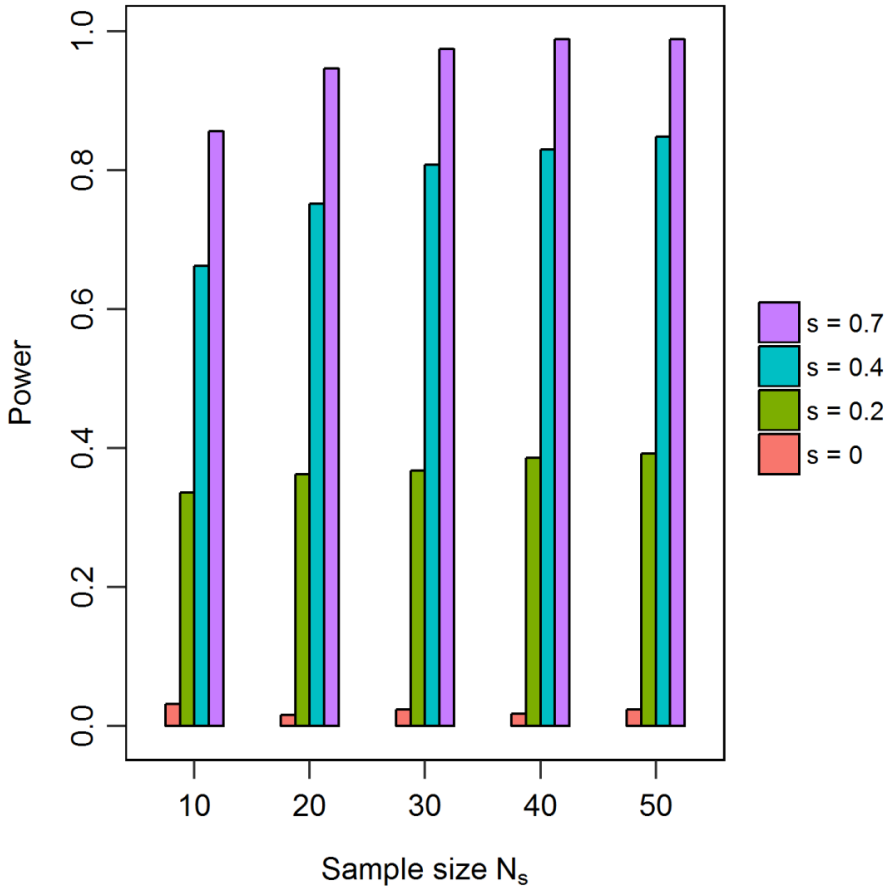



**Figure 7**



# Figure 8

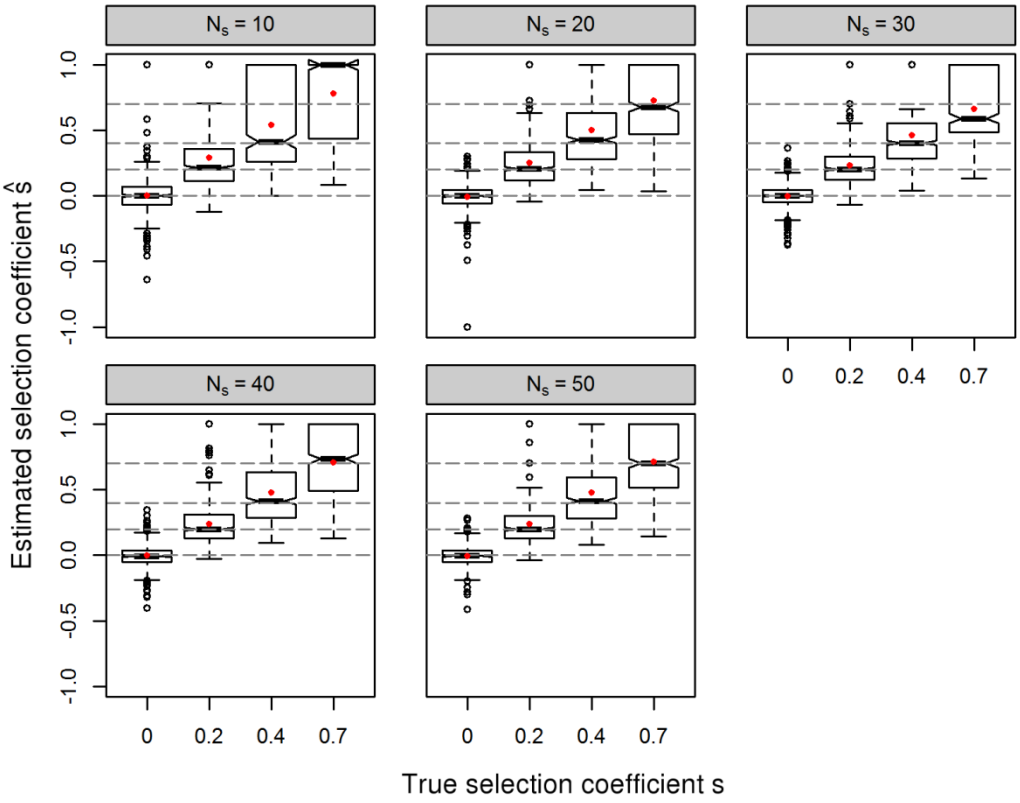

$N_s$ = 10
$N_s$ = 20
$N_s$ = 30
$N_s$ = 40
$N_s$ = 50
Estimated selection coefficient $\hat{s}$
-1.0
-0.5
0.0
0.5
1.0
0
0.2
0.4
0.7
True selection coefficient s

**Figure 9**

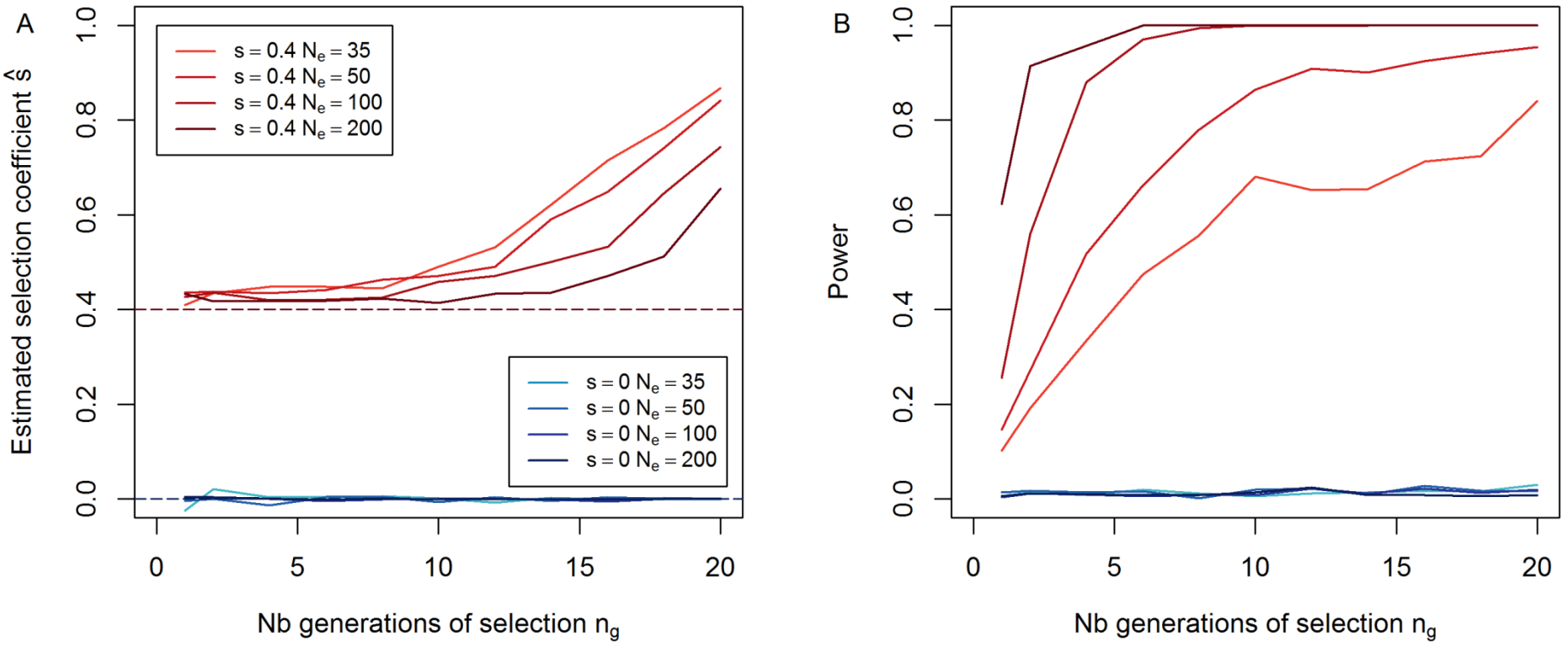

A
Estimated selection coefficient ŝ
s = 0.4 Ne = 35
s = 0.4 Ne = 50
s = 0.4 Ne = 100
s = 0.4 Ne = 200
s = 0 Ne = 35
s = 0 Ne = 50
s = 0 Ne = 100
s = 0 Ne = 200
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Nb generations of selection ng
B
Power
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Nb generations of selection ng

**Figure 10**

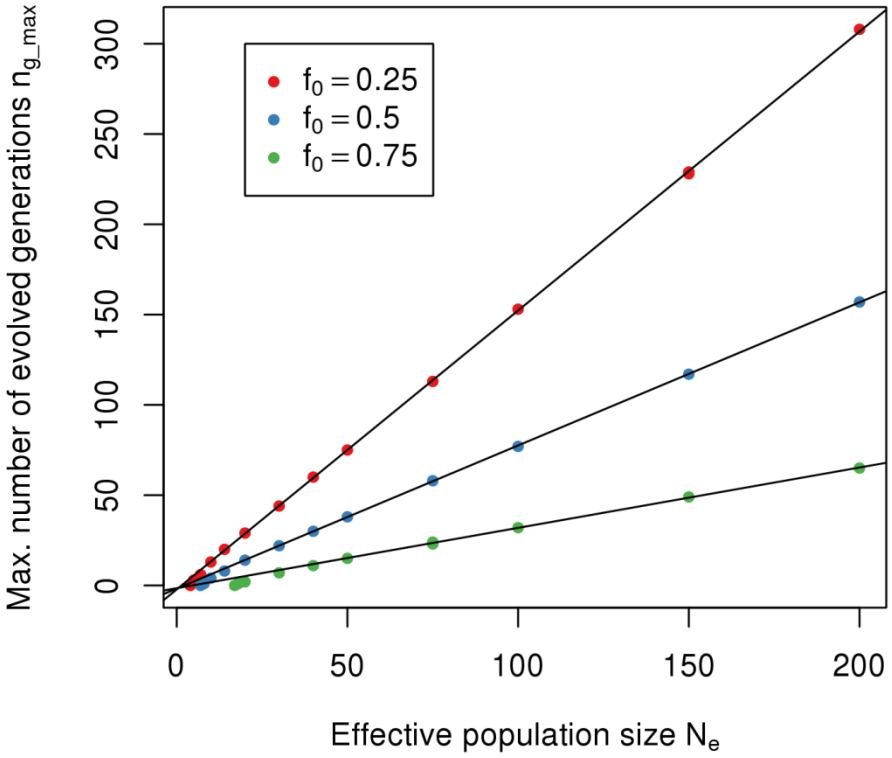



**Figure 11**

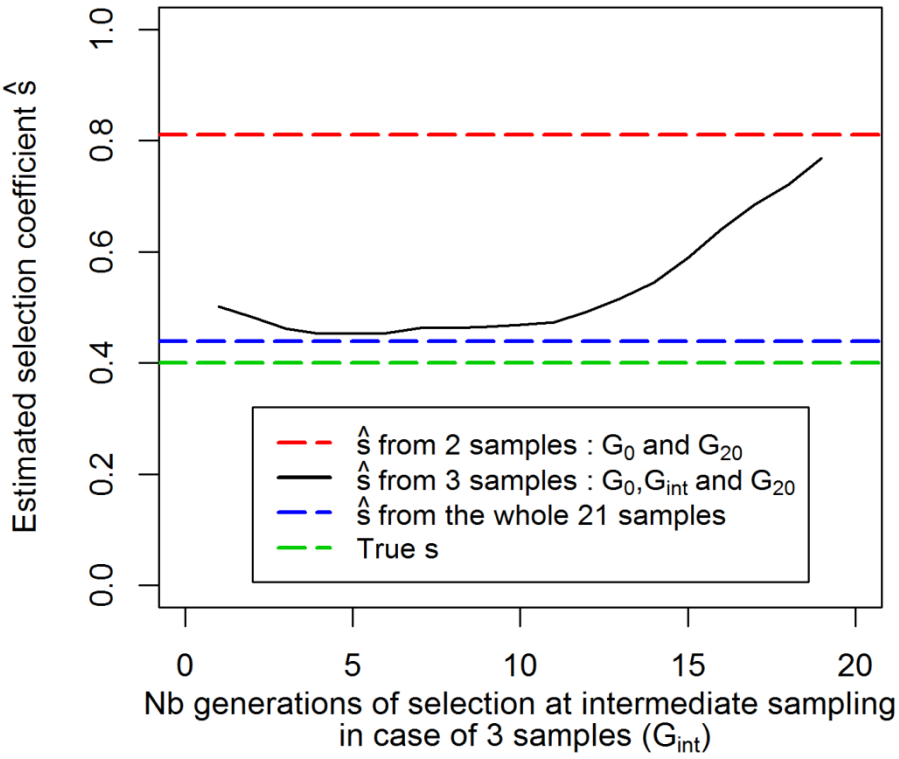


**Figure 12**

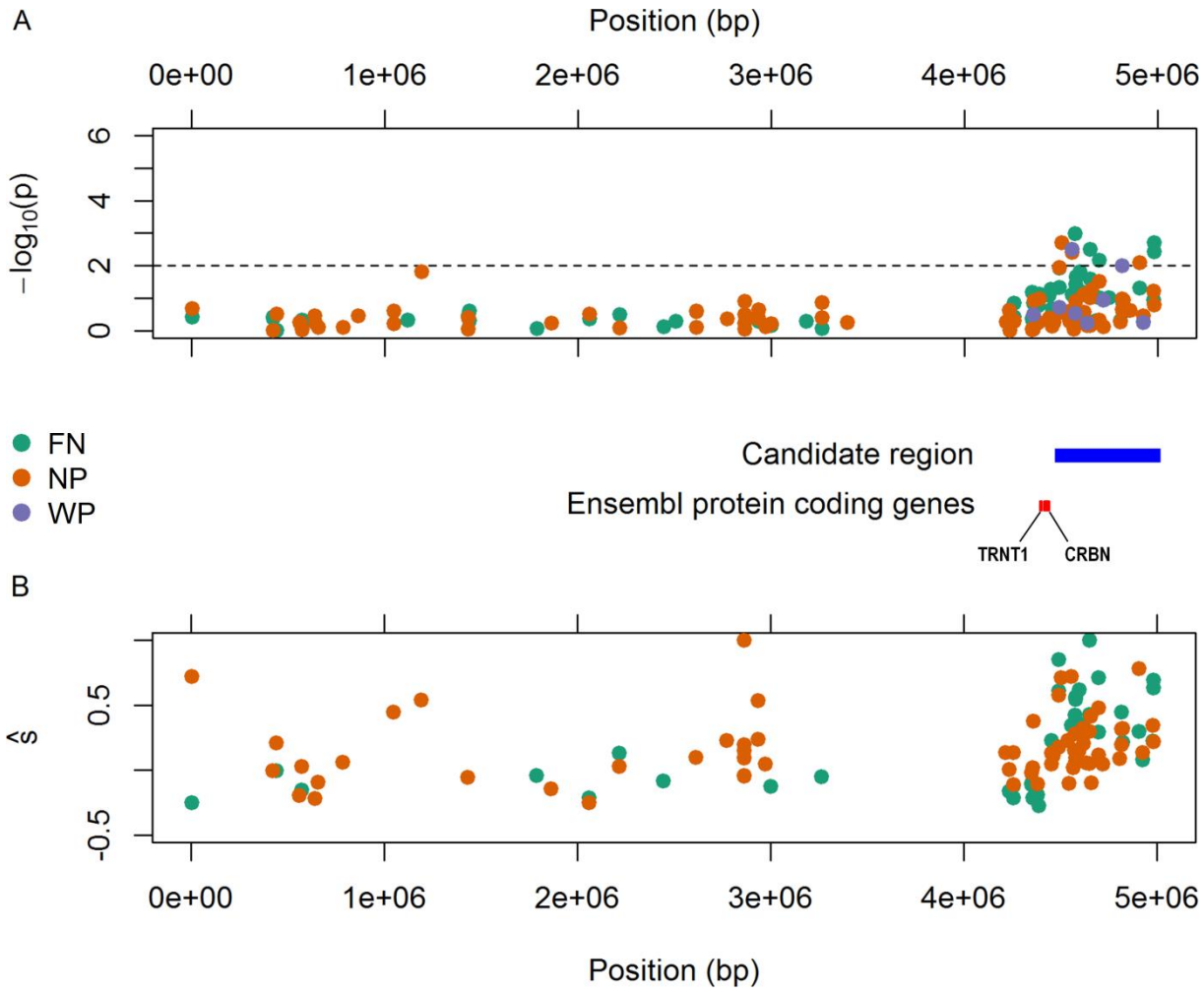

A
Position (bp)
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
$-\log_{10}(p)$
0
2
4
6
FN
NP
WP
Candidate region
Ensembl protein coding genes
TRNT1
CRBN
B
$\hat{s}$
-0.5
0.5
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Position (bp)

**Figure 13**

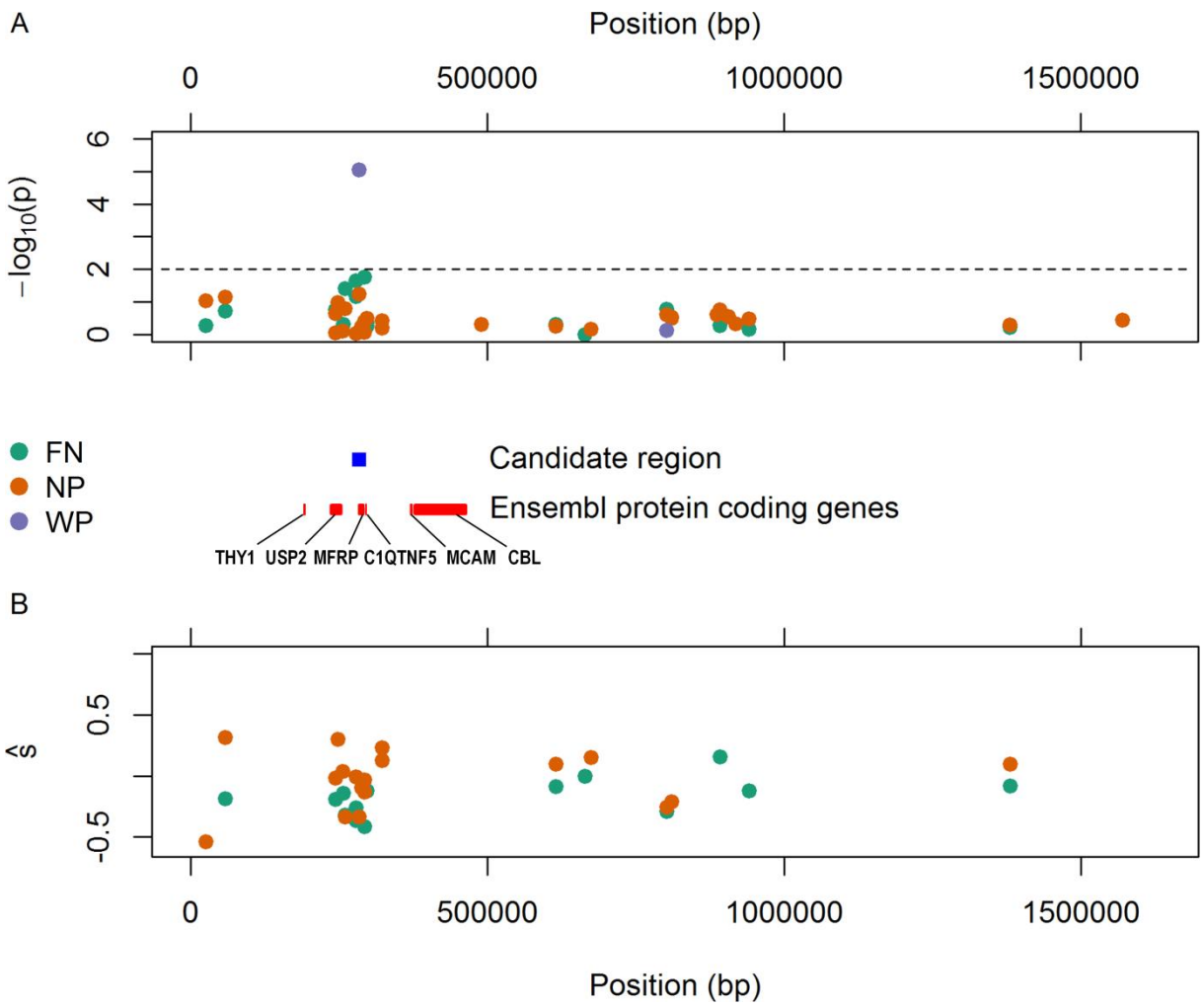

A
Position (bp)
0
500000
1000000
1500000
$-\log_{10}(p)$
0
2
4
6
FN
NP
WP
Candidate region
Ensembl protein coding genes
THY1
USP2
MFRP
C1QTNF5
MCAM
CBL
B
$\hat{s}$
0.5
-0.5
0
500000
1000000
1500000
Position (bp)

**Figure 14**

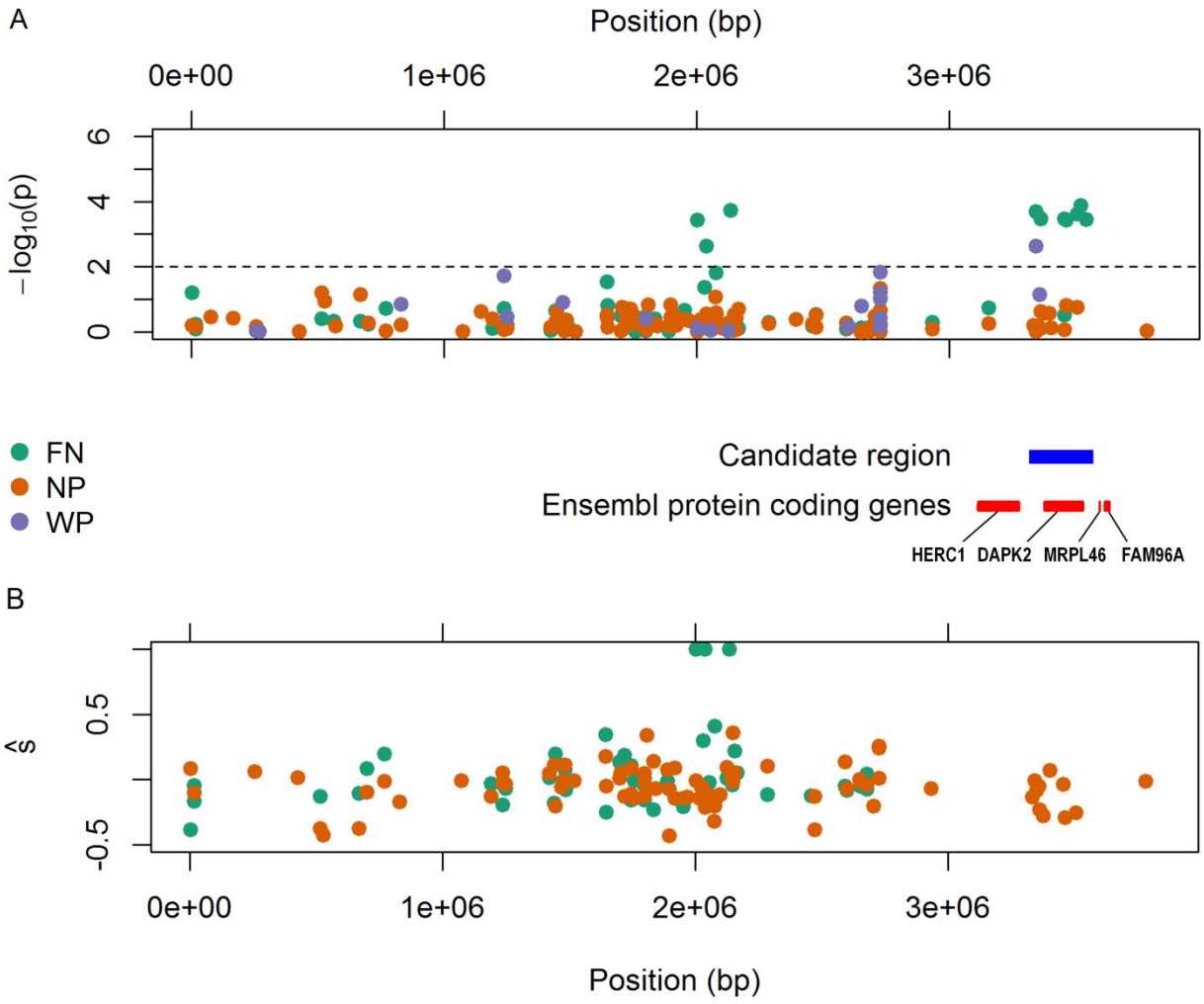
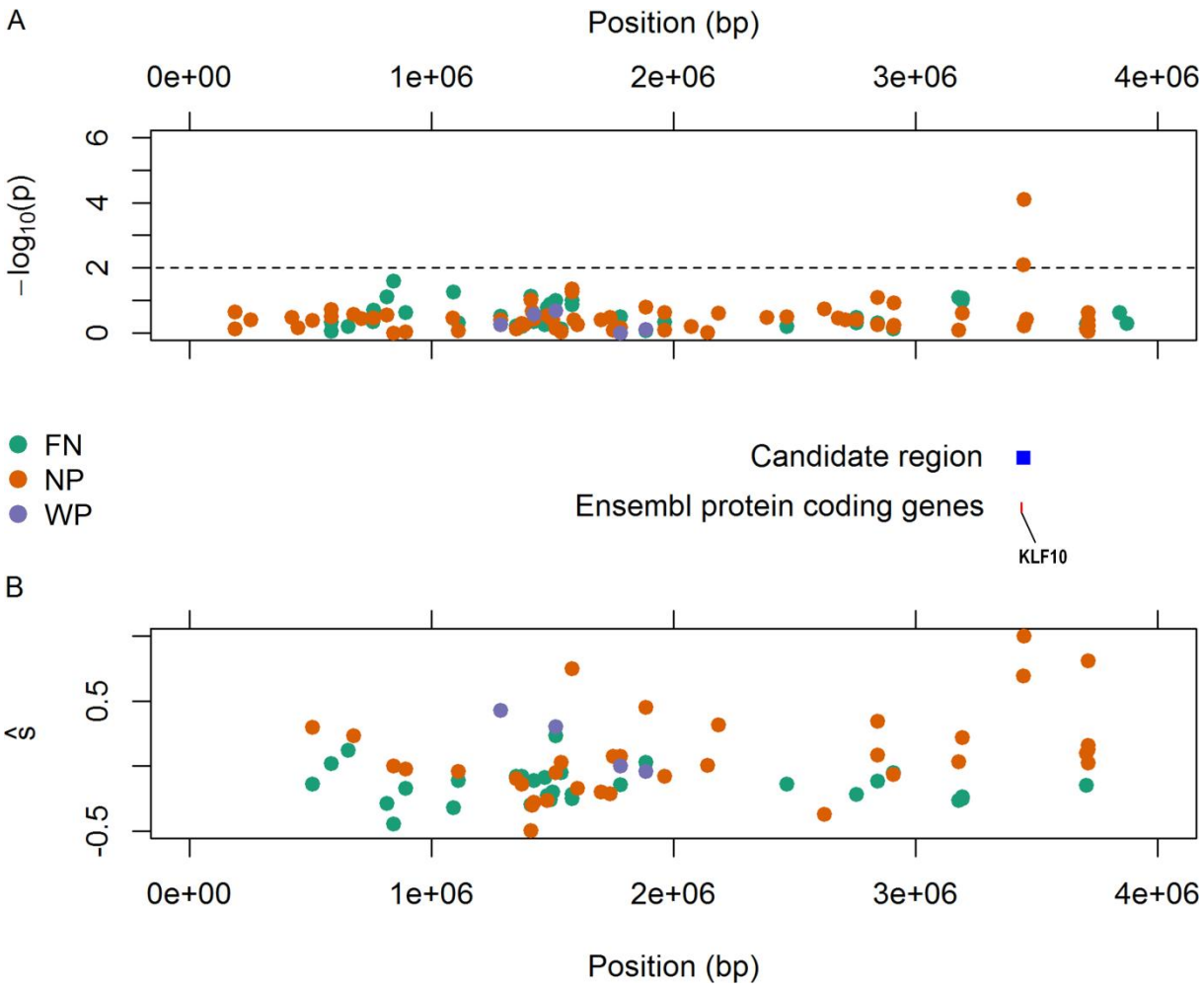

A
Position (bp)
0e+00
1e+06
2e+06
3e+06
$-\log_{10}(p)$
0
2
4
6
FN
NP
WP
Candidate region
Ensembl protein coding genes
HERC1
DAPK2
MRPL46
FAM96A
B
$\hat{s}$
0.5
-0.5
0e+00
1e+06
2e+06
3e+06
Position (bp)

**Figure 15**


A
Position (bp)
0e+00
1e+06
2e+06
3e+06
4e+06
$-\log_{10}(p)$
0
2
4
6
FN
NP
WP
Candidate region
Ensembl protein coding genes
KLF10
B
$\hat{s}$
-0.5
0.5
0e+00
1e+06
2e+06
3e+06
4e+06
Position (bp)

# Chapitre IV

# TABLE DES MATIERES

-

# Chapitre IV – La réponse évolutive du diable de Tasmanie au cancer

## — 1. Introduction

### 1.1. La généralisation de l'*Open data*[1] favorise la découvrabilité des données

En génomique, la question de l'accessibilité des données de séquence a émergé dès la fin des années 90 (Rowen *et al.,* 2000). Plus généralement, les possibilités offertes par l'évolution du matériel informatique depuis les années 2000, avec notamment l'avènement du *cloud computing*, ont initié de nouvelles pratiques dont l'encadrement légal est devenu une véritable question publique. La France a ainsi lancé du 26 septembre au 18 octobre 2015 une consultation publique dans le cadre du projet de loi « pour une République numérique ». Les débats menés sur Internet dans le cadre de cette consultation sont un bon marqueur de l'importance stratégique prise par les questions d'*Open data*. Même si la promulgation du texte de loi final est récente (octobre 2016), on perçoit dès maintenant que ce cadre juridique va renouveler/amplifier certaines pratiques dans l'enseignement supérieur et la recherche. En particulier, le texte de loi fixe des conditions d'exploitation des publications scientifiques, ce qui était un des points d'achoppement particulièrement visibles des débats. Sur cette question, la loi octroie aujourd'hui officiellement aux auteurs de travaux sur financement majoritairement public un droit d'exploitation secondaire. Cela signifie que l'auteur peut mettre la version finale de son manuscrit accepté par l'éditeur (c'est-à-dire la version obtenue au terme du processus de validation par les pairs) en accès libre sur une plateforme d'archivage ouverte (par exemple bioRxiv, HAL, prodinra…), moyennant un délai de six mois à compter de la date de la publication initiale et l'accord des co-auteurs. Cette disposition n'est peut-être pas la plus spectaculaire, mais elle est avant tout symbolique du fait de la reconnaissance officielle de droits de mise à disposition de leur production aux publiants. Voir les grandes questions de partage des données ainsi clarifiées par les pouvoirs publics incite les acteurs de la recherche (organismes de recherche, éditeurs) à réfléchir à leurs actions stratégiques en matière de gestion des données à l'ère de l'*Open Data*.

[1] L'anglicisme *Open data*, que l'on pourrait traduire par « données numériques ouvertes », est aujourd'hui un terme qui s'est imposé dans notre langue au point d'être conservé sur le portail de communication du gouvernement français. Nous avons conservé cet anglicisme dans ce paragraphe, au même titre que d'autres, comme *data article* ou *data journal*, qui sont également d'usage courant.

Dans ce contexte, les éditeurs contribuent à rendre les données explicites, ce qui peut consister à exiger des auteurs la mise à disposition des données sur lesquelles s'appuie leur publication dans un entrepôt dédié, comme *Dryad* (http://datadryad.org) ou *figshare* (https://figshare.com). Une fois archivé sous la forme d'un ou plusieurs fichiers informatiques et accompagné de métadonnées (le plus souvent un texte descriptif très succinct du contenu des fichiers), le jeu de données est librement accessible sur l'entrepôt et peut être cité de façon autonome. De plus en plus de journaux de premier plan, comme *Proceedings of the Royal Society* ou *PLoS One*, recommandent fortement voire imposent aux auteurs l'utilisation de l'un de ces deux entrepôts. Lancé en 2008, l'entrepôt de données *Dryad* héberge aujourd'hui près de 20000 jeux de données (Fig. IV-1), illustrant l'impact des politiques éditoriales actuelles en matière d'accès aux données. Le rôle de ces entrepôts de données est de garantir l'archivage et l'accès aux données associées aux publications. Ils offrent notamment aux revues des services personnalisés permettant de coordonner la soumission de manuscrits et celle des données correspondantes au sein d'un processus unique.

En parallèle, les éditeurs proposent de valoriser les jeux de données sous une forme nouvelle de publication, fréquemment désignée sous le terme de *data article* ou article de données. Les *data articles* sont des articles à part entière ayant pour objectif d'assurer une meilleure visibilité aux données en fournissant un cadre formalisé à leur partage (Austin *et al.,* 2015). A la différence des articles de recherche classiques, les *data articles* s'intéressent aux jeux de données afin de les rendre « accessibles, interprétables et réutilisables plutôt que de tester des hypothèses ou présenter de nouvelles analyses » (Austin *et al.,* 2015). Ils adoptent ainsi une structure un peu différente des articles classiques, avec des sections intitulées par exemple « *Specification table* », « *Value of data* », « *Dataset descriptors* », « *Experimental design* », ou encore « *Technical validation* », plus adaptées à la mise en avant des données que les habituels « *Introduction* », « *Results* », « *Discussion* » et « *Materials and Methods* ». Depuis 2014, le groupe *Nature Publishing* édite *Scientific Data*, un *data journal* (littéralement, un « journal de données ») destiné à valoriser des jeux de données de haute qualité. La même année, l'éditeur concurrent *Elsevier* lance *Data in Brief*, un *data journal* adoptant un format davantage calqué sur celui des communications courtes. Le succès est au rendez-vous pour ces deux premiers *data journals* généralistes qui sont indexés dans les moteurs de recherche de données bibliographiques et bénéficient aujourd'hui de métriques traditionnelles, comme les journaux conventionnels (Fig. IV-1). En moins de trois ans, *Data in Brief* a publié plus de 1000 *data articles*, tandis que *Scientific Data* a atteint un facteur d'impact de 4,8. En offrant des solutions pérennes autant à ceux qui produisent les données en leur permettant de publier, dans un contexte académique où la bibliométrie est perçue comme le premier indicateur de performance, qu'à ceux qui recherchent des données en leur fournissant un accès gratuit à des jeux de données ayant fait l'objet d'un contrôle par

les pairs, les *data journals* devraient devenir de plus en plus attractifs aux chercheurs et, dans leur sillage, les entrepôts de données devraient encore gagner en visibilité.

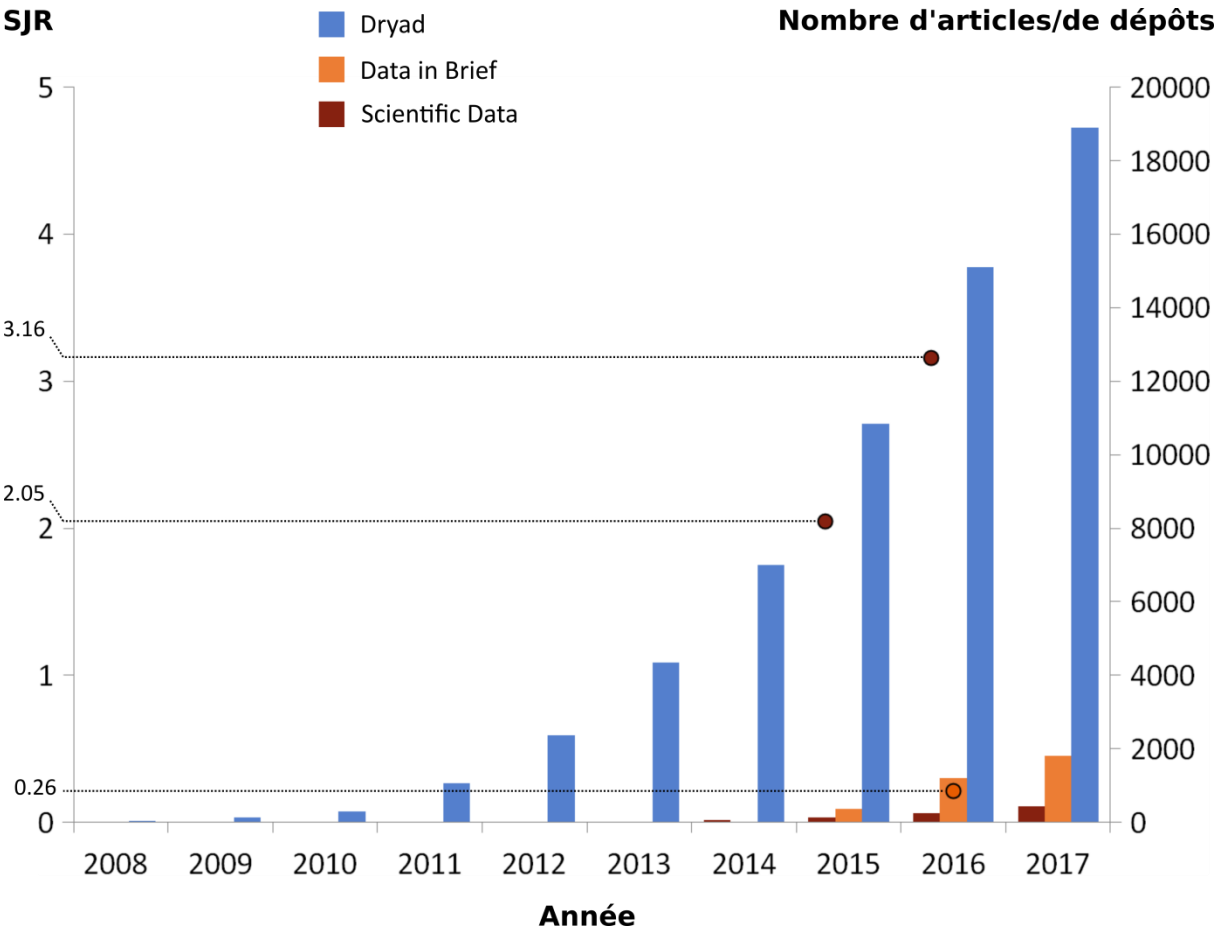



**Figure IV-1. Indicateurs bibliométriques de deux *data journals* (*Data in Brief* et *Scientific Data*) et d'un entrepôt de données (*Dryad*).** L'axe des ordonnées de gauche mentionne le *SJR* des journaux, disponible depuis 2015 pour *Scientific Data* et depuis 2016 pour *Data in Brief*. L'axe des ordonnées de droite indique le nombre cumulé d'articles (pour les journaux) ou de dépôts (pour *Dryad*) publiés. Pour 2017, il inclut les données disponibles sur les dix premiers mois de l'année. *SJR* = indicateur *SCImago Journal Rank*. Le *SJR* est une alternative *open access* au facteur d'impact (Falagas *et al.,* 2008).

Les organismes de recherche sont actuellement eux aussi très sensibles aux questions de valorisation et de réutilisation des données scientifiques. L'INRA mène par exemple une politique volontariste en matière d'*Open data.* L'accès aux données produites au sein de ses laboratoires est un chantier prioritaire pour l'institut qui a en fait (sous le terme « *Open Science* ») une de ses trois orientations politiques dans le cadre des choix stratégiques pour 2025 (INRA, 2017). Ce projet vise à installer une double dynamique : (i) favoriser la transparence des travaux auprès du public et (ii) promouvoir le partage des données entre personnels scientifiques dans le périmètre de l'institut en facilitant leur

réutilisation. Dans cette optique, le projet d'un portail de données propre à l'INRA vient d'aboutir à un service opérationnel (https://data.inra.fr/) au premier trimestre 2018. Ce type de dispositif s'inscrit dans une démarche de renforcement des pratiques collaboratives entre équipes de recherche et devrait faciliter les synergies entre modélisateurs et expérimentateurs.

Ainsi, ces dernières années ont initié un tournant important dans la gestion des données produites par la recherche académique. Les acteurs-clés de la recherche, publics comme privés, ont mis en place des politiques favorisant la visibilité de ces données ainsi que leur réutilisation. L'émergence des *data journals* et le succès naissant des entrepôts de données marquent notamment une évolution récente des pratiques éditoriales. Parmi les bénéfices attendus de ces nouvelles pratiques, on peut penser qu'elles faciliteront les méta-analyses ou encore qu'elles offriront des débouchés à des travaux théoriques ou méthodologiques (Thelwall *et al.,* 2017). Dans le cadre de cette thèse, nous avons bénéficié de ce contexte d'*Open Data* accru, et notamment de la mise à disposition des jeux de données adossés aux publications.

Un de nos objectifs était de valider nos travaux méthodologiques et de mettre en évidence leurs points forts au travers d'une application à un jeu de données réelles. Nous avions noué plusieurs contacts, essentiellement lors de congrès, avec des équipes désireuses de faire analyser des données. Cependant, ces discussions n'ont pas eu de perspectives concrètes. Nous avons en particulier participé à un appel à projets dans le but d'étudier la réponse adaptative de *Drosophila subobscura* à des perturbations environnementales soudaines, en collaboration avec Sofia Seabra (*Centre for Ecology, Evolution and Environmental Changes*, Lisbonne), mais le projet n'a finalement pas obtenu de financement. Nous nous sommes alors tournés vers les jeux de données déjà publiés et disponibles en accès libre sur des entrepôts de données comme *Dryad*. Nous avons identifié un jeu de données issu de populations de diable de Tasmanie (*Sarcophilus harrisii*) qu'il semblait intéressant de soumettre à notre méthode. Ce jeu de données comprend des séries génomiques temporelles acquises entre 1999 et 2014. Il est adossé à une publication mettant en évidence l'effet d'une sélection récente sur deux régions génomiques au sein des populations de diable de Tasmanie.

Nous avons donc bénéficié des politiques d'*Open data* émergentes pour réanalyser des données réelles correspondant à un cas de figure auquel notre méthode de détection de signatures de sélection semblait bien adaptée, ce qui devrait contribuer à une meilleure compréhension de la réponse évolutive du diable de Tasmanie à un cancer transmissible très agressif. A l'image de cette application, nous pensons que les travaux d'analyse des données publiées en accès libre vont devenir plus communs, stimulant la réflexion autour du développement de nouvelles méthodes, et permettant de mieux comprendre certains phénomènes évolutifs et génétiques importants.

## 1.2. Le diable de Tasmanie (*Sarcophilus harrisii*)

Le diable de Tasmanie est une espèce singulière à plus d'un titre. Il s'agit tout d'abord du plus gros marsupial carnivore encore en vie depuis l'extinction du thylacine (*Thylacinus cynocephalus*) au siècle précédent. L'histoire démographique du diable de Tasmanie a abouti à une situation aujourd'hui très particulière. La succession de plusieurs goulots d'étranglement, notamment du fait des changements climatiques du début de l'Holocène (Brüniche-Olsen *et al.,* 2014), est responsable d'un très faible niveau de diversité génétique chez les populations contemporaines (Jones, M.E., *et al.,* 2004 ; Storfer *et al.,* 2017). L'aire de répartition de ces populations est aujourd'hui limitée à la Tasmanie, une île d'un peu plus de 60000 $km^2$. En plus de son statut d'espèce insulaire, le diable de Tasmanie semble être particulièrement vulnérable au cancer (Madsen *et al.,* 2017), ce qui amplifie considérablement les préoccupations concernant le futur de l'espèce. Des travaux ont suggéré qu'une telle accumulation de facteurs défavorables pourrait conduire à la perte de populations locales dans les années à venir, voire à la disparition de l'espèce à l'état sauvage avant 2050 (McCallum *et al.,* 2007 ; McCallum 2008 ; McCallum *et al.,* 2009). Face à ce risque, le gouvernement australien a financé de nombreux projets visant à la sauvegarde du diable de Tasmanie (Lachish *et al.,* 2010 ; Gooley *et al.,* 2017 ; Tovar *et al.,* 2017), faisant de l'espèce un véritable cas d'étude en génétique de la conservation (McCallum, 2008 ; Hendricks *et al.,* 2017).

La vigilance actuelle quant au devenir de l'espèce a pour origine la découverte d'un cancer transmissible dans les populations du nord-est de la Tasmanie. Apparue à la fin des années 1990, la pathologie se transmet par morsure lors des contacts sociaux et se caractérise par l'apparition de tumeurs primitives bien visibles sur la face, le cou ou dans la cavité buccale des animaux infectés (Hawkins *et al.,* 2006 ; Hamede *et al.,* 2013), d'où son nom de tumeur faciale du diable de Tasmanie ou DFTD (*Devil Facial Tumor Disease*). Les cellules cancéreuses sont issues d'au moins deux lignées clonales différentes, DFT1 et DFT2, et se transmettent d'un individu à l'autre par allogreffe (Pye *et al.,* 2016a). La prévalence de ce type de phénomène, c'est-à-dire la survie d'un cancer au-delà du décès de son hôte initial du fait d'un transfert horizontal des cellules cancéreuses, est difficile à estimer au sein de la faune sauvage (Ujvari *et al.,* 2016). Certains auteurs ont suggéré que les populations de diable de Tasmanie étaient exposées à l'émergence de ce type de pathologie en raison d'une faible diversité des gènes composant le complexe majeur d'histocompatibilité (CMH), ce qui aurait limité leur capacité à produire une réponse immunitaire adaptée à un nouveau risque infectieux (Belov, 2011, 2012 ; Cheng, Y., *et al.,* 2012).

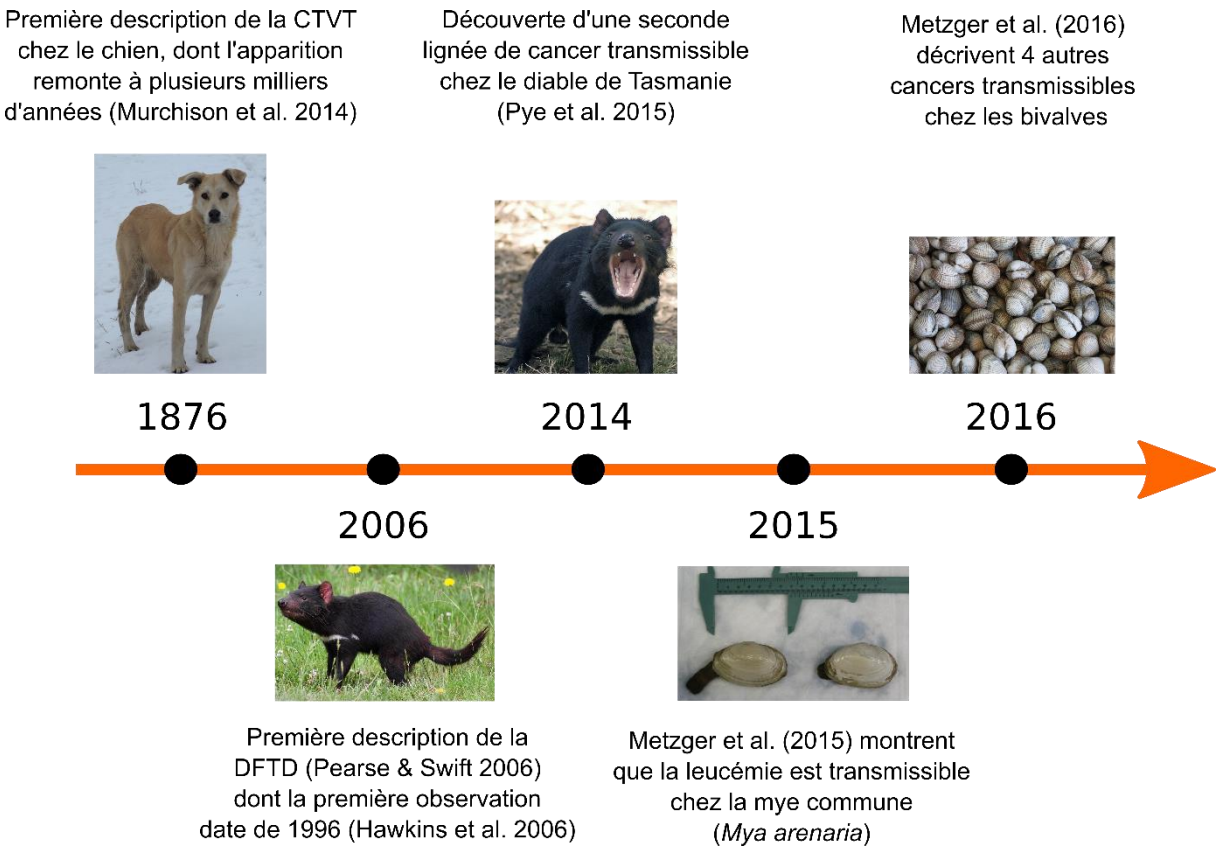



**Figure IV-2. Chronologie de la découverte de cancers transmissibles au sein des populations animales (modifié à partir de Caldwell & Siddle, 2017, Fig. 1).** A l'heure actuelle, huit cancers transmissibles ont été décrits dans la nature (un chez le chien, deux chez le diable de Tasmanie, et cinq chez les bivalves). CTVT : *Canine Transmissible Venereal Tumor* ; DFTD : *Devil Tumor Facial Dissease*.

Aujourd'hui, l'observation de cancers transmissibles est extrêmement rare au sein des populations naturelles (Fig. IV-2). Le seul autre cancer contagieux naturel observé chez un Mammifère est la tumeur vénérienne transmissible canine ou CTVT (*Canine Transmissible Venereal Tumor*), endémique dans de nombreux pays et dont l'apparition remonte à plusieurs milliers d'années (Ostrainder *et al.,* 2016). La CTVT occasionne des métastases dans seulement 7% des cas, alors que la DFTD est caractérisée par une croissance tumorale rapide et une forte fréquence (65%) des métastases (Belov, 2012). La DFTD est ainsi associée à un taux de mortalité de presque 100% dans les douze mois suivant la contamination de l'hôte (Belov, 2012 ; Pye *et al.,* 2016a). Cette pathologie exerce par conséquent une forte pression sélective sur les populations hôtes.

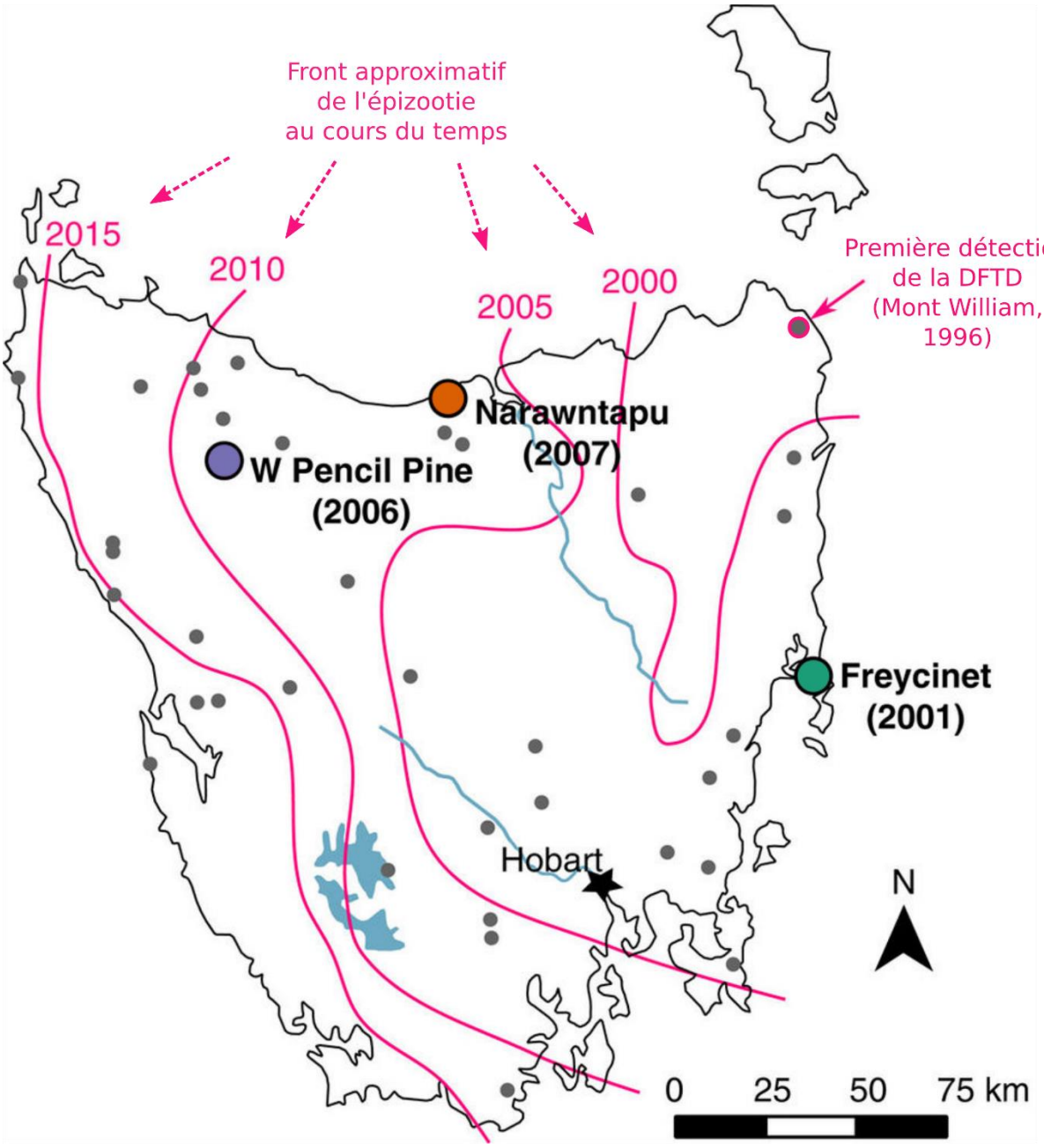



**Figure IV-3. Progression de l'épizootie de DFTD à travers la Tasmanie et sites d'échantillonnage des populations de diable de Tasmanie (modifié à partir d'Epstein *et al.*, 2016a, Fig. 1).** La DFTD a été détectée pour la première fois en 1996 dans le parc national du Mont William (Hawkins *et al.,* 2006). En l'espace d'une vingtaine d'années, ce cancer transmissible responsable d'une mortalité quasi-systématique s'est répandu à 95% de l'aire de répartition du diable de Tasmanie (Storfer *et al.,* 2017). Les positions successives du front de l'épizootie en 2000, 2005, 2010, et 2015, sont indiquées par des lignes magenta. Les petits disques gris identifient les sites de prélèvements des échantillons dont les génotypages sont disponibles sur *Dryad* (Epstein *et al.,* 2016b). Les trois populations analysées dans Epstein *et al.* (2016a) sont indiquées par les larges cercles remplis de couleurs diférentes. Le nom des trois sites de prélèvement correspondants est mentionné, ainsi que, entre parenthèses, la date supposée d'arrivée de la pathologie au sein de ces populations.

En plus de la métapopulation conservée en captivité plus ou moins stricte en vue de repeuplements ou de réintroductions (Hogg *et al.,* 2016), seules quelques petites populations situées à l'extrême nord-ouest ou à l'extrême sud-ouest de la Tasmanie seraient pour l'instant encore épargnées par la maladie (Storfer *et al.,* 2017). L'épizootie de DFTD s'est en effet rapidement propagée à la quasi-totalité de l'aire de répartition du diable de Tasmanie et y a eu des conséquences démographiques importantes. On estime que 80% des 130000 à 150000 individus présents sur l'île dans les années 1990 ont aujourd'hui disparu du fait de la DFTD, avec des pertes culminant localement à 95% (Storfer *et al.,* 2017). Malgré tout, les populations du nord-est persistent encore à une très faible densité 20 ans après l'apparition de la maladie (Epstein *et al.,* 2016a). De plus, certaines données récentes suggèrent que les populations de l'ouest seraient plus résilientes (Miller *et al.,* 2011 ; Hamede *et al.,* 2012 ; Pye *et al.,* 2015 ; Save the Tasmanian Devil Program, 2017). Ces éléments suggèrent la possibilité d'une réponse à la forte sélection exercée par la DFTD en dépit des faibles niveaux de diversité génétique affichés par les populations de diable de Tasmanie.

En septembre 2016, un article publié dans *Nature Communications* a fait état d'une réponse évolutive à la DFTD (Epstein *et al.,* 2016a). Les auteurs ont travaillé sur des séries génomiques temporelles acquises sur une durée de 7 à 14 ans au sein de trois sites de prélèvement (Fig. IV-3): Freycinet (FN), Narawntapu (NP) et West Pencil Pine (WP). Ils ont proposé deux régions génomiques comme candidates à la sélection en identifiant des SNP exhibant conjointement une forte variance temporelle de leur fréquence allélique et une forte augmentation de leur DL avec les locus voisins, le tout consécutivement à l'arrivée de la DFTD. Cette approche a permis de mettre en évidence sept gènes candidats dont les fonctions supposées sont principalement liées au risque de cancer et à l'immunité. De tels résultats constituent une étape importante dans la compréhension de la réponse d'un hôte à une pression de sélection forte générée par l'arrivée d'un nouveau pathogène dans l'environnement.

Ainsi, seulement une poignée de générations auraient suffi aux populations de diable de Tasmanie pour initier une réponse évolutive, alors même que les effectifs sont limités ($N_e \approx 30$). Les résultats issus de nos simulations (*cf.* Chapitre III) suggèrent que notre méthode de vraisemblance permet d'identifier des signatures de sélection lorsqu'une sélection forte est appliquée pendant quelques générations à une population de faible effectif. Conformément à l'actuelle politique de mise à disposition des données d'un journal comme *Nature Communications*, le jeu de données utilisé par Epstein *et al.* (2016a) a été déposé sur l'entrepôt public *Dryad*. Nous avons donc choisi de travailler sur les mêmes séries génomiques temporelles que celles analysées par Epstein *et al.* (2016a) afin d'illustrer sur un jeu de données réelles les possibilités offertes par notre méthode.

Nous avons ainsi pu montrer que notre méthode permettait non seulement de retrouver les deux régions candidates initialement détectées mais aussi et surtout d'identifier au total une centaine de signatures de sélection. Parmi ces régions candidates, soixante contiennent des gènes orthologues chez l'Homme. Nos résultats suggèrent par conséquent que la réponse évolutive à la DFTD se traduirait par une empreinte génomique plus étendue que précédemment envisagé. Ces résultats sont présentés dans le présent chapitre, mais ont aussi fait l'objet d'un manuscrit proposé à *PLoS One* (Article II), dont la version de soumission est donnée en annexe (Annexe I).

# — 2. Matériels et méthodes

## 2.1. Jeu de données

Les données brutes du génotypage par *RAD-seq* (voir Chapitre V pour une présentation du *RAD-seq*) de 360 individus prélevés en 38 différents sites balayant l'aire de répartition du diable de Tasmanie (Fig. IV-4) sont publiquement accessibles (Epstein *et al.,* 2016b) sur l'entrepôt de données *Dryad* (http://datadryad.org/). Nous avons contacté l'auteur de correspondance pour le prévenir que nous souhaitions analyser de nouveau ces données dans le cadre du test d'une nouvelle méthode, et pour obtenir un complément d'information, en particulier sur certaines étapes de filtrage peu détaillées dans Epstein *et al.* (2016a), mais nous n'avons pas obtenu de réponse. Nous avons donc essayé de reproduire fidèlement le traitement des données brutes sur la base des indications disponibles dans le manuscrit afin de travailler avec un jeu de données aussi proche que possible de celui qui a permis de détecter les deux régions candidates à la sélection.

Comme dans Epstein *et al.* (2016a), nous avons limité l'analyse aux trois populations arborant les plus grands échantillons, c'est-à-dire ceux prélevés au sein des localités de Freycinet, Narawntapu et West Pencil Pine. Dans ces populations, des séries génomiques temporelles ont pu être collectées sur une période allant de 1999 à 2014 (Fig. IV-4), ce qui correspond à de 3 à 6 générations avec un intervalle de génération communément estimé à 2 ans chez le diable de Tasmanie. La taille des échantillons est comprise entre 20 et 43 individus, et les estimations de l'effectif efficace ($N_e$) des populations FN, NP, et WP, atteignent 34, 37, et 26 individus efficaces, respectivement.

Les données se présentent sous la forme d'un fichier *.vcf* (*Variant Call Format*) version 4.2 qui répertorie les génotypes obtenus à l'issue de l'étape de constitution d'un catalogue de *RAD-tags* (*cf.* Chapitres V et VI) pour chaque locus et chaque individu. Conformément aux indications d'Epstein *et al.* (2016a), nous avons :

- écarté les SNP dont la MAF (*Minor Allele Frequency*) totale est inférieure à 0,01
- écarté les SNP dont l'hétérozygotie observée est supérieure à 0,5
- écarté les SNP génotypés chez moins du tiers des individus
- limité l'analyse à deux échantillons temporels dans les populations FN (1999 et 2012-2013) et WP (2006 et 2013-2014), et à trois échantillons temporels dans la population NP (1999, 2004 et 2009)
- écarté les SNP génotypés chez moins d'un tiers des individus au sein des échantillons retenus, sauf pour les échantillons WP1 et FN2 où un minimum de 10 individus a été requis

- écarté les SNP générant un trop fort DL local (nous avons utilisé *PLINK* (Purcell *et al.,* 2007) pour éliminer un SNP des paires dont le $R^2$ était supérieur à 0,99 dans un rayon de 20 SNP consécutifs et de 50 kb au sein d'une même *scaffold*).

Nous avons en outre écarté les SNP ne présentant pas de polymorphisme dans les échantillons précédant l'arrivée de la DFTD. Ainsi, nous avons soumis à notre méthode de détection de signatures de sélection un jeu de données composé de 16978, 27173 et 5401 SNP correspondant aux populations FN, NP et WP, respectivement.

## 2.2. Identification des signatures de sélection

Pour détecter des signatures de sélection, Epstein *et al.* (2016a) ont utilisé une approche empirique de type *outlier*, dans le sens où les SNP retenus appartiennent à une fraction arbitraire exhibant des statistiques extrêmes par rapport à la moyenne constatée sur l'ensemble du jeu de données. Afin de limiter le taux de faux-positifs, ils ont choisi de ne retenir comme candidates que les régions présentant des *outliers* communs aux trois populations analysées. Des résultats récents indiquent cependant que les populations de diable de Tasmanie présentent une structuration qui distingue notamment les populations de l'est de celles de l'ouest en deux principaux clusters génétiques (Hendricks *et al.,* 2017 ; Storfer *et al.,* 2017). En outre, les observations effectuées sur le terrain suggèrent que la DFTD n'a pas eu le même impact sur toutes les populations, notamment en ce qui concerne l'évolution des traits d'histoire de vie (Hamede *et al.,* 2012). Il faut aussi prendre en compte la variation de la densité de SNP d'une population à l'autre, ce qui limite la possibilité de détecter des signatures de sélection communes aux trois populations. Nous avons donc considéré que certaines signatures de sélection pouvaient être population-spécifiques, soit du fait d'un polymorphisme lui-même population-spécifique, soit du fait de la détection d'un SNP absent du génotypage d'une autre population.

Nous avons utilisé notre méthode de vraisemblance qui, comme nous l'avons étudié par simulation au cours du Chapitre III, permet de séparer les effets de la combinaison de la sélection et de la dérive de ceux de la dérive seule, y compris lorsque celle-ci est forte ($N_e \approx 30$) et lorsque le nombre de générations de sélection est limité ($n_g \approx 5$), à la condition que la sélection appliquée soit très forte. Cette situation correspond à celle du diable de Tasmanie, qui doit faire face à une très forte sélection imposée par l'émergence de la DFTD. Nous nous attendons donc à ce que cette sélection contemporaine très intense génère des signatures de sélection détectables au moyen de notre méthode, sous réserve que la variation génétique disponible soit suffisante pour permettre une réponse évolutive au cancer. Nous avons retenu comme candidates les régions génomiques affichant au moins un SNP dont la *p-value* est inférieure à $1.10^{-4}$ ou bien au moins deux SNP voisins dont les *p-value*s sont inférieures à $1.10^{-2}$, ce qui correspond à un FDR d'environ 13% (Benjamini & Hochberg,

1995). Le tableau IV-1 résume les principales informations quantitatives au sujet des séries génomiques temporelles analysées.

## 2.3. Annotation fonctionnelle des gènes candidats

Afin de faciliter le travail d'annotation individuelle des gènes et d'identifier des groupes fonctionnels significativement associés à notre liste de candidats, nous avons utilisé l'application *Ingenuity Pathway Analysis®* (IPA, https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis, *build version* 448560M, *content version* 36601845, distribuée à partir du 22/06/2017, QIAGEN Inc.). Celle-ci permet notamment de confronter des listes de gènes à une base de données actualisée de façon hebdomadaire à partir des résultats publiés dans près de 4000 journaux. Nous avons soumis la liste recensant les orthologues humains de nos gènes candidats à une « *Core Analysis* » en choisissant la base de données « *Ingenuity Knowledge Base* ». La significativité des résultats est évaluée par un test exact de Fisher.

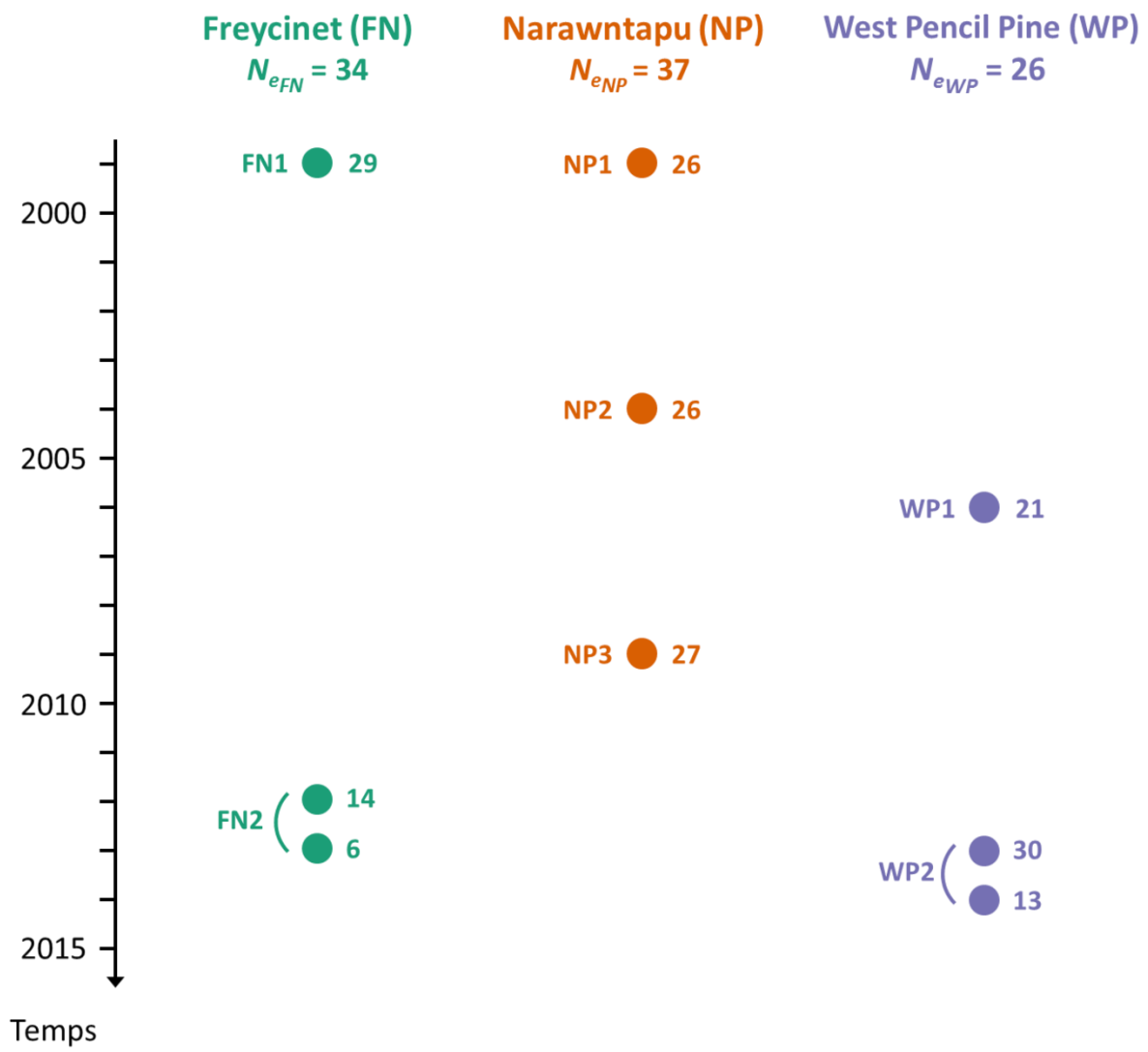



**Figure IV-4. Représentation graphique des séries génomiques temporelles étudiées afin d'identifier des traces de sélection au sein du génome du diable de Tasmanie.** L'axe vertical représente le temps en année, de haut en bas. La ligne supérieure indique les noms des trois populations examinées, avec entre parenthèses leur symbole et à la ligne suivante leur effectif efficace ($N_e$). Chaque disque coloré représente une date d'échantillonnage. Le nombre situé à droite des disques correspond au nombre d'individus échantillonnés. Au total, sept échantillons temporels sont considérés dans notre analyse : deux au sein des populations FN (FN1 et FN2) et WP (WP1 et WP2), et trois au sein de la population NP (NP1, NP2 et NP3). Les échantillons FN2 et WP2 sont chacun constitués de prélèvements effectués lors de deux années consécutives.

**Tableau IV-1.** Séries génomiques temporelles analysées à la recherche de signatures de sélection chez le diable de Tasmanie

| Symbole de population | Site d'échantillonnage | Arrivée de la DFTD | $N_e$ | $N_1$ ($t_1$)[a] | $N_2$ ($t_2$)[a] | $N_3$ ($t_3$)[a] | Nb. de SNP analysés | Nb. de SNP candidats | Prop. de SNP candidats (%) |
|---|---|---|---|---|---|---|---|---|---|
| FN | Freycinet | 2001 | 34 | 29 (1999) | - | 20 (2012-2013) | 16978 | 210 | 1,2 |
| NP | Narawntapu | 2007 | 37 | 26 (1999) | 26 (2004) | 27 (2009) | 27173 | 104 | 0,4 |
| WP | West Pencil Pine | 2006 | 26 | 21 (2006) | - | 43 (2013-2014) | 5401 | 88 | 1,6 |

$N_e$, taille efficace de population

[a] Les séries comprennent soit deux (cas des populations FN et WP) soit trois (cas de la population NP) échantillons temporels. $N_i$ désigne la taille de chaque échantillon temporel *i* numéroté dans l'ordre chronologique. L'année de prélèvement de l'échantillon, $t_i$, est précisée entre parenthèses.

*N.B.* Ce tableau correspond au tableau 1 de l'Article II. Les données ont été rendues publiques par Epstein *et al.* (2016b). Une information plus complète sur le jeu de données peut être trouvée dans Epstein *et al.* (2016a).

# — 3. Résultats

## 3.1. Le génome du diable de Tasmanie héberge une centaine de signatures de sélection

Nous avons utilisé notre méthode de détection de signatures de sélection afin de localiser dans le génome du diable de Tasmanie des empreintes laissées par la sélection exercée par la DFTD, à partir de données temporelles disponibles sur *Dryad* (Epstein *et al.,* 2016b). Notre analyse a permis l'identification de 97 signatures de sélection réparties sur l'ensemble des chromosomes, ce qui signifie qu'environ 0,3% du génome du diable de Tasmanie serait sous sélection du fait de la DFTD.

Une majorité (environ 2/3) de signatures de sélection semble population-spécifique (Fig. IV-5a), tandis qu'une trentaine est commune à deux populations. Une signature de sélection, située sur le chromosome 2 (sur la *scaffold* GL841593), a pu être identifiée dans chacune des trois populations examinées (Fig. IV-6a). Les populations WP, FN et NP présentent 54, 38 et 37 signatures de sélection, respectivement (Fig. IV-5b). Les conditions expérimentales propres à chaque population, et tout particulièrement le nombre de SNP disponibles et les dates d'échantillonnage (Tableau IV-1), affectent notre capacité à détecter la sélection. Le premier échantillon génétique de la population WP a été recueilli au moment de l'apparition de la DFTD, ce qui tend à maximiser la puissance de détection, d'où une plus forte proportion de SNP candidats que dans les autres populations (1,6%). Cet effet est dans une certaine mesure contrebalancé par une faible densité en SNP, ce qui se traduit au final par un plus faible nombre de SNP candidats à la sélection (88). La plus faible proportion de SNP candidats est trouvée dans la population NP (0,4%), malgré une densité en SNP plus importante que dans les autres populations. Ceci est dû à un dispositif expérimental non optimal pour identifier les traces de sélection dans le génome, avec un dernier échantillon temporel recueilli peu après l'arrivée supposée de la pathologie dans la population, ce qui a laissé peu de temps pour produire un effet visible. La population FN présente quant à elle une configuration plus équilibrée, avec une densité en SNP intermédiaire et des échantillons temporels bien positionnés par rapport à l'arrivée de la DFTD. Ceci a permis d'identifier le plus grand nombre de SNP candidats à la sélection (210) et les signatures de sélection les plus étendues. Par conséquent, la population FN arbore à l'issue de notre analyse la plus grande proportion (68%) de signatures de sélection situées à moins de 100 kb d'un gène codant pour une protéine (Fig. IV-5b). Au total, notre analyse a mis au jour soixante signatures de sélection à proximité de régions codantes, ce qui nous a permis de proposer une liste de 148 gènes candidats codant une

protéine (Tableau IV-3) à l'aide d'*Ensembl* v90 et du génome de référence du diable de Tasmanie (Murchison *et al.,* 2012).

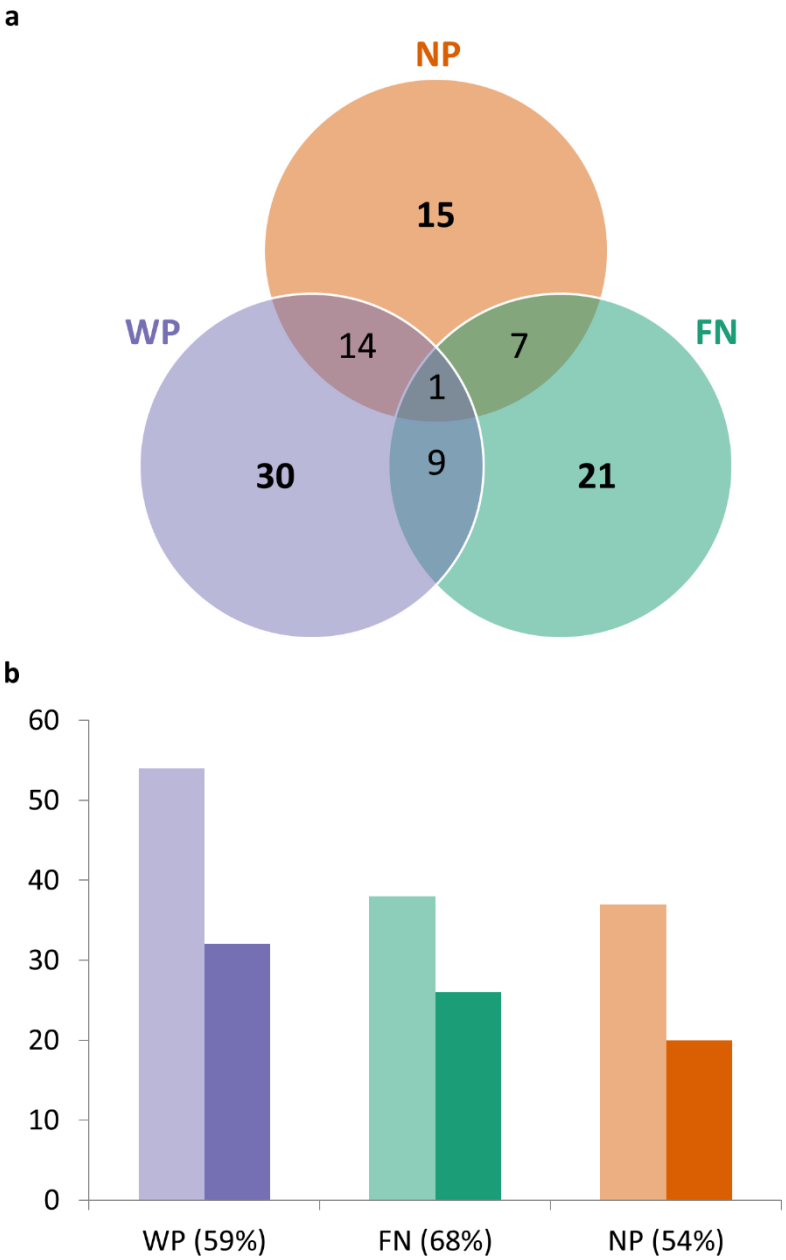



**Figure IV-5. Quatre-vingt-dix-sept signatures de sélection ont été identifiées dans le génome du diable de Tasmanie à l'aide de notre méthode de vraisemblance.** Leur répartition parmi les trois populations examinées est représentée sous la forme **(a)** d'un diagramme de Venn et **(b)** d'un diagramme en bâtons. Les bâtons clairs indiquent le nombre total de signatures de sélection identifiées au sein de chaque population. Les bâtons sombres (et les proportions entre parenthèses) indiquent la part de celles qui se trouvent à moins de 100 kb d'un gène codant une protéine et possédant un orthologue chez l'Homme d'après *Ensembl* v90. FN = Freycinet ; NP = Narawntapu ; WP = West Pencil Pine. *N.B.* Cette figure correspond à la figure 1 de l'Article II (*cf.* Annexe I).

## 3.2. Trois exemples de signatures de sélection identifiées au sein de régions codantes

Au total, soixante signatures de sélection hébergent des orthologues de gènes humains. Leur représentation graphique est disponible en annexe (Annexe II). Nous allons nous intéresser d'un peu plus près à trois d'entre elles (Fig. IV-6).

Parmi les régions identifiées à l'issue de notre recherche de signatures de sélection, seule celle qui se trouve à l'extrémité de la *scaffold* GL841593, sur le chromosome 2, est commune aux trois populations examinées (Fig. IV-6a). Cette région de 500 kb présente l'avantage d'avoir été bien génotypée dans l'ensemble des échantillons disponibles. Elle fait partie des deux régions candidates identifiées dans les travaux d'Epstein *et al.* (2016a). Deux gènes ayant chacun un orthologue humain, TRNT1 et CRBN, se trouvent à moins de 100 kb en amont des premiers SNP candidats de la signature de sélection. Le gène CRBN possède un lien avec le cancer, son expression étant utilisée comme marqueur prédictif du succès de certaines thérapies anticancéreuses (Schuster *et al.,* 2014). TRNT1 et CRBN sont aussi connus pour leur implication dans la déficience intellectuelle chez certains patients (Papuc *et al.,* 2015).

Les signatures de sélection les plus larges et les plus denses en SNP ont été identifiées au sein de la population FN. La signature de sélection présente sur le chromosome 1 dans la deuxième moitié de la *scaffold* GL834709 (Fig. IV-6b) en est une bonne illustration : le signal est constitué de six SNP tous issus de la population FN et régulièrement espacés sur 250 kb. Il s'agit d'une région orthologue à une portion de la région 8p11-12 chez l'Homme, laquelle consiste en un locus d'1 Mb amplifié dans de nombreux cancers (Garcia *et al.,* 2005). Un sous-ensemble de cette grande région incluant notamment des orthologues des gènes candidats identifiés dans notre analyse (LSM1, BAG4, DDHD2, PLPP5, NSD3, LETM2 et FGFR1) a aussi récemment été reconnu pour son implication dans certains troubles neuro-développementaux comme l'autisme (Autism Spectrum Disorders Working Group, 2017).

La plupart des signaux de sélection identifiés au sein des populations NP et WP ne sont pas aussi étendus du fait de conditions expérimentales dans l'ensemble moins favorables qu'avec la population FN. Par exemple, une signature de sélection intéressante a été détectée au sein de la population WP dans un gène orthologue à SRGAP1, sur le chromosome 5, et plus précisément sur la *scaffold* GL861740 (Fig. IV-6c). Ce signal est constitué de deux SNP dont l'un est associé à une très faible *p-value* ($p < 1.10^{-7}$). Le signal possède donc une amplitude *a priori* importante, mais la faible densité de SNP disponibles dans cette région nous prive d'une information plus précise quant à la localisation et à l'étendue de la signature de sélection. Le gène SRGAP1 encode une protéine activatrice de GTPase (GAP, *GTPase activating protein*) impliquée dans la motilité cellulaire (Ridley, 2015). Cette protéine possède en particulier un rôle inhibiteur de la migration des cellules nerveuses et de certaines cellules cancéreuses (Feng *et al.,* 2016).

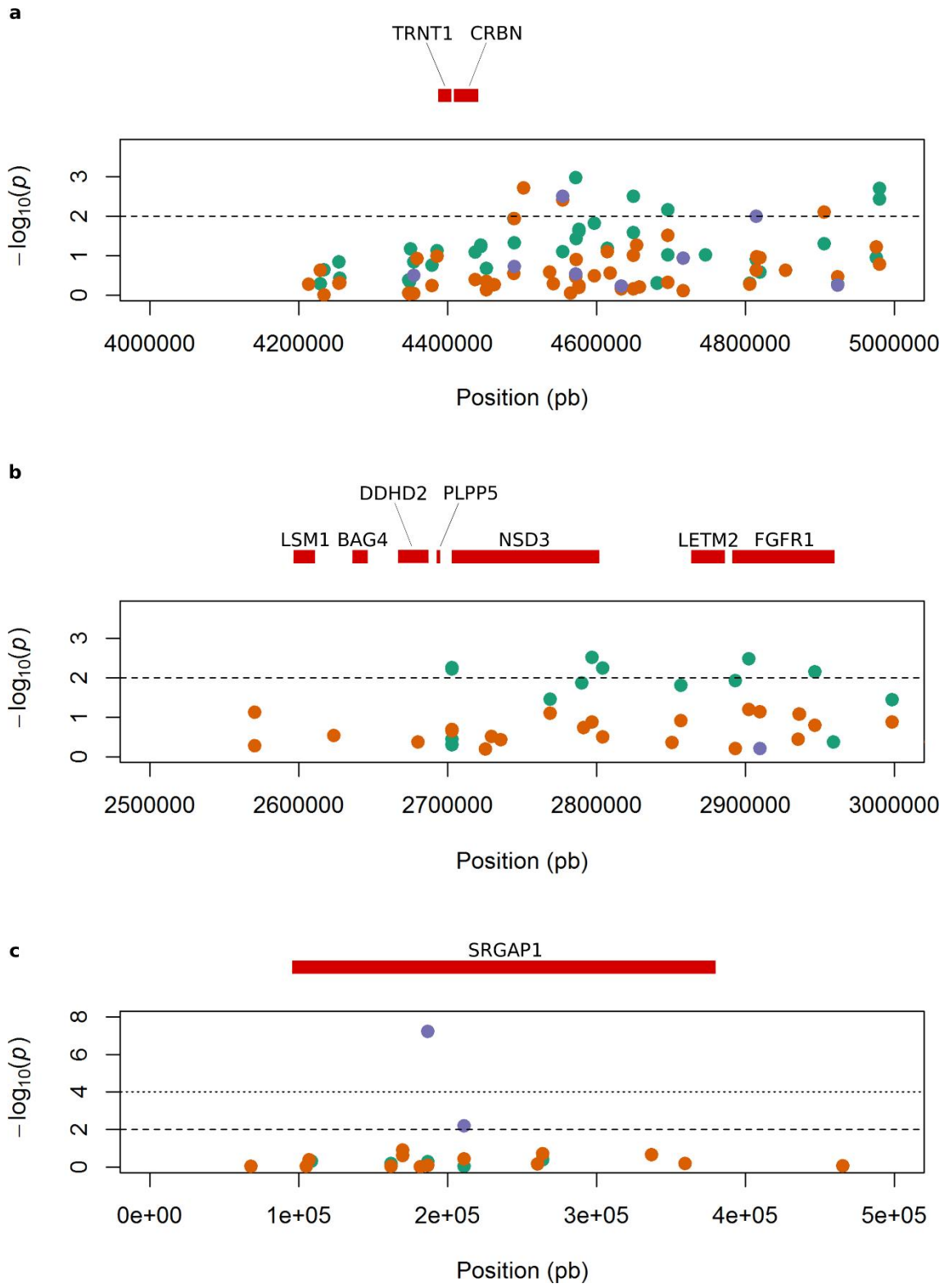



**Figure IV-6**. **Trois exemples de signatures de sélection identifiées chez le diable de Tasmanie.** Chaque graphique indique en ordonnée –log10(*p*) où *p* représente la *p-value* du LRT effectué avec notre méthode de détection de signatures de sélection (*cf.* Chapitre III). Les régions sont considérées comme candidates à la sélection si elles hébergent au moins un SNP avec $p < 1.10^{-4}$ (pointillés) ou bien au moins deux SNP voisins avec $p < 1.10^{-2}$ (tirets). Les gènes codant une protéine situés à moins de 100 kb des régions candidates et possédant un orthologue humain ont été identifiés à l'aide d'*Ensembl* v90. Ces gènes candidats sont représentés en rouge dans la marge supérieure de chaque graphique. L'axe des abscisses indique la position, en paire de bases (pb), par rapport au début des *scaffolds* GL841593, GL834709 et GL861740 pour **(a)**, **(b)** et **(c)**, respectivement.

Ces résultats illustrent des tendances générales quant au profil des signatures de sélection que nous avons identifiées. Nous allons voir que beaucoup de gènes candidats à la sélection sont impliqués dans la progression du cancer, et qu'un sous-ensemble d'entre eux entretient aussi un lien avec le fonctionnement du Système Nerveux Central (SNC).

## 3.3. La quasi-totalité des gènes candidats possède un lien avec le risque de cancer

Pour annoter la fonction des gènes candidats à la sélection chez le diable de Tasmanie, nous avons soumis la liste de leurs orthologues humains à l'application web Ingenuity Pathway Analysis® (*IPA*, https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis, QIAGEN Inc.). Parmi l'ensemble des fonctions et des pathologies examinées, le terme affichant la plus forte association avec notre liste de gènes est « tumeur solide » (*p-value* = 1,16 x $10^{-10}$), lequel a pu être relié à 138 gènes candidats sur les 147 soumis (Tableau IV-2). Les neuf gènes restant n'ont pu être reliés à aucune fonction, ce qui signifie que la totalité des gènes candidats ayant au moins une fonction renseignée dans la base de données IPA est associée au cancer. Ainsi, l'ensemble des soixante signatures de sélection proches d'un gène orthologue humain possède un lien supposé avec le cancer (Tableau IV-3). De plus, les trente premiers termes IPA associés à notre liste de gènes sont issus de la catégorie générale « Cancer » (Tableau IV-2). Les régions génomiques ciblées par la sélection chez le diable de Tasmanie semblent donc entretenir un lien fort avec les processus conférant un phénotype cancéreux aux cellules.

## 3.4. Des gènes candidats contrôlant la multiplication et la survie cellulaires sont liés au cancer

Un examen plus approfondi de leurs fonctions supposées au sein de la cellule montre que l'on retrouve les produits des gènes candidats dans l'ensemble des voies de signalisation classiquement altérées dans les cancers. Parmi les gènes les plus représentés au sein de notre liste de candidats, on trouve ceux qui participent au contrôle de la progression du cycle cellulaire, lesquels sont très fréquemment concernés par des dérégulations oncogéniques (Malumbres & Barbacid, 2009). En particulier, trois candidats (DTWD1, NEK6 et NSD3) sont connus pour leur implication dans le contrôle de la progression vers la mitose (transition G2/M). DTWD1 et NEK6 ont des rôles antagonistes en tant que régulateurs transcriptionnels de la cycline B qui favorise l'entrée en mitose (Zhang, B., *et al.,* 2014 ; Ma *et al.,* 2015). NSD3 amplifie l'expression de NEK7 et de la cycline G1, deux inducteurs de la transition G2/M, dans certains cancers (Vougiouklakis *et al.,* 2015). Ceci met en évidence la possible influence de plusieurs gènes candidats sur un point de contrôle critique du cycle cellulaire souvent dérégulé dans les cancers.

Une autre voie critique dans la régulation du destin cellulaire est l'apoptose, qui comprend les signaux de mort cellulaire programmée les mieux caractérisés (Okada & Mak, 2004). Les gènes associés à l'apoptose sont fréquemment altérés dans les tissus tumoraux. Autrefois nommé *Silencer Of Death Domain* (SODD), le candidat BAG4 est bien connu pour son rôle oncogénique empêchant l'apoptose extrinsèque. En se liant à leur domaine de mort intracellulaire, BAG4 maintient inactifs les récepteurs TNF-R1 (*Tumor Necrosis Factor Receptor* 1) qui sont chargés de transmettre les messages de mort cellulaire programmée à l'intérieur de la cellule (Jiang *et al.,* 1999). Par ailleurs, le candidat TRIM66 peut favoriser la progression du cancer, en inhibant cette fois la voie de l'apoptose intrinsèque. TRIM66 est capable d'inhiber P53, un célèbre suppresseur de tumeur qui a entre autres un rôle pro-apoptotique (Chen, Y., *et al.,* 2015). La liste de candidats héberge ainsi des gènes ayant un rôle dans les deux principales voies, apotposes extrinsèque et intrinsèque, de mort cellulaire programmée.

## 3.5. Des gènes candidats sont détournés de leur rôle développemental initial dans les cancers

Parmi les candidats identifiés, FGFR1 est un récepteur à activité tyrosine kinase (RTK) appartenant à la famille des récepteurs aux facteurs de croissance fibroblastiques (FGFR). A l'image de l'ensemble des RTK, les récepteurs FGFR1 contribuent normalement au développement, au maintien de l'homéostasie ou à la réponse inflammatoire, mais on les retrouve aussi fréquemment surexprimés ou mutés dans les cancers (Dieci *et al.,* 2013). Des altérations de FGFR1 ont été décrites dans de nombreuses tumeurs, notamment lorsque leur domaine kinase arbore une mutation activatrice (Jones, D.T., *et al.,* 2013 ; Cowell *et al.,* 2017). D'autres candidats sont connus pour être impliqués dans des voies de signalisation activées par des RTK et classiquement associées à la tumorigenèse. Le candidat SHC4 amplifie la voie des *MAP*-kinases activée par certains RTK, ce qui favorise la prolifération cellulaire et la métastase (Strub *et al.,* 2014). Le candidat ST5 (pour *Suppression of Tumorigenicity* 5) est lui aussi capable d'influencer les voies de signalisation oncogéniques activées par les RTK favorisant prolifération, croissance cellulaire et métastase (Ioannou & McPherson, 2016). Par ailleurs, le candidat CEP131 peut favoriser la prolifération et la migration cellulaires à travers l'activation de la voie de la phosphatidylinositol 3-kinase (PI3k/Akt), qui est une voie RTK-dépendante souvent affectée dans les cancers humains (Pal & Mandal, 2012).

De nombreux autres candidats sont impliqués dans des voies de signalisation contribuant au développement mais pouvant aussi être responsables de la progression du cancer. La voie du *Transforming Growth Factor* β (*TGF*β), qui met en jeu les facteurs de transcription de la famille SMAD, est caractérisée pour son rôle dans l'embryogenèse, mais aussi dans le cancer. Elle est capable d'un double-jeu vis-à-vis du cancer, avec une aptitude première consistant à inhiber l'oncogenèse précoce,

mais aussi une propension à favoriser la métastase (Syed, 2016). Le candidat SMAD3 est un messager-clé de la voie du *TGF*β. SMAD3 est capable d'interagir avec de nombreux partenaires, dont un grand nombre de coactivateurs et corépresseurs transcriptionnels, de sorte que son rôle est très largement dépendant du contexte cellulaire. Ses multiples contributions tumeur-suppressives et oncogéniques sont abondamment documentées dans la littérature (Tufegdzic Vidakovic *et al.,* 2015 ; Majumder *et al.,* 2016 ; Yang *et al.,* 2016 ; Thomas, A.L*., et al.,* 2017). En particulier, le candidat SPTBN1 est susceptible de coopérer avec SMAD3 afin de réprimer la transcription de STAT3 (*Signal Transducers and Activators of Transcription* 3), ce dernier étant un facteur de transcription célèbre pour son action en faveur de la prolifération (Lin, L., *et al.,* 2014). Le candidat NEK6 intervient dans l'oncogenèse, en empêchant la translocation de SMAD4 induite par l'activation de la voie du *TGF*β (Zuo *et al.,* 2015). D'une manière générale, la voie du *TGF*β implique une cascade de signalisation relativement simple mais connectée à de nombreuses autres voies (Ikushima & Miyazono, 2010 ; Dahl *et al.,* 2014 ; Shi & Chen, 2017). Certains candidats, comme PRRX2 ou TRIM66, sont impliqués dans des interactions avec la voie du *TGF*β susceptibles d'induire la métastase (Chen, Y., *et al.,* 2015 ; Juang *et al.,* 2016).

La voie *Wnt*/β-caténine est un autre bon exemple de ces voies de signalisation liées au développement mais aussi aux pathologies comme le cancer en raison de leur capacité à induire la prolifération cellulaire (Nusse & Clevers, 2017). Le candidat FOXN3 est un régulateur important de cette voie de transduction en empêchant la formation du complexe β-caténine/TCF4 responsable de la transcription des gènes de prolifération (Dai *et al.,* 2017). Le candidat SPTBN1 possède lui aussi un rôle tumeur-suppresseur via le contrôle de l'expression d'un inhibiteur de la voie *Wnt* (Zhi *et al.,* 2015). Il y a aussi des régulateurs positifs de la voie *Wnt* parmi les candidats, comme NSD3 (French *et al.,* 2014). Les gènes encodant des kinases cycline-dépendantes (CDK) sont habituellement décrits comme étant des régulateurs critiques du cycle cellulaire, mais certains d'entre eux assurent des fonctions non-canoniques (Hydbring *et al.,* 2016). C'est le cas du candidat CDK14 qui a un rôle régulateur important dans l'activation de la voie *Wnt* à travers la phosphorylation du corécepteur LRP6 (Davidson & Niehrs, 2010).

D'autres voies associées à l'embryogenèse sont ciblées par la sélection, comme l'indique par exemple l'identification du récepteur NOTCH2 parmi nos candidats. NOTCH2 est une protéine transmembranaire impliquée dans la signalisation issue des contacts cellule-cellule propres à la voie *Notch*, laquelle intervient notamment dans la différenciation des cellules souches (Nowell & Radtke, 2017). Des mutations de NOTCH2 ont été décrites dans plusieurs cancers humains (Nowell & Radtke, 2017). La voie *Hippo*, fortement conservée au cours de l'évolution (Zhao *et al.,* 2011), et dont l'importance dans le contrôle du développement des organes est bien illustrée (Boone *et al.,* 2016 ; Wang & Martin, 2017), peut aussi réguler plusieurs étapes-clés de la progression tumorale (Ehmer &

Sage, 2016 ; van Rensburg & Yang, 2016). La voie *Hippo* implique une série de sérine/thréonine-kinases qui contrôlent la transcription de gènes de prolifération. Le candidat NSUN6 peut favoriser la prolifération cellulaire et la métastase au travers de l'inactivation de MST1, une des sérine/thréonine-kinases de la voie *Hippo* (Li, C., *et al.,* 2017). Par ailleurs, le candidat TEAD4 appartient aux facteurs de transcription de la famille TEAD (*Transcriptional Enhancer Factor Domain*), qui sont nécessaires à l'activation des gènes de prolifération régulés par la voie *Hippo* (Harvey *et al.,* 2013).

Ainsi, la liste des gènes candidats comprend de nombreux régulateurs de voies de signalisation contribuant au développement – voies des RTK, du *TGF*β, *Wnt*, *Notch*, *Hippo*, *etc.* – pourvus de rôles importants dans différentes étapes de progression du cancer, comme le contrôle de la prolifération, de la croissance ou de la motilité cellulaires.

## 3.6. Des gènes candidats sont impliqués dans des processus oncogéniques importants mais encore mal compris

Au-delà des candidats impliqués dans les voies classiquement associées au développement et à la progression des cancers, nous avons pu identifier certains candidats participant à d'autres processus, aujourd'hui incomplètement caractérisés, mais dont l'impact semble déterminant dans l'acquisition de phénotypes cellulaires cancéreux. Par exemple, le candidat CNNM3 est un transporteur d'ions magnésium dont on a rapporté l'interaction avec des protéines tyrosine-phosphatases de la famille PRL (*Phosphatase Regenerating Liver*) (Labbé *et al.,* 2012). L'association de CNNM3 et PRL2 forme un complexe qui favorise le développement tumoral au travers de la modulation des niveaux de magnésium intracellulaire (Hardy *et al.,* 2015). L'incapacité de la cellule à réguler les flux de magnésium est connue pour être étroitement liée au cancer, mais certains aspects des mécanismes sous-jacents restent à éclaircir (Wolf & Trapani, 2012).

Les réactions de méthylation constituent un autre mécanisme important susceptible d'influencer la signalisation au sein des cellules tumorales (Biggar & Li, 2014 ; Klutstein *et al.,* 2016). Plusieurs candidats possèdent une activité méthyltransférase, comme NSD3, NSUN6 ou NTMT1. NSD3 fait partie d'un complexe qui engendre la prolifération cellulaire au travers d'une altération de la structure de la chromatine (French *et al.,* 2014). NSUN6 contribue à la méthylation de la sérine/thréonine-kinase MST1 contrôlant la métastase (Li, C., *et al.,* 2017). NTMT1 est une protéine-méthyltransférase possédant de multiples rôles dans la régulation des tumeurs (Bonsignore *et al.,* 2015).

Plus généralement, l'instabilité génomique est perçue comme une cause majeure du cancer (Hanahan & Weinberg, 2011 ; Burell *et al.,* 2013 ; Robert 2017). Les candidats CEP131 et PINX1 semblent avoir un rôle important dans le maintien de l'intégrité du génome. L'absence de la protéine centrosomale

CEP131 contribue à l'instabilité génomique (Staples *et al.,* 2012). PINX1 encode un inhibiteur de télomérase qui est inactivé dans certaines cellules cancéreuses (Li, H.L., *et al.,* 2016). La réactivation de la télomérase dans les cellules somatiques est la cause d'une large majorité de cancers. En permettant notamment aux cellules sujettes à l'instabilité génomique de maintenir leur potentiel réplicatif, la restauration de l'activité de la télomérase dans les cellules somatiques est à l'origine d'une large majorité de cancers (Maciejowski & de Lange, 2017). De plus, des résultats récents indiquent que la télomérase est également susceptible de réguler directement des voies de signalisation impliquées dans le développement tumoral (Terali & Yilmazer, 2016). Un inhibiteur de télomérase comme PINX1 a par conséquent un rôle central dans la carcinogenèse.

## 3.7. De nombreux gènes candidats sont associés à la métastase

De façon intéressante, la liste des gènes candidats comprend de nombreux médiateurs de la dissémination métastatique. Tout d'abord, plusieurs candidats encodent des protéines impliquées dans l'adhésion cellulaire, une fonction altérée lors d'étapes-clés de la métastase (Li, D.M., & Feng, 2011 ; Robert, 2013). Les candidats MCAM et THY1 produisent des molécules d'adhésion cellulaire de la superfamille des immunoglobulines (IgSF-CAM), que l'on retrouve associées à la progression du cancer (Rege & Hagood, 2006 ; Ouhtit *et al.,* 2009). Tous deux sont fréquemment surexprimés dans les tissus tumoraux métastatiques (Wu *et al.,* 2012 ; Zhang, D.H., *et al.,* 2016). Le candidat CDH8 appartient à la grande famille des cadhérines (CDH), qui ont un rôle majeur dans l'adhésion cellulaire et le cancer en cas de dérégulation (Kourtidis *et al.,* 2017). Une expression anormale de CDH8 a été rapportée dans les tissus tumoraux (Blaschke *et al.,* 2002). De plus, plusieurs travaux suggèrent que CDH8 est fortement associé au développement métastatique (Lu *et al.,* 2006 ; Lee *et al.,* 2013). Deux candidats, TSPAN9 et TSPAN11, appartiennent à la famille des tétraspanines (TSPAN), des molécules d'adhésion que l'on suspecte de jouer de nombreux rôles au niveau des assemblages protéiques de la surface cellulaire (Hemler, 2005). Les TSPAN sont encore trop peu étudiées pour avoir une vision de l'étendue de leurs rôles, mais il est avéré qu'elles sont capables d'influencer les processus liés à la métastase (Detchokul *et al.,* 2014 ; Hemler, 2014). Les récepteurs d'adhésion couplés aux protéines G (aGPCR) forment une autre famille de molécules d'adhésion cellulaire encore incomplètement caractérisée, du fait de sa diversité et de sa complexité, mais dont nous savons que les polymorphismes de certains membres sont associés au cancer et à la métastase (Langenhan *et al.,* 2013 ; Yona *et al.,* 2016). Deux représentants des aGPCR, ADGRA2 et ADGRD2, se trouvent parmi les candidats. Les fonctions de ADGRD2 ne sont pas encore bien caractérisées, mais l'implication de ADGRA2 dans l'angiogenèse tumorale, qui est un processus crucial pour la croissance et la métastase des tumeurs solides, a été établie (Vallon & Essler, 2006).

D'autres candidats sont capables de conférer des propriétés métastatiques aux cellules, comme PRRX2 et CST3, qui favorisent la migration et l'invasion ces cellules tumorales via la voie du *TGF*β (Juang *et al.,* 2016 ; Yan *et al.,* 2017). La surexpression du candidat MMP28, qui appartient à la famille des métalloprotéinases (MMP), induit la métastase (Jian *et al.,* 2011). Le candidat REG4 favorise l'invasion au travers de la régulation positive de deux MMP (He *et al.,* 2012). Nous pourrions aussi citer les candidats FOXN3 (Dai *et al.,* 2017), GLRX3 (He *et al.,* 2016), HMGCS2 (Chen, S.W., *et al.,* 2017), LSM1 (Little *et al.,* 2016), PHGDH (Samanta *et al.,* 2016), RIN2 (Sandri *et al.,* 2012), SHC4 (Strub *et al.,* 2011), TRPM8 (Yee 2015), NSUN6 (Li, C., *et al.,* 2017), USP2 (Qu *et al.,* 2015), et d'autres encore, pour leur rôle dans les processus métastatiques. Au total, plus d'une trentaine de candidats participent à des régulations connues de la métastase, sans compter les marqueurs de la métastase dont les rôles ne sont pas encore bien caractérisés, comme ITIH5 (Rose *et al.,* 2013), LAD1 (van Vlodrop *et al.,* 2016) ou THBS4 (Lin, X., *et al.,* 2016). Tout cela suggère que le contrôle de la métastase pourrait être une composante-clé de la réponse évolutive à la DFTD.

## 3.8. De nombreux gènes candidats ont aussi un rôle dans le développement et le fonctionnement du Système Nerveux Central

En plus d'une très forte association avec le cancer, notre liste de gènes candidats à la sélection chez le diable de Tasmanie montre un lien avec la neurogenèse, la neurotransmission et les troubles développementaux du SNC. L'analyse IPA a en particulier permis l'identification d'un groupe de 16 candidats associés à la fonction « Développement des neurones » (*p-value* ≈ 0,003 – *cf.* Fig. IV-7). Plus largement, plus d'une vingtaine de candidats sont capables de contribuer au développement et au maintien de l'homéostasie de plusieurs compartiments cellulaires importants dans le SNC. Par exemple, SYNDIG1 régule la synaptogenèse dans l'hippocampe (Lovero *et al.,* 2013). NKX2-2 est un régulateur-clé du développement des neurones sérotoninergiques (Cheng, L., *et al.,* 2003). SDK2 est une molécule d'adhésion nécessaire à la formation et au maintien des connections synaptiques (Krishnaswamy *et al.,* 2015). KIF13B appartient à la superfamille des kinésines, les célèbres « moteurs moléculaires » nécessaires au fonctionnement des neurones (Hirokawa *et al.,* 2009), et possède un rôle-clé dans la régulation du développement axonal (Nakata & Hirokawa, 2007). SLITRK5 et SLC38A10 jouent des rôles importants dans la neurotransmission centrale (Shmelkov *et al.,* 2010 ; Hellsten *et al.,* 2017). CDC42EP4 et HERC1 sont impliqués dans le maintien de l'homéostasie synaptique (Ageta-Ishihara *et al.,* 2015 ; Bachiller *et al.,* 2015).

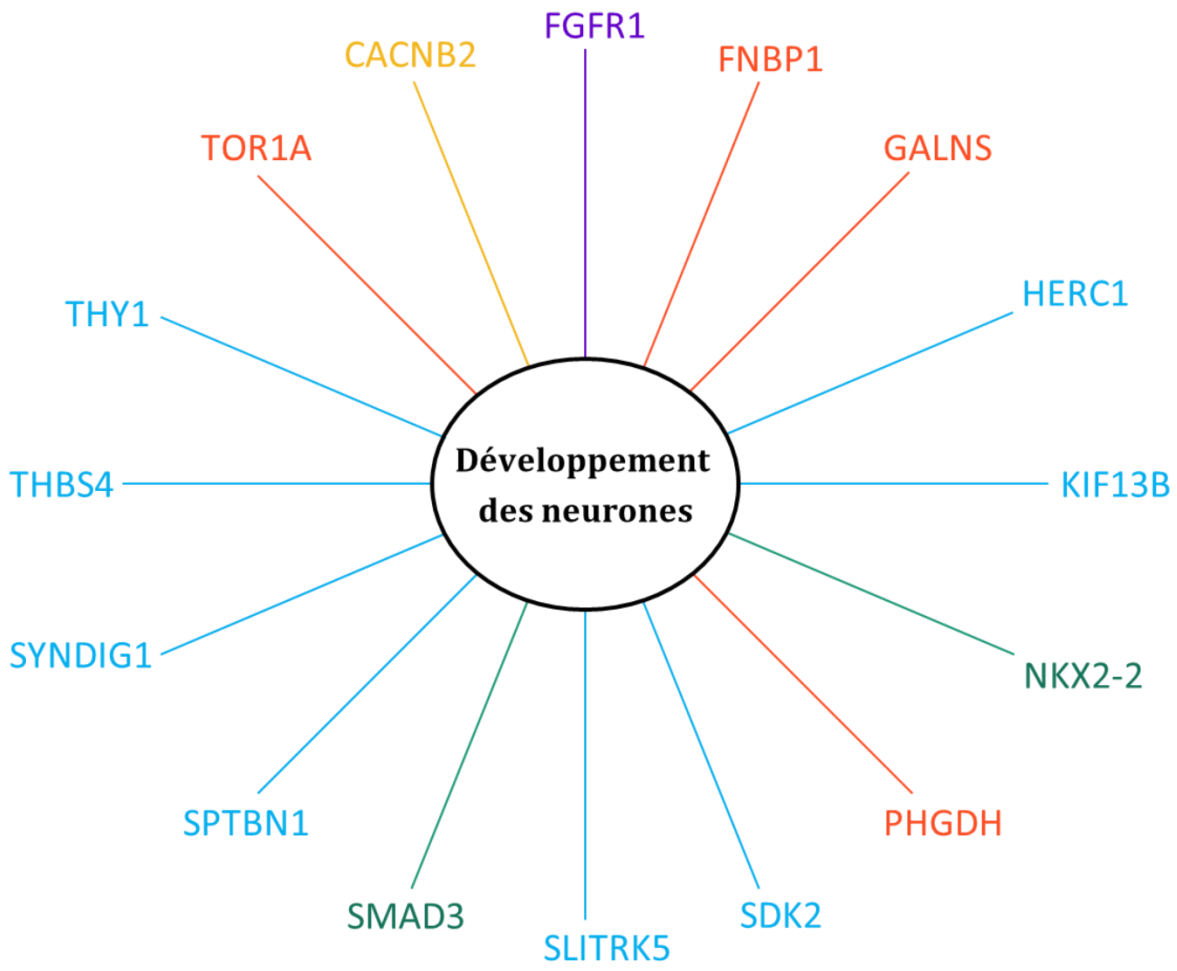



**Figure IV-7. Seize gènes candidats sont associés à la fonction « Développement des neurones » d'après *IPA*.** L'analyse IPA a permis d'identifier que la fonction « Développement des neurones » était surreprésentée dans notre liste de gènes candidats (*p-value* ≈ 0,003). Les couleurs indiquent le type de protéine encodée par chaque gène candidat. Orange = enzyme ; vert = facteur de transcription ; jaune = canal ionique ; violet = kinase ; bleu = autre. *N.B.* Cette figure correspond à la figure supplémentaire 2 de l'Article II.

En outre, plusieurs candidats correspondent à des gènes encodant des sous-unités de canaux ioniques (*e.g.,* CACNB2, KCNA4, KCNIP3 ou encore KCTD3) impliquées dans le fonctionnement du SNC chez l'Homme (voir par exemple Pruunsild & Timmusk, 2012, Cao-Ehlker *et al.,* 2013 ou Gonzalez *et al.,* 2014).

La présence de cet ensemble de candidats capables de réguler le fonctionnement du système nerveux, et en particulier la neurotransmission centrale, pourrait avoir un impact sur la prévalence des troubles comportementaux dans les populations de diable de Tasmanie. Rappelons que la large signature de sélection qui s'étend du nucléotide 2.702.597 au nucléotide 2.946.330 sur la *scaffold* GL834709 (Fig. IV-6b), et correspondant au locus 8p11.23, est fortement associée aux Troubles du Spectre Autistique (TSA) chez l'Homme (Autism Spectrum Disorders Working Group, 2017). Il est frappant de constater

que plusieurs autres locus candidats à la sélection sont associés à des troubles affectant le comportement social, et notamment les TSA. Le candidat NDUFAF5 est par exemple impliqué dans l'assemblage du complexe I de la chaîne respiratoire (Rhein *et al.,* 2016), qui est fréquemment altéré dans les troubles du SNC comme l'autisme (Hollis *et al.,* 2017). Plusieurs autres candidats, comme CACNB2 (Breitenkamp *et al.,* 2014), CDH8 (Pagnamenta *et al.,* 2011), HERC1 (Utine *et al.,* 2017), KCTD3 (Poot *et al.,* 2010), KIF13B (Li, J., *et al.,* 2016), SERINC2 (Hnoonual *et al.,* 2017) et SLC39A11 (Woodbury-Smith *et al.,* 2015) sont associés aux TSA. D'autres candidats sont connus pour leur association avec la déficience intellectuelle, comme CRBN (Kaufman *et al.,* 2010), GPKOW (Helsmoortel *et al.,* 2015), HERC1 (Utine *et al.,* 2017) et KCNA4 (Kaya *et al.,* 2016), ou encore à d'autres troubles du SNC, comme CDC42EP4 (Yan *et al.,* 2016) et SLITRK5 (Shmelkov *et al.,* 2010).

Ainsi, les gènes contribuant au développement et au fonctionnement du SNC représentent une catégorie fonctionnelle fortement associée à notre liste de candidats. Chez l'Homme, certains orthologues à ces candidats ne sont exprimés que dans le SNC et possèdent des rôles dans la neurotransmission, ce qui, couplé à l'identification de gènes associés à l'autisme et à la déficience intellectuelle, suggère une potentielle influence de la sélection sur des phénotypes complexes comme certains comportements sociaux.

## 3.9. Quelques gènes candidats pourraient être impliqués dans la surveillance immunitaire des tumeurs

Quelques candidats appartenant à la grande famille des TRIM (*tripartite motif family*) ont été identifiés (TRIM10, TRIM15, TRIM26, TRIM39-RPP21, TRIM62, et TRIM66). Il s'agit d'ubiquitines-ligases impliquées dans la réponse immunitaire innée mais aussi dans le contrôle de la progression du cancer (Hatakeyama *et al.,* 2017). A l'exception de TRIM39-RPP21, qui n'est associé à aucune fonction connue, l'ensemble de ces gènes de la famille TRIM sont connus pour être liés au cancer selon la base de données IPA (Tableau IV-2). De plus, les candidats TRIM10, TRIM15, TRIM26 et TRIM39-RPP21 appartiennent au locus HLA chez l'Homme, qui regroupe de nombreux médiateurs du système immunitaire (Shiina *et al.,* 2009). Un rôle dans l'immunité a aussi été suggéré pour les candidats TRIM62 et TRIM66 (Versteeg *et al.,* 2013 ; Cao *et al.,* 2015). Les candidats des familles TSPAN et aGPCR, que nous avions précédemment cités pour leur rôle dans la métastase, pourraient aussi être impliqués dans la réponse immunitaire (Veenbergen & van Spriel, 2011 ; Nijmeijer *et al.,* 2016). Il a récemment été montré que le candidat SMAD3, exprimé dans de nombreux tissus et impliqué dans de multiples voies de signalisation, régulait la surveillance immunitaire des tumeurs (Tang *et al.,* 2017). Ainsi, comme déjà suggéré par Epstein *et al.* (2016a), certains candidats ont le potentiel pour mettre en place une réponse immunitaire contre le cancer.

## 3.10. Des gènes candidats sont associés à des pathologies diverses

Certains candidats sont associés chez l'Homme à d'autres pathologies que le cancer, principalement des maladies ophtalmiques ou cardiaques. Les candidats C1QTNF5, CST3 et MFRP sont impliqués dans différents types de dégénérescences rétiniennes (Kameya *et al.,* 2002 ; Zurdel *et al.,* 2002 ; Stanton *et al.,* 2017). Des mutants de MFRP ont également été identifiés dans d'autres troubles ophtalmiques, dont la microphtalmie (Sundin *et al.,* 2005). D'autres candidats, tels que FRMD4B, KLF10 et TNNT2, sont eux impliqués dans le fonctionnement cardio-vasculaire et peuvent présenter des mutations exposant à un risque accru d'arrêt cardiaque (Cappola *et al.,* 2010 ; Subramaniam *et al.,* 2010 ; Wei & Jin, 2016). Le candidat KCNU1 encode un canal potassique pH- et voltage-dépendant impliqué dans la fertilité des mâles chez les Mammifères (Schreiber *et al.,* 1998). L'augmentation de la fréquence allélique de variants en ces locus pourrait représenter une réponse indirecte à la sélection.

# — 4. Discussion

Notre méthode de détection de signatures de sélection a permis d'identifier près d'une centaine d'empreintes laissées dans le génome du diable de Tasmanie sous l'effet de la forte sélection exercée par la DFTD. En recherchant la signification fonctionnelle de ces signaux de sélection, nous avons identifié près de 140 gènes candidats pouvant être reliés au risque de cancer. Une analyse approfondie des fonctions supposées de ces gènes candidats nous a permis de mettre en avant des mécanismes et des voies de signalisation susceptibles d'avoir été ciblés par la sélection en réponse à la DFTD. Nous avons ainsi pu souligner la présence de nombreux régulateurs et messagers critiques de voies classiquement associées au cancer, comme les voies apoptotiques, la voie *Wnt*/β-caténine, la voie des *MAP*-kinases, la voie du *TGF*β, *etc*. En plus des médiateurs de ces voies bien caractérisées et parmi les plus susceptibles de favoriser la prolifération et la croissance cellulaires dans les cancers, nous avons pu identifier des candidats prenant part à d'autres régulations étroitement associées au cancer (*e.g.,* altérations épigénétiques, maintien de l'intégrité du génome). Enfin, nous avons pu distinguer un groupe de candidats connus pour leurs rôles oncogéniques ou suppresseurs de tumeur affectant les fonctions contribuant à la métastase (*e.g.,* adhésion et motilité cellulaires, néo-angiogenèse). Ainsi, la sélection aurait ciblé un large ensemble de fonctions et de voies de signalisation contribuant à la progression du cancer. En nous référant aux travaux conceptuels de Hanahan & Weinberg (2011), il apparaît que nos candidats sont susceptibles de moduler l'acquisition des six *hallmarks* du cancer (Fig. IV-8), c'est-à-dire les six grands types de mécanismes contribuant à l'oncogenèse.

Le cancer est un phénotype complexe dont l'apparition et le maintien dépendent du détournement concomitant de plusieurs voies de leur fonction physiologique initiale. Nos résultats montrent ainsi que, plutôt que de cibler prioritairement une voie ou une fonction spécifique, la réponse à la sélection imposée par la DFTD repose sur des bases largement polygéniques reflétant les mécanismes conceptuels élémentaires impliqués dans l'oncogenèse. La létalité de la DFTD étant principalement due à la taille des tumeurs faciales et aux fréquentes métastases affectant les organes distants de la tumeur primitive (Pyecroft *et al.,* 2007), il n'est donc pas étonnant qu'une composante importante de la réponse à la sélection imposée par la DFTD semble résider dans le contrôle de la croissance cellulaire et de la métastase. Plus d'une trentaine de candidats sont en particulier connus pour prendre part aux mécanismes permettant l'émergence de la métastase. Ces candidats sont tout particulièrement susceptibles de moduler le microenvironnement tumoral afin de freiner la croissance et surtout la dissémination métastatique du cancer. Au bilan, la sélection a pu cibler, parmi la variation génétique disponible au sein de chaque population, une panoplie de gènes impliqués dans les circuits de prolifération, de survie, de motilité, *etc.*, qui contribueraient *in fine* à des mécanismes de résistance à la propagation de la DFTD.

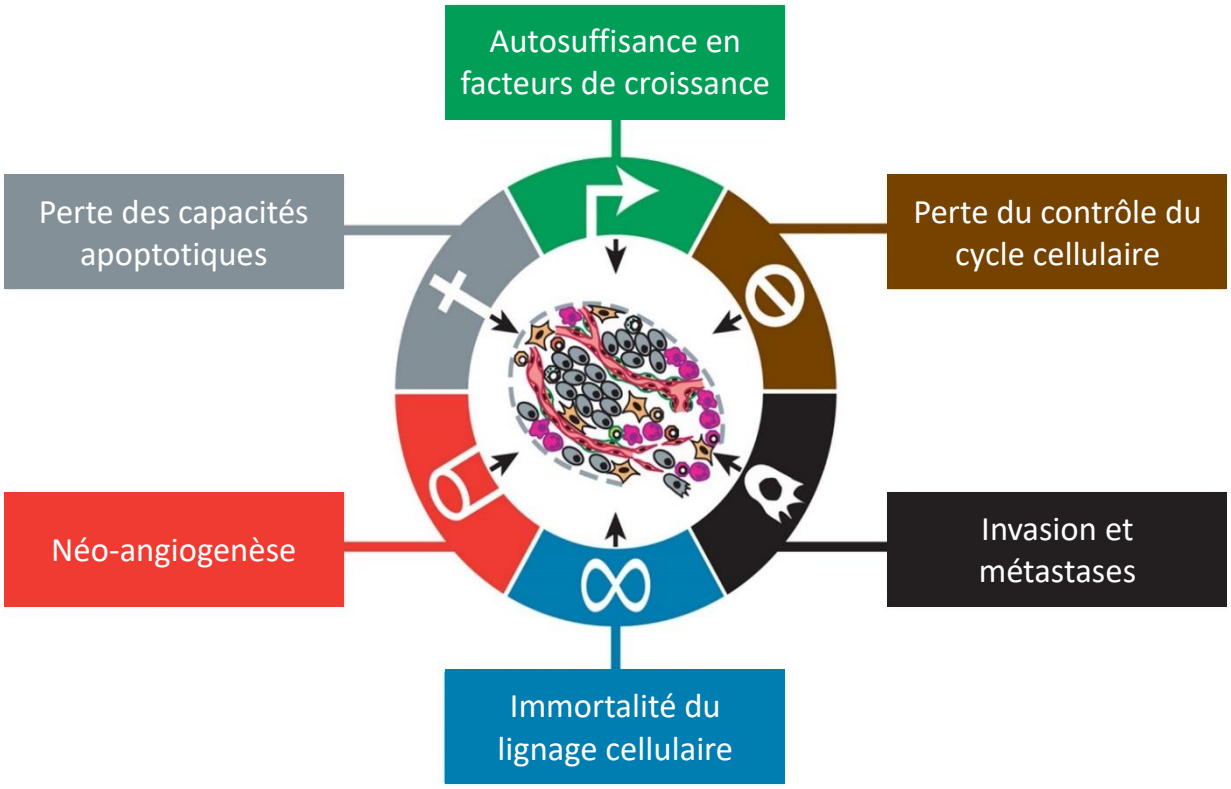



**Figure IV-8. *Hallmarks* du cancer proposées par Hanahan et Weinberg en 2000.** Cette figure, issue de Hanahan & Weinberg (2011), représente les *hallmarks* du cancer, telles que proposées par Hanahan et Weinberg dans leur publication séminale de l'an 2000 sur les mécanismes de l'oncogenèse et traduites d'après les propositions de Robert (2017). Le dessin situé au centre de la figure représente une tumeur solide, constituée d'un assemblage de plusieurs types de cellules.

Nos résultats confirment que l'étude de l'évolution récente du diable de Tasmanie, hôte d'un cancer agressif transmissible, est susceptible d'améliorer notre compréhension du cancer, et en particulier la biologie de la métastase.

Plus globalement, la question de l'importance écologique du cancer au sein des populations animales sauvages est prégnante. Le cancer est en effet de plus en plus perçu comme une composante significative de la valeur sélective des individus, au point que le terme « oncobiote », construit par analogie à « microbiote » et désignant l'ensemble des cellules cancéreuses présentes au sein d'un organisme, a été récemment introduit (Thomas, F., *et al.,* 2017). Malgré la difficulté à diagnostiquer le cancer au sein de la faune sauvage, l'importance de cette composante ne fait aucun doute d'après les données actuellement disponibles quant à sa prévalence (Madsen *et al.,* 2017) et ses conséquences sur la conservation de certaines populations (McAloose & Newton, 2009). En particulier, l'idée émerge que certains comportements puissent être ciblés par la sélection naturelle pour leur valeur adaptative face au risque de cancer (Vittecoq *et al.,* 2015).

Dans le cas de la sélection exercée par la DFTD, nos résultats suggèrent qu'une proportion non négligeable de gènes candidats est impliquée dans le développement et le fonctionnement du SNC. Certains de ces gènes sont associés chez l'Homme à l'autisme ou à la déficience intellectuelle, c'est-à-dire des pathologies caractérisées par un déficit de la communication et de l'interaction sociale (Orsmond *et al.,* 2004 ; Walton & Ingersoll, 2013 ; Taheri *et al.,* 2016). L'identification de gènes candidats contribuant à de tels phénotypes chez le diable de Tasmanie doit être rapprochée de résultats récents indiquant que la DFTD affecte davantage les individus qui possédaient une meilleure valeur sélective avant l'épizootie (Wells *et al.,* 2017). Depuis l'apparition de ce cancer, qui se transmet par morsure lors des contacts sociaux, plusieurs études ont suggéré que les individus socialement dominants – plus enclins aux contacts et plus fréquemment mordeurs – faisaient face à un risque d'infection plus grand (Hamede *et al.,* 2009, 2013 ; Wells *et al.,* 2017). Tout cela indique une sélection contre l'agressivité sociale, de sorte que les individus qui sont plus susceptibles d'éviter les conflits auraient mieux « résisté » à la DFTD, ce qui est cohérent avec l'implication des gènes du SNC dans la réponse à la sélection, tel que nos résultats le suggèrent.

Ainsi, nous pensons que plusieurs profils de variants ont pu être ciblés par la sélection chez le diable de Tasmanie : (i) d'une part, des gènes liés à la résistance au cancer et, dans une moindre mesure, à la réponse immunitaire, qui permettraient à leurs porteurs de se reproduire malgré l'infection si l'évolution des tumeurs est freinée, voire en cas de régression tumorale comme cela a pu être observé chez quelques individus (Pye *et al.,* 2016b) ; (ii) d'autre part, des gènes ayant une incidence sur le comportement (socialité/comportement reproducteur) favorisant les individus les moins susceptibles d'être infectés lors des interactions sociales (Wells *et al.,* 2017).

Au total, nous avons identifié plus d'une centaine de gènes candidats dont les rôles supposés ont un lien avec le cancer ou bien avec le risque d'exposition à la DFTD. Ces résultats suggèrent qu'il est concrètement possible d'identifier des signatures de sélection associées à un phénotype complexe à partir de séries génomiques temporelles, même si le nombre de générations de sélection (≈ 5 générations) et la taille efficace ($N_e$ ≈ 30) sont particulièrement faibles.

## — 5. Tableaux supplémentaires

**Tableau IV-2.** Les trente fonctions ou pathologies les plus fortement associées par IPA à notre liste de gènes candidats concernent le cancer

| Rang | Catégorie | Pathologie ou fonction associée | *p-value*[a] | *FDR*[b] | Nb. de gènes |
|---|---|---|---|---|---|
| 1 | Cancer | Tumeur solide | 1,16E-10 | 4,39E-07 | 138 |
| 2 | Cancer | Tumeur solide maligne | 2,37E-10 | 4,47E-07 | 137 |
| 3 | Cancer | Tumeur solide (sauf mélanome) | 3,07E-09 | 3,86E-06 | 131 |
| 4 | Cancer | Tumorigenèse des tissus | 5,54E-09 | 5,22E-06 | 130 |
| 5 | Cancer | Néoplasie du tissu épithélial | 9,10E-09 | 6,87E-06 | 129 |
| 6 | Cancer | Néoplasme abdominal | 1,66E-08 | 1,04E-05 | 128 |
| 7 | Cancer | Carcinome | 2,10E-08 | 1,05E-05 | 128 |
| 8 | Cancer | Cancer abdominal | 2,23E-08 | 1,05E-05 | 127 |
| 9 | Cancer | Cancer du sein ou colorectal | 2,09E-07 | 8,75E-05 | 84 |
| 10 | Cancer | Cancer du système digestif | 4,36E-07 | 1,64E-04 | 118 |
| 11 | Cancer | Adénocarcinome | 6,42E-07 | 2,20E-04 | 117 |
| 12 | Cancer | Tumeur du système digestif | 7,06E-07 | 2,22E-04 | 119 |
| 13 | Cancer | Carcinome gastro-intestinal | 2,48E-06 | 7,20E-04 | 102 |
| 14 | Cancer | Carcinome du gros intestin | 3,30E-06 | 8,88E-04 | 98 |
| 15 | Cancer | Cancer des structures sécrétrices | 3,60E-06 | 9,06E-04 | 72 |
| 16 | Cancer | Néoplasie colorectale | 3,88E-06 | 9,15E-04 | 74 |
| 17 | Cancer | Cancer colorectal | 4,32E-06 | 9,26E-04 | 73 |
| 18 | Cancer | Tumorigenèse ou tumeur maligne | 4,42E-06 | 9,26E-04 | 92 |
| 19 | Cancer | Néoplasie gastro-intestinale | 4,77E-06 | 9,47E-04 | 107 |
| 20 | Cancer | Cancer du colon | 6,67E-06 | 1,26E-03 | 66 |
| 21 | Cancer | Cancer du tractus gastro-intestinal | 1,07E-05 | 1,84E-03 | 104 |
| 22 | Cancer | Carcinome prostatique | 1,07E-05 | 1,84E-03 | 50 |
| 23 | Cancer | Néoplasme du gros intestin | 1,21E-05 | 1,95E-03 | 101 |
| 24 | Cancer | Tumorigenèse ou néoplasme épithélial | 1,24E-05 | 1,95E-03 | 91 |
| 25 | Cancer | Tumorigenèse ou tumeur du système digestif | 1,64E-05 | 2,48E-03 | 70 |
| 26 | Cancer | Adénocarcinome gastro-intestinal | 1,84E-05 | 2,60E-03 | 70 |
| 27 | Cancer | Cancer de la prostate et tumeurs | 1,86E-05 | 2,60E-03 | 55 |
| 28 | Cancer | Néoplasme malin du gros intestin | 1,93E-05 | 2,60E-03 | 100 |
| 29 | Cancer | Adénocarcinome du gros intestin | 2,10E-05 | 2,73E-03 | 95 |
| 30 | Cancer | Cancer du colon | 2,28E-05 | 2,82E-03 | 62 |

[a] IPA fournit la *p-value* d'un test exact de Fisher comme indication de la probabilité d'associer une fonction ou une pathologie par hasard à une liste de gènes.

[b] IPA fournit également une *p-value* ajustée selon la procédure de Benjamini-Hochberg permettant d'estimer le taux de fausses découvertes (*FDR*, Benjamini & Hochberg, 1995).

*N.B.* Ce tableau liste les 30 plus fortes associations entre les termes d'annotation IPA et notre liste de gènes candidats. Il s'agit d'une version simplifiée du Tableau supplémentaire 2 de l'Article II.

**Tableau IV-3.** Liste des gènes candidats à la sélection chez le diable de Tasmanie

| ID *Ensembl*[a] | *Scaffold* | Chr. | Pop. | HGNC[b] | Catégorie IPA[c] |
|---|---|---|---|---|---|
| ENSSHAG00000005817 | GL834412 | 1 | FN, WP | GLRX3 | Cancer |
| ENSSHAG00000009641 | GL834480 | 1 | WP | ARRDC4 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015699 | GL834484 | 1 | FN | TLN2 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015834 | GL834484 | 1 | FN | TLN2 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017933 | GL834484 | 1 | FN, WP | HERC1 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000018240 | GL834484 | 1 | FN, WP | DAPK2 | Cancer, Post-Translational Modification |
| ENSSHAG00000018436 | GL834484 | 1 | FN, WP | MRPL46 | Cancer |
| ENSSHAG00000018462 | GL834484 | 1 | FN, WP | FAM96A | Cancer |
| ENSSHAG00000015404 | GL834501 | 1 | FN | SHC4 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015542 | GL834501 | 1 | FN | NDUFAF5 | Cancer |
| ENSSHAG00000016105 | GL834501 | 1 | FN | SECISBP2L | Cancer, Gastrointestinal Disease |
| ENSSHAG00000001548 | GL834502 | 1 | FN | FAM227B | Cancer |
| ENSSHAG00000004048 | GL834502 | 1 | FN | DTWD1 | Cancer |
| ENSSHAG00000007149 | GL834502 | 1 | FN | AC092143.1 | Cancer, Cardiovascular disease, Connective Tissue Disorders, Gastrointestinal Disease |
| ENSSHAG00000007396 | GL834502 | 1 | FN | ATP8B4 | Cancer |
| ENSSHAG00000010850 | GL834528 | 1 | WP | FOXN3 | Cancer |
| ENSSHAG00000016400 | GL834603 | 1 | WP | PPARGC1B | Cancer, Cardiovascular disease, Cell cycle, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000015827 | GL834637 | 1 | WP | NEK6 | Cancer, Gastrointestinal Disease, Post-Translational Modification |
| *ENSSHAG00000016059* | *GL834637* | *1* | *WP* | *PSMB7* | *Unknown* |
| *ENSSHAG00000016480* | *GL834637* | *1* | *WP* | *ADGRD2* | *Unknown* |
| ENSSHAG00000010773 | GL834652 | 1 | FN | NTMT1 | Cancer |
| ENSSHAG00000010852 | GL834652 | 1 | FN | ASB6 | Cancer |
| ENSSHAG00000011321 | GL834652 | 1 | FN | PRRX2 | Cancer, Cellular Development |
| ENSSHAG00000011800 | GL834652 | 1 | FN | PTGES | Cancer, Cardiovascular disease, Cellular Development |
| ENSSHAG00000012185 | GL834652 | 1 | FN | TOR1B | Cancer, Cell Morphology |
| ENSSHAG00000012260 | GL834652 | 1 | FN | TOR1A | Cancer, Cell Morphology, Cellular Development |

| | | | | | |
|---|---|---|---|---|---|
| ENSSHAG00000012313 | GL834652 | 1 | FN | C9orf78 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000012417 | GL834652 | 1 | FN | USP20 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000012634 | GL834652 | 1 | FN | FNBP1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017342 | GL834671 | 1 | FN | ACTR1B | Cancer |
| ENSSHAG00000017358 | GL834671 | 1 | FN | COX5B | Cancer |
| ENSSHAG00000017454 | GL834671 | 1 | FN | ANKRD39 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017685 | GL834671 | 1 | FN | ANKRD23 | Cancer |
| ENSSHAG00000017697 | GL834671 | 1 | FN | CNNM3 | Cancer |
| ENSSHAG00000011303 | GL834709 | 1 | FN, NP | KCNU1 | Cancer |
| ENSSHAG00000016544 | GL834709 | 1 | NP, WP | ADGRA2 | Cancer, Embryonic Development |
| ENSSHAG00000016684 | GL834709 | 1 | NP, WP | RAB11FIP1 | Cancer |
| ENSSHAG00000016864 | GL834709 | 1 | NP, WP | GOT1L1 | Cancer |
| ENSSHAG00000016923 | GL834709 | 1 | NP, WP | ADRB3 | Cancer, Cardiovascular disease, Connective Tissue Disorders, Ophthalmic Disease |
| ENSSHAG00000017260 | GL834709 | 1 | FN | LSM1 | Cancer |
| ENSSHAG00000017318 | GL834709 | 1 | FN | BAG4 | Cancer, Cardiovascular disease |
| ENSSHAG00000017364 | GL834709 | 1 | FN | DDHD2 | Cancer |
| ENSSHAG00000017397 | GL834709 | 1 | FN | PLPP5 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017411 | GL834709 | 1 | FN | NSD3 | Cancer |
| ENSSHAG00000017634 | GL834709 | 1 | FN | LETM2 | Cancer |
| ENSSHAG00000017666 | GL834709 | 1 | FN | FGFR1 | Cancer, Cellular Development, Connective Tissue Disorders, Gastrointestinal Disease, Post-Translational Modification |
| ENSSHAG00000008229 | GL834715 | 1 | FN, WP | XKR6 | Cancer |
| ENSSHAG00000009238 | GL834715 | 1 | FN, WP | PINX1 | Cancer, Connective Tissue Disorders, Gastrointestinal Disease |
| ENSSHAG00000015443 | GL834715 | 1 | FN | KIF13B | Cancer, Cellular Development |
| ENSSHAG00000008576 | GL834716 | 1 | FN, NP | GPAT2 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000008968 | GL834716 | 1 | FN, NP | FAHD2A | Cancer |
| ENSSHAG00000009380 | GL834716 | 1 | FN, NP | KCNIP3 | Cancer |
| ENSSHAG00000010667 | GL834716 | 1 | FN, NP | PROM2 | Cancer, Gastrointestinal Disease |
| *ENSSHAG00000011131* | *GL834716* | *1* | *FN, NP* | *AC092835.1* | *Unknown* |
| ENSSHAG00000008950 | GL834718 | 1 | FN | SYNDIG1 | Cancer, Cellular Development, Connective Tissue Disorders |

| | | | | | |
|---|---|---|---|---|---|
| ENSSHAG00000007476 | GL834719 | 1 | FN, NP | CST3 | Cancer, Cellular Development, Gastrointestinal Disease, Ophthalmic Disease |
| ENSSHAG00000007906 | GL834719 | 1 | FN, NP | CST8 | Cancer |
| ENSSHAG00000009129 | GL834719 | 1 | FN | NAPB | Cancer |
| ENSSHAG00000009367 | GL834719 | 1 | FN | GZF1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000016889 | GL834719 | 1 | FN, WP | SMOX | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017556 | GL834719 | 1 | FN, WP | PAX1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000004915 | GL834720 | 1 | FN | NKX2-2 | Cancer, Cellular Development, Developmental Disorder, Embryonic Development |
| ENSSHAG00000014834 | GL834721 | 1 | FN | RIN2 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015606 | GL834721 | 1 | FN | SLC24A3 | Cancer |
| ENSSHAG00000010204 | GL834736 | 1 | WP | SPTBN1 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000017154 | GL834753 | 1 | FN | CDH8 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000008244 | GL834768 | 1 | FN | SMAD3 | Cancer, Cell-To-Cell Signaling and Interaction, Cellular Development, Connective Tissue Disorders, Gastrointestinal Disease, Inflammatory Response |
| ENSSHAG00000009521 | GL834768 | 1 | FN | GALNS | Cancer, Cellular Development |
| ENSSHAG00000009901 | GL834768 | 1 | FN | TRAPPC2L | Cancer |
| ENSSHAG00000010001 | GL834768 | 1 | FN | PABPN1L | Cancer, Gastrointestinal Disease |
| ENSSHAG00000010129 | GL834768 | 1 | FN | CBFA2T3 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000002820 | GL834783 | 1 | WP | SH2D6 | Cancer |
| ENSSHAG00000004491 | GL834783 | 1 | WP | CAPG | Cancer, Gastrointestinal Disease |
| ENSSHAG00000004885 | GL834783 | 1 | WP | ELMOD3 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000005344 | GL834783 | 1 | WP | RETSAT | Cancer, Gastrointestinal Disease |
| ENSSHAG00000001447 | GL835143 | 1 | FN | EIF2AK2 | Cancer, Cell cycle, Cellular Development, Gastrointestinal Disease, Inflammatory Response, Post-Translational Modification |
| ENSSHAG00000004594 | GL835143 | 1 | FN | SULT6B1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000010181 | GL841174 | 2 | WP | THBS4 | Cancer, Cardiovascular System Development and Function, Cellular Development, Connective Tissue Disorders |
| ENSSHAG00000010987 | GL841174 | 2 | WP | MTX3 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000013221 | GL841246 | 2 | WP | XKR9 | Cancer |
| ENSSHAG00000013334 | GL841246 | 2 | WP | LACTB2 | Cancer |
| ENSSHAG00000013652 | GL841246 | 2 | WP | TRAM1 | Cancer |

| | | | | | |
|---|---|---|---|---|---|
| ENSSHAG00000013805 | GL841246 | 2 | WP | ACER1 | Cancer |
| ENSSHAG00000018290 | GL841374 | 2 | NP | KLF10 | Cancer, Cardiovascular disease, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000013421 | GL841492 | 2 | WP | PTBP3 | Cancer |
| ENSSHAG00000013729 | GL841492 | 2 | WP | FRMD4B | Cancer, Cardiovascular disease, Developmental Disorder, Gastrointestinal Disease, Ophthalmic Disease |
| ENSSHAG00000017709 | GL841543 | 2 | WP | TLE6 | Cancer |
| ENSSHAG00000018861 | GL841593 | 2 | FN, NP, WP | TRNT1 | Cancer |
| ENSSHAG00000018867 | GL841593 | 2 | FN, NP, WP | CRBN | Cancer, Gastrointestinal Disease |
| ENSSHAG00000005543 | GL841951 | 2 | WP | FER1L6 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000005936 | GL849657 | 3 | WP | THY1 | Cancer, Cell-To-Cell Signaling and Interaction, Cellular Development, Gastrointestinal Disease, Post-Translational Modification |
| ENSSHAG00000006515 | GL849657 | 3 | WP | USP2 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000007088 | GL849657 | 3 | WP | MFRP | Cancer, Ophthalmic Disease |
| ENSSHAG00000007202 | GL849657 | 3 | WP | C1QTNF5 | Cancer, Ophthalmic Disease |
| ENSSHAG00000007967 | GL849657 | 3 | WP | MCAM | Cancer, Cell-To-Cell Signaling and Interaction |
| ENSSHAG00000008028 | GL849657 | 3 | WP | CBL | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000010022 | GL849681 | 3 | NP, WP | SLC36A4 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000006975 | GL849790 | 3 | WP | TRIM62 | Cancer, Cell-To-Cell Signaling and Interaction, Gastrointestinal Disease, Inflammatory Response |
| ENSSHAG00000007845 | GL849790 | 3 | WP | SERINC2 | Cancer, Cardiovascular disease, Developmental Disorder, Gastrointestinal Disease, Ophthalmic Disease |
| ENSSHAG00000009199 | GL849860 | 3 | NP, WP | SLITRK5 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000004719 | GL850047 | 3 | WP | GPKOW | Cancer, Gastrointestinal Disease |
| ENSSHAG00000002049 | GL856776 | 4 | FN | TNNT2 | Cancer, Cardiovascular disease, Gastrointestinal Disease |
| ENSSHAG00000003339 | GL856776 | 4 | FN | LAD1 | Cancer |
| ENSSHAG00000004519 | GL856776 | 4 | FN | TNNI1 | Cancer, Cardiovascular System Development and Function |
| ENSSHAG00000018757 | GL856785 | 4 | NP | KCTD3 | Cancer |
| ENSSHAG00000014799 | GL856833 | 4 | NP, WP | SDK2 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000015557 | GL856833 | 4 | NP, WP | CDC42EP4 | Cancer |
| ENSSHAG00000016333 | GL856833 | 4 | NP, WP | SLC39A11 | Cancer |

| | | | | | |
|---|---|---|---|---|---|
| ENSSHAG00000015864 | GL856846 | 4 | FN, WP | SLC38A10 | Cancer |
| ENSSHAG00000016077 | GL856846 | 4 | FN, WP | TEPSIN | Cancer, Gastrointestinal Disease |
| *ENSSHAG00000016121* | *GL856846* | *4* | *FN, WP* | *CEP131* | *Unknown* |
| ENSSHAG00000000170 | GL856873 | 4 | NP | TRIM10 | Cancer, Connective Tissue Disorders, Gastrointestinal Disease, Inflammatory Response |
| ENSSHAG00000002782 | GL856873 | 4 | NP | TRIM15 | Cancer, Inflammatory Response |
| ENSSHAG00000002914 | GL856873 | 4 | NP | TRIM26 | Cancer, Gastrointestinal Disease, Inflammatory Response |
| *ENSSHAG00000005422* | *GL856873* | *4* | *NP* | *TRIM39-RPP21* | *Unknown* |
| ENSSHAG00000014220 | GL856919 | 4 | NP | RRAGD | Cancer |
| ENSSHAG00000015016 | GL856919 | 4 | NP | ANKRD6 | Cancer |
| ENSSHAG00000006847 | GL856972 | 4 | WP | WDR27 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015157 | GL856995 | 4 | NP, WP | PHGDH | Cancer, Cellular Development, Developmental Disorder, Gastrointestinal Disease |
| ENSSHAG00000015388 | GL856995 | 4 | NP, WP | REG4 | Cancer |
| ENSSHAG00000015508 | GL856995 | 4 | NP, WP | HMGCS2 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015795 | GL856995 | 4 | NP, WP | ZNF697 | Cancer |
| ENSSHAG00000015913 | GL856995 | 4 | NP, WP | ADAM30 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015968 | GL856995 | 4 | NP, WP | NOTCH2 | Cancer, Cellular Development |
| *ENSSHAG00000001914* | *GL856999* | *4* | *NP* | *HEATR9* | *Unknown* |
| *ENSSHAG00000002449* | *GL856999* | *4* | *NP* | *P3H4* | *Unknown* |
| ENSSHAG00000003105 | GL856999 | 4 | NP | TAF15 | Cancer |
| ENSSHAG00000004324 | GL856999 | 4 | NP | MMP28 | Cancer |
| *ENSSHAG00000004686* | *GL856999* | *4* | *NP* | *C17orf50* | *Unknown* |
| *ENSSHAG00000005269* | *GL856999* | *4* | *NP* | *AP2B1* | *Unknown* |
| ENSSHAG00000000662 | GL857102 | 4 | FN | UGT1A1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000003180 | GL857102 | 4 | FN | TRPM8 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000010506 | GL861617 | 5 | WP | TSPAN11 | Cancer |
| ENSSHAG00000011343 | GL861617 | 5 | WP | TSPAN9 | Cancer |
| ENSSHAG00000012780 | GL861617 | 5 | WP | TEAD4 | Cancer, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000017819 | GL861623 | 5 | WP | ITIH5 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017960 | GL861623 | 5 | WP | ITIH2 | Cancer, Gastrointestinal Disease |

| | | | | | |
|---|---|---|---|---|---|
| ENSSHAG00000008315 | GL861686 | 5 | NP | NSUN6 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000008820 | GL861686 | 5 | NP | CACNB2 | Cancer, Cardiovascular disease, Cellular Development, Gastrointestinal Disease |
| ENSSHAG00000017835 | GL861688 | 5 | NP, WP | CDK14 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000000749 | GL861701 | 5 | NP | ZNF385D | Cancer |
| ENSSHAG00000004249 | GL861740 | 5 | WP | SRGAP1 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000015985 | GL864807 | 6 | NP | FSTL5 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000017222 | GL864880 | 6 | NP, WP | STK33 | Cancer, Gastrointestinal Disease, Post-Translational Modification |
| ENSSHAG00000017621 | GL864880 | 6 | NP, WP | TRIM66 | Cancer, Gastrointestinal Disease, Inflammatory Response |
| ENSSHAG00000017784 | GL864880 | 6 | NP, WP | ST5 | Cancer, Gastrointestinal Disease |
| ENSSHAG00000007961 | GL864888 | 6 | FN | KCNA4 | Cancer, Cardiovascular disease |
| ENSSHAG00000009491 | GL864888 | 6 | FN | FSHB | Cancer, Cellular Development |

[a] Les identifiants figurant dans la colonne « ID *Ensembl* » sont fournis sous la forme d'un lien redirigeant vers la page *Ensembl* du gène concerné.

[b] La colonne « HGNC » indique l'abréviation du Comité « *HUGO* » de Nomenclature du Gène (HGNC) correspondant à chaque orthologue humain.

[c] La colonne « Catégorie IPA » indique la (ou les) catégorie(s) fonctionnelle(s) IPA associées à chaque orthologue humain. Les entrées en gris italique mettent en évidence les orthologues qui ne sont associés à aucune fonction dans la base de données IPA.

*N.B.* Ce tableau est une version simplifiée du tableau supplémentaire 1 de l'Article II. Les représentations graphiques des soixante signatures de sélection associées à ces gènes sont disponibles en Annexe II.

# — 6. Références


Ageta-Ishihara, N., Yamazaki, M., Konno, K., Nakayama, H., Abe, M., Hashimoto, K., ... & Tanaka, K. (2015). A CDC42EP4/septin-based perisynaptic glial scaffold facilitates glutamate clearance. Nature communications, 6.

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., … Whyte, A. (2015). Key components of data publishing: Using current best practices to develop a reference model for data publishing. http://doi.org/10.5281/zenodo.34542

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Anney, R. J., Ripke, S., Anttila, V., Grove, J., Holmans, P., ... & Neale, B. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Molecular autism, 8, 1-17.

Bachiller, S., Rybkina, T., Porras-García, E., Pérez-Villegas, E., Tabares, L., Armengol, J. A., ... & Ruiz, R. (2015). The HERC1 E3 ubiquitin ligase is essential for normal development and for neurotransmission at the mouse neuromuscular junction. Cellular and molecular life sciences, 72(15), 2961-2971.

Belov, K. (2011). The role of the Major Histocompatibility Complex in the spread of contagious cancers. Mammalian Genome, 22(1-2), 83-90.

Belov, K. (2012). Contagious cancer: lessons from the devil and the dog. BioEssays, 34(4), 285-292.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Blaschke, S., Mueller, C. A., Markovic-Lipkovski, J., Puch, S., Miosge, N., Becker, V., ... & Klein, G. (2002). Expression of cadherin-8 in renal cell carcinoma and fetal kidney. International journal of cancer, 101(4), 327-334.

Bonsignore, L. A., Butler, J. S., Klinge, C. M., & Tooley, C. E. S. (2015). Loss of the N-terminal methyltransferase NRMT1 increases sensitivity to DNA damage and promotes mammary oncogenesis. Oncotarget, 6(14), 12248.

Boone, E., Colombani, J., Andersen, D. S., & Léopold, P. (2016). The *Hippo* signalling pathway coordinates organ growth and limits developmental variability by controlling dilp8 expression. *Nature communications*, *7*, 13505.

Breitenkamp, A. F., Matthes, J., Nass, R. D., Sinzig, J., Lehmkuhl, G., Nürnberg, P., & Herzig, S. (2014). Rare mutations of CACNB2 found in autism spectrum disease-affected families alter calcium channel function. PLoS One, 9(4), e95579.

Brüniche-Olsen, A., Jones, M. E., Austin, J. J., Burridge, C. P., & Holland, B. R. (2014). Extensive population decline in the Tasmanian devil predates European settlement and devil facial tumour disease. Biology letters, 10(11), 20140619.

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338-345.

Caldwell, A., & Siddle, H. V. (2017). The role of MHC genes in contagious cancer: the story of Tasmanian devils. *Immunogenetics*, *69*(8-9), 537-545.

Cao-Ehlker, X., Zong, X., Hammelmann, V., Gruner, C., Fenske, S., Michalakis, S., ... & Biel, M. (2013). Up-regulation of hyperpolarization-activated cyclic nucleotide-gated channel 3 (HCN3) by specific interaction with K+ channel tetramerization domain-containing protein 3 (KCTD3). *Journal of Biological Chemistry*, *288*(11), 7580-7589.

Cappola, T. P., Li, M., He, J., Ky, B., Gilmore, J., Qu, L., ... & Frackelton, E. (2010). Common variants in HSPB7 and FRMD4B associated with advanced heart failure. *Circulation: Cardiovascular Genetics*, CIRCGENETICS-109.

Chen, S. W., Chou, C. T., Chang, C. C., Li, Y. J., Chen, S. T., Lin, I. C., ... & Kuo, M. L. (2017). HMGCS2 enhances invasion and metastasis via direct interaction with PPARa to activate Src signaling in colorectal cancer and oral cancer. Oncotarget, 8(14), 22460.

Chen, Y., Guo, Y., Yang, H., Shi, G., Xu, G., Shi, J., ... & Chen, D. (2015). TRIM66 overexpression contributes to osteosarcoma carcinogenesis and indicates poor survival outcome. Oncotarget, 6(27), 23708.

Cheng, L., Chen, C. L., Luo, P., Tan, M., Qiu, M., Johnson, R., & Ma, Q. (2003). Lmx1b, Pet-1, and Nkx2.2 coordinately specify serotonergic neurotransmitter phenotype. Journal of Neuroscience, 23(31), 9961-9967.

Cheng, Y., Sanderson, C., Jones, M., & Belov, K. (2012). Low MHC class II diversity in the Tasmanian devil (Sarcophilus harrisii). Immunogenetics, 64(7), 525-533.

Cowell, J. K., Qin, H., Hu, T., Wu, Q., Bhole, A., & Ren, M. (2017). Mutation in the FGFR1 tyrosine kinase domain or inactivation of PTEN is associated with acquired resistance to FGFR inhibitors in FGFR1-driven leukemia/lymphomas. International Journal of Cancer.

Dahl, M., Maturi, V., Lönn, P., Papoutsoglou, P., Zieba, A., Vanlandewijck, M., ... & Heldin, C. H. (2014). Fine-tuning of Smad protein function by poly (ADP-ribose) polymerases and poly (ADP-ribose) glycohydrolase during transforming growth factor β signaling. *PloS one*, *9*(8), e103651.

Dai, Y., Wang, M., Wu, H., Xiao, M., Liu, H., & Zhang, D. (2017). Loss of FOXN3 in colon cancer activates beta-catenin/TCF signaling and promotes the growth and migration of cancer cells. Oncotarget, 8(6), 9783.

Detchokul, S., Williams, E. D., Parker, M. W., & Frauman, A. G. (2014). Tetraspanins as regulators of the tumour microenvironment: implications for metastasis and therapeutic strategies. British journal of pharmacology, 171(24), 5462-5490.

Dieci, M. V., Arnedos, M., Andre, F., & Soria, J. C. (2013). Fibroblast growth factor receptor inhibitors as a cancer treatment: from a biologic rationale to medical perspectives. Cancer discovery, 3(3), 264-279.

Ehmer, U., & Sage, J. (2016). Control of proliferation and cancer growth by the *Hippo* signaling pathway. *Molecular Cancer Research*, *14*(2), 127-140.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ... & Storfer, A. (2016a). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Nature communications, 7, 12684.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ... & Storfer, A. (2016b). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.r60sv

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. The FASEB journal, 22(8), 2623-2628.

Feng, Y., Feng, L., Yu, D., Zou, J., & Huang, Z. (2016). srGAP1 mediates the migration inhibition effect of Slit2-Robo1 in colorectal cancer. Journal of Experimental & Clinical Cancer Research, 35(1), 191.

French, C. A., Rahman, S., Walsh, E. M., Kühnle, S., Grayson, A. R., Lemieux, M. E., ... & Venkatramani, R. (2014). NSD3–NUT fusion oncoprotein in NUT midline carcinoma: implications for a novel oncogenic mechanism. Cancer discovery, 4(8), 928-941.

Garcia, M. J., Pole, J. C., Chin, S. F., Teschendorff, A., Naderi, A., Ozdag, H., ... & Ellis, I. (2005). A 1 Mb minimal amplicon at 8p11-12 in breast cancer identifies new candidate oncogenes. Oncogene, 24(33), 5235.

Gonzalez, W. G., Pham, K., & Miksovska, J. (2014). Modulation of the voltage-gated potassium channel (Kv4.3) and the auxiliary protein (KChIP3) interactions by the current activator NS5806. *Journal of Biological Chemistry*, *289*(46), 32201-32213.

Gooley, R., Hogg, C. J., Belov, K., & Grueber, C. E. (2017). No evidence of inbreeding depression in a Tasmanian devil insurance population despite significant variation in inbreeding. Scientific Reports, 7.

Hamede, R., Lachish, S., Belov, K., Woods, G., Kreiss, A., PEARSE, A., ... & McCallum, H. (2012). Reduced effect of Tasmanian devil facial tumor disease at the disease front. Conservation Biology, 26(1), 124-134.

Hamede, R. K., McCallum, H., & Jones, M. (2013). Biting injuries and transmission of Tasmanian devil facial tumour disease. *Journal of Animal Ecology*, *82*(1), 182-190.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, *144*(5), 646-674.

Hardy, S., Uetani, N., Wong, N., Kostantin, E., Labbe, D. P., Bégin, L. R., ... & Tremblay, M. L. (2015). The protein tyrosine phosphatase PRL-2 interacts with the magnesium transporter CNNM3 to promote oncogenesis. Oncogene, 34(8), 986-995.

Harvey, K. F., Zhang, X., & Thomas, D. M. (2013). The *Hippo* pathway and human cancer. Nature Reviews Cancer, 13(4), 246-257.

Hawkins, C. E., Baars, C., Hesterman, H., Hocking, G. J., Jones, M. E., Lazenby, B., ... & Restani, M. (2006). Emerging disease and population decline of an island endemic, the Tasmanian devil Sarcophilus harrisii. Biological Conservation, 131(2), 307-324.

Hellsten, S. V., Hägglund, M. G., Eriksson, M. M., & Fredriksson, R. (2017). The neuronal and astrocytic protein SLC38A10 transports glutamine, glutamate, and aspartate, suggesting a role in neurotransmission. FEBS open bio, 7(6), 730-746.

Helsmoortel, C., Vandeweyer, G., Ordoukhanian, P., Van Nieuwerburgh, F., Van der Aa, N., & Kooy, R. F. (2015). Challenges and opportunities in the investigation of unexplained intellectual disability using family-based whole-exome sequencing. Clinical genetics, 88(2), 140-148.

Hemler, M. E. (2005). Tetraspanin functions and associated microdomains. Nature reviews Molecular cell biology, 6(10), 801-811.

Hemler, M. E. (2014). Tetraspanin proteins promote multiple cancer stages. Nature reviews. Cancer, 14(1), 49.

Hendricks, S., Epstein, B., Schönfeld, B., Wiench, C., Hamede, R., Jones, M., ... & Hohenlohe, P. (2017). Conservation implications of limited genetic diversity and population structure in Tasmanian devils (Sarcophilus harrisii). Conservation Genetics, 1-6.

Hirokawa, N., Noda, Y., Tanaka, Y., & Niwa, S. (2009). Kinesin superfamily motor proteins and intracellular transport. *Nature reviews Molecular cell biology*, *10*(10), 682-696.

Hollis, F., Kanellopoulos, A. K., & Bagni, C. (2017). Mitochondrial dysfunction in Autism Spectrum Disorder: clinical features and perspectives. Current Opinion in Neurobiology, 45, 178-187.

Hydbring, P., Malumbres, M., & Sicinski, P. (2016). Non-canonical functions of cell cycle cyclins and cyclin-dependent kinases. Nature reviews. Molecular cell biology, 17(5), 280

INRA (2017). Une science ouverte grâce au numérique. [En ligne] http://2025.inra.fr/ (consulté le 07/01/2018).

Ikushima, H., & Miyazono, K. (2010). *TGF*β signalling: a complex web in cancer progression. *Nature reviews cancer*, *10*(6), 415-424.

Ioannou, M. S., & McPherson, P. S. (2016). Regulation of cancer cell behavior by the small GTPase Rab13. Journal of Biological Chemistry, 291(19), 9929-9937.

Jiang, Y., Woronicz, J. D., Liu, W., & Goeddel, D. V. (1999). Prevention of constitutive TNF receptor 1 signaling by silencer of death domains. Science, 283(5401), 543-546.

Jones, D. T., Hutter, B., Jäger, N., Korshunov, A., Kool, M., Warnatz, H. J., ... & Fontebasso, A. M. (2013). Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. Nature genetics, 45(8), 927-932.

Jones, M. E., Paetkau, D., Geffen, E. L. I., & Moritz, C. (2004). Genetic diversity and population structure of Tasmanian devils, the largest marsupial carnivore. Molecular Ecology, 13(8), 2197-2209.

Juang, Y. L., Jeng, Y. M., Chen, C. L., & Lien, H. C. (2016). PRRX2 as a novel *TGF*-β-induced factor enhances invasion and migration in mammary epithelial cell and correlates with poor prognosis in breast cancer. Molecular carcinogenesis, 55(12), 2247-2259.

Kaufman, L., Ayub, M., & Vincent, J. B. (2010). The genetic basis of non-syndromic intellectual disability: a review. Journal of neurodevelopmental disorders, 2(4), 182.

Kaya, N., Alsagob, M., D'adamo, M. C., Al-Bakheet, A., Hasan, S., Muccioli, M., ... & Mustafa, O. M. (2016). KCNA4 deficiency leads to a syndrome of abnormal striatum, congenital cataract and intellectual disability. Journal of medical genetics, 53(11), 786-792.

Kourtidis, A., Lu, R., Pence, L., & Anastasiadis, P. Z. (2017). A central role for cadherin signaling in cancer. Experimental Cell Research.

Krishnaswamy, A., Yamagata, M., Duan, X., Hong, Y. K., & Sanes, J. R. (2015). Sidekick 2 directs formation of a retinal circuit that detects differential motion. Nature, 524(7566), 466.

Labbé, D. P., Hardy, S., & Tremblay, M. L. (2012). Protein tyrosine phosphatases in cancer: friends and foes!. Progress in molecular biology and translational science, 106, 253-306.

Lachish, S., McCallum, H., & Jones, M. (2009). Demography, disease and the devil: life-history changes in a disease-affected population of Tasmanian devils (Sarcophilus harrisii). Journal of Animal Ecology, 78(2), 427-436.

Lachish, S., McCallum, H., Mann, D., Pukk, C. E., & Jones, M. E. (2010). Evaluation of selective culling of infected individuals to control Tasmanian devil facial tumor disease. Conservation Biology, 24(3), 841-851.

Langenhan, T., Aust, G., & Hamann, J. (2013). Sticky signaling—adhesion class G protein-coupled receptors take the stage. Sci Signal, 6(2), re3.

Lee, Y., Yoon, K. A., Joo, J., Lee, D., Bae, K., Han, J. Y., & Lee, J. S. (2013). Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study. Carcinogenesis, 34(2), 307-313.

Legoffic, A., Calvo, E., Cano, C., Folch-Puy, E., Barthet, M., Delpero, J. R., ... & Iovanna, J. (2009). The reg4 gene, amplified in the early stages of pancreatic cancer development, is a promising therapeutic target. PloS one, 4(10), e7495.

Li, C., Wang, S., Xing, Z., Lin, A., Liang, K., Song, J., ... & Hawke, D. H. (2017). A ROR1-HER3-LncRNA signaling axis modulates the *Hippo*-YAP pathway to regulate bone metastasis. Nature cell biology, 19(2), 106.

Li, D. M., & Feng, Y. M. (2011). Signaling mechanism of cell adhesion molecules in breast cancer metastasis: potential therapeutic targets. Breast cancer research and treatment, 128(1), 7.

Li, H. L., Song, J., Yong, H. M., Hou, P. F., Chen, Y. S., Song, W. B., ... & Zheng, J. N. (2016). PinX1: structure, regulation and its functions in cancer. Oncotarget, 7(40), 66267.

Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., ... & Xia, K. (2016). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. Molecular psychiatry, 21(2), 290.

Lin, L., Yao, Z., Bhuvaneshwar, K., Gusev, Y., Kallakury, B., Yang, S., ... & He, A. R. (2014). Transcriptional regulation of STAT3 by SPTBN1 and SMAD3 in HCC through cAMP-response element-binding proteins ATF3 and CREB2. Carcinogenesis, 35(11), 2393-2403.

Lin, X., Hu, D., Chen, G., Shi, Y., Zhang, H., Wang, X., ... & Luo, X. (2016). Associations of THBS2 and THBS4 polymorphisms to gastric cancer in a Southeast Chinese population. Cancer genetics, 209(5), 215-222.

Little, E. C., Camp, E. R., Wang, C., Watson, P. M., Watson, D. K., & Cole, D. J. (2016). The CaSm (LSm1) oncogene promotes transformation, chemoresistance and metastasis of pancreatic cancer cells. Oncogenesis, 5(1), e182.

Lu, Y., Lemon, W., Liu, P. Y., Yi, Y., Morrison, C., Yang, P., ... & Govindan, R. (2006). A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS medicine, 3(12), e467.

Ma, Y., Yue, Y., Pan, M., Sun, J., Chu, J., Lin, X., ... & Shin, V. Y. (2015). Histone deacetylase 3 inhibits new tumor suppressor gene DTWD1 in gastric cancer. American journal of cancer research, 5(2), 663.

Maciejowski, J., & de Lange, T. (2017). Telomeres in cancer: tumour suppression and genome instability. Nature Reviews Molecular Cell Biology.

Madsen, T., Arnal, A., Vittecoq, M., Bernex, F., Abadie, J., Labrut, S., ... & Roche, B. (2017). Cancer Prevalence and Etiology in Wild and Captive Animals. Ecology and Evolution of Cancer, 11.

Majumder, S., Bhowal, A., Basu, S., Mukherjee, P., Chatterji, U., & Sengupta, S. (2016). Deregulated E2F5/p38/SMAD3 Circuitry Reinforces the Pro-Tumorigenic Switch of *TGF*β Signaling in Prostate Cancer. *Journal of cellular physiology*, *231*(11), 2482-2492

Malumbres, M., & Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. Nature reviews. Cancer, 9(3), 153.

McAloose, D., & Newton, A. L. (2009). Wildlife cancer: a conservation perspective. *Nature reviews cancer*, *9*(7), 517-526.

McCallum, H., Tompkins, D. M., Jones, M., Lachish, S., Marvanek, S., Lazenby, B., ... & Hawkins, C. E. (2007). Distribution and impacts of Tasmanian devil facial tumor disease. EcoHealth, 4(3), 318.

McCallum, H. (2008). Tasmanian devil facial tumour disease: lessons for conservation biology. Trends in Ecology & Evolution, 23(11), 631-637.

McCallum, H., Jones, M., Hawkins, C., Hamede, R., Lachish, S., Sinn, D. L., ... & Lazenby, B. (2009). Transmission dynamics of Tasmanian devil facial tumor disease may lead to disease-induced extinction. Ecology, 90(12), 3379-3392.

Metzger, M. J., Reinisch, C., Sherry, J., & Goff, S. P. (2015). Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell*, *161*(2), 255-263.

Metzger, M. J., Villalba, A., Carballal, M. J., Iglesias, D., Sherry, J., Reinisch, C., ... & Goff, S. P. (2016). Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature*.

Miller, W., Hayes, V. M., Ratan, A., Petersen, D. C., Wittekindt, N. E., Miller, J., ... & Wang, Q. (2011). Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). Proceedings of the National Academy of Sciences, 108(30), 12348-12353.

Murchison, E. P., Schulz-Trieglaff, O. B., Ning, Z., Alexandrov, L. B., Bauer, M. J., Fu, B., ... & Ng, B. L. (2012). Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell*, *148*(4), 780-791.

Murchison, E. P., Wedge, D. C., Alexandrov, L. B., Fu, B., Martincorena, I., Ning, Z., ... & Donelan, E. M. (2014). Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. Science, 343(6169), 437-440.

Nowell, C. S., & Radtke, F. (2017). Notch as a tumour suppressor. Nature Reviews Cancer, 17(3), 145-159.

Nusse, R., & Clevers, H. (2017). *Wnt*/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell, 169(6), 985-999.

Okada, H., & Mak, T. W. (2004). Pathways of apoptotic and non-apoptotic death in tumour cells. Nature reviews. Cancer, 4(8), 592.

Orsmond, G. I., Krauss, M. W., & Seltzer, M. M. (2004). Peer relationships and social and recreational activities among adolescents and adults with autism. Journal of autism and developmental disorders, 34(3), 245-256.

Ouhtit, A., Gaur, R. L., Elmageed, Z. Y. A., Fernando, A., Thouta, R., Trappey, A. K., ... & Raj, M. G. (2009). Towards understanding the mode of action of the multifaceted cell adhesion receptor CD146. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1795(2), 130-136.

Pagnamenta, A. T., Khan, H., Walker, S., Gerrelli, D., Wing, K., Bonaglia, M. C., ... & Pinto, D. (2011). Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (CDH8) in susceptibility to autism and learning disability. Journal of medical genetics, 48(1), 48-54.

Pal, I., & Mandal, M. (2012). PI3K and Akt as molecular targets for cancer therapy: current clinical outcomes. Acta Pharmacologica Sinica, 33(12), 1441.

Papuc, S. M., Hackmann, K., Andrieux, J., Vincent-Delorme, C., Budişteanu, M., Arghir, A., ... & Di Donato, N. (2015). Microduplications of 3p26. 3p26. 2 containing CRBN gene in patients with intellectual disability and behavior abnormalities. *European journal of medical genetics*, *58*(5), 319-323.

Pearse, A. M., & Swift, K. (2006). Allograft theory: transmission of devil facial-tumour disease. Nature, 439(7076), 549-549.

Poot, M., Beyer, V., Schwaab, I., Damatova, N., van't Slot, R., Prothero, J., ... & Haaf, T. (2010). Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay and autism spectrum disorder. Neurogenetics, 11(1), 81-89.

Pruunsild, P., & Timmusk, T. (2012). Subcellular localization and transcription regulatory potency of KCNIP/Calsenilin/DREAM/KChIP proteins in cultured primary cortical neurons do not provide support for their role in CRE-dependent gene expression. *Journal of neurochemistry*, *123*(1), 29-43.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559-575.

Pye, R. J., Pemberton, D., Tovar, C., Tubio, J. M., Dun, K. A., Fox, S., ... & Siddle, H. V. (2016a). A second transmissible cancer in Tasmanian devils. Proceedings of the National Academy of Sciences, 113(2), 374-379.

Pye, R., Hamede, R., Siddle, H. V., Caldwell, A., Knowles, G. W., Swift, K., ... & Woods, G. M. (2016b). Demonstration of immune responses against devil facial tumour disease in wild Tasmanian devils. Biology letters, 12(10), 20160553.

Pyecroft, S. B., Pearse, A. M., Loh, R., Swift, K., Belov, K., Fox, N., ... & Boyle, D. (2007). Towards a case definition for devil facial tumour disease: what is it?. *EcoHealth*, *4*(3), 346.

Qu, Q., Mao, Y., Xiao, G., Fei, X., Wang, J., Zhang, Y., ... & Shen, K. (2015). USP2 promotes cell migration and invasion in triple negative breast cancer cell lines. Tumor Biology, 36(7), 5415-5423.

Rege, T. A., & Hagood, J. S. (2006). Thy-1 as a regulator of cell-cell and cell-matrix interactions in axon regeneration, apoptosis, adhesion, migration, cancer, and fibrosis. The FASEB journal, 20(8), 1045-1054.

Rhein, V. F., Carroll, J., Ding, S., Fearnley, I. M., & Walker, J. E. (2016). NDUFAF5 hydroxylates NDUFS7 at an early stage in the assembly of human complex I. Journal of Biological Chemistry, 291(28), 14851-14860.

Ridley, A. J. (2015). Rho GTPase signalling in cell migration. *Current opinion in cell biology*, *36*, 103-112.

Robert, J. (2013). Biologie de la métastase. *Bulletin du cancer*, *100*(4), 333-342.

Robert, J. (2017). *Principes généraux de la signalisation cellulaire* dans Robert, J. (dir.), Signalisation cellulaire et cancer : bases biologiques de la cancérologie 2e éd., Lavoisier, pp 1-5, ISBN 978-2-257-20708-1.

Rose, M., Gaisa, N. T., Antony, P., Fiedler, D., Heidenreich, A., Otto, W., ... & Knüchel, R. (2013). Epigenetic inactivation of ITIH5 promotes bladder cancer progression and predicts early relapse of pT1 high-grade urothelial tumours. Carcinogenesis, 35(3), 727-736.

Rowen, L., Wong, G. K., Lane, R. P., & Hood, L. (2000). Publication rights in the era of open data release policies. Science, 289(5486), 1881-1881.

Samanta, D., Park, Y., Andrabi, S. A., Shelton, L. M., Gilkes, D. M., & Semenza, G. L. (2016). PHGDH expression is required for mitochondrial redox homeostasis, breast cancer stem cell maintenance, and lung metastasis. Cancer research, 76(15), 4430-4442.

Sandri, C., Caccavari, F., Valdembri, D., Camillo, C., Veltel, S., Santambrogio, M., ... & Serini, G. (2012). The R-Ras/RIN2/Rab5 complex controls endothelial cell adhesion and morphogenesis via active integrin endocytosis and Rac signaling. Cell research, 22(10), 1479.

Santi, C. M., Martínez-López, P., de la Vega-Beltrán, J. L., Butler, A., Alisio, A., Darszon, A., & Salkoff, L. (2010). The SLO3 sperm-specific potassium channel plays a vital role in male fertility. *FEBS letters*, *584*(5), 1041-1046.

Save the Tasmanian Devil Program (2017). [en ligne] consulté le 15/01/2018 http://www.tassiedevil.com.au/tasdevil.nsf/news/30208C63F8C5338CCA258186000E5EB3

Schuster, S. R., Kortuem, K. M., Zhu, Y. X., Braggio, E., Shi, C. X., Bruins, L. A., ... & Mikhael, J. (2014). The clinical significance of cereblon expression in multiple myeloma. *Leukemia research*, *38*(1), 23-28.

Shi, Q., & Chen, Y. G. (2017). Interplay between *TGF*-β signaling and receptor tyrosine kinases in tumor development. *Science China Life Sciences*, 1-9.

Shmelkov, S. V., Hormigo, A., Jing, D., Proenca, C. C., Bath, K. G., Milde, T., ... & Murphy, A. J. (2010). Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive-like behaviors in mice. Nature medicine, 16(5), 598-602.

Stanton, C. M., Borooah, S., Drake, C., Marsh, J. A., Campbell, S., Lennon, A., ... & Dhillon, B. (2017). Novel pathogenic mutations in C1QTNF5 support a dominant negative disease mechanism in late-onset retinal degeneration. *Scientific Reports*, *7*(1), 12147.

Staples, C. J., Myers, K. N., Beveridge, R. D., Patil, A. A., Lee, A. J., Swanton, C., ... & Collis, S. J. (2012). The centriolar satellite protein Cep131 is important for genome stability. J Cell Sci, 125(20), 4770-4779.

Storfer, A., Epstein, B., Jones, M., Micheletti, S., Spear, S. F., Lachish, S., & Fox, S. (2017). Landscape genetics of the Tasmanian devil: implications for spread of an infectious cancer. Conservation Genetics, 1-11.

Strub, T., Kobi, D., Koludrovic, D., & Davidson, I. (2011). A POU3F2-MITF-SHC4 axis in phenotype switching of melanoma cells. In Research on Melanoma - A Glimpse into Current Directions and Future Trends, InTech, doi: 10.5772/19769.

Subramaniam, M., Hawse, J. R., Rajamannan, N. M., Ingle, J. N., & Spelsberg, T. C. (2010). Functional role of KLF10 in multiple disease processes. *Biofactors*, *36*(1), 8-18.

Sundin, O. H., Leppert, G. S., Silva, E. D., Yang, J. M., Dharmaraj, S., Maumenee, I. H., ... & DiBernardo, C. (2005). Extreme hyperopia is the result of null mutations in MFRP, which encodes a Frizzled-related protein. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9553-9558.

Syed, V. (2016). *TGF*-β Signaling in Cancer. Journal of cellular biochemistry, 117(6), 1279-1287.

Taheri, A., Perry, A., & Minnes, P. (2016). Examining the social participation of children and adolescents with Intellectual Disabilities and Autism Spectrum Disorder in relation to peers. Journal of Intellectual Disability Research, 60(5), 435-443.

Terali, K., & Yilmazer, A. (2016). New surprises from an old favourite: The emergence of telomerase as a key player in the regulation of cancer stemness. Biochimie, 121, 170-178.

Thelwall, M., Thelwall, M., Kousha, K., & Kousha, K. (2017). Do journal data sharing mandates work? Life sciences evidence from *Dryad*. *Aslib Journal of Information Management*, *69*(1), 36-45.

Thomas, A. L., Lind, H., Hong, A., Dokic, D., Oppat, K., Rosenthal, E., ... & Jeruss, J. S. (2017). Inhibition of CDK-mediated Smad3 phosphorylation reduces the Pin1-Smad3 interaction and aggressiveness of triple negative breast cancer cells. *Cell Cycle*, *16*(15), 1453-1464.

Thomas, F., Jacqueline, C., Tissot, T., Henard, M., Blanchet, S., Loot, G., ... & Beckmann, C. (2017). The importance of cancer cells for animal evolutionary ecology. *Nature Ecology & Evolution*, *1*(11), 1592.

Tovar, C., Pye, R. J., Kreiss, A., Cheng, Y., Brown, G. K., Darby, J., ... & Silva, A. (2017). Regression of devil facial tumour disease following immunotherapy in immunised Tasmanian devils. Scientific Reports, 7.

Tufegdzic Vidakovic, A., Rueda, O. M., Vervoort, S. J., Batra, A. S., Goldgraben, M. A., Uribe-Lewis, S., ... & Caldas, C. (2015). Context-specific effects of *TGF*-β/smad3 in cancer are modulated by the epigenome. *Cell reports*, *13*(11), 2480-2490.

Ujvari, B., Gatenby, R. A., & Thomas, F. (2016). Transmissible cancers, are they more common than thought?. Evolutionary applications, 9(5), 633-634.

Utine, G. E., Taşkıran, E. Z., Koşukcu, C., Karaosmanoğlu, B., Güleray, N., Doğan, Ö. A., ... & Alikaşifoğlu, M. (2017). HERC1 mutations in idiopathic intellectual disability. European Journal of Medical Genetics, 60(5), 279-283.

Vallon, M., & Essler, M. (2006). Proteolytically processed soluble tumor endothelial marker (TEM) 5 mediates endothelial cell survival during angiogenesis by linking integrin αvβ3 to glycosaminoglycans. Journal of Biological Chemistry, 281(45), 34179-34188.

van Rensburg, H. J. J., & Yang, X. (2016). The roles of the *Hippo* pathway in cancer metastasis. Cellular signalling, 28(11), 1761-1772.

van Vlodrop, I. J., Joosten, S. C., De Meyer, T., Smits, K. M., Van Neste, L., Melotte, V., ... & Yi, J. M. (2016). A four-gene promoter methylation marker panel consisting of GREM1, NEURL, LAD1, and NEFH predicts survival of clear cell renal cell cancer patients. Clinical Cancer Research.

Vittecoq, M., Ducasse, H., Arnal, A., Møller, A. P., Ujvari, B., Jacqueline, C. B., ... & Lemberger, K. (2015). Animal behaviour and cancer. *Animal Behaviour*, *101*, 19-26.

Vougiouklakis, T., Hamamoto, R., Nakamura, Y., & Saloura, V. (2015). The NSD family of protein methyltransferases in human cancer. Epigenomics 7:5.

Walton, K. M., & Ingersoll, B. R. (2013). Improving social skills in adolescents and adults with autism and severe to profound intellectual disability: A review of the literature. Journal of Autism and Developmental Disorders, 43(3), 594-615.

Wang, J., & Martin, J. F. (2017). *Hippo* Pathway: An Emerging Regulator of Craniofacial and Dental Development. *Journal of Dental Research*, *96*(11), 1229-1237.

Wei, B., & Jin, J. P. (2016). TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure–function relationships. *Gene*, *582*(1), 1-13.

Wolf, F. I., & Trapani, V. (2012). Magnesium and its transporters in cancer: a novel paradigm in tumour development. Clinical Science, 123(7), 417-427.

Woodbury-Smith, M., Paterson, A. D., Thiruvahindrapduram, B., Lionel, A. C., Marshall, C. R., Merico, D., ... & Chrysler, C. (2015). Using extended pedigrees to identify novel autism spectrum disorder (ASD) candidate genes. Human genetics, 134(2), 191-201.

Wu, Z., Wu, Z., Li, J., Yang, X., Wang, Y., Yu, Y., ... & Zhang, Z. (2012). MCAM is a novel metastasis marker and regulates spreading, apoptosis and invasion of ovarian cancer cells. Tumor Biology, 33(5), 1619-1628.

Yan, Z., Kim, E., Datta, D., Lewis, D. A., & Soderling, S. H. (2016). Synaptic actin dysregulation, a convergent mechanism of mental disorders?. *Journal of Neuroscience*, *36*(45), 11411-11417.

Yang, Y. A., Zhang, G. M., Feigenbaum, L., & Zhang, Y. E. (2006). Smad3 reduces susceptibility to hepatocarcinoma by sensitizing hepatocytes to apoptosis through downregulation of Bcl-2. Cancer cell, 9(6), 445-457.

Yona, S., Lin, H. H., Siu, W. O., Gordon, S., & Stacey, M. (2008). Adhesion-GPCRs: emerging roles for novel receptors. Trends in biochemical sciences, 33(10), 491-500.

Zhang, B., Zhang, H., Wang, D., Han, S., Wang, K., Yao, A., & Li, X. (2014). Never in mitosis gene A‐related kinase 6 promotes cell proliferation of hepatocellular carcinoma via cyclin B modulation. Oncology letters, 8(3), 1163-1168.

Zhang, D. H., Yang, Z. L., Zhou, E. X., Miao, X. Y., Zou, Q., Li, J. H., ... & Chen, S. L. (2016). Overexpression of Thy1 and ITGA6 is associated with invasion, metastasis and poor prognosis in human gallbladder carcinoma. Oncology letters, 12(6), 5136-5144.

Zhao, B., Tumaneng, K., & Guan, K. L. (2011). The *Hippo* pathway in organ size control, tissue regeneration and stem cell self-renewal. *Nature cell biology*, *13*(8), 877-883.

Zhi, X., Lin, L., Yang, S., Bhuvaneshwar, K., Wang, H., Gusev, Y., ... & Tian, X. (2015). βII-Spectrin (SPTBN1) suppresses progression of hepatocellular carcinoma and *Wnt* signaling by regulation of *Wnt* inhibitor kallistatin. *Hepatology*, *61*(2), 598-612.

Zurdel, J., Finckh, U., Menzer, G., Nitsch, R. M., & Richard, G. (2002). CST3 genotype associated with exudative age related macular degeneration. *British journal of ophthalmology*, *86*(2), 214-219.

# Chapitre V

# TABLE DES MATIERES

-

# Chapitre V – Le séquençage de marqueurs RAD

## — 1. Les techniques de séquençage de nouvelle génération (NGS)

La promesse d'un génome humain à moins de 1000 $, soutenue par les programmes de financement du *National Human Genome Research Institute* (NHGRI) démarrés en 2004, a rapidement catalysé le développement et la commercialisation de ce qu'il est coutumier d'appeler aujourd'hui les techniques de séquençage de nouvelle génération (NGS). Inspirées dans leur principe par le séquençage de type *shotgun* popularisé par Craig Venter, ces techniques s'appuient sur le séquençage parallèle de millions de fragments de génome (Shendure & Ji, 2008).

En augmentant la rapidité d'acquisition des données (d'un facteur 3500 au cours des dix dernières années selon Illumina, 2017) et en diminuant leur coût (d'un facteur supérieur à 10000 sur la même période d'après NHGRI, 2016), les NGS ont ouvert de très vastes perspectives, évidemment en médecine humaine (*e.g.*, Nelson *et al.*, 2015 ; Fontanges *et al.*, 2016), mais plus largement encore en écologie et évolution (*e.g.*, Feder *et al.*, 2012 ; Barba *et al.*, 2014 ; Ibaraki *et al.*, 2015). La possibilité de générer des milliers de marqueurs couvrant l'ensemble du génome en une seule expérience, chose qui représentait un idéal presque hors de portée au début des années 2000, permet en effet de travailler même chez les espèces pour lesquelles on ne dispose pas de données de cartographie (Ekblom & Galindo, 2011). Ceci a favorisé l'émergence de protocoles de plus en plus flexibles, adaptés au traitement de questions spécifiques (*e.g.*, Truong *et al.*, 2012 ; Jones & Good, 2016), mais a aussi amené un cortège nouveau d'interrogations, en particulier au sujet des capacités de traitement des données générées (Schmidt & Hildebrandt, 2017).

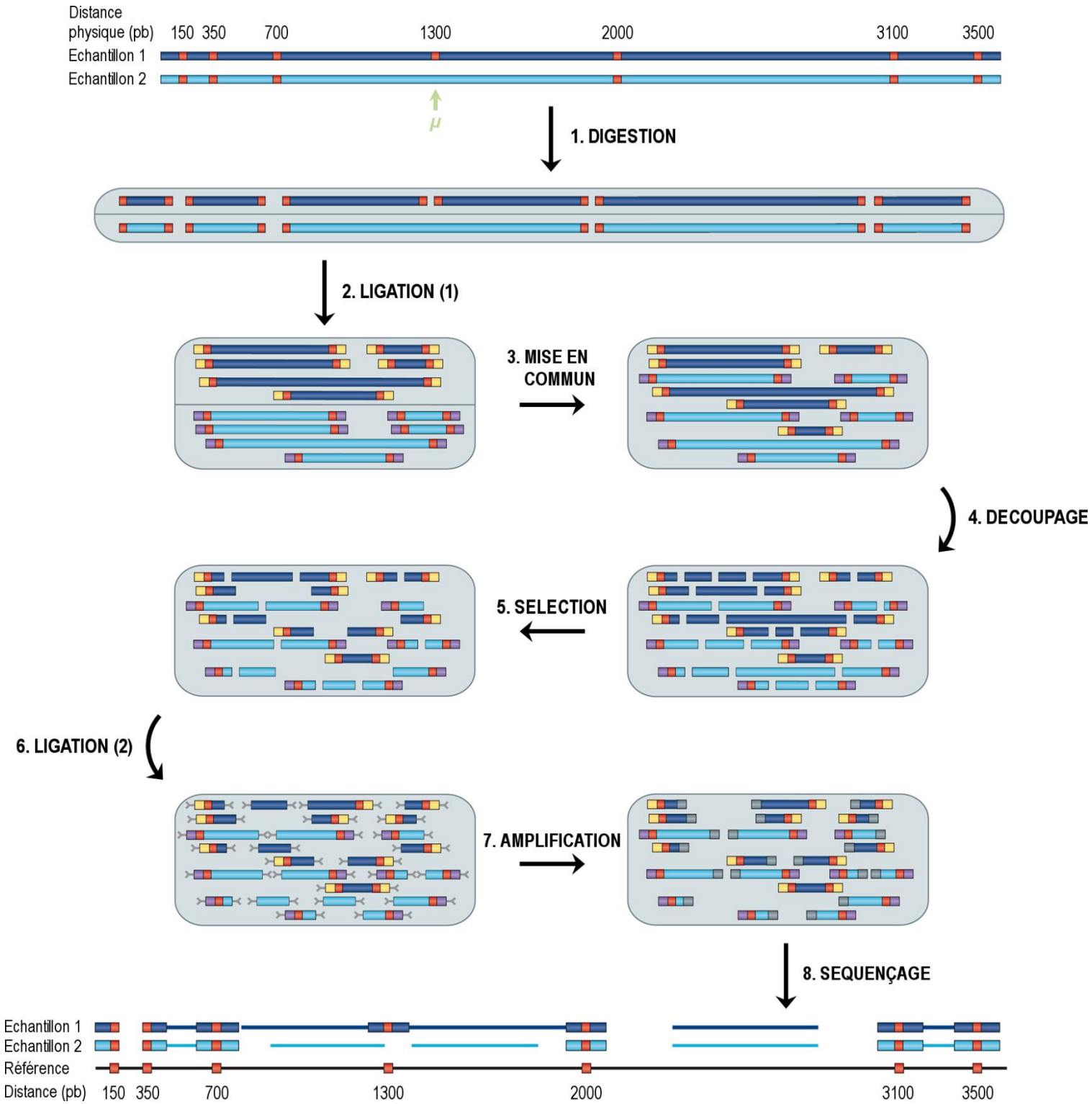



**Figure V-1. Principe général et grandes étapes d'un protocole de RAD-seq (modifié à partir de Davey *et al.*, 2011, Fig. 1).** Le génome de deux individus a été soumis à un protocole de RAD-seq. Deux échantillons (un par individu) de la même région génomique ont été représentés (l'échantillon 1 en bleu foncé et l'échantillon 2 en bleu clair). Les sites de restriction de cette région ont été indiqués en rouge. L'échantillon 2 présente une mutation ($\mu$) dans le site de restriction situé à une distance de 1300 nucléotides (pb) du début de la région examinée. 1. La première étape du protocole consiste en la digestion de l'ADN par une enzyme de restriction qui va générer un nombre de fragments dépendant du nombre de site de restriction accessibles. Dans notre exemple, il y aura moins de fragments issus de l'échantillon 2 du fait d'un site de restriction muté. 2. Des

adaptateurs P1 (représentés en jaune et composés d'une amorce d'amplification, d'une amorce de séquençage et d'un code-barres propre à chaque individu) sont liés au niveau des sites de restriction des fragments. 3. Les fragments ainsi marqués sont mis en commun (« *pooling* »), puis (4.) découpés en plus petits fragments de façon aléatoire par sonication (« *shearing* »). 5. Les fragments sont ensuite sélectionnés : seuls ceux dont la taille se situe entre 300 et 700 pb sont conservés. 6. Des adaptateurs P2 (représentés en forme de Y, dont la séquence de la terminaison divergente est telle qu'aucune amplification ne peut se produire en l'absence de P1) sont liés aux extrémités des fragments sélectionnés. 7. Ainsi, seuls les fragments sélectionnés incorporant P1 et P2 (*i.e.,* les fragments de 300 à 700 pb contenant les sites de restriction) sont amplifiés. 8. Les amplicons générés font l'objet d'un séquençage court (<1 kb) sur plateforme Illumina®. Les *reads* ainsi obtenus s'étendent sur quelques dizaines de paires de bases (il est techniquement possible d'obtenir aujourd'hui des *reads* de 300 pb de long), mais atteignent typiquement 100 pb (Davey *et al.*, 2013). La proportion de génome séquencée dépend aussi du protocole choisi : ici, les traits épais, adjacents aux sites de restriction, indiquent les zones de la région d'intérêt séquencées dans le cas d'un protocole *single-end* (SE RAD-seq). Les traits fins représentent les zones séquencées en supplément dans le cas d'un protocole *paired-end* (PE RAD-seq).

## — 2. *RAD-sequencing* (RAD-seq)

En s'imposant rapidement dans les études de génomique des populations menées chez les espèces sans génome de référence (Etter *et al.*, 2011 ; Narum *et al.*, 2013 ; Andrews *et al.*, 2016), le séquençage de marqueurs RAD (RAD-seq, pour *Restriction site-associated DNA sequencing*) est une méthode très représentative des changements provoqués par l'avènement des NGS.

Initialement décrit par Baird *et al.* (2008), le protocole de référence repose sur l'utilisation d'une enzyme de restriction. Après avoir soumis le génome à l'action de cette enzyme, la séquence des régions adjacentes aux sites de restriction est déterminée sur une longueur d'une centaine de nucléotides (Fig. V-1), produisant ce que l'on appelle des *reads* ou lectures. Le nombre de *reads* générés varie selon la population étudiée et l'enzyme choisie (Herrera *et al.*, 2015), suggérant la grande variété d'applications possibles du RAD-seq. Depuis son introduction, le RAD-seq a ainsi fait l'objet de nombreuses déclinaisons (Wang, S., *et al.*, 2012 ; Toonen *et al.*, 2013 ; Kess *et al.*, 2015 ; Ali *et al.*, 2016 ; Hoffberg *et al.*, 2016), ce qui permet le choix d'une stratégie ajustée à un type de projet de génotypage particulier.

Une des premières extensions de la méthode originale, déjà envisagée par Baird (2008) dans la publication princeps, a été proposée par Etter *et al.* (2011) avec le *paired-end* (PE RAD-seq). Le PE RAD-seq présente l'avantage d'augmenter la proportion de génome couverte par le séquençage, principalement du fait d'une modification des adaptateurs P2 du protocole de référence. Cette modification permet le séquençage des deux extrémités de chaque fragment, générant un couple de *reads* (Fig. V-2). Chaque région arborant un site de restriction peut donc faire l'objet d'un assemblage local formant un *contig* de plusieurs centaines de paires de bases. Au total, l'assemblage de ces *contigs* aboutit à une couverture importante du génome. Une des ses premières applications a montré, à travers l'assemblage *de novo* de 10% du génome d'un Vertébré (Willing *et al.*, 2011), l'intérêt du PE RAD-seq pour l'analyse de gros génomes eucaryotes.

En offrant la possibilité de générer rapidement un large nombre de marqueurs couvrant l'ensemble des régions du génome à moindre coût, le RAD-seq présente des atouts incontestables que ce soit pour des projets de reséquençage chez des espèces-modèles (ou apparentées à une espèce-modèle, voir par exemple Nevado *et al.*, 2014), ou bien chez des espèces dépourvues de référence (Willing *et al.*, 2011). Chez les poissons, les travaux s'appuyant sur des protocoles de RAD-seq ont trouvé un large écho, avec des résultats marquants, aussi bien chez les populations naturelles (Hohenlohe *et al.*, 2010 ; Willing *et al.*, 2011 ; Wagner *et al.*, 2013) qu'en aquaculture (Gonen *et al.*, 2014 ; Ao *et al.*, 2015). Les populations présentant un intérêt économique sur l'un ou l'autre continent sont nombreuses et ont souvent pour point commun la faible densité des cartes génétiques disponibles, initiées à partir d'un

nombre réduit de marqueurs (*e.g.,* quelques microsatellites ou AFLP, et éventuellement quelques SNP). Un génotypage par PE RAD-seq permet d'augmenter de façon substantielle la couverture de ces génomes, afin d'effectuer par exemple des travaux de cartographie fine de QTL pour des caractères d'intérêt agronomique (Wang, L., *et al.*, 2015).

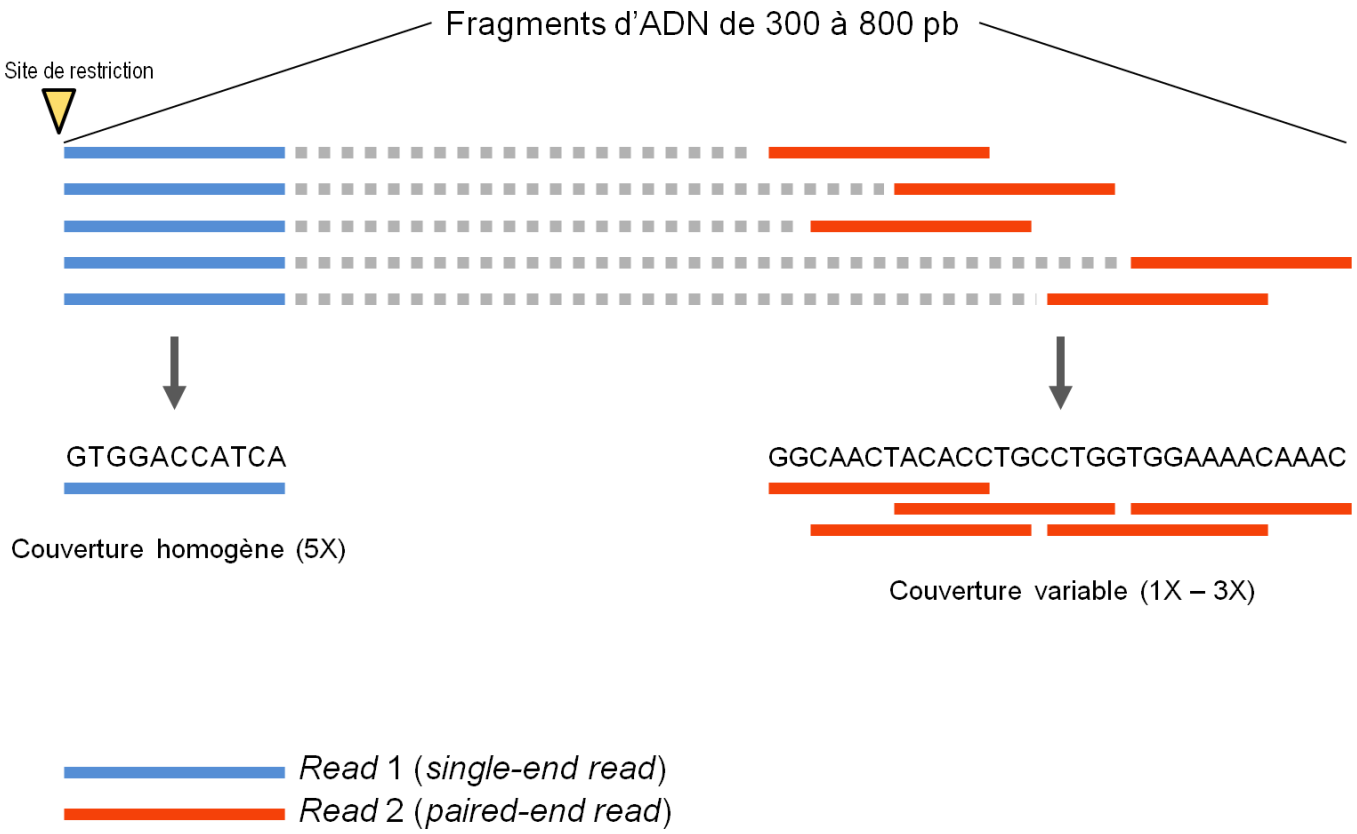



**Figure V-2. Principe général du *paired-end* RAD-seq (PE RAD-seq) (modifié d'après Catchen Lab, 2017).** Le *PE* RAD-seq génère deux *reads* de même longueur par fragment. Dans le sens *forward*, le *read* 1 (*single-end read)* débute au niveau du site de restriction. Dans le sens *reverse*, le *read* 2 (*paired-end read*) débute lui à une position déterminée par le découpage aléatoire, à l'autre bout du fragment. Dans cet exemple, la séquence des *reads* 1 est couverte 5 fois, tandis que les *reads* 2 s'échelonnent sur une région plus large. L'assemblage des *reads* 2 permet de former un *contig*, qui ne bénéficie pas d'une couverture d'aussi bonne qualité que la séquence adjacente au site de restriction, mais qui permet de cartographier une plus grande fraction du génome qu'avec les seuls *reads* 1. Cette stratégie permet de révéler un grand nombre de marqueurs, tout en améliorant la couverture du génome, et se révèle particulièrement intéressante chez les espèces disposant de cartes génétiques de faible densité.

## — 3. Sources d'erreur affectant les protocoles de RAD-seq

Le RAD-seq facilite la découverte de SNP et permet la représentation réduite d'un génome à partir d'un procédé simple dans son principe, rapide et personnalisable. En d'autres termes, cela signifie que l'étape d'acquisition des données n'est plus forcément limitante. Dès lors, les enjeux basculent dans le domaine de la prédiction des résultats attendus et du traitement des données générées, et il est pour cela nécessaire de prendre en compte certaines spécificités liées au RAD-seq.

Tout d'abord, comme tout protocole reposant sur les NGS, le RAD-seq est sujet à un taux d'erreur de séquençage non négligeable, principalement du fait de substitutions (Minoche *et al.*, 2014). De plus, l'étape d'amplification peut, de façon aléatoire, engendrer une meilleure représentation d'un allèle au détriment d'un autre. Ce phénomène semble fréquent et pourrait affecter plus d'un tiers des *reads* bruts (Schweyen *et al.*, 2014). D'autre part, la profondeur de séquençage (*i.e.,* le nombre de fois qu'une séquence est couverte par le séquençage) peut varier d'un allèle à l'autre. Un cas extrême est celui des allèles nuls, c'est-à-dire les allèles bien présents dans l'échantillon mais échappant au génotypage. C'est ce qui se produit lorsqu'un polymorphisme affectant le site de restriction est en DL avec un allèle particulier : l'enzyme ne se lie pas à l'ADN au site de restriction muté et l'allèle associé à la mutation n'est pas séquencé. Ce phénomène, désigné par l'expression *Allelic Dropout* (ADO), aboutit à considérer comme homozygotes des génotypes en réalité hétérozygotes. L'effet de l'ADO semble surtout perceptible dans les populations présentant un $N_e$ élevé (Gautier *et al.*, 2013). En outre, il peut y avoir d'autres causes à l'amplification préférentielle de certains allèles (Davey *et al.*, 2013), par exemple du fait de biais liés à l'étape de sonication (il y aurait une corrélation entre la longueur des fragments de restriction et la profondeur de séquençage pour les fragments de moins de 10 kb, *cf.* Davey *et al.*, 2013) ou bien à la composition en nucléotides GC de la séquence (la profondeur de séquençage dépend du taux en GC des fragments, sous l'influence du nombre de cycles d'amplification, *cf.* Davey *et al.*, 2013 et DaCosta & Sorenson, 2014).

Chaque combinaison modèle biologique/protocole va présenter une sensibilité différente à ces sources d'erreur. Des revues récentes compilent les informations à connaître lors de la planification d'un génotypage par RAD-seq (Herrera *et al.*, 2015 ; Andrews *et al.*, 2016). De plus, des outils d'analyse *in silico*, utilisables en amont du génotypage, voient aussi le jour (Lepais & Weir, 2014 ; Herrera *et al.*, 2015 ; Mora-Márquez *et al.*, 2017). Le développement de ce type d'outils facilite les prédictions et fournit donc une aide à l'optimisation du design expérimental.

## — 4. Le pipeline *Stacks*

Au paragraphe précédent, nous avons évoqué les principales sources d'erreur connues pouvant affecter une expérience de RAD-seq, ainsi que les appuis actuellement disponibles (revues, programmes) pour appréhender certaines situations en amont du séquençage. Compte tenu des conséquences potentielles de ces erreurs sur l'estimation de statistiques en aval, la phase de traitement des données brutes post-séquençage est clairement une étape-clé dans un projet impliquant un génotypage par RAD-seq.

Exploiter et organiser de façon optimale l'information brute demande du temps et des compétences en bioinformatique (démultipexage, détection des duplicats PCR, construction d'un catalogue de locus *de novo*, identification des allèles rares…), surtout en l'absence de ressources génomiques sur lesquelles s'appuyer. L'avènement des protocoles de RAD-seq a donc logiquement été accompagné du développement des analyses bioinformatiques requises. Plusieurs laboratoires ont proposé des pipelines destinés à faciliter le traitement des *reads* bruts (Catchen *et al.*, 2011 ; Chong *et al.*, 2012 ; Eaton *et al.*, 2014 ; Puritz *et al.*, 2014 ; Sovic *et al.*, 2015).

Parmi eux, *Stacks* (Catchen *et al.*, 2011 ; 2013) s'est imposé comme la solution la plus largement adoptée (la publication initiale de Julian Catchen a été citée près de 700 fois entre 2011 et fin 2017). *Stacks* est un outil présentant deux atouts majeurs pour la détection du polymorphisme disponible au sein d'échantillons génotypés par RAD-seq. Tout d'abord, il fédère une communauté très active autour de l'optimisation du pipeline et de son utilisation. Depuis sa présentation dans Catchen *et al.* (2011), *Stacks* est maintenu (correction de bugs, amélioration de l'efficacité, implémentation de fonctions nouvelles…) sur un rythme mensuel. L'autre atout majeur du pipeline est son principe de fonctionnement, très différent d'une simple « boîte noire », c'est-à-dire qu'il permet à l'utilisateur de faire varier plusieurs paramètres-clefs à chaque étape contrôle des *reads*, et d'en mesurer l'impact sur les statistiques de sortie.

# — 5. Références


Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. Genetics, 202(2), 389-400.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17(2), 81-92.

Ao, J., Li, J., You, X., Mu, Y., Ding, Y., Mao, K., ... & Chen, X. (2015). Construction of the high-density genetic linkage map and chromosome map of large yellow croaker (Larimichthys crocea). International journal of molecular sciences, 16(11), 26237-26248.

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses, 6(1), 106-136.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). *Stacks*: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1(3), 171-182.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

Chong, Z., Ruan, J., & Wu, C. I. (2012). Rainbow: an integrated tool for efficient clustering and assembling RAD-seq *reads*. *Bioinformatics*, *28*(21), 2732-2737.

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS One*, *9*(9), e106713.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12(7), 499-510.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. Molecular ecology, 22(11), 3151-3164.

Eaton, D. A. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, btu121.

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1), 1-15.

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., & Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing *reads*. PloS one, 6(4), e18561.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. Trends in Genetics, 28(7), 342-350.

Fontanges, Q., De Mendonca, R., Salmon, I., Le Mercier, M., & D'Haene, N. (2016). Clinical Application of Targeted Next Generation Sequencing for Colorectal Cancers. International Journal of Molecular Sciences, 17(12), 2117.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology, 22(11), 3165-3178.

Gonen, S., Lowe, N. R., Cezard, T., Gharbi, K., Bishop, S. C., & Houston, R. D. (2014). Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. BMC genomics, 15(1), 166.

Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting RAD-seq marker numbers across the eukaryotic tree of life. Genome biology and evolution, 7(12), 3207-3225.

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. Molecular Ecology Resources, 16(5), 1264-1278.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet, 6(2), e1000862.

Ibaraki, H., Wu, X., Uji, S., Yokoi, H., Sakai, Y., & Suzuki, T. (2015). Transcriptome analysis of vertebral bone in the flounder, Paralichthys olivaceus (Teleostei, Pleuronectiformes), using Illumina sequencing. Marine genomics, 24, 269-276.

Illumina (2017). Annual report for the U.S. Securities and Exchange Commission. Illumina, Inc.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. Molecular Ecology, 25(1), 185-202.

Kess, T., Gross, J., Harper, F., & Boulding, E. G. (2016). Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle Littorina saxatilis. Journal of Molluscan Studies, 82(1), 104-109.

Lepais, O., & Weir, J. T. (2014). SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Molecular ecology resources, 14(6), 1314-1321.

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome biology, 12(11), R112.

Mora-Márquez, F., García-Olivares, V., Emerson, B. C., & López de Heredia, U. (2017). ddradseqtools: a software package for in silico simulation and testing of double-digest RADseq experiments. Molecular ecology resources, 17(2), 230-246.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. Molecular ecology, 22(11), 2841-2847.

Nelson, I., Stojkovic, T., Allamand, V., Leturcq, F., Bécane, H. M., Babuty, D., ... & Eymard, B. (2015). Laminin α2 Deficiency-Related Muscular Dystrophy Mimicking Emery-Dreifuss and Collagen VI related Diseases. Journal of Neuromuscular Diseases, 2(3), 229-240.

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. Molecular ecology, 23(7), 1764-1779.

NHGRI (National Human Genome Research Institute, 2016). The Cost of Sequencing a Human Genome [En Ligne] https://www.genome.gov/sequencingcosts/ (consulté le 14/03/2018).

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ, 2, e431.

Schmidt, B., & Hildebrandt, A. (2017). Next-generation sequencing: big data meets high performance computing. Drug Discovery Today.

Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. The Biological Bulletin, 227(2), 146-160.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology, 26(10), 1135-1145.

Sovic, M. G., Fries, A. C., & Gibbs, H. L. (2015). AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. *Molecular ecology resources*, *15*(5), 1163-1171.

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ, 1, e203.

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J., Huvenaars, K. H., ... & van Eijk, M. J. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One, 7(5), e37565.

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., ... & Seehausen, O. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular ecology, 22(3), 787-798.

Wang, L., Wan, Z. Y., Bai, B., Huang, S. Q., Chua, E., Lee, M., ... & Sun, F. (2015). Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. Scientific reports, 5, 16358.

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature methods, 9(8), 808-810.

Willing, E. M., Hoffmann, M., Klein, J. D., Weigel, D., & Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics, 27(16), 2187-2193.

# Chapitre VI

# TABLE DES MATIERES

-

# Chapitre VI – Détection de signatures de sélection pour l'efficacité alimentaire chez la truite arc-en-ciel

## 1. Introduction

### 1.1. L'efficacité alimentaire : un caractère aux enjeux prégnants et étendus

Les questions associées à la durabilité des systèmes de production agricole s'imposent à la recherche agronomique comme des champs d'investigation prioritaires. On estime en effet que les activités humaines consomment aujourd'hui près de 30% de la biomasse aérienne primaire annuellement disponible (Haberl *et al.*, 2007 ; Allaire & Daviron, 2017) et émettent, majoritairement du fait de l'agriculture, plus d'azote réactif que l'ensemble des écosystèmes (plantes cultivées incluses) ne peuvent en capter (Galloway, 2008 ; Townsend & Howard, 2010 ; Peyraud *et al.*, 2012). Les impacts économiques et environnementaux défavorables des flux de matière imputables à l'agriculture sont parfois difficiles à quantifier mais de plus en plus largement documentés (Volk *et al.*, 2006 ; Atzori *et al.*, 2012 ; Davidson *et al.*, 2012 ; Herrero *et al.*, 2013 ; Besson *et al.*, 2016 ; Reichwaldt *et al.*, 2016), ce qui devrait permettre d'identifier des priorités pour l'action publique (Peyraud *et al.*, 2012 ; Allaire & Daviron, 2017). Ces enjeux s'accompagnent de réflexions sociales et éthiques de plus en plus poussées (Kaiser & Algers, 2016 ; Macdiarmid *et al.*, 2016), au moment où l'on prévoit que la consommation alimentaire de l'humanité représentera en 2050 près du double de ce qu'elle était en 2010 (von Braun, 2010).

Les communautés scientifiques des productions animales se positionnent par rapport à ces grandes questions en organisant leur communication et leurs travaux autour de thématiques comme l'élevage de précision ou l'agroécologie (GIS Elevages demain, 2017 ; INRA, 2017). Dans ce cadre, la génétique est perçue comme une discipline susceptible de fournir des solutions pérennes afin d'accompagner les systèmes d'élevage vers plus d'efficience et de durabilité. Ces objectifs ont été bien identifiés par les généticiens, qui ont vu en l'avènement de la génomique le saut technologique qui permettrait l'exploration détaillée de caractères complexes dont l'amélioration est souhaitable dans les populations animales d'intérêt agronomique (Institut de l'Elevage & INRA, 2011). Ainsi, certains travaux de recherche actuels s'appuient sur les NGS pour mieux comprendre l'architecture moléculaire de caractères liés à la santé animale ou à l'efficacité alimentaire (ABCIS & INRA, 2015).

L'efficacité alimentaire est un concept complexe se rapportant à l'efficacité d'utilisation de l'aliment en élevage. Elle peut être vue comme l'analogue d'un rendement qui caractériserait l'efficacité de transformation de l'aliment en produit à travers l'animal. Améliorer l'efficience des populations d'élevage passe entre autres par un progrès génétique sur ce caractère qui constitue un potentiel objectif de sélection commun à de nombreux acteurs, puisque des marges de progrès non négligeables ont été identifiées chez les principales espèces d'élevage (lapin de chair : Garreau *et al.*, 2008 ; volailles : Bordas *et al.*, 1992 ; poissons : Grima, 2010 ; porc : Gilbert, H., *et al.*, 2017 ; bovins : Brochard *et al.*, 2013 ; petits ruminants : Phocas *et al.*, 2013).

S'il existe des attentes évidentes autour de l'efficacité alimentaire, son amélioration génétique reste une question complexe qui se heurte aujourd'hui à des limites techniques. D'une part, nous avons affaire ici à un caractère polygénique et intégratif (Gilbert, H., *et al.*, 2017), formé de plusieurs composantes plus ou moins corrélées (Mauch *et al.*, 2017) et impliquant plusieurs processus biologiques majeurs (Phocas *et al.*, 2014). Mais, surtout, un enjeu important réside dans la mesure des quantités d'aliment ingérées par chaque individu, qui est une composante essentielle de l'efficacité alimentaire. Obtenir une mesure précise, peu coûteuse, répétable et extensible à un grand nombre d'individus de la quantité d'aliment ingérée représente une difficulté technique majeure pour l'ensemble des filières (Phocas *et al.*, 2014), même dans le cas de populations expérimentales (Aggrey & Rekaya, 2013 ; Brochard *et al.*, 2013). Ainsi, la mise en place d'une sélection sur l'efficacité alimentaire au sein de populations commerciales nécessiterait l'utilisation de prédicteurs fiables, permettant de s'affranchir de la mesure de l'ingéré individuel. Cependant, peu de marqueurs spécifiques de l'efficacité alimentaire sont actuellement disponibles, même dans les populations de monogastriques comme chez le porc ou le poulet de chair où le caractère a été bien étudié (Rotschild *et al.*, 2007 ; Gilbert, H., *et al.*, 2017 ; Sell-Kubiak *et al.*, 2017).

## 1.2. L'efficacité alimentaire en aquaculture

Le secteur aquacole est appelé à jouer un rôle important dans la fourniture de protéines pour l'alimentation humaine alors que la sécurité alimentaire est peut-être l'enjeu politique qui résume le mieux les grandes questions publiques évoquées au paragraphe précédent (FAO, 1996). Sous l'effet de l'augmentation de la demande mondiale, la production aquacole croît rapidement et devrait dépasser en volume l'ensemble des captures de pêche d'ici 2025 (OCDE/FAO, 2016). En particulier, la proportion d'espèces nourries continentales est en augmentation continue depuis le milieu des années 90, et représente déjà aujourd'hui 40% du volume de production total en aquaculture (FAO, 2016). Même si les poissons sont généralement considérés comme des convertisseurs d'énergie plus efficaces que les espèces homéothermes (Gjedrem *et al.*, 2012), ce sont surtout les herbivores comme le tilapia ou la

carpe qui offrent les meilleures performances dans ce domaine (Brown *et al.*, 2006). Le grand succès sur les marchés européens et nord-américains des poissons d'élevage carnivores comme les salmonidés appelle à identifier et actionner des leviers d'amélioration de l'efficience de ces productions (Aubin *et al.*, 2009 ; Le Boucher *et al.*, 2011).

La variabilité des caractéristiques physiologiques des plus de deux-cents espèces d'élevage (Naylor *et al.*, 2000), installées dans des bâtiments situés la plupart du temps en extérieur et nécessitant le maintien d'une eau de qualité suffisante, place les élevages aquacoles parmi les systèmes de production les plus complexes. L'alimentation des animaux y occasionne des dépenses particulièrement lourdes, pouvant atteindre 70% du coût de production total (de Verdal *et al.*, 2017). Chez la truite arc-en-ciel (*Oncorhynchus mykiss*), le coût de l'aliment, composé de 30% à 40% de ressources issues de la pêche minotière (CIPA, 2017), est estimé à 50% du coût de production total (Hardy & Barrows, 2002). Au plan environnemental, la question de l'impact de la consommation de ces ressources s'ajoute à celle de la gestion des effluents. En particulier, les déchets azotés (incluant les pertes dues aux aliments non consommés) issus des systèmes d'élevage de poissons carnivores comme *O. mykiss* sont responsables de l'eutrophisation des écosystèmes aquatiques (Westhoek *et al.*, 2011). Une meilleure gestion de l'aliment est susceptible de réduire le volume d'intrants coûteux en ressources naturelles et en énergie, mais aussi la charge nutritive des eaux d'élevage, liée à l'eutrophisation (Aubin *et al.*, 2009). En d'autres termes, une meilleure efficacité d'utilisation de l'aliment en aquaculture procurerait un bénéfice commun aux plans économiques et environnementaux. Il n'est donc pas étonnant de constater que l'efficacité alimentaire soit clairement identifiée par les aquaculteurs comme un trait dont l'amélioration est prioritaire (Sae-Lim *et al.*, 2012).

Dans ce contexte, de nombreux travaux (Grima *et al.*, 2008 ; Kause *et al.*, 2008 ; Aubin *et al.*, 2009 ; Grima, 2010 ; Grima *et al.*, 2010a, 2010b ; Besson *et al.*, 2016 ; Janssen *et al.*, 2017) soulignent les impacts positifs à attendre de l'amélioration génétique de l'efficacité alimentaire des poissons d'aquaculture. En couplant modélisation et Analyse de Cycle de Vie (ACV), Besson *et al.* (2016) ont pu mesurer que la sélection d'une meilleure efficacité alimentaire permettait invariablement d'améliorer l'efficacité de la production chez le poisson-chat africain (*Clarias gariepinus*). Leurs travaux ont montré que les gains économiques engendrés s'accompagnaient d'une réduction marquée des impacts environnementaux par tonne de biomasse de poisson produite. Ces résultats confirment l'intérêt de la mise en place d'une sélection sur l'efficacité alimentaire en aquaculture. Par ailleurs, il est admis que l'héritabilité de ce caractère est suffisamment importante dans les populations d'élevage pour y permettre sa sélection (Quinton *et al.*, 2007 ; Kause *et al.*, 2008 ; Grima, 2010).

Malgré cela, les déterminants de l'efficacité alimentaire demeurent très mal connus chez les poissons d'aquaculture car la mesure de ce caractère intégratif, déjà difficile à appréhender chez les Vertébrés terrestres, se complique chez les poissons (Grima, 2010). En effet, le comportement alimentaire des poissons dépend de beaucoup de variables externes (Fletcher, 1984), et notamment des relations sociales qui s'établissent entre individus au sein des groupes chez les salmonidés (McCarthy *et al.*, 1992). Accéder à l'ingéré individuel dans ce contexte, c'est-à-dire en milieu aquatique avec un aliment distribué de façon collective, est une tâche compliquée (Madrid *et al.*, 1997). Certaines approches visant à estimer cet ingéré individuel ont été développées (Kause *et al.*, 2006 ; Silverstein, 2006 ; de Verdal *et al.*, 2017), mais ont pour point commun de s'appuyer sur une logistique lourde (élevage en aquarium individuel, recours à l'imagerie ou à la vidéo nécessitant de nombreuses mesures individuelles) et parfois source de potentiels biais. Mesurer l'efficacité alimentaire en routine sur des milliers de poissons en passant par la mesure de l'ingéré individuel est donc clairement inenvisageable.

Pour contourner cette difficulté et mieux comprendre les mécanismes physiologiques et la base génétique qui gouvernent la variabilité de l'efficacité alimentaire, des approches indirectes ont été mises en place, lesquelles visent à étudier des caractères corrélés à des différences d'efficacité d'utilisation de l'aliment et plus faciles à explorer. Sont notamment étudiés des caractères impliqués dans la stratégie d'allocation des ressources énergétiques, comme la perte de poids au cours du jeûne, représentative des besoins de maintenance (Grima *et al.*, 2010a), la capacité de croissance compensatrice en période de réalimentation, dépendante de l'activité métabolique (Grima *et al.*, 2008) et la distribution des nutriments dans les différents compartiments corporels, en particulier les lipides (Quillet *et al.*, 2005 ; Grima *et al.*, 2010a), dont on sait qu'elle est une composante importante de l'efficacité globale de transformation de l'aliment.

En somme, les enjeux associés au contrôle de l'efficacité alimentaire sont particulièrement tangibles chez les poissons d'aquaculture, avec des implications fortes en termes économiques, environnementaux, zootechniques et scientifiques. Mieux comprendre le déterminisme génétique de ce caractère passe, au moins probablement dans un premier temps, par l'utilisation de mesures indirectes simples, peu invasives et répétables.

## 1.3. EFFICACE : un dispositif expérimental intéressant pour étudier l'efficacité alimentaire

Parmi les approches indirectes permettant d'estimer l'efficacité alimentaire, la mesure de la teneur en lipides du muscle est particulièrement intéressante. Tout d'abord, la littérature indique une héritabilité significative (avec des estimations comprises entre 0,12 et 0,72) des traits se rapportant à l'adiposité

corporelle chez les salmonidés (Quillet *et al.*, 2005 ; Quinton *et al.*, 2007 ; Kause *et al.*, 2016). Il a de plus a été mis en évidence que le contenu en lipides du muscle était génétiquement corrélé à l'efficacité alimentaire chez la truite arc-en-ciel ($r_g$ = 0,68 ± 0,24, Kause *et al.*, 2016). D'autre part, la mesure de ce phénotype à l'échelle individuelle ne pose pas problème : il est possible d'utiliser une approche simple, précise et non invasive, compatible avec le traitement d'un grand nombre de poissons (Quillet *et al.*, 2005). Explorer le déterminisme génétique de la teneur en lipides des muscles serait ainsi un bon moyen de mieux comprendre les mécanismes impliqués dans le contrôle de l'efficacité alimentaire, mais aussi d'en identifier des prédicteurs génétiques qui font aujourd'hui défaut. Nous nous sommes engagés dans cette voie en recherchant des signatures de sélection pour la teneur en lipides musculaires chez la truite *O. mykiss* dans le cadre du projet EFFICACE.

Le projet EFFICACE (pour EFFICacité Alimentaire : allocation des ressourCes Energétiques) fait l'objet d'une collaboration entre l'UMR Intrépid (B. Chatain) et notre unité. L'objectif est de permettre l'identification de souches présentant une meilleure efficacité alimentaire afin d'améliorer durablement ce caractère chez les poissons d'aquaculture. Participer à ce projet nous permet en particulier d'exploiter des données de génotypage acquises au terme d'une expérience de sélection divergente menée chez la truite *O. mykiss* et débutée il y a plus de 20 ans au sein d'installations expérimentales de l'INRA. Les données génomiques collectées dans le cadre de cette expérience de sélection se prêteraient bien à l'utilisation de notre méthode de détection de signatures de sélection. En effet, les simulations ont montré que notre méthode offrait de bonnes performances sur les petites populations soumises à une forte sélection sur une dizaine de générations (*cf.* Chapitre III). L'expérience de sélection menée sur les truites dans le cadre du projet EFFICACE correspond en tout point à cette description. La démographie est contrôlée : il n'y a pas de chevauchement entre générations, pas de migration, et la gestion des reproducteurs suggère *a priori* un effectif efficace ni trop faible ni trop élevé, de l'ordre d'une centaine d'individus. La sélection imposée est intense, avec une pression de sélection d'environ 10%, ce qui correspond aux ordres de grandeur testés avec succès par simulation. Par ailleurs, d'après les travaux de cartographie menés chez la truite arc-en-ciel, la taille minimale du génome est de l'ordre de 3000 cM (Young *et al.*, 1998 ; Rexroad & Vallejo, 2009 ; Palti *et al.*, 2011). Dans le cadre d'un scénario de génotypage raisonnable où nous disposerions de 15000 SNP informatifs, la densité moyenne serait donc de cinq marqueurs par cM. Les estimations menées au sein de populations américaines rapportent un DL important à cette échelle (Rexroad & Vallejo, 2009). Tout cela suggère la possibilité d'identifier des empreintes laissées par la sélection avec notre méthode de vraisemblance dans le cadre du projet EFFICACE.

Toutefois, comme nous le verrons dans le détail au cours des sections suivantes, il a été nécessaire de tenir compte de certaines particularités du dispositif expérimental. En premier lieu, un des risques

inhérents aux expériences de sélection menées plusieurs années durant sur des animaux installés en extérieur comme les poissons d'aquaculture est que ceux-ci peuvent être exposés à des perturbations inattendues. Ce type de situation s'est présenté dans le cadre du projet EFFICACE : la population Témoin de quatrième génération (G3) a subi un aléa expérimental avant que les individus n'aient pu être reproduits, ce qui a abouti à la perte de la lignée Témoin. Pour faire face à cet imprévu, la lignée Témoin a été répliquée en s'appuyant sur une cohorte de truites issue de la même souche et maintenue dans les mêmes conditions que la lignée Témoin initiale. Un autre élément important à prendre en compte est la stratégie d'acquisition des données génomiques, obtenues via un génotypage par séquençage. En effet, le polymorphisme disponible au sein des populations n'a pas été identifié au travers d'une puce à ADN dédiée à l'espèce étudiée, mais par séquençage de marqueurs RAD (*cf.* Chapitre V). Enfin, nous souhaitons effectuer ici une analyse de signatures de sélection tirant parti de la dynamique des fréquences alléliques au cours du temps. Une telle analyse nécessite évidemment d'avoir accès à des données longitudinales, c'est-à-dire des échantillons génétiques distants dans le temps. Cependant, dans le cas présent, le génotypage a été effectué pour plusieurs populations mais au même stade de l'expérience, après sept générations de sélection. Dans l'idéal, nous aurions espéré bénéficier en plus des génotypes parentaux (*i.e.,* en G0), mais aucun échantillon de la population parentale n'a été conservé.

# — 2. Matériels et méthodes

## 2.1. Evolution expérimentale et reséquençage de populations de truite INRA

### 2.1.a. Sélection divergente sur la teneur en lipides musculaires

Une expérience de sélection sur la teneur en lipides musculaires chez la truite arc-en-ciel a été initiée en 1995 au sein d'unités expérimentales INRA, en vue de tester la faisabilité et l'effet d'une sélection massale sur ce type de dépôts corporels. Deux lignées divergentes pour le caractère, bien corrélé à l'efficacité alimentaire chez les salmonidés, ont ainsi été dérivées à partir d'une population de truites issue de la souche INRA à ponte printanière (INRA-PRN) (Fig. VI-1).

A chaque génération, entre 800 et 1600 individus ont été élevés pour chaque lignée. A l'âge d'un an, les poissons sont mesurés individuellement pour le taux lipidique du muscle dorsal ajusté au poids corporel (Quillet *et al.*, 2005). Les individus affichant une mesure comprise dans la plage des 10% de valeurs extrêmes supérieures (resp. inférieures) de la distribution du caractère ont été conservés pour la reproduction de la lignée muscle gras (resp. maigre). Cette sélection divergente a été imposée pendant sept générations, au travers du croisement panmictique de 64 à 201 reproducteurs. En parallèle, une lignée Témoin (T), entretenue sans sélection, a été maintenue à partir de la même souche.

Au plan phénotypique, les deux lignées, muscle maigre (M) et muscle gras (G), diffèrent fortement pour la teneur en lipides du muscle au terme de la sélection (deux fois plus élevée chez les animaux G que chez les animaux M dès les jeunes stades). Elles se distinguent également par des localisations différentes des dépôts lipidiques globaux, la lignée M déposant davantage de lipides en zone périviscérale, ainsi que par une différence d'efficacité alimentaire en faveur de la lignée M (E. Quillet, comm. pers.).

### 2.1.b. Réplication de la lignée Témoin

Initialement, une lignée Témoin, reproduite par panmixie dans les mêmes conditions que les lignées sélectionnées, avait aussi été dérivée à partir de la même population parentale (G0). Cependant, cette lignée a été perdue en cours d'expérience. Les poissons du lot Témoin en G3 sont morts à 1 an, avant que les géniteurs ne puissent reproduire. La lignée Témoin a donc été répliquée à partir de poissons d'une cohorte issue de la même souche INRA-PRN que la population parentale. Cette cohorte était entretenue de la même façon que la lignée Témoin initiale, mais en décalage d'une demi-génération (*i.e.,* les poissons étaient reproduits un an plus tôt dans la lignée qui a permis de reconstituer un lot Témoin).

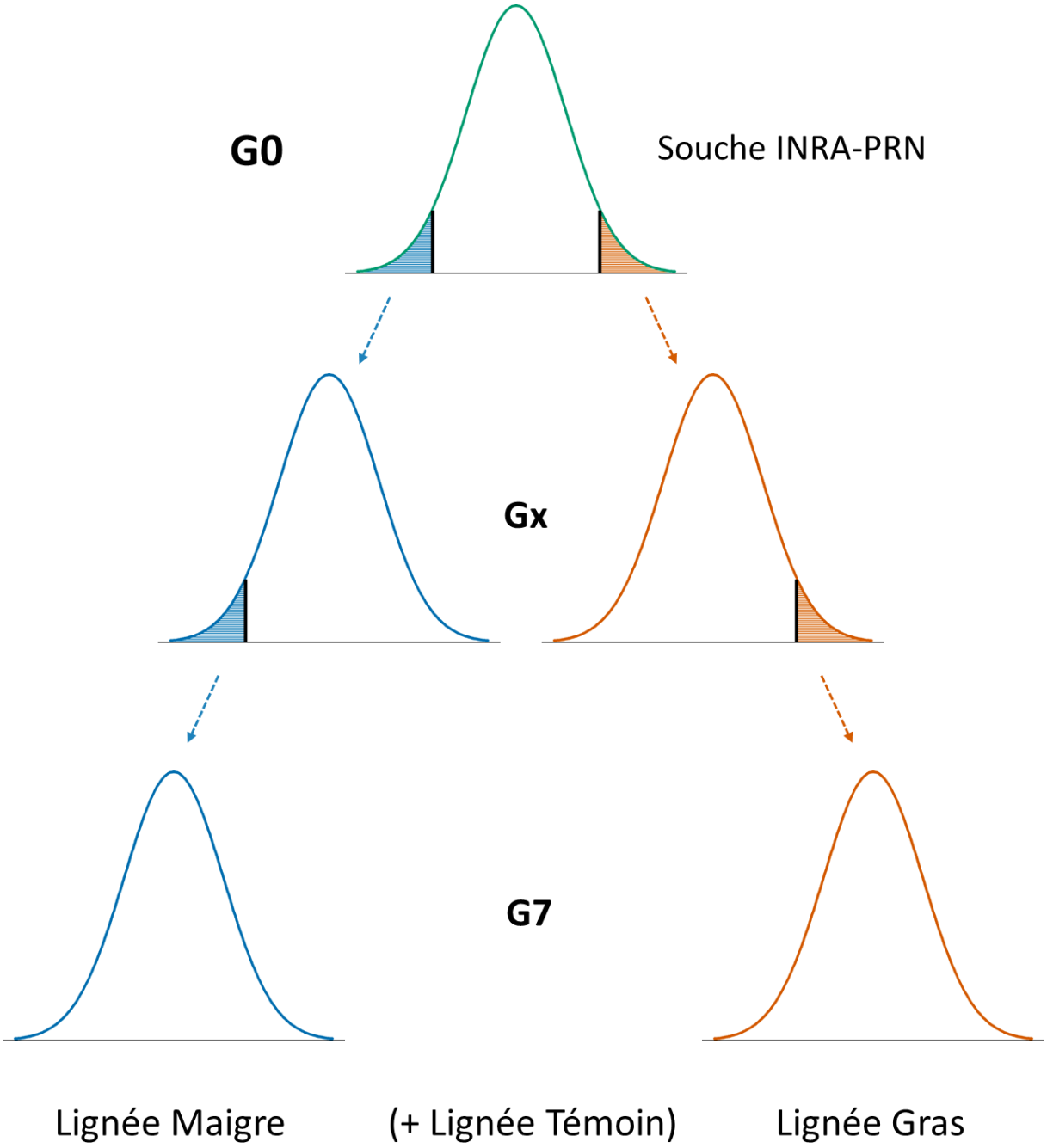



**Figure VI-1. Schéma du principe de l'expérience de sélection artificielle exploitée dans le cadre du projet EFFICACE.** Une population de truite (*Oncorhynchus mykiss*) de la souche INRA à ponte printanière (INRA-PRN) a été soumise à une sélection par troncature bidirectionnelle sur la teneur en lipides du muscle, générant deux lignées divergentes (M et G) pour ce caractère. Sept générations de sélection divergente ont pu être réalisées par croisement factoriel intra-population d'une centaine de reproducteurs. La souche INRA-PRN a par ailleurs été maintenue par panmixie sur la même période, produisant une lignée Témoin soumise uniquement à la dérive génétique.

Compte tenu de cet aléa expérimental, l'histoire démographique complète des lignées expérimentales de truite étudiées dans le cadre du projet EFFICACE peut être représentée par un arbre phylogénétique à quatre branches (Fig. VI-2). Par commodité, nous nommerons populations M, G et T les populations génotypées au terme de l'évolution expérimentale dans les lignées M, G et T, respectivement, telles que représentées par cet arbre.

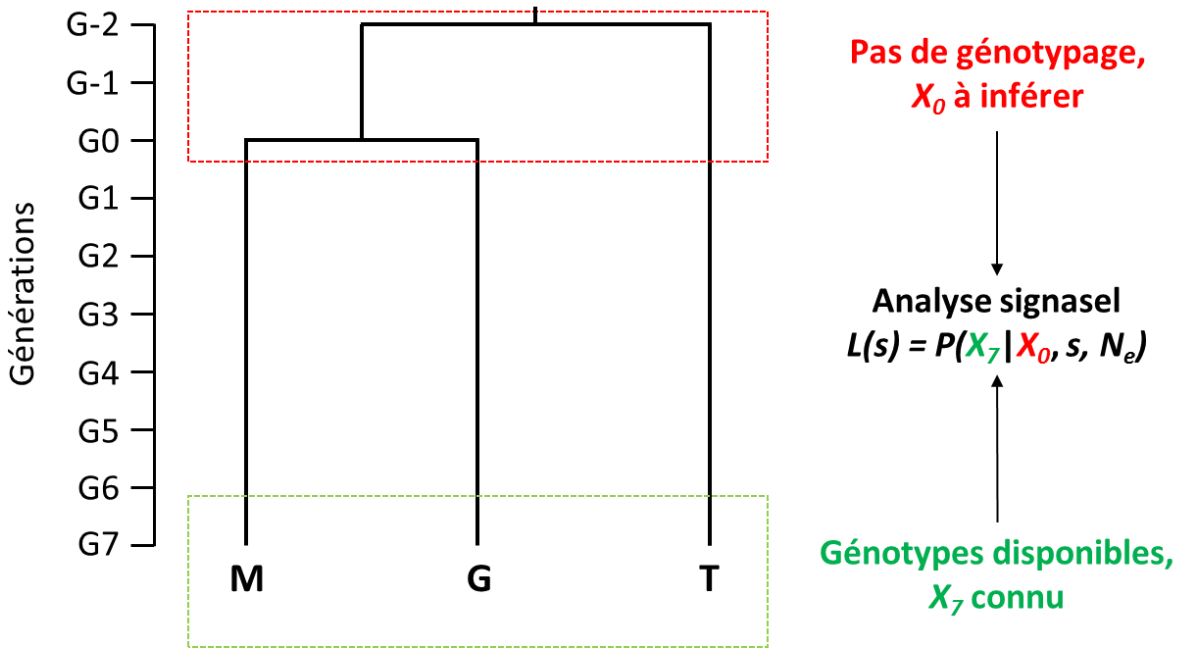



**Figure VI-2. Arbre phylogénétique des populations M, G et T d'après les informations disponibles sur l'historique des lignées.** Deux lignées sélectionnées (M et G) sont issues de la même population, tandis que la lignée T, entretenue sans sélection, dérive de la même souche. La génération G0 correspond au début de l'expérience de sélection. Seules les populations contemporaines (*i.e.,* en génération G7) ont pu être génotypées, ce qui signifie que les fréquences alléliques des populations ancestrales devront être inférées afin d'effectuer une analyse du polymorphisme sélectionné avec notre méthode temporelle de détection de signatures de sélection (« analyse *signasel* »).

### 2.1.c. Echantillonnage et génotypage

Afin d'explorer la base génétique de la divergence phénotypique constatée entre lignées M et G, un fragment de nageoire a été collecté sur 70 individus des populations M, G et T. De l'ADN a été extrait à partir des échantillons de tissu collectés. Le contrôle de la concentration et de la qualité de l'ADN extrait a permis de retenir 58 individus M, 60 individus G et 58 individus T en vue du génotypage. Ces premières étapes ont été réalisées par des techniciens en biologie moléculaire de notre laboratoire. Au total, 180 échantillons (176 individus uniques et 4 réplicats) issus de l'expérience d'évolution ont ensuite été génotypés par RAD-seq (Eurofins MWG Operon) selon la procédure décrite par Etter *et al.* (2011). Un lot de 22 échantillons, issus de 18 individus haploïdes doublés, a également été introduit dans l'analyse. Il s'agit d'individus issus de lignées clonales initiées et maintenues par gynogenèse

(Verrier, 2013). Chez la truite arc-en-ciel, il est en effet possible de provoquer expérimentalement une gynogenèse, c'est-à-dire le développement d'un zygote diploïde viable sans le matériel génétique du mâle. Totalement homozygotes, les individus HD ainsi produits facilitent la détection et l'élimination des locus paralogues résultant de la tétraploïdie résiduelle chez la truite (Palti et al., 2014).

### 2.1.d. Génération et traitement bioinformatique des *reads*

Une bibliothèque de fragments barcodés a été construite (Eurofins MWG Operon) à partir d'un protocole de PE RAD-seq (Etter *et al.*, 2011) utilisant l'enzyme de restriction *Sbf*I. Au total, le séquençage sur trois files d'une plateforme Illumina HiSeq 2000 a produit 250 millions de *reads* d'une longueur moyenne de 100 pb chez 202 échantillons (176 individus issus de l'expérience de sélection, 18 individus haploïdes doublés et 8 réplicats).

Nous avons utilisé *Stacks* v1.44 pour identifier un jeu de marqueurs à partir des séquences obtenues par PE RAD-seq. Au moment où nous avons débuté ces travaux, il n'y avait que peu de revues abordant l'impact du choix des paramètres permettant de construire un catalogue de locus, malgré la perception largement partagée de l'importance des étapes de traitement des données brutes et de leurs effets potentiels sur les analyses ultérieures. Afin d'identifier une stratégie pertinente de découverte de marqueurs en vue des analyses de génomique des populations, nous avons testé l'effet de la variation de différents paramètres de traitement des *reads* bruts, notamment avec *Stacks*, et nous nous sommes pour cela largement appuyés sur l'expertise et le travail de Maria Bernard, bioinformaticienne au sein de notre équipe, et sur la connaissance du modèle biologique d'Edwige Quillet, animatrice de l'équipe *Génétique et Aquaculture*, et coordinatrice du projet EFFICACE au sein de notre laboratoire.

Les opérations visant à révéler le polymorphisme disponible au sein de nos échantillons sont menées uniquement sur les *reads* 1, adjacents au site de restriction. Nous avons essayé d'incorporer le polymorphisme identifié via les *reads* 2, mais la faible couverture de ces derniers ne permet pas de récupérer une information suffisante en vue des analyses de génomique des populations. Les *reads* 2 seront toutefois utiles dans un but de cartographie. La stratégie de détection du polymorphisme détaillée dans cette section ne s'applique donc qu'aux *reads* 1.

Les *reads* bruts font tout d'abord l'objet d'une phase de préparation. Cette étape de pré-traitement débute par le démultiplexage, effectué à l'aide d'un script maison (M. Bernard), qui permet de rassembler les séquences propres à chaque individu. Nous avons ensuite utilisé les programmes *clone_filter* et *process_radtags* du pipeline *Stacks* pour éliminer les duplicats de PCR et les séquences de mauvaise qualité, respectivement. La combinaison de ces filtres préliminaires aboutit à l'élimination d'environ 30% des *reads* Illumina bruts.

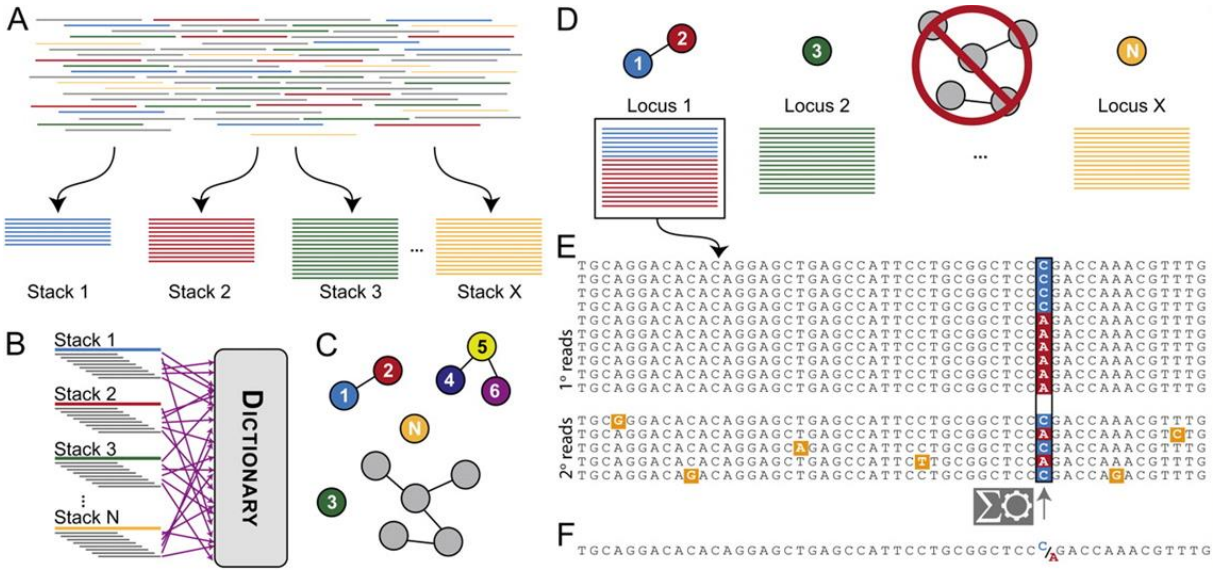



**Figure VI-3. Représentation conceptuelle du fonctionnement du programme *ustacks* (modifié d'après Catchen et al., 2011, Fig. 1).** Le programme *ustacks* est au cœur du pipeline *Stacks*. Il aligne les *reads* obtenus à l'issue de la phase de nettoyage afin d'identifier le polymorphisme spécifique à chaque individu. **(A)** Le programme *ustacks* construit des piles (*Stacks*) d'allèles putatifs pour chaque individu à partir des *reads* dont la séquence est rigoureusement identique. **(B)** L'algorithme de *ustacks* décompose la séquence de chaque pile en *k-mers* (*i.e.,* des fragments chevauchants de *k* pb), et interroge le dictionnaire de *k-mers* afin de créer une liste de piles dont la séquence est potentiellement compatible. **(C)** Les piles compatibles peuvent être vues comme les nœuds d'un graphe reliés par la distance nucléotidique qui les sépare. **(D)** *ustacks* combine les piles compatibles pour former des locus putatifs. **(E)** *ustacks* incorpore sous condition les *reads* secondaires (2° *reads*), qui avaient initialement été mis de côté, afin d'augmenter la profondeur de pile. La présence d'erreurs de séquençage est testée par maximisation de la vraisemblance des données (Hohenlohe *et al.*, 2010), afin d'attribuer en chaque site polymorphe le génotype le plus probable. **(F)** *ustacks* détermine une séquence consensus et archive les SNP et les haplotypes.

L'étape suivante occupe une place centrale dans le traitement bioinformatique des *reads* issus de RAD-seq. Il s'agit de la phase de découverte du polymorphisme à proprement parler, nécessitant de ce fait un travail attentif d'optimisation des paramètres de détection. L'alignement des *reads* nettoyés va permettre de révéler le polymorphisme propre à chaque individu. Nous avons construit *de novo,* c'est-à-dire sans nous appuyer sur une séquence de référence, un catalogue de locus rendant compte du polymorphisme disponible au sein des nos échantillons. Les marqueurs candidats à la sélection seront alignés ultérieurement sur l'assemblage du génome de la truite arc-en-ciel.

Avec *Stacks*, l'identification *de novo* d'allèles et de locus putatifs intra-individus est dévolue au programme *ustacks* (pour *unique Stacks*). Son principe repose sur la constitution de piles (*stacks* en anglais) de séquences identiques et suffisamment nombreuses pour être considérées comme des allèles, avant de rassembler les piles susceptibles d'appartenir à un même locus (Fig. VI-3). L'utilisateur contrôle certains critères permettant de définir les allèles et les locus qui seront conservés par *ustacks*.

Ce faisant, *ustacks* implémente de façon automatique plusieurs algorithmes correctifs chargés de distinguer les erreurs de séquençage et les séquences répétées du polymorphisme d'un locus.

A ce stade, trois principaux paramètres définis par l'utilisateur affectent fortement l'estimation du polymorphisme disponible : la profondeur minimale de pile (*m*), la distance maximale autorisée entre les piles (*M*) et la distance maximale autorisée pour l'incorporation des *reads* secondaires (*N*). L'influence du premier cité (*m*) ne doit pas être sous-estimée car elle peut se révéler sensible sur l'ensemble des traitements ultérieurs. Le paramètre *m* renseigne le critère de définition d'un allèle putatif, c'est-à-dire le nombre de séquences identiques requises pour que l'information qu'elles portent soit conservée. Les séquences conservées sont nommées *reads* primaires et sont utilisées pour rechercher des locus polymorphes. Les séquences dont la profondeur est insuffisante constituent les *reads* secondaires, qui sont mis de côté en attendant de voir si elles peuvent malgré tout être utilisées. Nous avons souhaité un minimum de trois séquences identiques pour ce paramètre de profondeur (*m* = 3). Ce choix représente un compromis raisonnable entre le risque d'incorporer des erreurs dans les séquences retenues et celui d'introduire un biais dans la détermination des génotypes.

Un seuil plus strict (*m* = 5) aurait une incidence défavorable sur la capacité de *Stacks* à rendre compte du polymorphisme observé. En effet, changer *m* ne modifie pas le nombre total de marqueurs (moins de 2% de marqueurs supplémentaires sont identifiés avec *m* = 3 par rapport au cas où *m* = 5) mais la proportion d'individus génotypés pour chaque marqueur (Fig. VI-4). Ces marqueurs supplémentaires ne présentent pas vraiment d'intérêt, car on les retrouve chez relativement peu d'individus et ils seront pour la plupart éliminés lors des étapes ultérieures de tri. Cependant, être moins strict sur *m* garantit une meilleure couverture en individus pour les marqueurs intéressants (Fig. VI-4). Par exemple, un peu plus de 27% des marqueurs sont génotypés chez au moins 55 individus dans chaque population lorsque *m* = 3, contre seulement 7,5% si *m* = 5, soit un écart de vingt points en faveur d'un critère assoupli (*m* = 3). Il y a donc moins de données manquantes sur les marqueurs qui seront *in fine* retenus avec *m* = 3 plutôt qu'avec *m* = 5. Par ailleurs, plus d'un marqueur sur deux identifié avec *m* = 5 présente un déficit en hétérozygotes incompatible avec les proportions de Hardy-Weinberg. Ainsi, un paramètre *m* trop strict est susceptible de biaiser la représentation du polymorphisme disponible.

Un second critère important est le paramètre *M*, qui définit la distance maximale autorisée (en nombre de nucléotides) entre chaque paire de piles retenues (*i.e.,* chaque couple d'allèles putatifs) pour considérer qu'elles appartiennent à un locus commun. Comme ce processus de combinaison de piles est itératif, la distance totale combinée pour un locus peut être supérieure à la valeur du paramètre *M*. Nous avons choisi de conserver la valeur par défaut de ce paramètre (*M* = 2), qui permet de repérer

des locus multialléliques (environ 15% des marqueurs identifiés ont au moins trois allèles) mais aussi d'éviter le regroupement injustifié d'allèles.

Enfin, *ustacks* permet de définir via le paramètre *N* un critère d'acceptation des *reads* secondaires. Il définit une distance maximale entre les *reads* primaires (*i.e.,* les piles déjà formées) et les *reads* secondaires pour accepter ou non d'aligner ces derniers sur les *reads* primaires. L'objectif est d'améliorer la couverture (nombre de *reads*) de certains allèles et, *in fine*, le génotypage. Cependant, nous avons pu vérifier que l'ajout des *reads* secondaires générait plus de bruit que d'information utile. Nous avons donc choisi d'ignorer les *reads* secondaires (*N* = 0). L'ensemble des opérations effectuées à l'aide du programme *ustacks* permet de valider un ensemble de locus, chacun étant formé d'une séquence consensus et incluant éventuellement des haplotypes, pour chaque individu.

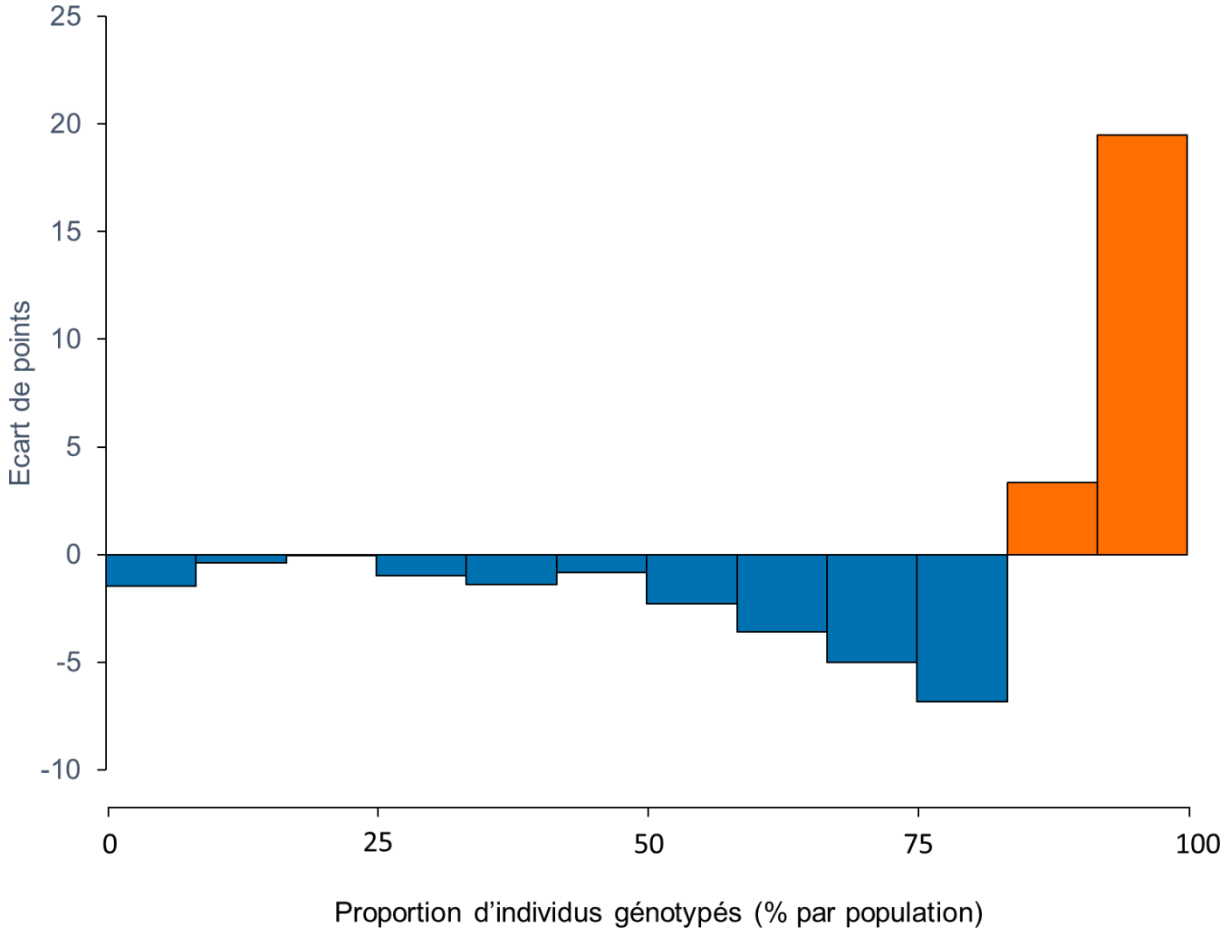



**Figure VI-4. Distribution de l'écart de points pour la proportion de marqueurs identifiés entre les cas *m* = 3 et *m* = 5**. Avec *ustacks*, le choix du paramètre *m* (profondeur minimale de pile) a un impact évident sur le polymorphisme détecté par *Stacks*. L'histogramme montre la différence en proportion de marqueurs détectés lorsque l'on choisit *m* = 3 plutôt que *m* = 5. Cette différence est en faveur d'un paramètre plus souple (*m* = 3), qui améliore nettement la couverture des « meilleurs » marqueurs, c'est-à-dire ceux qui sont génotypés chez au moins 55 individus (≈ 92 %) par population. Les couleurs permettent d'identifier les classes pour lesquelles l'écart est en faveur (orange) ou en défaveur (bleu) du choix de *m* = 3.

L'étape suivante de traitement des *reads* a été menée avec le programme *cstacks* (*catalog Stacks*), qui permet de compiler l'ensemble des séquences consensus des différents individus dans un catalogue unique. *cstacks* fonctionne selon le même algorithme de comparaison de *k-mers* que *ustacks*, et regroupe en un locus unique les séquences inter-individus suffisamment proches selon un critère spécifique, renseigné par le paramètre *n*. Afin de permettre l'identification des locus bialléliques dont un allèle différent serait fixé dans les populations divergentes, nous avons choisi *n* = 1, c'est-à-dire autorisé une distance maximale d'un nucléotide pour rassembler en un locus unique le polymorphisme interindividuel, ce qui a abouti à un catalogue de 130942 locus uniques.

Les génotypes définitifs de chaque individu en chaque locus ont ensuite été fixés par les programmes *sstacks* (*search Stacks*), qui identifie les génotypes individuels à partir des données du catalogue, et *populations*, qui formate le jeu de marqueurs de façon population-spécifique en vue des analyses ultérieures. Ces programmes implémentent en outre des procédures de contrôle afin d'écarter les locus ambigus (*e.g.,* séquences répétées) ou peu plausibles (*e.g.,* haplotypes très peu couverts). Ces contrôles automatiques représentent un prérequis minimal pour faire de la génomique des populations. Bien évidemment, nous filtrerons plus avant ce jeu de données, qui est composé de 120318 marqueurs en sortie du pipeline *Stacks* (Fig. VI-5).

### 2.1.e. Filtres additionnels

Une deuxième étape de traitement des données a été effectuée pour finalement aboutir à un jeu de 16370 SNP qui seront utilisés pour la recherche de signatures de sélection (Fig. VI-6).

Cette deuxième étape de tri des variants a débuté par l'élimination du polymorphisme résultant plus vraisemblablement de régions dupliquées que de locus uniques. La duplication ancestrale du génome chez les salmonidés a en effet favorisé la présence de locus paralogues (Glasauer & Neuhauss, 2014) dont les nucléotides différents peuvent être attribués à tort au polymorphisme d'un seul locus lors du traitement des *reads* bruts avec *Stacks*. Ces « faux-SNP » sont détectables au moyen des échantillons HD. Le génotypage par RAD-seq de 18 individus HD, dont les *reads* ont été intégrés à l'analyse *Stacks*, nous permet de détecter les paralogues présents au sein de notre catalogue de locus. Nous avons considéré qu'un locus était dupliqué lorsque celui-ci était identifié comme hétérozygote chez au moins deux individus HD. Ainsi, 3605 locus paralogues, et donc en réalité monomorphes, ont pu être identifiés et écartés des analyses ultérieures. Une fois l'ensemble des locus monomorphes éliminés, il reste 43886 SNP polymorphes dans le jeu de données, dont 30018 sont bialléliques.

Un peu moins de la moitié des 30018 SNP bialléliques a ensuite été écartée. Nous avons tout d'abord écarté les SNP dont la fréquence de l'allèle mineur (MAF) était inférieure à 1% dans l'ensemble du jeu

de données. Nous avons aussi écarté les locus ne respectant pas les proportions attendues sous l'équilibre de Hardy-Weinberg selon le test exact de Wigginton *et al.* (2005), qui est bien adapté au traitement de petits échantillons. Enfin, nous avons écarté les SNP génotypés chez moins de 20 individus dans les trois populations (M, G et T), ce qui représente une couverture minimale d'au moins un tiers des individus par marqueur au sein de chaque population.

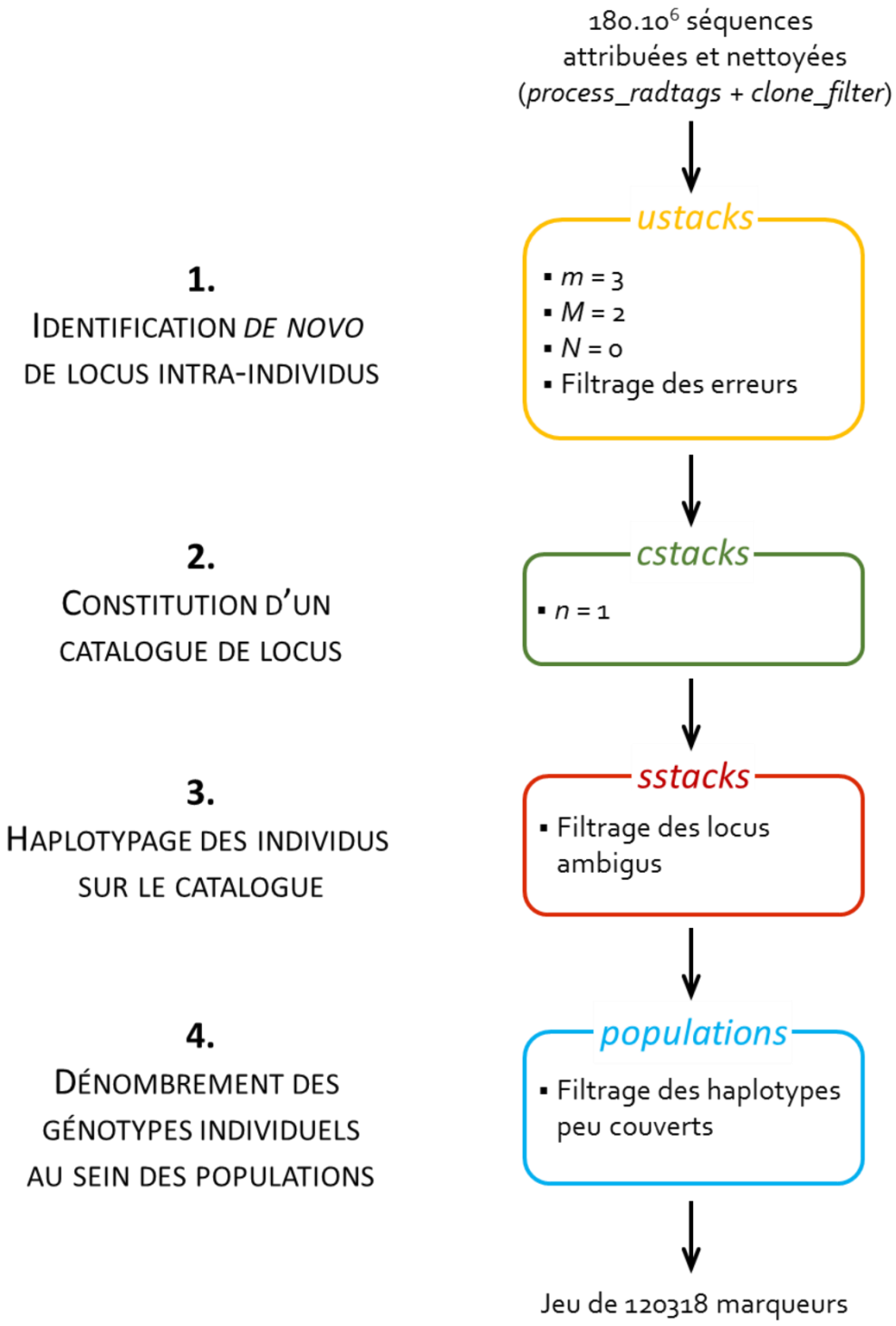



**Figure VI-5. Résumé des principales étapes du traitement bioinformatique des séquences brutes *de novo* avec *Stacks*.** La figure indique les principaux paramètres et types de filtres appliqués aux *reads* 1 à l'aide de quatre programmes (*ustacks*, *cstacks*, *sstacks*, *populations*) du pipeline principal (*core Stacks*, Catchen *et al.*, 2011, 2013).

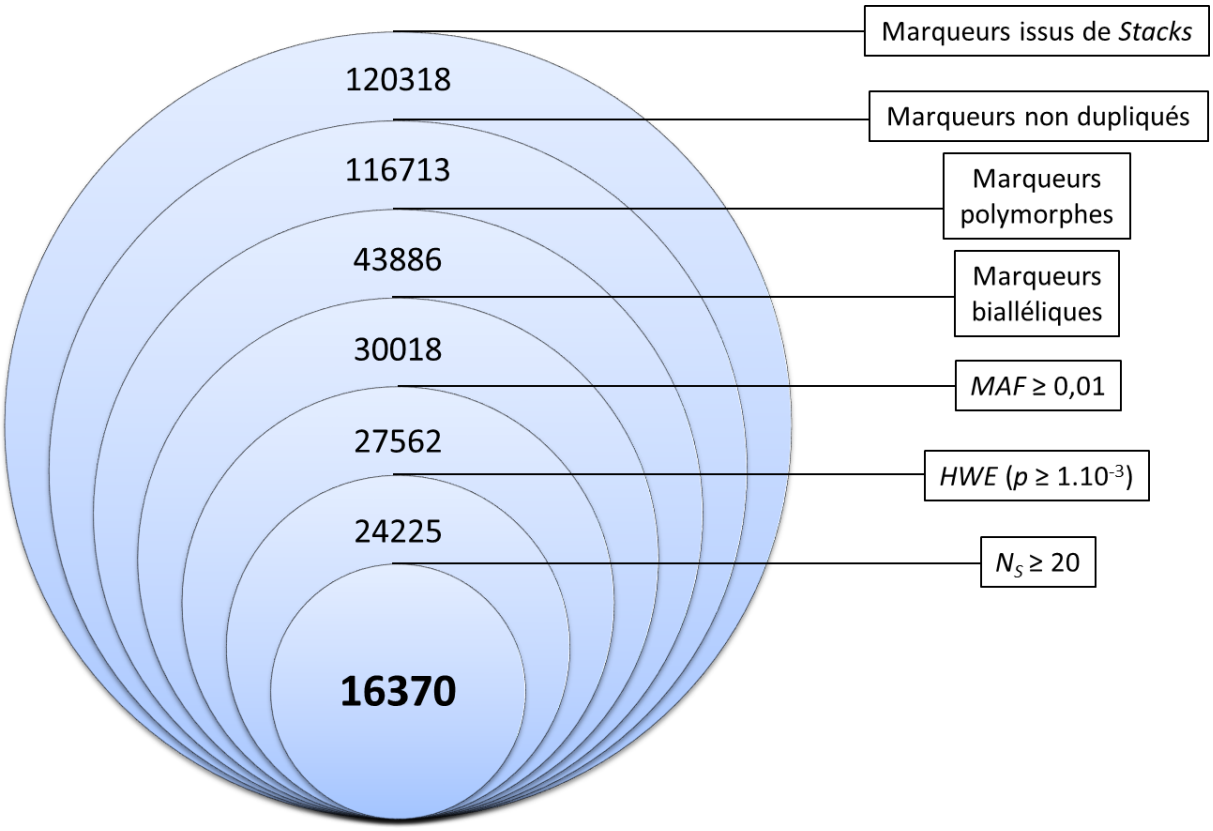



**Figure VI-6. Stratégie de filtrage des variants issus de *Stacks*.** Au total, 43886 sites polymorphes ont pu être identifiés à partir des locus proposés par *Stacks*, dont 30018 sont bialléliques. Parmi ceux-ci, nous avons conservé pour les analyses ultérieures les 16370 marqueurs qui vérifiaient les trois conditions suivantes : un génotype disponible chez au moins 20 individus ($N_s \geq 20$), des proportions d'hétérozygotes et d'homozygotes respectant l'équilibre de Hardy-Weinberg (test exact de Wigginton *et al*. (2005), $p \geq 1.10^{-3}$) et une *MAF* supérieure ou égale à 0,01.

## 2.2. Inférence des fréquences alléliques ancestrales

### 2.2.a. Echantillons génétiques temporels

Notre méthode de détection de signatures de sélection nécessite de disposer d'au moins deux échantillons génétiques temporels, c'est-à-dire distants dans le temps. Idéalement, il est souhaitable d'avoir un échantillon issu de la population immédiatement avant que la sélection débute, la population parentale (G0), en plus de l'échantillon disponible au terme de la sélection (G7).

L'expérience de sélection menée sur les truites INRA a abouti au génotypage de trois populations de truite (Fig. VI-2), dont deux (M et G) correspondent au produit de 7 générations de sélection divergente sur la teneur en lipides du muscle à partir de la même population parentale (G0). La troisième population génotypée est la population T, entretenue sans sélection pendant neuf générations à partir d'une cohorte issue d'une souche commune (G-2).

Comme nous ne disposons que d'échantillons génétiques des populations contemporaines (G7), il a été nécessaire d'inférer les fréquences alléliques des deux populations « ancestrales », la population parentale (G0) et la population souche (G-2), afin de constituer des séries génétiques temporelles complètes pour chaque lignée expérimentale.

Aujourd'hui, les outils de la génétique des populations permettent d'estimer les fréquences alléliques sur l'ensemble du génome de populations auxquelles nous n'avons plus accès, à condition que les génotypes de populations apparentées soient disponibles. Nous nous proposons donc d'estimer les fréquences alléliques des populations ancestrales (Fig. VI-2) à l'aide du programme *KimTree* (Gautier & Vitalis, 2013).

### 2.2.b. KimTree

Il existe aujourd'hui des méthodes permettant d'inférer l'histoire démographique des populations à partir des fréquences alléliques de marqueurs génétiques, en s'appuyant sur des modèles classiques de génétique des populations. Ainsi, Tataru *et al.* (2016) ont récemment montré que l'approximation de diffusion permettait d'obtenir une inférence précise et robuste de la distribution des fréquences alléliques.

Parmi les méthodes disponibles implémentant l'approximation de diffusion, celle proposée par Gautier & Vitalis (2013), distribuée sous le nom de *KimTree*, présente des caractéristiques intéressantes pour traiter le cas des populations de truite génotypées dans le cadre du projet EFFICACE. Les auteurs utilisent un modèle bayésien hiérarchique reposant sur la solution proposée par Kimura (1964) pour estimer les temps de divergence entre populations et la distribution des fréquences alléliques. Leurs résultats suggèrent que ce modèle est bien adapté à l'analyse de populations expérimentales, dont la démographie est contrôlée (générations non chevauchantes, absence de migration). Dans le cadre du projet EFFICACE, non seulement la démographie est contrôlée, mais l'arbre phylogénétique des populations est *a priori* connu. De plus, le modèle de *KimTree* serait utilisable même si $N_e$ est faible, et il donne de bons résultats même si le nombre de marqueurs disponibles est restreint (5000 SNP suffiraient pour l'analyse d'un dispositif où trois populations apparentées sont génotypées comme c'est le cas avec le projet EFFICACE). Recourir à *KimTree* apparaît tout à fait indiqué dans le cas présent.

En statistique bayésienne, les paramètres d'intérêt sont estimés à partir de leur distribution *a posteriori*. Il est donc nécessaire de générer des échantillons de cette distribution afin d'avoir accès aux mesures qui nous intéressent. *KimTree* utilise un échantillonnage de Gibbs pour estimer les distributions des fréquences alléliques et les temps de divergence. Ce type d'algorithme de Monte-Carlo par Chaînes de Markov (MCMC) est très utilisé, notamment dans le cadre de modèles

hiérarchiques. Si l'échantillonnage de Gibbs offre des propriétés très intéressantes, même dans des cas complexes, il est cependant nécessaire de s'assurer que l'échantillon obtenu est assez représentatif de la distribution *a posteriori* pour que l'on puisse considérer que les estimations sont fiables. Nous mènerons donc en premier lieu une analyse de convergence des MCMC obtenues avec *KimTree*.

### 2.2.c. Analyse de la convergence des MCMC

Lorsque l'on utilise un algorithme MCMC pour échantillonner une distribution *a posteriori*, on s'attend à ce que la distribution des points échantillonnés converge vers la distribution véritable quand le nombre de réalisations tend vers l'infini. On désigne donc par analyse de convergence la démarche empirique qui consiste à vérifier que le nombre d'itérations parcourues par une MCMC dans un contexte donné est suffisamment grand pour avoir une bonne approximation de la distribution cible.

Dans la pratique, une MCMC va comporter deux portions fonctionnellement distinctes. Le début de la chaîne constitue la zone dite « d'approche » (ou « période de chauffe ») et ne sera pas conservé. Le but de cette opération est de s'assurer que les valeurs initiales de l'algorithme MCMC n'influenceront pas les estimations effectuées à partir de l'échantillon. Seules seront donc conservées pour le calcul des quantités d'intérêt les itérations de la portion « principale » de la chaîne, au sein de laquelle sera tiré l'échantillon. Par ailleurs, la pratique du *thinning* (que l'on pourrait traduire par « amincissement »), qui consiste à n'échantillonner qu'une fraction des itérations de chaîne principale, peut être employée pour réduire l'autocorrélation entre les tirages. Cette stratégie peut toutefois demander beaucoup de ressources supplémentaires puisqu'elle nécessite d'augmenter la taille des MCMC générées.

Nous avons appliqué le modèle de *KimTree* à un jeu de données réduit comprenant les SNP les mieux couverts au sein des trois populations M, G et T du dispositif EFFICACE. Pour diagnostiquer la convergence, nous avons utilisé des outils classiques et complémentaires permettant de comparer l'effet de différentes longueurs de chaîne et de différentes répartitions des itérations au sein des chaînes sur leur capacité à converger.

De façon plus précise, nous avons utilisé le package CODA (Plummer, 2006) en nous appuyant sur la démarche générique proposée par Cowles & Carlin (1996), qui recommandent d'évaluer la convergence de $m$ chaînes indépendantes ($3 \leq m \leq 5$) en combinant plusieurs outils de diagnostic. Conformément à ces indications, nous avons lancé quatre chaînes en parallèle pour chaque test de convergence, en prenant soin de faire débuter l'algorithme avec des valeurs initiales distinctes choisies au hasard. Les échantillons MCMC ont fait l'objet d'une inspection visuelle sur graphiques usuels et

d'un test de Gelman-Rubin (Gelman & Rubin, 1992). En outre, les corrélations entre variables et au sein des chaînes ont été examinées.

La comparaison des principaux résultats du diagnostic montre que plusieurs options d'organisation des chaînes permettent d'atteindre la convergence (Tableau VI-1). Parmi les essais effectués, la combinaison la plus rationnelle est de recourir à des MCMC comprenant $10^5$ itérations en phase d'approche, suivies à nouveau de $10^5$ itérations qui formeront l'échantillon *a posteriori* (sans *thinning*). Dans cette configuration, le facteur de réduction d'échelle potentiel ($\hat{R}$) se stabilise autour de 1, comme attendu sous l'hypothèse de stationnarité. Ce facteur est en effet fondé sur un rapport de variances qui tend vers 1 lorsque la distribution stationnaire est atteinte (Gelman & Rubin, 1992). Dans la pratique, on considère qu'un facteur de réduction compris entre 1 et 1,05 est un indicateur satisfaisant de la convergence des chaînes MCMC. Toujours dans la même configuration, nous avons mesuré un facteur de réduction multivarié $\hat{R}^p$ de 1,01 (Brooks & Gelman, 1998), confirmant la convergence des quatre chaînes pour l'ensemble des paramètres.

On constate souvent dans les faits que les auteurs souhaitent minimiser les corrélations entre les réalisations de l'échantillon, et préconisent ou effectuent à ce dessein un *thinning* des chaînes. La pertinence de cette pratique fait actuellement l'objet de discussions parmi les utilisateurs de méthodes bayésiennes. Link *et al.* (2012) ont notamment montré que les estimations étaient souvent meilleures en l'absence de *thinning* des échantillons issus de MCMC. Ces auteurs rappellent qu'observer une meilleure précision, de même que constater des corrélations importantes au sein des chaînes, est normal et découle des caractéristiques des MCMC. Ils recommandent d'éviter un *thinning* systématique lors des analyses de convergence, et de réserver cette pratique aux cas qui le nécessitent. Trop souvent, le *thinning* aboutirait selon eux à un « gaspillage » de chaînes dont la production peut être lente, pour un gain de précision nul ou dérisoire. Nos comparaisons *thinning* c. *absence de thinning* vont dans ce sens et montrent que le ratio efficace et le temps de calcul plaident en la défaveur du *thinning* (Tableau VI-1). Nous avons donc considéré que le *thinning* des chaînes ne se justifiait pas dans notre cas.

Au bilan, les résultats de l'analyse de convergence indiquent que les tailles de la zone d'approche ($10^5$ itérations) et de la chaîne principale ($10^5$ itérations et sans *thinning*) sont suffisantes pour que les valeurs initiales de l'algorithme de *KimTree* n'affectent pas la distribution postérieure. Nous considérerons donc ces modalités d'obtention des échantillons MCMC pour mener les inférences ultérieures.

Tableau VI-1. Analyse de la convergence des chaînes MCMC permettant l'estimation des temps de divergence entre populations M, G et T

| $i_{total}$ (Nb. total d'itérations) | $i_{burn}$ (Rang de la dernière itération écartée) | $i_{post}$ (Nb. itérations conservées) | $thin$ (*Thinning*) | $n_{post}$ (Taille de l'échantillon MCMC) | $n_{eff}$ (Taille efficace de l'échantillon MCMC)[a] | $n_{eff}/n_{post}$ (Ratio efficace) | $\hat{R}$ (Facteur de réduction univarié)[bc] | $\hat{R}_{0,95}$ (Borne sup. IC95)[c] | $\hat{R}^p$ (Facteur de réduction multivarié)[d] | $ACF$ (Auto-corrélation)[e] | $t$ (Temps d'exécution)[f] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $1{,}01.10^5$ | $10^3$ | $10^5$ | 20 | $5.10^3$ | 406 | $4.10^{-3}$ | 3,83 | 6,63 | 4,73 | 0,45 (*k* = 20) | 3 h 56 m |
| $1{,}1.10^5$ | $10^5$ | $10^4$ | 1 | $10^4$ | 218 | $2{,}2.10^{-2}$ | 1,03 | 1,09 | 1,03 | 0,93 (*k* = 1) | 3 h 42 m |
| $1{,}1.10^5$ | $10^4$ | $10^5$ | 1 | $10^5$ | 1224 | $1{,}2.10^{-2}$ | 1,01 | 1,04 | 1,02 | 0,93 (*k* = 1) | 4 h 13 m |
| $1{,}1.10^5$ | $10^4$ | $10^5$ | 20 | $5.10^3$ | 586 | $6.10^3$ | 1 | 1,01 | 1,01 | 0,49 (*k* = 20) | 4 h 31 m |
| $2.10^5$ | $10^5$ | $10^5$ | 1 | $10^5$ | 1243 | $1{,}2.10^{-2}$ | 1 | 1,01 | 1 | 0,93 (*k* = 1) | 6 h 39 m |
| $2.10^5$ | $10^5$ | $10^5$ | 20 | $5.10^3$ | 618 | $6.10^{-3}$ | 1 | 1,01 | 1,01 | 0,47 (*k* = 20) | 7 h 21 m |
| $1{,}1.10^6$ | $10^6$ | $10^5$ | 1 | $10^5$ | 1141 | $1{,}1.10^{-2}$ | 1,01 | 1,02 | 1,01 | 0,93 (*k* = 1) | 34 h 10 m |
| $1{,}35.10^6$ | $10^5$ | $1{,}25.10^6$ | 250 | $5.10^3$ | 2994 | $2.10^{-3}$ | 1 | 1,01 | 1 | 0,16 (*k* = 250) | 39 h 35 m |

[a] La taille efficace correspond ici au nombre d'itérations que l'on peut considérer comme indépendantes dans l'échantillon de la distribution *a posteriori*.

[b] Sous l'hypothèse de stationnarité, on s'attend à ce que le facteur de réduction d'échelle $\hat{R}$ (Gelman & Rubin, 1992) soit le plus proche possible de 1.

[c] Les estimations concernent le paramètre dont l'établissement de la convergence nécessite le plus d'itérations.

[d] Le facteur de réduction multivarié $\hat{R}^p$ résume la qualité de la convergence pour l'ensemble des paramètres estimés (Brooks & Gelman, 1998).

[e] L'autocorrélation mesure la corrélation intra-chaîne avec un décalage *k* équivalent à la valeur de *thinning*.

[f] *KimTree* a été compilé avec Intel®Fortran Compiler sur la plateforme Genotoul-bioinfo.

*N.B.* Les chaînes MCMC ont été obtenues avec *KimTree* (Gautier & Vitalis, 2013) à partir des 7601 SNP les plus couverts.

### 2.2.d. Comparaison de modèles avec *KimTree*

Un des objectifs de l'utilisation de *KimTree* est de retrouver l'histoire démographique des populations analysées. Afin de mesurer les performances prédictives du logiciel pour cette tâche, nous disposons principalement de deux critères de comparaison de modèles, obtenus à partir de la distribution *a posteriori*, et communément utilisés en analyse bayésienne (voir par exemple Guo *et al.*, 2009 ou Jiang *et al.*, 2013) : le *DIC* (*Deviance Information Criterion*) et le *LPML* (*Logarithm of the PseudoMarginal Likelihood*).

Dans leur publication présentant *KimTree*, Gautier & Vitalis (2013) ont montré que le *DIC* était un critère pertinent pour identifier l'arbre phylogénétique rendant le mieux compte de la véritable histoire démographique des populations (Gautier & Vitalis, 2013). Cette statistique fondée sur la déviance (*i.e.*, moins deux fois la log-vraisemblance des données) a été introduite pour faciliter la comparaison de modèles hiérarchiques complexes (Spiegelhalter *et al.,* 2002). Considérant que le choix d'un modèle doit être le fruit d'un compromis entre son pouvoir explicatif et sa complexité, Spiegelhalter *et al.* (2002) ont défini le *DIC* comme la somme d'une mesure de la qualité de l'ajustement et d'un terme de pénalité (Equation VI-1).

$$DIC = \overline{D} + p_D \qquad \text{Equation VI-1}$$

Ce critère utilise la moyenne *a posteriori* de la déviance $\overline{D}$ afin de quantifier la qualité de l'ajustement des données au modèle. Plus $\overline{D}$ est faible, meilleur est l'ajustement. L'autre composante du *DIC* est le terme de pénalité, $p_D$, qui estime le nombre efficace de paramètres d'un modèle bayésien. L'idée directrice est d'identifier le moment à partir duquel le gain en qualité d'ajustement n'est plus suffisant pour justifier d'un plus grand dimensionnement du modèle. En d'autres termes, le *DIC* fournit un critère de choix du modèle le plus parcimonieux : plus le *DIC* est faible, plus le modèle est parcimonieux.

Le *LPML* s'appuie sur une approche de type validation croisée pour comparer les modèles bayésiens. Dans les années 80, plusieurs auteurs ont montré l'intérêt d'une quantité usuellement nommée CPO (*Conditional Predictive Ordinate*), laquelle peut être estimée à partir d'un échantillon MCMC pour mesurer la précision d'un modèle (Gelfand *et al.*, 1992). La CPO calcule la probabilité de chaque observation $y_k$ en se fondant sur un échantillon de ($k$ - 1) observations $y_{(-k)}$ d'où la $k$-ième observation a été retirée (Equation VI-2).

$$CPO_k = p\left(y_k \middle| y_{(-k)}\right) \qquad \text{Equation VI-2}$$

On bénéficie ainsi d'une mesure de l'ajustement du modèle pour chaque *k*. Le *LPML* résume l'information en sommant les *log-CPO*$_k$ afin de fournir un critère d'évaluation des capacités prédictives d'un modèle (Equation VI-3). Plus le *LPML* est grand, meilleur est le modèle (Lewis *et al.*, 2014).

$$LPML = \sum_{k=1}^{n} log\widehat{CPO_k} \qquad \text{Equation VI-3}$$

Nous nous sommes appuyés sur ces deux critères, *DIC* et *LMPL*, pour identifier, parmi les histoires démographiques compatibles avec un dispositif expérimental comportant trois populations apparentées, celle qui rendait le mieux compte des données acquises à partir des lignées de truite EFFICACE.

## 2.3. Détection de signatures de sélection

### 2.3.a. Méthodes basées sur la mesure de la différenciation entre populations

Dans cette section, nous considérerons que l'échantillonnage de $n$ populations contemporaines apparentées (*i.e.,* ayant divergé à partir d'une même population dite « ancestrale ») a permis d'obtenir un jeu de SNP bialléliques neutres dans une large majorité. Soit $p$ le vecteur des fréquences d'un allèle d'intérêt en un locus donné, la divergence entre populations en ce locus est classiquement mesurée par l'indice de fixation de Wright ou $F_{ST}$ (Equation VI-4).

$$F_{ST} = \frac{s_p^2}{\bar{p}(1-\bar{p})} \qquad \text{Equation VI-4}$$

$s_p^2$ et $\bar{p}$ représentent la variance et la moyenne, respectivement, des fréquences alléliques.

Les tests de détection de signatures de sélection effectués à partir de données issues de populations sélectionnées de façon artificielle et divergente sont fréquemment basés sur l'utilisation du $F_{ST}$ (voir par exemple Akey *et al.*, 2010 ; Moradi *et al.*, 2012 ; Petersen *et al.*, 2013 ; Rothammer *et al.*, 2013 ; Yang *et al.*, 2014). Ces approches ont en commun l'idée proposée en 1973 par Lewontin et Krakauer que les locus sous forte sélection directionnelle doivent montrer d'une population à l'autre une différenciation visible de fréquence allélique par rapport au reste du génome (Equation VI-5) (Lewontin & Krakauer, 1973).

$$T_{LK} = \frac{n-1}{\bar{F}_{ST}} F_{ST} \qquad \text{Equation VI-5}$$

- $n$ est le nombre de populations incluses dans l'analyse,
- $F_{ST}$ est calculé pour le locus testé (Equation VI-4),
- $\bar{F}_{ST}$ est la moyenne des $F_{ST}$ sur l'ensemble des locus inclus dans l'analyse.

Plus de trois décennies après la suggestion initiale de Lewontin et Krakauer, l'accès aux données génomiques à haut débit a donc suscité un important engouement pour les approches de type $F_{ST}$ *outlier*, c'est-à-dire proposant d'identifier des candidats à la sélection parmi les SNP affichant des valeurs de $F_{ST}$ extrêmes. Des développements récents ont permis l'amélioration de la statistique de test de Lewontin-Krakauer ($T_{LK}$, équation VI-5). En particulier, Bonhomme *et al.* (2010) ont proposé d'inclure l'information disponible sur les apparentés en incorporant la matrice de parenté dans une nouvelle statistique de test ($T_{FLK}$, équation VI-6).

$$T_{FLK} = (p - p_0 1_n)^T V(p)^{-1} (p - p_0 1_n) \qquad \text{Equation VI-6}$$

- $p$ est le vecteur des fréquences alléliques au SNP testé,
- $p_0$ est la fréquence allélique du SNP testé à la génération ancestrale,
- $1_n$ désigne le vecteur unité de taille $n$,
- $V(p)$ est la variance du vecteur $p$, dépendant de la matrice de parenté $\mathcal{F}$ (Equation VI-7).

$$V(p) = \mathcal{F} p_0 (1 - p_0) \qquad \text{Equation VI-7}$$

$p_0$ et $\mathcal{F}$ sont les paramètres du modèle et sont estimés à partir de la mesure des fréquences alléliques de chaque SNP disponible au sein des populations contemporaines, sous l'hypothèse nulle d'évolution neutre ($H_0$).

Sous $H_0$, les deux statistiques de test, $T_{LK}$ et $T_{FLK}$, suivent une loi de $\chi^2$ à $(n - 1)$ degrés de liberté. En incluant l'information sur les apparentés, $T_{FLK}$ prend en compte l'hétérogénéité de l'effectif d'une population à l'autre et la non-indépendance de la divergence des populations étudiées, ce qui offre de meilleures performances que $T_{LK}$ au plan statistique dans une majorité de cas (Bonhomme *et al.*, 2010). Son usage est adapté à la détection du polymorphisme sélectionné au sein des populations d'intérêt agronomique, car celles-ci sont en général génétiquement isolées (absence de migration) et leur divergence de la population ancestrale est relativement récente, ce qui signifie que le polymorphisme ségrégeant dans les populations contemporaines était déjà présent dans la population ancestrale (Bonhomme *et al.*, 2010). Son application a permis d'identifier des gènes candidats à la sélection au sein de plusieurs espèces, y compris l'Homme (Bonhomme *et al.*, 2010 ; Fariello *et al.*, 2014 ; Barson *et al.*, 2015 ; Gholami *et al.*, 2015 ; Schaschl *et al.*, 2015).

Nous avons calculé $T_{LK}$ et $T_{FLK}$ pour chacun des 16370 SNP acquis dans le cadre du dispositif expérimental EFFICACE. Nous avons pour cela utilisé les fonctions disponibles dans le script FLK.R distribué par l'INRA via la *Quantitative Genetics Software Platform* (https://qgsp.jouy.inra.fr). La

matrice de parenté des populations analysées a été obtenue en estimant les distances de Reynolds à partir des fréquences alléliques et en enracinant l'arbre des populations avec la population Témoin. Les *p-values* obtenues à l'issue des tests ($p_{LK}$ et $p_{FLK}$) ont en outre été corrigées avec la fonction *p.adjust* du *package* '*stats*' selon la procédure de Benjamini & Hochberg (1995), afin d'évaluer la proportion fausses découvertes attendue liée à la réalisation de tests multiples.

### 2.3.b. *PPP-values*

L'approche bayésienne implémentée dans *KimTree* permet aussi de mesurer, pour chaque SNP, la dispersion des données observées par rapport au modèle d'évolution neutre sous-jacent. Il est ainsi possible de repérer des *outliers* à travers le calcul de *p-values* prédictives *a posteriori* (*PPP-values*), lesquelles sont un équivalent bayésien des *p-values* fréquentistes (Gelman *et al.*, 1996 ; Gautier *et al.*, 2010).

La distribution des fréquences alléliques des populations contemporaines est prédite en considérant les paramètres estimés sous le modèle de dérive pure de *KimTree*. L'écart entre les distributions prédites et les observations peut être testé en comparant les critères de divergence $T_i^{rep}$ et $T_i^{obs}$, propres à chaque SNP *i* (Equation VI-8). Ces critères rendent compte de la différenciation des SNP au cours du temps, telle que prédite par le modèle d'évolution neutre ($T_i^{rep}$) et telle qu'estimée à partir des observations ($T_i^{obs}$), sachant les paramètres du modèle. Sous l'hypothèse de neutralité, les critères de divergence $T_i^{rep}$ et $T_i^{obs}$ sont supposés proches et la *PPP-value* est attendue autour de 0,5. Une faible *PPP-value*, c'est-à-dire une forte probabilité d'avoir $T_i^{obs}$ supérieur à $T_i^{rep}$, indique que la variance estimée à partir des observations est trop importante pour être compatible avec un effet de la seule dérive, et suggère la signature d'une sélection directionnelle au locus examiné (Gautier *et al.*, 2010).

$$PPP_i = \Pr[T_i^{(rep)} \geq T_i^{(obs)} | f^{(obs)}] \qquad \text{Equation VI-8}$$

— $T_i^{rep}$ (resp. $T_i^{obs}$) est le critère de divergence prédit (resp. observé) au SNP *i*,

— $f^{(obs)}$ désigne l'ensemble des fréquences alléliques observées.

En résumé, les *PPP-values* fournissent un moyen d'estimer la qualité de l'ajustement du modèle d'évolution neutre de *KimTree* aux données observées. Les *outliers* associés à de faibles *PPP-values* individuelles sont des SNP candidats à une sélection directionnelle. En cas de forte sélection positive ($s$ = 0,1) sur 10 générations, Gautier *et al.* (2010) ont montré par simulation qu'un seuil de 0,2 ($PPP_i$ < 0,2) permettait d'obtenir une puissance d'environ 45% (en contrepartie d'un *FDR* d'environ 10%). Nous avons calculé les *PPP-values* associées à chacun des 16370 SNP biallélliques générés chez la truite dans le cadre du dispositif EFFICACE. Les SNP présentant une *PPP-value* inférieure à 0,2 ont été identifiés

comme de potentiels candidats à une sélection directionnelle sur la teneur en lipides du muscle chez la truite.

### 2.3.c. Application de notre méthode de détection (*signasel*)

Au plan méthodologique, un objectif est de tester le comportement de notre méthode temporelle de détection de signatures de sélection en conditions réelles, et plus particulièrement dans une situation (sélection directionnelle intense pendant sept générations) où les simulations suggèrent des performances intéressantes. Par commodité, nous ferons souvent référence à notre méthode de détection de signatures de sélection sous le terme *signasel* (du nom du programme associé) dans la suite de ce chapitre.

Comme notre méthode s'intéresse à la dynamique temporelle des fréquences alléliques, une série composée d'un minimum de deux échantillons génétiques temporels est nécessaire pour exploiter chaque lignée examinée. En l'absence d'un échantillon initial, une stratégie consiste à inférer les fréquences alléliques manquantes à l'aide de *KimTree* v1, comme nous l'avons proposé. Nous considérerons que la moyenne *a posteriori* de la distribution inférée par *KimTree* estime correctement la fréquence allélique dans la population ancestrale correspondante. La reconstitution de séries génétiques temporelles permet ainsi de tester la présence de sélection directionnelle au sein de chaque lignée. Chaque SNP du jeu de données fera l'objet de trois tests avec *signasel*, soit un par lignée. Deux tests vont examiner si l'évolution des fréquences entre les échantillons collectés dans les populations sélectionnées et la population parentale inférée, distantes de 7 générations, est localement compatible avec une action de la sélection. Le troisième test sera mené sur la lignée Témoin, soumise à la seule dérive.

Une autre inférence est préalablement nécessaire à l'utilisation de *signasel*. Notre méthode de détection des signatures de sélection suppose en effet que l'effectif efficace ($N_e$) des populations analysées est connu. Cette information n'est pas disponible *a priori* dans de nombreux cas, et il est alors nécessaire d'estimer $N_e$ à partir du même jeu de données que celui qui fera l'objet de la recherche de signatures de sélection. De récents travaux comparatifs (Gilbert, K.J., & Whitlock 2015 ; Jiménez-Mena, 2016 ; Wang, 2016 ; Waples, R.S., 2016) et implémentations au sein de packages ou de logiciels d'usage simple (Jones & Wang, 2010 ; Sheehan *et al.*, 2013 ; Do *et al.*, 2014 ; Nikolic & Chevalet, 2014) apportent aujourd'hui une aide notable dans le choix et l'utilisation d'un estimateur adapté pour inférer $N_e$ à partir de données moléculaires.

Nous avons estimé le $N_e$ des populations Maigre, Gras et Témoin en appliquant la méthode de Waples, R.S. (2006), distribuée sous le nom de *LDNe* (Waples, R.S., & Do 2008) et basée sur la mesure du DL

des marqueurs polymorphes disponibles au sein de la population examinée. Cette méthode offre une estimation relativement précise de $N_e$ lorsque l'on dispose d'un unique échantillon génétique temporel (Wang, 2016). Nous avons soumis le jeu de 16370 SNP à l'implémentation de *LDNe* proposée dans *NeEstimator* v2.01 (Do *et al.*, 2014). Conformément aux recommandations de Waples, R.S., & Do (2010), nous avons écarté les SNP dont la MAF était inférieure à 5%, car ils sont peu informatifs et susceptibles d'induire une surestimation du $N_e$.

De plus, nous avons ajusté les estimations de façon à tenir compte du nombre de SNP testés. Inclure plusieurs milliers de SNP dans l'analyse peut en effet conduire à sous-estimer $N_e$ (Waples, R.K., *et al.* 2016). Nous avons par conséquent appliqué à nos estimations une correction préconisée par Waples, R.K., *et al.* (2016), prenant en considération la taille du génome (Equation VI-9) pour traiter ce biais.

$$\frac{\widehat{N}_e}{N_e} = -0{,}91 + 0{,}219 \times \ln(\mathrm{cM}) \qquad \text{Equation VI-9}$$

- $\widehat{N}_e$ est l'estimation brute (*i.e.,* avant correction) de l'effectif efficace,
- $N_e$ est la véritable valeur (*i.e.,* inconnue) de l'effectif efficace,
- cM désigne la longueur totale du génome, exprimée en centimorgan (cM).

Chez *O. mykiss*, on estime que le génome a une longueur moyenne de 3346 cM (Palti *et al.*, 2011), ce qui provoquerait une sous-estimation de $N_e$ de 13%. Ainsi, nos estimations finales de l'effectif efficace sont $\widehat{N}_{e_M}$ = 106, $\widehat{N}_{e_G}$ = 130 et $\widehat{N}_{e_T}$ = 183 pour les populations M, G et T, respectivement.

Nous avons considéré comme candidats les SNP associés par *signasel* à une *p-value* inférieure ou égale à $10^{-4}$, qui correspond à un FDR d'environ 2% selon la procédure de Benjamini-Hochberg (Benjamini & Hochberg, 1995). Le choix d'un seuil de significativité plutôt strict (malgré un $N_e$ relativement élevé à l'échelle des cas que notre méthode peut traiter) nous paraît justifié en premier lieu parce que nos estimations des fréquences alléliques dans les populations ancestrales reposent sur une inférence et non directement sur le génotypage d'un échantillon. En outre, nous ne connaissons pas *a priori* la localisation des SNP analysés, ce qui nous a également incités à la stringence.

# — 3. Résultats

## 3.1. Inférence de l'histoire démographique des populations de truite EFFICACE

Nous avons utilisé l'approche bayésienne hiérarchique implémentée dans *KimTree* (Gautier & Vitalis, 2013), qui exploite un modèle de diffusion formalisé par Kimura (1964) pour rendre compte de l'effet de la dérive génétique sur les fréquences alléliques au cours du temps. *KimTree* permet d'inférer l'histoire démographique des populations et d'estimer d'autres paramètres d'intérêt comme les fréquences alléliques ancestrales, à partir du génotypage de populations contemporaines apparentées.

Utiliser *KimTree* sur les données acquises chez la truite dans le cadre du projet EFFICACE présente plusieurs intérêts. Son emploi est tout d'abord motivé par la possibilité de constituer une série génomique temporelle complète en inférant les fréquences alléliques ancestrales manquantes, ce qui est indispensable pour utiliser ensuite *signasel*. Un autre aspect intéressant est lié au jeu de données, qui a fait l'objet de plusieurs étapes de traitement bioinformatique à partir des données de RAD-seq brutes. Retrouver une histoire démographique cohérente avec les informations dont nous disposons à partir du jeu de SNP final serait une indication que les choix effectués lors de sa constitution n'introduisent pas de biais majeur dans une analyse de génomique des populations. Il s'agit enfin de tester la capacité de *KimTree* à retrouver une histoire démographique *a priori* connue à partir de données réelles dans une configuration (absence de mutation, absence de migration, phylogénie simple, faible $N_e$, quelques marqueurs probablement sous sélection directionnelle…) qui convient au modèle sous-jacent.

*KimTree* résume l'histoire démographique des populations analysées sous la forme d'un arbre phylogénétique dont la topologie *a priori*, c'est-à-dire l'ordre supposé du branchement des nœuds, doit être renseignée par l'utilisateur. Nous avons testé la capacité de *KimTree* à identifier, à partir des 16370 marqueurs générés par RAD-seq, la topologie attendue (*i.e.,* celle associée aux informations dont nous disposions sur l'histoire des populations) parmi les quatre topologies envisageables lorsque l'on travaille avec trois populations apparentées (Fig. VI-7).

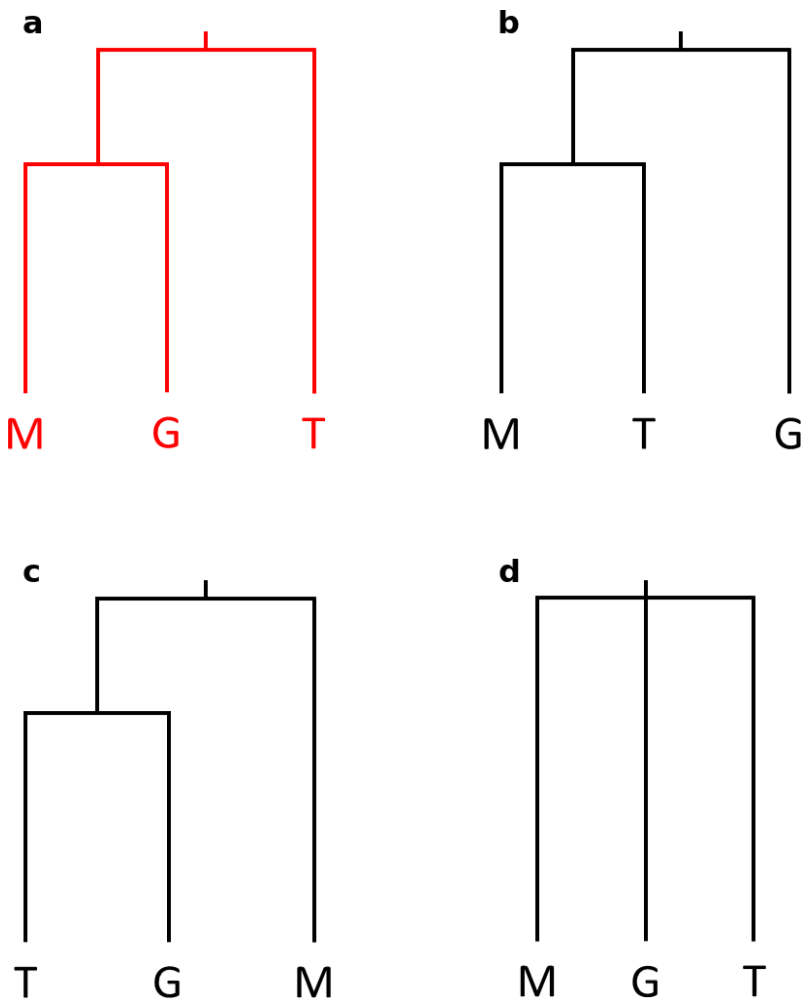



**Figure VI-7. Quatre topologies concurrentes sont susceptibles de rendre compte de l'histoire démographique des populations M, G et T.** La topologie représentée en rouge **(a)** est la véritable : les populations Maigre (M) et Gras (G) divergent à partir de la même population parentale, tandis que la population Témoin (T) permet d'enraciner l'arbre phylogénétique des populations. Les topologies alternatives considèrent un arbre enraciné avec la population Gras **(b)**, Maigre **(c)** ou bien une phylogénie en étoile **(d)**.

Les quantités $\bar{D}$, $DIC$ et $LMPL$, permettant de mesurer l'ajustement du modèle aux données, ont été calculées à partir des échantillons MCMC obtenus avec les 16370 marqueurs du jeu de données pour les quatre topologies renseignées *a priori* (Tableau V-2).

Le critère $DIC$, testé avec succès par Gautier & Vitalis (2013) afin d'identifier la topologie attendue sur données simulées lorsque le génotypage est suffisamment dense (au moins 5000 SNP), a ici permis d'écarter la topologie **(d)** rendant compte d'une phylogénie en étoile avec une unique population ancestrale. Néanmoins, le $DIC$ n'a pas suffi à distinguer les trois autres topologies, dont les scores sont similaires. Il y a en effet au maximum 3 unités de différence entre les scores obtenus pour les topologies **(a)** – l'attendu – **(b)** et **(c)**, alors que Gautier & Vitalis (2013) ont considéré que l'écart devait être supérieur à 10 unités pour conclure à la supériorité d'une topologie sur une autre. L'incapacité du

$DIC$ à trancher en faveur de la topologie véritable est due au terme de pénalité $p_D$. Le nombre efficace de paramètres $p_D$ suggère en effet qu'opter pour la topologie **(a)** serait plus coûteux, mais la plus faible déviance $\overline{D}$ constatée dans ce cas indique que c'est aussi celui qui offre le meilleur ajustement aux données dans l'absolu (Tableau V-2). Ainsi, le $DIC$ ne permet pas de conclure, mais les deux autres mesures des performances du modèle, $\overline{D}$ et $LMPL$, identifient sans ambiguïté que la topologie **(a)** rend mieux compte des données que les autres. *KimTree* permet donc de retrouver l'histoire démographique connue des lignées expérimentales EFFICACE.

Tableau VI-2. Comparaison des performances prédictives de *KimTree* pour quatre topologies concurrentes

| **Topologie**[a] | $\overline{D}$ [b] | $p_D$ [b] | $DIC$ [b] | $LMPL$ [c] |
|---|---|---|---|---|
| **(a)** | 215144 | 35999 | 251143 | -146173 |
| **(b)** | 215258 | 35881 | 251140 | -146222 |
| **(c)** | 215235 | 35906 | 251141 | -146311 |
| **(d)** | 215258 | 35932 | 251190 | -146364 |

[a] Chaque topologie est désignée par une lettre (a, b, c ou d) en référence à la Figure VI-7.
[b] $DIC = \overline{D} + p_D$ (*cf.* équation VI-1). Plus $\overline{D}$ et $DIC$ sont faibles, meilleur est l'ajustement du modèle aux données.
[c] Plus *LMPL* est grand, meilleur est l'ajustement des données au modèle.
Les meilleurs scores pour les trois mesures d'ajustement utilisées ($\overline{D}$, $DIC$ et $LMPL$) sont indiqués en vert.
*DIC = Deviance Information Criterion*, *LMPL = Logarithm of the PseudoMarginal Likelihood*
*N.B. KimTree* (Gautier & Vitalis, 2013) a été utilisé sur 16370 SNP issus des lignées EFFICACE.

*KimTree* infère un arbre *a posteriori* faisant apparaître les temps de divergence estimés ($\hat{\tau}$) entre populations (Fig. VI-8). Cet arbre phylogénétique est cohérent avec l'histoire des populations et avec les estimations de $N_e$. Comme attendu, les deux branches menant aux populations Maigre ($\widehat{N}_{e_M}$ = 106) et Gras ($\widehat{N}_{e_G}$ = 130), soumises à des sélections divergentes similaires, affichent des longueurs comparables, tandis que la branche menant à la population Témoin ($\widehat{N}_{e_T}$ = 183), soumise à la seule dérive, est plus courte.

Ainsi, les estimations des paramètres rendant compte de la structure hiérarchique du modèle de *KimTree* (*i.e.,* les temps de divergence), obtenues à parti des 16370 SNP du jeu de données EFFICACE, sont compatibles avec les informations dont nous disposons par ailleurs sur les populations analysées. Ce résultat montre que le modèle de *KimTree* peut être appliqué au cas des petites populations soumises à une sélection directionnelle sur une période de quelques générations. En outre, nous avons utilisé les fréquences alléliques estimées par *KimTree* en chaque nœud ancestral pour constituer les

séries génomiques temporelles nécessaires à l'identification de signatures de sélection avec notre méthode de détection (*signasel*).

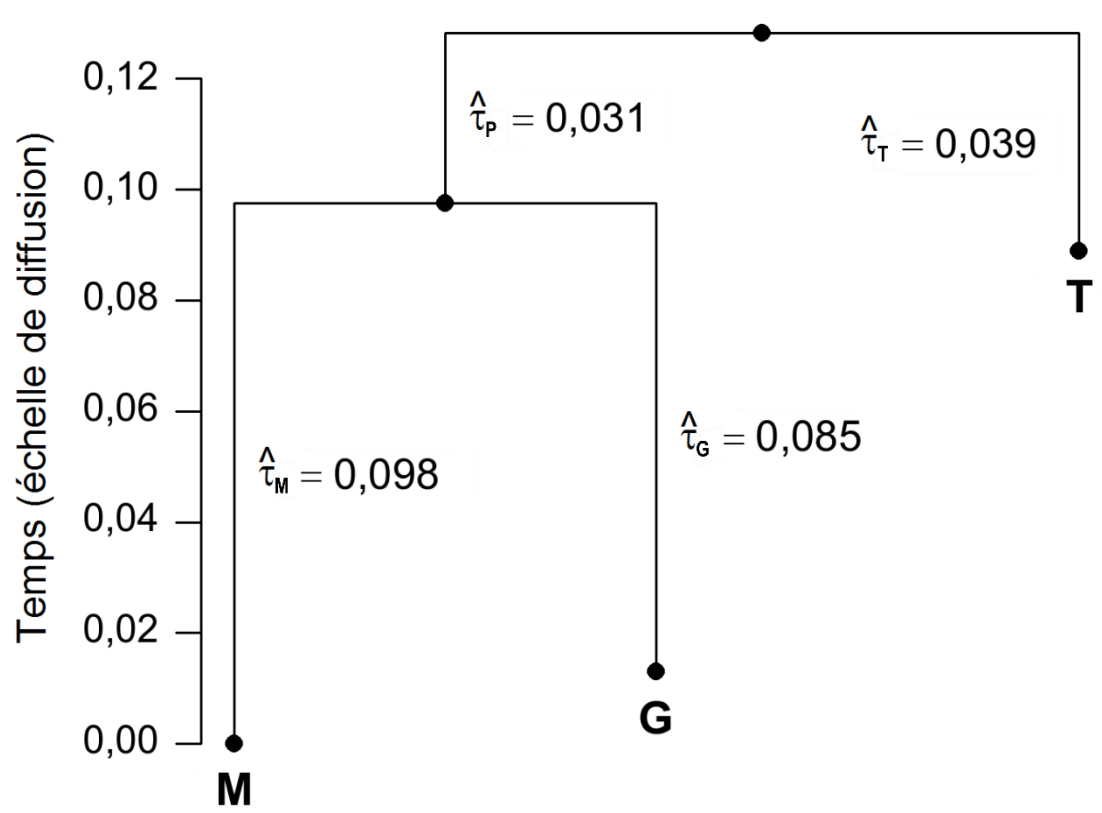



**Figure VI-8. Arbre phylogénétique des populations M, G et T tel qu'inféré par *KimTree* à partir de 16370 SNP.** Les longueurs de branche sont données par les temps de divergence ($\hat{\tau}$), qui estiment la divergence génétique entre chaque population (M, G, T ou Parentale (P)) et la population ancestrale dont elle est issue, et qui sont exprimés en unités de dérive génétique.

## 3.2. Environ 150 SNP sont candidats à la sélection pour la teneur lipidique du muscle

Nous nous sommes appuyés sur la combinaison des résultats de l'application de trois méthodes de détection de signatures de sélection récentes au même jeu de marqueurs afin de proposer une liste de locus candidats à la sélection sur le taux lipidique du muscle chez la truite. Chacun des 16370 marqueurs identifiés par RAD-seq a ainsi été associé à un ensemble de *p-values* fournies par *signasel* (pour chacune des trois populations), *KimTree* (*PPP-value*) et *FLK* ($p_{FLK}$) permettant de quantifier la plausibilité d'un rôle de la sélection sur le polymorphisme observé après 7 générations d'évolution expérimentale.

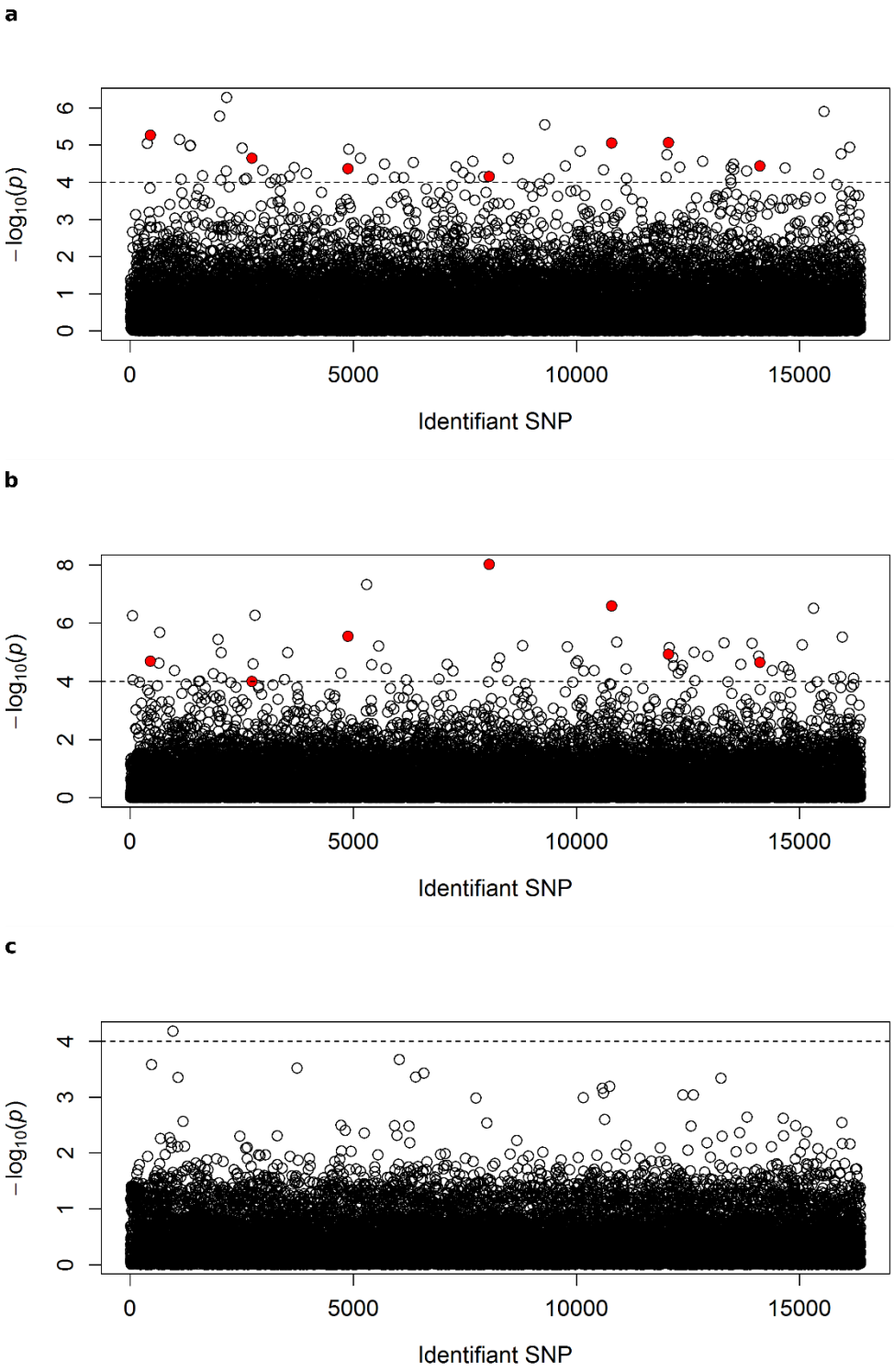



**Figure VI-9. Cent-dix-sept SNP candidats à la sélection ont été identifiés au moyen de *signasel* au sein des lignées expérimentales du dispositif EFFICACE.** Chaque graphique indique en ordonnée $-\log_{10}(p)$ où $p$ représente la *p-value* du LRT effectué dans *signasel*, le programme implémentant notre méthode de détection de signatures de sélection (*cf.* Article I). Pour être candidat, un SNP doit être associé à $p \leq 1.10^{-4}$ (tirets). Parmi les 117 candidats retenus, 59 ont été identifiés dans la population Maigre **(a)** et 65 dans la population Gras **(b)**. La population Témoin **(c)** ne présente qu'un seul SNP associé à $p \leq 1.10^{-4}$. Les SNP identifiés en rouge sont les candidats communs aux deux populations sélectionnées. *N.B.* Les SNP sont ordonnés selon un identifiant qui ne préjuge en rien de leur position dans le génome.

L'utilisation de *signasel* a permis d'identifier au total 117 locus candidats à la sélection (Fig. VI-9). Pour limiter le taux de faux-positifs dans une situation où la fréquence allélique à la génération parentale et la position des marqueurs dans le génome sont *a priori* inconnues, nous avions fixé $\alpha = 1.10^{-4}$. A ce niveau strict de contrôle du risque de première espèce, le taux de fausses-découvertes (*FDR*) au sens de Benjamini & Hochberg (1995) est de 2%. De plus, nous n'avons pu identifier qu'un seul marqueur au-delà du seuil de $\alpha$ ($p \approx 7.10^{-5}$, *cf.* Fig. VI-9c) dans la population Témoin. Si on rapporte ce potentiel faux-positif aux 117 candidats détectés dans les lignées sélectionnées et aux 16370 locus testés, ce résultat confirme la bonne performance statistique de *signasel* dans un cas réel.

Il est intéressant de noter que la proportion de candidats détectés par *signasel* est quasiment identique au sein des lignées divergentes Maigre (0,036%) et Gras (0,039%), qui ont été soumises à une pression de sélection similaire au cours de l'expérience. Sept marqueurs sont identifiés comme candidats dans les deux lignées sélectionnées. De façon intéressante, l'ensemble des 14 allèles concernés présentent des trajectoires opposées d'une lignée à l'autre, signifiant qu'un ou plusieurs locus pourraient eux-mêmes avoir fait l'objet d'une sélection divergente. Nous avons essayé de mapper en priorité ces sept SNP candidats sur l'assemblage du génome de la truite. Trois d'entre eux sont localisées sur le chromosome 26. Parmi ces trois candidats, deux SNP ne sont distants que de 50 kb. Au milieu de cette petite région de 50 kb, on trouve le gène LOC110506579, encodant une aspartate aminotransférase. Cette enzyme impliquée dans le métabolisme est bien connue chez l'Homme, où elle est très fortement exprimée dans les tissus musculaires et hépatiques (Papatheodorou *et al.*, 2017). L'identification d'un gène impliqué dans le métabolisme à proximité de deux SNP ayant des trajectoires divergentes dans les lignées M et G en fait un candidat très intéressant dans le contexte d'une sélection sur la teneur en lipides musculaires.

Ainsi, cette application aux données générées dans le cadre du projet EFFICACE montre que notre méthode permet de distinguer les signaux laissés par la sélection du bruit de fond produit par la dérive au sein de petites populations faisant l'objet une expérience E&R sur une période de moins de 10 générations. Plus généralement, nos résultats soulignent que le génotypage de lignées expérimentales sélectionnées de façon divergente à court terme, comme c'est souvent le cas dans le cadre des expériences de sélection menées chez les espèces d'intérêt agronomique, constitue un matériel pertinent pour examiner l'architecture moléculaire de caractères complexes.

Nous avons affiné la liste de marqueurs candidats par le biais d'autres stratégies de détection de marqueurs potentiellement ciblés par une sélection directionnelle. Un indicateur est le critère bayésien de Gautier *et al.* (2010), qui évalue la compatibilité de la variance des fréquences alléliques observées avec le modèle d'évolution neutre de *KimTree* (Gautier & Vitalis, 2013). Des *PPP-values*

(Gautier *et al.*, 2010) rendent compte de la qualité de l'ajustement du modèle en chaque marqueur. Les 16370 *PPP-values* du jeu de marqueurs examiné sont dans l'ensemble neutres, leur moyenne affichant une valeur très proche de 0,5 (Fig. VI-10a). Néanmoins, leur distribution fait apparaître des *outliers* possédant de faibles valeurs, ce qui est attendu en présence de locus soumis à une sélection directionnelle (Gautier *et al.*, 2010). Soixante-seize marqueurs présentent une *PPP-value* compatible avec une forte sélection directionnelle (Fig. VI-10b).

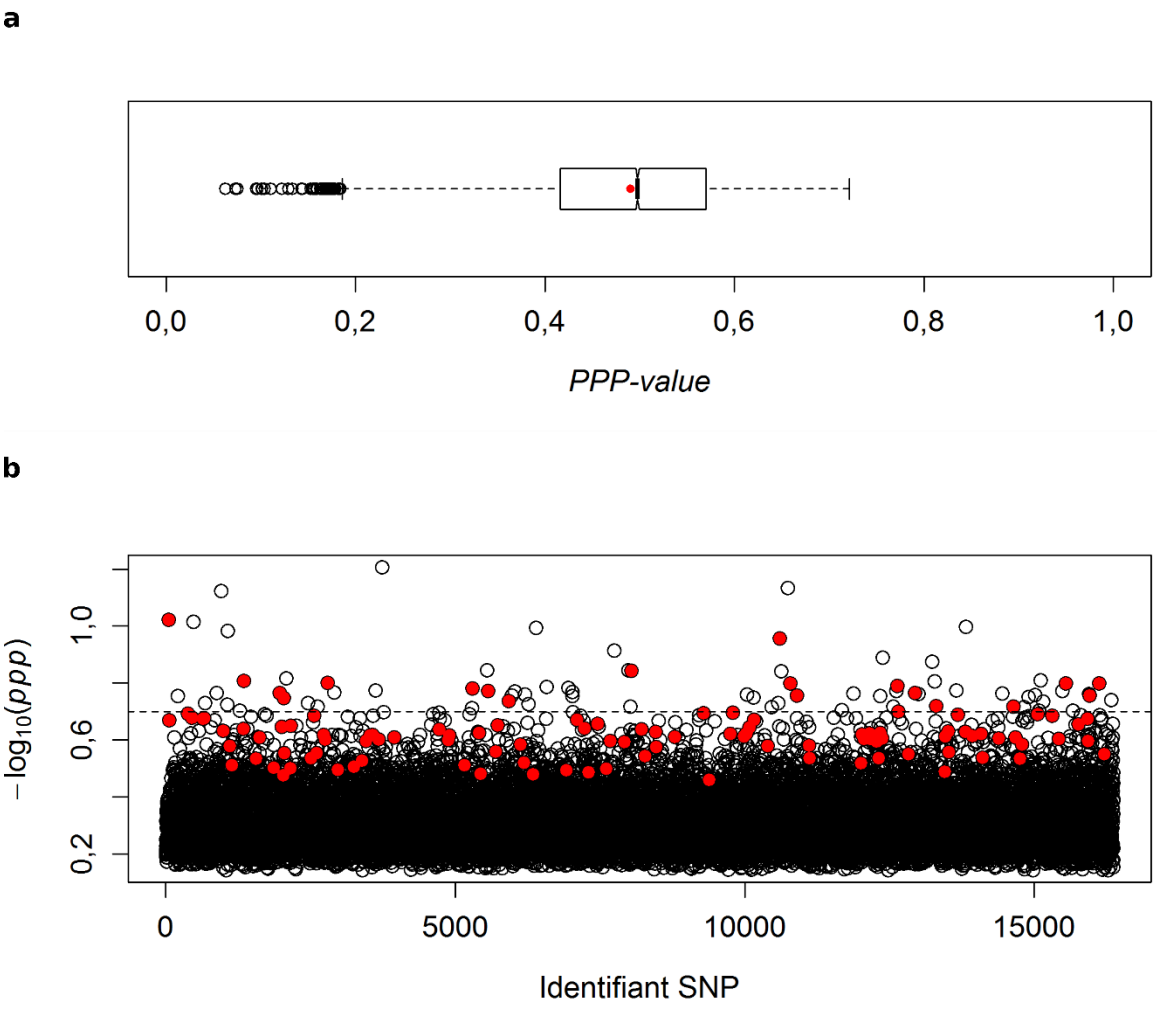



**Figure VI-10. Soixante-seize SNP candidats à la sélection ont été identifiés au moyen de *KimTree* au sein des lignées expérimentales du dispositif EFFICACE.** Les *p-values* postérieures prédictives (*PPP-values*), qui permettent de quantifier la probabilité d'un rôle de la sélection au sein de populations apparentées (Gautier *et al.*, 2010), ont été calculées au moyen de *KimTree* (Gautier & Vitalis, 2013). **(a)** Diagramme de Tukey des 16370 *PPP-values* associées au jeu de marqueurs obtenu par génotypage des lignées expérimentales du dispositif EFFICACE. Le point rouge indique la moyenne des *PPP-values*. **(b)** Le graphique indique en ordonnée $-\log_{10}(ppp)$ où *ppp* désigne la *PPP-value*. Pour être candidat, un SNP doit être associé à $ppp \leq 0,2$ (tirets, $-\log_{10}(ppp) \approx 0,7$), un seuil de significativité qui semble constituer un bon compromis entre puissance et contrôle du taux de faux-positifs dans des conditions proches des celles rencontrées ici (Gautier *et al.*, 2010). Les 117 SNP identifiés par *signasel* sont représentés en rouge. *N.B.* Les SNP sont ordonnés selon un identifiant qui ne préjuge en rien de leur position dans le génome.

Parmi ces candidats, vingt marqueurs avaient été identifiés par *signasel*, dont sept au sein de la population *Maigre* et quinze au sein de la population *Gras* (un seul de ces vingt marqueurs ayant été détecté dans les deux populations). Aucun des 117 candidats identifiés par *signasel* n'a de *PPP-value* supérieure à 0,34. Le critère bayésien de Gautier *et al.* (2010) permet donc d'identifier des candidats à la sélection dans le cadre d'une expérience de sélection divergente de quelques générations.

Le test $F_{LK}$ apporte lui aussi une information intéressante. Dans un premier temps, il s'avère que les statistiques de test $T_{LK}$ et $T_{FLK}$ plafonnent à un niveau similaire et modéré, ce qui se traduit par des *p-values* relativement élevées ($p_{FLK} \geq 2{,}3.10^{-3}$, *cf.* Fig. VI-11) associées à un *FDR* de 100% selon la procédure d'ajustement des *p-values* de Benjamini & Hochberg (1995). La distribution des *p-values* ne permet donc pas de distinguer *a priori* les marqueurs sélectionnés des faux-positifs.

Cependant, il apparaît aussi que les meilleurs candidats selon $T_{FLK}$ appuient les résultats obtenus avec les deux autres méthodes. Ainsi, une vingtaine de marqueurs, tous déjà identifiés comme candidats à la sélection, peuvent être détectés avec un risque de première espèce $\alpha = 0{,}01$ pour le test $F_{LK}$ (Fig. VI-11). Huit d'entre eux sont même détectés de façon conjointe par les trois méthodes (*signasel*, bayésien et $F_{LK}$). La convergence des trois méthodes est intéressante et conforte l'intérêt de ce petit groupe de marqueurs.

En outre, les résultats au test $F_{LK}$ incitent à reconsidérer la priorité accordée à certains candidats possédant une faible *PPP-value*, car ils affichent par ailleurs une faible valeur de $T_{FLK}$, ce qui est incompatible avec une action de la sélection. Nous avons donc écarté de la liste des candidats les marqueurs qui avaient été dans un premier temps signalés sur la base d'une faible *PPP-value* si leur $p_{FLK}$ était supérieure à 0,1 et si la *p-value* du LRT (*signasel*) était supérieure à 0,01 dans les deux lignées divergentes. Après élimination de ces probables faux-positifs, nous proposons une liste de 154 marqueurs candidats dont trente-trois ont été détectés par au moins deux tests.

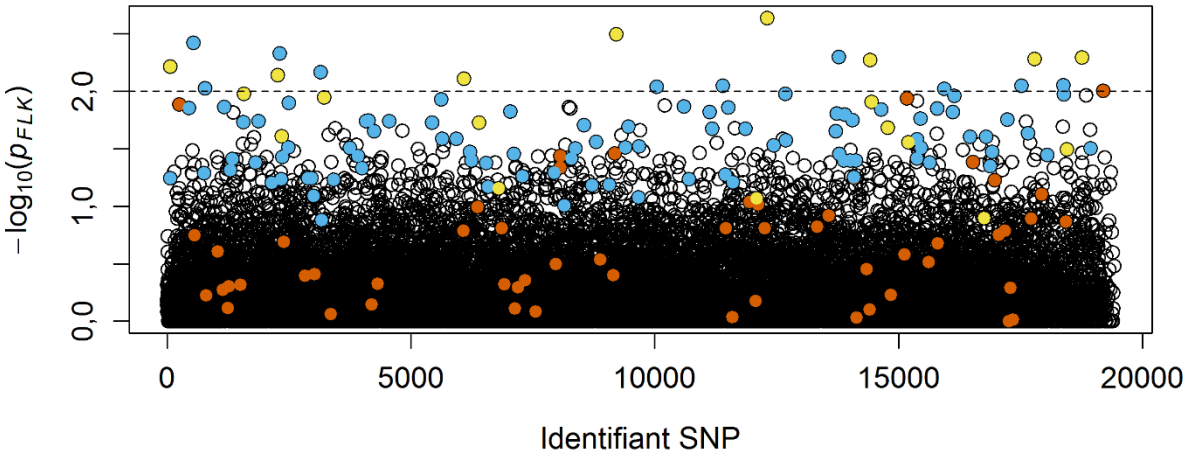



**Figure VI-11. Le test $F_{LK}$ permet d'affiner la liste des candidats à la sélection au sein des lignées expérimentales du dispositif EFFICACE.** Le graphique indique en ordonnée $-\log_{10}(p_{FLK})$ où $p_{FLK}$ désigne la *p-value* du test $F_{LK}$ (Bonhomme *et al.*, 2010). Les tirets représentent le seuil $p_{FLK}$ = 0,01. Les SNP sont colorés selon les résultats des autres tests, ce qui permet de distinguer les candidats détectés par *signasel* (en bleu) de ceux identifiés par *KimTree* (orange) ou conjointement par *signasel* et *KimTree* (jaune). *N.B.* Les SNP sont ordonnés selon un identifiant qui ne préjuge en rien de leur position dans le génome.

# — 4. Discussion

## 4.1. Découverte de marqueurs par RAD-seq

Dans un projet s'appuyant sur un génotypage par RAD-seq, il semble de façon générale plus efficace de former un catalogue de locus à partir de quelques principes directeurs clairs choisis en fonction des contraintes propres à chaque étude. Le jeu de données parfait n'existe pas, et chercher à augmenter coûte que coûte la quantité de marqueurs retenus peut se révéler chronophage et, surtout, contre-productif, car le prix de quelques marqueurs supplémentaires peut être la diminution de la qualité de l'ensemble du jeu de marqueurs.

Dans le cas d'un séquençage de profondeur moyenne comme avec les populations expérimentales de truite du projet EFFICACE, vouloir limiter l'incorporation d'erreurs de séquençage peut inciter à la mise en place de filtres inadaptés. La proportion d'erreurs de séquençage reste la même quelle que soit la profondeur de séquençage. Il est nécessaire d'en tenir compte lors du choix du critère de définition des allèles. Par exemple, être trop strict sur le critère de définition des allèles (*i.e.,* le paramètre *m* de *ustacks*) n'est pas justifié en cas de faible profondeur de séquençage. Nous avons en particulier montré qu'être trop strict sur le choix de *m* pouvait introduire un biais important dans la détection du polymorphisme. Dans ces conditions, notre stratégie de détection du polymorphisme a consisté à choisir un seuil de détection des lectures primaires relativement bas, quitte à ne pas incorporer les lectures secondaires.

D'autres étapes de tri ont été effectuées en aval de l'utilisation *Stacks*. A ce niveau, un choix important concerne la proportion minimale de données manquantes que l'on considère comme acceptable en vue des inférences ultérieures. Nous avons jugé nécessaire d'écarter les marqueurs absents chez au moins deux tiers des individus au sein d'une population ($N_s \geq 20$), afin de limiter le risque d'erreur d'estimation des fréquences alléliques, ce qui aurait un impact sur le résultat de la détection de signatures de sélection en augmentant notamment le taux de faux-positifs. L'ensemble des opérations de contrôle de la qualité des *reads* a abouti à un jeu de données de 16370 marqueurs bialléliques, soit l'ordre de grandeur minimal (≈ 15000 SNP) que nous souhaitions pour tester la présence de signatures de sélection au sein des lignées expérimentales EFFICACE.

## 4.2. Détection de signatures de sélection

Les approches de type $F_{ST}$ *outliers*, fondées sur l'idée qu'une forte différenciation génomique locale peut signaler un évènement sélectif passé, sont fréquemment employées pour la recherche de candidats à la sélection au sein de populations divergentes. Nous avons essayé d'exploiter une

expérience de sélection divergente de sept générations où trois lignées apparentées ont été génotypées en utilisant notamment le test $F_{LK}$ (Bonhomme *et al.*, 2010). Ce dernier améliore le test original de Lewontin & Krakauer (1973) et offre un modèle intéressant dans le cas de nos lignées expérimentales, qui sont de taille relativement petite ($N_e$ < 200), mais qui ont divergé récemment de la population ancestrale. Si les marqueurs associés aux statistiques de test les plus élevées sont certainement de bons candidats dans la mesure où ils sont aussi détectés par d'autres approches, nos résultats mettent cependant en évidence une puissance insuffisante du test $F_{LK}$. Dans la configuration proposée ici, les statistiques de test $T_{LK}$ (Lewontin-Krakauer) et $T_{FLK}$ présentent des performances similaires, avec une puissance légèrement meilleure pour $T_{LK}$. Plusieurs facteurs sont susceptibles de compliquer la détection de signatures de sélection à partir de méthodes de différenciation. Dans le cas présent, la faiblesse de l'effectif efficace ($\widehat{N}_e$ = 106 dans la plus petite population analysée) et le faible nombre de populations analysées (trois populations) expliquent probablement les difficultés rencontrées et la relative supériorité du test de Lewontin-Krakauer sur le $F_{LK}$. Ainsi, nos résultats confirment que le $F_{LK}$ est un test spécifique, utilisable sur des lignées d'intérêt agronomique, mais dont la puissance est limitée si l'effectif efficace est faible.

En plus d'inférer les fréquences alléliques ancestrales*, KimTree* a permis d'évaluer l'ajustement du modèle d'évolution neutre aux données, et d'identifier parmi un grand nombre de SNP les *outliers* qui s'écartent significativement de l'hypothèse de neutralité grâce au calcul des *PPP-values* (Gautier *et al.*, 2010). Une faible *PPP-value* suggère la présence d'une sélection directionnelle. Le choix d'un seuil de significativité n'est pas évident, car, à l'image du test $F_{LK}$, de nombreux facteurs (intensité de la sélection, nombre de populations analysées, effectif efficace, temps de divergence, niveaux de différenciation) peuvent affecter les performances de la stratégie de détection de signatures de sélection implémentée dans *KimTree*. Pour appuyer le choix du seuil de *PPP-value*, Gautier *et al.* (2010) ont examiné l'impact de plusieurs combinaisons de facteurs-clés sur différentes valeurs de seuil (de 0,05 à 0,3) et recommandent un seuil de 0,1 pour contrôler efficacement le taux de faux-positifs dans une grande gamme de scénarios. Nous avons choisi un seuil de 0,2 en nous appuyant sur les simulations disponibles (Gautier *et al.*, 2010), ce qui semble générer davantage de faux-positifs qu'attendu. Il se peut que ces difficultés soient liées au faible $N_e$ des populations analysées. D'après nos résultats, un bon compromis entre puissance et contrôle du taux de faux-positifs serait de choisir un seuil de significativité compris entre 0,1 et 0,2 dans les conditions examinées ici. Le critère bayésien implémenté dans *KimTree* peut ainsi fournir une indication de la proportion de marqueurs sous sélection et du type de sélection agissant sur ces marqueurs. En somme, *KimTree* est un outil complet et intéressant pour exploiter les expériences de sélection divergente comme celles dont il est ici question : il permet, à partir du génotypage d'au moins 5000 marqueurs dans les populations

contemporaines, de retrouver leur histoire démographique, d'estimer les fréquences alléliques de la (ou des) population(s) ancestrale(s), mais aussi de fournir une première idée des locus potentiellement ciblés par la sélection et du type de sélection appliquée.

Les simulations ont montré (chapitre III) que *signasel* permettait de détecter la sélection au sein de populations présentant un faible $N_e$, ce qui semble être le principal facteur limitant les performances des autres approches utilisées dans ce chapitre. L'utilisation de *signasel* nécessite de disposer d'au moins deux échantillons génétiques temporels, idéalement collectés au début et au terme d'une expérience de sélection d'une dizaine de générations. Si la durée de l'expérience de sélection menée dans le cadre du projet EFFICACE semble adaptée (sept générations), nous n'avons accès qu'aux génotypes des populations dérivées. Il a donc été dans un premier temps nécessaire d'inférer l'échantillon temporel ancestral manquant.

Le modèle de *KimTree* offre une solution presque sur mesure pour travailler sur les SNP acquis dans ces populations. Son approche bayésienne hiérarchique a permis de retrouver la généalogie des populations et de proposer une estimation de la fréquence allélique des 16370 SNP analysés en chaque nœud de l'arbre inféré. Disposer d'estimations des fréquences ancestrales a permis d'utiliser *signasel* et d'identifier 117 marqueurs présentant une très faible *p-value* après sept générations de sélection au sein des lignées divergentes.

Ainsi, nous avons pu identifier des patrons de variation génétique correspondant bien à l'histoire évolutive de chaque lignée : une soixantaine de marqueurs sont détectés dans chaque lignée sélectionnée alors que l'on a détecté qu'un seul marqueur dans la lignée T soumise à la seule dérive. Ces résultats confirment que notre méthode est particulièrement intéressante pour identifier des candidats à la sélection dans les petites populations expérimentales. Ils suggèrent également que suivre la dynamique temporelle des fréquences alléliques au sein de chaque lignée expérimentale offrirait de meilleures performances que les méthodes fondées sur des mesures de différenciation. Disposer de séries génétiques temporelles acquises au cours d'expériences de sélection artificielle où un phénotype particulier fait l'objet d'une sélection divergente, un *design* expérimental relativement classique en sciences agronomiques, est donc un avantage pour explorer l'architecture moléculaire des caractères complexes. Dans le cas où aucun échantillon de la population ancestrale ne serait disponible, la combinaison de *KimTree*, pour inférer les fréquences alléliques ancestrales, et de *signasel*, pour sa capacité à détecter des SNP sous sélection positive même lorsque $N_e$ est faible, donne des résultats intéressants.

## 4.3. Conclusion

Nous avons soumis le jeu de données de 16370 SNP générés *de novo* par RAD-seq dans le cadre du dispositif expérimental EFFICACE à trois méthodes permettant d'identifier les traces laissées par une sélection directionnelle récente. Deux méthodes, l'une dans un cadre fréquentiste (*FLK*) et l'autre dans un cadre bayésien (*KimTree*), évaluent dans quelle mesure la divergence entre populations peut être adaptative. La troisième est la stratégie que nous proposons, couplant un modèle de Wright-Fisher à un test du rapport de vraisemblance (*signasel*). En combinant les résultats issus de l'ensemble de ces tests, nous avons identifié 154 SNP candidats à une sélection directionnelle pour la teneur en lipides du muscle, un bon prédicteur de l'efficacité alimentaire chez la truite arc-en-ciel.

Dans le cadre de ce projet, notre rôle se limitait à la fourniture de la liste des marqueurs candidats. Ceux-ci vont être alignés sur l'assemblage du génome de la truite arc-en-ciel et une analyse de la fonction putative des gènes identifiés en leur voisinage sera menée par des spécialistes de l'efficacité alimentaire chez les poissons d'aquaculture. Ces résultats seront comparés aux données transcriptomiques acquises dans les lignées M et G, et, plus largement, aux données fonctionnelles obtenues pour des caractères corrélés à l'efficacité alimentaire chez une espèce marine, le bar (*Dicentrarchus labrax*). A moyen terme, l'ambition est de cibler les mécanismes impliqués dans les variations d'efficacité alimentaire chez les poissons d'élevage et, *in fine*, de proposer des marqueurs utilisables en sélection.

# — 5. Références


ABCIS & INRA (2015). Outils et leviers pour favoriser le développement d'une génétique animale adaptée aux enjeux de l'agroécologie (dir. Brochard, M., & Phocas, F.). Rapport final de l'étude n° SSP-2014-061.

Aggrey, S. E., & Rekaya, R. (2013). Dissection of Koch's residual feed intake: Implications for selection. Poultry science, 92(10), 2600-2605.

Akey, J. M., Ruhe, A. L., Akey, D. T., Wong, A. K., Connelly, C. F., Madeoy, J., ... & Neff, M. W. (2010). Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences*, *107*(3), 1160-1165.

Allaire, G. & Daviron, B. (2017). Transformations agricoles et agroalimentaires: Entre écologie et capitalisme. 429 p. Editions Quae. ISBN 978-2-7592-2615-3.

Atzori, A.S., Ford, D., Tedeschi, L., Cannas, C. (2012). Policy modeling for greenhouses gas emissions on dairy cattle sector: the importance of milk improvement, 30th International Conference of the Sustem Dynamics Society, St. Gallen, Switzerland, July 22-26, 2012.

Aubin, J., Papatryphon, E., Van der Werf, H. M. G., & Chatzifotis, S. (2009). Assessment of the environmental impact of carnivorous finfish production systems using life cycle assessment. Journal of Cleaner Production, 17(3), 354-361.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3(10), e3376.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., ... & Kent, M. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, *528*(7582), 405-408.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

Besson, M., Aubin, J., Komen, H., Poelman, M., Quillet, E., Vandeputte, M., & De Boer, I. J. M. (2016). Environmental impacts of genetic improvement of growth rate and feed conversion ratio in fish farming under rearing density and nitrogen output limitations. Journal of Cleaner Production, 116, 100-109.

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

Bordas, A., Tixier-Boichard, M., & Mérat, P. (1992). Direct and correlated responses to divergent selection for residual food intake in Rhode Island Red laying hens. *British poultry science*, *33*(4), 741-754.

Brochard, M., Boichard, D., Ducrocq, V., & Fritz, S. (2013). La sélection pour des vaches et une production laitière plus durables: acquis de la génétique et opportunités offertes par la sélection génomique. INRA Prod. Anim, 26(2), 145-156.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7(4), 434-455.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). *Stacks*: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1(3), 171-182.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

Catchen Lab (2017). Tutorial: building mini-contigs from paired-end sequences [En Ligne] http://catchenlab.life.illinois.edu/stacks/param_tut.php (consulté le 14/03/2018).

CIPA (Comité Interprofessionnel des Produits de l'Aquaculture, 2017). Les essentiels de l'aquaculture. L'alimentation. [En ligne] http://www.poisson-aquaculture.fr/l%E2%80%99alimentation (consulté le 15/03/2018).

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434), 883-904.

Davidson, E. A., David, M. B., Galloway, J. N., Goodale, C. L., Haeuber, R., Harrison, J. A., ... & Peel, J. L. (2012). Excess nitrogen in the US environment: trends, risks, and solutions. Issues in Ecology 15, Ecological Society of America, Ithaca, New York.

de Verdal, H., Mekkawy, W., Lind, C. E., Vandeputte, M., Chatain, B., & Benzie, J. A. (2017). Measuring individual feed efficiency and its correlations with performance traits in Nile tilapia, Oreochromis niloticus. Aquaculture, 468, 489-495.

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources, 14(1), 209-214.

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., & Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing *reads*. PloS one, 6(4), e18561.

FAO (1996). Déclaration de Rome sur la sécurité alimentaire mondiale. Sommet mondial de l'alimentation, 23-27 Nov. 1996. [En Ligne] Archives de la FAO http://www.fao.org/docrep/003/w3613f/w3613f00.htm (consulté le 30/03/2018).

FAO (2016). The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all. Rome. 200 p. ISBN 978-92-5-109185-2.

Fariello, M. I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., San Cristobal, M., ... & International Sheep Genomics Consortium. (2014). Selection signatures in worldwide sheep populations. *PLoS One*, *9*(8), e103813.

Fletcher, D. J. (1984). The physiological control of appetite in fish. Comparative Biochemistry and Physiology Part A: Physiology, 78(4), 617-628.

Galloway, J. N., Townsend, A. R., Erisman, J. W., Bekunda, M., Cai, Z., Freney, J. R., ... & Sutton, M. A. (2008). Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. Science, 320(5878), 889-892.

Garreau, H., Brun, J. M., Theau-Clement, M., & Bolet, G. (2008). Evolution des axes de recherche à l'INRA pour l'amélioration génétique du lapin de chair. INRA Prod. Anim, 21(3), 269-276.

Gautier, M., Hocking, T. D., & Foulley, J. L. (2010). A Bayesian outlier criterion to detect SNPs under selection in large data sets. PloS one, 5(8), e11913.

Gautier, M., & Vitalis, R. (2013). Inferring population histories using genome-wide allele frequency data. Molecular biology and evolution, 30(3), 654-668.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report No. 462, Dept. of Statistics, Stanford Univ.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science, 457-472.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, 733-760.

Gholami, M., Reimer, C., Erbe, M., Preisinger, R., Weigend, A., Weigend, S., ... & Simianer, H. (2015). Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. *PloS one*, *10*(7), e0130497.

Gilbert, H., Billon, Y., Brossard, L., Faure, J., Gatellier, P., Gondret, F., ... & Louveau, I. (2017). Review: divergent selection for residual feed intake in the growing pig. Animal, 1-13.

Gilbert, K. J., & Whitlock, M. C. (2015). Evaluating methods for estimating local effective population size with and without migration. Evolution, 69(8), 2154-2166.

GIS Elevages demain (2017). [En ligne] https://www.gis-elevages-demain.org/Actions-thematiques (consulté le 30/03/2018).

Gjedrem, T., Robinson, N., & Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: a review. Aquaculture, 350, 117-129.

Glasauer, S. M., & Neuhauss, S. C. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular genetics and genomics*, *289*(6), 1045-1060.

Grima, L., Quillet, E., Boujard, T., Robert-Granié, C., Chatain, B., & Mambrini, M. (2008). Genetic variability in residual feed intake in rainbow trout clones and testing of indirect selection criteria. Genet. Sel. Evol, 40, 607-624.

Grima, L. (2010). Vers une amélioration de l'efficacité alimentaire chez le poisson. Thèse de doctorat en Sciences de la vie et de la santé. AgroParisTech.

Grima, L., Vandeputte, M., Ruelle, F., Vergnet, A., Mambrini, M., & Chatain, B. (2010a). In search for indirect criteria to improve feed utilization efficiency in sea bass (Dicentrarchus labrax): Part I: Phenotypic relationship between residual feed intake and body weight variations during feeding deprivation and re-feeding periods. Aquaculture, 300(1), 50-58.

Grima, L., Chatain, B., Ruelle, F., Vergnet, A., Launay, A., Mambrini, M., & Vandeputte, M. (2010b). In search for indirect criteria to improve feed utilization efficiency in sea bass (Dicentrarchus labrax): Part II: Heritability of weight loss during feed deprivation and weight gain during re-feeding periods. Aquaculture, 302, 169-174.

Guo, F., Dey, D. K., & Holsinger, K. E. (2009). A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. Journal of the American Statistical Association, 104(485), 142-154.

Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., ... & Fischer-Kowalski, M. (2007). Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. Proceedings of the National Academy of Sciences, 104(31), 12942-12947.

Hardy, R. W., & Barrows, F. T. (2002). Chapter 9 - Diet Formulation and Manufacture, dans Fish Nutrition 3e édition (dir. Halver, J. E., & Hardy, R. W.), pp. 505-600. Academic press. ISBN 978-0-12-319652-1.

Herrero, M., Havlík, P., Valin, H., Notenbaert, A., Rufino, M. C., Thornton, P. K., ... & Obersteiner, M. (2013). Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. Proceedings of the National Academy of Sciences, 110(52), 20888-20893.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet, 6(2), e1000862.

Institut de l'Elevage (Idele) & INRA (2011). La révolution génomique animale, 161 pp, France Agricole, 978-2-85557-181-2.

INRA (2017). Dictionnaire d'agroécologie (dir. Dedieu, B., & Batifol-Garandel, V.). [En ligne] http://dicoagroecologie.fr/ (consulté le 30/03/2018).

Janssen, K., Berentsen, P., Besson, M., & Komen, H. (2017). Derivation of economic values for production traits in aquaculture species. Genetics Selection Evolution, 49(1), 5.

Jiang, X., Dey, D. K., Prunier, R., Wilson, A. M., & Holsinger, K. E. (2013). A new class of flexible link functions with application to species co-occurrence in Cape floristic region. The Annals of Applied Statistics, 7(4), 2180-2204.

Jiménez-Mena, B. (2016). Estimation of the effective population size (Ne) and its application in the management of small populations. Thèse de Doctorat en Génétique Animale. AgroParisTech.

Kaiser, M., & Algers, A. (2016). Food ethics: a Wide Field in Need of Dialogue. Food Ethics, 1(1), 1-7.

Kause, A., Tobin, D., Dobly, A., Houlihan, D., Martin, S., Mäntysaari, E. A., Ritola, O., & Ruohonen, K. (2006). Recording strategies and selection potential of feed intake measured using the X-ray method in rainbow trout. Genetics Selection Evolution, 38(4), 1-21.

Kause, A., Quinton, C., Ruohonen, K., & Koskela, J. (2008). Selection potential for feed efficiency in farmed salmonids. Back to Nature, 13(3), 20.

Kause, A., Kiessling, A., Martin, S. A., Houlihan, D., & Ruohonen, K. (2016). Genetic improvement of feed conversion ratio via indirect selection against lipid deposition in farmed rainbow trout (Oncorhynchus mykiss Walbaum). British Journal of Nutrition, 116(9), 1656-1665.

Kimura, M. (1964). Diffusion models in population genetics. Journal of Applied Probability, 1(2), 177-232.

Le Boucher, R., Quillet, E., Vandeputte, M., Lecalvez, J. M., Goardon, L., Chatain, B., ... & Dupont-Nivet, M. (2011). Plant-based diet in rainbow trout (Oncorhynchus mykiss Walbaum): Are there genotype-diet interactions for main production traits when fish are fed marine vs. plant-based diets from the first meal ? Aquaculture, 321, 41-48.

Lewis, P. O., Xie, W., Chen, M. H., Fan, Y., & Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. Systematic biology, 63(3), 309-321.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. Methods in Ecology and Evolution, 3(1), 112-115.

Macdiarmid, J. I., Douglas, F., & Campbell, J. (2016). Eating like there's no tomorrow: Public awareness of the environmental impact of food and reluctance to eat less meat as part of a sustainable diet. Appetite, 96, 487-493.

Madrid, J. A., Azzaydi, M., & Zamora, S. (1997). Continuous recording of uneaten food pellets and demand-feeding activity: a new approach to studying feeding rhythms in fish. Physiology & behavior, 62(4), 689-695.

Mauch, E., Serão, N., Young, J., Patience, J., Gabler, N., & Dekkers, J. (2017). Diet by Genotype Interaction in Yorkshire Pigs Divergently Selected for Feed Efficiency. Animal Industry Report, 663(1), 57.

McCarthy, I. D., Carter, C. G., & Houlihan, D. F. (1992). The effect of feeding hierarchy on individual variability in daily feeding of rainbow trout, Oncorhynchus mykiss (Walbaum). Journal of Fish Biology, 41(2), 257-263.

Moradi, M. H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K. G., & McEwan, J. C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC genetics*, *13*(1), 10.

Naylor, R. L., Goldburg, R. J., Primavera, J. H., Kautsky, N., Beveridge, M. C., Clay, J. & Troell, M. (2000). Effect of aquaculture on world fish supplies. Nature, 405(6790), 1017-1024.

Nikolic, N., & Chevalet, C. (2014). Detecting past changes of effective population size. Evolutionary applications, 7(6), 663-681.

OCDE/FAO (2016). OECD-FAO Agricultural Outlook 2016-2025, OECD Publishing, Paris. http://dx.doi.org/10.1787/agr_outlook-2016-en ISBN 978-92-64-25323-0.

Palti, Y., Genet, C., Luo, M. C., Charlet, A., Gao, G., Hu, Y., ... & Vallejo, R. L. (2011). A first generation integrated map of the rainbow trout genome. *BMC genomics*, *12*(1), 180.

Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., ... & Huerta, L. (2017). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research*, *46*(D1), D246-D251.

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., ... & Brama, P. (2013). Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS genetics*, *9*(1), e1003211.

Peyraud, J.-L., Cellier, P., Donnars, C., Réchauchère, O. (2012). Les flux d'azote liés aux élevages, réduire les pertes, rétablir les équilibres. Expertise scientifique collective, synthèse du rapport, INRA (France), 68 p.

Phocas, F., Brochard, M., Larroque, H., Lagriffoul, G., Labatut, J., & Guerrier, J. (2013). Etat actuel et perspectives d'évolution des objectifs de sélection chez les ruminants. Rencontres autour des recherches sur les ruminants, 129-132.

Phocas, F., Agabriel, J., Dupont-Nivet, M., Geurden, I., Medale, F., Mignon-Grasteau, S., ... & Dourmad, J. Y. (2014). Le phénotypage de l'efficacité alimentaire et de ses composantes, une nécessité pour accroître l'efficience des productions animales. INRA Productions Animales, 27(3), 235-248.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. R news, 6(1), 7-11.

Quillet, E., Le Guillou, S., Aubin, J., & Fauconneau, B. (2005). Two-way selection for muscle lipid content in pan-size rainbow trout (Oncorhynchus mykiss). *Aquaculture*, *245*(1-4), 49-61.

Quillet, E., Le Guillou, S., Aubin, J., Labbé, L., Fauconneau, B., & Médale, F. (2007). Response of a lean muscle and a fat muscle rainbow trout (Oncorhynchus mykiss) line on growth, nutrient utilization, body composition and carcass traits when fed two different diets. Aquaculture, 269(1), 220-231.

Quinton, C. D., Kause, A., Ruohonen, K., & Koskela, J. (2007). Genetic relationships of body composition and feed utilization traits in European whitefish (L.) and implications for selective breeding in fishmeal-and soybean meal-based diet environments. Journal of animal science, 85(12), 3198-3208.

Reichwaldt, E. S., Stone, D., Barrington, D. J., Sinang, S. C., & Ghadouani, A. (2016). Development of toxicological risk assessment models for acute and chronic exposure to pollutants. Toxins, 8(9), 251.

Rexroad, C. E., & Vallejo, R. L. (2009). Estimates of linkage disequilibrium and effective population size in rainbow trout. BMC genetics, 10(1), 83.

Rothammer, S., Seichter, D., Förster, M., & Medugorac, I. (2013). A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC genomics*, *14*(1), 908.

Rothschild, M. F., Hu, Z. L., & Jiang, Z. (2007). Advances in QTL mapping in pigs. International journal of biological sciences, 3(3), 192.

Sae-Lim, P., Komen, H., Kause, A., Van Arendonk, J., Barfoot, A. J., Martin, K. E., & Parsons, J. E. (2012). Defining desired genetic gains for rainbow trout breeding objective using analytic hierarchy process. Journal of animal science, 90(6), 1766-1776.

Schaschl, H., Huber, S., Schaefer, K., Windhager, S., Wallner, B., & Fieder, M. (2015). Signatures of positive selection in the cis-regulatory sequences of the human oxytocin receptor (OXTR) and arginine vasopressin receptor 1a (AVPR1A) genes. *BMC evolutionary biology*, *15*(1), 85.

Sell-Kubiak, E., Wimmers, K., Reyer, H., & Szwaczkowski, T. (2017). Genetic aspects of feed efficiency and reduction of environmental footprint in broilers: a review. Journal of Applied Genetics, 1-12.

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics, 194(3), 647-662

Silverstein, J. T. (2006). Relationships among feed intake, feed efficiency, and growth in juvenile rainbow trout. North American Journal of Aquaculture, 68(2), 168-175.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583-639.

Tataru, P., Simonsen, M., Bataillon, T., & Hobolth, A. (2016). Statistical inference in the Wright-Fisher model using allele frequency data. Systematic Biology, syw056.

Townsend, A., & Howarth, R. (2010). Réduire la pollution par l'azote. Pour la science, 393, 54-60.

Verrier, E. (2013). Bases génétiques de la résistance aux rhabdovirus et réponse cellulaire chez la truite arc-en-ciel : importance des mécanismes de défense innés. [en ligne] Thèse de doctorat en Sciences de la Vie et de la Santé, AgroParisTech, 2013, 282 p. https://pastel.archives-ouvertes.fr/pastel-00914894

Volk, M., Bungener, P., Contat, F., Montani, M., & Fuhrer, J. (2006). Grassland yield declined by a quarter in 5 years of free-air ozone fumigation. Global Change Biology, 12(1), 74-83.

von Braun, J., (2010) The role of livestock production for a growing world population, 9th World Congress of Genetics Applied to Livestock Production, Leipzig, Germany, August 1-6, 2010.

Wang, J. (2016). A comparison of single-sample estimators of effective population sizes from genetic marker data. Molecular ecology, 25(19), 4692-4711.

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. Heredity.

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics, 7(2), 167.

Waples, R. S., & Do, C. (2010). Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*(3), 244-262.

Waples, R. S. (2016). Making sense of genetic estimates of effective population size. Molecular Ecology, 25(19), 4689-4691.

Westhoek, H., Rood, T., van den Berg, M., Janse, J., Nijdam, D., Reudink, M. & Stehfest, E. (2011). The Protein Puzzle: The consumption and production of meat, dairy and fish in the European Union (No. 500166001). Netherlands Environmental Assessment Agency. ISBN: 978-90-78645-61-0.

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, *76*(5), 887-893.

Yang, S., Li, X., Li, K., Fan, B., & Tang, Z. (2014). A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC genetics*, *15*(1), 7.

Young, W. P., Wheeler, P. A., Coryell, V. H., Keim, P., & Thorgaard, G. H. (1998). A detailed linkage map of rainbow trout produced using doubled haploids. Genetics, 148(2), 839-850.

# Chapitre VII

# TABLE DES MATIERES

-

# Chapitre VII – Discussion générale

## — 1. Inférer la sélection en temps réel au sein de petites populations

Si depuis le début de la décennie 2010 le potentiel des expériences E&R a été largement souligné (Turner *et al.*, 2011 ; Long *et al.*, 2015 ; Schlötterer *et al*., 2015), peu de travaux ont été spécifiquement consacrés au cas des petites populations soumises à une sélection directionnelle intense sur une période de quelques générations à quelques dizaines de générations. Pourtant, ce régime de sélection concerne des populations dont le suivi temporel est potentiellement riche en enseignements (*e.g.,* les lignées de populations domestiques sélectionnées en conditions contrôlées pour des caractères complexes, les populations férales ou naturelles soumises à une nouvelle donne écologique). Dans cette thèse, nous avons proposé, caractérisé et appliqué une méthode permettant de détecter des signatures de sélection au sein de ces populations.

Nous avons proposé une méthode de maximum de vraisemblance couplée à un modèle de Wright-Fisher. L'application de cette méthode à des données simulées puis réelles a montré qu'elle permettait de détecter les locus sous sélection directionnelle – ou liés à des locus sélectionnés – même parmi les variants déjà communs dans la population au moment où la sélection a débuté. Cette caractéristique est très importante, car elle indique une capacité de notre méthode à traiter avec succès les cas d'adaptation rapide, où la sélection s'appuie exclusivement sur la variation génétique préexistante (Barrett & Schluter, 2008). Ainsi, nos analyses de données simulées et réelles montrent que notre méthode peut détecter des empreintes laissées par la sélection sur une période très courte (moins de dix générations), si la sélection est particulièrement forte (cas du diable de Tasmanie, *cf.* Chapitre IV) ou si la dérive est relativement modérée (cas de la truite arc-en-ciel, *cf.* Chapitre VI).

Pris ensemble, nos résultats soulignent l'intérêt de collecter des données génomiques longitudinales. Celles-ci offrent la possibilité d'explorer le déterminisme génétique de caractères complexes, même si la durée de l'expérience n'est que de quelques générations et même si l'effectif efficace est très faible, des points qui peuvent décourager à tort l'examen de petites populations expérimentales présentant par ailleurs un fort intérêt scientifique. Notre méthode a permis l'identification de plusieurs centaines de SNP sous sélection directionnelle pour des caractères quantitatifs : la résistance au cancer chez le diable de Tasmanie et la teneur en lipides musculaires chez la truite arc-en-ciel. Ces résultats indiquent

que les données E&R permettent d'identifier la présence de sélection dans le contexte d'une expérience planifiée de sélection artificielle, mais aussi lorsqu'elles ont été acquises au cours d'une « expérience de sélection naturelle ».

Si ces résultats sont prometteurs, il faut tout de même rappeler que les régions génomiques identifiées lors de nos analyses de données réelles ne sont à ce stade que des régions candidates, qui présentent une évolution à court terme statistiquement imputable à une sélection directionnelle, mais qui doivent être validées en s'appuyant sur d'autres études (*e.g.,* analyses comparatives, reséquençage de régions d'intérêt ou du transcriptome, cartographie fine, validation fonctionnelle) afin d'éventuellement identifier un polymorphisme causal.

## — 2. Des petites populations fortement contraintes mais évolvables

L'évolvabilité désigne un concept équivoque rassemblant des chercheurs d'horizons très divers (Hendrikse *et al.*, 2007 ; Pigliucci, 2008 ; Brookfield, 2009 ; Agarwal, 2013 ; Brown, 2013 ; Lehman, 2017). Dans une acception large, l'évolvabilité peut être vue comme la capacité des organismes à répondre à la sélection. Chercher à comprendre dans quelle mesure les populations sont capables de tirer parti d'une variation génétique préexistante pour s'adapter aux nouvelles conditions imposées par leur environnement a été un fil conducteur durant mon parcours étudiant. C'est de plus une thématique qui montre à quel point les travaux, y compris fondamentaux, menés en génétique évolutive sont en phase avec les attentes du public, dans un contexte où l'on s'interroge à la fois sur la capacité des organismes à faire face à un environnement changeant et sur notre capacité à prédire leur destin (Chevin *et al.*, 2010). A court terme, l'évolvabilité dépend « de la variation génétique préexistante sur laquelle la sélection peut agir » (Hansen & Pélabon, 2018). Nos résultats montrent qu'une sélection directionnelle intense pourrait cibler de nombreux gènes ayant une valeur adaptative et ségrégeant au sein de petites populations.

La capacité de populations expérimentales de truite, comme celles étudiées dans le cadre du projet EFFICACE, à répondre à une sélection sur la teneur en lipides du muscle est documentée au niveau phénotypique (Quillet *et al.*, 2005 ; Kause *et al.*, 2016). Si l'on sait que l'héritabilité des caractères reliés à l'adiposité corporelle est élevée chez la truite (Kause *et al.*, 2016 ; de Verdal *et al.*, 2017), l'impact d'une sélection sur ce type de caractère à l'échelle génomique est en revanche peu documenté. Notre analyse de signatures de sélection a montré qu'environ 150 SNP pouvaient être ciblés par une sélection artificielle de moins de dix générations sur la teneur en lipides musculaires chez la truite, ce qui est en faveur d'un déterminisme polygénique. Un tel caractère est probablement régulé par des voies de synthèse des lipides et de l'ATP, c'est-à-dire un ensemble de processus biochimiques ramifiés et hautement interconnectés impliquant des nombreux gènes. Identifier un grand nombre de SNP dans ce contexte n'est donc pas surprenant, mais signifie toutefois que la variation génétique dont bénéficie du point de vue fonctionnel la sélection sur des caractères quantitatifs peut être importante et disponible immédiatement, contrairement à ce que l'on peut supposer pour des petites populations. Ces résultats viennent étoffer une littérature en expansion sur l'identification de signatures de sélection pour des caractères quantitatifs consécutivement à l'action d'une sélection expérimentale. Nous retrouvons ici les mêmes résultats que chez des populations animales de laboratoire plus classiques (*e.g.,* drosophile, rat) présentant des effectifs efficaces et un nombre de générations d'évolution généralement supérieurs (Remolina *et al.*, 2012 ; Lo *et al.*, 2016).

Les perturbations environnementales subies par les populations sont plus complexes en milieu naturel qu'en laboratoire. C'est pourquoi l'étude des expériences de sélection naturelle, qui intègrent cette complexité, est complémentaire des travaux d'évolution expérimentale effectués au laboratoire, mieux contrôlés. En particulier, l'introduction, intentionnelle ou non, de populations dans un nouvel environnement naturel, auquel elles vont devoir rapidement s'adapter, représente un « point de comparaison écologique aux études de laboratoire qui manipulent artificiellement la dynamique des populations » (Irschick & Reznick, 2009). J'ai eu l'opportunité d'étudier un tel cas d'adaptation rapide en milieu naturel consécutivement à l'introduction de la carpe commune (*Cyprinus carpio*) à Madagascar. Ce travail a commencé lors de mon Master 2, puis a été poursuivi durant ma thèse en parallèle des travaux sur la détection de signatures de sélection pour aboutir à une publication dans *Proceedings B* (Hubert *et al.*, 2016, *cf.* Annexe III).

La carpe a été initialement introduite sur l'île de Madagascar en 1912 à partir d'individus issus de pisciculture françaises dont une particularité phénotypique visible était l'absence d'écailles – la présence d'écailles a été contre-sélectionnée dans les populations domestiques de carpe, voir par exemple Balon (2014). Au bout de quelques dizaines de générations dans les eaux malgaches, une forme écailleuse est devenue prédominante (Kiener, 1958) et les populations contemporaines que nous avons pu observer sont constituées en quasi-totalité d'individus « néo-écaillés », qui arborent une écaillure aux motifs irréguliers mais recouvrant leurs flancs. Par ailleurs, tous les poissons prélevés au sein du milieu naturel que nous avons génotypés étaient porteurs d'une mutation connue (Rohner *et al.*, 2009) pour induire l'absence d'écailles. Tout cela suggère que la néo-écaillure des carpes férales malgaches est fonctionnellement comparable à celle des poissons sauvages, mais repose sur une base génétique distincte – et probablement polygénique – du système de détermination de l'écaillure précédemment décrit chez la carpe (Kirpitchnikov, 1999). Nous avons effectué des analyses complémentaires qui indiquent que les carpes férales de Madagascar forment une population génétiquement isolée et, surtout, que leur nombre d'écailles est un caractère quantitatif dont la base génétique est héritable. Pris ensemble, nos résultats appuient l'idée d'une compensation polygénique rapide (en l'espace de moins de 40 générations) de la fixation d'un gène majeur non fonctionnel dans la population fondatrice. La restauration d'une couverture écaillée fonctionnelle suit donc une nouvelle trajectoire adaptative chez la carpe férale malgache, ce qui signifie que la sélection naturelle a pu cibler, parmi la variation génétique cryptique présente dans les populations, un groupe de locus permettent de compenser les effets délétères de l'absence d'écailles en milieu naturel.

Chez le diable de Tasmanie, la réponse évolutive à la sélection imposée par la DFTD représente un cas unique d'expérience de sélection naturelle fortuite. L'apparition d'un cancer transmissible au sein de populations naturelles de très petite taille efficace, présentes uniquement sur une île d'une superficie

comparable à celle d'une région française, est une situation singulière à plus d'un titre. Malgré un déclin démographique brutal et considérable, nous avons pu identifier un grand nombre de traces de sélection fonctionnellement intéressantes chez les trois populations étudiées, en suivant l'évolution des fréquences alléliques de quelques milliers de SNP depuis l'émergence de la DFTD. Ceci indique que chacune de ces petites populations de diable de Tasmanie hébergeait une variation génétique suffisante pour permettre une action de la sélection à très court terme dans le contexte d'une perturbation écologique majeure (*i.e.,* l'émergence d'un cancer infectieux). Il est notamment intéressant de constater que la sélection naturelle, qui est le produit de compromis fonctionnels complexes en milieu naturel (Kokko *et al*., 2017), a pu mobiliser en réponse à la DFTD une large base génétique susceptible d'influencer plusieurs phénotypes complexes.

Les expériences de sélection que nous avons étudiées montrent que des populations de petite taille parviennent à s'adapter rapidement, en utilisant la variation génétique disponible, à des conditions nouvelles affectant certains caractères quantitatifs. Les données empiriques fournies par ces expériences sont importantes pour améliorer notre compréhension de l'architecture moléculaire de caractères complexes et de la façon dont la sélection intègre les contraintes endogènes (*e.g.,* variation génétique disponible, épistasie) et exogènes (*e.g.,* dynamique des populations, interactions interspécifiques).

## — 3. Des bisbilles autour du RAD-seq…

Un point commun aux jeux de données réelles examinés dans cette thèse est qu'ils dérivent d'un génotypage par RAD-seq. Comme nous l'avons vu au Chapitre V, les techniques de séquençage d'un grand nombre de *reads* courts comme le RAD-seq mettent en théorie la génomique au service d'un grand nombre de populations, mais se traduisent dans la pratique par des protocoles expérimentaux qui peuvent être lourds, et par des analyses bioinformatiques n'ayant rien de trivial et nécessitant des compromis (*cf.* Chapitre VI). Le jeu de données utilisé pour la recherche de signatures de sélection chez le diable de Tasmanie (*cf.* Chapitre IV) présentait par exemple une variation importante (jusqu'à un facteur 5) en densité de SNP d'une population à l'autre, illustrant bien les difficultés que peut occasionner l'utilisation de données dérivées de RAD-seq.

Les difficultés techniques rencontrées par les utilisateurs de protocoles de RAD-seq ont alimenté des discussions passionnées dans les journaux spécialisés en génétique moléculaire (*e.g.,* Molecular Ecology Resources) entre les tenants d'une vision très critique des performances de ces protocoles (*e.g.,* Puritz *et al.*, 2014 ; Tiffin & Ross-Ibara, 2014 ; Lowry *et al.*, 2017a, 2017b) et ceux qui soulignent leurs apports (*e.g.,* Andrews *et al.*, 2014 ; Catchen *et al.*, 2017 ; McKinney *et al.*, 2017). Si les termes employés ont parfois été rudes, les uns titrant « Démystifier l'engouement pour le RAD » ou « *Breaking RAD* » et étalant leurs réticences vis-à-vis du RAD-seq, tandis que les autres insistaient sur les erreurs ou le caractère « anecdotique » des griefs formulés par les premiers, ces débats ont aussi eu le mérite de porter sur la place publique certaines questions d'intérêt général pour la communauté des généticiens.

Les arguments échangés de part et d'autre ont tout d'abord eu le mérite de rappeler que les études génomiques ayant pour objectif la recherche de signatures de sélection sont des entreprises qui restent imparfaites, d'autant plus lorsque les conditions ne sont pas favorables (*e.g.,* faible taille des blocs de DL, génome de référence indisponible ou insuffisamment annoté). Si ces conditions ne sont pas imputables aux méthodes de génotypage, il est tout de même important de reconnaître que le RAD-seq produit une quantité de marqueurs variant grandement d'une expérience à l'autre (Catchen *et al.*, 2017) et de s'interroger en amont du séquençage sur ce que cela peut impliquer (Lowry *et al.*, 2017b). Il est ainsi légitime d'attendre des équipes amenées à rechercher des empreintes génomiques laissées par la sélection à partir de marqueurs dérivés de RAD-seq le choix d'un protocole *a priori* compatible avec ce que l'on sait ou ce que l'on peut estimer du DL ou bien, *a minima*, une communication transparente sur la fraction du génome que l'on s'attend à échantillonner. Des outils comme les scripts R proposés dans les données supplémentaires de Lowry *et al.* (2017a) permettent d'estimer simplement la fraction de génome couverte en prévision d'une expérience donnée. Le

résultat de ce type de simulations ne doit toutefois pas conduire à proscrire l'usage du RAD-seq chez les populations dans lesquelles le DL est insuffisant. Il faut en particulier noter que le DL peut être plus important au sein des régions sélectionnées, ce qui peut faciliter leur découverte (McKinney *et al.* 2017).

Si le RAD-seq ne permet pas une couverture suffisante pour échantillonner de façon exhaustive les très gros génomes soumis à des évènements de recombinaison fréquents, un de ses principaux avantages est sa flexibilité (Andrews *et al.*, 2016 ; McKinney *et al.*, 2017), ce qui autorise malgré tout l'examen de systèmes *a priori* peu faciles d'accès. Les exemples ne manquent pas de protocoles ajustés aux besoins de systèmes particuliers dont on peut s'inspirer lors de la planification d'une expérience. Par exemple, Yang, G.Q., *et al.* (2016) suggèrent des protocoles facilitant l'étude des populations d'angiospermes, Grewe *et al.* (2017) proposent une stratégie adaptée au cas des analyses métagénomiques, *etc.* La littérature laisse apparaître de nombreuses expériences bien conduites, qui ont permis de générer des groupes de quelques dizaines de milliers de marqueurs biologiquement robustes issus de RAD-seq chez des espèces non-modèles, et qui ont entre autres débouché sur la détection de régions génomiques vraisemblablement impliquées dans l'adaptation locale (Chaves *et al.*, 2016 ; Babin *et al.*, 2017 ; Zhao *et al.*, 2018). Moyennant un protocole et une analyse adaptés, le RAD-seq est donc bien une solution pertinente pour rechercher des signatures de sélection.

Par ailleurs, il existe des approches complémentaires au RAD-seq. Il est de notre point de vue contre-productif d'opposer le RAD-seq à des stratégies de séquençage dont la mise en œuvre nécessite des temps de développement et un coût par individu séquencé supérieurs, et apportant une information différente (*e.g., Whole-Exome Sequencing*, Pool-seq). Le RAD-seq représente un saut quantitatif important pour le développement de ressources génomiques chez les nombreuses espèces non-modèles (ou en passe de devenir des modèles) ne pouvant pas bénéficier à l'heure actuelle du financement d'un génome complet, de la mise au point de kits de séquençage d'exome ou de puces de génotypage haute densité, *etc*. Les apports du RAD-seq sont très visibles chez ces espèces, puisqu'il permet en une expérience d'augmenter d'un facteur 10 ou 100 la quantité de marqueurs disponibles par rapport au cas où des ressources génétiques plus traditionnelles (*e.g.,* AFLP, microsatellites) avaient déjà été développées (Bourgeois *et al.*, 2013 ; Reitzel *et al.*, 2013 ; Fu *et al.*, 2016 ; Shirasawa *et al.*, 2017 ; Rahman *et al.*, 2018), voire tout simplement d'initier leur collecte à moindre coût (Torres-Martínez & Emery 2016 ; Gailing *et al.*, 2017). De très nombreuses populations ont ainsi pu bénéficier des apports du RAD-seq, que ce soit pour la construction de cartes génétiques, l'amélioration de la qualité des assemblages ou la cartographie de QTL (Nadeau *et al.*, 2014 ; Zhou *et al.*, 2014 ; Li *et al.*, 2015 ; Shao *et al.*, 2015 ; Fu *et al.*, 2016 ; Kanamori *et al.*, 2016 ; Gailing *et al.*, 2017 ; Pan *et al.*, 2017 ; Pyne et *al.*, 2017 ; Bai *et al.*, 2018).

Comparer les cartes génétiques générées grâce au RAD-seq à celles d'espèces apparentées mieux caractérisées permet d'améliorer notre compréhension de l'évolution et de l'organisation de génomes parfois complexes (*e.g.,* chez les salmonidés). Cette stratégie est en plein essor chez les poissons d'aquaculture, qui comportent de nombreuses espèces non-modèles ayant un intérêt économique (Kakioka *et al.*, 2013 ; Shao *et al.*, 2015 ; Fu *et al.*, 2016 ; Manousaki *et al.*, 2016 ; McKinney *et al.*, 2016), mais aussi chez les plantes cultivées (Barchi *et al.*, 2012 ; Kundu *et al.*, 2015 ; Pan *et al.*, 2017 ; Zhong *et al.*, 2017). Chez ces espèces, les possibilités de croisement permettent d'obtenir des lignées pures recombinantes (RIL) ou d'haploïdes doublés (HD) avec des tailles de famille importantes, associées à des niveaux d'homologie entre génomes parfois élevés (Kai *et al.*, 2014 ; Pan *et al.*, 2017). Ceci favorise l'exploration fonctionnelle des génomes, pour peu qu'il existe un accès à des ressources génomiques. En fournissant cet accès, le RAD-seq représente un moyen de mieux comprendre l'architecture génomique des phénotypes complexes d'intérêt agronomique (*e.g.,* résistance aux pathologies, morphologie) chez des espèces non-modèles, afin d'éventuellement promouvoir leur amélioration génétique au travers de la mise en place d'une sélection assistée par marqueurs. Le RAD-seq a donc toute sa place en tant que stratégie de séquençage contribuant à la caractérisation des relations entre génotype et phénotype, au côté d'autres approches plus coûteuses en ressources matérielles et/ou humaines.

Comme nous l'avons vu au cours du Chapitre VI, recourir à un génotypage par RAD-seq nécessite d'accomplir une phase de préparation des séquences brutes à l'aide d'outils bioinformatiques afin d'obtenir des génotypes fiables. Cette étape passe par l'application d'une combinaison de filtres et de tests adaptés au système étudié, c'est-à-dire tenant compte des contraintes biologiques (taille et ploïdie du génome, étendue du DL, fréquence des sites de restriction, variation génétique disponible, distance entre individus…) et techniques (protocole, type de séquenceur, longueur des *reads*, nombre d'individus séquencés, profondeur de séquençage…) (Lowry *et al.*, 2017b ; Paris *et al.*, 2017).

La littérature abordant la planification des aspects expérimentaux et analytiques liés au RAD-seq est en train de s'étoffer, depuis le choix du protocole (Recknagel *et al.*, 2015 ; Andrews *et al.*, 2016) à l'optimisation des paramètres de l'analyse bioinformatique (Mastretta-Yanes *et al.*, 2015 ; Paris *et al.*, 2017 ; Rochette *et al.*, 2017), en passant par la prise en compte de l'effet du pipeline (Fitz-Gibbon *et al.*, 2017 ; Shafer *et al.*, 2017) ou les possibilités d'imputation des génotypes manquants (Money *et al.*, 2017). Nous souhaitons attirer en particulier l'attention sur deux publications très récentes et complémentaires (Paris *et al.*, 2017 ; Rochette *et al.*, 2017) qui vont grandement faciliter la détermination de génotypes fiables à partir d'expériences de RAD-seq.

Tout d'abord, la publication de Paris *et al.* (2017) apporte un appui bienvenu aux utilisateurs du RAD-seq en leur permettant de mieux appréhender les données brutes et de choisir une combinaison de paramètres pertinente pour un assemblage *de novo*. Paris *et al.* (2017) proposent pour cela une stratégie d'optimisation, dite *r80*, consistant à choisir les trois principaux paramètres dans *Stacks*, *m* (le critère de définition d'un allèle), *M* (le critère de définition d'un locus intra-individu), et *n* (le critère de définition d'un locus inter-individus), de façon à maximiser la quantité de locus polymorphes découverts chez 80% des individus. Cette stratégie permet à un utilisateur familier d'un modèle ou d'une question biologique donnés, mais ne disposant d'aucune expérience préalable d'alignement de *reads* courts, d'explorer rapidement un espace de paramètres pertinent et de bénéficier d'une règle de décision simple lui garantissant au final un assemblage fiable. Dans les faits, cette stratégie aboutit fréquemment à une configuration où le paramètre *m* optimal vaut 3, *M* étant compris entre 2 et 5 (selon un compromis entre le risque de passer à côté de SNP réels et celui d'assembler des paralogues ou des séquences répétées ; c'est donc avant tout pour le choix du paramètre *M* qu'une bonne appréciation de la question biologique posée est prépondérante), et *n* étant proche de *M* ($n = M \pm 1$) (Paris *et al.* 2017). Le choix des paramètres définitifs reste bien sûr à l'appréciation de l'expérimentateur et doit être ajusté à chaque cas particulier, mais la publication de Paris *et al.* (2017) offre une stratégie simple pour au moins aider à identifier un espace de paramètres pertinent.

Rochette *et al.* (2017) ont quant à eux publié un protocole complet d'analyse des *reads* bruts issus de RAD-seq avec *Stacks* au format *Nature Protocols*, tenant compte des résultats les plus récents – dont la stratégie *r80* de Paris *et al.* (2017) – et incluant de nombreuses recommandations précises pour préparer l'analyse bioinformatique (*e.g.,* estimation du temps d'exécution de chaque étape, identification des étapes critiques, vérifications à prévoir) à destination d'un public de biologistes aussi large que possible. A la manière d'instructeurs lors d'un *workshop*, les auteurs guident le lecteur pas à pas au travers d'un exemple réel en commentant à chaque étape une liste de commandes *shell* adaptées. De nombreux conseils pratiques (*e.g.,* traiter les données de préférence compressées pour préserver l'espace disque, utiliser à bon escient les fichiers journaux, choisir des filtres additionnels pour améliorer la qualité des génotypes) sont fournis tout au long du protocole. Des points de discussion importants sont également abordés (*e.g.,* quelle profondeur de séquençage minimale exiger, pourquoi opter pour une approche *de novo*, quelle méthode d'alignement choisir, quelles analyses préliminaires de génomique des populations envisager). Les équipes impliquées dans un projet de RAD-seq trouveront dans cette publication une ressource de tout premier plan pour les aider à prendre en charge l'analyse de leurs données, depuis les séquences brutes jusqu'à l'obtention de statistiques fiables permettant d'explorer le polymorphisme observé.

A la lumière de l'ensemble de ces éléments, notre opinion est que le contexte n'a jamais été aussi favorable à une utilisation raisonnée du RAD-seq. Nous disposons aujourd'hui d'un corpus de publications cohérent : pour planifier une nouvelle expérience, voir Davey *et al.*, 2013 ; Andrews *et al.*, 2016 ; Lowry *et al.*, 2017b ; McKinney *et al.* 2017 ; pour préparer l'analyse bioinformatique des données brutes, voir Mastretta-Yanes *et al.*, 2015 ; Paris *et al.*, 2017 ; Rochette *et al.*, 2017. Ces publications fournissent une information complète et synthétique, qui faisait défaut il y a encore peu de temps et qui permet la mise en œuvre de bonnes pratiques malgré les contraintes externes (*e.g.,* budget et temps disponibles, niveau d'expertise en bioinformatique) et internes (*e.g.,* complexité du système étudié, absence de ressources génomiques préalables) pouvant peser sur le design expérimental. La large diffusion et la consultation de ces ressources bibliographiques en amont du lancement d'un projet, lors de la relecture d'un travail des pairs, ou lors de la réanalyse de données publiées devrait permettre de maintenir un niveau élevé de qualité des résultats obtenus grâce à un génotypage par RAD-seq.

En conclusion, pour de très nombreuses populations chez lesquelles on ne dispose que de peu ou pas de ressources génomiques, il n'y a à nos yeux pas encore d'alternative aussi peu coûteuse, simple dans son principe, flexible, et bien caractérisée que ne l'est le RAD-seq aujourd'hui. La communauté des généticiens a en effet produit un arsenal de connaissances et d'outils qui lui permet d'avoir aujourd'hui du recul sur son utilisation. Une deuxième version de *Stacks* est actuellement en bêta-test, et apportera à terme une meilleure ergonomie et de nouvelles fonctions (dont un nouveau modèle de détermination des génotypes). La prise en compte de certains points spécifiques peut encore être améliorée, comme la gestion des paralogues dans les gros génomes polyploïdes (Stobie *et al.*, 2018) ou l'adaptation des protocoles au type de plateforme de séquençage utilisé (Boubli *et al.*, 2018). Tout cela ne doit en outre pas nous dispenser de réfléchir à des stratégies alternatives au RAD-seq qui pourraient aussi permettre d'étudier les populations non-modèles à moindre coût, comme par exemple le protocole d'isolation d'exome actuellement développé par Jon Puritz et Katie Lotterhos (Puritz & Lotterhos, 2017, disponible à l'heure actuelle uniquement en version *preprint*).

## — 4. Du bon usage des ressources informatiques dans le contexte à venir

Les plateformes Illumina ont dominé le marché du séquençage ces dix dernières années – la plateforme concurrente Ion Torrent PGM n'a par exemple été utilisée qu'à la marge dans les protocoles de RAD-seq (Recknagel *et al.*, 2015 ; Kearns *et al.*, 2018). Illumina, qui propose un grand nombre d'appareils, générant jusqu'à 2 x 300 pb de longueur de *reads* (avec un séquenceur MiSeq, Schirmer *et al.*, 2015), possède une avance très confortable sur le segment des *reads* courts (< 1 kb). Cependant, d'autres acteurs ont fait émerger des technologies alternatives qui commencent à prendre de l'importance, mais plutôt sur le segment des *reads* plus longs. Des technologies de séquençage de molécule unique (*Single Molecule Sequencing* ou SMS) ont ainsi été développées commercialement au cours des années 2010 par des entreprises de biotechnologie comme Pacific Biosciences ou Oxford Nanapore Technologies. Celles-ci permettent de travailler avec des *reads* de plusieurs dizaines de kilobases de long en moyenne (Lee *et al.*, 2016), mais sont encore sujettes à des taux d'erreur très élevés (da Fonseca *et al.*, 2016).

Le marché du séquençage connaît une évolution rapide, avec des acteurs rivalisant d'idées nouvelles et de promesses (Carrasco-Ramiro *et al.*, 2017 ; Loose *et al.*, 2017), et il est difficile d'anticiper précisément sur ce que seront les usages des généticiens dans les années 2020. Il semble toutefois acquis que le futur des NGS sera marqué par une tendance à la complexification des données de séquence, avec la coexistence de différents types de séquenceurs dont les produits afficheront une sensibilité différente aux risques d'erreur. Les avancées technologiques actuelles ouvrent en effet la voie à des projets de génotypage hybrides, combinant différentes techniques.

Malgré un coût qui peut être important du fait de la prise en compte d'un risque d'erreur accru, l'avènement commercial des *reads* longs apporte une solution à des questions difficiles à résoudre. La génération de longs fragments d'ADN rend en effet visibles des informations qui passent souvent inaperçues tant que l'on n'a accès qu'à des fragments de quelques centaines de paires de bases (Chaisson *et al.*, 2017). Le séquençage de *reads* longs permet par exemple de caractériser des variants structuraux associés à des phénotypes particuliers, qu'il s'agisse d'indels (Merker *et al.*, 2018) ou de réarrangements plus complexes (McGinty *et al.*, 2017 ; Sanchis-Juan *et al.*, 2018). En permettant le traitement des régions génomiques riches en séquences répétées, l'accès à des fragments de séquence de plus en plus longs présente un avantage évident pour améliorer la qualité et la contiguïté des assemblages (Koren & Phillippy 2015). Il est même possible d'obtenir des *reads* ultra-longs (> 100 kb) à partir de l'appareil portatif MinION (Oxford Nanopore Technologies), ce qui a permis de phaser les haplotypes du Complexe Majeur d'Histocompatibilité humain (Jain *et al.*, 2018).

La complémentarité entre les *reads* courts, plus précis et relativement peu coûteux à générer en grande quantité, et les *reads* longs, éclairant sur l'organisation du génome à une échelle d'investigation plus large, ouvre la voie à la généralisation d'approches combinées d'assemblage *de novo* de génomes de haute qualité à un coût modéré (Bickhart *et al.*, 2017 ; Daccord *et al.*, 2017). La possibilité d'associer des *reads* issus de différents types de plateformes laisse entrevoir un énorme potentiel applicatif, notamment auprès des thématiques abordées à l'INRA (*e.g.,* assemblage *de novo* du génome de certaines souches d'intérêt (de Been *et al.*, 2014 ; Magnan *et al.*, 2016 ; Sohn *et al.*, 2017), de génomes polyploïdes chez les plantes cultivées (Yang, J., *et al.*, 2016 ; Hatakeyama *et al.*, 2017). De tels projets d'assemblage hybride pourront donc être étendus, à l'image des méthodes de représentation réduite du génome (*e.g.,* RAD-seq), à de très nombreuses populations, ce qui augmentera la charge au plan bioinformatique : les jeux de données composites générés devront être traités au moyen de pipelines complexes nécessitant la coordination de plusieurs programmes aux objectifs différents (Daccord *et al.*, 2017).

La coexistence de plateformes produisant différents types de séquence à faible coût rend les projets d'assemblage *de novo* applicables à toute population présentant un intérêt scientifique. Le développement de quelques compétences en bioinformatique peut par conséquent représenter un bon investissement pour le généticien, afin *a minima* de comprendre l'effet du choix de certains logiciels. Actuellement, les personnels académiques ont la possibilité de bénéficier d'appuis (ressources, formations) facilitant la prise en main d'outils de base en bioinformatique. En particulier, recourir aux services de plateformes de calcul à haute performance (HPC) dès le début de la thèse permet de gagner un temps précieux (*e.g.,* pour explorer rapidement un espace de paramètres d'intérêt, pour bénéficier d'un grand nombre de logiciels déjà installés et utiles à la communauté des généticiens). Plusieurs plateformes HPC académiques permettent le partage de moyens de calcul et d'outils scientifiques puissants. Développer des compétences d'utilisation de ces plateformes HPC est un investissement qui devient rapidement rentable.

L'Institut Français de Bioinformatique recense des informations détaillées, dont les ressources disponibles et les modalités d'accès, au sujet d'une trentaine de plateformes HPC orientées sur un usage bioinformatique (https://www.france-bioinformatique.fr/fr/plateformes). A l'INRA, deux plateformes à vocation nationale (Migale : http://migale.jouy.inra.fr et Genotoul-bioinfo : http://bioinfo.genotoul.fr) sont ouvertes aux personnels académiques, y compris non permanents. Le fonctionnement de ces plateformes HPC nécessite l'utilisation d'un système d'ordonnancement des tâches, avec une interface SGE (Sun Grid Engine) ou SLURM (Simple Linux Utility for Resource Management), la seconde étant en général plus répandue sur les plateformes les plus puissantes. La plateforme Migale utilise SGE, tandis que la plateforme Genotoul-bioinfo effectue une migration vers

SLURM dans le cadre d'une amélioration de son infrastructure et devrait avoir totalement abandonné SGE en 2019. Des exemples de scripts simples pour débuter avec SGE ou SLURM sont donnés en annexe (Annexe IV). Pour aller plus loin, il est possible d'utiliser des outils de gestion de pipelines, dont l'usage est appelé à se généraliser afin d'améliorer la reproductibilité des analyses omiques (Di Tommaso *et al.*, 2017). Certaines plateformes HPC proposent régulièrement des sessions de formation aux outils de la bioinformatique. La plateforme Migale organise par exemple au premier semestre de chaque année un « Cycle de bioinformatique par la pratique » se déroulant sur le campus de l'INRA de Jouy-en-Josas et proposant plusieurs modules spécifiquement adressés aux débutants, ce qui peut également être un bon investissement en début de thèse.

# — 5. Références


Andrews, K. R., Hohenlohe, P. A., Miller, M. R., Hand, B. K., Seeb, J. E., & Luikart, G. (2014). Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular ecology*, *23*(24), 5943-5946.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81.

Agarwal, S. (2013). Systems approaches in understanding evolution and evolvability. *Progress in biophysics and molecular biology*, *113*(3), 369-374.

Babin, C., Gagnaire, P. A., Pavey, S. A., & Bernatchez, L. (2017). RAD-Seq reveals patterns of additive polygenic variation caused by spatially-varying selection in the American Eel (Anguilla rostrata). *Genome biology and evolution*, *9*(11), 2974-2986.

Bai, B., Wang, L., Zhang, Y. J., Lee, M., Rahmadsyah, R., Alfiko, Y., ... & Yue, G. H. (2018). Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. *Scientific reports*, *8*(1), 691.

Balon, E. K. (2004). About the oldest domesticates among fishes. *Journal of fish Biology*, *65*(s1), 1-27.

Barchi, L., Lanteri, S., Portis, E., Valè, G., Volante, A., Pulcini, L., ... & Rotino, G. L. (2012). A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS one*, *7*(8), e43740.

Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in ecology & evolution*, *23*(1), 38-44.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., ... & Burton, J. N. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature genetics*, *49*(4), 643.

Boubli, J. P., da Silva, M. N., Rylands, A. B., Nash, S. D., Bertuol, F., Nunes, M., ... & Sampaio, I. (2018). How many Pygmy Marmoset (Cebuella Gray, 1870) species are there? A taxonomic re-appraisal based on new molecular evidence. *Molecular phylogenetics and evolution*, *120*, 170-182.

Bourgeois, Y. X., Lhuillier, E., Cézard, T., Bertrand, J. A., Delahaie, B., Cornuault, J., ... & Thébaud, C. (2013). Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources*, *13*(5), 899-907.

Brookfield, J. F. (2009). Evolution and evolvability: celebrating Darwin 200. *Biology letters*, *5*(1), 44-46.

Brown, R. L. (2013). What evolvability really is. *The British Journal for the Philosophy of Science*, *65*(3), 549-572.

Carrasco-Ramiro, F., Peiró-Pastor, R., & Aguado, B. (2017). Human genomics projects and precision medicine. *Gene therapy*, *24*(9), 551.

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular ecology resources*, *17*(3), 362-365.

Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., ... & Fan, X. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*, 193144.

Chaves, J. A., Cooper, E. A., Hendry, A. P., Podos, J., De León, L. F., Raeymaekers, J. A., ... & Uy, J. A. C. (2016). Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular ecology*, *25*(21), 5282-5295.

Chevin, L. M., Lande, R., & Mace, G. M. (2010). Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS biology*, *8*(4), e1000357.

Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., ... & Di Pierro, E. A. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics*, *49*(7), 1099.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics*, *9*(5-6), 416-423.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., ... & Pereira, R. J. (2016). Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine genomics*, *30*, 3-13.

de Been, M., Lanza, V. F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., ... & Heederik, D. J. (2014). Dissemination of cephalosporin resistance genes between Escherichia coli strains from farm animals and humans by specific plasmid lineages. *PLoS genetics*, *10*(12), e1004776.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, *35*(4), 316.

Fitz-Gibbon, S., Hipp, A. L., Pham, K. K., Manos, P. S., & Sork, V. L. (2017). Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (Quercus section Quercus). *Genome*, *60*(9), 743-755.

Fu, B., Liu, H., Yu, X., & Tong, J. (2016). A high-density genetic map and growth related QTL mapping in bighead carp (Hypophthalmichthys nobilis). *Scientific reports*, *6*, 28679.

Gailing, O., Staton, M. E., Lane, T., Schlarbaum, S. E., Nipper, R., Owusu, S. A., & Carlson, J. E. (2017). Construction of a Framework Genetic Linkage Map in Gleditsia triacanthos L. *Plant Molecular Biology Reporter*, *35*(2), 177-187.

Grewe, F., Huang, J. P., Leavitt, S. D., & Lumbsch, H. T. (2017). Reference-based RADseq resolves robust relationships among closely related species of lichen-forming fungi using metagenomic DNA. *Scientific reports*, *7*(1), 9884.

Hansen, T., & Pélabon, C. (2018). Evolvability: a unifying concept in evolutionary biology. Second Joint Congress of Evolution Biology, Symposium number 50, Montpellier. http://evolutionmontpellier2018.org/symposia

Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., ... & Nataraja, K. N. (2017). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Research*.

Hendrikse, J. L., Parsons, T. E., & Hallgrímsson, B. (2007). Evolvability as the proper focus of evolutionary developmental biology. *Evolution & development*, *9*(4), 393-401.

Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting RAD-seq marker numbers across the eukaryotic tree of life. Genome biology and evolution, 7(12), 3207-3225.

Hubert, J. N., Allal, F., Hervet, C., Ravakarivelo, M., Jeney, Z., Vergnet, A., ... & Vandeputte, M. (2016). How could fully scaled carps appear in natural waters in Madagascar?. *Proc. R. Soc. B*, *283*(1837), 20160945.

Irschick, D. J., & Reznick, D. (2009). Field experiments, introductions, and experimental evolution. *Experimental evolution: concepts, methods, and applications of selection experiments*, 173-194.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... & Malla, S. (2018). Nanopore sequencing and assembly of a human genome with ultra-long *reads*. *Nature biotechnology*.

Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., ... & Nagao, J. (2014). A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (Anguilla japonica) provides insights into genome evolution after the teleost-specific genome duplication. *BMC genomics*, *15*(1), 233.

Kakioka, R., Kokita, T., Kumada, H., Watanabe, K., & Okuda, N. (2013). A RAD-based linkage map and comparative genomics in the gudgeons (genus Gnathopogon, Cyprinidae). *BMC genomics*, *14*(1), 32.

Kanamori, A., Sugita, Y., Yuasa, Y., Suzuki, T., Kawamura, K., Uno, Y., ... & Postlethwait, J. H. (2016). A genetic map for the only self-fertilizing vertebrate. *G3: Genes, Genomes, Genetics*, *6*(4), 1095-1106.

Kearns, A. M., Restani, M., Szabo, I., Schrøder-Nielsen, A., Kim, J. A., Richardson, H. M., ... & Omland, K. E. (2018). Genomic evidence of speciation reversal in ravens. *Nature communications*, *9*(1), 906.

Kiener, A. (1958). Intérêt et perspectives de la pisciculture de la carpe à Madagascar. *Bulletin de Madagascar*, *8*(147), 693-702.

Kirpitchnikov, V. S. (1999). *Genetics and breeding of common carp*. INRA éditions. 98 p. ISBN 978-2-7380-0869-5.

Kokko, H., Chaturvedi, A., Croll, D., Fischer, M. C., Guillaume, F., Karrenberg, S., ... & Stapley, J. (2017). Can evolution supply what ecology demands?. *Trends in ecology & evolution*, *32*(3), 187-197.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, *23*, 110-120.

Kundu, A., Chakraborty, A., Mandal, N. A., Das, D., Karmakar, P. G., Singh, N. K., & Sarkar, D. (2015). A restriction-site-associated DNA (RAD) linkage map, comparative genomics and identification of QTL for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (Corchorus olitorius L., Malvaceae sl). *Molecular breeding*, *35*(1), 19.

Lehman, J. (2017). Analyzing deception, evolvability, and behavioral rarity in evolutionary robotics. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 1479-1486). ACM.

Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., ... & Sehgal, D. (2015). A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC genomics*, *16*(1), 216.

Lo, C. L., Lossie, A. C., Liang, T., Liu, Y., Xuei, X., Lumeng, L., ... & Muir, W. M. (2016). High resolution genomic scans reveal genetic architecture controlling alcohol preference in bidirectionally selected rat model. *PLoS genetics*, *12*(8), e1006178.

Long, A., Liti, G., Luptak, A., & Tenaillon, O. (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, *16*(10), 567.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017a). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular ecology resources*, *17*(2), 142-152.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular ecology resources*.

Loose, M. W. (2017). The potential impact of nanopore sequencing on human genetics. *Human molecular genetics*, *26*(R2), R202-R207.

McGinty, R. J., Rubinstein, R. G., Neil, A. J., Dominska, M., Kiktev, D., Petes, T. D., & Mirkin, S. M. (2017). Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair. *Genome research*, *27*(12), 2072-2082.

McKinney, G. J., Seeb, L. W., Larson, W. A., Gomez-Uchida, D., Limborg, M. T., Brieuc, M. S. O., ... & Seeb, J. E. (2016). An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (Oncorhynchus tshawytscha). *Molecular ecology resources*, *16*(3), 769-783.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). *Molecular ecology resources*, *17*(3), 356-361.

Manousaki, T., Tsakogiannis, A., Taggart, J. B., Palaiokostas, C., Tsaparis, D., Lagnel, J., ... & Tsigenopoulos, C. S. (2016). Exploring a nonmodel teleost genome through rad sequencing—linkage mapping in Common Pandora, Pagellus erythrinus and comparative genomic analysis. *G3: Genes, genomes, genetics*, *6*(3), 509-519.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28-41.

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., ... & Montgomery, S. B. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*, *20*(1), 159.

Money, D., Migicovsky, Z., Gardner, K., & Myles, S. (2017). LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC genomics*, *18*(1), 523.

Nadeau, N. J., Ruiz, M., Salazar, P., Counterman, B., Medina, J. A., Ortiz-Zuazaga, H., ... & Papa, R. (2014). Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato. *Genome research*, *24*(8), 1316-1333.

Pan, L., Wang, N., Wu, Z., Guo, R., Yu, X., Zheng, Y., ... & Chen, C. (2017). A High Density Genetic Map Derived from RAD Sequencing and Its Application in QTL Analysis of Yield-Related Traits in Vigna unguiculata. *Frontiers in plant science*, *8*, 1544.

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360-1373.

Pigliucci, M. (2008). Is evolvability evolvable?. *Nature Reviews Genetics*, *9*(1), 75.

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, *23*(24), 5937-5942.

Puritz, J. B., & Lotterhos, K. E. (2017). Expressed Exome Capture Sequencing (EecSeq): a method for cost-effective exome sequencing for all organisms with or without genomic resources. *bioRxiv*, 223735.

Pyne, R., Honig, J., Vaiciunas, J., Koroch, A., Wyenandt, C., Bonos, S., & Simon, J. (2017). A first linkage map and downy mildew resistance QTL discovery for sweet basil (Ocimum basilicum) facilitated by double digestion restriction site associated DNA sequencing (ddRADseq). *PloS one*, *12*(9), e0184319.

Rahman, S., Schmidt, D., & Hughes, J. (2018). De novo SNP discovery and strong genetic structuring between upstream and downstream populations of Paratya australiensis Kemp, 1917 (Decapoda: Caridea: Atyidae). *Journal of Crustacean Biology*.

Reitzel, A. M., Herrera, S., Layden, M. J., Martindale, M. Q., & Shank, T. M. (2013). Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular ecology*, *22*(11), 2953-2970.

Remolina, S. C., Chang, P. L., Leips, J., Nuzhdin, S. V., & Hughes, K. A. (2012). Genomic basis of aging and life-history evolution in Drosophila melanogaster. *Evolution*, *66*(11), 3390-3403.

Rohner, N., Bercsényi, M., Orbán, L., Kolanczyk, M. E., Linke, D., Brand, M., ... & Harris, M. P. (2009). Duplication of fgfr1 permits Fgf signaling to serve as a target for selection during domestication. *Current Biology*, *19*(19), 1642-1647.

Sanchis-Juan, A., Stephens, J., French, C. E., Gleadall, N., Mégy, K., Penkett, C., ... & Erwood, M. (2018). Complex Structural Variants Resolved by Short-Read and Long-Read Whole Genome Sequencing in Mendelian Disorders. *bioRxiv*, 281683.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*, *43*(6), e37-e37.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., & Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity*, *114*(5), 431.

Shafer, A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907-917.

Shao, C., Niu, Y., Rastas, P., Liu, Y., Xie, Z., Li, H., ... & Sakamoto, T. (2015). Genome-wide SNP identification for the construction of a high-resolution genetic map of Japanese flounder (Paralichthys olivaceus): applications to QTL mapping of Vibrio anguillarum disease resistance and comparative genomic analysis. *DNA Research*, *22*(2), 161-170.

Shirasawa, K., Tanaka, M., Takahata, Y., Ma, D., Cao, Q., Liu, Q., ... & Lee, H. U. (2017). A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (Ipomoea batatas). *Scientific reports*, *7*, 44207.

Stobie, C. S., Oosthuizen, C. J., Cunningham, M. J., & Bloomer, P. (2018). Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing. *Ecology and Evolution*.

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in ecology & evolution*, *29*(12), 673-680.

Torres-Martínez, L., & Emery, N. C. (2016). Genome-wide SNP discovery in the annual herb, Lasthenia fremontii (Asteraceae): genetic resources for the conservation and restoration of a California vernal pool endemic. *Conservation genetics resources*, *8*(2), 145-158.

Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., & Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS genetics*, *7*(3), e1001336.

Yang, G. Q., Chen, Y. M., Wang, J. P., Guo, C., Zhao, L., Wang, X. Y., ... & Guo, Z. H. (2016). Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. *Plant methods*, *12*(1), 39.

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., ... & Yao, P. (2016). The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature genetics*, *48*(10), 1225.

Zhao, Y., Peng, W., Guo, H., Chen, B., Zhou, Z., Xu, J., ... & Xu, P. (2018). Population Genomics Reveals Genetic Divergence and Adaptive Differentiation of Chinese Sea Bass (Lateolabrax maculatus). *Marine Biotechnology*, *20*(1), 45-59.

Zhong, Y. J., Zhou, Y. Y., Li, J. X., Yu, T., Wu, T. Q., Luo, J. N., ... & Huang, H. X. (2017). A high-density linkage map and QTL mapping of fruit-related traits in pumpkin (Cucurbita moschata Duch.). *Scientific reports*, *7*(1), 12785.

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., ... & Jiang, H. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC genomics*, *15*(1), 351.

# Annexes

Version de soumission de l'Article II

# Cancer- and behavior-related genes are targeted by selection in the Tasmanian devil (*Sarcophilus harrisii*)

Jean-Noël Hubert*, Tatiana Zerjal and Frédéric Hospital

GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

*Correspondence: jean-noel.hubert@inra.fr

## Abstract

Devil Facial Tumor Disease (DFTD) is an aggressive cancer notorious for its rare etiology and its impact on Tasmanian devil populations. Two regions underlying an evolutionary response to this cancer were recently identified using genomic time-series pre- and post-DTFD arrival. Here, we support that DFTD shaped the genome of the Tasmanian devil in an even more extensive way than previously reported. We detected 97 signatures of selection, including 148 protein coding genes having a human orthologue, linked to DFTD. Most candidate genes are associated with cancer progression, and an important subset of candidate genes has additional influence on social behavior. This confirms the influence of cancer on the ecology and evolution of the Tasmanian devil. Our work also demonstrates the possibility to detect polygenic footprints of short-term selection in very small populations.

**Keywords** : Tasmanian devil ; Devil Facial Tumor Disease (DFTD) ; rapid natural selection ; cancer genetics ; behavioral disorders ; complex phenotype.

# Introduction

Understanding the role of selection in the resistance against cancer should benefit from the study of animal populations. Domestic populations that develop spontaneous neoplasms similar to those encountered in humans have been increasingly investigated to unravel the complex genetic determinism of some cancers (Schiffman and Breen 2015). Studies that examine cancer in natural populations are, however, quite rare, despite their potential for improving our knowledge of cancer resistance mechanisms in an ecological context. Observations made in this field are limited to a small number of unique cancer evolution cases, as those described in the naked-mole rat, the elephant, and the Tasmanian devil (Roche et al. 2017). In particular, horizontally transmitted cancers are rarely observed events in nature, which would be more likely to happen in bivalve species (Metzger et al. 2016). The best-known case of transmissible cancer is the one affecting the Tasmanian devil (*Sarcophilus harrisii*), the largest extant marsupial carnivore. This cancer, known as the Devil Facial Tumor Disease (DFTD), is characterized by a recent arrival, a high propensity to metastasize, and a mortality rate close to 100% within 12 months after infection (Belov et al. 2012 ; Pye et al. 2016a, 2016b). The disease was first detected in North-Eastern Tasmania in 1996 and has since spread across 95% of the species' range (Storfer et al. 2017) through biting injuries during social contacts (Hamede et al. 2013). DFTD has exerted a very strong selective pressure on the Tasmanian devil, which has translated into a rapid and extensive decline in population size, with local population losses over 90%, giving rise to serious concerns regarding the species survival in the short term (Hendricks et al. 2017).

DFTD has become a very well documented case study over the past decade, due to the scarcity of transmissible cancer examples in nature and to the population management implications for conservation biology. This amount of data is a valuable resource to better understand cancer biology, and is expected to provide insight into the mechanisms underlying

46 immunosurveillance of cancer initiation and metastatic spreading (Belov et al. 2012). A recent
47 analysis of Tasmanian devil genomic time-series (Epstein et al. 2016a) led to the identification
48 of two regions exhibiting signatures of selection in response to DFTD which contained genes
49 related to cancer risk and immune function in humans. Here, we extend this research by using
50 a customized maximum-likelihood method that has been shown efficient for investigating rapid
51 selection in experimental populations when genomic time-series data are available (Thépot et
52 al. 2015). In comparison to the approach used in Epstein et al. (2016a), our method has in
53 particular the advantage of efficiently disentangling the effects of strong selection from those
54 of strong drift in a population-specific manner. We show that 97 genomic regions, including
55 the two previously identified, are candidate targets of selection. A total of 148 protein coding
56 genes stand in the vicinity of these signatures of selection. The functional analysis of the
57 candidate genes revealed that almost all of them have a link with cancer, including over 30
58 genes mediating metastasis. Furthermore, around 15% of the candidate genes have orthologues
59 involved in the functioning of the Central Nervous System (CNS), including a dozen loci
60 associated with behavioral disorders, indicating that natural selection has probably favored
61 particular behaviors.

62 Our data support the view that the evolutionary response to DFTD consists in several biological
63 and behavioral strategies that rely on a larger range of genetic variants than previously thought.
64 These findings should contribute to a better understanding of the ecology and the evolution of
65 both the Tasmanian devil and cancer. In particular, this should help to pinpoint some genes that
66 may oppose to cancer progression.

67

# Results

## Signatures of selection in the Tasmanian devil genome

Our analysis resulted in the identification of 97 signatures of selection dispersed throughout the chromosomes, accounting for about 0.3% of the genome. Most signatures of selection were found to be population-specific, and only one was common to the three investigated populations (Fig 1A). The West Pencil Pine (WP), Freycinet (FN) and Narawntapu (NP) populations displayed 54, 38 and 37 signatures of selection, respectively (Fig 1B). The ability to detect signatures of selection from the dataset was influenced by the experimental sampling protocol and SNP detection, which was specific to each population (Table 1). In the WP population, characterized by an adequate time-series sampling but by a low SNP density, a large number of small signatures of selection were detected. In the NP population, despite a high SNP density, a low proportion of candidate SNP (0.4%) was detected. This can be explained by the non-optimal sampling time-series (Table 1), with the latest data point too close to the date of arrival of the disease, which left little time for selection to produce visible effects. The FN population presented a much better SNP density and sampling time-series, compared to the other populations, which allowed detecting the highest number of candidate SNP (210) and larger signatures of selection. As a consequence, a larger proportion (68%) of candidate regions were located less than 100 kb from a protein coding gene in the FN population (Fig 1B). Among the three populations, 60 signatures of selection were detected in the vicinity of protein coding genes (see S1 Fig and S1 Table for the details). In total, we can draw up a list of 148 candidate genes according to Ensembl genome browser 90 (S1 Table).

**Fig 1.** Signatures of selection in the Tasmanian devil genome.

Ninety-seven signatures of selection were identified in the Tasmanian devil (*Sarcophilus harrisii*). Their distribution among the three investigated populations is provided in the form of **(A)** a Venn diagram and **(B)** bar plots. Light bars show the total number of signatures of selection identified in each population, among which dark bars (and the proportions between brackets) represent those standing at less than 100 kb from a protein-coding gene having a human orthologue according to Ensembl 90. Populations: FN = Freycinet ; NP = Narawntapu ; WP = West Pencil Pine.

Table 1. Genomic time-series investigated to identify signatures of selection in the Tasmanian devil (*Sarcophilus harrisii*)

| **Population symbol** | **Sampling location** | **DFTD arrival** | $N_e$ | $N_1$ ($t_1$)[a] | $N_2$ ($t_2$)[a] | $N_3$ ($t_3$)[a] | **Nb. of analyzed SNPs** | **Nb. of candidate SNPs** | **Prop. of candidate SNPs (%)** |
|---|---|---|---|---|---|---|---|---|---|
| FN | Freycinet | 2001 | 34 | 29 (1999) | - | 20 (2012-2013) | 16978 | 210 | 1.2 |
| NP | Narawntapu | 2007 | 37 | 26 (1999) | 26 (2004) | 27 (2009) | 27173 | 104 | 0.4 |
| WP | West Pencil Pine | 2006 | 26 | 21 (2006) | - | 43 (2013-2014) | 5401 | 88 | 1.6 |

$N_e$, effective population size

[a] Time-series include either two (for FN and WP) or three (for NP) temporal samples. $N_i$ denotes the size of each temporal sample $i$ indexed in chronological order. The year of sampling, $t_i$, is provided between brackets.

N.B. Data were made publicly available by Epstein et al. (2016b). More information about the dataset can be found in Epstein et al. (2016a).

99 **The functional annotation of candidate genes reveals a strong link with cancer**

100 We used the knowledge base analysis tool IPA (Ingenuity Systems®, www.ingenuity.com) to
101 identify biological functions and disease-related categories associated with our candidate genes.
102 The top molecular and cellular functions identified by IPA included several "cell-related
103 functions" such as cell cycle, morphology and organization. Other functions were related to
104 nucleic acids, primary metabolism and immune system, as shown in Fig 2. Among the top 100
105 disease-related categories, 73 were associated with "cancer" (S2 Table). The "solid tumor"
106 category was the most overrepresented (p-value = $1.16\times10^{-10}$, FDR = $4.39\times10^{-7}$) with 138 genes
107 out of 147 associated with this term. According to IPA, all the 60 signatures of selection located
108 in the vicinity of coding sequence host genes potentially related to cancer.

109

110 **Fig 2.** Molecular and functional categories associated with protein coding candidate genes.
111 Representation of overrepresented molecular and functional categories obtained by functional
112 analysis in IPA. Related functional categories are represented with the same color code. Only
113 functions with a $-\log_{10}$(p-value) $\geq 1.3$ (orange line in the graph, corresponding to a p-value $\leq$
114 0.05) were considered.

115

116 Many candidate genes are key regulators of signaling pathways mediating cancer progression.
117 For example, three candidates (DTWD1, NEK6 and NSD3) are regulators of the G2/M
118 transition, a critical cell cycle checkpoint that is often deregulated in cancer (Zhang et al. 2014
119 ; Ma et al. 2015 ; Vougiouklakis et al. 2015). Another key pathway in the regulation of cell fate
120 is apoptosis (Okada and Mak 2004). The candidate BAG4 is an oncogene that prevents extrinsic
121 apoptosis (Jiang et al. 1999) and the candidate TRIM66 intervenes in intrinsic apoptosis by
122 negatively regulating P53 (Chen et al. 2015), a central tumor-suppressor gene. The candidates

CEP131 and PINX1 are important to prevent genome instability, which is seen as a crucial feature underlying the mechanisms of carcinogenesis (Hanahan and Weinberg 2011). The depletion of the centrosomal protein CEP131 is a factor of genome instability (Staples et al. 2012). PINX1 has been identified as a telomerase inhibitor whose role is documented in a large number of cancers (Li et al. 2016), in particular through the prevention of telomerase reactivation in somatic cells (Maciejowski and de Lange 2017).

A great number of candidates were associated with signaling pathways that usually mediate embryonic development, but may also have a role in cancer progression. Alterations of the candidate FGFR1, which belongs to a family of tyrosine-kinase receptors (RTK), have been described in many tumors, in particular when the kinase domain displays activating mutations (Jones et al. 2013 ; Cowell et al. 2017). Other candidates are known to be involved in oncogenic RTK pathways, such as SHC4 (Strub et al. 2011) and ST5 (Ioannou and McPherson 2016). The candidate CEP131 has also been reported to promote cell proliferation and migration through the activation of the phosphoinositide 3-kinase (PI3K/Akt) signaling pathway (Liu et al. 2017), which is one of the downstream cascades of RTK most often affected in human cancers (Pal and Mandal 2012). As an important member of the SMAD family of transcription factors, the candidate SMAD3 is a key mediator of the transforming growth factor-β (TGF-β) signaling pathway involved in cancer progression (Syed 2016). The ability of SMAD3 to interact with many transcriptional regulators confers it several roles in carcinogenesis (Yang et al. 2006 ; Tang et al. 2017). The famous Wnt/β-Catenin pathway regulates cell proliferation and is altered in many diseases, particularly cancers (Nusse and Clevers 2017). The candidate FOXN3 is a crucial regulator in this pathway by preventing the formation of the β-Catenin/TCF4 complex that promotes the transcription of proliferation genes (Dai et al. 2017). SPTBN1 has been suggested as a tumor-suppressor through the modulation of an inhibitor of the Wnt pathway (Zhi et al. 2015). Both candidates NSD3 (Vougiouklakis et al. 2015) and CDK14 (Davidson

and Niehrs 2010) may promote Wnt signaling. Key mediators of other pathways related to both development and cancer, such as Notch and Hippo, are among our candidates. The candidate NOTCH2 is one of the four Notch receptors involved in direct cell-cell interactions in the Notch pathway. Both activating and inactivating mutations of NOTCH2 have been reported in human cancers (Nowell and Radtke 2017). The candidate TEAD4 belongs to the TEAD (transcriptional enhancer factor domain) transcription factors that are required for activating the proliferation genes targeted by Hippo signaling in cancer (Harvey et al. 2013). The functional annotation of our candidates have therefore allowed the identification of putative key mediators of various signaling events that may lead to cancer.

**An important subset of candidate genes are mediators of metastasis**

It is noteworthy that a great number of candidate genes identified in the present study have the potential to influence invasiveness and metastasis of cancer cells. Several candidates encode proteins involved in cell adhesion processes. In particular, we identified seven metastasis-related genes that encode adhesion molecules (ADGRA2, ADGRD2, CDH8, MCAM, THY1, TSPAN9, and TSPAN11). For example, both MCAM and THY1, which encode cell adhesion molecules of the immunoglobulin superfamily (IgSF-CAMs), are frequently overexpressed in metastatic tumor tissues (Wu et al. 2012 ; Zhang et al. 2016). Other candidates, such as PRRX2 and CST3, confer metastatic properties to cancer cells through the TGF-β pathway (Juang et al. 2016 ; Yan et al. 2017). Similarly, FOXN3 (Dai et al. 2017), HMGCS2 (Chen et al. 2017), LSM1 (Little et al. 2016), PHGDH (Samanta et al. 2016), RIN2 (Sandri et al. 2012), TRPM8 (Yee 2015), are known regulators of metastatic processes. In total, over 30 selection candidates identified in this study are known to participate in metastasis-related mechanisms. All this suggests that the control of metastasis may be a key component of the evolutionary response to DFTD.

**Behavior-related genes among the identified candidate genes**

Although the "cancer" category is strongly overrepresented, the results of IPA analysis also display a link between the candidate list and functions associated with development and the nervous system (Table S2) that may influence behavior. In particular, 16 candidates are gathered within the annotation term "development of neurons" (S2 Fig). Several candidates may contribute to the development and the homeostasis of important cellular compartments in the central nervous system (CNS). For example, SYNDIG1 is essential for the formation of excitatory synapses in the hippocampus (Lovero et al. 2013). NKX2-2 is a key regulator of serotonergic neuron development (Cheng et al. 2003). SDK2 is an adhesion molecule that regulates synaptic connections, thereby influencing the arrangement of neural circuits in the CNS (Krishnaswamy et al. 2015). KIF13B, a member of the kinesin motor protein superfamily, has a key role in the regulation of axon development (Nakata and Hirokawa 2007). Other candidates have putative roles at the synapse level. SLITRK5 and SLC38A10 play important roles in neurotransmission (Shmelkov et al. 2010 ; Hellsten et al. 2017). CDC42EP4 and HERC1 are involved in synapse homeostasis (Ageta-Ishihara et al. 2015 ; Bachiller et al. 2015). In addition, the candidate list harbors several genes encoding subunits of ion channels (*e.g.* CACNB2, KCNA4, KCNIP3, or KCTD3) that are involved in signal transmission in the CNS.

This may have an impact on the prevalence of behavioral disorders in Tasmanian devil populations. For instance, the large signature of selection displayed on scaffold GL834709 (nucleotides 2,702,597 to 2,946,330) corresponds to human locus 8p11.23 that has been proposed as a 'neurodevelopmental hub' associated with autism spectrum disorders (ASD) according to a recent meta-analysis (Autism Spectrum Disorders Working Group 2017). The NDUFAF5 candidate is involved in the assembly of the first complex of the respiratory chain (Rhein et al. 2016), which is frequently impaired in CNS disorders like ASD (Hollis et al. 2017). Several other candidates, such as CACNB2 (Breitenkamp et al. 2014), CDH8 (Pagnamenta et al. 2011), HERC1 (Utine et al. 2017), KCTD3 (Poot et al. 2010), KIF13B (Li et al. 2016b),

198 SERINC2 (Hnoonual et al. 2017), and SLC39A11 (Woodbury-Smith et al. 2015) have been
199 associated with ASD. Some candidates have been linked to intellectual disability (ID), such as
200 CRBN (Kaufman et al. 2010), GPKOW (Helsmoortel et al. 2015), HERC1 (Utine et al. 2017)
201 and KCNA4 (Kaya et al. 2016), or to other behavioral disorders, such as CDC42EP4 (Yan et
202 al. 2016) and SLITRK5 (Shmelkov et al. 2010).

203 Genes involved in the architecture of complex behaviors therefore represent the second major
204 functional category associated with our candidate list after cancer. Overall, some candidates
205 may influence neurotransmission and thereby complex phenotypes such as life-history traits or
206 social behaviors.

207 **A few candidates may contribute to immunosurveillance of cancer**

208 Immunity, a function that may be related to cancer progression, is also represented through a
209 few genes. In particular, six candidates (TRIM10, TRIM15, TRIM26, TRIM39-RPP21,
210 TRIM62, and TRIM66) are members of the large tripartite motif (TRIM) family that consists
211 of ubiquitin ligases involved in both innate immunity and cancer progression (Hatakeyama et
212 al. 2017). With the exception of TRIM39-RPP21, all the TRIM genes identified here are known
213 to be related to cancer according to IPA (Table S1). TRIM10, TRIM15, TRIM26, and TRIM39-
214 RPP21 belong to the human leukocyte antigen (HLA) region, which gathers many immunity
215 regulators (Shiina et al. 2009). TRIM62 and TRIM66 have also been suggested to have a role
216 in immunity (Versteeg et al. 2013 ; Cao et al. 2015). Candidates from the TSPAN (TSPAN9
217 TSPAN11) and the aGPCR (ADGRA2, ADGRD2) families, which were cited above for their
218 putative roles in metastasis, may also be implicated in immune response (Veenbergen and van
219 Spriel 2011 ; Nijmeijer et al. 2016). The versatile SMAD3 candidate has been shown to
220 influence the immunosurveillance of cancer (Tang et al. 2017). As already suggested by Epstein
221 et al. (2016a), some candidates have the potential to provide an immune response against
222 cancer.

## Discussion

The Tasmanian devil dataset investigated here was initially presented and analyzed in Epstein et al. (2016a), who reported two putatively selected regions harboring seven candidate genes. These two regions were also detected in our analysis and corresponded to signatures of selection located in scaffolds GL841593 (nucleotides 4,501,785 to 4,979,756) and GL849657 (nucleotides 283,671 to 283,701). Epstein et al. (2016a) performed their analysis with a composite test statistic that took into account temporal changes in both allele frequency and integrated extended haplotype homozygosity of an individual SNP site (iES) (Tang et al. 2007). Such an approach is based on the idea that an SNP undergoing strong selection would rapidly rise in frequency, while the haplotypic diversity of the surrounding region would suddenly decrease. To avoid false-positives, Epstein et al. (2016a) restricted their analysis to the candidate regions shared by the three populations. The disadvantage of this approach is that the total number of candidate genes detectable is highly dependent on the SNP density, especially as the West Pencil Pine (WP) population presented a low genotyping density with only ~ 5000 available SNP (Table 1).

In our analysis, we considered the possibility that some observed polymorphisms might be population-specific, as supported by recent results (Hendricks et al. 2017 ; Storfer et al. 2017). Our method relies on a Wright-Fisher model coupled to maximum-likelihood computations. This provides an objective criterion, tailored to the amount of drift estimated in each examined population, to determine whether the observed variation in SNP frequency is more likely to be caused by selection and drift than by drift alone. In particular, simulations indicated that our method generated very few false-positives (from <1% to 2%) in the presumed tricky conditions where $N_e$ is as low as 30, as in the Tasmanian devil.

Most candidate genes identified in our genome-wide scan for DFTD driven selection in the Tasmanian devil have roles in cancer progression, showing that the evolutionary response to DFTD relies on a large number of mediators of the multistep processes of carcinogenesis. Some of them are crucial regulators, which are together able to influence the cellular mechanisms central to cancer progression, such as altered cell signaling, neoangiogenesis, metastatic spreading, and genetic instability. The large spectrum of biological functions played by the candidate genes in the Tasmanian devil reflects the cancer hallmarks proposed as a conceptual framework for understanding the complex biology of cancer (Hanahan and Weinberg 2011). Our analysis therefore suggests that natural selection may target multiple cellular circuits to limit the acquisition of the cancer hallmarks, which should prevent or at least delay cancer growth and metastatic processes.

Given the complexity of a phenotype like cancer and the strength of the selection imposed by DFTD, a response to selection involving the contribution of many genes is likely. What was more difficult to anticipate, however, was that such a highly polygenic response was detectable despite the very low effective population sizes ($N_e$ ~ 30) in the Tasmanian devil. It is worth noting that the majority of the signatures of selection identified in the present study were population-specific. In other words, the evolutionary response to DFTD may rely on different genes, depending on the population specific genetic variation. This possibility was supported by recent analyses reporting the existence of several genetic clusters among Tasmanian devil populations (Hendricks et al. 2017 ; Storfer et al. 2017). All this explains why we could identify a large number of candidate genes from the Tasmanian devil genomic time-series.

Noteworthily, the candidate list includes genes that have the potential to influence behavior. Diseases such as ASD and ID, which are seen as developmental disorders affecting synaptic connections, are especially associated with human orthologues of our candidate genes. In humans, these behavioral disorders are characterized by deficits in communication and in social

interaction (Orsmond et al. 2004 ; Walton and Ingersoll 2013 ; Taheri et al. 2016). The identification of candidate genes that contribute to such phenotypes in the Tasmanian devil must be juxtaposed to recent results indicating that DFTD affects primarily individuals that exhibited a greater fitness before the arrival of the disease (Wells et al. 2017). In particular, several studies have suggested that socially dominant individuals characterized by aggressive behavior, such as biting, were at greater risk of infection (Hamede et al. 2009 ; Hamede et al. 2013 ; Wells et al. 2017). This suggests that individuals that are more likely to avoid conflicts are expected to better 'resist' to DFTD, hence an indirect selection against social aggressiveness.

Indeed, the extended functional annotation of candidate genes revealed genes involved in the architecture of behaviors. We must not lose sight of the fact that cancer and behavior are both complex phenotypes that may rely on the same types of genes, as exemplified by the signature of selection in scaffold GL834709 (nucleotides 2,702,597 to 2,946,330), which has been shown to be associated with both cancer (Garcia et al. 2005) and behavioral disorders (Autism Spectrum Disorders Working Group 2017). Further analysis is therefore required to get a more precise insight into the function(s) targeted by selection in each candidate region and in each population.

Overall, our results suggest that a wide range of functions have been targeted by selection in the Tasmanian devil. For candidate genes that are regulators of cancer progression, the benefit is evident because individuals with a slow DFTD progression will be able to reproduce despite the infection (Pye et al. 2016b ; Wells et al. 2017). Some other candidate genes seem to affect behavior by acting on the CNS. In this case, the advantaged individuals will be those showing less aggressive behavior, which are less likely to be infected due to reduced physical interactions with other Tasmanian devils (Wells et al. 2017) or those individuals showing an early reproduction (Lashich et al. 2009). This gives an empirical support to the expectations

296 recently detailed by Roche et al. (2017) that an evolutionary response to cancer should rely not
297 only on cellular pathways involved in carcinogenesis, but also on adjustments of life-history
298 traits and behavior. Even if evolutionary costs driven by sexual selection may be opposed,
299 Roche et al. notably suggest that "avoidance of contagious cancers could be a selective force
300 for specific behavior", which is in line with our results.

301 In total, we identified almost 150 candidate genes with annotation functions related to strategies
302 that may deal with the DFTD infection risk. DFTD therefore has extensively shaped the genome
303 of the Tasmanian devil, which highlights the evolvability of contemporary populations despite
304 limited genetic variation, but also raises potential long-term impacts of DFTD on both the social
305 and genetic structures in the species. Importantly, our results show the possibility of detecting
306 signatures of natural selection associated with complex phenotypes from genomic time-series,
307 even if the number of generations of selection investigated (~ 5 generations) and the effective
308 population size ($N_e$ ~ 30) are very small.

309 ## Methods

310 **Tasmanian devil dataset.** The data analyzed in the present article were initially reported in
311 Epstein et al. (2016a) and made publicly available as a *Dryad* data package (Epstein et al.
312 2016b). This data package consists of SNP genotyping data produced by Stacks (Catchen et al.
313 2013) following RAD-seq (Restriction-site Associated DNA sequencing) assays (Etter et al.
314 2011) from tissue samples collected in 360 Tasmanian devils across Tasmania. Samples were
315 collected at different time points between 1999 and 2014, allowing the analysis of genomic
316 time-series that take into account the impact of DFTD through time on Tasmanian devil
317 populations. We restricted our analysis to the same samples as in Epstein et al. (2016a), from
318 the localities of Freycinet (FN), Narawntapu (NP) and West Pencil Pine (WP). We reproduced
319 the SNP filtering strategy reported by Epstein et al. (2016a) in their Methods section. In brief,

we performed filtering according to (i) MAF computed over the whole dataset (SNP with MAF less than 0.01 were discarded), (ii) observed heterozygosity computed over the whole dataset (SNP with heterozygosity over 0.5 were discarded), (iii) the proportion of missing genotypes (SNP with less than one-third of genotypes either in the whole dataset or in a sample were discarded, except for the two smallest samples in which SNP with less than half genotypes were removed), (iv) the linkage disequilibrium between neighboring SNP (using PLINK (Purcell et al. 2007), we removed one SNP from pairs of SNP harboring $R^2 > 0.99$ over 20 successive SNP and 50 kb of distance in any sample). We obtained filtered datasets of 16978, 27173 and 5401 SNP for the FN, NP and WP populations, respectively. Relevant information about the dataset for the present work is summarized in Table 1. Further information about sample collection, genotyping and data processing can be found in Epstein et al. (2016a).

**Signatures of selection.** We submitted the genomic time-series of the three investigated Tasmanian devil populations to a customized method for detecting footprints of selection (available at https://github.com/hubert-pop/signasel). This method was initially described and successfully implemented to detect genomic regions targeted by short-term selection in experimental wheat populations in Thépot et al. (2015). Briefly, this method compares two Wright-Fisher models, one including drift and the other including drift plus selection, in a maximum-likelihood framework. The model that best fits the SNP frequency variation observed over time is identified through a Likelihood Ratio Test (LRT). Each SNP is individually tested and associated with a p-value that quantifies to what extent the temporal variation in SNP frequency may be due to selection, under a null hypothesis postulating an effect of drift only. This method has proven to be efficient for detecting SNP under selection from genomic temporal samples separated by a few generations in small populations undergoing intense selection, as in the Tasmanian devil. We performed an extensive characterization of the performances of our method through forward simulations (Hubert and

Hospital, submitted). In the present study, we considered as relevant signatures of selection the genomic regions that included at least one SNP with p-value < 0.0001 or at least two neighboring pruned SNP with p-values < 0.01, which corresponds to a FDR of 13% (Benjamini & Hochberg, 1995). We looked for candidates for selection by applying our method to two temporal samples in the FN (the sample from 1999 and the combination of those from 2012 and 2013) and WP (the sample from 2006 and the combination of those from 2013 and 2014) populations, and to three temporal samples (samples from 1999, 2004 and 2009) in the NP population (Table 1). Given the sampling times and an assumed generation time of 2 years in the Tasmanian devil, we considered that the time-series covered a period of 6, 5, and 3 complete generations in the FN, NP, and WP populations, respectively. As our method relies on Wright-Fisher models, the effective size ($N_e$) of each investigated population must be provided. We used the values of $N_e$ suggested in Epstein et al. (2016a), that is, 34, 37 and 26 for the FN, NP and WP populations, respectively.

**Candidate genes identification.** We identified as candidate genes all the protein coding genes standing at less than 100 kb of the detected signatures of selection and having a human orthologue according to Ensembl genome browser 90 (https://www.ensembl.org). Ensembl stable ID of candidate genes and corresponding orthologues were retrieved using the Ensembl Genes 90 database and the Devil_ref v7.0 Tasmanian devil reference assembly (Murchison et al. 2012) obtained from BioMart.

**Ingenuity Pathway Analysis.** We used the manually curated database IPA® (Ingenuity Pathway Analysis, QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis) to identify the diseases and developmental disorder as well as the molecular and cellular functions associated with our candidate genes. We submitted the list of 147 human orthologues of our candidate genes to a “Core Analysis” with the “Ingenuity Knowledge Base” reference set to find overrepresented functions and diseases in our gene set. This is achieved by

IPA by applying a right-tailed Fisher's Exact test to estimate the likelihood that the overlap between the set of genes and a given function or disease is due to random chance.

## Acknowledgements

The authors wish to thank Wendy Brand-Williams for linguistic revision of the manuscript. This work was supported by the French Ministry of Higher Education and Research.

## References

Ageta-Ishihara, N., Yamazaki, M., Konno, K., Nakayama, H., Abe, M., Hashimoto, K., ...& Tanaka, K. (2015). A CDC42EP4/septin-based perisynaptic glial scaffold facilitates glutamate clearance. Nature communications, 6.

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Anney, R. J., Ripke, S., Anttila, V., Grove, J., Holmans, P., ... & Neale, B. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. Molecular autism, 8, 1-17.

Bachiller, S., Rybkina, T., Porras-García, E., Pérez-Villegas, E., Tabares, L., Armengol, J. A., ... & Ruiz, R. (2015). The HERC1 E3 ubiquitin ligase is essential for normal development and for neurotransmission at the mouse neuromuscular junction. Cellular and molecular life sciences, 72(15), 2961-2971.

Belov, K. (2012). Contagious cancer: lessons from the devil and the dog. BioEssays, 34(4), 285-292.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

392 Breitenkamp, A. F., Matthes, J., Nass, R. D., Sinzig, J., Lehmkuhl, G., Nürnberg, P., & Herzig,
393 S. (2014). Rare mutations of CACNB2 found in autism spectrum disease-affected families alter
394 calcium channel function. PLoS One, 9(4), e95579.

395 Cao, Z., Conway, K. L., Heath, R. J., Rush, J. S., Leshchiner, E. S., Ramirez-Ortiz, Z. G., ... &
396 Cheng, S. C. (2015). Ubiquitin ligase TRIM62 regulates CARD9-mediated anti-fungal
397 immunity and intestinal inflammation. Immunity, 43(4), 715-726.

398 Cheng, L., Chen, C. L., Luo, P., Tan, M., Qiu, M., Johnson, R., & Ma, Q. (2003). Lmx1b, Pet-
399 1, and Nkx2.2 coordinately specify serotonergic neurotransmitter phenotype. Journal of
400 Neuroscience, 23(31), 9961-9967.

401 Chen, Y., Guo, Y., Yang, H., Shi, G., Xu, G., Shi, J., ... & Chen, D. (2015). TRIM66
402 overexpression contributes to osteosarcoma carcinogenesis and indicates poor survival
403 outcome. Oncotarget, 6(27), 23708.

404 Chen, S. W., Chou, C. T., Chang, C. C., Li, Y. J., Chen, S. T., Lin, I. C., ... &Kuo, M. L. (2017).
405 HMGCS2 enhances invasion and metastasis via direct interaction with PPARa to activate Src
406 signaling in colorectal cancer and oral cancer. Oncotarget, 8(14), 22460.

407 Cowell, J. K., Qin, H., Hu, T., Wu, Q., Bhole, A., & Ren, M. (2017). Mutation in the FGFR1
408 tyrosine kinase domain or inactivation of PTEN is associated with acquired resistance to FGFR
409 inhibitors in FGFR1-driven leukemia/lymphomas. International Journal of Cancer.

410 Dai, Y., Wang, M., Wu, H., Xiao, M., Liu, H., & Zhang, D. (2017). Loss of FOXN3 in colon
411 cancer activates beta-catenin/TCF signaling and promotes the growth and migration of cancer
412 cells. Oncotarget, 8(6), 9783.

413 Davidson, G., &Niehrs, C. (2010). Emerging links between CDK cell cycle regulators and Wnt
414 signaling. Trends in cell biology, 20(8), 453-460.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ...&Storfer, A. (2016a). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Nature communications, 7, 12684.

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., ...&Storfer, A. (2016b). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.r60sv

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., &Cresko, W. A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Molecular methods for evolutionary genetics, 157-178.

French, C. A., Rahman, S., Walsh, E. M., Kühnle, S., Grayson, A. R., Lemieux, M. E., ...&Venkatramani, R. (2014). NSD3–NUT fusion oncoprotein in NUT midline carcinoma: implications for a novel oncogenic mechanism. Cancer discovery, 4(8), 928-941.

Garcia, M. J., Pole, J. C., Chin, S. F., Teschendorff, A., Naderi, A., Ozdag, H., ... & Ellis, I. (2005). A 1 Mb minimal amplicon at 8p11-12 in breast cancer identifies new candidate oncogenes. Oncogene, 24(33), 5235.

Hamede, R. K., Bashford, J., McCallum, H., & Jones, M. (2009). Contact networks in a wild Tasmanian devil (Sarcophilus harrisii) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. Ecology letters, 12(11), 1147-1157.

Hamede, R. K., McCallum, H., & Jones, M. (2013). Biting injuries and transmission of Tasmanian devil facial tumour disease. Journal of Animal Ecology, 82(1), 182-190.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. cell, 144(5), 646-674.

438 Harvey, K. F., Zhang, X., & Thomas, D. M. (2013). The Hippo pathway and human cancer.
439 Nature Reviews Cancer, 13(4), 246-257.

440 Hatakeyama, S. (2017). TRIM family proteins: roles in autophagy, immunity, and
441 carcinogenesis. Trends in biochemical sciences.

442 Hellsten, S. V., Hägglund, M. G., Eriksson, M. M., &Fredriksson, R. (2017). The neuronal and
443 astrocytic protein SLC38A10 transports glutamine, glutamate, and aspartate, suggesting a role
444 in neurotransmission. FEBS open bio, 7(6), 730-746.

445 Helsmoortel, C., Vandeweyer, G., Ordoukhanian, P., Van Nieuwerburgh, F., Van der Aa, N.,
446 &Kooy, R. F. (2015). Challenges and opportunities in the investigation of unexplained
447 intellectual disability using family-based whole-exome sequencing. Clinical genetics, 88(2),
448 140-148.

449 Hendricks, S., Epstein, B., Schönfeld, B., Wiench, C., Hamede, R., Jones, M., ...& Hohenlohe,
450 P. (2017). Conservation implications of limited genetic diversity and population structure in
451 Tasmanian devils (Sarcophilus harrisii). Conservation Genetics, 1-6.

452 Hollis, F., Kanellopoulos, A. K., &Bagni, C. (2017). Mitochondrial dysfunction in Autism
453 Spectrum Disorder: clinical features and perspectives. Current Opinion in Neurobiology, 45,
454 178-187.

455 Hnoonual, A., Thammachote, W., Tim-Aroon, T., Rojnueangnit, K., Hansakunachai, T.,
456 Sombuntham, T., ... & Wattanasirichaigoon, D. (2017). Chromosomal microarray analysis in a
457 cohort of underrepresented population identifies SERINC2 as a novel candidate gene for autism
458 spectrum disorder. *Scientific reports*, *7*(1), 12096.

459 Ioannou, M. S., & McPherson, P. S. (2016). Regulation of cancer cell behavior by the small
460 GTPase Rab13. Journal of Biological Chemistry, 291(19), 9929-9937.

461 Jiang, Y., Woronicz, J. D., Liu, W., &Goeddel, D. V. (1999). Prevention of constitutive TNF
462 receptor 1 signaling by silencer of death domains. Science, 283(5401), 543-546.

463 Jones, D. T., Hutter, B., Jäger, N., Korshunov, A., Kool, M., Warnatz, H. J., ...&Fontebasso, A.
464 M. (2013). Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma.
465 Nature genetics, 45(8), 927-932.

466 Juang, Y. L., Jeng, Y. M., Chen, C. L., & Lien, H. C. (2016). PRRX2 as a novel TGF-β-induced
467 factor enhances invasion and migration in mammary epithelial cell and correlates with poor
468 prognosis in breast cancer. Molecular carcinogenesis, 55(12), 2247-2259.

469 Kaufman, L., Ayub, M., & Vincent, J. B. (2010). The genetic basis of non-syndromic
470 intellectual disability: a review. Journal of neurodevelopmental disorders, 2(4), 182.

471 Kaya, N., Alsagob, M., D'adamo, M. C., Al-Bakheet, A., Hasan, S., Muccioli, M., ...& Mustafa,
472 O. M. (2016). KCNA4 deficiency leads to a syndrome of abnormal striatum, congenital cataract
473 and intellectual disability. Journal of medical genetics, 53(11), 786-792.

474 Krishnaswamy, A., Yamagata, M., Duan, X., Hong, Y. K., &Sanes, J. R. (2015). Sidekick 2
475 directs formation of a retinal circuit that detects differential motion. Nature, 524(7566), 466.

476 Lachish, S., McCallum, H., & Jones, M. (2009). Demography, disease and the devil: life-history
477 changes in a disease-affected population of Tasmanian devils (Sarcophilus harrisii). Journal of
478 Animal Ecology, 78(2), 427-436.

479 Li, H. L., Song, J., Yong, H. M., Hou, P. F., Chen, Y. S., Song, W. B., ... & Zheng, J. N. (2016a).
480 PinX1: structure, regulation and its functions in cancer. Oncotarget, 7(40), 66267.

481 Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., ...& Xia, K. (2016b). Genes with de novo
482 mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database.
483 Molecular psychiatry, 21(2), 290.

Little, E. C., Camp, E. R., Wang, C., Watson, P. M., Watson, D. K., & Cole, D. J. (2016). The CaSm (LSm1) oncogene promotes transformation, chemoresistance and metastasis of pancreatic cancer cells. Oncogenesis, 5(1), e182.

Liu, X. H., Yang, Y. F., Fang, H. Y., Wang, X. H., Zhang, M. F., & Wu, D. C. (2017). CEP131 indicates poor prognosis and promotes cell proliferation and migration in hepatocellular carcinoma. The International Journal of Biochemistry & Cell Biology.

Lovero, K. L., Blankenship, S. M., Shi, Y., & Nicoll, R. A. (2013). SynDIG1 promotes excitatory synaptogenesis independent of AMPA receptor trafficking and biophysical regulation. PLoS One, 8(6), e66171.

Ma, Y., Yue, Y., Pan, M., Sun, J., Chu, J., Lin, X., ... & Shin, V. Y. (2015). Histone deacetylase 3 inhibits new tumor suppressor gene DTWD1 in gastric cancer. American journal of cancer research, 5(2), 663.

Maciejowski, J., & de Lange, T. (2017). Telomeres in cancer: tumour suppression and genome instability. Nature Reviews Molecular Cell Biology.

Metzger, M. J., Villalba, A., Carballal, M. J., Iglesias, D., Sherry, J., Reinisch, C., ...& Goff, S. P. (2016). Widespread transmission of independent cancer lineages within multiple bivalve species. Nature, 534(7609), 705-709.

Murchison, E. P., Schulz-Trieglaff, O. B., Ning, Z., Alexandrov, L. B., Bauer, M. J., Fu, B., ... & Ng, B. L. (2012). Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. Cell, 148(4), 780-791.

Nakata, T., &Hirokawa, N. (2007). Neuronal polarity and the kinesin superfamily proteins. Sci. STKE, 2007(372), pe6-pe6.

Nijmeijer, S., Vischer, H. F., & Leurs, R. (2016). Adhesion GPCRs in immunology. Biochemical pharmacology, 114, 88-102.

Nowell, C. S., &Radtke, F. (2017). Notch as a tumour suppressor. Nature Reviews Cancer, 17(3), 145-159.

Nusse, R., &Clevers, H. (2017). Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell, 169(6), 985-999.

Okada, H., &Mak, T. W. (2004). Pathways of apoptotic and non-apoptotic death in tumour cells. Nature reviews. Cancer, 4(8), 592.

Orsmond, G. I., Krauss, M. W., & Seltzer, M. M. (2004). Peer relationships and social and recreational activities among adolescents and adults with autism. Journal of autism and developmental disorders, 34(3), 245-256.

Pagnamenta, A. T., Khan, H., Walker, S., Gerrelli, D., Wing, K., Bonaglia, M. C., ...& Pinto, D. (2011). Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (CDH8) in susceptibility to autism and learning disability. Journal of medical genetics, 48(1), 48-54.

Pal, I., & Mandal, M. (2012). PI3K and Akt as molecular targets for cancer therapy: current clinical outcomes. ActaPharmacologicaSinica, 33(12), 1441.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ...& Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559-575.

Pye, R. J., Woods, G. M., &Kreiss, A. (2016a). Devil facial tumor disease. *Veterinary pathology*, *53*(4), 726-736.

528 Pye, R., Hamede, R., Siddle, H. V., Caldwell, A., Knowles, G. W., Swift, K., ... & Woods, G.
529 M. (2016b). Demonstration of immune responses against devil facial tumour disease in wild
530 Tasmanian devils. Biology letters, 12(10), 20160553.

531 Poot, M., Beyer, V., Schwaab, I., Damatova, N., van't Slot, R., Prothero, J., ...&Haaf, T. (2010).
532 Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay
533 and autism spectrum disorder. Neurogenetics, 11(1), 81-89.

534 Rhein, V. F., Carroll, J., Ding, S., Fearnley, I. M., & Walker, J. E. (2016). NDUFAF5
535 hydroxylates NDUFS7 at an early stage in the assembly of human complex I. Journal of
536 Biological Chemistry, 291(28), 14851-14860.

537 Roche, B., Moller, A.P., de Gregori, J., & Thomas, F. (2017). Cancer in animals: reciprocal
538 feedbacks between evolution of cancer resistance and ecosystem functioning. In "Ecology and
539 Evolution of Cancer", edited by Ujvari, B., Roche, B. and Thomas, F. Elsevier. pp. 180-188.

540 Samanta, D., Park, Y., Andrabi, S. A., Shelton, L. M., Gilkes, D. M., & Semenza, G. L. (2016).
541 PHGDH expression is required for mitochondrial redox homeostasis, breast cancer stem cell
542 maintenance, and lung metastasis. Cancer research, 76(15), 4430-4442.

543 Sandri, C., Caccavari, F., Valdembri, D., Camillo, C., Veltel, S., Santambrogio, M., ... &Serini,
544 G. (2012). The R-Ras/RIN2/Rab5 complex controls endothelial cell adhesion and
545 morphogenesis via active integrin endocytosis and Rac signaling. Cell research, 22(10), 1479.

546 Schiffman, J. D., & Breen, M. (2015). Comparative oncology: what dogs and other species can
547 teach us about humans with cancer. Phil. Trans. R. Soc. B, 370(1673), 20140231.

548 Shiina, T., Hosomichi, K., Inoko, H., &Kulski, J. K. (2009). The HLA genomic loci map:
549 expression, interaction, diversity and disease. Journal of human genetics, 54(1).

Shmelkov, S. V., Hormigo, A., Jing, D., Proenca, C. C., Bath, K. G., Milde, T., ... & Murphy, A. J. (2010). Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive-like behaviors in mice. Nature medicine, 16(5), 598-602.

Staples, C. J., Myers, K. N., Beveridge, R. D., Patil, A. A., Lee, A. J., Swanton, C., ... & Collis, S. J. (2012). The centriolar satellite protein Cep131 is important for genome stability. J Cell Sci, 125(20), 4770-4779.

Storfer, A., Epstein, B., Jones, M., Micheletti, S., Spear, S. F., Lachish, S., & Fox, S. (2017). Landscape genetics of the Tasmanian devil: implications for spread of an infectious cancer. Conservation Genetics, 1-11.

Strub, T., Kobi, D., Koludrovic, D., & Davidson, I. (2011). A POU3F2-MITF-SHC4 axis in phenotype switching of melanoma cells. In Research on Melanoma - A Glimpse into Current Directions and Future Trends, InTech, doi: 10.5772/19769.

Syed, V. (2016). TGF-β Signaling in Cancer. Journal of cellular biochemistry, 117(6), 1279-1287.

Taheri, A., Perry, A., &Minnes, P. (2016). Examining the social participation of children and adolescents with Intellectual Disabilities and Autism Spectrum Disorder in relation to peers. Journal of Intellectual Disability Research, 60(5), 435-443.

Tang, K., Thornton, K. R., &Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS biology, 5(7), e171.

Tang, P. M. K., Zhou, S., Meng, X. M., Wang, Q. M., Li, C. J., Lian, G. Y., ...& To, K. F. (2017). Smad3 promotes cancer progression by inhibiting E4BP4-mediated NK cell development. Nature Communications, 8, 14677.

572 Thépot, S., Restoux, G., Goldringer, I., Gouache, D., Hospital, F., Mackay, I., & Enjalbert, J.
573 (2015). Efficiently tracking selection in a multiparental population: the case of earliness in
574 wheat. Genetics, 199(2), 609-623.

575 Utine, G. E., Taşkıran, E. Z., Koşukcu, C., Karaosmanoğlu, B., Güleray, N., Doğan, Ö. A., ...
576 &Alikaşifoğlu, M. (2017). HERC1 mutations in idiopathic intellectual disability. European
577 Journal of Medical Genetics, 60(5), 279-283.

578 Veenbergen, S., & van Spriel, A. B. (2011). Tetraspanins in the immune response against
579 cancer. Immunology letters, 138(2), 129-136.

580 Versteeg, G. A., Rajsbaum, R., Sánchez-Aparicio, M. T., Maestre, A. M., Valdiviezo, J., Shi,
581 M., ...&García-Sastre, A. (2013). The E3-ligase TRIM family of proteins regulates signaling
582 pathways triggered by innate immune pattern-recognition receptors. Immunity, 38(2), 384-398.

583 Vougiouklakis, T., Hamamoto, R., Nakamura, Y., &Saloura, V. (2015). The NSD family of
584 protein methyltransferases in human cancer. Epigenomics 7:5.

585 Walton, K. M., & Ingersoll, B. R. (2013). Improving social skills in adolescents and adults with
586 autism and severe to profound intellectual disability: A review of the literature. Journal of
587 Autism and Developmental Disorders, 43(3), 594-615.

588 Wells, K., Hamede, R. K., Kerlin, D. H., Storfer, A., Hohenlohe, P. A., Jones, M. E., &
589 McCallum, H. I. (2017). Infection of the fittest: devil facial tumour disease has greatest effect
590 on individuals with highest reproductive output. Ecology Letters, 20(6), 770-778.

591 Woodbury-Smith, M., Paterson, A. D., Thiruvahindrapduram, B., Lionel, A. C., Marshall, C.
592 R., Merico, D., ...& Chrysler, C. (2015). Using extended pedigrees to identify novel autism
593 spectrum disorder (ASD) candidate genes. Human genetics, 134(2), 191-201.

Wu, Z., Wu, Z., Li, J., Yang, X., Wang, Y., Yu, Y., ...& Zhang, Z. (2012). MCAM is a novel metastasis marker and regulates spreading, apoptosis and invasion of ovarian cancer cells. Tumor Biology, 33(5), 1619-1628.

Yan, Z., Kim, E., Datta, D., Lewis, D. A., &Soderling, S. H. (2016). Synaptic actin dysregulation, a convergent mechanism of mental disorders?.*Journal of Neuroscience*, *36*(45), 11411-11417.

Yan, Y., Fan, Q., Wang, L., Zhou, Y., Li, J., & Zhou, K. (2017). LncRNA Snhg1, a non-degradable sponge for miR-338, promotes expression of proto-oncogene CST3 in primary esophageal cancer cells. Oncotarget, 8(22), 35750.

Yang, Y. A., Zhang, G. M., Feigenbaum, L., & Zhang, Y. E. (2006). Smad3 reduces susceptibility to hepatocarcinoma by sensitizing hepatocytes to apoptosis through downregulation of Bcl-2. Cancer cell, 9(6), 445-457.

Yee, N. S. (2015). Roles of TRPM8 ion channels in cancer: proliferation, survival, and invasion. Cancers, 7(4), 2134-2146.

Zhang, B., Zhang, H., Wang, D., Han, S., Wang, K., Yao, A., & Li, X. (2014). Never in mitosis gene A‐related kinase 6 promotes cell proliferation of hepatocellular carcinoma via cyclin B modulation. Oncology letters, 8(3), 1163-1168.

Zhang, D. H., Yang, Z. L., Zhou, E. X., Miao, X. Y., Zou, Q., Li, J. H., ... & Chen, S. L. (2016). Overexpression of Thy1 and ITGA6 is associated with invasion, metastasis and poor prognosis in human gallbladder carcinoma. Oncology letters, 12(6), 5136-5144.

Zhi, X., Lin, L., Yang, S., Bhuvaneshwar, K., Wang, H., Gusev, Y., ...& Tian, X. (2015). βII-Spectrin (SPTBN1) suppresses progression of hepatocellular carcinoma and Wnt signaling by regulation of Wnt inhibitor kallistatin. Hepatology, 61(2), 598-612.

## Supporting information captions

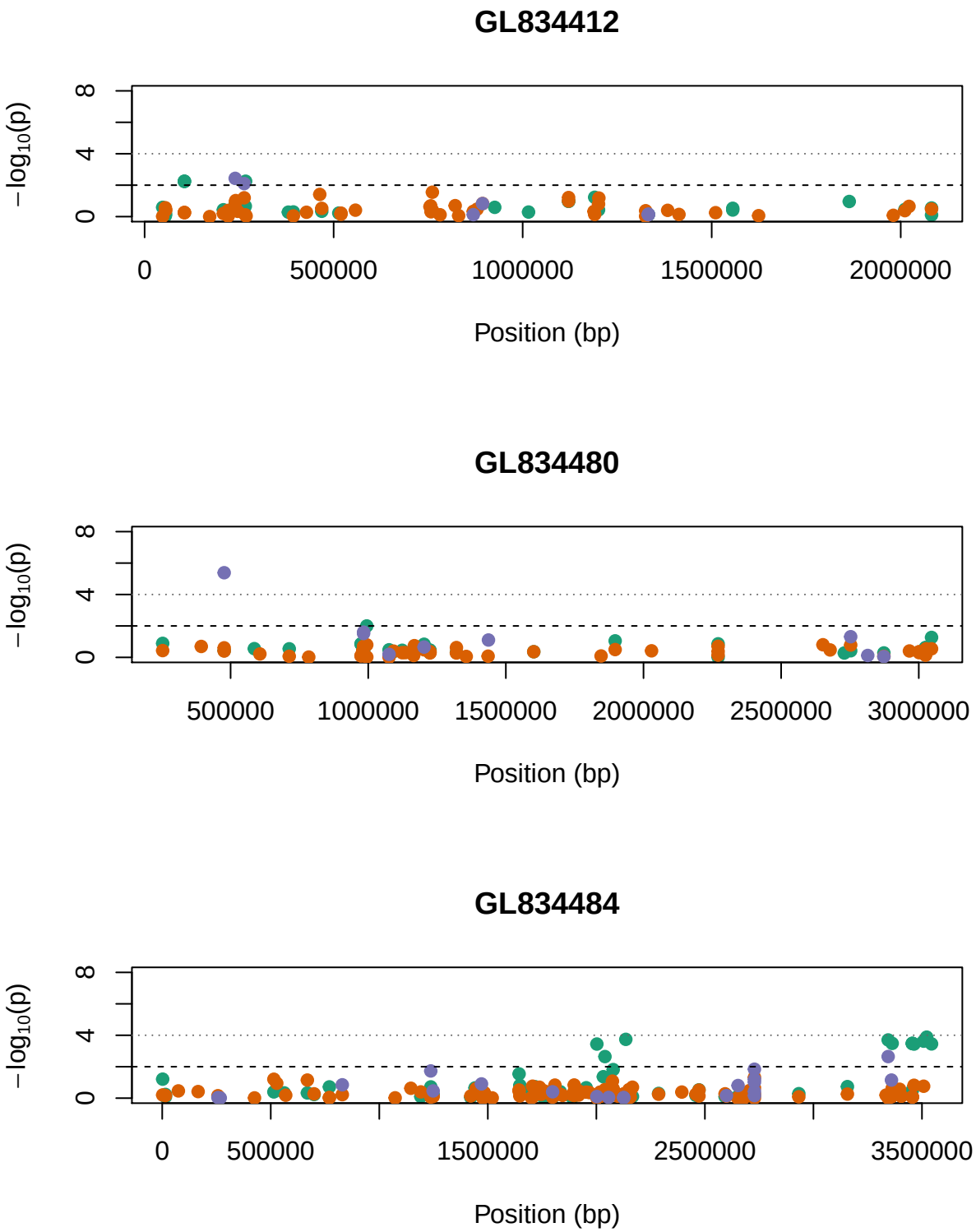
**S1 Fig.** Sixty signatures of selection were identified in 53 scaffolds in the Tasmanian devil genome within 100 kb of a protein coding gene having a human orthologue.

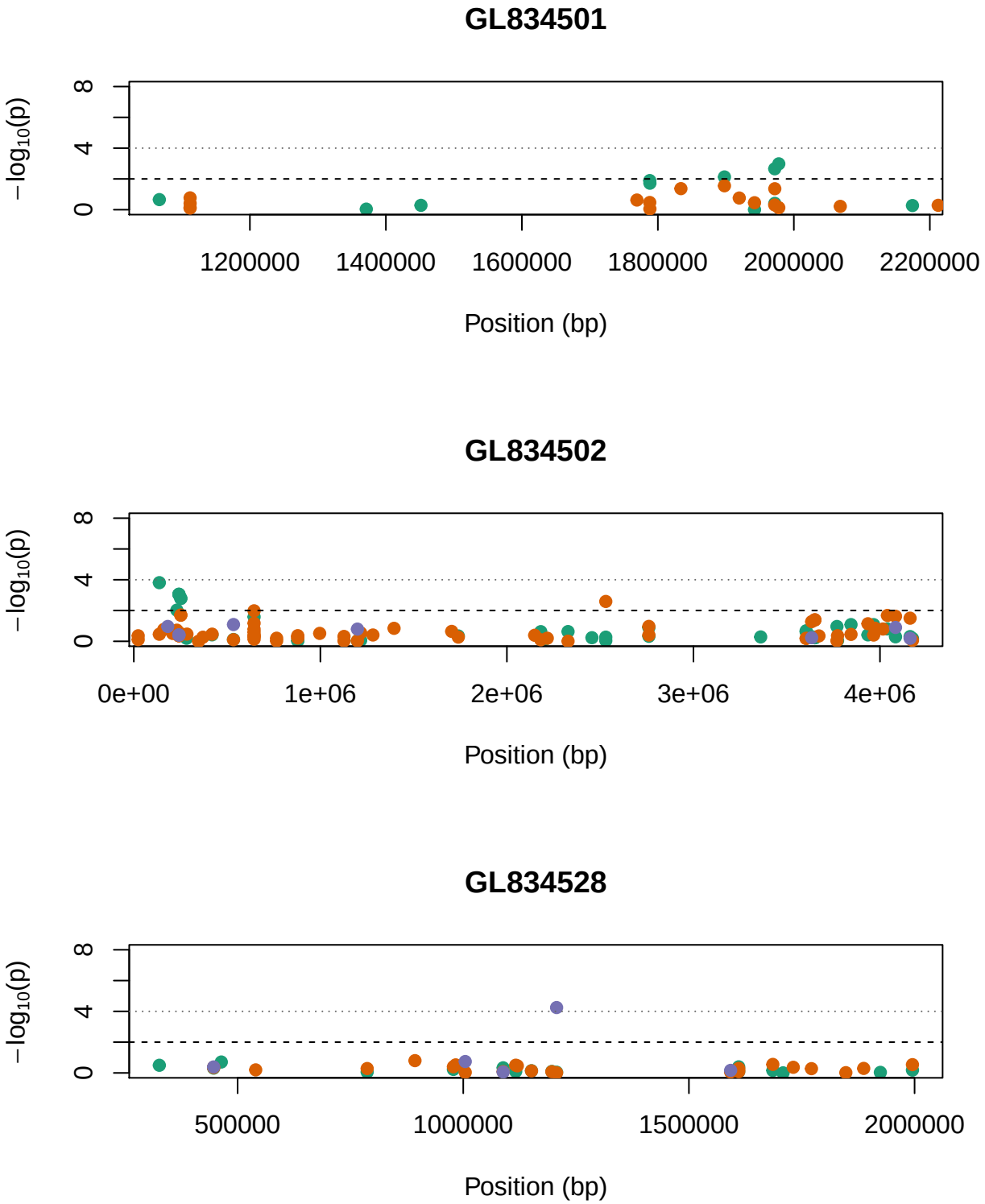
**S1 Table.** List of candidate genes for DFTD-driven selection in the Tasmanian devil.

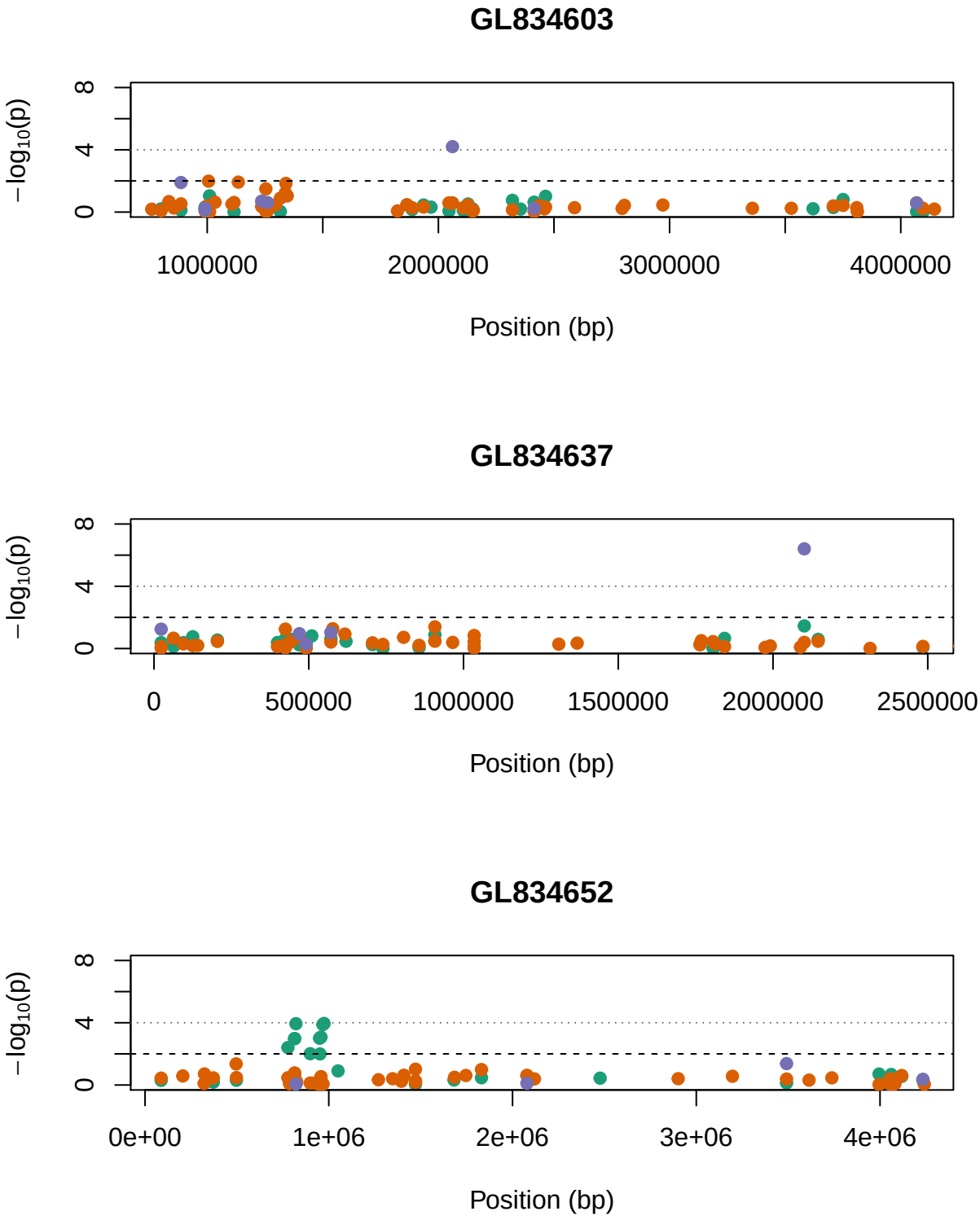
**S2 Table.** Top annotation terms associated by IPA with the list of candidate genes.

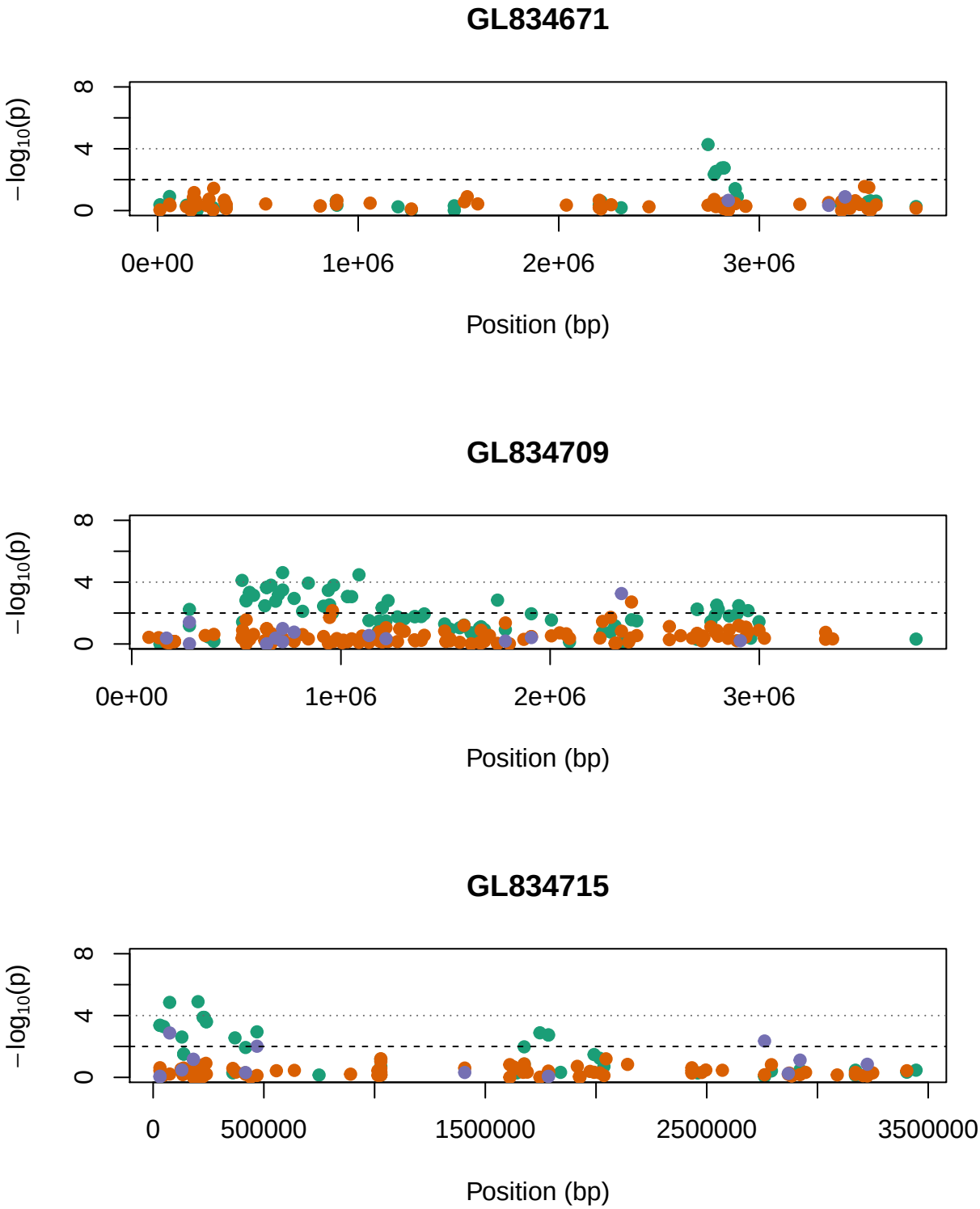
**S2 Fig.** Sixteen candidate genes are associated with the annotation term 'Development of neurons'.

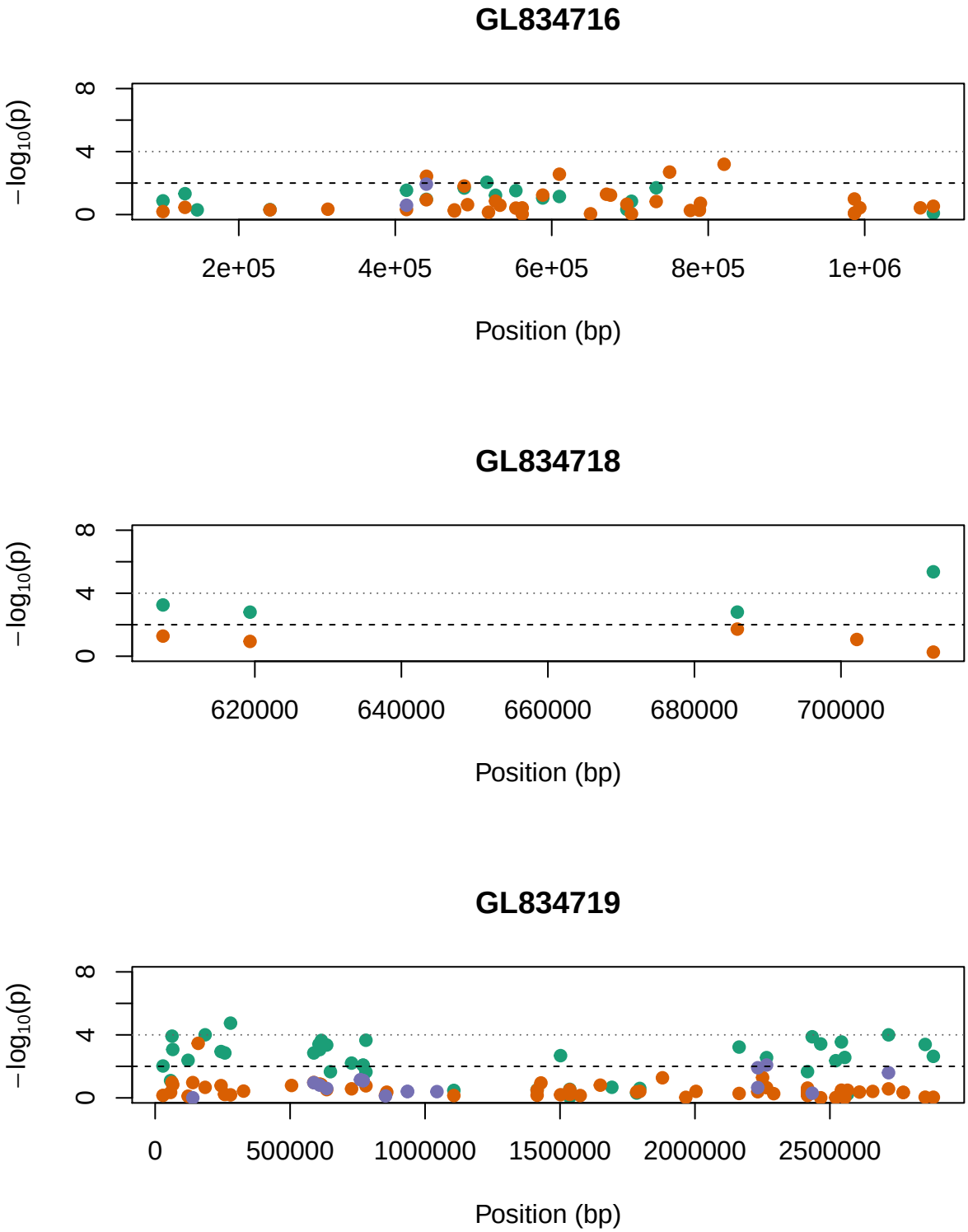
Figure supplémentaire 1 de l'Article II

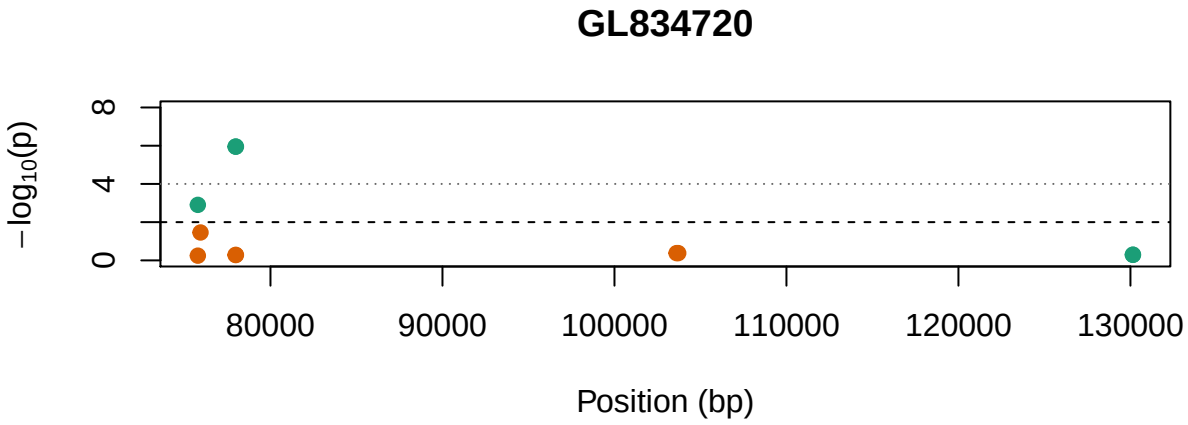


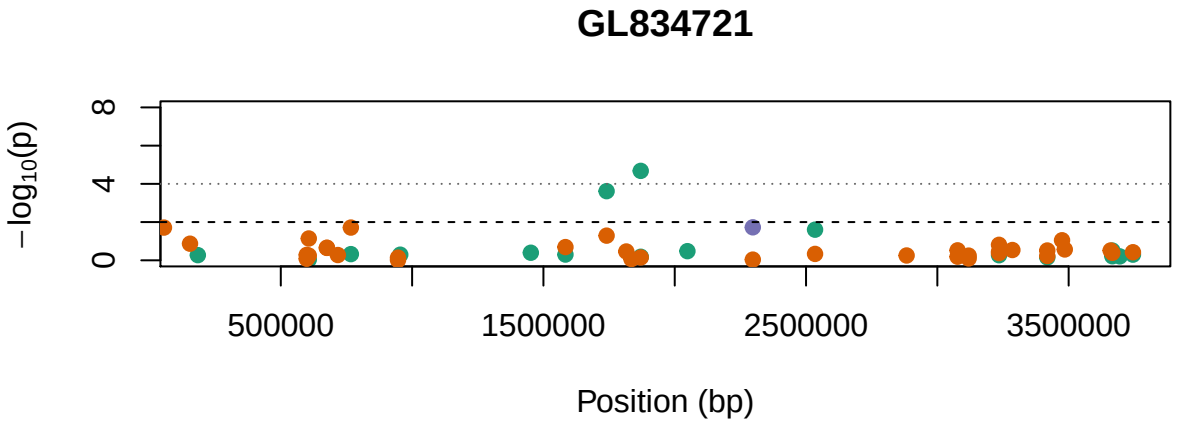

GL834501
−log10(p)
0
4
8
1200000
1400000
1600000
1800000
2000000
2200000
Position (bp)
GL834502
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
4e+06
Position (bp)
GL834528
−log10(p)
0
4
8
500000
1000000
1500000
2000000
Position (bp)

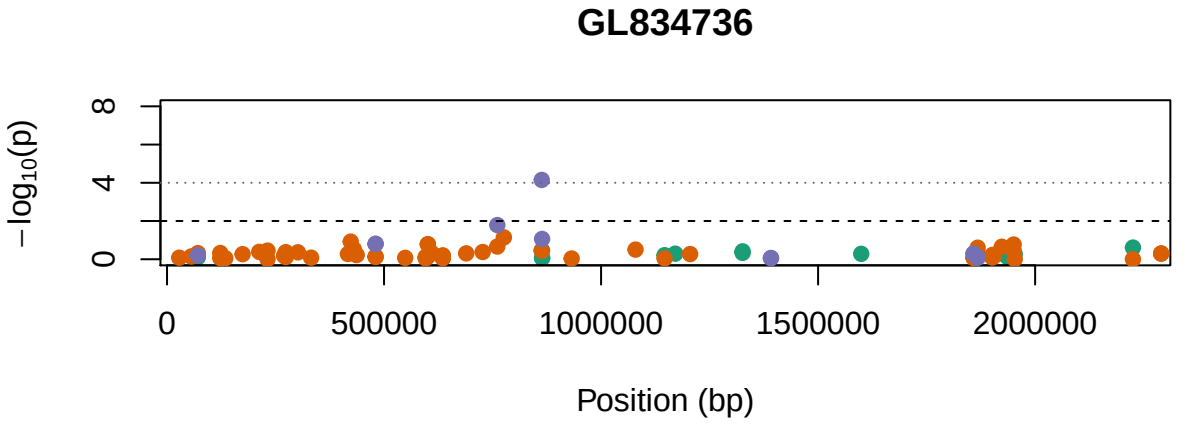

GL834603
−log10(p)
0
4
8
1000000
2000000
3000000
4000000
Position (bp)
GL834637
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
2500000
Position (bp)
GL834652
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
4e+06
Position (bp)

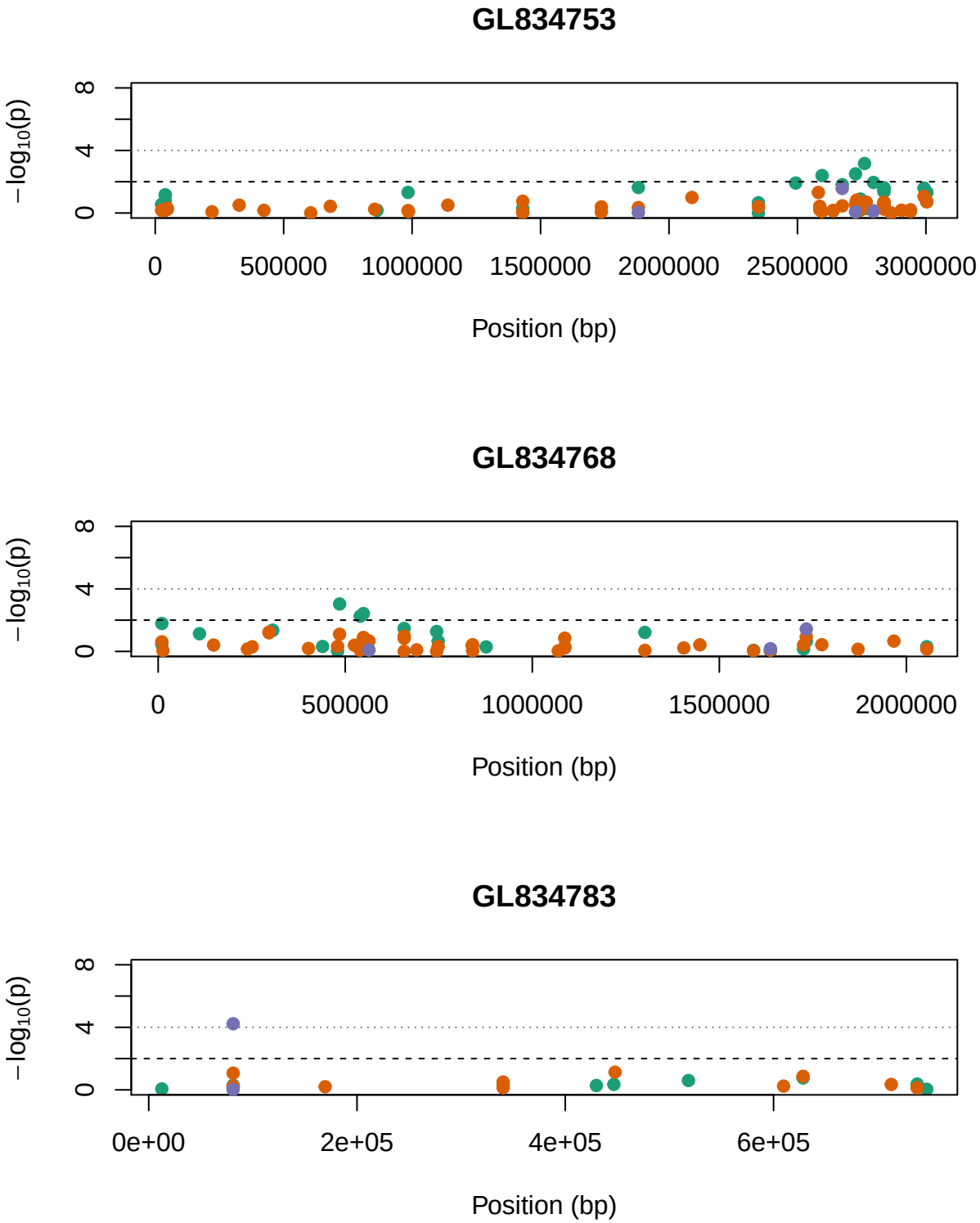

GL834671
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
Position (bp)
GL834709
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
Position (bp)
GL834715
−log10(p)
0
4
8
0
500000
1500000
2500000
3500000
Position (bp)

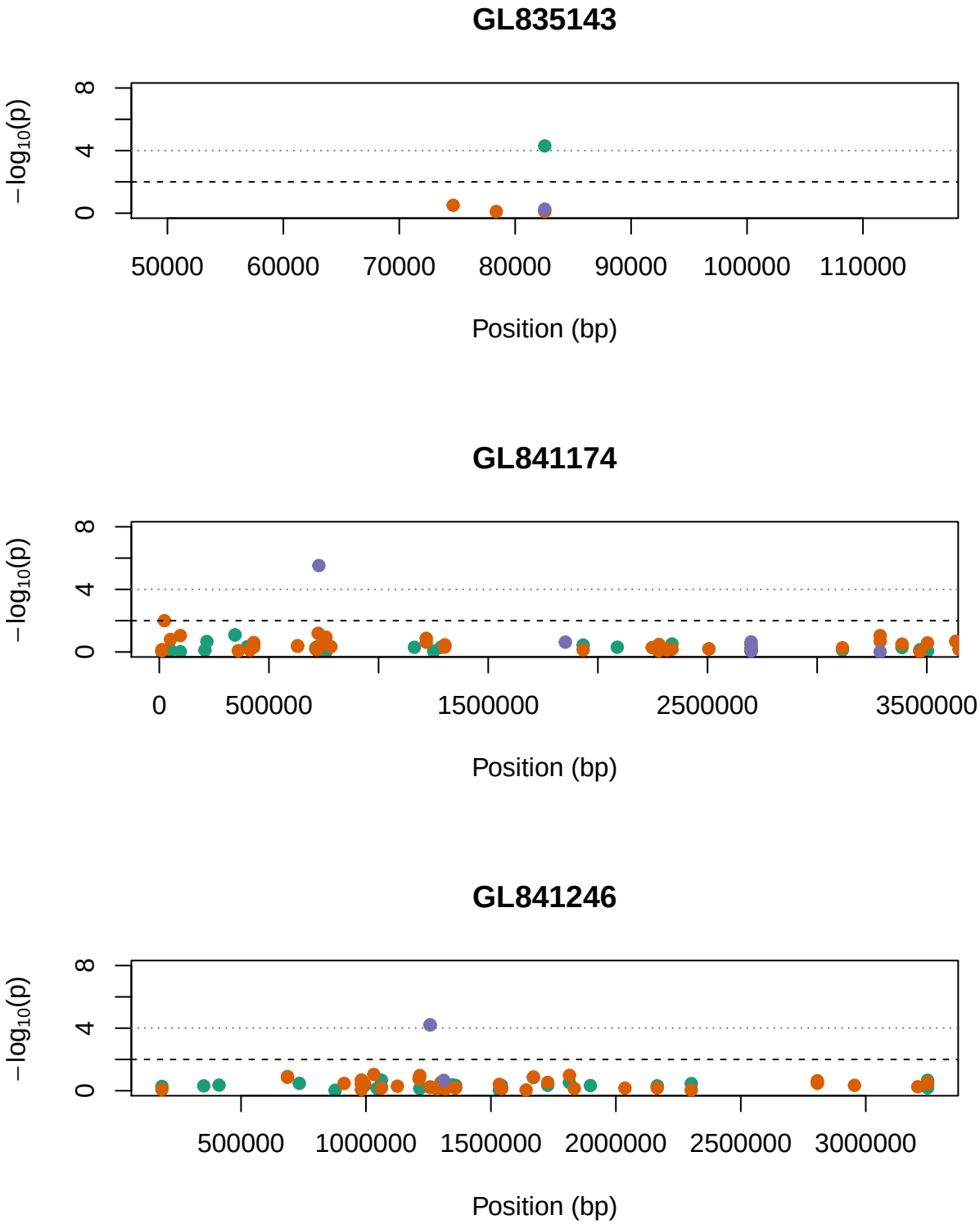

GL834716
−log10(p)
0
4
8
2e+05
4e+05
6e+05
8e+05
1e+06
Position (bp)
GL834718
−log10(p)
0
4
8
620000
640000
660000
680000
700000
Position (bp)
GL834719
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
2500000
Position (bp)

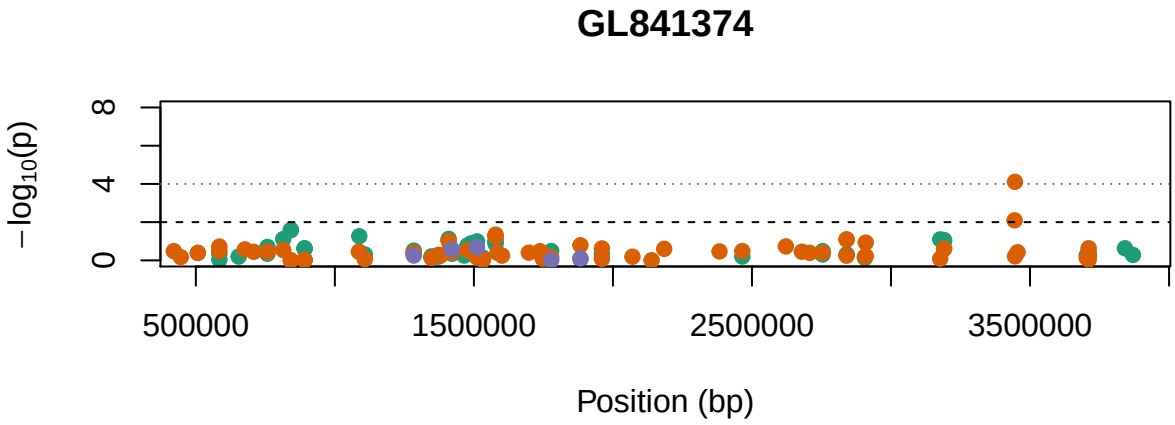

GL834720
−log10(p)
0
4
8
80000
90000
100000
110000
120000
130000
Position (bp)

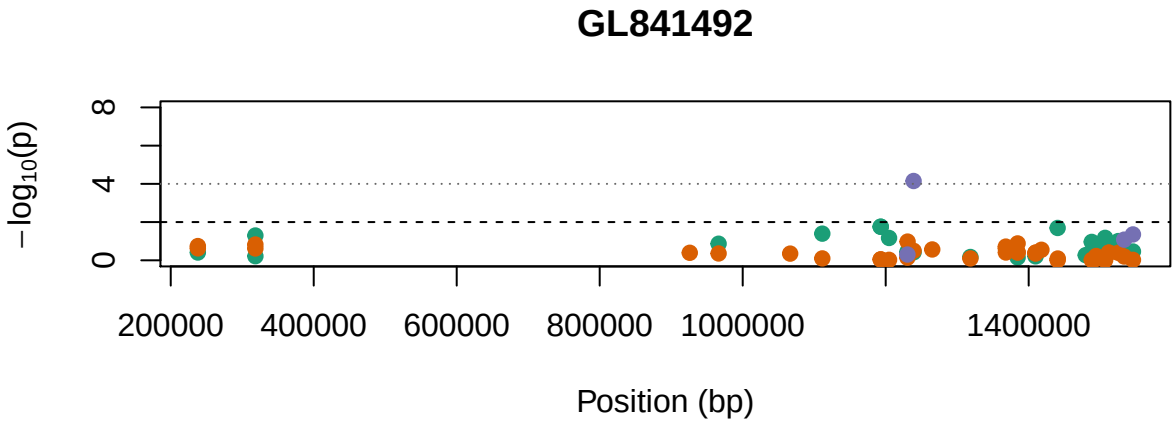

GL834721
−log10(p)
0
4
8
500000
1500000
2500000
3500000
Position (bp)

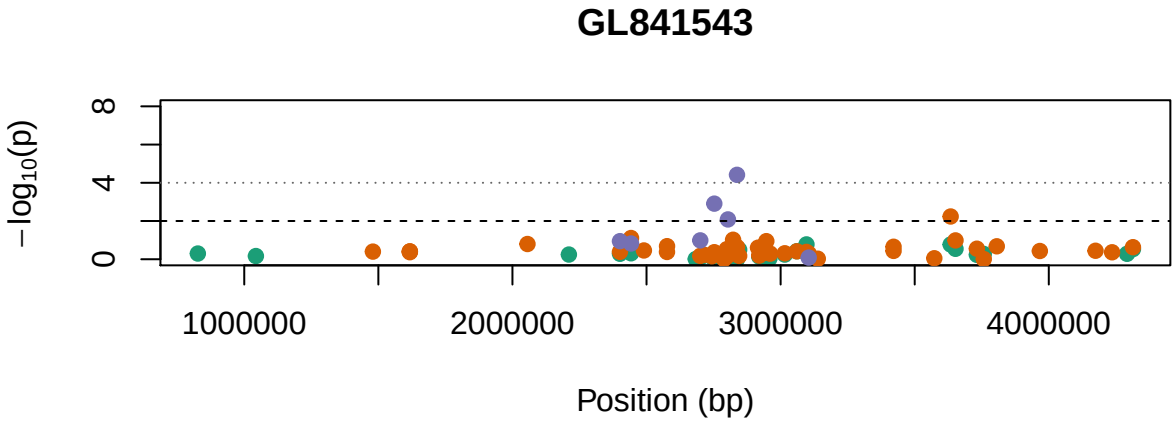

GL834736
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
Position (bp)

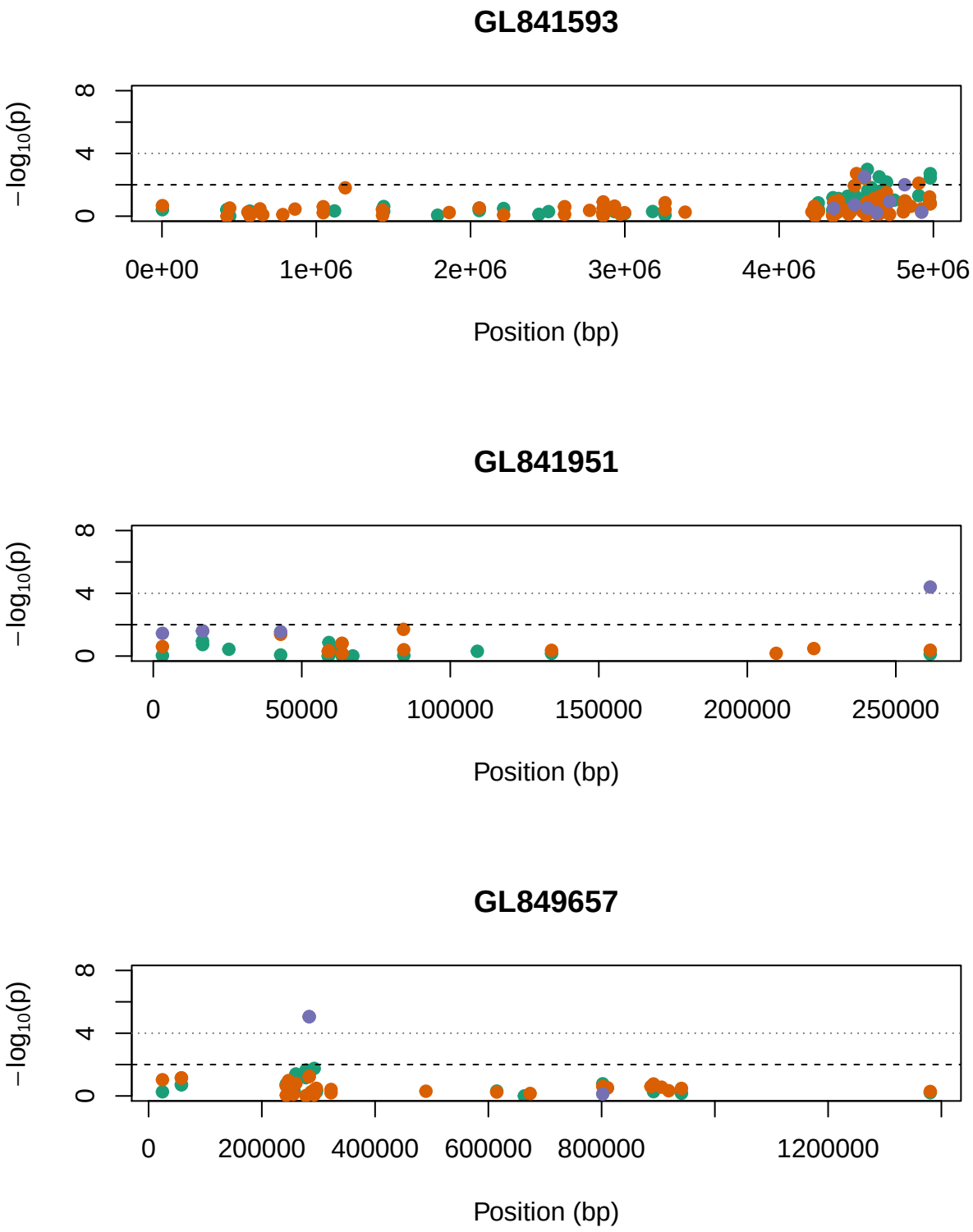

GL834753
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
2500000
3000000
Position (bp)
GL834768
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
Position (bp)
GL834783
−log10(p)
0
4
8
0e+00
2e+05
4e+05
6e+05
Position (bp)

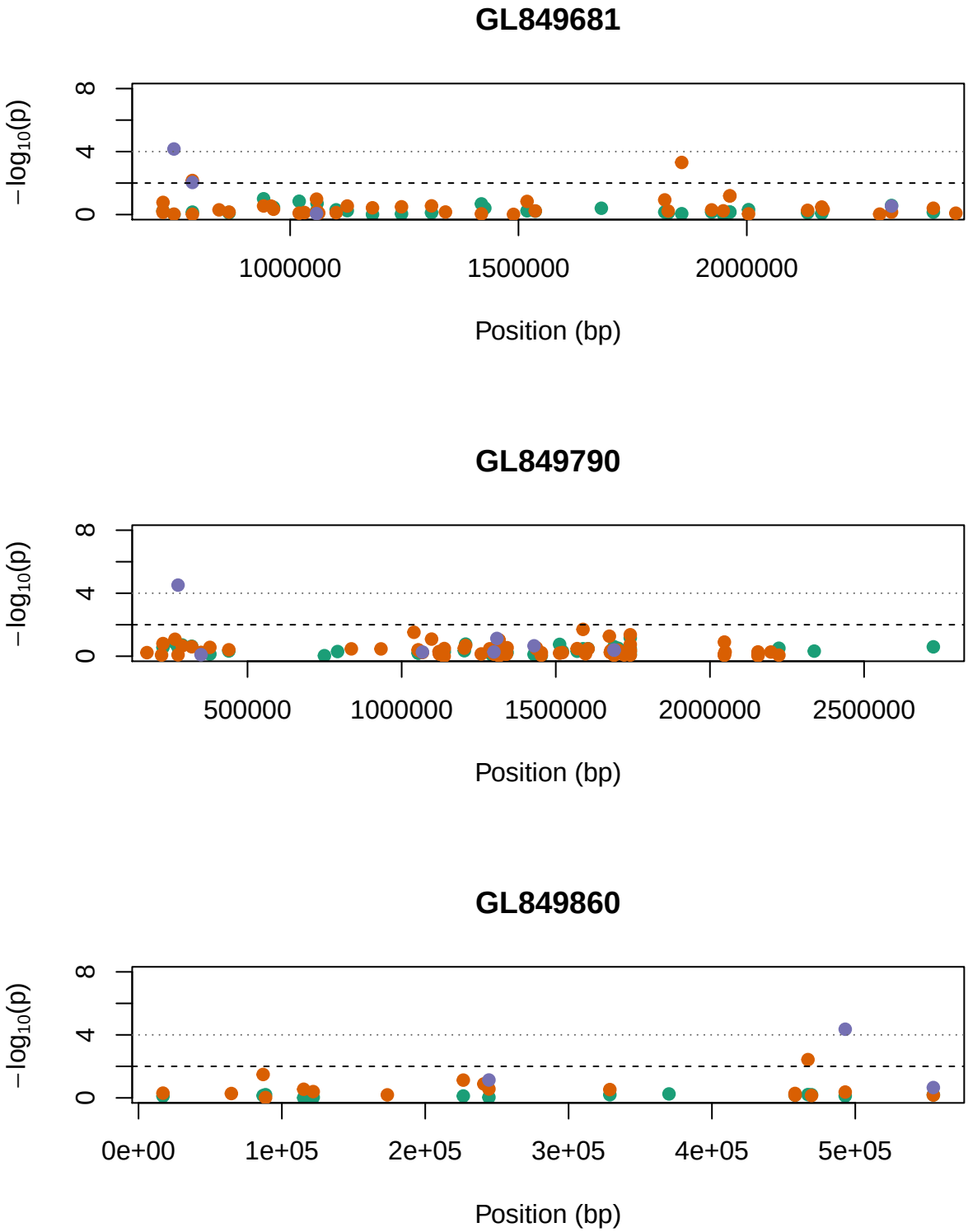

GL835143
−log10(p)
0
4
8
50000
60000
70000
80000
90000
100000
110000
Position (bp)
GL841174
−log10(p)
0
4
8
0
500000
1500000
2500000
3500000
Position (bp)
GL841246
−log10(p)
0
4
8
500000
1000000
1500000
2000000
2500000
3000000
Position (bp)

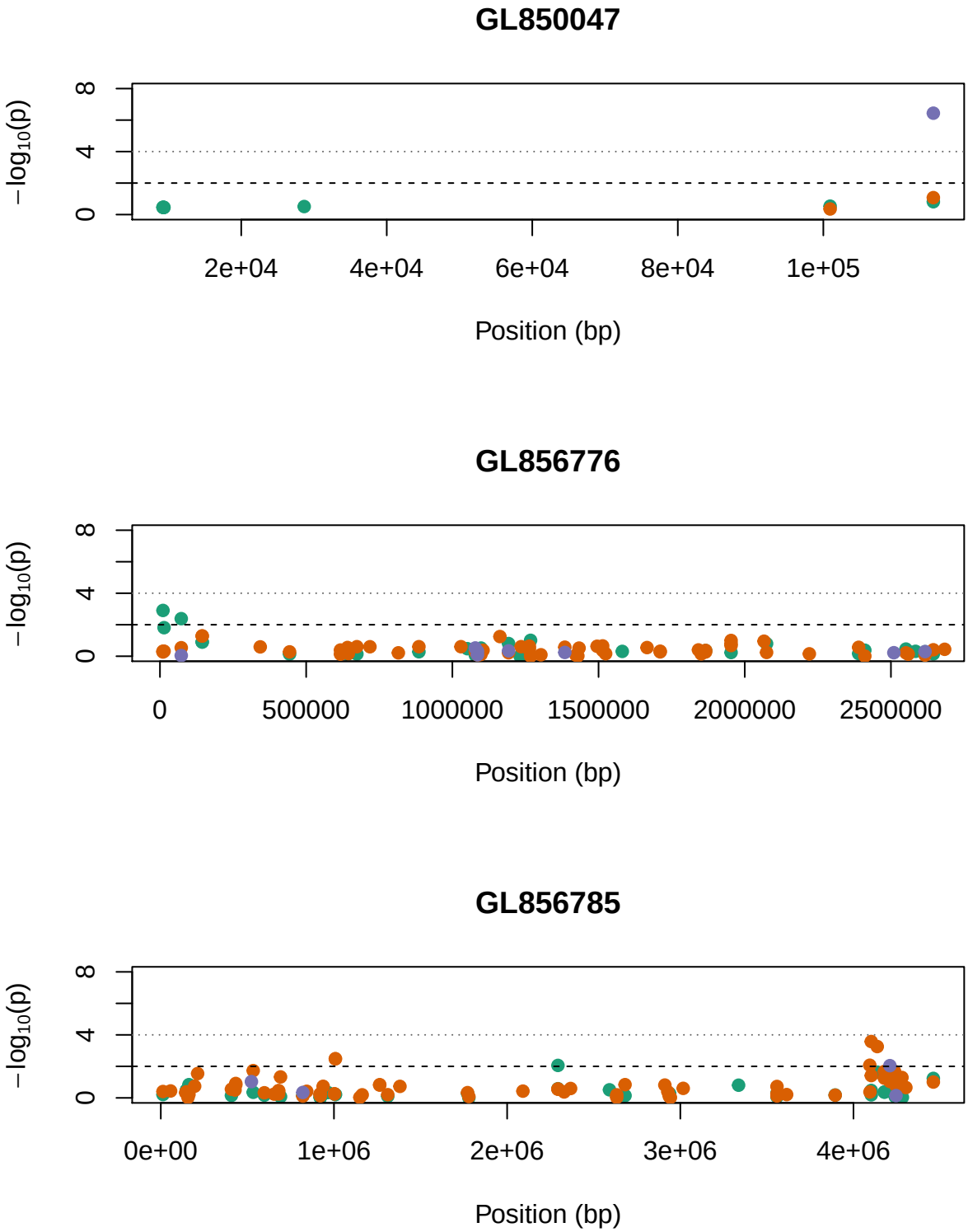

GL841374
−log10(p)
0
4
8
500000
1500000
2500000
3500000
Position (bp)

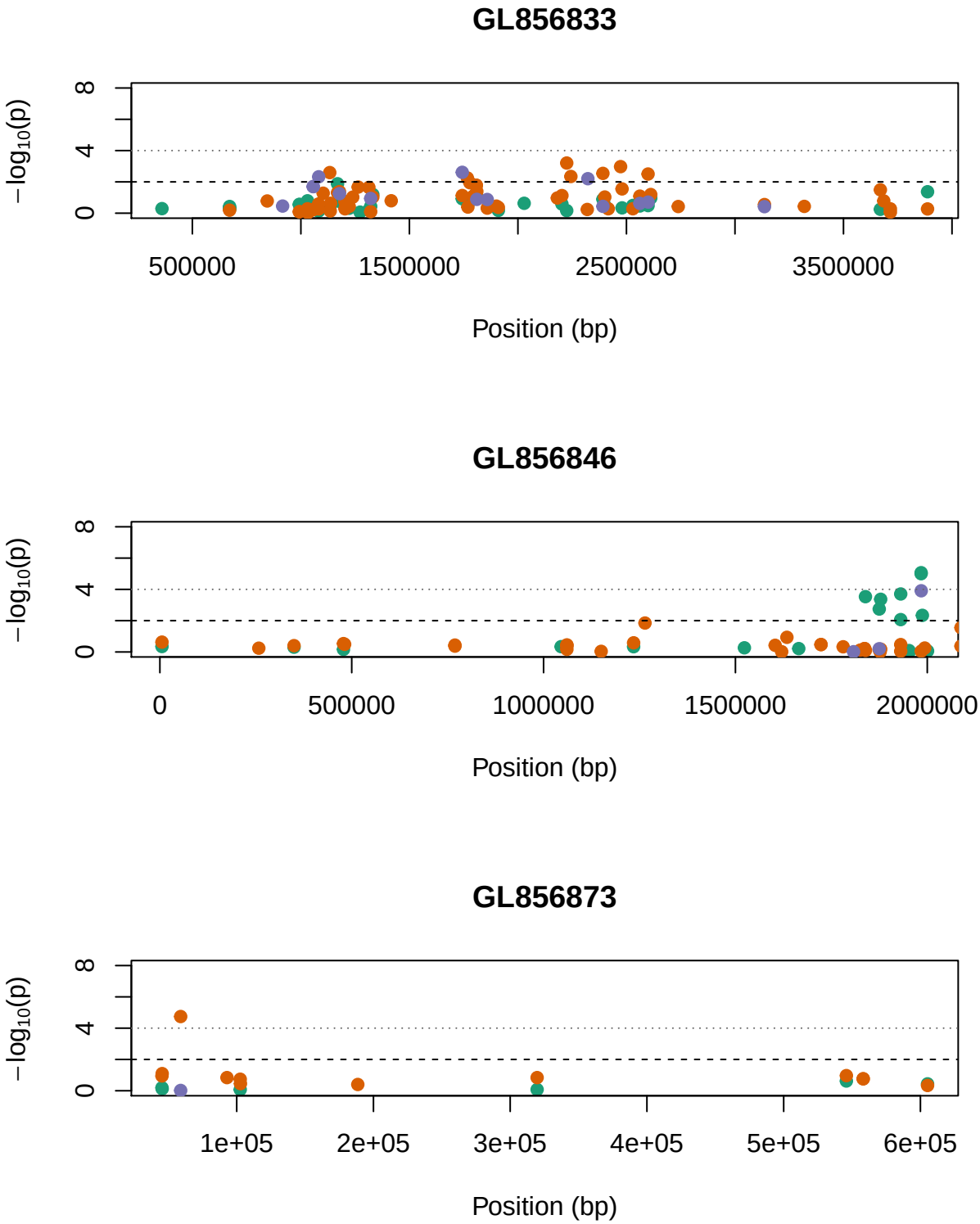

GL841492
−log10(p)
0
4
8
200000
400000
600000
800000
1000000
1400000
Position (bp)

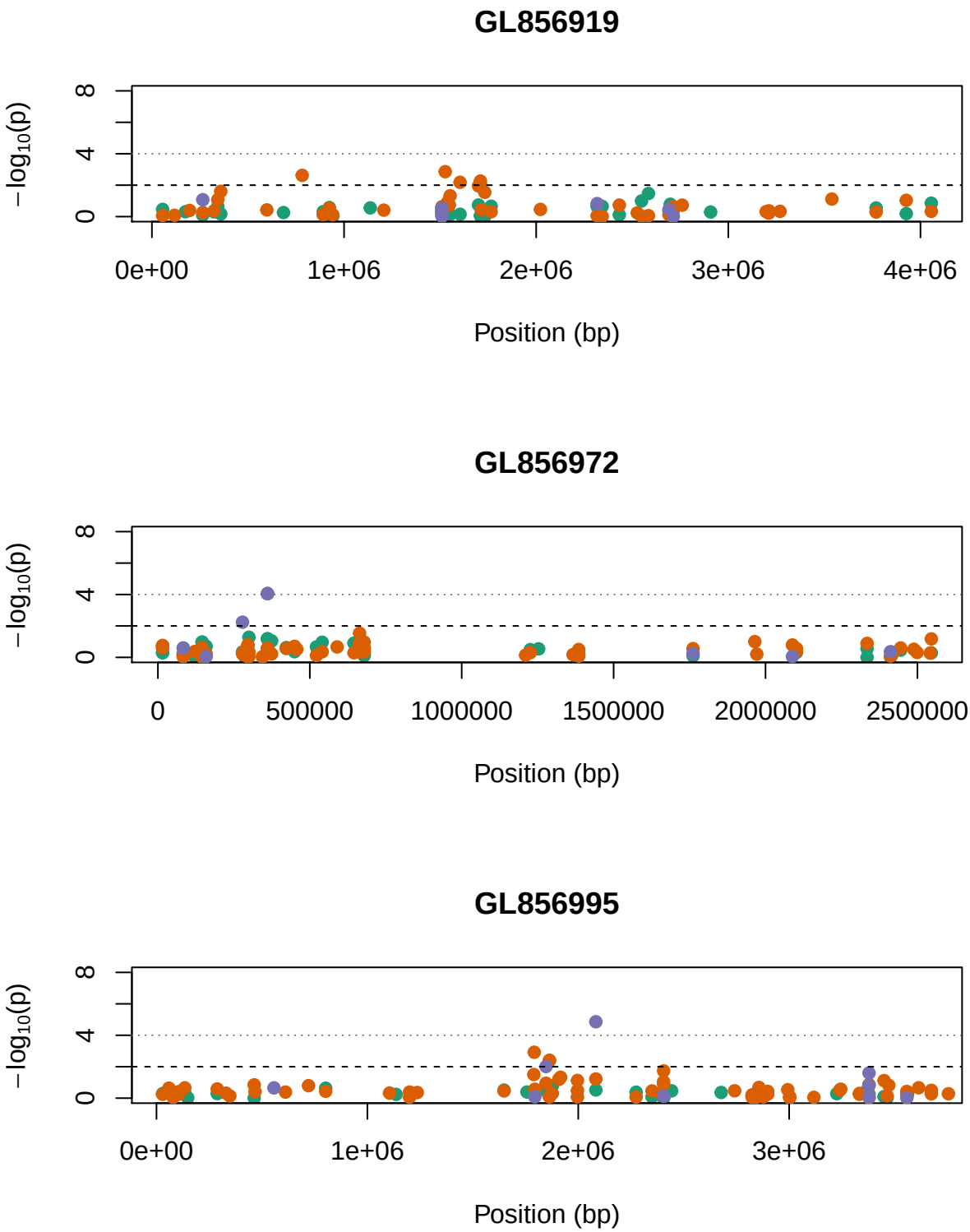

GL841543
−log10(p)
0
4
8
1000000
2000000
3000000
4000000
Position (bp)

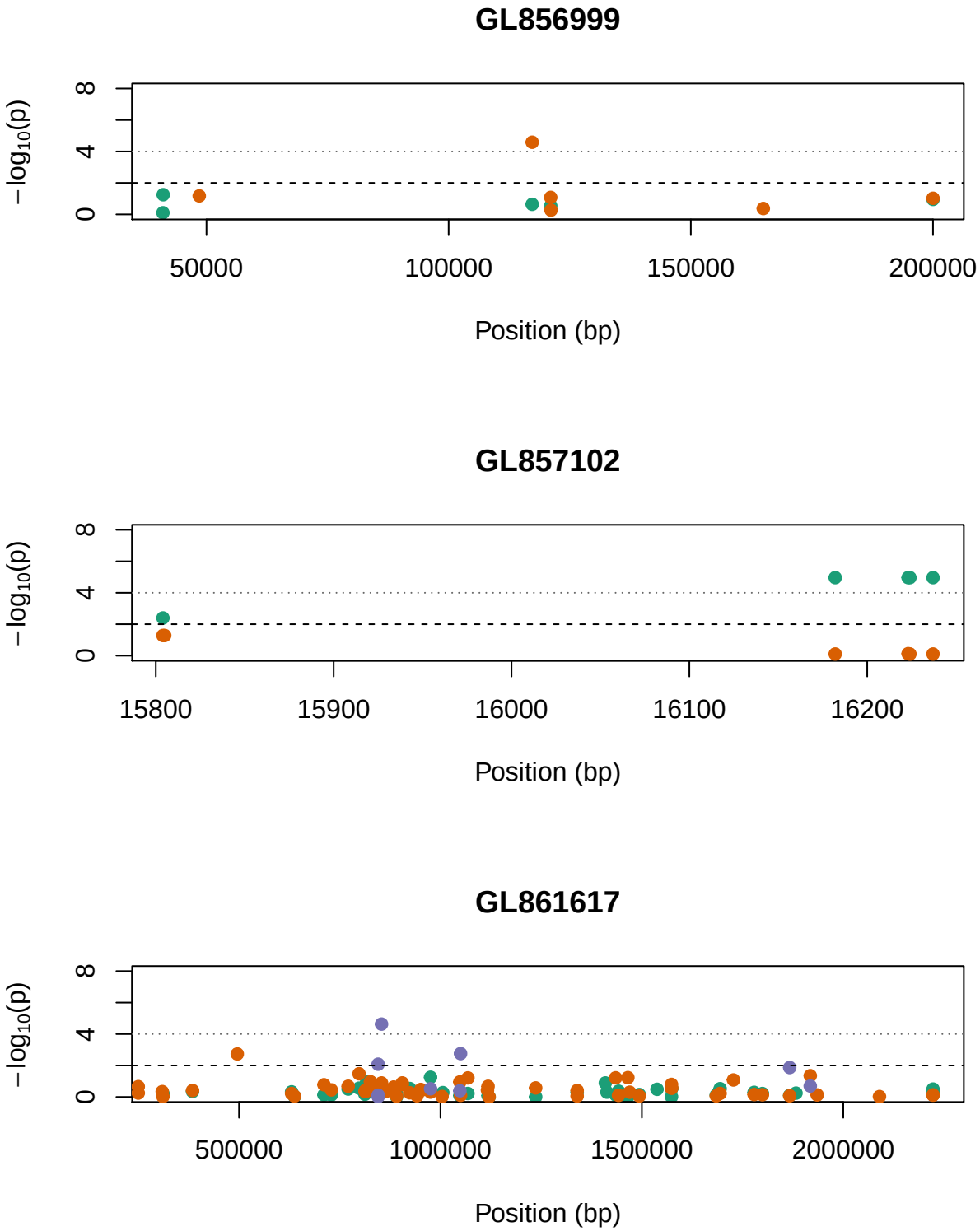

GL841593
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Position (bp)
GL841951
−log10(p)
0
4
8
0
50000
100000
150000
200000
250000
Position (bp)
GL849657
−log10(p)
0
4
8
0
200000
400000
600000
800000
1200000
Position (bp)

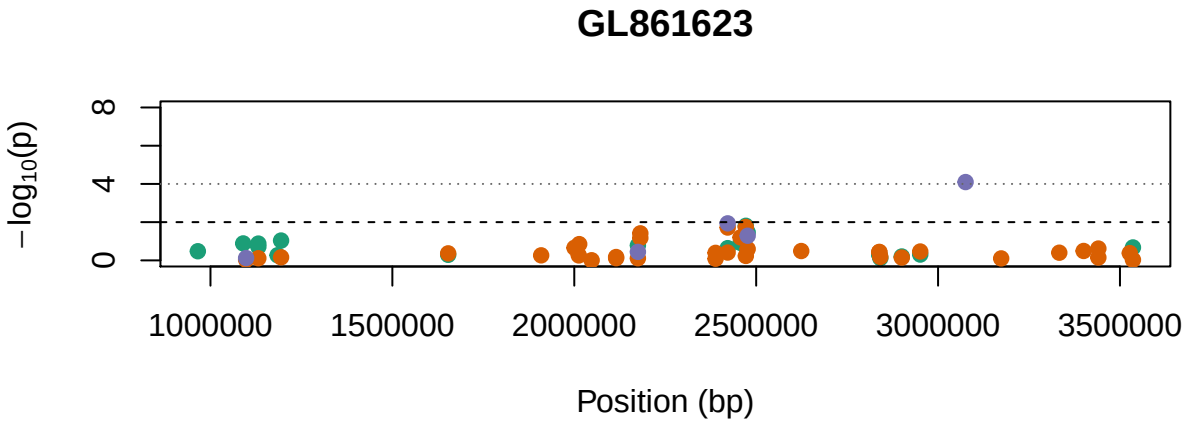

GL849681
−log10(p)
0
4
8
1000000
1500000
2000000
Position (bp)
GL849790
−log10(p)
0
4
8
500000
1000000
1500000
2000000
2500000
Position (bp)
GL849860
−log10(p)
0
4
8
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Position (bp)

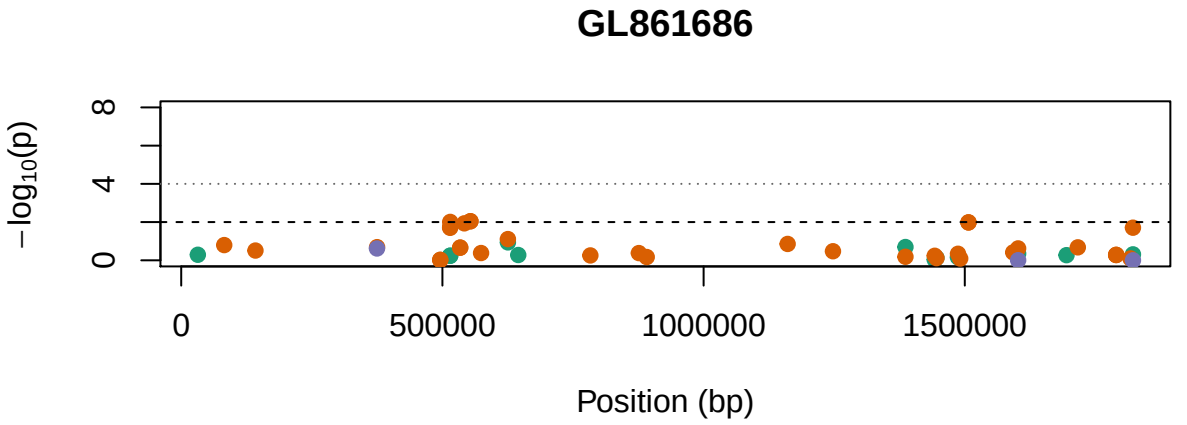

GL850047
−log10(p)
0
4
8
2e+04
4e+04
6e+04
8e+04
1e+05
Position (bp)
GL856776
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
2500000
Position (bp)
GL856785
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
4e+06
Position (bp)

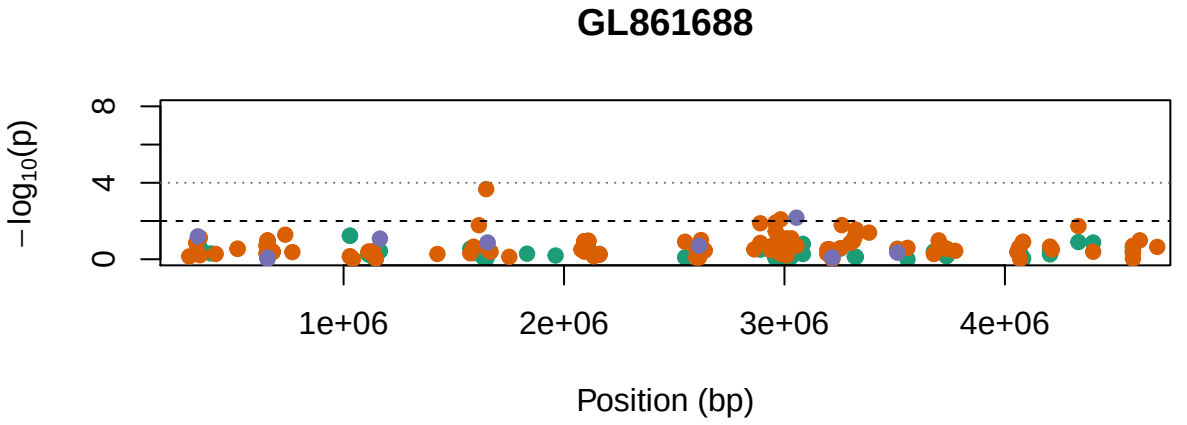

GL856833
−log10(p)
0
4
8
500000
1500000
2500000
3500000
Position (bp)
GL856846
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
Position (bp)
GL856873
−log10(p)
0
4
8
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Position (bp)

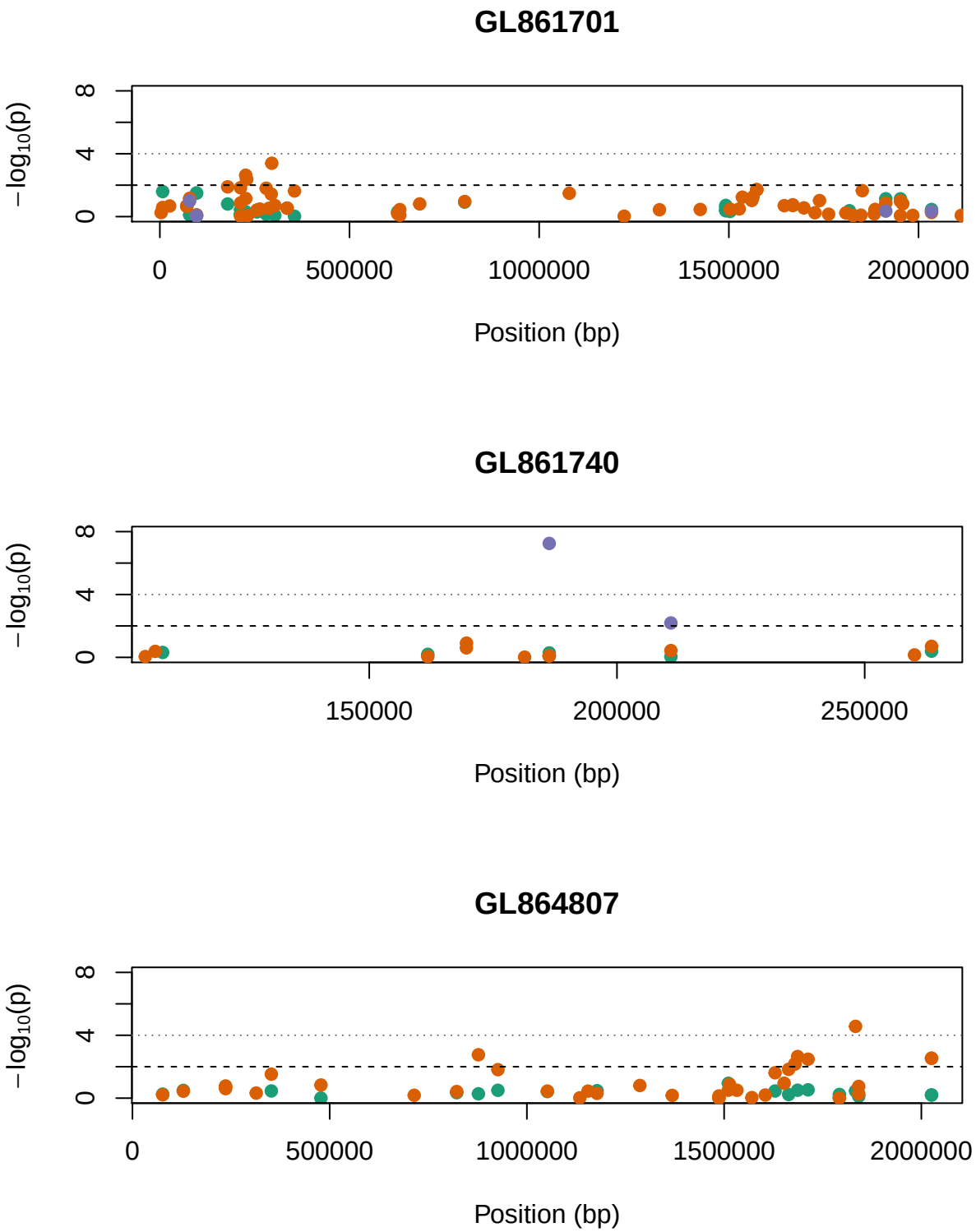

GL856919
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
4e+06
Position (bp)
GL856972
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
2500000
Position (bp)
GL856995
−log10(p)
0
4
8
0e+00
1e+06
2e+06
3e+06
Position (bp)


GL856999
−log10(p)
0
4
8
50000
100000
150000
200000
Position (bp)
GL857102
−log10(p)
0
4
8
15800
15900
16000
16100
16200
Position (bp)
GL861617
−log10(p)
0
4
8
500000
1000000
1500000
2000000
Position (bp)


GL861623
−log10(p)
0
4
8
1000000
1500000
2000000
2500000
3000000
3500000
Position (bp)


GL861686
−log10(p)
0
4
8
0
500000
1000000
1500000
Position (bp)


GL861688
−log10(p)
0
4
8
1e+06
2e+06
3e+06
4e+06
Position (bp)


GL861701
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
Position (bp)
GL861740
−log10(p)
0
4
8
150000
200000
250000
Position (bp)
GL864807
−log10(p)
0
4
8
0
500000
1000000
1500000
2000000
Position (bp)

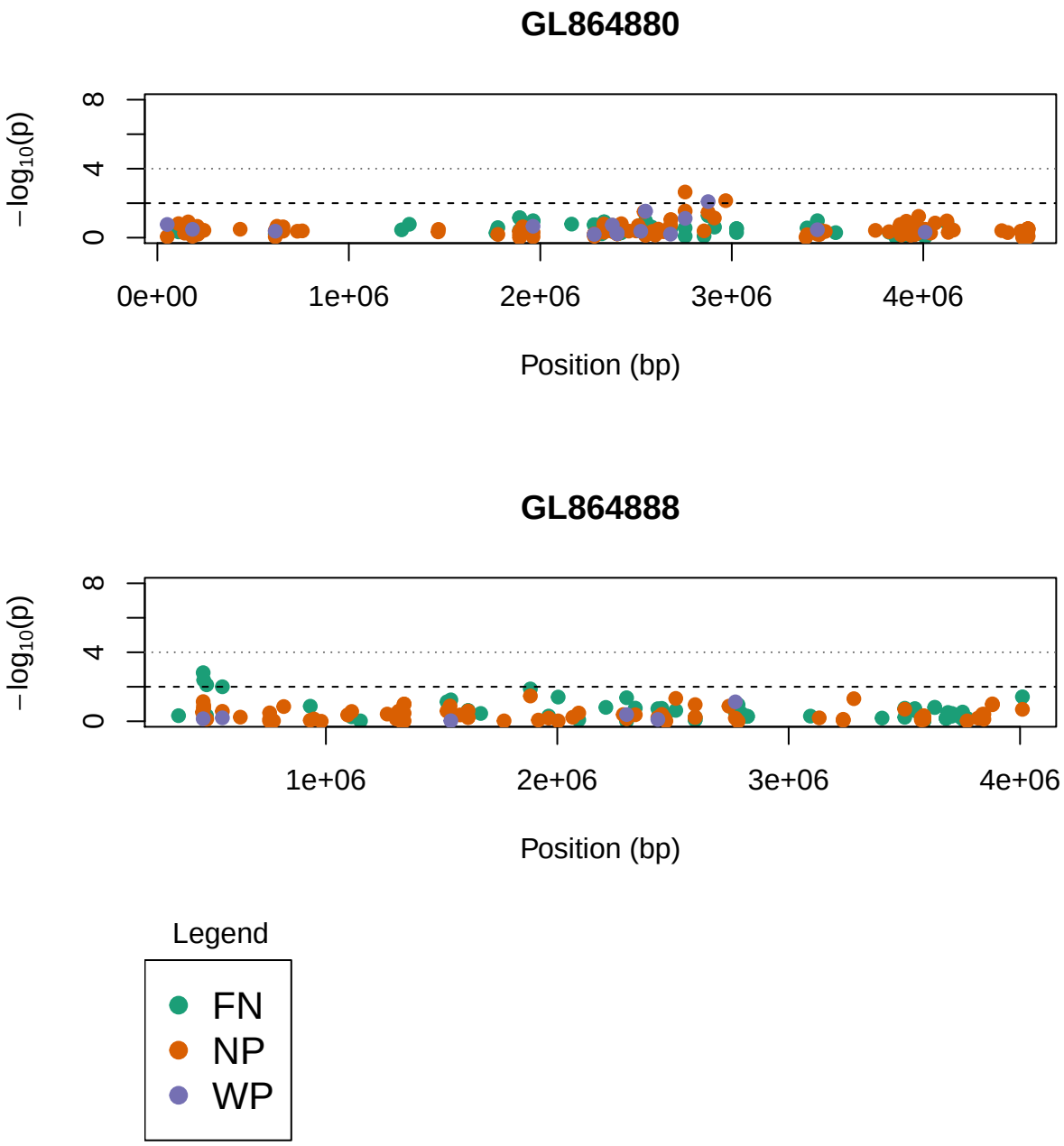



**Figure S1. Sixty signatures of selection were identified in 53 scaffolds in the Tasmanian devil genome at less than 100kb of a protein coding gene having a human orthologue.**

Filtered SNP were given a p-value (p) with *signasel* (https://github.com/hubert-pop/signasel), an R programme that is adapted to infer selection from genomic time-series (see Methods). To define a signature of selection, we required at least either one SNP with $p < 10^{-4}$ (dotted line) or two close SNP with $p < 10^{-2}$ (dashed line). Physical positions (in bp) are relative to the scaffold whose name is reported above the plot. Analyses were performed using the same samples, from three populations (see Table 1), as in Epstein et al. (2016a). Populations: FN = Freycinet ; NP = Narawntapu ; WP = West Pencil Pine

# Article III

## How could fully scaled carps appear in natural waters in Madagascar?

(Published 24 August 2016 in Proc. R. Soc. B., DOI: 10.1098/rspb.2016.0945)

Jean-Noël Hubert[1], François Allal[2], Caroline Hervet[1], Monique Ravakarivelo[3], Zsigmond Jeney[4], Alain Vergnet[5], René Guyomard[1] and Marc Vandeputte[1,5]

[1]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[2]Ifremer, UMR 9190 MARBEC, 34250 Palavas-les-Flots, France

[3]FOFIFA, DRZV, Antananarivo, Madagascar

[4]NAIK HAKI, 5540 Szarvas, Hungary

[5]Ifremer, 34250 Palavas-les-Flots, France

Keywords : evolution, mutation compensation, heritability

Author for correspondence : Marc Vandeputte (marc.vandeputte@jouy.inra.fr)

## Abstract

The capacity of organisms to rapidly evolve in response to environmental changes is a key feature of evolution, and studying mutation compensation is a way to evaluate whether alternative routes of evolution are possible or not. Common carps (*Cyprinus carpio*) carrying a homozygous loss-of-function mutation for the scale cover gene *fgfr1a1*, causing the 'mirror' reduced scale cover, were introduced in Madagascar a century ago. Here we show that carps in Malagasy natural waters are now predominantly covered with scales, though they still all carry the homozygous mutation. We also show that the number of scales in mutated carps is under strong polygenic genetic control, with a heritability of 0.49. As a whole, our results suggest that carps submitted to natural selection could evolve a wild-type-like scale cover in less than 40 generations from standing polygenic genetic variation, confirming similar findings mainly retrieved from model organisms.

## Introduction

Reversibility of evolution is a long studied and questioned aspect of evolutionary biology [1]. Especially in small populations, slightly deleterious mutations may accumulate and become fixed by genetic drift [2]. To which extent and by which mechanisms the phenotypic effects of these mutations could be counteracted by natural selection is important in conservation biology, where individuals from small populations are re-introduced in nature, but also has more general implications on the mechanisms of adaptation. Although in some cases reversion of mutations has been observed [3], the general picture is more that mutations themselves are irreversible. However, reverse adaptation by compensation of deleterious mutations has been demonstrated experimentally, essentially in microorganisms [4–7] and model invertebrate species [8,9]. The mechanisms invoked are diverse, but the appearance and/or selection of intermediate fitness compensatory mutations, not necessarily acting on the same biological pathways, seems more likely than reversion [5–7]. Many mutations remain cryptic (*i.e.* with little or no effect) in the absence of the deleterious mutation, and those cryptic mutations generate usable standing genetic variation that can be co-opted by natural selection, once revealed by the deleterious mutation [10]. Studies on the evolutionary basis of adaptation typically make use of organisms with rapid generation times and small physical size—in many cases, microbes in a controlled environment [4,5,8,11]. Still, studies of natural populations remain particularly attractive as they can show evolution in action on macroscopic, easily scored traits, in complex organisms such as vertebrates [12–14], providing an ecological point of comparison for the artificial set-ups used in laboratory studies [15]. The common carp, *Cyprinus carpio L.*, is a cyprinid fish species originating from the Eurasian continent, which has a long history of domestication [16], and has also been introduced in many areas throughout the world [17]. While wild-type common carp are typically exhibiting a full-scale cover, mutants in two independent bi-allelic Mendelian systems (S/s and N/n) have been selected during the domestication process, leading to four different scale patterns, the wild-type scaled phenotypes (genotypes SSnn or Ssnn), and reduced scale cover phenotypes identified as scattered or mirror (genotype ssnn), linear (genotypes SSNn or SsNn) and nude (genotype ssNn), all NN homozygotes being lethal [18]. The s allele of the S/s system has recently been shown to be a loss of function mutation in a kinase domain of *fgfr1a1*, which could be either a 310 bp deletion or a missense point mutation that encodes Lys-664 (AAA) instead of Glu-664 (GAA) in the *fgfr1a1* gene [19]. The N/n system has not been identified to date [20], although it shows some similarities with mutations in ectodysplasin *eda* and its receptor *edar*, found in the zebrafish *Danio rerio* [21] and in the

stickleback *Gasterosteus aculeatus* [22–24]. The N/n system may also be more complex than initially thought, and a gradation of phenotypes caused by dose-dependent signalling rather than a simple Mendelian bi-allelic system has been postulated [20]. Common carp of French origin was introduced to Madagascar in 1912 for fish farming purposes, and only the mirror phenotype was introduced, as the most valued for carp farming [25]. Between 1920 and 1950, carps quickly spread to most rivers and lakes, especially in the highlands and in the tropical lowlands of the Western coast of Madagascar [25,26]. At the end of the 1950s, field records showed many carps had 'degenerated' to a scaled phenotype [25], and a new introduction of mirror carps from France took place in 1959 to 'refresh' farmed stocks [27,28]. It is only in 1979 that carps with the wild-type scaled phenotype (strain Szarvas P33—genotype SSnn) were introduced in Madagascar, together with other mirror carps (strain Szarvas 215—genotype ssnn), from Hungary [29] (J Bakos 2016, personal communication). In this study, we investigated the scale cover phenotype and the S/s genotype at *fgfr1a1* of common carps collected in fish farms and in nature, in different areas of Madagascar. We also performed a controlled-breeding common garden experiment to investigate the segregation of scale cover phenotypes in the progeny of selected broodstock fish and estimate scale cover heritability.

## Results

### Occurrence of different scale patterns in feral and cultured carp populations

Of the 686 carps sampled from four river basins in Madagascar (406 from farms and 280 from the wild—see electronic supplementary material, figure S1), 439 (64.0%) had incomplete scale cover and 247 (36.0%) were fully covered with scales. In farms, full-scale cover was relatively uncommon (16.0%), while it clearly dominated (65.0%) in the wild. The situation in the wild varied greatly among sampling regions, with 93.2% scaled carps in lake Alaotra (n = 44), 89.0% in the Tsiribihina drainage basin (n = 136), 40.0% in the Betsiboka drainage basin (n = 25) and 13.3% in the Mangoky drainage basin (n = 75). In farms, there was little variation between regions, with 15.0 –17.5% scaled carps in the farms from the three drainage basins samples.

### Genetic and phenotypic diversity of Malagasy common carp for scale cover

Among the 236 individual samples sequenced for the S/s point mutation of *fgfr1a1*, surprisingly, the whole set of 103 feral individuals with a fully scaled phenotype were found homozygous s/s (table 1), showing that a full-scale cover can be achieved even though these fish carry a loss-of-function mutation, which normally implies the mirror (incomplete) phenotype. As expected, the 89 mirror carps analysed also showed this homozygous s/s genotype. Scaled carps sampled from fish farms predominantly (35/44) carried at least one S (wild-type) allele at the investigated locus.

**Table 1** Type of scale cover – scaled or mirror – determined for 236 carps sampled from different regions of Madagascar together with their genotype at the polymorphic loss-of-function point mutation site identified by Rohner *et al.* [19] within the *fgfr1a1* gene.

| | | | ***Ffgfr1a1* Genotype** | | |
|---|---|---|---|---|---|
| | | | S/S | S/s | s/s |
| **Phenotype** | Scaled | Farmed | 5 | 30 | 9 |
| | | Feral | 0 | 0 | 103 |
| | Mirror | Farmed | 0 | 0 | 32 |
| | | Feral | 0 | 0 | 57 |

The 236 Malagasy carps sequenced for *fgfr1a1* were also genotyped for nine microsatellite markers, together with 72 Hungarian carps representative of the strains introduced in 1979. Clustering analysis [37,38] showed that two clusters were present among these fish.

Fully scaled carps with mirror (s/s) genotype almost exclusively belonged to the first cluster, whereas regular scaled carps (S/s or S/S genotype) all belonged to the second cluster (figure 1a). All Hungarian carps, irrespective of their scale cover, belonged to the second cluster, which we then qualify as 'Hungarian farmed'. Together with the history of introductions, and the fact that fully scaled carps were observed in the wild before the second introduction from France in 1959, this suggests that feral carps (mostly fully scaled) derive from the first introduction of carps in 1912. We refer to this first cluster as the 'pioneer' cluster. Mirror carps could belong to any of both clusters. Most fish from the wild belonged to the pioneer cluster, with the notable exception of 22 fish caught from the Mandarano River, in the Mangoky drainage basin (figure 1b). The mean allelic richness was greater in Malagasy cultured populations ($A_{R,mean} = 7.8$) than in feral populations ($A_{R,mean} = 5.2$) suggesting a higher genetic diversity in cultured populations. Analysis of private alleles showed only three alleles specific to feral populations (all with a frequency of less than 5%) versus 26 (of which seven had a frequency of more than 5%) alleles which were present only in cultured populations. Twelve of those 26 private alleles (including five of the seven most frequent ones) were found in the Hungarian samples. This suggests that the greater allelic richness of cultured populations is partly owing to introgression of Hungarian genes, but that there is a low gene flow from cultured to feral populations. The population from Mandarano river that showed introgression from cluster 2 (figure 1) was also shown to be intermediate between Hungarian, Malagasy farmed and feral populations, suggesting a possible effect of restocking in this area (see electronic supplementary material, figure S2).

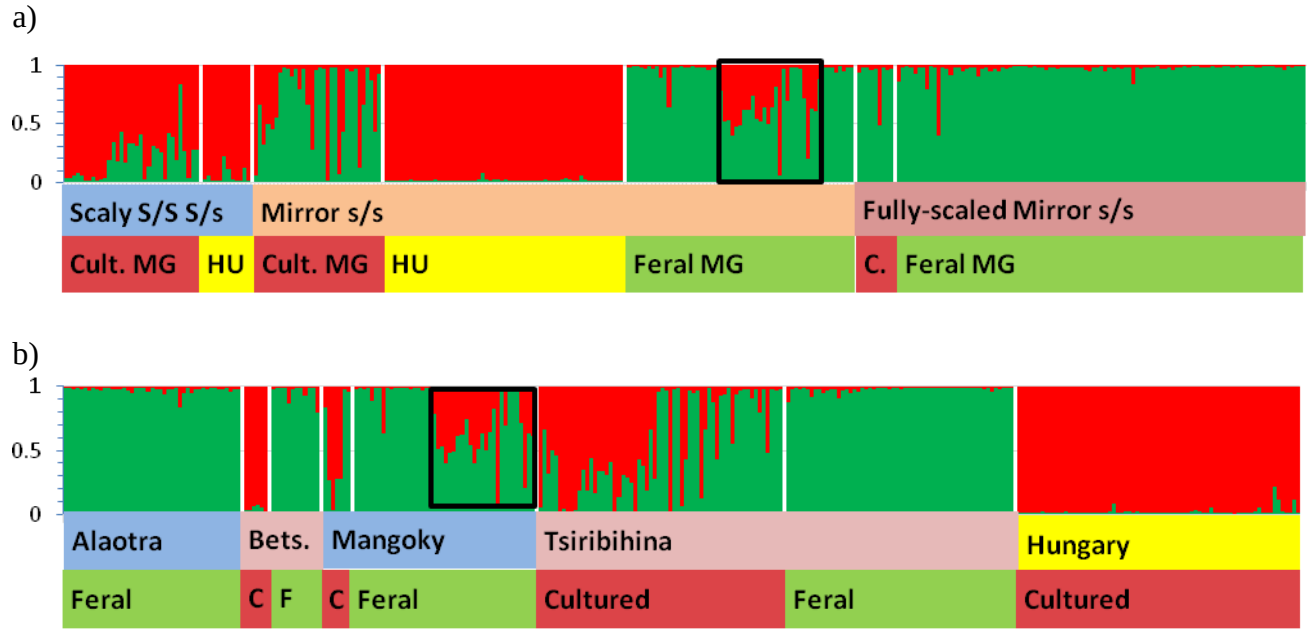
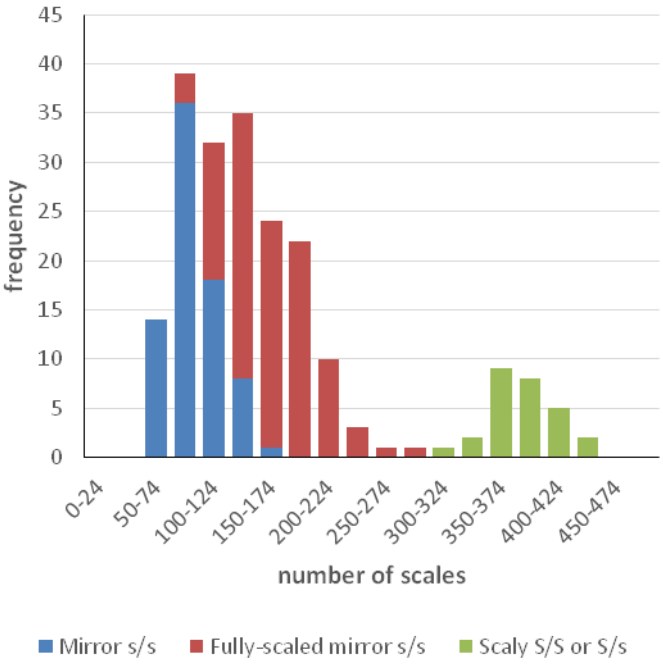



**Figure 1:** Population clusters inferred by STRUCTURE from genotypes at 11 microsatellite markers in 236 feral and cultured carps sampled in 2012 in Madagascar, as well as 72 samples from Hungary, ordered either a) according to their scale cover phenotype and their genotype for the loss of function mutation of *fgfr1a1*, or b) to the sampling region. Cluster 1 (pioneer) in green, cluster 2 (Hungarian farmed) in red. Samples from Mandarano river in a black frame.

The pictures of 209 out of the 236 carps enabled quantification of the number of scales on one flank. Scales were counted on 105 scaled carps that were homozygous for the loss-of-function mutation (s/s genotype), 76 carps exhibiting a mirror phenotype (also with s/s genotype) and 27 scaled carps carrying the functional S/s or S/S genotype. The number of scales had a bimodal distribution (figure 2). The lower mode corresponded to fish carrying an s/s (loss of function) genotype, whereas the upper mode corresponded to fish carrying at least one functional allele (S/S or S/s). The 27 S/S or S/s scaled carps had between 300 and 448 scales (383 on average), which was consistently greater than the scale number of the 105 homozygous s/s fully scaled carps (90–280 scales, 161 on average). This highlights that the mechanism leading to full-scale cover is not the same in S-carrying carps and s/s homozygotes. The 181 s/s carps, including mirror and fully scaled ones, displayed a wide range of scale covers, with a number of scales varying from 55 to 280 scales. This shows that both for carps carrying one functional *fgfr1a1* allele and for carps homozygous for the loss-of-function mutation, the number of scales presents significant phenotypic variation (figure 2).



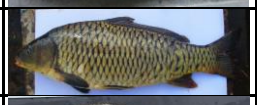

| Pheno | Geno | N scales | |
|---|---|---|---|
| M | s/s | 65 |  |
| M | s/s | 82 |  |
| FSM | s/s | 118 |  |
| FSM | s/s | 184 |  |
| FSM | s/s | 263 |  |
| S | S/s | 393 |  |

**Figure 2:** Left panel: Distribution of the number of scales quantified on one flank for 208 carps which were genotyped for the loss-of-function mutation in *fgfr1a1*. Colors indicate the type of scale cover followed by *fgfr1a1* genotype [blue: mirror (s/s), red: fully-scaled mirror (s/s), green: scaled (S/S or S/s)]. Right panel: Phenotype (Mirror; Fully-Scaled Mirror; Scaled), genotype for *fgfr1a1* for 6 carps representative of the distribution of scale numbers.

**Estimation of the heritability of scale cover in a controlled experiment**

Scales were counted on 196 offspring from a factorial crossing of four s/s males (two fully scaled, two with mirror phenotype) with four s/s females (all mirror). The scale counts of the males ranged from 100 to 206, and those of the females from 87 to 123. The offspring had 96.9 scales on average (range 34–197). The effect of the male parent on scale number was highly significant ($F_{3,178} = 12.37$, $p < 0.0001$), that of the female parent was also significant ($F_{3,178} = 3.09$, $p = 0.03$) while their interaction was not ($F_{9,178} = 1.77$, $p = 0.08$). The distributions of scale counts in the offspring of the four males largely overlapped with each other (electronic supplementary material, figure S3). Only male FSM2, which had the highest scale number (n = 206), produced some fully scaled offspring (n = 12 out of 68, having between 122 and 197 scales). This shows that that there is no secondary dominant locus substituting *fgfr1a1* in fully scaled s/s fish. The fully scaled offspring of male FSM2 were observed in its crosses with females M2 (5/15), M3 (5/15) and M4 (2/28), but not M1 (0/10). If full-scale cover in s/s fish was governed by a recessive compensatory mutation, this would mean that females M2–M4

carry it and that male FSM2 is homozygous. This would lead to an expectation of 50% fully scaled offspring in these crosses, but this was not observed ($\chi 2 = 22.3$, 1 d.f., $p < 10^{-5}$). This would also imply the presence of fully scaled fish in the offspring of male FSM1, and this was also not observed. Using the same data, heritability of scale number was estimated to be 0.49 ± 0.16, showing that this trait is under strong quantitative genetic control in s/s individuals. The distribution of the number of scales could be considered normal in the offspring of two dams and three sires (Shapiro–Wilk test, $p > 0.05$), whereas it significantly departed from normality for two (mirror) dams (Dam M2, $p < 0.01$; Dam M4, $p < 0.04$) and one fully scaled sire (Sire FSM2, $p < 0.02$). All distributions were however unimodal (Dip test [48], $p > 0.70$—see electronic supplementary material, figure S3), so that there was no apparent segregation of a major gene for scale cover in the s/s progenies tested, contrary to what was observed when s/s and S-carrying fish were compared (figure 2). In the light of these results, we conclude that pre-existing polygenic variation was selected and compensated for the absence of any functional version of the major gene *fgfr1a1* in Malagasy feral carp populations.

**Effect of scale cover on survival**

The survival from the larval stage to 109 days post-fertilization was 51.8% in a pond where offspring from S/s males mated with s/s females were stocked. Out of the 312 offspring assigned to their parents, 169 were scaled and 143 mirror. These numbers did not significantly depart from the 1:1 ratio expected under equal survival and Mendelian segregation of S ($\chi 2 = 2.17$, 1 d.f., $p > 0.14$). When differential survival was evaluated until the age of 11 months in six rice fields and four ponds, the average survival was 84.8% in ponds and 28.0% in rice fields, but in all rearing units except one rice field where scaled carp survival was higher ($\chi 2 = 3.846$, 1 d.f., $p = 0.049$), survival was similar between scaled and mirror carps ($p > 0.6$).

**Pace of evolution**

We calculated the rate of evolution of scale cover considering that scale counts in contemporary mirror carps (mean = 94.6, s.d. = 22.7) were representative of the carps introduced in 1912. The present scale count in fully scaled mirror carps (mean = 161.0 s.d. = 36.9) was supposed to be reached in 100 years (1912–2012; 40 generations), giving estimates of 5.3 kilodarwins or 0.056 haldanes, which is in the 3% highest rates of genetic change [49], implying a strong natural selection intensity.

## Discussion

In this study, we examined the present phenotypic and genotypic status of common carp in Madagascar, following an introduction in 1912 of fish homozygous for a loss-of-function mutation in *fgfr1a1* causing a reduction in scale cover, *i.e.* the mirror phenotype.

We showed that feral carps were predominantly fully scaled, whereas farmed ones were mostly mirror, although some fish with full-scale cover could also be found in farms. However, in farms, most scaled fish carried a functional S allele for *fgfr1a1*, which presumably originated from Hungarian carps introduced in 1979, or from later introductions. Most of the feral fish belonged to the same 'pioneer' cluster, and some fish from this cluster were also present in farmed populations. The reappearance of the wild-type-like scaled phenotype was first described in 1958 [25]. At that time, it was mostly recorded in the lowlands of the west (more than 95% scaled carps) while in lake Alaotra the percentage of scaled carps was 85%, and only 70% in the highlands [27]. This trend was also found in this study, with the exception of the Mangoky drainage basin where the proportion of scaled carps was only 13.3%. The first remarkable finding of the present study is that all scaled carps from the wild are still homozygous for the loss-of-function mutation of *fgfr1a1*, like their mirror ancestors. Hence, despite the mutation, they were able to evolve a compensatory mechanism to produce a full-scale cover. This also indirectly confirms that the carps introduced in 1912 (a time where the genetic basis of the mirror phenotype was unknown) were true mirrors, homozygous for the loss-of-function allele.

We could not show any effect of scale cover on survival in the environment tested, but previous data show that scaled carps have a higher survival than mirror carps [18], *e.g.* with a higher resistance to parasitic infections [50]. An interesting parallel example is that of the USA, where both mirror (n = 227) and scaled (n = 118) carps were introduced in 1887, and where scaled carps now represent more than 98% of the wild individuals [51]. In the USA case however, the scaled fish introduced seemingly possessed the functional *fgfr1a1* allele, as first-generation offspring of the introduced broodstock were nearly all fully scaled [51]. In the case of Madagascar, regular scaled (SSnn) fish were introduced in 1979, but we did not find them in the wild. Even though escapes from farms cannot be avoided, and restocking operations with farmed fish are common, the lack of any wild SSnn scaled carp here suggests that they do not have a decisive competitive advantage with the ssnn 'neoscaled' fish. Interestingly, irrespective of scale cover, fish from the 'Hungarian farmed' cluster were not found in the wild, except in the Mangoky drainage basin, which is specific in the fact that scaled fish are rare (13.3%), even among fish from the 'pioneer' cluster. Taken together, this indicates that natural selection

against incomplete scale cover must be high in most (maybe not all) natural environments, precluding gene flow from farms to the wild. The nature of the selecting agent(s), however, remains unknown. The inferred evolutionary rate for scale number of 5.3 kilodarwins or 0.056 haldanes is high [47,49], but subject to caution for two reasons: first, we do not know how many scales the mirror fish introduced in 1912 carried, and the possible presence of fully scaled mirror fish in this initial stock (or further undocumented introductions) cannot be excluded with certainty, although fully scaled mirror carps are not normally seen in European farmed carp populations. Second, the present predominance of mirror fish in farms, and their relatively low average scale number may also be the result of introductions of new mirror fish in 1959, 1979 or later. Additionally, negative selection for scale number by fish farmers is likely to happen, as scales are seen as a complication for cooking. Conversely, as a majority of scaled fish was already present in the wild at the end of the 1950s, the time for selection to operate was probably shorter than what we estimated.

Although we cannot totally exclude an input from environmental effects on the observed scale cover patterns, our results clearly support a polygenic control of the number of scales in Malagasy feral carps. First, phenotypic plasticity for scale number would not lead to the observed very low level of gene flow from farms to the wild, this being especially true in zones where fully scaled feral carps are frequent or dominate (all except the Mangoky drainage basin). In addition, our controlled breeding assay showed a significant heritable component of scale number in s/s fish ($h^2 = 0.49$), and the absence of a simple Mendelian system underlying this variation. The parental scale numbers were shown to significantly influence the phenotypes in the F1 individuals, which also displayed a large variance in scale number. For example, the progeny from male M1 (100 scales) exhibited from 34 to 143 scales, whereas from 63 to 197 scales were identified in the progeny of male FSM2 (206 scales). Such a pattern suggests that the number of scales is likely to be affected by several genes, as suggested earlier [19,20], not precluding the possibility that this variation may be governed by a few major QTLs as seen in lateral plate number variation in the threespine stickleback [12]. Indeed, though the Mendelian S/s system still explains most of the variation in farmed populations (table 1), polygenic (or at least oligogenic) genetic determinism would fit well with the recent hypothesis that some variation in scale phenotypes of common carp or zebrafish would be caused by dose-dependent signalling [20,21]. Overall, the most likely explanation for the rapid evolution of scale cover in Madagascar feral carps is natural selection of pre-existing polygenic variation, uncovered by the homozygous s/s genotype of *fgfr1a1* in the carps introduced in 1912. Genetic variation for

scale cover may exist in wild-type (S-carrying) fully scaled carp (for which we showed at least phenotypic variation in scale numbers—figure 2), but have little effect on fitness in such fully scaled carps—and in this sense remain cryptic and unselected [10].

Studies performed on a range of laboratory models have addressed the topic of compensatory evolution in response to gene loss. Mutated populations were shown to evolve towards better fitness without necessarily resorting to molecular convergence, even when the wild-like phenotype was restored. Instead, compensatory mechanisms predominantly relied on alternative pathways involving several genes [6,7] and resulted in strong fitness benefits that could arise very rapidly [9]. Similarities with such experimental settings and outcomes can be found in the history of the common carp in Madagascar. The initial introduction of a population that remained shielded from migration at least during the first generations, and that was fixed for a mutation affecting fitness, resulted in a rapid adaptation to Malagasy waters. Under the adaptive landscape concept, the pioneer mirror carp population was possibly located in a fitness valley and subsequent generations climbed an alternative route through standing genetic variation to gain fitness. Finally, this serendipitous experiment on carp illustrates in nature the observations previously reported from laboratory evolution and highlights the potential of 'unplanned natural' experiments to help better understand rapid evolution in an ecological context [15].

This work brings evidence for a rescue of the wild-type-like scale cover through the likely selection of polygenes from standing genetic variation. This provides a visible and striking example that evolutionary convergence (*i.e.* to wild-type-like scale cover) can use other routes than reversion mutation, and suggests that natural populations can host enough capacity for adaptation on the short-term to face a sudden environmental change, even if a harmful mutation was formerly fixed.

## Materials and methods

### Fish sampling

A total of 686 carps were sampled from eight regions in Madagascar (406 from farms and 280 from the wild—see electronic supplementary material, figure S1). For each fish sampled, origin (farmed/wild) and location were recorded, a digital picture was taken, and a fin sample was collected in 90% ethanol for further DNA extraction and genotyping. Phenotyping of scale cover was assessed on each picture first using a binary score: scaled when the whole body was covered with scales, irrespective of their number, and incomplete (mirror) when only part of the body was covered with scales. Scaled carps may in reality encompass different phenotypes, the real wild-type scaly carp with more than 300 scales on one side, when fish carry at least one functional S allele, and a 'fully scaled mirror' type with less than 300 scales when fish are s/s homozygotes.

### *Fgfr1a1* genotyping

Both regions from the kinase domain of *fgfr1a1* containing the two mutations—the EK substitution and the 310-bp deletion—previously reported as associated with a mirror phenotype in carp were investigated using the primer sets designed by Rohner et al. [19]. DNA was extracted with Wizard® Genomic DNA purification kit (Promega). PCRs were done for each sample in four wells with 2 µl of DNA solution (20 ng.µl$^{-1}$), 3.63 µl of water, 2 µl of 5X GoTaq Flexi buffer (Promega), 0.6 µl of $MgCl_2$ (25 mM), 0.7 µl of dNTP (1 mM), 0.5 µl of each primer (10 µM) and 0.07 µl of GoTaq polymerase (5 U.µl$^{-1}$, Promega). Thermocycling consisted of 5 min at 96°C, then five initial cycles of 30 s at 96°C (denaturation), 30 s at 58°C (annealing) and 1 min at 72°C (extension), followed by 25 cycles of 30 s at 96°C, 30 s at 58°C and 30 s at 72°C, and a final period of 5 min at 72°C. PCR products from 236 Malagasy carp samples were sent to Eurofins MWG (Germany) in order to perform a sequencing assay of the genomic region including the EK mutation at the first base of codon 664 in *fgfr1a1*. A subset of 45 samples was also explored for the presence of the 310-bp deletion that shortens intron 10 and exon 11 in some mirror carps. DNA was visualized under UV light after migration at 130 V for 45 min in a 2% agarose ethidium bromide-stained gel. The whole amplicons had the expected size of around 550 bp, indicating the absence of deletion. In addition, three samples from one mirror and two fully scaled mirror feral carps were submitted to sequencing and confirmed the absence of any polymorphism for the deletion allele. All sequences were analysed using NovoSNP software [30].

**Genetic structure of Malagasy carp populations**

The 236 carps sequenced for *fgfr1a1*, as well a sample of 72 carps from Hungary, which were a contemporary sample of the populations introduced in 1979, were genotyped for 11 microsatellite markers HLJE265, HLJ2241, HLJ2346, HLJ2382, HLJ2465, HLJ2544, HLJ334, HLJ526, HLJ534 [31], J58 [32] and MFW16 [33] by Labogena (Jouy-en-Josas, France). To account for the validity of the set of microsatellite markers, FIS [34] departure from Hardy–Weinberg equilibrium was tested through allele randomizations (10000 permutations per test) using Fstat [35] within the five population samples with $n \geq 30$. Two of them (HLJ2346 and HLJ534) significantly departed from the Hardy–Weinberg equilibrium in at least one population after Bonferroni correction for multiple testing [36] (electronic supplementary material, table S1), and were excluded from further analyses. Microsatellite genotypes were used for a clustering analysis with STRUCTURE [37], with an admixture model (default setting), correlated allele frequencies (default setting), 20000 burn-in repetitions and 20000 repetitions after burn-in. The most likely number of clusters K was assessed with the deltaK method [38], testing values of K ranging from one to five with 20 replicate simulations for each level of K. The allelic richness was computed with Fstat [35], and the number of private alleles was estimated in the farmed and feral samples from Madagascar and in the Hungarian farmed sample. All allelic data including those of the Hungarian carps were used to produce an unrooted tree, using an unweighted neighbour joining (NJ) clustering method for a dissimilarity matrix calculated by the simple matching method [39] with 1000 bootstrap iterations implemented in DARwin6 [40]. Genotype data are available in electronic supplementary material, data S2.

**Controlled breeding experiments**

We also performed a controlled breeding experiment using carp broodstock collected in farms in the Vakinankaratra region (Tsiribihina drainage basin). Four males, two of which were fully scaled mirror (FSM1, FSM2) and two of which were standard mirror (M1, M2) were mated to four mirror females in a full-factorial mating design, and 10 heterozygous S/s scaled males were mated to the same four mirror (s/s) females. Female ovulation was induced with Ovopel (d-Ala6, Pro9-Net-mGnRH, Unic-trade, Hungary) homogenized using 1 pellet.ml$^{-1}$ in 0.9% NaCl solution [41], using a first injection of 0.1 ml solution per kg of fish, and a second injection of 0.9 ml solution per kg of fish 12 h later. Before any manipulation, the fish were anaesthetized with 2-phenoxyethanol (0.3 ml.l$^{-1}$). Spawning occurred 12 h after the second injection, and the

spawns of the four females were stripped by gentle abdominal pressure and mixed in equal volumes to produce a pool of eggs. The sperm of the four males had been collected 12 h in advance by stripping, and was stored at 4°C in 5 ml syringes (max. 1 ml sperm per syringe). Fifty grams of eggs from the pool were split into four equal parts of 12.5 g, each being gently mixed with 0.1 ml sperm from one male, and activated with 15 ml of activation solution (3 $g.l^{-1}$ urea, 4 $g.l^{-1}$ NaCl). The operation was repeated with 50 g of eggs from the pool that were split into 10 equal aliquots of 5 ml, each fertilized with one heterozygous S/s male. One minute after activation, fertilization batches were mixed by sire type (mirror and fully scaled mirror on the one side, heterozygous scaled on the other side) and manually agitated with a semi-skimmed milk:water solution (1:4) for 30 min to avoid egg sticking, after which they were rinsed with hatchery water and each egg group incubated in a McDonald jar at an average temperature of 24°C. Hatching occurred at 47 h post-fertilization, and larvae were transferred to a resorption tank with flow-through water. Two ponds were stocked with larvae, pond 1 (25 $m^2$) with 800 larvae from the 4×4 cross and pond 2 (100 $m^2$) with 1700 larvae, 425 from the 4×4 cross and 1275 from the 10×4 cross. At 109 days post-fertilization, the ponds were drained, 363 fish were collected in pond 1 and 881 in pond 2. They were first anaesthetized with 2-phenoxyethanol (0.3 $ml.l^{-1}$), then a subsample was collected in each pond for further characterization: 74 fish in pond 1 and 486 in pond 2 were individually photographed (Canon Powershot S50), and a piece of fin was collected in 90% ethanol for further DNA extraction and parentage reconstruction.

Offspring and parents were genotyped for 14 microsatellite loci CCE46 [42], HLJE265, HLJ2241, HLJ2346, HLJ2382, HLJ2465, HLJ2544, HLJ526, HLJ534, HLJ334 [31], J58 [32], KOI 57–58 [43], MFW16 and MFW40 [33] by Labogena (Jouy-en-Josas, France). Parentage was assessed by exclusion with VITASSIGN [44], allowing for up to two allelic mismatches. Five hundred and thirteen offspring out of 560 (91.6%) were assigned to a unique parental pair.

Numbers of scaled and mirror fish were counted in the offspring of the 10 heterozygous S/s males, and departure from 1:1 tested with a $\chi 2$ test (data available in electronic supplementary material, data S4). Six rice fields and four ponds were further stocked each with 50 scaled fish and 50 mirror fish at 110 days post-fertilization, and survivors were counted and classified for scale cover at 11 months post-fertilization. Departure from 1:1 was tested with a $\chi 2$ test.

**Acquisition and interpretation of scale numbers**

Scales were counted on one flank on the digital pictures of 208 adult fish (feral and farmed) out of 236 that were genotyped and of 196 of the 197 offspring from the 4×4 controlled breeding experiment that were unambiguously assigned to their parents. Individuals for which scales were not counted were removed due to low picture quality. This was done with ImageJ [51] using the Cell Counter plugin (http://rsbweb.nih.gov/ij/plugins/cell-counter.html).
In the controlled breeding experiment, effects of sire, dam and their interaction were tested with the following mixed model in SAS:

$$Y_{ijkl} = \mu + P_i + s_j + d_k + sd_{jk} + e_{ijkl}$$

With $Y_{ijkl}$ the number of scales in offspring $l$, $\mu$ the overall mean, $P_i$ the fixed effect of pond $i$, $s_j$ the random effect of sire $j$, $d_k$ the random effect of dam $k$, $sd_{jk}$ the random interaction term between sire $j$ and dam $k$ and $e_{ijkl}$ the random residual. The same data were also used to estimate the heritability of scale cover using an animal model, with pond as a fixed effect, using VCE6 [52].

Rates of evolution in KDarwins and haldanes were computed as proposed by Kinnison and Hendry [36]:

$$kd = \frac{\ln(x_2) - \ln(x_1)}{t}$$

Where $kd$ is the rate of evolution in kilodarwins, $\ln(x_2)$ is the natural logarithm of the average number of scales in fully-scaled mirror carps, $\ln(x_1)$ is the natural logarithm of the average number of scales in mirror carps (taken as a surrogate for the initial number of scales in the mirror carps introduced in 1912), and $t$ is the time interval in thousand years (here 0.1, or 100 years between 1912 and 2012)

$$h = \frac{x_2/s_p - x_1/s_p}{g}$$

Where $h$ is the rate of evolution in haldanes $x_1$ and $x_2$ are as before, $s_p$ is the pooled standard deviation of scale number across mirror and fully-scales mirror groups, and $g$ is the number of generations (here 40, taking a mean generation interval of 2.5 years).

## Acknowledgements

The authors thank Harena Rasamoelina, Fabien Cousseau, Marc Oswald and the APDRA team in Madagascar for help in sample collection, Bertrand Pajon, Diana Andria-Mananjara, Rojomalala Razafimandimby for following field experiments, CRFPA Antanetimboahanghy for logistical support, Alex Vasilescu at Labogena for genotyping, Andras Woynarowitch and Janos Bakos for documenting carp introduction from Hungary, Bruno Guinand for comments on an earlier version of the manuscript. The project was partly financed by the French Ministry of Foreign affairs (project PARRUR Madapisci) and the French Development Agency (AFD).

## Authors contributions

M.V. and R.G. conceived, designed and supervised the study. R.G., M.R., Z.J. coordinated sample collection. J.N.H., C.H., A.V., M.R. performed experiments. J.N.H, F.A., M.R., M.V. analysed data and drafted the manuscript. All authors critically reviewed the manuscript and approved the final version. M.V. has access to all data and is responsible for the scientific integrity of this work.

## References

1. Teotónio H, Rose MR. 2001 Perspective: reverse evolution. Evolution (N.Y.) 55, 653. (doi:10.1554/0014-3820(2001)055[0653:PRE]2.0.CO;2)

2. Lande R. 1994 Risk of population extinction from fixation of new deleterious mutations. Evolution (N.Y.) 48, 1460–1469. (doi:10.2307/2410240)

3. Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997 Exceptional convergent evolution in a virus. Genetics 147, 1497–1507.

4. Moore FB-G, Rozen DE, Lenski RE. 2000 Pervasive compensatory adaptation in Escherichia coli. Proc. R. Soc. Lond. B 267, 515–522. (doi:10.1098/rspb.2000.1030)

5. Levin BR, Perrot V, Walker N. 2000 Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. Genetics 154, 985–997.

6. Harcombe WR, Springman R, Bull JJ. 2009 Compensatory evolution for a gene deletion is not limited to its immediate functional network. BMC Evol. Biol. 9, 106. (doi:10.1186/1471-2148-9-106)

7. Szamecz B et al. 2014 The genomic landscape of compensatory evolution. PLoS Biol. 12, e1001935. (doi:10.1371/journal.pbio.1001935)

8. Estes S, Lynch M. 2003 Rapid fitness recovery in mutationally degraded lines of Caenorhabditis elegans. Evolution 57, 1022–1030. (doi:10.1111/j.0014-3820.2003.tb00313.x)

9. Chandler CH, Chadderdon GE, Phillips PC, Dworkin I, Janzen FJ. 2012 Experimental evolution of the Caenorahabditis elegans sex determination pathway. Evolution (N.Y.) 66, 82–93.

10. Rajon E, Masel J. 2013 Compensatory evolution and the origins of innovations. Genetics 193, 1209–1220. (doi:10.1534/genetics.112.148627)

11. Wiser MJ, Ribeck N, Lenski RE. 2013 Long-term dynamics of adaptation in asexual populations. Science 342, 1364–1367. (doi:10.1126/science.1243357)

12. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM. 2004 The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. 2,635–641. (doi:10.1371/journal.pbio.0020109)

13. Reznick DN, Shaw F, Rodd FH, Shaw RG. 1997 Evaluation of the rate of evolution in natural populations of guppies (Poecilia reticulata). Science 275, 1934–1937. (doi:10.1126/science.275.5308.1934)

14. Steiner CC, Weber JN, Hoekstra HE. 2007 Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol. 5, e219. (doi:10.1371/journal.pbio.0050219)

15. Irschick DJ, Reznick D. 2009 Field experiments, introductions, and experimental evolution. In Experimental evolution: concepts, methods, and applications of selection experiments (eds T Jr Garland, MR Rose), pp. 173–194. Berkeley, CA: University of California Press.

16. Balon EK. 2004 About the oldest domesticates among fishes. J. Fish Biol. 65, 1–27. (doi:10.1111/j.1095-8649.2004.00563.x)

17. Casal C. 2006 Global documentation of fish introductions: the growing crisis and recommendations for action. Biol. Invasions 8,3–11. (doi:10.1007/s10530-005-0231-3)

18. Kirpichnikov VS. 1999 Genetics and breeding of common carp. Paris, France: INRA Editions.

19. Rohner N, Bercse´nyi M, Orba´n L, Kolanczyk ME, Linke D, Brand M, Nu¨sslein-Volhard C, Harris MP. 2009 Duplication of fgfr1 permits Fgf signaling to serve as a target for selection during domestication. Curr. Biol. 19, 1642–1647. (doi:10.1016/j.cub.2009.07.065)

20. Casas L, Szűcs R, Vij S, Goh CH, Kathiresan P, Németh S, Jeney Z, Bercsényi M, Orbán L. 2013 Disappearing scales in carps: re-visiting Kirpichnikov's model on the genetics of scale pattern formation. PLoS ONE 8, e83327. (doi:10.1371/journal.pone.0083327)

21. Harris MP, Rohner N, Schwarz H, Perathoner S, Konstantinidis P, Nu¨sslein-Volhard C. 2008 Zebrafish eda and edar mutants reveal conserved and ancestral roles of ectodysplasin signaling invertebrates. PLoS Genet. 4, e1000206. (doi:10.1371/journal.pgen.1000206)

22. Colosimo PF et al. 2005 Widespread parallel evolution in sticklebacks by repeated fixation ofectodysplasin alleles. Science 307, 1928–1933. (doi:10.1126/science.1107239)

23. Knecht AK, Hosemann KE, Kingsley DM. 2007 Constraints on utilization of the EDA-signalingpathway in threespine stickleback evolution. Evol. Dev. 9, 141–154. (doi:10.1111/j.1525-142X.2007.00145.x)

24. O'Brown NM, Summers BR, Jones FC, Brady SD, Kingsley DM. 2015 A recurrent regulatory change underlying altered expression and Wnt response of the stickleback armor plates gene EDA. Elife 2015,1–17. (doi:10.7554/eLife.05290)

25. Kiener A. 1958 Intérêt et perspectives de la pisciculture de la carpe à Madagascar. Bull. Madagascar 8,693–702.26. Lemasson L. 1957 Réflexions sur la pêche et la pisciculture à Madagascar. Bois Forêts des Trop. 52,57–61.

27. Kiener A. 1963 Poissons, pêche et pisciculture à Madagascar. Nogent sur Marne, France: Editions du CTFT.

28. Moreau J, Arrignon J, Jubb RA. 1988 Les introductions d'espèces étrangères dans les eaux continentales africaines: intérêt et limites. In Biologie et écologie des poissons d'eau douce africains (eds C Lévèque, MN Bruton), pp. 329–425. Paris, France: ORSTOM.

29. Láng WM. 1980 Transport of fingerlings to Africa. Halászat XXVI, 43.

30. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P. 2005 novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 15, 436–442. (doi:10.1101/gr.2754005)

31. Zheng X, Kuang Y, Zhang X, Lu C, Cao D, Li C, Sun X. 2011 A genetic linkage map and comparative genome analysis of common carp (Cyprinus carpio L.) using microsatellites and SNPs. Mol. Genet. Genomics 286,261–277. (doi:10.1007/s00438-011-0644-x)

32. Yue GH, Orban L. 2002 Polymorphic microsatellites from silver crucian carp (Carassius auratus gibelio Bloch) and cross-amplification in common carp (Cyprinus carpio L.). Mol. Ecol. Notes 2, 534–536. (doi:10.1046/j.1471-8286.2002.00308.x)

33. Crooijmans RPMA, Van der Poel JJ, Groenen MAM, Bierbooms VAF, Komen J. 1997 Microsatellitemarkers in common carp (Cyprinus carpio L.). Anim. Genet. 28, 129–134. (doi:10.1111/j.1365-2052.1997.00097.x)

34. Weir BS, Cockerham CC. 1984 Estimating F-statistics for the analysis of population structure. Evolution (N.Y). 38, 1358–1370. (doi:10.2307/2408641)

35. Goudet J. 1995 Fstat (version 1.2): a computer program to calculate F-statistics. J. Hered. 86,485–486. (doi:10.1093/jhered/esg066)

36. Rice WR. 1989 Analyzing tables of statistical tests. Evolution (N.Y). 43, 223–225. (doi:10.2307/2409177)

37. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

38. Evanno G, Regnaut S, Goudet J. 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611–2620. (doi:10.1111/j.1365-294X.2005.02553.x)

39. Sokal RR, Michener CD. 1958 A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38, 1409–1438.

40. Perrier X, Jacquemoud-Collet JP. 2006 DARwin software. See http://darwin.cirad.fr.

41. Horvath L, Szabo T, Burke J. 1997 Hatchery testing of GnRH analogue-containing pellets on ovulation in four cyprinid species. Pol. Arch. Hydrobiol. 44,221–226.

42. Wang D, Liao X, Cheng L, Yu X, Tong J. 2007 Development of novel EST-SSR markers in common carp by data mining from public EST sequences. Aquaculture 271, 558–574. (doi:10.1016/j.aquaculture.2007.06.001)

43. David L, Rajasekaran P, Fang J, Hillel J, Lavi U. 2001 Polymorphism in ornamental and common carp strains (Cyprinus carpio L.) as revealed by AFLP analysis and a new set of microsatellite markers. Mol. Genet. Genomics 266, 353–362. (doi:10.1007/s004380100569)

44. Vandeputte M, Mauger S, Dupont-Nivet M. 2006 An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Mol. Ecol. Notes 6, 265–267. (doi:10.1111/j.1471-8286.2005.01167.x)

45. Abràmoff MD, Magalhães PJ, Ram SJ. 2004 Image processing with ImageJ. Biophotonics Int. 11, 36–42. (doi:10.1117/1.3589100)

46. Groeneveld E, Kovac M, Mielenz N. 2008 VCE user's guide and reference manual version 6.0. Neustadt, Germany: Friedrich Loeffler Institute.

47. Kinnison M, Hendry A. 2001 The pace of modern life II: from rates of contemporary microevolution to pattern and process. Genetica 112–113, 145–164. (doi:10.1023/A:1013375419520)

48. Hartigan J, Hartigan P. 1985 The dip test of unimodality. Ann. Stat. 13, 70–84. (doi:10.1214/aos/1176346577)

49. Bell G. 2008 Selection: the mechanism of evolution, 2nd edn. Oxford, UK: Oxford University Press.

50. Price DJ, Clayton GM. 1999 Genotype environment interactions in the susceptibility of the common carp, Cyprinus carpio, to Ichthyophthirius multifiliis infections. Aquaculture 173, 149–160. (doi:10.1016/S0044-8486(98)00483-9)

51. Walden HT. 1964 Familiar freshwater fishes of America. New York, NY: Harper & Row.

# Exemples de scripts de soumission de tâches avec SGE ou SLURM

Voici un exemple de script minimaliste, nommé *task-0.sh*, permettant de lancer une tâche en mode *batch* (*i.e.,* les instructions de la tâche forment un lot traité de manière autonome) sur une plateforme HPC bénéficiant d'une interface SGE.

```
#$ -cwd
#$ -o output-file.out
#$ -e error-file.err
#$ -M utilisateur@domaine.fr
#$ -m abe
i=0
echo $i
```

Les lignes débutant par « *#$* » sont destinées au gestionnaire de soumission de tâches SGE. Dans l'exemple ci-dessus, nous précisons vouloir travailler dans le répertoire courant, créer un fichier journal des sorties nommé *output-file.out* et un autre *error-file.err* pour les erreurs, et notifier par mail à utilisateur@domaine.fr le lancement et la fin de la tâche, ainsi que son éventuel arrêt avant la fin en cas d'erreur. La suite des instructions correspond au corps du script, c'est-à-dire à l'ensemble des opérations qui constitueront la tâche (*e.g.,* lancement d'un script écrit dans un langage tiers, d'un programme installé sur la plateforme). Pour les besoins de l'exemple, nous avons utilisé l'affichage d'une variable *i* initialisée à 0. La soumission du script s'effectue avec la commande *qsub*.

```
qsub task-0.sh
```

Dans cet exemple, le lancement du script aboutira à la production d'un fichier de sortie affichant comme demandé le contenu de la variable *i* (*i.e.,* 0) et l'heure de la fin du traitement.

```
0
Epilog : job finished at xxx
```

Nous avons vu qu'il pouvait être intéressant d'explorer rapidement un espace de paramètres, par exemple pour tester par simulation les performances d'une méthode en génomique des populations en fonction de la variation de certains paramètres (*e.g.,* déséquilibre de liaison, effectif efficace, taille d'échantillon) ou pour comprendre l'impact de paramètres d'alignement de séquence sur le polymorphisme observé. C'est pour la gestion de ce genre de tâches où l'on souhaite pouvoir contrôler un grand ensemble de combinaisons, en faisant varier successivement plusieurs paramètres, que les

moyens de calcul et les possibilités de parallélisation des traitements offertes par les plateformes HPC sont intéressantes.

Voici un autre exemple de script, que nous nommerons *multiple-sub.sh*, permettant de lancer un grand nombre de tâches similaires au sens du paragraphe précédent (*i.e.,* identiques, à la valeur d'un paramètre près).

```
for i in `seq 1 10`
do
cp task-0.sh task-$i.sh
sed –i "s/0/$i/" task-$i.sh
qsub task-$i.sh
done
```

Ce script de soumission multiple permet de créer dix fichiers identiques au script de soumission unique *task-0.sh*, à l'exception des occurrences de 0 qui sont remplacées par le contenu de la variable *i*. La soumission du script de soumission multiple à l'aide de la commande *qsub* aboutira donc à la production de dix scripts, *task-1.sh*, *task-2.sh*, …, et *task-10.sh*, qui seront eux-mêmes automatiquement soumis à l'interface SGE.

```
qsub multiple-sub.sh
```

Comme attendu, le fichier de sortie affichera les dix valeurs successivement prises par *i* au cours de l'ensemble des opérations, ainsi que l'heure de la fin de chaque tâche effectuée sur la plateforme.

```
1
Epilog : job finished at xxx
2
Epilog : job finished at xxx
(...)
10
Epilog : job finished at xxx
```

L'utilisation de deux scripts, un script de soumission unique (*task-0.sh*) et un script de soumission multiple (*multiple-sub.sh*), permet donc de générer autant de tâches que l'utilisateur juge nécessaires, ne se distinguant les unes des autres que par la valeur d'une variable. Un résultat similaire pourrait être obtenu plus simplement en utilisant les options de la commande *qsub*, ce qui éviterait notamment la création de la dizaine de fichiers *task-i.sh*.

Les variables d'environnement permettent d'échanger les informations de façon plus efficiente entre les scripts. Par exemple, le script de soumission unique nommé *task-var.sh* utilise le contenu de la variable d'environnement *VARIABLE* (les noms de variables d'environnement sont par convention écrits en majuscules).

```
#$ -cwd
#$ -o output-file.out
#$ -e error-file.err
#$ -M utilisateur@domaine.fr
#$ -m abe
echo $VARIABLE
```

Un nouveau script de soumission multiple nommé *multiple-sub-2.sh* permet de contrôler la variation de la variable d'environnement *VARIABLE* et de passer cette information au script *task-var.sh* via l'option *–v* de la commande *qsub*.

```
for i in `seq 1 10`
do
qsub –v  "VARIABLE=$i" task-var.sh
done
```

L'utilisation de variables d'environnement permet d'atteindre le même objectif que précédemment, mais de façon plus économe, puisque l'écriture est plus simple et occasionne la création de seulement deux scripts *shell* quel que soit le nombre total de tâches à exécuter. Le script de soumission unique *task-var.sh* est utilisé à chaque itération parcourue dans le script de soumission multiple *multiple-sub-2.sh*.

Une autre solution du même type consiste à utiliser un tableau multitâche. La contrôle du script de soumission, nommé *task-t.sh*, se fait dans ce cas au moyen de la variable d'environnement *SGE_TASK_ID* et grâce à l'option *–t* de la commande *qsub*.

```
#$ -cwd
#$ -o output-file.out
#$ -e error-file.err
#$ -M utilisateur@domaine.fr
#$ -m abe
echo $SGE_TASK_ID
```

Il n'y a alors besoin que d'un seul script de soumission, celui-ci ne demandant qu'une seule ligne de *shell* pour être lui-même soumis.

```
qsub -t 1-10 task-t.sh
```

Le fichier de sortie *output-file.out* contient les mêmes informations que dans les exemples précédents.

Nos exemples montrent le fonctionnement de base la commande *qsub*, qui peut être optimisé pour répondre à des besoins spécifiques, notamment en ajustant les ressources (*e.g.,* temps d'exécution, mémoire, nombre de cœurs) aux besoins d'un projet.

Avec un gestionnaire de soumission SLURM, quelques adaptations syntaxiques sont nécessaires pour soumettre les tâches que nous avons données en exemple. En particulier, les directives destinées au gestionnaire de soumission sont précédées de la syntaxe « #SBATCH », et SLURM travaille par défaut dans le répertoire courant. Le script *shell* suivant, nommé *task-t-slurm.sh*, effectue les mêmes opérations que *task-t.sh*, mais dans le contexte de l'environnement SLURM.

```
#SBATCH -o output-file.out
#SBATCH -e error-file.err
#SBATCH --mail-user=utilisateur@domaine.fr
#SBATCH --mail-type=BEGIN,END,FAIL
echo $SLURM_ARRAY_TASK_ID
```

La soumission d'un script *shell* se fait alors au moyen de la commande *sbatch*.

```
sbatch -a 1-10 task-t-slurm.sh
```

Dans l'exemple ci-dessus, l'option *–a* est l'équivalent SLURM de l'option *–t* de SGE permettant de traiter un tableau de tâches. Plusieurs ressources sont disponibles sur Internet pour optimiser ses scripts de soumission de tâches et trouver les correspondances entre les syntaxes SGE et SLURM (*e.g.,* http://bioinfo.genotoul.fr/index.php/faq/job_submission_faq/ ou https://srcc.stanford.edu/sge-slurm-conversion).

*N.B.* Outre l'ouverture d'un compte auprès des gestionnaires et l'écriture d'un couple de scripts *shell* de soumission de tâches, l'utilisation d'une plateforme HPC nécessite en pratique l'installation de deux logiciels sur l'ordinateur de l'utilisateur : un logiciel client FTP permettant le transfert de fichiers entre l'ordinateur et la plateforme (*e.g.,* FileZilla), et un logiciel client SSH pour se connecter à la plateforme (*e.g.,* PuTTY). FileZilla et PuTTY présentent l'avantage d'être des logiciels libres et dotés d'une interface pratique sous Windows.

université PARIS-SACLAY | ÉCOLE DOCTORALE Agriculture, alimentation, biologie, environnement, santé (ABIES)

**Titre :** Génomique des populations appliquée : détection de signatures de sélection au sein de populations expérimentales.

**Mots clés :** Signatures de sélection, séries génétiques temporelles, expériences de sélection, tumeur faciale transmissible du diable de Tasmanie (DFTD), truite arc-en-ciel, RAD-sequencing (RAD-seq).

**Résumé :** La génomique des populations rend possible la mise en évidence de traces de sélection dans le génome. Les travaux effectués considèrent en général une échelle de temps longue (~ $10^3$ générations). En comparaison, peu d'intérêt a été porté aux études expérimentales de court terme (~ 10 générations). De telles expériences sont pourtant susceptibles de nous renseigner sur la base génétique de caractères complexes. Nous proposons une méthode de vraisemblance basée sur un modèle de Wright-Fisher pour détecter la sélection à partir d'échantillons génétiques temporels acquis sur une période de dix générations. Nous montrons par simulation que notre méthode permet de différencier les signaux dus à la combinaison de la sélection et de la dérive génétique de ceux dus à la dérive seule. Nous montrons également par simulation qu'il est possible d'estimer le coefficient de sélection appliqué à un locus testé. De plus, nous illustrons l'intérêt de notre méthode pour la détection de marqueurs candidats à la sélection au travers de deux études génomiques sur données réelles, chez le diable de Tasmanie (*Sarcophilus harrisii*) et chez la truite arc-en-ciel (*Oncorhynchus mykiss*). Ces applications mettent en évidence des régions génomiques candidates pour des phénotypes complexes dans des contextes différents. Dans l'ensemble, nos résultats montrent qu'il est possible de détecter des gènes sujets à une sélection directionnelle intense à partir d'échantillons génétiques temporels, même si la sélection est de courte durée et si les populations examinées ont un faible effectif.

**Title :** Applied population genomics : detection of signatures of selection in experimental populations.

**Keywords :** Signatures of selection, genetic time-series, selection experiments, Devil Facial Tumor Disease (DFTD), rainbow trout, RAD-sequencing (RAD-seq).

**Abstract :** Population genomics makes it possible to detect traces of selection in the genome. Studies in this field have mainly focused on long time scale (~ $10^3$ generations). In comparison, short-term experimental studies (~ 10 generations) have attracted much less interest. Such experiments are, however, likely to inform us about the genetic basis of complex characters. We propose a likelihood method based on a Wright-Fisher model to detect selection from genetic temporal samples collected over ten generations. We show through simulation that our method can disentangle signals due to the combination of genetic drift and selection to those due to drift alone. We also show through simulation that it is possible to estimate the selection coefficient applied to a tested locus. In addition, we illustrate the interest of our method for the detection of candidate markers for selection through two genome scans performed on real data, in the Tasmanian devil (*Sarcophilus harrisii*) and in the rainbow trout (*Oncorhynchus mykiss*). These practical applications highlight candidate genomic regions for complex phenotypes in different contexts. Collectively, our results show the possibility of detecting genes submitted to strong directional selection from genetic time-series, even if selection is applied on a short time period and if the examined populations are small.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France