Origines et conséquences des variants régulateurs chez l'humain

Martin Silvert

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

Spécialité : Génétique des populations humaines

École doctorale n°515 : Complexité du Vivant

présentée par

Martin SILVERT

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

Sujet de la thèse :

Origines et conséquences des variants régulateurs chez l'humain

soutenue le 27 septembre 2019

devant le jury composé de :

Mme	Christine DILLMANN	Rapporteure
M.	Ludovic ORLANDO	Rapporteur
M.	Guillaume ACHAZ	Représentant de l'Université
Mme	Camille BERTHELOT	Examinatrice
M.	David-Alexandre TRÉGOUËT	Examinateur
M.	Lluis QUINTANA-MURCI	Directeur de thèse
M.	Maxime ROTIVAL	Membre invité

Remerciements

En tout premier lieu, je tiens à remercier les membres du jury : Mme. Christine Dillmann, M. Ludovic Orlando, M. Guillaume Achaz, Mme Camille Berthelot et M. David-Alexandre Trégouët qui me font l'honneur et le plaisir d'évaluer ce travail de thèse et de participer à cette soutenance.

Je tiens aussi à remercier Lluis Quintana-Murci, mon directeur de thèse. Merci de m'avoir accueilli au sein de ton laboratoire et de m'avoir encadré. Ça a été un privilège pour moi de réaliser ma thèse dans ce laboratoire, et j'en ressors grandi d'un point de vue scientifique et personnel. Merci notamment d'avoir activement cherché des solutions à nos conflits, et d'avoir été ouvert aux solutions que j'ai moi même proposées. Tu as tout fait pour me donner un environnement ou je pouvais produire la meilleure science possible, et pour cela je te remercie et te souhaite le meilleur pour la suite.

Mes remerciements les plus sincères à Maxime Rotival, qui a été mon encadrant direct pendant ces années au laboratoire. Ce n'est pas une exagération de dire que ce manuscrit n'aurais jamais été écrit sans ton aide. Merci de tes nombreux conseils professionnels, mais également ton soutien personnel. Merci de tout ce que tu m'as appris, je ressors toujours de nos échanges grandi.

Merci également à tous les membres, de l'unité de Génétique Évolutive Humaine, que j'ai côtoyé durant cette thèse. Merci de m'avoir supporté. Merci pour ces moments de partages, les joies comme les moments difficiles. Merci pour les bières, les blagues et les bizarreries. Merci d'avoir été là.

Merci à mes amis, surtout Renaud, Alexandre, Pierre et Cécile. Je manque de mots pour vous exprimer à quel point vous m'avez aidé, et je ne peux qu'espérer vous renvoyer l'ascenseur.

Et à ce paragraphe je sais que mes mots ne peuvent pas refléter tout ce que je ressens, donc beaucoup trop brièvement : merci Maman.

Table des figures

Liste des tableaux

Liste des abréviations

BDM	Bateson–Dobzhansky–Muller
CAGE	Cap Analysis of Gene Expression
CDC	Center for Diseases Control
CpG	Dinucléotide Cytosine Guanine
CSdCA	Cellules Souches dérivées de Cellules Adipeuses
HAM	Humain Anatomiquement Moderne
MAF	Minor Allele Frequency
miARN	Micro ARN
MMC	Modèle de Markov Caché
SNP	Single Nucleotide Polymorphism

Table des matières

Introduction

L'expression des gènes est un phénotype fascinant. Elle est, non seulement, une des briques sur laquelle reposent de nombreux autres phénotypes, et donc responsable d'une grande partie de la diversité phénotypique du vivant, mais elle est également très variable en fonction du tissu, de l'environnement ou des individus. La plasticité, cependant contrôlée, de ce phénotype est permise par la combinaison de nombreuses régions régulatrices dont l'activation, variable selon les tissus et les stimuli auxquels sont soumises les cellules, permet une régulation fine de l'expression des gènes en fonction du contexte. De plus, la redondance des ces éléments régulateurs leur permet d'absorber l'effet des variations génétiques et d'accumuler une grande diversité génétique au sein d'une même population.

L'expression des gènes est également un phénotype qui peut servir de levier de sélection, notamment chez l'humain, où les variants sous sélection affectent plus l'expression qu'attendu. L'exemple le plus parlant est probablement celui de la persistance de la lactase, où des mutations permettant l'expression à l'âge adulte du gène de la lactase, responsable de la digestion du lactose, ont été sélectionnées indépendamment dans plusieurs populations humaines, en même temps que se développait la pratique de l'élevage.

Ainsi le sujet de l'évolution des variants régulateurs est à la fois passionnant et fortement complexe. Il serait impossible pour moi de le traiter dans sa totalité au cours de cette thèse, considérez plutôt ceci comme le fil directeur de ce manuscrit, où ce thème sera abordé sous deux angles différents, (i) en se concentrant sur un évènement démographique particulier, l'introgression néandertalienne dans les populations eurasiennes, et son impact sur la régulation génique (ii) en étudiant la contributions des micro ARN à la régulation de l'expression génique dans un contexte immunitaire, et la variabilité de cette régulation au sein des populations humaines.

1. Sources de la diversité génétique humaine

Les caractéristiques physiques présentes chez les humains sont extrêmement variées. Un regard rapide dans une foule nous permet d'observer une grande variété de phénotypes, taille, couleurs et formes, et il ne s'agit que de ceux visibles à l'œil nu. Parmi la grande diversité de phénotypes observables chez l'humain, la majorité est influencée non seulement par notre environnement, mais également par la génétique. Et si, à l'exception des jumeaux homozygotes, l'ADN de chaque être humain est unique, la variabilité génétique au sein de l'humanité reste faible. Deux humains pris au hasard dans la population ont ainsi un ADN identique à 99.9% (Lander et al., 2001 ; Venter et al., 2001). Le petit pourcentage restant permet malgré tout d'expliquer une partie des variabilités des phénotypes présents dans l'espèce humaine (Lakhani et al., 2019 ; Polderman et al., 2015).

Lors de cette thèse, je me suis intéressé à la variabilité génétique présente dans les régions qui régulent l'expression des gènes (régions régulatrices), les conséquences phénotypiques des variants génétiques qui y sont présents, mais aussi les dynamiques sous-tendant leur propagation dans les populations.

1.1 Forces génomiques

Les forces génomiques, notamment la mutation et la recombinaison, sont les forces qui créent la diversité génétique. Les différents types de mutations sont très divers, couvrant à la fois les remplacements ponctuels de nucléotides, ou *Single Nucleotide Polymorphism* (SNP), et des changements plus généraux incluant des insertions ou délétions d'un ou plusieurs nucléotides, ou encore des duplications de séquences existantes pouvant mener à une variation du nombre de copies de certaines régions géniques. Cependant, les SNP sont de loin le type de mutation le plus fréquent, ainsi que le plus facile à détecter à l'échelle génomique. Dans le cadre de cette thèse, nous nous focaliserons donc sur ceux-ci.

Certains types de SNP sont plus probables que d'autres, ainsi les transitions, c'est à dire les mutations d'une purine à l'autre (A $\leftrightarrow$ G) ou d'une pyrimidine à l'autre (T $\leftrightarrow$ C), sont plus fréquentes que les transversions, qui consistent en une mutation d'une purine par une pyrimidine, et vice et versa. Les SNP peuvent apparaître suite à une erreur lors de la réplication ou de la réparation de l'ADN suite à la dégradation par un facteur extérieur, pouvant aller de facteurs ionisants à certains produits chimiques mutagènes. Il est également

notable que la probabilité de mutation d'un nucléotide peut également augmenter par rapport à son environnement direct. Ainsi les dinucléotides cytosine guanine, souvent notés CpG, ont un taux de mutation vers les dinucléotides TpG et CpA très élevé. Ceci est dû au fait que les cytosines des sites CpG peuvent être méthylés, et sont alors plus susceptibles d'être remplacées par des thymines, créant des sites TpG si la mutation se trouve sur le brin direct, ou CpA sur le brin indirect. Ainsi le taux de mutations est très variable le long du génome, et si la moyenne est estimée à 10^{-8} par base et par génération, celui ci peut être 10 fois plus grand dans les régions riches en sites CpG (Campbell et al., 2012). Si une mutation est présente dans les lignées germinales, cette dernière peut être transmise à la descendance, avec les autres mutations présentes sur le même chromosome.

Le contexte génomique dans lequel une mutation apparaît est également important. En effet, puisqu'un parent transmet un de ses deux chromosomes à sa descendance, le contexte génétique dans lequel la mutation est apparue est transmis avec elle. Ainsi, en l'absence de recombinaison, chaque nouvel allèle est transmis avec l'intégralité des allèles présents sur le chromosome dans lequel il est apparu. Ces allèles forment alors un "bloc" génétique appelé haplotype. La recombinaison des chromosomes homologues pendant la méiose va alors casser ces blocs génétiques, créant ainsi de nouveaux haplotypes. La recombinaison va donc permettre un forme de mélange des chromosomes parentaux et créer de nouvelles combinaisons d'allèles dans la population, tout en réduisant la taille des haplotypes qui ségrègent dans la population. Le taux de recombinaison est variable tout au long du génome humain, et peut même varier d'une population à l'autre. Parmi d'autres facteurs, on sait que ce dernier est variable notamment en fonction du taux de CG dans la région ainsi qu'à la présence de gènes (Altshuler, Donnelly, & The International HapMap, 2005 ; The International HapMap et al., 2007, 2010).

1.2 Dérive génétique

Sous neutralité, c'est à dire sous l'hypothèse qu'aucun variant génétique n'impacte la reproduction des individus, la fréquence d'un variant dans la population suit une progression stochastique appelée dérive génétique (Wright, 1931). En effet, tous les individus d'une population n'ont pas le même nombre de descendants, certains décédant sans aucune descendance et d'autres en ayant un grand nombre. De plus, un variant présent de manière hétérozygote chez un individu peut ne pas être transmis à un descendant en particulier ($\sim$50% de chances) en raison de la diploïdie humaine.

Suivant ce modèle, un variant peut soit disparaître au fur et à mesure des générations, soit augmenter en fréquence jusqu'à fixation, c'est à dire que tous les chromosomes présents dans cette population soient porteurs de ce variant. Alternativement, celui-ci peut rester à des fréquences intermédiaires pendant une longue durée, cependant, les états de disparition et de fixation étant des états absorbants (i.e. définitifs une fois atteint) la probabilité qu'un variant reste à une fréquence intermédiaire diminue avec le nombre de génération observé (Fig 1.1A).

L'aléa de ce phénomène est très dépendant de la taille effective de population : N_e définie comme le nombre d'individu pouvant participer à la reproduction de la population, en effet la probabilité d'une mutation présente sur un seul chromosome (après apparition par exemple) de ne pas être transmise à la génération suivante est de $1/2N_e$, l'aléa et la diversité génétique diminuent donc avec la taille de population.

1.3 Forces démographiques

Le modèle précédemment décrit concerne une population sans limites de reproduction entre individus. Or dans la réalité, les populations humaines peuvent ne pas avoir de contact entres elles pendant plusieurs générations, suite à une séparation géographique par exemple. Dans ce cas, chaque population va avoir des dynamiques propres, avec certains variants atteignant de hautes fréquences dans une population, en étant absents dans d'autres.

Une limitation à la différentiation génétique des populations peut provenir des migrations entre les deux populations. Si les deux populations échangent régulièrement des individus, une mutation présente dans la population x peut alors être transmise à la population y, et symétriquement. Les migrations dans ce cas n'affectent pas la fréquence globale des variants, mais peuvent faire varier leurs fréquences au sein de chaque population.

Cependant, une séparation de longue durée peut mener à l'apparition d'incompatibilités génétiques entre les populations, ce phénomène, appelé spéciation, est discuté plus en détail dans les sections 1.5 et 2.3.1 de ce manuscrit.

Il est intéressant de noter que dans le cas des populations humaines, il n'existe pas de variant complètement fixé dans une population donnée, et totalement absent des autres.

1.4 Sélection naturelle

1.4.1 Concept de sélection naturelle

L'hypothèse de neutralité des variants génétiques, c'est à dire qu'un variant génétique n'affecte pas le succès reproductif des individus, n'est que partiellement vérifiée. Et si la plupart des variants observés chez l'humain n'a pas d'effet mesurable sur les phénotypes macroscopiques, la présence de régions sous sélection a des effets importants sur la diversité génétique et la répartition des variants génétiques le long du génome. Il est important d'observer que l'effet d'un variant génétique sur le succès reproducteur d'un individu est relatif à son environnement, un changement dans ce dernier pouvant altérer, voire inverser, l'effet d'un variant sur la reproduction. L'environnement est ici défini de manière très large, comportant non seulement la zone géographique, notamment ses prédateurs, ses pathogènes et l'alimentation possible, mais également les effets sociaux.

1.4.2 Sélection positive

On appelle sélection positive les cas où un variant augmente le succès reproductif de ses porteurs. Au niveau de la population, la sélection positive augmente la vitesse de propaga-

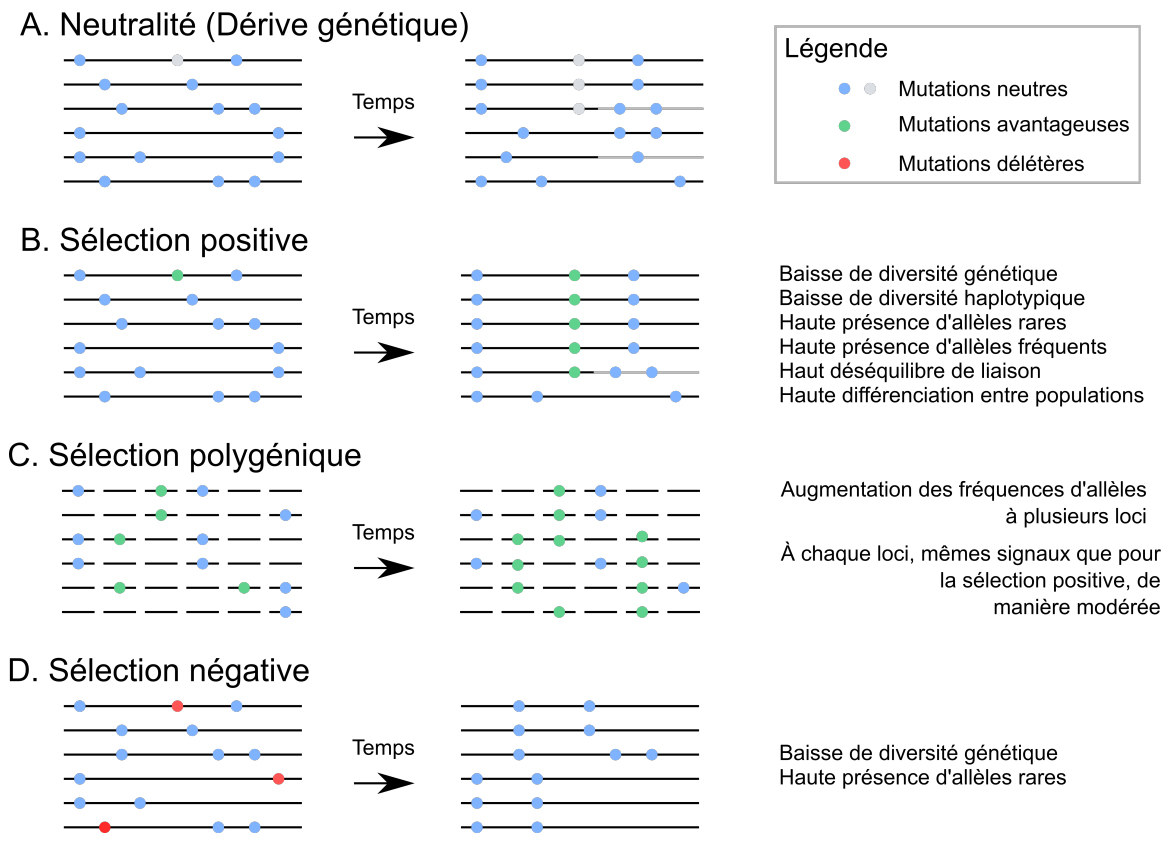


Fig. 1.1 ***Types de sélection : Représentation des effets de chaque type de sélection dans le temps***

tion du variant dans la population, jusqu'à une éventuelle fixation. La grande vitesse de la propagation d'un variant dans la population va laisser plusieurs traces dans les génomes, notamment une grande taille d'haplotype et une réduction locale de la diversité (Fig 1.1B). De plus, si le phénomène de sélection est propre à une population, on s'attend à observer une divergence accrue avec les autres populations (Nielsen, 2005 ; Lohmueller et al., 2011). Ce modèle de balayage sélectif correspond à l'idée d'un *hard sweep*. Cependant, la sélection positive à un locus ne rentre pas toujours dans le cadre de ce modèle. En effet, des cas plus complexes ,des *soft sweeps*, peuvent correspondre à des cas où plusieurs mutations d'un même locus sont sous sélection, chacune empêchant la montée en fréquence des autres. Une autre possibilité peut provenir de cas ou un variant auparavant neutre, voire légerement délétère, devient avantageux suite à un changement d'environnement. On parle alors de sélection sur un variant pré-existant. Ces cas de *soft sweep* sont plus difficiles à détecter, car, même s'ils présentent des signatures similaires aux *hard sweeps*, ces dernières sont beaucoup moins nettes (Pritchard, Pickrell, & Coop, 2010).

1.4.3 Sélection polygénique

Il faut cependant se souvenir qu'un grand nombre de phénotypes sont influencés par de nombreux variants à de nombreux loci. Ainsi, si un phénotype particulier devient avantageux

dans une population, chacun des variants influençant ce phénotype va alors être faiblement favorisé, on parle alors de sélection polygénique sur variants pré-existants, ou plus généralement de sélection polygénique. L'effet cumulé des légères augmentations en fréquence de chacun de ces variants pourra alors mener à des changements importants du phénotype dans la population. Les traces de la sélection polygénique sont délicates à identifier, car il est nécessaire de connaître au préalable quels sont les variants qui influencent le phénotype étudié, par exemple suite à une *Genome Wide Association Study* (GWAS). Chacun de ces variants devrait, dans le cas d'une sélection polygénique, présenter des signatures similaires à la sélection positive, mais dans des amplitudes plus faibles (Fig 1.1C) (Berg & Coop, 2014 ; Pritchard et al., 2010).

1.4.4 Sélection négative

Le cas de la sélection négative est l'exact inverse de celui de la sélection positive. Un variant sous sélection négative tend à réduire le succès reproductif de son porteur. Il va ainsi baisser en fréquence jusqu'à disparition, bien que dans le cas des variants faiblement délétères, ces derniers puissent persister à basse fréquence dans la population pendant de nombreuses générations (Eyre-Walker & Keightley, 1999 ; Kryukov, Pennacchio, & Sunyaev, 2007). Certaines régions ont tendance à avoir beaucoup de variants sous sélection négative, notamment les régions codantes, on y retrouve donc une diversité génétique diminuée, et un excès d'allèles rares (Fig 1.1D).

1.4.5 Sélection de fond

L'existence de régions évolutivement contraintes, c'est à dire où les nouveaux variants ont de fortes chances d'être délétères, et donc sous sélection négative, façonne également la variabilité des régions voisines. En effet, un haplotype qui acquière un variant délétère va avoir tendance à disparaître, la proximité d'une région évolutivement contrainte va donc avoir l'effet de réduire la diversité haplotypique locale, et d'augmenter la dérive génétique locale entraînant donc une perte de diversité ainsi qu'un excès de variants à des fréquences intermédiaires. On parle dans ce cas de sélection de fond ou *background selection*. Ces effets diminuent lorsque la distance génétique à une région sous contrainte augmente. La sélection de fond peut aussi mener à une forte divergence entre espèces dans ces régions (McVicker, Gordon, Davis, & Green, 2009).

1.5 La spéciation

Un cas particulier qui ne rentre pas dans les définitions de sélections décrites ici est le cas de la spéciation. En effet, comme mentionné plus tôt, si deux groupes d'une même espèce se reproduisant de manière sexuée sont génétiquement séparés, c'est à dire qu'ils n'échangent pas ou très peu de matériel génétique entre eux, pendant de nombreuses générations, il arrive un moment où les deux groupes deviennent incapables de produire une descendance fertile

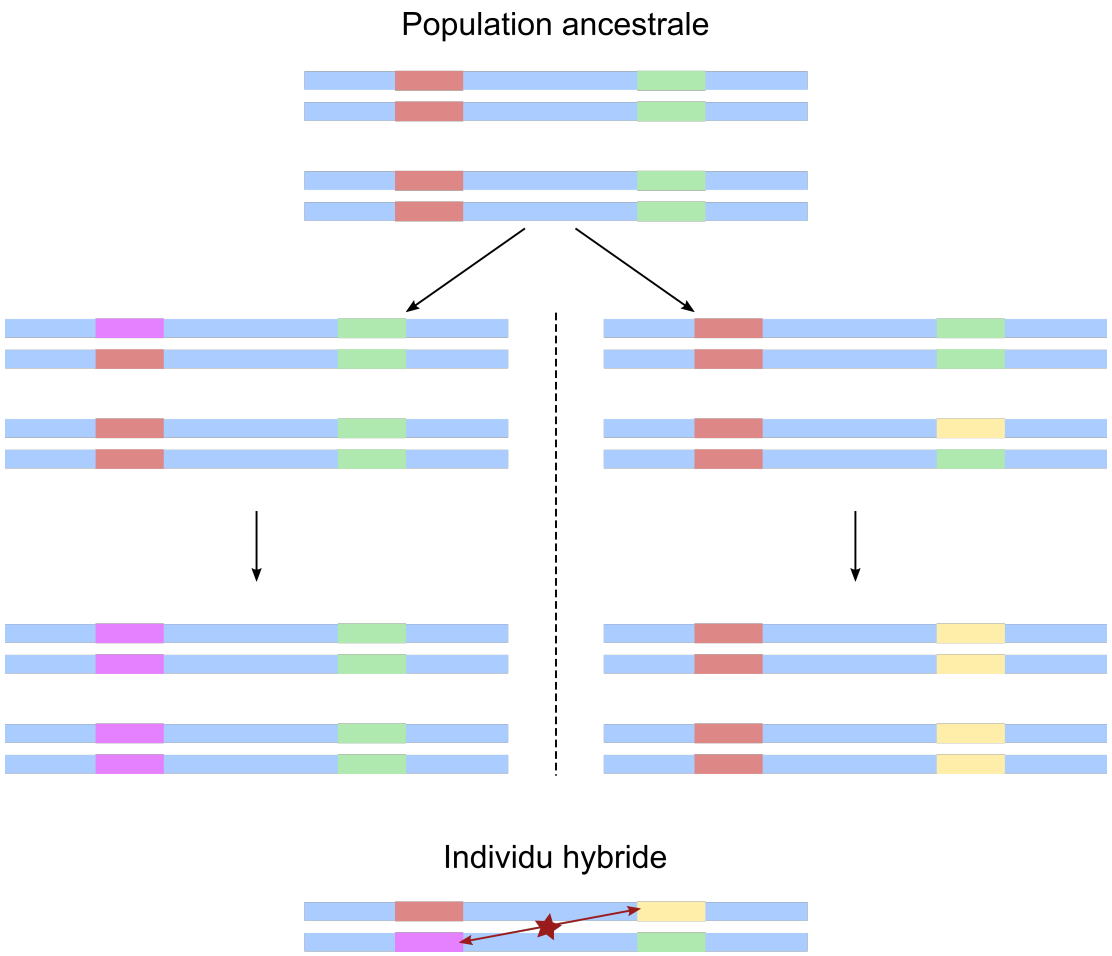


Fig. 1.2 ***Incompatibilités de Dobzhansky–Muller : Représentation des incompatibilités de Dobzhansky-Muller. Deux groupes provenant d'une même population voient différents variants apparaître, puis se fixer dans la population (de manière neutre ou sous sélection). Cependant, ces variants peuvent interagir et être délétères à un hybride ou sa descendance. Figure adaptée de Mack et al. (Mack & Nachman, 2017)***

entre eux. Les deux groupes sont alors considérés comme deux espèces à part entière, c'est le phénomène de spéciation.

Il est important de comprendre que ce procédé est graduel, et que les hybrides vont devenir de moins en moins viables au fur et à mesure que les deux groupes divergent en deux espèces différentes. Cependant, le faible succès reproductif des hybrides dans le cadre de la spéciation n'est pas strictement applicable au cas de la sélection négative décrit plus tôt. En effet, même si les deux parents sont parfaitement viables au sein de leur groupe respectif, l'interaction de leurs génomes pose des problèmes à leur descendance. C'est un cas d'épistasie, c'est à dire ou l'effet d'un variant est dépendant de la présence d'un second variant. Dans le cas particulier d'une épistasie menant à un hybride moins performant que les populations parentes, on parle du modèle Bateson–Dobzhansky–Muller (BDM) (Bateson, 1909 ; Dobzhansky, 1982 ; Muller, 1942).

2. Expression génique

2.1 L'expression génique comme phénotype

Une question essentielle en génétique est le lien entre diversité génétique et variation phénotypique. Ainsi de nombreuses études ont analysé la fréquence des variants génétiques afin d'identifier les variants dont la fréquence change en fonction du phénotype des individus, et donc jouant potentiellement un rôle causal dans ce phénotype. Notons notamment l'existence aujourd'hui du *GWAS Catalog* qui rassemble les résultats d'un grand nombre de ces *Genome Wide Association Studies* et est régulièrement mis à jour (Buniello et al., 2018)

Ces études ont permis d'étudier la localisation des variants qui influencent le plus les phénotypes humains. De manière intéressante, très peu d'entre elles (9%) perturbent la séquence codante d'un gène (Hindorff et al., 2009). Au contraire, la majorité des loci identifiés est dans des régions régulant la transcription (Maurano et al., 2012 ; Nicolae et al., 2010). De plus, ces mutations peuvent être à l'origine de changements phénotypiques majeurs, telle que la persistance de la lactase dans certaines populations humaines. Ces observations suggèrent non seulement que la majorité des variabilités de phénotypes prennent place au niveau du contrôle de l'expression, mais que ces derniers peuvent aboutir à un changement de phénotype majeur.

La plasticité de l'expression génique est également remarquable, cette dernière étant variable non seulement du tissu étudié (GTEx Consortium, 2013 ; GTEx Consortium et al., 2017), mais également en réponse à des stimulis externes (Quach et al., 2016 ; Piasecka et al., 2018). Ainsi l'étude de la régulation génique est une étape nécessaire afin de comprendre à la fois comment celle-ci peut affecter les phénotypes, mais également comment une certaine variabilité dans les profils d'expression est contrôlée entre tissus, et tolérée au sein d'une même espèce.

2.2 Régulation de l'expression génique

2.2.1 Régulation de la transcription

Un moyen d'étudier l'expression génique est de s'intéresser directement aux différents mécanismes de régulation de la transcription, c'est à dire la synthèse même de l'ARN.

Les promoters et les enhancers sont deux types de régions régulatrices de la transcription des gènes. Au niveau du site d'initiation de la transcription, le promoter recrute l'ARN polymérase, ainsi que d'autres facteurs de transcription, afin d'initier la transcription du

gène. Les enhancers sont quant à eux éloignés du site d'initiation de la transcription, ils peuvent en être situés jusqu'à 1 Mb, et lorsqu'ils sont activés par des facteurs de transcription spécifiques au tissu, sont mis au contact du promoter correspondant par un repliement de l'ADN afin de moduler son activité.

En pratique, les promoters et les enhancers sont détectés par différentes méthodes expérimentales, qui peuvent malgré tout se contredire (Benton, Talipineni, Kostka, & Capra, 2018)

CAGE

Le *Cap Analysis of Gene Expression* est une méthode de séquençage qui cible l'extrémité 5' des fragments d'ARN. Cette méthode a été notamment utilisée lors du projet FANTOM5 pour créer un atlas des promoters des mammifères et enhancers chez les humains (Fantom Consortium et al., 2014 ; Andersson et al., 2014).

Les pics de signaux CAGE proches de sites de début de transcription connus sont alors considérés comme des promoters des gènes (Fantom Consortium et al., 2014). Les enhancers sont cartographiés grâce à une de leurs particularités : ceux-ci montrent une faible activité de transcription (eARN) qui s'effectue de manière bidirectionelle (Andersson et al., 2014).

Utilisation de marqueurs biologiques

Plusieurs approches utilisent différentes informations telles que les marques d'histones, la méthylation de l'ADN, l'accessibilité de l'ADN ou encore l'expression génique pour classer chaque région du génome en fonction de son activité régulatrice dans un ou plusieurs tissus. Par exemple, les régions régulatrices peuvent être caractérisées par une sensibilité de la région à la DNAse due à l'ouverture de la chromatine lors de la transcription, différentes marques d'histones sont également utilisées pour caractériser spécifiquement les promoters et les enhancers.

Une approche plus générale agrège de nombreuses marques biologiques et classe chaque région du génome en fonction des ses particularités épigénétiques. Chaque classe est par la suite associée à une fonction biologique connue auparavant. Cela permet notamment d'étendre les définitions habituelles d'enhancers et promoters (Roadmap Epigenomics Consortium et al., 2015 ; Backenroth et al., 2018). Ainsi dans le cas du modèle à 15 catégories utilisé dans le cadre du Roadmap Epigenomic Project, la catégorie correspondant à la présence de H3K4me1 et l'absence des autres marques étudiées correspond aux enhancers, tandis que les sites de début de transcription sont caractérisés par la présence unique de H3K4me3.

Intéraction ADN-ADN

Une manière de détecter les enhancers est d'identifier les régions du génome qui interagissent avec chaque site de transcription. La méthode de *Promoter Capture Hi-C*(PCHi-C) permet désormais de faire ce genre d'études à grande échelle. Malgré l'efficacité de ces dernières pour assigner les enhancers à un ou plusieurs gènes, les études de ce type chez l'humain

restent limitées à l'heure actuelle (Javierre et al., 2016 ; Pan et al., 2018 ; Law et al., 2019).

2.2.2 Régulation par les miARN

Au delà de l'effet de la transcription sur l'expression génique, deux autres aspects sont importants : la vitesse de dégradation de l'ARN affecte également l'expression génique, de plus, la traduction de ce dernier va affecter l'impact de l'expression sur le phénotype final. Ces deux aspects peuvent être affectés par de courtes séquences (∼22 nucléotides) nommées micro ARN (miARN).

a) Expression des miARN

Les miARN peuvent être localisés soit dans des régions intergéniques, soit être transcrits en même temps que les gènes codants. De plus un même transcrit primaire de micro ARN (pri-miARN) peut servir de précurseur à un (mono-cistronique) ou à plusieurs miARN (poly-cistronique). La transcription du pri-miARN peut être déclenchée par la transcription puis maturation d'un transcrit classique lorsque le pri-miARN est intronique, ou bien être déclenchée indépendamment par des promoters dédiés.

Le pri-miARN va ensuite être clivé pour former une épingle à cheveux de 70 à 80 nucléotides de long, nommée micro ARN précurseur (pre-miARN). Le précurseur est ensuite exporté hors du noyau cellulaire jusque dans le cytoplasme, ou la boucle de l'épingle sera clivée, aboutissant à un complexe de deux brins d'ARN, imparfaitement complémentaires.

Une variabilité supplémentaire des miARN provient également du fait de l'existence d'isoformes, appelés isomiR. En effet, les miARN sont sensibles à des transformations post-transcriptionnelles telles que des substitutions de nucléotides (Li et al., 2018 ; de Hoon et al., 2010), l'ajout d'adénine ou d'uracyle à l'extrémité 3' du transcrit (Jones et al., 2009 ; Katoh et al., 2009), ou le raccourcissement de l'extrémité 3' du transcrit (Lee, Park, Park, Kim, & Shin, 2019). Un déplacement de leur extrémité 5' a également été décrit (Tan et al., 2014) (Fig 2.1). Il est également intéressant de noter que la quantité relative de certains isomiR est dépendante de l'environnement, et peut être affectée, par exemple, par la présence de bactéries (Siddle et al., 2015).

b) Régulation des gènes par les miARN

Un des deux brins est ensuite chargé dans le complexe RISC (*RNA Induced Silencing Complex*). La reconnaissance de l'ARN ciblé se fait principalement par complémentarité de séquence avec les nucléotides 2 à 8 de la partie 5' du miARN (appelée *seed sequence*), mais le reste du miARN participe également à la fixation. La plupart des sites de fixation des miARN se trouve dans la partie 3' non codante des transcrits, mais il s'agit plus d'une tendance que d'une règle absolue. La fixation peut impacter l'expression de l'ARNm en clivant et dégradant ce dernier, ou en perturbant la queue polyadénlyée de ce dernier, entraînant là aussi une dégradation. Cependant, les miARN peuvent aussi affecter la traduction de l'ARNm sans en affecter l'expression (Bartel, 2009).

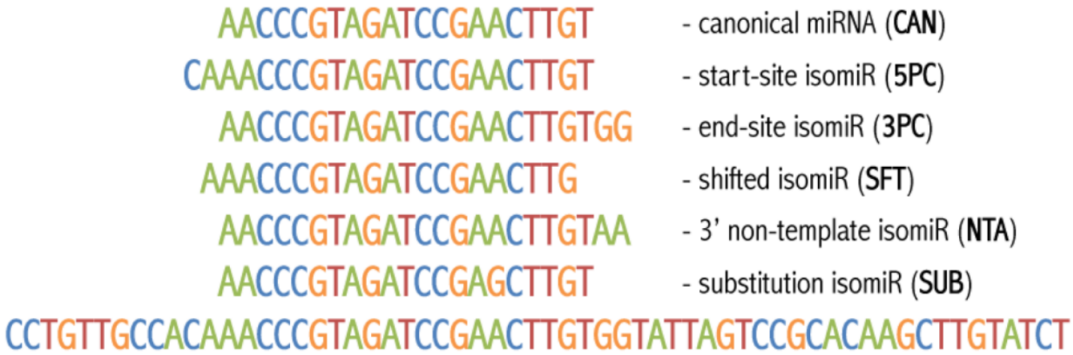


Fig. 2.1 ***Exemples d'isomiR : figure provenant de (Siddle et al., 2015)***

c) Détection de la régulation d'un miARN sur un gène

c).1 Corrélation des expressions Bien que les miARN participent à la dégradation de certains transcrits, la plupart du temps, une corrélation négative entre l'expression d'un gène régulé par un miARN, et l'expression du miARN en question est loin d'être systématique (Siddle et al., 2014 ; Lappalainen et al., 2013 ; Parts et al., 2012 ; Rantalainen et al., 2011), plusieurs effets de co-transcription rentrant également en jeu. De plus le contrôle des miARN sur l'expression des gènes est très subtile de manière individuelle, même si il peut être très remarqué en groupe. Une manière de se défaire du problème de co-transcription est de sur-exprimer, ou sous-exprimer le miARN étudié afin d'observer les changements d'expression des gènes résultant de cette variabilité.

c).2 Prédiction d'interaction Une autre manière d'identifier si un miARN régule la dégradation d'un gène est de comparer les séquences du miARN et du transcrit afin de prévoir si le miARN peut se fixer sur le transcrit et engendrer le clivage et la dégradation du transcrit génique. Plusieurs méthodes existent aujourd'hui, qui peuvent être basées sur la complémentarité des séquences au niveau de la *seed*, la conservation évolutive de la séquence entre espèces, ou encore une forte stabilité thermodynamique entre miARN et transcrit génique (Enright et al., 2003 ; Marco, 2018). Il est cependant notable que ces méthodes sont critiquées pour leurs propensions aux faux positifs, mais peuvent servir en tant que premier filtre (Pinzon et al., 2017).

2.2.3 Analyses d'association entre génétique et expression

Un moyen beaucoup plus direct d'étudier la régulation de l'expression génique est d'effectuer une étude *expression Quantitative Trait Loci* (eQTL), c'est-à-dire d'identifier quels variants sont associés à une expression plus élevée (ou plus basse) d'un certain gène (Quach et al., 2016 ; Mogil et al., 2018 ; Stranger et al., 2012 ; Battle et al., 2014 ; Kelly, Hansen, & Tishkoff, 2017). Cette approche est aveugle au mode de régulation génique perturbé par les variants en question, cependant il a été observé que la majorité des eQTL correspond

à des régions promotrices des gènes. Une limite importante de cette approche est la puissance nécessaire. Afin d'obtenir des associations fines ou concernant des mutations rares, il est nécessaire d'utiliser une cohorte de très grande ampleur, rendant ces études coûteuses. Cependant de nombreuses banques de données sur ce type d'association existent maintenant, notamment GTEx (GTEx Consortium, 2013 ; GTEx Consortium et al., 2017) comprenant de nombreux tissus, mais aussi eQTLGen (Võsa et al., 2018), une métanalyse des études eQTL dans le sang profitant donc d'une puissance inégalée jusqu'alors.

2.3 Expression génique et évolution

2.3.1 Rôle dans la spéciation

La spéciation, c'est à dire la différenciation de deux groupes d'une même espèce en deux espèces distinctes passe en grande partie par des divergences au niveau de la régulation de l'expression. Ainsi de nombreux cas d'hybrides stériles exhibent des expressions de certains gènes hors des normes attendues chez l'un ou l'autre groupe parental (Michalak & Noor, 2003 ; Haerty & Singh, 2006 ; Good, Giger, Dean, & Nachman, 2010).

Cette observation est cohérente avec les effets épistatiques nécessaires aux incompatibilités du modèle DBM (cf partie 1.5). En effet, les interactions entre plusieurs régions du génome nécessaires à la régulation de l'expression impliquent la potentialité d'une épistasie. Ainsi, l'accumulation progressive de modifications dans les éléments régulateurs peut aboutir à un état ou deux groupes parentaux présentent une expression des gènes similaires, mais avec des systèmes de régulation différents.

Dans le cas particulier des miARN, une divergence plus forte qu'attendue a été observée dans leurs sites de fixation entre certaines espèces de poissons, suggérant qu'ils ont joué un rôle (Loh, Yi, & Streelman, 2011). De plus, les enhancers chez les mamifères évoluent rapidement (Villar et al., 2015), suggérant que des derniers peuvent participer au processus de spéciation. Tout type d'interactions dans le cadre de la régulation génique peut, en théorie, être impliqué dans ce modèle, cependant, l'intitialisation de la transcription, principalement par les évolutions de facteurs de transcription et leurs sites de fixation ont été le plus étudiés. Je renvoie le lecteur au travail de Mack et collègues pour une revue plus détaillée (Mack & Nachman, 2017).

2.3.2 L'expression génique comme levier de sélection

L'expression génique est un levier de sélection majeur chez l'humain, en effet, les SNP sous sélection ont été montrés comme affectant plus l'expression que prévu, à la fois de manière générale (Kudaravalli, Veyrieras, Stranger, Dermitzakis, & Pritchard, 2009), mais également par rapport à des sélections sur des phénotypes précis telle que l'exposition aux rayons ultraviolets (Fraser, 2013) ou encore les réponses immunitaires (Quach et al., 2016). Tous ces effets polygéniques montrent à quel point, de manière générale, le contrôle de l'expression de gènes agissant sur un phénotype particulier peut être sélectionné. Cependant, il existe aussi des cas dus à un seul locus.

Mutation	Populations avec effet notable
-13 910 :T (associée à -22 018 :A)	Eurasie, Afrique du Nord, Afrique Centrale
-13 915 :G	Moyen-Orient
-13 907 :G	Afrique de l'Est (Éthiopie et Soudan)
-14 009 :G	Afrique de l'Est (Éthiopie et Soudan)
-14 010 :C	Afrique de l'Est (Kenya et Tanzanie) et Afrique du Sud

Table 2.1 – Mutations associées à la persistance de la lactase, table adaptée de de Ségurel et al. (Ségurel & Bon, 2017)

a) La persistance de la lactase

L'enzyme de la lactase est exprimée par la majorité des êtres humains lorsqu'ils sont très jeunes. Cependant, celle-ci n'est exprimée à l'âge adulte que dans un tiers de la population (Ingram, Mulcare, Itan, Thomas, & Swallow, 2009). Cette enzyme est exprimée dans l'intestin grêle et permet une bonne digestion du lactose, un sucre présent dans le lait. En Europe, ce phénotype est fortement associé au variant -13 910 :T, qui est localisé dans un enhancer intronique et permet une interaction enhancer-promoter même à l'âge adulte (Fang, Ahn, Wodziak, & Sibley, 2012 ; Lewinsky et al., 2005). De plus cette région de génome montre de fortes traces de sélection positive, principalement par la taille anormalement grande de l'haplotype portant ce variant, mais aussi car sa répartition dans les populations ne suit pas les attendus (Bersaglieri et al., 2004). Nous avons donc ici un cas (extrême) de sélection d'une mutation affectant la régulation d'un enhancer. Il est à noter que de nombreuses mutations ont été associées au phénotype de persistance de la lactase dans le monde (voir table 2.1). Et que plusieurs d'entre elles montrent des signaux de sélection, la persistance de la lactase est un exemple frappant de convergence évolutive (i.e. plusieurs mutations étant sélectionnées pour un même phénotype dans diverses populations de manière indépendante) (Tishkoff et al., 2007).

b) Le Locus *OAS1-3*

Le locus OAS a été identifié comme un cas particulier d'introgression archaïque adaptative, sujet qui sera développé plus en détail dans le prochain chapitre. C'est également un exemple de sélection positive d'un variant participant à la régulation post-transcriptionelle (Sams et al., 2016). L'haplotype archaïque introgressé à ce locus est associé à une expression différenciée de *OAS3*, mais également à un épissage alternatif de *OAS1* et *OAS2* dans les macrophages non-stimulés, ou infectés par *Salmonella typhimurium*. Cet exemple montre un cas de sélection d'un variant ayant un effet sur la régulation post-transcriptionelle.

3. L'introgression néandertalienne

3.1 Homo Neanderthalensis

3.1.1 Description physique

L'homme de Néandertal tire son nom de la vallée de Néander où le premier fossile attribué à ce groupe a été découvert (*Tal*, anciennement écrit *Thal*, signifiant vallée en allemand) (King, 1864). Par la suite de nombreux fossiles ont été ajoutés au groupe désormais nommé *homo néanderthalensis*, plus couramment désigné comme les néandertaliens.

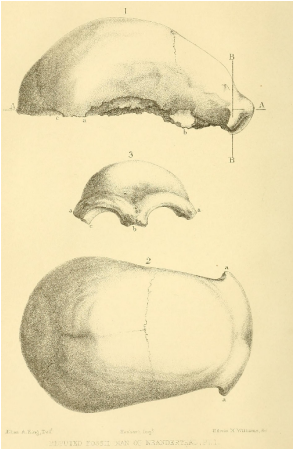


Fig. 3.1 ***Premier fossile de l'Homme de Néandertal décrit. Figure provenant de King et al. 1864***

La répartition spatiale de ces fossiles nous permet aujourd'hui d'avoir une idée de l'espace géographique occupé autrefois par ce groupe. Ainsi, on observe une grande concentration de fossiles en Europe (Mellars, 2004), avec plusieurs fossiles plus à l'est, jusque dans les montagnes d'Altaï (Krause et al., 2007). Cela suggère une présence des néandertaliens principalement en Europe, avec des périodes d'habitat en Asie.

Ces fossiles nous permettent également de savoir durant quelles périodes vivaient les néandertaliens. Les sites de fouilles connus ont été occupés pendant des dates très diverses, ainsi, il semblerait que les néandertaliens aient habité l'Eurasie depuis au moins 400 000 ans (Stringer & Hublin, 1999 ; Bischoff et al., 2007) et aient disparus il y a 40 000 ans (Hublin, 2017 ; Higham et al., 2014). La présence des néandertaliens en Eurasie pendant ces périodes signifie qu'ils y ont connu plusieurs périodes de glaciation. Il est ainsi probable que leur mode de vie leur ait permis de survivre dans un environnement froid.

Cela est soutenu par plusieurs éléments de leur physiologie. En effet, il semblerait que leur cavité nasale ait été adaptée à une respiration dans un milieu sec et froid, une adaptation aussi observée chez les européens modernes, mais qui est arrivée de manière indépendante (de Azevedo et al., 2017). Leurs membres courts comparés à leur tronc correspondent aux physiologies attendues dans l'hypothèse de la règle d'Allen.

3.1.2 Mode de vie des néandertaliens

Une des potentielles raisons pour les différences phénotypiques entre néandertaliens et humains anatomiquement modernes (HAM) peut se trouver dans le mode de vie. Il est donc important de lister rapidement ce qui est connu aujourd'hui sur ce sujet.

a) Alimentation

Il est possible d'obtenir les informations sur l'alimentation des néandertaliens à partir des fossiles. En effet, la concentration de certains isotopes dans le collagène des os nous renseigne sur les types d'isotopes consommés par ces individus, ainsi les concentrations de carbone 13 vont dépendre des types de végétaux consommés, et les niveaux d'azote 15 vont dépendre de la quantité de viande, ainsi que de son niveau trophique (i.e. sa place dans la chaîne alimentaire). Les analyses du site de Jonzac en France ont ainsi montré de haut niveaux d'azote 15 chez les néandertaliens, suggérant un régime riche en viande provenant de grands herbivores (Richards et al., 2008). Cependant le régime des néandertaliens n'était pas homogène. Ainsi, les fossiles de El Sidron en Espagne, exhibent des marques d'usure de leurs dents cohérents avec un régime omnivore, mais plus riche en plantes (Estalrrich, El Zaatari, & Rosas, 2017). Il semblerait donc que les néandertaliens suivaient un régime omnivore dépendant des ressources locales de chaque groupe.

b) Utilisation du feu

Il est également intéressant de noter que les preuves de l'utilisation du feu par les néandertaliens sont très éparses comparées aux HAM. Ceci suggère que, à l'inverse des humains anatomiquement modernes, les néandertaliens n'avaient recours au feu que de manière occasionnelle. Sous cette hypothèse, il est probable que les néandertaliens avaient un système digestif assez différent de celui des HAMs, afin de leur permettre de mieux digérer ces aliments, ou bien utilisaient des moyens alternatifs, telle que la putréfaction de la viande, afin de la rendre plus digeste (Dibble, Sandgathe, Goldberg, McPherron, & Aldeias, 2018).

c) Comportements complexes

Malgré l'image populaire des néandertaliens comme primitifs, plusieurs preuves indirectes laissent à penser qu'ils étaient capables de comportements complexes. Ainsi, des traces suggérant des enterrements rituels sont encore débattues (Sommer, 2009). De plus la construction d'un muret dans la grotte de Bruniquel, datée à 176 000 ans avant l'ère moderne, a été attribuée aux néandertaliens, seuls hominines de la région à cette époque. Ce muret, situé très en profondeur dans la grotte, aurait nécessité un éclairage artificiel pour sa construction, ainsi que le rassemblement en amont d'un grand nombre de pièces (morceaux de stalactites), suggérant une capacité de néandertal à planifier des ouvrages complexes. Un deuxième élément provient de le réévalution de la datation d'une peinture en Espagne. Cette peinture est datée d'au moins 64 000 ans, époque où les néandertaliens étaient les seuls habitants de la région (Hoffmann et al., 2018), indiquant la capacité à l'abstraction des concepts nécessaire

à la production d'art rupestre. Cependant ces deux interprétations doivent désormais être revues suite à la découverte récente de fossiles attribués aux HAM en Grèce datant de 200 000 ans avant l'ère moderne (Harvati et al., 2019).

3.2 Premières études de l'introgression archaïque

3.2.1 Premières hypothèses et preuves

L'hypothèse d'une potentielle hybridation entre les néandertaliens et les HAM n'est pas jeune, elle a en effet été soulevée sur la base de comparaison de crânes. Cependant, les premières réponses définitives sont arrivées par le biais de la génétique. Dans un premier temps, l'étude de l'ADN mitochondrial néandertalien a prouvé que la lignée néandertalienne n'avait pas transmis d'ADN mitochondrial aux populations modernes (Krings et al., 1997 ; Krings, Geisert, Schmitz, Krainitzki, & Pääbo, 1999 ; Serre et al., 2004 ; Green et al., 2008).

En 2006, une analyse des structures haplotypiques des humains modernes révèle la présence de plusieurs haplotypes long en Europe, cohérent avec une introgression il y a 50 000 ans (Wall, 2000 ; Plagnol & Wall, 2006 ; Wall, Lohmueller, & Plagnol, 2009). Les néandertaliens sont une hypothèse de choix. Ce qui sera confirmé par un premier séquençage à faible couverture du génome des néandertaliens basé sur trois individus différents (Green et al., 2010) montrant notamment que les néandertaliens partagent plus de variants génétiques avec les eurasiens qu'avec les populations sub-sahariennes, ce qui est suggestif d'un mélange génétique entre les néandertaliens et les eurasiens.

Cependant, le premier génome de néandertalien séquencé avec une haute couverture sera publié 4 ans plus tard (Prufer et al., 2014).

3.2.2 Caractérisations des génomes archaïques

Dans la grotte de Denisova, située dans les montagnes Altaï, en Sibérie, plusieurs ossements hominines archaïques ont été découverts, notamment la phalange qui a permis l'identification du groupe des dénisoviens (Krause et al., 2010 ; Reich et al., 2010 ; Meyer et al., 2012), un groupe d'hominines proche des néandertaliens, mais également l'os d'un orteil d'une femme néandertalienne, qui a été le premier ossement de néandertalien séquencé à haute couverture (Prufer et al., 2014). Ce séquençage a permis entre autre la première étude précise des échanges de gènes entre dénisoviens, néandertaliens et les humains modernes (Fig 3.2A). Ces deux génomes ont été complétés par un génome à haute couverture d'un individu néandertalien provenant de la grotte de Vindija, en Croatie (Prufer et al., 2017). Aujourd'hui, la connaissance croissante de la séquence et de la diversité des génomes néandertaliens a ouvert la porte à l'étude et la cartographie détaillée des segments néandertaliens introgressés dans le génome des populations eurasiennes contemporaines.

Nom	Taille d'haplotype	Groupe non métissé	Génome archaïque
S*	Oui	Non	Non
CRF	Indirecte	Oui	Oui
S prime	Oui	Relaxée	Non
HMM	Indirecte	Oui	Oui

Table 3.1 – Liste de méthodes permettant de cartographier l'introgression archaïque et les hypothèses associées

3.2.3 Méthodes d'identification de segments introgressés

Le métissage des premiers eurasiens avec les néandertaliens a laissé plusieurs traces spécifiques dans les génomes de leurs descendants. Depuis 2006, diverses méthodes ont permis l'identification de segments issus d'ancêtres néandertaliens dans les génomes modernes, (CF table 3.1), ces méthodes reposent sur 3 principes qui sont, la présence d'haplotypes longs et fortement divergés, la comparaison des haplotypes à un groupe non métissé, et la proximité des haplotypes à un génome archaïque de référence.

a) Présence d'haplotypes longs et fortement divergés

Dans le cas d'une introgression, les haplotypes introgressés ont une taille fortement dépendante du nombre de générations écoulées depuis cette introgression. En effet, à chaque génération, la recombinaison va réduire la taille des fragments introgressés en les mélangeant de plus en plus à l'ADN des HAM.

Ce raisonnement est à la base de la création de la statistique S*, qui identifie les longs haplotypes portant des mutations spécifiques (Plagnol & Wall, 2006 ; Wall et al., 2009). L'utilisation de cette statistique a permis l'identification d'une introgression d'un hominidé inconnu dans les populations africaines, et d'une introgression possiblement néandertalienne en Europe.

Des études plus récentes ont estimé la taille moyenne des haplotypes néandertaliens introgressés à 0,05cM (approximativement 50kb) (Sankararaman et al., 2014).

b) Distance à un groupe de référence non métissé (Africains)

Un moyen de reconnaître une introgression est également de connaître et d'utiliser un groupe d'individus dans lequel l'introgression n'a pas eu lieu. Dans le cas de l'introgression néandertalienne, on peut, pour se faire, utiliser les populations sub-sahariennes, en général représentées par une population de Yoruba. En effet, en plus de ne pas avoir de traces archéologiques de la présence des néandertaliens en Afrique, les études basées sur la taille des haplotypes ont rejeté l'hypothèse d'une introgression néandertalienne dans les populations africaines (Plagnol & Wall, 2006 ; Wall et al., 2009). Cette information est aujourd'hui utilisée dans de nombreuses méthodes, notamment dans la méthode Sprime -une amélioration de la méthode S* qui prend également en compte la variabilité de la recombinaison- et dans certaines études se basant sur des modèles de Markov cachés (MMC) (Browning, Browning, Zhou, Tucci, & Akey, 2018 ; Skov et al., 2018). Se baser uniquement sur un groupe de contrôle

non introgressé et ne pas utiliser de génome de référence permet notamment d'étudier plus efficacement les cas où une population a introgressé des haplotypes provenant de plusieurs hominines différents, comme c'est le cas en Océanie (Browning et al., 2018).

c) Proximité avec le génome archaïque

La proximité avec les génomes introgressant est probablement la méthode la plus directe pour identifier les segments introgressés dans les génomes modernes. Celle-ci couplée aux deux autres méthodes mentionnées a notamment été utilisée par plusieurs équipes pour cartographier l'introgression néandertalienne et dénisovienne en Europe, Asie, et Océanie (Sankararaman et al., 2014 ; Sankararaman, Mallick, Patterson, & Reich, 2016 ; Vernot & Akey, 2014 ; Vernot et al., 2016).

L'utilisation de cette information a cependant ses limites. Outre la nécessité d'avoir un génome de référence disponible, ce qui n'est par exemple pas le cas pour l'introgression archaïque africaine, si le génome de référence est trop éloigné du génome des individus introgressant, cela peut poser des problèmes. Ainsi après le séquençage à haute couverture d'un deuxième individu néandertalien, Néandertal Vindija, qui est un meilleur représentant de la population introgressante, la quantité d'introgression néandertalienne a été revue à la hausse, permettant de détecter 4,1 Mb introgressées supplémentaires (Prufer et al., 2017).

3.3 État de l'art sur l'introgression néandertalienne

Les cartographies de l'introgression ont ouvert de nouvelles branches d'étude. Il était notamment devenu possible d'étudier en détail à la fois quelles sont les pressions sélectives qui ont influencé cette introgression, mais également d'étudier quelles sont les phénotypes les plus impactés par cette dernière.

3.3.1 Dynamique d'introgression

a) Purges locales et générales des haplotypes introgressés

Les hybrides entre deux populations très avancées sur le chemin de la spéciation présentent des incompatibilités génétiques, nous avons déjà mentionné le cas des incompatibilités de Dobzhansky-Muller (Figure 1.2). Mais de manière plus générale, une infertilité des hybrides mâles est attendue dans ce cas (Coyne & Orr, 1989). Plusieurs éléments de la répartition des mutations néandertaliennes introgressées suggèrent que les néandertaliens et les HAMs ont rencontré ces difficultés pendant leur hybridation. Ainsi, les gènes sur-exprimés dans les testicules sont dépletés en ascendance néandertalienne dans les génomes eurasiens (Sankararaman et al., 2014). De plus, chez les hétérozygotes, les transcrits portant une mutation néandertalienne ont tendance à être moins exprimés que les transcrits classiques dans les testicules et le cerveau (McCoy, Wakefield, & Akey, 2017). Ces éléments, et leur présence au niveau des testicules, sont réguliérement interprétés comme un signe d'infertilité chez les hybrides masculins, ce qui est attendu dans un cas de spéciation. De plus, on observe dans

les génomes modernes des déserts d'ascendance néandertalienne (Sankararaman et al., 2014, 2016), c'est-à-dire de larges régions complètement dépourvues d'haplotypes néandertaliens. Ceci est suggestif d'une forte sélection négative sur les haplotypes introgressés à ces régions, il est cependant incertain si ce phénomène est dû à une incompatibilité ou à une forte sélection négative à ces loci.

Au delà de l'absence totale d'introgression néandertalienne à certains endroits du génome et des effets dus à la spéciation, il semblerait que les haplotypes néandertaliens aient été, de manière générale, sous sélection négative après l'introgression.

En effet, une des découvertes liées aux génomes des néandertaliens connus est leur petite taille de population effective (Prufer et al., 2014, 2017), due soit à une petite taille de population, soit à une forte structure à l'intérieur de celle-ci (Prufer et al., 2014 ; Kuhlwilm et al., 2016 ; Rogers, Bohlender, & Huff, 2017). Cette taille de population effective réduite à long terme a permis l'accumulation de mutations légèrement délétères dans les génomes néandertaliens. Plusieurs études basées sur des simulations suggèrent que cela a pu avoir un impact majeur sur les dynamiques de l'introgression néandertalienne (Harris & Nielsen, 2016 ; Petr, Pääbo, Kelso, & Vernot, 2019 ; Kim, Huber, & Lohmueller, 2018), le scénario le plus accepté étant celui d'une brutale baisse de l'ascendance néandertalienne moyenne de la population pendant les 100 premières générations, passant de $\sim$10% à 3%, puis une stabilité qui s'installe. Cette purge généralisée est extrêmement cohérente avec la corrélation négative observée entre l'ascendance Néandertalienne moyenne d'une région génomique, et la force de la sélection de fond qui y est exercée (Sankararaman et al., 2014, 2016).

b) Introgression adaptative

Le cas de l'introgression néandertalienne est également une opportunité d'adaptation à l'environnement via l'introgression. En effet, la population néandertalienne ayant vécu en Eurasie pendant $\sim$ 400 000 ans, cette dernière a eu le temps de s'adapter à cet environnement (température, pathogènes etc ...), ainsi, certaines mutations avantageuses dans cet environnement ont pu être transmises aux hybrides entre les néandertaliens et les premiers HAMs eurasiens, et leurs être bénéfiques. On parle dans ce cas d'introgression adaptative.

Ainsi, bien que la majorité de l'introgression néandertalienne ait été sous sélection négative, plusieurs exemples d'haplotypes introgressés ayant fourni un avantage évolutif aux populations eurasiennes ont été décrits. Ainsi, le cas du cluster *OAS* décrit au chapitre précédent est un exemple d'haplotype néandertalien sous sélection positive.

L'exemple le plus iconique d'introgression archaïque adaptative ne provient pas des néandertaliens, mais des dénisoviens. Ainsi, une mutation dans le gène *EPAS1* est sous forte sélection positive dans la population tibétaine, et provient de l'introgression avec les dénisoviens (Huerta-Sanchez et al., 2014). Ce gène est impliqué dans la régulation de la concentration de globules rouges dans le sang en fonction de la pression partielle d'oxygène dans l'environnement, la mutation en question permet à la population tibétaine de mieux vivre à haute altitude.

Cette mutation a été prouvée d'être sous sélection avant de savoir qu'elle provenait d'une

introgression archaïque, cependant de nouvelles méthodes peuvent détecter l'introgression archaïque adaptative spécifiquement (Racimo, Marnetto, & Huerta-Sanchez, 2017). Bien que cette méthode soit basée sur une approche par valeurs extrêmes, et n'identifie donc que les candidats les plus probables d'une introgression adaptative, elle a été fortement utile pour identifier à la fois quelles statistiques étaient les plus aptes à détecter l'introgression adaptative, mais aussi en fournissant une liste de candidats à explorer.

Il est intéressant de noter qu'il est extrêmement difficile d'obtenir une preuve claire d'introgression adaptative. En effet, les exemples fournis sont soit des approches par valeurs extrêmes, qui ne font qu'identifier les candidats les plus probables, ou sont basés sur des simulations sous hypothèse de neutralité, et sont donc fortement dépendants des modèles d'introgression considérés. Nous aborderons les difficultés inhérentes à ce problème plus en profondeur dans la discussion.

3.3.2 Conséquences phénotypiques

Pour plusieurs cas d'introgression adaptative le phenotype sous sélection est connu, comme par exemple le cas du locus *EPAS1* (Huerta-Sanchez et al., 2014). Cependant, il s'agit d'un impact phénotypique créé par un seul haplotype. L'introgression néandertalienne peut également impacter certains phénotypes de manière polygénique. C'est-à-dire que certains phénotypes ont été impactés par des haplotypes introgressés indépendants, et ce de manière excessive comparé à l'attendu sous neutralité.

Au niveau génétique

Il a été observé que l'ascendance néandertalienne moyenne est plus élevée dans les gènes de l'immunité innée chez les individus européens et asiatiques (Deschamps et al., 2016). Bien que cela n'ait pas été relié à un phénotype précis, cela reste suggestif d'un impact de l'introgression néandertalienne sur l'immunité innée des humains modernes. L'impact potentiel de cette introgression sur l'immunité des eurasiens a été confirmé par la découverte de l'excès de mutations néandertaliennes dans les transcrits des protéines interagissant avec les virus, en particulier les virus à ARN (Enard & Petrov, 2018).

De plus, comme cité précédemment, les gènes sur-exprimés dans les testicules sont déplétés en ascendance néandertalienne (Sankararaman et al., 2014), ce qui suggère un impact réduit de l'introgression néandertalienne sur la fonction testiculaire chez les humains modernes.

Au niveau de l'expression

L'introgression néandertalienne a fortement impacté l'expression des gènes. Il est également notable que, de manière générale, les haplotypes néandertaliens à haute fréquence (supérieure à 5%) semblent contrôler l'expression plus qu'attendu, à la fois de manière générale, mais également dans certains tissus particuliers tels que les tissus adipeux et les poumons (Dannemann, Prufer, & Kelso, 2017). Ce contrôle génétique par les haplotypes

néandertaliens peut également être dépendant de l'environnement, ainsi il a été prouvé que les haplotypes néandertaliens contrôlent la réponse transcriptionnelle à l'infection grippale de manière disproportionnée (Quach et al., 2016).

Comme cité précédemment, l'expression dans les testicules et le cerveau a également été affectée, quoique de manière différente. En effet, les transcrits dans le cerveau et dans les testicules qui portent une mutation néandertalienne sont moins exprimés que leur version moderne chez les individus hétérozygotes (McCoy et al., 2017).

Au niveau macroscopique

Plusieurs groupes ont également étudié directement si les mutations néandertaliennes étaient plus souvent associées à certains phénotypes qu'attendu. Ainsi les mutations néandertaliennes introgressées ont été associées, entre autres, à la kératose actinique et la dépression (Simonti et al., 2016). Une autre approche a notamment étudié à quels phénotypes étaient reliées les mutations néandertaliennes sous sélection positive, et les ont relié, notamment, au fait d'être fumeur et à la couleur des cheveux (Dannemann & Kelso, 2017). Sur ce dernier point, il est intéressant de constater que les mutations néandertaliennes présentes dans les génomes modernes affectent la couleur des cheveux à la fois vers des tons plus clairs, mais pour certaines, également vers des tons foncés. Cela permet de supposer que la population néandertalienne avait également une grande diversité de couleurs de cheveux.

3.3.3 Complexité des introgressions archaïques

Le modèle des introgressions archaïques présenté par Prûfer et collègues en 2014 a beaucoup évolué en cinq ans, et a gagné en précision et en complexité (Figure 3.2). Ici, nous discutons de quelques éléments importants découverts ces dernières années.

a) Différences entre Europe et Asie

Dès 2014, il a été observé que les populations de l'est de l'Asie ont une plus grande partie de leur génome d'ascendance néandertalienne que les populations européennes (Vernot & Akey, 2014 ; Sankararaman et al., 2014). Ces différences ont été attribuées à une deuxième vague d'introgression après la séparation entre les populations européennes et asiatiques (Vernot & Akey, 2014 ; Vernot et al., 2016), cependant il a également été théorisé que la plus faible ascendance néandertalienne des populations européennes proviendrait d'une hybridation avec une population qui ne se serait pas mélangée avec les néandertaliens (Lazaridis et al., 2016). Enfin très récemment, Mikkel Schierrup a présenté des résultats préliminaires expliquant que cette différence peut être expliquée par un temps de génération différent entre populations (données non publiées). L'origine de ces différences reste donc fortement débattue aujourd'hui.

b) Introgression dans les populations archaïques

Si les génomes néandertaliens et dénisoviens ont été introgressés à l'intérieur des génomes modernes, ils ont également reçu du matériel génétique de groupes divers. Ainsi les néandertaliens ont introgréssé une partie de génome provenant d'HAM (Kuhlwilm et al., 2016). De plus, si l'introgression des néandertaliens vers les dénisoviens avait été observée dès 2014 (Prufer et al., 2014), il est possible que les relations entre ces deux groupes d'hominines aient été sous-estimées au vu de la découverte récente d'un hybride de première génération entre les deux espèces (Slon et al., 2018). Il est intéressant de noter que la moitié dénisovienne du génome montre également des traces d'une introgression néandertalienne passée, solidifiant l'hypothèse que les échanges génétiques entre ces deux groupes n'ont pas été rares dans leur histoire. A ce jour, aucune trace d'introgression dénisovienne dans un génome néandertalien n'a été trouvée.

c) Introgression dans les populations africaines

Il serait également réducteur de penser que les introgressions archaïques se limitent à l'Eurasie, les néandertaliens et les dénisoviens. En effet, les génomes de plusieurs populations africaines portent des traces d'un potentiel hominine archaïque inconnu. Cette introgression n'a pas été étudiée à grande échelle, notamment car l'absence d'un génome de référence rend cela extrêmement difficile. Notons tout de même une étude en cours d'écriture qui s'applique à étudier ce phénomène en prenant la population néandertalienne comme population n'ayant pas reçu cette introgression (Durvasula & Sankararaman, 2019).

Ces éléments font partie d'une myriade d'indices différents qui témoignent de la diversité et de la complexité des échanges génétiques entre populations pendant le Pleistocène supérieur. Cependant, nous avons ici toutes les informations nécessaires au lecteur afin d'aborder les aspects recherchés et nouveaux présentés durant cette thèse.

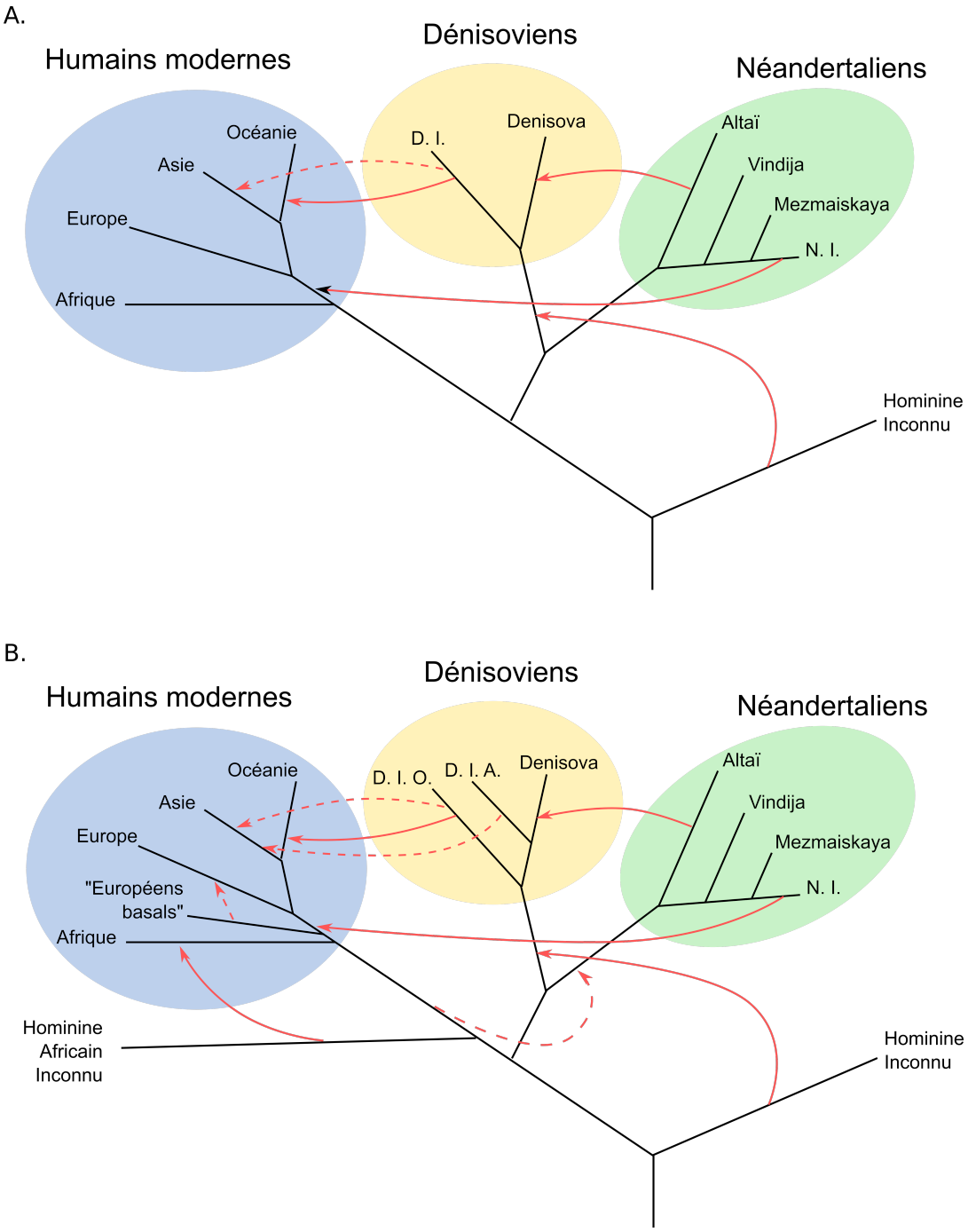


Fig. 3.2 ***Échanges génétiques horizontaux entre hominidés : A. Modèle proposé en 2014 des échanges génétiques entre homininés. Les dénisoviens et néandertaliens introgressant sont notés N.I. et D.I., B. Modèle mis à jour incluant plusieurs théories non complètement confirmées. La séparation des dénisoviens introgressant en deux groupes, introgressant en Asie et en Océanie n'est par exemple par certaine. Figures adaptées des articles suivants : (Prufer et al., 2014 ; Dannemann & Racimo, 2018)***

4. Objectifs de la thèse

La variabilité des profils d'expression visibles au sein de la population humaine a de multiples sources. Si ceux-ci peuvent être influencés par les environnements occupés par différentes populations, et par des comportements individuels, une partie de cette variabilité trouve son origine dans la diversité génétique. Ainsi, les profils d'expression visibles sont en partie dus au hasard inhérent à la diversité génétique, l'apparition des mutations et la dérive génétique. Cependant une partie d'entre eux peut être une cible de sélection, positive comme négative.

De plus l'expression des gènes est le résultat de nombreuses interactions. De la régulation de la transcription par les promoters et les enhancers à de nombreuses formes de régulations post-transcriptionelles. Cependant, ces régions régulatrices et mécanismes post-transcriptionnels peuvent être soumis à des pressions sélectives différentes, ils sont impactés de manière distincte par l'histoire démographique des populations.

Dans ce contexte, l'étude de l'évolution des différentes régions régulatrices au sein des populations humaines, ainsi que l'évaluation de leurs contributions respectives à la variabilité de l'expression génique sont essentielles à la compréhension de la variabilité phénotypique humaine. Ce manuscrit se propose donc d'étudier la contribution de la variabilité génétique à la régulation de l'expression génique sous deux angles différents.

Dans un premier temps, je me suis penché sur les conséquences de l'introgression néandertalienne sur la diversité au sein des régions régulatrices dans les populations eurasiennes. En effet, si les effets de l'introgression néandertalienne sur de nombreux phénotypes et sur l'expression des gènes avaient déjà été décrits par le passé, une comparaison de l'impact observé sur les différentes régions régulatrices n'avait pas été entreprise. De plus, à part quelques éléments soulignant un début de spéciation probable entre les néandertaliens et les HAM, les causes évolutives de ces différences d'expressions restaient un mystère. Pour cela, mon travail dans cette thèse a utilisé de nombreuses données publiquement disponibles afin de déterminer non seulement quelles sont les régions régulatrices dont la diversité provient de l'introgression néandertalienne de manière disproportionnée, mais également d'identifier si le probable évènement de sélection associé a eu lieu dans la population néandertalienne, ou bien après l'introgression.

Dans un second temps, je me suis intéressé plus précisément au fonctionnement d'un type de régulation particulier, la régulation par les miARN. En utilisant les résultats de séquençage des petits ARN dans les monocytes, immuno-stimulés ou non, de 100 individus d'ascendance européenne et 100 individus d'ascendance africaine, et en croisant ces résultats

avec les données d'expression et de transcription chez les mêmes individus, j'ai pu étudier à la fois la diversité de l'expression des miARN au sein de ces individus, mais également comment ceux-ci participent à la régulation de l'expression des gènes dans un contexte immunitaire.

5. Résultats I : origines et conséquences de la diversité néandertalienne dans les régions régulatrices

5.1 Contexte

Comme nous l'avons discuté dans les chapitres précédents, l'impact de l'introgression néandertalienne sur les phénotypes des populations modernes avait déjà été étudié par le passé (Simonti et al., 2016 ; Dannemann & Kelso, 2017). Le fait que cet évènement démographique avait impacté l'expression génique plus qu'attendu, au vu de la répartition des haplotypes néandertaliens, avait également été observé (Quach et al., 2016 ; Dannemann et al., 2017), et le fait que ces effets pouvaient varier selon le tissu ou un environnement particulier avait également été décrit (Quach et al., 2016 ; Dannemann et al., 2017 ; McCoy et al., 2017).

Cependant de nombreuses questions restaient sans réponse. En effet, les sources mécanistiques de ces effets n'avaient pas été étudiées. Notamment, ces études montraient un effet disproportionné des haplotypes néandertaliens sur l'expression, cependant il n'était pas clair si les effets de ces haplotypes étaient dus à des mutations néandertaliennes ou à des mutations présentes simultanément dans les populations humaines et néandertaliennes, et qui auraient été ré-introgressées dans les populations eurasiennes. De plus, les régions régulatrices les plus impactées par ces changements n'avaient pas été identifiées.

En croisant plusieurs banques de données publiques, notamment une cartographie des promoters et enhancers dans de nombreux tissus humains (Roadmap Epigenomics Consortium et al., 2015), les données concernant les miARNs humains (Kozomara & Griffiths-Jones, 2010), ainsi que les génomes de populations européennes et asiatiques (The 1000 Genomes Project Consortium et al., 2015), nous avons étudié, en premier lieu, la diversité due aux mutations néandertaliennes dans différentes classes régulatrices, puis nous nous sommes penchés plus en avant sur chaque classe, notamment en étudiant la participation des mutations néandertaliennes communes aux promoters et aux enhancers de 127 tissus différents. Enfin, dans le cadre de deux exemples, nous avons souligné qu'une forte densité de mutations néandertaliennes dans les régions régulatrices de différents tissus pouvait provenir de différents scénarios démographiques, et nous avons exploré les potentielles conséquences phénotypiques

de ces dernières en utilisant des données publiques de PC-HiC (Pan et al., 2018 ; Javierre et al., 2016).

5.2 Article

Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation

Martin Silvert,[1,2] Lluis Quintana-Murci,[1,3,*] and Maxime Rotival[1,3,*]

Archaic admixture is increasingly recognized as an important source of diversity in modern humans, and Neanderthal haplotypes cover 1%–3% of the genome of present-day Eurasians. Recent work has shown that archaic introgression has contributed to human phenotypic diversity, mostly through the regulation of gene expression. Yet the mechanisms through which archaic variants alter gene expression and the forces driving the introgression landscape at regulatory regions remain elusive. Here, we explored the impact of archaic introgression on transcriptional and post-transcriptional regulation. We focused on promoters and enhancers across 127 different tissues as well as on microRNA (miRNA)-mediated regulation. Although miRNAs themselves harbor few archaic variants, we found that some of these variants may have a strong impact on miRNA-mediated gene regulation. Enhancers were by far the regulatory elements most affected by archaic introgression: up to one-third of the tissues we tested presented significant enrichments. Specifically, we found strong enrichments of archaic variants in adipose-related tissues and primary T cells, even after accounting for various genomic and evolutionary confounders such as recombination rate and background selection. Interestingly, we identified signatures of adaptive introgression at enhancers of some key regulators of adipogenesis, raising the interesting hypothesis of a possible adaptation of early Eurasians to colder climates. Collectively, this study sheds new light on the mechanisms through which archaic admixture has impacted gene regulation in Eurasians and, more generally, increases our understanding of the contribution of Neanderthals to the regulation of acquired immunity and adipose homeostasis in modern humans.

The sequencing of the genomes of extinct human forms, such as Neanderthals or Denisovans, has enabled the mapping of archaic variants in the genomes of modern humans.[1–7] This archaic introgression has functional consequences today, as introgressed variants have been reported to alter a variety of phenotypes ranging from skin pigmentation to sleeping patterns and mood disorders.[8,9] Furthermore, several studies have shown that Neanderthal haplotypes are enriched in regulatory variants, with respect to non-archaic haplotypes,[10,11] suggesting that archaic introgression might impact complex, organismal phenotypes through changes in gene expression. Indeed, up to one-quarter of Neanderthal-introgressed haplotypes have been estimated to present *cis*-regulatory effects across tissues and there is a bias toward downregulation of Neanderthal alleles in brain and testes.[12] Furthermore, genes involved in innate immunity and interactions with RNA viruses have been reported to be enriched in Neanderthal ancestry.[13,14] Archaic variants affect, in particular, transcriptional responses to viral challenges.[11,15] A depletion of Neanderthal introgression has recently been documented in conserved coding regions and, surprisingly, in promoters,[16] suggesting that archaic introgression could affect gene expression through promoter-independent mechanisms. One such example is found in post-transcriptional regulation by miRNAs; such regulation has been reported to contribute to phenotypic differences between archaic and modern humans.[17,18] Thus, the relative contributions of transcriptional and post-transcriptional mechanisms to the effects of archaic variants on gene expression remain to be determined.

Our understanding of the selective forces that shaped the landscape of archaic introgression is also rapidly growing. In most cases, archaic variants were selected against, and regions of higher selective constraint, in particular those that are X-linked or contain testis-expressed and meiotic-related genes, were depleted in archaic ancestry.[1,2,19] Some studies have also suggested that Neanderthals had a reduced effective population size[6,7] because of a prolonged bottleneck or a deeply structured population.[6,20,21] Natural selection in Neanderthals would thus have been less efficient in purging deleterious mutations,[22,23] a large proportion of which were removed from the genome of modern humans after their admixture with Neanderthals.[16] However, archaic variants have also contributed, in some cases, to human adaptation[15,24–28] shortly after their introduction into modern humans or after an initial period of genetic drift.[29,30] Given the rapid evolution of regulatory regions and their potential adaptive nature,[31,32] the evolutionary dynamics of Neanderthal introgression at regulatory elements needs to be explored in further detail.

In this study, we aimed to increase knowledge about the impact archaic introgression has had on transcriptional and post-transcriptional mechanisms; we focused on promoter-, enhancer-, and microRNA (miRNA)-mediated regulation.[33,34] To this end, we first characterized the set

[1]Human Evolutionary Genetics Unit, Institut Pasteur, Centre National de la Recherche Scientifique, UMR 2000, 75015 Paris, France; [2]Sorbonne Universités, École Doctorale Complexité du Vivant, 75005 Paris, France
[3]These authors contributed equally to this work
*Correspondence: quintana@pasteur.fr (L.Q.-M.), maxime.rotival@pasteur.fr (M.R.)
https://doi.org/10.1016/j.ajhg.2019.04.016.

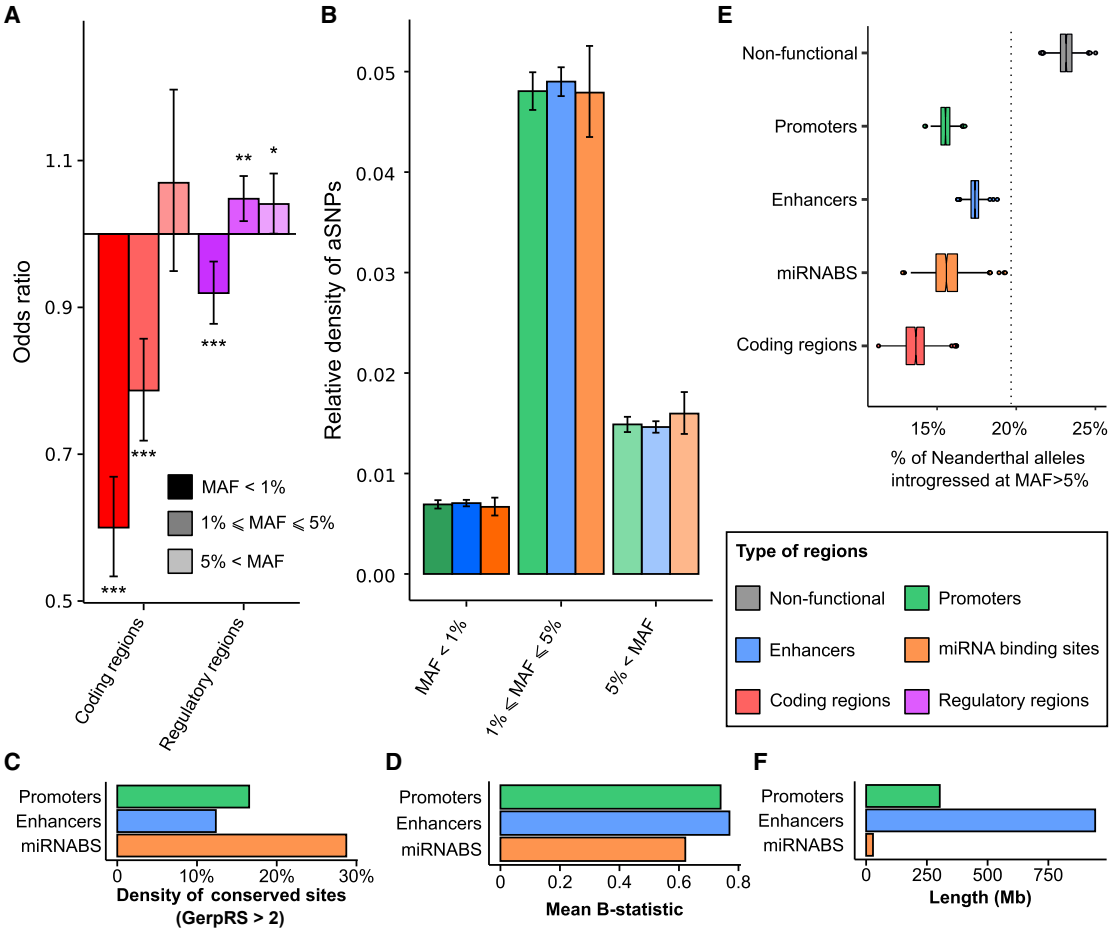


Figure 1. Enrichment of Neanderthal Variants in Regulatory Regions
(A) Odds ratio depicting the excess or depletion of Neanderthal variants in coding regions and regulatory elements (promoters, enhancers, and miRNA binding sites) compared to the remainder of the genome. Enrichments are shown for three bins of minor-allele frequencies (MAFs), together with 95% bootstrap confidence intervals: *p value < 0.05, **p value < 0.01, ***p value < 0.001.
(B) Relative density of aSNPs in promoters, enhancers, and miRNA binding sites in different MAF bins, with 95% bootstrap confidence intervals.
(C and D) Comparison of density of conserved sites (GerpRS > 2) and mean B statistic of promoters, enhancers, and miRNA binding sites.
(E) Percentage of alleles that are fixed in Neanderthal, absent from the African Yoruba from Nigeria (YRI), and introgressed at a MAF > 5% in Eurasians. For each type of region, box plots show the variability of the estimates based on 1,000 bootstrap resamples of 100 kb genomic windows. The dashed vertical line indicates the genome-wide average.
(F) Total length of promoters, enhancers, and miRNA binding sites.

of variants of putative Neanderthal origin — archaic SNPs (aSNPs) — as those for which one allele is both present in the Neanderthal Altai genome[6] and absent in the Yoruba African population of the 1000 Genomes Project[35] (see Supplemental Data). We further required aSNPs to be located in genomic regions where Neanderthal introgression has already been detected in Europe or Asia.[1] We then investigated deviations in the presence or absence of aSNPs among specific classes of functional elements by measuring the density of aSNPs with respect to that of non-aSNPs in the European (CEU) and Asian (CHB) populations of the 1000 Genomes Project.[35] We then compared the relative density of aSNPs at specific functional regions to that in the rest of the genome. Genomic regions were considered as enriched or depleted in aSNPs if the resulting odds ratio (OR) was significantly different from 1.

Overall, we observed a strong depletion of aSNPs in coding regions (OR = 0.71, p value < 10^{-4}) and similar levels of introgression at regulatory regions compared to those of non-functional elements. We then divided genetic variants according to the frequency of their minor allele, which corresponds to the Neanderthal allele in 99.8% of all aSNPs. Minor-allele frequency (MAF) was computed in Eurasian populations combined, and variants were split into three bins (rare – MAF < 1%, low frequency – 1% ≤ MAF < 5%, and common – MAF ≥ 5%). In doing so, we found that the depletion in coding regions was driven by rare and low-frequency variants (OR < 0.79, p value < 2×10^{-5}, Figure 1A), whereas regulatory regions were weakly enriched in low-frequency and common aSNPs (OR > 1.04, p value < 0.05). We then used the ancestral or derived state of each aSNP to distinguish between derived alleles

that originated in the Neanderthal lineage (i.e., derived aSNPs, 91% of all aSNPs) and ancestral alleles that were re-introduced by Neanderthals after the fixation of the derived allele in the human lineage (i.e., ancestral aSNPs, 9% of all aSNPs, Figure S1). When comparing derived aSNPs to non-archaic variants of similar derived allele frequency (DAF), we observed a depletion of both coding and regulatory regions in rare archaic variants (DAF < 1%, OR < 0.91, $p < 2 \times 10^{-4}$), and there was no significant enrichment of common and low-frequency aSNPs in regulatory regions. Interestingly, when focusing on variants presenting a DAF ≥ 95% (i.e., ancestral aSNPs at ≤ 5% frequency), we observed an enrichment of archaic variants in regulatory regions (OR > 1.25, $p < 2 \times 10^{-5}$), highlighting the contribution that ancestral alleles re-introgressed by Neanderthals make to gene regulation in humans.

To understand how introgression has impacted genetic diversity across various types of regulatory elements, we then investigated the relative density of aSNPs across promoters, enhancers, and miRNA binding sites. Although this metric differed markedly across frequency bins, it did not differ across categories of regulatory elements, despite their important differences in strength of negative or background selection (Figure 1B–1D). However, when measuring the rate at which Neanderthal alleles were introgressed in Europe or Asia, we found that they were less likely to reach high frequency (MAF > 5%) in coding or regulatory regions with respect to non-functional regions ($p < 10^{-10}$, Figures 1E and S2). This effect was less marked among enhancers, and that, together with the larger size of enhancers (Figure 1F), suggests that Neanderthal variants are quantitatively more likely to affect gene regulation via modification of enhancer activity than though changes of promoter or miRNA binding sites.

Given the low fraction of the genome that is covered by miRNAs and miRNA-binding sites (miRNABS) (Figure 1F), they are expected to be, quantitatively, the least affected by archaic introgression. Indeed, we only found six aSNPs that overlap the sequence of mature miRNAs, two of which alter the seed region (Figure 2A): rs74904371 in miR-2682-3p ($MAF_{CHB} = 0$, $MAF_{CEU} = 3\%$) and rs12220909 in miR-4293 ($MAF_{CHB} = 17\%$, $MAF_{CEU} = 0$). The presence of aSNPs in four of these miRNAs, particularly those located in seed regions, affected the set of genes they bind (Figure 2B and Table S1). We also detected 2,909 aSNPs in miRNABS, 29% of which were common (Table S2). We found a direct linear relationship between the number of genes bound by a miRNA and the number of aSNPs in its binding sites (r = 0.56, p value $< 10^{-10}$, Figure 2C), suggesting that introgression affected miRNABS independently of their cognate miRNAs. As a pertinent example, the *ONECUT2* locus (MIM: 604894) presents the highest number of aSNPs that alter conserved miRNABS (Figure 2D) and has been previously reported to be a likely target of adaptive introgression.[24] This gene, which encodes a member of the onecut family of transcription factors, contains 13 aSNPs that alter miRNABS, six of which are highly conserved (GerpRS > 2). Interestingly, these aSNPs fall within the 0.4% most-differentiated aSNPs between Europeans and Asians at the genome-wide level ($F_{ST} > 0.38$). We also detected aSNPs, mostly population specific, that alter conserved miRNABS at several key immune genes, including *CXCR5* (MIM: 601613; $MAF_{CHB} = 16\%$, $MAF_{CEU} = 1\%$), *TLR6* (MIM: 605403; $MAF_{CHB} = 8\%$, $MAF_{CEU} = 0$), *IL7R* (MIM: 146661; $MAF_{CHB} = 8\%$, $MAF_{CEU} = 0$), and *IL21* (MIM: 605384; $MAF_{CHB} = 0$, $MAF_{CEU} = 8\%$).

Next, we focused on how archaic introgression has affected promoters and enhancers. Given the tissue-specific impact of archaic introgression on gene regulation,[10,12] we searched for enrichments in Neanderthal ancestry across regulatory elements in 127 different tissues.[33] The impact of archaic introgression in promoters was similar to that in the remainder of the genome in all tissues and frequency bins (Table S3). Conversely, we found that enhancers are enriched in common aSNPs in 42 tissues (FDR < 5%, Figure 3A and Table S4), and we detected similar patterns in CEU and CHB populations (r = 0.62, Figure S3). Among the 42 tissues presenting significant enrichments, adipose-derived mesenchymal stem cells (AdMSCs) and mesenchymal stem-cell-derived adipocytes were the most enriched (OR > 1.13, p value $< 3 \times 10^{-5}$), followed by fetal heart (OR = 1.15, p value $= 8 \times 10^{-5}$), small intestine (OR = 1.21, p value $= 2 \times 10^{-4}$), and different T cell tissues (OR > 1.14, p value $< 1.5 \times 10^{-2}$). When restricting our analyses to derived aSNPs (and using SNPs with DAF < 50% as background set), we replicated the enrichments at enhancers for 27 tissues (FDR < 5%, Tables S3 and S4), indicating that the impact of archaic introgression for these tissues is driven by Neanderthal-derived variants.

Focusing on circulating immune cell types (Figure 3B), we found enrichments among enhancers of various types of primary T cells, the most significant being $CD4^+/CD25^-$ memory T cells (OR = 1.21, p value $= 2.2 \times 10^{-4}$), whereas enhancers of B cells, monocytes, and natural killer cells exhibited a density of common aSNPs similar to genome-wide expectations. We also observed that shared enhancers across different T cell subtypes (i.e., active in more than half of T cell subtypes, "core T cell enhancers") display an enrichment in aSNPs (OR = 1.22, p value $= 5 \times 10^{-4}$, Figure 3C) with respect to more specialized enhancers that are only active in a small fraction of T cell subtypes.

We sought to assess whether the enrichment in aSNPs detected in enhancers resulted from an excessive divergence of these elements in the Neanderthal lineage or from a higher rate of archaic introgression at enhancers. We quantified the number of fixed human-Neanderthal differences at enhancers across the 127 tissues and focused on sites where both the Altai and Vindija Neanderthal genomes[6,7] differ from the ancestral sequence. We uncovered large tissue variability; we found that enhancers active in induced pluripotent stem cells presented the highest

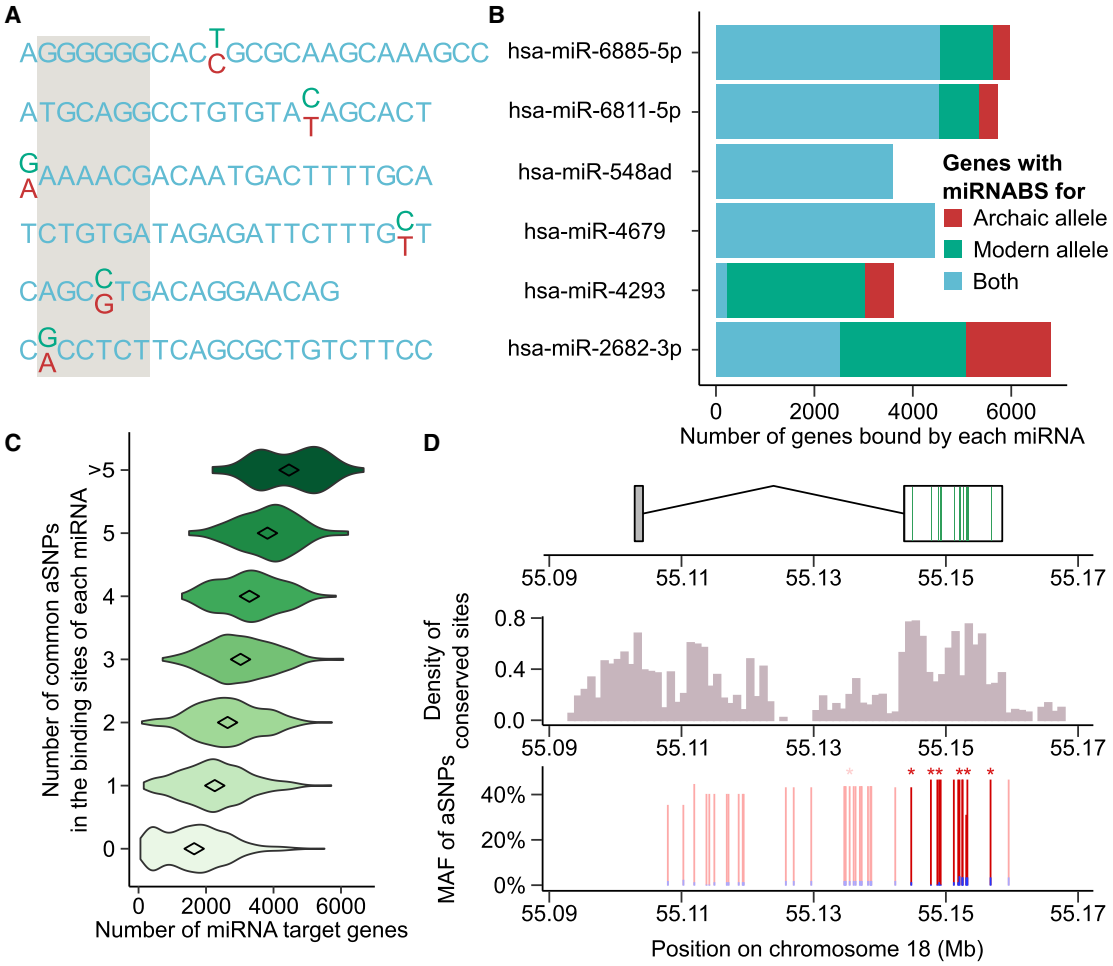


Figure 2. Effects of Archaic Introgression on miRNA-Mediated Regulation

(A) Representation of the archaic (red) and modern (green) human alleles for the six miRNAs presenting a Neanderthal-introgressed variant in their mature sequence. The seed region of the miRNAs is shaded in gray.

(B) Total number of genes bound by the archaic and/or modern human allele of each of the six miRNAs harboring a Neanderthal variant in their mature sequence.

(C) Relationship between the number of targets of each miRNA and the number of common aSNPs in the corresponding miRNA binding sites.

(D) Introgression of aSNPs altering the miRNABS at the *ONECUT2* locus (MIM: 604894). Gene structure is shown in the upper panel, and miRNA binding sites that are altered by archaic introgression are highlighted in green. The middle panel represents the density of conserved sites (GerpRS > 2) in 1,000 bp windows, and the bottom panel represents the repartition and frequency of archaic alleles at the locus (blue for CEU, red for CHB). aSNPs that overlap miRNABS are represented with a darker shade, and aSNPs that disrupt a conserved site are marked with stars.

divergence (290 differences/Mb) and that those active in pancreas cells showed the lowest (220 differences/Mb). However, given that the number of fixed differences strongly correlates with genetic diversity (i.e., density of common variants, $r = 0.71$, p value $< 10^{-20}$), we measured the ratio of the number of fixed differences between humans and Neanderthals differences to that of common, segregating SNPs in the region. Using this metric, we found that enhancers of T cells displayed the strongest divergence (7% increase compared to the mean across tissues, Wilcoxon p value $< 2 \times 10^{-8}$), whereas stem cells showed the lowest (4% decrease, Wilcoxon p value $< 7 \times 10^{-6}$) (Figure 4A). Focusing on the rate of introgression, which is defined as the proportion of Neanderthal-descended alleles that are present in the human genome at a MAF > 5%, we found that enhancers of T cells showed the highest percentage (5% increase, Wilcoxon p value $< 2 \times 10^{-5}$), whereas brain cells showed the lowest percentage (7% decrease, Wilcoxon p value $< 4 \times 10^{-5}$) (Figure 4A).

We then explored the factors that might drive, at the genome-wide level, the detected variation in Neanderthal divergence and archaic introgression. Using 100 kb windows, we correlated divergence and introgression with metrics that capture local variation in neutral (mutation and recombination) and selected (negative and background selection) diversity. Specifically, we measured the percentage of guanine-cytosine (GC) to account for their higher mutability, genetic size as measure of

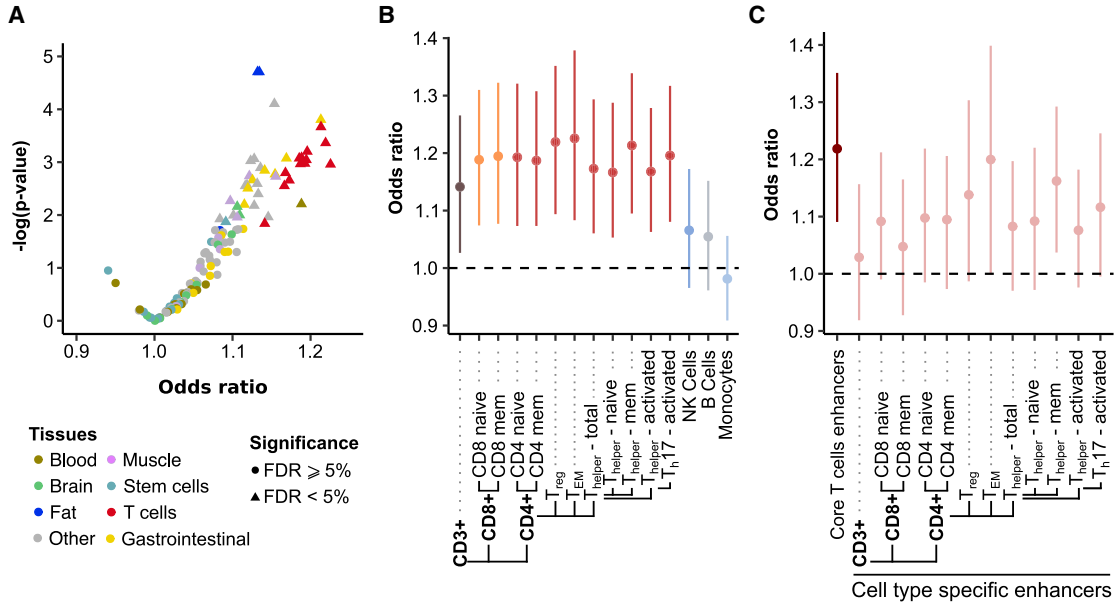


Figure 3. Effects of Archaic Introgression at Enhancers

(A) Volcano plot illustrating the enrichment of common aSNPs in the enhancers of 127 different tissues from the Epigenomic Roadmap Consortium. Tissues with FDR < 5% (triangles) are significantly enriched.

(B) Enrichments of common aSNPs in the enhancers of different immune tissues. Vertical bars indicate 95% confidence intervals computed by bootstrap analysis.

(C) Enrichment of common aSNPs in the enhancers that are active in more than half of the investigated T cell subtypes (dark red, referred to as "core T cells") and in enhancers that are active in each T cell subtype and are not part of core T cell enhancers (light red, referred to as "cell-type-specific enhancers"). Vertical bars indicate 95% confidence intervals computed by bootstrap.

(B and C) Note that $CD4^+$ T cells are separated on the basis of CD25 so that T_{reg} ($CD25^+$), T_{EM} ($CD25^{low}$), and T_{helper} ($CD25^-$) are distinguished from one another.

recombination rate, density of conserved sites (GerpRS > 2) as a measure of negative selection, and background selection derived as (1 − B), where B is the mean B statistic in the window.[36] We found that background selection correlated with a lower rate of archaic introgression (r = −0.049, p value < 10^{-15}, Figure 4B), consistent with previous findings,[1,2] but also correlated with increased local divergence (r = 0.22, p value < 10^{-20}, Figure 4C) and reduced density of both common variants and fixed differences (r = −0.46 and −0.05, respectively, p value < 5×10^{-20}, Figure S4). We also found that negative selection and recombination rate correlated with both divergence and introgression, even after we adjusted for background selection (Figure S5).

To understand further how these factors could account for the variation in divergence and introgression detected at enhancers (Figure S6), we focused on three model tissues: T cells (enhancers with high divergence and introgression), AdMSCs (enhancers with low divergence and high introgression), and prefrontal cortex (enhancers with high divergence and low introgression) (Figure 4D). When correcting for the various neutral and selective factors, we found that introgression at T cell enhancers did not exceed that of other tissues (p value > 0.11), but the high divergence and relative density of aSNPs remained significant (p value < 8×10^{-3}). For AdMSCs, introgression remained higher than expected (p value = 4×10^{-3}), leading to an excess of aSNPs despite their depletion in divergence (p value = 3.8×10^{-2}). For enhancers active at the prefrontal cortex, all variables were within expected bounds. Collectively, these analyses indicate that variation of several neutral and selective factors is not sufficient to explain the excess of Neanderthal introgression detected at enhancers. Some enhancers might have undergone past adaptation in the Neanderthal lineage or adaptive introgression in modern humans, as illustrated by T cells and AdMSCs, respectively.

Finally, we explored the impact of archaic introgression at enhancers on gene expression. To identify genes whose expression is altered by Neanderthal introgression at enhancers, we focused on tissues where data from promoter capture Hi-C were available[37,38] and assigned each enhancer located in a promoter-interacting region to the corresponding gene(s). Archaic variants at enhancers predicted to interact with a gene were strongly enriched in eQTLs (OR = 2.6, p value < 10^{-3}, Supplemental Note 1 and Figure S7), further supporting the regulatory potential of aSNPs. Genes interacting with T cell enhancers that harbor common aSNPs (n = 1,629, Table S5) were not enriched in any specific biological function. However, 285 of these genes are highly expressed in T cells (fragments per kilobase of transcript per million mapped reads [FPKM] > 100) and include known regulators of the immune response (e.g., *CXCR4* [MIM: 162643], *IL7R* [MIM: 146661], *IL10RA* [MIM: 146933], *NFKBIA*

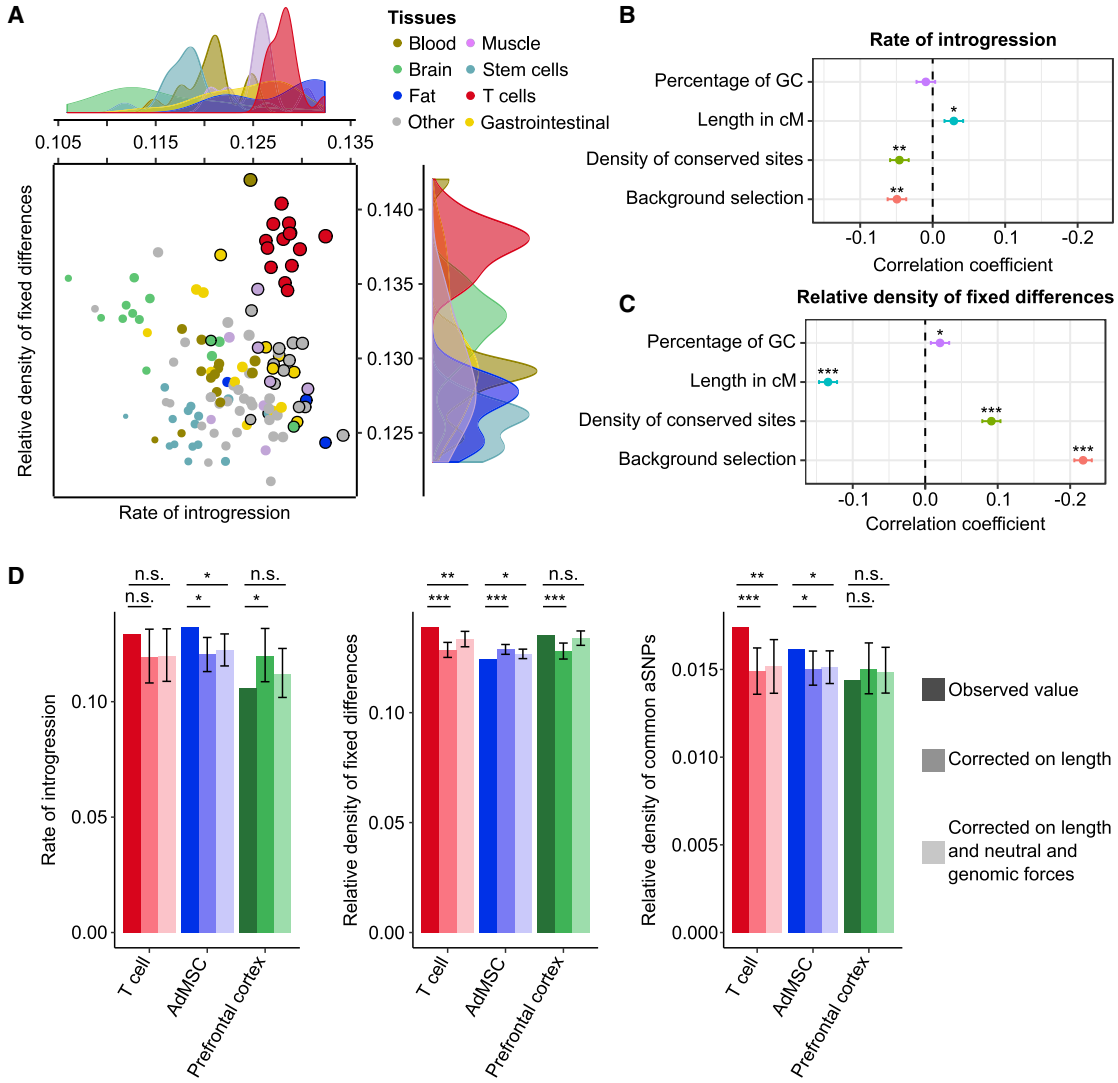


Figure 4. Factors Shaping Human-Neanderthal Divergence and Archaic Introgression at Enhancers
(A) Comparison of the relative density of fixed human-Neanderthal differences and rate of introgression in the enhancers of the 127 tissues studied. The size of the circles is proportional to the relative density of common aSNPs in the enhancers of the corresponding tissue; a black circle is added when the relative density of common aSNPs is significantly higher in these enhancers (FDR < 5%) than in the rest of the genome. The density of each tissue category along the two axes is also presented.
(B and C) Genome-wide correlations, using 100 kb windows, between either the rate of Neanderthal introgression (B) or the relative density of fixed human-Neanderthal differences (C) and neutral and selective forces. *p value < 10^{-2}, **p value < 10^{-10}, and *** p value < 10^{-20}. For each correlation, horizontal lines indicate 95% confidence interval.
(D) Observed values of rate of introgression and relative density of fixed differences and common aSNPs at the enhancers of core T cells, AdMSCs, and prefrontal cortex, with respect to expectations based on 100 kb windows matched for length of enhancers alone or for length of enhancers, percentage of GC, recombination rate, density of conserved sites, and mean B statistic of their enhancers (see Supplemental Data). n.s. = not significant; *p value < 0.05, **p value < 0.01, and ***p value < 10^{-3}. Errors bars indicate 95% confidence intervals of the expected values obtained by resampling.

[MIM: 164008], and *PTPRC* [MIM: 151460]). We found 14 loci presenting signatures of adaptive introgression; i.e., these were genes that interact with enhancers harboring very-high-frequency aSNPs (99th percentile of MAF: $MAF_{CEU} > 0.29$ or $MAF_{CHB} > 0.35$; Figures 5A and 5B). Among these, we found *ANKRD27*, which is associated with eosinophilic esophagitis (MIM: 610247),[39] and *MED15* (MIM: 607372), which is involved in several cancers.[40–42] With respect to adipose-related tissues, we identified 690 genes — 43 of which were highly expressed (FPKM > 100) in the adipose tissue — interacting with AdMSC enhancers that contain common aSNPs (Table S6). These genes were enriched in functions related to the regulation of cell motility (GO: 2000145, p value < 2.0×10^{-8}) and insulin-like growth factor binding protein complex (GO: 0016942, p value < 2.7×10^{-5}) (Table S7). We detected 16 aSNPs at AdMSC enhancers that present strong signatures of adaptive introgression (Figures 5C and 5D).

This study reconstructs the history of how Neanderthal introgression has affected various types of regulatory

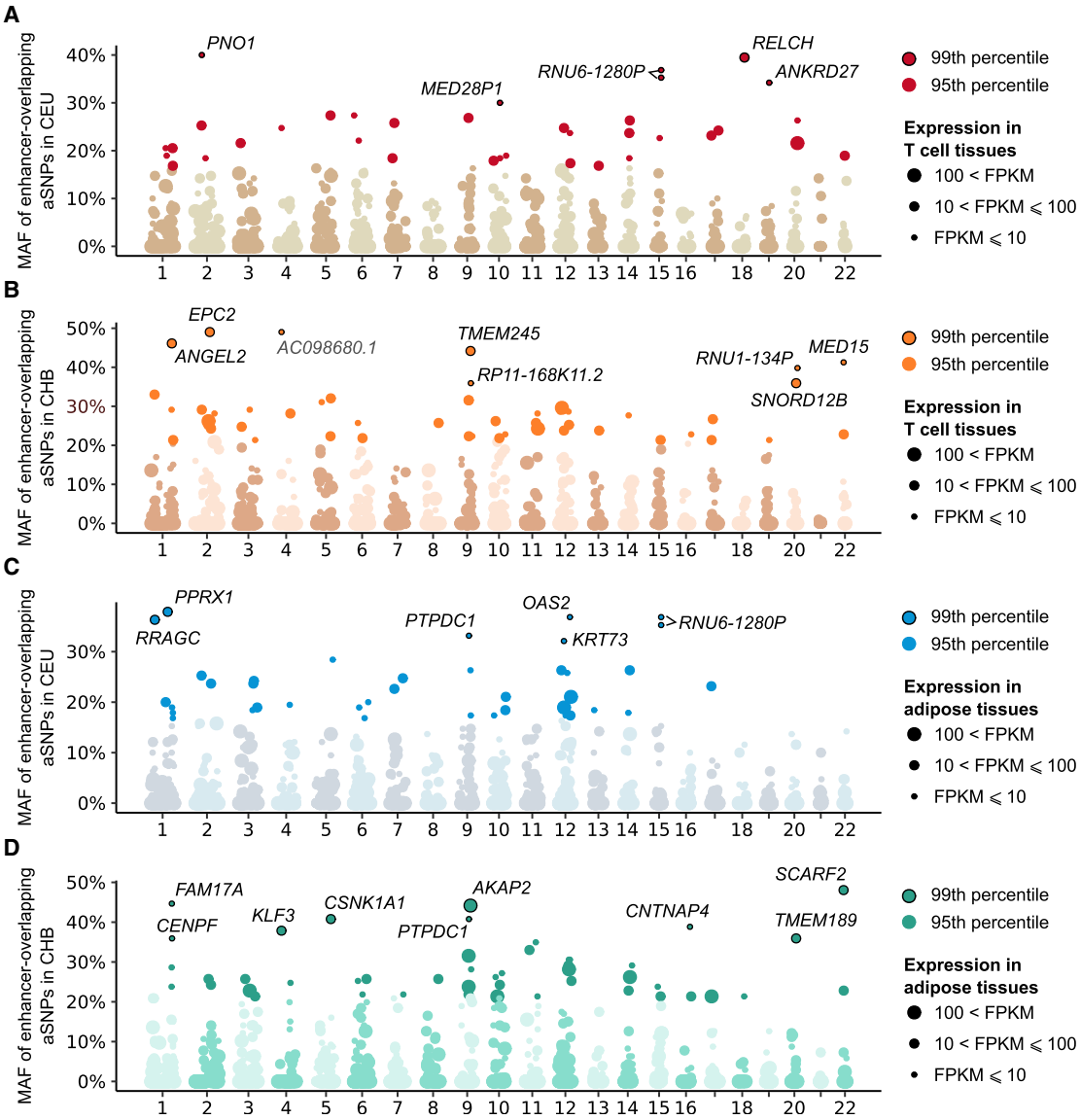


Figure 5. Manhattan Plots of Genes Interacting with Enhancers That Contain Archaic Variants
(A and B) Genome-wide distribution of MAFs in CEU or CHB at aSNPs that overlap enhancers active in T Cells (core T cell enhancers). For each window of 1 Mb along the genome, only the aSNP with the highest MAF is shown. Point sizes reflect FPKM of the most expressed genes (max FPKM across T lymphocytes from Blueprint database[58]) among genes interacting with the enhancer in T cells.[37]
(C and D) Similar plots for enhancers active in AdMSCs. Point sizes reflect the FPKM of the most expressed gene (max FPKM in GTEx tissues *Adipose—Subcutaneous* and *Adipose—Visceral [Omentum]*[59]) among genes interacting with the enhancer in adipose tissue.[38]

elements as well as the mechanistic bases through which archaic variants have altered gene regulation. Previous studies have shown that archaic variants are more likely to correlate with gene expression than non-archaic variants segregating at the same frequency.[10,11] Our approach differs from these studies in that it excludes indirect effects from non-archaic variants segregating on introgressed haplotypes and that it focuses on the direct regulatory potential of archaic alleles. In doing so, we find little evidence for an enrichment of common archaic variants in regulatory regions taken as a whole; these results might seem at odds with previous studies. Yet one should note that the functional impact of the archaic material we measure can be decomposed in two separate components: (1) the frequency at which Neanderthal haplotypes are introgressed into the human lineage, which corresponds to the rate of introgression measured by the f4-ratio statistics,[16] and (2) the degree of human-Neanderthal divergence at regulatory elements, which determines the probability that introgressed haplotypes carry a functional variant. Indeed, when focusing on the rate of introgression, we find that Neanderthal alleles were introgressed at a lower rate in regulatory regions, consistent with recent findings for promoter regions.[16] This lower introgression rate is nevertheless compensated by a higher human-Neanderthal divergence at regulatory regions (Figure S8), which is

consistent with an increased probability that Neanderthal haplotypes are associated with gene expression.[10,11]

We also explored how Neanderthal introgression has impacted miRNA-mediated regulation, and we showed that although miRNAs harbor few archaic variants per se, some of them might impact strongly miRNA-mediated gene regulation and disease risk. For example, the archaic allele at miR-4293 (rs12220909) is responsible for the loss of 95% of its targets and has been associated with diminished cancer susceptibility.[43,44] Archaic introgression has also affected miRNA binding sites, as illustrated by *ONECUT2* (MIM: 604894), where an archaic haplotype that is present at a high frequency in Asia ($MAF_{CHB} = 0.49$) alters multiple conserved miRNA binding sites. *ONECUT2* is involved in liver, pancreas, and nervous-system development[45] and has recently been proposed as a regulator of tumor growth in ovarian cancer (MIM: 167000).[46]

Finally, our study reveals that archaic introgression has impacted enhancers in a tissue-specific manner, reflecting either high human-Neanderthal differentiation, as observed in T cell enhancers, or increased archaic introgression, as detected in AdMSCs. Interestingly, the AdMSC enhancers impacted by archaic introgression interact preferentially with genes involved in the regulation of adipocyte differentiation and adipogenesis. These include receptors such as *PDGFRB* (MIM: 173410) and *TGFBR2* (MIM: 190182), the insulin growth factor *IGF1* (MIM: 147440) and its binding partners *IGFBP2* (MIM: 146731) and *IGFBP3* (MIM: 146732), and the *CXCR4* chemokine (MIM: 162643).[47–53] Furthermore, two of the enhancers harboring archaic variants at the highest frequencies interact with key adipocyte-differentiation regulators, such as *KLF3* (MIM: 609392) and *PRRX1* (MIM: 167420),[54,55] suggesting that introgression at AdMSC might have been adaptive in humans. In support of this notion, Dannemann and colleagues have found that more than half of aSNPs associated with gene expression in subcutaneous adipose tissue had increased in frequency over the last 10,000 years, whereas the majority of aSNPs had decreased in frequency over the same period of time.[10] Given the proposed adaptation of Neanderthals to cold environments,[56] it is tempting to speculate that archaic alleles at enhancers of AdMSCs provided a selective advantage to early modern humans during their migration out of Africa. This hypothesis becomes particularly interesting in light of previously reported cases of adaptive introgression at the *LEPR* (MIM: 601007) locus and at the locus of the *WARS2* (MIM: 604733) and *TBX15* (MIM: 604127) genes, both loci being involved in the regulation of adipose tissue differentiation and body-fat distribution.[25,57] Further studies aiming to functionally characterize the regulatory effects of Neanderthal variants on adipocyte differentiation and fat distribution are now required, as these archaic variants might have contributed to the adaptation of early Eurasians to colder climates.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.04.016.

Acknowledgments

This work was supported by the Institut Pasteur, the Centre National de la Recherche Scientifique (CNRS), and the Agence Nationale de la Recherche (ANR) grants: "IEIHSEER" ANR-14-CE14-0008-02 and "TBPATHGEN" ANR-14-CE14-0007-02. The laboratory of L.Q.M. has received funding from the French government's Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant no. ANR-10- LABX-62-IBEID). M.S. was funded by the Ecole Doctorale "Complexité du vivant," Sorbonne Université (contract n°2532/2016).

Declaration of Interests

The authors declare no competing interests.

Received: January 12, 2019
Accepted: April 23, 2019
Published: May 30, 2019

Web Resources

1000 Genomes Project, http://www.internationalgenome.org/
Epigenomic Roadmaps, https://egg2.wustl.edu/roadmap/web_portal/
Genotype-Tissue Expression (GTEx) project, https://gtexportal.org/home/
miRanda Software, http://www.microrna.org/microrna/
Neanderthal Genomes, http://cdna.eva.mpg.de/neandertal/
OMIM, http://www.omim.org

References

1. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.
2. Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr. Biol. *26*, 1241–1247.
3. Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.
4. Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., et al. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science *352*, 235–239.
5. Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell *173*, 53–61.e9.
6. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

7. Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655–658.
8. Dannemann, M., and Kelso, J. (2017). The contribution of Neanderthals to phenotypic variation in modern humans. Am. J. Hum. Genet. *101*, 578–589.
9. Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Jarvik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science *351*, 737–741.
10. Dannemann, M., Prüfer, K., and Kelso, J. (2017). Functional implications of Neandertal introgression in modern humans. Genome Biol. *18*, 61.
11. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell *167*, 643–656.e17.
12. McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. Cell *168*, 916–927.e12.
13. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. *98*, 5–21.
14. Enard, D., and Petrov, D.A. (2018). Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. Cell *175*, 360–371.e13.
15. Sams, A.J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J.E., Messer, P.W., and Barreiro, L.B. (2016). Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome Biol. *17*, 246.
16. Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. Proc. Natl. Acad. Sci. USA *116*, 1639–1644.
17. Lopez-Valenzuela, M., Ramírez, O., Rosas, A., García-Vargas, S., de la Rasilla, M., Lalueza-Fox, C., and Espinosa-Parrilla, Y. (2012). An ancestral miR-1304 allele present in Neanderthals regulates genes involved in enamel formation and could explain dental differences with modern humans. Mol. Biol. Evol. *29*, 1797–1806.
18. Gunbin, K.V., Afonnikov, D.A., Kolchanov, N.A., Derevianko, A.P., and Rogaev, E.I. (2015). The evolution of Homo sapiens denisova and Homo sapiens neanderthalensis miRNA targeting genes in the prenatal and postnatal brain. BMC Genomics *16* (*Suppl 13*), S4.
19. Jégou, B., Sankararaman, S., Rolland, A.D., Reich, D., and Chalmel, F. (2017). Meiotic genes are enriched in regions of reduced archaic ancestry. Mol. Biol. Evol. *34*, 1974–1980.
20. Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de la Rasilla, M., et al. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. Nature *530*, 429–433.
21. Rogers, A.R., Bohlender, R.J., and Huff, C.D. (2017). Early history of Neanderthals and Denisovans. Proc. Natl. Acad. Sci. USA *114*, 9859–9863.
22. Harris, K., and Nielsen, R. (2016). The genetic cost of Neanderthal Introgression. Genetics *203*, 881–891.
23. Juric, I., Aeschbacher, S., and Coop, G. (2016). The strength of selection against Neanderthal introgression. PLoS Genet. *12*, e1006340.
24. Racimo, F., Marnetto, D., and Huerta-Sánchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. Mol. Biol. Evol. *34*, 296–317.
25. Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E., and Nielsen, R. (2017). Archaic adaptive introgression in TBX15/WARS2. Mol. Biol. Evol. *34*, 509–524.
26. Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akey, J.M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. Curr. Biol. *26*, 3375–3382.
27. Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. *16*, 359–371.
28. Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature *512*, 194–197.
29. Jagoda, E., Lawson, D.J., Wall, J.D., Lambert, D., Muller, C., Westaway, M., Leavesley, M., Capellini, T.D., Mirazón Lahr, M., Gerbault, P., et al. (2017). Disentangling immediate adaptive introgression from selection on standing introgressed variation in humans. Mol. Biol. Evol. *35*, 623–630.
30. Dannemann, M., and Racimo, F. (2018). Something old, something borrowed: admixture and adaptation in human evolution. Curr. Opin. Genet. Dev. *53*, 1–8.
31. Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009). Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. *26*, 649–658.
32. Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.
33. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
34. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol. *5*, R1.
35. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
36. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. *5*, e1000471.
37. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al.; BLUEPRINT Consortium (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell *167*, 1369–1384.e19.
38. Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek,

M., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nat. Commun. *9*, 1512.
39. Sleiman, P.M., Wang, M.L., Cianferoni, A., Aceves, S., Gonsalves, N., Nadeau, K., Bredenoord, A.J., Furuta, G.T., Spergel, J.M., and Hakonarson, H. (2014). GWAS identifies four novel eosinophilic esophagitis loci. Nat. Commun. *5*, 5593.
40. Syring, I., Weiten, R., Müller, T., Schmidt, D., Steiner, S., Kristiansen, G., Müller, S.C., and Ellinger, J. (2018). The knockdown of the Mediator complex subunit MED15 restrains urothelial bladder cancer cells' malignancy. Oncol. Lett. *16*, 3013–3021.
41. Weiten, R., Müller, T., Schmidt, D., Steiner, S., Kristiansen, G., Müller, S.C., Ellinger, J., and Syring, I. (2018). The Mediator complex subunit MED15, a promoter of tumour progression and metastatic spread in renal cell carcinoma. Cancer Biomark. *21*, 839–847.
42. Shaikhibrahim, Z., Offermann, A., Halbach, R., Vogel, W., Braun, M., Kristiansen, G., Bootz, F., Wenzel, J., Mikut, R., Lengerke, C., et al. (2015). Clinical and molecular implications of MED15 in head and neck squamous cell carcinoma. Am. J. Pathol. *185*, 1114–1122.
43. Fan, L., Chen, L., Ni, X., Guo, S., Zhou, Y., Wang, C., Zheng, Y., Shen, F., Kolluri, V.K., Muktiali, M., et al. (2017). Genetic variant of miR-4293 rs12220909 is associated with susceptibility to non-small cell lung cancer in a Chinese Han population. PLoS ONE *12*, e0175666.
44. Zhang, P., Wang, J., Lu, T., Wang, X., Zheng, Y., Guo, S., Yang, Y., Wang, M., Kolluri, V.K., Qiu, L., et al. (2015). miR-449b rs10061133 and miR-4293 rs12220909 polymorphisms are associated with decreased esophageal squamous cell carcinoma in a Chinese population. Tumour Biol. *36*, 8789–8795.
45. Kropp, P.A., and Gannon, M. (2016). Onecut transcription factors in development and disease. Trends Dev. Biol. *9*, 43–57.
46. Lu, T., Wu, B., Yu, Y., Zhu, W., Zhang, S., Zhang, Y., Guo, J., and Deng, N. (2018). Blockade of ONECUT2 expression in ovarian cancer inhibited tumor cell proliferation, migration, invasion and angiogenesis. Cancer Sci. *109*, 2221–2234.
47. Gao, Z., Daquinag, A.C., Su, F., Snyder, B., and Kolonin, M.G. (2018). PDGFRα/PDGFRβ signaling balance modulates progenitor cell differentiation into white and beige adipocytes. Development *145*, dev155861.
48. Onogi, Y., Wada, T., Kamiya, C., Inata, K., Matsuzawa, T., Inaba, Y., Kimura, K., Inoue, H., Yamamoto, S., Ishii, Y., et al. (2017). PDGFRβ regulates adipose tissue expansion and glucose metabolism via vascular remodeling in diet-induced obesity. Diabetes *66*, 1008–1021.
49. Kim, Y.J., Hwang, S.J., Bae, Y.C., and Jung, J.S. (2009). MiR-21 regulates adipogenic differentiation through the modulation of TGF-beta signaling in mesenchymal stem cells derived from human adipose tissue. Stem Cells *27*, 3093–3102.
50. Chang, H.R., Kim, H.J., Xu, X., and Ferrante, A.W., Jr. (2016). Macrophage and adipocyte IGF1 maintain adipose tissue homeostasis during metabolic stresses. Obesity (Silver Spring) *24*, 172–183.
51. Yau, S.W., Russo, V.C., Clarke, I.J., Dunshea, F.R., Werther, G.A., and Sabin, M.A. (2015). IGFBP-2 inhibits adipogenesis and lipogenesis in human visceral, but not subcutaneous, adipocytes. Int. J. Obes. *39*, 770–781.
52. Chan, S.S., Schedlich, L.J., Twigg, S.M., and Baxter, R.C. (2009). Inhibition of adipocyte differentiation by insulin-like growth factor-binding protein-3. Am. J. Physiol. Endocrinol. Metab. *296*, E654–E663.
53. Yao, L., Heuser-Baker, J., Herlea-Pana, O., Zhang, N., Szweda, L.I., Griffin, T.M., and Barlic-Dicen, J. (2014). Deficiency in adipocyte chemokine receptor CXCR4 exacerbates obesity and compromises thermoregulatory responses of brown adipose tissue in a mouse model of diet-induced obesity. FASEB J. *28*, 4534–4550.
54. Sue, N., Jack, B.H., Eaton, S.A., Pearson, R.C., Funnell, A.P., Turner, J., Czolij, R., Denyer, G., Bao, S., Molero-Navajas, J.C., et al. (2008). Targeted disruption of the basic Krüppel-like factor gene (Klf3) reveals a role in adipogenesis. Mol. Cell. Biol. *28*, 3967–3978.
55. Du, B., Cawthorn, W.P., Su, A., Doucette, C.R., Yao, Y., Hemati, N., Kampert, S., McCoin, C., Broome, D.T., Rosen, C.J., et al. (2013). The transcription factor paired-related homeobox 1 (Prrx1) inhibits adipogenesis by activating transforming growth factor-β (TGFβ) signaling. J. Biol. Chem. *288*, 3036–3047.
56. Steegmann, A.T., Jr., Cerny, F.J., and Holliday, T.W. (2002). Neandertal cold adaptation: Physiological and energetic factors. Am. J. Hum. Biol. *14*, 566–583.
57. Sazzini, M., Schiavo, G., De Fanti, S., Martelli, P.L., Casadio, R., and Luiselli, D. (2014). Searching for signatures of cold adaptations in modern and archaic humans: hints from the brown adipose tissue genes. Heredity (Edinb) *113*, 259–267.
58. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell *167*, 1398–1414.e24.
59. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

The American Journal of Human Genetics, Volume *104*

Supplemental Data

Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation

Martin Silvert, Lluis Quintana-Murci, and Maxime Rotival

Supplemental Figures

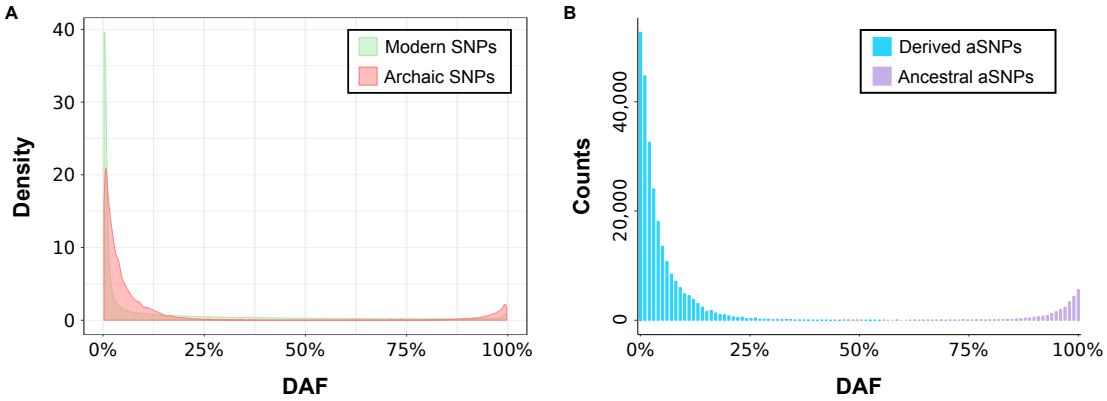


Figure S1. Allele frequency spectrum of modern and archaic alleles. (A) Densities of derived allele frequency (DAF) of modern and archaic variants. (B) Number of aSNPs within each bin of DAF, for derived-aSNPs and ancestral-aSNPs separately.

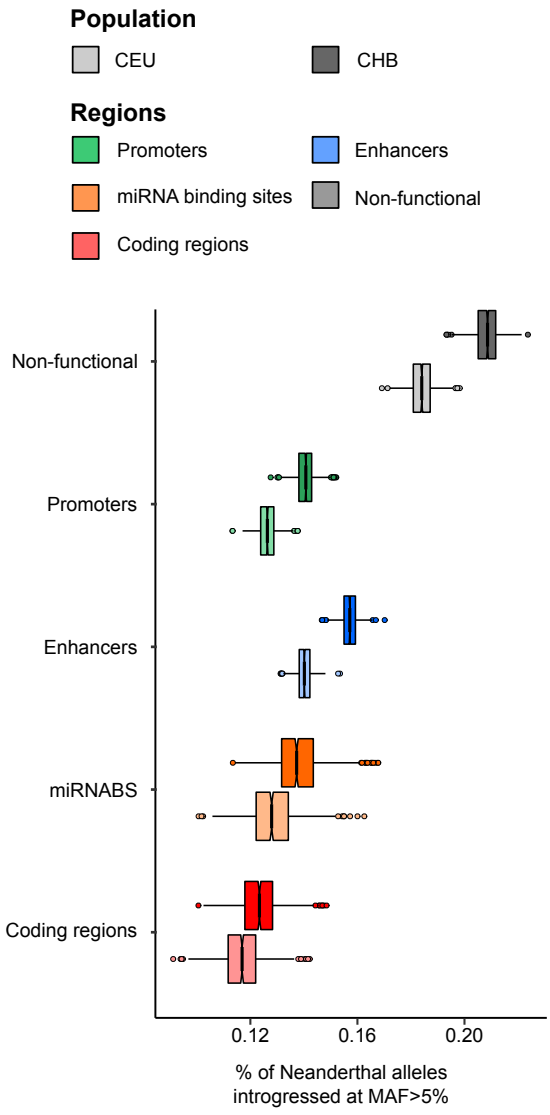


Figure S2. Percentage of Neanderthal alleles introgressed at a MAF > 5% across different genomic regions, when considering CEU and CHB separately. For each type of region, boxplots show the variability of the estimates based on 1,000 bootstrap resamples of 100kb genomic windows.

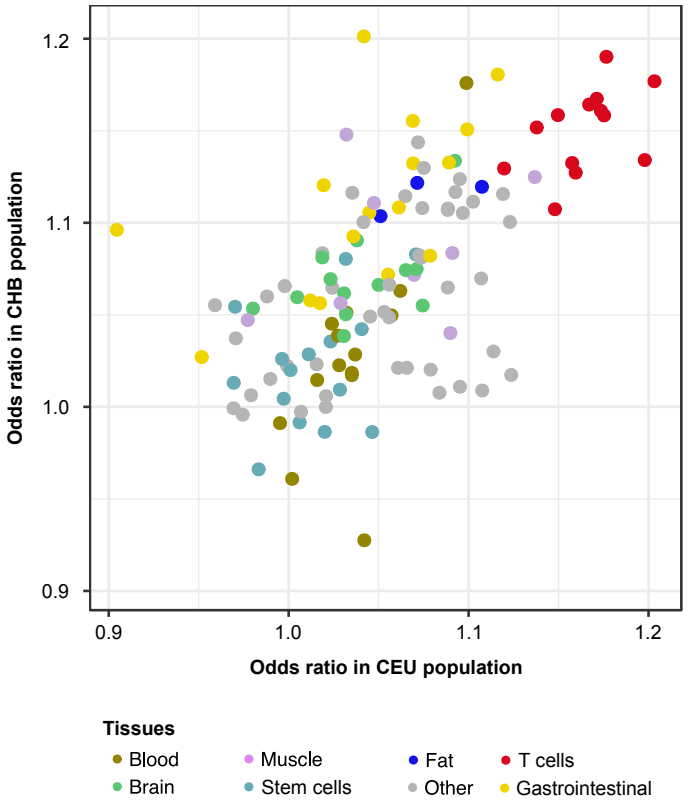


Figure S3. Comparison of enrichments in common aSNPs at enhancers between European and Asian populations. For each of the 127 tissues tested, odds ratios are computed in European and Asian samples separately, based on archaic and modern variants that segregate at MAF > 5% in each population.

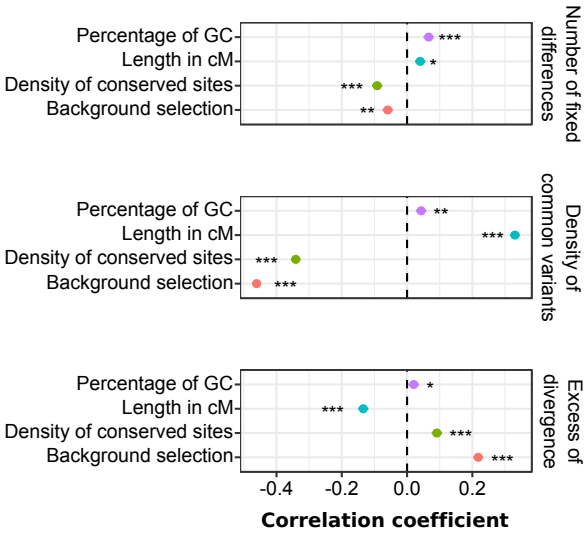


Figure S4. Effects of neutral and selective factors on the density of common variants and fixed human-Neanderthal differences. Genome-wide correlations, using 100kb-windows, between the density of fixed human-Neanderthal differences, the density of common variants in Eurasia or the ratio of these metrics (i.e., excess of divergence), and several proxies of neutral and selective factors. **p*-value < 10^{-2}, ***p*-value < 10^{-10}, ****p*-value < 10^{-20}.

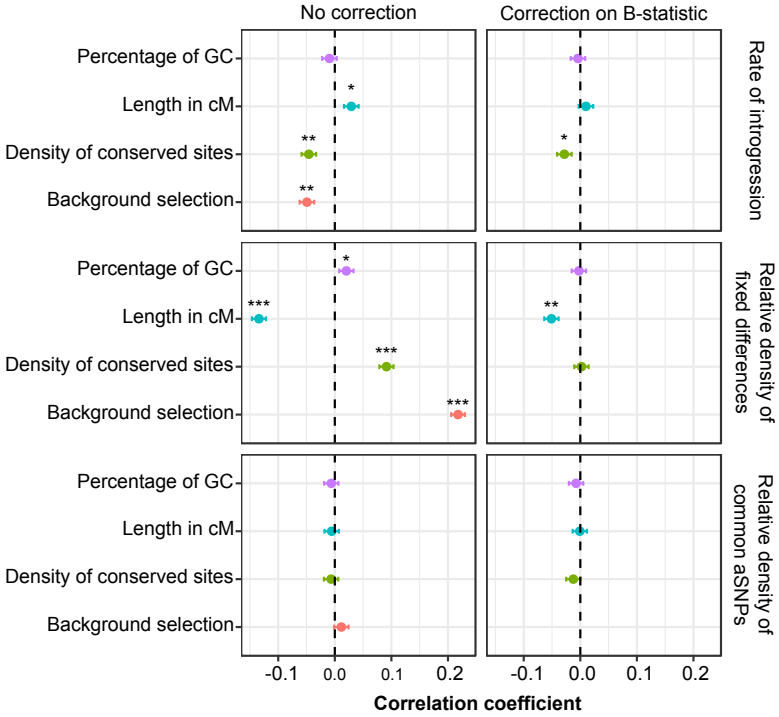


Figure S5. Effects of neutral and selective factors on rate of introgression and relative density of fixed human-Neanderthal differences, conditional on background selection. Correlations, computed in 100kb windows along the genome, between the rate of introgression, the relative density of fixed human-Neanderthal differences and common aSNPs, and several proxies of neutral and selective factors. **p*-value < 10^{-2}, ***p*-value < 10^{-10}, ****p*-value < 10^{-20}. For each correlation, horizontal lines indicate 95% confidence interval.

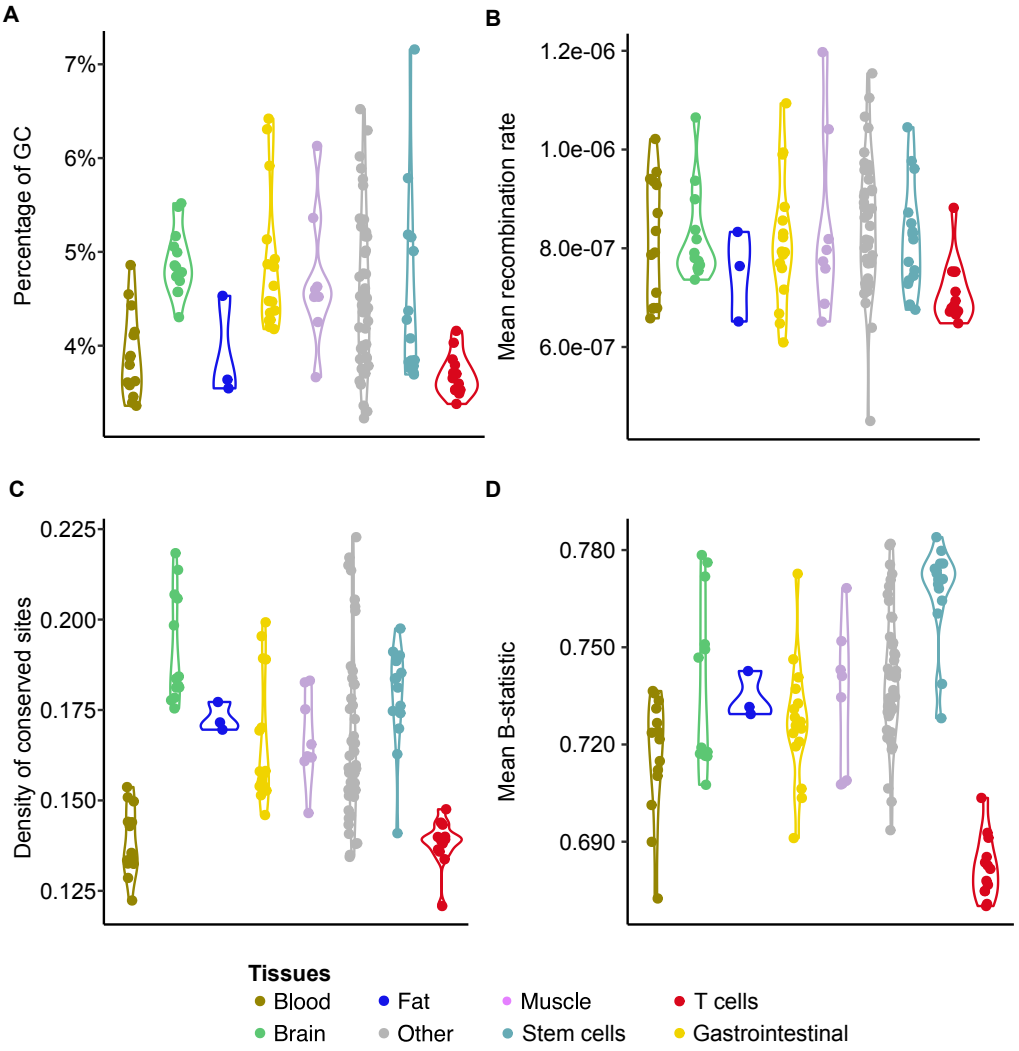


Figure S6. Intensity of neutral and selective factors at enhancers across tissues. Values, in the enhancers of the 127 tissues studied, of the percentage of GC, the mean recombination rate, the density of conserved sites (GerpRS > 2), and the mean B-statistic.

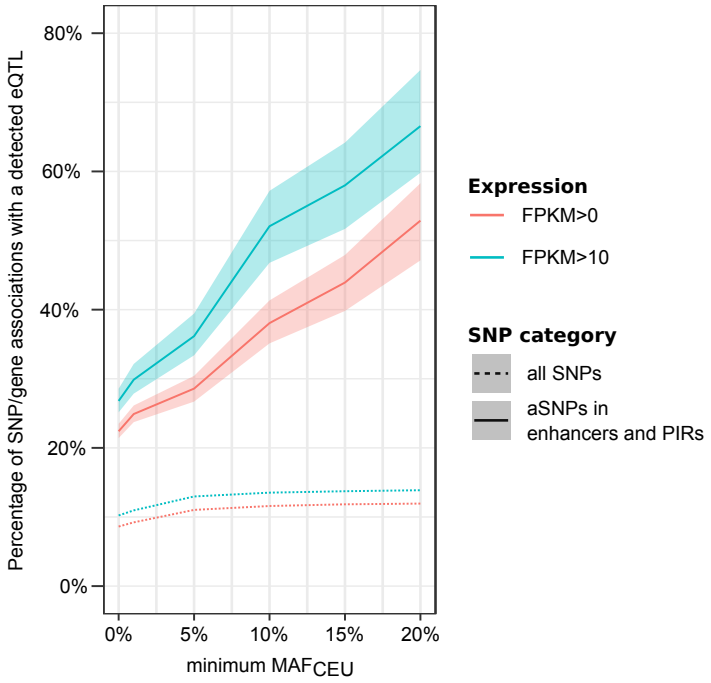


Figure S7. Effects of enhancer variants on gene expression. Comparison of the proportion of SNP/gene pairs with an eQTL in the eQTLGen Consortium data,[1] as a function of the frequency of the SNP in the CEU population, and the expression of the gene in whole blood[2] for two classes of SNPs: all SNPs tested in eQTLGen dataset (dotted line), and aSNPs that are in T cell enhancers that interact with the gene promoter (i.e. promoter interacting region – PIR) based on T cell contact maps[3] (plain line). Shaded regions indicate 95% confidence intervals computed by bootstrap.

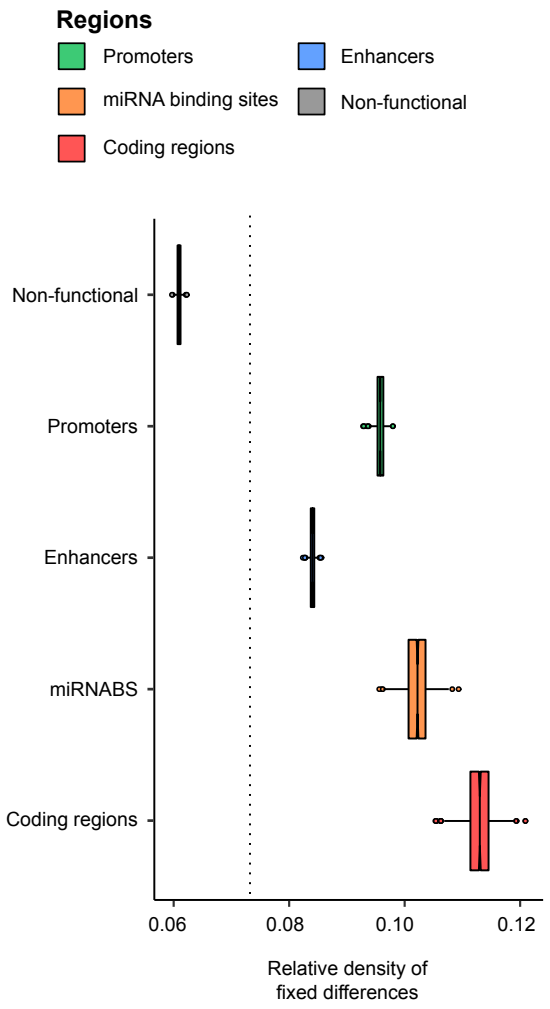


Figure S8. Relative density of fixed human-Neanderthal differences across genomic regions. For each type of region, boxplots show the distribution of the estimates of the relative density of fixed human-Neanderthal differences obtained across 1,000 bootstrap resamples of 100kb genomic windows.

Supplemental Methods

Definition of archaic SNPs (aSNPs)

We considered all SNPs present in the European (CEU) and Asian (CHB) populations of the 1000 Genomes Consortium phase 3 (ref.[4]). Among them, aSNPs were defined as SNPs that (i) have an allele for which the Neanderthal Altai is homozygous,[5] (ii) are absent from the African Yoruba population, and (iii) are located in a region in which Neanderthal introgression has already been detected in Eurasia (probability of Neanderthal introgression > 0.9) (ref.[6]). To distinguish alleles that originated in the Neanderthal lineage from loci where an ancestral allele was re-introduced by Neanderthal into the modern human lineage, we inferred ancestral/derived states based on the 6EPO ancestor sequence. aSNPs where the Neanderthal allele matches the derived allele were considered as derived-aSNPs, while aSNPs where the Neanderthal allele matches the ancestral state were classified as ancestral-aSNPs. Because variants due to incomplete lineage sorting are more likely to segregate at high frequency, and to minimize false positives among signals of adaptive introgression, we took additional steps to filter out such variants when considering aSNPs at high frequencies (**Figure 5** and **Tables S5** and **S6**). Specifically, we retrieved for each aSNP the set of all aSNPs that are in high linkage disequilibrium ($r^2 > 0.8$) in either CEU or CHB. We then required variants to have at least one linked aSNP at a distance of >10 kb, thus filtering likely cases of incomplete lineage sorting.

Relative density of aSNPs and enrichments

To measure the impact of Neanderthal introgression on a specific region, or set of regions, we measured the density of Neanderthal variants, as the number of aSNPs in the region, divided by the length (in bp) of the study region. Likewise, the density of non-archaic variants was computed as a measure of the overall diversity of the region. We then measured the excess or depletion of archaic variants in a region by computing the ratios of these densities (i.e. relative density of aSNPs) in the region, which were compared with those of the rest of the genome. In doing so, we obtained an odds ratio that is significantly higher than 1 if the region presents an excess of aSNPs, and significantly lower than 1 if the region is depleted in aSNPs. We also used this statistic considering only aSNPs and SNPs within a given range of frequencies f, based either on MAF, when considering all aSNPs (MAF<1%, 21% of aSNPs; 1% ≤ MAF ≤ 5%, 48% of aSNPs; or MAF > 5%, 31% of aSNPs*)* or DAF, when considering derived and ancestral alleles separately (DAF<1%, 19% of aSNPs; 1% ≤ DAF ≤ 5%, 44% of aSNPs; or 5% < DAF < 50%, 28% of aSNPs for derived alleles; DAF>95%, 6% of aSNPs for ancestral alleles).

$$\frac{\#aSNPs^{f}_{In\ Region} \Big/ \#non\ aSNPs^{f}_{in\ Region}}{\#aSNPs^{f}_{Outside\ Region} \Big/ \#non\ aSNPs^{f}_{Oustide\ Region}}$$

To compute the significance of the odds ratio, while considering both the haplotype structure of Neanderthal variants and the local structure of the study regions, we divided the genome into windows of 100kb and performed 10,000 bootstrap resamples of these windows, recomputing the odds ratio for each bootstrap sample. We then computed enrichment/depletion *p*-values as the percentage of bootstrap resamples where the odds ratio is lower/higher than 1. Bidirectional *p*-values were then obtained as $2\times\min(p_{\text{enrichment}}, p_{\text{depletion}})$

Definition of regulatory regions

Human miRNA sequences and their locations were obtained from the miRbase database, version 20 (ref.[7]). We used the miRanda software[8] version 3.3a, to predict miRNA binding sites in the 3'UTR of coding genes, as defined in Ensembl Annotation GRcH37.70. Defaults cutoffs were used. Promoters and enhancers were defined based on chromatin marks in the 127 tissues of the Roadmap Epigenomics Consortium.[9] The calling of promoters and enhancers was performed based on 15-state ChromHMM.[10] We considered the union of the *Active TSS* and *Flanking* TSS as "promoters", and the union of the *Enh* (enhancers) and *EnhG* (enhancers genic) categories as "enhancers".

Characterizing the impact of introgression on regulatory regions

To dissect the relative contribution of Human-Neanderthal divergence, and post-admixture removal of Neanderthal introgressed variants in shaping the current landscape of introgressed regulatory variants, we first searched for fixed differences between the genomes of Neanderthals and modern humans. Namely, we considered as a fixed difference any variant (i) where both Neanderthal Altai[5] and Neanderthal Vindija[11] were homozygous for an allele, (ii) absent in 6EPO ancestor sequence and (iii) absent in the Yoruba population.[4] We then defined the density of fixed differences in a region as the number of fixed differences over the number of sites in that region, where sequence information was available for Altai, Vindija and 6EPO genomes. This density was further divided by the density of common variants in the region to yield a 'relative density of fixed differences', which measures the excess of divergence in a study region given its overall diversity. Reciprocally, we considered as the rate of introgression, the percentage of fixed differences

that were introgressed into modern humans and reach a MAF of at least 5%. With these definitions, the product of the rate of introgression and the relative density of fixed differences is equal to the relative density of common aSNPs in the region.

Impact of neutral and selective factors on introgression-related metrics

We investigated the effects of mutation, recombination, and negative selection (directly or indirectly through background selection) on various introgression-related metrics, including the rate of introgression, the relative density of fixed differences, as well as the density of archaic variants segregating in CEU and CHB populations. To do so, we split the human genome into 100kb windows, and focused on sites where sequence information was available for Altai, Vindija and 6EPO genomes, excluding windows where sequence information was available for less than 50% of the window. We then computed, for each window, the percentage of GC or CG dinucleotide in the sequence, the mean recombination rate, the proportion of conserved sites (GerpRS > 2) and the mean B-statistic. For each of these metrics, the Pearson correlations with each introgression-related metric were computed across all windows.

Next, similarly to what we performed genome-wide, we subdivided the genome in 100kb windows and, for each tissue, we computed, at windows containing enhancers, the total enhancer's length and the percentage of GC, mean recombination rate, percentage of conserved sites (GerpRS >2) and mean B-statistic in the corresponding enhancers. For each tissue, we then assembled enhancers from randomly sampled windows and tissues to create a pseudo-tissue, for which we can compute the relative density of fixed differences, the rate of introgression and relative density of common aSNPs. To ensure that the reconstructed tissues had an enhancer structure that is comparable to the original tissue, each resampled pair (window and tissue) was selected so that the length of their enhancers matched that of the enhancers from the original tissue.

To evaluate the contribution of neutral and selective forces to the relative density of fixed differences and rate of introgression, we performed additional resamples matching enhancers simultaneously for their percentage of GC, mean recombination rate, percentage of conserved sites and mean B-statistic, in addition to their length. For each tested tissue and matching, a total of 1,000 resamplings was performed and a *p*-value was computed as the number of resamplings for which the relative density of fixed differences or rate of introgression at enhancers of the tested tissue exceeded that of enhancers in the reconstructed tissue. When resampling, we used the following bins for matching: (i) total enhancer length: 20 bins defined as follows [0-200 bp],]200-400 bp],]400-600 bp],]600-800 bp],]800 bp-1kb],]1-1.5 kb],]1.5-2 kb],]2-3 kb],]3-4 kb],]4-5 kb],]5-7.5 kb],]7.5-10 kb],]10-20 kb],]20-30 kb],]30-40 kb],]40-50 kb],]50-75 kb],]75-100 kb], and]100-200 kb], (ii)

percentage of GC and percentage of sites with GerpRS > 2: 20 uniformly distributed bins of 5% width, (iii) B-statistic: 10 uniform bins of width 0.1, and (iv) mean recombination rate: 10 bins, based on deciles.

Identification of enhancer-interacting genes

To assign genes to the enhancers detected that are active in AdMSC, we used promoter-capture HiC (PC-HiC) data obtained from adipose tissue,[12] and assigned each promoter to a gene when it is located within 100 bp of its TSS. We then selected all interactions with a CHiCAGO score above 5, where the promoter-interacting region overlapped an enhancer in AdMSC, and assigned the corresponding genes as targets of the enhancer. For primary T cells, we used PC-HiC data obtained from Javierre *et al*.[3] We selected interactions with CHiCAGO score above 5 in the total $CD8^+$ T cells, as promoter interacting regions in this cell type showed the strongest overlap with core T cell enhancers (Jaccard Index = 9.7%).

GO Enrichments

To assess whether specific biological functions had been preferentially affected by archaic introgression at enhancers, we considered both tissues where PC-HiC was available, and assigned each enhancer to a gene based on promoter interactions. As enhancers can control multiple genes (22% of core T Cell enhancers are associated to more that 5 genes, with up to 73 associated genes for the same enhancer), and genes that share a common biological function tend to be found in clusters along the genome, we filtered out enhancers with more than 3 target genes from our enrichment analysis, thus reducing the risk of spurious enrichments due to clusters of co-regulated genes. We then used the GOseq package[13] to search for biological functions overrepresented among genes with aSNPs in their enhancers, using the set of all genes with a SNP in their enhancers as background and adjusting on total enhancer length of each gene.

Supplemental Note 1: Effect of aSNPs in enhancers on gene expression

To assess the impact on gene expression of aSNPs that overlap enhancer regions, we first considered, for each gene, the set of aSNPs that overlap promoter-interacting enhancers in primary T-cells (focusing on core T cell enhancers). We then assessed the frequency at which such aSNPs were associated with changes in gene expression, based on GTEx eQTLs and whole blood eQTLs identified by the eQTLGen consortium.[1,2] We found that while only ~1% of aSNPs that overlap core T cell enhancers regulate their associated gene in GTEx tissues (FDR <5%), this figure reaches 22% when considering eQTLs obtained through metanalysis of whole blood samples from over 30,000 donors.[1] This suggests that while enhancer-overlapping aSNPs contribute to gene expression variability, large sample sizes are required to assess their true effects. Consistent with this notion, we observed that the proportion of enhancer aSNPs that control the expression of their associated genes increases with median gene expression and allele frequency (**Figure S7**), reaching 67% for genes with FPKM>10 and aSNPs with a MAF >20% in Europe. Our data suggests that while >60% of enhancer-overlapping aSNPs are significantly associated with gene expression variation, many of these associations are usually missed by eQTL studies due to low power or under-representation of individuals of non-European ancestry.

Supplemental References

1. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv, doi.org/10.1101/447367
2. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.
3. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384.
4. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68-74.
5. Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43-49.
6. Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354-357.
7. Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., Huang, W.C., Sun, T.H., Tu, S.J., Lee, W.H., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res *46*, D296-D302.
8. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol *5*, R1.
9. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.
10. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc *12*, 2478-2492.
11. Prufer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlevic, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyregne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655-658.
12. Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nat Commun *9*, 1512.
13. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol *11*, R14.

5.3 Résumé des résultats

Nous avons conçu une statistique, la densité relative de variants néandertaliens permettant de comparer la participation des variants néandertaliens à la diversité de différents types de régions. Ainsi, nous observons que les régions régulatrices en général, ont un excès de variants néandertaliennes communs (MAF $> 5\%$). Cependant cet effet est majoritairement porté par les enhancers, qui représentent une grande partie du génome.

Si aucun enrichissement n'est visible dans les sites de fixation des miARN, nous avons identifié plusieurs exemples de variants néandertaliens affectant la régulation via miARN, soit en perturbant un site de fixation, soit en changeant directement le miARN exprimé.

Nous nous sommes également intéressés aux variations inter tissus, et n'avons trouvé aucun enrichissement de variants néandertaliens dans les promoters des 127 tissus étudiés. En revanche, un tiers des tissus testés ont un excès de variants néandertaliens dans leurs enhancers. De nombreux types de tissus sont concernés, notamment plusieurs tissus gastro-intestinaux, les lymphocytes T, et les cellules souches dérivées de cellules adipeuses (CSdCA).

Nous avons par la suite étudié les raisons évolutives de la présence excessive des variants néandertaliens dans les enhancers de différents tissus, en nous concentrant sur les lymphocytes T et les CSdCA. Nous soulignons deux raisons évolutives différentes pouvant mener à un même enrichissement, soit les haplotypes néandertaliens sont plus fréquents dans les enhancers du tissu étudié : ce qui est le cas observé dans les CSdCA, soit les haplotypes présents comportaient plus de divergences entre les HAM et les néandertaliens qu'attendu, ce qui est le cas présent dans les lymphocytes T.

5.4 Discussion

La nouvelle métrique que nous avons introduit dans cette étude, qui mesure la densité des variants néandertaliens, relativement à la diversité globale de la région, diffère sensiblement des mesures précédemment utilisées pour quantifier l'impact de l'introgression néandertalienne sur l'expression génique. Auparavant, plusieurs études s'étaient basées sur des méthodes de ré-échantillonage de variants non-néandertaliens, pour comparer leurs effets à ceux des variants introduits lors de l'introgression néandertalienne (Quach et al., 2016 ; Dannemann et al., 2017). De plus ces études se sont principalement intéressées aux variants ayant un fort effet sur l'expression (eQTL). Ce choix de focaliser sur les eQTL biaise les résultats vers des effets trouvés au sein des promoters puisque ces régions sont les plus à même d'avoir des effets forts sur l'expression. De plus ces études sont sensibles au déséquilibre de liaison et ne peuvent pas différencier les effets d'un variant néandertalien des effets de variants qui ne sont pas spécifiques de la population néandertalienne, mais qui sont présents uniquement sur les haplotypes introgressés dans les populations étudiées. Ces deux points participent très certainement à la différence de résultats observée. Ainsi si les haplotypes néandertaliens influencent grandement l'expression dans les monocytes (Quach et al., 2016), les variants néandertaliens ne sont pas enrichis dans les promoters et les enhancers de ces derniers. La distinction entre les effets des mutations néandertaliennes et les effets des ha-

plotypes néandertaliens n'a pas encore été beaucoup explorée, mais constitue une voie de recherche fascinante (Rinker et al., 2019).

Un développement supplémentaire pourrait être trouvé dans la distinction entre mutations néandertaliennes dérivées, c'est à dire le cas où l'allèle dérivé est apparu dans la lignée néandertalienne, et mutations néandertaliennes ancestrales, c'est à dire dans le cas ou l'allèle dérivé a atteint fixation dans les populations humaines ancestrales. Nous trouvons que ces dernières semblent avoir un fort impact fonctionnel, soulignant que les variants néandertaliens pourraient être classifiés en deux catégories distinctes, en fonction de leur histoire évolutive.

Une autre question soulevée par notre étude est celle de la comparabilité des régions fonctionnelles entre espèces. En effet, nous avons étudié ici les régions qui sont des enhancers chez l'humain moderne, mais nous n'avons pas d'assurance que ces dernières étaient également des enhancers chez les néandertaliens. Si cela n'a pas d'importance lorsque l'on cherche à évaluer les conséquences de l'introgression chez l'humain moderne, cela peut impacter la manière dont on interprète l'histoire évolutive. Prenons l'exemple du cas de l'excès de divergence chez la population néandertalienne dans les enhancers modernes des lymphocytes T, Nous avons montré que, *par rapport aux enhancers modernes des autres tissus étudiés* et compte tenu des forces évolutives et génomiques qui agissent sur ces régions, les néandertaliens ont accumulé, dans les régions qui sont les enhancers modernes des lymphocytes T, un nombre inattendu de divergences. Ainsi, dans l'hypothèse ou les enhancers des lymphocytes T évolueraient plus rapidement que ceux d'autres tissus chez les primates, et qu'une grande part de ces derniers aient été perdus ou gagnés entre les néandertaliens et les HAM, il est possible que le résultat observé provienne du fait que les régions étudiées n'étaient majoritairement pas fonctionnelles chez les néandertaliens, permettant donc une plus grande divergence. Bien que l'évolution des régions régulatrices entre mammifères ait été étudiée dans le foie (Villar et al., 2015), une étude de la vitesse relative d'évolution des enhancers en fonction des tissus dans lesquels ils sont actifs serait à présent nécessaire pour confirmer cette hypothèse.

6. Résultats II : variabilité des miARN entre populations dans un contexte immunitaire

6.1 Contexte

La réaction immunitaire est un phénotype variable simultanément au sein de la population et au niveau de l'individu. Et si beaucoup d'aspects environnementaux, tels que les précédentes infections par des pathogènes, influencent cette réponse immunitaire, la génétique des individus affecte également cette réponse, notamment dans le cadre de l'immunité innée.

Ainsi, l'étude de la variabilité de l'expression des miARN dans le contexte de l'immunité permet non seulement d'étudier la diversité présentée par ces petits ARN, mais également d'identifier quels sont les gènes dont l'expression ou la traduction est impactée par ce mécanisme lors d'une stimulation du système immunitaire inné. Et si l'expression des miARN dans ce contexte a déjà fait l'objet d'études (Siddle et al., 2014, 2015), la variabilité exacte, à la fois inter et intra population reste peu définie, à la fois en termes d'abondance de chaque miARN, mais également en termes d'abondance de chacun de leurs isomiR.

Pour remédier à cela, nous utilisons les résultats de séquençage de petits ARN dans les monocytes provenant de 100 personnes d'ascendance africaine et 100 personne d'ascendance européenne, suite à l'activation de trois voies TLR (TLR4, TLR1/2, TLR7/8), des capteurs essentiels de l'immunité innée, ou suite à une infection grippale. Nous utilisons également les données d'expression et de transcription de plus de 10 000 gènes des mêmes échantillons afin d'étudier l'impact relatif de la transcription et de la dégradation médiée par les miARN sur l'expression finale des gènes.

6.2 Article

The contribution of miRNA regulation to inter-individual and inter-population variability in immune responses

Maxime Rotival[1], Martin Silvert[1,2], Katherine J Siddle[3,4], Julien Pothlichet[1], Helene Quach[1], Lluis Quintana-Murci[1]

[1]Human Evolutionary Genetics Unit, Institut Pasteur, CNRS UMR 2000, 75015 Paris, France

[2]Sorbonne Universités, École Doctorale Complexité du Vivant, 75005 Paris, France

[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

Abstract

While the regulatory role of microRNAs (miRNAs) in the immune response is increasingly recognized, their contribution to the intra- and inter-population differences of immune responses, both in terms of abundance and isoform diversity (isomiRs), is poorly characterized. Here, we generated RNA-sequencing data of small RNAs from monocytes, originating from 200 individuals of African and European ancestry, following activation of three major TLR pathways (TLR4, TLR1/2 and TLR7/8) or infection with Influenza A virus (IAV). We show that monocytes display a strong shift in their miRNA profiles upon immune activation, with more than a third of miRNAs (233 miRNAs) being up-regulated. Furthermore, in response to viral stimuli, we observed a strong reduction in 3' uridylation that leads to shorter isomiRs. From a population perspective, individuals of African and European ancestry show marked differences in miRNA expression patterns, particular for miRNAs that respond to TLR activations. Among these miRNAs, we find key modulators of the immune response, such as miR-155, miR-146a or miR-222. When mapping the genetic basis of miRNA expression variability, we find that miR-QTLs are largely shared across conditions, and account for up to 60% of population differences in expression of the miRNAs they regulate. Finally, integrating miRNA and mRNA data from the same individuals, we show that miRNA-mediated degradation accounts for <0.2% of the inter-individual variation in gene expression, suggesting that their consequences occur primarily at the protein level. Overall, this study sheds new light onto the factors driving population differences in miRNA abundance and isomiR ratios, and how in turn these differences may ultimately affect immune response variation, both between individuals and populations.

Introduction

Since their discovery in 1993[1], microRNAs (miRNAs) — short, evolutionary conserved RNA sequences of ~22 nucleotides — have emerged as key epigenetic regulators, involved in a large variety of developmental and cellular processes such as cell differentiation, proliferation and homeostasis [2]. There is also increasing evidence that support their key role in immune responses [3-8], with miRNAs such as miR-146b or miR-155 acting to promote and stabilize the inflammatory response. Furthermore, multiple studies have reported strong shifts in miRNA profiles in response to infectious agents, for example, upon infection with *Mycobacterium tuberculosis* [9, 10], *Salmonella* [11] or Influenza A virus [12].

Studies of miRNA abundance across various cell types and tissues have allowed to characterize the extent of genetic regulation of miRNA expression variability, i.e., miRNA expression quantitative trait loci (miR-QTLs), and highlighted the role of genetic variants located in the promoters of pri-miRNAs in shaping inter-individual differences in miRNA expression [9, 13-20]. In the context of immunity, despite increasing evidence for extreme population variability of mRNA expression in response to immune challenges [21, 22], the extent to which the miRNA response to infection varies across individuals of different ancestry remains largely unknown.

Fuelled by the advent of deep sequencing technologies, growing evidence has emerged that mature miRNAs undergo important post-transcriptional modifications [23-27]. These include nucleotide substitutions (miRNA editing) [28, 29], 3' adenylation or urydilation by terminal nucleotidyl transferases (TENT) [30, 31], shortening of their 3' end by poly(A)-specific ribonuclease (PARN) [32], and, more rarely, shifts in their 5' start sites[27]. The diversity of miRNA isoforms (known as isomiRs) was initially proposed to increase the robustness of miRNA-mediated regulation, by fine-tuning the binding of miRNAs to their target-sites [33]. However, there is now growing support to the idea that miRNA

modifications may act as a conserved, additional layer of regulation of their activity [34], as illustrated by the case of miR-222. Upon stimulation with interferon or *Salmonella*, shortening of the 3' end of miR-222 occurs, and leads to a decreased apoptotic action of the miRNA, while maintaining an anti-proliferative effect through the binding of its canonical targets [35, 36]. However, our understanding of the variability of isomiR expression across individuals and populations remains largely incomplete.

Regulation of gene expression by miRNAs, following the canonical model, is achieved through the recognition of conserved target sites, which are mostly located in the 3' UTR of protein-coding genes [37-41]. This binding typically results in the repression of target expression, by inducing mRNA deadenylation and degradation, or by inhibiting translation [42]. Furthermore, a strong body of evidence highlights the importance of strong sequence complementarity between the seed region of the miRNA — located at nucleotides 2-7 of the mature miRNA [37, 43] — and its target site in determining miRNA-binding. Nonetheless, identifying which mRNAs are actively targeted by a given miRNA remains challenging [44, 45]. Previous studies of the regulatory impact of miRNAs on gene expression have reported conflicting results [9, 15, 18, 46], possibly due to difficulties in disentangling the direct effects of miRNAs on mRNA degradation from co-transcription between miRNAs and their targets. Because RNA-seq can capture both (i) steady state gene expression levels via the analysis of exonic reads and (ii) dynamic rate of transcription through the quantification of intronic reads [47], it offers a unique opportunity to determine the relative contribution of transcription and post-transcriptional regulation by miRNAs to inter-individual variability in gene expression.

Here, we sequenced genome-wide small RNAs from primary monocytes originating from 200 individuals of African and European ancestry (100 individuals from each background), at the basal state or upon stimulation with three TLR ligands (LPS for TLR4,

Pam3CSK4 for TLR1/2 and R848 for TLR7/8), or Influenza A virus (H1N1) (**Figure 1**). In doing so, we characterized the miRNA landscape, both in their abundance and isomiR diversity, in response to immune stimuli. Using whole-genome genotyping and exome sequencing data, we then assessed the genetic bases of inter-individual and inter-population differences in miRNA expression, both at basal state and in response to stimulation. Finally, by integrating these data with RNA-seq data from protein-coding genes from the same individuals, we quantify the relative impact of transcription and miRNAs on gene expression variability.

Results

Exploring the landscape of miRNA expression in human monocytes

We generated 977 small RNA sequencing profiles, in resting and activated cells, from 200 healthy individuals of African and European-descent. Activation was performed for 6h with three different Toll-like receptor (TLR) ligands (LPS, Pam_3CSK_4 and R848 activating TLR4, TLR1/2 and TLR7/8 pathways, respectively), as well as with a live strain of Influenza A Virus (IAV). Small RNAs were then separated from mRNAs, and sequenced at a mean depth of 12.4 million reads per samples (**Methods and Supplementary Figure 1A-C**). After excluding reads outside the 18-26 nucleotides range and low-quality samples (**Supplementary Figure 1C**), we obtained an average of 5 million reads aligned to miRNAs. To correct for cross-mapping artefacts between similar miRNAs, multiply-mapped reads were assigned to each possible locus using an Expectation-Maximization strategy [29]. Library size was normalized across samples, and miRNAs with an average of <1 read per million (RPM) were discarded. This yielded a final set of 675 loci, encoding for 658 distinct miRNA products (**Supplementary Table 1A**).

Focusing on unique sequences, we identified a total of 23,447 putative isomiRs, the vast majority (90%) of which were either low abundant (<1 Read Per Million [RPM]; 14,277 isomiRs) or extremely rare (<1% of the reads; 6,811 isomiRs). Focusing on the remaining 2,359 unique miRNA sequences (corresponding to a total of 492 loci encoding 451 distinct miRNAs, **Supplementary Table 1B**), we found that 86% of miRNAs expressed one or more isomiR(s), beside their canonical form, with a single miRNA expressing up to 8 frequent isomiRs (>5% of the reads) (**Figure 2A, Supplementary Figure 2A,B**). Interestingly, for more than 65% of miRNAs, the canonical isomiR, as defined in miRbase, accounted for less than half of the copies of the miRNA (**Figure 2B**). Among the 138 miRNAs where the canonical isomiR is in minority, only a third had a seed – defined as the sequence located at

position 2-7 from the 5' end of the miRNA – that differed from that of the canonical isomiR in more than 20% of their copies (**Figure 2C**).

To dissect the processes leading to such a high isomiR diversity, we next quantified each type of miRNA modification separately (i.e. shifts in start/end site, non-template additions [NTA] and substitutions). Consistently with previous results[10], we found that shifts in 3' end site of miRNAs were the most frequent type of modification, with 79% of miRNAs presenting a shift of their 3' end site in over 5% of the reads, even after exclusion of non-template additions (**Supplementary Figure 2C,D**, 87% including NTA). Conversely, less than 32% of miRNAs presented a frequent shift of their 5' start site (>5% of the reads, **Supplementary figure 2E**), consistent with strong constraints on the miRNA seed across isomiRs. Focusing on substitutions, we found a strong enrichment of nucleotide substitutions at 3'end of the miRNA (binomial $p < 3.7\times10^{-38}$**, Figure 2D**), recapitulating known patterns of 3' terminal uridylation and adenylation, but also a strong enrichment of substitutions at the 5' end of the miRNA, as well as the seed-altering positions 2 and 4 of the miRNA (binomial $p < 2.2\times10^{-5}$). All three positions (i.e. nucleotides 1, 2 and 4) presented a strong bias (binomial $p < 9.8\times10^{-31}$) toward G->U substitutions, as well as low frequency U->G changes at positions 1 and 4 (binomial $p < 0.003$). While the frequency of terminal substitutions was stable across all sequencing batches ($R^2 < 0.1\%$), substitutions at positions 1, 2 and 4 were strongly dependent on the sequencing lane ($R^2 > 48\%$), suggesting a technical bias rather than a biological mechanism. These substitutions were thus ignored and isomiR abundances were recomputed taking only into account shifts in the start or end of the miRNA, as well as terminal uridylation and adenylation. After removing miRNAs with a single isomiR, the final dataset consisted of 2,049 common isomiRs, measured across 435 miRNAs.

We compared the frequency of miRNA modifications across both arms of the pre-miRNA hairpin (**Supplementary Table 1C**). We observed a stronger degree of 3' terminal uridylation at -3p miRNAs (+12% of uridylated miRNAs on 3p arm compared to 5p; Wilcoxon $p < 1.7\times10^{-8}$, **Supplementary Figure 2F**), consistent with the reported role of uridyl transferases in pre-miRNA maturation[48-50]. This increased uridylation was not associated to a higher rate of 3' extensions among miRNAs located on the 3p arm (p=0.65), due to a higher rate of template extensions among 5p miRNAs (+12% on 5p arm compared to 3p; Wilcoxon $p < 0.005$, **Supplementary Figure 2G,H**). Finally, we also detected a higher usage of non-canonical, downstream start sites among 3p miRNAs (+3% compared to 5p miRNAs; Wilcoxon $p < 0.003$), consistent with a regulation of isomiR variability through the tuning of DICER positioning on the pre-miRNA [51]. Together, these results reveal a wide variety of isomiRs, and highlight the complexity of the landscape of miRNA modifications in primary human monocytes.

Variability of miRNA expression upon innate immunity activation

After adjusting miRNA and isomiR expression for batch effects (date of experiment, date of library preparation and sequencing lane) and technical confounders (GC content and mean read length), principal component (PC) analysis of miRNA abundances revealed a clear separation by stimulation conditions (**Figure 3A**). PC1, which accounted for by 11.9% of variance, opposed TLR-activated from IAV-infected samples, while PC2, which explained 4.8% of the variance, captured the shared effect of all 4 immune stimuli on gene expression. Interestingly, significant shifts between populations were also visible on both PCs (PC1, $p < 1.0\times10^{-79}$; PC2, $p < 1.2\times10^{-11}$), suggesting differences in the intensity of immune response between African- and European-ancestry individuals.

At FDR < 1%, we identified 340 miRNAs that presented differential expression upon stimulation (DE miRNAs, 30 with $\log_2$FC > 1), 233 of which were up-regulated in at least one condition (58-74% per condition; **Figure 3B** and **Supplementary Table 2A**). Using a Likelihood-based Model Selection framework [52] (**Figure 3C**), we estimated that 90% of DE miRNAs respond in a stimulus-dependant manner. The most frequent patterns of response were (i) a TLR-specific response (N = 65, 19% of DE miRNAs), as in the case of the NF-κB inhibitors miR-9-5p (**Figure 3D**) [53] and miR-155-5p [5-7], (ii) a viral-stimuli specific response (R848 and IAV, N = 55, 16% of DE miRNAs), such as miR-3614-5p recently involved in Crohn's disease susceptibility [54] (**Figure 3E**), and (iii) an IAV-specific response (N = 78, 23% of DE miRNAs), as observed for the pro-inflammatory mir-429[55], or the TRIM22 repressor mir-215-5p[56] (**Figure 3F**).

Focussing on isomiR ratios, IAV infection clearly had the strongest impact (PC1, 4.7% of variance explained), followed by TLR7/8 activation (PC2, 2.6% of variance explained), with LPS and PAM3CSK4 showing a limited impact on isomiRs (**Supplementary Figure 3A**). We identified 316 miRNAs that changed their isomiR ratios upon stimulation (**Supplementary Table 2B)**. Among these, 212 (67%) changed their ratio of canonical isomiR upon stimulation, with this ratio decreasing in 56% to 70% of cases (**Figure 3G**). Interestingly, most miRNAs only displayed moderate shifts in isomiR ratios, with changes in isomiR ratio exceeding 5% in only 5-11% of miRNAs (LPS and IAV, respectively; **Supplementary Figure 3B**). Notable exceptions included mir-155-5p, shifting toward extended 3' isomiRs upon stimulation (13-33% increase in longer isomiRs, $p < 2.7\times10^{-68}$), and miR-429, showing an IAV-specific shift toward a 3'-shortened isomiR (42% increase in shorter isomiR upon infection, $p < 7.5\times10^{-93}$, **Figure 3H**). Overall, changes in isomiRs were the most frequent following treatment with viral ligands (R848 and IAV), with 36% of isomiRs changes being shared between these stimuli (**Supplementary Figure 3C**).

We next investigated whether this observation could be explained by changes in frequency of specific types of miRNA modifications (**Supplementary Table 2C**). We found a significant trend toward reduction of miRNAs at their 3'end site for both R848 and IAV conditions ($p < 4.6\times10^{-14}$, **Supplementary Figure 3D**). This trend was observed for miRNAs located on both arms of their pri-miRNA hairpins, but was largely restricted to the IAV condition among -5p miRNAs (p-value = 0.014 and *p*-value $< 8.1\times10^{-24}$ for R848 and IAV respectively, **Supplementary Figure 3D)**, suggesting a combination of multiple effects rather than a single mechanism. Furthermore, while non-template additions accounted for only minor fraction (~6%) of the observed shift in miRNA end site, we observed a significant shift from uridylation to adenylation at the 3' end of miRNAs that appeared slightly more pronounced upon R848 stimulation (**Supplementary Figure 3E-F).** Altogether, these results highlight significant shifts in miRNAs expression upon immune stimulation, and indicate a high rate of isomiR modifications in the response to viral stimuli.

Genetic basis of inter-individual variation in miRNA response to stimulation

To assess the contribution of genetics to miRNA expression variability, we next searched for genetic variants associated with changes in miRNA abundances (miRNA Quantitative Trait Loci, or miR-QTLs) and isomiR ratios (isomiR-QTLs). At 5% FDR, we identified 122 miRNAs associated with at least one miR-QTL (**Supplementary Table 3A**). miR-QTLs were strongly enriched in a 20kb window around either the transcription start site (TSS) of the pri-miRNA or the pre-mRNA hairpin that contains the mature mRNA (**Figure 4A**). Nevertheless, 46% of miR-QTLs were located >20kb away from the pri-miRNA. Furthermore, we observed that miRNAs with conserved promoters (mean phastCons > 20%) were depleted in miR-QTLs, with respect to miRNAs with less conserved promoters (OR = 0.54, $p < 0.008$), and

their miR-QTLs were located further away from the TSS on average (+ 3.5 kb, Wilcoxon *p*-value < .03, **Supplementary Figure 4**).

We estimated that 85% of all miR-QTLs were shared across conditions of stimulation (**Figure 4B, Supplementary Table 3B**), with only a small minority (N=18) displaying condition-specific effects. Among these, we detected 4 response miR-QTLs, i.e., genetic variants that manifest their effects on miRNA abundance only in the presence of immune stimulation ($p_{\text{Interaction}} < 0.001$, corresponding to 5%FDR). For example, the African-specific rs75335466 has a derived allele (Derived Allele Frequency (DAF) = 7.5% in Africans) that is associated, upon stimulation by LPS and R848, to a reduced expression of the dominant arm of miR-146a (miR-146a-5p, $p_{\text{interaction}} < 5.6\times10^{-4}$, **Figure 4C**), which acts as an inhibitor of TRAF6 and IRAK1[7].

Focusing on isomiRs, we found only 25 isomiRs that were associated with at least one isomiR-QTL, involving 13 miRNAs (**Supplementary Table 3C)**, with 84% of these being shared across conditions. Note that because we ignore non-terminal substitutions in our definition of isomiRs, these numbers do not take into consideration genetic variants that directly alter the miRNA sequence, unless they also alter the start/end site of the miRNA. An interesting case of isomiR-QTL is provided by the rs2910164 variant (DAF: 49% in Africans and 21% in Europeans), which disrupts the seed of the passenger arm of miR-146a (miR-146a-3p, **Figure 4D**). The derived allele of rs2910164 is associated to both an increase in expression of miR-146a-3p ($|\beta_{\text{miR-QTL}}| > 0.31$, $p < 3.1\times10^{-7}$) and a shift of both the start and end sites of the mature miRNA ($|\beta_{\text{isomiR-QTL}}| > 0.15$, $p < 2.1\times10^{-9}$, **Figure 4E**). This shift leads to a complete redefining of the miR-146a-3p targets, with 2,273 predicted targets being lost (73%) and 2,352 novel targets being predicted (**Figure 4F**). These results underline how genetic variants can alter the impact of miRNAs on immune response, by changing either their abundance in response to stimulation or the set of genes that they target.

Sources of population differences in miRNA response to stimulation

We subsequently explored the extent to which miRNA response to stimulation may differ between individuals of African and European ancestry. We identified a total of 351 miRNAs whose transcriptional profiles differed between populations, either in abundance (pop-DE-miR, N=244, including 141 with $|\log_2 FC|>.2$, **Figure 5A** and **Supplementary Table 4A**), or in isomiR ratios (pop-DE-isomiR, N=188, including 148 with $\Delta_{isomiR\text{-}ratio} > 1\%$, **Figure 5B** and **Supplementary Table 4B**), with 81 miRNAs differing in both expression and isomiRs. Such population differences were largely shared across conditions, with 67% of popDE-miR and 61% of pop-DE-isomiRs being shared across all conditions (**Figure 5C**). Yet, we identified 8 miRNAs that displayed population differences only upon stimulation (**Supplementary Table 4C**), including key immune modulators such as the pro-inflammatory miR-155-5p, which showed marked population differences in expression upon TLR1/2 stimulation ($p_{interaction} < 1.0 \times 10^{-9}$, **Figure 5D**). Looking at the rate of miRNA modifications, we found that 3' end shortening of miRNAs located in 3p arm was more frequent in Africans with respect to Europeans ($p < 5.0\times10^{-4}$). In addition, Africans presented, upon stimulation, an increased rate of 3' adenylation compared to Europeans, regardless of the arm of miRNAs are located ($p < 4.5\times10^{-5}$), partially compensating the detected shortening of the 3p arm miRNAs.

More generally, we also found that population differences were more frequent among miRNAs and isomiRs that respond to stimulation ($OR > 1.5$, $p < 9.9\times10^{-4}$, **Figure 5E** and **5F**), even after accounting for the impact of miRNA/isomiR expression levels. Focusing on miRNAs and isomiRs that show the strongest population differences in expression ($\log_2 FC>.2$ and $\Delta_{isomiR\ ratio} > 1\%$) revealed a stronger enrichment among miRNAs that respond to TLR stimulation ($OR > 3.5$, $p < 5.1\times10^{-8}$), consistent with the detection of population differences at modulators of TLR activation such as miR-155-5p, miR-146a-3p and miR-222-5p.

We also observed a significant enrichment of popDE-miR and -isomiR in miRNAs that have a miR- or isomiR-QTL (OR > 1.7, $p < 1.1\times10^{-2}$). We thus computed the fraction of population differences in miRNA expression that is attributable to genetic factors, and estimated that, among the 57 popDE-miR with a miR-QTL, ~60% of population differences could be attributable to genetics. Across all miR-QTLs, the strongest differences in frequency between Africans and Europeans was observed at the rs12881760 on chromosome 14, which is associated with the expression of 12 miRNAs, located in a cluster of 97 miRNAs spanning over 250kb (**Figure 5G**). The derived allele (C) of this SNP, which disrupts a CTCF binding site located ~200kb upstream of the miRNA cluster, is found at high frequency in Europe (73%, **Figure 5H**), while is virtually absent from Africa (frequency ~3%). Interestingly, the C allele is associated to a strong signature of positive selection in Europe (iHS = -3.10, p_{emp}=0.002, 31% of SNPs with |iHS|>99th percentile in a 100 SNP window around the locus, p_{enrich}=0.003, **Figure 5I**), supporting a history of recent adaptation targeting this locus. Overall, these results show that while a substantial fraction of population differences may be due non-genetic factors, genetic differentiation at miR-QTLs has largely contributed to ancestry-related differences in miRNA expression.

Impact of miRNA variation on immune responses

Finally, we sought to quantify the extent to which miRNAs contribute to the regulation of immune-related gene expression. To do so, we used stability selection (see **Methods**) to identify for each gene and condition, the set of miRNAs whose patterns of expression most strongly correlate with gene expression, while adjusting for population. At an 80% probability threshold (corresponding to a 5% FDR in simulated data), we found that 25-48% of genes were significantly associated to at least one miRNA, with a single gene being independently associated to up to 8 different miRNAs per condition (**Figure 6A**). Surprisingly, among the

7,354 miRNA-gene associations detected at the basal state, 47% displayed negative associations, of which 14% presented a known binding site for their associated miRNA. Surprisingly, we found that predicted miRNA targets were depleted in negative correlations with their cognate miRNAs (OR = 0.85, $p < 0.02$). This suggests that, despite the potential caveats related to target prediction, miRNA-driven transcript degradation has only a minor impact on correlations between miRNAs and gene expression.

To test the extent to which such empirical correlations may be driven by co-transcription rather than miRNA-mediated degradation, we next quantified intronic reads that derive from unspliced, nascent transcripts, as a measure of transcription rate. Correlating these to miRNA expression, we identified widespread gene-miRNA correlations, with each miRNA correlating with transcription rate of ~100 genes at the basal state (min:1, max: 2,680) and up to 173 upon stimulation (min:1, max: 4,137). Interestingly, we found 7 miRNAs that are co-transcribed with their target genes, i.e., there is an enrichment among the targets of the miRNAs of genes that are positively correlated at the transcription level (**Figure 6B**). Among these, the regulator of cholesterol homeostasis miR-33a-5p (OR=4.3, Fisher's $p < 1.7\times10^{-4}$) was previously shown to balance the effect of its host gene, the transcription factor *SREBF2*, on fatty acid synthesis/uptake by repressing the cholesterol transporter ABCA1 [57]. Likewise, we identified 7 miRNAs that are negatively correlated to the transcription of their target genes (**Figure 6B**), suggesting that they act to promote rapid expression changes of their targets, through a feed forward loop mechanism. These include key regulators of immune response such as the NF-κB inhibitors miR-9-5p (OR=1.2, $p <5.8\times10^{-4}$) and miR-155-5p (OR=1.3, $p <1.0\times10^{-5}$).

Using a variance partitioning approach, we quantified the amount of inter-individual variation in gene expression that could be attributed to either transcription or miRNA expression [58]. We found that, on average, transcription accounts for 25% of the variance in

gene expression at the basal state, with this amount decreasing upon stimulation (min: R848-21%, max LPS-24%, Wilcoxon $p < 4.1\times10^{-4}$, **Figure 6C-E**). In contrast, miRNAs accounted for only 3.6% of the total variance of expression of their associated genes, and between 2.6% (Pam_3CSK_4) and 6.5% (R848) upon stimulation (**Figure 6C,D** and **6F**). This figures decreased to ~0.2% when focusing only on negative associations, and disregarding miRNAs with no predicted targets for the gene under consideration (**Figure 6C,D**). Overall, these results indicate that while miRNAs have a sizeable impact on gene expression, only a small fraction of this effect can be attributed to direct regulation of miRNA degradation.

Discussion

In this study, we characterized the diversity of miRNA response to immune stimuli in human populations, using 977 small RNA-sequencing profiles obtained from 200 individuals of African and European ancestry. Integrating this data with high-density genotyping and whole-exome sequencing, as well as mRNA-sequencing data from the same individuals, we define the sources of variation in miRNA expression across human populations, and quantify the downstream consequences of these differences on the variability of human immune responses.

Several important insights can be drawn from our study. First, we show that upon immune stimulation or infection the miRNA repertoire is subject to important modifications that are not only quantitative, through the modulation of miRNA expression, but also qualitative, through changes in isomiR proportions [10, 36]. Although isomiR modifications can be confounded by cross-mapping artifacts and sequencing errors[29], we reduced here the impact of such technical biases by focusing on frequent, biologically-plausible modifications, and excluding those that correlate with technical covariates (i.e., substitutions at the 5' end of miRNAs correlate with lane effects). In doing so, we detected systematic shifts in isomiR proportions that occur in a stimuli dependent manner. While most changes in isomiR usage are of modest effect size, it is possible that they anticipate more drastic modifications occurring at later time points, as previously shown in the case of miRNA abundances [10]. We also show that changes in isomiR usage, whether induced by stimulation or associated with genetic variants, can lead to a complete rewiring of miRNA-gene interactions. This is clearly illustrated by the case of miR-146a-3p, where a genetic variant induces a shift in miRNA boundaries and a broad redefinition of miRNA targets. Thus, our work provides further support to the importance of considering isomiR changes when studying the impact of miRNAs on immune response [35, 36].

This study also sheds new light, for the first time, on the contribution of miRNAs to population differences in immune responses. We find widespread differences in the expression of miRNAs between populations, although these were generally of moderate amplitude. The observation that miRNAs that change their expression upon stimulation are more likely to differ between populations at the basal state could indicate either that (i) these miRNAs are more prone to accumulate genetic variation (miR-QTL) that differ in frequency between populations, leading to increased expression divergence or (ii) that populations differ in their basal activation state, suggesting a role of miRNAs as mediators of innate immune memory [59]. Our analyses support the second scenario as the most likely, given that the enrichment of miRNAs that respond to stimulation among those that differ between populations is robust to adjustment on genetics (miR-QTLs). Yet, we find cases where population differences in miRNA or isomiR expression is driven by genetic differentiation between populations. For example, the variant rs290164 presents modest population differences (Delta DAF=28%), but these are sufficient to explain ~76% of population differences in isomiR ratios of miR-146a-3p. Furthermore, we identify a variant (rs12881760) that controls a cluster of 12 miRNAs in *cis,* which is among the top 0.2% most extreme F_{ST} between Africans and Europeans at the genome-wide level, and detect strong enrichment of |iHS| outliers at the locus. Moreover, this variant displays an extreme differentiation between Europeans and East Asians (F_{ST}=0.74, top 0.004% genome wide), which supports the adaptive role of the variant in Europeans. Interestingly, the cluster of miRNAs regulated by the rs12881760 variant has also been reported as a candidate of positive selection in Asians [60], more generally highlighting the adaptive nature of the whole locus among non-Africans.

Finally, our design combining miRNA with protein-coding gene expression data allows us to assess the relative contribution of transcriptional regulation and miRNA-mediated degradation on the variability of the immune response. Intriguingly, while we do find a strong

effect of transcription rate on gene expression, our model predicts that miRNA-mediated degradation accounts for <0.2% of the inter-individual variation in gene expression, suggesting a very limited effect of miRNAs on mRNA stability. Although this is at odds with the canonical model of miRNA activity[42], our model is consistent with previous reports of low levels of miRNA-mRNA correlations, and the frequent occurrence of positive correlations between expression of miRNAs and their predicted targets [9, 15, 18, 61], Furthermore, a recent study has shown that DGCR8 knock-out embryonic stem cells, which are unable to process miRNAs, display an increased translation rate with no change in stability of their mRNAs {Freimer, 2018 #114, 62].

In this context of these observations, our study extends previous findings by providing a model that could explain such positive corrections. Indeed, we observe several cases where miRNA expression is correlated with the transcription of their targets, creating either feedback loops, as in the case of miR-33a-5p mediated cholesterol homeostasis[57], or feed-forward loops, as in the case of the TLR-induced miR-9-5p and miR-155-5p. Furthermore, and interestingly, when adjusting for transcription rate, we find that miRNA expression capture from 3 to 6%, of variation in gene expression on average, possibly due to their contribution on immune response variability though translation inhibition of key immune genes. Together, this study shows that both environmental and genetic factors contribute to population-differences in miRNA abundance and isomiR ratios, in particular for miRNAs that respond to immune stimulation. Yet, we also show that differences in miRNA expression have only a moderate impact on the transcriptional landscape of immune cells, suggesting that their consequences occur primarily at the protein level, a hypothesis that now needs to be tested experimentally.

Methods

Ethics statement. Human primary monocytes were obtained from healthy volunteers who gave informed consent. All experiments were approved by the Ethics Board of Institut Pasteur (EVOIMMUNOPOP-281297) and the relevant French authorities (CPP, CCITRS and CNIL).

Samples and dataset. The high-density genotyping and RNA sequencing data used in this study were generated as part of the EvoImmunoPop project[22]. The EvoImmunoPop cohort is composed of 200 healthy, male participants of self-reported African and European descent, recruited in Belgium (100 individuals of each population). For all individuals, genotyping data was obtained using both Illumina HumanOmni5-Quad BeadChips and whole-exome sequencing with the Nextera Rapid Capture Expanded Exome kit. Stringent quality control procedures were applied[22] to obtain a set of 3,782,260 high quality SNPs. These SNPs were then used for imputation based on the 1,000 Genomes Project imputation reference panel (Phase 1 v3.2010/11/23), leading to a final set of 19,619,457 SNPs, of which 7,650,709 SNPs had a minor allele frequency (MAF) greater than 5% in our cohort. Details on quality control and imputation have been provided elsewhere [22].

Library preparation and sequencing. Total RNA was extracted with the Nucleospin miRNA kit from Macherey Nagel, which removes genomic DNA through enzymatic digestion. Extractions were performed in batches of 30 samples (i.e. 5 conditions for 3 Africans and 3 Europeans), and RNA quality and quantity were assessed with a Nanodrop spectrometer and the Agilent Bioanalyzer RNA 6000 nano kit. We generated a set of 978 samples, from the 200 donors, fulfilling the criteria for high-throughput RNA-sequencing (RIN > 7, quantity > 2.5 mg). These included 200, 188, 197, 193 and 200 samples for the non-simulated, LPS, Pam_3CSK_4, R848 and IAV conditions, respectively. Each of these 978

samples was split in two, and separate protocols were applied for the sequencing of both poly adenylated mRNAs (described in[22]) and miRNAs (this study). Briefly, for mRNAs, library preparation and sequencing were done using TruSeq RNA Sample Prep Kit v2 for mRNA library construction, TruSeq SR Cluster Kit v3-HS for cluster generation, and TruSeq SBS kit v3-HS for sequencing. mRNA libraries were sequenced using Illumina HiSeq2000 and six samples were pooled within each lane to generate an average of 34.4 million 101-bp single-end reads per sample (min : 27.7 max : 94.8 million reads). For miRNAs, low molecular weight RNA fragments were selected by gel excision (targeting fragments of ~22 bp), and sequencing libraries were prepared using the Illumina TruSeq small RNA library prep Kit. Indexed cDNA libraries were then pooled by groups of 18 (in equimolar amounts), and sequenced with single-end 50bp reads on the Illumina HiSeq2000. After exclusion of one sample that yielded less than 1.8 million read count even after repeating the library preparation step, we obtained an average of 12.4 million raw reads per sample with a minimum yield of 8.0 million reads.

Pre-processing of raw sequencing reads. Sequences matching the 3' adaptor sequence were identified and trimmed, using fastx_clipper version 0.0.13 with the following options –l 0 –n –M 10, to require a minimum adapter alignment length of 10 base pairs, while keeping all sequences regardless of their length or presence of unknown nucleotides. This led to the exclusion of ~2% of reads per sample, and final read lengths ranging from 1 to 42 bases. We confirmed that all samples had average base quality (Q) values >30 at all positions and that per-base GC distributions were within expected ranges. We further checked that read length distributions showed an enrichment of ~22 bases long reads for all samples, consistent with expectations for mammalian miRNAs (~22 bases), and discarded reads shorter than 18 or longer than 26 bases. After these filtering steps, we obtained an average of 8.8 million

(minimum 4.1 million) short reads per sample that were used for small RNA quantification.

Sequence alignment. Sequences were aligned to the human reference genome (build GRCh37/hg19) using bowtie (version 1.1.1) [63]. We mapped reads allowing for 2 mismatches (-v 2) and reported all best alignments for reads that mapped equally well to more than one genomic location (-a —best —strata). We suppressed reads with more than 50 possible alignments (-m 50). On average, ~97% of reads aligned to the genome (min 90%), of which 59% overlapped a known miRNA. Due to their reduced size, miRNAs are known to be susceptible to cross-mapping, i.e. spurious read alignments to other related miRNAs with strong sequence similarity [29]. In the present dataset, around 65% of reads aligning to known miRNAs had more than one possible alignment on the genome. To mitigate the impact of such cross mapping on miRNA quantification, we used a correction strategy that assigns weights to each of the candidate mapping loci of multiply aligning reads, based on local expression levels and mismatches in the alignment [29], allowing to distinguish true miRNA reads from likely alignment errors.

Quantification of miRNA expression. We extracted reads aligning to annotated mature miRNA sequences (miRBase v20) with at least 75% overlap using BEDTools[64] and divided counts per million associated of each miRNA by the total number of miRNA mapping reads to obtain comparable numbers across all libraries. In addition, we used DESeq2 (version 1.20) [65] to compute size factors associated to each library and normalize miRNA counts per million across libraries. We then removed lowly, or sporadically, expressed miRNAs by keeping only those with counts of greater than 1 read per million on average across all experimental conditions, leading to a final set of 658 miRNAs across 675 loci. We then added a pseudo-count of 1 RPM to all miRNAs, and $\log_2$ transformed the data to stabilize the variance of miRNA expression. Linear models were then used to adjust $\log_2$ transformed counts for technical confounders such as mean read length of the library (after clipping), or mean GC content of miRNA-aligned reads, and batch effect induced by date of experiment

and library preparation were sequentially removed using ComBat[66].

Assessment of isomiR diversity. For the analyses at the isomiR level, reads aligning to annotated mature miRNA sequences were extracted as described above, and each unique sequence with a mean expression of >1 count per sample was treated as a separate isomiR. MiRNA sequences presenting less than 1 count per sample on average were discarded, and read counts were normalised using the same approach applied for total miRNA expression. We also removed reads where at least one nucleotide could not be called. For each miRNA, the canonical sequence was defined according to miRBase v.20 and similarity with canonical sequence at nucleotides 2-7 was used to distinguish *canonical seed* isomiRs from *non-canonical seed* isomiRs. We further classified isoform modifications into four main categories, each subdivided into subtypes of miRNA modifications: (i) *changes in start site*, subdivided in *5' extension* and *5' reduction*; (ii) *template changes in end site*, subdivided in *3' extension* and *3' reduction*; (iii) *non-template 3' additions*, subdivided into *3' adenylation* and *3'uridylation*, and *other 3' additions*, and (iv) *internal nucleotide substitutions*, subdivided by position and nucleotide change. For each miRNA, we then quantified the frequency of each type of modifications, and used these quantities for all downstream analyses. These frequencies were then averaged across miRNAs, to provide global estimates of the frequency of miRNA modification events across samples.

Differential expression/isomiR analysis. To identify miRNAs that are differentially expressed upon stimulation, we transformed miRNAs counts using an inverse normal rank-transformation, and fitted a linear mixed model of the form $\{ y_{ij} = a_i + b.\mathbb{I}_j^{stim} + \epsilon_{ij} \}$, where y_{ij} is the transformed counts of individual *i* in condition *j*, a_i is a random effect capturing the inter-individual variability in miRNA expression, b is the effect of stimulation on miRNA expression, $\mathbb{I}_j^{stim}$ is an indicator variable equal to 0 for the non stimulated samples and 1 for stimulated samples, and ϵ_{ij} are the residuals. Significance was assessed by maximum likelihood test, and a global Benjamini and Hochberg FDR-correction was applied

across all 4 stimuli. Only changes in miRNA expression with corrected p-value < 0.01 were considered as significant. To detect significant change in isomiR ratios, we employed a similar approach using isomiRs ratios, instead of miRNA read counts.

Sharing of effects across conditions. To assess the similarity of miRNA response across stimuli, we focused on all miRNAs and isomiRs that were detected to respond to stimulation in at least one condition. We then used a Likelihood-based Model Selection framework [52], assuming that miRNAs respond to only a subset of stimuli, and identified the most likely subset of stimuli by jointly modelling rank-transformed miRNA expression, or isomiR ratios, across all 5 conditions. Specifically, for each stimulus j, we assign an indicator variable γ_j equal to 1 if a given miRNA responds to the stimulus and 0 otherwise. Then, for each of the 15 non-null combinations of stimuli $(\gamma_j)_{j\in\{1,2,3,4\}}$, we fitted a linear mixed model, as done previously in each condition $\{ y_{ij} = a_i + b.\gamma_j + \epsilon_{ij} \}$, with a_i a random effect capturing the inter-individual variability in miRNA expression or isomiR ratio, b is the effect of stimulation and ϵ_{ij} the residuals, and assigned a probability to each model m as

$$Prob(\text{Model } m) = \frac{Likelihood_m}{\sum_{k=1}^{15} Likelihood_k}$$

Detection of miRNA/isomiR-QTLs. To identify genetic variants associated with miRNA expression level or isomiR ratios, i.e., miRNA- and isomiR- quantitative trait loci (miR-QTLs and isomiR-QTLs), we focussed on the 598 miRNAs that could be uniquely assigned to a single genomic locus and considered a set of 10,261,270 genetic variants with a minor allele frequency (MAF) ≥ 0.05 in either Europeans or Africans, of which 1,981,401 were located < 1Mb from one the 598 mature miRNAs. We used *MatrixEQTL*[67] to map miR-QTLs within a 1Mb window on each side of mature miRNAs. miRQTL mapping was performed separately for each condition, merging both populations and including an indicator variable to control for the effect of population on miRNA expression. MiRNA counts per million values and isomiR ratios were rank-transformed to a normal distribution before

mapping, to reduce the impact of outliers. FDR was computed by mapping sQTL on 100 permuted datasets, in which genotypes were randomly permuted within each population. We then kept, for each permutated dataset, the most significant *p*-value per miRNA or isomiR, across all conditions, and computed the FDR associated with various *p*-value thresholds ranging from 10^{-3} to 10^{-50}. We subsequently selected the *p*-value threshold that provided a 5% FDR ($p < 10^{-6}$).

When comparing miR-QTLs across conditions, we used a likelihood-based model selection framework to increase power for detection of shared effects. Namely, for each SNP-miRNA pair, rank-transformed miRNA expression, or isomiR ratios, y_{ij} are modelled jointly across all 5 conditions. An indicator variable γ_j is created that is equal to 1 if the miRNA is under genetic control in condition j and 0 otherwise. Then, for each of the 31 non-null combinations of stimuli $(\gamma_j)_{j\in\{1,2,3,4,5\}}$, we fitted a linear model of the form $\{ y_{ij} = a_{jp} + b.\gamma_j SNP_i + \epsilon_{ij} \}$, with a_{jp} the mean expression of the miRNA or isomiR in condition j and population *p*, SNP_i the number of minor alleles carried by individual *i*, *b* the mean effect of the SNP in conditions where it is active and ϵ_{ij} the residuals. Each model is then assigned a probability as :

$Prob(\text{Model } m) = \frac{Likelihood_m}{\sum_{k=1}^{31} Likelihood_k}$ and the model with the highest probability is retained.

In addition, to identify response-miRQTLs, we also tested for significant differences in effect size of miR-QTLs between the stimulated and non-stimulated state using an interaction test. Rank-transformed miRNA expression, or isomiR ratios, y_{ij} are decomposed between a_{jp} the mean expression of the miRNA or isomiR in condition *j* and population *p*, the effect of the SNP at basal state *b*, and the differences in effect size between basal and stimulated state *c*.

$$y_{ij} = a_{jp} + b.SNP_i + c.SNP_i.\mathbb{I}_j^{stim} + \epsilon_{ij}$$

Significance of the interaction is then tested by a Student t test for H_0: $\{c=0\}$.

Annotation of miRNA Transcription start site and miR-QTLs. Transcription start site (TSS) of miRNAs were obtained from Fantom5 data, based on [68], together with their conservation levels (mean PhastCons of promoter region). Hairpin coordinates were retrieved from mirBase V20. MiR-QTLs for which TSS information was available were then classified based on their location relative to the TSS and hairpin. Namely miR-QTLs were first classified as *miRNA-altering* or *hairpin-altering* if they overlapped the sequence of the mature miRNA or its associated hairpin. Then, we computed for each miR-QTL the distance between the SNP and both the hairpin and the TSS of the associated pri-miRNA. miR-QTLs that were located less than 10kb from the TSS or the hairpin were annotated as *hairpin-* or *TSS-flanking* according to the feature from which they were the closest. Finally miR-QTLs that were located > 10kb from both TSS and hairpin, were annotated as *Distant*.

Population differences in miRNA/isomiR expression. To identify miRNAs that are differentially expressed between populations, we applied Student's t-test to inverse normal rank-transformed miRNAs counts within each condition separately, comparing African- to European- ancestry individuals. A global Benjamini and Hochberg FDR-correction was applied across all 5 conditions to evaluate significance. Only changes in miRNA expression with corrected p-value < 0.01 were considered as significant. A similar approach was used to test for population differences in isomiR levels, using isomiRs ratios, instead of miRNA read counts. Sharing of population differences among conditions was assessed using a model selection framework similar to the one used to assess sharing of mir-QTLs. For each individual *i* and condition *j*, we assigned an indicator variable γ_j equal to 1 if a the miRNA is differentially expressed between populations in that condition and 0 otherwise. Then, for each of the 31 non-null combinations of conditions $(\gamma_j)_{j\in\{1,2,3,4,5\}}$, we fitted a linear model$\{ y_{ij} = a_j + b.\gamma_j.\mathbb{I}_{ji}^{pop} + \epsilon_{ij} \}$, with a_j a the mean expression of the miRNA in condition *j* across African individuals, *b* the mean difference in miRNA expression between European- and African- ancestry individuals, and ϵ_{ij} a normally distributed residual. Each possible model is then assigned a probability *m* as and the most likely model is retained.

$$Prob(\text{Model } m) = \frac{Likelihood_m}{\sum_{k=1}^{31} Likelihood_k}$$

Assignment of miRNA targets. miRNA targets were predicted using miRanda v3.3a, providing canonical sequences obtained from miRBase V20 as input and 3'UTR sequence of known transcripts based on Ensembl V70. Defaults settings were used for target prediction, and a gene was considered as targeted by a miRNA if at least one of its annotated transcripts had a predicted binding site for the miRNA. Prediction of isomiR targets was performed in a similar manner, using the isomiR sequence instead of the canonical sequence.

Quantification of gene expression levels. RNA-seq reads were aligned to hg19 using Tophat2 [69] and gene expression values (FPKM) were computed with CuffDiff[69] based on Ensembl v70. Samples with uneven gene coverage were excluded leaving a total 969 samples with both miRNAs and gene expression. Gene expression values were log transformed (with an offset of A and corrected for GC content and 5'/3'coverage biais, as well as experiment and library preparation date using linear models and ComBat [70]. Further details on Gene expression quantification, QC and normalization can be found elsewhere [22]. To avoid indirect correlations between miRNA and gene expression that result from co-transcription, we estimated transcription rate based on the number of nascent unspliced transcripts [47]. Namely, for each gene we used HT-Seq [71] to compute the average number of reads mapping to the gene after exclusion of all exonic regions. This number of intronic read was then divided by the total length of introns to yield a mean intronic coverage that was used as a proxy of the transcription rate. For each gene, inverse-normal rank transformation was applied to the gene expression levels and transcription rate to reduce the impact of outlier values in downstream analyses.

Assessment of miRNA-gene correlation.

To identify likely miRNA-gene interactions occurring in each condition, we modelled gene expression as a function of miRNA levels, using population as covariate. All miRNAs were introduced simultaneously in the model and an elastic net penalty[72] was set on the miRNA effects to make the model identifiable, leading to the following model

$$\boldsymbol{Expr} = a + b\,.\boldsymbol{population} + \sum_{j=1}^{n} c_j.\boldsymbol{miRNA}_j + \boldsymbol{\epsilon}$$

with $\left(\sum_{j=1}^{n}|c_j| + \sum_{j=1}^{n} {c_j}^2\right) < \lambda.$

Here, $\boldsymbol{Expr}$ is the vector of gene expression across all samples from the condition under study, $\boldsymbol{population}$ is an indicator variable representing population of origin of the individual, $(\boldsymbol{miRNA}_j)_{j=1..n}$ are the vectors of expression of the 658 expressed miRNAs, and $\boldsymbol{\epsilon}$ is a random Gaussian noise. *a* denotes the mean expression in the reference population, *b* and $(c_j)_{j=1..n}$ are parameters capturing the effect of population, and miRNAs on gene expression. λ is a constant value that captures the amount on constraint on the effect of miRNAs included in the model.

Using this model, we performed stability selection [73] to select miRNAs that have a significant effect on gene expression with high probability. Briefly, stability selection consist in performing repeated subsamplings of the data, typically considering only half of the initial data, and selecting the first Q miRNAs with non null c_j coefficients across increasing values of λ. Under the reasoning that only miRNAs with a true effect on gene expression will be consistently selected across subsamplings, we can then use the frequency at which a miRNA was kept in the model as the posterior probability that this miRNA has a significant impact on gene expression. Here, we performed 100 resamplings with Q=30 and considered as significant only miRNAs that reach a posterior probability of 0.8, which is equivalent to a 5% FDR based on simulations.

Correlation between miRNAs and transcription.

Correlation between miRNAs and transcription rate was obtained using the MatrixeQTL package, providing miRNAs instead of genotypes and adjusting for population.

All associations where the miRNA was located less than 1kb away from the gene were discarded and 5% FDR was used to declare associations as significant. For each miRNA, significantly associated genes were split into *negatively* and *positively* correlated genes according to the sign of the corresponding β parameter. We then tested each set of associated genes for enrichment in predicted binding sites obtained from miRanda compared to the set of all transcribed genes with at least one predicted miRNA binding site. Benjamini Hochberg correction was applied across all miRNAs for both positive and negative corrleations, and only enrichments passing a 5% FDR were retained.

Relative contribution of transcription and miRNAs to the variability of gene expression.

To account for co-trancription when assessing miR-gene correlations, we repeated our stability selection approach adding transcription as a covariate in the model. The final model can thus be written as

$$\boldsymbol{Expr} = a_p + b\,.\boldsymbol{transcription} + \sum_{j=1}^{n} c_j.\boldsymbol{miRNA}_j + \boldsymbol{\epsilon}$$

with $\left(\sum_{j=1}^{n}|c_j| + \sum_{j=1}^{n} {c_j}^2\right) < \lambda.$

Here, $\boldsymbol{Expr}$ and $\boldsymbol{transcription}$ are the vectors of gene expression and transcription rate across all samples from the condition under study, $(\boldsymbol{miRNA}_j)_{j=1..n}$ are the vectors of expression of the 658 expressed miRNAs, and $\boldsymbol{\epsilon}$ is a random Gaussian noise. a_p denotes the mean expression in population p and b and $(c_j)_{j=1..n}$ are parameters capturing the effect of transcription and miRNAs on gene expression. λ is a constant value that captures the amount on constraint on miRNAs that are included in the model.

After identifying miRNAs that have a significant effect on gene expression, miRNA effect sizes were assessed using CAR scores as implemented in the care package [58]. Briefly, CAR scores (noted ω) are a variation of partial correlations allow to measure correlation between 1 or more covariates and response variable, while adjusting each covariate for the

effect all other covariates. More importantly, the squared CAR scores (ω^2) sum to the total percentage of variance explained by the model (R^2), allowing to interpret the square of each individual CAR score as the percentage of variance explained by the associated covariate, when adjusting for all other covariates. To evaluate the variance explained by a subset of miRNAs (i.e. negatively correlated miRNAs or negatively correlated miRNAs with a known binding site to the gene), we simply consider the sum of ω^2 over all miRNAs of that subset (using the sign of ω, to identify negative correlations).

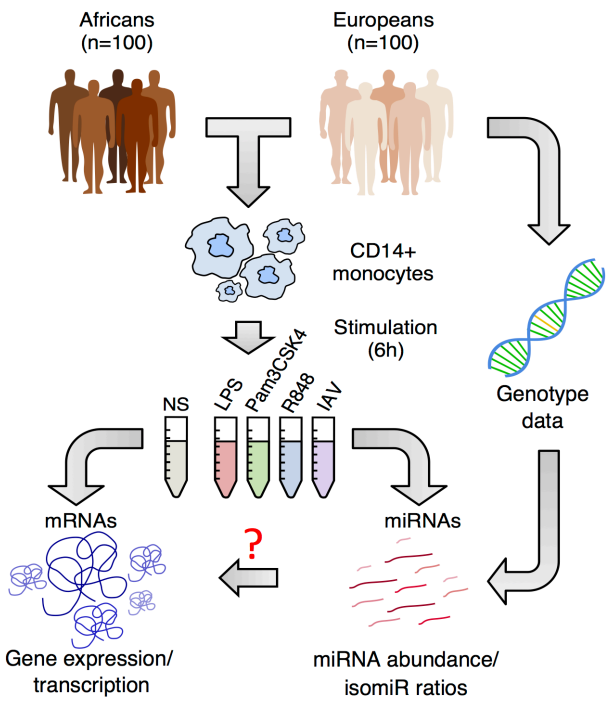


Figure 1. Assessing the causes and consequence of population variation in miRNA response to immune activation. To understand the contribution of miRNAs to population differences in immune responses, we stimulated monocytes from 200 healthy individuals (100 of African-descent and 100 of European-descent), using 3 TLR ligands (LPS activating TLR4, Pam_3CSK_4 activating TLR1/2 and R848 activating TLR7/8) as well as a live strain of influenza A virus (A/USSR/90/77(H1N1), denoted as IAV thereafter). For each individual,

RNAs were extracted after 6h of stimulation, and sequenced mRNAs and small RNAs, from both stimulated cells and non-stimulated cells (NS) kept resting for the same amount of time. The integration of small RNA sequencing data with genetic data, obtained through whole genome genotyping, whole exome sequencing and imputation, allows assessing the genetic bases of population differences in miRNA responses to stimulation, both quantitatively (miRNA abundance) and qualitatively (isomiR ratios). Furthermore, the availability of mRNA sequencing data from the same individuals allows quantifying both total gene expression levels (exonic reads) and transcription rate (intronic reads derived from nascent mRNAs). These data are then used to assess the impact of miRNAs and their isomiRs on immune response, both at the transcriptional and post-transcriptional level.

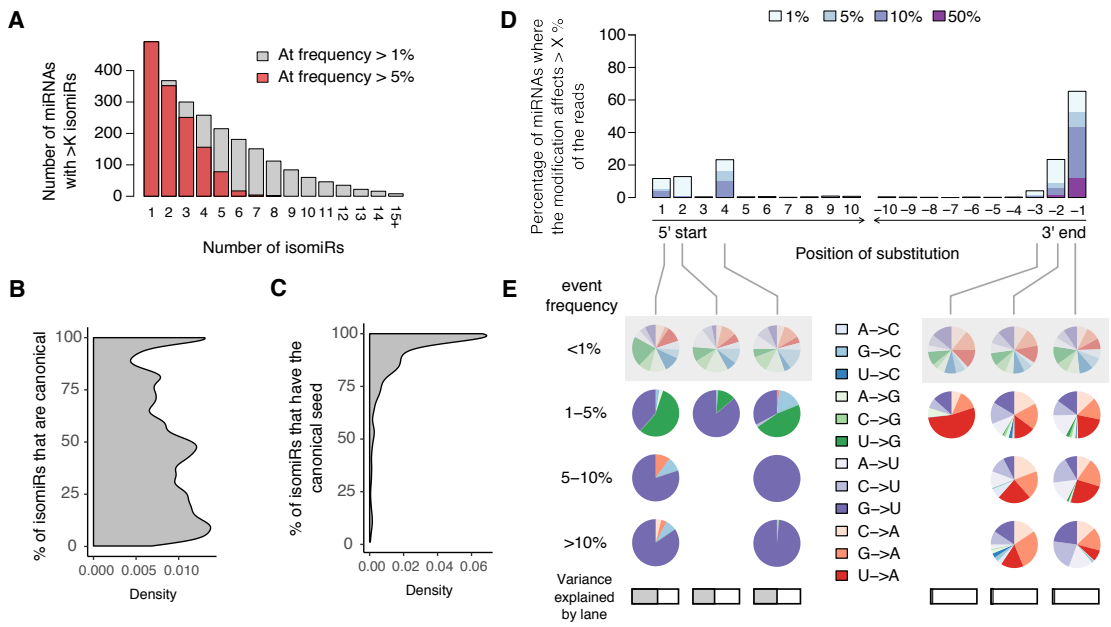


Figure 2. The landscape of miRNA diversity in human primary monocytes. (**A**) Reverse cumulative distribution function of the number of isomiRs per miRNA for various frequency thresholds. For each possible number of isomiRs *K*, the plot shows the number of miRNAs with more than *K* isomiRs at frequency of either >5% or >1%. (**B**) Distribution of the percentage of canonical isomiRs among all detected miRNAs. (**C**) Distribution of percentage of isomiRs with canonical seed among all detected miRNAs. (**D**) Distribution of edited nucleotides along miRNAs. For each nucleotide position, counting either from the 5' start site (position 1 to 10) or from the 3' end site (positions -10 to -1), we report the percentage of miRNAs that present an editing event accounting for at least 1% of edited reads (light blue). Similarly, we quantify the fraction of miRNAs where the editing event accounts for more than 5% (blue), 10% (indigo) or 50% (deep purple) of the reads. For instance, 20% of miRNAs have a substitution at their 4th nucleotide in over 1% of their reads, and 8% have a substitution in over 10% of their reads. (**E**) Frequency of each type of substitution, according to the percentage of miRNA reads that are edited. Results are shown only for positions where events are detected in >1% of miRNAs. At each position, distribution of substitution types among low frequency editing events (<1%), which we expect to be enriched in false positives, is provided as a reference (gray shadow).

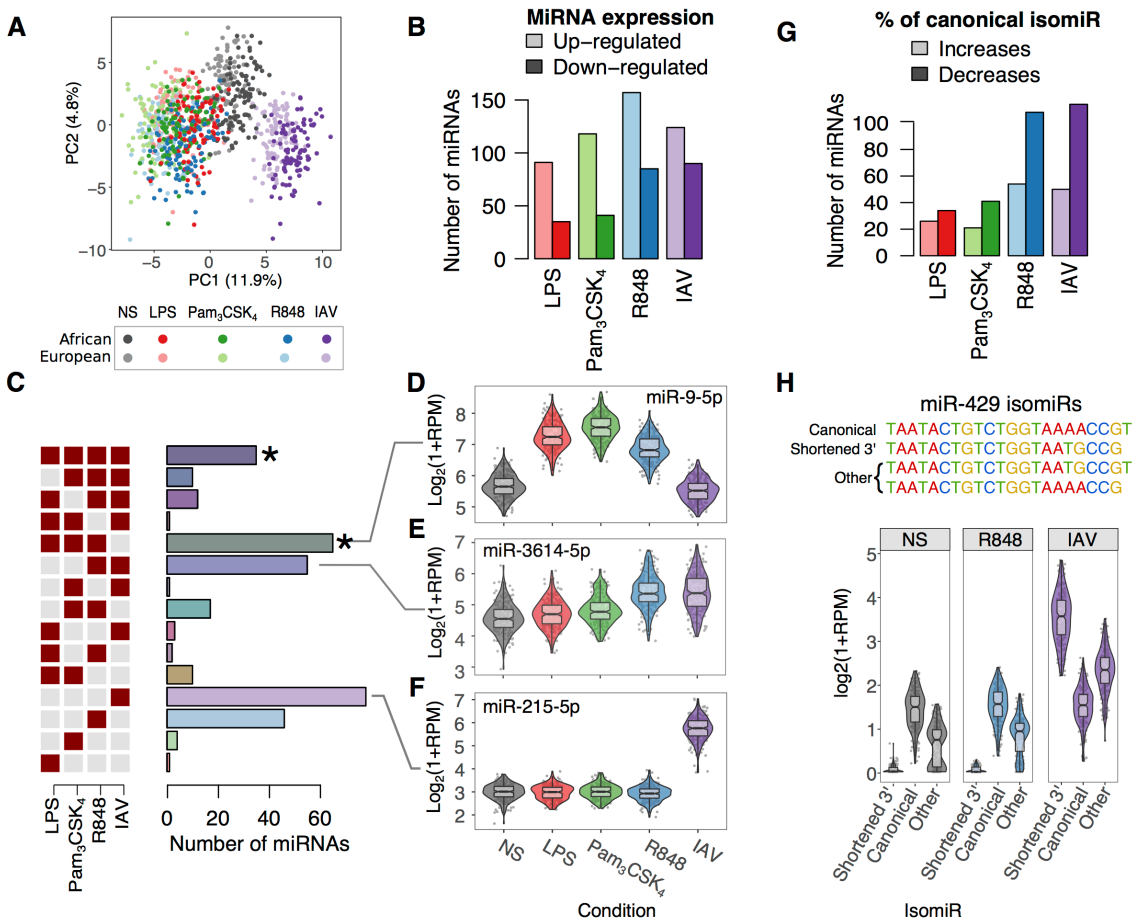


Figure 3. Dissecting stimulus-specific miRNA responses to immune activation. (A) PCA of log transformed miRNA abundances. Each dot represents a sample, colored according to the condition of stimulation (grey – Non stimulated, red – LPS, green – Pam_3CSK_4, blue – R848, purple – IAV). The same color code for conditions is used throughout the manuscript, and light and dark shades indicate European and African ancestry, respectively. (**B**) For each condition, number of DE miRNAs that are either up- (light shade) or down-regulated (dark shade). (**C**) Number of miRNAs that are differentially expressed (compared to NS) in a single condition or a combination of immune stimulations (*binomial $p < 0.001$, significance of overlap between stimuli). (**D-F**) Examples of DEs miRNAs for the three most frequent patterns of differential expression across stimuli (**D**) miR-9-5p, exhibiting a TLR-specific response. (**E**) miR-3614-5p, responding specifically to viral stimuli. (**F**) miR-215-5p showing an IAV-specific response. (**G**) For each condition, number of miRNAs where the canonical isomiR is either up- (light shade) or down-regulated (dark shade). (**H**) Example of isomiR-levels response to IAV of miR-429. miR-429 expresses 4 isomiRs that differ in their 3' end and at nucleotides 17-18. Violin plots show the expression of miR-429 isomiRs at the basal state, after R848 stimulation (as an example of a TLR-ligand), and IAV. The two least frequent isomiRs were grouped.

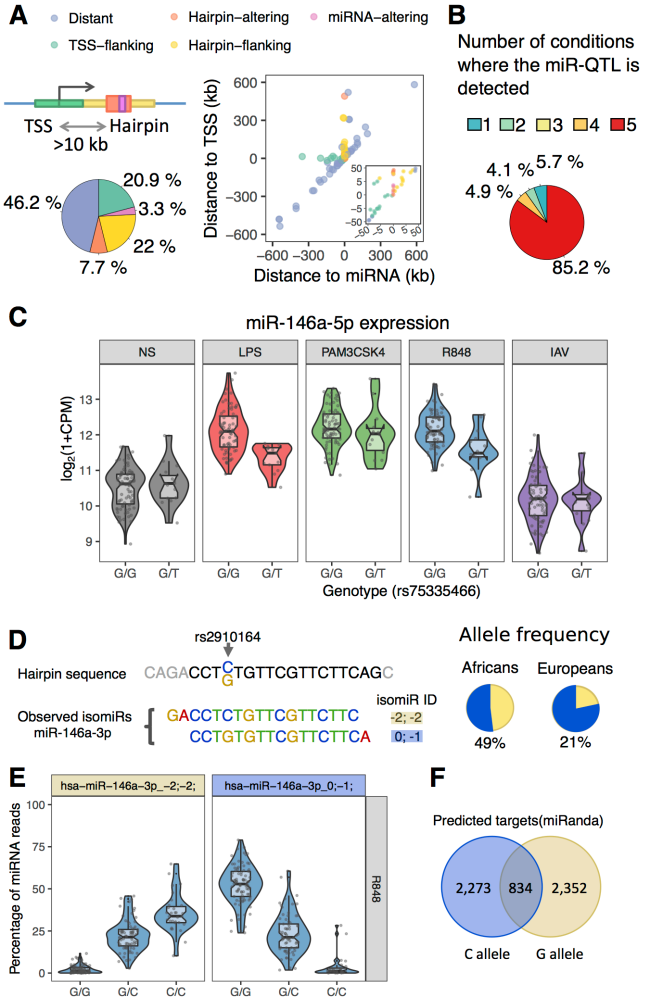


Figure 4. Genetic basis of miRNA expression upon immune activation. (A) Localization of miR-QTLs. (*left*) Frequency of miR-QTL that either overlap the mature miRNA (*miRNA-altering*, pink) or its hairpin (*hairpin-altering*, orange*)*, or are located <10kb away from the miRNA hairpin (*hairpin-flanking*, yellow) or TSS (*TSS-flanking*, green). Remaining miR-QTLs are annotated as *Distant* (blue). (*right*) Distance of mirQTLs from the mature miRNAs (*x*-axis) and its associated TSS (*y*-axis). Each miR-QTL is shown as a separate dot, colored according to its localization. Negative distance indicate that the miR-QTL is located upstream of the miRNA and/or TSS. Close up view is shown for miR-QTLS located < 50 kb from the miRNA or TSS. (**B)** Sharing of miRQTL. Sharing of miR-QTLs across conditions. For the 122 miR-QTLs, number of conditions where sQTLs is identified. (**C)** Exemple of an african-specific response miR-QTL. The rs75335466-T allele is associated with reduced expression of miR-146a-5p specifically upon stimulation by LPS and R848. For clarity, data is shown only for African individuals. (**D-F**) Impact of rs2910164 variant on miR-146a-3p isomiRs. (**D**) Genomic context and frequency of the rs2910164 variant. The rs2910164 C/G variant is shown with its neighbouring hairpin sequence. Sequence of the canonical miRNA defined in miRbase is displayed in black. The 2 most common isomiRs of miR-146a-3p are displayed below and denoted as (-2; -2), and (0; -1) based on the coordinates of their start/end site relative to the canonical miRNA sequence. Note that the C/G substitution is not taken into account for the quantification of (-2; -2) and (0; -1) isomiRs. Frequency of C/G alleles in our sample is shown in Africans and Europeans individuals separately. (**E**) IsomiR-QTL of miR-146a-3p. Ratios of (-2;-2) and (0; -1) isomiRs are shown for each genotype, in the R848

condition where the isomiR-QTL is the strongest. For clarity, other isomiRs are not displayed. (**F**) Overlap of miRNA targets predicted by miRNA for each possible isomiRs.

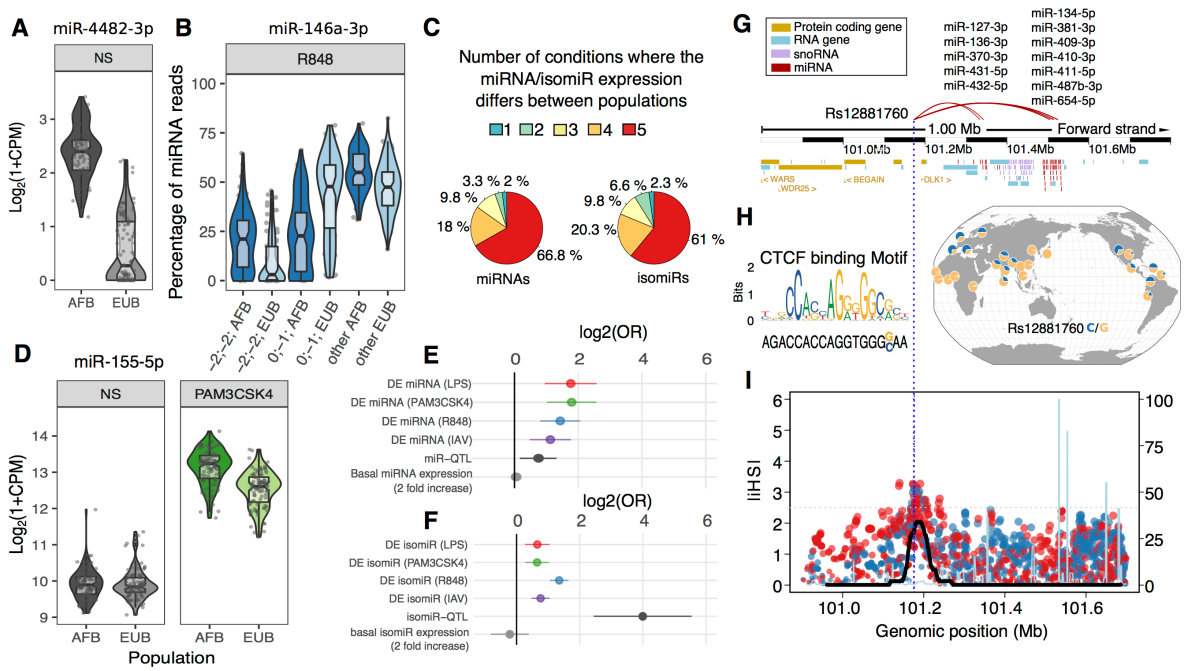


Figure 5. Population differences in miRNA expression. (A) Example of a miRNA (miR-4482-3p) differentially expressed between populations. Expression of miR-4482-3p is shown separately for African (AFB) and European individuals (EUB). (**B)** IsomiRs of miR-146a-3p are differentially expressed between populations. For each population and isomiR, isomiR ratios are shown in the R848 condition where the difference is the strongest. All isomiRs with <1 RPM on average, are pooled and annotated as *other*. (**C)** Sharing of popDE miRNA and popDE isomiRs across conditions. For the 244 popDE-miR and 188 popDE isomiRs, number of conditions where we observe a difference between populations. (**D**) Expression of miR-155-5p is differential between Europeans and Africans specifically in TLR-stimulated conditions. For simplicity, only Pam_3CSK_4 condition is shown here. (**E-F)** Enrichment of popDE-miRs (**E**) and isomiRs (**F**) in miRNAs/isomiRs that change their expression upon stimulation or have a QTL. All Odds ratio are adjusted for mean expression of the miRNA/isomiR at basal sate. Odds ratio associated to mean expression at basal state are provided for reference. (**G-I)** Evidence of selection on the miR-QTL hotspot rs12881760. (**G**) Genomic context of the miR-QTL displaying the protein coding genes (yellow), RNA genes (cyan), snoRNAs (purple) and miRNAs (red) in & 1Mb window around the locus. Red lines link the miR-QTL to its target miRNAs, the name of which are indicated above. (**H**) Impact of the rs12881760 variant on formation of a CTCF motif, and worldwide frequency of the motif disrupting C allele. (**I**) Evidence for positive selection at the rs12881760 locus. |iHS| are displayed for all SNPS with MAF > 5% in Europeans, and dots are colored according to the sign of the iHS statistic (red - positive, blue - negative). Black line indicates the percentage of outliers (|iHS|>2.5) on a sliding window of 100 consecutive SNPs with MAF>5% (right axis). Recombination rate is overlayed in light blue and normalized to the maximum recombination rate in the region (peak :152 cM/Mb).

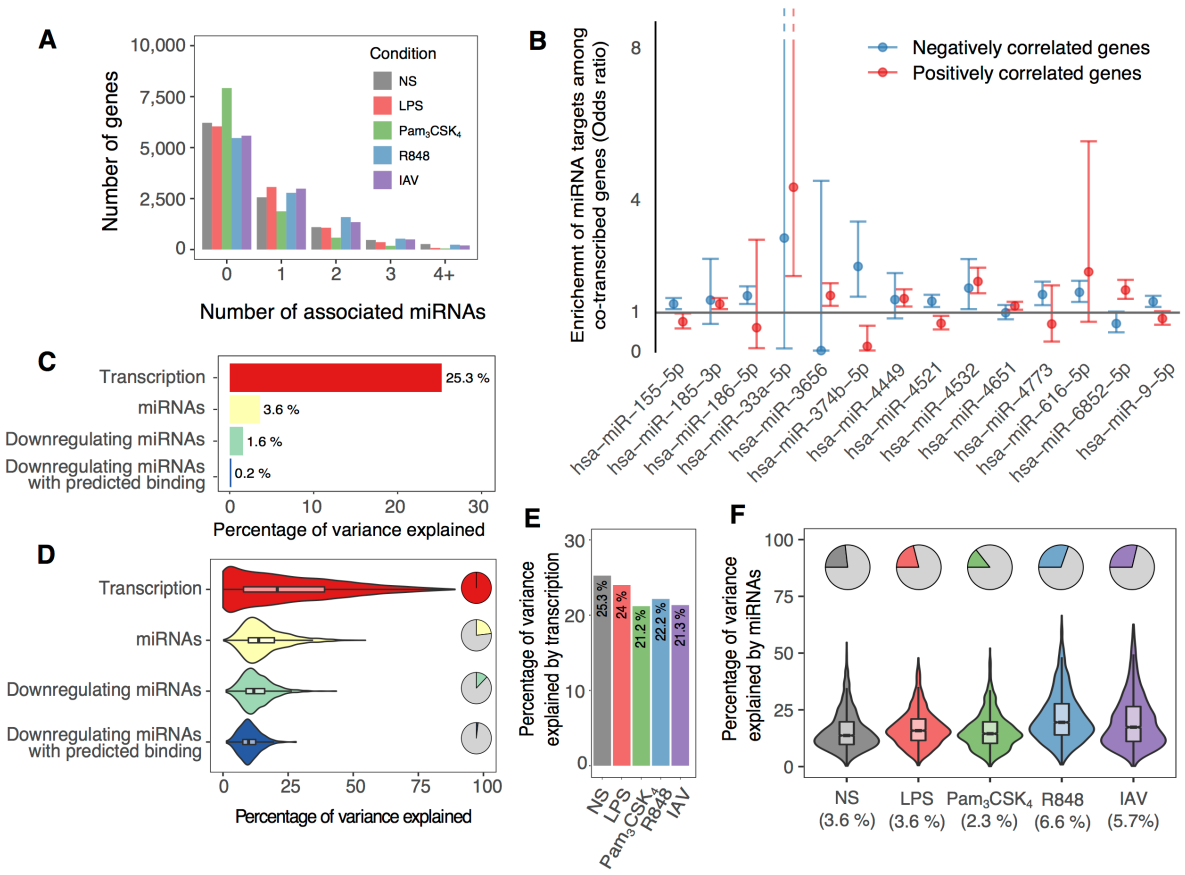
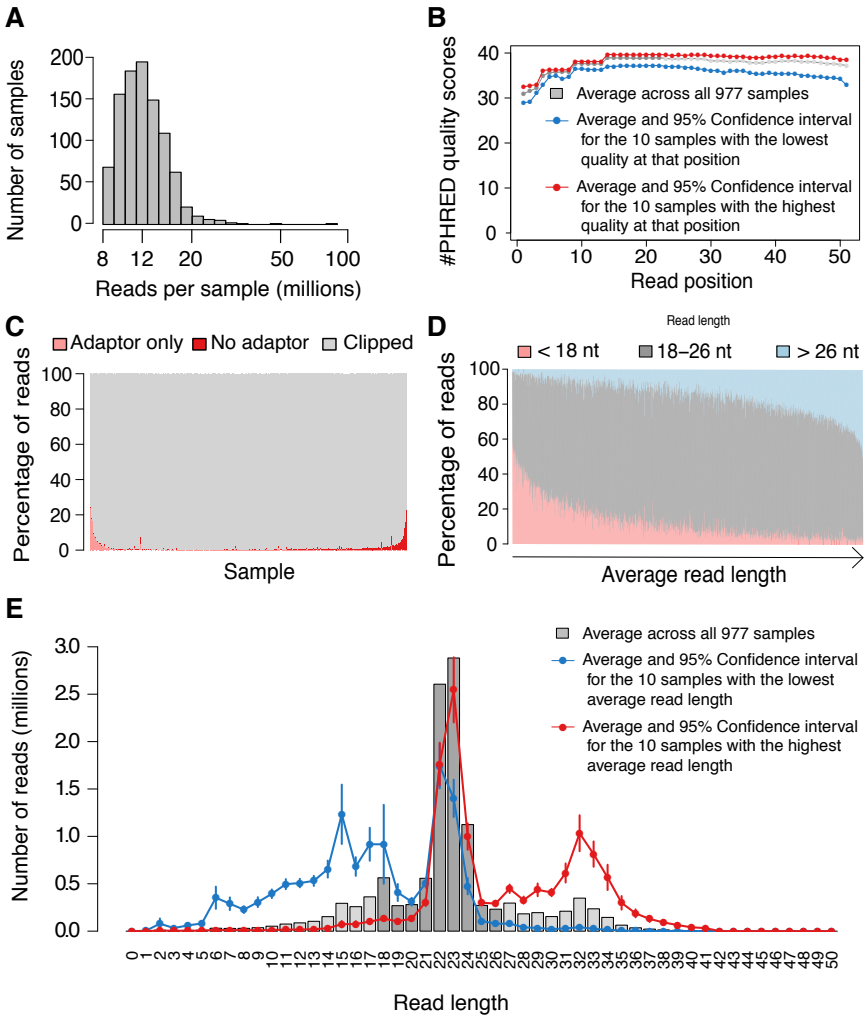


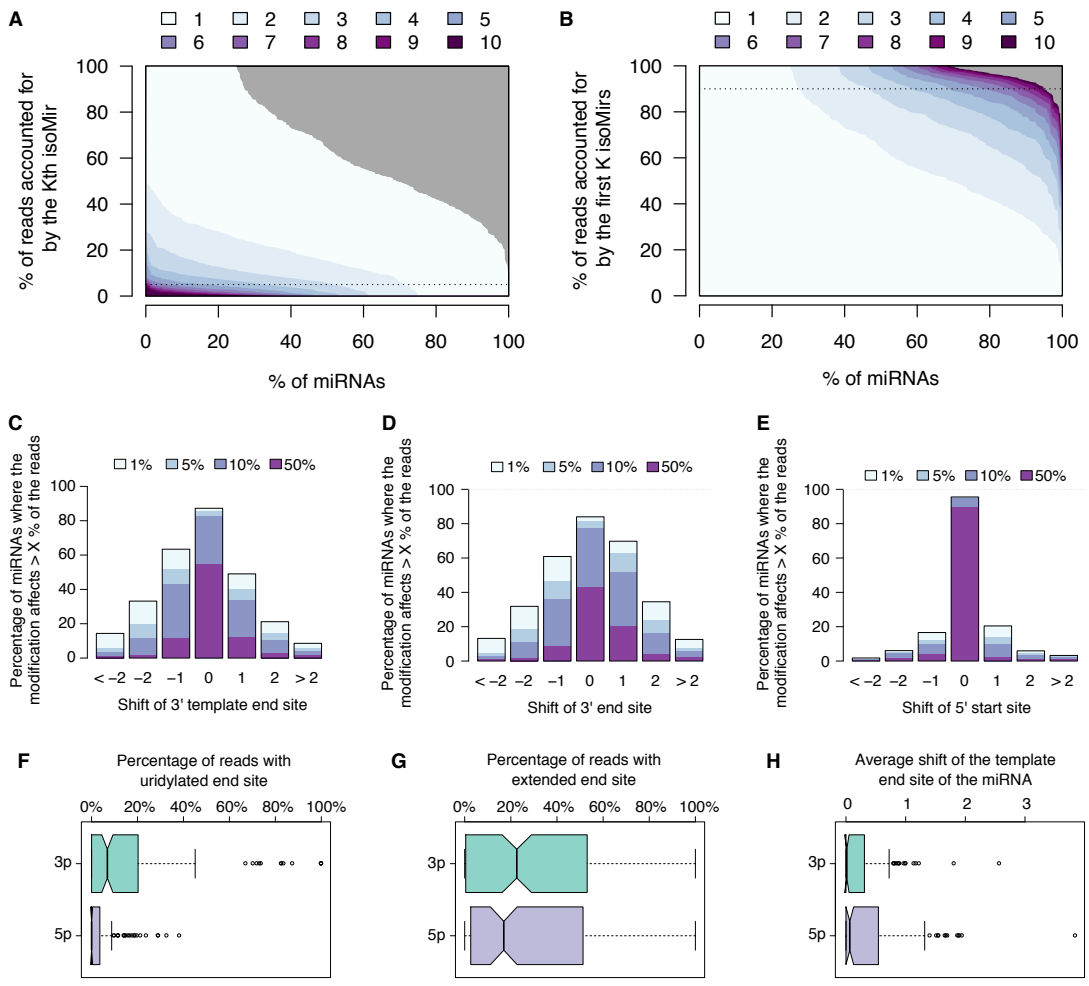
Figure 5. Impact of miRNA levels on gene expression.
(A) Distribution of the number of associated miRNAs per gene according to the condition of stimulation. (**B**) Enrichment or depletion of miRNA targets among genes whose transcription level correlates with miRNA expression (co-transcribed genes). For each miRNA, odds ratios are reported separately for genes whose transcription is positively (red) or negatively (blue) correlated to miRNA expression. Enrichments are displayed only for miRNAs that have a significant enrichment of their targets in either positively or negatively correlated genes (5% FDR). (**C-D**) Percentage of gene expression variance that is attributable, at basal state, to either transcription or miRNA variation. For miRNAs, attributable variance is also reported considering only negative associations, or negative associations with predicted binding between the gene and the miRNA. (**C**) Global percentages (Average values across all genes). (**D**) Distribution at gene level. Violin plots display the distribution of the variance attributable to transcription or miRNAs, among genes with at least one associated miRNA. Pie charts indicate the percentage of genes associated to transcription or miRNAs. **(E)** Global percentage of gene expression variance that is attributable to transcription according to the condition of stimulation. **(F)** Distribution of the percentage of gene expression variance that is attributable to miRNAs according to the condition of stimulation. Violin plots display the distribution of the variance attributable to miRNAs among genes with at least one associated miRNA. Pie charts indicate the percentage of genes associated to at least one miRNA.

Supplementary Figures

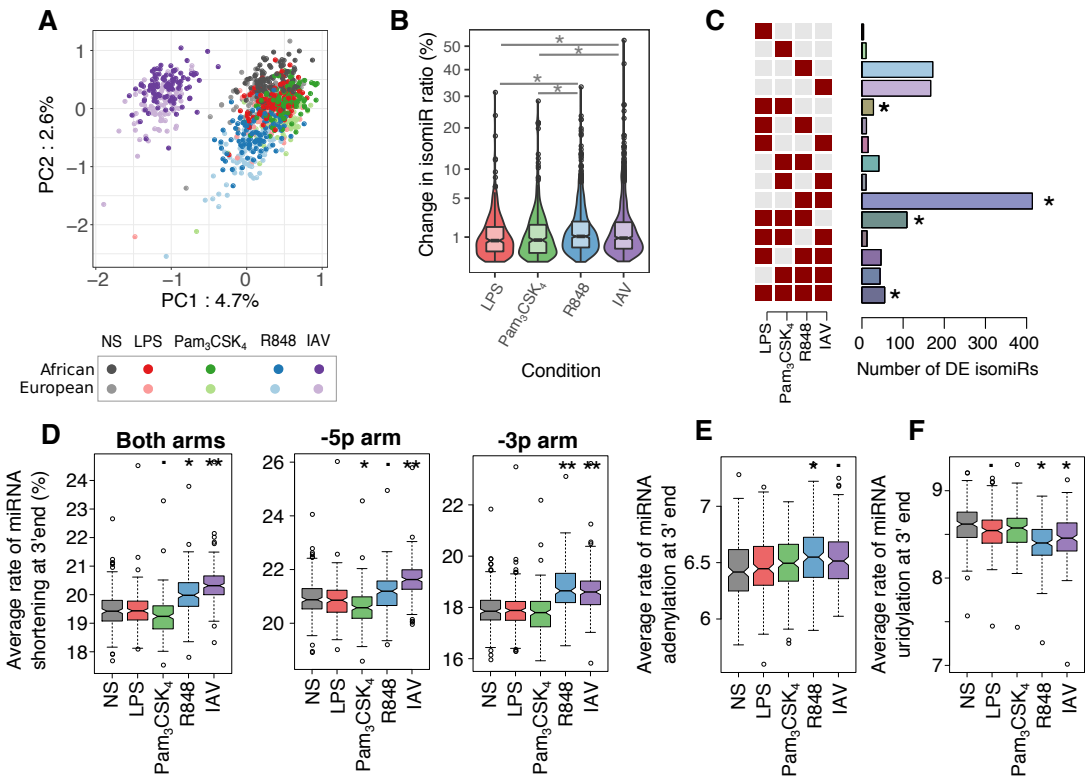


Supplementary Figure 1. Quality and pre-processing of small RNA sequencing data. (A) Histogram of the total number of sequenced reads per sample. (**B**) Mean quality of sequenced reads. For each position along the 50bp of the sequenced reads, the mean quality at that position (across all 977 samples) is reported in grey. In addition, at each position, the mean quality among the 10 samples with highest and lowest quality at that position is show in blue and red, respectively. Vertical bars indicate 2 standard errors from the mean, among these samples. (**C**) Distribution, among the 977 high quality samples, of the fraction of reads removed during the adaptor trimming. Each column is one sample. Reads where only the adaptor was sequenced are displayed in red, and reads that lack of adaptor sequence and displayed in pink. Correctly clipped reads, are displayed in grey. (**D**) Distribution, among the 977 high quality samples, of the fraction of reads removed due based on length. Each column is one sample. Reads that are shorter than 18 nucleotides are displayed in red. Reads that are longer than 26 nucleotides are displayed in blue. Reads that are kept for downstream analyses are shown in grey. Samples are sorted according to their average read length. (**E**) Distribution

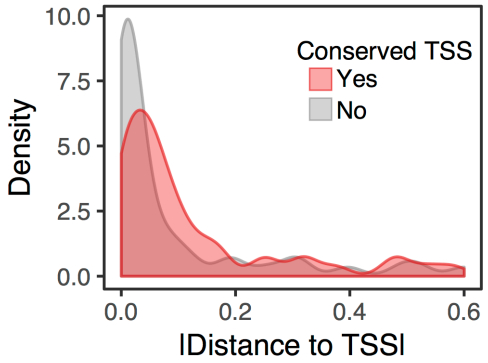
of read lengths after adaptor trimming. For each possible read length, the mean number of read per sample is displayed as a grey bar (darker grey is used for the 18-26 nucleotide range). In addition, at each possible length, the mean number of reads among the 10 samples with highest and lowest average read length is shown in blue and red, respectively. Vertical bars indicate 2 standard errors from the mean, among these samples.



Supplementary Figure 2. IsomiR diversity and rate of miRNA modifications. (A) Distribution of the percentage of reads accounted by the first 10 isomiRs, across miRNAs. (**B**) Distribution of the cumulative percentage of reads accounted by the first 10 isomiRs. (**C**) Distribution of shifts in 3' end site, excluding non-template additions. For each possible 3' end site, we report the percentage of miRNAs for which the 3' end site accounts for at least 1% (light blue), 5% (blue), 10% (indigo) and 50% (deep purple) of edited reads. (**D**) Distribution of shifts in 3' end site, including non-template additions. (**E**) Distribution of shifts in 5' start site, including non-template additions. (**F**) Difference in average percentage of miRNA reads uridylated at their 3' end, between -3p miRNAs (green) and -5p miRNAs (purple). (**G**) Difference in average percentage of miRNA reads with an extended 3' end site, between -3p miRNAs (green) and -5p miRNAs (purple). (**H**) Differences in shift of the 3' end site, between -3p miRNAs (green) and -5p miRNAs (purple), site excluding non-template additions.



Supplementary Figure 3. Sources of isomiR variation upon immune activation. (A) PCA of the ratios of commons isomiR (RPM>1, >1% of miRNA reads). Each dot represents a sample, coloured according to the condition of stimulation (grey – Non stimulated, red – LPS, green – Pam_3CSK_4, blue – R848, purple – IAV; The same color code for conditions is used throughout the manuscript), light and dark shades indicate African and European ancestry respectively. (**B)** For each condition, violin plot showing the distribution of absolutes changes in isomiR ratios for DE isomiRs. (* Wilcoxon p<.01) (**C)** Number of IsomiR that change their ratio (compared to NS) in a single stimulus or a combination of stimuli (*binomial $p < 0.001$, significance of overlap between stimuli). (**D)** For each sample, the rate of miRNA shortening at 3' end is obtained as the average across all miRNAs of the fractions of isomiRs that have a shortened 3' end. This rate is measured either across all miRNAs (*left*), or separately among miRNAs present from the -5p (*middle*) and -3p arm (*right*) of the hairpin loop. (symbols indicate significant deviations from the non-stimulated state based on Wilcoxon rank test; • P_{adj}<0.05, * P_{adj}<0.01, ** $P_{adj}<10^{-20}$) (**E-F)** For each sample, the rate of miRNA adenylation (**E**) and uridylation (**F**) at 3' end is obtained as the average across all miRNAs of the fraction of isomiRs that are adenylated or uridylated at their terminal site, and is plotted as a function of the condition of stimulations (symbols indicate significant deviations from the non-stimulated state based on Wilcoxon rank test; (• P_{adj}<0.05, * P_{adj}<0.01, ** $P_{adj}<10^{-20}$).



Supplementary Figure 4. Distance of miR-QTLs from their associated transcription start site (TSS). Density plots showing the distribution of distance in Megabases between miR-QTLs and the transcription start site of their associated pri-miRNA. Distribution is shown separately for miRNAs with conserved (red, mean phastCons>20%) and non-conserved promoters (grey, mean phastCons<20%).

List of Supplementary Tables

Tables are available upon request

Supplementary table 1

(**A**) List of all 675 loci encoding a miRNA, with their coordinates and the name of the associated miRNA. Coordinates and ID of the associated hairpin and transcription start site (TSS) are also provided. (**B**) List of all 2,359 frequent isomiRs sequences with their associated miRNA and hairpin. Details of observed deviations from the canonical isomiR are also provided. (**C**) For each type of miRNA modification, average rate of occurrence of the modification among -3p and -5p miRNAs and significance of the differences in rate between both arms (Wilcoxon test).

Supplementary table 2

(**A**) List of 340 miRNAs that are change their expression in response to at least one stimulus. For each miRNA, basal expression and fold changes upon stimulation are provided together with significance of expression changes and the inferred model of sharing across conditions. (**B**) List of isomiRs that are change their ratios upon stimulation. For each isomiR, the name of the associated miRNA is reported together with the % of miRNA reads that it accounts for at basal state. For each stimulus, changes in expression ratio are indicated together with their associated p-value. The inferred model of sharing across conditions is also provided. (**C**)
For each type of miRNA modification, changes in rate of occurrence of the modification upon stimulation, and their significance. Average rate at basal state is also reported. Results are provided for all miRNAs together (any) or focussing on -3p or -5p arm miRNAs.

Supplementary table 3

(**A**) List of 122 miRNAs associated with at least one miR-QTL, and their most strongly associated variant. Association statistics (β coefficient, *p*-value, R^2) are provided for each condition. For each SNP, genomic coordinates are reported, together with derived allele frequency in Europeans and Africans. Distance from the mature miRNA, pre-miRNA and TSS (when available) are also provided as well as miR-QTL classification (miRNA-altering, hairpin-altering, hairpin-flanking, TSS-flanking, and Distant). (**B**) Tests for genotype × stimulation interactions across all 4 stimuli for the 122 miR-QTLs. Association statistics (interaction coefficient, *p*-value) are provided for each condition and the best model of sharing of the miR-QTL is reported. (**C**) List of 25 isomiRs associated with at least one isomiR-QTL, and their most strongly associated SNP. For each SNP, genomic coordinates are reported, together with derived allele frequency in Europeans and Africans.

Supplementary table 4

(**A**) List of 244 miRNAs that are differentially expressed between populations in at least one condition. For each miRNA and condition, average expression in Africans and Europeans is provided together with significance of differences in expression. The inferred model of sharing of these differences across conditions is also provided (**B**) List of 188 isomiRs that are differentially expressed between populations. For each isomiR and condition, average ratio in Africans and Europeans is provided together with significance of differences between populations. The inferred model of sharing of these differences across conditions is also provided. (**C**) Tests for population × stimulation interactions across all 4 stimuli for the 244 that are differentially expressed between populations together with the associated. Association statistics (interaction coefficient, *p*-value) are provided for each condition and the best model of sharing of the popDE is reported. (**D**) For each type of miRNA modification and

conditions, changes in rate of occurrence of the modification between populations and their significance. Results are provided for all miRNAs together (any) or focussing on -3p or -5p arm miRNAs.

References

1. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
2. Krol, J., I. Loedige, and W. Filipowicz, *The widespread regulation of microRNA biogenesis, function and decay.* Nat Rev Genet, 2010. **11**(9): p. 597-610.
3. O'Connell, R.M., D.S. Rao, and D. Baltimore, *microRNA regulation of inflammatory responses.* Annu Rev Immunol, 2012. **30**: p. 295-312.
4. Li, Y. and X. Shi, *MicroRNAs in the regulation of TLR and RIG-I pathways.* Cell Mol Immunol, 2013. **10**(1): p. 65-71.
5. Luo, X., et al., *microRNA-mediated regulation of innate immune response in rheumatic diseases.* Arthritis Res Ther, 2013. **15**(2): p. 210.
6. Vigorito, E., et al., *miR-155: an ancient regulator of the immune system.* Immunol Rev, 2013. **253**(1): p. 146-57.
7. Taganov, K.D., et al., *NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses.* Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12481-6.
8. Xiao, C. and K. Rajewsky, *MicroRNA control in the immune system: basic principles.* Cell, 2009. **136**(1): p. 26-36.
9. Siddle, K.J., et al., *A genomic portrait of the genetic architecture and regulatory impact of microRNA expression in response to infection.* Genome Res, 2014. **24**(5): p. 850-9.
10. Siddle, K.J., et al., *bacterial infection drives the expression dynamics of microRNAs and their isomiRs.* PLoS Genet, 2015. **11**(3): p. e1005064.
11. Pai, A.A., et al., *Widespread Shortening of 3' Untranslated Regions and Increased Exon Inclusion Are Evolutionarily Conserved Features of Innate Immune Responses to Infection.* PLoS Genet, 2016. **12**(9): p. e1006338.
12. Zhang, S., et al., *Up-regulation of microRNA-203 in influenza A virus infection inhibits viral replication by targeting DR1.* Sci Rep, 2018. **8**(1): p. 6797.
13. Borel, C., et al., *Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts.* Genome Res, 2011. **21**(1): p. 68-73.
14. Civelek, M., et al., *Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits.* Hum Mol Genet, 2013. **22**(15): p. 3023-37.
15. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

16. Budach, S., M. Heinig, and A. Marsico, *Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci.* Genetics, 2016. **203**(4): p. 1629-40.
17. Huan, T., et al., *Genome-wide identification of microRNA expression quantitative trait loci.* Nat Commun, 2015. **6**: p. 6601.
18. Parts, L., et al., *Extent, causes, and consequences of small RNA expression variation in human adipose tissue.* PLoS Genet, 2012. **8**(5): p. e1002704.
19. Gamazon, E.R., et al., *A genome-wide integrative study of microRNAs in human liver.* BMC Genomics, 2013. **14**: p. 395.
20. Gamazon, E.R., et al., *Genetic architecture of microRNA expression: implications for the transcriptome and complex traits.* Am J Hum Genet, 2012. **90**(6): p. 1046-63.
21. Nedelec, Y., et al., *Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens.* Cell, 2016. **167**(3): p. 657-669 e21.
22. Quach, H., et al., *Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations.* Cell, 2016. **167**(3): p. 643-656 e17.
23. Li, S.C., et al., *miRNA arm selection and isomiR distribution in gastric cancer.* BMC Genomics, 2012. **13 Suppl 1**: p. S13.
24. Neilsen, C.T., G.J. Goodall, and C.P. Bracken, *IsomiRs--the overlooked repertoire in the dynamic microRNAome.* Trends Genet, 2012. **28**(11): p. 544-9.
25. Backes, S., et al., *Degradation of host microRNAs by poxvirus poly(A) polymerase reveals terminal RNA methylation as a protective antiviral mechanism.* Cell Host Microbe, 2012. **12**(2): p. 200-10.
26. Ameres, S.L. and P.D. Zamore, *Diversifying microRNA sequence and function.* Nat Rev Mol Cell Biol, 2013. **14**(8): p. 475-88.
27. Tan, G.C., et al., *5' isomiR variation is of functional and evolutionary importance.* Nucleic Acids Res, 2014. **42**(14): p. 9424-35.
28. Li, L., et al., *The landscape of miRNA editing in animals and its impact on miRNA biogenesis and targeting.* Genome Res, 2018. **28**(1): p. 132-143.
29. de Hoon, M.J., et al., *Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries.* Genome Res, 2010. **20**(2): p. 257-64.
30. Jones, M.R., et al., *Zcchc11-dependent uridylation of microRNA directs cytokine expression.* Nat Cell Biol, 2009. **11**(9): p. 1157-63.
31. Katoh, T., et al., *Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2.* Genes Dev, 2009. **23**(4): p. 433-8.
32. Lee, D., et al., *Poly(A)-specific ribonuclease sculpts the 3' ends of microRNAs.* RNA, 2019. **25**(3): p. 388-405.
33. Cloonan, N., et al., *MicroRNAs and their isomiRs function cooperatively to target common biological pathways.* Genome Biol, 2011. **12**(12): p. R126.
34. Fernandez-Valverde, S.L., R.J. Taft, and J.S. Mattick, *Dynamic isomiR regulation in Drosophila development.* RNA, 2010. **16**(10): p. 1881-8.
35. Yu, F., et al., *Naturally existing isoforms of miR-222 have distinct functions.* Nucleic Acids Res, 2017. **45**(19): p. 11371-11385.
36. Nejad, C., et al., *miR-222 isoforms are differentially regulated by type-I interferon.* RNA, 2018. **24**(3): p. 332-341.
37. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

38. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
39. Burroughs, A.M., et al., *A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness.* Genome Res, 2010. **20**(10): p. 1398-410.
40. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.
41. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.
42. Pasquinelli, A.E., *MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship.* Nature Reviews Genetics, 2012. **13**: p. 271.
43. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
44. Gumienny, R. and M. Zavolan, *Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G.* Nucleic Acids Res, 2015. **43**(3): p. 1380-91.
45. Khorshid, M., et al., *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets.* Nat Methods, 2013. **10**(3): p. 253-5.
46. Wang, L., et al., *Correlation analyses revealed global microRNA-mRNA expression associations in human peripheral blood mononuclear cells.* Mol Genet Genomics, 2018. **293**(1): p. 95-105.
47. Gaidatzis, D., et al., *Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation.* Nat Biotechnol, 2015. **33**(7): p. 722-9.
48. Kim, H., et al., *Bias-minimized quantification of microRNA reveals widespread alternative processing and 3' end modification.* Nucleic Acids Res, 2019. **47**(5): p. 2630-2640.
49. Heo, I., et al., *TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation.* Cell, 2009. **138**(4): p. 696-708.
50. Kim, B., et al., *TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms.* EMBO J, 2015. **34**(13): p. 1801-15.
51. Zhu, L., S.K. Kandasamy, and R. Fukunaga, *Dicer partner protein tunes the length of miRNAs using base-mismatch in the pre-miRNA stem.* Nucleic Acids Res, 2018. **46**(7): p. 3726-3741.
52. Ding, J., V. Tarokh, and Y. Yang, *Model Selection Techniques: An Overview.* IEEE Signal Processing Magazine, 2018. **35**(6): p. 16-34.
53. Bazzoni, F., et al., *Induction and regulatory function of miR-9 in human monocytes and neutrophils exposed to proinflammatory signals.* Proc Natl Acad Sci U S A, 2009. **106**(13): p. 5282-7.
54. Wohlers, I., L. Bertram, and C.M. Lill, *Evidence for a potential role of miR-1908-5p and miR-3614-5p in autoimmune disease risk using integrative bioinformatics.* J Autoimmun, 2018. **94**: p. 83-89.
55. Xiao, J., et al., *miR-429 regulates alveolar macrophage inflammatory cytokine production and is involved in LPS-induced acute lung injury.* Biochem J, 2015. **471**(2): p. 281-91.
56. Tian, H. and Z. He, *miR-215 Enhances HCV Replication by Targeting TRIM22 and Inactivating NF-kappaB Signaling.* Yonsei Med J, 2018. **59**(4): p. 511-518.
57. Najafi-Shoushtari, S.H., et al., *MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis.* Science, 2010. **328**(5985): p. 1566-9.

58. Zuber, V. and K. Strimmer, *High-Dimensional Regression and Variable Selection Using CAR Scores*, in *Statistical Applications in Genetics and Molecular Biology*. 2011.
59. Seeley, J.J., et al., *Induction of innate immune memory via microRNA targeting of chromatin remodelling factors.* Nature, 2018. **559**(7712): p. 114-119.
60. Quach, H., et al., *Signatures of purifying and local positive selection in human miRNAs.* Am J Hum Genet, 2009. **84**(3): p. 316-27.
61. Rantalainen, M., et al., *MicroRNA expression in abdominal and gluteal adipose tissue is associated with mRNA expression levels and partly genetically driven.* PLoS One, 2011. **6**(11): p. e27338.
62. Freimer, J.W., T.J. Hu, and R. Blelloch, *Decoupling the impact of microRNAs on translational repression versus RNA degradation in embryonic stem cells.* Elife, 2018. **7**.
63. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
64. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
65. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
66. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.
67. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.
68. de Rie, D., et al., *An integrated expression atlas of miRNAs and their promoters in human and mouse.* Nat Biotechnol, 2017. **35**(9): p. 872-878.
69. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature Protocols, 2012. **7**: p. 562.
70. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-3.
71. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
72. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software; Vol 1, Issue 1 (2010), 2010.
73. Hofner, B., L. Boccuto, and M. Goker, *Controlling false discoveries in high-dimensional situations: boosting with stability selection.* BMC Bioinformatics, 2015. **16**: p. 144.

6.3 Résumé des résultats

Nous observons de nombreux effets des stimulations immunitaires sur la diversité des miARN, au point que les deux premières composantes principales de l'expression des miARN, représentant respectivement 11,9% et 4,8% de la variabilité observée, correspondent respectivement à la différence entre une stimulation par les TLR ou par la grippe et à la présence de stimulation. De manière intéressante, ces deux PCs corrèlent également avec la population, soulignant une potentielle différence de réponse entre les deux groupes testés. Près d'un tiers des miARN observés montrent une augmentation de leur expression dans au moins une condition. La variabilité d'usage des isomiR a également été impactée par les stimulations.

Nous avons également identifié de nombreux miARN dont l'expression est influencée par la présence d'un SNP. De manière intéressante, dans la majorité des cas, l'effet de la mutation est présente dans les 5 conditions d'observation. Pour ces miARN, les différences de fréquence de ces mutations entre populations expliquent ∼60% des différences entre populations, cependant les miARN les plus différenciés entre populations sont également enrichis en réponse dans les conditions TLR.

Sur le point de la régulation des gènes par les miARN, nous observons que la majorité des corrélations entre l'expression de gènes et l'expression des miARN proviendrait d'une co-transcription entre les miARN et les gènes, ainsi les miARN réguleraient plus la traduction que l'expression, bien que ce dernier point demande des recherches supplémentaires.

6.4 Discussion

La différence de réponse entre populations est un point intéressant. Si pour une partie des miARN différemment exprimés entre populations nous avons identifié les bases génétiques de ces différences d'expression, de nombreux miARN différentiellement exprimés ne sont pas sous contrôle génétique. Ces différences d'expression non-expliquées par la génétique pourraient découler de différences d'exposition aux pathogènes ou à d'autres facteurs environnementaux entre les individus d'ascendance européenne et africaine. De telles différences pourraient affecter la pré-activation des monocytes par des mécanismes d'entraînements immunitaires et de mémoire epigénétique (Quintin, Cheng, van der Meer, & Netea, 2014 ; Pacis et al., 2015).

L'observation d'un faible effet des miARNs sur l'expression des gènes, au moins dans un cadre général, est un point délicat, mais qui n'est pas incohérent avec plusieurs aspects de la littérature. Ainsi, il avait été observé que les corrélations entre l'expression des gènes et des miARN n'étaient pas nécessairement négatives (Siddle et al., 2014 ; Lappalainen et al., 2013 ; Parts et al., 2012 ; Rantalainen et al., 2011). De plus, il a été observé que la désactivation de la régulation des miARN dans des cellules souches embryonnaires aboutit à une augmentation de la traduction, sans impact sur l'expression des gènes (Freimer, Hu, & Blelloch, 2018). Il semblerait ainsi que les impacts des miARN sur l'expression sont beaucoup moins fréquents qu'il n'était supposé, bien que ceux-ci aient déjà été observés dans certains cas particuliers.

7. Discussion générale

7.1 La diversité de l'expression et de sa régulation

Sur de nombreux points, notre connaissance de la diversité de l'expression augmente énormément. La plasticité de l'expression au sein même d'un individu commence à être beaucoup plus étudiée. En effet, le nombre de tissus où l'expression est mesurée, ainsi que différents environnements pouvant affecter leur expression, a grandement augmenté (GTEx Consortium, 2013 ; Quach et al., 2016). De plus, plusieurs études se penchent vers une compréhension plus fine encore que le niveau du tissu en allant étudier les variations d'expression au niveau cellulaire (Liu & Trapnell, 2016), parfois associées à des méthodes de déconvolution afin d'extraire plus d'informations d'une base de données n'ayant pas utilisé cette technologie (Donovan, Antonio-Chronowska, Antonio, & Frazer, 2019) .

La variabilité de l'expression entre individus quant à elle est de plus en plus étudiée, notamment avec l'inclusion de multiples populations dans les études (Quach et al., 2016 ; Mogil et al., 2018 ; Kelly et al., 2017) permettant non seulement une meilleure caractérisation de la variabilité des profils d'expression entre populations humaines, mais permettant aussi, dans le cas des études eQTL, d'identifier plus précisément les variants qui contrôlent l'expression des gènes.

Il est cependant crucial de préciser que la diversité de l'expression chez l'humain reste sous-estimée sur tous les fronts. En effet, non seulement, les études étudiant la transcription ne peuvent pas couvrir la totalité des individus (même si certains tissus ont été étudiés dans de nombreuses populations (Kelly et al., 2017)), mais il est bien évidemment impossible d'étudier la variabilité inter tissus chez des êtres humains vivants. Ainsi la base de données GTEx a été construite en utilisant les tissus de donneurs récemment décédés, le plus souvent de maladie cardiaque (GTEx Consortium, 2013 ; GTEx Consortium et al., 2017).

Enfin, il est également délicat d'étudier la diversité de l'expression en fonction de l'environnement. Si des projets tels que le Milieu Intérieur visent à mieux comprendre l'interjeu entre l'environnement et la biologie (Piasecka et al., 2018 ; Thomas et al., 2015), les choix de sélection des cohortes peuvent limiter la variabilité observée. Ainsi, dans le cas du Milieu Intérieur, le fait que tous les individus concernés soient en bonne santé restreint les observations, et donc sous-estime la véritable diversité de l'expression présente au sein de la population.

Ces limitations sont également valables au sujet de la diversité des régions régulatrices. Ainsi, de la même manière, la diversité entre tissus (Fantom Consortium et al., 2014 ; Road-

map Epigenomics Consortium et al., 2015 ; Backenroth et al., 2018), due à des situations particulières, telles que les infections (Alasoo et al., 2019, 2018), et inter-individuelles sont étudiées (Helmy, Hatlen, & Marco, 2019), mais ne peuvent pas représenter la totalité de la diversité présente au sein de l'humanité.

Il est important de noter que ces limitations sont nécessaires à la création de résultats interprétables, et ce paragraphe n'a pas pour but de dénigrer tous les efforts qui sont fournis dans la formation d'une image précise de la variabilité d'expression génique humaine. Cependant, nous devons également garder en tête que des mesures telles que l'héritabilité dépendent de la variabilité présente dans la cohorte observée, et peuvent être sur-estimées si la variabilité globale est sous-estimée.

7.2 Difficultés d'étude de l'introgression archaïque

7.2.1 Des signatures similaires à celles de la sélection

Une des difficultés majeures lors de l'étude de l'introgression archaïque est la nature des traces qu'elle laisse dans le génome, et le fait que ces marques sont souvent utilisées comme marqueurs de sélection positive.

En effet, les haplotypes de grandes tailles sont également une marque d'un balayage sélectif et sont notamment utilisés par certaines statistiques telles que l'iHS (Voight, Kudaravalli, Wen, & Pritchard, 2006). Ainsi dans le cas d'une introgression archaïque sans génome archaïque de référence ni groupe non-introgressé, comme c'est le cas pour l'introgression archaïque dans les populations africaines, il n'est pas possible de conclure efficacement à un cas d'introgression adaptative car la taille anormale de l'haplotype pourrait être expliquée par un simple balayage sélectif, sans que l'introgression y joue un rôle. Pour contourner ce problème, Durvasula et collègues (Durvasula & Sankararaman, 2019) utilisent les néandertaliens comme groupe non-introgressé.

La présence d'un génome de référence ou d'un groupe non-introgressé limite grandement ce problème en permettant d'ajouter des preuves d'introgressions indépendantes de la taille de l'haplotype. Cependant d'autres statistiques utilisées pour la sélection peuvent être impactées par une introgression mal caractérisée.

En particulier, les tests qui comparent deux populations entre elles (tels le FST) sont sujets à de mauvaises interprétations si l'histoire exacte de ses deux populations et leur relation à cette introgression n'est pas connue. Et comme le montrent les débats sur les sources de la présence de plus d'haplotypes néandertaliens dans les populations asiatiques que dans les populations européennes, cet historique est extrêmement difficile à obtenir.

Ainsi il n'y a aujourd'hui que les méthodes spécialement dédiées à la détection de l'introgression archaïque (Racimo et al., 2017) ou bien les études extrêmement détaillées d'une région spécifique du génome (Sams et al., 2016) qui peuvent réfuter que l'haplotype concerné évolue de manière neutre, cependant, dans le cas spécifique de l'introgression néandertalienne, cela ne devrait pas être suffisant à prouver une sélection positive.

7.2.2 À la limite des modèles

Depuis 2014, il a été observé que les régions étant sujettes à la sélection de fond ont moins d'haplotypes néandertaliens que les autres (Sankararaman et al., 2014), cela étant dû à l'accumulation dans la population néandertalienne de mutations faiblement délétères (Harris & Nielsen, 2016 ; Petr et al., 2019).

Cependant, malgré le fait que l'impact de la sélection de fond sur l'introgression est accepté par la communauté, celle-ci est très rarement prise en compte lors des études de l'introgression archaïque. Ce n'est pas sans raison : il est difficile de simuler la sélection de fond de manière précise, notamment car cela demanderait une connaissance de la répartition des coefficients de sélection des allèles de chaque région du génome.

Au delà des simulations, la seule mesure disponible de la sélection de fond, la statistique B (McVicker et al., 2009), est basée sur la conservation des gènes au niveau des mammifères. Ainsi, au niveau des régions où la pression de sélection n'est pas la même chez les hominines que chez la majorité des mammifères, cette mesure est biaisée. Une étude précise nécessiterait une mesure de la sélection de fond au niveau des hominines, ou hominidés.

Une difficulté supplémentaire provient du fait que les effets de la sélection de fond sur les divergences entre les HAM et les néandertaliens n'ont pas été étudiés. Cependant, la sélection de fond va créer, sur une échelle de temps "courte", une absence de divergence, tandis que sur une échelle de temps "longue" elle va au contraire conduire à une accumulation de divergences (McVicker et al., 2009). Ainsi, non seulement la fréquence des haplotypes néandertaliens est impactée par cette force, mais leur diversité l'est également.

Malgré ces difficultés majeures rendant une étude précise compliquée, cet effet devrait être pris en compte dans les possibles interprétations des études de l'introgression archaïque. Ainsi, une région pourrait montrer un comportement incompatible avec la neutralité, par exemple une fréquence élevée (Sams et al., 2016), à cause d'un effet faible de la sélection de fond dans cette région particulière, ou plus généralement, à une purge moins efficace des haplotypes néandertaliens locaux, potentiellement due à une densité de divergence avec l'humain différente.

De manière plus générale, il semblerait que l'introgression néandertalienne soit un phénomène qui pousse les limites des modèles de génétique humaine. Non seulement l'hypothèse de neutralité ne peut pas être faite, mais plusieurs hypothèses simplificatrices telles que les introgressions en une génération ont également été questionnées (Sams et al., 2016). Plus récemment, l'étude des différences d'introgression néandertalienne entre les populations européennes et asiatiques a conduit à la conclusion préliminaire d'un temps de génération différent entre ces deux populations (résultats temporaires présentés par Mikkel Schierrup à la conférence Biology of Genomes 2019). Si ces résultats sont confirmés, et que leur impact s'avère non négligeable, cela impliquerait une modification complète des modèles mathématiques sous-tendant la majorité des études de génétique des populations humaines.

Ces difficultés rendent le sujet de l'introgression néandertalienne intéressant au delà même des implications directes du sujet. Ces obstacles sont également des opportunités d'améliorer la manière dont la génétique des populations humaines se construit, illustrés par les

conséquences de la petite taille de population néandertalienne sur l'introgression (Harris & Nielsen, 2016) qui ont pavé le chemin vers une étude plus systématique des effets des tailles de populations sur les dynamiques d'introgression (Kim et al., 2018).

7.3 De la simplification à l'erreur

L'introgression néandertalienne est aussi un sujet particulier car il a entraîné une forme de "mode". Cependant, de manière intéressante, plusieurs fausses notions ont circulé sur plusieurs aspects de cet évènement démographique. En tant qu'étude de cas, je vais revenir sur les résultats concernant la présence d'introgression néandertalienne dans les régions géniques, pour récapituler les connaissances actuelles, avant de montrer comment des glissements de sens progressifs ont pu être sources d'erreurs dans la communauté scientifique à ce sujet, et mener à d'apparentes contradictions.

7.3.1 Présence d'haplotypes néandertaliens dans les régions géniques

En 2014 lorsque les premières cartes de l'introgression néandertalienne ont été publiées, Sankararaman et ses collègues ont observé l'association entre l'hérédité néandertalienne et la sélection de fond (Sankararaman et al., 2014). On notera qu'à aucun moment dans l'article, un lien entre la sélection de fond et une quelconque fonctionnalité ou le fait d'être dans une région génique n'est ne serait-ce que supposé.

Parallèlement, Vernot et collègues commentent la présence d'introgression néandertalienne dans les gènes codants des protéines : "*∼26% of all protein-coding genes had one or more exons that overlapped a Neandertal sequence*" (Vernot & Akey, 2014). On notera que cette mesure est très différente de celle utilisée par Sankararaman et collègue, mais la comparaison avec le fait que ∼20% du génome néandertalien peut être retrouvé dans les génomes modernes, présenté dans le même article, suggère une forte présence d'introgression néandertalienne dans les régions exoniques.

En 2016, Fu et collègues ont commenté, dans un article abordant l'histoire de l'Europe : *"We [observe] that the decrease in Neanderthal-derived alleles [through time] is more marked near genes than in less constrained regions of the genome"* (Fu et al., 2016). Bien que la mesure utilisée soit la même mesure de la sélection de fond que Sankararaman et collègues, les auteurs ici font une opposition entre les gènes et les régions moins contraintes, inappropriée au vu du test utilisé (McVicker et al., 2009).

En 2017, Danneman et collègues étudient, entre autre, l'évolution des fréquences de mutations non-synonymes introgressées dans les populations eurasiennes. Bien que le texte précise que près de 2/3 d'entre elles montrent une tendance non significative à l'augmentation de fréquence, la figure associée souligne le fait qu'aucun allèle non-synonyme introgressé n'ait d'augmentation de fréquence significative sur la période de temps étudié. (Dannemann et al., 2017).

Enfin en 2019, Petr et collègues ont estimé l'ascendance néandertalienne par type de régions fonctionnelle dans les génomes du *Simons Genome Diversity Project* (à l'exception

des génomes Océaniens) (Petr et al., 2019 ; Mallick et al., 2016), et commentent la présence d'haplotypes néandertaliens dans les régions codantes du génome ainsi : *"in seeming contrast with previous studies, we observed no significant depletion of Neandertal ancestry in CDS compared with intronic and intergenic regions"*, les études citées étant celles de Sankararaman et collègues, et de Fu et collègues, précédemment décrites.

7.3.2 Explication des confusions

Avant la clarification de Petr, il y avait effectivement l'idée dans la communauté scientifique que l'introgression néandertalienne était peu présente dans les régions génique. Il est donc intéressant, voire nécessaire, de comprendre comment a-t-on pu prendre pour acquis un résultat, non seulement faux, mais tout simplement jamais testé et incohérent avec plusieurs papiers.

Une partie de réponse est due à une sur-interprétation des résultats : l'observation entre ascendance néandertalienne et sélection de fond qui a pris, petit à petit, des notions de fonctionnalités.

Une deuxième partie provient de l'utilisation de métriques différentes, interprétées de manière similaire. Ainsi nous avons des études s'intéressant à l'ascendance néanderthalienne moyenne avec Petr et Sankararaman, mais également des études se concentrant sur la présence ou l'absence d'haplotypes avec Akey. Enfin Danneman qui, soit étudie directement les mutations néandertaliennes, soit les compare à des ensembles similaires de mutations non introgressées. Chacune de ces statistiques mesure un aspect différent de l'introgression néandertalienne, cependant, un haut score de ces statistiques sera trop souvent simplifié en "excès d'introgression néandertalienne".

Ces deux causes d'erreurs ont un point en commun intéressant, elles proviennent toutes deux d'une simplification des résultats. Et bien que la simplification des concepts soit nécessaire à la communication, dans ce cas précis, elle a mené à des interprétations erronées.

7.4 Communication scientifique et pseudo-science

Cette confusion sur les résultats des articles scientifiques m'amène à un point plus général, la manière dont les résultats scientifiques sont interprétés, et repris par le public. Pour ce dernier point de discussion, je vais donc m'éloigner des chemins parcourus au sein du laboratoire, et aborder un autre aspect de la vie de chercheur : le contact avec le public. Et quitte à rédiger une pièce d'opinion, j'aimerais préfacer cette partie par ces mots : la science est politique.

7.4.1 Liens entre science et politique

Les liens entre politique et science sont extrêmement nombreux. Les gouvernements ont un impact massif sur la création de la recherche, mais également sur sa diffusion. Une quantité non négligeable de recherche, y compris toute la recherche effectuée au cours de cette thèse, est financée par des organisations gouvernementales. Un exemple de recherche fortement

financée pour des raisons politiques est la course spatiale qui a eu lieu pendant la guerre froide, mais plusieurs exemples de considérations politiques conduisant à un ralentissement de la recherche existent.

Un exemple récent est apparu pendant l'épidémie de SIDA aux États-Unis dans les années 80. A cette période, la maladie étaient comprise comme étant une maladie de personnes homosexuelles, principalement des hommes gays, et en réponse à cette épidémie le gouvernement a demandé au *Center for Disease Control* : "Look pretty and do as little as possible" (Francis, 2012). Plusieurs médecins, dont notamment Everett Koop se sont opposés à cette décision, cependant cette décision a sans aucun doute eu des conséquences sur la diffusion du VIH dans les premières années de l'épidémie.

Cependant, la science est aussi politique parce qu'elle informe des décisions politiques. Que ce soit au niveau gouvernemental comme au niveau personnel. La question du changement climatique, de plus en plus présente en politique est un bon exemple, ce phénomène étant difficilement identifiable au niveau individuel.

Ce dernier point est ainsi crucial, en effet, si la représentation du monde d'une partie du public dépend des résultats scientifiques, il est indispensable de communiquer clairement les résultats, mais également de vérifier que ceux ci ne sont pas détournés afin de servir une idéologie.

7.4.2 Le cas de la génétique

La génétique a été très reliée au racisme et à l'eugénisme de manière historique, et ce spectre est toujours assez présent dans la sphère publique. Et je ne vise pas ici à faire un historique, ou même une étude détaillée sur le sujet, mais de montrer que l'argument génétique est encore aujourd'hui utilisé pour soutenir des agendas politiques ou une stratégie de communication, souvent de manière eugéniste ou raciste.

Récemment, plusieurs personnes clamant la supériorité de la "race blanche", ont été filmés, en groupe, à boire une grande quantité de lait. Le lien, très bancal, avec la génétique ici serait fait avec la persistance de la lactase, supposément un trait présent "chez les blancs". Cette lecture simpliste demande d'oublier que beaucoup de populations, y compris certaines n'étant pas "blanches" peuvent consommer du lait, et que beaucoup de "blancs" ne le peuvent pas ce qui nous informe sur un point très important : les faits comptent moins que l'air vaguement scientifique qui les entoure.

L'actuel président des États Unis, Donald Trump, utilise régulièrement l'excuse de ses "bons gènes" afin de justifier de nombreuses choses, notamment le succès familial, en parlant du "gène gagnant". Au vu de l'énorme impact médiatique de ce dernier, je pense pouvoir déclarer que la simplification énorme de la génétique est un problème répandu.

7.4.3 Prises de positions

Il est important de noter que dans les dernières années, plusieurs institutions scientifiques se sont clairement positionnées contre le racisme. Ainsi plusieurs éditoriaux sur le blog de *Nature* ont clairement adressé le problème, déclarant : "*Two recommendations can be made*

for the public behaviour of scientists and scholars. The first : give ample credit to the insight of complementary disciplines. The second : refute statements that misconstrue what your insights actually reveal and that can be used politycally to justify direspect, or worse, to groups of people.", et en allant plus loin en aout 2017 en déclarant "*There is nothing in any data anywhere that can excuse or justify policies that discriminate against the potentials of individuals or that systematicaly reinforce different roles and status in society for people of any gender or ethnic group*". l'American Society of Human Genetics à également déclarée qu'elle est "*alarmed to see a societal resurgence of groups rejecting the value of genetic diversity and using discredited or distorted genetic concepts to bolster bogus claims of white supremacy. ASHG denounces this misuse of genetics to feed racist ideologies. In public dialog, our research community should be clear about genetic knowledge related to ancestry and genomic diversity*".

Plusieurs personnes, au delà des grandes institutions, ont également cherché à combattre l'utilisation de la science comme justification du racisme. C'est notamment le cas d'Adam Rutherford qui a tenu plusieurs conférences abordant le sujet, ou encore d'Angela Saini qui a publié récemment un livre sur le sujet (Saini, 2019).

À un niveau plus accessible à chaque chercheur, lors de la publication d'un article traitant de sélection polygénique, Racimo et collègues ont également publié une Foire Aux Questions simplifiant les résultats et les possibles implications de manière compréhensible à tous (Racimo, Berg, & Pickrell, 2018), et remettant en contexte les signaux de sélection polygénique sur le phénotype *Educational Attainment* dans des populations d'Asie de l'est. Ce document a l'intérêt majeur de ne pas éviter le sujet et de répondre aux questions que peuvent soulever ce genre de découvertes. Il est cependant regrettable qu'un lien à ce document ne soit pas présent dans l'article scientifique même, mais qu'il ait été diffusé indépendamment, notamment via Twitter.

7.4.4 Ce qu'il reste à faire

Je ne vais pas apporter ici une solution à l'utilisation de la génétique à des fins racistes ou eugénistes. Ces problèmes perdureront probablement pendant plusieurs décennies. Cependant, il me semble important de souligner plusieurs points où les institutions scientifiques permettent involontairement la propagation de ces idées.

La revue par les pairs

Ce point est délicat, car la solidité des publications scientifiques repose aujourd'hui sur la revue par les pairs. Cependant, le système permet aux éditeurs de choisir les personnes évaluant la qualité d'une publication scientifique, et ainsi de créer une communauté au sein d'un journal, où les chercheurs s'entre-évaluent. Dans ce système, un groupe de publication politiquement orienté peut voir le jour, et si la notion de facteur d'impact peut nous permettre de comprendre les différences de qualité entre différentes revues, cette distinction n'est pas nécessairement connue du public.

La vulgarisation scientifique imprudente

Que penser lorsque l'on tombe sur le titre suivant : "In the Nature–Nurture War, Nature Wins" ? Dans cet article de la partie blog de *Scientific American*, Robert Plomin déclare "*DNA differences account for about 50 percent of the differences between us*". Cette déclaration est accompagnée de la citation suivante (la seule de l'article de blog) d'un article de Polderman et collègues, une meta-analyse des études menées sur les jumeaux (Polderman et al., 2015). Je suppose que Plomin fait référence au fait que l'héritabilité moyenne des traits observés est de 49%, cependant, il ne mentionne pas plusieurs limites de cette étude, notamment le manque de données dans certaines parties du monde, ou encore le biais vers certains types de traits (psychologiques).

Même en étant généreux et en oubliant ces problèmes, notons que l'article scientifique comporte une partie nommée "*Equal contribution of genes and environment*", et est utilisé pour justifier un titre clamant "Nature wins". Il s'agit donc, au minimum, d'une exagération des résultats.

Cette exagération peut-être reprise, voire internalisée par plusieurs lecteurs, renforçant une vision déterministique de la nature des individus, vision assez présente dans les mouvements racistes.

Face à ces simplifications, nous devons répondre, et nous devons répondre rapidement, ce qui a été fait, à cet article comme à d'autres déclarations du même auteur. Notons cependant que l'article de blog de Robert Plomin est tout de même mieux référencé par google, étant le premier résultat à une recherche sur le terme exact "Nature-Nurture War".

Un devoir d'éducation

L'impétus derrière l'écriture de cette partie sur l'utilisation de la science dans les mouvements racistes provient d'une expérience personnelle. Quelques mois avant l'écriture de ce manuscrit, lors d'une intervention de vulgarisation sur le sujet de l'Homme de Néanderthal, en abordant le sujet de la disparition des néandertaliens et la colonisation de l'Eurasie par les HAM, j'ai été interrompu par la question suivante : "Comme le Grand Remplacement ?".

Le grand remplacement est une théorie raciste (souvent accompagnée d'une couche de complotisme) selon laquelle il existerait un processus délibéré de remplacement des personnes blanches par des personnes non-blanches, en influençant les changements démographiques.

Le Grand Remplacement était également le titre du manifeste du meutrier de Christchurch, qui a tué 51 personnes en attaquant deux mosquées. Et au "Unite the Right rally" de Charlottesville, en Virginie, où une femme a trouvé la mort, les membres du rally scandaient "Jews will not replace us", en directe référence à cette théorie.

Me rendre compte qu'une personne de mon audience aurait pu interpréter ce que je considérais comme un simple fait de cette manière m'a choqué, mais m'a permis de réaliser que, si les racistes ne sont intéressés par la science que dans ce qu'elle peut avancer leurs opinions politiques, d'autres personnes doutent véritablement, et cherchent véritablement à savoir. Pour ces personnes, nous nous devons qu'elles puissent facilement trouver une information complète et contextualisée du savoir scientifique sur un sujet. Cela passe par l'éducation directe,

mais aussi et surtout en mettant à disposition des ressources compréhensibles, complètes, et facilement partageables.

Lorsque quelqu'un décidera de faire une recherche Google pour répondre à une question, le premier résultat est peut être bien ce qui aura le plus d'impact.

Références

Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S., & Gaffney, D. J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife*, *8*.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., Consortium, H., Hale, C., Dougan, G., & Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet*, *50*(3), 424-431.

Altshuler, D., Donnelly, P., & The International HapMap, C. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299-1320.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jorgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Muller, F., Forrest, A. R. R., Carninci, P., Rehli, M., & Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455-461.

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., Christiano, A., Buxbaum, J. D., & Ionita-Laza, I. (2018). Fun-lda : A latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation : Methods and applications. *Am J Hum Genet*, *102*(5), 920-942.

Bartel, D. P. (2009). Micrornas : target recognition and regulatory functions. *Cell*, *136*(2), 215-233.

Bateson, W. (1909). Heredity and variation in modern lights. *Darwin and modern science*.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., & Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome Res*, *24*(1), 14-24.

Benton, M. L., Talipineni, S. C., Kostka, D., & Capra, J. A. (2018). Genome-wide enhancer maps differ significantly in genomic distribution, evolution, and function. *bioRxiv*, 176610.

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS genetics*, *10*(8), e1004412-e1004412.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*, *74*(6), 1111-20.

Bischoff, J. L., Williams, R. W., Rosenbauer, R. J., Aramburu, A., Arsuaga, J. L., García, N., & Cuenca-Bescós, G. (2007). High-resolution u-series dates from the sima de los huesos hominids yields 600-66+inf kyrs : implications for the evolution of the early

neanderthal lineage. *Journal of Archaeological Science*, *34*(5), 763-770.

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., & Akey, J. M. (2018). Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell*, *173*(1), 53-61 e9.

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., & Parkinson, H. (2018). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005-D1012.

Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., Abney, M., Ober, C., & Eichler, E. E. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, *44*, 1277.

Coyne, J. A., & Orr, H. A. (1989). Patterns of speciation in drosophila. *Evolution*, *43*(2), 362-381.

Dannemann, M., & Kelso, J. (2017). The contribution of neanderthals to phenotypic variation in modern humans. *Am J Hum Genet*, *101*(4), 578-589.

Dannemann, M., Prufer, K., & Kelso, J. (2017). Functional implications of neandertal introgression in modern humans. *Genome Biol*, *18*(1), 61.

Dannemann, M., & Racimo, F. (2018). Something old, something borrowed : admixture and adaptation in human evolution. *Curr Opin Genet Dev*, *53*, 1-8.

de Azevedo, S., González, M. F., Cintas, C., Ramallo, V., Quinto-Sánchez, M., Márquez, F., Hünemeier, T., Paschetta, C., Ruderman, A., Navarro, P., Pazos, B. A., Silva de Cerqueira, C. C., Velan, O., Ramírez-Rozzi, F., Calvo, N., Castro, H. G., Paz, R. R., & González-José, R. (2017). Nasal airflow simulations suggest convergent adaptation in neanderthals and modern humans. *Proceedings of the National Academy of Sciences*, *114*(47), 12442.

de Hoon, M. J. L., Taft, R. J., Hashimoto, T., Kanamori-Katayama, M., Kawaji, H., Kawano, M., Kishima, M., Lassmann, T., Faulkner, G. J., Mattick, J. S., Daub, C. O., Carninci, P., Kawai, J., Suzuki, H., & Hayashizaki, Y. (2010). Cross-mapping and the identification of editing sites in mature micrornas in high-throughput sequencing libraries. *Genome Research*, *20*(2), 257-264.

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J. L., Patin, E., & Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am J Hum Genet*, *98*(1), 5-21.

Dibble, H. L., Sandgathe, D., Goldberg, P., McPherron, S., & Aldeias, V. (2018). Were western european neandertals able to make fire ? *Journal of Paleolithic Archaeology*, *1*(1), 54-79.

Dobzhansky, T. (1982). *Genetics and the origin of species* (Vol. 11). Columbia university press.

Donovan, M. K. R., Antonio-Chronowska, A., Antonio, M., & Frazer, K. A. (2019). Cellular deconvolution of gtex tissues powers eqtl studies to discover thousands of novel disease and cell-type associated regulatory variants. *bioRxiv*, 671040.

Durvasula, A., & Sankararaman, S. (2019). Recovering signals of ghost archaic introgression in african populations. *bioRxiv*, 285734.

Enard, D., & Petrov, D. A. (2018). Evidence that rna viruses drove adaptive introgression between neanderthals and modern humans. *Cell*, *175*(2), 360-371 e13.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). Microrna

targets in drosophila. *Genome Biol*, *5*(1), R1.

Estalrrich, A., El Zaatari, S., & Rosas, A. (2017). Dietary reconstruction of the el sidrón neandertal familial group (spain) in the context of other neandertal and modern hunter-gatherer groups. a molar microwear texture analysis. *Journal of Human Evolution*, *104*, 13-22.

Eyre-Walker, A., & Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, *397*(6717), 344-347.

Fang, L., Ahn, J. K., Wodziak, D., & Sibley, E. (2012). The human lactase persistence-associated snp -13910*t enables in vivo functional persistence of lactase promoter-reporter transgene expression. *Hum Genet*, *131*(7), 1153-9.

Fantom Consortium, the, R. P., Clst, Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jorgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. A., Plessy, C., Vitezic, M., Severin, J., Semple, C., Ishizu, Y., Young, R. S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. A., Arner, P., Babina, M., Rennie, S., Balwierz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A. M., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. A., Detmar, M., Diehl, A. D., Dohi, T., Drablos, F., Edge, A. S., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462-70.

Francis, D. P. (2012). Deadly aids policy failure by the highest levels of the us government : A personal look back 30 years later for lessons to respond better to future epidemics. *Journal of Public Health Policy*, *33*(3), 290-300.

Fraser, H. B. (2013). Gene expression drives local adaptation in humans. *Genome Res*, *23*(7), 1089-96.

Freimer, J. W., Hu, T. J., & Blelloch, R. (2018). Decoupling the impact of micrornas on translational repression versus rna degradation in embryonic stem cells. *Elife*, *7*.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwangler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., Skoglund, P., Derevianko, A. P., Drozdov, N., Slavinsky, V., Tsybankov, A., Cremonesi, R. G., Mallegni, F., Gely, B., Vacca, E., Morales, M. R., Straus, L. G., Neugebauer-Maresch, C., Teschler-Nicola, M., Constantin, S., Moldovan, O. T., Benazzi, S., Peresani, M., Coppola, D., Lari, M., Ricci, S., Ronchitelli, A., Valentin, F., Thevenet, C., Wehrberger, K., Grigorescu, D., Rougier, H., Crevecoeur, I., Flas, D., Semal, P., Mannino, M. A., Cupillard, C., Bocherens, H., Conard, N. J., Harvati, K., Moiseyev, V., Drucker, D. G., Svoboda, J., Richards, M. P., Caramelli, D., Pinhasi, R., Kelso, J., Patterson, N., Krause, J., Paabo, S., & Reich, D. (2016). The genetic history of ice age europe. *Nature*, *534*(7606), 200-5.

Good, J. M., Giger, T., Dean, M. D., & Nachman, M. W. (2010). Widespread over-expression of the x chromosome in sterile f1 hybrid mice. *PLoS genetics*, *6*(9), e1001148.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen,

J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., & Paabo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*(5979), 710-722.

Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L. F., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajković, D., Kućan, Z., Gusić, I., Wikström, M., Laakkonen, L., Kelso, J., Slatkin, M., & Pääbo, S. (2008). A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, *134*(3), 416-426.

GTEx Consortium. (2013). The genotype-tissue expression (gtex) project. *Nat Genet*, *45*(6), 580-5.

GTEx Consortium, Laboratory, D. A., Coordinating Center Analysis Working, G., Statistical Methods groups Analysis Working, G., Enhancing, G. g., Fund, N. I. H. C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E. B. I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U. o. C. S. C., Lead, a., Laboratory, D. A., Coordinating, C., management, N. I. H. p., Biospecimen, c., Pathology, e, Q. T. L. m. w. g., Battle, A., Brown, C. D., Engelhardt, B. E., & Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204-213.

Haerty, W., & Singh, R. S. (2006). Gene regulation divergence is a major contributor to the evolution of dobzhansky–muller incompatibilities between species of drosophila. *Molecular Biology and Evolution*, *23*(9), 1707-1714.

Harris, K., & Nielsen, R. (2016). The genetic cost of neanderthal introgression. *Genetics*, *203*(2), 881-91.

Harvati, K., Röding, C., Bosman, A. M., Karakostis, F. A., Grün, R., Stringer, C., Karkanas, P., Thompson, N. C., Koutoulidis, V., Moulopoulos, L. A., Gorgoulis, V. G., & Kouloukoussa, M. (2019). Apidima cave fossils provide earliest evidence of homo sapiens in eurasia. *Nature*.

Helmy, M., Hatlen, A., & Marco, A. (2019). The impact of population variation in the analysis of microrna target sites. *Noncoding RNA*, *5*(2).

Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruiz, C., Bergman, C., Boitard, C., Boscato, P., Caparros, M., Conard, N. J., Draily, C., Froment, A., Galvan, B., Gambassini, P., Garcia-Moreno, A., Grimaldi, S., Haesaerts, P., Holt, B., Iriarte-Chiapusso, M. J., Jelinek, A., Jorda Pardo, J. F., Maillo-Fernandez, J. M., Marom, A., Maroto, J., Menendez, M., Metz, L., Morin, E., Moroni, A., Negrino, F., Panagopoulou, E., Peresani, M., Pirson, S., de la Rasilla, M., Riel-Salvatore, J., Ronchitelli, A., Santamaria, D., Semal, P., Slimak, L., Soler, J., Soler, N., Villaluenga, A., Pinhasi, R., & Jacobi, R. (2014). The timing and spatiotemporal patterning of neanderthal disappearance. *Nature*, *512*(7514), 306-9.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, *106*(23),

9362-7.

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., Alcolea-González, J. J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz, J., Weniger, G. C., & Pike, A. W. G. (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, *359*(6378), 912.

Hublin, J. J. (2017). The last neanderthal. *Proc Natl Acad Sci U S A*, *114*(40), 10520-10522.

Huerta-Sanchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., & Nielsen, R. (2014). Altitude adaptation in tibetans caused by introgression of denisovan-like dna. *Nature*, *512*(7513), 194-7.

Ingram, C. J., Mulcare, C. A., Itan, Y., Thomas, M. G., & Swallow, D. M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet*, *124*(6), 579-91.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Varnai, C., Thiecke, M. J., Burden, F., Farrow, S., Cutler, A. J., Rehnstrom, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., Consortium, B., Stunnenberg, H. G., Todd, J. A., Zerbino, D. R., Stegle, O., Ouwehand, W. H., Frontini, M., Wallace, C., Spivakov, M., & Fraser, P. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, *167*(5), 1369-1384 e19.

Jones, M. R., Quinton, L. J., Blahna, M. T., Neilson, J. R., Fu, S., Ivanov, A. R., Wolf, D. A., & Mizgerd, J. P. (2009). Zcchc11-dependent uridylation of microrna directs cytokine expression. *Nat Cell Biol*, *11*(9), 1157-63.

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S., Baba, T., & Suzuki, T. (2009). Selective stabilization of mammalian micrornas by 3' adenylation mediated by the cytoplasmic poly(a) polymerase gld-2. *Genes Dev*, *23*(4), 433-8.

Kelly, D. E., Hansen, M. E. B., & Tishkoff, S. A. (2017). Global variation in gene expression and the value of diverse sampling. *Curr Opin Syst Biol*, *1*, 102-108.

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2018). Deleterious variation shapes the genomic landscape of introgression. *PLoS Genet*, *14*(10), e1007741.

King, W. (1864). The reputed fossil man of the neanderthal. *The Quarterly journal of science.*, *1*, 88-97.

Kozomara, A., & Griffiths-Jones, S. (2010). mirbase : integrating microrna annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(suppl1), D152-D157.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., & Pääbo, S. (2010). The complete mitochondrial dna genome of an unknown hominin from southern siberia. *Nature*, *464*, 894.

Krause, J., Orlando, L., Serre, D., Viola, B., Prufer, K., Richards, M. P., Hublin, J. J., Hanni, C., Derevianko, A. P., & Paabo, S. (2007). Neanderthals in central asia and siberia. *Nature*, *449*(7164), 902-4.

Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H., & Pääbo, S. (1999). Dna sequence of the mitochondrial hypervariable region ii from the neandertal type specimen. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(10), 5581-5585.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., & Paabo, S. (1997). Neandertal dna sequences and the origin of modern humans. *Cell*, *90*(1), 19-30.

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans : Implications for complex disease and association studies.

American Journal of Human Genetics, *80*(4), 727-739.

Kudaravalli, S., Veyrieras, J. B., Stranger, B. E., Dermitzakis, E. T., & Pritchard, J. K. (2009). Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol*, *26*(3), 649-58.

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Marques-Bonet, T., Andres, A. M., Viola, B., Paabo, S., Meyer, M., Siepel, A., & Castellano, S. (2016). Ancient gene flow from early modern humans into eastern neanderthals. *Nature*, *530*(7591), 429-33.

Lakhani, C. M., Tierney, B. T., Manrai, A. K., Yang, J., Visscher, P. M., & Patel, C. J. (2019). Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nature Genetics*, *51*(2), 327-334.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis, C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Hasler, R., Syvanen, A. C., van Ommen, G. J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I. G., Estivill, X., & Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506-11.

Law, P. J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., Farrington, S., Svinti, V., Palles, C., Orlando, G., Sud, A., Holroyd, A., Penegar, S., Theodoratou, E., Vaughan-Shaw, P., Campbell, H., Zgaga, L., Hayward, C., Campbell, A., Harris, S., Deary, I. J., Starr, J., Gatcombe, L., Pinna, M., Briggs, S., Martin, L., Jaeger, E., Sharma-Oates, A., East, J., Leedham, S., Arnold, R., Johnstone, E., Wang, H., Kerr, D., Kerr, R., Maughan, T., Kaplan, R., Al-Tassan, N., Palin, K., Hänninen, U. A., Cajuso, T., Tanskanen, T., Kondelin, J., Kaasinen, E., Sarin, A.-P., Eriksson, J. G., Rissanen, H., Knekt, P., Pukkala, E., Jousilahti, P., Salomaa, V., Ripatti, S., Palotie, A., Renkonen-Sinisalo, L., Lepistö, A., Böhm, J., Mecklin, J.-P., Buchanan, D. D., Win, A.-K., Hopper, J., Jenkins, M. E., Lindor, N. M., Newcomb,

P. A., Gallinger, S., Duggan, D., Casey, G., Hoffmann, P., Nöthen, M. M., Jöckel, K.-H., Easton, D. F., Pharoah, P. D. P., Peto, J., Canzian, F., Swerdlow, A., Eeles, R. A., Kote-Jarai, Z., Muir, K., Pashayan, N., Henderson, B. E., Haiman, C. A., Schumacher, F. R., Al Olama, A. A., Benlloch, S., Berndt, S. I., Conti, D. V., Wiklund, F., Chanock, S., Gapstur, S., Stevens, V. L., Tangen, C. M., Batra, J., Clements, J., Gronberg, H., Schleutker, J., Albanes, D., Wolk, A., West, C., Mucci, L., Cancel-Tassin, G., Koutros, S., et al. (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nature Communications*, *10*(1), 2154.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E. R., Roodenberg, S. A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J. M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S. M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D. A., O'Reilly, S., Richards, M. B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J. F., Yengo, L., Hovhannisyan, N. A., Patterson, N., Pinhasi, R., & Reich, D. (2016). Genomic insights into the origin of farming in the ancient near east. *Nature*, *536*, 419.

Lee, D., Park, D., Park, J. H., Kim, J. H., & Shin, C. (2019). Poly(a)-specific ribonuclease sculpts the 3' ends of micrornas. *RNA*, *25*(3), 388-405.

Lewinsky, R. H., Jensen, T. G., Moller, J., Stensballe, A., Olsen, J., & Troelsen, J. T. (2005). T-13910 dna variant associated with lactase persistence interacts with oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet*, *14*(24), 3945-53.

Li, L. S., Song, Y. L., Shi, X. R., Liu, J. H., Xiong, S. L., Chen, W. Y., Fu, Q., Huang, Z. C., Gu, N. N., & Zhang, R. (2018). The landscape of mirna editing in animals and its impact on mirna biogenesis and targeting. *Genome Research*, *28*(1), 132-143.

Liu, S., & Trapnell, C. (2016). Single-cell transcriptome sequencing : recent advances and remaining challenges. *F1000Res*, *5*.

Loh, Y. H., Yi, S. V., & Streelman, J. T. (2011). Evolution of micrornas and the diversification of species. *Genome Biol Evol*, *3*, 55-65.

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A. F., Grarup, N., Jørgensen, T., Jiang, T., Witte, D. R., Sandbæk, A., Hellmann, I., Lauritzen, T., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS genetics*, *7*(10), e1002326-e1002326.

Mack, K. L., & Nachman, M. W. (2017). Gene regulation and speciation. *Trends Genet*, *33*(1), 68-80.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J. P., Song, Y. S., Poletti, G., Balloux, F., van Driem, G., de Knijff, P., Romero, I. G., Jha, A. R., Behar, D. M., Bravi, C. M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O. L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M. S., Ruiz-Linares, A., Beall, C. M., Di Rienzo, A., Jeong, C., Starikovskaya, E. B., Metspalu, E., Parik, J., Villems, R., Henn, B. M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J. T. S., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M. F., Kivisild, T., Klitz, W., Winkler, C. A., Labuda, D., Bamshad, M., Jorde, L. B., Tishkoff, S. A., Watkins, W. S., Metspalu, M., Dryomov, S., Sukernik, R.,

Singh, L., Thangaraj, K., Pääbo, S., Kelso, J., Patterson, N., & Reich, D. (2016). The simons genome diversity project : 300 genomes from 142 diverse populations. *Nature*, *538*, 201.

Marco, A. (2018). Seedvicious : Analysis of microrna target and near-target sites. *PLoS One*, *13*(4), e0195532.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., & Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, *337*(6099), 1190-5.

McCoy, R. C., Wakefield, J., & Akey, J. M. (2017). Impacts of neanderthal-introgressed sequences on the landscape of human gene expression. *Cell*, *168*(5), 916-927 e12.

McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, *5*(5), e1000471-e1000471.

Mellars, P. (2004). Neanderthals and the modern human colonization of europe. *Nature*, *432*(7016), 461-5.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., & Pääbo, S. (2012). A high-coverage genome sequence from an archaic denisovan individual. *Science*, *338*(6104), 222.

Michalak, P., & Noor, M. A. (2003). Genome-wide patterns of expression in drosophila pure species and hybrid males. *Molecular biology and evolution*, *20*(7), 1070-1076.

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., Johnson, W. C., Im, H. K., Liu, Y., & Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genet*, *14*(8), e1007586.

Muller, H. (1942). Recessive genes causing interspecific sterility and other disharmonies between drosophila melanogaster and simulans. *Genetics*, *27*, 157.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls : annotation to enhance discovery from gwas. *PLoS Genet*, *6*(4), e1000888.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics*, *39*(1), 197-218.

Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., Dumaine, A., Danckaert, A., Luca, F., Grenier, J. C., Hansen, K. D., Gicquel, B., Yu, M., Pai, A., He, C., Tung, J., Pastinen, T., Kobor, M. S., Pique-Regi, R., Gilad, Y., & Barreiro, L. B. (2015). Bacterial infection remodels the dna methylation landscape of human dendritic cells. *Genome Res*, *25*(12), 1801-11.

Pan, D. Z., Garske, K. M., Alvarez, M., Bhagat, Y. V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C. K., Cantor, R. M., Civelek, M., Glastonbury, C. A., Small, K. S., Boehnke, M., Lusis, A. J., Sinsheimer, J. S., Mohlke, K. L., Laakso, M., Pajukanta, P., & Ko, A. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from gwas. *Nat Commun*, *9*(1), 1512.

Parts, L., Hedman, A. K., Keildson, S., Knights, A. J., Abreu-Goodger, C., van de Bunt, M., Guerra-Assuncao, J. A., Bartonicek, N., van Dongen, S., Magi, R., Nisbet, J., Barrett,

A., Rantalainen, M., Nica, A. C., Quail, M. A., Small, K. S., Glass, D., Enright, A. J., Winn, J., Mu, T. C., Deloukas, P., Dermitzakis, E. T., McCarthy, M. I., Spector, T. D., Durbin, R., & Lindgren, C. M. (2012). Extent, causes, and consequences of small rna expression variation in human adipose tissue. *PLoS Genet*, *8*(5), e1002704.

Petr, M., Pääbo, S., Kelso, J., & Vernot, B. (2019). Limits of long-term selection against neandertal introgression. *Proceedings of the National Academy of Sciences*, *116*(5), 1639.

Piasecka, B., Duffy, D., Urrutia, A., Quach, H., Patin, E., Posseme, C., Bergstedt, J., Charbit, B., Rouilly, V., MacPherson, C. R., Hasan, M., Albaud, B., Gentien, D., Fellay, J., Albert, M. L., Quintana-Murci, L., & Milieu Interieur, C. (2018). Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc Natl Acad Sci U S A*, *115*(3), E488-E497.

Pinzon, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., & Seitz, H. (2017). microrna target prediction programs predict many false positives. *Genome Res*, *27*(2), 234-245.

Plagnol, V., & Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genet*, *2*(7), e105.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*, 702.

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation : hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, *20*(4), R208-15.

Prufer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlevic, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyregne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., Dannemann, M., Nelson, B. J., Key, F. M., Rudan, P., Kucan, Z., Gusic, I., Golovanova, L. V., Doronichev, V. B., Patterson, N., Reich, D., Eichler, E. E., Slatkin, M., Schierup, M. H., Andres, A. M., Kelso, J., Meyer, M., & Paabo, S. (2017). A high-coverage neandertal genome from vindija cave in croatia. *Science*, *358*(6363), 655-658.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., & Paabo, S. (2014). The complete genome sequence of a neanderthal from the altai mountains. *Nature*, *505*(7481), 43-9.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y. E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., Deschamps, M., Naffakh, N., Duffy, D., Coen, A., Leroux-Roels, G., Clement, F., Boland, A., Deleuze, J. F., Kelso, J., Albert, M. L., & Quintana-Murci, L. (2016). Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell*, *167*(3), 643-656 e17.

Quintin, J., Cheng, S.-C., van der Meer, J. W. M., & Netea, M. G. (2014). Innate immune memory : towards a better understanding of host defense mechanisms. *Current Opinion in Immunology*, *29*, 1-7.

Racimo, F., Berg, J. J., & Pickrell, J. K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics*, *208*(4), 1565-1584.

Racimo, F., Marnetto, D., & Huerta-Sanchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. *Mol Biol Evol*, *34*(2), 296-317.

Rantalainen, M., Herrera, B. M., Nicholson, G., Bowden, R., Wills, Q. F., Min, J. L., Neville, M. J., Barrett, A., Allen, M., Rayner, N. W., Fleckner, J., McCarthy, M. I., Zondervan, K. T., Karpe, F., Holmes, C. C., & Lindgren, C. M. (2011). Microrna expression in abdominal and gluteal adipose tissue is associated with mrna expression levels and partly genetically driven. *PLoS One*, *6*(11), e27338.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J. J., Kelso, J., Slatkin, M., & Paabo, S. (2010). Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, *468*(7327), 1053-60.

Richards, M. P., Taylor, G., Steele, T., McPherron, S. P., Soressi, M., Jaubert, J., Orschiedt, J., Mallye, J. B., Rendu, W., & Hublin, J. J. (2008). Isotopic dietary analysis of a neanderthal and associated fauna from the site of jonzac (charente-maritime), france. *Journal of Human Evolution*, *55*(1), 179-185.

Rinker, D. C., Simonti, C., McArthur, E., Shaw, D., Hodges, E., & Capra, J. A. (2019). Neanderthal introgression reintroduced functional alleles lost in the human out of africa bottleneck. *bioRxiv*, 533257.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., & Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317-30.

Rogers, A. R., Bohlender, R. J., & Huff, C. D. (2017). Early history of neanderthals and denisovans. *Proc Natl Acad Sci U S A*, *114*(37), 9859-9863.

Saini, A. (2019). *Superior : The return of race science.*

Sams, A. J., Dumaine, A., Nedelec, Y., Yotova, V., Alfieri, C., Tanner, J. E., Messer, P. W., & Barreiro, L. B. (2016). Adaptively introgressed neandertal haplotype at the oas locus functionally impacts innate immune responses in humans. *Genome Biol*, *17*(1), 246.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354-7.

Sankararaman, S., Mallick, S., Patterson, N., & Reich, D. (2016). The combined landscape of denisovan and neanderthal ancestry in present-day humans. *Curr Biol*, *26*(9), 1241-7.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G., & Pääbo, S. (2004). No evidence of neandertal mtdna contribution to early modern humans. *PLoS biology*, *2*(3), E57-E57.

Siddle, K. J., Deschamps, M., Tailleux, L., Nedelec, Y., Pothlichet, J., Lugo-Villarino, G., Libri, V., Gicquel, B., Neyrolles, O., Laval, G., Patin, E., Barreiro, L. B., & Quintana-Murci, L. (2014). A genomic portrait of the genetic architecture and regulatory impact of microrna expression in response to infection. *Genome Res*, *24*(5), 850-9.

Siddle, K. J., Tailleux, L., Deschamps, M., Loh, Y. H., Deluen, C., Gicquel, B., Antoniewski, C., Barreiro, L. B., Farinelli, L., & Quintana-Murci, L. (2015). bacterial infection drives the expression dynamics of micrornas and their isomirs. *PLoS Genet*, *11*(3), e1005064.

Simonti, C. N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D. S., Chisholm, R. L., Crosslin, D. R., Hebbring, S. J., Jarvik, G. P., Kullo, I. J., Li, R., Pathak, J., Ritchie, M. D., Roden, D. M., Verma, S. S., Tromp, G., Prato, J. D., Bush, W. S., Akey, J. M., Denny, J. C., & Capra, J. A. (2016). The phenotypic legacy of admixture between modern humans and neandertals. *Science*, *351*(6274), 737-41.

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M. H., & Durbin, R. (2018). Detecting archaic introgression using an unadmixed outgroup. *PLOS Genetics*, *14*(9), e1007641.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyrégne, S., Nagel, S., Brown, S., Douka, K., Higham, T., Kozlikin, M. B., Shunkov, M. V., Derevianko, A. P., Kelso, J., Meyer, M., Prüfer, K., & Pääbo, S. (2018). The genome of the offspring of a neanderthal mother and a denisovan father. *Nature*, *561*(7721), 113-116.

Sommer, J. D. (2009). The shanidar iv 'flower burial' : a re-evaluation of neanderthal burial ritual. *Cambridge Archaeological Journal*, *9*(1), 127-129.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P., & Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, *8*(4), e1002639.

Stringer, C. B., & Hublin, J. (1999). New age estimates for the swanscombe hominid, and their significance for human evolution. *J Hum Evol*, *37*(6), 873-7.

Ségurel, L., & Bon, C. (2017). On the evolution of lactase persistence in humans. *Annual Review of Genomics and Human Genetics*, *18*(1), 297-319.

Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., Robinson, S., Zhang, S. C., Ellis, P., Langford, C. F., Guillot, P. V., Chandrashekran, A., Fisk, N. M., Castellano, L., Meister, G., Winston, R. M., Cui, W., Baulcombe, D., & Dibb, N. J. (2014). 5 ' isomir variation is of functional and evolutionary importance. *Nucleic Acids Research*, *42*(14), 9424-9435.

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R.,

Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., et al. (2015). A global reference for human genetic variation. *Nature*, *526*, 68.

The International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Marie Muzny, D., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarroll, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Cristina Manca, M., Marshall, P. A., Matsuda, I., Ngare, D., Ota Wang, V., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., & McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*, 52.

The International HapMap, C., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Qiang Song, Y., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., et al. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, *449*, 851.

Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L. G., Fauchoux, N., Sayegrih, S., Vray, M., Duffy, D., Quintana-Murci, L., Albert, M. L., & Milieu Interieur, C. (2015). The milieu interieur study - an integrative approach for study of human immunological variance. *Clin Immunol*, *157*(2), 277-93.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in africa and europe. *Nature Genetics*, *39*(1), 31-40.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz,

S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304.

Vernot, B., & Akey, J. M. (2014). Resurrecting surviving neandertal lineages from modern human genomes. *Science*, *343*(6174), 1017-21.

Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., Dannemann, M., Grote, S., McCoy, R. C., Norton, H., Scheinfeldt, L. B., Merriwether, D. A., Koki, G., Friedlaender, J. S., Wakefield, J., Paabo, S., & Akey, J. M. (2016). Excavating neandertal and denisovan dna from the genomes of melanesian individuals. *Science*, *352*(6282), 235-9.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flicek, P., & Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554-66.

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

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-Guzman, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Arindrarto, W., Beutner, F., Brugge, H., Dmitreva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Pritchard, J., Raitakari, O., Rotzchke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., Hoen, P. A. C. t., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J., Völker, U., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B., Lehtimäki, T., Boomsma, D., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W., Downes, K., Stegle, O., Battle, A., Yang, J., Visscher, P. M., Scholz, M., Gibson, G., Esko, T., & Franke, L. (2018). Unraveling the polygenic architecture of complex traits using blood eqtl metaanalysis. *bioRxiv*, 447367.

Wall, J. D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, *154*(3), 1271-9.

Wall, J. D., Lohmueller, K. E., & Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*, *26*(8), 1823-7.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *16*(2), 97-159.

Annexe

Report

Current Biology

Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest

Highlights

- A strong selective sweep at *TRPS1* occurred in African rainforest hunter-gatherers
- Pleiotropic height genes lead to polygenic selection signals for reproductive age
- Pathogen-driven selection, mostly viral, has been pervasive among hunter-gatherers
- Post-admixture selection has maintained adaptive variation in hunter-gatherers

Authors

Marie Lopez, Jeremy Choin, Martin Sikora, ..., Paul Verdu, Etienne Patin, Lluís Quintana-Murci

Correspondence

etienne.patin@pasteur.fr (E.P.), quintana@pasteur.fr (L.Q.-M.)

In Brief

Lopez et al. search for genomic evidence of local adaptation of hunter-gatherers to the African rainforest. They find signals of classic sweeps, polygenic adaptation, and post-admixture selection at height, development, and immune response genes. They show that pleiotropy of height genes leads to polygenic selection signals for life-history traits.

Lopez et al., 2019, Current Biology *29*, 1–10
September 9, 2019
https://doi.org/10.1016/j.cub.2019.07.013

CellPress

Please cite this article in press as: Lopez et al., Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest, Current Biology (2019), https://doi.org/10.1016/j.cub.2019.07.013

Current Biology
Report

Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest

Marie Lopez,[1,2] Jeremy Choin,[1] Martin Sikora,[3] Katherine Siddle,[1,11] Christine Harmant,[1] Helio A. Costa,[4] Martin Silvert,[1,2] Patrick Mouguiama-Daouda,[5] Jean-Marie Hombert,[6] Alain Froment,[7] Sylvie Le Bomin,[8] George H. Perry,[9] Luis B. Barreiro,[10] Carlos D. Bustamante,[4] Paul Verdu,[8] Etienne Patin,[1,12,*] and Lluís Quintana-Murci[1,12,13,*]
[1]Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris 75015, France
[2]Sorbonne Universités, Ecole Doctorale Complexité du Vivant, 75005 Paris, France
[3]Centre for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark
[4]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA
[5]Laboratoire Langue, Culture et Cognition (LCC), Université Omar Bongo, 13131 Libreville, Gabon
[6]CNRS UMR 5596, Université Lumière-Lyon 2, 69007 Lyon, France
[7]Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, 75005 Paris, France
[8]UMR7206, Muséum National d'Histoire Naturelle, CNRS, Université Paris Diderot, Paris 75016, France
[9]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA
[10]Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
[11]Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[12]Senior author
[13]Lead Contact
*Correspondence: etienne.patin@pasteur.fr (E.P.), quintana@pasteur.fr (L.Q.-M.)
https://doi.org/10.1016/j.cub.2019.07.013

SUMMARY

African rainforests support exceptionally high biodiversity and host the world's largest number of active hunter-gatherers [1–3]. The genetic history of African rainforest hunter-gatherers and neighboring farmers is characterized by an ancient divergence more than 100,000 years ago, together with recent population collapses and expansions, respectively [4–12]. While the demographic past of rainforest hunter-gatherers has been deeply characterized, important aspects of their history of genetic adaptation remain unclear. Here, we investigated how these groups have adapted—through classic selective sweeps, polygenic adaptation, and selection since admixture—to the challenging rainforest environments. To do so, we analyzed a combined dataset of 566 high-coverage exomes, including 266 newly generated exomes, from 14 populations of rainforest hunter-gatherers and farmers, together with 40 newly generated, low-coverage genomes. We find evidence for a strong, shared selective sweep among all hunter-gatherer groups in the regulatory region of *TRPS1*—primarily involved in morphological traits. We detect strong signals of polygenic adaptation for height and life history traits such as reproductive age; however, the latter appear to result from pervasive pleiotropy of height-associated genes. Furthermore, polygenic adaptation signals for functions related to responses of mast cells to allergens and microbes, the IL-2 signaling pathway, and host interactions with viruses support a history of pathogen-driven selection in the rainforest. Finally, we find that genes involved in heart and bone development and immune responses are enriched in both selection signals and local hunter-gatherer ancestry in admixed populations, suggesting that selection has maintained adaptive variation in the face of recent gene flow from farmers.

RESULTS

Exome Sequencing Dataset and Population Structure

African rainforest hunter-gatherers (RHGs)—historically grouped under the term "Pygmies"—live along the dense tropical rainforests of central Africa, in the western and eastern part of the Congo Basin [1–3]. Genetic studies have deeply investigated the demographic history of these groups, characterized by long-term isolation since the Upper Paleolithic and substantial admixture with neighboring Bantu-speaking farmers in the last 1,000 years [4–12]. However, their adaptive history has received less attention. Natural selection studies in RHGs have primarily focused on small adult body size as the only trait characterizing the "pygmy" phenotype [13–20], and used SNP genotyping data [14, 15, 19–21] or whole-genome/exome sequencing of a few individuals or populations [4, 6, 18, 22, 23].

To understand human genetic adaptation to the rainforest, we generated and analyzed whole-exome sequencing data (~40× coverage) for seven RHG groups from Cameroon, Gabon, and Uganda, as well as, for comparison purposes, seven sedentary groups of Bantu-speaking agriculturalists (AGRs) (Figure 1A; Table S1). After quality filters, we obtained a final dataset of 566 individuals (298 RHGs and 268 AGRs), consisting of 266 newly generated exomes that were analyzed with 300 previously reported exomes [4] (Figure S1).

Genetic differentiation among RHG groups was higher than that between RHGs and AGRs (among-RHG, F_{ST} = 0.025; among-western RHG, F_{ST} = 0.021; RHG-AGR, F_{ST} = 0.017;

CURBIO 15656

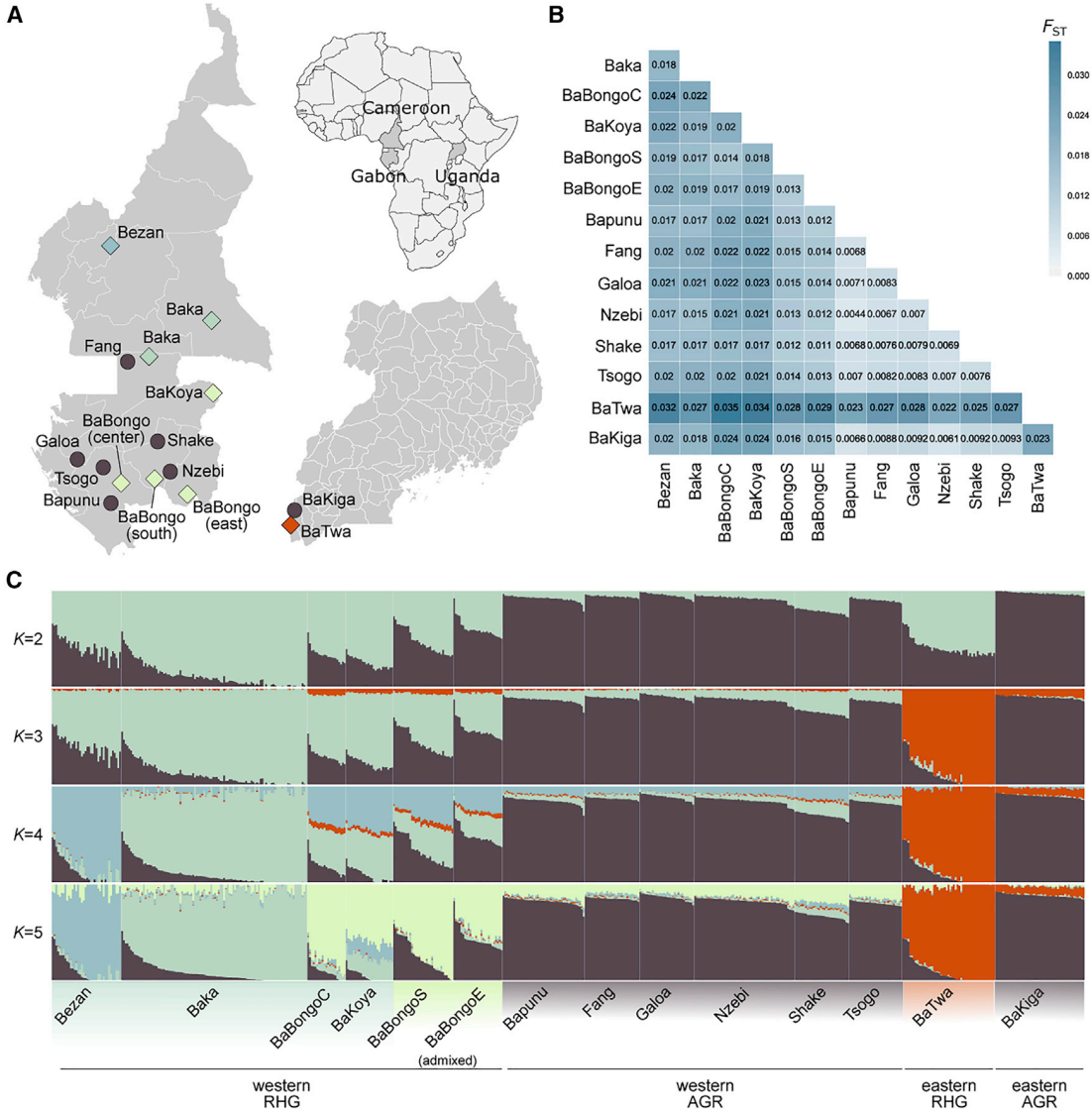


Figure 1. Location, Genetic Differentiation, and Structure of Central African Populations

(A) Geographic location of the populations analyzed. Populations of rainforest hunter-gatherers (diamonds) and neighboring farmers (circles) originating from the three countries are shown in the map of Africa. Colors indicate the dominant membership in each population, based on ADMIXTURE results (C).

(B) Levels of genetic differentiation between populations measured by pairwise F_{ST} calculated on the exome data.

(C) Cluster membership proportions estimated by ADMIXTURE on the merged exome and SNP array data. Cross-validation values were lowest at $K = 5$ clusters.

(B and C) BaBongoC, BaBongoS, and BaBongoE stand for BaBongo populations from the center, south, and east of Gabon, respectively.

See also Figure S1 and Table S1.

among-AGR, $F_{ST} = 0.007$; Figure 1B). To increase SNP density, particularly in the non-coding genome, we combined the exome data with SNP array data for the same individuals [12, 24, 25], yielding a total of 1,253,548 SNPs. When using ADMIXTURE [26] on the dataset pruned for allele frequency (MAF > 5%) and linkage disequilibrium ($r^2 < 0.5$), RHGs separated into four clusters at $K = 5$ (Figure 1C), corresponding to Bezan, Baka, BaBongo and BaKoya, and BaTwa groups. As previously observed [5, 12, 14, 24], membership proportions to the cluster assigned to AGRs were non-negligible and similar among RHG groups (~4%–9%; Table S1), with the exception of the BaBongo of east and south Gabon, who presented high AGR proportions (~43% [SD = 11%] and ~24% [SD = 17%], respectively). Membership proportions to the cluster assigned to RHGs were also non-negligible among AGRs (~10%–30%). Our results show that RHG populations are highly structured, emphasizing the importance of considering these groups separately in subsequent analyses.

Searching for Signals of Local Genetic Adaptation in Central Africans

For all natural selection analyses, we increased SNP density to 9,129,103 high-quality variants (MAF > 1%), through genotype imputation using (1) newly generated whole genomes from

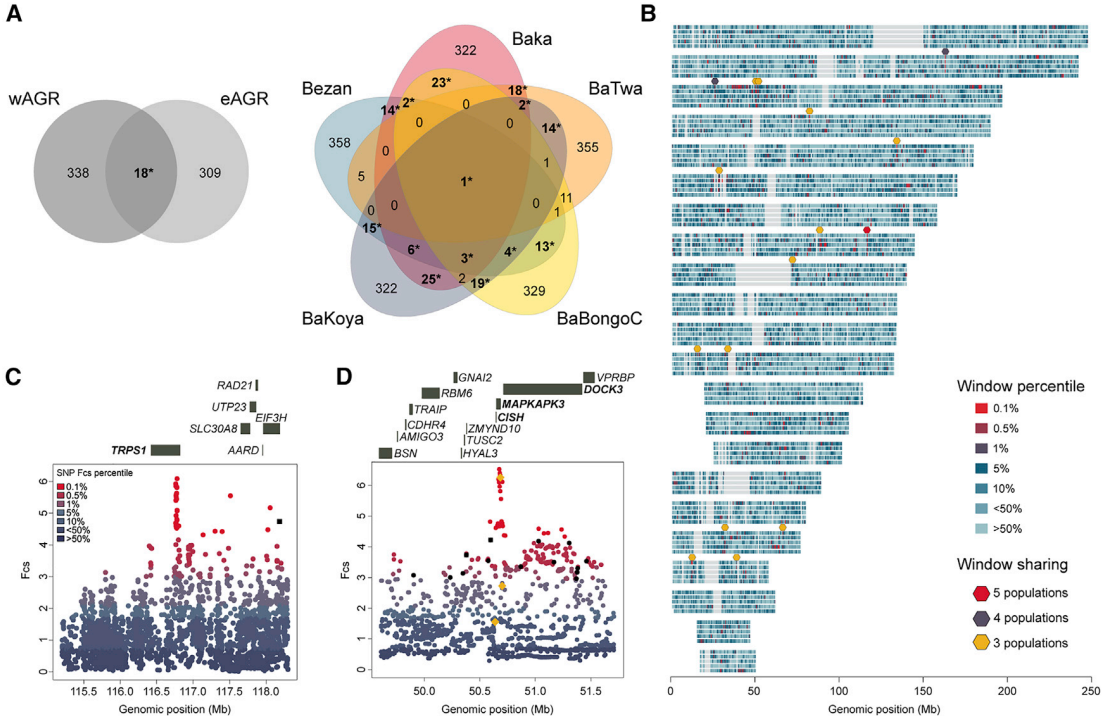


Figure 2. Shared Signals of Classic Sweeps among Rainforest Hunter-Gatherers

(A) Number of candidate windows for classic sweeps (i.e., windows with proportions of outlier SNPs among the 1% highest of the genome) common to western and eastern AGR populations (wAGR and eAGR), as well as common to RHG populations. p values obtained based on 10,000 resamples are shown: *$p < 10^{-4}$.

(B) Genome-wide map of classic sweep signals in RHG groups. The autosomes of each of the five RHG populations (from top to bottom: Bezan, Baka, lowly admixed BaBongo, BaKoya, and BaTwa) are shown. Colored dots indicate genomic regions that are common to at least three RHG populations.

(C) Selective sweep signal at the locus containing the *TRPS1* gene (chr8:116702422-116802422) in the Baka RHGs.

(D) Selective sweep signal at the locus containing *CISH*, *MAPKAPK3*, and *DOCK3* genes (chr3:50610197-50710197 and chr3:50660197-50760197) in the BaTwa RHGs.

(C and D) Dot colors indicate SNP F_{CS} percentiles, black squares indicate non-synonymous mutations, and black dots indicate eQTLs (q value < 0.005) [33]. eQTLs of *MAPKAPK3* (rs107457 and rs9879397) and *DOCK3* (rs12629788) are shown as yellow diamonds. Not all genes of the genomic region are shown for convenience.

See also Figures S2 and S3 and Data S1.

20 RHG Baka and 20 AGR Nzébi from Gabon (5–6× coverage) and (2) the 1000 Genomes Phase 3 panel [27] (STAR Methods; Figure S1). We focused on the five RHG populations presenting the lowest average levels of AGR ancestry and analyzed the highly admixed RHG groups differently (see Recent Genetic Adaptation of Admixed Rainforest Hunter-Gatherers). To identify signals of strong sweeps, we searched for variants with both high allele frequency and extended haplotype homozygosity in RHGs, relative to AGRs (STAR Methods). Genome-wide ranks of PBS [28] and XP-EHH [29] were combined into a Fisher's score (F_{CS}), and to reduce false positives, candidate regions were defined as 100-kb windows with the 1% highest proportion of outlier SNPs of the genome.

We first scanned the genomes of AGR populations (Figure S2), the evolutionary history of whom is well characterized [24, 29–32]. We found 18 candidate regions for positive selection in both western and eastern AGRs, while only ~3.5 were expected to be shared if candidate loci were false positives (10,000 random samples; resampling $p < 10^{-4}$) (Figure 2A; Data S1). Among candidates, we replicated, for example, the signal encompassing the *LARGE* gene, involved in Lassa virus infectivity [34]. These results provide evidence that the genomic regions detected by our approach are enriched in true signals.

A Strong, Shared Selective Sweep at *TRPS1* across All Hunter-Gatherer Groups

Our search for sweeps in RHGs identified candidates that were shared by RHG groups more than expected by chance (resampling $p < 10^{-4}$) (Figure 2A; Data S1). Remarkably, we identified a single genomic region that exhibits sweep signals in all RHG populations, but not in AGRs (Figures 2A–2C and S3). This region lies upstream of the 5′ UTR of *TRPS1*, which encodes a transcription factor (TF) with multiple pleiotropic effects, including skeletal development and inflammatory T_H17 cell differentiation [35–37]. The six variants presenting the highest frequency differences between RHGs and AGRs (Data S1) define a 5,777 bp region that contains a primate-specific THE1B endogenous retrovirus sequence, known to control the expression of nearby genes [38]. Given the high expression of *TRPS1* in monocytes [39], we analyzed published RNA sequencing (RNA-seq) data from monocytes of individuals of central African ancestry to test if candidate variants affect *TRPS1* expression

[40]. A highly differentiated variant that falls within the THE1B fragment was associated with increased expression of a short, non-canonical *TRPS1* transcript upon immune stimulation (rs111351287; regression $p = 5 \times 10^{-6}$). These findings suggest that the most robust signal of adaptation to the African rainforest can be ascribed to *TRPS1*, possibly in relation with variation in morphological and/or immunological traits.

Detection of Other Classic Sweep Signals in Rainforest Hunter-Gatherers

Other selective sweep signals were specific to a smaller number of RHG groups (Figure S2; Data S1). These include the known 150-kb region encompassing *CISH*, *MAPKAPK3*, and *DOCK3* [6, 14], which we show here to be shared among western and eastern RHGs (Baka, BaKoya, and BaTwa). We searched the GTEx database [33] for regulatory variation at these genes (eQTLs) and found two *cis*-eQTLs for *MAPKAPK3* (rs107457 and rs9879397), one for *DOCK3* (rs12629788), and none for *CISH* (Data S1). Selection scores at these eQTLs were among the highest of the region, particularly for *MAPKAPK3* (Figure 2D), which affects hepatitis C virus (HCV) infectivity [41].

We also detected two contiguous regions at the *IFIH1* locus [18], which present strong enrichments in selection scores that are shared by all western RHG groups. Candidate variation at this locus (rs12479043) controls the expression of the nearby *FAP* gene [33], which regulates fibroblast and myofibroblast growth and wound healing during chronic inflammation [42]. We also identified two windows—shared by Bezan, Baka, and BaKoya—encompassing *RASGEF1B*, whose expression is induced in macrophages by lipopolysaccharide, a membrane component of Gram-negative bacteria [43]. Finally, we found a window in the Bezan, BaBongo, and BaKoya that overlaps *PITX1*, recently identified as a selection candidate in RHGs [22]. PITX1 modulates the core development of limb [44], is associated with height variation [45], acts as an early TF in the developing pituitary gland [46], and regulates interferon-α virus induction [47]. These results support the hypothesis that development and immunity are key traits in local adaptation to the rainforest.

Evidence for Polygenic Selection Favoring the "Pygmy" Phenotype

Given the polygenic nature of most adaptive traits [48, 49], we searched for evidence of polygenic adaptation focusing on 12 candidate quantitative traits. These include height, body mass index, skin pigmentation, life history traits, and immune cell counts, the genetic architectures of which have been extensively studied [50]. We compared the distribution of mean F_{CS} scores in non-overlapping, 100-kb genomic windows containing trait-associated SNPs to that of randomly sampled windows, accounting for SNP density, LD levels, and background selection (STAR Methods). Stature-related traits showed the most significant polygenic selection signals, in all RHG groups (adjusted $p < 0.05$) while being non-significant in AGRs (Figure 3A). Life-history traits related to reproduction also exhibited selection signals in various RHG groups, consistent with the proposed adaptive nature of early reproduction in RHGs [51, 52]. Furthermore, we replicated selection signals for cardiovascular traits in the BaTwa (adjusted $p < 0.001$) [23]. Notably, we found significant signals in "Leukocyte count" in the Baka and the BaBongo (adjusted $p < 0.05$), suggesting polygenic adaptation related to immunity.

We next examined whether signals of polygenic selection could result from pleiotropy; e.g., advantageous height-associated variants affect other correlated traits [49]. Using the UK Biobank dataset [50], we computed the genetic correlations from LD-score regressions between "Standing height" and the remaining traits, and found significant correlations for eight of them (STAR Methods; Data S1). For these, we repeated the analysis after excluding windows associated with "Standing height" or "Comparative height at age 10," and the significance of selection signals was lost or dramatically reduced (Figures 3B and S4). Conversely, when excluding windows associated with non-height traits (e.g., reproduction-related traits), we found that "Standing height" was still significant in four RHG populations (adjusted $p < 0.05$) (Figure 3C). These results show that height has been an adaptive trait in RHGs, resulting in spurious polygenic selection signals for other correlated traits because of pleiotropy.

Evidence of Pervasive Pathogen-Driven Selection in the Equatorial Rainforest

We further investigated genomic signatures of polygenic adaptation, by searching for excesses in mean F_{CS} among windows related to 5,354 gene ontology (GO) terms [53] (STAR Methods). We detected 38 terms that were significant in at least three RHG groups, but not in AGRs (Figure 3D; Data S1). Among these, we found positive regulation of "mast cell degranulation" and "the phosphatidylinositol 3-kinase (PI3K) pathway" (false discovery rate [FDR] $p < 5\%$). Recognition by mast cells of allergens and antigens induces degranulation, a process mediated by the PI3K pathway that results in inflammation and allergy [54]. Enrichments were also found in the IL-2 signaling pathway, which activates the PI3K pathway and regulates immune tolerance [55]. All enrichments remained significant after removing windows associated with height (FDR $p < 5\%$), excluding potential pleiotropic effects. To gain further insights into pathogen-driven selection, we next focused on 1,553 innate immunity genes (IIGs) [56] and 1,257 genes encoding virus-interacting proteins (VIPs) [57]. We found significant enrichments in selection signals for both gene sets in RHGs, but not in AGRs, in particular for VIPs interacting with double-stranded DNA (dsDNA) and single-stranded RNA (ssRNA) viruses (FDR $p < 5\%$; Table S2; Data S1). These results collectively support the notion that pathogens have been a major driver of local adaptation in the African rainforest.

Recent Genetic Adaptation of Admixed Rainforest Hunter-Gatherers

To search for evidence of recent selection in RHG since their admixture with AGRs, we focused on the highly admixed BaBongo (Figure 1C) and performed local ancestry inference with RFMix [58], using as putative parental populations western RHG and AGR individuals with the lowest AGR and RHG membership proportions, respectively (STAR Methods). Six contiguous windows on chromosome 1 showed both evidence of selection (i.e., top 1% of the proportion of outlier SNPs) and an excess of RHG local ancestry (i.e., higher than the genome-wide average + 2 SD) in admixed RHG (Figures 4A and S2; Data S1). Among the

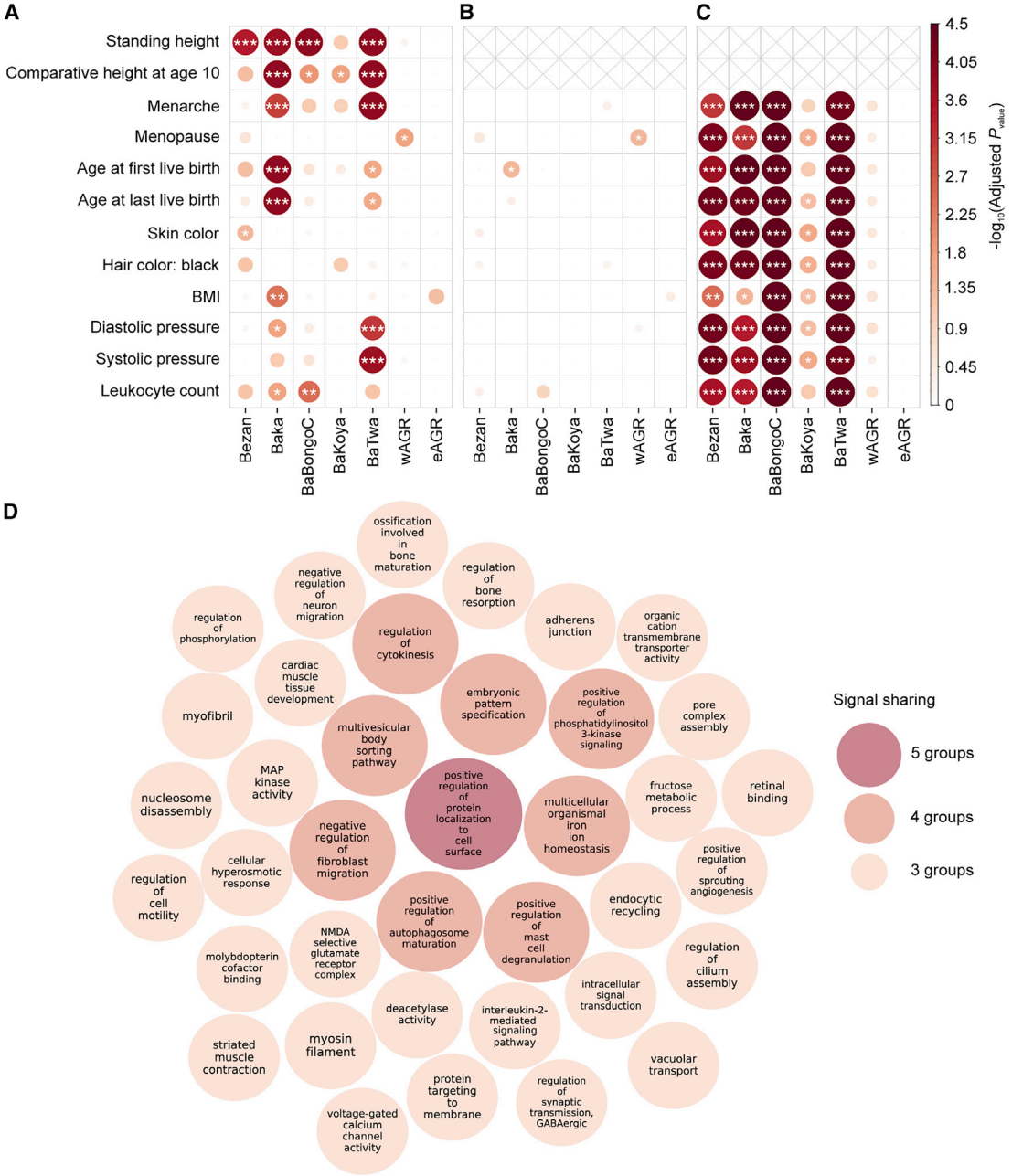


Figure 3. Signals of Polygenic Selection in African Rainforest Hunter-Gatherers

(A) Signals of polygenic selection for 12 candidate quantitative traits, based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations.

(B) Signals of polygenic selection for the candidate quantitative traits, based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations, after removing windows associated with "Standing height" and "Comparative height at age 10." Loss of significance was not explained by the reduced number of windows tested (Figure S4).

(C) Signals of polygenic selection for "Standing height," based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations, after removing windows associated with each of the remaining quantitative traits.

(A–C) Color gradient and circle sizes are proportional to $-\log_{10}$(adjusted p) with adjusted *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$. Multiple testing corrections were performed using the Benjamini-Hochberg method. wAGR and eAGR stand for western and eastern AGR groups. Signals were generally stronger in Baka and BaTwa RHGs, probably because of their larger sample size.

(D) Gene Ontology (GO) terms enriched in selection scores (FDR $p < 5\%$) in RHG, but not in AGR, populations, considering the window mean F_{CS} as selection score. Circle color and size indicate the number of RHG populations that show significant evidence of polygenic selection for a given GO term.

See also Figure S4, Table S2, and Data S1.

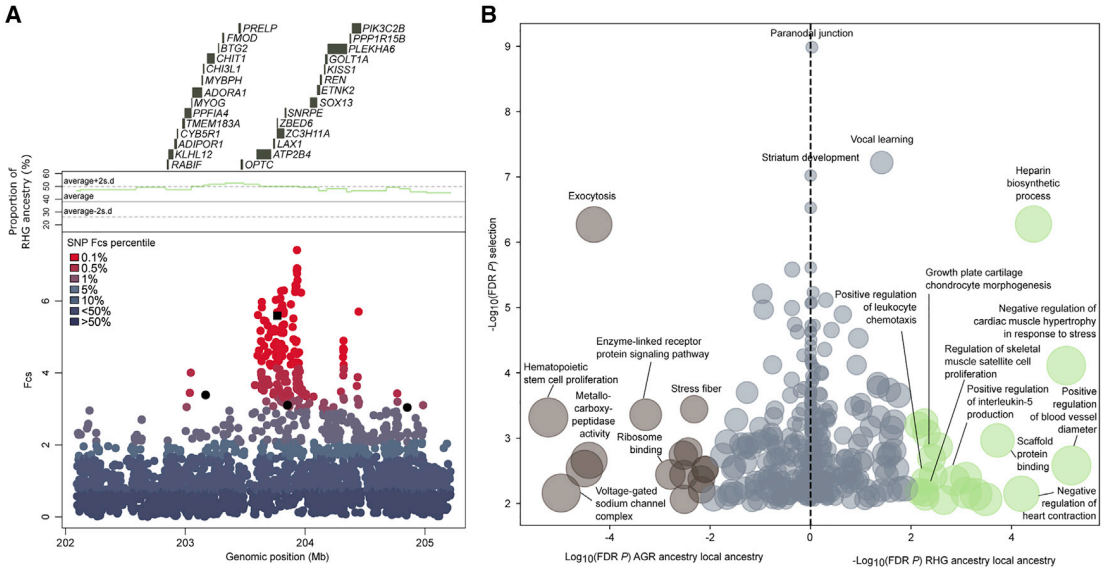


Figure 4. Selection Signals in Highly Admixed Rainforest Hunter-Gatherers

(A) Selective sweep signal and average local RHG ancestry at the chr1:203564464-203764464 locus in the highly admixed RHG BaBongo. Dot colors indicate SNP F_{CS} percentiles, the black square indicates the non-synonymous variant (rs6697388) at *ZBED6*, and black dots indicate eQTLs (q value < 0.005) [33].

(B) GO terms enriched in both local ancestry in the highly admixed RHG BaBongo, and selection scores in each of the two putative parental populations, with respect to the rest of the genome (FDR p < 5%). Green (brown) dots indicate GO terms enriched in both western RHG (western AGR) local ancestry and selection scores in parental western RHG (western AGR) populations (FDR p < 5%). Enrichments were assessed using the Mann-Whitney-Wilcoxon rank-sum test. See also Data S1.

strongest candidate variants, we found a non-synonymous mutation (rs6697388) in *ZBED6*, which encodes a TF that controls muscle growth through *IGF2* repression [59]. *ZBED6* is located within the intron of the *ZC3H11A* gene, whose product is required for the efficient growth of several nuclear-replicating viruses [60]. The rs6697388 G allele (p.Leu391Arg) is present at the highest frequency in admixed BaBongo (51%), with lower frequencies in parental RHG (42%) and AGR (15%) groups. With respect to the strong, shared selective sweep detected at *TRPS1* (Figure 2C), the locus also presented selection signals in the BaBongo but no excess of RHG or AGR ancestry (Figures S2 and S3), suggesting weaker or no positive selection at *TRPS1* since admixture.

Finally, we searched for evidence of polygenic selection since admixture, by testing for excesses in AGR or RHG local ancestry in genomic windows related to GO terms in the admixed BaBongo (STAR Methods). We found 21 GO terms that were enriched in both RHG local ancestry and selection signals in the parental RHGs (Figure 4B; Data S1), an overlap that was significantly larger than expected (7.3% versus 4.7%, χ^2 test, p = 0.042). These terms were mostly related to cardiac and skeletal development and immune functions, and included "heparin biosynthetic process," which participates in mast cell-mediated immune and inflammatory responses [61], echoing the signals detected for "mast cell degranulation" in weakly admixed RHGs (Figure 3D). We also found 16 GO terms that were enriched in both AGR local ancestry and selection signals in the parental AGRs (Figure 4B; Data S1), including stem cell proliferation, exocytosis, and muscle composition. Together, these results support further the notion that heart and bone development as well as immune responses have been an important substrate of selection in RHGs, before and after their admixture with neighboring farmers.

DISCUSSION

Here we present the first exome-based survey of multiple geographically dispersed groups of African rainforest hunter-gatherers, with the aim of investigating how populations have adapted to the challenging habitats of the equatorial rainforest. Because positive selection often targets regulatory regions [62], we combined the exome dataset with SNP array data, to cover both genic and intergenic regions. In doing so, we found evidence of a unique, strong sweep that is shared by all RHG groups, targeting the regulatory region of *TRPS1*, mutations in which can cause growth retardation, distinctive craniofacial features [63], and hypertrichosis [64]. Furthermore, the transcription factor TRPS1 regulates STAT3, a mediator of inflammation and immunity [65], and RUNX2, controlling facial features and viral clearance [66, 67]. Interestingly, *TRPS1* has been recently shown to carry signals of archaic introgression in western Africans [68]. Functional studies should help determine the adaptive nature—developmental and/or immune-related—of variation at this locus, which possibly introgressed from extinct African hominins [18, 68, 69].

This study also extends previous findings of a sweep targeting the *CISH-MAPKAPK3-DOCK3* region [6, 14], by delineating *MAPKAPK3* as the most likely target. *MAPKAPK3* expression is regulated by two eQTLs that are among the strongest candidates for positive selection at the locus in RHG populations. MAPKAPK3 directly interacts with HCV and regulates cell infectivity [41]. A lower prevalence of HCV infection has been reported in RHG, with respect to AGR [70, 71]. Our results strengthen the evolutionary importance of the *CISH-MAPKAPK3-DOCK3* region in both western and eastern RHGs, and pinpoint

MAPKAPK3 variation as a putative, additional risk factor for HCV infection in Africans.

Our analyses provide robust evidence for polygenic selection of height, which we replicate in various RHG groups. Importantly, our results are not affected by biased genome-wide association study (GWAS) summary statistics due to partial control for population stratification, which can result in spurious polygenic selection signals [72, 73]. Our approach tests for the co-localization of selection signals and trait-associated genes; thus, it does not depend on effect size estimates and does not assume that associated variants are the same across populations. More generally, polygenic selection of height is unlikely to result from sexual selection [74] but from genetic adaptation to equatorial forest environments [75]. Our study sheds new light onto the debated adaptive nature of height, and supports that the early reproductive age of RHGs is not the cause of their small body size, as previously suggested [51, 52]. Instead, our results suggest that directional selection of height has resulted in changes in life-history traits because of pervasive pleiotropy of height-associated genes.

We also found signals of polygenic selection in RHGs at functions related to the IL-2 pathway, the sensing of allergens and microbes, and interactions with dsDNA and ssRNA viruses. Interestingly, higher seropositivity for more than 30 viruses has been reported in the BaTwa from Uganda, with respect to AGRs, particularly for dsDNA viruses [76]. That we also found an excess of RHG ancestry related to heparin biosynthesis, interleukin production, and leukocyte chemotaxis in highly admixed RHGs suggests preferential retention of RHG variation at immune-related functions. This finding supports a long-standing history of adaptation of RHGs to high pathogen pressures. This contrasts with a study in southern Africa, which reported a low exposure and adaptation to pathogens of hunter-gatherers of the Kalahari Desert, except for those who recently came in contact with other populations [77].

Collectively, our analyses uncover height, development, and immune response as iconic adaptive traits of African RHGs. It is interesting to note that the PI3K signaling pathway—under polygenic selection in four RHG populations—modulates inflammatory responses [78], body energy homeostasis [79, 80], and insulin secretion [81]. Several studies have highlighted the reciprocal relationship between proinflammatory cytokines and the regulation of the growth hormone through the IGF-1 axis [82]. It is thus tempting to speculate that pleiotropic effects between developmental growth and immunity could have further participated in the "pygmy" phenotype. Epidemiological work on the infectious disease burden in hunter-gatherers should increase our understanding of how historical high pathogen-driven selection has contributed to the reduced stature characterizing populations of the rainforest.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection
- METHOD DETAILS
  - Exome Sequencing
  - SNP Array Data
  - Merging Exome and SNP Array Data
  - Whole-Genome Sequencing
  - Imputation of SNP Array and Exome Data
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome Scans for Selective Sweeps
  - Polygenic Selection of Complex Traits
  - Polygenic Selection of Gene Ontologies
  - Local Ancestry Inference
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.07.013.

ACKNOWLEDGMENTS

We thank all the participants for providing the DNA samples used in this study. We also thank Guillaume Laval and Maxime Rotival for useful discussions, and Muh-Ching Yee for laboratory assistance. This work was supported by the Institut Pasteur, the Centre National de la Recherche Scientifique, the "Histoire du Génome des Populations Humaines Gabonaises" project (Institut Pasteur/Republic of Gabon), and an Agence Nationale de la Recherche grant "AGRHUM" (ANR-14-CE02-0003-01) to L.Q.-M. M.L. was supported by the Fondation pour la Recherche M´edicale (FDT20170436932), and J.C. by the INCEPTION program and the "Ecole Doctorale FIRE - Programme Bettencourt."

AUTHOR CONTRIBUTIONS

E.P. and L.Q.-M. conceived and supervised the study. M.L. conducted all the analyses and analyzed the data, with contributions from J.C., M. Silvert, and E.P. C.H. performed laboratory work. M. Sikora, K.S., H.C., and C.D.B. generated and/or analyzed whole-genome data. P.M.-D., J.-M.H., A.F., S.L.B., G.H.P., L.B.B., and P.V. assembled the samples. M.L., E.P., and L.Q.-M. wrote the manuscript, with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 14, 2019
Revised: June 26, 2019
Accepted: July 4, 2019
Published: August 8, 2019

REFERENCES

1. Hewlett, B.S. (2014). Hunter-Gatherers of the Congo Basin: Cultures, Histories and Biology of African Pygmies (Transaction Publishers).
2. Perry, G.H., and Verdu, P. (2017). Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. Quat. Int. *448*, 150–157.
3. Bahuchet, S. (2012). Changing language, remaining pygmy. Hum. Biol. *84*, 11–43.
4. Lopez, M., Kousathanas, A., Quach, H., Harmant, C., Mouguiama-Daouda, P., Hombert, J.M., Froment, A., Perry, G.H., Barreiro, L.B., Verdu, P., et al. (2018). The demographic history and mutational load of African hunter-gatherers and farmers. Nat Ecol Evol *2*, 721–730.
5. Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.M., et al. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr. Biol. *19*, 312–318.

6. Hsieh, P., Veeramah, K.R., Lachance, J., Tishkoff, S.A., Wall, J.D., Hammer, M.F., and Gutenkunst, R.N. (2016). Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. Genome Res. *26*, 279–290.
7. Patin, E., Laval, G., Barreiro, L.B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K.K., Kidd, J.R., Van der Veen, L., Hombert, J.M., et al. (2009). Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet. *5*, e1000448.
8. Batini, C., Lopes, J., Behar, D.M., Calafell, F., Jorde, L.B., van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D. (2011). Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol. Biol. Evol. *28*, 1099–1110.
9. Veeramah, K.R., Wegmann, D., Woerner, A., Mendez, F.L., Watkins, J.C., Destro-Bisol, G., Soodyall, H., Louie, L., and Hammer, M.F. (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol. Biol. Evol. *29*, 617–630.
10. Aimé, C., Laval, G., Patin, E., Verdu, P., Ségurel, L., Chaix, R., Hegay, T., Quintana-Murci, L., Heyer, E., and Austerlitz, F. (2013). Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. Mol. Biol. Evol. *30*, 2629–2644.
11. Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S., et al. (2008). Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc. Natl. Acad. Sci. USA *105*, 1596–1601.
12. Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., et al. (2014). The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat. Commun. *5*, 3163.
13. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
14. Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G., et al. (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet. *8*, e1002641.
15. Migliano, A.B., Romero, I.G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., Huang, D.W., Sherman, B.T., Siddle, K., Scholes, C., et al. (2013). Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. Hum. Biol. *85*, 251–284.
16. Becker, N.S., Verdu, P., Georges, M., Duquesnoy, P., Froment, A., Amselem, S., Le Bouc, Y., and Heyer, E. (2013). The role of GHR and IGF1 genes in the genetic determination of African pygmies' short stature. Eur. J. Hum. Genet. *21*, 653–658.
17. Pemberton, T.J., Verdu, P., Becker, N.S., Willer, C.J., Hewlett, B.S., Le Bomin, S., Froment, A., Rosenberg, N.A., and Heyer, E. (2018). A genome scan for genes underlying adult body size differences between Central African hunter-gatherers and farmers. Hum. Genet. *137*, 487–509.
18. Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell *150*, 457–469.
19. Mendizabal, I., Marigorta, U.M., Lao, O., and Comas, D. (2012). Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum. Genet. *131*, 1305–1317.
20. Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nédélec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S.L., et al. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proc. Natl. Acad. Sci. USA *111*, E3596–E3603.
21. Amorim, C.E., Daub, J.T., Salzano, F.M., Foll, M., and Excoffier, L. (2015). Detection of convergent genome-wide signals of adaptation to tropical forests in humans. PLoS ONE *10*, e0121557.
22. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., et al. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. *20*, 82.
23. Bergey, C.M., Lopez, M., Harrison, G.F., Patin, E., Cohen, J.A., Quintana-Murci, L., Barreiro, L.B., and Perry, G.H. (2018). Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. Proc. Natl. Acad. Sci. USA *115*, E11256–E11263.
24. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543–546.
25. Fagny, M., Patin, E., MacIsaac, J.L., Rotival, M., Flutre, T., Jones, M.J., Siddle, K.J., Quach, H., Harmant, C., McEwen, L.M., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat. Commun. *6*, 10047.
26. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
28. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.
29. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
30. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., et al.; 1000 Genomes Project (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.
31. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
32. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
33. Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
34. Andersen, K.G., Shylakhter, I., Tabrizi, S., Grossman, S.R., Happi, C.T., and Sabeti, P.C. (2012). Genome-wide scans provide evidence for

positive selection of genes implicated in Lassa fever. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 868–877.

35. Fantauzzo, K.A., and Christiano, A.M. (2012). Trps1 activates a network of secreted Wnt inhibitors and transcription factors crucial to vibrissa follicle morphogenesis. Development *139*, 203–214.
36. Wuelling, M., Kaiser, F.J., Buelens, L.A., Braunholz, D., Shivdasani, R.A., Depping, R., and Vortkamp, A. (2009). Trps1, a regulator of chondrocyte proliferation and differentiation, interacts with the activator form of Gli3. Dev. Biol. *328*, 40–53.
37. Yosef, N., Shalek, A.K., Gaublomme, J.T., Jin, H., Lee, Y., Awasthi, A., Wu, C., Karwacz, K., Xiao, S., Jorgolli, M., et al. (2013). Dynamic regulatory network controlling TH17 cell differentiation. Nature *496*, 461–468.
38. Dunn-Fletcher, C.E., Muglia, L.M., Pavlicev, M., Wolf, G., Sun, M.A., Hu, Y.C., Huffman, E., Tumukuntala, S., Thiele, K., Mukherjee, A., et al. (2018). Anthropoid primate-specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. PLoS Biol. *16*, e2006337.
39. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell *175*, 1701–1715.e16.
40. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell *167*, 643–656.e17.
41. Ngo, H.T., Pham, L.V., Kim, J.W., Lim, Y.S., and Hwang, S.B. (2013). Modulation of mitogen-activated protein kinase-activated protein kinase 3 by hepatitis C virus core protein. J. Virol. *87*, 5718–5731.
42. Tillmanns, J., Hoffmann, D., Habbaba, Y., Schmitto, J.D., Sedding, D., Fraccarollo, D., Galuppo, P., and Bauersachs, J. (2015). Fibroblast activation protein alpha expression identifies activated fibroblasts after myocardial infarction. J. Mol. Cell. Cardiol. *87*, 194–203.
43. Andrade, W.A., Silva, A.M., Alves, V.S., Salgado, A.P., Melo, M.B., Andrade, H.M., Dall'Orto, F.V., Garcia, S.A., Silveira, T.N., and Gazzinelli, R.T. (2010). Early endosome localization and activity of RasGEF1b, a toll-like receptor-inducible Ras guanine-nucleotide exchange factor. Genes Immun. *11*, 447–457.
44. Nemec, S., Luxey, M., Jain, D., Huang Sung, A., Pastinen, T., and Drouin, J. (2017). *Pitx1* directly modulates the core limb development program to implement hindlimb identity. Development *144*, 3325–3335.
45. Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. *14*, e1007371.
46. Szeto, D.P., Rodriguez-Esteban, C., Ryan, A.K., O'Connell, S.M., Liu, F., Kioussi, C., Gleiberman, A.S., Izpisúa-Belmonte, J.C., and Rosenfeld, M.G. (1999). Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. Genes Dev. *13*, 484–494.
47. Island, M.L., Mesplede, T., Darracq, N., Bandu, M.T., Christeff, N., Djian, P., Drouin, J., and Navarro, S. (2002). Repression by homeoprotein pitx1 of virus-induced interferon a promoters is mediated by physical interaction and trans repression of IRF3 and IRF7. Mol. Cell. Biol. *22*, 7120–7133.
48. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.
49. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.
50. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
51. Migliano, A.B., Vinicius, L., and Lahr, M.M. (2007). Life history trade-offs explain the evolution of human pygmies. Proc. Natl. Acad. Sci. USA *104*, 20216–20219.
52. Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., De Souza, R., Djurovic, G., Hames, R., Hurtado, A.M., Kaplan, H., et al. (2006). Growth rates and life histories in twenty-two small-scale societies. Am. J. Hum. Biol. *18*, 295–311.
53. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
54. Kim, M.S., Rådinger, M., and Gilfillan, A.M. (2008). The multiple roles of phosphoinositide 3-kinase in mast cell biology. Trends Immunol. *29*, 493–501.
55. Malek, T.R., and Castro, I. (2010). Interleukin-2 receptor signaling: at the interface between tolerance and immunity. Immunity *33*, 153–165.
56. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. *98*, 5–21.
57. Enard, D., and Petrov, D.A. (2018). Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. Cell *175*, 360–371.e13.
58. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
59. Younis, S., Schönke, M., Massart, J., Hjortebjerg, R., Sundström, E., Gustafson, U., Björnholm, M., Krook, A., Frystyk, J., Zierath, J.R., and Andersson, L. (2018). The ZBED6-IGF2 axis has a major effect on growth of skeletal muscle and internal organs in placental mammals. Proc. Natl. Acad. Sci. USA *115*, E2048–E2057.
60. Younis, S., Kamel, W., Falkeborn, T., Wang, H., Yu, D., Daniels, R., Essand, M., Hinkula, J., Akusjärvi, G., and Andersson, L. (2018). Multiple nuclear-replicating viruses require the stress-induced protein ZC3H11A for efficient growth. Proc. Natl. Acad. Sci. USA *115*, E3808–E3816.
61. Humphries, D.E., Wong, G.W., Friend, D.S., Gurish, M.F., Qiu, W.T., Huang, C., Sharpe, A.H., and Stevens, R.L. (1999). Heparin is essential for the storage of specific granule proteases in mast cells. Nature *400*, 769–772.
62. Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009). Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. *26*, 649–658.
63. Momeni, P., Glöckner, G., Schmidt, O., von Holtum, D., Albrecht, B., Gillessen-Kaesbach, G., Hennekam, R., Meinecke, P., Zabel, B., Rosenthal, A., et al. (2000). Mutations in a new gene, encoding a zinc-finger protein, cause tricho-rhino-phalangeal syndrome type I. Nat. Genet. *24*, 71–74.
64. Fantauzzo, K.A., Tadin-Strapps, M., You, Y., Mentzer, S.E., Baumeister, F.A., Cianfarani, S., Van Maldergem, L., Warburton, D., Sundberg, J.P., and Christiano, A.M. (2008). A position effect on TRPS1 is associated with Ambras syndrome in humans and the Koala phenotype in mice. Hum. Mol. Genet. *17*, 3539–3551.
65. Hillmer, E.J., Zhang, H., Li, H.S., and Watowich, S.S. (2016). STAT3 signaling in immunity. Cytokine Growth Factor Rev. *31*, 1–15.
66. Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R.B., Pérez, G.M., et al. (2016). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat. Commun. *7*, 11616.
67. Chopin, M., Preston, S.P., Lun, A.T.L., Tellier, J., Smyth, G.K., Pellegrini, M., Belz, G.T., Corcoran, L.M., Visvader, J.E., Wu, L., and Nutt, S.L. (2016). RUNX2 mediates plasmacytoid dendritic cell egress from the bone marrow and controls viral immunity. Cell Rep. *15*, 866–878.

68. Durvasula, A., and Sankararaman, S. (2019). Recovering signals of ghost archaic introgression in African populations. bioRxiv. https://doi.org/10.1101/285734.

69. Hsieh, P., Woerner, A.E., Wall, J.D., Lachance, J., Tishkoff, S.A., Gutenkunst, R.N., and Hammer, M.F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res. *26*, 291–300.

70. Foupouapouognigni, Y., Mba, S.A., Betsem à Betsem, E., Rousset, D., Froment, A., Gessain, A., and Njouom, R. (2011). Hepatitis B and C virus infections in the three Pygmy groups in Cameroon. J. Clin. Microbiol. *49*, 737–740.

71. Kowo, M.P., Goubau, P., Ndam, E.C., Njoya, O., Sasaki, S., Seghers, V., and Kesteloot, H. (1995). Prevalence of hepatitis C virus and other blood-borne viruses in Pygmies and neighbouring Bantus in southern Cameroon. Trans. R. Soc. Trop. Med. Hyg. *89*, 484–486.

72. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.

73. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, e39702.

74. Becker, N., Touraille, P., Froment, A., Heyer, E., and Courtiol, A. (2012). Short stature in African pygmies is not explained by sexual selection. Evol. Hum. Behav. *33*, 615–622.

75. Perry, G.H., and Dominy, N.J. (2009). Evolution of the human pygmy phenotype. Trends Ecol. Evol. *24*, 218–225.

76. Harrison, G.F., Sanz, J., Boulais, J., Mina, M.J., Grenier, J.-C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C.M., Elledge, S.J., et al. (2019). Natural selection contributed to immunological differences between human hunter-gatherers and agriculturalists. Nat Ecol Evol, in press.

77. Owers, K.A., Sjödin, P., Schlebusch, C.M., Skoglund, P., Soodyall, H., and Jakobsson, M. (2017). Adaptation to infectious disease exposure in indigenous Southern African populations. Proc. Biol. Sci. *284*, 20170226.

78. Hawkins, P.T., and Stephens, L.R. (2015). PI3K signalling in inflammation. Biochim. Biophys. Acta *1851*, 882–897.

79. Hopkins, B.D., Pauli, C., Du, X., Wang, D.G., Li, X., Wu, D., Amadiume, S.C., Goncalves, M.D., Hodakoski, C., Lundquist, M.R., et al. (2018). Suppression of insulin feedback enhances the efficacy of PI3K inhibitors. Nature *560*, 499–503.

80. Fruman, D.A., Chiu, H., Hopkins, B.D., Bagrodia, S., Cantley, L.C., and Abraham, R.T. (2017). The PI3K pathway in human disease. Cell *170*, 605–635.

81. Odegaard, J.I., and Chawla, A. (2013). Pleiotropic actions of insulin resistance and inflammation in metabolic homeostasis. Science *339*, 172–177.

82. Smith, T.J. (2010). Insulin-like growth factor-I regulation of immune function: a potential therapeutic target in autoimmune diseases? Pharmacol. Rev. *62*, 199–236.

83. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

84. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

85. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

86. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

87. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

88. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

89. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

90. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., et al. (2018). GOATOOLS: a Python library for Gene Ontology analyses. Sci. Rep. *8*, 10872.

91. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

92. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

93. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 1–33.

94. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

95. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

96. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

97. Cavalli-Sforza, L.L. (1986). African Pygmies (Academic Press).

98. Diamond, J.M. (1991). Anthropology. Why are pygmies small? Nature *354*, 111–112.

99. Shea, B.T., and Bailey, R.C. (1996). Allometry and adaptation of body proportions and stature in African pygmies. Am. J. Phys. Anthropol. *100*, 311–340.

100. Froment, A. (2001). Hunter-gatherers: an interdisciplinary perspective. In Hunter-Gatherers: An Interdisciplinary Perspective, L.R.-C. Panter-Brick, ed. (Cambridge University Press), pp. 239–266.

101. Becker, N.S., Verdu, P., Hewlett, B., and Pavard, S. (2010). Can life history trade-offs explain the evolution of short stature in human pygmies? A response to Migliano et al. (2007). Hum. Biol. *82*, 17–27.

102. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Subheading		
Nextera Rapid Capture Expanded Exome kit	Illumina	Cat#FC-140-1006
HumanOmniExpress-24 v1.1 DNA Analysis Kit	Illumina	N/A
Deposited Data		
Exome and whole-genome sequencing	This paper	EGAS00001003722
Software and Algorithms		
PLINK v1.9	[83]	http://www.cog-genomics.org/plink/1.9/
KING v1.4	[84]	http://people.virginia.edu/~wc9c/KING/history.htm
ADMIXTURE	[26]	http://software.genetics.ucla.edu/admixture/download.html
BWA v.0.7.7	[85]	http://bio-bwa.sourceforge.net/
Picard Tools v.1.94	N/A	http://broadinstitute.github.io/picard
GATK v3.5	[86]	https://software.broadinstitute.org/gatk/download/
SHAPEIT2	[87]	http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
IMPUTE v.2	[88]	http://mathgen.stats.ox.ac.uk/impute/impute_v2.1.0.html
LDSC	[89]	https://data.broadinstitute.org/alkesgroup/LDSCORE/
GOATOOLS	[90]	https://github.com/tanghaibao/goatools
RFMIX v1.5.4	[58]	https://sites.google.com/site/rfmixlocalancestryinference/
BEAGLE	[91]	https://faculty.washington.edu/browning/beagle/beagle.html
GERP++	[92]	http://mendel.stanford.edu/SidowLab/downloads/gerp/

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents. Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Lluís Quintana-Murci (quintana@pasteur.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection

Sampling consisted in human saliva or blood from 157 rainforest hunter-gatherers and 120 farmers from western and eastern central Africa (Figure S1), including 208 males and 69 females. Informed consent was obtained from all participants in this study, which was overseen by the institutional review board of Institut Pasteur (2011-54/IRB/8), the *Comité National d'Ethique du Gabon* (0016/2016/SG/CNE), the University of Chicago (IRB 16986A) and Makerere University, Kampala, Uganda (IRB 2009-137). The 277 new samples collected for exome sequencing were analyzed together with 317 exomes of central Africans from Lopez et al. 2018 [4] and 101 Europeans from Quach et al. 2016 [40] (Table S1).

METHOD DETAILS

Exome Sequencing

Sample libraries were prepared with the Nextera Rapid Capture Expanded Exome Kit, which delivers 62Mb of genomic content per individual, including exons, untranslated regions and microRNAs, and were sequenced on Illumina HiSeq2500 machines. Using the GATK Best Practices recommendations [93], pairs of 101-bp reads were mapped onto the human reference genome (GRCh37) with Burrows-Wheeler Aligner (BWA) version 0.7.7 [85], using 'bwa mem -M -t 4 -R', and reads duplicating the start position of another read were marked as duplicates with Picard Tools version 1.94 (http://broadinstitute.github.io/picard/), using 'MarkDuplicates'. We used GATK version 3.5 [86] for base quality score recalibration ('Base Recalibrator'), insertion/deletion (indel) realignment ('IndelRealigner'), and SNP and indel discovery ('Haplotype Caller') for each sample. Individual variant files were combined with 'GenotypeGVCFs' and filtered with 'VariantQualityScoreRecalibration'. We used high confidence variants from the 1000G Phase 1 and HapMap 3 projects [94, 95] as VQSR training callsets, and applied a tranche sensitivity threshold of 99.5%. From the 947,523 sites detected, we removed indels as well as SNPs that (i) were located on the sex chromosomes,

(ii) were not biallelic, (iii) were monomorphic in our total sample, (iv) had a depth of coverage < 5 ×, (v) had a genotype quality score (GQ) < 20, (vi) presented missingness > 15%, and (vii) presented a Hardy-Weinberg test $p < 10^{-6}$ in at least one of population. As criteria to remove low-quality samples, we required a total genotype missingness < 15% (21 excluded samples). In addition, we checked for unexpectedly high or low heterozygosity values, suggesting high levels of inbreeding or DNA contamination, and excluded 3 individuals presenting heterozygosity levels 4 SD higher than their population average. We thus retained exome data for 671 individuals, with an average depth of coverage after duplicate removal of 38 × (SD: 9 ×), ranging from 25 × to 95 ×. The application of these quality-control filters resulted in a final dataset of 682,468 SNPs (Figure S1), of which 107,621 SNPs were polymorphic only in the 268 newly-sequenced individuals.

SNP Array Data

In addition to exome sequencing, we retrieved the genotyping data of the same 671 individuals from Quach et al. 2016 [40], Patin et al. 2014 [12], Patin et al. 2017 [24] and Fagny et al. 2015 [25] (Figure S1; Table S1). We removed SNPs located on the X and Y chromosomes, problematic genotype clustering profiles (i.e., Illumina GenTrain score < 0.35) or with call rate < 95%. We kept 599,559 SNPs common to different genotyping SNP arrays. We removed a total of 53 C/G or A/T SNPs to prevent misaligned SNPs, and excluded a total 5 additional SNPs that were under Hardy-Weinberg disequilibrium in at least one of the populations ($p < 10^{-6}$) using PLINK [96], leading to a final dataset of 559,501 SNPs.

We applied additional filters on the genotyping dataset of the 671 individuals retained for exome sequencing. We removed two individuals with heterozygosity levels higher or lower than the population mean ± 4 SD Although related individuals were avoided during the sampling and for exome sequencing (based on published SNP array data) [5, 12, 17, 24, 25], we sought to exclude possibly remaining pairs of cryptically related individuals. Indeed, RHG populations are small isolated communities, where individuals can be related to many others. We considered that two individuals were strongly (cryptically) related if they presented a first-degree relationship (kinship coefficient > 0.177), as inferred by KING [84]. Following this criterion, only one individual was removed. Additionally, we removed another individual who did not present any first-degree relatedness but was related in second-degree to many others. After removing these two individuals, the dataset included 77 and 232 pairs of second-degree (kinship coefficient > 0.0884) and third-degree (kinship coefficient > 0.0442) related RHG individuals, respectively. The application of these quality-control filters resulted in a final genotyping dataset of 667 individuals and 599,501 SNPs (Figure S1).

Merging Exome and SNP Array Data

Before merging the genotyping array and the exome data from the 667 high-quality individuals in common, we flipped alleles for 8,393 SNPs with incompatible allelic states, and removed 9 SNPs with alleles that remained incompatible after allele flipping from the genotyping dataset. The total concordance rate was evaluated on 28,403 SNPs common to both datasets. The concordance rates for each of the 667 individuals exceeded 98%, confirming an absence of errors during DNA sample processing. The entire genotyping and exome datasets (599,492 and 682,468 SNPs, respectively) were then merged, yielding a final dataset of 1,253,548 SNPs for 667 individuals, 566 of whom were African farmers or hunter-gatherers (Figure S1).

Whole-Genome Sequencing

We generated whole genomes of 20 RHG Baka and 20 AGR Nzébi of Gabon, which were also part of the exome and SNP array datasets. All the samples were processed using the paired-end library preparation protocol from Illumina. Libraries were sequenced on Illumina HiSeq 2000 machines at the Stanford Center for Genomics and Personalized Medicine. 101-pb reads were aligned to the human reference genome (GRCh37) using BWA [85], followed by base quality recalibration and realignment around known indels with GATK [86]. Genotyping was carried out across all 40 individuals jointly using GATK 'UnifiedGenotyper', and called variants were stratified into variant quality tranches using 'VariantQualityScoreRecalibration' tool (VQSR) from GATK. SNPs with a VQSR tranche > 99.0 were considered as confidently called. Genotype calls were refined and improved based on LD using BEAGLE [91], yielding a final dataset of 17,687,206 variants (Figure S1). All individuals presented very low rates of missing values ranging from 0.5% to 4%, and a mean depth of coverage of 6.5 × (ranging from 4 × to 13 ×).

Imputation of SNP Array and Exome Data

Before imputation, we phased the data with SHAPEIT2 using 100 states, 20 MCMC main steps, 7 burnin and 8 pruning steps [87]. SNPs and allelic states were then aligned with the 1000 Genomes Project imputation reference panel (Phase 3 [27]), referred to as 'reference panel 1', as well as the 40 whole genomes of Baka RHG and Nzébi AGR of Gabon, referred to as 'reference panel 2' (Figure S1). We removed from the reference panels SNPs with MAF < 1%, SNPs with C/G or A/T alleles and 414,679 multiallelic SNPs in the reference panel 1. We evaluated the allelic concordance between the two reference panels and excluded 9,649 additional sites from the reference panel 2, yielding to final datasets of 11,501,018 SNPs in the reference panel 1 and 14,252,666 SNPs in the reference panel 2.

Genotype imputation was performed with IMPUTE v.2 [88] considering 1-Mb windows and both reference panels simultaneously, with the '-merge_ref_panels' option. We used genotype calls instead of genotype probabilities, which are not handled by downstream programs, and considered as confident genotype calls genotypes with posterior probability > 0.8. Of the 13,092,258 SNPs obtained after imputation, we removed SNPs that: (i) presented an information metric < 0.8, (ii) had a duplicate, (iii) presented a call rate < 95%, and (iv) were monomorphic. The final imputed dataset included 10,262,236 SNPs, and 9,129,103 after filtering

SNPs with MAF < 1%. To evaluate imputation accuracy, we estimated correlation coefficients r^2 between true genotypes (i.e., obtained by Illumina genotyping array or exome sequencing) and imputed genotypes for the same SNPs (i.e., obtained by artificially removing genotyped SNPs from the data before imputation and then imputing them). The average correlation coefficient across all genotyped SNPs with information metric > 0.8 were 0.86 and 0.85 for reference panels 1 and 2, respectively, showing that our quality filters ensure to keep accurately imputed SNPs for further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome Scans for Selective Sweeps

Genomic regions candidate for positive selection were detected in seven populations of RHG (Bezan, Baka, BaBongo of central Gabon, BaKoya, BaBongo of south and east Gabon and BaTwa) and two populations of AGR (western and eastern AGR), with an outlier approach that considers two interpopulation statistics: PBS (Population Branch Score [28]), and XP-EHH [29]. We combined these scores into a Fisher's score (F_{CS}) equal to the sum, over the two statistics, of $-\log_{10}$(rank of the statistic for a given SNP/number of SNPs). Interpopulation statistics require a reference population, and PBS statistics an outgroup population. We performed separate scans of classic sweeps for each population, using Europeans as outgroup, and different reference populations: western AGR for each western RHG population, eastern AGR for eastern RHG, pooled western RHG for western AGR, and eastern RHG for eastern AGR. PBS was calculated for each SNP using AMOVA-based F_{ST} values computed with home-made scripts (available upon request). The derived allele of each SNP was defined based on the 6-EPO alignment. XP-EHH was computed in 100-kb sliding windows with a 50-kb pace, with home-made scripts (available upon request). Only SNPs with a derived allele frequency (DAF) between 10% and 90% were analyzed further. XP-EHH scores were normalized in 40 separate bins of DAF. An outlier SNP was defined as a SNP with an F_{CS} among the 1% highest of the genome. A putatively selected genomic region was defined as a 100-kb window presenting a proportion of outlier SNPs among the 1% highest of all windows, in five bins of SNP numbers. Windows containing less than 50 SNPs were discarded as well as 500-kb regions around gaps, to avoid biases in the outlier enrichment scores.

Polygenic Selection of Complex Traits

We retrieved the results of the Genome Wide Association studies from UK BIOBANK (round 2, http://www.nealelab.is/uk-biobank/) of 12 complex traits that we selected as candidates for adaptation of RHG, based on previous hypotheses from biological anthropology studies [51, 73, 97–101]. Our genomic dataset was split into non-overlapping 100-kb windows. We considered a window as associated with a trait if it included a SNP with a genome-wide significant association with this trait ($P_{assoc}<5 \times 10^{-8}$). We computed for each genomic window, associated or not with the trait, the average F_{CS}, the proportion of conserved SNP positions based on GERP scores > 2 [92], and the recombination rate using the combined HapMap genetic map [102], to account for the confounding effects of background selection.

In order to test for polygenic selection, we generated a null distribution by randomly sampling *x* windows (*x* being the number of windows associated with a tested trait) among windows with a similar number of SNPs, proportion of GERP > 2 sites and recombination rate observed in the trait-associated windows. We then calculated the average of the mean of the F_{CS} across the *x* resampled windows. We resampled 100,000 sets of *x* windows for each trait. To test for significance, we computed a resampling *P*-value by calculating the proportion of resampled windows which mean F_{CS} was higher than that observed for the tested trait. All *P*-values for polygenic adaptation were then adjusted for multiple testing by the Benjamini-Hochberg method, to account for the number of traits tested, and traits with an adjusted $p < 0.05$ were considered as candidates for polygenic selection.

To test if polygenic selection signals are due to pleiotropy of height-associated genes, we first estimated genetic correlations between candidate traits from LD-score regression using the ldsc tool [89]. We used precomputed European LD-scores (https://data.broadinstitute.org/alkesgroup/LDSCORE/). *P*-values were corrected for multiple testing using the Bonferroni correction, and adjusted *P*-values < 0.05 were considered as significant.

To correct for pleiotropy for each trait genetically correlated with height, we removed windows significantly associated with 'Standing Height' and 'Comparative height at age 10' in both windows associated with the candidate trait and resampled windows. Similarly, we re-tested for polygenic adaptation on ''Standing height'' and ''Comparative height at age 10'' associated regions using the same approach, but by removing all trait-associated windows, except height-associated windows. To test if loss of significance was due to a decrease in power, we down-sampled the number of tested trait-associated windows to the same number as after removing height-associated windows. We down-sampled a 100 times trait-associated windows, and estimated a hundred *P*-values as described above. We finally compared the distribution of the 100 obtained *P*-values with the estimated *P*-value (non-adjusted for multiple testing) both before and after removing height-associated windows.

Polygenic Selection of Gene Ontologies

To detect enrichment of F_{CS} scores in sets of genes corresponding to a given biological pathway, we compared the distributions of F_{CS} between genes that were part of the gene ontology (GO) term tested, relative to the rest of the genes of the genome, using a Mann-Whitney-Wilcoxon rank-sum test. To limit the effect of clusters of genes on the enrichment calculation, we assigned to each 100-kb non-overlapping genomic window both a GO term, based on the presence of at least one gene from the corresponding term, and a mean F_{CS} score. We tested if mean F_{CS} of windows assigned to a given GO term were different from genome-wide expectations, accounting for multiple testing. We restricted the enrichment analysis to 5,354 GO terms with levels comprised between

levels 3 and 7 [53], using the python library goatools [90], and that include at least 5 genes. We examined a total of 15,503 windows and determined *P*-values corresponding to 5% and 1% of false discoveries, FDR p = 9.24×10^{-3} and FDR p = 4.03×10^{-4}, respectively, by randomly resampling *y* genes (*y* being sampled from the distribution of the number of genes assigned to each GO term). We also studied additional gene sets, including 1,553 manually-curated genes involved in innate immunity [56] and 1,257 genes encoding proteins known to have physical interactions with multiple families of viruses [57].

Local Ancestry Inference

To perform local ancestry inference in the genomes of the highly-admixed BaBongo RHG from south and east Gabon, we first constituted putative parental populations that were representative of RHG and AGR ancestry. We considered as the parental AGR population, 163 individuals with less than 20% of their ancestry assigned to the RHG component, based on the ADMIXTURE analysis at $K = 5$. Likewise, we considered as the parental RHG population, 101 individuals with less than 5% AGR ancestry. The genomes of the highly-admixed BaBongo were decomposed into segments of RHG or AGR ancestry with RFMix v.1.5.4 [58], including two EM steps. We excluded 2-Mb regions from the telomeres of each chromosome. Based on RFMix ancestry estimations, the mean AGR ancestry was 94% [SD = 1.6%] in the parental AGR population, 62% [SD = 5.9%] in the highly-admixed BaBongo, and 27% [SD = 3.7%] in the parental RHG population. These ancestry proportions were highly correlated with ADMIXTURE membership proportions at $K = 2$ (Pearson's correlation coefficient $R^2 = 0.99$). We then searched for excesses in RHG or AGR ancestry in pathways by assigning ancestry proportions to 100-kb windows across the genome, with the same approach used for GO enrichments.

DATA AND CODE AVAILABILITY

The newly generated exomes (n = 266) and genomes (n = 40) of central African rainforest hunter-gatherers and agriculturalists have been deposited in the European Genome-phenome Archive under accession code EGAS00001003722. Data accessibility is restricted to academic research on human genetic history and adaptation. Exome sequencing data for the remaining, previously published samples are available under accession codes EGAS00001002457 and EGAS00001001895.