



HAL
open science

Comment les espèces s'adaptent génétiquement à leur environnement ? Sélection positive, balancée, et microbiote intestinal

Laure Ségurel

► To cite this version:

Laure Ségurel. Comment les espèces s'adaptent génétiquement à leur environnement ? Sélection positive, balancée, et microbiote intestinal. Evolution [q-bio.PE]. Université Claude Bernard Lyon 1, 2025. <tel-05261586>

HAL Id: tel-05261586

<https://hal.science/tel-05261586v1>

Submitted on 15 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



Comment les espèces s'adaptent génétiquement à leur environnement ?

Sélection positive, balancée, et microbiote intestinal

Mémoire présenté pour l'obtention du diplôme
d'Habilitation à Diriger les Recherches

Université Claude Bernard Lyon 1

Ecole doctorale 341 - E2M2

Laure Ségurel

Chargée de recherche CNRS - Section 29 et CID 52

LBBE, UMR 5558, Lyon (depuis 2020)

Musée de l'Homme, UMR 7206, Paris (2013-2020)

Soutenue le 10/02/2025, devant le jury composé de :

Frédéric Delsuc (DR, ISEM, Montpellier)
Violaine Llaurens (DR, CIRB, Paris)
Franck Prugnolle (DR, REHABS, Afrique du Sud)
Patricia Balaesque (CR, CRBE, Toulouse)
Julien Varaldi (MCU, LBBE, Lyon)

Rapporteur
Rapporteuse
Rapporteur
Examinatrice
Examineur

Remerciements

J'ai eu la chance d'être extrêmement bien entourée tout au long de mon parcours, tant sur le plan professionnel qu'humain. Je souhaite tout d'abord remercier les personnes qui m'ont formée à la recherche, en premier lieu Evelyne Heyer, pour la confiance qu'elle m'a accordé, pour la curiosité et l'imagination scientifique dont elle regorge, et pour le plaisir à travailler ensemble dans une ambiance toujours très chaleureuse. J'aimerais ensuite remercier Molly Przeworski, pour tout ce qu'elle m'a appris, pour sa passion et sa rigueur scientifique, et pour m'avoir permis de me lancer dans mes projets avec un soutien sans faille. J'aimerais également remercier Renaud Vitalis pour avoir passé autant de temps à me former pendant ma thèse, toujours avec bienveillance, et pour m'avoir aidé à me construire en tant que chercheuse.

J'aimerais ensuite remercier les personnes qui m'ont formée en tant qu'encadrante, que j'ai eu la chance d'accompagner un peu ou beaucoup, et avec qui les discussions m'ont toujours beaucoup apporté, particulièrement Nina Marchi, Elise Lucotte, Léa Bouteille, Ana Lokmer et Johanne Adam au musée de l'Homme, et Anthony Antos, Dominique Hicks, Manuel Barrientos et Célia Lacomme au LBBE. C'est une vraie chance que de pouvoir confronter ses idées avec des personnes aussi diverses, et chaque rencontre m'a fait évoluer à sa manière.

J'aimerais également remercier chaleureusement mes collègues du Musée de l'Homme : entre autres Paul Verdu, Raphaëlle Chaix, Bruno Toupance, Céline Bon, Shelly Masi et Marie-Claude Kergoat ; de l'Université de Chicago : Timothée Flutre, Joanna Kelley, Joseph Pickrell et Ellen Leffler; et du LBBE : Carina Mugal, Annabelle Haudry, Damien de Vienne, Julien Joseph, Sylvain Charlat, Julien Varaldi, et d'autres membres de l'équipe BPGE et du Cocon, pour m'avoir transmis leurs savoirs, pour avoir partagé leurs opinions, pour m'avoir ouvert l'esprit, pour m'avoir écoutée et conseillée, pour m'avoir pleinement intégrée, et pour avoir largement égayé mon quotidien. Sorry Carina that I wrote this HDR in French.

Je tiens à remercier tout particulièrement Fabrice Vavre et Emmanuel Desouhant, ainsi que Laurent Duret et Bastien Boussau pour m'avoir accueillie à bras ouverts au LBBE, pour leur bienveillance sans limite, leur modestie, leur immense savoir scientifique et leur exemplarité tant au niveau humain que professionnel. Je n'avais pas bien pris la mesure de la chance que j'ai eu de rejoindre ce laboratoire, ces équipes, ce bureau.

J'aimerais aussi remercier Magali Chabé, Mathieu Groussin, William Helbert et Virginie Rougeron pour m'avoir embarqué dans leurs projets ou accepté que je les embarque dans les miens.

J'aimerais également remercier mes acolytes de terrain, ceux de la première heure qui m'ont vu m'essayer à la guimbarde et au russe, et me débattre avec la vodka kirghize et les yeux de mouton (Evelyne, Philippe, Patrick), ceux qui ont accepté de m'accompagner chercher des fèces humains dans la forêt camerounaise et ont été récompensés par un ragoût de singe (Alain, Fabrice, Rihlat, Mathieu, Sophie, Francis), celle qui m'a permis de découvrir l'intimité d'une yourte et a traduit mes nombreuses questions sans queue ni tête aux éleveurs mongols

(Sandrine), et ceux qui m'ont initié à la traque de crottes fraîches de babouins en pleine ville (Célia, Virginie, Franck et Benjamin). Je remercie également les personnes qui nous ont ouvert leurs portes et nous ont donné de leur temps ou d'eux-mêmes pour que nos projets de recherche puissent se faire.

Je souhaite également remercier les personnes qui, plus discrètement, ont permis que mon travail se passe dans les meilleures conditions possibles, et ce parfois malgré mon manque d'anticipation ou de compréhension, notamment tous les personnels administratifs et techniques qui m'ont accompagnée au musée de l'Homme et au LBBE.

Je remercie enfin le jury d'avoir accepté de lire et d'évaluer ce mémoire. Leurs retours me seront précieux.

Plus personnellement, je remercie ma famille et mes ami.e.s pour toutes les questions qu'ils me posent et pour leurs efforts répétés pour essayer de comprendre (et retenir) ce que je fais. Fabrice, merci d'avoir supporté de garder les échantillons fécaux dans notre chambre d'hôtel au Cameroun malgré l'odeur.

Tables des matières

RESUME	5
INTRODUCTION	6
<i>Boite 1. Sélection naturelle : définitions et détection</i>	<i>11</i>
PARTIE 1- DETECTION DE SELECTION POSITIVE DANS LES POPULATIONS HUMAINES.....	14
1. ADAPTATION ET MALADAPTATION : LE CAS DU DIABETE DE TYPE 2	14
2. ADAPTATION A LA CONSOMMATION DE LAIT	16
3. ADAPTATION POLYGENIQUE EN ASIE INTERIEURE	23
4. TRAVAIL EN COURS	25
PARTIE 2- DETECTION DE SELECTION BALANCEE CHEZ LES PRIMATES	30
1. GROUPE SANGUIN ABO	30
2. SCAN GENOMIQUE DE SELECTION BALANCEE PARTAGEE ENTRE HUMAINS ET CHIMPANZES	36
3. TRAVAIL EN COURS	37
PARTIE 3- CO-EVOLUTION HOTE-MICROBIOTE INTESTINAL	41
1. MICROBIOTE INTESTINAL HUMAIN ET MODE DE VIE	41
<i>Boite 2. Indices de diversité écologique.....</i>	<i>43</i>
<i>Boite 3. 16S, métagénomique, et MAGs (metagenome-assembled genomes).....</i>	<i>44</i>
2. COEVOLUTION HUMAIN-BACTERIE A UNE ECHELLE MONDIALE	51
3. MICROBIOTE ET TRANSFERT HORIZONTAL DE GENES	53
4. TRAVAIL EN COURS	60
AUTRES PROJETS NON DEVELOPPES	62
PROJET: IMPORTANCE DE LA SELECTION BALANCEE DANS L'ADAPTATION	64
1. PUISSANCE DES METHODES EXISTANTES ET POTENTIEL DEVELOPPEMENT METHODOLOGIQUE	65
2. JEU DE DONNEES EMPIRIQUES.....	66
3. PREVALENCE DE LA SELECTION BALANCEE ET MECANISMES SOUS-JACENTS.....	67
4. ROLE DE LA SELECTION BALANCEE DANS LE MAINTIEN DE VARIANTS STRUCTURAUX	70
5. PERSPECTIVES / CONCLUSIONS.....	71
CONCLUSION	72
REFLEXIONS SUR L'ENCADREMENT	74
REFLEXIONS SUR L'IMPACT ECOLOGIQUE ET L'ETHIQUE DE LA RECHERCHE	76
RÉFÉRENCES BIBLIOGRAPHIQUES.....	78

Résumé

Mes travaux de recherche ont pour objectif général de retracer les dynamiques évolutives des populations, en particulier les dynamiques adaptatives, grâce aux outils de la génétique des populations. Afin d'étudier comment les organismes s'adaptent génétiquement à leur environnement, je cherche à identifier les signatures génomiques de la sélection naturelle, afin de comprendre quels mécanismes adaptatifs sont les plus prévalents, quels sont les traits phénotypiques les plus soumis à des contraintes sélectives, et quels principes généraux gouvernent l'évolution adaptative des populations. Il s'agit aussi d'avoir accès à une fenêtre sur le passé évolutif des espèces pour comprendre quelles ont été les transitions majeures de leur environnement ou de leur mode de vie, et comment l'histoire évolutive des populations peut expliquer la diversité génétique et phénotypique observée aujourd'hui.

Je me suis plus précisément focalisée sur les trajectoires évolutives des populations humaines, avec un intérêt particulier pour l'influence de la diversité culturelle sur la diversité génétique. J'ai en effet étudié pendant ma thèse comment l'organisation sociale des populations (choix matrimoniaux, patrons de migration) et leur régime alimentaire (éleveurs, agriculteurs) influence la diversité génétique (neutre et adaptative) en Asie Centrale. Pendant mon postdoctorat, je me suis intéressée à un mode d'adaptation particulier, la sélection balancée, à une échelle taxonomique plus large, au niveau des primates, pour i) mieux comprendre l'histoire évolutive du gène ABO dont les polymorphismes sont maintenus depuis des millions d'années, et ii) identifier des régions du génome dont les variations génétiques sont partagées entre humains et chimpanzés. Suite à mon recrutement au CNRS à Paris, j'ai cherché à comprendre en quoi le microbiote intestinal pouvait contribuer à l'adaptation des populations humaines à des régimes alimentaires différents. Après avoir étudié la structure microbienne intestinale et orale de chasseur-cueilleurs, agriculteurs et pêcheurs du Cameroun, je me suis focalisée sur le contraste entre populations rurales et urbaines, en considérant à la fois les procaryotes et les eucaryotes intestinaux. J'ai également contribué à caractériser les transferts horizontaux de gènes au sein du microbiote intestinal humain, et à identifier les souches bactériennes dont la phylogénie suit celles de leurs hôtes à une échelle globale. Depuis ma mutation au LBBE (Lyon) en 2020, j'ai continué à m'intéresser à la coévolution hôte-microbiote en étudiant un cas de transfert horizontal de gènes depuis des bactéries marines vers le microbiote de populations asiatiques. J'ai également poursuivi mes recherches sur l'adaptation locale de populations humaines à leur environnement (alimentaire et climatique). J'ai enfin rejoint un projet visant à comprendre en quoi l'anthropisation de populations de babouins influence leur microbiote intestinal, et plus largement leur état de santé.

Le projet de recherche que je souhaite désormais développer vise à évaluer l'importance de la sélection balancée dans l'adaptation des espèces, en m'appuyant sur des simulations et des données de séquençage, afin d'estimer la prévalence de la sélection balancée dans des taxons contrastés (primates, oiseaux, insectes), de comprendre l'importance relative de différents mécanismes évolutifs aboutissant au maintien de polymorphismes, et de tester si la sélection balancée cible préférentiellement certains variants structuraux comme les inversions.

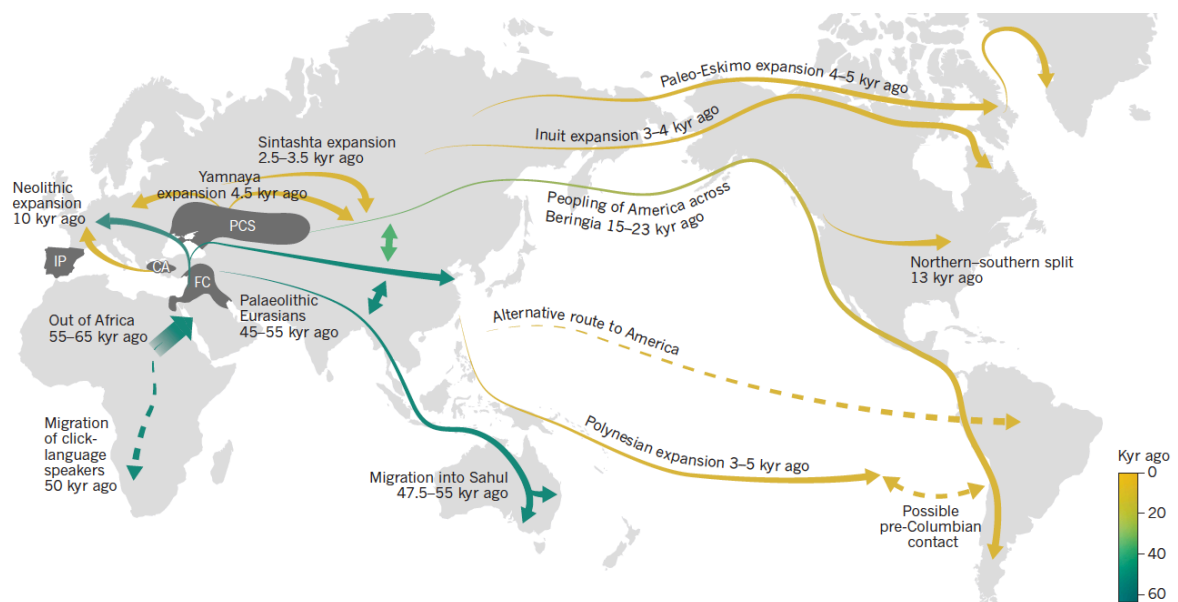
Introduction

Chaque individu porte, au sein de son génome, de l'information *collective* sur l'histoire de son espèce (Li et al. 2011). Un individu est en effet une mosaïque génétique de nombreux ancêtres pouvant appartenir à plusieurs populations, et parfois à plusieurs espèces. La génétique des populations est une discipline qui permet, à partir de l'étude de la diversité génétique des populations, de retracer les dynamiques évolutives relativement récentes de ces différents ancêtres. En effet, cet outil donne accès à l'histoire démographique et adaptative des populations sur une échelle de temps qui correspond à celle où ségrégent des polymorphismes génétiques. A titre d'exemple, pour l'espèce humaine, l'analyse de la diversité génétique actuelle permet d'inférer les dynamiques évolutives passées jusqu'à il y a environ un million d'années (Albers et al. 2020). Au-delà de ce temps, toutes les séquences génétiques humaines actuelles ont un seul et même ancêtre commun et nous avons donc perdu toute trace de la diversité génétique antérieure. En parallèle, la phylogénie, de par l'étude des liens de parenté génétiques entre espèces, nous permet de connaître l'histoire évolutive plus ancienne des espèces. Ces données sont informatives sur une échelle de temps qui correspond au temps de divergence entre espèces ; chez l'humain, nos plus proches cousins étant le chimpanzé et le bonobo, nous pouvons étudier l'évolution de nos lignées respectives, qui sont longues d'environ 7 millions d'années (Langergraber et al. 2012).

Ces deux approches donnent accès à des temporalités différentes et nous ouvrent des fenêtres complémentaires sur notre passé, questionnant par exemple pourquoi et comment les espèces naissent et s'éteignent, quand et dans quelle mesure les populations s'agrandissent, se métissent, et/ou se structurent, ou encore comment les populations s'adaptent à diverses contraintes de leur environnement. Malgré le fait que l'évolution ne soit, tout du moins à mes yeux, pas une science ayant un fort pouvoir prédictif, probablement du fait de l'importance de nombreux phénomènes aléatoires sous-jacents et de la labilité des contraintes environnementales, un objectif important de la recherche en évolution est tout de même d'établir des principes généraux qui pourront nous aider à mieux anticiper comment les espèces vont répondre aux changements environnementaux en cours et à venir, et par exemple à adapter les mesures de gestion des espèces en biologie de la conservation. Cette question est d'autant plus cruciale que l'ensemble du vivant est actuellement confronté à une crise climatique et anthropique sans précédents. Mais il s'agit aussi d'une quête de nos origines : comprendre d'où nous venons, pour mieux comprendre qui nous sommes. Et ce tout particulièrement quand l'on s'intéresse à l'histoire évolutive de l'espèce humaine.

La génétique des populations humaines s'est en effet attachée à décrire de manière très précise quand *Homo sapiens* a émergé et a colonisé différents territoires, quels ont ensuite été les mouvements des populations et les taux de métissage entre différentes populations (mais aussi avec des espèces archaïques), et comment les populations se sont adaptées à diverses contraintes environnementales (Cavalli-Sforza et al. 2003; Nielsen et al. 2017; Wohns et al. 2022). Ces questions de recherche fondamentale permettent ainsi de mieux comprendre

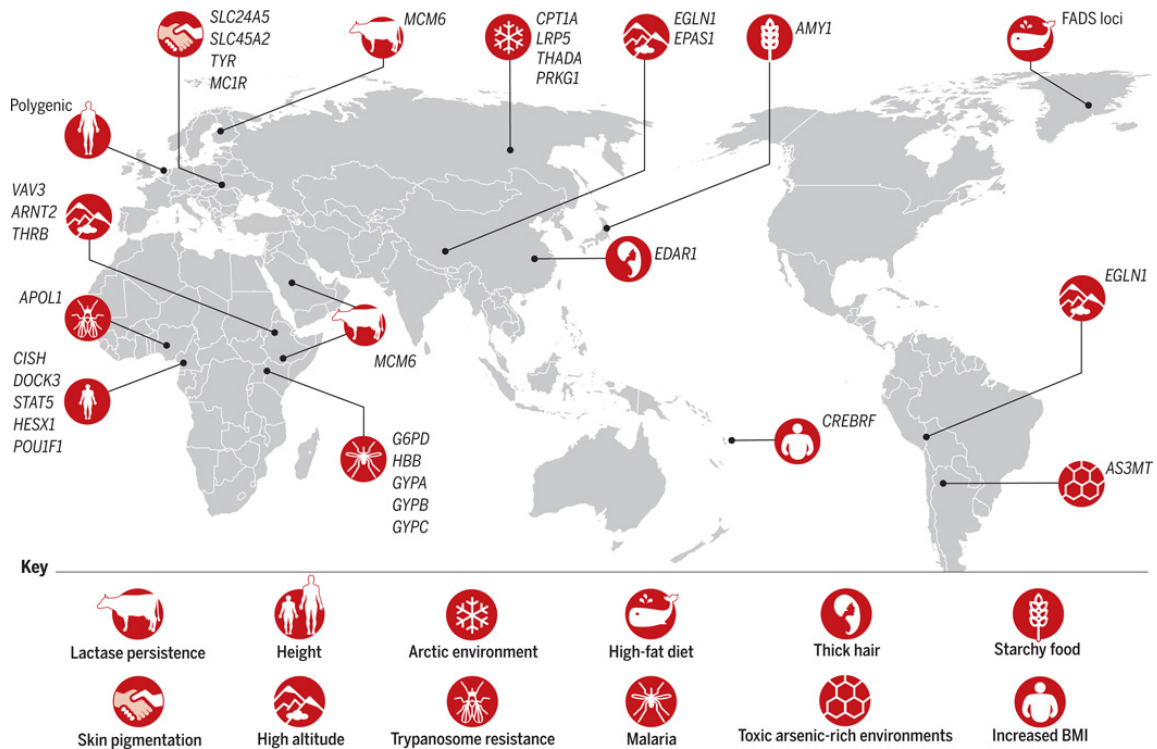
comment les populations se sont succédées les unes aux autres pour donner lieu à la diversité observée aujourd'hui, et entre autres de nourrir les débats sociétaux liés aux problématiques de discriminations raciales. Il a notamment découlé de ces études que les populations humaines sont le fruit de nombreux mouvements de populations et de métissages complexes, de telle sorte qu'aucune population ne résulte d'une simple continuité génétique locale. A titre illustratif, les populations européennes actuelles descendent de populations africaines d'*Homo sapiens* chasseurs-cueilleurs qui ont migré en Europe il y a environ 50 000 ans et qui se sont métissées au passage avec Néandertal, puis qui se sont faites remplacées en grande partie, il y a environ 7 500 ans, par des populations migrantes d'agriculteurs du Proche-Orient, pour ensuite se mélanger de manière complexe il y a environ 4 500 ans avec des populations agropastorales des steppes de l'Est (Haak et al. 2015). N'en déplaise à certains, les populations humaines ont toujours été en mouvement, et se sont toujours mélangées.



Age et trajectoires des principales migrations humaines inférées à partir de données génétiques.

Nielsen et al. (2017)

L'histoire évolutive humaine est également particulièrement intéressante à étudier sous un angle adaptatif, du fait que notre espèce a colonisé, dans un intervalle de temps relativement court (~60 000 ans), des milieux climatiques extrêmement diversifiés, et a développé en parallèle une grande diversité culturelle (notamment alimentaire). Cette variabilité des modes de vie représente à la fois une capacité accrue d'adaptation, au travers d'innovations culturelles, mais également une source de nouvelles contraintes (certains choix culturels n'étant pas favorables à la bonne survie ou reproduction des individus). L'humain représente ainsi une espèce singulière qui a évolué, comme toutes les autres espèces, en étroite interaction avec son environnement, mais qui a également été moteur de sa propre évolution, de par la modification de son milieu. L'étude de la diversité génétique adaptative nous permet donc d'avoir accès aux contraintes, naturelles et/ou culturelles, auxquelles les populations humaines ont fait face dans le passé, et donc indirectement aux changements de modes de vie de nos ancêtres.



Exemples connus d'adaptation génétique locale humaine, regroupés par le type de phénotype et/ou de contrainte sélective, ainsi que les gènes impliqués. Fan et al. (2016)

Les études récentes de génétique des populations nous ont ainsi appris que les populations humaines se sont localement génétiquement adaptées à la présence de certains pathogènes (paludisme, trypanosomiase), à certaines contraintes climatiques (manque d'oxygène en altitude, faibles températures et baisse d'ensoleillement dans le Nord de l'Eurasie), mais également à des contraintes liées à l'alimentation, en particulier à la transition d'un mode de vie basé sur la chasse et la cueillette à un mode de vie agropastoral (consommation de céréales, de lait à l'âge adulte) (Fan et al. 2016; Rees et al. 2020). Dans le cas emblématique de l'éclaircissement de la couleur de la peau en Eurasie, sélectionnée pour favoriser une synthèse cutanée plus importante de vitamine D (permise par l'action des rayons UV), il semble que la pression sélective ait découlé d'une combinaison de contraintes climatiques et alimentaires¹. Il est cependant important de noter qu'une difficulté majeure des études de génétique des populations est de connaître les *phénotypes* sous sélection, ainsi que les contraintes sélectives agissant sur ces phénotypes ; en conséquence, dans de nombreux cas, la pression de sélection n'est pas clairement identifiée (par exemple la réduction de la taille dans les forêts tropicales humides, l'augmentation de la densité des glandes sudoripares ou encore la mauvaise détoxification de l'alcool en Asie de l'Est).

¹ La vitamine D est essentielle pour la santé puisqu'elle participe à l'absorption du calcium et du phosphore, stimule les cellules du système immunitaire, et est primordiale pour la croissance osseuse. Elle peut être obtenue par l'alimentation (huile de foie de morue, poissons gras, œufs, champignons, produits laitiers) ou bien être synthétisée par l'organisme à partir du cholestérol lorsque la peau est soumise aux rayons UVB du soleil. La baisse d'ensoleillement dans les latitudes élevées, combinée à des changements d'alimentation chez les néo-agriculteurs du Néolithique, semblent avoir entraîné des carences en vitamine D, résultant, entre autres, en du rachitisme, et favorisant ainsi au cours du temps les peaux plus claires. Encore actuellement, les carences en vitamine D sont fréquentes en France (60-80%), particulièrement en hiver, et surtout chez les femmes enceintes, les personnes âgées et les personnes ayant une couleur de peau foncée.

Ces histoires adaptatives ne sont pas seulement intéressantes pour retracer les changements de modes de vie des populations du passé, mais également pour mieux comprendre pourquoi les populations actuelles peuvent présenter des différences physiologiques significatives pour certains traits d'importance en santé publique (maladies métaboliques, auto-immunes, prédispositions allergiques, réponses à certains médicaments...). Par exemple, les différences de prévalence d'hypertension et de diabète de type 2 observées entre populations pourraient en partie être le résultat d'adaptations génétiques locales à des environnements climatiques et culturels différents (Hoh et al. 2019; Meeks et al. 2021). Cependant, de manière générale, nous ne comprenons encore qu'assez mal si les différences physiologiques observées entre populations humaines sont plutôt la conséquence de facteurs génétiques ou au contraire plutôt de déterminants culturels, et donc dans quelle mesure l'origine ethnique des individus est informative sur leurs facteurs de risque (Lu et al. 2014).

Mes travaux de recherche se situent à l'intersection de ces différentes questions de recherche, et ont pour objectif général d'utiliser les outils de la génétique des populations pour retracer les dynamiques évolutives des populations, notamment des populations humaines. Bien que cette discipline ne soit pas nouvelle, les outils moléculaires et informatiques qui la rendent possible ont connu des développements majeurs ces dernières années, lui permettant de répondre à autant de questions qu'elle en a soulevé de nouvelles. Mes travaux reposent également sur l'intégration de plusieurs autres disciplines comme l'écologie, l'anthropologie et les sciences médicales. En effet, la compréhension de l'évolution des organismes et des espèces dans leur ensemble nécessite un dialogue interdisciplinaire, particulièrement enrichissant.

Je me suis plus précisément intéressée à l'adaptation des populations humaines en réponse aux changements de leur alimentation. Pour cela, j'ai choisi de mener en parallèle deux approches différentes, tant dans leurs concepts théoriques que dans leurs outils méthodologiques : une perspective humaine et une perspective microbienne. En effet, il apparaît de plus en plus clairement que les organismes ne peuvent plus être vus comme des entités indépendantes, mais doivent être considérés comme des écosystèmes en tant que tels, entraînant un important changement de paradigme (McFall-Ngai et al. 2013). Les humains sont en effet hôtes d'une grande diversité de micro-organismes (bactéries, archées, protozoaires), et parmi eux, le microbiote intestinal représente l'écosystème le plus densément peuplé de la planète. L'approche humaine et microbienne sont donc complémentaires, étant donné que l'adaptation génétique humaine se fait sur un temps long (du fait de son temps de génération d'environ 30 ans), tandis que l'adaptation microbienne est beaucoup plus rapide, permettant de se placer à des échelles de temps très courtes (de l'ordre de quelques années).

Dans ce mémoire, j'ai abordé dans une première partie la question de l'adaptation des populations humaines à leur environnement local (notamment alimentaire) par recherche de *sélection positive* (définie comme l'augmentation en fréquence de mutations avantageuses). Il s'est principalement agi d'étudier l'adaptation génétique des populations humaines suite à la révolution néolithique, une transition culturelle majeure qui a eu lieu il y a 10 000 à 4 000 ans (Diamond et al. 2003). Cette période a entre autres été marquée par d'importantes

transformations des modes de productions alimentaires, avec le passage d'un mode de vie nomade basé sur la chasse, la pêche, et/ou la cueillette, à un mode de vie plus sédentaire basé sur l'agriculture et/ou l'élevage. C'est en effet à partir de ce moment que les populations humaines se sont mises à domestiquer un certain nombre de plantes et d'animaux et ont commencé à consommer en quantité importante des céréales et des produits laitiers. Ces changements alimentaires ont eu des répercussions majeures sur la physiologie et le métabolisme des individus, et ont généré d'importantes pressions de sélection dans le génome humain (Brown 2012). Pour cette partie, j'ai concentré mon travail sur l'analyse de données biologiques et culturelles collectées sur le terrain en Asie Centrale, où cohabitent des éleveurs de langue Turco-Mongole et des agriculteurs de langue Indo-Iranienne.

Dans une deuxième partie, je me suis intéressée à un autre mode d'adaptation des espèces à leur environnement, la *sélection balancée* (définie comme toute forme d'adaptation qui entraîne un maintien de diversité génétique à l'échelle d'une espèce) (Charlesworth 2006). Ce mode de sélection a été proposé comme étant à l'origine de la grande diversité phénotypique observée dans la nature mais il est complexe à étudier car il recouvre différents mécanismes sous-jacents laissant des signatures génomiques assez différentes (Fijarczyk et al. 2015). La question de la prévalence de la sélection balancée à l'échelle du génome, des mécanismes principaux entraînant un tel maintien de polymorphismes génétiques, ainsi que des cibles génétiques de ce mode d'adaptation, reste donc ouverte. Pour cette partie, je me suis située principalement à l'échelle des primates, espèces pour lesquelles des données de diversité génomique se sont accumulées ces dernières années, permettant ainsi de tirer avantage à la fois du signal présent au sein des espèces et entre espèces proches.

Enfin, dans une troisième partie, j'ai adopté une perspective microbienne pour étudier comment les micro-organismes associés aux humains ont coévolué avec leurs hôtes dans le contexte des changements de mode de vie des populations humaines. Ce projet avait pour ambition initiale de comprendre si (et comment) le microbiote intestinal s'est modifié en réponse aux changements alimentaires liés à la transition néolithique, mais il s'est rapidement avéré que le déterminant principal du microbiote intestinal actuel est le degré d'urbanisation et d'industrialisation des populations, et non le fait d'être chasseur-cueilleur, agriculteur ou pêcheur. Comme dit précédemment, ces modifications récentes mais profondes de notre rapport à l'environnement n'ont pas eu le temps d'impacter la diversité génétique humaine, mais elles ont d'ores et déjà eu d'importantes conséquences sur la diversité et composition microbienne intestinale, et donc indirectement sur la physiologie humaine. Je me suis donc focalisée sur l'évolution du microbiote face aux changements *récents* de mode de vie de leurs hôtes. Pour ces recherches, j'ai mis en place des campagnes d'échantillonnage au Cameroun et analysé des données biologiques et culturelles chez des populations de chasseurs-cueilleurs, d'agriculteurs et de pêcheurs, en contexte rural et urbain.

En plus de ces projets, qui ont pour point commun de s'interroger principalement sur les dynamiques *adaptatives* des populations et espèces, j'ai mené d'autres projets (notamment sur les dynamiques *démographiques* des populations) que j'aborderai beaucoup plus succinctement

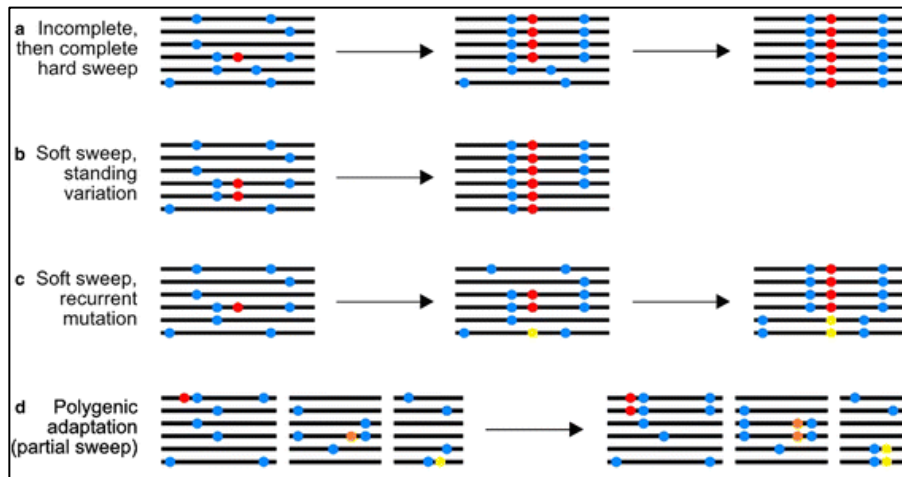
dans une quatrième partie. C'est en effet dans ce cadre que j'ai eu l'opportunité d'effectuer mon premier encadrement doctoral officiel, celui de la thèse de Nina Marchi (2014-2017). Cette thèse a eu pour objectif de comprendre en quoi la diversité socio-culturelle des populations humaines a affecté leur diversité et leur évolution génétique. Plus précisément, Nina s'est attachée à mesurer les taux de migration sexe-spécifique et la consanguinité génétique de populations ayant des organisations sociales différentes (populations patrilocales versus cognatiques), en Asie Centrale. Elle a également contribué à améliorer la compréhension du peuplement de cette région centrale en participant à l'analyse conjointe de données d'ADN moderne et ancien.

Boite 1. Sélection naturelle : définitions et détection

Il existe dans la nature d'importantes variations phénotypiques entre individus, populations, et espèces, résultant de différences génétiques, environnementales et/ou culturelles. Quand cette diversité phénotypique est le fruit d'une diversité génétique, elle est soumise à des processus évolutifs qui peuvent être regroupés en des processus neutres (aléatoires) qui touchent l'ensemble du génome, et des processus adaptatifs qui touchent les polymorphismes génétiques en fonction de leurs effets sur la valeur sélective (fitness) des individus. Bien que l'adaptation soit clairement un moteur à l'origine de l'évolution des espèces, quantifier son rôle et son importance dans la diversité phénotypique et génétique des espèces, et connaître les traits phénotypiques influençant la fitness des individus, reste depuis longtemps une question d'intérêt fondamental en biologie évolutive (Nei 2005).

Parmi les différents processus adaptatifs, on peut distinguer la sélection directionnelle incluant la sélection positive et négative (l'augmentation / diminution en fréquence de mutations avantageuses / désavantageuses, respectivement), qui toutes deux engendrent une baisse de diversité génétique autour du locus sous sélection ; et la sélection balancée (également appelée équilibrante), qui regroupe toutes les formes de sélection naturelle entraînant un maintien de diversité génétique au cours du temps autour du locus sous sélection.

La sélection positive entraîne donc un balayage sélectif autour de la mutation avantageuse, du fait du déséquilibre de liaison entre loci, qui peut être plus ou moins complet. En effet, si la sélection agit sur une mutation unique qui vient de se produire (mutation *de novo*), la mutation augmentera en fréquence jusqu'à atteindre la fixation, et le balayage sera donc maximal et complet (« hard sweep »). Par contre, si la mutation sous sélection était déjà présente dans la population avant de devenir avantageuse (« standing variation »), et/ou si la sélection agit sur un trait polygénique, c'est-à-dire encodé par plusieurs mutations, le phénotype optimal sera atteint en augmentant la fréquence d'allèles à plusieurs loci en parallèle, et donc chacun augmentera seulement partiellement en fréquence. Dans ce cas, le balayage sélectif sera incomplet (« soft » ou « partial sweep »). Évidemment, plus le trait sous sélection est codé par un grand nombre de mutations (donc fortement polygénique), plus le changement en fréquence allélique sera subtil, et plus la signature génomique associée sera difficile à identifier.



Effet de balayages sélectifs plus ou moins complets sur la diversité génétique. Booker et al. (2017)

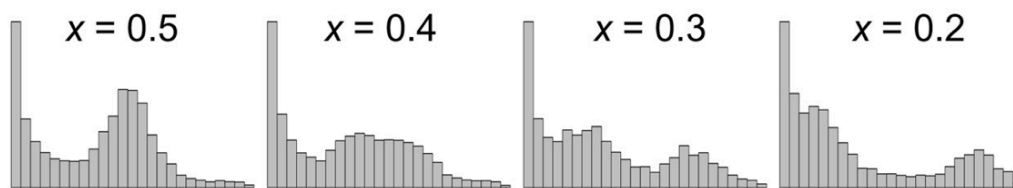
Bien que la sélection positive entraîne des signatures plus ou moins intenses dans le génome, et donc nécessite de déployer des outils différents pour mettre en évidence des balayages complets ou incomplets, elle reste caractérisée par un seul mécanisme sous-jacent (l'augmentation en fréquence de mutations avantageuses), qui peut se détecter à partir de trois signatures génomiques : i) l'augmentation du nombre de polymorphismes à faibles fréquences, entraînant une distorsion du spectre de fréquences alléliques, ii) l'augmentation du déséquilibre de liaison et de la taille des régions d'homozygotie, et iii) l'augmentation du niveau de différenciation entre populations (correspondant spécifiquement à un scénario d'adaptation locale) (Weigand et al. 2018).

A l'inverse, la sélection balancée recouvre des mécanismes causaux assez différents : l'avantage à l'hétérozygote (AH ; l'hétérozygote présente une valeur sélective plus importante que les deux types d'homozygotes, respectivement), la sélection fréquence-dépendante négative (SFDN ; les allèles les plus rares sont avantagés), la sélection fluctuante du fait d'environnements hétérogènes dans le temps et/ou l'espace, la sélection sexuellement antagoniste (quand les deux sexes ont des optimums différents) ou la pléiotropie antagoniste (quand les mutations ont des effets contrastés à différents moments du cycle de vie) (Charlesworth 2006; Llaurens et al. 2017; Bitarello et al. 2023). Le point commun de ces mécanismes est que plusieurs allèles sont maintenus dans la population, tandis que sous neutralité, les mutations sont vouées, pour la majorité, à disparaître (c'est-à-dire à revenir à une fréquence de 0), et, pour une minorité, à se fixer (c'est-à-dire à atteindre une fréquence de 1).

Une signature classique de sélection balancée est donc une augmentation de la diversité génétique. Cependant, cette augmentation de diversité génétique peut se traduire par une augmentation du nombre de polymorphismes génétiques par unité de distance physique (particulièrement s'il y a un grand nombre d'haplotypes différents maintenus par la sélection), ou par une augmentation du nombre moyen de différences entre haplotypes (particulièrement si la fréquence des différents haplotypes maintenus est intermédiaire). Or, tandis que l'AH prédit plutôt la coexistence de deux haplotypes avec des fréquences alléliques stables et probablement intermédiaires (qui peuvent cependant en théorie prendre toutes les valeurs, en fonction de la différence de fitness entre les deux types d'homozygotes), la SFDN prédit plutôt

un grand nombre d'haplotypes et des fréquences alléliques faibles (égales à $1 / \text{le nombre d'allèles}$ qui, lui, peut tendre vers l'infini) ; la sélection fluctuante, quant à elle, prédit des fréquences alléliques instables pouvant prendre toutes les valeurs. Par conséquent, les différentes mesures de diversité génétique ne présentent pas nécessairement la même puissance de détection en fonction des mécanismes de sélection balancée à l'œuvre.

De la même manière, l'autre signature très utilisée de la sélection balancée, la distorsion du spectre de fréquences alléliques (SFS, voir ci-dessous) du fait d'un excès de polymorphismes à fréquence intermédiaire, ne permet pas de détecter avec la même puissance l'ensemble des mécanismes cités plus haut, étant donné qu'ils ne produisent pas les mêmes attendus.



Effet de la sélection balancée sur le spectre de fréquence allélique (SFS), en considérant différentes fréquences alléliques d'équilibre (x). Cheng et DeGiorgio (2020)

Une autre caractéristique commune à tous ces mécanismes est la réduction de divergence entre espèces (du fait que pour devenir un site divergent, un polymorphisme sous sélection balancée doit être transmis sur l'haplotype complémentaire par un événement de recombinaison). Cette baisse de divergence est d'autant plus forte qu'il y a de nombreux haplotypes maintenus dans la population. De manière intéressante, la divergence est classiquement utilisée non pas comme une signature de sélection balancée en tant que telle, mais pour corriger l'effet des variations du taux de mutation le long du génome, qui représente un facteur confondant pour l'étude des fluctuations de diversité génétique. La difficulté avec la divergence, cependant, est que sa réduction met du temps à devenir significative, et la force de ce signal est donc négativement corrélée avec la longueur du segment le portant (qui lui, diminue avec le temps).

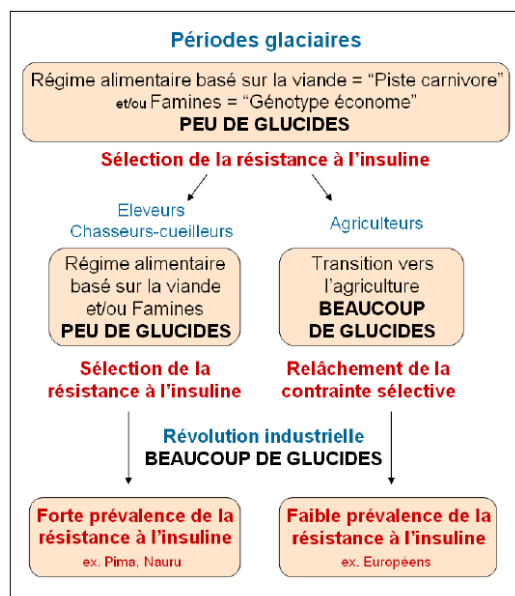
Pour résumer, la sélection balancée est donc particulièrement difficile à identifier dans son ensemble du fait de la diversité de ces mécanismes sous-jacents et de leurs effets contrastés sur les patrons de diversité génomique (Fijarczyk et al. 2015; Bitarello et al. 2023), mais également à cause de l'effet confondant de l'histoire démographique qui est souvent complexe (et difficile à inférer avec précision). Ceci explique certainement pourquoi la sélection balancée a été largement moins étudiée que la sélection positive.

Enfin, il est intéressant de noter que sélection positive et balancée ne constituent pas deux entités fondamentalement distinctes, étant donné que l'adaptation locale, par sélection positive d'un allèle dans une population donnée et d'un autre allèle dans une autre population, peut entraîner un maintien de diversité génétique à l'échelle de l'espèce, et donc être considérée comme une forme de sélection balancée, à l'image des motifs mimétiques chez *Heliconius* (Joron et al. 1999).

Partie 1- Détection de sélection positive dans les populations humaines

1. Adaptation et maladaptation : le cas du diabète de type 2

Aux Etats-Unis, l'incidence du diabète de type 2, caractérisé par un excès de sucre dans le sang non insulino-dépendant, est 19 fois plus forte dans les populations Pima (Amérindiens) que dans celles d'origine européenne. Les Pima ont par ailleurs la prévalence la plus forte au monde (50% chez les plus de 35 ans, Knowler et al. 1978, 1983). Ces différences entre groupes ethniques peuvent résulter de facteurs culturels et/ou de facteurs génétiques. Certaines hypothèses évolutives ont proposé que l'adaptation génétique des populations humaines à différents régimes alimentaires soit à l'origine des inégalités observées en terme de prévalence de diabète de type 2 (et plus largement de prévalence du syndrome métabolique).



Hypothèses combinées de la "piste carnivore" et du "genotype économe". Figure inspirée de Miller & Colagiuri (1994)

L'hypothèse du « génotype économe » (Neel 1962; Neel et al. 1998) et de « l'hypothèse carnivore » (Miller et al. 1994; Colagiuri et al. 2002) ont ainsi proposé que les populations ayant eu accès à une quantité limitée de glucides (les chasseurs-cueilleurs, éleveurs, ou toutes populations ayant connu des famines) ont sélectionné des polymorphismes génétiques favorisant une utilisation plus efficace des ressources disponibles.

La résistance à l'insuline, qui permet d'éviter l'utilisation des glucides par le foie et les muscles mais favorise son stockage dans la masse grasseuse, pourrait ainsi avoir été avantageuse dans le passé, et continuerait à l'être dans certaines populations où les glucides viennent toujours à manquer. Cependant, dans une grande majorité de

populations où l'apport en sucres n'est actuellement plus limitant, ce phénotype est devenu délétère. D'après ces hypothèses, les populations descendant de chasseurs-cueilleurs ou d'éleveurs devraient alors présenter une fréquence élevée de résistance à l'insuline et montrer des signes de sélection positive d'allèles à risque pour le diabète de type 2 (l'expression symptomatique de la résistance à l'insuline sur le long terme). A contrario, ces pressions de sélection devraient s'être relâchées dans les populations ayant adopté des pratiques agricoles depuis la révolution néolithique (du fait de l'apport important de glucides procuré par les céréales). Ces hypothèses n'ont cependant pas reçu de réel support des données génétiques (Vander Molen et al. 2005; Helgason et al. 2007; Hancock et al. 2008; Klimentidis et al. 2011).

Pour comprendre les facteurs à l'origine des différences de prévalence de diabète de type 2 entre populations, j'ai participé pendant ma thèse à une campagne d'échantillonnage en anthropologie évolutive visant à disséquer l'histoire adaptative des polymorphismes associés

au diabète de type 2 dans des populations aux modes de vie contrastés (des éleveurs Kirghizes et des agriculteurs Tadjiks), en Asie Centrale. Grâce à un financement ANR, nous avons effectué cinq missions de terrain, en milieu rural et urbain, consistant à récolter pour environ 700 individus des données génétiques, physiologiques, anthropométriques, ethnolinguistiques et alimentaires. Les résultats décrits ci-dessous ont été publiés dans Ségurel et al. (2013).

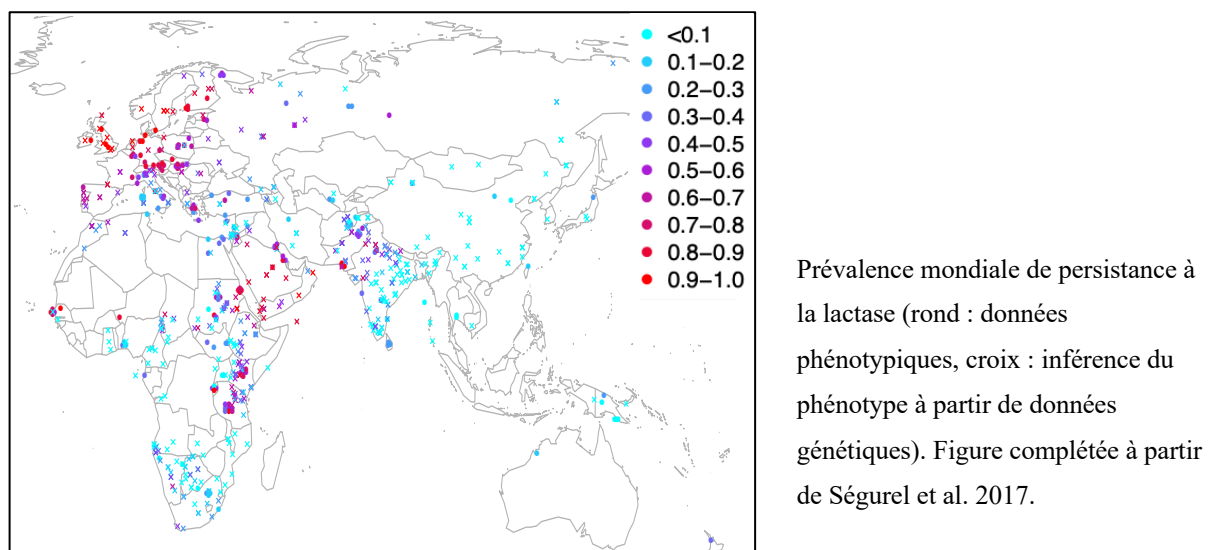
Les données phénotypiques ont tout d'abord permis de montrer, grâce à des régressions linéaires et logistiques, que l'ethnicité a un effet significatif sur la résistance à l'insuline (facteur de risque 1.8 fois plus élevé chez les éleveurs Kirghizes), en contrôlant pour diverses variables confondantes (sexe, âge, poids, environnement urbain vs rural, activité physique). J'ai ensuite séquencé dix régions d'environ 1.1 kb autour de mutations associées au risque de diabète de type 2 et vingt régions d'environ 1.3 kb choisies pour ne pas être a priori sous l'action de la sélection, chez 40 individus de chaque groupe ethnique, afin de caractériser la fréquence d'allèles protecteurs versus à risque dans chaque population. Ces analyses ont montré que, malgré leur plus grand risque phénotypique, les éleveurs Kirghizes ne présentent pas d'allèles à risque en plus forte fréquence que les agriculteurs Tadjiks.

J'ai ensuite effectué des tests de sélection pour les dix régions candidates basés sur la distorsion du spectre de fréquence allélique et la différenciation entre populations. Nous avons également génotypé tous les individus sur une puce à ADN Illumina d'environ 660 000 SNPs, afin de pouvoir estimer les coefficients de sélection et dater des potentiels événements sélectifs à partir de données haplotypiques. J'ai ainsi trouvé qu'un gène, *FABP2*, semble sous sélection balancée dans les deux populations (excès d'allèles en fréquence intermédiaires) tandis que deux gènes, *LEPR* et *KCNQ1*, présentent des polymorphismes significativement plus différenciés qu'attendu entre Kirghizes et Tadjiks (ainsi que des traces de sélection positive récente chez les Kirghizes pour *LEPR*). Grâce à l'analyse de populations supplémentaires de la littérature (base de données HGDP-CEPH), nous avons pu déduire que le signal trouvé sur le gène *KCNQ1* semble lié au mode de vie pastoral, étant donné qu'il est retrouvé dans d'autres populations d'éleveurs (Hancock et al. 2010). A l'inverse, le signal sur *LEPR* semble correspondre à une adaptation différentielle entre populations asiatiques et européennes (Voight et al. 2006). De manière importante, les signatures de sélection positive identifiées correspondent dans tous les cas à la sélection d'allèles protecteurs par rapport au diabète de type 2 (qu'il s'agisse d'allèles ancestraux ou dérivés) ayant été sélectionnés durant le Néolithique (il y a entre 5 500 et 12 000 ans). Ce même résultat a été trouvé pour des gènes associés à des maladies cardiaques (Fullerton et al. 2000) et à de l'hypertension (Thompson et al. 2004). Nous n'avons donc trouvé aucun signal de sélection positive d'allèles à risque (contrairement aux prédictions des hypothèses précédemment décrites), bien que nous ne puissions pas exclure que nous n'avons pas assez de puissance pour détecter de tels événements plus anciens. Par contre, la révolution néolithique semble bien s'être accompagnée d'événements d'adaptations locales différentes entre populations pour ces gènes. Il est par contre assez surprenant que ce soit chez les Kirghizes, les plus à risque de développer de la résistance à l'insuline, que les allèles protecteurs aient été identifiés comme étant sous sélection

positive. Ceci peut s'expliquer par une mauvaise prise en compte de facteurs confondants difficiles à quantifier correctement (notamment alimentaires et d'activité physique), et/ou par une faible corrélation entre génotype et phénotype dans ces populations (les polymorphismes associés au diabète ayant été principalement identifiés dans des populations européennes, bien que nous ayons été justement vigilants à choisir des polymorphismes dont l'association a été répliquée en Asie de l'Est où il y a des données d'association génotype-phénotype).

2. Adaptation à la consommation de lait

Un des exemples les plus emblématiques d'adaptation locale chez l'humain est certainement celui de l'adaptation à la consommation de lait chez certaines populations agropastorales, aliment nouvellement introduit à l'âge adulte durant le Néolithique, avec la domestication de certains animaux (vache, mouton et chèvre en Europe ; vache en Afrique ; cheval, chameau et buffle en Asie) (Ségurel et al. 2017). J'ai eu l'occasion de travailler sur cette question en thèse, dans le cadre de la comparaison entre éleveurs Turco-Mongols et agriculteurs Indo-Iraniens d'Asie Centrale (Heyer et al. 2011). J'ai plus tard ré-examiné cette question en collaboration avec deux étudiantes en thèse, Perle Guarino-Vignon (co-encadrée par Evelyne Heyer et Céline Bon) et Nina Marchi (co-encadrée par E Heyer et moi-même), ce qui a permis de combiner des données d'ADN moderne et d'ADN ancien pour améliorer notre compréhension de cette histoire adaptative encore bien mystérieuse (Ségurel et al. 2020).



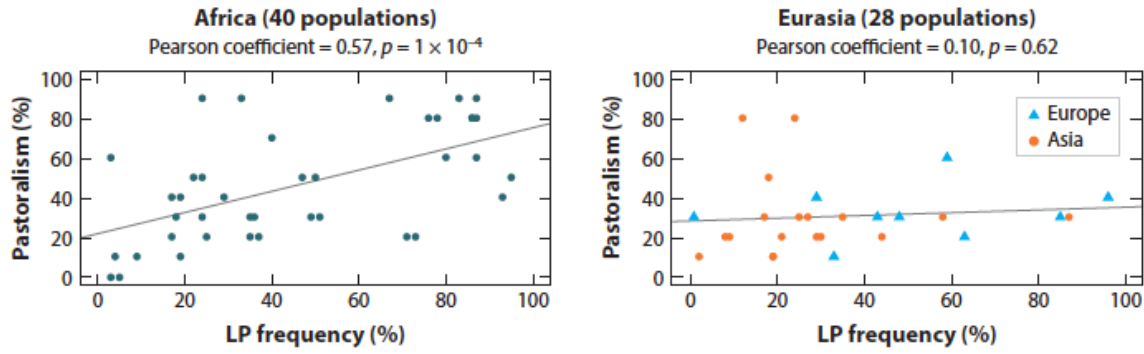
Il a en effet été proposé que, dans les populations pastorales consommant du lait animal depuis le Néolithique, plusieurs mutations associées au phénotype de persistance de la lactase (PL) aient été sélectionnées. Ce phénotype est défini par la capacité des adultes à exprimer l'enzyme de la lactase (normalement uniquement exprimée jusqu'au sevrage), et donc à pouvoir continuer à digérer le seul sucre contenu dans le lait, le lactose (c'est à dire le transformer en glucose et galactose digestes), tout au long de leur vie. En l'absence de cette enzyme, les adultes consommant trop de lactose souffrent de diarrhées et de crampes abdominales dues, entre autres, à la fermentation du lactose par certaines bactéries du microbiote intestinal. Cette

histoire adaptative représente un remarquable cas de convergence évolutive (avec au moins 5 mutations différentes associées à ce trait et sous sélection naturelle) (Ségurel et al. 2017), un exemple unique de coévolution entre gène et culture (Holden et al. 1997), ainsi qu'un des signaux les plus forts de sélection dans le génome humain (Bersaglieri et al. 2004; Tishkoff et al. 2007). Cependant, de nombreuses questions restent en suspens.

Notamment, quelle est précisément la nature des pressions de sélection ayant entraîné l'augmentation en fréquence des mutations associées à la PL ? S'agit-il d'un avantage lié à un apport calorique, à un évitement de carences en calcium, à une meilleure hydratation, à une résistance accrue au paludisme, ou encore à une réduction du risque de mortalité maternelle grâce à un accroissement de taille du bassin (Flatz et al. 1973; Cook et al. 1975; Gerbault et al. 2011; Cordain et al. 2012; Wells et al. 2021; Evershed et al. 2022; Stock et al. 2023) ? La plupart de ces hypothèses reposent sur le fait que les individus non-persistents ont un coût important suite à la consommation de lait, du fait des symptômes associés à la présence de lactose non digéré dans le colon (notamment des diarrhées). Cependant, ces symptômes ne semblent apparaître que suite à la consommation d'assez grandes quantités de lait frais (c'est-à-dire plus d'un verre de lait par jour, ou encore plus si le lait est consommé avec d'autres aliments qui ralentissent le transit) (Shaukat et al. 2010). Les individus non-persistents peuvent donc à priori bénéficier des protéines, graisses, vitamines et minéraux (comme le calcium) présents dans le lait s'ils en consomment en quantité raisonnable. Ces hypothèses sont donc difficilement réconciliables avec les forts coefficients de sélection observés (aussi élevés que ceux associés au paludisme).

Avec mon collègue pédiatre Alexandre Fabre (PUPH à l'université Aix Marseille), nous avons récemment proposé que la sélection pour la PL puisse avoir eu lieu non pas chez les adultes, mais chez les jeunes enfants (Fabre et al. 2023). En effet, l'activité de la lactase commence à décroître dès deux ans chez les individus non-persistents, âge qui correspond à l'âge moyen du sevrage dans les populations contemporaines préindustrielles (Sellen 2001). Notre méta-analyse montre que 50% des enfants non-persistents n'expriment plus de lactase dès l'âge de cinq ans, tandis que l'existence de contenants préhistoriques semblant servir à nourrir des enfants (sortes de biberons) et ayant des traces de lait de ruminants atteste bien du fait que les jeunes enfants du néolithique et de l'âge du Bronze consommaient du lait animal. A cet âge, l'écart calorique dû à l'absorption de glucose entre individus exprimant ou non la lactase est considérable et représente 12% des apports journaliers recommandés des enfants (versus 6.7% chez des adultes). De plus, la jeune enfance représente une période cruciale pour la croissance et l'exposition aux pathogènes, augmentant ainsi encore plus la pression de sélection sur la quantité et la qualité de leurs apports nutritionnels.

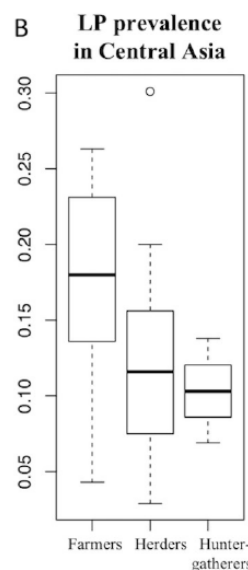
Une autre zone d'ombre de cette histoire évolutive réside dans l'absence de corrélation entre l'intensité du pastoralisme (la dépendance des populations aux ressources animales) et la fréquence de PL en Eurasie, tandis qu'en Afrique, la corrélation est plus forte et significative.



Corrélation entre pastoralisme et fréquence de PL en Afrique et en Eurasie. Ségurel et al. (2017)

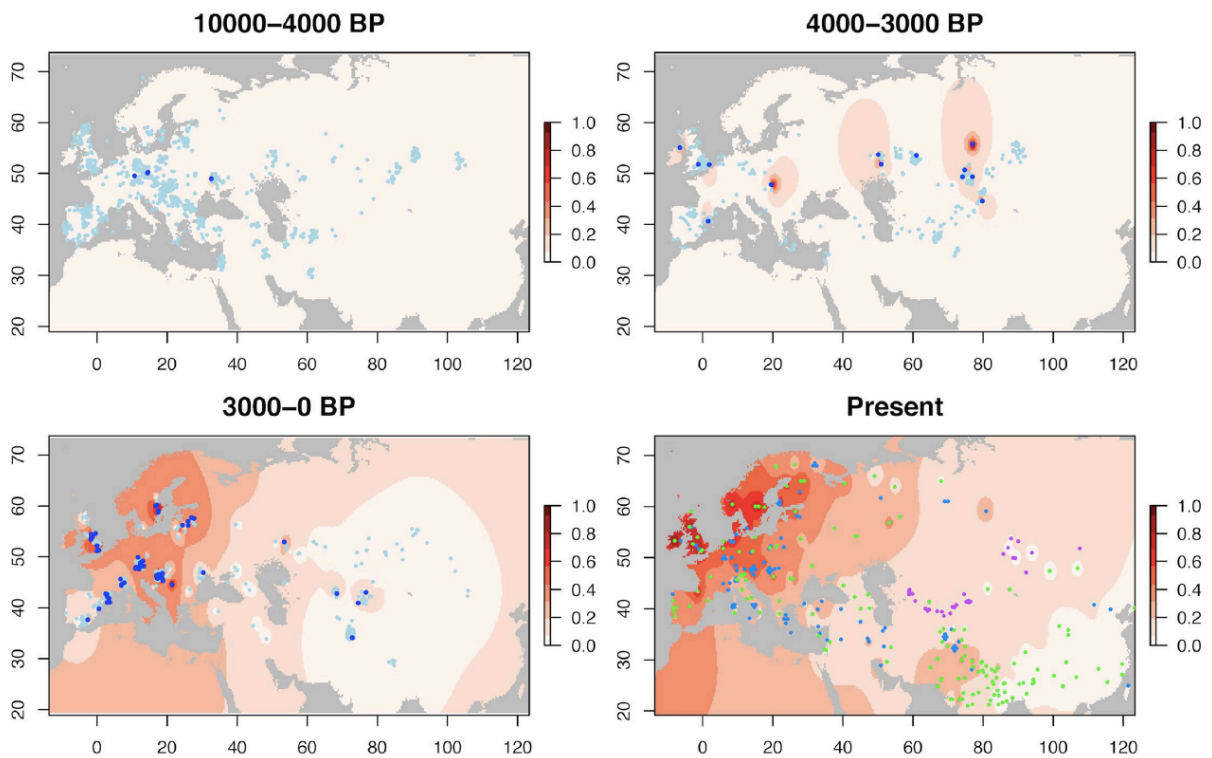
En effet, les fréquences les plus élevées de PL sont observées en Europe du Nord dans des populations plus agricoles que pastorales, tandis que les populations d'éleveurs des steppes d'Asie Centrale, fortement pastorales, semblent n'avoir que des fréquences modérées ou faibles (Ségurel et al. 2017). Pourquoi une telle contradiction ? Pour mieux décrire la prévalence de PL dans les populations des steppes d'Asie, une région assez peu étudiée, nous avons commencé par compiler un large jeu de données génétiques obtenues au laboratoire à partir de diverses sources (RFLP, PCR, puce à ADN), provenant de 963 individus correspondant à 13 groupes ethniques différents en Asie Centrale et en Asie du Nord. Nous avons ainsi obtenu le génotype de la mutation -13.910:C/T et inféré le phénotype de PL pour chaque individu. En effet, malgré le caractère convergent de la PL et sa nature polygénique dans les populations africaines, le phénotype de PL semble monogénique et fortement pénétrant à travers toute l'Eurasie (Enattah et al. 2007; Heyer et al. 2011). Nous avons ainsi confirmé que les fréquences de PL sont faibles dans les steppes asiatiques (autour de 10-20%), avec des fréquences même légèrement plus élevées dans les populations d'agriculteurs (17.5%) par rapport aux éleveurs (12.2%) et aux chasseurs-cueilleurs (10.3%). Ce même constat a d'ailleurs été obtenu dans d'autres populations, en Iran (Charati et al. 2019). Pourquoi une telle absence de PL en Asie, où les éleveurs des steppes sont pourtant très dépendants du pastoralisme et consomment du lait de cheval depuis au moins 5 500 ans (Outram et al. 2009)?

Ethnic group	Ancestral subsistence mode	Location	LP prevalence	N
Kazakh	Herders	NE + SW	0.208	159
Kyrgyz	Herders	NE + SW	0.119	201
Karakalpak	Herders	SW	0.067	45
Buryat	Herders	NE	0.179	28
Telengit	Herders	NE	0.034	29
Altai Kizhi	Herders	NE	0.121	33
Mongol	Herders	NE	0.125	32
Uzbek	Unclear*	SW	0.244	45
Turkmen	Farmers [#]	SW	0.200	50
Tajik	Farmers	SW	0.177	254
Khakas	Hunter-gatherers	NE	0.138	29
Shors	Hunter-gatherers	NE	0.103	29
Tubalar	Hunter-gatherers	NE	0.069	29



Prévalence de persistance à la lactase dans des populations ayant différents modes de vie ancestraux en Asie Centrale. Ségurel et al. (2020).

Une caractéristique intéressante des steppes asiatiques est que cette région a été le théâtre de nombreux mouvements et mélanges de populations, tout particulièrement durant l'Age du Bronze (environ 5 000-3 000 BP) et l'Age du Fer (environ 3 000-2 000 BP) (Allentoft et al. 2015; Haak et al. 2015). Il est donc possible que des remplacements de populations aient pu brouiller les signatures de sélection. Afin de mieux comprendre la dynamique évolutive de la PL, à la fois au niveau temporel et au niveau spatial, nous avons décidé de compiler les données existantes d'ADN ancien, au vu de l'accumulation récente et spectaculaire de données génétiques dans des populations humaines du passé. En effet, nous avons pu extraire le génotype de la mutation -13.910:C/T pour 1 434 individus anciens d'Eurasie (où la LP est monogénique) datant d'il y a 10 000 ans à aujourd'hui (en se basant principalement sur une base de données déjà existante (Reich 2019) que nous avons actualisée). Nous avons ainsi pu représenter l'évolution de la fréquence de PL à travers le temps en Eurasie.



Les trois premières cartes représentent la fréquence extrapolée ancienne de PL (dégradé de rouge en fond), avec les individus porteurs d'allèle C/T représentés par des points en bleu clair/foncé, à différentes périodes anciennes. La quatrième carte représente la fréquence extrapolée moderne de PL (dégradé de rouge en fond), avec les individus représentés par des points de différentes couleurs selon la source des données (en violet celles produites par notre laboratoire). Ségurel et al. (2020).

Nous notons ainsi que le premier individu à être génétiquement lactase-persistant est un individu vivant en Ukraine il y a 5 960 ans, comme précédemment observé (Mathieson et al. 2018), suivi de quelques autres individus Européens. La fréquence de l'allèle -13.910:T reste cependant faible en Europe entre 5 000 et 4 000 BP (1.2%), tandis qu'elle est nulle en Asie. L'allèle -13.910:T émerge ensuite rapidement dans une large zone d'Eurasie allant du Royaume-Uni au Kazakhstan, et atteint une fréquence de 4.6% entre 4 000 et 3 000 BP (sans

différence significative entre l'Ouest et l'Est). Enfin, durant l'Age du Fer, l'allèle -13.910:T augmente en fréquence en Europe (30.8%) tandis qu'elle stagne en Asie (5.6%), un patron semblable à celui observé actuellement. Le fait que la mutation soit déjà présente au Kazakhstan à l'Age du Bronze nous indique que, s'il y avait eu une forte pression de sélection pour la PL dans la région (comme en Europe), il y aurait eu assez de temps pour que la mutation atteigne une prévalence plus élevée (d'au moins 50%). Il semble donc que la PL n'a pas été fortement sélectionnée dans les populations pastorales des steppes de l'Age du Bronze. Pourquoi ?

Une spécificité des éleveurs actuels d'Asie Centrale est qu'ils boivent traditionnellement du lait de jument, un lait très riche en lactose (et donc qui devrait créer encore plus de pression pour être lactase-persistant), mais qui, également en conséquence, fermente spontanément. Ces éleveurs ne consomment en effet jamais de lait cru, mais l'incorporent en petites quantités chauffé dans le lait, et surtout le consomment en grande quantité sous forme transformée: yaourt, beurre, crème, fromage et lait fermenté (Ruhlmann 2019). Ces produits transformés contiennent plus ou moins de lactose, mais ont la caractéristique de contenir pour la plupart des bactéries lactiques vivantes. Or l'ingestion simultanée de lactose et de bactéries lactiques permet de réduire considérablement les symptômes associés à la consommation de lactose (Hertzler et al. 2003). Il semble donc que la transformation du lait représente une adaptation culturelle et microbiologique qui ait permis à ces populations de dépendre des produits laitiers sans adaptation génétique. La question devient alors : pourquoi certaines populations ont sélectionné génétiquement aussi fortement la PL, sachant que la transformation du lait en fromage a été attestée très tôt, depuis au moins 9,000 ans, en Europe (Salque et al. 2013) ?

Un des éléments de réponse repose certainement sur les espèces laitières exploitées par les populations : en Europe du Nord, en Afrique et en Arabie où la vache ou le chameau (et le lait frais) tient une place centrale, la PL a été fortement sélectionnée. A l'inverse, en Europe du Sud, au Proche-Orient ou en Asie Centrale où chèvres, moutons, et/ou chevaux sont majoritaires (et le lait est principalement dérivé en fromages et/ou yaourts), la PL reste minoritaire. Il est donc possible que les propriétés des laits de différents animaux soit associées à des usages différents. Cependant, peu d'études d'anthropologie comparative permettent de comparer formellement et quantitativement les pratiques associées aux produits laitiers entre animaux, populations et continents.

Afin de continuer à explorer ces questions et formellement tester ces hypothèses, j'ai mis en place en 2019 un nouveau projet en Mongolie à l'interface entre génétique humaine, génétique microbienne et anthropologie évolutive, en collaboration avec Sandrine Ruhlmann (ethnologue du CNRS spécialiste des techniques en Mongolie) et l'équipe de microbiologie de Battogtokh Chimeddorj à la MNUMS (Mongolian National University of Medical Sciences) en Mongolie. J'ai ainsi obtenu deux financements en 2020 : une bourse IEA du CNRS pour financer des échanges et formaliser la coopération entre pays et une bourse ATM du Muséum pour effectuer les missions de terrain et les analyses nécessaires. Les objectifs principaux étaient i) de quantifier la prévalence de PL en Mongolie (étant donné que la seule étude phénotypique est assez ancienne, Yongfa et al. 1984), ii) de quantifier la consommation de

produits laitiers en prenant en compte les différents usages de transformation et iii) d'étudier l'influence du microbiote intestinal sur les symptômes ressentis associés à la consommation de lactose. Il n'y a en effet à ce jour aucune étude considérant simultanément le phénotype de PL et le microbiote intestinal. J'ai organisé une première mission préliminaire en Mongolie avec Sandrine Ruhlmann en juillet 2019 afin d'établir un contact avec la MNUMS, de demander les permis éthiques nécessaires et de récolter certaines données anthropologiques préliminaires. Cependant, le terrain prévu en avril 2020 a dû être annulé du fait de la COVID-19 et je n'ai par la suite pas relancé ce projet pour lequel nous avons dû rendre les financements obtenus. Suite à mon déménagement en 2020 au LBBE à Lyon, j'ai en effet préféré me recentrer sur des projets impliquant des membres du LBBE mais aussi minimisant l'impact écologique de ma recherche.

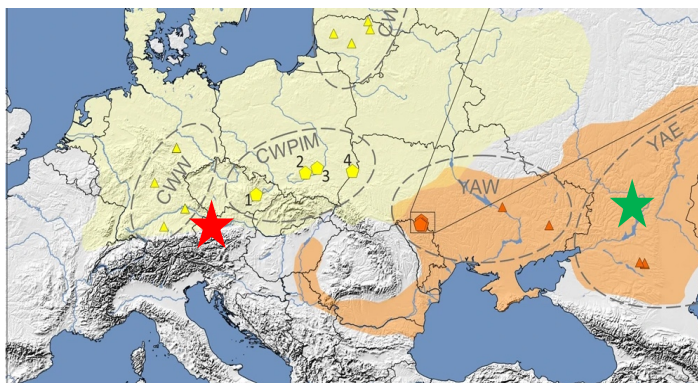
En parallèle, en 2021, j'ai été contactée par Fernando Racimo (Université de Copenhague) et son étudiante en thèse, Rasa Muktopavela, car ils étaient en train de développer une nouvelle méthode pour caractériser plus finement dans le temps et l'espace des événements de sélection positive en combinant des données d'ADN ancien et moderne. Nous avons ainsi collaboré pour appliquer cette méthode sur les données de la lactase, non seulement pour mieux cerner les forces et les limites de leur méthode, mais également pour voir ce qu'on pouvait apprendre de plus sur l'évolution de la lactase (Muktopavela et al. 2020).

En effet, jusqu'à maintenant, il n'existait pas de cadre formel pour analyser des génomes récents et anciens de manière continue à la fois dans le temps et dans l'espace. A l'image de l'analyse effectuée dans notre papier de 2020, nous étions contraints de regrouper les échantillons ensemble, à la fois spatialement et temporellement, en espérant que nous capturions bien l'hétérogénéité sous-jacente d'intérêt. Ainsi, pour le papier de 2020, nous avons considéré l'Europe et l'Asie de l'Ouest comme deux zones distinctes, et nous avons séparé les échantillons anciens en trois pas de temps (10 000-4 000 BP ; 4 000-3 000 BP ; 3 000 BP-présent) en fonction des connaissances actuelles. L'intérêt de la méthode nouvellement développée est de pouvoir traiter les échantillons individuellement sans les regrouper, et d'inférer a posteriori les dynamiques temporelles et spatiales de la sélection naturelle, en estimant, pour chaque locus sous sélection : i) le coefficient de sélection, ii) le coefficient de diffusion (quantifiant dans quelle mesure l'allèle sous sélection se diffuse en continu dans l'espace à partir de son point d'origine), et iii) le coefficient d'advection (quantifiant dans quelle mesure le point d'origine de l'allèle change dans l'espace, correspondant à la migration des individus le portant). Cette méthode s'appuie sur la modélisation de l'évolution des fréquences alléliques dans un continuum spatio-temporel et permet de prendre en compte la vraisemblance des génotypes. Ce deuxième point est particulièrement pertinent dans le contexte de l'ADN ancien : en effet, comme les génomes anciens sont souvent mal couverts, du fait de la faible quantité d'ADN endogène, la majorité des polymorphismes n'ont pas assez de lectures de séquençage pour pouvoir déterminer s'ils sont homozygotes ou hétérozygotes. En conséquence, les données génétiques sont fréquemment pseudo-haploïdisées, c'est-à-dire qu'un seul allèle est pioché au hasard pour constituer le génotype, ce qui entraîne une importante perte d'information. Le fait de prendre en compte la vraisemblance des différents génotypes permet

a contrario de garder toute l'information présente dans les données, et donc de mieux estimer les variations en fréquences alléliques dans le temps et l'espace.

Les simulations ont montré que la méthode est robuste pour l'estimation des coefficients de sélection, de l'origine géographique de la mutation et des paramètres de diffusion en deux dimensions. Par contre, en présence d'advection (i.e., migration des populations), l'inférence de l'origine géographique de l'allèle sous sélection devient moins précise. Sans surprise, dans ce cas, la densité de l'échantillonnage est cruciale pour bien identifier la dynamique ancienne de la mutation.

Concernant la lactase, l'origine géographique de l'allèle sous sélection a été inférée comme étant plus à l'Est (Nord Caucase) que celle inférée par les données d'ADN moderne seules (Autriche) (Itan et al. 2009) (voir ci-dessous), récapitulant bien les résultats obtenus par l'approche plus grossière des analyses précédentes d'ADN ancien (Ségurel et al. 2020). Ce point (qui peut paraître du détail) est en réalité important pour le débat sur l'origine de la PL : certains chercheurs ont en effet émis l'hypothèse que la PL a émergé dans les populations agropastorales d'Europe Centrale (la culture LBK, Linearbandkeramik) (Itan et al. 2009; Key et al. 2016), tandis que d'autres ont proposé que la PL vienne des populations pastorales des steppes (les Yamnaya) qui ont par la suite migré vers l'Europe (Allentoft et al. 2015; Mathieson et al. 2015). Ce dernier résultat conforte donc plutôt la deuxième hypothèse.



Aire de répartition de la culture Corded Ware (en jaune) et Yamnaya (en orange). L'étoile rouge montre le point d'origine de la PL selon les données d'ADN moderne (Itan et al. 2009), et l'étoile verte montre celui estimé par nos analyses (modèle basé sur les données d'ADN ancien). Figure modifiée d'après Juras et al. (2018).

Concernant les coefficients de sélection, de diffusion et d'advection, nous les avons estimés avant et après 5 000 ans, du fait des changements de dynamique des populations au moment de l'âge de Bronze. La méthode estime des coefficients de sélection très forts ($s=0.10$ avant 5 000 ans et $s=0.03$ après, en considérant que la mutation date d'il y a environ 7 500 ans), en accord avec les données de la littérature (entre 0.01 et 0.19) (Ségurel et al. 2017). Ce n'est pas clair si la différence entre les deux périodes reflète une réalité biologique en terme de changement de contraintes sélectives, ou si elle reflète une moins bonne estimation des coefficients de sélection anciens qui i) sont basés sur un plus petit nombre d'individus, et ii) sont plus sensibles à la date inférée d'apparition de la mutation (si l'on considère une apparition de la mutation il y a environ 20 000 ans, les coefficients de sélection avant/après 5 000 ans deviennent $s=0.029$ versus $s=0.026$).

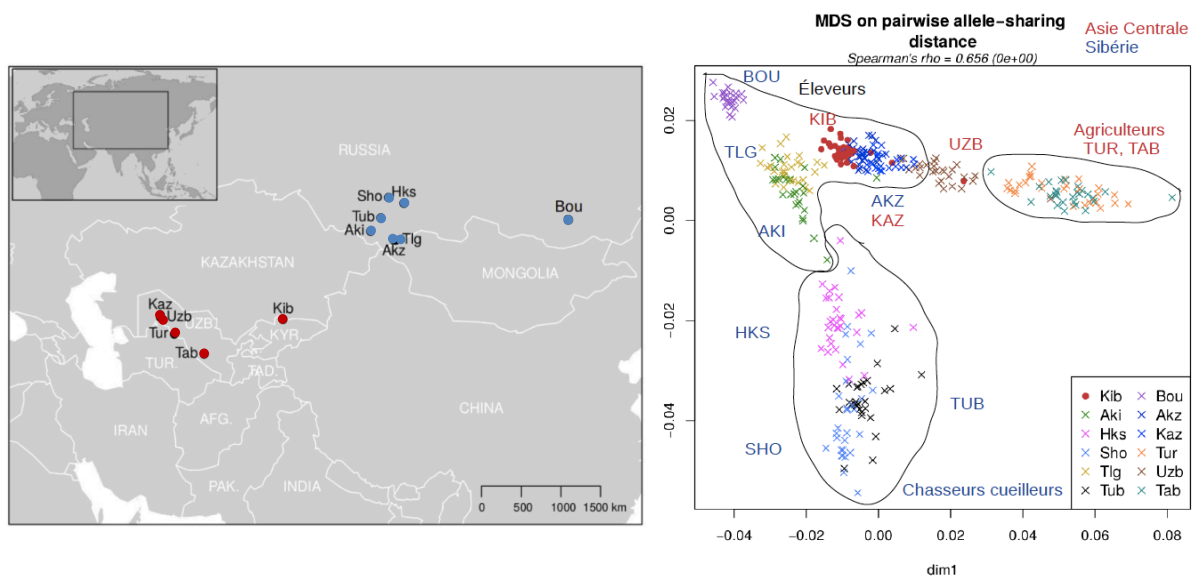
Les coefficients de diffusion sont, eux, plus importants après 5 000 ans, attestant d'une expansion rapide, avec des coefficients d'advection montrant un net déplacement de l'allèle vers l'Ouest. Ceci corrobore à la fois l'absence de pressions sélectives en Asie et les mouvements massifs des populations des steppes vers l'ouest de l'Europe aux alentours de 5.000 BP. Même si l'origine de la PL n'est pas résolue, les peuples Yamnaya semblent clairement avoir contribué à la diffusion rapide de cet allèle à travers toute l'Europe et même au-delà (notamment en Inde).

3. Adaptation polygénique en Asie intérieure

En dehors de ces études ciblées sur un gène candidat ou un phénotype candidat, je me suis également intéressée à la détection de sélection « à l'aveugle » à l'échelle du *génom*e (scan génomique). En effet, comme introduit plus tôt, et à l'instar de la persistance de la lactase, la plupart des cas connus d'adaptation des populations à leur environnement concernent des traits phénotypiques monogéniques (encodés par un seul gène). Ceci est cependant le résultat de biais méthodologiques, puisque ceux-ci laissent des signatures génétiques plus marquées et sont donc plus faciles à détecter que les cas polygéniques (voir Boîte 1). Ainsi, par simplicité, la grande majorité des tests statistiques de sélection naturelle qui ont été développés font l'hypothèse que la sélection agit sur un seul polymorphisme génétique. Cependant, la plupart des caractères, notamment ceux liés à des fonctions métaboliques ou immunologiques, sont polygéniques : ils sont codés par plusieurs gènes ayant des effets plus ou moins forts sur le caractère (Boyle et al. 2017). Ces gènes peuvent de plus interagir de manière complexe, par exemple dans le cadre de réseaux de gènes. Il semble d'ailleurs que ce mode de sélection soit plus prévalent que la sélection monogénique, chez l'humain comme chez d'autres espèces (Pritchard et al. 2010).

La thèse de Léa Bouteille, démarrée en 2017 en co-encadrement avec Frédéric Austerlitz (UMR7206) et Marianne Elias (UMR7205), avait justement pour objectif de s'intéresser à ce type de sélection encore assez mal compris, en combinant des approches méthodologiques (simulations et tests de méthodes supervisés par F Austerlitz) et de l'analyse de données, avec une partie supervisée par M Elias sur des données de papillons (le long d'un gradient altitudinal) et une partie supervisée par moi-même sur des données humaines ayant des modes de subsistance contrastés (chasseurs-cueilleurs, agriculteurs et éleveurs).

En ce qui concerne la partie humaine, l'objectif était d'appliquer des tests de sélection sur un large jeu de données génomiques obtenu au fil des années par l'UMR 7206, plus précisément des données de puce à ADN obtenus chez 336 individus venant de 12 populations différentes d'Asie Centrale et d'Asie du Nord, une région centrale en terme de peuplement humain mais assez mal représentée dans les bases de données publiques existantes. En effet, la plupart des études précédentes se sont basées sur des jeux de données très larges en terme de couverture génomique, mais très incomplets en terme géographique. Ainsi, en Eurasie, nos connaissances viennent presque exclusivement de populations d'Europe de l'Ouest et d'Asie de l'Est (Chine, Japon), faute d'accessibilité (Frazer et al. 2007; 1000 Genomes Project Consortium et al. 2015).



Carte montrant la localisation géographique des 12 populations étudiées en Asie Centrale / du Nord, et MDS des distances génétiques à l'échelle du génome montrant un regroupement des individus par mode de subsistance (3 groupes : agriculteurs, éleveurs et chasseurs-cueilleurs) et par géographique (Asie Centrale en rouge, Asie du Nord en bleu) (© Léa Bouteille)

La délimitation géographique des événements adaptatifs précédemment identifiés dans la littérature n'est donc pas claire, et il reste très certainement de nouveaux cas d'adaptation à identifier entre ces deux zones géographiques assez éloignées. En effet, les modes de vie et les conditions climatiques de cette région sont très diversifiés, donc propices à l'étude de la sélection locale divergente sur des traits complexes. De précédentes études ont notamment montré des signaux de sélection en Sibérie sur des gènes liés au métabolisme, bien que l'on ne sache pas si cela correspond à de l'adaptation au froid ou à une alimentation riche en protéines d'origine marine (Cardona et al. 2014; Hallmark et al. 2019).

Léa a commencé par tester la puissance et la spécificité de différents tests de sélection existants basés sur la différenciation entre populations et cherchant à détecter de la sélection divergente monogénique (Fdist, Bayecan, PCAdapt, OutFLANK). Pour ce faire, elle a généré des données simulées dans un contexte polygénique de sélection stabilisatrice ou directionnelle, et a montré que PCAdapt et OutFLANK étaient plus performants dans ces conditions.

Elle a alors appliqué ces deux méthodes sur les données humaines du labo (puce Omni1 passée sur 12 populations, 336 individus, plus de 800,000 SNPs après contrôle qualité). L'avantage de ces méthodes est qu'elles cherchent des outliers de différenciation en prenant en compte l'ensemble des populations et ne nécessite donc pas de définir des groupes à contraster a priori. En se focalisant sur les signaux en commun entre les deux approches utilisées et en cherchant des outliers sur des fenêtres glissantes de 50kb ou 100kb, elle a trouvé cinq régions significatives à l'échelle du génome : quatre différenciant certains chasseurs-cueilleurs des autres populations (agriculteurs, éleveurs), et une différenciant l'Asie Centrale de l'Asie du Nord. Léa a ensuite regardé s'il y avait des eQTLs dans les régions identifiées et a trouvé certains variants associés à la tension artérielle, la répartition de la graisse corporelle, ou encore des comportements addictifs.

En parallèle, Léa a regardé spécifiquement les trois SNPs identifiés par Hallmark et al. (2019) comme étant sous sélection en Sibérie (dans les gènes *PLA2G2A*, *PLINI* et *ANGPTL8*, respectivement), pour savoir si nous pouvions distinguer un effet « climat » d'un effet « alimentation ». Aucun de ces SNPs n'a été trouvé comme significatif par PCAdapt ou OutFLANK dans notre jeu de données, mais les fréquences observées montrent : i) pour *PLA2G2A*, une fréquence plus élevée chez les chasseurs-cueilleurs (19%) par rapport aux agriculteurs et éleveurs (8%) (entraînant une légère différence entre Asie du Nord et Asie Centrale) ; ii) pour *PLINI*, pas de différence entre modes de subsistance mais une fréquence plus forte en Asie du Nord (18%) par rapport à l'Asie Centrale (10%) ; et pour *ANGPTL8*, une fréquence plus élevée chez les chasseurs-cueilleurs et éleveurs (27%) par rapport aux agriculteurs (13%), entraînant une fréquence plus importante en Asie du Nord par rapport à l'Asie Centrale. Les mécanismes à l'œuvre ne semblent donc pas être les mêmes pour ces trois signaux adaptatifs, mais grâce aux données analysées ici, nous pouvons proposer que les signaux sur *PLA2G2A* et *ANGPTL8* soient plutôt dus à des phénotypes métaboliques liés au mode de subsistance, tandis que le signal sur *PLINI* pourrait être associé à une variable climatique.

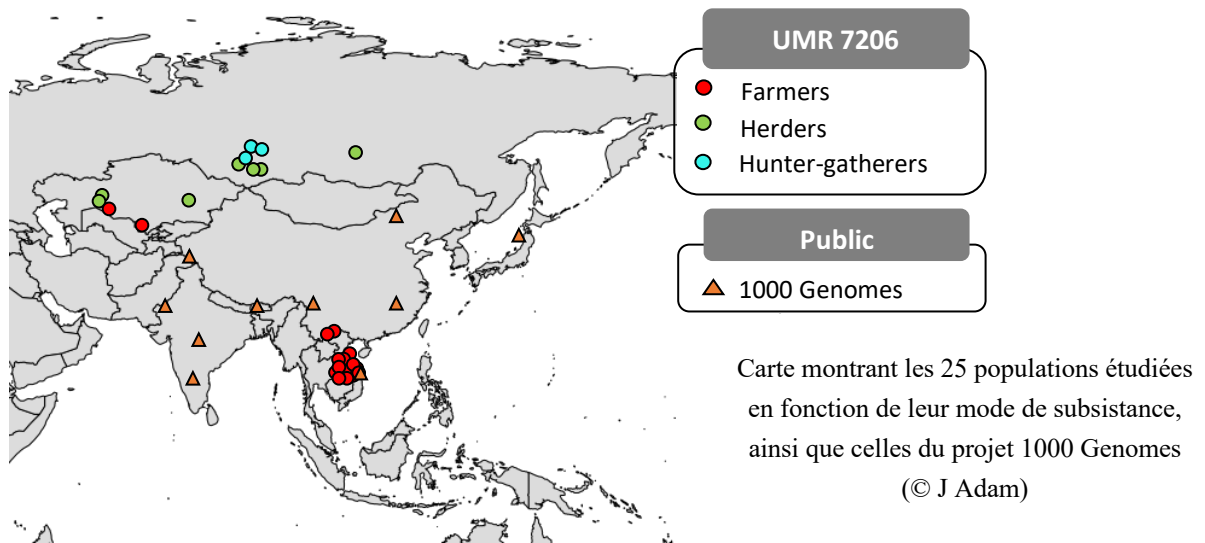
Les analyses en cours ont cependant été arrêtées début 2020 du fait que Léa a attrapé la COVID-19 puis s'est mise en arrêt maladie jusqu'en mai 2021 suite à une dépression, après quoi elle a finalement décidé d'arrêter sa thèse. Ses travaux sont donc restés en suspens le temps de savoir si elle allait les reprendre, et n'ont malheureusement pas été valorisés.

4. Travail en cours

Adaptation polygénique en Asie (thèse de J. Adam 2022-2025)

En 2022, cette thématique de recherche a été reprise par Johanne Adam, qui a obtenu une bourse de thèse auprès du MNHN en co-direction avec Evelyne Heyer, Raphaëlle Chaix et moi-même, pour continuer à explorer ces questions d'adaptation polygénique en Asie, en rajoutant un ensemble de données venant du laboratoire pour des populations d'Asie du Sud-Est (Cambodge et Laos). Son projet doctoral s'appuie donc au final sur l'analyse de plus de 800 individus venant de 25 populations réparties dans 5 pays, et a pour objectif de mieux comprendre les contraintes écologiques et culturelles qui ont façonné l'évolution génétique du continent eurasiatique ces derniers millénaires.

Plus précisément, elle cherche à comprendre : i) quelles sont les régions du génome et les phénotypes sous sélection monogénique et polygénique en Asie, ii) si ces signaux sont principalement dus à des contraintes climatiques, alimentaires, ou pathogéniques, et iii) quelle est la dynamique temporelle et spatiale des événements sélectifs.



Les données génétiques utilisées pour la thèse de Johanne consistent en trois puces de génotypage : une puce Omni1 (1 million de SNPs, celle précédemment utilisée par Léa) sur des populations d'Asie Centrale et d'Asie du Nord, une puce Omni1 et une puce Omni2.5 (2.5 millions de SNPs) sur des populations d'Asie du Sud-Est collectées par R. Chaix. Ces populations ont été précédemment étudiées au laboratoire en terme d'histoire démographique (Palstra et al. 2015; Marchi et al. 2017, 2018; Ly et al. 2018, 2019), mais pas en terme d'histoire adaptative. Après avoir effectué les contrôles qualité et enlevé les individus apparentés, il reste, pour chacune des puces, respectivement : 756k SNPs pour 325 individus (Asie Centrale / du Nord, Omni 1), 741k SNPs pour 382 individus (Asie du Sud-Est, Omni 1), et 1.2M SNPs pour 155 individus (Asie du Sud-Est, Omni 2.5).

Dans un premier temps, Johanne s'est focalisée sur des tests de sélection intra-population basés sur la diversité haplotypique afin de ne pas avoir à fusionner les trois puces (ce qui entraîne une forte diminution du nombre final de SNPs). Elle a d'abord phasé les données en prenant un panel de référence d'haplotypes mondiaux venant du projet 1000 Génomes (Delaneau et al. 2014), étant donné qu'un large panel de populations donne de meilleurs résultats qu'un panel trop restreint (Delaneau et al. 2013). Elle a utilisé une carte de recombinaison soit de LD (1000 Genomes Project Consortium et al. 2015), soit de pedigree (Halldorsson et al. 2019), pour voir l'influence du phasage sur les signaux obtenus. Elle a ensuite obtenu des scores d'iHS pour chaque population (et deux conditions de phasage) avec le package selscan (Szpiech et al. 2014).

Johanne a d'abord observé que les scores d'iHS sont bien corrélés entre les SNPs phasés avec la carte LD versus la carte de pedigree (0.969 en moyenne). Elle a ensuite obtenu les régions « outliers » du génome, définies comme des fenêtres de 100kb ayant un nombre de SNPs significatifs (ayant un iHS dans le top 1%) dans le top 1% des fenêtres, et les a comparées entre populations.

Parmi les 55.366 fenêtres analysées, Johanne en a identifié quatre significatives étant partagées par un grand nombre de populations (17 ou plus sur les 25), dont une région sur le chromosome 5 partagée par 19 populations, qui contient un long ARN non-codant (*LINC02147*), et une région sur le chromosome 2 partagée par 19 populations, qui contient le gène *ERBB4* et qui a déjà été montré comme étant sous sélection positive chez toutes les populations humaines non-africaines, et plus particulièrement en Asie de l'Est (Pickrell et al. 2009).

Elle a ensuite, parmi les 1212 fenêtres significatives chez au moins deux populations, comparé le nombre de SNPs significatifs par fenêtre entre populations du Nord et du Sud (test de rang, 12 versus 13 populations), ainsi qu'entre populations de chasseurs-cueilleurs, d'agriculteurs et d'éleveurs au sein de l'Asie Centrale (tests de rang contrastant un mode de vie aux deux autres, avec respectivement 3, 3 et 6 populations). Les résultats montrent qu'aucune fenêtre n'est significativement associée au mode de subsistance (possiblement à cause d'un manque de puissance du au faible nombre de populations dans chaque catégorie). Par contre, 71 fenêtres sont significativement associées à la latitude (de manière catégorielle, Nord versus Sud), dont 41 correspondent à de la sélection au Nord et 30 à de la sélection au Sud. Une analyse d'enrichissement fonctionnel est en cours sur ces deux ensembles de SNPs.

En parallèle, Johanne a fusionné les SNPs des trois puces pour pouvoir effectuer des tests de sélection inter-population basés sur le Fst. La même approche que précédemment a été appliquée, pour identifier des fenêtres de 200kb ayant un nombre de SNPs significatifs (ayant un Fst moyen des paires de populations Nord versus Sud dans le top 1%) dans le top 1% des fenêtres. La fenêtre utilisée est ici plus grande car la densité des SNPs est plus faible (329k SNPs au total). Elle a ainsi trouvé 144 régions ayant des Fst Nord/Sud significatifs.

Parmi toutes ces régions candidates, 7 régions présentent à la fois un iHS significativement différent entre Nord et Sud et un Fst significatif entre Nord et Sud et correspondent donc aux signaux les plus forts. Johanne est en train d'analyser ces régions pour identifier leur propriétés fonctionnelles (présence de gènes, de marques régulatrices, d'eQTLs) et essayer de comprendre quelles pourraient être les pressions de sélections associées.

En parallèle, l'idée est d'ajouter des données publiques (1000G, HGDP) sur les fréquences alléliques de ces SNPs candidats dans des populations plus diverses géographiquement, afin de tester si les pressions sélectives identifiées sont bien compatibles avec un effet de la latitude, ou si elles seraient plutôt dues à d'autres facteurs confondants.

Les régions candidates obtenues pourront également être analysées plus précisément en terme de dynamique temporelle, en utilisant la base de données d'ADN ancien AADR (Reich 2019) régulièrement actualisée et qui combine les génotypes de plus de 9 900 individus datés de 60 000 ans à aujourd'hui (avec une majorité d'individus venant d'Europe ou d'Asie, et datant d'il y a 6 000 ans à il y a 500 ans). Nous pourrions également regarder si les régions candidates sont issues de métissage avec des espèces archaïques (introgression adaptative),

particulièrement étant donné que Neandertal et de Denisova étaient présents en Asie pendant la préhistoire et que plusieurs génomes anciens sont disponibles pour ces espèces.

En plus de ce scan de sélection à l'échelle du génome, nous souhaitons également étudier de plus près certaines histoires adaptatives d'intérêt. Notamment, une étude précédente a montré l'existence d'un signal adaptatif témoignant d'une ancienne épidémie virale dans certaines populations d'Asie de l'Est (Chine et Vietnam, et dans une moindre mesure, Japon), datant d'il y a environ 25.000 ans et impliquant les gènes d'interaction avec le coronavirus actuel (Souilmi et al. 2021). Nous nous attacherons à définir plus précisément dans quelles populations asiatiques ce signal est détectable, et à caractériser plus finement sa dynamique spatiale et temporelle grâce à notre large jeu de données.

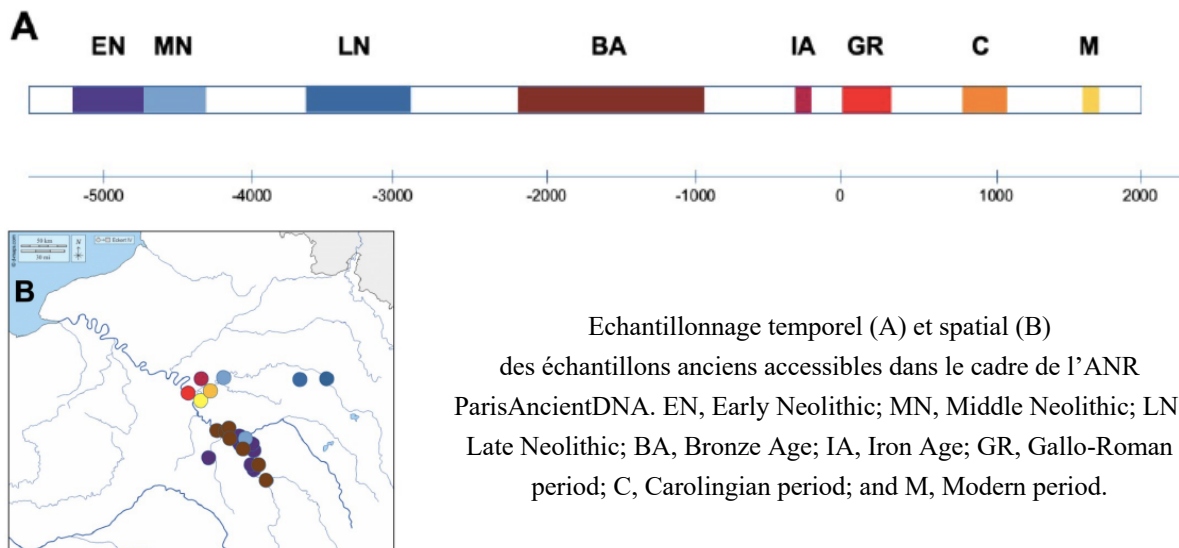
L'objectif suivant sera d'identifier des signatures de sélection polygénique. Pour cela, Johanne va regrouper les polymorphismes en fonction de leur annotation fonctionnelle afin d'obtenir des scores de sélection pour des groupes de gènes associés à des traits phénotypiques d'intérêt (paramètres métaboliques, sanguins, immunité...). Il nous faudra donc récupérer la liste des polymorphismes associés à ces traits par des GWAS (Genome-Wide Association Studies) faites en Asie, puis combiner les scores de sélection par trait et tester leur significativité.

Evolution des populations du bassin parisien ces sept derniers millénaires

En parallèle de ces projets, je participe à l'ANR ParisAncientDNA (2021-2025) portée par Marie-Claude Kergoat (UMR 7206), qui a pour objectif de caractériser la diversité génétique de restes humains du bassin parisien entre le Néolithique et la période moderne. En effet, de nombreuses études de paléogénomique ont vu le jour en Europe du Nord et du Sud notamment en Angleterre, Irlande ou Espagne (Cassidy et al. 2016; Mitnik et al. 2018; Antonio et al. 2019; Olalde et al. 2019) et ont permis de caractériser les mouvements des populations du passé et de mieux comprendre comment les grandes transitions culturelles se sont effectuées. Cependant, peu de données sont disponibles pour la France, qui occupe pourtant une place centrale à la croisée de différentes routes migratoires.

Le premier objectif de ce projet est d'obtenir des données génomiques anciennes de bonne qualité (couverture d'au moins 0.5X) pour 15-20 individus correspondant à huit périodes archéologiques différentes (Néolithique précoce/moyen/tardif, Age de Bronze, Age du Fer, période gallo-romaine, période carolingienne, et 17^{ème} siècle).

Grâce à ces données, les objectifs sont ensuite i) d'étudier la structure sociale et les relations entre caractéristiques archéologiques et génétiques pour chaque site ; ii) de caractériser l'histoire démographique et migratoire des individus à travers le temps et notamment questionner la continuité génétique de ces populations ; et enfin iii) de détecter des signatures de sélection naturelle positive le long de ce transect temporel, pour lequel nous savons que les contraintes alimentaires et pathogéniques ont été très fortes et variables dans le temps et l'espace.



Cette dernière partie, que je supervise, ne pourra commencer que quand l'histoire démographique de cette région aura bien été inférée. En effet, pour pouvoir construire au mieux l'approche méthodologique et formuler les hypothèses précises à tester, il nous faut d'abord caractériser précisément les dynamiques migratoires et de métissage des populations. D'une manière générale, l'idée est de chercher des endroits du génome où la variation en fréquence allélique au cours du temps diverge du reste du génome. De manière intéressante, il a été montré que, chez l'humain, l'étude des populations anciennes permet d'identifier un plus grand nombre de balayages sélectifs que l'étude des populations modernes, car les populations actuelles sont le fruit d'un important métissage datant environ de l'âge de Bronze, qui a dilué les signaux sélectifs (Souilmi et al. 2022).

De nombreuses études ont commencé à exploiter des données d'ADN ancien pour avoir accès plus directement aux événements adaptatifs passés (Dehasque et al. 2020). Cependant, la plupart de ces études contrastent les fréquences actuelles avec un seul point dans le temps ou se focalisent sur un polymorphisme donné (Kerner et al. 2021). De notre côté, nous tirerons avantage du fait que nous avons un grand nombre de points temporels. Cela nous permettra notamment de chercher des régions du génome où les pressions sélectives n'ont pas été constantes dans le temps, et ainsi d'identifier des cas où les contraintes sélectives ont apparu et/ou disparu au cours du temps, mais aussi des cas de sélection fluctuante (c'est-à-dire où le coefficient de sélection change au cours du temps, soit parce qu'il y a des phénomènes de fréquence-dépendance, soit du fait de dynamiques coévolutives). Nous serons donc, à notre connaissance, la première étude à faire une étude de sélection basée sur la différenciation génétique à l'échelle du génome en utilisant un grand nombre de périodes temporelles.

Partie 2- Détection de sélection balancée chez les primates

Comme introduit plus tôt (Boite 1), les signatures génomiques de la sélection balancée sont assez difficiles à détecter, non seulement car la sélection balancée recouvre plusieurs mécanismes évolutifs produisant des attendus partiellement différents, mais également car la diversité génétique neutre fluctue le long du génome, de par la nature stochastique des événements de coalescence. En conséquence, le rôle de la sélection balancée en tant que moteur de la diversité génétique et phénotypique des espèces, sa prévalence globale ainsi que l'importance relative des différents mécanismes sous-jacents restent largement méconnues (Bubb et al. 2006; Charlesworth 2006; Fijarczyk et al. 2015; Bitarello et al. 2023).

Certains cas de sélection balancée ont cependant été bien décrits : le maintien chez l'humain du polymorphisme du gène *Hbb* en réponse à la malaria par avantage à l'hétérozygote (avec les hétérozygotes sS étant avantagés face aux homozygotes ss souffrant d'anémie falciforme et aux homozygotes SS étant plus susceptibles d'attraper la malaria) (Hedrick 2011) ; le maintien de différents allèles du gène des opsines également par avantage à l'hétérozygote chez certains singes du nouveau monde puisque les hétérozygotes perçoivent un plus large spectre de couleurs (vision trichromate) (Hiwatashi et al. 2010) ; le complexe majeur d'histo-incompatibilité (CMH) chez les vertébrés qui est une famille de gènes jouant un rôle essentiel dans l'immunité et la reconnaissance des pathogènes et qui semble à la fois sous sélection liée aux parasites et au choix du partenaire (Klein et al. 1993); le locus d'auto-incompatibilité (locus *S*) sous sélection fréquence-dépendante négative chez les plantes et les champignons empêchant l'auto-fécondation et la reproduction entre individus apparentés et permettant donc aux allèles rares d'avoir accès à un nombre plus important de partenaires (Goldberg et al. 2010; Roux et al. 2013); ou encore les gènes *R* de résistance aux pathogènes chez les plantes qui résultent d'une co-évolution entre hôtes et pathogènes et évoluent sous une forme de sélection fluctuante dans le temps (Karasov et al. 2014).

Dans tous les cas cités précédemment (sauf *Hbb*), la sélection balancée est tellement ancienne que les mêmes polymorphismes sont observés dans différentes espèces ; on parle alors de polymorphisme trans-spécifique. Etudier de tels polymorphismes partagés est une manière efficace de s'affranchir des deux problèmes méthodologiques soulevés plus tôt : en effet, tous les mécanismes de sélection balancée, s'ils sont assez anciens, prédisent des polymorphismes trans-spécifiques, et de tels partages de polymorphismes ne sont pas attendus sous neutralité si les espèces sont suffisamment distantes. De manière contre-intuitive, il y a donc plus de signaux très anciens de sélection balancée qui ont été détectés que de signaux récents ou d'âge intermédiaires.

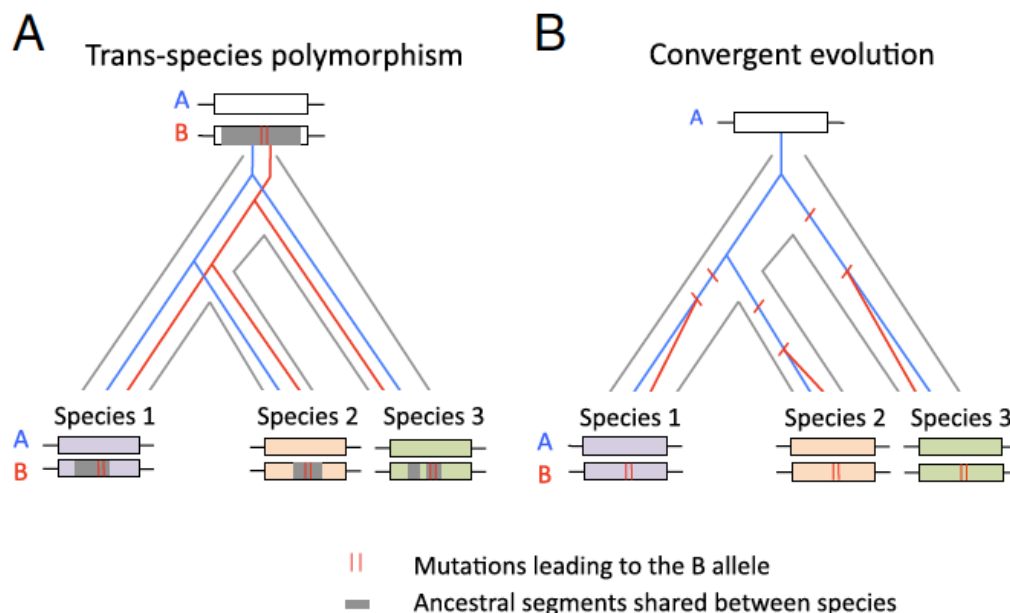
1. Groupe sanguin ABO

Un autre exemple emblématique de sélection balancée partagée entre espèces, dont le mécanisme sous-jacent et la pression de sélection restent cependant obscurs, est le système ABO. Ce système correspond à l'expression d'antigènes A, B et/ou H à la surface de certaines

cellules, et par effet miroir, la présence d'anticorps contre les antigènes non exprimés. Chez l'humain, le gène *ABO* présente des patrons de diversité génétique atypiques par rapport au reste du génome, avec une diversité jusqu'à six fois plus élevée dans l'exon 6 et 7 (où sont codées les différences entre les activités enzymatiques A et B) que la moyenne du génome (Calafell et al. 2008). Ces observations ont fait émerger l'idée que les différents groupes ABO humains ont persisté dans le temps du fait de la sélection balancée.

De manière remarquable, le polymorphisme d'antigènes (et d'anticorps) ABO se retrouve chez un grand nombre de primates non-humains, ainsi que plus largement chez différents mammifères. Les cochons présentent ainsi un allèle A et O (Yamamoto et Yamamoto 2001), et les souris ont un allèle cis-AB ayant une double capacité enzymatique A et B (Yamamoto, Lin, et al. 2001). Des antigènes de type A/B sont également trouvés chez différents microbes (Yamamoto et al. 2014), ce qui explique d'ailleurs pourquoi nous développons, dès la naissance, et en l'absence de transfusion sanguine, des anticorps ABO (par contact avec des micro-organismes extérieurs).

Cependant, ce partage de polymorphismes entre espèces (quand bien même nous prenons pour acquis qu'il soit adaptatif car très improbable sous neutralité) ne signifie pas que la sélection balancée soit ancienne. En effet, il peut s'agir de mutations récurrentes, apparues indépendamment chez plusieurs espèces du fait d'évolution convergente. Le seul fait que différents primates partagent le système ABO ne permet donc pas d'attester que ces phénotypes étaient déjà présents chez leur ancêtre commun. La question de savoir si le système ABO résulte d'une sélection balancée ancienne (i.e. avec des polymorphismes ancestraux) ou d'une évolution convergente récente est justement celle que j'ai exploré pendant mon post-doctorat.

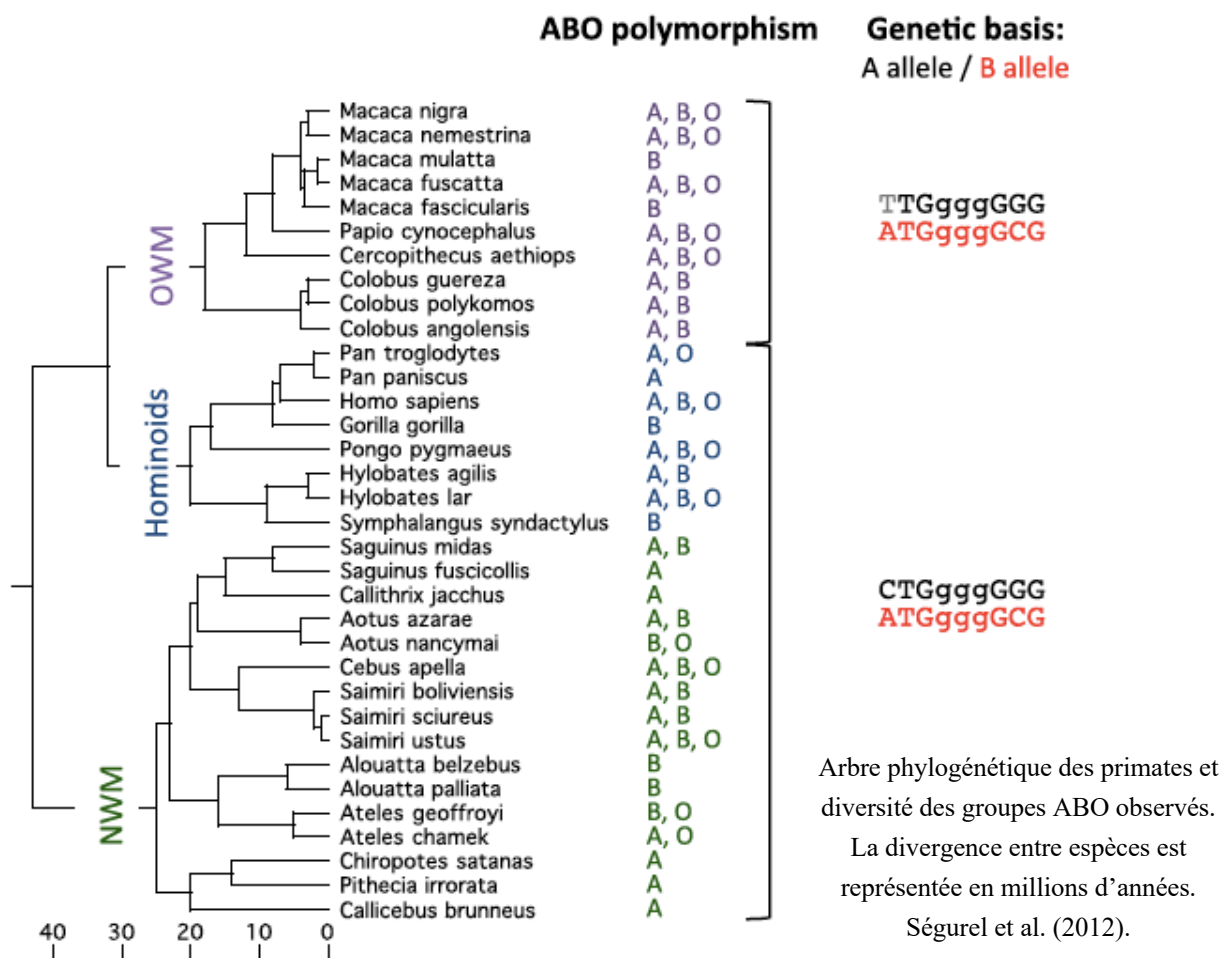


Arbre phylogénétique représentant un polymorphisme trans-spécifique (A), où un segment ancestral en gris est partagé entre espèces, *versus* un scénario d'évolution convergente (B) où deux mutations en rouge apparaissent indépendamment sur trois lignages. Ségurel et al. (2012).

Ces deux modèles explicatifs ont été longtemps débattus, avec un consensus en faveur de l'introduction indépendante de l'allèle B plusieurs fois dans l'évolution (Saitou et al. 1997; Kermarrec et al. 1999; Calafell et al. 2008; Kitano et al. 2009).

Afin de réexaminer la trajectoire évolutive du système ABO, une collecte de données génétiques a été initiée dans un large échantillon de grands singes (chimpanzés, bonobos, gorilles et orang-outans) par Carole Ober (Université de Chicago, USA), en collaboration avec le zoo de Lincoln Park à Chicago. Grâce à ces données, j'ai pu, sous la supervision de Molly Przeworski, étudier les signatures moléculaires présentes dans l'exon 7 du gène ABO, où se situent les deux mutations fonctionnelles différenciant les allèles A et B (Ségurel et al. 2012). J'ai également récupéré les données de la littérature pour des gibbons, macaques et babouins, et contacté d'autres institutions pour obtenir d'autres échantillons ADN, ajoutant ainsi à notre liste des colobes, des vervets, et pour la première fois, des singes du Nouveau Monde.

Nous avons tout d'abord montré que chez les primates, le polymorphisme A/B est retrouvé chez au moins 19 espèces appartenant à 10 genres différents, avec le reste des espèces équitablement réparties entre celles portant l'allèle A mais non l'allèle B (10 espèces), ou inversement l'allèle B mais non l'allèle A (11 espèces) (Ségurel et al. 2012). Au sein des hominoïdes (ou grands singes), les humains, les orang-outans et les gibbons sont polymorphes pour A/B, tandis que les chimpanzés et les bonobos portent seulement l'allèle A, et les gorilles portent seulement l'allèle B.



Il est important de noter que la détection des allèles A et B se fait plus facilement que la détection des allèles O, que ce soit par des approches génétiques ou enzymatiques, du fait que les allèles O sont des allèles nuls, c'est-à-dire perte de fonction. En conséquence, dans un certain nombre de cas, nous savons si les espèces portent des allèles A et/ou B, mais nous ne savons pas si elles portent des allèles O.




Ces nouvelles données génétiques m'ont tout d'abord permis de montrer que les singes du Nouveau Monde et les hominoïdes (espèces ayant divergé il y a 43 millions d'années) ne partagent pas seulement leur polymorphisme enzymatique ABO, mais également la base génétique de ce système : l'allèle A est en effet codé par les nucléotides CTG/GGG et l'allèle B par les nucléotides ATG/GCG (acides aminés 266 et 268 de l'exon 7 du gène ABO) chez toutes ces espèces. Cette observation laisse ouverte la possibilité que le système ABO soit aussi vieux que la divergence de ces espèces. Chez les singes de l'Ancien Monde, comme précédemment observé (Kominato et al. 1992; Saitou et al. 1997), une mutation a été observée, avec l'allèle A étant codé par les nucléotides TTG/GGG. Cette mutation code cependant pour le même acide aminé, donc il s'agit vraisemblablement d'un changement sans effet phénotypique survenu chez l'ancêtre des singes de l'Ancien Monde (« turnover »). En parallèle, les allèles O étant des pertes de fonctions, de nouveaux allèles nuls peuvent apparaître régulièrement et remplacer les allèles existants, et donc, comme attendu, ces allèles O ne sont pas partagés entre espèces (Kermarrec et al. 1999).

Pour distinguer entre le maintien à long terme d'un polymorphisme trans-spécifique et l'évolution convergente plus récente des deux mêmes mutations, j'ai ensuite obtenu des arbres phylogénétiques (en masquant le site fonctionnel). Ces analyses ont révélé que les haplotypes B des humains sont génétiquement plus proches des gorilles B ou des gibbons B que des haplotypes provenant d'autres humains de type A, et similairement pour les colobes, macaques et babouins. Cette signature suggère le maintien d'un polymorphisme ancestral depuis l'origine des hominoïdes (il y a 18 millions d'années) et des singes de l'Ancien Monde (il y a 20 millions d'années). J'ai ensuite montré que les différences génétiques entre allèles A et B sont dix fois plus élevées qu'attendues au sein des espèces polymorphes. Finalement, j'ai développé un test basé sur le nombre de substitutions observé sur chaque branche de la phylogénie des hominoïdes, qui a permis de confirmer que plusieurs haplotypes ont été maintenus depuis au moins l'origine des hominoïdes, si ce n'est plus. Ces résultats excluent donc l'hypothèse d'une évolution convergente des groupes sanguins ABO au sein de ces groupes de primates (Ségurel et al. 2012). Un tel partage de diversité génétique entre grands singes n'avait été jusqu'alors observé que pour le CMH (Hedrick 1998).

Il est fort probable que le polymorphisme A/B soit encore plus ancien que 20 millions d'années, mais les données génétiques ne permettent pas de l'affirmer car le signal de partage entre espèces disparaît avec le temps. Par exemple, on peut calculer qu'en théorie, la région du génome qui porte un signal de partage entre espèces ne fait plus que 300 paires de bases au bout de 7 millions d'années (comme humain-chimpanzé), et seulement environ 60 paires de bases au bout de 30 millions d'années (comme humain-macaque) (Ségurel et al. 2012).

Quelles contraintes sélectives sont responsables d'un tel maintien de diversité génétique au cours du temps par sélection balancée ? Bien qu'à partir des données génomiques, il ne soit pas évident de faire la distinction entre différents mécanismes sélectifs, les cibles génomiques de la sélection balancée semblent souvent associées à des fonctions immunitaires, à l'image du CMH (Spurgin et al. 2010). Les antigènes des groupes sanguins sont d'ailleurs connus pour jouer le rôle de récepteurs cellulaires au travers desquels les pathogènes peuvent initier des infections (Moulds et al. 1996). Une étude chez l'humain a également montré que les gènes codant pour les groupes sanguins sont fortement sous sélection naturelle et semblent soumis à des dynamiques de course à l'armement entre hôtes et pathogènes (Fumagalli et al. 2009).

Alors qu'en est-il pour le système ABO ? Quels sont les mécanismes sélectifs permettant d'expliquer son maintien chez de nombreuses espèces depuis des millions d'années ? Cette question est débattue depuis déjà plusieurs décennies (Livingstone 1960), avec à l'origine l'hypothèse d'une sélection liée à son rôle dans les incompatibilités Rhésus foeto-maternelle (Vos 1965). Cependant, bien que le système ABO soit principalement connu pour son implication dans les problématiques de transfusion sanguine, son rôle évolutif ne semble pas lié à son expression à la surface des globules rouges (Ségurel et al. 2013). En effet, tandis que chez l'humain (et autres grands singes), les antigènes ABO sont trouvés non seulement libres dans la circulation de certains fluides, mais également à la surface des cellules rouges, de l'endothélium vasculaire et à la surface de diverses cellules épithéliales, ces antigènes ne sont pas exprimés à la surface des globules rouges chez les singes de l'ancien monde, et ne sont pas non plus exprimés dans l'endothélium vasculaire chez les singes du nouveau monde. Par conséquent, le seul point commun entre tous les primates polymorphes pour A/B est l'expression d'ABO sur les cellules épithéliales digestives et respiratoires, lieu par excellence de contact avec des micro-organismes extérieurs.

	Species	Body fluids (e.g., saliva)	Respiratory, digestive epithelium	Vascular endothelium	Red blood cells
	New World monkeys	+	+	-	-
	Old World monkeys	+	+	+	-
	Apes	+	+	+	+
		(humans: only in secretors)	(humans: only in secretors)		

Patterns d'expression des antigènes ABO chez trois grands groupes de primates (grands singes, singes de l'Ancien Monde et du Nouveau Monde) montrant que le dénominateur commun (et donc probablement ancestral) est l'épithélium digestif et respiratoire. Ségurel et al. (2013).

Il a d'ailleurs été montré que les variations du système ABO sont impliquées dans la susceptibilité à un certain nombre de maladies infectieuses n'impliquant pas les globules rouges

(Garratty 2005; Yamamoto et al. 2012; Abegaz 2021). A titre illustratif, la variation génétique d'ABO a été montrée comme impliquée dans la réponse au choléra (Glass et al. 1985), au norovirus (Huang et al. 2003), à *Helicobacter pylori* (Aspholm-Hurtig et al. 2004), ou encore au coronavirus (Pereira et al. 2022), avec dans chaque cas des allèles différents étant à risque ou protecteurs. De manière intéressante, une partie de ces pathogènes sont également connus pour infecter d'autres primates non-humains, permettant possiblement d'expliquer le partage de ces pressions de sélection entre espèces. La nature des pressions sélectives sur ABO n'est donc toujours pas claire, tout comme pour le CMH (Spurgin et al. 2010), mais il semble possible que ces deux cas représentent une interaction avec de nombreux pathogènes, et non un agent unique comme dans le cas de *Hbb* et la malaria.

En termes de mécanisme sous-jacent, dans le cas d'ABO, nous pouvons exclure un avantage aux hétérozygotes étant donné que des allèles cis-AB permettant d'exprimer A et B sur le même allèle existent mais restent très rares chez l'humain (Yamamoto et al. 1993). Si le phénotype AB était avantageux, ces allèles devraient en effet se fixer dans les populations. Certains auteurs ont proposé qu'il s'agisse principalement d'un avantage aux rares, avec les pathogènes arrivant mieux à infecter les individus portant les antigènes ABO les plus fréquemment rencontrés dans la population (Gagneux et al. 1999), mais il est difficile d'exclure la possibilité d'un modèle de sélection fluctuante qui résulterait d'interactions hôtes-pathogènes où la virulence des pathogènes et la résistance de l'hôte sont toutes deux coûteuses.

Afin de mieux comprendre quelles sont les dynamiques adaptatives à l'œuvre, nous pouvons également nous poser la question à l'envers, c'est à dire nous demander non pas pourquoi le polymorphisme ABO a été maintenu depuis des millions d'années chez de nombreuses espèces, mais pourquoi certaines espèces ont perdu l'allèle A ou B, à l'instar de nos proches cousins le chimpanzé, le bonobo et le gorille. L'étude de ces différentes espèces dans l'arbre phylogénétique des primates ne permet pas de trouver de points communs évidents pour élaborer une hypothèse évolutive claire : les espèces ayant perdu A ou B ont à la fois des petites et grandes tailles de population, sont trouvées sur tous les continents, et chez des espèces occupant la même aire de répartition que d'autres espèces polymorphes pour ABO. Il est notable qu'aucune espèce ne semble avoir perdu à la fois A et B au profit de l'allèle O, bien que comme dit précédemment, nous ne pouvons pas en être complètement sûrs du fait de la plus grande difficulté de caractériser génétiquement les allèles O.

Une possibilité intrigante est que la caractérisation du système ABO en allèles A, B et O, qui découle des phénotypes des groupes sanguins, ne soit pas appropriée pour décrire correctement les polymorphismes qui sont perçus par les pathogènes, et donc par la sélection naturelle. Ainsi, des espèces que l'on considère comme n'ayant que l'allèle A pourraient en réalité porter plusieurs polymorphismes fonctionnellement importants non reconnus. L'existence de différences antigéniques importantes entre les allèles A1 et A2 non seulement chez l'humain mais également chez les grands singes (Socha et al. 1979 ; Chester et al. 2001) et l'observation d'un pic de diversité génétique au niveau de l'exon 4 du chimpanzé malgré le fait que cette espèce ne porte que l'allèle A (Ségurel et al. 2013) pourrait conforter cette piste

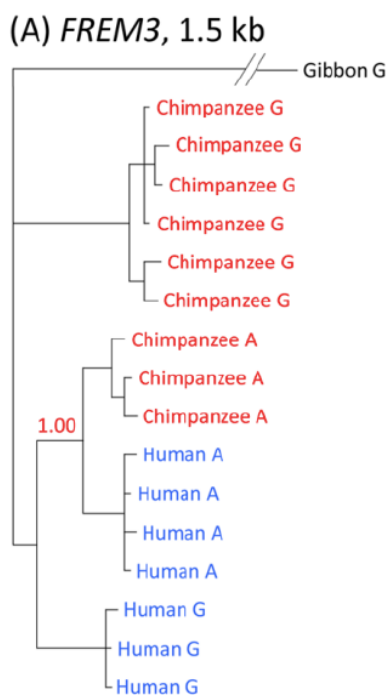
qui reste à explorer. Il semble donc bien que le système ABO n'a pas encore livré tous ses secrets évolutifs.

2. Scan génomique de sélection balancée partagée entre humains et chimpanzés

Dans un autre projet initié lors de mon post-doctorat par Molly Przeworski (Université de Chicago) et Gil McVean (Université d'Oxford), nous avons scanné le génome entier de populations humaines et de chimpanzés pour trouver d'autres cas de polymorphismes partagés entre espèces (Leffler et al. 2013). Le projet PanMap représentait en effet le premier jeu de données de polymorphismes à l'échelle du génome de grands singes, ici 10 chimpanzés de l'Ouest (Auton et al. 2012). Nous les avons ainsi combiné avec des données de polymorphismes obtenues par séquençage génomique de 59 individus Yoruba d'origine Africaine (1000 Genomes Project Consortium 2010).

Après un filtrage intensif des données, nous avons listé tous les loci partageant les deux mêmes allèles entre humains et chimpanzés. Ces polymorphismes peuvent être ancestraux et avoir été maintenus par sélection balancée ou correspondre à des mutations récurrentes dans les deux espèces (neutres ou sous sélection balancée). Afin de nous focaliser sur les signaux les plus intéressants, nous avons regroupés les cas où au moins deux polymorphismes partagés étaient situés à moins de 4kb les uns des autres, et dont le déséquilibre de liaison était identique entre espèces (haplotypes partagés), une situation qui, d'après nos simulations, devrait exclure la majorité des mutations récurrentes. Nous avons trouvé 125 haplotypes partagés entre espèces.

Pour trois régions (vérifiées par reséquençage), j'ai de plus trouvé un arbre phylogénétique regroupant les allèles par classe haplotypique et non par espèce, ainsi que des niveaux de diversité élevés au sein de chaque espèce.



Arbre phylogénétique des haplotypes à proximité du gène *FREM3*, où nous voyons que les haplotypes humains et chimpanzés portant un « A » sont plus proches entre eux que des haplotypes conspécifiques portant un « G ». Leffler et al. (2013)

Ces trois régions comprennent de plus des eQTL (polymorphismes d'expression génique), montrant que ces cas de sélection balancée sont associés à la régulation de l'expression des gènes. Une de ces régions a d'ailleurs par la suite été montrée comme associée à la résistance à la malaria chez l'humain (Malaria Genomic Epidemiology Network et al. 2015). Il serait intéressant de savoir si cette région est également associée à un phénotype immunitaire chez le chimpanzé, et si oui, en association avec quel pathogène.

Afin d'évaluer les points communs à toutes les régions identifiées, nous avons ensuite effectué une analyse d'enrichissement fonctionnel en partant d'un côté des SNPs non-synonymes partagés et de l'autre des haplotypes partagés ; ces deux analyses ont abouti à un enrichissement en glycoprotéines. Etant donné que les virus et les bactéries utilisent fréquemment ces glucides complexes pour infecter les cellules des hôtes, il semble que les candidats identifiés dans cette étude reflètent bien l'association connue entre les cibles de la sélection balancée et les interactions hôtes-pathogènes (Ferrer-Admetlla et al. 2008; Andrés et al. 2009; Fumagalli et al. 2009; Spurgin et al. 2010; DeGiorgio et al. 2014; Karasov et al. 2014; Key et al. 2014; Croze et al. 2016; Koenig et al. 2019; Alharbi et al. 2022; Lawal et al. 2023). De manière plus générale, cette étude a montré qu'il existe plus de cibles de la sélection balancée que ce qui était suggéré dans la littérature.

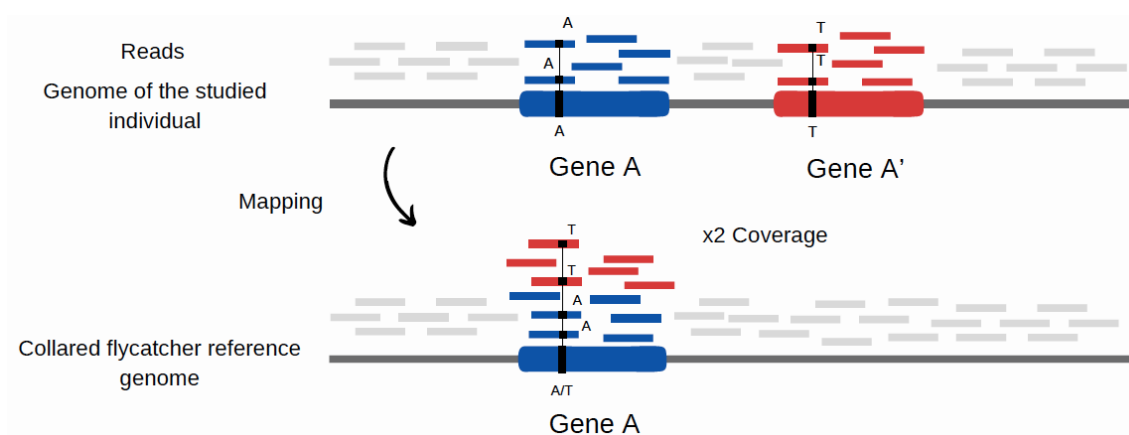
3. *Travail en cours*

Afin de continuer à étudier et identifier les cibles génomiques de la sélection balancée chez les grands singes, j'ai encadré le stage de master 2 de Dominique Hicks (en 2023) qui a consisté à appliquer la méthode de Cheng & DeGiorgio (2022), appelée BalLeRMix+, sur des données publiques de bonobos et chimpanzés (de Manuel et al. 2016), espèces qui ont divergé il y a 1 millions d'années (Prado-Martinez et al. 2013).

BalLeRMix est une méthode qui utilise un modèle mixte pour différencier les sites sous sélection balancée des sites neutres. Brièvement, pour chaque site focal, le logiciel calcule la probabilité d'observer les fréquences alléliques des sites avoisinants sous deux spectres de fréquence alléliques (SFS) : celui correspondant au SFS neutre (calculé sur l'ensemble du génome), et celui attendu sous sélection balancée, en prenant en compte différentes fréquences alléliques possibles à l'équilibre, sous un scénario d'avantage à l'hétérozygote. Ce deuxième SFS correspond à un modèle mixte car il est un mélange de l'attendu sous neutralité et sous sélection balancée, en fonction de la distance entre le site d'étude et le site focal. Le logiciel établit ensuite le score de chaque site focal en pondérant les probabilités des sites avoisinants par leur distance au site focal. Ainsi, le logiciel n'est pas très sensible à la taille de la fenêtre considérée, et ne fait pas non plus d'a priori sur la fréquence d'équilibre attendue sous sélection balancée. Cette méthode permet ainsi de capturer un plus large spectre de signal. L'avantage de cette méthode, par rapport notamment au scan que nous avons effectué en 2013 (Leffler et al. 2013) est qu'elle est appliquée à toutes les espèces (et populations) séparément, et permet donc de détecter des régions sous sélection balancée dans plusieurs espèces, sans nécessairement faire l'hypothèse que les polymorphismes soient partagés (trans-spécifiques).

Bien que les données de chimpanzés aient été largement utilisées en combinaison avec des données humaines, aucun scan de sélection balancée au sein du genre *Pan* n'a été effectué auparavant. Avant de faire tourner BalLeRMix sur les données, la première étape a été de les filtrer correctement. En effet, nous utilisons ici des données de lectures courtes de séquençage qui ont été alignées contre un génome de référence. Ainsi, toute erreur d'assemblage du génome de référence, mais aussi toute erreur d'alignement des lectures courtes sur le génome de référence, ou tout simplement le fait que les lectures courtes (ici venant de chimpanzés et bonobos) aient été alignées sur le génome d'une autre espèce que la leur (ici, l'humain), peut avoir des conséquences importantes pour les analyses en aval.

Notamment, au niveau des régions répétées du génome, qu'il s'agisse d'éléments répétés ou de variants structuraux (duplication/délétion), les erreurs d'assemblage ou d'alignement sont fréquentes. Or si des lectures de séquençage venant en réalité de deux régions paralogues (gène A et gène A' sur la figure ci-dessous) se retrouvent alignées à un endroit unique du génome de référence (gène A), toutes les positions où les régions paralogues sont différentes (divergentes) sont alors perçus (à tort) comme des polymorphismes (SNP A/T sur la figure). Dans de telles régions appelées 'collapsed', on s'attend à observer une augmentation artificielle de la densité de SNPs et un excès d'allèles à fréquences intermédiaires, des signatures typiques de la sélection balancée. L'étape de filtrage est donc cruciale pour pouvoir interpréter correctement les sorties de BalLeRMix.



Représentation schématique d'une région 'collapsed' où des erreurs de mapping de lectures de séquençage ou d'assemblage du génome de référence entraînent un excès de couverture, d'hétérozygotie, et la présence de SNPs erronés à fréquences intermédiaires. (© M Barrientos)

Dominique a ainsi utilisé la couverture moyenne des individus et l'hétérozygotie par site comme moyens complémentaires de détecter ces régions problématiques. En effet, si les lectures de deux régions se retrouvent alignées contre une seule région du génome, cette région aura une couverture de 2X par rapport au reste du génome. Egalement, si ces deux régions divergent à une position donnée, tel que les copies du gène A portent un A et que les copies du gène A' portent un T, la région 'collapsed' présentera un polymorphisme A/T, et ce *pour tous les individus*, entraînant donc une hétérozygotie à ce locus de 100%. Or, dans une population

panmictique, la fréquence maximale d'hétérozygotes doit être de 50%. Observer un SNP avec une fréquence d'hétérozygotes significativement supérieure à 50% est donc un indice que la région est problématique.

Après avoir filtré les données, Dominique a identifié huit régions candidates partagées entre populations de chimpanzés ou entre chimpanzés et bonobos, dont quatre correspondent au CMH et quatre sont nouvelles. Malheureusement, après inspection du signal le plus fort, il semble que le signal dans cette région soit également artefactuel, avec les individus hétérozygotes ayant une couverture plus élevée que les individus homozygotes. Les filtres mis en place par Dominique n'ont donc pas suffi à identifier toutes les régions problématiques, notamment celles qui sont polymorphes pour le nombre de copies. En effet, si un individu porte seulement le gène A sur son haplotype paternel, et le gène A et A' sur son haplotype maternel, sa couverture attendue ne sera plus de 2X, mais de 1.5X, et le pourcentage d'hétérozygotie attendu ne sera pas de 100%, mais dépendra alors de la fréquence de ce variant structural.

Afin d'établir une pipeline d'analyse plus sensible, et dans l'optique d'appliquer ces mêmes analyses sur des données de séquençage d'un large ensemble de gobemouches en collaboration avec Carina Mugal (LBBE), nous avons alors proposé un autre sujet de stage visant à utiliser la couverture par site *par individu* pour identifier de façon plus robuste ces régions 'collapsed'. Cela a été le travail de stage de master 2 de Manuel Barrientos en 2024. En utilisant des fenêtres chevauchantes de 1kb, il a identifié chez ces oiseaux les régions ayant une couverture supérieure à 1.5X celle du génome, et ce chez au moins 4/20 individus de la population. Ces régions 'collapsed' représentent au final environ 0.2% du génome. Bien que ces paramètres aient été choisis pour maximiser notre sensibilité à ces erreurs d'alignement sans pour autant enlever trop de régions non problématiques, il s'est avéré qu'encore un grand nombre d'outliers de BalLeRMix (environ la moitié des signaux) correspondait à des régions suspectes. Il s'agit par exemple de régions pour lesquelles la couverture est problématique sur seulement 300bp, ou de régions à proximité de celles que Manuel a identifiées mais qui n'étaient pas elles-mêmes considérées comme problématiques. Il va désormais s'agir de combiner l'information de couverture avec celle d'hétérozygotie, et d'établir des scores quantitatifs (plutôt que binaires) pour essayer de gagner encore plus en puissance de détection.

Le travail de Manuel a cependant permis de détecter un certain nombre de régions candidates sous sélection balancée, dont certaines partagées entre populations et/ou espèces, notamment une région dans le gène *ABCA13* qui avait été également trouvée dans un scan de sélection balancée entre humains et chimpanzés (Cheng et al. 2019), sans que le phénotype associé ne soit connu. Il sera donc intéressant d'étudier plus en détail la diversité et structure génétique de ce locus, notamment en phasant les données et en obtenant un arbre phylogénétique des haplotypes identifiés.

De manière générale, ce travail a aussi permis de mettre en lumière que les erreurs d'alignement des lectures courtes, qui peuvent être sans grande conséquence si l'on s'intéresse à l'histoire démographique de la population, deviennent cruciales à prendre en compte quand

on s'intéresse à la sélection balancée. J'ai donc à cœur de perfectionner ces filtres et de les appliquer sur les jeux de données publiques d'humains et de primates non-humains qui ont été largement utilisés pour identifier précédemment des cibles de sélection balancée. Cependant, la sensibilité de ces filtres dépend fortement de la couverture moyenne des jeux de données, et du nombre d'individu par population, et dans certains cas, il ne sera pas possible d'identifier systématiquement ces régions *a priori*, et donc des vérifications devront être effectuées *a posteriori*.

En parallèle, une approche intéressante (mais couteuse) consiste à obtenir des données de lectures longues (d'environ 10kb), ce qui permet de s'affranchir de la plupart des erreurs d'alignement. J'ai donc également pour ambition d'obtenir de telles données pour le projet détaillé dans la dernière partie.

Partie 3- Co-évolution hôte-microbiote intestinal

1. *Microbiote intestinal humain et mode de vie*

Les êtres humains, longtemps considérés comme des organismes singuliers, sont en réalité des écosystèmes en tant que tels composés non seulement de cellules humaines, mais également de communautés microbiennes complexes, indispensables à leur survie. L'importance de ces interactions hôte-microbe chez les plantes et les animaux, ainsi que leur rôle dans la physiologie de l'hôte, a récemment été mise en évidence et a entraîné un changement important de paradigme aboutissant au concept d'holobionte, défini comme l'ensemble formé par un hôte et tous les micro-organismes qui lui sont associés (McFall-Ngai et al. 2013; Hassani et al. 2018). Le microbiote intestinal humain a reçu particulièrement beaucoup d'attention de la part de la communauté scientifique ces dernières années car il semble être un facteur clef en santé humaine, effectuant des fonctions digestives et immunitaires essentielles et étant impliqué dans un certain nombre de maladies métaboliques et inflammatoires (Kinross et al. 2011). Il présente de plus une grande variabilité taxonomique entre individus (Gevers et al. 2012). Ces différences de composition ont initialement été utilisées pour définir trois grands entérotypes (des groupes d'individus présentant une composition microbienne proche, avec une prévalence caractéristique de *Ruminococcus*, de *Bacteroides*, ou de *Prevotella*, respectivement) (Arumugam et al. 2011), mais il a ensuite été argumenté que ce regroupement ne reflète pas la complexité de ces écosystèmes qui sont mieux caractérisés par une variabilité continue (Arumugam et al. 2014; Falony et al. 2016).

Cette variabilité inter-individuelle semble de manière générale largement influencée par des facteurs environnementaux externes (Rothschild et al. 2018), tandis que seulement une minorité de taxons sont influencés par la génétique de l'hôte (Bonder et al. 2016; Goodrich et al. 2016). Une large étude de deux cohortes urbaines européennes a ainsi montré que l'alimentation, les pratiques médicales, le métabolisme et l'état de santé de l'hôte sont les facteurs les plus significativement associés aux variations du microbiote intestinal (Falony et al. 2016). Cependant, bien que la structure du microbiote soit dynamique, elle présente aussi une certaine stabilité, entraînant une réelle spécificité individuelle dont les déterminants restent méconnus (Faith et al. 2013; Truong et al. 2017; Mehta et al. 2018; Chen et al. 2021). En effet, un individu partage la même souche que lui-même (pris à quelques mois d'écart) pour 80% des espèces bactériennes partagées, la même souche que sa mère pour 50% des espèces partagées si l'individu a moins de 1 an, et pour 15-20% des espèces si l'individu a 3-80 ans, approchant la valeur de 12% trouvée entre individus vivant sous le même toit (Valles-Colomer et al. 2023).

Bien que la caractérisation du microbiome humain (l'ensemble des génomes des microorganismes associés à l'humain) ait engendré de nombreuses questions de recherche en sciences médicales, et que sa structure représente un phénotype d'intérêt potentiellement ciblé par la sélection naturelle, peu d'études l'ont considéré dans une perspective écologique et évolutive. Ainsi, nous ne savons encore que peu de choses sur l'histoire évolutive du microbiote et sur les interactions qu'il existe au sein de cet écosystème microbien dans un contexte

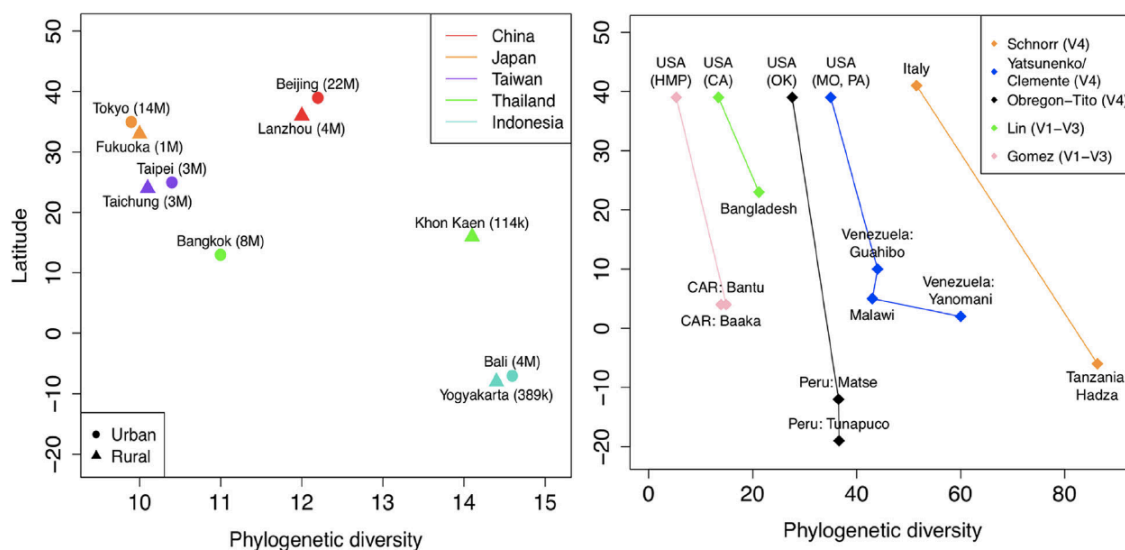
« naturel ». En effet, la majorité des études se sont intéressées au microbiome de populations urbaines fortement industrialisées (principalement aux États-Unis, en Europe et en Asie de l'Est), ce qui représente un environnement particulièrement perturbé. La compréhension de la biodiversité du microbiome dans des environnements et milieux plus diversifiés constitue donc une thématique de recherche importante.

Les études qui se sont intéressées à la diversité et composition du microbiome dans des populations rurales aux modes de vie « traditionnels » vivant dans des pays moins industrialisés (en Afrique, en Amérique du Sud, et en Asie) les ont comparé à des populations urbaines de pays fortement industrialisés (De Filippo et al. 2010; Yatsunenko et al. 2012; Lin et al. 2013; Schnorr et al. 2014; Martinez et al. 2015; Clemente et al. 2015; Obregon-Tito et al. 2015; Gomez et al. 2016) et ont observé qu'il existe une perte systématique de diversité du microbiote intestinal dans ces dernières, ainsi que des changements importants de composition (baisse de *Prevotella* et de *Treponema*, et augmentation de *Bacteroides*) (Segata 2015; Mancabelli et al. 2017; Pasolli et al. 2019). Ces modifications de structures microbiennes ont des implications potentiellement importantes en sciences médicales, étant donné qu'une moindre diversité bactérienne intestinale a également été associée à un certain nombre de maladies métaboliques et inflammatoires (Mosca et al. 2016). Cependant, interpréter les différences observées entre des populations culturellement et géographiquement si différentes n'est pas trivial, car il existe de nombreux facteurs confondants, d'autant que la plupart de ces études n'ont pas collecté de métadonnées associées (ou très peu). En conséquence, les causes de cette réduction de biodiversité (ainsi que des changements de composition) sont mal connues. S'agit-il d'une réponse des microbes intestinaux à une diminution de la consommation en fibres alimentaires (liée à une alimentation plus grasse et plus sucrée en ville) (Deehan et al. 2016), aux changements sanitaires liés à l'industrialisation tels que la stérilisation accrue de notre environnement, ou encore aux nouvelles pratiques médicales telles que la consommation massive d'antibiotiques, de lait artificiel ou la césarienne (Blaser et al. 2009; Cho et al. 2012) ? De plus, l'influence de ces changements culturels sur les autres microbiomes humains reste méconnue. Le microbiome intestinal est-il le seul à perdre en diversité ou est-ce un trait commun à tous les micro-organismes associés aux humains ? Les quelques précédentes études sur le microbiome oral n'ont pas été concluantes (Li et al. 2014; Clemente et al. 2015). Les réponses à ces questions sont cependant importantes car elles peuvent permettre d'informer les politiques de santé publique afin de savoir quels aspects de nos modes de vie urbains et industrialisés sont les plus néfastes pour la bonne santé de notre écosystème microbien.

Effet de la latitude ? (Dikongué et al. 2017)

Avec une étudiante de master 1 (Emma Dikongué, 2016), nous avons tout d'abord effectué une méta-analyse de données publiées de diversité de microbiote intestinal et comparé les différences de diversité entre populations et entre individus sains et malades. Nous avons montré que la perte de diversité chez les individus malades est en moyenne de 22% tandis que celle associée à l'industrialisation est de 37%. Nous avons de plus reanalysé des données existantes pour des populations urbaines et rurales dans plusieurs pays d'Asie (Nakayama et al.

2015) afin de tester l'hypothèse d'un effet du climat sur la diversité du microbiote intestinal. Les populations industrialisées étudiées sont en effet toutes à des latitudes plus élevées que les populations non-industrialisées et il est connu que la diversité des espèces est inversement corrélée à la latitude (Mittelbach et al. 2007; Fuhrman et al. 2008). Nous avons en effet bien trouvé une corrélation significative entre latitude et diversité dans ce jeu de données, suggérant que le climat pourrait être une variable confondante dans ces études. De plus, des études plus récentes sur le microbiote de populations aux modes de vie « traditionnels » vivant à des hautes latitudes (Inuit) en ou altitude (Népalais) ont montré que ces populations ont un niveau de diversité microbienne proche de celui des populations urbaines industrialisées, confortant l'idée que les variables climatiques sont potentiellement un facteur important à prendre en compte.



Diversité phylogénétique du microbiome intestinal en fonction de la latitude chez des enfants asiatiques ruraux et urbains (gauche); et chez des adultes de pays plus ou moins industrialisés (droite). Dikongué et Ségurel (2017)

Boite 2. Indices de diversité écologique

La diversité taxonomique d'un écosystème est une métrique complexe pouvant se calculer de différentes manières. On peut en effet compter le nombre d'espèces présentes dans cet écosystème (pour un microbiome, il s'agit du nombre d'OTUs [Operational Taxonomical Units] ou d'ASVs [Amplicon Sequence Variants]), il s'agit alors de la *richesse spécifique*. Mais on peut corriger cette richesse pour la *distance phylogénétique* entre espèces. Observer quatre espèces phylogénétiquement très proches n'est en effet pas équivalent à en observer quatre très différentes, du fait qu'a priori, des espèces phylogénétiquement proches auront des fonctions plus redondantes dans l'écosystème. Mais la richesse peut également être pondérée par la prévalence des espèces, il s'agit alors d'une valeur d'*abondance*. En effet, un écosystème ayant une espèce majoritaire et trois espèces minoritaires n'est pas équivalent à un écosystème avec quatre espèces aux prévalences bien équilibrées. L'abondance peut être calculée de différentes manières, en fonction du poids donné aux fréquences observées, entraînant un continuum de valeurs d'abondance possibles (Jost 2006; Chao et al. 2010). Ainsi, chaque indice de diversité taxonomique capture des aspects différents et complémentaires des données.

Boite 3. 16S, métagénomique, et MAGs (metagenome-assembled genomes)

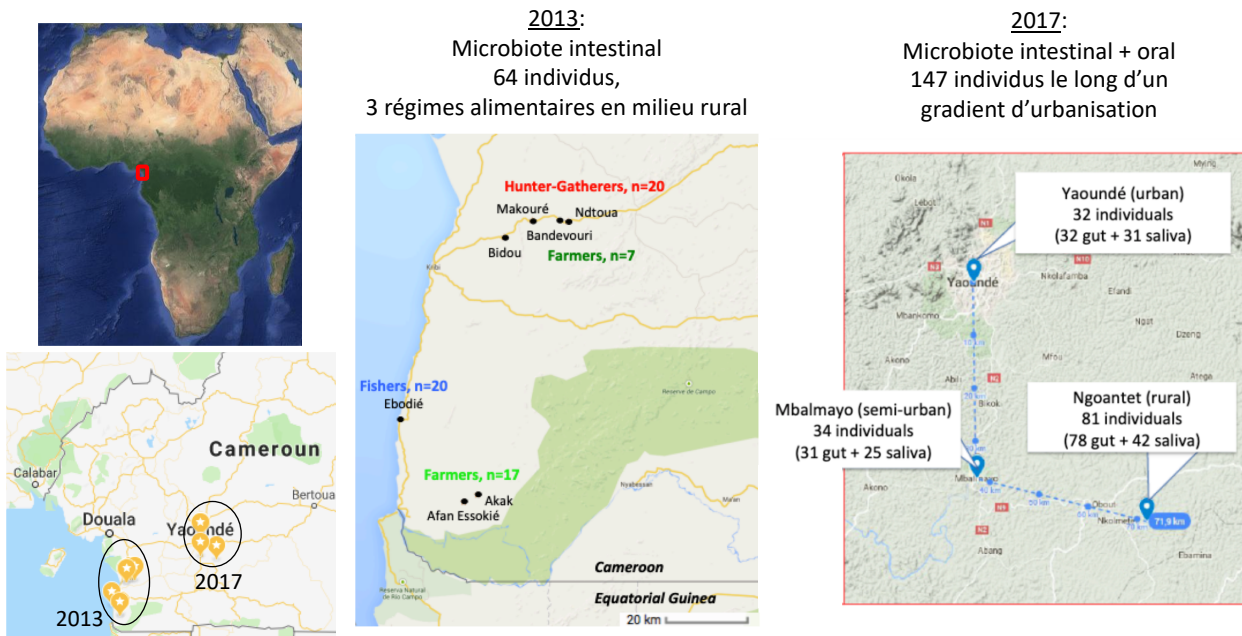
Il me semble nécessaire de faire un rapide point sur les différents types de données génomiques qu'il est possible d'acquérir sur une communauté d'espèces, c'est-à-dire quand les données sont obtenues sur un écosystème et non sur un individu ou une espèce unique.

Tout d'abord, il est possible de séquencer un gène (ou une partie de gène) donné, ce qui nécessite donc d'effectuer une amplification par PCR au préalable, et est communément appelé « metabarcoding ». L'avantage de cette approche est qu'elle permet d'obtenir une idée de la diversité phylogénétique de l'échantillon (par comparaison de toutes les séquences obtenues) à moindre coût (il est par exemple possible d'amplifier seulement 250bp sur un microbiote). L'inconvénient est i) qu'il y a forcément des biais d'amplification : même si les gènes choisis, à l'instar du 16S, sont caractérisés par une alternance de séquences conservées « universelles » pour situer les primers et de séquences hyper variables pour maximiser les différences entre espèces, toutes les espèces ne s'amplifient pas équitablement, entraînant des déformations des abondances réelles ; ii) conséquence directe du premier point, le metabarcoding ne cible toujours qu'une partie de la diversité du vivant (procaryote avec le 16S, eucaryote avec le 18S...) car il n'existe pas de primers réellement universel, donc l'écosystème n'est pas exhaustivement séquencé ; iii) conséquence également directe du premier point, la résolution taxonomique n'est pas très bonne, car plusieurs espèces, voir genres, peuvent partager les mêmes séquences 16S (par manque de diversité génétique) ; iv) par définition, un seul gène est séquencé donc l'information sur le reste du génome (et notamment la présence/absence d'autres gènes d'intérêt) est manquante. Nous l'aurons compris, l'approche de métagénomique, c'est-à-dire le séquençage de l'ensemble des génomes d'un échantillon sans amplification ciblée présente de nombreux avantages face aux approches de metabarcoding. Le seul inconvénient : le prix, et peut être aussi l'expertise bioinformatique (de stockage et d'analyse) requise pour le traitement de ces données.

Concernant les métagénomiques, il s'agit ensuite soit de mapper les lectures de séquençage contre des génomes de référence obtenus le plus souvent à partir d'isolats (ces génomes ont l'avantage de ne venir, pour sûr, que d'une seule espèce), ou d'assembler les génomes *de novo* (aboutissant ainsi à des MAGs, i.e. Metagenomes-Assembled Genomes). Les méthodes d'assemblage sont loin d'être triviales, mais permettent de s'affranchir des biais liés aux bases de données de référence, d'autant plus si l'on travaille sur des écosystèmes peu/mal couverts dans les bases de données (comme c'est notamment le cas pour les populations humaines peu industrialisées, bien que la situation soit encore pire pour d'autres espèces non modèles).

Il est intéressant de noter que, bien qu'actuellement, l'écrasante majorité des données de séquençage correspondent à des lectures courtes, ce qui aboutit à une réelle difficulté (voire parfois impossibilité) d'assembler correctement les génomes, le développement récent du séquençage en lectures longues, bien qu'encore assez restreint du fait de son coût élevé, promet d'apporter des améliorations notables pour ces problématiques méthodologiques.

Pour répondre aux questions préalablement soulevées, j'ai mis en place, en collaboration avec Alain Froment (IRD), des collectes de données biologiques et culturelles au Cameroun. La première campagne d'échantillonnage (effectuée en 2013) a ciblé le microbiote **intestinal**, dans des populations partageant un même environnement rural peu industrialisé, mais différant de par leur **régime alimentaire** (chasseurs-cueilleurs, agriculteurs, pêcheurs). La deuxième campagne d'échantillonnage (effectuée en 2017) a permis de collecter des données de microbiote **intestinal et oral** le long d'un **gradient d'urbanisation** (dans la capitale, 2 millions d'habitants ; dans une ville de 60 000 habitants ; et dans un village de 1 500 habitants). Pour ces deux missions, nous avons également collecté de nombreuses métadonnées associées : alimentation, pratiques médicales, habitat, variables socio-économiques. Nous avons obtenu, en collaboration avec Ran Blekhman de l'Université de Minnesota (USA), des données 16S sur les échantillons fécaux et salivaires et des données de métagénomique sur les échantillons fécaux. Ce projet a été possible grâce à l'obtention d'une ANR JCJC (MICROREGAL, 2015-2020), ainsi que d'un PEPS CNRS (2017) et d'un prix de recherche IDF-FRM (2018).



Cartes montrant la localisation géographique des populations échantillonnées lors des campagnes de terrain de 2013 et 2017. Morton et al (2015) ; Lokmer et al (2020)

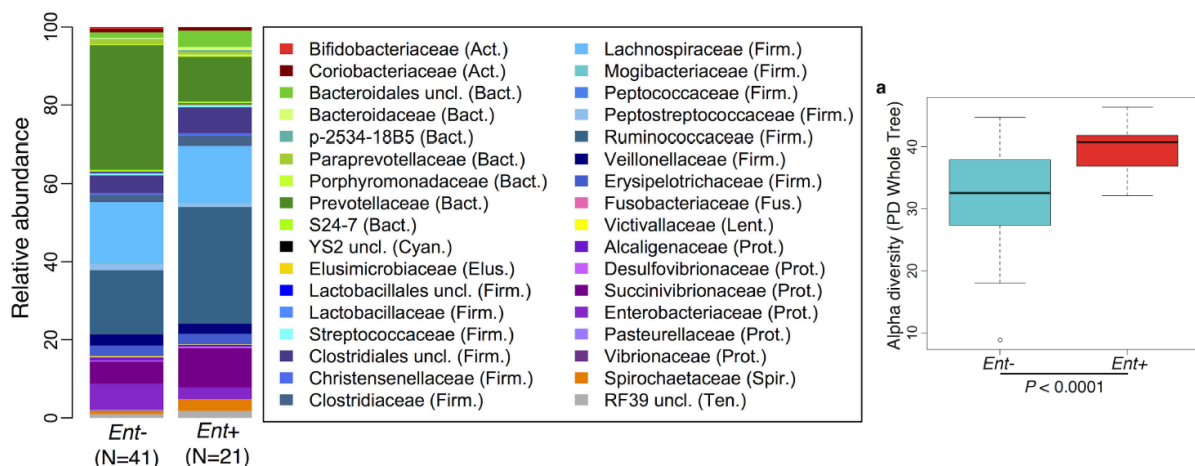
L'objectif général de ces campagnes de terrain était de se placer à une échelle locale et d'étudier le microbiote en milieu peu industrialisé, afin de pouvoir faire la part des choses entre l'importance respective des facteurs alimentaires et sanitaires. Les données de 2013 ont été analysées pour regarder l'effet du régime alimentaire et des vers intestinaux (vers), tandis que les données de 2017 ont été analysées pour regarder l'effet de l'urbanisation. Les deux jeux de données ont été utilisés pour étudier l'effet des protozoaires intestinaux.

Effet du régime alimentaire et des parasites intestinaux (Morton et al. 2015)

Il est important de noter tout d'abord qu'un facteur confondant de notre comparaison chasseurs-cueilleurs / agriculteurs / pêcheurs est le fond génétique. En effet, tous les chasseurs-

cueilleurs sont de groupe ethnique (et de fond génétique) Pygmés, tandis que les agriculteurs et pêcheurs sont de groupe ethnique (et de fond génétique) Bantu. Il a été précédemment estimé que ces populations ont divergé il y a environ 60,000 ans (Verdu et al. 2009). Cependant, comme mentionné plus haut, il semble que la génétique de l'hôte n'a qu'une influence minime sur le microbiote, avec seulement quelques taxons héritables (Bonder et al. 2016; Goodrich et al. 2016). L'objectif de notre étude était donc d'estimer l'influence du régime alimentaire sur le microbiome intestinal, et d'identifier les taxons qui diffèrent significativement entre modes de subsistance. Nos résultats ont montré que, bien que nous ayons identifié des taxons précis qui diffèrent significativement entre modes de subsistance (notamment une plus grande prévalence chez les chasseurs-cueilleurs du phylum Proteobacteria et du genre *Sarcina*, tous deux également montrés comme différenciant des populations rurales traditionnelles de populations urbaines industrialisées, ainsi que la présence de *Bifidobacterium* exclusivement chez les pêcheurs, reflétant leur consommation occasionnelle de yaourts), le régime alimentaire n'a pas un effet majeur sur la composition globale du microbiome à cette échelle d'étude.

En parallèle du régime alimentaire, nous avons également exploré l'influence des parasites intestinaux sur le microbiome intestinal, avec l'idée qu'il faut regarder de la manière la plus exhaustive possible l'écosystème intestinal dans son ensemble. Nous avons donc observé les échantillons fécaux sous microscope pour identifier les parasites présents. En collaboration avec le Centre Pasteur du Cameroun, nous avons ainsi identifié trois espèces de vers intestinaux (*Trichuris*, *Ancylostoma* et *Ascaris*) et une espèce de protozoaire (*Entamoeba*), toutes les quatre présentes à des fréquences non négligeables (10-30% par population). Contrairement au régime alimentaire, la présence de parasites intestinaux s'est avérée être un facteur déterminant pour le microbiote. Nous avons en effet montré que les vers intestinaux sont significativement associés à une plus forte diversité microbienne, mais pas à des changements majeurs de composition. L'effet le plus fort était celui du protozoaire *Entamoeba*, associé à la fois à une augmentation de diversité et un important changement en composition. Cet effet est si fort que nous pouvons prédire avec une précision de 79% si un individu est porteur ou non d'*Entamoeba* à partir de son microbiote.



Composition et diversité du microbiote d'individus porteurs ou non d'*Entamoeba*.

Morton et al. (2015)

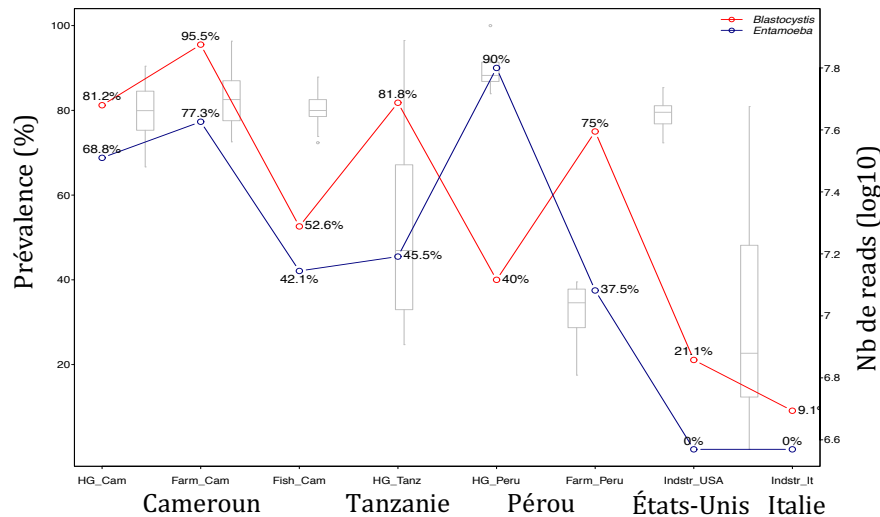
Ces résultats étaient assez inattendus. En effet, contrairement aux vers intestinaux dont l'effet, passant a priori par un remodelage de l'état inflammatoire de l'hôte, avait précédemment déjà été montré, l'effet d'*Entamoeba* était nouveau, étant donné qu'aucune étude n'avait encore pris en compte l'effet de protozoaires sur le microbiome intestinal humain. Ne sachant pas si l'espèce d'*Entamoeba* présente dans nos données était *E. histolytica*, une espèce clairement pathogénique responsable de l'amibiase, ou une des autres espèces non pathogéniques (dont la plus courante est *E. dispar*), du fait qu'elles ne sont pas distinguables par microscopie, nous avons ensuite effectué des PCR ciblées et ainsi montré qu'aucun individu n'était porteur d'*E. histolytica* mais plusieurs étaient positifs pour *E. dispar* (résultats non publiés du stage d'E. Dikongué). L'effet observé dans notre étude d'*Entamoeba* sur la diversité et la composition du microbiote intestinal était donc bien celui d'un protozoaire a priori commensal.

En parallèle de nos résultats, deux autres études, une en milieu urbain en France (Audebert et al. 2016) et l'autre en milieu rural au Pérou (Tito et al. 2019), ont également analysé l'effet sur le microbiote intestinal d'un autre protozoaire intestinal assez prévalent, *Blastocystis sp.*, dont le statut infectieux est toujours débattu. A l'instar de nos résultats sur *Entamoeba*, ces études ont trouvé des effets positifs sur la diversité et des changements importants de composition du microbiote en présence de *Blastocystis*. J'ai rencontré Magali Chabé, chercheuse à l'institut Pasteur de Lille et co-auteurice de la première étude, à un congrès où nous avons toutes deux présenté nos résultats, similaires dans le fond (effet d'un protozoaire sur le microbiote intestinal) mais différents dans la forme (*Entamoeba* en milieu rural au Cameroun versus *Blastocystis* en milieu urbain en France). Nous avons ainsi tout naturellement initié une collaboration à partir de ce moment-là, en commençant par écrire un papier d'opinion s'interrogeant sur la nature des interactions entre protozoaires et bactéries intestinales (Chabe et al. 2017). Dans leur ensemble, nos résultats suggèrent en effet que, à l'instar des vers intestinaux, certains protozoaires pourraient avoir un effet positif sur la diversité microbienne intestinale, et donc indirectement sur la santé de leurs hôtes. Ces résultats ont été à mes yeux importants car les protozoaires intestinaux ont été jusqu'alors très largement ignorés, et leur diversité reste assez mal connue (Parfrey et al. 2014), tandis que nos résultats indiquent qu'il est important d'être attentif à leur diversité.

Identification des protozoaires intestinaux (Lokmer et al. 2019)

Pour explorer plus en détail cette diversité de protozoaires intestinaux, nous avons, avec Ana Lokmer (post-doctorante recrutée sur mon ANR JCJC), obtenu et analysé des données de métagénomique (donnant ainsi accès aux eucaryotes) sur les échantillons de 2013. Ana a tout d'abord mis au point une méthodologie pour mapper les données métagénomiques sur des génomes complets ou des données 18S pour dix protozoaires présents dans les bases de données publiques et identifier de manière robuste la présence de ces protozoaires intestinaux. Nous avons appliqué cette méthode à nos données ainsi qu'à des données publiques (d'individus vivant en Tanzanie, au Pérou, en Italie, ou aux États-Unis).

Nous avons trouvé que 72% des individus présentent au moins un protozoaire intestinal, mais seulement trois genres de protozoaires sont en prévalence importante dans les données : *Blastocystis spp.* (61%, avec cinq sous-types différents a priori non-pathogènes), *Entamoeba sp.* (46%, avec cinq espèces différentes a priori non-pathogènes) et *Enteromonas hominis* (11%).



Prévalence de deux protozoaires intestinaux majeurs (*Blastocystis* et *Entamoeba*) dans des populations aux modes de vie contrastés. Lokmer et al. (2019)

Le mode de vie des populations influence clairement la prévalence de ces protozoaires puisque dans les populations urbaines industrialisées, nous avons trouvé une fréquence de 15% de *Blastocystis*, de 0% d'*Entamoeba* et de 0% d'*Enteromonas*, par rapport à 40-96%, 38-90% et 0-50% en milieu rural traditionnel, respectivement. Ainsi, les bactéries ne sont pas les seules à avoir perdu de la diversité dans l'écosystème intestinal industrialisé. De manière intéressante, il y a une corrélation positive entre la présence de *Blastocystis* et d'*Entamoeba*, donc les individus les plus à même d'être colonisés par l'un le sont aussi par l'autre (Lokmer et al. 2019).

Ces prévalences élevées sont à mes yeux difficilement conciliables avec un effet pathogénique de ces espèces, et je pense qu'il serait important de cesser de considérer ces organismes systématiquement comme étant pathogéniques uniquement du fait que ce sont des eucaryotes. Encore aujourd'hui, de nombreuses équipes cherchent « désespérément » à mettre en évidence les effets délétères de ces micro-eucaryotes (malgré l'absence de résultats allant dans ce sens), sans remettre en question ce dogme. C'est ce point de vue que nous avons essayé de défendre dans notre papier d'opinion (Chabe et al. 2017).

Effet conjoint d'*Entamoeba* et de *Blastocystis* (Even et al. 2021)

Grâce aux données de métagénomique obtenues sur les données du gradient d'urbanisation de 2017 (dont les données 16S sont analysées plus bas en terme d'effet de l'urbanisation), nous avons, en collaboration avec Magali mais également avec Gaël Even de l'institut Pasteur de Lille, et grâce au travail de master 2 de Jules Rodrigues que j'ai encadré, pu regarder l'effet couplé d'*Entamoeba* et de *Blastocystis* sur le microbiote intestinal (tandis que les précédentes études, y compris la nôtre, n'avaient étudié l'effet que de l'un ou l'autre de

ces protozoaires). Appliquant la méthodologie de détection mise au point précédemment par Ana (Lokmer et al. 2019), nous avons tout d'abord pu montrer que la prévalence de ces protozoaires intestinaux ne diffère pas significativement entre milieu rural et urbain (prévalence globale d'*Entamoeba* de 45.8% et *Blastocystis* de 75.4%), malgré une tendance vers une prévalence un peu plus faible en milieu urbain. Ceci montre que l'écosystème eucaryote intestinal est relativement bien préservé à Yaoundé, la capitale du Cameroun. Nous avons également caractérisé ces protozoaires au niveau de l'espèce et/ou du sous-type et montré que les protozoaires les plus fréquents sont *Entamoeba coli* et *Entamoeba hartmanni*, ainsi que *Blastocystis* ST1, ST2 et ST3. De plus, la plupart des colonisations sont mixtes.

Nous avons ensuite pour la première fois étudié l'influence conjointe d'*Entamoeba* et *Blastocystis* sur le microbiote intestinal. Nous avons trouvé que *Blastocystis* a un effet plus important qu'*Entamoeba* sur la structure microbienne, bien que les deux protozoaires aient des effets cumulatifs. En effet, *Blastocystis* influence significativement plusieurs indices de diversité (richesse spécifique et abondance), tandis qu'*Entamoeba* influence seulement la richesse spécifique. En parallèle, *Blastocystis* influence la prévalence d'un plus grand nombre de taxons qu'*Entamoeba*. Ces résultats suggèrent que ces protozoaires interagissent potentiellement avec le microbiote bactérien par l'intermédiaire de différents mécanismes qu'il reste à caractériser.

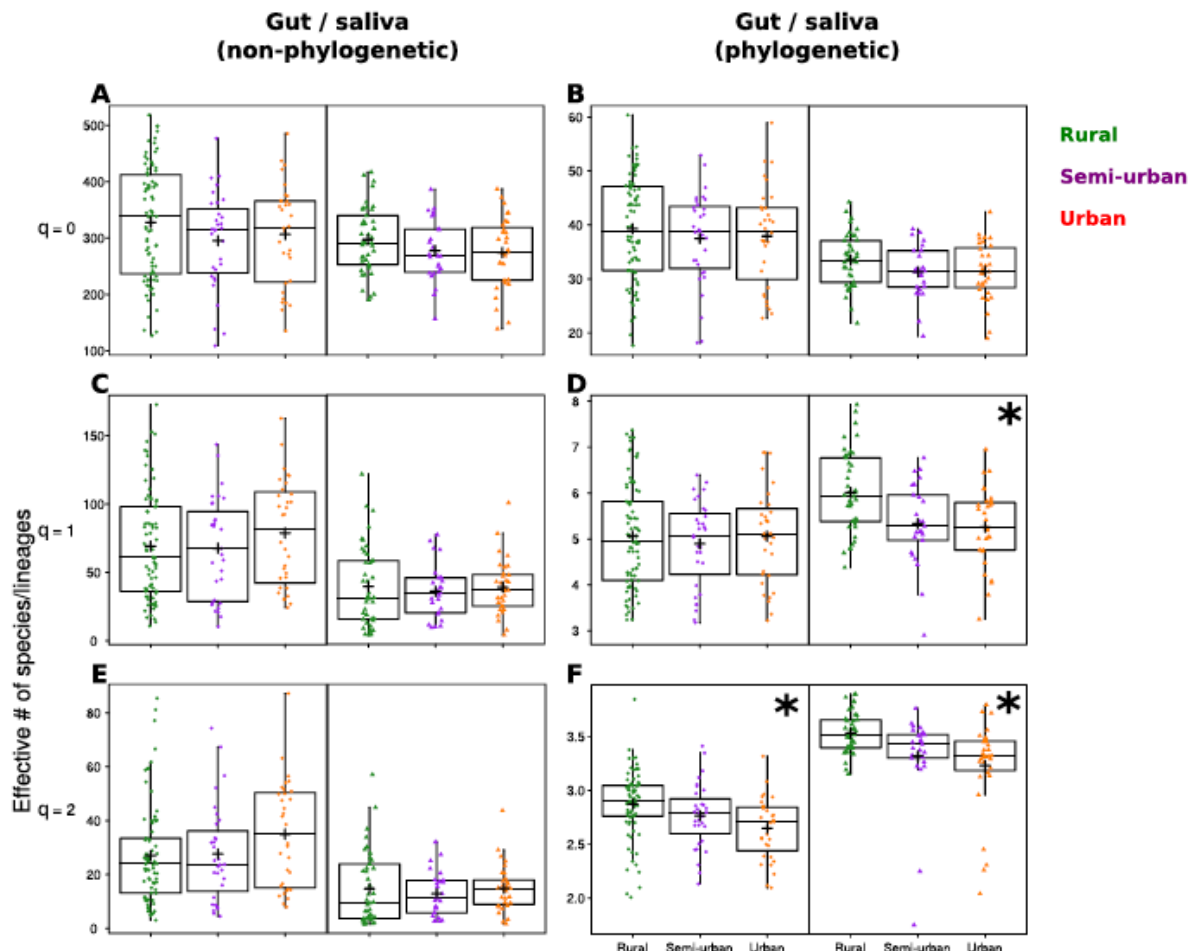
Effet de l'urbanisation (Lokmer et al. 2020)

L'objectif premier des données de 2017 était d'étudier la contribution relative des facteurs alimentaires et sanitaires à une échelle locale, sur le microbiote intestinal et oral. Nous avons ainsi, avec Ana Lokmer et Sophie Aflalo, une étudiante que j'ai encadrée en Master 2, tout d'abord caractérisé culturellement ces populations et mis en évidence des différences significatives entre groupes : très peu d'accès à une eau potable, pas d'électricité, plus d'animaux domestiques, plus de fumeurs, plus d'enfants et un moindre niveau d'éducation en milieu rural. Également, nous avons trouvé des différences alimentaires significatives : moins de fibres, plus de protéines animales et de sucre en milieu urbain. Nous avons également effectué des analyses parasitologiques des selles au Centre Pasteur du Cameroun et mis en évidence des différences significatives de prévalence de vers intestinaux entre populations : 17.5% en milieu rural, versus 6.3% et 3.1% en milieu semi-urbain et urbain, respectivement. Ces résultats témoignent des différences sanitaires importantes entre milieu urbain et rural. Cependant, nous n'avons pas trouvé de différences en termes de consommation d'antibiotiques (les trois populations en ayant une consommation occasionnelle non supervisée par un médecin) ou d'autres pratiques médicales (la quasi-totalité des individus sont nés par voie basse, ont été allaités et ne côtoient pas régulièrement de structures de santé). Par ailleurs, les trois populations présentent une proportion similaire d'individus maigres et en surpoids.

Concernant le microbiote (données 16S), Ana a identifié quatre types de communautés microbiennes différentes au niveau intestinal (enterotypes) et deux au niveau oral, avec une prévalence différente de chaque type dans chaque population. Plus précisément, au niveau

intestinal, nous avons observé des changements importants de composition entre milieux semi-urbain et rural d'une part et milieu urbain d'autre part (augmentation de *Bacteroides* et diminution de *Prevotella*). Ces bactéries sont connues pour digérer les graisses et les fibres, respectivement, et ces différences attestent donc très probablement des changements d'alimentation. Les différences de type salivaire sont plus difficiles à interpréter.

Par contre, malgré ces différences de composition, nous n'avons pas trouvé de différences significatives de diversité bactérienne entre les trois populations (à part pour un des sept estimateurs de diversité, significatif à la fois pour le milieu intestinal et oral), suggérant que la diminution de consommation de fibres et les meilleures conditions sanitaires en milieu urbain n'ont pas (ou pas encore) entraîné de perte de diversité bactérienne (faisant écho également aux résultats sur les procaryotes). Ces résultats suggèrent que la perte de diversité observée en milieu urbain industrialisé est peut-être due à d'autres changements de pratiques médicales ou sanitaires propres au milieu industrialisé. D'autres études contrastant milieu rural et urbain dans d'autres contextes culturels seront cependant nécessaires pour valider cette hypothèse.



Diversité bactérienne intestinale et salivaire en milieu rural (vert), semi-urbain (violet) et urbain (orange).

A-C-E : indices de diversité non-phylogénétique ; B-D-F : indices de diversité phylogénétique.

A-B : indices de richesse spécifique ; C-F : indices d'abondance plus ou moins pondérés par les espèces dominantes (voir Boite 2). Lokmer et al. (2020).

Enfin, nous avons montré que la richesse du microbiome intestinal est faiblement mais significativement corrélée à celle du microbiome salivaire à l'échelle de l'individu, et que les microbiomes salivaire et intestinal partagent plus d'espèces au sein des individus qu'entre individus, et ce d'autant plus en milieu urbain. Ces résultats pourraient être le reflet de différences d'exposition des individus aux bactéries environnementales, ou bien être le fruit d'une certaine perméabilité entre ces deux microbiotes.

2. Coévolution humain-bactérie à une échelle mondiale

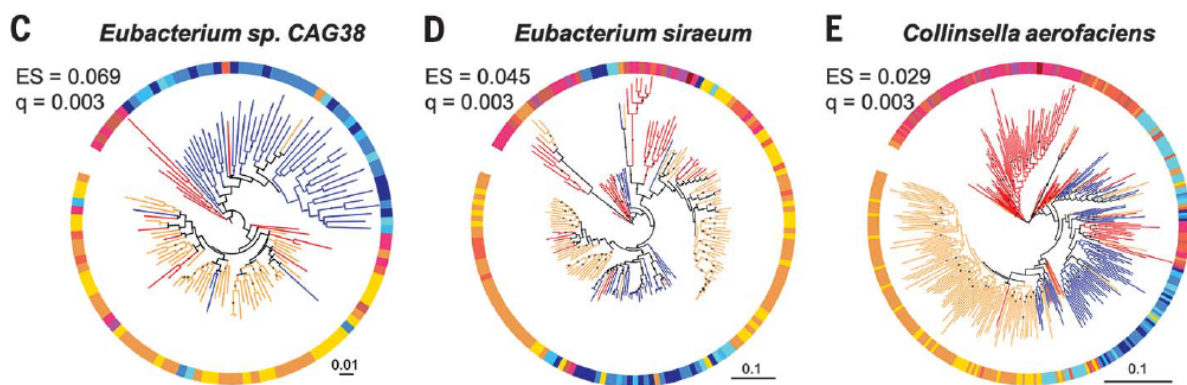
Une des motivations initiales de mon projet de recherche était d'étudier si certaines bactéries intestinales se sont adaptées au régime alimentaire (ou de manière générale au mode de vie) de leurs hôtes humains, c'est-à-dire de trouver des exemples de coévolution entre humains et bactéries. Une manière de répondre à cette question est de chercher s'il existe des associations significatives entre souches bactériennes et populations humaines, autrement dit si certaines espèces bactériennes présentent des patrons de cophylogénie avec leurs hôtes (i.e., un arbre phylogénétique des souches bactériennes d'une espèce reflétant l'arbre de ses hôtes). De tels patrons ont notamment été observés à l'échelle des mammifères (Groussin et al. 2017; Gaulke et al. 2018; Youngblut et al. 2019) et des primates (Moeller et al. 2016), mais pas à une échelle intra-spécifique.

Mais ces patrons de cophylogénie ne résultent pas nécessairement de processus coévolutifs. En effet, la dispersion microbienne limitée entre populations/espèces peut résulter de leur isolation géographique et entraîner de la cophylogénie (Groussin et al. 2020). Cependant, ce qui nous intéresse ici n'est pas tant le niveau global de cophylogénie entre hôte et bactéries intestinales, mais sa variance (c'est-à-dire l'observation de bactéries qui ont des patrons de cophylogénie particulièrement forts par rapport aux autres bactéries). De telles observations peuvent résulter de plusieurs mécanismes : i) un déterminisme génétique (par l'hôte) plus fort pour certaines espèces bactériennes, ii) une transmission verticale plus forte pour certaines espèces bactériennes (par rapport à la transmission horizontale), entraînant une isolation de ces espèces qui suit encore plus l'arbre des hôtes ; iii) une coévolution à proprement parler entre hôtes et bactéries ayant généré une dépendance réciproque entre ces organismes ; ou iv) particulièrement pour le cas de l'espèce humaine, une adaptation locale des souches bactériennes à des facteurs climatiques, alimentaires et/ou culturels qui suivent également l'arbre phylogénétique des hôtes, étant donné que cet arbre reflète lui-même en partie la différenciation géographique et culturelle entre populations humaines.

Tandis que j'étais de moins en moins convaincue que de tels patrons puissent être observés à l'échelle intra-spécifique, et encore moins à l'échelle du Cameroun, j'ai suivi une conférence de Taichi Suzuki (à l'époque post-doctorant dans l'équipe de Ruth Ley, Max Planck Tübingen, Allemagne, et actuellement professeur à l'université d'Arizona, USA) testant précisément s'il trouvait des patrons de co-phylogénie chez l'humain à une échelle mondiale. Pour cela, il a compilé des données appariées de métagénomique bactérienne et de diversité génétique humaine chez les mêmes individus. Ayant moi-même collecté de telles données

appariées (justement à terme pour tester cette idée), j'ai contacté Taichi en 2021 pour lui proposer de collaborer. J'ai ainsi eu l'opportunité de rajouter mes données de 115 individus au Cameroun aux 171 Gabonais, 284 individus asiatiques (Vietnam et Corée) et 269 individus européens (Allemagne et Royaume-Uni) de cette analyse (Suzuki et al. 2022).

Taichi a tout d'abord sélectionné les 59 espèces de bactéries ou d'archées présentes chez au moins 100 individus à l'échelle mondiale, et chez au moins 10 individus par population. Contrastant les phylogénies bactériennes (obtenues avec un ensemble de gènes spécifiques, ou à partir de MAGs) et humaines (obtenues avec plus de 20,000 SNPs), Taichi a montré que 36 des 59 espèces testées présentent des cophylogénies significatives à l'échelle mondiale. De manière intéressante, il y a un enrichissement de Firmicutes (et une déplétion de Bacteroidetes) parmi ces espèces, et plusieurs espèces d'*Eubacterium* présentent les scores les plus élevés.



Arbre phylogénétique bactérien pour trois espèces ayant de forts coefficients de cophylogénie (valeur d'ES). Les souches bactériennes asiatiques sont en bleu, africaines en rose et européennes en jaune/orange. Suzuki et al. (2022)

Taichi a également obtenu des indices de cophylogénie pour chaque pays et montré que 20 des 59 espèces étaient significatives pour au moins un pays, mais après correction pour tests multiples, seul *Prevotella copri* reste significative, au sein du Gabon. Pour cette espèce, il est donc possible d'exclure un effet confondant de l'environnement.

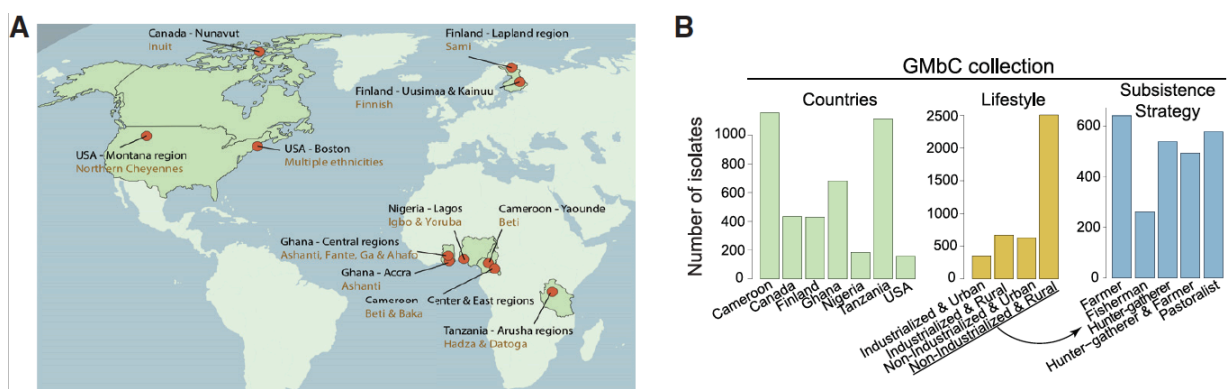
A mes yeux, le résultat peut être le plus intéressant de cette étude est la corrélation négative observée entre les indices de cophylogénie et la taille des génomes bactériens, comme précédemment observé chez les symbiontes de certains insectes (McCutcheon et al. 2011). Cette diminution de taille de génome, couplée également à une augmentation de la proportion du génome dédié à des fonctions essentielles et une moins bonne survie des bactéries en présence d'oxygène et à 27°C semble indiquer que ces espèces ont évolué vers une dépendance accrue à l'hôte. Reste à savoir, pour chacune de ces espèces bactériennes, si les patrons de cophylogénie observés sont dus à une transmission horizontale réduite ou à une possible co-évolution avec l'hôte.

3. Microbiote et transfert horizontal de gènes

Les micro-organismes peuvent évoluer face à des changements de leur environnement en modifiant la composition et la diversité de leur écosystème, et ainsi être *globalement* mieux adaptés à leur milieu. Mais les bactéries peuvent également évoluer *individuellement* par **transfert horizontal de gènes**, un mécanisme par lequel un organisme intègre du matériel génétique venant d'un autre organisme, sans qu'il n'y ait de relation de descendance entre eux. Ces transferts horizontaux de gènes (THGs) sont connus pour avoir un impact majeur sur l'évolution des espèces bactériennes (Boto 2010; Smillie et al. 2011). Explorer ces mécanismes de transferts, leur dynamique et leur rôle adaptatif est essentiel pour mieux comprendre le fonctionnement des écosystèmes bactériens. Cependant, caractériser les THGs entre bactéries intestinales à partir de données de métagénomique est particulièrement difficile, puisque par définition, une séquence transférée est trouvée identique sur deux fonds génomiques différents et est donc particulièrement difficile à assembler correctement.

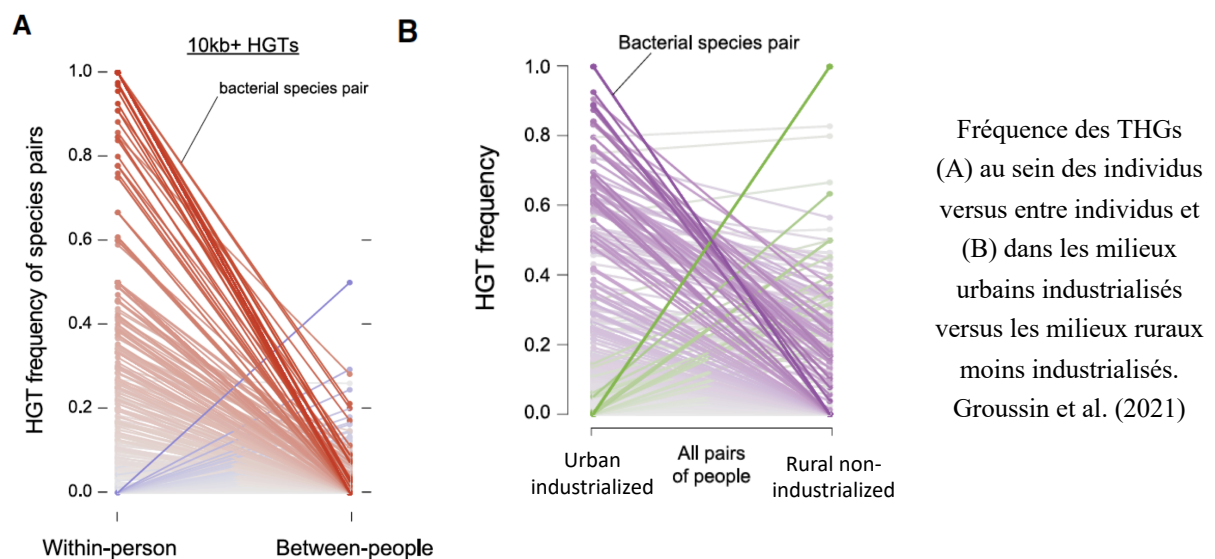
Transfert horizontal de gènes et industrialisation (Groussin et al. 2021)

Afin de quantifier les THGs au sein et entre microbiotes intestinaux, Mathieu Groussin et Mathilde Poyet (tous deux postdoc au MIT, USA, dans le laboratoire d'Eric Alm au moment de cette collaboration, et actuellement chercheurs à l'université de Kiel, Allemagne), ont mis en place un projet visant à collecter et congeler immédiatement des échantillons fécaux dans divers environnements plus ou moins industrialisés pour pouvoir ensuite les cultiver, les isoler, et ainsi séquencer le génome de souches individuelles (isolats). C'est dans ce contexte que nous sommes rentrés en contact en 2016 et avec cet objectif en tête que Mathieu a fait sa première mission de terrain avec moi au Cameroun, en 2017. Par la suite, Mathieu et Mathilde ont obtenu des données similaires dans plusieurs autres pays, leur permettant ainsi de constituer une banque de données génomiques sur 4,149 souches venant de 37 individus de 14 populations différentes. Ces données ont servi de base au Global Microbiome Conservancy Project, <https://openbiome.org/the-global-microbiome-conservancy-gmbc/>.



Répartition géographique des échantillons collectés et caractérisation géographique et culturelle des souches obtenues. Groussin et al. (2021)

Mathieu a ainsi cherché des segments génomiques strictement identiques entre paires d'espèces (segments récents de 10kb ou plus, et segments plus anciens de 500bp ou plus) et a montré que les THGs de 10kb (correspondant à des transferts datant d'il y a 0-100 ans) sont très fréquents et sont notamment observés plus souvent au sein des individus qu'entre individus (en corrigeant pour la phylogénie qui influence fortement la probabilité de transferts entre espèces). De manière intéressante, ces THGs ont été trouvés à plus grande fréquence dans les populations urbaines industrialisées par rapport aux populations rurales moins industrialisées (là encore corrigeant pour des facteurs confondants tels que la phylogénie ou l'abondance des espèces), suggérant que l'industrialisation génère de nouvelles contraintes sélectives.



Fréquence des THGs (A) au sein des individus versus entre individus et (B) dans les milieux urbains industrialisés versus les milieux ruraux moins industrialisés. Groussin et al. (2021)

En terme fonctionnel, les gènes transférés sont dans leur ensemble les plus différents entre les populations rurales non-industrialisées et les trois autres (urbaines industrialisées ou non, rurales industrialisées). Tandis que les populations rurales non-industrialisées sont caractérisées par plus de transfert de gènes de digestion de carbohydrates (CAZymes), les populations urbaines industrialisées sont marquées par une prévalence plus forte de transferts correspondant à des éléments mobiles (phages, plasmides, transposons). Il est également intéressant de noter que les transferts de gènes de résistance aux antibiotiques ne sont pas les plus fréquents en milieu urbain industrialisé mais chez les Dagota, une population pastorale rurale africaine qui utilise des antibiotiques pour leur bétail.

Consommation d'algues et transfert horizontal de gènes (Mathieu et al. 2023)

Bien que la majorité des THGs s'effectuent entre espèces bactériennes phylogénétiquement proches occupant les mêmes niches écologiques, comme au sein du microbiote intestinal, des transferts peuvent aussi avoir lieu entre bactéries écologiquement distinctes (Smillie et al. 2011). C'est ce qui a été observé pour un ensemble de gènes de digestion d'un sucre complexe, le porphyran, un polysaccharide qui constitue la paroi cellulaire de l'algue rouge *Porphyra sp.* et qui se retrouve dans le nori, préparation culinaire utilisée pour faire les makis et les sushis. Il s'avère que cet ensemble d'une quarantaine de gènes, appelé

PUL porphyran (pour Polysaccharide-Utilization Loci), a été transféré depuis une bactérie marine vivant à la surface de ces algues vers une bactérie intestinale de populations japonaises (Hehemann et al. 2010). Ces bactéries porteuses du PUL porphyran (PUL-por) ont ainsi acquis la capacité de dégrader en sucres simples la paroi de ces algues rouges consommées dans ces populations.

William Helbert (CERMAV, Grenoble), biochimiste spécialiste des enzymes de digestion des sucres ayant participé à cette étude, m'a contactée en 2020 pour initier une collaboration afin de caractériser plus finement ce cas de transfert horizontal. William a en effet étudié expérimentalement les enzymes du PUL-por et montré que le PUL-por contient tout l'ensemble des gènes nécessaires pour digérer le porphyran de façon autonome, incluant des enzymes capables de dégrader non seulement la fraction non-méthylée du porphyran mais également la fraction méthylée (Mathieu et al. 2023). Il a ensuite effectué une recherche systématique (par blast) de ces gènes dans un grand nombre de données métagénomiques publiques, puis a souhaité mettre en place une collaboration pour automatiser cette approche par des outils bio-informatiques. Afin d'intégrer l'analyse d'un large catalogue de données métagénomiques mondiales que William n'avait pas interrogé (Almeida et al. 2021), j'ai co-supervisé le stage de master 1 de Mickael Ivanoff avec Damien de Vienne (LBBE) et Vincent Navratil (PRABI). Les résultats présentés ici viennent donc à la fois de ce que William a fait en amont, de ce qui a été effectué par Mickael pendant son stage, et de ce que j'ai fait pour refaire et compléter ces analyses.

Cette étude sur l'histoire évolutive de ce transfert singulier de gènes avait pour objectif de répondre à deux questions principales :

i) Combien d'espèces bactériennes intestinales portent le PUL-por, et quelles sont-elles ? La première étude de Hehemann et al. (2010) avait identifié *Bacteroides plebeius* comme réceptrice du transfert, puis l'étude de Pudlo et al (2022) sur la croissance in vitro de bactéries de 240 donneurs d'échantillons fécaux a montré que quatre espèces de *Bacteroides* avaient la capacité de croître en présence de porphyran. Cependant, aucune étude à large échelle des génomes bactériens existants n'avait encore été réalisée.

ii) Quelles populations humaines, en dehors des populations japonaises, portent le PUL-por dans son microbiote intestinal, et en quelles proportions ? En effet, bien que ces algues soient aujourd'hui principalement consommées par les populations asiatiques, la preuve la plus ancienne que les humains ont récolté et stocké *Porphyra sp.* a été trouvée dans le site archéologique de Monte Verde au Chili (12,300BP) (Dillehay et al. 2008). Les pré-incas et incas intégraient également les algues rouges à leur alimentation (Michanek 2013). De plus, il a récemment été découvert que les populations européennes (de l'Espagne à l'Ecosse en passant par la Lituanie) consommaient des algues depuis le Néolithique jusqu'au Moyen-Âge (Buckley et al. 2023). La consommation d'algues est donc attestée depuis des millénaires en Amérique du Sud et en Europe. Jusqu'à présent, le PUL-por a uniquement été observé dans des populations japonaises et chinoises, mais n'a pas été trouvé au Danemark, en Espagne, aux Etats-Unis, en Tanzanie ou au Brésil (étude d'un ensemble de 2,440 individus) (Pudlo et al.

2022). La question de l'étendue géographique et de la prévalence du PUL-por dans différentes populations humaines mondiales reste donc en partie ouverte.

Afin de répondre à ces questions et d'obtenir le nombre et l'identité taxonomique des espèces bactériennes portant PUL-por, ainsi que sa prévalence à échelle mondiale, nous avons utilisé plusieurs types de données :

1- Les génomes d'isolats bactériens, qui présentent l'avantage d'avoir été séquencé après culture et isolation, et donc qui ne nécessitent pas d'étape d'assemblage des génomes pouvant entraîner des erreurs (voir Boite 3). Le fait d'identifier le PUL-por dans un isolat est donc une preuve formelle de sa présence chez cette espèce.

2- Les génomes bactériens assemblés à partir des métagénomiques (MAGs, voir Boite 3), qui présentent l'intérêt d'être en plus grand nombre et de représenter une plus large diversité géographique que les isolats ci-dessus, mais dont l'étape d'assemblage peut avoir entraîné des erreurs.

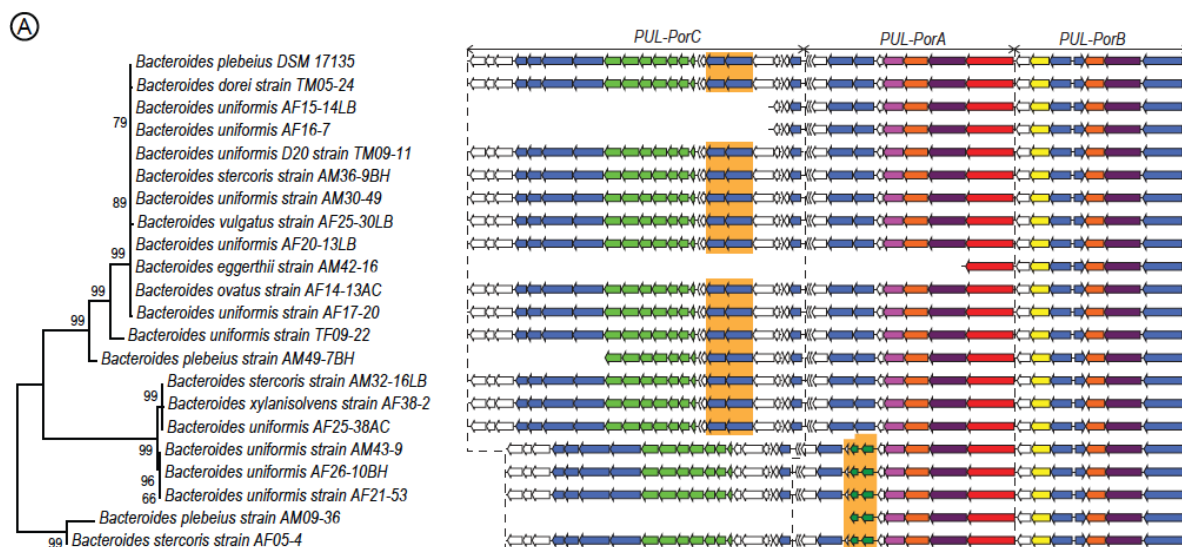
3- Les données métagénomiques non assemblées (SRA, pour short reads), présentant l'avantage d'être en encore plus grand nombre que les MAGs (car seulement une minorité d'études métagénomiques obtiennent des MAGs à partir de leur données), mais ne donnant pas accès à l'information taxonomique de quelle bactérie est porteuse au sein d'un individu.

Concernant des aspects pratiques, les données SRA sont en très grande quantité dans la littérature, correspondant à des volumes de données astronomiques, et ne sont pas compilées, donc non disponibles pour téléchargement à un endroit unique. Pour cette raison, pour le scan de la présence du PUL-por à large échelle géographique, nous avons privilégié l'utilisation des MAGs, également motivés par le fait qu'une méta-analyse venait justement de compiler plus de 278,000 MAGs venant de 13,587 individus différents, ainsi que plus de 10,000 isolats bactériens (Almeida et al. 2021). Il s'agit d'une partie de la base de données MGnify (<https://www.ebi.ac.uk/metagenomics>). Le fait d'interroger les MAGs par rapport aux SRA permet également d'être plus efficace d'un point de vue bio-informatique puisque les MAGs sont non redondants entre eux (tout du moins au sein des individus) et correspondent donc à une quantité de données bien moindre à interroger que les SRA.

La première étape a été de scanner les 10,969 isolats bactériens du catalogue d'Almeida pour étudier la distribution phylogénétique du PUL-por. Nous avons ainsi blasté les séquences codantes de trois parties différentes du PUL-por (PUL-porA, PUL-porB et PUL-porC, définies par leur propriété transcriptomique, je ne rentrerai pas dans les détails ici mais cela correspond à 13-19kb chacune). Nous avons trouvé 22 souches bactériennes positives, correspondant au moins à 8 espèces différentes, toutes isolées à partir d'individus est-asiatiques. De manière intéressante, ces mêmes espèces existent également sans PUL-por, donc la présence de PUL-por est polymorphe dans les populations bactériennes. Toutes ces espèces correspondent à des *Bacteroides*, mais toutes les espèces de *Bacteroides* ne portent pas le PUL-por, sans qu'il ne semble y avoir de corrélation entre la présence de PUL-por et leur abondance dans l'écosystème intestinal. Cette observation suggère qu'il y a eu (et continue certainement d'y avoir) de

nombreux transferts du PUL-por entre bactéries du microbiote depuis sa première acquisition à partir d'une bactérie marine. Une conséquence de ces observations est que la capacité métabolique d'un individu à digérer le porphyran ne peut pas être prédite à partir de sa composition taxonomique.

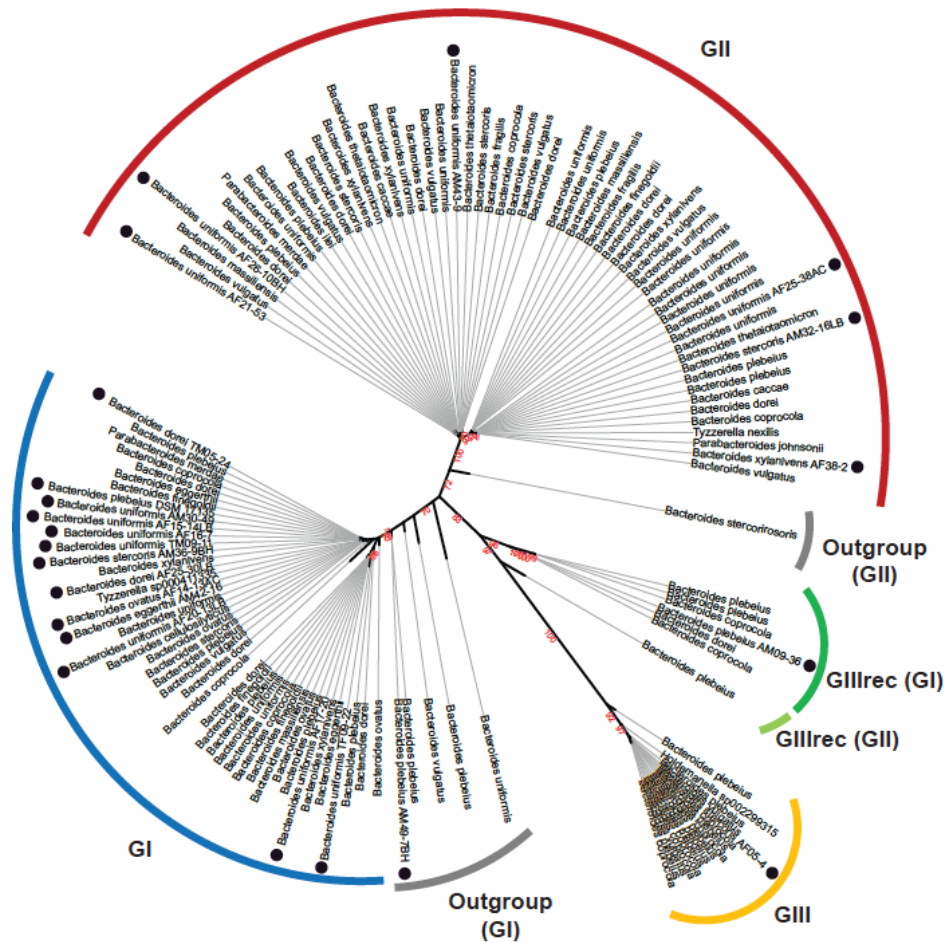
William a ensuite étudié la conservation des gènes de PUL-por entre bactéries et observé que l'organisation de PUL-porB était conservée chez toutes les isolats bactériens, tandis que celle de PUL-porA et PUL-porC était conservée seulement dans 16/21 et 14/18 souches, respectivement (les autres souches n'ayant pas de séquences, ou des événements de recombinaison par rapport à la référence de *B. plebeius*).



Organisation des gènes du système de dégradation du porphyran trouvés dans 22 isolats bactériens par rapport à la souche référence de *Bacteroides plebeius*. L'arbre phylogénétique a été obtenu à partir des séquences géniques concaténées du PUL-por (40 gènes). (Mathieu et al. 2023)

Nous avons ensuite blasté les séquences codantes des six gènes de PUL-porB (le plus conservé) pour scanner le catalogue de 278,263 MAGs d'Almeida (2021). Nous y avons également rajouté 29,082 MAGs venant de l'étude de Kim et collaborateurs (2021) incluant des individus d'origine japonaise, indienne et coréenne, pays sous-représentés dans le catalogue d'Almeida. Nous avons trouvé au total, après fusion avec les isolats bactériens, 130 souches positives non redondantes, correspondant à 12 espèces de *Bacteroides*, ainsi que deux nouveaux genres appartenant au phylum Bacillota (*Tyzzarella* sp. et *Holdemanella* sp.).

Cet arbre phylogénétique a révélé une structuration du PUL-porB en trois grands clades, GI, GII et GIII. Des formes recombinantes sont également observées à différents niveaux de l'arbre (notamment GI outgroup et GII outgroup). Du fait de la plus grande quantité de souches recombinantes entre GIII et GI/GII, nous avons créé un quatrième groupe : GIIIrec. Il ne semble pas y avoir de corrélation entre ces groupes et l'identité taxonomique des bactéries porteuses.



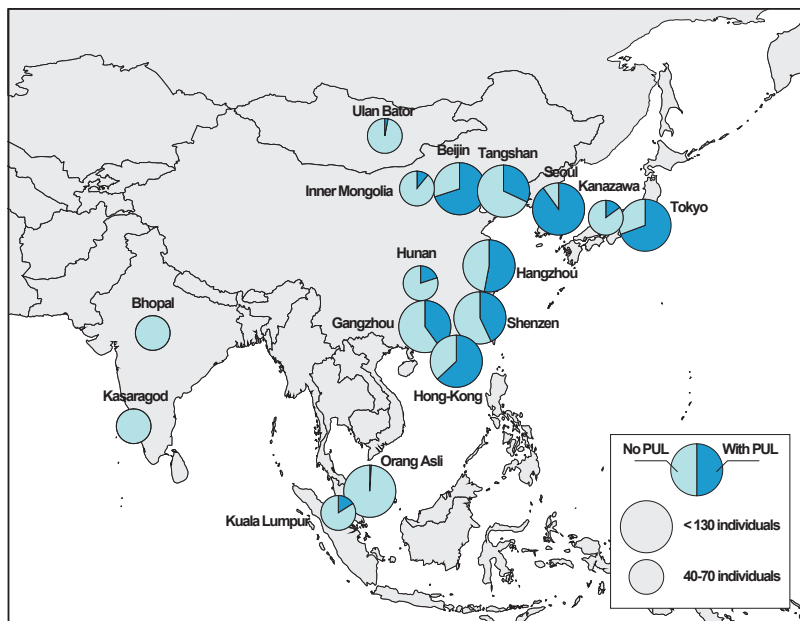
Arbre phylogénétique à partir des séquences géniques concaténées du PUL-porB (6 gènes). (Mathieu et al. 2023)

Nous avons ensuite cherché à estimer la prévalence du PUL-porB par population à partir du catalogue d'Almeida. Nous avons ainsi observé que sur les 30 pays inclus (13 en Europe, 7 en Asie, 6 en Amérique, 2 en Océanie et 2 en Afrique), seul le Japon et la Chine avaient des prévalences supérieures à 1%. Cependant, sans rentrer dans les détails, nous nous sommes rendus compte que ces données, dans l'optique d'être une base de données de référence des MAGs intestinaux, avait appliqué des filtres de qualité très stricts sur les données originales (notamment en termes de complétude des génomes et de score de qualité), diminuant d'un facteur 10 le nombre de contigs final moyen par individu. Cette réduction drastique de données aboutit à une sous-estimation de la prévalence du PUL-porB, donc nous avons conclu que le catalogue d'Almeida n'est pas approprié pour notre analyse.

Nous sommes donc repartis des données initiales, en téléchargeant manuellement les différents projets de séquençage disponibles contenant des MAGs. Nous avons ainsi observé des prévalences de 33% en Chine, 25% au Japon et 26% en Corée du Sud, tandis que tous les autres pays avaient des prévalences nulles ou inférieures à 1% (Etats-Unis, Italie, Tanzanie, Madagascar, Pérou, Inde, Bangladesh).

Pour avoir une meilleure couverture géographique, nous avons ensuite téléchargé et scanné des données de SRA (short reads, donc non assemblées). Pour les interroger, William a conçu une sonde de 54pb permettant de discriminer si les individus appartiennent au groupe

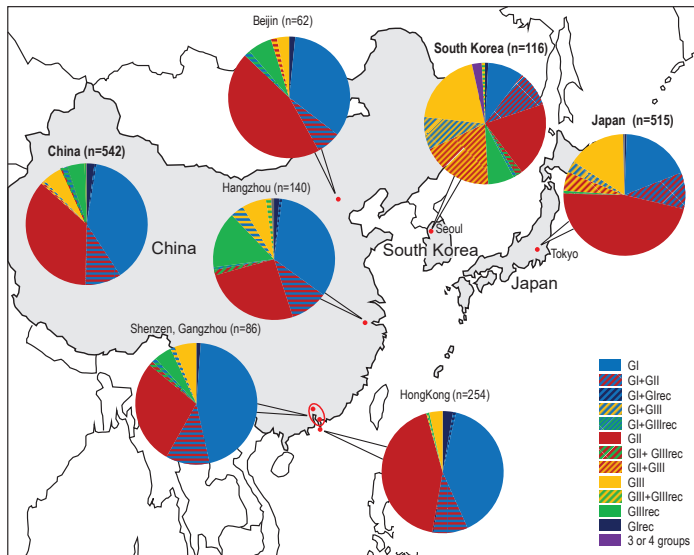
GI, GII, GIII ou GIIIrec. Nous avons ainsi blasté les données SRA de 4,617 individus venant de 11 pays différents, et obtenu la carte de prévalence ci-dessous.



Proportion d'individus ayant des SRA mappant sur le PUL-porB, regroupés par ville. (Mathieu et al. 2023)

Nous pouvons d'abord remarquer que la prévalence est plus importante à partir des SRA que des MAGs : 50% (versus 33%) en Chine, 72% (versus 25%) au Japon, et 89% (versus 26%) en Corée du Sud. Ce dernier pays, jusqu'alors non analysé pour sa prévalence de PUL-por, représente donc le pays avec la valeur la plus élevée de tout le jeu de données. Dans l'ensemble, la présence du PUL-por semble bien se limiter à l'Asie de l'Est côtière (Japon, Chine, Corée) et ne pas être présente dans d'autres pays asiatiques (3% tout au plus en Mongolie) ou ailleurs dans le monde. Notons toutefois que la prévalence à Kuala Lumpur est de 16 % (et de 0 % dans les zones rurales de Malaisie), reflétant probablement des flux migratoires venant de Chine.

Nous avons ensuite étudié la diversité génétique au sein du PUL-porB en Asie de l'Est en utilisant les sondes de 54pb. Pour ces analyses, nous n'avons conservé que les individus ayant plus de 10 lectures alignées sur PUL-porB afin de ne pas créer de biais en fonction de la couverture. Nous avons ainsi observé que le groupe GI est majoritaire en Chine (52% des individus), tandis que le groupe GII est majoritaire au Japon (62% des individus), et le groupe GIII est majoritaire en Corée du Sud (60% des individus). Les différences de prévalence des différents groupes se sont avérées être significatives entre pays et entre certaines villes, montrant que le PUL-porB est structuré géographiquement. Le fait que tous les groupes PUL-porB (GI, GII, GIII et GIIIrec) soient trouvés dans toutes les villes étudiées montrent cependant que les souches bactériennes circulent assez largement à cette échelle géographique.



Proportion des différents groupes de PUL-porB au sein des individus positifs. (Mathieu et al. 2023)

De manière intéressante, nous avons observé que parmi les individus positifs, 60% avaient des lectures attribuées à un seul groupe, et 21% avaient un groupe dominant, défini comme ayant une prévalence égale ou supérieure à 80% des lectures. Cette observation recoupe les résultats obtenus plus largement dans le microbiote intestinal, avec 80% des espèces bactériennes présentant une souche unique ou dominante (Truong et al. 2017).

En conclusion, ce cas intéressant de transfert horizontal depuis l'environnement vers des bactéries intestinales puis au sein des bactéries intestinales soulève encore de nombreuses questions. L'origine du PUL-por n'a notamment pas encore été établie (ni le nombre de transferts ayant eu lieu), et nous explorons la possibilité d'interroger des bases de données plus larges pour trouver quelle bactérie environnementale porte le PUL porphyran le plus proche de celui trouvé dans le microbiote humain. Pour cela, nous privilégierons une approche de blast protéique et non nucléique. Des méthodes de réconciliation (comparaison de l'histoire évolutive du PUL-porphyrane et de l'histoire évolutive des espèces qui le portent) pourraient également être utilisées pour mieux caractériser les dynamiques d'acquisition, de transferts et de pertes de ces gènes. Enfin, le cas du porphyran a été identifié de façon fortuite et il n'est certainement pas isolé. L'alimentation met en permanence en contact des bactéries de l'environnement et du microbiote intestinal, et il serait intéressant d'estimer la prévalence et les fonctions des gènes transférés de l'environnement vers l'intestin humain. L'accumulation rapide de données métagénomique d'environnements de plus en plus diversifiés peut permettre de poser cette question de manière plus systématique. Le cas du PUL-por servirait alors de témoin positif.

4. Travail en cours

La question de l'évolution du microbiote en réponse à des changements d'alimentation ne se pose pas seulement chez l'humain, mais également chez certaines espèces animales dont les régimes alimentaires ont été récemment modifiés par l'anthropisation grandissante de leurs habitats (White et al. 2023). C'est dans ce contexte scientifique que Virginie Rougeron (IRL REHABS) m'a proposé de collaborer sur la question de l'effet de l'anthropisation sur l'état de

santé des babouins chacma (*Papio ursinus*) en Afrique du Sud. Nous avons ainsi, avec Virginie, mais également avec Franck Prugnolle (IRL REHABS) et Benjamin Rey (LBBE), démarré la co-supervision de la thèse de Célia Lacomme fin 2023.

Cette thèse a pour objectif général d'évaluer l'impact de l'anthropisation, caractérisée non seulement par des changements alimentaires, mais aussi par une plus forte pollution de l'environnement et une plus forte densité d'infrastructures (entraînant des contacts accrus avec les humains), sur l'état de santé des babouins chacma au sens large (charge et diversité parasitaire, microbiote intestinal, marqueurs de stress et d'inflammation). Cette espèce est en effet présente dans la région de la Garden Route en Afrique du Sud, avec certaines troupes vivant dans un environnement encore très naturel (en forêt), certaines populations vivant en milieu naturel ou péri-urbain mais consommant occasionnellement des aliments d'origine anthropique (notamment en se nourrissant dans les poubelles de parcs nationaux), et certaines troupes faisant des incursions régulières dans des maisons en milieu urbain. Ce gradient d'anthropisation a été montré comme associé à un changement de composition isotopique des fèces, avec une alimentation enrichie en produits protéinés et sucrés en milieu urbain (Kennedy Overton et al. 2023), montrant ainsi qu'il peut s'agir d'une quantité non négligeable de nourriture d'origine anthropique.

Ce projet est particulièrement interdisciplinaire et intégratif, et parmi toutes les données collectées (génomiques, hormonales, micro-plastiques, isotopes), je serai plus précisément en charge de superviser Célia sur l'analyse du microbiote intestinal (16S), cherchant à caractériser l'influence de diverses variables d'anthropisation. Nous pouvons notamment nous demander si c'est l'alimentation anthropique (caractérisée par l'analyse isotopique des fèces) ou le degré d'altération de l'habitat par urbanisation (béton et/ou pollution caractérisés par l'étude du paysage écologique et l'analyse de la quantité de microplastique dans les fèces) qui affecte le plus le microbiote intestinal. Une difficulté de ce projet sera de prendre en compte le fait que ces troupes ne sont pas habituées, et donc nous n'avons pas de données individualisées.

Célia a ainsi collecté courant 2024 des fèces pour environ 500 individus dans une trentaine de troupes sauvages correspondant à des environnements contrastés (naturel, à proximité de fermes, peri-urbain et urbain), représentant ainsi un gros travail de collecte. J'ai eu la chance, avec Benjamin, de participer à une partie de la campagne d'échantillonnage en avril 2024. Célia a ensuite effectué le traitement moléculaire des échantillons et les analyses génétiques viennent de débiter.

Autres projets non développés

Dans la continuité de précédents travaux dont je n'ai pas parlé ici sur l'influence de l'organisation sociale sur la diversité génétique, j'ai co-encadré la thèse de Nina Marchi (2014-2017), intitulée : « A la croisée de l'anthropologie et de la biologie évolutive : diversité génétique et comportements migratoires en Asie intérieure ». Cette thèse a permis de décrire l'histoire démographique de populations nouvellement échantillonnées par l'équipe, en Sibérie et en Mongolie, et de poser de nouvelles questions sur les relations entre comportements culturels et diversité génétique.

Dans une première partie, Nina a exploré le rôle respectif de la géographie, l'ethnicité et l'organisation sociale sur les patrons de différenciation génétique du chromosome Y et de l'ADN mitochondrial (Marchi et al. 2017). Cette étude a confirmé que, du fait de la patrilocalité (règle de résidence commune à toutes les populations étudiées et entraînant principalement des migrations de femmes entre populations), la différenciation des populations est réduite pour l'ADN mitochondrial par rapport à celle du chromosome Y. D'autre part, les populations patrilinéaires présentent une diversité du chromosome Y plus faible que les cognatiques, ainsi qu'une structuration génétique particulière. La diversité mitochondriale est par contre similaire entre patrilinéaires et cognatiques, illustrant que la patrilinéarité influence exclusivement la diversité génétique masculine. Nina a également calculé l'âge génétique des groupes ethniques patrilinéaires (à partir de marqueurs du chromosome Y) et montré que cet âge biologique est plus ancien que les âges historiques, suggérant que l'ethnie, du moins en Asie intérieure, est en partie une construction sociale, plutôt qu'une entité entièrement biologique.

Dans une deuxième partie, Nina s'est intéressée à l'influence des pratiques matrimoniales sur la consanguinité génétique (Marchi et al. 2018). Elle a notamment cherché à comprendre quelle relation existe entre le taux d'exogamie géographique des populations (en d'autres mots, la distance à laquelle les individus choisissent leurs conjoints) et la consanguinité génétique des individus. Elle a testé si : i) les populations les plus exogames sont les moins consanguines car elles choisissent des conjoints distants géographiquement et donc moins similaires génétiquement, ou au contraire ii) si les populations les plus consanguines mettent en place des pratiques d'exogamie afin d'éviter les effets néfastes de la consanguinité. Cette étude a donc la particularité d'avoir utilisé conjointement des données ethnologiques et génétiques, ceci pour un grand nombre d'individus. De manière surprenante, nous n'avons pas trouvé de différences de consanguinité entre populations, malgré des taux d'exogamie variables. A l'échelle des individus, nous avons observé une plus grande consanguinité des individus descendants de couples exogames (pour des distances inférieures à 40km) par rapport aux descendants de couples endogames. Ce résultat peut s'expliquer soit par une mauvaise corrélation entre distance génétique et géographique à l'échelle considérée, soit par un biais comportemental des individus consistant à choisir parmi les individus géographiquement éloignés des individus apparentés (notamment par la lignée maternelle). Ce biais d'apparentement maternel est assez

probable car il a déjà été décrit par les ethnologues (alliances croisées). Nous avons en effet trouvé qu'un certain nombre d'individus descendants de couples exogames étaient consanguins, favorisant la deuxième hypothèse. Ces résultats illustrent que, chez l'humain, des comportements culturels d'alliance peuvent s'opposer aux attendus de la biologie évolutive.

En dehors de ces questions spécifiques, ce travail a permis de décrire les attendus de diversité génétique neutre dans ces populations d'Asie intérieure et donc de fournir une hypothèse nulle contre laquelle tester l'existence de sélection naturelle. Ainsi, les résultats de cette thèse sont précieux pour l'analyse de sélection naturelle effectuée sur ces mêmes populations pour la thèse de Johanne Adam.

En parallèle, j'ai continué ma collaboration avec Shelly Masi, de l'équipe de primatologie de l'UMR7206, afin d'obtenir des données génétiques microsatellites pour des gorilles de l'Ouest habitués et étudier les profils d'apparentements entre individus (Masi et al. 2021). En contrastant ces données avec des données observationnelles (et en collaboration également avec Frédéric Austerlitz, UMR7206), nous avons ainsi montré que, bien que la majorité des enfants des quatre groupes étudiés sont bien les enfants du dos argenté du groupe, deux cas de paternité extra-groupe ont été observés. Ces résultats ouvrent ainsi la possibilité jusqu'alors inimaginable que les femelles puissent occasionnellement copuler avec des mâles adultes en dehors du groupe. Nous avons également étudié l'apparentement entre femelles pour tester l'hypothèse que les femelles d'un groupe se regroupent (c'est-à-dire ont tendance à migrer dans le même groupe) du fait de leur apparentement. Malgré certaines relations de sœurs observées, l'apparentement des femelles intra-groupes n'est pas plus important que celui des femelles inter-groupes, contrairement à des résultats obtenus précédemment sur des gorilles non habitués (Bradley et al. 2007). Enfin, nous avons estimé (et comparé) les taux de dispersion mâle et femelle et confirmé que les mâles semblent disperser plus loin que les femelles.

Suite à nos résultats de paternité extra-groupe, d'autres échantillons sont en cours d'analyse pour voir si d'autres cas sont identifiés sur un plus large panel d'individus.

Projet: Importance de la sélection balancée dans l'adaptation

Comme vu plus tôt (Boite 1), la sélection balancée est une forme de sélection naturelle qui aboutit à un maintien de diversité génétique et qui peut résulter de mécanismes évolutifs sous-jacents variés (avantage à l'hétérozygote [AH], sélection fréquence-dépendante négative [SFDN], sélection fluctuante dans le temps et/ou l'espace [SF], sélection sexuellement antagoniste [SSA] ou pleiotropie antagoniste). La question de l'importance de la sélection balancée en tant que mode d'adaptation des espèces à leur environnement est une question ancienne pour laquelle nous n'avons cependant toujours pas de réponse claire. En effet, historiquement, comme précédemment résumé (Nei 2005; Bitarello et al. 2023), la sélection balancée a d'abord été proposée comme une force importante responsable du maintien des polymorphismes dans les populations (« Balance hypothesis », Dobzhansky 1955). L'observation des niveaux importants de diversité dans les données d'allozymes (Lewontin et al. 1966), et la découverte du cas d'avantage à l'hétérozygote maintenant un allèle responsable de l'anémie falciforme chez l'humain (Livingstone 1964) a contribué encore plus à l'idée que la sélection balancée était une force majeure d'évolution des génomes. Cependant, peu de temps après, la théorie neutraliste de Kimura (1968) a pris le dessus et a semblé suffisante pour expliquer l'écrasante majorité des patrons de diversité observés dans la nature. De plus, en dehors des signaux clairs au niveau du CMH (Klein et al. 1993) et du gène ABO (Calafell et al. 2008), les premiers scans de sélection balancée chez l'humain n'ont pas réussi à mettre en évidence de nouveaux candidats (Asthana et al. 2005; Bubb et al. 2006), suggérant que la sélection balancée est un mécanisme instable, et/ou très rare. En conséquence, l'écrasante majorité des études se sont alors focalisées sur la détection et la quantification de la sélection positive et négative dans les génomes. Cependant, ces dernières années, plusieurs études semblent remettre en question cette affirmation, avec l'identification de nouveaux cas convaincants chez l'humain (Andrés et al. 2009; Leffler et al. 2013; DeGiorgio et al. 2014; Teixeira et al. 2015). Une étude récente a ainsi estimé que la sélection balancée touche 0.8% des sites génomiques humains (Bitarello et al. 2018) et une autre 2 à 4% des sites non synonymes partagés entre populations humaines (Soni et al. 2022). Chez les plantes, un certain nombre de phénotypes semble également évoluer sous sélection balancée (Delph et al. 2014; Karasov et al. 2014). En parallèle, des études chez la drosophile ont montré que plusieurs centaines de loci sont soumis à de la forte sélection fluctuante dans le temps (en réponse aux variations climatiques), entraînant un maintien important de polymorphismes à long terme (Bergland et al. 2014; Machado et al. 2021). Récemment, un scan génomique dans un complexe d'espèces de Daphnies a montré que le nombre de polymorphismes partagés entre espèces (des dizaines de milliers) ne peut pas être expliqué seulement par de l'hybridation ou du tri de lignées incomplet, mais suggère l'action de la sélection balancée (principalement par convergence récente) (Murray et al. 2024). Plusieurs auteurs ont ainsi dernièrement avancé l'idée que la sélection balancée pourrait être plus répandue que jusqu'alors considéré, bien que l'importance relative des différents mécanismes sous-jacents reste mal connue (Key et al. 2014; Llaurens et al. 2017; Bitarello et al. 2023; Johnson et al. 2023). La plupart des études précédentes se sont

cependant concentrées sur la détection de polymorphismes *anciens et partagés*, et sur des espèces *modèles*. La question reste donc ouverte de la prévalence de la sélection balancée à d'autres échelles de temps, et chez des organismes non-modèles.

Au vu de ces résultats récents qui ont relancé l'intérêt de la communauté pour la sélection balancée, mais également étant donné l'accumulation de données de polymorphismes dans une large gamme d'organismes, permettant de combiner des données génomiques intra- et inter-espèces, ainsi que le développement récent de nombreuses méthodes pour détecter la sélection balancée de façon plus robuste par rapport à l'effet confondant de la démographie (Siewert et al. 2017, 2020; Cheng et al. 2019, 2020, 2022; Isildak et al. 2021; Soni et al. 2022; Braichenko et al. 2024; Hayeck et al. 2024), il me semble opportun de revisiter la question de l'importance de la sélection balancée dans l'adaptation des espèces en étudiant : i) sa prévalence à l'échelle des génomes et ses cibles fonctionnelles, ii) la relative importance des différents mécanismes évolutifs sous-jacents, et iii) son rôle dans le maintien de variant structuraux, notamment d'inversions.

1. Puissance des méthodes existantes et potentiel développement méthodologique

De manière problématique, la plupart des méthodes ayant été développées pour détecter la sélection balancée ont été évaluées à l'aide de simulations basées sur des scénarios d'AH (DeGiorgio et al. 2014; Siewert et al. 2017, 2020; Bitarello et al. 2018; Cheng et al. 2019, 2020, 2022; Braichenko et al. 2024; Hayeck et al. 2024), le plus souvent avec des fréquences d'équilibre des mutations sous sélection à 50% (en partie car ce mécanisme est plus simple à simuler, mais aussi parce que les signatures associées sont plus stables dans le temps et donc plus faciles à détecter). Une conséquence importante est que la puissance des méthodes existantes dans des cas de SFDN, de SF ou de SSA n'est tout simplement pas connue (à noter, l'exception de Soni et al. 2022 qui ont évalué leur méthode dans la cadre de SFDN).

Pourtant, un certain consensus semble émerger pour favoriser l'idée que la SFDN et/ou la SF participerait plus que l'AH au maintien de polymorphismes génétiques (Hedrick 2007; Delph et al. 2014; Ashby et al. 2017; Brisson 2018). Il est en effet considéré que l'AH est un mécanisme transitoire, et donc plus rare, car il implique un lourd fardeau génétique du fait que, à chaque génération, la moitié de la population est homozygote et a donc une valeur sélective non optimale par rapport aux hétérozygotes ('segregation load') (Fijarczyk et al. 2015). Ce coût semble de plus pouvoir être évité grâce à la duplication du polymorphisme en question, comme observé pour la vision trichromate obtenue chez certains primates par la duplication d'un gène des opsines (Dulai et al. 1999). De son côté, la SFDN est prédite comme étant un mécanisme pouvant se maintenir à long terme, mais qui peut se accompagner d'un remplacement important d'allèles ('turnover'), aboutissant à des signaux qui, au final, semblent plus récents (Fijarczyk et al. 2015). En parallèle, de nombreuses études ont montré que les cibles de la sélection balancée sont souvent liées aux interactions hôtes-pathogènes (Andrés et al. 2009; Spurgin et al. 2010; Brown et al. 2011; Leffler et al. 2013; DeGiorgio et al. 2014; Key et al. 2014; Croze et al. 2016; Koenig et al. 2019; Bitarello et al. 2023; Lawal et al. 2023), qui semblent plutôt

résulter de SFDN et/ou de SF dans le temps (Carius et al. 2001; Ashby et al. 2017). Il est donc possible que la prévalence de la sélection balancée dans la nature soit globalement sous-estimée du fait de biais méthodologiques (Bitarello et al. 2023).

Le premier objectif de ce projet, en collaboration avec Sylvain Mousset (LBBE) et Sylvain Glémin (ECOBIO, Rennes), sera donc de simuler des données génomiques sous des scénarios d'AH, de NFDS, de SF et de SSA, grâce au logiciel SLIM, pour évaluer la puissance relative de différentes méthodes sous des dynamiques contrastées. L'idée est d'explorer une large gamme de paramètres pour chacune de ces dynamiques (taille efficace de population, structuration spatiale, coefficient de dominance, valeur sélective, taux de recombinaison), afin de déterminer si le biais de détection est important ou non, et dans quelle situation. Etant donné que sous sélection liée aux pathogènes, comme dans le cas du CMH, différents mécanismes peuvent agir conjointement (Spurgin et al. 2010), nous essaierons aussi de combiner certains de ces mécanismes pour étudier les signatures génomiques résultantes.

En parallèle, un deuxième objectif de cette partie sera d'utiliser les signatures génomiques observées sous ces différents scénarios pour trouver des statistiques qui permettent justement de distinguer les différents mécanismes causaux. Là où la plupart des méthodes existantes s'appuient sur des données génomiques *intra-population*, nous nous attacherons à combiner des statistiques *intra-* et *inter-population* pour gagner en puissance. Par exemple, nous pourrions considérer la différenciation génétique entre populations de la même espèce et entre espèces proches, pour distinguer la sélection fluctuante de l'AH, la NFDS ou la SSA, l'idée étant de regarder si les fréquences alléliques sont globalement stables ou non entre populations et espèces proches. En effet, sous sélection fluctuante dans l'espace, nous nous attendons à ce que la différenciation génétique entre populations soit élevée. Egalement, sous sélection fluctuante dans le temps, dans l'hypothèse que la source des fluctuations (par exemple la co-évolution avec un agent infectieux) ne soit pas synchronisée entre populations, nous nous attendons à une différenciation génétique élevée. A l'inverse, sous AH, NFDS et SSA, la différenciation génétique entre populations devrait être faible.

En parallèle, nous explorerons non seulement les signatures génomiques à l'échelle des SNVs (single-nucleotide variants), mais aussi à l'échelle des haplotypes, pour incorporer de l'information sur le déséquilibre de liaison et gagner en puissance afin de mieux discriminer les mécanismes évolutifs à l'œuvre.

2. *Jeu de données empiriques*

Pour étudier la prévalence et les mécanismes évolutifs associés à la sélection balancée dans la nature, je souhaite m'appuyer sur quatre systèmes d'étude complémentaires pour lesquels des données de reséquençage (de lectures courtes) ont été obtenues dans plusieurs populations de la même espèce et plusieurs espèces proches :

- 70 grands singes du genre *Pan*, incluant 10 bonobos (*P. paniscus*) et 60 chimpanzés (*P. troglodytes*) venant de quatre sous-espèces différentes et correspondant à un gradient Ouest-

Est en Afrique sub-saharienne (de Manuel et al. 2016), en collaboration avec Peter Sudmant (UC Berkeley, USA). Des données sont également disponibles pour différentes populations humaines (1000 Genomes Project Consortium 2010).

- 280 gobemouches du genre *Ficedula* correspondant à deux groupes d'espèces proches, incluant deux espèces pour lesquelles nous disposons de plusieurs populations à travers l'Europe (Burri et al. 2015; Chase et al. 2021), en collaboration avec Carina Mugal (LBBE).

- 160 mouches du varek du genre *Coelopa* correspondant à trois populations de *C. frigida* (Amérique du Nord, Europe, Asie ; la dernière étant probablement une espèce différente), ainsi qu'une population de l'espèce soeur *C. nebularum* (données non publiées), en collaboration avec Claire Mérot (ECOBIO, Rennes).

- 60 papillons du genre *Morpho* correspondant à trois populations de *M. helenor* (Forêt atlantique, Amazonie et Amérique centrale), ainsi qu'une population de l'espèce soeur *M. achilles* (données non publiées), en collaboration avec Violaine Llaurens (CIRB, Paris).

Ces systèmes d'étude sont intéressants à étudier et à comparer car i) ils sont taxonomiquement et écologiquement variés, permettant de nous interroger sur l'universalité des résultats obtenus, ii) ils correspondent à des données intra- et inter-population, ouvrant la possibilité de mieux comprendre les dynamiques micro- et macro-évolutives en jeu, et iii) ils proviennent d'ensembles d'espèces pour lesquelles l'histoire démographique et les contextes écologiques sont connus, nous permettant d'avoir des attendus précis sous neutralité et de guider l'interprétation des résultats.

L'objectif de cette partie est d'acquérir des données de lectures longues (PACBIO) pour chaque complexe d'espèces, en ciblant 40 individus au total (10 individus par population, pour 4 populations, incluant à chaque fois deux espèces proches différentes). Pour les grands singes, des données sont déjà disponibles (non publiées) et seront complétées, en collaboration avec Peter Sudmant. Pour les humains, nous utiliserons des données publiques (Liao et al. 2023). Les données de lectures longues sont essentielles, non seulement pour s'assurer de la fiabilité des génotypes pour les SNVs (particulièrement étant donné les résultats obtenus par D Hicks et M Barrientos décrits dans la partie 2), mais aussi pour obtenir une résolution à l'échelle haplotypique, nous permettant de gagner en puissance pour reconstruire l'histoire adaptative et généalogique de chaque haplotype, et enfin pour pouvoir identifier les variants structuraux (SVs), une classe de variants très intéressantes pour l'adaptation à l'environnement mais inaccessibles à partir de lectures courtes.

3. *Prévalence de la sélection balancée et mécanismes sous-jacents*

En combinant des données de lectures courtes et longues grâce à des approches basées sur des graphes de pangéome, nous pourrions obtenir un jeu de données de SNVs de bonne qualité pour chacun des quatre systèmes d'espèces étudiés. Ces jeux de données seront utilisés pour évaluer i) la prévalence à l'échelle du génome de la sélection balancée (considérant tous les possibles mécanismes évolutifs), ii) la contribution relative de chaque mécanisme sous-jacent, et iii) la consistance des résultats entre groupes taxonomiques.

De manière générale, pour connaître la prévalence globale de la sélection naturelle, c'est-à-dire le nombre de loci évoluant en réponse à une dynamique adaptative, il faudrait pouvoir prédire les attendus (de la statistique résumée d'intérêt) sous neutralité, pour ensuite identifier les loci qui dévient de cet attendu. Cela nécessite de connaître précisément l'histoire démographique des populations/espèces étudiées, c'est à dire la fluctuation de leur taille efficace dans le temps, ainsi que la structuration de ces populations dans l'espace, incluant des événements d'introggression possibles avec d'autres espèces et d'extinction de certaines populations. Un tel niveau de détail, intégrant des variations spatiales et temporelles complexes, est presque hors d'atteinte. L'approche fréquemment utilisée est alors d'identifier des loci 'outliers' présentant une (ou des) statistiques résumées extrêmes (par exemple top 1% ou 5%), et de se satisfaire d'avoir une liste de candidats *enrichis* pour des loci sous sélection naturelle, mais ne permettant pas d'obtenir une estimation de leur nombre absolu.

Une solution pour pallier à ce problème est de tirer avantage du fait que la sélection naturelle ne peut, par définition, toucher que des loci ayant un effet sur le phénotype, et donc qu'il est possible d'obtenir une idée de la distribution d'une statistique donnée sous neutralité en sélectionnant les régions du génome n'ayant *a priori* pas d'effet sur le phénotype (par exemple régions éloignées des gènes, dénuées d'éléments régulateurs, et non conservées entre espèces). Cette approche ne permet cependant pas de bien contrôler pour différents facteurs influençant les patrons de diversité génétique à une échelle *locale*, comme le taux de mutation, le taux de recombinaison ou encore la composition en bases. Une possibilité est alors de contraster les patrons génétiques observés entre des loci synonymes (SYN) et non-synonymes (NS) du même gène, puisqu'ils sont dans les mêmes régions génomiques mais que les premiers sont censés être invisibles aux yeux de la sélection, contrairement aux deuxièmes. Cette approche a été largement utilisée pour obtenir des mesures d'intensité de l'adaptation, principalement en s'intéressant à la sélection positive (McDonald et al. 1991).

Récemment, un test s'appuyant sur ce principe et calculant un Z-score, le ratio de SNPs partagés entre populations entre sites SYN et NS a été développé afin d'obtenir une idée de la prévalence de la sélection balancée à l'échelle du génome (Soni et al. 2022). En effet, l'idée est que le partage de SNPs SYN reflète seulement l'histoire démographique des populations, tandis que le partage de SNPs NS reflète l'histoire démographique *et* d'éventuels événements de partage par sélection balancée, avec donc l'idée implicite que la sélection balancée favorise toujours le partage de variants fonctionnels entre populations. Ce ratio est ensuite divisé par le ratio de SNPs SYN/NS non partagés par les populations, pour prendre en compte l'effet de la sélection négative.

$$Z = \frac{S_N/S_S}{R_N/R_S}$$

S : Polymorphismes partagés entre populations ; R : polymorphismes privés
 N : NS ; s : SYN. (Soni et al. 2022)

Ce test a été appliqué chez l'humain et a montré que de 2 à 4% des sites NS partagés entre populations sont soumis à la sélection balancée, correspondant à des centaines de sites NS (Soni et al. 2022). Cependant, ce test n'a, à ma connaissance, jamais été utilisé sur d'autres espèces, donc il serait intéressant de l'appliquer aux différents jeux de données mentionnés plus

haut. Pour que ce test soit applicable, il faut que les populations/espèces comparées soient assez distantes pour ne pas partager trop de leur diversité génétique, mais assez proches pour encore en partager, donc ce test convient à des paires de populations structurées d'une même espèce, ou des paires d'espèces proches, cas qui correspond exactement à nos données empiriques.

Ce test est, par construction, limité aux régions géniques et ne considère donc pas les événements de sélection balancée dans des régions régulatrices. Cependant, il pourrait être adapté pour contraster non pas que des sites SYN et NS, mais par exemples des sites dans des exons versus dans des introns, ou des sites dans des régions codantes versus dans des régions régulatrices. Nous pourrions également l'appliquer (comme fait par les auteurs) sur des catégories fonctionnelles, par exemple pour tester l'enrichissement de signal en sélection balancée dans des sous-ensembles de gènes impliqués dans des fonctions immunitaires, métaboliques, ou autre.

Afin de contrôler pour l'effet de la conversion génique biaisée pour GC (gBGC), nous pourrions également analyser séparément les mutations A/G, A/C, T/G et T/C des autres. En effet, la gBGC correspond à un biais de transmission des allèles G et C face aux allèles A et T résultant d'un biais de réparation des hétéroduplex d'ADN pendant la recombinaison méiotique (Duret et al. 2009). Ce biais est non-adaptatif, mais mime l'effet de la sélection positive pour les allèles G/C face aux allèles A/T, et semble à l'œuvre assez largement chez les eucaryotes (Pessia et al. 2012). Par conséquence, quand un allèle A/T est avantageux par rapport à un allèle G/C (en terme de valeur sélective du phénotype), il va mettre plus de temps à se fixer dans la population qu'attendu sans gBGC (et vice versa, si un allèle G/C est avantageux, il va se fixer plus rapidement qu'attendu). De tels polymorphismes (ayant un allèle A ou T et un allèle G ou C) peuvent donc rester dans la population plus longtemps qu'attendu sous neutralité si les allèles A/T sont avantageux, du fait de ces deux forces qui s'opposent, et ainsi imiter également l'effet de la sélection balancée.

En parallèle, en fonction des résultats des simulations de la partie 1, nous utiliserons la méthode (ou la combinaison de méthodes) optimale pour détecter le plus large spectre de dynamiques évolutives sous-jacentes. Pour chaque région identifiée comme potentiellement sous sélection balancée, nous pourrions ensuite étudier quel mécanisme causal est le plus probable, en s'appuyant sur les statistiques les plus discriminantes.

Le fait d'analyser en parallèle quatre jeux de données variés, et donc de comparer les résultats entre primates, oiseaux et insectes (diptères et lépidoptères), nous permettra de savoir si les prévalences obtenues et les mécanismes à l'œuvre diffèrent ou non entre organismes ayant des caractéristiques génomiques et des traits d'histoire de vie différents (notamment tailles efficaces, temps de génération, paysages de recombinaison, systèmes de détermination du sexe et architectures génomiques). Nous pouvons en effet par exemple nous attendre à ce que la prévalence de la SF soit plus forte chez des organismes ayant des temps de génération courts. En fonction des observations, nous pourrions ensuite formuler de nouvelles hypothèses de travail qui pourront faire l'objet de projets futurs plus ciblés.

Une autre perspective intéressante pour mieux comprendre les implications fonctionnelles des différentes régions candidates identifiées serait de chercher quels gènes sont différemment exprimés entre individus homozygotes et hétérozygotes pour chaque locus d'intérêt. De telles données d'expression sont notamment disponibles pour les gobemouches (C Mugal, résultats non publiés), mais également pour les humains (notamment le projet GTEx, GTEx Consortium 2013) et pourraient donc être explorées à court terme.

4. Rôle de la sélection balancée dans le maintien de variants structuraux

En parallèle de l'étude des SNVs, l'obtention de lectures longues permettra de caractériser les variants structuraux du génome, et donc de poser la question de la prévalence de la sélection balancée non seulement sur des polymorphismes nucléotidiques, mais également sur des polymorphismes de nombre de copies, ou sur des inversions, variants génétiques qui sont restés longtemps sous-étudiés et semblent pourtant cruciaux pour l'adaptation des espèces à leur environnement et plus largement pour la spéciation (Kirkpatrick et al. 2006; Faria et al. 2019; Wellenreuther et al. 2019; Durmaz et al. 2020; Mérot et al. 2020a).

En effet, les inversions peuvent créer une association stable entre loci co-adaptés (de par la diminution de la recombinaison entre haplotypes inversés et non-inversés) et ainsi permettre aux populations de s'adapter rapidement à des environnements hétérogènes ou fluctuants (Hoffmann et al. 2004; Llaurens et al. 2017; Wellenreuther et al. 2018; Faria et al. 2019). Mais les inversions peuvent également être maintenues du fait de pseudo-overdominance (également appelée overdominance associative), c'est-à-dire à cause de l'accumulation de mutations délétères récessives sur chaque fond haplotypique, ce qui peut entraîner une diminution de fitness quand ces mutations se retrouvent à une fréquence suffisante pour être observées à l'état homozygote (Ohta et al. 1969). En effet, si une inversion se produit et l'haplotype nouvellement inversé est avantageux, il va commencer à augmenter en fréquence par sélection positive. Mais du fait de l'absence de recombinaison entre haplotypes inversés et non-inversés, toute mutation récessive délétère qui était présente sur l'haplotype nouvellement inversé ne pourra pas être purgée par la sélection négative. Par conséquent, ces mutations récessives délétères vont se révéler chez les individus homozygotes pour l'inversion et entraîner un coût important sur la fitness, empêchant alors l'haplotype inversé initialement avantageux de se fixer.

Récemment, plusieurs exemples d'inversions polymorphes ont été mis en évidence chez différentes espèces, avec des patrons génétiques compatibles avec de la sélection balancée (Jay et al. 2018; Wellenreuther et al. 2018; Faria et al. 2019; Kapun et Flatt 2019; Mérot et al. 2020b; Berdan et al. 2023; Kapun et al. 2023; Nosil et al. 2023). Cependant, l'importance des deux mécanismes discutés plus haut dans le maintien d'inversions à l'état polymorphe sur le long-terme représente une question émergente de recherche. L'objectif de cette partie est donc de tester si les inversions se retrouvent préférentiellement ou non sous sélection balancée, étant

donné qu'aucun scan génomique systématique n'a jamais été effectué, et d'explorer quel rôle jouent les mutations délétères récessives dans le maintien de ces inversions.

Pour se faire, nous pourrions prendre en compte la taille des inversions pour étudier leur maintien par sélection balancée, puisque cette caractéristique influence leur trajectoire évolutive. En effet, Connallon et Olito (2022) ont montré à l'aide de modèles théoriques que les scénarios d'adaptation locale (inversions indirectement bénéfiques car diminution de la recombinaison entre loci adaptés au même environnement) devraient favoriser la fixation d'inversions de plus grande taille, car elles ont plus de chance de capturer plusieurs loci co-adaptés, tandis que les scénarios de sélection directe (avantage du point de cassure de l'inversion) ou d'underdominance (désavantage des hétérozygotes face aux homozygotes) devraient aboutir à des inversions de plus petites tailles. De manière intéressante, les grandes inversions sont également plus à même de contenir des mutations récessives délétères et donc d'empêcher l'augmentation en fréquence et la fixation de telles inversions même si elles sont avantageuses. Dans un cadre diploïde, leurs modèles montrent également que le fardeau de mutations délétères fortement récessives diminue la taille attendue des inversions maintenues à l'état polymorphe, ainsi que leur fréquence à l'équilibre. Ces statistiques pourraient donc servir pour déterminer quel rôle joue la pseudo-overdominance dans le maintien d'inversions (par exemple en comparant la taille d'inversions fixées versus polymorphes). Plus généralement, leurs travaux pourraient être étendus pour étudier les attendus sur la taille des inversions maintenues dans des cas d'AH, de SFDN ou de SF.

5. Perspectives / Conclusions

Une autre piste intéressante à explorer serait d'inclure des jeux de données d'espèces haploïdes dans l'étude, ce qui permettrait d'exclure l'action de l'AH (et de caractériser les patrons génétiques observés en son absence). L'étude d'espèces haploïdes permettrait également d'observer la dynamique des inversions en l'absence de pseudo-overdominance due à l'accumulation de mutations délétères récessives, et donc de pouvoir indirectement quantifier cet effet par comparaison avec les patrons observés chez des espèces diploïdes. Une telle étude pourrait se faire sur des moustiques du genre *Anopheles* qui présentent un cycle de vie principalement haploïde, en collaboration avec Michael Fontaine (MIVEGEC, Montpellier) et/ou sur *Plasmodium spp.* en collaboration avec Ellen Leffler (University of Utah, USA).

En conclusion, j'espère que ce projet combinant simulations et données empiriques permettra d'y voir plus clair dans les biais associés aux différentes méthodes de détection de la sélection balancée, de mieux comprendre le rôle que joue cette force évolutive dans la diversité génomique et l'adaptation des espèces, d'identifier de nouveaux signaux de sélection balancée dans des espèces non-modèles variées et de savoir comment l'architecture génétique interagit avec la sélection balancée, notamment pour voir si les inversions sont globalement préférentiellement maintenues à l'état polymorphe. A plus long terme, je souhaite que les données nouvellement acquises puissent également servir à poser d'autres questions de recherche en génomique des populations plus spécifiques à chacun des systèmes étudiés.

Conclusion

Comme vous avez peut-être pu le constater en parcourant ce manuscrit, mes projets de recherche, bien que toujours ciblés autour de questions de biologie évolutive, sont assez diversifiés, tant dans les échelles de temps considérées (maintien de sélection balancée depuis des millions d'années, adaptation humaine depuis le Néolithique ou adaptation microbienne depuis la révolution industrielle), que par les données traitées (génotypage par puce à ADN, séquençage de génomes complets ou données de métagénomique), ou encore les disciplines mobilisées (évolution, génomique, anthropologie, écologie, bioinformatique, microbiologie, sciences médicales). C'est en partie ainsi que je trouve du plaisir à travailler, en renouvelant sans cesse les questions à explorer, et en apprenant en permanence de nouveaux concepts et outils. Cependant, cela vient avec le coût de ne pas avoir une trajectoire scientifique très linéaire, de ne pas être facilement « identifiable » thématiquement, et de sortir fréquemment de ma zone de confort. Et en découle également la difficulté d'écrire une conclusion générale...

Ma mutation en 2020 depuis le laboratoire d'anthropologie génétique au musée de l'Homme de Paris vers le laboratoire de Biométrie et Biologie Evolutive de Lyon a fait voler en éclat le dénominateur commun qui jusqu'alors était resté stable : le fait de travailler sur l'humain. J'ai notamment récemment initié des collaborations pour étudier l'adaptation de populations de gobemouches, de babouins, de souris, et de punaises de lit à leurs environnements. Mon fil rouge, cependant, est l'adaptation. Je suis ravie de pouvoir allier des questions d'évolution à des questions d'écologie, et curieuse de voir quelles connaissances peuvent contribuer à répondre à des préoccupations sociétales actuelles. Il me semble en effet avoir fait un tournant important pour porter mon regard non plus vers le passé et la compréhension de nos racines évolutives, mais plus vers le présent, et notamment celui de nombreuses espèces confrontées à des défis environnementaux importants. D'un autre côté, l'étude d'espèces diverses me donne accès à des événements évolutifs beaucoup plus lointains et m'a amené à réfléchir à comment concilier les patrons de micro-évolution (à l'échelle intra-spécifique, grâce aux outils de génétique des populations) avec ceux de macro-évolution (à l'échelle inter-spécifique, grâce aux outils de la phylogénie).

Face aux défis environnementaux contemporains, toutes les espèces ne sont pas égales, certaines évoluent suffisamment vite pour avoir le temps de trouver des parades adaptatives à court terme, d'autres sont plus contraintes par leur long temps de génération et seront probablement obligées d'adapter leur comportement. Parmi les questions que je trouve intéressantes et que j'ai peu explorées jusqu'ici, figure celle de la plasticité : comment s'articulent la plasticité phénotypique et l'adaptation moléculaire ? Quelles interactions existent-ils entre les changements épigénétiques et génétiques ? Ces deux mécanismes répondent-ils aux mêmes contraintes, sont-ils corrélés dans leur taux et leur tempo d'évolution, ou bien au contraire sont-ils inversement corrélés, avec l'idée que la plasticité vient « soulager » les contraintes environnementales, et donc réduire les pressions de sélection ? Une réflexion similaire peut être menée chez l'humain sur les interactions entre évolution culturelle et évolution génétique : la culture permet-elle de contourner les contraintes environnementales

(on peut par exemple penser au rôle des innovations culturelles face aux conditions climatiques, aux pathogènes, à la toxicité de certains aliments) ou au contraire la culture vient-elle créer de nouvelles contraintes auxquelles les humains doivent s'adapter (consommation de nouveaux aliments, pollution, perturbateurs endocriniens).

Une autre question à laquelle je me suis confrontée au cours de certains de mes projets est liée au concept d'holobionte : l'holobionte représente-t-il une unité de sélection sur laquelle agit la sélection naturelle, ou une somme d'acteurs essayant, individuellement, de maximiser leur valeur sélective ? Dans le cadre du microbiote, cela me paraît encore bien compliqué d'arriver à estimer quels sont les interactions entre valeurs sélectives humaines et bactériennes. En d'autres mots, dans quels cas les évolutions microbiennes sont sans effets, bénéfiques ou coûteuses pour leurs hôtes ? Et comment le coût potentiel pour l'hôte influence-t-il en retour l'évolution bactérienne ?

En conclusion, en dehors de l'intérêt personnel que je trouve à questionner comment les individus / populations / espèces s'adaptent à diverses contraintes, y compris celles qu'ils ont eux-mêmes créées, je pense que notre espèce a la quasi obligation de se préoccuper de l'impact de son propre comportement non pas sur sa survie, mais sur celle de l'ensemble des écosystèmes. Je n'ai pas l'illusion de penser que la recherche scientifique constitue une réponse satisfaisante aux enjeux de demain, mais je suis fière de contribuer à ce travail collectif de construction de connaissances fondamentales.

Réflexions sur l'encadrement

Ma motivation principale pour écrire ce document est de pouvoir encadrer des thèses de manière autonome. Il me semble donc naturel d'y inclure une ébauche de réflexion sur l'encadrement d'étudiant.e.s, bien que (et c'est sûrement regrettable), cela ne soit ni obligatoire, ni une grande partie de mon HDR. Depuis le début de mon parcours, j'ai co-encadré une thèse soutenue (N Marchi), une thèse non aboutie (L Bouteille) et je co-encadre actuellement deux thèses (J Adam et C Lacomme). Je souhaite désormais pouvoir déposer des sujets de manière indépendante, même si j'affectionne particulièrement le co-encadrement, car je trouve cela plus riche scientifiquement et humainement.

En réfléchissant à ce que j'ai envie de dire sur le sujet, j'ai trouvé utile de lister ce que je pense être important dans l'encadrement d'étudiant.e.s en thèse (mais aussi en master), afin de clarifier les objectifs à atteindre :

- i) d'être présent.e (même si cela peut sembler trivial, il ne me semble pas inutile de le rappeler),
- ii) de fournir les bases théoriques nécessaires pour comprendre les questions de recherche et interpréter les résultats,
- iii) de former à la logique scientifique,
- iv) d'apprendre à synthétiser, reformuler et vulgariser les objectifs fixés et les résultats obtenus,
- v) de permettre à l'autre de s'emparer de son sujet et d'explorer de nouveaux chemins,
- vi) de transmettre, le cas échéant, sa passion de la science,
- vii) de donner l'exemple en terme d'éthique de la recherche, et
- viii) d'apprendre à prendre du recul et à accepter la remise en cause.

Ces éléments de réflexion ne sont pas hiérarchisés, et clairement non exhaustifs (et je serai ravie de savoir lesquels pourraient être rajoutés). Évidemment, de manière plus large, il s'agit de donner les conditions matérielles et humaines pour que la thèse puisse aboutir, et pour que les étudiant.e.s puissent atteindre à la fois des objectifs professionnels leur permettant de pouvoir continuer dans la voie de la recherche si tel est leur souhait, mais aussi d'atteindre des objectifs personnels comme la reconnaissance du fruit de leur travail.

La principale difficulté à mes yeux est de trouver un juste équilibre entre une transmission trop asymétrique (et hiérarchisée) de l'encadrant.e vers l'encadré.e et une relation complètement symétrique entre encadrant.e et encadré.e. Quels sont les aspects pour lesquels je me permets de me positionner en tant que directeur.rice, et ceux pour lesquels je me positionne plutôt en tant que collègue ? Un sujet de thèse appartient-il plus à l'encadrant.e ou à l'encadré.e ? La difficulté est d'arriver à individualiser ce travail d'encadrement, pour comprendre, pour chaque étudiant.e, les aspects sur lesquels il.elle a besoin d'être plus accompagné.e et ceux sur lesquels il.elle a besoin d'être plus libre. Certaines personnes vont en effet être plus épanouies si on les accompagne de très près, et qu'on délimite clairement les objectifs précis, et d'autres s'épanouiront mieux si on les laisse explorer par eux-mêmes. Cette fine ligne de crête est évidemment vouée à évoluer tout au long de la thèse, avec l'expérience acquise par l'étudiant.e. Je dois dire que j'ai moi-même un peu changé de perspective sur cette question depuis mon arrivée au LBBE, et je remercie les différents étudiant.e.s et les collègues

qui m'ont élargi l'esprit à ce sujet. Mes expériences d'encadrement, thèses et masters, m'ont en effet permis de réajuster mes objectifs : arriver à faire ressortir le meilleur de chaque personne, avec ses singularités, plutôt que d'arriver à faire avancer au mieux le projet de recherche fixé. Parfois, les deux vont de pair, parfois non.

A propos de cette notion du juste cadre, je pense que le terme de mentor serait peut-être plus approprié que celui de directeur.rice de thèse, qui implique une forme de hiérarchie pas nécessairement constructive. Le mentorat peut se définir par une « relation interpersonnelle de soutien, d'aide, d'échanges et d'apprentissage, dans le cadre de laquelle une personne d'expérience, le mentor, offre sa sagesse acquise et son expertise dans le but de favoriser l'évolution d'une autre personne, le mentoré, ayant des compétences ou des connaissances à acquérir et des objectifs professionnels à atteindre ». J'ai le sentiment que la notion de soutien et d'aide passe trop souvent au deuxième plan par rapport à celle d'atteindre des objectifs professionnels.

Une autre question que je trouve intéressante, liée à cette HDR, est comment évaluer la qualité du travail d'encadrement ? L'objectif est-il de publier un certain nombre d'articles de qualité avec l'étudiant.e, de garder le cap sur la rigueur scientifique, quitte à passer beaucoup de temps en chemin à explorer une méthodologie infinie, de rendre la personne autonome dans sa réflexion, d'arriver à donner du sens à cette recherche et in fine, un sentiment de satisfaction à l'étudiant.e ? Encore une fois, évidemment, idéalement tous ses objectifs se nourrissent les uns les autres, mais concrètement, il n'est pas rare qu'ils s'affrontent.

L'encadrement est en tout cas aujourd'hui une des tâches que je préfère du métier de chercheuse, et j'espère que je saurai être une encadrante bienveillante malgré mon exigence parfois élevée et ma difficulté à lâcher prise. En conclusion, je pense me fixer comme objectif d'effectuer un travail rigoureux et éthique en favorisant la qualité sur la quantité et en respectant le rythme et le bien-être de chacun.e, tout en préparant au mieux les étudiant.e.s à pouvoir être compétitif.ve.s sur le marché du travail qui les attend par la suite. J'aime à croire que les processus de recrutement évoluent favorablement vers une évaluation plus qualitative de la recherche, même si le nombre de postes dans la recherche a une sérieuse tendance à diminuer.

Réflexions sur l'impact écologique et l'éthique de la recherche

Mon intégration au LBBE en 2020, ainsi que ma participation au CoNRS depuis 2021, m'ont amenée à m'interroger d'avantage à la fois sur l'impact environnemental des activités de recherche, mais aussi sur les enjeux éthiques liés à la publication de la recherche.

Concernant la partie 'impact environnemental', je suis convaincue de la nécessité et de l'utilité de réfléchir et agir face aux conséquences environnementales de nos activités de recherche. Bien que je n'ai pas encore réussi à trouver l'énergie de m'investir plus profondément dans le groupe de travail « Empreinte environnementale » du laboratoire, il m'a permis de faire murir ma réflexion, et j'espère sincèrement que l'avenir me permettra d'être plus active de ce côté-là.

La première étape, amorcée par ce groupe de travail pour le laboratoire et guidée en partie par les retours d'expérience et les outils mis en place par l'initiative Labos 1.5 (<https://labos1point5.org/>), me semble d'estimer les coûts environnementaux associés aux différentes activités de recherche (que l'on peut grossièrement catégoriser en achats, missions, et calculs informatiques). Cette quantification n'est pas triviale, et probablement pas parfaite, mais permet tout de même à chacun.e de voir quelles activités sont les plus gourmandes. Si l'on accepte l'idée que les coûts sont quantifiables, une question à mes yeux plus délicate est de savoir comment intégrer ces coûts dans la construction de nos projets de recherche. Une solution logique serait de comparer ces coûts aux bénéfices des projets, et de ne retenir que ceux pour lesquels le bénéfice est supérieur au coût. Cependant, cela signifie de pouvoir quantifier les bénéfices des projets de recherche, ce qui me semble loin d'être évident. En dehors de la faisabilité, je ne sais même pas si cela serait vraiment souhaitable, car cela revient à considérer que l'on connaît les bénéfices de la recherche, or une partie importante de notre travail est d'approfondir des connaissances fondamentales sans nécessairement savoir, a priori, à qui, à quoi, quand et dans quel contexte elles seront utiles/utilisées. Le fait même de devoir quantifier des bénéfices, dans un monde où les services deviennent de plus en plus une somme de métriques souvent absurdes, me dérange. Une autre option serait de proscrire tous les projets qui ont un coût considéré comme trop élevé (prohibitif), en dehors de leurs bénéfices potentiels. En effet, à quoi bon étudier le vivant si l'intérêt qu'on lui porte le rend trop vulnérable ? Cette idée me semble plus acceptable, mais nécessite que les coûts soient particulièrement bien estimés (Quels aspects de l'impact environnemental, multifactoriel par définition, sont pris en compte ? Comment ces différents impacts sont combinés ? Comportent-ils des biais ? Quels contournements et dérives pourraient évoluer en réponse à de tels choix ?). Dans ce contexte, il serait peut-être intéressant de s'appuyer d'avantage sur l'expérience des comités éthiques pour l'échantillonnage humain et animal, qui ont l'habitude d'évaluer si les coûts d'un projet sont justifiés, et qui proscrivent certaines pratiques considérées trop coûteuses (par exemple invasive) pour les animaux. Nous pourrions élargir ces comités pour y considérer systématiquement les coûts environnementaux de nos projets de recherche. Il serait également souhaitable qu'à l'avenir, cela soit incontournable pour les demandes de financement.

Entre temps, une autre option (certes, moins ambitieuse, mais qui me semble plus à portée de main), est d'essayer de se fixer des objectifs de réduction relative ou absolue, pour permettre d'effectuer à peu près les mêmes projets mais à moindre cout, en d'autres mots d'optimiser nos stratégies d'échantillonnage, de séquençage, et de calcul pour minimiser leurs effets. C'est à cette échelle que je me situe pour l'instant, en essayant de réfléchir systématiquement aux couts et bénéfices de chaque déplacement en avion, de chaque hébergement en hôtel, de chaque acquisition de données, de chaque stockage de matériel (numérique, congélateurs), et de chaque calcul informatique. Cela ne veut pas dire qu'il faut complètement se restreindre, mais que chaque décision soit prise de manière éclairée sur ces aspects. Nous avons souvent peur de perdre des données précieuses, mais je pense qu'il est important de savoir faire régulièrement du tri, informatique comme biologique.

En terme d'éthique scientifique, je suis heureuse de voir que notre communauté scientifique est, dans son ensemble, de plus en plus sensible aux questions d'intégrité, de qualité (plutôt que de quantité) et de libre accès de la science. J'essaye d'aller personnellement dans cette direction en questionnant le modèle de financement et l'éthique des journaux dans lesquels je publie, en contribuant à rendre mes données accessibles et réutilisables par d'autres, et en participant, au sein du CoNRS, à une évaluation de la recherche d'avantage portée sur la qualité des travaux et prenant en compte les principes FAIR (Findable, Accessible, Interoperable, Reusable). Dans le champ de la génétique humaine, ces injonctions se heurtent parfois au besoin d'anonymat des données, et je trouve cela regrettable que les chercheur.se.s se retrouvent responsables de décider eux-mêmes dans quel cadre ils acceptent de partager leurs données (mise à disposition des données 'upon request').

Pour ces enjeux environnementaux et éthiques, il me semble inévitable, en tant que scientifique dans les sciences du vivant et de l'environnement, de montrer l'exemple au reste des communautés scientifiques, car nous sommes censés être les plus sensibles et les mieux informés pour porter ces changements. Je pense sincèrement que, même si le changement est lent, notre communauté porte globalement des valeurs avec lesquelles je me sens alignée.

Références bibliographiques

- 1000 Genomes Project Consortium. 2010. « A Map of Human Genome Variation from Population-Scale Sequencing ». *Nature* 467(7319): 1061-73. doi:10.1038/nature09534.
- 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. « A Global Reference for Human Genetic Variation ». *Nature* 526(7571): 68-74. doi:10.1038/nature15393.
- Abegaz, Silamlak Birhanu. 2021. « Human ABO Blood Groups and Their Associations with Different Diseases ». *BioMed Research International* 2021: 6629060. doi:10.1155/2021/6629060.
- Albers, Patrick K., et Gil McVean. 2020. « Dating genomic variants and shared ancestry in population-scale sequencing data ». *PLoS Biology* 18(1): e3000586. doi:10.1371/journal.pbio.3000586.
- Alharbi, Adel F., Nongfei Sheng, Katie Nicol, Nicklas Strömberg, et Edward J. Hollox. 2022. « Balancing Selection at the Human Salivary Agglutinin Gene (DMBT1) Driven by Host-Microbe Interactions ». *iScience* 25(5): 104189. doi:10.1016/j.isci.2022.104189.
- Allentoft, Morten E., Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, et al. 2015. « Population genomics of Bronze Age Eurasia ». *Nature* 522: 167-72. doi:10.1038/nature14507.
- Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. « A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome ». *Nature Biotechnology* 39(1): 105-14. doi:10.1038/s41587-020-0603-3.
- Andrés, Aida M., Melissa J. Hubisz, Amit Indap, Dara G. Torgerson, Jeremiah D. Degenhardt, Adam R. Boyko, Ryan N. Gutenkunst, et al. 2009. « Targets of Balancing Selection in the Human Genome ». *Molecular Biology and Evolution* 26(12): 2755-64. doi:10.1093/molbev/msp190.
- Antonio, Margaret L., Ziyue Gao, Hannah M. Moots, Michaela Lucci, Francesca Candilio, Susanna Sawyer, Victoria Oberreiter, et al. 2019. « Ancient Rome: A genetic crossroads of Europe and the Mediterranean ». *Science* 366(6466): 708-14. doi:10.1126/science.aay6826.
- Arumugam, Manimozhayan, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, et al. 2011. « Enterotypes of the Human Gut Microbiome ». *Nature* 473(7346): 174-80. doi:10.1038/nature09944.
- Arumugam, Manimozhayan, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, et al. 2014. « Addendum: Enterotypes of the Human Gut Microbiome ». *Nature* 506(7489): 516-516. doi:10.1038/nature13075.
- Ashby, Ben, et Mike Boots. 2017. « Multi-Mode Fluctuating Selection in Host-Parasite Coevolution ». *Ecology Letters* 20(3): 357-65. doi:10.1111/ele.12734.
- Aspholm-Hurtig, M., G. Dailide, M. Lahmann, A. Kalia, D. Ilver, N. Roche, S. Vikstrom, et al. 2004. « Functional Adaptation of BabA, the H. Pylori ABO Blood Group Antigen Binding Adhesin ». *Science* 305(5683): 519-22. doi:10.1126/science.1098801.
- Asthana, Saurabh, Steffen Schmidt, et Shamil Sunyaev. 2005. « A Limited Role for Balancing Selection ». *Trends in genetics: TIG* 21(1): 30-32. doi:10.1016/j.tig.2004.11.001.
- Audebert, Christophe, Gaël Even, Amandine Cian, Blastocystis Investigation Group, Alexandre Loywick, Sophie Merlin, Eric Viscogliosi, et Magali Chabé. 2016. « Colonization with the Enteric Protozoa Blastocystis Is Associated with Increased Diversity of Human Gut Bacterial Microbiota ». *Scientific Reports* 6: 25255. doi:10.1038/srep25255.
- Auton, A., A. Fledel-Alon, S. Pfeifer, O. Venn, Laure Ségurel, T. Street, E. M. Leffler, et al. 2012. « A fine-scale chimpanzee genetic map from population sequencing ». *Science* 336(6078): 193-98. doi:10.1126/science.1216872.
- Berdan, Emma L., Nicholas H. Barton, Roger Butlin, Brian Charlesworth, Rui Faria, Inês Fragata, Kimberly J. Gilbert, et al. 2023. « How Chromosomal Inversions Reorient the Evolutionary Process ». *Journal of Evolutionary Biology* 36(12): 1761-82. doi:10.1111/jeb.14242.
- Bergland, Alan O., Emily L. Behrman, Katherine R. O'Brien, Paul S. Schmidt, et Dmitri A. Petrov. 2014. « Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila* ». *PLoS genetics* 10(11): e1004775. doi:10.1371/journal.pgen.1004775.
- Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, et J. N. Hirschhorn. 2004. « Genetic signatures of strong recent positive selection at the lactase gene ». *Am J Hum Genet* 74(6): 1111-20. doi:10.1086/421051.
- Bitarello, Bárbara D, Débora Y C Brandt, Diogo Meyer, et Aida M Andrés. 2023. « Inferring Balancing Selection From Genome-Scale Data ». *Genome Biology and Evolution* 15(3): evad032. doi:10.1093/gbe/evad032.

- Bitarello, Bárbara D, Cesare de Filippo, João C Teixeira, Joshua M Schmidt, Philip Kleinert, Diogo Meyer, et Aida M Andrés. 2018. « Signatures of Long-Term Balancing Selection in Human Genomes ». *Genome Biology and Evolution* 10(3): 939-55. doi:10.1093/gbe/evy054.
- Blaser, M. J., et S. Falkow. 2009. « What are the consequences of the disappearing human microbiota? » *Nat Rev Microbiol* 7(12): 887-94. doi:10.1038/nrmicro2245.
- Bonder, M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, A. V. Vila, P. Deelen, et al. 2016. « The effect of host genetics on the gut microbiome ». *Nat Genet* 48(11): 1407-12. doi:10.1038/ng.3663.
- Booker, Tom R., Benjamin C. Jackson, et Peter D. Keightley. 2017. « Detecting Positive Selection in the Genome ». *BMC biology* 15(1): 98. doi:10.1186/s12915-017-0434-y.
- Boto, Luis. 2010. « Horizontal Gene Transfer in Evolution: Facts and Challenges ». *Proceedings. Biological Sciences* 277(1683): 819-27. doi:10.1098/rspb.2009.1679.
- Boyle, Evan A., Yang I. Li, et Jonathan K. Pritchard. 2017. « An Expanded View of Complex Traits: From Polygenic to Omnigenic ». *Cell* 169(7): 1177-86. doi:10.1016/j.cell.2017.05.038.
- Bradley, Brenda J, Diane M Doran-Sheehy, et Linda Vigilant. 2007. « Potential for Female Kin Associations in Wild Western Gorillas despite Female Dispersal ». *Proceedings of the Royal Society B: Biological Sciences* 274(1622): 2179-85. doi:10.1098/rspb.2007.0407.
- Braichenko, Svitlana, Rui Borges, et Carolin Kosiol. 2024. « Polymorphism-Aware Models in RevBayes: Species Trees, Disentangling Balancing Selection and CG-Biased Gene Conversion ». *Molecular Biology and Evolution: msae138*. doi:10.1093/molbev/msae138.
- Brisson, Dustin. 2018. « Negative Frequency-Dependent Selection Is Frequently Confounding ». *Frontiers in Ecology and Evolution* 6: 10. doi:10.3389/fevo.2018.00010.
- Brown, Elizabeth A. 2012. « Genetic Explorations of Recent Human Metabolic Adaptations: Hypotheses and Evidence ». *Biological Reviews of the Cambridge Philosophical Society* 87(4): 838-55. doi:10.1111/j.1469-185X.2012.00227.x.
- Brown, et Aurélien Tellier. 2011. « Plant-Parasite Coevolution: Bridging the Gap between Genetics and Ecology ». *Annual Review of Phytopathology* 49: 345-67. doi:10.1146/annurev-phyto-072910-095301.
- Bubb, K. L., D. Bovee, D. Buckley, E. Haugen, M. Kibukawa, M. Paddock, A. Palmieri, et al. 2006. « Scan of Human Genome Reveals No New Loci under Ancient Balancing Selection ». *Genetics* 173(4): 2165-77. doi:10.1534/genetics.106.055715.
- Buckley, Stephen, Karen Hardy, Fredrik Hallgren, Lucy Kubiak-Martens, Žydrūnė Miliauskienė, Alison Sheridan, Iwona Sobkowiak-Tabaka, et Maria Eulalia Subirà. 2023. « Human Consumption of Seaweed and Freshwater Aquatic Plants in Ancient Europe ». *Nature Communications* 14(1): 6192. doi:10.1038/s41467-023-41671-2.
- Burri, Reto, Alexander Nater, Takeshi Kawakami, Carina F. Mugal, Pall I. Olason, Linnea Smeds, Alexander Suh, et al. 2015. « Linked Selection and Recombination Rate Variation Drive the Evolution of the Genomic Landscape of Differentiation across the Speciation Continuum of *Ficedula* Flycatchers ». *Genome Research* 25(11): 1656-65. doi:10.1101/gr.196485.115.
- Calafell, F., F. Roubinet, A. Ramirez-Soriano, N. Saitou, J. Bertranpetit, et A. Blancher. 2008. « Evolutionary Dynamics of the Human ABO Gene ». *Hum Genet* 124(2): 123-35. doi:10.1007/s00439-008-0530-8.
- Cardona, Alexia, Luca Pagani, Tiago Antao, Daniel J. Lawson, Christina A. Eichstaedt, Bryndis Yngvadottir, Ma Than Than Shwe, et al. 2014. « Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations ». *PLOS ONE* 9(5): e98076. doi:10.1371/journal.pone.0098076.
- Carius, H. J., T. J. Little, et D. Ebert. 2001. « Genetic Variation in a Host-Parasite Association: Potential for Coevolution and Frequency-Dependent Selection ». *Evolution; International Journal of Organic Evolution* 55(6): 1136-45. doi:10.1111/j.0014-3820.2001.tb00633.x.
- Cassidy, Lara M., Rui Martiniano, Eileen M. Murphy, Matthew D. Teasdale, James Mallory, Barrie Hartwell, et Daniel G. Bradley. 2016. « Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome ». *Proceedings of the National Academy of Sciences of the United States of America* 113(2): 368-73. doi:10.1073/pnas.1518445113.
- Cavalli-Sforza, L. Luca, et Marcus W. Feldman. 2003. « The Application of Molecular Genetic Approaches to the Study of Human Evolution ». *Nature Genetics* 33 Suppl: 266-75. doi:10.1038/ng1113.
- Chabe, M., A. Lokmer, et L. Ségurel. 2017. « Gut Protozoa: Friends or Foes of the Human Gut Microbiota? » *Trends Parasitol.* doi:10.1016/j.pt.2017.08.005.
- Chao, Anne, Chun-Huo Chiu, et Lou Jost. 2010. « Phylogenetic Diversity Measures Based on Hill Numbers ». *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365(1558): 3599-3609. doi:10.1098/rstb.2010.0272.
- Charati, H., M. S. Peng, W. Chen, X. Y. Yang, R. Jabbari Ori, M. Aghajanjpour-Mir, A. Esmailzadeh, et Y. P. Zhang. 2019. « The evolutionary genetics of lactase persistence in seven ethnic groups across the Iranian plateau ». *Hum Genomics* 13(1): 7. doi:10.1186/s40246-019-0195-5.

- Charlesworth, Deborah. 2006. « Balancing Selection and Its Effects on Sequences in Nearby Genome Regions ». *PLoS genetics* 2(4): e64. doi:10.1371/journal.pgen.0020064.
- Chase, Madeline A., Hans Ellegren, et Carina F. Mugal. 2021. « Positive Selection Plays a Major Role in Shaping Signatures of Differentiation across the Genomic Landscape of Two Independent *Ficedula* Flycatcher Species Pairs ». *Evolution; International Journal of Organic Evolution* 75(9): 2179-96. doi:10.1111/evo.14234.
- Chen, Lianmin, Daoming Wang, Sanzhima Garmeva, Alexander Kurilshikov, Arnau Vich Vila, Ranko Gacesa, Trishla Sinha, et al. 2021. « The long-term genetic stability and individual specificity of the human gut microbiome ». *Cell* 184(9): 2302-2315.e12. doi:10.1016/j.cell.2021.03.024.
- Cheng, Xiaoheng, et Michael DeGiorgio. 2019. « Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism ». *Molecular Biology and Evolution* 36(1): 177-99. doi:10.1093/molbev/msy202.
- Cheng, Xiaoheng, et Michael DeGiorgio. 2020. « Flexible Mixture Model Approaches That Accommodate Footprint Size Variability for Robust Detection of Balancing Selection ». *Molecular Biology and Evolution* 37(11): 3267-91. doi:10.1093/molbev/msaa134.
- Cheng, Xiaoheng, et Michael DeGiorgio. 2022. « BaLeRMix +: Mixture Model Approaches for Robust Joint Identification of Both Positive Selection and Long-Term Balancing Selection ». *Bioinformatics (Oxford, England)* 38(3): 861-63. doi:10.1093/bioinformatics/btab720.
- Chester, M. A., et M. L. Olsson. 2001. « The ABO Blood Group Gene: A Locus of Considerable Genetic Diversity ». *Transfus Med Rev* 15(3): 177-200. doi:10.1053/tmrv.2001.24591.
- Cho, I., et M. J. Blaser. 2012. « The human microbiome: at the interface of health and disease ». *Nat Rev Genet* 13(4): 260-70. doi:10.1038/nrg3182.
- Clemente, J. C., E. C. Pehrsson, M. J. Blaser, K. Sandhu, Z. Gao, B. Wang, M. Magris, et al. 2015. « The microbiome of uncontacted Amerindians ». *Sci Adv* 1(3). doi:10.1126/sciadv.1500183.
- Colagiuri, S., et J. Brand Miller. 2002. « The 'Carnivore Connection'--Evolutionary Aspects of Insulin Resistance ». *European Journal of Clinical Nutrition* 56 Suppl 1: S30-35. doi:10.1038/sj.ejcn.1601351.
- Connallon, Tim, et Colin Olito. 2022. « Natural Selection and the Distribution of Chromosomal Inversion Lengths ». *Molecular Ecology* 31(13): 3627-41. doi:10.1111/mec.16091.
- Cook, G. C., et M. T. al-Torki. 1975. « High intestinal lactase concentrations in adult Arabs in Saudi Arabia ». *Br Med J* 3(5976): 135-36.
- Cordain, L., M.SS Hickey, et K. Kim. 2012. « Malaria and rickets represent selective forces for the convergent evolution of adult lactase persistence ». In *Biodiversity in Agriculture: Domestication, Evolution, and Sustainability*, éd. P. Gepts, Famula T.R., et Bettinger R.L. Cambridge University Press, 299-308.
- Croze, Myriam, Daniel Živković, Wolfgang Stephan, et Stephan Hutter. 2016. « Balancing Selection on Immunity Genes: Review of the Current Literature and New Analysis in *Drosophila Melanogaster* ». *Zoology (Jena, Germany)* 119(4): 322-29. doi:10.1016/j.zool.2016.03.004.
- De Filippo, C., D. Cavalieri, M. Di Paola, M. Ramazzotti, J. B. Poullet, S. Massart, S. Collini, G. Pieraccini, et P. Lionetti. 2010. « Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa ». *Proc Natl Acad Sci U S A* 107(33): 14691-96. doi:10.1073/pnas.1005963107.
- Deehan, E. C., et J. Walter. 2016. « The Fiber Gap and the Disappearing Gut Microbiome: Implications for Human Nutrition ». *Trends Endocrinol Metab* 27(5): 239-42. doi:10.1016/j.tem.2016.03.001.
- DeGiorgio, Michael, Kirk E. Lohmueller, et Rasmus Nielsen. 2014. « A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data ». *PLoS genetics* 10(8): e1004561. doi:10.1371/journal.pgen.1004561.
- Dehasque, Marianne, María C. Ávila-Arcos, David Díez-del-Molino, Matteo Fumagalli, Katerina Guschanski, Eline D. Lorenzen, Anna-Sapfo Malaspinas, et al. 2020. « Inference of natural selection from ancient DNA ». *Evolution Letters* 4(2): 94-108. doi:10.1002/evl3.165.
- Delaneau, Olivier, Jonathan Marchini, 1000 Genomes Project Consortium, et 1000 Genomes Project Consortium. 2014. « Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel ». *Nature Communications* 5: 3934. doi:10.1038/ncomms4934.
- Delaneau, Olivier, Jean-Francois Zagury, et Jonathan Marchini. 2013. « Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies ». *Nature Methods* 10(1): 5-6. doi:10.1038/nmeth.2307.
- Delph, Lynda F., et John K. Kelly. 2014. « On the Importance of Balancing Selection in Plants ». *The New Phytologist* 201(1): 45-56. doi:10.1111/nph.12441.
- Diamond, Jared, et Peter Bellwood. 2003. « Farmers and Their Languages: The First Expansions ». *Science (New York, N.Y.)* 300(5619): 597-603. doi:10.1126/science.1078208.
- Dikongué, E., et Laure Ségurel. 2017. « Latitude as a co-driver of human gut microbial diversity? » *Bioessays* 39(3). doi:10.1002/bies.201600145.
- Dillehay, Tom D., C. Ramírez, M. Pino, M. B. Collins, J. Rossen, et J. D. Pino-Navarro. 2008. « Monte Verde: Seaweed, Food, Medicine, and the Peopling of South America ». *Science (New York, N.Y.)* 320(5877): 784-86. doi:10.1126/science.1156533.

- Dobzhansky, T. 1955. « A Review of Some Fundamental Concepts and Problems of Population Genetics ». Cold Spring Harbor Symposia on Quantitative Biology 20: 1-15. doi:10.1101/sqb.1955.020.01.003.
- Dulai, K. S., M. von Dornum, J. D. Mollon, et D. M. Hunt. 1999. « The Evolution of Trichromatic Color Vision by Opsin Gene Duplication in New World and Old World Primates ». *Genome Research* 9(7): 629-38.
- Duret, Laurent, et Nicolas Galtier. 2009. « Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes ». *Annual Review of Genomics and Human Genetics* 10: 285-311. doi:10.1146/annurev-genom-082908-150001.
- Durmaz, Esra, Envel Kerdaffrec, Georgios Katsianis, Martin Kapun, et Thomas Flatt. 2020. « How Selection Acts on Chromosomal Inversions ». In *eLS*, John Wiley & Sons, Ltd, 307-15. doi:10.1002/9780470015902.a0028745.
- Enattah, N. S., A. Trudeau, V. Pimenoff, L. Maiuri, S. Auricchio, L. Greco, M. Rossi, et al. 2007. « Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans ». *Am J Hum Genet* 81(3): 615-25. doi:10.1086/520705.
- Even, Gaël, Ana Lokmer, Jules Rodrigues, Christophe Audebert, Eric Viscogliosi, Laure Ségurel, et Magali Chabé. 2021. « Changes in the Human Gut Microbiota Associated With Colonization by *Blastocystis* Sp. and *Entamoeba* Spp. in Non-Industrialized Populations ». *Frontiers in Cellular and Infection Microbiology* 11: 533528. doi:10.3389/fcimb.2021.533528.
- Evershed, Richard P., George Davey Smith, Mélanie Roffet-Salque, Adrian Timpson, Yoan Diekmann, Matthew S. Lyon, Lucy J. E. Cramp, et al. 2022. « Dairying, Diseases and the Evolution of Lactase Persistence in Europe ». *Nature* 608(7922): 336-45. doi:10.1038/s41586-022-05010-7.
- Fabre, Alexandre, Anne Fabre, Céline Bon, Paul Guerry, et Laure Ségurel. 2023. « Proposed Mechanism for the Selection of Lactase Persistence in Childhood ». *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 45(7): e2200243. doi:10.1002/bies.202200243.
- Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, et al. 2013. « The long-term stability of the human gut microbiota ». *Science* 341(6141): 1237439. doi:10.1126/science.1237439.
- Falony, G., M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, et al. 2016. « Population-level analysis of gut microbiome variation ». *Science* 352(6285): 560-64. doi:10.1126/science.aad3503.
- Fan, Shaohua, Matthew E. B. Hansen, Yancy Lo, et Sarah A. Tishkoff. 2016. « Going global by adapting local: A review of recent human adaptation ». *Science (New York, N.Y.)* 354(6308): 54-59. doi:10.1126/science.aaf5098.
- Faria, Rui, Kerstin Johannesson, Roger K. Butlin, et Anja M. Westram. 2019. « Evolving Inversions ». *Trends in Ecology & Evolution* 34(3): 239-48. doi:10.1016/j.tree.2018.12.005.
- Ferrer-Admetlla, Anna, Elena Bosch, Martin Sikora, Tomàs Marquès-Bonet, Anna Ramírez-Soriano, Aura Muntasell, Arcadi Navarro, et al. 2008. « Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes ». *Journal of Immunology (Baltimore, Md.: 1950)* 181(2): 1315-22. doi:10.4049/jimmunol.181.2.1315.
- Fijarczyk, Anna, et Wiesław Babik. 2015. « Detecting Balancing Selection in Genomes: Limits and Prospects ». *Molecular Ecology* 24(14): 3529-45. doi:10.1111/mec.13226.
- Flatz, G., et H. W. Rotthauwe. 1973. « Lactose nutrition and natural selection ». *Lancet* 2(7820): 76-77.
- Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, et al. 2007. « A Second Generation Human Haplotype Map of over 3.1 Million SNPs ». *Nature* 449(7164): 851-61. doi:10.1038/nature06258.
- Fuhrman, J. A., J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green, et J. H. Brown. 2008. « A latitudinal diversity gradient in planktonic marine bacteria ». *Proc Natl Acad Sci U S A* 105(22): 7774-78. doi:10.1073/pnas.0803070105.
- Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H. Stengård, V. Salomaa, et al. 2000. « Apolipoprotein E Variation at the Sequence Haplotype Level: Implications for the Origin and Maintenance of a Major Human Polymorphism ». *American Journal of Human Genetics* 67(4): 881-900. doi:10.1086/303070.
- Fumagalli, M., U. Pozzoli, R. Cagliani, G. P. Comi, S. Riva, M. Clerici, N. Bresolin, et M. Sironi. 2009. « Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions ». *J Exp Med* 206(6): 1395-1408. doi:10.1084/jem.20082779.
- Gagneux, P., et A. Varki. 1999. « Evolutionary Considerations in Relating Oligosaccharide Diversity to Biological Function ». *Glycobiology* 9(8): 747-55. doi:10.1093/glycob/9.8.747.
- Garratty, G. 2005. « Relationship of blood groups to disease: Do blood group antigens have a biological role? » *Rev Med Inst Mex Seguro Soc* 43(Suppl 1): 113-21.
- Gaulke, Christopher A., Holly K. Arnold, Ian R. Humphreys, Steven W. Kembel, James P. O'Dwyer, et Thomas J. Sharpton. 2018. « Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota ». *mBio* 9(5): e01348-18. doi:10.1128/mBio.01348-18.

- Gerbault, P., A. Liebert, Y. Itan, A. Powell, M. Currat, J. Burger, D. M. Swallow, et M. G. Thomas. 2011. « Evolution of lactase persistence: an example of human niche construction ». *Philos Trans R Soc Lond B Biol Sci* 366(1566): 863-77. doi:10.1098/rstb.2010.0268.
- Gevers, D., R. Knight, J. F. Petrosino, K. Huang, A. L. McGuire, B. W. Birren, K. E. Nelson, et al. 2012. « The Human Microbiome Project: a community resource for the healthy human microbiome ». *PLoS Biol* 10(8): e1001377. doi:10.1371/journal.pbio.1001377.
- Glass, R. I., J. Holmgren, C. E. Haley, M. R. Khan, A. M. Svennerholm, B. J. Stoll, K. M. Belayet Hossain, et al. 1985. « Predisposition for Cholera of Individuals with O Blood Group. Possible Evolutionary Significance ». *American Journal of Epidemiology* 121(6): 791-96. doi:10.1093/oxfordjournals.aje.a114050.
- Goldberg, Emma E., Joshua R. Kohn, Russell Lande, Kelly A. Robertson, Stephen A. Smith, et Boris Igić. 2010. « Species Selection Maintains Self-Incompatibility ». *Science (New York, N.Y.)* 330(6003): 493-95. doi:10.1126/science.1194513.
- Gomez, A., K. J. Petrzalkova, M. B. Burns, C. J. Yeoman, K. R. Amato, K. Vlckova, D. Modry, et al. 2016. « Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns ». *Cell Rep* 14(9): 2142-53. doi:10.1016/j.celrep.2016.02.013.
- Goodrich, J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, C. Ober, T. D. Spector, et al. 2016. « Genetic Determinants of the Gut Microbiome in UK Twins ». *Cell Host Microbe* 19(5): 731-43. doi:10.1016/j.chom.2016.04.017.
- Groussin, Mathieu, Florent Mazel, Jon G. Sanders, Chris S. Smillie, Sébastien Lavergne, Wilfried Thuiller, et Eric J. Alm. 2017. « Unraveling the Processes Shaping Mammalian Gut Microbiomes over Evolutionary Time ». *Nature Communications* 8: 14319. doi:10.1038/ncomms14319.
- Groussin, Mathieu, Florent Mazel, et Eric J. Alm. 2020. « Co-Evolution and Co-Speciation of Host-Gut Bacteria Systems ». *Cell Host & Microbe* 28(1): 12-22. doi:10.1016/j.chom.2020.06.013.
- Groussin, Mathieu, Mathilde Poyet, Ainara Sistiaga, Sean M. Kearney, Katya Moniz, Mary Noel, Jeff Hooker, et al. 2021. « Elevated Rates of Horizontal Gene Transfer in the Industrialized Human Microbiome ». *Cell* 184(8): 2053-2067.e18. doi:10.1016/j.cell.2021.02.052.
- GTEx Consortium. 2013. « The Genotype-Tissue Expression (GTEx) Project ». *Nature Genetics* 45(6): 580-85. doi:10.1038/ng.2653.
- Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. « Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe ». *Nature* 522(7555): 207-11. doi:10.1038/nature14317.
- Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. « Characterizing mutagenic effects of recombination through a sequence-level genetic map ». *Science* 363(6425): eaau1043. doi:10.1126/science.aau1043.
- Hallmark, Brian, Tatiana M Karafet, PingHsun Hsieh, Ludmila P Osipova, Joseph C Watkins, et Michael F Hammer. 2019. « Genomic Evidence of Local Adaptation to Climate and Diet in Indigenous Siberians ». *Molecular Biology and Evolution* 36(2): 315-27. doi:10.1093/molbev/msy211.
- Hancock, Angela M., Gorka Alkorta-Aranburu, David B. Witonsky, et Anna Di Rienzo. 2010. « Adaptations to New Environments in Humans: The Role of Subtle Allele Frequency Shifts ». *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365(1552): 2459-68. doi:10.1098/rstb.2010.0032.
- Hancock, Angela M., David B. Witonsky, Adam S. Gordon, Gidon Eshel, Jonathan K. Pritchard, Graham Coop, et Anna Di Rienzo. 2008. « Adaptations to Climate in Candidate Genes for Common Metabolic Disorders ». *PLoS genetics* 4(2): e32. doi:10.1371/journal.pgen.0040032.
- Hassani, M. Amine, Paloma Durán, et Stéphane Hacquard. 2018. « Microbial Interactions within the Plant Holobiont ». *Microbiome* 6(1): 58. doi:10.1186/s40168-018-0445-0.
- Hayeck, Tristan J., Yang Li, Timothy L. Mosbrugger, Jonathan P. Bradfield, Adam G. Gleason, George Damianos, Grace Tzun-Wen Shaw, et al. 2024. « The Impact of Patterns in Linkage Disequilibrium and Sequencing Quality on the Imprint of Balancing Selection ». *Genome Biology and Evolution* 16(2): evae009. doi:10.1093/gbe/evae009.
- Hedrick, P. W. 1998. « Balancing Selection and MHC ». *Genetica* 104(3): 207-14. doi:10.1023/a:1026494212540.
- Hedrick, P. W. 2007. « Balancing Selection ». *Current biology: CB* 17(7): R230-231. doi:10.1016/j.cub.2007.01.012.
- Hedrick, P. W. 2011. « Population genetics of malaria resistance in humans ». *Heredity (Edinb)* 107(4): 283-304. doi:10.1038/hdy.2011.16.
- Hehemann, Jan-Hendrik, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, et Gurvan Michel. 2010. « Transfer of Carbohydrate-Active Enzymes from Marine Bacteria to Japanese Gut Microbiota ». *Nature* 464(7290): 908-12. doi:10.1038/nature08937.

- Helgason, Agnar, Snaebjörn Pálsson, Gudmar Thorleifsson, Struan F. A. Grant, Valur Emilsson, Steinunn Gunnarsdóttir, Adebawale Adeyemo, et al. 2007. « Refining the Impact of TCF7L2 Gene Variants on Type 2 Diabetes and Adaptive Evolution ». *Nature Genetics* 39(2): 218-25. doi:10.1038/ng1960.
- Hertzler, S. R., et S. M. Clancy. 2003. « Kefir improves lactose digestion and tolerance in adults with lactose maldigestion ». *J Am Diet Assoc* 103(5): 582-87. doi:10.1053/jada.2003.50111.
- Heyer, E., L. Brazier, L. Ségurel, T. Hegay, F. Austerlitz, L. Quintana-Murci, M. Georges, P. Pasquet, et M. Veuille. 2011. « Lactase persistence in central Asia: phenotype, genotype, and evolution ». *Hum Biol* 83(3): 379-92. doi:10.3378/027.083.0304.
- Hiwatashi, Tomohide, Yugo Okabe, Toko Tsutsui, Chihiro Hiramatsu, Amanda D. Melin, Hiroki Oota, Colleen M. Schaffner, et al. 2010. « An Explicit Signature of Balancing Selection for Color-Vision Variation in New World Monkeys ». *Molecular Biology and Evolution* 27(2): 453-64. doi:10.1093/molbev/msp262.
- Hoffmann, Ary A., Carla M. Sgrò, et Andrew R. Weeks. 2004. « Chromosomal Inversion Polymorphisms and Adaptation ». *Trends in Ecology & Evolution* 19(9): 482-88. doi:10.1016/j.tree.2004.06.013.
- Hoh, Boon-Peng, Thuhairah Abdul Rahman, et Khalid Yusoff. 2019. « Natural selection and local adaptation of blood pressure regulation and their perspectives on precision medicine in hypertension ». *Hereditas* 156(1): 1. doi:10.1186/s41065-019-0080-1.
- Holden, C., et R. Mace. 1997. « Phylogenetic analysis of the evolution of lactose digestion in adults ». *Hum Biol* 69(5): 605-28.
- Huang, P., T. Farkas, S. Marionneau, W. Zhong, N. Ruvoen-Clouet, A. L. Morrow, M. Altaye, et al. 2003. « Noroviruses Bind to Human ABO, Lewis, and Secretor Histo-Blood Group Antigens: Identification of 4 Distinct Strain-Specific Patterns ». *J Infect Dis* 188(1): 19-31. doi:10.1086/375742.
- Isildak, Ulas, Alessandro Stella, et Matteo Fumagalli. 2021. « Distinguishing between Recent Balancing Selection and Incomplete Sweep Using Deep Neural Networks ». *Molecular Ecology Resources* 21(8): 2706-18. doi:10.1111/1755-0998.13379.
- Itan, Y., A. Powell, M. A. Beaumont, J. Burger, et M. G. Thomas. 2009. « The origins of lactase persistence in Europe ». *PLoS Comput Biol* 5(8): e1000491. doi:10.1371/journal.pcbi.1000491.
- Jay, Paul, Annabel Whibley, Lise Frézal, María Ángeles Rodríguez de Cara, Reuben W. Nowell, James Mallet, Kanchon K. Dasmahapatra, et Mathieu Joron. 2018. « Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion ». *Current Biology* 28(11): 1839-1845.e3. doi:10.1016/j.cub.2018.04.072.
- Johnson, Olivia L., Raymond Tobler, Joshua M. Schmidt, et Christian D. Huber. 2023. « Fluctuating Selection and the Determinants of Genetic Variation ». *Trends in genetics: TIG* 39(6): 491-504. doi:10.1016/j.tig.2023.02.004.
- Joron, Mathieu, Ian R. Wynne, Gerardo Lamas, et James Mallet. 1999. « Variable Selection and the Coexistence of Multiple Mimetic Forms of the Butterfly *Heliconius Numata* ». *Evolutionary Ecology* 13(7): 721-54. doi:10.1023/A:1010875213123.
- Jost, Lou. 2006. « Entropy and Diversity ». *Oikos* 113(2): 363-75. doi:10.1111/j.2006.0030-1299.14714.x.
- Juras, Anna, Maciej Chyleński, Edvard Ehler, Helena Malmström, Danuta Żurkiewicz, Piotr Włodarczyk, Stanisław Wilk, et al. 2018. « Mitochondrial Genomes Reveal an East to West Cline of Steppe Ancestry in Corded Ware Populations ». *Scientific Reports* 8(1): 11603. doi:10.1038/s41598-018-29914-5.
- Kapun, Martin et Thomas Flatt. 2019. « The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster* ». *Molecular Ecology* 28(6): 1263-1282. doi:10.1111/mec.14871
- Kapun, Martin, Esra Durmaz Mitchell, Tadeusz J. Kawecki, Paul Schmidt, et Thomas Flatt. 2023. « An Ancestral Balanced Inversion Polymorphism Confers Global Adaptation ». *Molecular Biology and Evolution* 40(6): msad118. doi:10.1093/molbev/msad118.
- Karasov, Talia L., Joel M. Kniskern, Liping Gao, Brody J. DeYoung, Jing Ding, Ullrich Dubiella, Ruben O. Lastra, et al. 2014. « The Long-Term Maintenance of a Resistance Polymorphism through Diffuse Interactions ». *Nature* 512(7515): 436-40. doi:10.1038/nature13439.
- Kennedy Overton, Elizabeth, Alice Bernard, Pierre-Cyril Renaud, Grant Hall, Chloe Guerbois, Hervé Fritz, Franck Prugnolle, et Virginie Rougeron. 2023. Land use influence on chacma baboon (*Papio ursinus*) diet in South Africa using stable isotopes. doi:10.21203/rs.3.rs-3427689/v1.
- Kermarrec, N., F. Roubinet, P. A. Apoil, et A. Blancher. 1999. « Comparison of Allele O Sequences of the Human and Non-Human Primate ABO System ». *Immunogenetics* 49(6): 517-26.
- Kerner, Gaspard, Guillaume Laval, Etienne Patin, Stéphanie Boisson-Dupuis, Laurent Abel, Jean-Laurent Casanova, et Lluís Quintana-Murci. 2021. « Human Ancient DNA Analyses Reveal the High Burden of Tuberculosis in Europeans over the Last 2,000 Years ». *The American Journal of Human Genetics* 108(3): 517-24. doi:10.1016/j.ajhg.2021.02.009.
- Key, Felix M., Qiaomei Fu, Frédéric Romagné, Michael Lachmann, et Aida M. Andrés. 2016. « Human adaptation and population differentiation in the light of ancient genomes ». *Nature Communications* 7: 10775. doi:10.1038/ncomms10775.

- Key, Felix M., João C. Teixeira, Cesare de Filippo, et Aida M. Andrés. 2014. « Advantageous Diversity Maintained by Balancing Selection in Humans ». *Current Opinion in Genetics & Development* 29: 45-51. doi:10.1016/j.gde.2014.08.001.
- Kim, Chan Yeong, Muyeong Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, et Insuk Lee. 2021. « Human Reference Gut Microbiome Catalog Including Newly Assembled Genomes from Under-Represented Asian Metagenomes ». *Genome Medicine* 13(1): 134. doi:10.1186/s13073-021-00950-7.
- Kimura, Motoo. 1968. « Evolutionary Rate at the Molecular Level ». *Nature* 217(5129): 624-26. doi:10.1038/217624a0.
- Kinross, J. M., A. W. Darzi, et J. K. Nicholson. 2011. « Gut microbiome-host interactions in health and disease ». *Genome Med* 3(3): 14. doi:10.1186/gm228.
- Kirkpatrick, Mark, et Nick Barton. 2006. « Chromosome Inversions, Local Adaptation and Speciation ». *Genetics* 173(1): 419-34. doi:10.1534/genetics.105.047985.
- Kitano, T., R. Noda, O. Takenaka, et N. Saitou. 2009. « Relic of Ancient Recombinations in Gibbon ABO Blood Group Genes Deciphered through Phylogenetic Network Analysis ». *Mol Phylogenet Evol* 51(3): 465-71. doi:10.1016/j.ympev.2009.02.023.
- Klein, J., Y. Satta, C. O’Huin, et N. Takahata. 1993. « The Molecular Descent of the Major Histocompatibility Complex ». *Annu Rev Immunol* 11: 269-95. doi:10.1146/annurev.iy.11.040193.001413.
- Klimentidis, Yann C., Marshall Abrams, Jelai Wang, Jose R. Fernandez, et David B. Allison. 2011. « Natural Selection at Genomic Regions Associated with Obesity and Type-2 Diabetes: East Asians and Sub-Saharan Africans Exhibit High Levels of Differentiation at Type-2 Diabetes Regions ». *Human Genetics* 129(4): 407-18. doi:10.1007/s00439-010-0935-z.
- Knowler, W. C., P. H. Bennett, R. F. Hamman, et M. Miller. 1978. « Diabetes Incidence and Prevalence in Pima Indians: A 19-Fold Greater Incidence than in Rochester, Minnesota ». *American Journal of Epidemiology* 108(6): 497-505. doi:10.1093/oxfordjournals.aje.a112648.
- Knowler, W. C., D. J. Pettitt, P. H. Bennett, et R. C. Williams. 1983. « Diabetes Mellitus in the Pima Indians: Genetic and Evolutionary Considerations ». *American Journal of Physical Anthropology* 62(1): 107-14. doi:10.1002/ajpa.1330620114.
- Koenig, Daniel, Jörg Hagmann, Rachel Li, Felix Bemm, Tanja Slotte, Barbara Neuffer, Stephen I. Wright, et Detlef Weigel. 2019. « Long-Term Balancing Selection Drives Evolution of Immunity Genes in *Capsella* ». *eLife* 8: e43606. doi:10.7554/eLife.43606.
- Kominato, Y., P. D. McNeill, M. Yamamoto, M. Russell, S. Hakomori, et F. Yamamoto. 1992. « Animal Histo-Blood Group ABO Genes ». *Biochem Biophys Res Commun* 189(1): 154-64.
- Langergraber, Kevin E., Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, et al. 2012. « Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution ». *Proceedings of the National Academy of Sciences* 109(39): 15716-21. doi:10.1073/pnas.1211740109.
- Lawal, Raman Akinyanju, et Beth L. Dumont. 2023. « Ancestral Variation and Its Impact on Wild House Mouse Genomes ». doi:10.1101/2023.11.09.566486.
- Leffler, E. M., Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, O. Venn, R. Bowden, et al. 2013. « Multiple instances of ancient balancing selection shared between humans and chimpanzees ». *Science* 339(6127): 1578-82. doi:10.1126/science.1234070.
- Lewontin, R. C., et J. L. Hubby. 1966. « A Molecular Approach to the Study of Genic Heterozygosity in Natural Populations. II. Amount of Variation and Degree of Heterozygosity in Natural Populations of *Drosophila Pseudoobscura* ». *Genetics* 54(2): 595-609. doi:10.1093/genetics/54.2.595.
- Li, et R. Durbin. 2011. « Inference of human population history from individual whole-genome sequences ». *Nature* 475(7357): 493-96. doi:10.1038/nature10231.
- Li, J., D. Quinque, H. P. Horz, M. Li, M. Rzhetskaya, J. A. Raff, M. G. Hayes, et M. Stoneking. 2014. « Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa ». *BMC Microbiol* 14: 316. doi:10.1186/s12866-014-0316-1.
- Liao, Wen-Wei, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, et al. 2023. « A Draft Human Pangenome Reference ». *Nature* 617(7960): 312-24. doi:10.1038/s41586-023-05896-x.
- Lin, A., E. M. Bik, E. K. Costello, L. Dethlefsen, R. Haque, D. A. Relman, et U. Singh. 2013. « Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States ». *PLoS One* 8(1): e53838. doi:10.1371/journal.pone.0053838.
- Livingstone, F. B. 1960. « Natural Selection, Disease, and Ongoing Human Evolution, as Illustrated by the ABO Blood Groups ». *Human Biology* 32: 17-27.
- Livingstone, F. B. 1964. « Aspects of the Population Dynamics of the Abnormal Hemoglobin and Glucose-6-Phosphate Dehydrogenase Deficiency Genes ». *American Journal of Human Genetics* 16(4): 435-50.

- Llaurens, Violaine, Annabel Whibley, et Mathieu Joron. 2017. « Genetic Architecture and Balancing Selection: The Life and Death of Differentiated Variants ». *Molecular Ecology* 26(9): 2430-48. doi:10.1111/mec.14051.
- Lokmer, Ana, Sophie Aflalo, Norbert Amougou, Sophie Lafosse, Alain Froment, Francis Ekwin Tabe, Mathilde Poyet, et al. 2020. « Response of the Human Gut and Saliva Microbiome to Urbanization in Cameroon ». *Scientific Reports* 10(1): 2856. doi:10.1038/s41598-020-59849-9.
- Lokmer, Ana, Amandine Cian, Alain Froment, Nausicaa Gantois, Eric Viscogliosi, Magali Chabé, et Laure Ségurel. 2019. « Use of Shotgun Metagenomics for the Identification of Protozoa in the Gut Microbiota of Healthy Individuals from Worldwide Populations with Various Industrialization Levels ». *PloS One* 14(2): e0211139. doi:10.1371/journal.pone.0211139.
- Lu, Yi-Fan, David B. Goldstein, Misha Angrist, et Gianpiero Cavalleri. 2014. « Personalized Medicine and Human Genetic Diversity ». *Cold Spring Harbor Perspectives in Medicine* 4(9): a008581. doi:10.1101/cshperspect.a008581.
- Ly, Goki, Bérénice Alard, Romain Laurent, Sophie Lafosse, Bruno Toupance, Chou Monidarin, Gérard Diffloth, et al. 2018. « Residence Rule Flexibility and Descent Groups Dynamics Shape Uniparental Genetic Diversities in South East Asia ». *American Journal of Physical Anthropology* 165(3): 480-91. doi:10.1002/ajpa.23374.
- Ly, Goki, Romain Laurent, Sophie Lafosse, Chou Monidarin, Gérard Diffloth, Frédéric Bourdier, Olivier Evrard, et al. 2019. « From matrimonial practices to genetic diversity in Southeast Asian populations: the signature of the matrilineal puzzle ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 374(1780): 20180434. doi:10.1098/rstb.2018.0434.
- Machado, Heather E, Alan O Bergland, Ryan Taylor, Susanne Tilk, Emily Behrman, Kelly Dyer, Daniel K Fabian, et al. 2021. « Broad Geographic Sampling Reveals the Shared Basis and Environmental Correlates of Seasonal Adaptation in *Drosophila* ». *eLife* 10: e67577. doi:10.7554/eLife.67577.
- Malaria Genomic Epidemiology Network, Gavin Band, Kirk A. Rockett, Chris C. A. Spencer, et Dominic P. Kwiatkowski. 2015. « A Novel Locus of Resistance to Severe Malaria in a Region of Ancient Balancing Selection ». *Nature* 526(7572): 253-57. doi:10.1038/nature15390.
- Mancabelli, L., C. Milani, G. A. Lugli, F. Turroni, C. Ferrario, D. van Sinderen, et M. Ventura. 2017. « Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations ». *Environ Microbiol* 19(4): 1379-90. doi:10.1111/1462-2920.13692.
- de Manuel, Marc, Martin Kuhlwilm, Peter Frandsen, Vitor C. Sousa, Tariq Desai, Javier Prado-Martinez, Jessica Hernandez-Rodriguez, et al. 2016. « Chimpanzee Genomic Diversity Reveals Ancient Admixture with Bonobos ». *Science (New York, N.Y.)* 354(6311): 477-81. doi:10.1126/science.aag2602.
- Marchi, N., T. Hegay, P. Menecier, M. Georges, R. Laurent, M. Whitten, P. Endicott, et al. 2017. « Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations ». *Am J Phys Anthropol* 162(4): 627-40. doi:10.1002/ajpa.23151.
- Marchi, N., P. Menecier, M. Georges, S. Lafosse, T. Hegay, C. Dorzhu, B. Chichlo, Laure Ségurel, et E. Heyer. 2018. « Close inbreeding and low genetic diversity in Inner Asian human populations despite geographical exogamy ». *Sci Rep* 8(1): 9397. doi:10.1038/s41598-018-27047-3.
- Martinez, I., J. C. Stegen, M. X. Maldonado-Gomez, A. M. Eren, P. M. Siba, A. R. Greenhill, et J. Walter. 2015. « The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes ». *Cell Rep* 11(4): 527-38. doi:10.1016/j.celrep.2015.03.049.
- Masi, Shelly, Frédéric Austerlitz, Chloé Chabaud, Sophie Lafosse, Nina Marchi, Myriam Georges, Françoise Dessarps-Freichy, et al. 2021. « No Evidence for Female Kin Association, Indications for Extragroup Paternity, and Sex-Biased Dispersal Patterns in Wild Western Gorillas ». *Ecology and Evolution* 11(12): 7634-46. doi:10.1002/ece3.7596.
- Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szecsenyi-Nagy, N. Rohland, S. Mallick, I. Olalde, et al. 2018. « The genomic history of southeastern Europe ». *Nature* 555(7695): 197-203. doi:10.1038/nature25778.
- Mathieson, Iain, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, et al. 2015. « Genome-wide patterns of selection in 230 ancient Eurasians ». *Nature*: 1-16. doi:10.1038/nature16152.
- Mathieu, Sophie, Mélanie Touvrey, Laurent Poulet, Sophie Drouillard, Thirumalai Selvi Ulaganathan, Laure Ségurel, Miroslav Cygler, et William Helbert. 2023. « The Porphyran Degradation System of the Human Gut Microbiota Is Complete, Phylogenetically Diverse and Geographically Structured across Asian Populations ». : 2023.03.30.534863. doi:10.1101/2023.03.30.534863.
- McCutcheon, John P., et Nancy A. Moran. 2011. « Extreme Genome Reduction in Symbiotic Bacteria ». *Nature Reviews. Microbiology* 10(1): 13-26. doi:10.1038/nrmicro2670.
- McDonald, J. H., et M. Kreitman. 1991. « Adaptive Protein Evolution at the Adh Locus in *Drosophila* ». *Nature* 351(6328): 652-54. doi:10.1038/351652a0.

- McFall-Ngai, M., M. G. Hadfield, T. C. Bosch, H. V. Carey, T. Domazet-Loso, A. E. Douglas, N. Dubilier, et al. 2013. « Animals in a bacterial world, a new imperative for the life sciences ». *Proc Natl Acad Sci U S A* 110(9): 3229-36. doi:10.1073/pnas.1218525110.
- Meeks, Karlijn A C, Amy R Bentley, Adebowale A Adeyemo, et Charles N Rotimi. 2021. « Evolutionary forces in diabetes and hypertension pathogenesis in Africans ». *Human Molecular Genetics* 30(R1): R110-18. doi:10.1093/hmg/ddaa238.
- Mehta, Raaj S., Galeb S. Abu-Ali, David A. Drew, Jason Lloyd-Price, Ayshwarya Subramanian, Paul Lochhead, Amit D. Joshi, et al. 2018. « Stability of the Human Faecal Microbiome in a Cohort of Adult Men ». *Nature Microbiology* 3(3): 347-55. doi:10.1038/s41564-017-0096-0.
- Mérot, Claire, Rebekah A. Oomen, Anna Tigano, et Maren Wellenreuther. 2020a. « A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation ». *Trends in Ecology & Evolution* 35(7): 561-72. doi:10.1016/j.tree.2020.03.002.
- Mérot, Claire, Violaine Llaurens, Eric Normandeau, Louis Bernatchez, et Maren Wellenreuther. 2020b. « Balancing Selection via Life-History Trade-Offs Maintains an Inversion Polymorphism in a Seaweed Fly ». *Nature Communications* 11(1): 670. doi:10.1038/s41467-020-14479-7.
- Michanek, G. 2013. « Seaweed resources for pharmaceutical uses ». *De Gruyter*, 203-36. doi:10.1515/9783110882049.203.
- Miller, J. C., et S. Colagiuri. 1994. « The Carnivore Connection: Dietary Carbohydrate in the Evolution of NIDDM ». *Diabetologia* 37(12): 1280-86. doi:10.1007/BF00399803.
- Mittelbach, G. G., D. W. Schemske, H. V. Cornell, A. P. Allen, J. M. Brown, M. B. Bush, S. P. Harrison, et al. 2007. « Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography ». *Ecol Lett* 10(4): 315-31. doi:10.1111/j.1461-0248.2007.01020.x.
- Mittnik, Alissa, Chuan-Chao Wang, Saskia Pfrengle, Mantas Daubaras, Gunita Zariņa, Fredrik Hallgren, Raili Allmäe, et al. 2018. « The genetic prehistory of the Baltic Sea region ». *Nature Communications* 9: 442. doi:10.1038/s41467-018-02825-9.
- Moeller, Andrew H., Alejandro Caro-Quintero, Deus Mjungu, Alexander V. Georgiev, Elizabeth V. Lonsdorf, Martin N. Muller, Anne E. Pusey, et al. 2016. « Cospeciation of Gut Microbiota with Hominids ». *Science (New York, N.Y.)* 353(6297): 380-82. doi:10.1126/science.aaf3951.
- Morton, E. R., J. Lynch, A. Froment, S. Lafosse, E. Heyer, M. Przeworski, R. Blekhan, et L. Ségurel. 2015. « Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by Entamoeba and Subsistence ». *PLoS Genet* 11(11): e1005658. doi:10.1371/journal.pgen.1005658.
- Mosca, A., M. Leclerc, et J. P. Hugot. 2016. « Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? ». *Front Microbiol* 7: 455. doi:10.3389/fmicb.2016.00455.
- Moulds, J. M., S. Nowicki, J. J. Moulds, et B. J. Nowicki. 1996. « Human Blood Groups: Incidental Receptors for Viruses and Bacteria ». *Transfusion* 36(4): 362-74.
- Muktupavela, Rasa A, Martin Petr, Laure Ségurel, Thorfinn Korneliussen, John Novembre, et Fernando Racimo. 2022. « Modeling the spatiotemporal spread of beneficial alleles using ancient genomes » éd. Aida M Andrés, George H Perry, Isabel Alves, et Anders Eriksson. *eLife* 11: e73767. doi:10.7554/eLife.73767.
- Murray, Connor S., Madison Karram, David J. Bass, Madison Doceti, Dörthe Becker, Joaquin C. B. Nunez, Aakrosh Ratan, et Alan O. Bergland. 2024. « Balancing Selection and the Functional Effects of Shared Polymorphism in Cryptic Daphnia Species ». *bioRxiv: The Preprint Server for Biology*: 2024.04.16.589693. doi:10.1101/2024.04.16.589693.
- Nakayama, J., K. Watanabe, J. Jiang, K. Matsuda, S. H. Chao, P. Haryono, O. La-Ongkham, et al. 2015. « Diversity in gut bacterial community of school-age children in Asia ». *Sci Rep* 5: 8397. doi:10.1038/srep08397.
- Neel, J. V. 1962. « Diabetes Mellitus: A “Thrifty” Genotype Rendered Detrimental by “Progress”? » *American Journal of Human Genetics* 14(4): 353-62.
- Neel, J. V., A. B. Weder, et S. Julius. 1998. « Type II Diabetes, Essential Hypertension, and Obesity as “Syndromes of Impaired Genetic Homeostasis”: The “Thrifty Genotype” Hypothesis Enters the 21st Century ». *Perspectives in Biology and Medicine* 42(1): 44-74. doi:10.1353/pbm.1998.0060.
- Nei, Masatoshi. 2005. « Selectionism and Neutralism in Molecular Evolution ». *Molecular Biology and Evolution* 22(12): 2318-42. doi:10.1093/molbev/msi242.
- Nielsen, Rasmus, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, et Eske Willerslev. 2017. « Tracing the Peopling of the World through Genomics ». *Nature* 541(7637): 302-10. doi:10.1038/nature21347.
- Nosil, Patrik, Victor Soria-Carrasco, Romain Villoutreix, Marisol De-la-Mora, Clarissa F. de Carvalho, Thomas Parchman, Jeffrey L. Feder, et Zachariah Gompert. 2023. « Complex Evolutionary Processes Maintain an Ancient Chromosomal Inversion ». *Proceedings of the National Academy of Sciences of the United States of America* 120(25): e2300673120. doi:10.1073/pnas.2300673120.

- Obregon-Tito, A. J., R. Y. Tito, J. Metcalf, K. Sankaranarayanan, J. C. Clemente, L. K. Ursell, Z. Zech Xu, et al. 2015. « Subsistence strategies in traditional societies distinguish gut microbiomes ». *Nat Commun* 6: 6505. doi:10.1038/ncomms7505.
- Ohta, T., et M. Kimura. 1969. « Linkage Disequilibrium at Steady State Determined by Random Genetic Drift and Recurrent Mutation ». *Genetics* 63(1): 229-38. doi:10.1093/genetics/63.1.229.
- Olalde, Iñigo, Swapan Mallick, Nick Patterson, Nadin Rohland, Vanessa Villalba-Mouco, Marina Silva, Katharina Dulias, et al. 2019. « The genomic history of the Iberian Peninsula over the past 8000 years ». *Science* 363(6432): 1230-34. doi:10.1126/science.aav4040.
- Outram, A. K., N. A. Stear, R. Bendrey, S. Olsen, A. Kasparov, V. Zaibert, N. Thorpe, et R. P. Evershed. 2009. « The earliest horse harnessing and milking ». *Science* 323(5919): 1332-35. doi:10.1126/science.1168594.
- Palstra, Friso P., Evelyne Heyer, et Frédéric Austerlitz. 2015. « Statistical Inference on Genetic Data Reveals the Complex Demographic History of Human Populations in Central Asia ». *Molecular Biology and Evolution* 32(6): 1411-24. doi:10.1093/molbev/msv030.
- Parfrey, L. W., W. A. Walters, C. L. Lauber, J. C. Clemente, D. Berg-Lyons, C. Teiling, C. Kodira, et al. 2014. « Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity ». *Front Microbiol* 5: 298. doi:10.3389/fmicb.2014.00298.
- Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. « Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle ». *Cell* 176(3): 649-662.e20. doi:10.1016/j.cell.2019.01.001.
- Pereira, Eric, Stela Felipe, Raquel de Freitas, Valdevane Araújo, Paula Soares, Jannison Ribeiro, Luiz Henrique Dos Santos, et al. 2022. « ABO Blood Group and Link to COVID-19: A Comprehensive Review of the Reported Associations and Their Possible Underlying Mechanisms ». *Microbial Pathogenesis* 169: 105658. doi:10.1016/j.micpath.2022.105658.
- Pessia, Eugénie, Alexandra Popa, Sylvain Mousset, Clément Rezvoy, Laurent Duret, et Gabriel A. B. Marais. 2012. « Evidence for Widespread GC-Biased Gene Conversion in Eukaryotes ». *Genome Biology and Evolution* 4(7): 675-82. doi:10.1093/gbe/evs052.
- Pickrell, Joseph K., Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z. Li, Devin Absher, Balaji S. Srinivasan, et al. 2009. « Signals of Recent Positive Selection in a Worldwide Sample of Human Populations ». *Genome Research* 19(5): 826-37. doi:10.1101/gr.087577.108.
- Prado-Martinez, Javier, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, et al. 2013. « Great Ape Genetic Diversity and Population History ». *Nature* 499(7459): 471-75. doi:10.1038/nature12228.
- Pritchard, J. K., J. K. Pickrell, et G. Coop. 2010. « The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation ». *Curr Biol* 20(4): R208-15. doi:10.1016/j.cub.2009.11.055.
- Pudlo, Nicholas A., Gabriel Vasconcelos Pereira, Jaagni Parnami, Melissa Cid, Stephanie Markert, Jeffrey P. Tingley, Frank Unfried, et al. 2022. « Diverse events have transferred genes for edible seaweed digestion from marine to human gut bacteria ». *Cell Host & Microbe* 30(3): 314-328.e11. doi:10.1016/j.chom.2022.02.001.
- Rees, Jasmin S., Sergi Castellano, et Aida M. Andrés. 2020. « The Genomics of Human Local Adaptation ». *Trends in genetics: TIG* 36(6): 415-28. doi:10.1016/j.tig.2020.03.006.
- Reich, D. 2019. « Downloadable genotypes of present-day and ancient DNA data (compiled from published papers) ». <https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers>.
- Rothschild, Daphna, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, et al. 2018. « Environment Dominates over Host Genetics in Shaping Human Gut Microbiota ». *Nature* 555(7695): 210-15. doi:10.1038/nature25973.
- Roux, Camille, Maxime Pauwels, Maria-Valeria Ruggiero, Deborah Charlesworth, Vincent Castric, et Xavier Vekemans. 2013. « Recent and Ancient Signature of Balancing Selection around the S-Locus in *Arabidopsis Halleri* and *A. Lyrata* ». *Molecular Biology and Evolution* 30(2): 435-47. doi:10.1093/molbev/mss246.
- Ruhlmann, S. 2019. *11 Inviting happiness: Food sharing in Mongolia*. Leiden/Boston: Brill.
- Saitou, N., et F. Yamamoto. 1997. « Evolution of Primate ABO Blood Group Genes and Their Homologous Genes ». *Mol Biol Evol* 14(4): 399-411.
- Salque, Mélanie, Peter I. Bogucki, Joanna Pyzel, Iwona Sobkowiak-Tabaka, Ryszard Grygiel, Marzena Szmyt, et Richard P. Evershed. 2013. « Earliest evidence for cheese making in the sixth millennium BC in northern Europe. » *Nature* 493: 522-25. doi:10.1038/nature11698.
- Schnorr, S. L., M. Candela, S. Rampelli, M. Centanni, C. Consolandi, G. Basaglia, S. Turrone, et al. 2014. « Gut microbiome of the Hadza hunter-gatherers ». *Nat Commun* 5: 3654. doi:10.1038/ncomms4654.
- Segata, N. 2015. « Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity ». *Curr Biol* 25(14): R611-3. doi:10.1016/j.cub.2015.05.040.

- Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, Moyses J, et al. 2012. « The ABO Blood Group Is a Trans-Species Polymorphism in Primates ». *Proc Natl Acad Sci U S A* 109(45): 18493-98. doi:10.1073/pnas.1210603109.
- Ségurel L, Gao Z, et Przeworski M. 2013. « Ancestry Runs Deeper than Blood: The Evolutionary History of ABO Points to Cryptic Variation of Functional Importance ». *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35(10): 862-67. doi:10.1002/bies.201300030.
- Ségurel L, Austerlitz F, Toupance B, Gautier M, Kelley JL, Pasquet P, Lonjou C, et al. 2013. « Positive Selection of Protective Variants for Type 2 Diabetes from the Neolithic Onward: A Case Study in Central Asia ». *European journal of human genetics: EJHG* 21(10): 1146-51. doi:10.1038/ejhg.2012.295.
- Ségurel L, et Bon C. 2017. « On the Evolution of Lactase Persistence in Humans ». *Annu Rev Genomics Hum Genet* 18: 297-319. doi:10.1146/annurev-genom-091416-035340.
- Ségurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, Fabre A, Hegay T, et Heyer E. 2020. « Why and When Was Lactase Persistence Selected for? Insights from Central Asian Herders and Ancient DNA ». *PLOS Biology* 18(6): e3000742. doi:10.1371/journal.pbio.3000742.
- Sellen, Daniel W. 2001. « Comparison of Infant Feeding Patterns Reported for Nonindustrial Populations with Current Recommendations ». *The Journal of Nutrition* 131(10): 2707-15. doi:10.1093/jn/131.10.2707.
- Shaukat, A., M. D. Levitt, B. C. Taylor, R. MacDonald, T. A. Shamliyan, R. L. Kane, et T. J. Wilt. 2010. « Systematic review: effective management strategies for lactose intolerance ». *Ann Intern Med* 152(12): 797-803. doi:10.7326/0003-4819-152-12-201006150-00241.
- Siewert, Katherine M., et Benjamin F. Voight. 2017. « Detecting Long-Term Balancing Selection Using Allele Frequency Correlation ». *Molecular Biology and Evolution* 34(11): 2996-3005. doi:10.1093/molbev/msx209.
- Siewert, Katherine M., et Benjamin F. Voight. 2020. « BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data ». *Genome Biology and Evolution* 12(2): 3873-77. doi:10.1093/gbe/evaa013.
- Smillie, Chris S., Mark B. Smith, Jonathan Friedman, Otto X. Cordero, Lawrence A. David, et Eric J. Alm. 2011. « Ecology Drives a Global Network of Gene Exchange Connecting the Human Microbiome ». *Nature* 480(7376): 241-44. doi:10.1038/nature10571.
- Socha, W. W., et J. Moor-Jankowski. 1979. « Blood groups of anthropoid apes and their relationship to human blood groups ». *J Hum Evol* 8(4): 453-65.
- Soni, Vivak, Michiel Vos, et Adam Eyre-Walker. 2022. « A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection ». *PLoS Biology* 20(6): e3001645. doi:10.1371/journal.pbio.3001645.
- Souilmi, Yassine, M. Elise Lauterbur, Ray Tobler, Christian D. Huber, Angad S. Johar, Shayli Varasteh Moradi, Wayne A. Johnston, et al. 2021. « An Ancient Viral Epidemic Involving Host Coronavirus Interacting Genes More than 20,000 Years Ago in East Asia ». *Current Biology* 31(16): 3504-3514.e9. doi:10.1016/j.cub.2021.05.067.
- Souilmi, Yassine, Raymond Tobler, Angad Johar, Matthew Williams, Shane T. Grey, Joshua Schmidt, João C. Teixeira, et al. 2022. « Admixture Has Obscured Signals of Historical Hard Sweeps in Humans ». *Nature Ecology & Evolution* 6(12): 2003-15. doi:10.1038/s41559-022-01914-9.
- Spurgin, L. G., et D. S. Richardson. 2010. « How Pathogens Drive Genetic Diversity: MHC, Mechanisms and Misunderstandings ». *Proc Biol Sci* 277(1684): 979-88. doi:10.1098/rspb.2009.2084.
- Stock, Jay T, et Jonathan C K Wells. 2023. « Dairying and the evolution and consequences of lactase persistence in humans ». *Animal Frontiers* 13(3): 7-13. doi:10.1093/af/vfad022.
- Suzuki, Taichi A., J. Liam Fitzstevens, Victor T. Schmidt, Hagay Enav, Kelsey E. Huus, Mirabeau Mbong Ngwese, Anne Grieffhammer, et al. 2022. « Codiversification of Gut Microbiota with Humans ». *Science (New York, N.Y.)* 377(6612): 1328-32. doi:10.1126/science.abm7759.
- Szpiech, Zachary A., et Ryan D. Hernandez. 2014. « Selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection ». *Molecular Biology and Evolution* 31(10): 2824-27. doi:10.1093/molbev/msu211.
- Teixeira, João C., Cesare de Filippo, Antje Weilmann, Juan R. Meneu, Fernando Racimo, Michael Dannemann, Birgit Nickel, et al. 2015. « Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos ». *Molecular Biology and Evolution* 32(5): 1186-96. doi:10.1093/molbev/msv007.
- Thompson, E. E., H. Kuttub-Boulos, D. Witonsky, L. Yang, B. A. Roe, et A. Di Rienzo. 2004. « CYP3A Variation and the Evolution of Salt-Sensitivity Variants ». *American Journal of Human Genetics* 75(6): 1059-69. doi:10.1086/426406.
- Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, et al. 2007. « Convergent adaptation of human lactase persistence in Africa and Europe ». *Nat Genet* 39(1): 31-40. doi:10.1038/ng1946.

- Tito, Raul Y., Samuel Chaffron, Clara Caenepeel, Gipsi Lima-Mendez, Jun Wang, Sara Vieira-Silva, Gwen Falony, et al. 2019. « Population-Level Analysis of Blastocystis Subtype Prevalence and Variation in the Human Gut Microbiota ». *Gut* 68(7): 1180-89. doi:10.1136/gutjnl-2018-316106.
- Truong, D. T., A. Tett, E. Pasolli, C. Huttenhower, et N. Segata. 2017. « Microbial strain-level population structure and genetic diversity from metagenomes ». *Genome Res* 27(4): 626-38. doi:10.1101/gr.216242.116.
- Valles-Colomer, Mireia, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Leonard Dubois, Davide Golzato, Federica Armanini, et al. 2023. « The Person-to-Person Transmission Landscape of the Gut and Oral Microbiomes ». *Nature* 614(7946): 125-35. doi:10.1038/s41586-022-05620-1.
- Vander Molen, J., L. M. Frisse, S. M. Fullerton, Y. Qian, L. Del Bosque-Plata, R. R. Hudson, et A. Di Rienzo. 2005. « Population Genetics of CAPN10 and GPR35: Implications for the Evolution of Type 2 Diabetes Variants ». *American Journal of Human Genetics* 76(4): 548-60. doi:10.1086/428784.
- Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Thery, A. Froment, et al. 2009. « Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa ». *Curr Biol* 19(4): 312-18. doi:10.1016/j.cub.2008.12.049.
- Voight, Benjamin F., Sridhar Kudaravalli, Xiaquan Wen, et Jonathan K. Pritchard. 2006. « A Map of Recent Positive Selection in the Human Genome ». *PLoS biology* 4(3): e72. doi:10.1371/journal.pbio.0040072.
- Vos, G. H. 1965. « The Frequency of Abo-Incompatible Combinations in Relation to Maternal Rhesus Antibody Values in Rh Immunized Women ». *Am J Hum Genet* 17: 202-11.
- Weigand, Hannah, et Florian Leese. 2018. « Detecting Signatures of Positive Selection in Non-Model Species Using Genomic Data ». *Zoological Journal of the Linnean Society* 184(2): 528-83. doi:10.1093/zoolinnean/zly007.
- Wellenreuther, Maren, et Louis Bernatchez. 2018. « Eco-Evolutionary Genomics of Chromosomal Inversions ». *Trends in Ecology & Evolution* 33(6): 427-40. doi:10.1016/j.tree.2018.04.002.
- Wellenreuther, Maren, Claire Mérot, Emma Berdan, et Louis Bernatchez. 2019. « Going beyond SNPs: The Role of Structural Genomic Variants in Adaptive Evolution and Species Diversification ». *Molecular Ecology* 28(6): 1203-9. doi:10.1111/mec.15066.
- Wells, Jonathan C. K., Emma Pomeroy, et Jay T. Stock. 2021. « Evolution of Lactase Persistence: Turbo-Charging Adaptation in Growth Under the Selective Pressure of Maternal Mortality? ». *Frontiers in Physiology* 12: 696516. doi:10.3389/fphys.2021.696516.
- White, Joël, Katherine R. Amato, Ellen Decaestecker, et Valerie J. McKenzie. 2023. « Editorial: Impact of Anthropogenic Environmental Changes on Animal Microbiomes ». *Frontiers in Ecology and Evolution* 11. doi:10.3389/fevo.2023.1204035.
- Wohns, Anthony Wilder, Yan Wong, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, et al. 2022. « A Unified Genealogy of Modern and Ancient Genomes ». *Science (New York, N.Y.)* 375(6583): eabi8264. doi:10.1126/science.abi8264.
- Yamamoto, Emili Cid, Miyako Yamamoto, et Antoine Blancher. 2012. « ABO Research in the Modern Era of Genomics ». *Transfusion Medicine Reviews* 26(2): 103-18. doi:10.1016/j.tmr.2011.08.002.
- Yamamoto, Emili Cid, Miyako Yamamoto, Naruya Saitou, Jaume Bertranpetit, et Antoine Blancher. 2014. « An Integrative Evolution Theory of Histo-Blood Group ABO and Related Genes ». *Scientific Reports* 4: 6601. doi:10.1038/srep06601.
- Yamamoto, F., P. D. McNeill, Y. Kominato, M. Yamamoto, S. Hakomori, S. Ishimoto, S. Nishida, M. Shima, et Y. Fujimura. 1993. « Molecular Genetic Analysis of the ABO Blood Group System: 2. Cis-AB Alleles ». *Vox Sang* 64(2): 120-23.
- Yamamoto, X. H. Lin, Y. Kominato, Y. Hata, R. Noda, N. Saitou, et F. Yamamoto. 2001. « Murine Equivalent of the Human Histo-Blood Group ABO Gene Is a Cis-AB Gene and Encodes a Glycosyltransferase with Both A and B Transferase Activity ». *J Biol Chem* 276(17): 13701-8. doi:10.1074/jbc.M010805200.
- Yamamoto, et Yamamoto. 2001. « Molecular Genetic Basis of Porcine Histo-Blood Group AO System ». *Blood* 97(10): 3308-10. doi:10.1182/blood.v97.10.3308.
- Yatsunenkov, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, et al. 2012. « Human gut microbiome viewed across age and geography ». *Nature* 486(7402): 222-27. doi:10.1038/nature11053.
- Yongfa, W., Y. Yongshan, X. Jinjin, D. Ruofu, S.D. Flatz, W. Kühnau, et G. Flatz. 1984. « Prevalence of primary adult lactose malabsorption in three populations of northern China ». *Human Genetics* 67(1): 103-6.
- Youngblut, Nicholas D., Georg H. Reischer, William Walters, Nathalie Schuster, Chris Walzer, Gabrielle Stalder, Ruth E. Ley, et Andreas H. Farnleitner. 2019. « Host Diet and Evolutionary History Explain Different Aspects of Gut Microbiome Diversity among Vertebrate Clades ». *Nature Communications* 10(1): 2200. doi:10.1038/s41467-019-10191-3.