Sélection naturelle et adaptation aux changements rapides de pressions environnementales chez l'Homme

Marie Lopez

▶ **To cite this version:**

Marie Lopez. Sélection naturelle et adaptation aux changements rapides de pressions environnementales chez l'Homme. Génétique des populations [q-bio.PE]. Sorbonne Université, 2018. Français. NNT : 2018SORUS605 . tel-02968134

HAL Id: tel-02968134
https://theses.hal.science/tel-02968134v1

Submitted on 15 Oct 2020

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

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

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

Spécialité : Génétique des populations humaines

École doctorale n°515: Complexité du Vivant

présentée par

Marie LOPEZ

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

Sujet de la thèse :

Sélection naturelle et adaptation aux changements rapides de pressions environnementales chez l'Homme

soutenue le 27 septembre 2018

devant le jury composé de :

M.	François-Xavier RICAUT	Rapporteur
M.	Vincent CASTRIC	Rapporteur
M^{me}	Molly PRZEWORSKI	Examinatrice
M^{me}	Evelyne HEYER	Examinatrice
M.	Dominique HIGUET	Représentant de l'Université
M.	Lluis QUINTANA-MURCI	Directeur de thèse

Remerciements

En tout premier lieu, je tiens à remercier les membres du jury : M. François-Xavier Ricaut, M. Vincent Castric, Mme. Evelyne Heyer, Mme. Molly Przeworski, et M. Dominique Higuet qui me font l'honneur et le plaisir d'évaluer ce travail de thèse et de participer à cette soutenance. Parmi eux, j'aimerais tout particulièrement exprimer ma gratitude à Molly qui, malgré ma grande inexpérience, m'a donné l'opportunité de découvrir la génétique des populations humaines et de partager la vie de son laboratoire au cours de mon stage de Master 1.

Mes remerciements les plus sincères vont ensuite à Lluis Quintana-Murci, le directeur de ma thèse. Lluis, je te remercie pour la confiance que tu m'as accordée en me laissant rejoindre ton équipe, ainsi que pour ton encadrement attentif tout au long de ces quatre années. C'est une vraie chance que d'avoir pu réaliser ce travail de thèse dans ton laboratoire, duquel je ressors immensément grandie, tant sur le plan scientifique que personnel. Je te remercie d'avoir tout mis en œuvre pour assurer le succès des différents projets en me donnant toujours l'opportunité de mettre en avant mon travail. Merci de m'avoir écoutée, encouragée à chacune des étapes et d'avoir fait naître en moi l'envie de faire toujours mieux. Même si nos chemins ne se recroisent pas, je te remercie et te souhaite le meilleur pour la suite des projets actuels et futurs du laboratoire.

Mes pensées se tournent à présent vers l'ensemble des membres de l'unité de Génétique Évolutive Humaine. Tout d'abord, j'aimerais remercier chaleureusement Etienne Patin pour tout le temps qu'il a consacré à mon encadrement, et dont la très grande pertinence scientifique et l'extrême bienveillance ont largement éclairé le ou plutôt les chemins à suivre au cours de ce travail. Je te remercie sincèrement et te souhaite une réussite à la hauteur de la générosité et de l'énergie que tu déploies pour la réussite de chacun. Un grand merci également à Guillaume Laval et à Maxime Rotival pour l'aide qu'ils m'ont si gentiment et si patiemment apporté, en particulier en statistique. Enfin, j'aimerais spécialement remercier Hélène Quach pour ses encouragements et pour toutes les petites attentions du quotidien, ainsi que mes compagnons de thèse et les autres membres du laboratoire pour leur bonne humeur et leur assistance.

Cette thèse doit également beaucoup à l'ensemble des collaborateurs sans lesquels ce travail n'aurait pas été possible. En premier lieu, merci à Athanasios Kousathanas qui a largement contribué au succès de la première partie de ce projet et avec qui le travail fut intense et extrêmement enrichissant. Je tiens également à remercier l'ensemble des collaborateurs qui ont contribué mon travail de thèse : Paul Verdu, Georges Perry, Christina Bergey, Luis Barreiro, Frédéric Austerlitz, Michael Blum, Laurent Abel et Antonio Rausell ainsi que les membres de mon comité de thèse : Dominique Higuet, Laure Ségurel et Guillaume Achaz pour le regard critique qu'ils ont porté sur mon travail et leur souci du bon déroulement de cette thèse.

Un grand merci aux miens, à la tribu dont j'ai l'immense chance d'appartenir et qui remplit chaque jour ma vie d'un amour inconditionnel et véritable. Merci à mes parents et à ma sœur, qui ont partagé chaque joie et chaque doute de ce travail. Je ne pourrais être plus reconnaissante pour votre soutien sans faille, quoi que je choisisse d'entreprendre. Merci à mes grands-parents adorés dont le courage et la résilience forcent l'admiration, ce travail vous est dédié puisqu'il existe pour vous rendre, en partie, ce que vous avez choisi de nous transmettre. Merci à ma tante et à mes presque-sœurs de croire en moi, tout ce que je vis seule, nous le vivons aussi tous les neuf.

Enfin, Pierrick, pour m'avoir épaulée et encouragée, relevée et valorisée, aimée chaque instant, je te remercie du plus profond de mon âme car sans toi à mes côtés je ne serais sans doute pas la même personne aujourd'hui.

Table des figures

Liste des tableaux

Liste des abréviations

ABC	Approximate Bayesian Computation
A	Adénosine
ADN	Acide Désoxyribonucléique
C	Cytosine
CMS	Composite of Multiple Signals
DFE	Distribution of Fitness Effects
DGVa	Database of Genomic Variants archive
DL	Déséquilibre de Liaison
EHH	Extended Haplotype Homozygosity
eQTL	expression Quantitative Trait Loci
G	Guanine
GWAS	Genome-Wide Association Study
iHS	integrated Haplotype Score
IMC	Indice de Masse Corporelle
kb	Kilobases
Mb	Mégabases
pb	Paires de bases
RCA	République Centrafricaine
RDC	République Démocratique du Congo
SFS	Site Frequency Spectrum
SNP	Single Nucleotide Polymorphism
XP-EHH	Cross-Population Extended Haplotype Homozygosity
T	Thymine

Table des matières

Introduction

L'Afrique centrale, qui s'étend du Bassin du Congo jusqu'au Lac Victoria, est une région couverte de forêts équatoriales denses qui abrite une biodiversité parmi les plus riches au monde. Cette région clé de la préhistoire africaine a conditionné les dynamiques et les modes de vie des populations humaines depuis l'émergence d'*Homo sapiens* en Afrique il y a environ 200 000 ans. Ce territoire abrite aujourd'hui le plus grand groupe de chasseurs-cueilleurs forestiers au monde, historiquement appelés *Pygmées* et considérés comme l'un des derniers exemples d'un mode de vie ayant prédominé au cours de notre évolution.

Les communautés de Pygmées occupent traditionnellement des campements semi-nomades au sein des forêts équatoriales, cet habitat singulier se caractérise par un climat chaud et humide, des ressources alimentaires qui fluctuent avec les saisons et la présence de nombreux pathogènes. Par ailleurs les Pygmées entretiennent des relations socio-économiques étroites avec des populations sédentaires "non-Pygmées", arrivées dans les zones de savane d'Afrique centrale avec l'émergence de l'agriculture il y a environ 5 000 ans. Ainsi, les modes de subsistance traditionnels des Pygmées et non-Pygmées (chasseurs-cueilleurs vs agriculteurs) sont le reflet de leurs habitats différents (forêt vs savane) et s'accompagnent de différences socio-culturelles. Cette diversité de pratiques interroge sur l'histoire plus ancienne de ces populations, en particulier sur leurs passés démographiques et leurs modes de vie ancestraux.

En raison de la quasi-absence de données archéologiques dans les zones forestières d'Afrique centrale, il est extrêmement difficile d'obtenir des informations concernant la distribution géographique et temporelle des populations ancestrales de Pygmées et de non-Pygmées. Cependant, cette histoire démographique a conditionné la diversité génétique des groupes humains et l'action sur leurs génomes de différents mécanismes de sélection naturelle.

L'étude de l'histoire évolutive des Pygmées et non-Pygmées à partir de données génétiques offre donc la possibilité de reconstituer l'histoire démographique de ces populations aux modes de vie différents, mais aussi de mieux comprendre comment les fluctuations récentes des tailles de populations ont influencé les mécanismes de sélection naturelle, en particulier la sélection négative qui élimine les mutations délétères dans le génome. De plus, l'étude du génome de ces groupes permet de mieux comprendre les mécanismes de sélection positive et les fonctions biologiques ayant permis leur adaptation génétique à leurs habitats distincts.

Chapitre 1

Approche multidisciplinaire de l'histoire des populations de Pygmées d'Afrique centrale

1.1 Les populations de Pygmées du Bassin congolais

1.1.1 Découverte des Pygmées et difficultés de nomenclature

Le terme "Pygmée", qui signifie littéralement « haut d'une coudée », est dérivé du grec ancien $\pi\upsilon\gamma\mu\alpha\iota\sigma\sigma$ et apparaît pour la première fois dans l'*Iliade* d'Homère pour désigner un peuple d'une taille excessivement réduite entretenant une lutte perpétuelle pour sa survie contre des nuées migratoires de grues. Ces Pygmées évoqués par les récits mythologiques mesurent une trentaine de centimètres et vivent dans les contrées lointaines de l'Inde ou du sud de l'Égypte. Au milieu du XIX$^{\text{ème}}$ siècle, les premiers explorateurs du Bassin congolais font la rencontre de groupes humains de faible stature habitant les forêts d'Afrique centrale. Ces individus physiquement semblables entre eux mais différents d'autres populations africaines ont été indifféremment qualifiés de *Nains*, *Dwarfs*, *Négrilles*, *Zwergen* selon la nationalité des explorateurs puis baptisés *Pygmées* pour la première fois en 1873 en référence au récit d'Homère (S. Bahuchet, 1993). L'inventaire des populations de Pygmées s'est poursuivi dans les moindres recoins des forêts du Bassin du Congo dans le but de découvrir les populations aux statures les plus faibles et aux modes de vie les plus « primitifs », c'est-à-dire des populations forestières nomades subsistant grâce à la chasse et la cueillette et habitant des huttes faites de branches et de feuilles.

Ce n'est que plus tard, au cours du XX$^{\text{ème}}$ siècle, qu'apparaissent les premiers travaux ethnographiques et scientifiques d'observation des populations de Pygmées en Afrique centrale. Ces études ont mis en évidence l'existence de nombreux sous-groupes de populations au sein de la communauté jusqu'alors indistinctement nommée Pygmées, en soulignant les différences linguistiques, culturelles et technologiques entre les groupes (Schebesta, 1938, 1941, 1948, 1952). Ces travaux ont également mis en garde contre la généralisation parfois abusive et fallacieuse véhiculée par le terme unique « Pygmées » (Turnbull, 1965). L'emploi

de ce terme fait référence à la stature des individus, supposée plus faible que dans d'autres populations africaines, mais varie pourtant selon les groupes (Hewlett, 2014). Ce critère physique est parfois délaissé au profit d'autres caractéristiques liées à l'habitat et au mode de vie. Cependant, une caractérisation des populations de Pygmées uniquement basée sur un mode de vie forestier, commun à la plupart des groupes, ne ferait pas état de certaines communautés qui vivent dans la savane comme au Cameroun ou en République Démocratique du Congo (RDC) et exclurait également les populations de Pygmées qui habitent les montagnes comme au Rwanda et au Burundi. De la même manière, les modes de subsistance des Pygmées actuels sont pluriels et varient selon les régions car certains groupes pratiquent aujourd'hui la pêche ou encore l'agriculture.

Il existe donc une réelle controverse autour de la catégorisation et de la nomenclature des populations forestières d'Afrique centrale. Le terme historique de « Pygmées » ne rend pas toujours compte de la réalité complexe des populations mais il conserve néanmoins l'avantage de rassembler les différentes communautés ethniques appartenant à l'identité «Pygmées» en Afrique équatoriale et sera utilisé dans ce texte pour des raisons pratiques lorsque les précisions ethniques ne sont pas disponibles (Perry & Dominy, 2009 ; Hewlett, 2014).

1.1.2 Répartition géographique

Les Pygmées d'Afrique centrale constituent le plus grand groupe de chasseurs-cueilleurs au monde mais ne représentent qu'une fraction limitée des populations africaines actuelles (Figure 1.1 ; Table1.1). On compte au Cameroun environ 40 000 Pygmées pour 17 millions d'habitants et les densités de populations ne sont que d'un seul individu Pygmée pour 7 à 10 individus non-Pygmées par kilomètre carré (Hewlett, 2014). Les groupes de Pygmées sont réunis en campements qui dépassent rarement 40 individus alors que les villages non-Pygmées regroupent en moyenne 230 à 330 individus (Hewlett, 2014). Il existe en Afrique d'autres groupes de chasseurs-cueilleurs, distincts phénotypiquement et linguistiquement des Pygmées, qui occupent les savanes de l'est et zones désertiques du sud du continent (Figure 1.1).

La vingtaine d'ethnies que regroupe l'appellation « Pygmées » est répartie d'ouest en est de la forêt équatoriale dans le Bassin du Congo (Figure 1.1 ; Table 1.1) et toutes sont physiquement, culturellement et linguistiquement différentes. On notera que, dans la plupart des noms de populations, le préfixe Ba- indique le pluriel en langue bantoue alors que Mo- indique le singulier (MoAka : singulier, BaAka : pluriel). Dans ce texte les noms de populations seront indiqués au pluriel en utilisant une majuscule pour séparer le préfixe de la racine du nom. Les groupes de Pygmées les plus représentés dans la littérature sont les communautés semi-nomades telles que : les Baka, les BaAka et les BaMbuti séparés en trois groupes ethniques : les Efe, les Asua, et les BaSua (Hewlett, 2014). Ces populations établissent des campements temporaires au sein de la forêt en construisant des huttes en

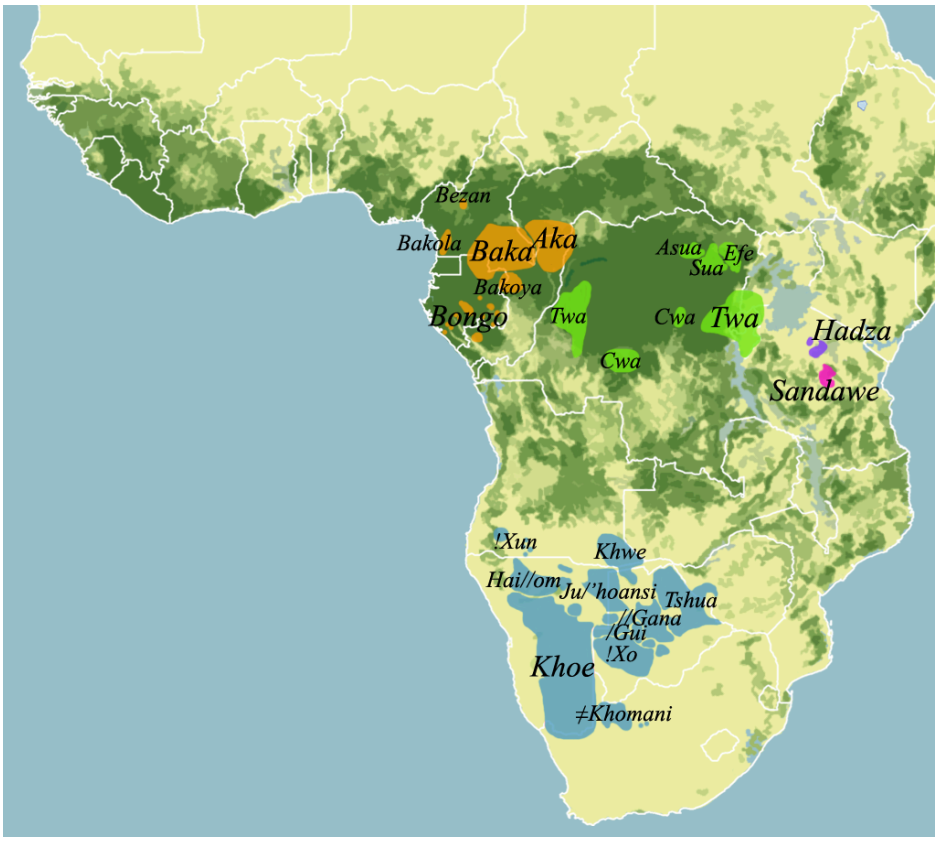


Fig. 1.1 ***Répartition géographique des populations de chasseurs-cueilleurs en Afrique sub-Saharienne.*** *Répartition actuelle des groupes de chasseurs-cueilleurs Pygmées à l'ouest (orange) et à l'est (vert) du Bassin du Congo. Les noms de populations sont indiqués au singulier et la description détaillée des différents sous-groupes est disponible dans les Tables 1.1 et 1.2. Les groupes de chasseurs-cueilleurs africains non-Pygmées sont également représentés.*

forme de dômes (Figure 1.1 ; Table 1.1). Ils subsistent principalement grâce à la chasse et la cueillette mais échangent également des produits de la forêt, comme le miel ou la viande, avec les agriculteurs voisins (Hewlett, 1996).

Il existe d'autres groupes de Pygmées qui occupent des villages en bordure de forêt et pratiquent une agriculture plus ou moins efficace en supplément de la chasse et de la cueillette (Figure 1.1 ; Table1.1). Ces groupes maintiennent des relations étroites avec les agriculteurs dont ils sont phénotypiquement proches, bien qu'ils appartiennent à deux groupes ethniques distincts. Ces groupes sont d'ouest en est : les BaKola, les BaBongo, les BaKoya, et les BaTwa (Hewlett, 2014).

Enfin, on trouve des groupes de Pygmées qui s'établissent dans la savane tels que les Bedzan, plusieurs groupes de BaCwa ainsi que les BaTwa (Figure 1.1 ; Table 1.1), une appellation qui désigne également des groupes de potiers dispersés au Rwanda et au Burundi (Hewlett, 2014).

1.1.3 Populations non-Pygmées

Les groupes de Pygmées du Bassin du Congo vivent à proximité de populations sédentaires avec qui ils établissent des échanges commerciaux de nature et d'intensité variables. Ces groupes ethniques « non-Pygmées » occupent les savanes et se distinguent des Pygmées à la fois par leur habitat, leurs modes de subsistance et leur morphologie. Ces individus « agriculteurs » sont tous des locuteurs de langues bantoues, une famille linguistique à laquelle appartient la majorité des langues parlées en Afrique subsaharienne (310 millions de locuteurs et 450 langues et dialectes) (Figure 1.2) (Klieman, 2003). Les bantous se répartissent sur environ un tiers du continent Africain, formant un trapézoïde entre le nord-ouest du Cameroun, les côtes du Kenya, la pointe de l'Afrique du Sud et le nord-ouest de la Namibie (Klieman, 2003). Au sein du Bassin du Congo, on trouve près de 155 groupes ethniques bantouphones (Joiris & Bahuchet, 1994) et l'appellation « agriculteur » peut là encore faire figure d'une généralisation ne faisant pas état des différences culturelles entre les communautés, et renforce parfois l'opposition binaire aux groupes de Pygmées.

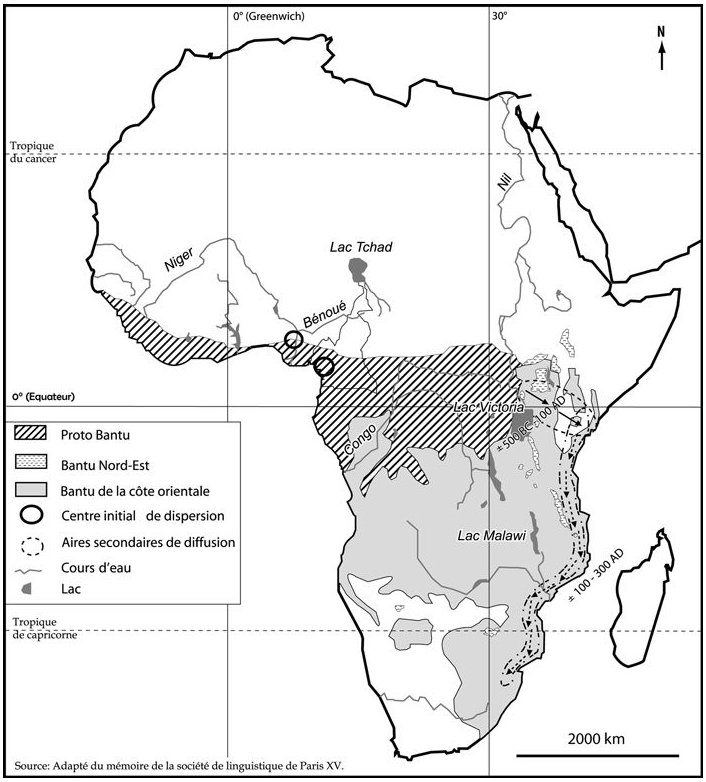


Fig. 1.2 ***Répartition géographique des locuteurs bantous en Afrique sub-Saharienne.*** *Les zones grisées indiquent la répartition des locuteurs bantous et les cercles indiquent la localisation des populations à l'origine des migrations bantoues il y a environ 5 000 ans.*

Des reconstructions linguistiques ont identifié un vocabulaire partagé entre tous les groupes d'agriculteurs qui met en évidence l'existence d'une langue ancestrale proto-bantoue et l'origine commune récente de ces langues. Leur diffusion spatiale a été permise par les ex-

pansions bantoues, des migrations ayant eu lieu au moment de l'émergence de l'agriculture au nord-ouest du Cameroun entre -5 000 et -4 000 ans (Figure 1.2). Des études archéologiques ont corroboré l'existence de ces migrations auxquelles s'est associée la diffusion de la métallurgie et de la céramique.

1.2 Unité et diversité des groupes de Pygmées

1.2.1 Diversité culturelle et de modes de vie

Les difficultés de nomenclatures des populations de Pygmées sont liées à l'absence de caractéristiques physiques, de modes de vie, ou d'habitats unissant tous les groupes. Les différences qui existent entre les communautés sont le reflet de leur grande diversité culturelle. Une approche multidisciplinaire permet de faire apparaître les éléments d'unité et de diversité de ces groupes.

Habitations

Il existe une relative diversité de formes et d'organisations des habitations construites par les Pygmées au sein du Bassin congolais. Les groupes mobiles qui subsistent dans la forêt construisent des huttes caractéristiques en forme de dômes ou de "ruches" (beehive-shape) à partir de branches pliées et tressées ensemble puis recouvertes de feuilles de Marantacées. Dans le cas des BaAka, Baka et BaMbuti, ces huttes accueillent des familles et sont placées en cercle afin de former des campements temporaires. A première vue, ces habitations sont très différentes de celles construites par les groupes moins mobiles, dont les villages semi-permanents sont faits de maisons rectangulaires plus robustes et organisées linéairement. Cependant, certains groupes aujourd'hui sédentarisés ne l'étaient pas au moment des premières descriptions datant du XIX$^{\text{ème}}$ siècle. Par exemple, les BaBongo du centre du Gabon, qui vivent aujourd'hui dans des hameaux sédentaires, occupaient des campements faits de huttes en forme de dômes à l'époque des descriptions faites par Paul Du Chaillu en 1867 (Du Chaillu & Owen., 1867). Les différences de forme et d'organisation des habitations Pygmées aujourd'hui sont donc le reflet de transformations récentes plutôt que représentatives d'une diversité culturelle.

Chasse, cueillette et autres modes de subsistance

Bien que certains groupes de Pygmées pratiquent aujourd'hui l'agriculture ou l'artisanat de façon plus ou moins permanente, il existe toujours une grande variété de techniques de chasse, de cueillette et d'autres moyens de subsistance entre les communautés. Les groupes de Pygmées dont le mode de subsistance dépend principalement de la chasse déploient un large éventail de stratégies, de techniques et d'armes en fonction des régions. C'est généralement une activité collective où les femmes sont des partenaires égales aux hommes (S. Bahuchet, 1987 ; Harako, 1976 ; Tanno, 1976 ; Terashima, 1980). La pratique de la chasse au filet est la

technique la plus employée au sein du Bassin du Congo à l'exception de deux populations : les Efe de l'est de la RDC qui pratiquent la chasse à l'arc, et les Baka du Cameroun et du Gabon qui pratiquent la chasse collective avec des lances. Le gibier le plus communément ciblé lors des battues collectives est le céphalophe bleu, un animal de taille moyenne de la famille des bovidés. D'autres animaux comme les singes et les oiseaux sont également chassés à l'aide d'arcs et d'arbalètes. Les Pygmées sont considérés par les populations de non-Pygmées comme des experts de la chasse à l'éléphant, qu'ils traquent à l'aide de longues lances munies de lames. Tous les groupes de Pygmées pratiquent la cueillette de produits de la forêt comme certains ignames, les noix, les champignons, les feuilles mais aussi les insectes tels que les termites et les chenilles. Ce sont des spécialistes de la collecte du miel qu'ils utilisent comme monnaie d'échange lors des transactions avec les non-Pygmées. Bien que la pêche ne soit pas un mode de subsistance majoritaire, elle est pratiquée par les Pygmées de façon saisonnière en complément d'autres ressources. Pour cela, des végétaux ichtyotoxiques peuvent être répandus dans les cours d'eau ou de faibles portions de rivière peuvent être asséchées afin de piéger les poissons dans des zones de faible profondeur où ils seront plus faciles à capturer.

Musique

Les Pygmées sont reconnus pour leurs talents musicaux auprès de leurs voisins non-Pygmées et il existe dans ce domaine une grande diversité de styles, de constructions musicales et d'instruments. Les groupes de BaMbuti, BaAka et Baka possèdent un style musical très caractéristique avec de nombreux chanteurs, des chorales polyphoniques, des instruments rythmiques et une forme de yodel (Arom, 1987 ; Arom & Fürniss, 1992). Les instruments mélodiques tels que l'arc musical, la harpe ou la flûte jouent un rôle secondaire. Les BaAka et les Baka, à l'ouest du Bassin du Congo, partagent l'utilisation d'un arc musical à deux cordes joué uniquement par les femmes. Ces instruments sont aussi retrouvés dans des populations de l'est de l'Afrique centrale proche de l'Ouganda et de la RDC, parmi les groupes de Pygmées Efe (Demolin, 1990) et certains agriculteurs (S. Bahuchet, 1992). Les différences et similitudes des activités culturelles des Pygmées sont probablement le signe de relations passées et d'une d'histoire commune difficile à retracer en l'absence de données génétiques (Fürniss & Bahuchet, 1995).

1.2.2 Diversité linguistique

La situation linguistique complexe des Pygmées du Bassin congolais révèle une grande hétérogénéité, à l'image de leurs modes de subsistance, leurs habitats et leurs pratiques culturelles (Table 1.2). De plus, les études linguistiques réalisées chez les Pygmées sont moins nombreuses que les études anthropologiques et la majorité des langues Pygmées ne sont pas documentées, à l'exception des langues Efe, Gyeli, Aka et Baka. Il n'existe pas de « famille » de langues Pygmées à proprement dit. Les langues parlées par les différentes communautés sont reliées aux autres langues africaines et appartiennent à deux phyla d'Afrique centrale : les langues Nigero-kordofaniennes (branche Nigero-congolaise, familles bantoues et

oubanguiennes) et les langues Nilo-Sahariennes (branche Soudanique-centrale) (Table 1.2). Les BaKola, les BaBongo, les BaTwa et les BaMbuti/BaSua parlent des langues apparentées aux langues bantoues. Les Bedzan parlent une langue bantoïde proche du Tikar parlée par les agriculteurs voisins et les Efe parlent un dialecte de la famille Soudanique-centrale très proche de la langue des agriculteurs Lefe. D'autres groupes de Pygmées parlent des langues apparentées à celles des agriculteurs qui ne sont pas mutuellement intelligibles comme c'est le cas des Baka qui parlent une langue Oubanguienne, des BaAka qui parlent une langue bantoue et des Asua qui parlent une langue Soudanique-centrale (Table 1.2). Cette proximité linguistique entre Pygmées et agriculteurs bantouphones est le témoin d'une longue histoire partagée et souligne l'importance des interactions entre ces deux groupes. Ces langues apparentées dérivent d'une langue ancestrale que les ancêtres des Pygmées ont certainement acquise au contact des ancêtres des agriculteurs actuels. Malgré ce basculement linguistique vers la grande famille des langues africaines, les modes de vie, les habitats, la technologie et la culture "Pygmées" sont restés intacts, signe que ce changement linguistique ne s'est pas accompagné d'un changement culturel : ce phénomène a été nommé le « Paradoxe Pygmée » (S. Bahuchet, 1993).

1.2.3 Relations inter-ethniques avec les non-Pygmées

Même si la nature et l'intensité des interactions socio-économiques entre Pygmées et non-Pygmées fluctuent en fonction des régions, ces deux populations entretiennent des relations très étroites dans tout le Bassin du Congo. La longue histoire des interactions entre ces populations apparaît dans leurs affinités linguistiques, autant que par la complémentarité de leurs techniques et de leurs activités. Si tous les groupes de Pygmées interagissent avec des agriculteurs voisins, le nombre d'agriculteurs qui n'interagissent pas avec les Pygmées est supérieur à ceux qui le font. Il existe certaines constantes dans les interactions qui unissent les Pygmées et les agriculteurs, la première étant la reconnaissance de l'existence de ces deux groupes distincts. Les agriculteurs attribuent aux Pygmées des caractéristiques qui sont à la fois négatives et positives. Les Pygmées sont reconnus comme d'excellents musiciens et sont invités par les groupes d'agriculteurs à participer et jouer aux festivités. Ils sont également sollicités pour leurs dons de guérisseur et jugés capables de traiter les maladies physiques et mentales. Cependant, il existe des principes stricts régissant ces relations inter-ethniques, et dans la majorité des cas, les mariages entre Pygmées et non-Pygmées sont prohibés. Peu de groupes dans le Bassin congolais ont levé cette interdiction et seuls les mariages entre des hommes agriculteurs et des femmes Pygmées sont tolérés. Les unions d'hommes Pygmées et de femmes non-Pygmées ne sont possibles que dans certaines populations comme les Ba-Bongo du Gabon. Ces normes établies entre les deux communautés participent à la mise en place d'une hiérarchie claire où les groupes d'agriculteurs considèrent parfois les Pygmées comme leur propriété.

Au niveau économique, les dynamiques entre Pygmées et non-Pygmées sont interdépendantes. Les Pygmées échangent des produits de la forêt tels que le miel, la viande ou des traitements médicinaux et participent aux travaux d'agriculture sur brûlis. En échange, les

agriculteurs leur fournissent des outils en fer, des poteries, ainsi que le fruit des récoltes. Ces interactions permettent aux deux partis de profiter de l'écosystème exploité par la population voisine, à savoir la forêt pour les agriculteurs et les zones cultivées pour les Pygmées. Les deux groupes sont interdépendants et bénéficient du travail de chacune des communautés. C'est particulièrement le cas des Pygmées au style de vie semi-nomade tels que les BaMbuti, les Efe, les BaMbuti, les BaAka et les Baka, même si la nature des échanges entre les groupes de Pygmées moins mobiles, comme les BaKola et les BaBongo, et les agriculteurs voisins est similaire. Les groupes de Pygmées qui dépendent le moins de la chasse et de la cueillette échangent des services tels que des performances musicales ou de l'artisanat, comme c'est le cas des BaTwa du Rwanda.

La dimension sociale des interactions entre les Pygmées et les non-Pygmées est relativement complexe (S. Bahuchet & Guillaume, 1982 ; Delobeau, 1989 ; Grinker, 1994 ; Joiris, 2003 ; Turnbull, 1965) et implique une indépendance interne, une dépendance externe et des interactions régies par de nombreuses règles. Tout d'abord les deux communautés sont socialement indépendantes concernant la filiation, l'organisation sociale et la religion. Cependant, les deux communautés sont construites autour d'une forte interdépendance. Les populations d'agriculteurs font partie intégrante du système social des Pygmées et inversement. Chez les Pygmées BaAka et Baka, le mariage, qui est au centre du système socio-économique, est indissociable des outils en fer fournis par les agriculteurs et qui constituent le «prix de la fiancée », c'est-à-dire le don fait à la famille de l'épouse à l'occasion de son mariage. Pour les agriculteurs, la grande quantité de viande fournie par les chasseurs Pygmées leur permet d'approvisionner les rassemblements cérémoniels, comme par exemple les festivités de fin de deuil et les rites d'initiation des jeunes garçons. Il arrive également que les membres d'une communauté assistent aux événements cérémoniels de l'autre, favorisant alors les liens entre les groupes autour de rites de «fraternité » ou de «parenté fictive » (Joiris, 2003 ; Robillard, 2012).

Les constructions sociales dans lesquelles les agriculteurs se placent comme supérieurs aux Pygmées marquent une dépendance externe, bien que le degré de cette hiérarchie varie de relations commerciales libres à l'intégration des Pygmées dans un système de castes (Hewlett, 2014). Bien souvent, les groupes de Pygmées sont intégrés dans des relations poly-ethniques complexes et interagissent avec de multiples groupes d'agriculteurs. C'est le cas des BaAka qui entretiennent des relations avec une vingtaine d'ethnies d'agriculteurs au nord du Congo et au sud de la République Centrafricaine (RCA). On trouve d'autres exemples de systèmes poly-ethniques au Cameroun avec les Baka ou au Gabon avec les BaBongo.

1.3 Occupation préhistorique de la forêt

1.3.1 Mythes et traditions orales

La perception que les agriculteurs non-Pygmées ont des Pygmées est souvent mitigée et ambiguë puisqu'ils sont à la fois méprisés, admirés et parfois craints (S. Bahuchet & Guillaume, 1982). Les Pygmées sont présents dans la mythologie et les traditions orales de

quasiment toutes les populations d'agriculteurs d'Afrique centrale (Klieman, 2003). De nombreux documents et récits attestent de l'importance symbolique des Pygmées et de leur statut quasi-surnaturel, à la frontière du monde des humains (S. Bahuchet, 1993 ; S. Bahuchet & Guillaume, 1982 ; Grinker, 1994 ; Joiris, 2003). La plupart des traditions orales évoquent la rencontre des bantous avec les Pygmées lors des premières migrations et les reconnaissent comme les habitants historiques des régions forestières qu'ils ont colonisées (S. Bahuchet & Guillaume, 1982). Dans ces récits, les Pygmées étaient des guides qui ont initié les agriculteurs au monde de la forêt en leur enseignant les rites et les techniques à observer. Cependant, certaines techniques rapportées par les mythes, comme le travail du fer, ne sont traditionnellement pas spécifiques des chasseurs-cueilleurs mais placent les Pygmées comme éléments «civilisateurs» et porteurs de connaissance.

1.3.2 Pléistocène (-2,58 millions d'années à -11 700 ans)

L'histoire de l'occupation des forêts d'Afrique centrale semble précéder l'apparition de l'agriculture et pose de nombreuses questions sur l'histoire des populations de Pygmées et de non-Pygmées du Bassin du Congo. Même si la fin du Pléistocène est considérée comme la période de diversification de l'Homme moderne en Afrique, (Vansina, 1984) il n'existe que peu de restes archéologiques attestant de l'occupation des forêts équatoriales au Pléistocène Moyen (-781 000 à -126 000 ans) et au Pléistocène Supérieur (-126 000 à -11 700 ans) (Mercader, 2002). Des données archéologiques provenant de complexes industriels datant du Lupembien (Figure 1.3 ; (Taylor, 2011)), préhistoire de l'Afrique centrale, permettent de retracer la présence des premiers chasseurs-cueilleurs du Bassin du Congo (Barham, 2001 ; Taylor, 2011). La datation de ces restes lithiques, qui repose sur l'observation des séquences stratigraphiques ou des méthodes de datation radiométrique, reste complexe et varie d'environ 42 000 à 300 000 ans. (Barham, 2001 ; Taylor, 2011)

la difficulté il y a environ 300 000 ans (Barham, 2001 ; Taylor, 2011).

Cette période se caractérise par la manufacture d'outils fixés à des manches ou des poignées pour en faciliter l'usage tels que des haches. Les dates d'apparition de ces technologies coïncident avec une période interglaciaire plus chaude qui aurait permis aux chasseurs-cueilleurs d'exploiter les ressources de la forêt équatoriale. D'autres preuves anthropologiques et archéologiques suggèrent la présence d'humains dont le mode de subsistance était la chasse et la cueillette au sein du bassin congolais dès la fin du Pléistocène, avant l'arrivée des premiers fermiers bantous au cours de l'Holocène (Barham, 2001). Cependant, la nature acide des sols de la forêt équatoriale entraîne une dégradation rapide des restes archéologiques, rendant difficile la reconstruction de l'histoire de son occupation par l'Homme.

1.3.3 Holocène (-11 700 ans à nos jours)

De nombreux travaux archéologiques, linguistiques et paléoclimatiques ont permis de mieux comprendre les interactions entre Pygmées et non-Pygmées au cours du peuplement préhistorique des forêts africaines. Tout d'abord, l'Holocène (-11 700 ans à nos jours) a été caractérisé en Afrique par un épisode de grande sécheresse (Figure 1.4 ; (Compton, 2011)).

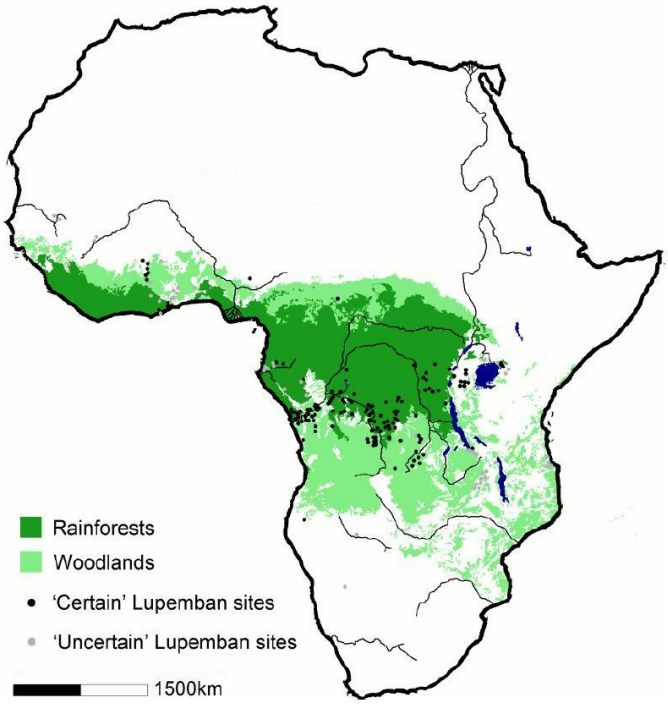


Fig. 1.3 ***Sites archéologiques de la période du Lupembien (adaptée de Taylor, 2011).*** *Localisation des complexes industriels datant du Lupembien dans les zones de forêt du Bassin du Congo.*

Des travaux de palynologie (étude du pollen fossile), de diatomées et d'éléments géochimiques en Afrique de l'ouest et en Afrique centrale, ont mis en évidence une période aride s'étalant entre -4 500 et -2 000 ans. Cet intervalle climatique a fait régresser les limites de la forêt équatoriale, les lacs et les zones humides entraînant une modification des espèces végétales endémiques et une intensification de dépôts de poussières (Figure 1.4) (Delegue et al., 2001 ; deMenocal et al., 2000). Cette transition paléo-environnementale a transformé la forêt, dense et continue, en blocs fragmentés laissant apparaître des couloirs de savane qui auraient facilité les migrations humaines entre -2 800 et -2 000 ans (Figure 1.4) (Schwartz & Lanfranchi, 1990).

C'est au cours de cette période que l'agriculture s'est diffusée en Afrique avec les expansions bantoues. Des reconstructions linguistiques ont identifié des migrations de langues proto-bantoues depuis une terre commune à l'est du Nigeria et à l'ouest du Cameroun vers l'est et le sud de l'Afrique, il y a 3 000 à 3 500 ans (Figure 1.2). Ces observations sont corroborées par la diffusion de céramiques vers la côte ouest, le centre ouest et le centre de l'Afrique entre -2 500 et -2 000 (J. Diamond & Bellwood, 2003) ainsi que par la découverte d'espèces végétales domestiquées et adaptées au biome sec de la savane, comme des graines de mil à chandelles datant de 2 200 ans et des pois de bambara datant de 1 700 ans à l'ouest du Cameroun. La présence massive de pollen de palmiers à huile datant de l'Holocène Supérieur est également un bon indicateur de l'apparition de l'arboriculture et de la paraculture

de palmiers indigènes autour de 5 000 ans qui s'intensifie autour de 3 000 ans (Sowunmi, 1999).

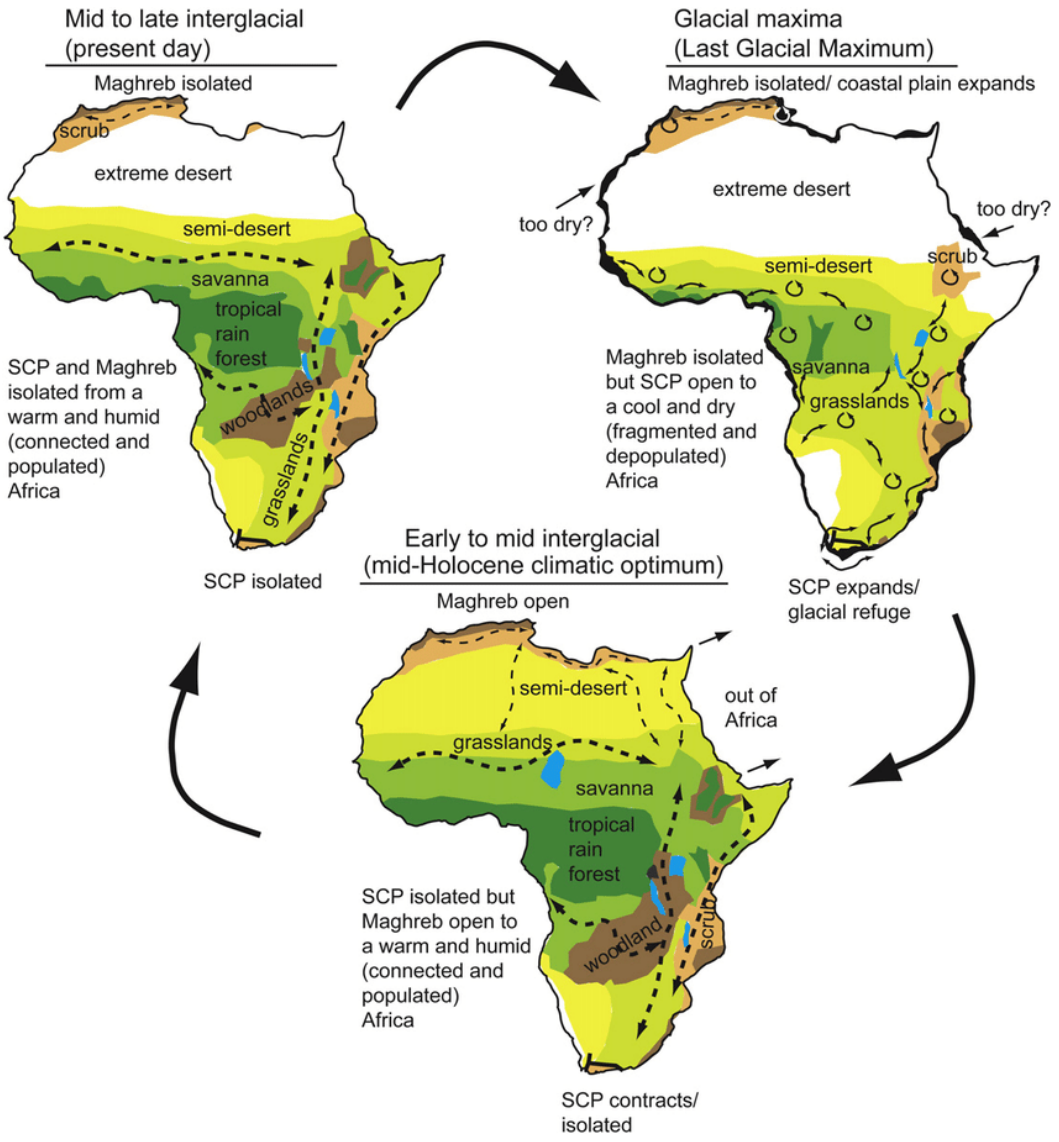


Fig. 1.4 ***Biomes africains au cours des cycles glaciaires et interglaciaires (Compton, 2011).*** *Répartition de zones de forêts, de savanes, de désert et de la plaine côtière du sud (Southern Coast Plain, SCP) en fonction des époques climatiques (périodes glaciaires et interglaciaires) en Afrique. Les flèches indiquent les routes de migrations possibles et les cercles les refuges glaciaires.*

Certaines hypothèses suggèrent qu'avant l'émergence de l'agriculture et l'arrivée des populations de pré-agriculteurs de langues bantoues il y a 2 000 à 3 500 ans, les chasseurs-cueilleurs préhistoriques n'auraient occupé les zones forestières que par intermittence. Selon ces hypothèses, l'apparition de nouvelles technologies (céramiques, maîtrise du fer et d'outils en pierre) et de ressources stables (animaux et plantes domestiqués) (Philipson, 2005) aurait permis aux ancêtres des chasseurs-cueilleurs d'établir des échanges commerciaux avec les proto-agriculteurs, et de compléter les ressources disponibles en forêt. L'agriculture sur

brûlis et la demande en métallurgie, de concert avec des changements climatiques auraient donc été à l'origine des premières interactions entre les proto-agriculteurs et les ancêtres des chasseurs-cueilleurs du Bassin du Congo (Bailey & Zechenter., 1989 ; Vansina, 1990). Par la suite, un retour à des conditions plus humides il y a 2 000 ans aurait permis la résurgence des forêts, isolant les populations de Pygmées et pérennisant alors les relations avec les non-Pygmées.

Cependant, de nombreux travaux ont remis en question cette hypothèse en supposant que les ancêtres des chasseurs-cueilleurs se procuraient déjà les ressources nécessaires pour vivre en autonomie dans la forêt avant l'arrivée des agriculteurs (D. M. Bahuchet S. & de Garine, 1991 ; S. Bahuchet, 1993 ; Hladik & Dounias, 1993). Des études de botanique et d'ethnobotanique ont mis en évidence des zones très riches en ignames comestibles dans les forêts d'Afrique centrale (D. M. Bahuchet S. & de Garine, 1991 ; Hladik & Dounias, 1993), dont la récolte pendant la période humide peut excéder la quantité de féculent produite via l'agriculture (Dounias, 2001). Ces ignames sauvages constituent la première source saisonnière d'amidon pour de nombreux chasseurs-cueilleurs de la forêt (Kitanishi, 1995). De plus, des études ethnographiques ont mis en évidence des pratiques de paraculture d'ignames sauvages chez certains groupes de chasseurs-cueilleurs qui replantent les tiges après cueillette (Dounias, 2001). Enfin, les recherches archéologiques soutiennent également l'hypothèse d'une occupation des forêts d'Afrique centrale précédant l'émergence de l'agriculture (Barker et al., 2007 ; Barton & Arroyo-Kalin., 2012 ; Mercader, 2002).

Name	Country	Other Names	Population Size and Environment	Main References
Baka	Cameroon, Congo, Gabon	Bangombe, Bibayak, Babinga	± 30,000–40,000, rainforest	Joiris 1998; Leclerc 2001, 2012: Sato 1992; Tsuru 1998; Vallois and Marquer 1976
Bedzan	Cameroon	Tikar Pygmies	± 400, rainforest, savannah margins	Leclerc 1999; Mebenga Tamba 1998
(Ba)Kola	Cameroon	(Ba)Gyeli	± 4,000, rainforest	Joiris 1994; Koppert et al. 1997; Loung 1959, 1987, 1996; Ngima Mawoung 1996
(Ba)Koya	Gabon, Congo	(Ba)Kola	± 2,600, rainforest	Soengas 2009, 2010, 2012; Tilquin 1997
(Ba)Bongo	Gabon	Akoa, Barimba	± 3,000, rainforest, savannah margins	Andersson 1983; Knight 2003; Matsuura 2006; Le Bomin and Mbot 2012a
Mikaya	Congo	Bambenga	No data, rainforest	Boyi s.d.
(Ba)Aka	CAR, Congo	Bayaka, Biaka, Babinga, Bambenga, BaMbenzele	± 30,000–50,000, rainforest	Arom 1987; Bahuchet 1985; Demesse 1980; Hewlett 1991; Kitanishi 1995; Lewis 2002
(Ba)Twa	DRC	Konda Twa	No data, rainforest, swamps	Elshout 1963; Pagezy 1986, 1988; Schultz 1986; Sulzmann 1986; Van Everbroeck 1974
(Ba)Cwa	DRC	Bushong Twa, Kuba Cwa	± 6,000, savannah	Hiernaux 1966; Vansina 1954
(Ba)Cwa/(Ba) Tembo	DRC	Luba Cwa, Batwa, Bambote	No data, savannah	Kazadi 1981
(Ba)Twa	Uganda	Batwa	± 2000?, mountain forest	Kenrick 2000, UNEP 2005
(Ba)Twa-(Ba) Rhwa	DRC	Batwa, Kivu Twa, western Twa, Barhwa	± 6.000, mountain forest	Schumacher 1949, 1950; Seitz 1993
(Ba)Twa	Rwanda, Burundi	Batwa, eastern Twa	± 70,000?, savannah, mountains	Lewis 2000; Lewis and Knight 1996
(Ba)Sua	DRC	(Ba)Mbuti, Kango	± 26,000, rainforest	Harako 1981; Hart and Hart 1986; Ichikawa 1978; Tanno 1976; Turnbull 1965a
Asua	DRC	(Ba)Mbuti, Akka, Tikki-tikki	± 10,000, rainforest	Demolin 1992; Schebesta 1952
Efe	DRC	(Ba)Mbuti	± 10,000, rainforest	Bailey 1991; Bailey and Peacock 1988; Demolin 1993; Terashima 1983, 1985

Note: CAR = Central African Republic, DRC = Democratic Republic of Congo

Table 1.1 – Inventaire des populations de Pygmées du Bassin du Congo (Hewlett, 2014)

Pygmy Group	Language Family	Sub-branch	Closest Farmer Language	Status of the Pygmy Language
Niger-Kordofanian, Niger-Congo				
Baka	Ubangi	Gbanzili-Sere	Ngbaka	Language
Bedzan	Bantoid	Non-Bantu	Tikar	Dialect
Kola	Bantoid	Northwest Bantu A80	Mvumbo	Dialect
Koya	Bantoid	Northwest Bantu B20	Ngom	Dialect
Bongo	Bantoid	Northwest Bantu B30	Tsogho	Dialect
Bongo	Bantoid	Northwest Bantu B60	Kaningi	Dialect
Bongo	Bantoid	Northwest Bantu B70	Tege	Dialect
Mikaya	Bantoid	Northwest Bantu C10	Ngundi?	Language
Aka	Bantoid	Northwest Bantu C10	Ngando	Language
Twa	Bantoid	Northwest Bantu C60	Konda	Dialect
Cwa	Bantoid	Central Bantu L30	Luba	Dialect?
Cwa	Bantoid	Central Bantu L30	Hemba	Dialect
Twa	Bantoid	Central Bantu JE10	Kiga	Dialect
Twa	Bantoid	Central Bantu D50	Shi?	Dialect
Twa	Bantoid	Central Bantu JD60	Rundi	Dialect
Sua	Bantoid	Central Bantu D30	Bila	Dialect
Nilo-Saharan, Central Sudanic				
Asua	Moru-Mangbetu	Mangbetu-Asua	Mangbetu	Language
Efe	Moru-Mangbetu	Mangbutu-Efe	Lese	Dialect

Table 1.2 – Classification généalogique des langues Pygmées (Hewlett, 2014)

Chapitre 2

Apport des données génomiques à l'histoire démographique des Pygmées

L'anthropologie, l'ethnographie, la linguistique et la paléoclimatologie ont permis de caractériser de nombreux aspects de la vie des populations de chasseurs-cueilleurs forestiers et d'agriculteurs non-Pygmées en Afrique centrale. Cependant, la quasi-absence de restes archéologiques dans les forêts tropicales limite considérablement la reconstruction de l'histoire de ces populations et des dynamiques de l'occupation préhistorique des forêts équatoriales. De nombreuses questions restent donc en suspens comme la date d'arrivée des ancêtres des Pygmées dans les forêts tropicales du Bassin du Congo, les fluctuations démographiques de ces populations, ainsi que la nature des relations qu'ils entretenaient avec les ancêtres des groupes non-Pygmées.

La génétique des populations, parce qu'elle s'intéresse à l'histoire des marqueurs génétiques portés par les individus, permet de retracer les dynamiques démographiques et migratoires des populations humaines. Dans le contexte multi-ethnique de l'Afrique centrale, ces approches sont primordiales pour reconstruire directement ou indirectement l'histoire des populations de Pygmées et de non-Pygmées.

2.1 Diversité génétique et structure des populations

2.1.1 Forces génomiques à l'origine de la diversité génétique

Il existe deux forces à l'origine de la diversité génétique : les forces génomiques de mutation à l'origine de nouveaux polymorphismes et les forces génomiques de recombinaison méiotique à l'origine de nouvelles combinaisons d'allèles, ou haplotypes.

Polymorphismes génétiques

L'ADN ou acide désoxyribonucléique est la macromolécule stable du support de l'information génétique. Il est transmis verticalement d'une génération à la suivante et subit des modifications aléatoires de sa séquence nucléotidique. Lorsque ces mutations apparaissent dans l'ADN des cellules de la lignée germinale (à l'origine des gamètes), elles sont transmises à la descendance et participent à la diversité génétique de la population.

Parmi ces modifications, les substitutions simples d'un nucléotide par un autre (SNP, *Single Nucleotide Polymorphism*) sont les plus nombreuses et les plus étudiées en génétique des populations. Ces mutations proviennent majoritairement de deux types de mécanismes : une erreur de la machinerie de réplication de l'ADN entraînant l'insertion d'un nucléotide erroné ou une erreur du système de réparation de l'ADN suite à une altération physique ou chimique de l'intégrité de la molécule. On constate que les transitions, c'est à dire le remplacement d'une purine (adénosine (A) <-> guanine(G)) ou d'une pyrimidine (cytosine (C) <-> thymine (T)) par une autre sont bien plus fréquentes que les transversions. Le taux de mutation de l'espèce humaine est en moyenne de 10^{-8} mutations par site et par génération mais varie selon les régions du génome (C. D. Campbell et al., 2012). Chez l'Homme, la base de données dbSNP (build 151, juin 2018, (Sherry et al., 2001)) recense environ 114 millions de SNPs validés répartis sur le génome.

D'autres polymorphismes appelés "variations structurelles" désignent les réarrangements génomiques de plus de 50 pb (paires de bases) et représentent environ 1.2% des variations entre individus (contre 0,1% pour les SNP) (Tattini et al., 2015). Ces polymorphismes de séquence d'une taille moyenne de 8 kb (kilobases) (Auton et al., 2015) regroupent les inversions de séquences, les insertions, les délétions d'un ou plusieurs nucléotides, la transposition d'éléments mobiles ainsi que l'amplification d'un motif répété de nucléotides (Tattini et al., 2015). Elles proviennent d'erreurs du mécanisme de réplication, de la réparation des cassures double-brin de l'ADN et de la recombinaison méiotique (Conrad et al., 2010) et présentent un taux de mutation bien supérieur à celui des substitutions, allant de 10^{-5} à 10^{-3} mutations par site et par génération (Lupski, 2007). Ces variations peuvent avoir des conséquences majeures sur le phénotype, en particulier si elles interrompent un gène ou affectent le dosage génique (Conrad et al., 2010 ; Cooper et al., 2007 ; Hurles et al., 2008 ; Stankiewicz & Lupski, 2010 ; Tattini et al., 2015). Afin d'identifier les implications cliniques de ces polymorphismes, plusieurs bases de données répertorient les variations structurelles telles que la Database of Genomic Variants archive (DGVa) (Lappalainen et al., 2013), dbVAR (https ://www.ncbi.nlm.nih.gov/dbvar/). Cependant, même si les technologies de séquencage de l'ADN à haut débit ont facilité l'identification de ces variations, leur détection reste limitée par la longueur des fragments séquencés.

Recombinaison méiotique

En plus de l'apparition aléatoire de mutations dans la séquence d'ADN, la recombinaison méiotique constitue une seconde source de variabilité génétique. Le phénomène de recombinaison méiotique correspond à l'échange de matériel génétique entre deux chromosomes homologues au cours de la gamétogénèse. En intervertissant les variations génétiques portées par deux chromosomes homologues, la recombinaison diminue la probabilité qu'une mutation soit transmise à la descendance en même temps que les autres mutations déjà présentes dans la même région génomique, ce qui a pour conséquence de diminuer le déséquilibre de liaison (DL). Ces combinaisons d'allèles sont appelées des haplotypes et de nouveaux haplotypes,

dits haplotypes recombinants, apparaissent à chaque génération suite aux événements de recombinaison.

Le taux de recombinaison, comme le taux de substitution, est variable le long du génome humain avec en moyenne un taux de recombinaison plus élevé à un distance d'environ 10^3 à 10^4 pb du codon d'initiation de la transcription (Myers et al., 2005). La reconstruction de cartes génétiques a permis d'identifier des zones de "points froids" ou "déserts" de recombinaison qui couvrent environ un tiers du génome humain, où le taux de recombinaison est inférieur à la moyenne génomique (Hussin et al., 2015). A l'inverse, il existe des régions de "points chauds", couvrant un total de 634 Mb (mégabases), où le taux de recombinaison est supérieur à la moyenne du génome de plusieurs ordres de grandeurs (Hussin et al., 2015 ; McVean et al., 2004 ; Myers et al., 2005 ; International HapMap et al., 2007). Environ 25 000 points chauds de recombinaison ont été identifiés dans le génome humain (soit environ un tous les 50kb) mais leur distribution et leur intensité varient en fonction du sexe des individus (Bherer et al., 2017) ainsi qu'en fonction des polymorphismes génétiques qu'ils portent (Pratto et al., 2014).

2.1.2 Dérive génétique, isolement et flux géniques

Lorsqu'une mutation apparaît dans la séquence d'ADN d'un individu, celle-ci a trois devenirs différents dans la population : elle peut disparaître, rester à une fréquence intermédiaire ou se fixer dans la population. En considérant que la mutation est neutre, c'est-à-dire qu'elle ne confère ni avantage, ni désavantage sélectif à l'individu qui la porte (respectivement augmentation ou diminution du succès reproductif), les variations de sa fréquence allélique dans la population sont stochastiques et ce phénomène est qualifié de dérive génétique (Nei, 1987 ; Wright, 1931).

Sous neutralité et en absence de migration (isolement génétique), les fréquences alléliques observées à une génération t proviennent de l'échantillonnage aléatoire des allèles de la génération *t-1*. La probabilité de transmission d'un allèle d'une génération *t-1* à t dépend donc du nombre d'allèles portés par l'ensemble des individus de la génération *t-1*. Ainsi, le nombre d'individus qui participent à la diversité allélique d'une population est appelé la taille efficace (ou taille effective) de la population, notée N_e.

Une mutation neutre qui apparaît dans le génome humain (diploïde) a une probabilité de fixation égale à $\frac{1}{2N_e}$ et mettra environ $4N_e$ générations à se fixer. Dans ce modèle, les allèles restent à des fréquences stables pendant de nombreuses générations dans les populations de grandes tailles efficaces alors que les fluctuations des fréquences alléliques sont plus importantes dans les populations avec un N_e faible. La taille efficace N_e d'une population détermine donc son niveau de diversité génétique et sous neutralité, ces deux variables sont proportionnelles. La diversité génétique (θ) s'exprime alors comme le produit du taux de mutation (μ) et de la taille effective de la population (N_e) soit :

$$\theta = 4N_e\mu$$

L'isolement de deux populations, qui désigne l'absence d'individus dont les parents sont

issus des deux populations, peut naître de l'isolement géographique (isolement par distance) ou reproductif. L'absence de migration entre les groupes augmente leurs niveaux de différenciation génétique, c'est-à-dire les différences de fréquences alléliques observées dans les populations. Cependant, les populations humaines sont rarement isolées génétiquement et les échanges génétiques sont d'autant plus probables que les populations sont proches géographiquement. La plupart des groupes humains répondent donc à un modèle d'isolement avec migration dans lequel deux populations se séparent à un temps donné puis échangent des migrants à un taux variable. Ainsi, une faible différenciation génétique entre deux populations peut être due à une séparation récente des groupes ou à une séparation ancienne suivie d'un fort taux de migration.

Les niveaux de différenciation génétique entre les populations humaines sont très informatifs sur le degré d'isolement et les dynamiques migratoires. Les données génétiques permettent alors de mieux comprendre l'histoire des populations.

2.1.3 Structure des populations de chasseurs-cueilleurs et flux migratoires

A la fin des années 60, Cavalli-Sforza et ses collaborateurs ont mené les premiers travaux caractérisant la diversité génétique des populations de Pygmées en Afrique centrale. Grâce à l'étude de polymorphismes protéiques présents dans des populations de Pygmées de l'ouest (RCA et Cameroun) et de l'est (RDC), les auteurs ont montré que les groupes de Pygmées étaient génétiquement très différenciés des groupes de non-Pygmées et qu'ils présentaient un profil de diversité génétique indiscutablement africain. Les auteurs ont également identifié des différences génétiques majeures entre populations de Pygmées à l'ouest et à l'est de l'Afrique centrale, ainsi qu'au sein des populations de Pygmées de l'ouest (L. L. Cavalli-Sforza et al., 1969 ; Santachiara-Benerecetti et al., 1980). Ces travaux basés sur un nombre limité de polymorphismes ont amené Cavalli-Sforza et son équipe à suggérer une origine commune de toutes les populations de Pygmées suivie d'une séparation plus récente des groupes de l'ouest et de l'est sans pouvoir tester formellement cette hypothèse.

D'autres travaux sur la différenciation et la structure génétique des populations de Pygmées et de non-Pygmées ont permis d'étendre ces résultats grâce à des marqueurs obtenus par séquençage de régions neutres du génome ou par génotypage (Verdu et al., 2013 ; Patin et al., 2014, 2009 ; Veeramah et al., 2012). Ces études ont montré que les niveaux de différenciation génétique entre les groupes de Pygmées de l'ouest et de l'est sont plus importants que les niveaux de différenciation génétique observés au sein des groupe Pygmées de l'ouest (Patin et al., 2009, 2014 ; Tishkoff et al., 2009 ; Verdu et al., 2009). De plus, malgré la proximité géographique de certains groupes de chasseurs-cueilleurs Pygmées et d'agriculteurs non-Pygmées au Cameroun et au Gabon, ces populations sont génétiquement très différenciées (Verdu et al., 2009 ; Patin et al., 2014, 2009). Il existe donc une structure génétique des populations d'Afrique centrale en fonction de leurs modes de subsistance.

Les niveaux de différenciation génétique peuvent être mesurés à l'aide du F_{ST} qui compare

les fréquences alléliques de deux populations (Wright, 1943, 1965). En considérant les données de génotypage de 9 populations de Pygmées et 9 populations de non-Pygmées à l'ouest et a l'est de l'Afrique, l'étude de Patin et al., en 2014 a quantifié leurs niveaux de différenciation génétique. Les distances génétiques séparant les populations de Pygmées suivent un modèle d'isolement par la distance, à l'exception des populations de Pygmées de l'est (BaTwa et Mbuti). En effet, ces deux populations présentent un F_{ST} égal à 0.036 malgré leur relative proximité géographique (environ 400km). Ce niveau de différentiation est comparable à celui des Pygmées de l'ouest et de l'est (F_{ST}=0.038), distants d'environ 1 500km, et suggère un fort isolement entre populations de Pygmées à l'est de l'Afrique centrale. Ces résultats contrastent avec les niveaux de différenciation observés au sein des groupes de Pygmées de l'ouest dont le F_{ST} moyen est égal à 0.014. Cependant, trois sous-structures de populations apparaissent à l'ouest : les Baka du Cameroun et du Gabon, les Biaka de RCA et un groupe constitué des Bedzan du Cameroun et des BaBongo du Gabon (Patin et al., 2014).

La différenciation génétique observée au sein des populations d'agriculteurs non-Pygmées est très faible avec un $F_{\mathrm{ST}} < 0.01$ malgré des distances géographiques considérables entre les populations (Patin et al 2017). De plus, bien que de nombreux groupes de Pygmées soient génétiquement très différenciés des groupes de non-Pygmées, deux populations du Gabon, les BaBongo du sud et de l'est, sont génétiquement plus semblables aux populations voisines de non-Pygmées (Verdu et al., 2009).

Bien qu'une analyse de la structure génétique des populations permette de caractériser de nombreux aspects de leur passé et de leurs interactions, elle ne permet pas d'estimer précisément certains paramètres clés de leurs passés démographiques tels que les temps de divergence entre populations, les variations de (N_e) et les taux de migrations.

2.2 Reconstruction de la démographie

2.2.1 Impact de la démographie sur la distribution des fréquences alléliques

Les fluctuations des tailles de populations au cours du temps et les événements de migration ont un impact direct sur le niveau de diversité génétique. Ces variations de N_e définissent l'histoire démographique de la population.

L'histoire démographique des populations laisse des signatures caractéristiques dans leurs distributions des fréquences alléliques ou SFS (Site Frequency Spectrum). Ainsi, une réduction de taille efficace provoquée par un effondrement du nombre d'individus dans la population ou un effet fondateur (population issue d'un petit nombre d'individus), augmente les effets de la dérive génétique, favorise la disparition ou la fixation d'allèles rares et provoque une diminution de la diversité génétique globale. Les SFS des populations ayant subi de telles réductions de N_e présentent un excès de mutations fixées et un déficit de mutations à faibles fréquences en comparaison à une population dont la taille serait restée constante.

A l'inverse, une augmentation de la taille efficace de la population à la suite d'une ex-

pansion démographique diminue les effets de la dérive en augmentant le N_e, et le nombre accru d'individus entraîne l'apparition d'un plus grand nombre de nouvelles mutations. Les SFS de ces populations se caractérisent par un nombre important de mutations à faibles ou très faibles fréquences alléliques et un nombre limité de mutations fixées.

Dans le cas particulier d'un goulot d'étranglement (réduction de la taille de la population suivie d'une expansion rapide), les SFS des populations présentent à la fois des signes de la réduction passée de la taille des populations avec de nombreux allèles fixés ou à fréquences élevées, ainsi que les marques de l'expansion caractérisée par un excès de mutations à fréquences faibles.

2.2.2 Méthodes de reconstruction de la démographie

Les modifications des tailles de populations au cours de l'histoire influencent la diversité génétique et la distribution des fréquences alléliques observables au temps présent. L'utilisation des données génomiques permettent souvent de compléter les données archéologiques en faisant le chemin inverse et en estimant leurs histoires démographiques, ainsi que les dates de divergence des populations et les dynamiques de migration.

Il existe plusieurs méthodes permettant d'analyser la diversité génétique des populations dans le but de retracer leurs histoires démographiques. Tout d'abord, les données de polymorphismes génétiques permettent de reconstruire des arbres de coalescence en suivant, en sens inverse, l'évolution des allèles de tous les individus d'une population jusqu'à leurs copies ancestrales. Ces arbres de coalescence obtenus à partir de données empiriques sont ensuite comparés à des arbres simulés pour une population de taille constante (Pybus et al., 2000). Les méthodes de coalescence sont généralement très fiables mais le temps nécessaire à ces reconstructions augmente rapidement avec le nombre de sites considérés et peuvent être biaisées par la recombinaison (Lapierre et al., 2016).

Les informations contenues dans les distributions des fréquences alléliques peuvent également être utilisées pour retracer l'histoire démographique des populations. Bien que le SFS ne soit pas altéré par les forces de recombinaison (Wall, 1999) il est néanmoins sensible aux forces de sélection naturelle qui provoquent des excès ou des déficits locaux de variants à forte fréquence (Fay & Wu, 2000). Cependant, la démographie affecte le génome dans sa globalité alors que la sélection naturelle agit principalement sur les parties codantes et régulatrices du génome. Par conséquent, les changements de fréquences alléliques induits par la sélection naturelle sont évités si on ne considère que les mutations les plus "neutres" du génome pour inférer l'histoire démographique des populations. Il existe plusieurs méthodes d'inférence démographique à partir du SFS. Par exemple, il est possible de déterminer les scénarios démographiques qui expliquent le mieux le SFS observé grâce à des méthodes de maximum de vraisemblance. La validité du modèle est estimée par comparaison du SFS observé dans les données empiriques avec le SFS estimé par simulation de coalescence (Nielsen, 2000 ; Coventry et al., 2010 ; Nielsen et al., 2012). Cette méthode peut être étendue à l'es-

timation des histoires démographiques conjointes de plusieurs populations à partir de leurs SFS joints (Excoffier et al., 2013).

Une autre famille de méthodes est couramment utilisée pour l'inférence de l'histoire démographique à partir de données génétiques. Les méthodes basées sur l'Approximate Bayesian Computation (ABC) utilisent le théorème de Bayes pour inférer les paramètres démographiques à partir de statistiques résumant les données (Tavare et al., 1997 ; Pritchard et al., 1999 ; Beaumont et al., 2002). Tout d'abord les méthodes d'ABC simulent des données génétiques sous plusieurs modèles démographiques plus ou moins complexes. Ces données sont ensuite résumées en calculant un ensemble de statistiques pour chaque simulation. Les statistiques simulées sont comparées aux mêmes statistiques calculées sur les données empiriques. Ces comparaisons permettent d'établir une distribution 'a posteriori' des paramètres du modèle en utilisant les valeurs des paramètres utilisées pour produire les simulations qui se rapprochent le plus des empiriques données. La précision des estimations des paramètres démographiques dépend alors de la pertinence des statistiques résumées utilisées.

2.2.3 Histoire démographique des populations humaines

Les études d'inférence démographique des populations humaines ont permis de reconstruire l'histoire d'*Homo sapiens* depuis son apparition il y a environ 200 000 ans en Afrique (Figure 2.1). Ces travaux ont retracé les événements de sortie d'Afrique, il y a environ 40 000 à 80 000 ans, ainsi que la dispersion des populations humaines en Asie du sud, en Australie, en Europe et en Asie de l'est (Malaspinas et al., 2016 ; Nielsen et al., 2017 ; Novembre & Ramachandran, 2011 ; Veeramah & Hammer, 2014). Les premiers groupes d'*Homo sapiens* ont ensuite atteint le continent américain il y a 15 000 à 35 00 ans et les différentes régions d'Océanie il y a 1 000 à 4 000 ans. Ces mouvements de populations se sont accompagnés d'une succession de goulots d'étranglement, aussi appelés effets fondateurs.
La diversité génétique des populations hors d'Afrique est donc un échantillonnage de la diversité observée au sein de l'Afrique et diminue progressivement avec la distance au continent africain (Genomes Project et al., 2010, 2012 ; M. C. Campbell et al., 2014 ; Jakobsson et al., 2008 ; Henn et al., 2012). De récentes études basées sur l'analyse de données de séquençage de génomes complet ont montré que les populations non-africaines modernes dériveraient d'une unique sortie d'Afrique il y a moins de 60 000 ans (Malaspinas et al., 2016 ; Fumagalli et al., 2011). Une autre étude suggère que les populations Océaniennes présentent des signatures d'une sortie d'Afrique très ancienne, remontant à 120 000 ans (Pagani et al., 2016), mais ayant peu contribué aux populations eurasiennes actuelles.

D'autres études ont reconstitué les événements de migration et de métissage ayant eu lieu au cours du peuplement du monde par l'Homme. Ces résultats indiquent que la structure des populations humaines suit généralement un modèle d'isolement par distance qui corrèle avec la géographie, les familles linguistiques et les modes de subsistance des populations (Novembre & Di Rienzo, 2009 ; Henn et al., 2012 ; M. C. Campbell et al., 2014 ;

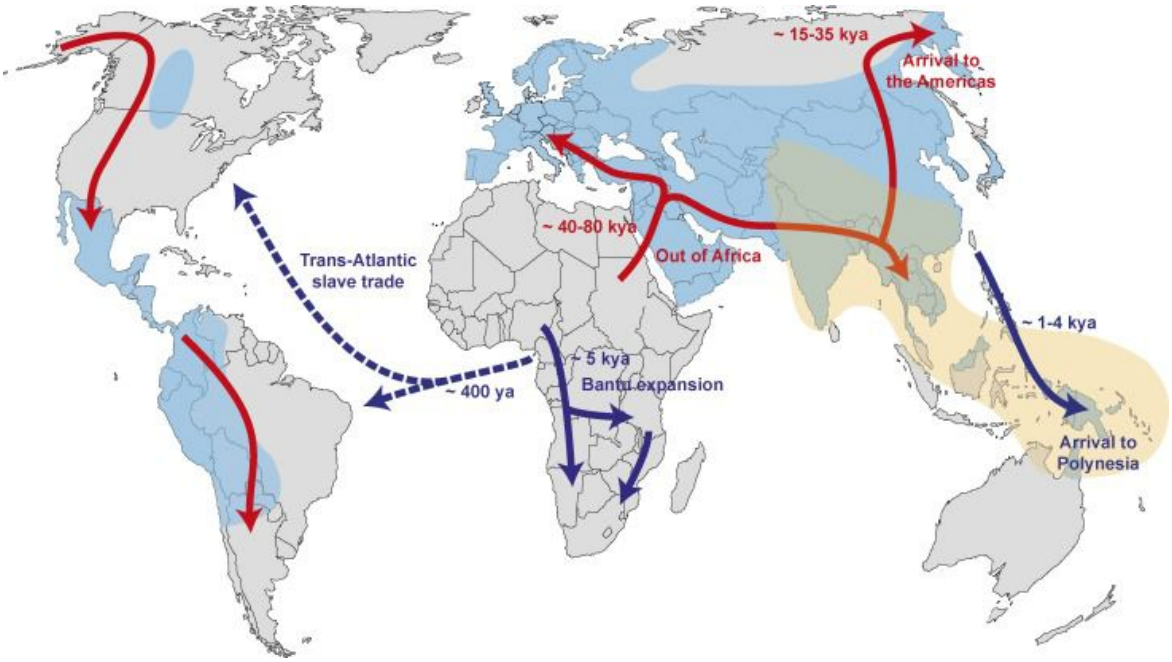


Fig. 2.1 ***Migrations majeures des populations humaines (Quach et al., 2016).*** *Les flèches indiquent les migrations anciennes (rouges) et récentes (< 5 000 ans, bleues) des populations humaines. Les zones géographiques où les populations humaines modernes présentent un mélange génétique avec les hominidés archaïques de Néandertal ou de Dénisova sont représentées en bleu et orange respectivement.*

Patin et al., 2014). Ces études ont également mis en évidence l'échange de migrants entre les populations à des taux variables. Les progrès réalisés en matière de séquencage ont permis d'obtenir des séquences d'ADN d'homininés archaïques tels que l'Homme de Néandertal et l'Homme de Denisova, présents en Eurasie autour de 30 000 à 50 000 ans (Figure 2.1). L'analyse de ces séquences a permis de mettre en évidence l'existence de mélanges génétiques entre ces populations archaïques et les populations eurasiennes modernes (Kelso & Prufer, 2014 ; Vattathil & Akey, 2015). Les segments génomiques résiduels résultant de l'introgression archaïque avec l'homme de Néandertal constituent environ 2% du génome des Européens et 4% du génome des populations asiatiques (Sankararaman et al 2012, Green et al 2010, Prufer et al 2014). Les fragments d'ADN issus d'événements de métissage avec l'Homme de Denisova sont principalement retrouvés dans les populations Australo-Mélanesiennes, où ils représentent jusqu'à 6% de leur génome, ainsi que dans les populations d'Asie du sud-est dans de plus faibles proportions (Reich et al 2010, 2011, Meyer et al 2012, Vernot2016). De plus, bien que les restes fossiles d'homininés archaïques ne soient pas disponibles en Afrique, des analyses génétiques ont suggéré que les populations africaines se seraient elles aussi métissées avec des homininés archaïques encore inconnu (Hammer et al 2011, Lachance et al 2012, Plagnol and Wall 2006).

2.3 Histoire démographique des Pygmées et non-Pygmées d'Afrique centrale

2.3.1 Études phylogéographiques à partir de marqueurs à hérédité monoparentale

Comme nous l'avons vu dans le premier chapitre, les travaux d'ethnographes et d'anthropologues ont montré l'existence de plusieurs groupes de Pygmées à l'ouest et à l'est du Bassin du Congo. Ils ont également mis en évidence les interactions sociales et économiques complexes qu'entretiennent les Pygmées avec les non-Pygmées (S. Bahuchet, 1992 ; Hewlett, 1996 ; Joiris, 2003). Cependant, ces travaux n'ont pas pu déterminer si les Pygmées d'Afrique centrale et les non-Pygmées possédaient une origine commune ou indépendante. De la même manière, l'étude de la structure génétique des populations de Pygmées à l'ouest et à l'est du Bassin du Congo n'a pas pu conclure sur l'origine de ces populations et sur la date à laquelle elles ont divergé l'une de l'autre.

Plusieurs études ont tenté de répondre à ces questions en utilisant les marqueurs génétiques à hérédité monoparentale portés par l'ADN mitochondrial et le chromosome Y. En effet, une petit génome circulaire est présent dans la mitochondrie, un organite cellulaire responsable de la production d'énergie nécessaire au métabolisme. Le génome mitochondrial a la particularité d'être transmis uniquement par les mères au cours des générations. Par conséquent, il est donc possible de retracer l'histoire des lignées maternelles des populations en étudiant les mutations présentes sur cette molécule d'ADN. De la même manière, les hommes possèdent un chromosome X et un chromosome Y sur la paire de chromosomes sexuels et les chromosomes Y sont transmis de père en fils au cours des générations. A l'inverse de la mitochondrie, les mutations portées par le chromosome Y permettent de retracer l'histoire des lignées paternelles des populations. Le génome mitochondrial et le chromosome Y ont l'autre particularité de ne pas recombiner, ce qui permet de facilement retracer l'évolution de mutations qui sont apparues sur ces marqueurs, mais sont cependant plus influencés par la sélection naturelle et ne donnent qu'une image partielle de l'histoire génétique des populations.

L'étude de l'ADN mitochondrial de BaAka, Mbuti et BaAka Mbenzele (sud de la RDC) a révélé que ces trois groupes présentaient des ADN mitochondriaux fortement différenciés mais partageaient des marqueurs suggérant une origine maternelle commune (Destro-Bisol et al., 2004). En supposant l'origine commune des Pygmées de l'ouest et de l'est, des résultats obtenus par simulation ont estimé une divergence des lignées maternelles des Pygmées et non-Pygmées à environ 70 000 ans et ont également estimé la divergence des Pygmées de l'ouest et de l'est entre 3 000 et 18 000 ans (Destro-Bisol et al., 2004). D'autres études se sont concentrées sur l'estimation et le raffinement de ces temps de divergence à partir de marqueurs à hérédité monoparentale mais ces estimations peuvent être biaisées par des contributions différentes des lignées maternelles et paternelles à la diversité génétique (Batini et al., 2007, 2011 ; Quintana-Murci et al., 2008 ; Destro-Bisol et al., 2004).

Ces marqueurs permettent en revanche de séparer la contribution des lignées maternelles et paternelles aux dynamiques de migration entre les groupes. Les interactions génétiques complexes entre Pygmées et non-Pygmées conduisent à des distances génétiques différentes pour les lignées maternelles et paternelles. Tout d'abord, bien que fortement différenciés, les marqueurs portés par l'ADN mitochondrial présentent un certain degré de partage entre Pygmées et non-Pygmées (Verdu et al., 2013). Ces signatures génétiques résultent des flux de gènes maternels entre ces populations via des systèmes patrilocaux, c'est-à-dire des mariages entre Pygmées et non-Pygmées où les épouses vont vivre avec la famille des époux après le mariage. Les mélanges génétiques seraient apparus avec les premières migrations d'agriculteurs de langues bantoues (Salas et al., 2002 ; Seielstad et al., 1998) et ces résultats ont été corroborés par d'autres études incluant différentes populations de Pygmées (Batini et al., 2007 ; Destro-Bisol et al., 2004 ; Quintana-Murci et al., 2008).

A l'inverse, le travaux menés sur le chromosome Y ont montré une faible différenciation des marqueurs génétiques entre les populations, ce qui est cohérent avec une forte migration des hommes non-Pygmées vers les groupes Pygmées (Verdu et al., 2013). Ce déséquilibre du sexe ratio de migration (jusqu'à 3 fois supérieur pour les hommes) n'est pas attendu au vu des coutumes de patrilocalité dans la région (Hammer et al., 2011). Ces résultats s'expliquent par des mariages inter-ethniques entre des hommes non-Pygmées et des femmes Pygmées mais surtout par le retour fréquent des femmes Pygmées dans leurs communautés accompagnées des enfants métissés suite au décès de leur mari non-Pygmée, ou d'un divorce souvent lié aux fortes discriminations dont elles sont victimes (S. Bahuchet & Guillaume, 1982 ; Hewlett, 1996 ; Joiris, 2003). La mobilité des épouses et des enfants issus des mariages inter-ethniques pourrait expliquer le flux de gènes portés par le chromosome Y des populations non-Pygmées vers les populations de Pygmées dans un contexte de patrilocalité et de discrimination sociale contre les Pygmées.

2.3.2 Reconstruction de la démographie à partir de données autosomales

La disponibilité croissante de données de génotypage et de séquencage partiel du génome au cours des années 2000 ont facilité la reconstruction de l'histoire démographique des populations de Pygmées et de non-Pygmées à partir de données autosomales (Figure 2.2). De nombreux travaux se sont concentrés sur l'estimation des temps de divergence, des tailles effectives de populations et des taux de migrations entre les groupes de chasseurs-cueilleurs Pygmées et agriculteurs non-Pygmées à partir du génotypage de microsatellites ou du séquencage de plusieurs dizaines de régions "neutres" du génome (hors des régions géniques) (Patin et al., 2009 ; Verdu et al., 2009 ; Veeramah et al., 2012). L'utilisation des méthodes d'ABC sur ces données ont montré pour la première fois que la diversité génétique neutre des populations de Pygmées du Bassin de Congo était mieux expliquée par une origine commune que par une origine indépendante des groupes (Patin et al., 2009 ; Veeramah et al., 2012 ; Batini et al., 2011). Ces travaux ont testé formellement l'arbre de divergence des populations et estimé la divergence des populations de Pygmées et non-Pygmées à environ 60 000 ans. Ils ont

également conclu à l'origine commune des Pygmées de l'ouest et de l'est dont la population ancestrale s'est séparée il y a environ 20 000 ans (Patin et al., 2009), puis s'est à nouveau séparée à l'ouest il y a environ 3 000 ans (Verdu et al., 2009) (Figure 2.2). Ainsi, la différenciation génétique rapide des groupes de Pygmées et la structure des populations pourraient être expliquées par l'isolement reproductif et la dérive génétique. Même s'il est impossible pour les généticiens de déterminer les causes de cette divergence des populations, une approche multidisciplinaire intégrant des données archéologiques et paléo-climatologiques permettent de proposer certaines hypothèses. Les dates de divergence des populations de Pygmées de l'ouest et de l'est entre -70 000 et -60 000 ans, correspondent au Dernier Maximum Glaciaire (Last Glacial Maximum) (Figure 1.4), caractérisé par une fragmentation des forêts et l'apparition de petits refuges forestiers dans le Bassin du Congo. Cette réduction des habitats forestiers aurait provoqué l'isolement des populations ancestrales des Pygmées de l'ouest et de l'est de l'Afrique centrale (Batini et al., 2011 ; Patin et al., 2009) (Figure 2.2).

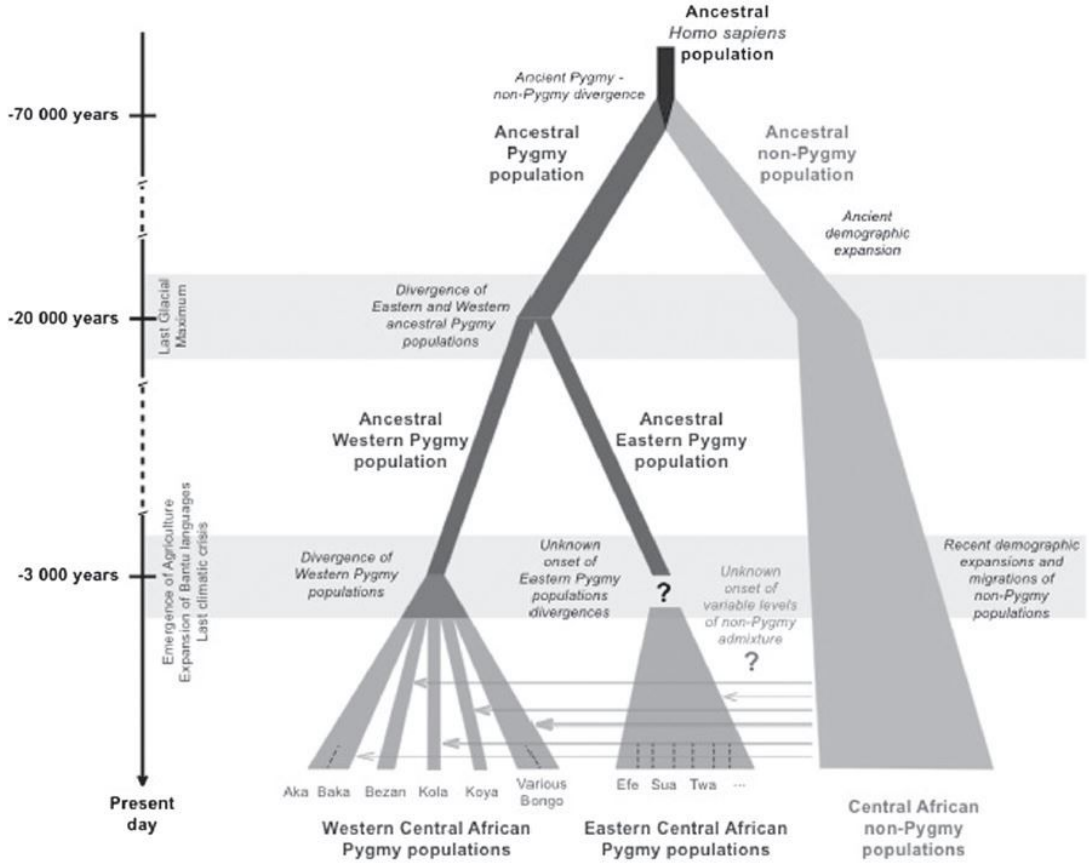


Fig. 2.2 ***Modèle démographique récapitulatif des populations de Pygmées et de non-Pygmées d'Afrique centrale (Hewlett, 2014)*** *Reconstruction de la démographie des populations de Pygmées et de non-Pygmées inférée à partir de données autosomales, selon Verdu et al., 2009, Patin et al., 2009 et Batini et al., 2011.*

A l'ouest, des données archéologiques et archéolinguistiques ont montré respectivement que l'agriculture et les langues bantoues sont apparus il y environ 4 000 à 5 000 ans, et

se sont propagés en Afrique centrale et du sud (Philipson, 2005) (Figure 2.1). Les travaux de génétique des populations ont montré que ces diffusions culturelles et technologiques ont été accompagnées d'une forte expansion démographique des populations d'agriculteurs, débutant il y a 10 000 à 30 000 ans (Verdu et al., 2009 ; Patin et al., 2009, 2014). A l'inverse, les chasseurs-cueilleurs Pygmées ont connu à la même époque des événements de contraction des tailles populations (Patin et al., 2014) (Figure 2.2).

L'émergence de l'agriculture et les migrations des populations de langues bantoues ont été à l'origine de profonds changements sociaux et démographiques chez les populations ancestrales des Pygmées. Les études génétiques indiquent que des événements de migration entre les populations ont eu lieu dès l'arrivée des proto-agriculteurs, mais se sont intensifiés au cours des 1 000 dernières années (Patin et al., 2009, 2014 ; Verdu et al., 2009). Les groupes de Pygmées de l'ouest et de l'est de l'Afrique centrale présentent aujourd'hui des niveaux de métissage avec les non-Pygmées inférieurs à 6% chez les Mbuti et les Biaka mais qui atteignent jusqu'à 39% et 48% chez les Bezan et les BaBongo respectivement (Patin et al., 2014). A l'inverse, le mélange génétique des agriculteurs n'est en moyenne que de 10% avec les populations de Pygmées (Patin et al., 2014).

2.3.3 Intérêts des données de séquençage à haut débit

Bien que la reconstruction de l'histoire démographique des populations de chasseurs-cueilleurs Pygmées et d'agriculteurs non-Pygmées ait été facilitée par le séquencage de quelques régions neutres et par les données de génotypage en génome entier, ces données manquent de résolution en raison du faible nombre de polymorphismes étudiés, ou de la sous-représentation de mutations rares dans les puces de génotypage, respectivement. Les données de séquencage de génomes complets permettent d'éviter ces problèmes et d'obtenir une distribution complète et non biaisée des fréquences alléliques. Les SFS obtenus à partir de ces données permettent alors une estimation fine des paramètres démographiques.

La première analyse réalisée à partir de données de séquençage de génomes complets de Pygmées (3 Baka, 1 BaKola et 1 Bedzan) a estimé la chronologie des divergences de ces populations avec d'autres groupes de chasseurs-cueilleurs africains, des populations pastoralistes et des populations eurasiennes (Figure 1.1) (Lachance et al., 2012). Ces travaux ont révélé que les groupes de Pygmées sont les premiers à diverger du reste des populations et présentent une diversité génétique élevée ainsi qu'une forte sous-structure de population (Lachance et al., 2012). D'autres travaux par Hsieh et al. en 2016 ont estimé les temps de divergence, les variations de tailles effectives et les taux de migration entre les populations de Pygmées de l'ouest et non-Pygmées (Hsieh et al., 2016). L'étude des génomes complets de 3 individus Baka, 4 individus Biaka et 9 individus Yoruba (non-Pygmées) ont permis d'identifier environ 1.58 millions de SNPs autosomaux. L'inférence de la démographie réalisée à partir du SFS indique que les deux populations de Pygmées ont divergé entre -5 000 et -4 000 ans. Ces travaux estiment également des tailles effectives de populations plus élevées chez les non-Pygmées que chez les Pygmées et un taux de migration des populations Pygmées vers les

populations de non-Pygmées environ 10 fois supérieur aux taux de migrations inverses. Il est intéressant de noter que ces travaux estiment une divergence des ancêtres des populations de Pygmées et de non-Pygmées il y a environ 90 000 à 150 000 ans. Ces estimations dépassent les dates estimées par le passé (Verdu et al., 2009 ; Patin et al., 2009 ; Batini et al., 2011 ; Veeramah et al., 2012) et peuvent être expliquées par une meilleure résolution des données de génome entier pour estimer les évènements démographiques anciens, bien que les faibles tailles d'échantillons réduisent la précision d'estimation d'événements démographiques.

Chapitre 3

Sélection naturelle chez les Pygmées

L'inférence de l'histoire démographique récente des populations humaines a mis en évidence des variations considérables de leurs tailles de populations (Nielsen et al., 2017 ; Henn et al., 2012). Comme nous l'avons vu dans le chapitre précédent, l'augmentation ou la diminution de la taille effective des populations fait varier l'intensité de la dérive génétique et modifie la distribution des fréquences alléliques. Cependant, on compte parmi les mutations du génome une fraction non négligeable de variants dont les effets sont délétères (Henn et al., 2015 ; Fu et al., 2013 ; Gazave et al., 2013) et participent à la susceptibilité génétique aux maladies (Lohmueller, 2014b). Ainsi, dans les populations de faible N_e, la fréquence de ces mutations délétères sera principalement contrôlée par les effets stochastiques de la dérive génétique alors que l'augmentation du N_e augmentera aussi l'efficacité de la sélection négative à les purger de la population. Au cours des dernières années ces résultats théoriques ont été évalués dans les populations humaines à l'aide de données empiriques, menant parfois à des conclusions contradictoires (Simons & Sella, 2016 ; Henn et al., 2015). Ces observations soulignent l'importance de comprendre les interactions complexes entre l'histoire démographique et les effets de la sélection négative afin de déterminer si la fluctuation des tailles effectives ont influencé ou non les fardeaux de mutations délétères portés par les différentes populations humaines.

De plus, l'histoire récente des populations humaines a été marquée par la colonisation de nouveaux environnements variés, conduisant à l'adaptation locale de ces populations via des mécanismes de sélection positive (Fan et al., 2016). L'étude de ces mécanismes adaptatifs ont permis de mettre en évidence plusieurs gènes et fonctions biologiques particulièrement important au cours du passé évolutif des populations humaines. Cependant, les mécanismes de sélection naturelle permettant cette adaptation sont variés et restent parfois difficiles à détecter (Fan et al., 2016 ; Pritchard et al., 2010 ; Messer & Petrov, 2013). L'histoire récente des populations humaines est donc largement influencée par les forces de sélection naturelle, à la fois négative et positive, qui contribuent à la valeur sélective des individus de populations différentes. Dans le contexte de l'Afrique centrale, l'étude des populations de chasseurs-cueilleurs Pygmées et d'agriculteurs non-Pygmées, qui présentent des modes de subsistance, des habitats et des régimes démographiques différents permet de mieux caractériser les forces de sélection naturelle qui s'appliquent sur le génome humain.

3.1 Histoire démographique et fardeau de mutations délétères

3.1.1 Mutations délétères dans le génome humain

Parmi les nouvelles mutations introduites dans le génome à chaque génération, la plupart sont délétères (Keightley & Lynch, 2003) et susceptibles de diminuer le succès reproducteur de l'individu qui les porte (aussi appelé valeur sélective ou *fitness* en anglais), limitant ainsi sa contribution moyenne à la diversité génétique de la prochaine génération. A l'exception des allèles dont les effets sont létaux à l'état hétérozygote, les allèles délétères peuvent persister dans les populations pendant de nombreuses générations avant d'être éliminés par la sélection négative, aussi appelée sélection purificatrice. Ainsi, les mutations délétères du génome sont en moyenne plus récentes que les variations neutres (Kiezun et al., 2013) et ont des fréquences alléliques relativement faibles (Henn et al., 2015). Ces mutations sont d'une importance capitale dans la compréhension des facteurs génétiques de susceptibilité aux maladies complexes (Agarwala et al., 2013 ; Maher et al., 2012 ; Lohmueller, 2014b). L'avènement des technologies de séquençage à haut débit a grandement contribué à la détection des mutations délétères dans les populations humaines (Tennessen et al., 2012 ; Charlesworth, 2009 ; Auton et al., 2015). Chez l'Homme, chaque individu porte en moyenne entre 250 et 300 mutations conduisant à la perte de fonction du gène muté (Auton et al., 2015), ainsi que quelques centaines de mutations modifiant la séquence d'acide aminés (non-synonymes) dont les effets sont plus ou moins sévères (Eyre-Walker, 2006 ; Charlesworth, 2010). En considérant la fraction du génome sous contrainte sélective, on peut supposer que le nombre de variants délétères ayant un rôle dans la régulation génique est au moins aussi élevé que le nombre de mutations non-synonymes (Eory et al., 2010). En définitive, chaque individu porte en moyenne plusieurs milliers de mutations délétères, la plupart à l'état hétérozygote, qui ne sont pas éliminées par la sélection négative et qui peuvent affecter sa valeur sélective.

3.1.2 Efficacité de la sélection négative

Le temps nécessaire à l'élimination des mutations délétères dans le génome dépend de son coefficient de sélection standardisé noté $\gamma{=}N_e s$, avec N_e la taille effective de la population et s le coefficient de sélection. Cela indique que la purge des mutations délétères dans une population dépend à la fois de l'effet délétère de la mutation, mesuré par s, et la taille de la population N_e.

La théorie neutraliste prédisant l'évolution aléatoire des fréquences alléliques dans la population a été établie grâce à des travaux théoriques et empiriques au milieu du XX$^{\text{ème}}$ siècle. A cette époque tous les polymorphismes génétiques étaient considérés fonctionnels. Cependant, à mesure que de nouveaux polymorphismes ont été détectés, la diversité génétique existant entre les espèces et les individus s'est avérée plus importante que prédite au départ. Tomoko Ohta et Motoo Kimura ont étendu les principes de la théorie neutraliste et statué que les mutations ayant un faible impact sur la valeur sélective évoluent aussi de manière neutre donc stochastique dans les populations. Plus particulièrement, si une mutation a un coefficient de sélection $|s| << \frac{1}{N_e}$, elle évoluera dans la population de la même manière

qu'une mutation neutre. Les mutations quasi-neutres, c'est-à-dire les mutations faiblement délétères dont la fréquence évolue de façon stochastique, constituent une classe de mutations dont le coefficient de sélection est environ égal à $\frac{1}{N_e}$. Ces mutations pourront évoluer de façon neutre dans une population de faible taille effective, être quasi-neutre dans une population de taille intermédiaire et être éliminée par la sélection négative dans une population de grande taille où la dérive génétique a un effet faible. Les fréquences alléliques des mutations délétères sont donc conditionnée par l'histoire démographique de la population.

Le nombre de mutations ayant un effet faiblement ou moyennement délétère dans un population va donc dépendre de son N_e. Les variations de N_e observées au cours de l'histoire évolutive d'une population conditionnent son N_e global défini comme la moyenne harmonique des tailles effectives au cours des générations. Ainsi, une population qui connaîtrait une expansion démographique rapide après un épisode de contraction de sa taille de population aura une taille effective plus faible qu'une population de même taille dans laquelle la contraction n'aurait pas eu lieu car les faibles valeurs de N_e vont fortement influencer la moyenne harmonique.

3.1.3 Fardeau de mutations délétères

L'effet des mutations délétères dans une population peut être évalué en quantifiant leur contribution à la réduction de la valeur sélective des individus. La diminution du succès reproducteur résultant de la combinaison de mutations délétères portées par un individu est appelé le fardeau de mutations délétères (*mutational load* en anglais). Le fardeau de mutations délétères est la composante du fardeau génétique associée à la réduction de la valeur sélective liée aux effets des mutations. D'autres composantes du fardeau génétique incluent la réduction de la valeur sélective associée à la consanguinité de la population (*inbreeding load*), à un génotype hétérozygote ayant une valeur sélective plus élevée que chacun des génotypes homozygotes (*segregation load*), à des allèles devenus délèteres suite à un changement de l'environnement (*transitory load*) ou à la dérive génétique (*drift load*). En effet, la dérive génétique peut occasionnellement mener à la fixation d'allèles délétères dans une population de petite taille effective, réduisant alors la valeur sélective de tous les individus. Ce chapitre se concentre sur les effets de la réduction de la valeur sélective due aux mutations.

L'étude du fardeau de mutations délétères a fait l'objet de nombreux travaux au cours de la deuxième moitié du XX$^{\text{ème}}$ siècle. Des travaux datant de 1950 par H.J. Muller ont posé les fondements de la contribution des variations délétères à la mortalité et aux risques de maladies chez l'homme. Bien que le terme de fardeau de mutations délétères soit souvent employé dans un contexte général et parfois à tort pour designer l'ensemble des conséquences associées aux mutations délétères présentes dans le génome, il sera utilisé ici dans son sens formel, c'est-à-dire la réduction de la valeur sélective d'un individu attribuable aux mutations délétères qu'il porte dans son génome par rapport au génotype optimal dont la valeur sélective est maximum. Afin de comparer les effets de la démographie sur l'efficacité de la sélection entre les populations humaines, il est nécessaire de quantifier le fardeau de muta-

tions délétères noté L :

$$L = \frac{W_{\max} - W_{\text{mean}}}{W_{\max}}$$

W_{max} représente la valeur sélective maximum notée 1 pour des raisons algébriques et W_{mean} est la valeur sélective moyenne de l'individu sur tous les loci. La plupart des travaux théoriques ont utilisé des hypothèses simplificatrices concernant le coefficient de dominance h des mutations. De plus, même si W_{max} est facile à identifier dans un modèle idéal, il reste complexe à estimer dans un modèle empirique.

La proportion de mutations délétères récessives ou dominantes dans le génome humain reste une question ouverte. Le coefficient de dominance h a cependant un effet majeur sur la quantification du fardeau de mutations délétères d'un individu. A un locus pour lequel les génotypes seraient AA, Aa et aa, leurs valeurs sélectives respectives sont W_{AA}=1, W_{Aa}=1-hs et W_{aa}=1-s. Si l'on souhaite étendre ce résultat à tous les loci, il est nécessaire de disposer d'une distribution de h supposée ou bien faire l'hypothèse d'une complète additivité ou récessivité des mutations délétères. Ainsi, h est égal 1 si la mutation est dominante et h est égal à 0 si la mutation est récessive (c'est-à-dire que la mutation doit être à l'état homozygote pour qu'elle ait un impact sur la valeur sélective de l'individu). Des coefficients de dominance intermédiaires entre 0 et 1 seraient également à considérer dans le génome. Le coefficient de dominance est donc probablement le facteur ayant le plus d'impact sur la quantification du fardeau de mutations délétères pour un individu. Cette observation est particulièrement vraie pour les populations humaines dont la proportion de sites homozygotes, et donc le fardeau de mutations délétères récessives, varie au cours de leur histoire démographique (goulots d'étranglement, effets fondateurs).

3.1.4 Quantification du fardeau de mutations délétères dans les populations humaines

Les populations humaines ont subi des variations majeures de tailles de populations au cours des 100 000 dernières années (Tennessen et al., 2012 ; Malaspinas et al., 2016 ; Mallick et al., 2016 ; Nielsen et al., 2017) en particulier des événements de réduction des tailles de populations suivis d'expansions ou d'effets fondateurs le long des routes de migration (Figure 3.1). Afin de comprendre si ces variations de N_e ont entraîné des différences d'efficacité de la sélection négative, plusieurs travaux se sont concentrés sur la comparaison du fardeau de mutations délétères dans les populations humaines (Lohmueller et al., 2008 ; Simons et al., 2014 ; Henn et al., 2016 ; Do et al., 2015 ; Casals et al., 2013 ; Pedersen et al., 2017 ; Peischl et al., 2016).

Bien que la valeur exacte du fardeau de mutations délétères d'un individu ne puisse pas être calculée en raison de la distribution inconnue de h chez l'Homme, il est néanmoins possible d'utiliser des statistiques corrélées au fardeau de mutations délétères pour l'estimer. L'étude du fardeau de mutations délétères chez l'Homme a suscité à la fois intérêt et confusion

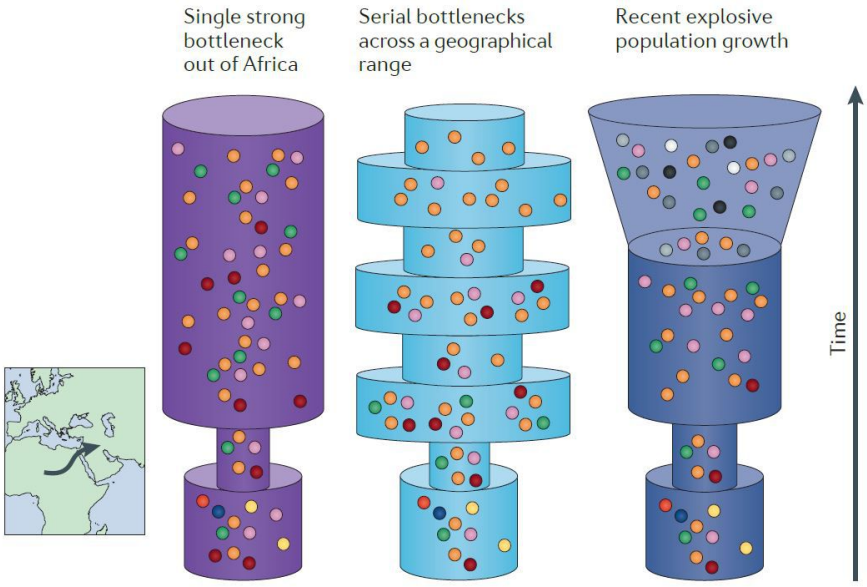


Fig. 3.1 ***Représentation schématique de l'impact de la sortie d'Afrique sur la diversité génétique. (Henn et al., 2015).*** *Représentation schématique de trois modèles démographiques associés à la sortie d'Afrique dans les populations humaines, de gauche à droite : fort goulot d'étranglement, effets fondateurs, goulot d'étranglement suivi d'une expansion. Les points de couleurs représentent la diversité allélique des populations. L'époque contemporaine correspond à la partie supérieure des modèles.*

au cours des dernières années (Lohmueller, 2014a ; Simons & Sella, 2016). Selon les études, les auteurs ont conclu que la démographie avait, ou n'avait pas, de conséquences sur l'efficacité de la sélection purificatrice chez l'Homme (Lohmueller et al., 2008 ; Simons et al., 2014 ; Henn et al., 2016 ; Do et al., 2015 ; Casals et al., 2013 ; Pedersen et al., 2017 ; Peischl et al., 2016). Ces conclusions contradictoires proviennent de différences dans le choix des statistiques mesurant le fardeau de mutations délétères, le calcul des intervalles de confiance ainsi que les hypothèses faites sur l'additivité ou la recessivité des mutations (Simons & Sella, 2016).

Des simulations qui suivent l'évolution du fardeau de mutations délétères en fonction de différents modèles démographiques ont montré que sous un modèle additif, la seule statistique empirique directement corrélée au fardeau de mutations délétères est le nombre d'allèles délétères portés par un individu (Simons & Sella, 2016). Lors de la réanalyse de certaines données, les résultats interprétés comme des différences d'efficacité de la sélection entre les populations africaines et européennes n'apparaissent plus en mesurant le compte d'allèles délétères par individu $N_{\text{allèles}} = 2N_{\text{homozygotes}} + N_{\text{hétérozygotes}}$ (Lohmueller et al., 2008 ; Simons et al., 2014 ; Do et al., 2015). En utilisant cette statistique, la seule diminution de l'efficacité de la sélection qui ait été détectée concerne l'Homme de Denisova (Do et al., 2015). Dans le cas d'une complète récessivité des variants délétères, la corrélation de cette statistique avec le fardeau de mutations délétères est plus complexe (Simons et al., 2014 ; Simons & Sella, 2016).

3.2 Sélection positive et adaptation à l'environnement

3.2.1 Différents types de sélection positive

La plupart des fréquences alléliques d'une population fluctuent de manière neutre ou quasi-neutre. Cependant, il arrive que certaines mutations génétiques apparues au hasard dans le génome confèrent un avantage sélectif aux individus et évoluent sous sélection positive. Avec l'augmentation du succès reproducteur des porteurs de la mutation sélectionnée, celle-ci va augmenter en fréquence dans la population plus rapidement qu'elle ne le ferait sous neutralité.

La sélection positive entraîne des signatures moléculaires locales spécifiques autour de l'allèle sélectionné, dont : la diminution de la diversité génétique, l'augmentation de la présence d'allèles rares et très fréquents, l'augmentation du déséquilibre de liaison et l'augmentation de la différenciation entre populations (Figure 3.2) (Lohmueller et al., 2011 ; Nielsen, 2005 ; Pritchard et al., 2010 ; Scheinfeldt & Tishkoff, 2013). On distingue plusieurs sous-types de sélection positive qui laissent des signatures légèrement différentes sur le génome (Figure 3.2) (Pritchard et al., 2010 ; Scheinfeldt & Tishkoff, 2013 ; Wollstein & Stephan, 2015). Dans le modèle classique, la sélection positive cible une mutation dès son apparition (*selective sweep*). On parle alors de balayage sélectif complet si l'allèle est fixé (*classical sweep*) ou incomplet si l'augmentation en fréquence dans la population est toujours en cours (*ongoing sweep*) (Figure 3.2) (Pritchard et al., 2010 ; Vitti et al., 2013 ; Fan et al., 2016 ; Hernandez et al., 2011).

Il arrive cependant qu'une nouvelle mutation commence à évoluer en fréquence dans la population sous neutralité, ou sous sélection négative faible, avant de devenir avantageuse suite à un changement d'environnement. On parle alors de sélection sur variant préexistant (*selection on standing variation*) (Figure 3.2). Par ailleurs l'avantage sélectif peut être conféré non pas par une seule mutation mais par un ensemble de mutations à plusieurs loci ayant une contribution faible au phénotype adaptatif (Novembre & Di Rienzo, 2009). Dans ce cas, la sélection positive va entraîner une faible augmentation en fréquence simultanée de ces loci, que l'on qualifie de sélection polygénique (*polygenic selection*) (Figure 3.2). Enfin, il est possible que les populations acquièrent des mutations conférant un avantage sélectif par mélange génétique avec des populations déjà adaptées à leur environnement. Cette adaptation à l'environnement médiée par l'acquisition d'un allèle présent dans une autre population s'appelle introgression adaptative (*adaptive introgression*, s'il s'agit d'espèces différentes) ou métissage adaptatif (*adaptive admixture*, s'il s'agit de populations d'une même espèce).

3.2.2 Détecter la sélection positive dans le génome

Tous les régimes de sélection positive détaillés dans le paragraphe précédent ont des conséquences différentes sur la diversité génétique locale (Bamshad & Wooding, 2003 ; Lohmueller et al., 2011 ; Nielsen, 2005 ; Vitti et al., 2013 ; Przeworski et al., 2005 ; Pritchard et

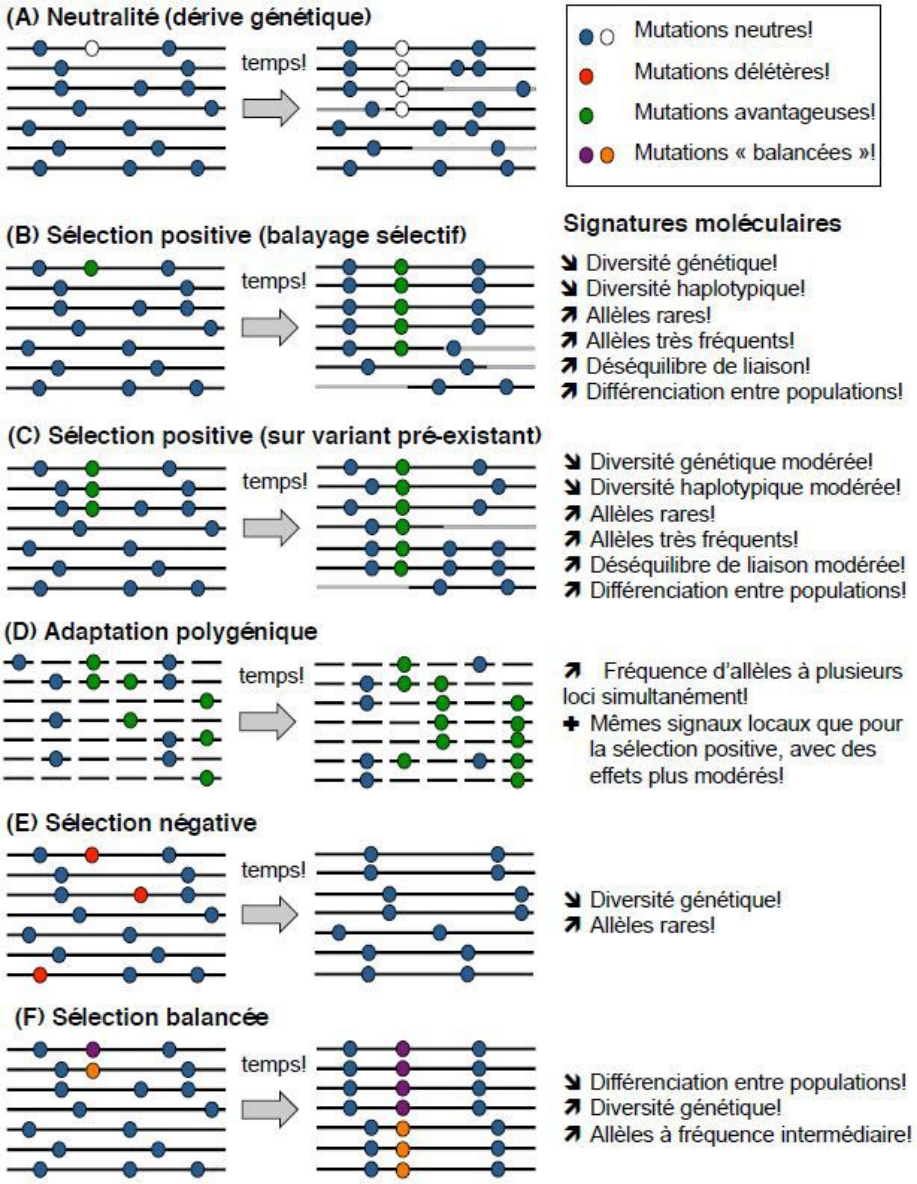


Fig. 3.2 ***Les différents régimes de sélection positive et leurs signatures moléculaires (M.Fagny).*** *Evolution d'une région génomique* ***(A)*** *sous neutralité* ***(B)*** *sous sélection positive à partir d'une mutation de novo* ***(C)*** *Sous sélection positive à partir d'un variant pré-existant* ***(D)*** *sous sélection polygénique* ***(E)*** *sous sélection négative* ***(F)*** *sous sélection balancée.Les points bleu ou blanc sont des mutations neutres et les traits grisés ou noirs indiquent les événements de recombinaison.*

al., 2010). Ces signatures moléculaires peuvent être détectées à l'aide de différentes statistiques. Il existe deux grands types de méthodes permettant de détecter les signatures de sélection positive dans le génome humain : les approches inter-spécifiques et les approches intra-spécifiques. Bien que ces deux approches soient basées sur l'étude des mutations, elles ne détectent pas les mêmes signatures moléculaires de la sélection positive et peuvent par conséquent détecter des événements plus ou moins anciens.

Méthodes inter-spécifiques

Pour détecter les événements de sélection positive ayant eu lieu dans la lignée humaine, on peut comparer certaines régions du génome humain à des régions homologues dans d'autres espèces comme le chimpanzé. Les tests inter-spécifiques utilisent le nombre de mutations fixées entre les espèces et comparent le nombre de mutations fonctionnelles (non-synonymes) au nombre de mutations non fonctionnelles (synonymes). Le ratio $\frac{d_N}{d_S}$ (Yang & Bielawski, 2000) et ses dérivés comparent ainsi le nombre de divergences, synonymes et non-synonymes entre espèces. Le test HKA (Hudson et al., 1987) compare le nombre de divergences au nombre de polymorphismes. Enfin, le test MK (McDonald & Kreitman, 1991) se concentre sur la comparaison du nombre total de divergences et de polymorphismes synonymes et non-synonymes. Pour tous les tests inter-spécifiques, un excès de divergences non-synonymes dans une région génomique peut être le signe d'un événement de sélection positive. En revanche, il faut être vigilant à l'interprétation d'un excès de polymorphisme fonctionnel, car cet excès peut aussi être le signe d'une pseudogénéisation (accumulation de mutations dans un gène qui a ou qui est en train de perdre sa fonction) ou d'un effet de la "sélection d'arrière plan" (*background selection*) (Eyre-Walker & Keightley, 2009 ; Hernandez et al., 2011).

Méthodes intra-spécifiques

Lorsqu'une mutation est la cible de sélection positive dans une population exposée à un environnement particulier, celle-ci augmente en fréquence dans la population mais continue d'évoluer sous neutralité dans les autres populations. Cette évolution différentielle des fréquences alléliques entre populations va entraîner une augmentation de la différenciation génétique à ce locus. Cette signature de différenciation peut être mesurée à l'aide du F_{ST} (Wright, 1943, 1965). Un fort F_{ST} peut résulter d'un événement de sélection positive qui provoque l'augmentation de la fréquence allélique dans une population mais également être le reflet d'une différenciation entre populations due à l'isolement (Hutchison & Templeton, 1999). De plus, même si le F_{ST} détecte le locus pour lequel les différences alléliques entre deux populations sont les plus grandes, il ne permet pas d'identifier la population cible de la sélection positive. Pour cela, il existe des tests basés sur le F_{ST} entre populations qui utilisent une troisième population comme référence. L'avantage de l'utilisation d'une population supplémentaire, souvent éloignée génétiquement, est de tracer un arbre de distances génétiques et d'identifier la population pour laquelle la différenciation est maximale. Parmi ces tests on compte le *PBS* (*population branch statistics*, (Yi et al., 2010 ; Zhang et al., 2005)) combinant les F_{ST} de trois populations et le *LSBL* (*locus-specific branch lengths*, (Shriver et al., 2004)) combinant le F_{ST} de multiples populations.

Afin de détecter les régions génomiques potentiellement ciblées par la sélection positive, il est possible d'utiliser les informations contenues dans les haplotypes. Chaque mutation qui apparaît au sein d'un haplotype est en déséquilibre de liaison avec les mutations qui l'entourent, c'est-à-dire qu'il existe une forte probabilité que cette mutation soit transmise à la descendance avec l'ensemble des mutations de l'haplotype. Au cours des générations et sous neutralité, ce déséquilibre de liaison va disparaître sous l'effet de la recombinaison. Une

mutation neutre à haute fréquence sera en moyenne ancienne et portée par un haplotype "court" (peu conservé, comparé à l'haplotype ancestral sur lequel la mutation est apparue), alors qu'une nouvelle mutation avantageuse à forte fréquence sera portée par un haplotype "long" (conservé). Parce qu'une mutation sous sélection positive augmente rapidement en fréquence, la recombinaison n'aura pas le temps de casser l'haplotype sur lequel elle est apparue. Il existe de nombreuses statistiques de détection de la sélection naturelle qui utilisent ces signatures haplotypiques. Toutes ces statistiques sont basées sur le calcul de l'*EHH* (*extended haplotype homozygosity*, (Sabeti et al., 2002)) qui mesure l'homozygotie autour de chaque allèle à un SNP coeur. En comparant l'*EHH* autour des allèles ancestraux et dérivés à un locus donné, l'*iHS* (*integrated haplotypse score*, (Voight et al., 2006)) permet la détection d'haplotypes anormalement longs. De la même manière, l'*XP-EHH* (*cross population extended haplotype homozygosity*, (Sabeti et al., 2007)) compare l'homozygotie des haplotypes autour des allèles dérivés de deux populations. Ces méthodes qui utilisent les informations haplotypiques permettent de détecter des événements de sélection très récents, environ inférieurs à 30 000 ans (Sabeti et al., 2007).

Il existe donc une grande diversité de statistiques et de méthodes permettant de détecter les signatures moléculaires de la sélection positive mais chaque statistique possède une puissance maximale pour des événements sélectifs ayant lieu à une échelle de temps bien définie. Afin de combler ces lacunes, il existe des méthodes de détection basées sur l'utilisation de plusieurs statistiques combinées ensemble, telles que le *CMS* (*composite of multiple signals*, (Grossman et al., 2010, 2013)).

3.2.3 Approche de génome entier

Des travaux de détection de la sélection positive ont d'abord été menés sur des données de génotypage qui identifient les variations génétiques portées par les individus d'une population à des SNPs prédéfinis. En génotypant plus de 3 millions de SNPs dans onze populations humaines, *The International HapMap Project* (International HapMap et al., 2007) a permis de réaliser les premières études de détection de la sélection naturelle à grande échelle. Cependant, les technologies de génotypage ne donnent accès qu'à un nombre limité de variations génétiques et ces variations, définies au préalable, ne rendent pas compte de la distribution réelle des fréquences alléliques dans la population. En particulier, les mutations à faible fréquence sont absentes des jeux de données de génotypage. Les données de séquençage permettent d'établir la liste exhaustive des mutations portées par un individu. Ces données permettent d'obtenir une distribution non biaisée des fréquences alléliques dans la population et des projets comme *The 1 000 Genomes Project* (Auton et al., 2015) ont permis d'identifier près de 38 millions de variants dans plus de 2 000 individus provenant de 26 populations humaines.

Cependant, malgré les progrès réalisés pour détecter de nouveaux gènes soumis à des pressions adaptatives, les signaux de sélection désignent parfois des régions candidates très larges où la mutation sélectionnée est difficile à identifier, ou bien des régions non géniques dont la fonction n'est pas connue. De plus, certains travaux ont suggéré que la détection géno-

mique de balayages sélectifs présentait un fort taux de faux négatifs et faux positifs (Teshima et al., 2006 ; Kelley et al., 2006), et que de nombreuses régions détectées comme candidates pour la sélection positive pourraient en réalité porter de signatures moléculaires confondues avec celles de la sélection négative (Coop et al., 2009 ; Hernandez et al., 2011 ; Pritchard et al., 2010). De nombreuses études suggèrent également que le modèle classique du balayage sélectif complet est extrêmement rare dans le génome humain et a très peu participé à l'évolution récente des populations humaines (Granka et al., 2012 ; Hernandez et al., 2011). Des mécanismes de sélection moins drastiques ou "soft sweeps" comme la sélection polygénique, la sélection sur des variants préexistants ou le métissage adaptatif pourraient jouer des rôles prépondérants dans l'adaptation, mais leurs signatures moléculaires sont subtiles et restent difficiles à détecter (Figure 3.2).

3.2.4 Exemples d'adaptation chez l'Homme

Les pressions environnementales appliquées à l'Homme au cours de l'évolution peuvent être de différentes natures. De plus, la sélection sexuelle (choix du partenaire sur des caractères phénotypiques morphologiques) a probablement joué un rôle prépondérant dans la distribution et l'augmentation en fréquence de certains variants dans les populations. Hormis la sélection sexuelle, de nombreux travaux indiquent que les agents pathogènes (virus, bactéries, parasites), le climat et le régime alimentaire font parti des pressions sélectives majeures appliquées à l'Homme au cours de son évolution (Figure 3.3) (Akey, 2009 ; Barreiro & Quintana-Murci, 2010 ; Fan et al., 2016 ; Vitti et al., 2013).

Adaptation aux pathogènes

L'exposition aux pathogènes est probablement la pression de sélection la plus importante sur le génome humain (Figure 3.4) (Barreiro & Quintana-Murci, 2010 ; Fumagalli et al., 2011 ; Quintana-Murci & Clark, 2013 ; Siddle & Quintana-Murci, 2014) et environ 200 gènes liés à l'immunité montrent des signatures de sélection positive (Barreiro & Quintana-Murci, 2010). De nombreux pathogènes semblent avoir joué un rôle dans l'évolution des populations humaines : *Plasmodium falciparum*, le parasite responsable du paludisme, les bactéries responsables de la lèpre, de la tuberculose, du choléra, et de nombreux virus (Karlsson et al., 2014). Certaines mutations portées par des gènes impliqués dans la résistance au paludisme sont sous forte sélection positive dans les zones où le parasite est endémique comme *DARC* et *CR1* en Afrique subsaharienne et *G6PD* en Afrique et en Asie du sud-est (Barreiro et al., 2008 ; Hamblin & Di Rienzo, 2000 ; Louicharoen et al., 2009 ; Sabeti et al., 2006 ; Tishkoff et al., 2009). De même, une mutation de *LARGE* qui permet de diminuer la sensibilité des cellules à l'infection par le virus de la fièvre de Lassa (Andersen et al., 2015 ; Sabeti et al., 2007) est sous sélection positive en Afrique de l'ouest, et des gènes de la famille TLR tels que *TLR5* en Afrique et le cluster *TLR1*, *TLR6*, *TLR10* en Europe et en Asie (Wlasiuk et al., 2009 ; Barreiro & Quintana-Murci, 2009 ; Grossman et al., 2013) ont été la cible de la sélection positive, probablement à cause de leur rôle de senseurs de bactéries à la surface des cellules.

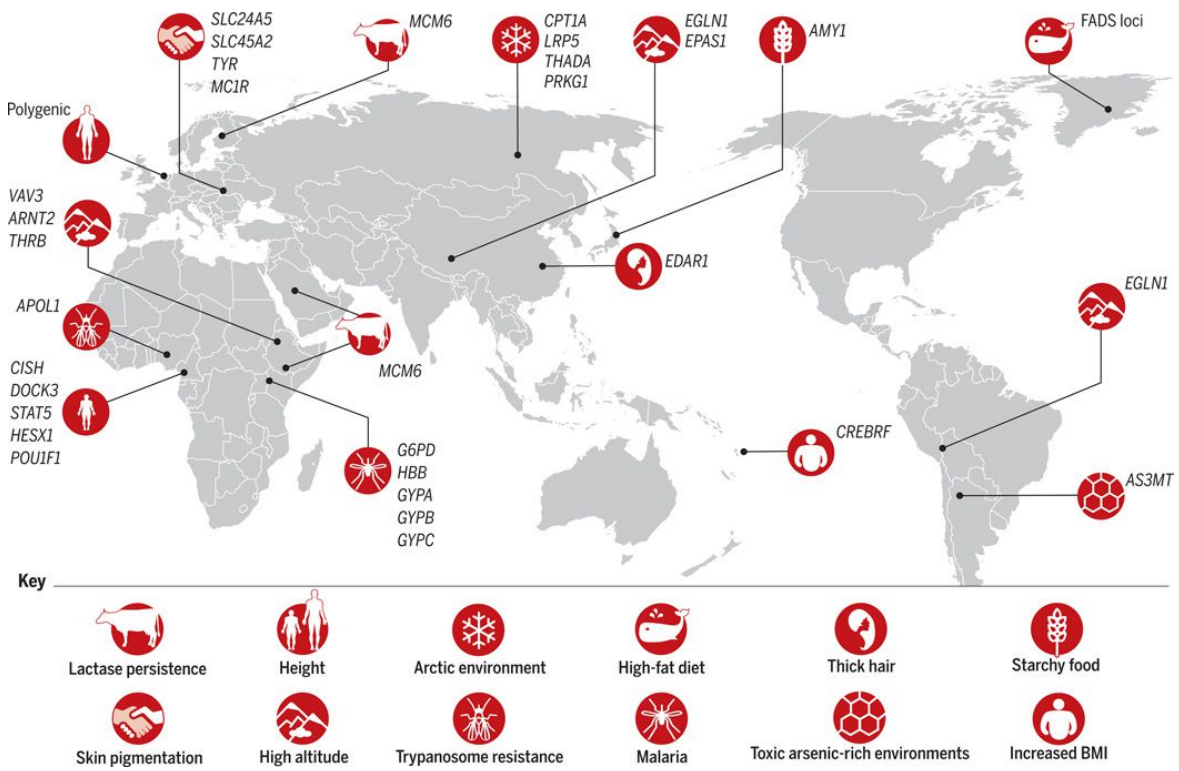


Fig. 3.3 ***Adaptation locale des populations humaines à leur environnement (Fan et al., 2016))*** *Exemples de gènes et de phénotypes ayant permis l'adaptation des populations humaines aux contraintes sélectives de leur environnement local. Chaque phénotype adaptatif est légendé en fonction la nature du trait sélectionné et indique le/les gènes sous sélection.*

Adaptation au climat

Il existe un certain nombre d'exemples d'adaptation des populations humaines au climat (Hancock et al., 2011 ; Wollstein & Stephan, 2015). La plus connue est l'adaptation à l'ensoleillement et à l'exposition aux UV. On trouve en effet parmi les gènes sous sélection en Europe et en Asie, des gènes associés à une pigmentation plus claire de la peau, des cheveux et des yeux comme *SLC24A5*, *MATP*, *KITLG*, *TYR*, *HERC2*, *OCA2*, *TPCN2*, *IRF4* et *ASIP*. Ces gènes ont probablement été sélectionnés car ils confèrent un avantage dans le cadre d'une faible exposition aux UV (Izagirre et al., 2006 ; Sabeti et al., 2007 ; Sulem et al., 2008 ; Wilde et al., 2014), bien qu'on ne puisse pas exclure le rôle de la sélection sexuelle. Au contraire, en Afrique, le gène *MC1R*, impliqué dans la production de mélanine, est sous forte sélection purificatrice, probablement car la mélanine est indispensable à la protection contre les dommages provoqués par l'exposition aux rayons UV sur l'ADN et la lyse du folate (Harding et al., 2000 ; Jablonski and Chaplin 2000). Il existe d'autres exemples d'adaptation au climat comme l'adaptation à l'hypoxie causée par la haute altitude (>2 500m au dessus du niveau de la mer). Ainsi, les allèles de *EGLN1*, *EPAS1* et *PPARA*, entraînant une diminution de la concentration en hémoglobine, sont sous balayage sélectif dans des populations vivant en altitude comme au Tibet (Beall et al., 2010 ; Simonson et al., 2010 ; Yi et al., 2010).

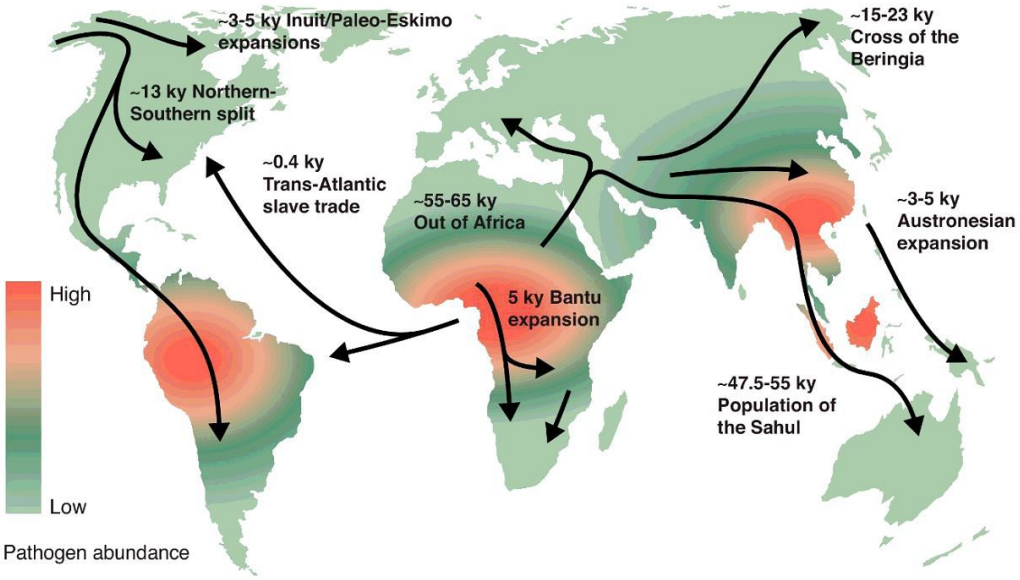


Fig. 3.4 ***Répartition géographique des densités des pathogènes (adapté de (Sanz et al., 2018))*** *Représentation de la densité de pathogènes dans les régions du monde colonisées par les populations humaines au cours 200 000 dernières.*

Adaptation aux changements de régimes alimentaires

Le régime alimentaire a également joué un rôle important sur la diversité phénotypique humaine (Luca et al., 2010). Il semble notamment que l'Homme soit adapté à un mode de vie où les sources de nourriture sont incertaines. L'invention de l'agriculture et sa diffusion dans la quasi-totalité des populations humaines au cours des dernières 10 000 années semble par ailleurs avoir exercé de nombreuses pressions de sélection sur le génome des populations. Les spécificités alimentaires liées à la culture et l'élevage ont notamment été la source d'adaptations locales. Par exemple, la persistance de la lactase à l'âge adulte, dont l'expansion est concomitante à celle de l'élevage au Moyen-orient, en Europe et en Afrique de l'Est, est causée par une mutation dans le gène *LCT* sous sélection positive forte. Il existe également d'autres adaptations du métabolisme au changement de régime alimentaire. Ainsi, la mutation rs1229984 du gène *ADH1B* qui altère la capacité à digérer l'alcool (Dick and Foroud 2003), présente des signatures de balayage sélectif en Asie de l'est et en Europe (Barreiro et al., 2008 ; Galinsky et al., 2016), et son augmentation en fréquence dans les populations suit l'expansion de la culture du riz à l'est de l'Asie. De même, diverses mutations de *NAT2*, causant une acétylation plus lente, sont sous sélection positive dans les populations pratiquant l'agriculture ou le pastoralisme, probablement en réponse à la diminution des folates dans le régime alimentaire (Patin et al., 2006 ; Sabbagh et al., 2011). Un certain nombre de mutations impliquées dans le goût et l'olfaction, deux sens particulièrement importants dans l'alimentation, portent également des signatures de sélection positive. Par exemple, une mutation non-synonyme dans *TAS2R16* provoquant une plus grande sensibilité à certains glycosides, est sous sélection dans l'espèce humaine. Ce phénotype a probablement été sé-

lectionné chez les ancêtres chasseurs-cueilleurs des populations humaines car il confère une protection contre les toxines cyanogènes présentes dans certaines plantes (Soranzo et al. 2005). Un certain nombre de récepteurs d'olfaction sont également sous sélection positive dans plusieurs populations humaines (Gilad and Lancet 2003 ; Gilad et al., 2003 ; Williamson et al., 2007). De façon générale, les gènes sous sélection sont enrichis en gènes impliqués dans le métabolisme des glucides, des graisses et de l'alcool, la perception du goût et l'olfaction (Barreiro et al., 2008 ; Voight et al., 2006), ce qui indique que de nombreux événements d'adaptation pourraient être liés au régime alimentaire.

3.3 Sélection négative et positive chez les Pygmées

3.3.1 Modes de subsistance, histoire démographique et vie dans la forêt

Au sein de la forêt équatoriale, les Pygmées sont exposés à de nombreuses contraintes sélectives puisque cet habitat conjugue des contraintes liées au climat chaud et humide, à la densité accrue en pathogènes (Figure 3.4) et aux ressources alimentaires qui fluctuent drastiquement en fonction des saisons.

Le statut nutritionnel des Pygmées est le résultat des interactions complexes entre la disponibilité des ressources dans leur milieu, leurs dépenses énergétiques, la qualité et la quantité de nourriture ainsi que les maladies et les règles sociales concernant le partage de la nourriture. Un questionnaire alimentaire réalisé chez les BaKola a montré que leur consommation quotidienne de viande est en moyenne de 200 grammes (Koppert & Pasquet., 1993) et ce régime riche en protéines animales et en lipides est visible par l'observation du profil d'acides aminés dans leur sérum (Paolucci & Pennetti., 1973). Cependant, les variations de la quantité de nourriture disponible en forêt existent, en particulier quand le gibier est moins abondant (Garine, 1990). On observe d'ailleurs une forte corrélation entre les variations saisonnières du statut nutritionnel des enfants des populations forestières et les épidémies infectieuses (Pagezy & Hauspie, 1989). De plus, malgré une consommation de viande importante, la prévalence des cas d'anémie est élevée chez les individus Pygmées et sont principalement causées par des maladies parasitaires (Mann & Merrill., 1962).

Il existe un forte corrélation entre le climat et la diversité virale, bactérienne, parasitaire et fongique (Guernier et al., 2004) car la chaleur et l'humidité agissent comme incubateur des maladies transmissibles (Figure 3.4). Alors que certains chasseurs-cueilleurs tels que les aborigènes d'Australie ou les San du Kalahari ne sont porteurs que de deux ou trois types de parasites en moyenne, les Pygmées portent environ une vingtaine de parasites différents (Dunn, 1977). Le milieu forestier est particulièrement favorable à l'infection par les vers intestinaux dont la prévalence de l'infection est de 98% près de l'équateur (Froment & Koppert., 1999). Dans la foret de l'Ituri, 36% des individus sont infectés par des amibes pathogènes et 13% des enfants sont infectés par d'autres protozoaires intestinaux (Mann & Merrill., 1962). La présence de ces parasites et d'autre pathogènes potentiellement virulents a été quantifiée ré-

cemment grâce à l'étude du microbiome intestinal de populations baAka (Gomez et al., 2016).

La foret équatoriale est également considérée comme une forêt très riche en virus (Zerner, 2003) et les niveaux d'immunoglobulines des Pygmées montrent qu'ils sont particulièrement exposés aux infections virales et bactériennes. Les contacts entre les Pygmées, tels que les Baka et les BaKola, et les animaux sauvages, en particulier les primates sont fréquents. Les blessures et morsures lors de la chasse ou du dépeçage de la viande (Aghokeng et al., 2010) peuvent favoriser la transmission de virus tel que le virus spumeux qui affecte 5% des Pygmées (Calattini et al., 2007). Environ 10% des BaKola présentent également des anticorps contre le virus Ebola, probablement dû au contact avec des souches peu pathogènes, et l'infection par HTLV-2 et HTLV-3 est également fréquente. Cependant, le relatif isolement des populations de chasseurs-cueilleurs et leur mode de vie semi-nomade devrait les prémunir contre les virus dont la propagation nécessite une densité d'hôtes importante, ayant sévi dans les groupes sédentaires telles que la variole, la rougeole, les oreillons, la rubéole, ou la grippe, mais peu de données permettent de valider cette hypothèse (Ohenjo & Mugarura, 2006).

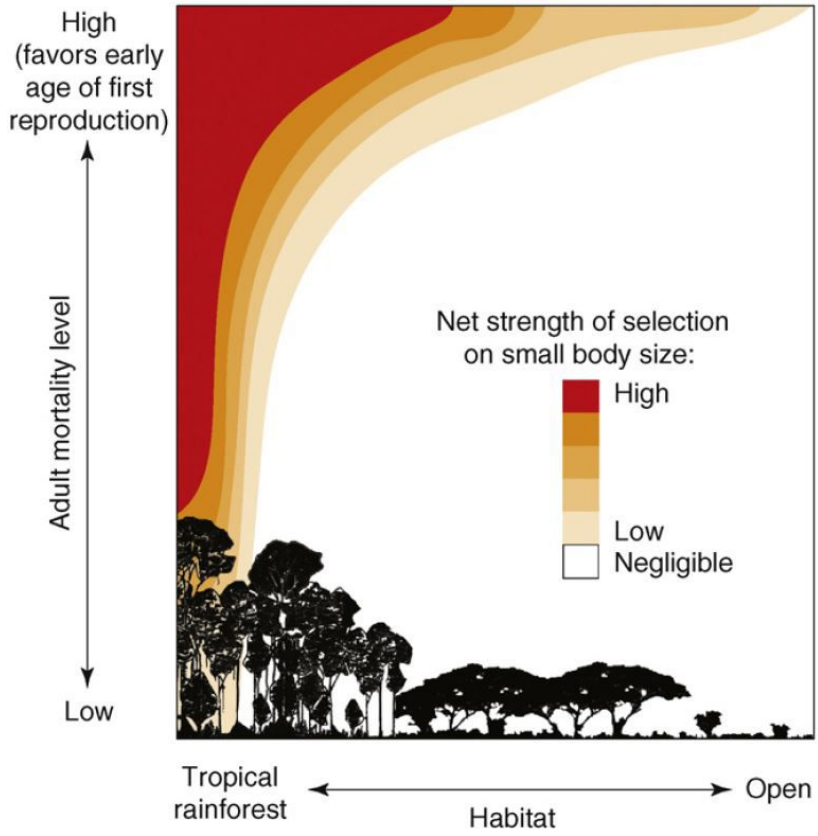


Fig. 3.5 ***Modèle intégratif du phénotype pygmée (Perry & Dominy, 2009)*** *Modèle schématisant l'intensité de la sélection du phénotype pygmée en fonction de l'habitat (densité de la foret tropicale qui conditionnent des facteurs comme les ressources nutritionnelles, la température et l'humidité) et du taux de mortalité (un taux de mortalité élevé pourrait favoriser un âge de reproduction plus jeune et l'arrêt précoce de la croissance).*

La faible stature des Pygmées, en comparaison aux non-Pygmées, semble fortement corrélée à leur habitat (Figure 3.5) et ce phénotype a fait l'objet de nombreuses études et mesures anthropométriques. La taille des Pygmées atteint environ 93% de celle des non-Pygmées mais leurs poids est proportionnellement plus faible, et n'atteint que 79% de celui des non-Pygmées. L'IMC (indice de masse corporelle) des BaKola est environ 91% de celui des non-Pygmées et les masses graisseuses rapportées pour les femmes BaKola ne sont que de 19% contre 25% pour les femmes non-Pygmées (Hewlett, 2014).

Malgré ces observations, on ne connaît pas les mécanismes endocriniens responsables de la morphologie des populations de Pygmées, même si la thyroïde et la glande hypophysaire jouent probablement un rôle majeur. Des expériences réalisées sur des lignées cellulaires de Pygmées Efe (Hattori et al., 1996) suggèrent que la base moléculaire de la faible stature des Pygmées serait liée à une résistance d'un récepteur au facteur de croissance IGF-1 (insuline-like growth factor 1) libéré sous l'action de l'hormone de croissance (GH, growth hormone).

3.3.2 Fardeau de mutations délétères chez les Pygmées

Comprendre les effets des variations démographiques telles que les expansions ou les réductions de tailles de populations sur la capacité des populations à éliminer les mutations délétères est d'une importance capitale et permet de mieux identifier les mutations ayant un rôle dans les maladies complexes (Lohmueller, 2014b ; Agarwala et al., 2013 ; Maher et al., 2012). Cette question est particulièrement intéressante dans le cas des populations de chasseurs-cueilleurs dont la plupart n'ont pas subi d'expansion récente, souvent associée à la transition Néolithique (Pedersen et al., 2017). Cependant l'étude de l'efficacité de la sélection négative des populations de chasseurs-cueilleurs n'a été abordée que pour les populations de San du Kalahari et de Mbuti (Henn et al., 2016 ; Do et al., 2015). En raison des grandes différences de statistiques mesurées, du calcul des intervalles de confiance et du faible nombre d'échantillons, ces travaux ont abouti à des conclusions opposées et cette question reste donc ouverte (Henn et al., 2016 ; Do et al., 2015).

3.3.3 Sélection positive chez les Pygmées et non-Pygmées

Les populations de chasseurs-cueilleurs Pygmées et d'agriculteurs non-Pygmées d'Afrique centrale ont divergés il y a plus de 60 000 ans (Hsieh et al., 2016a ; Patin et al., 2009 ; Verdu et al., 2009) suggérant une longue histoire adaptative dans des environnements différents.

Le phénotype "pygmée", c'est à dire une stature moyenne réduite (Perry & Dominy, 2009 ; Perry et al., 2014) a probablement une base génétique en Afrique (Jarvis et al., 2012 ; Perry et al., 2014) et semble être associé à une meilleure valeur adaptative dans un environnement de forêt tropicale (Figure 3.6) en permettant une meilleure thermorégulation, des besoins réduits en nourriture, une meilleure mobilité et/ou une capacité à se reproduire plus tôt (Perry & Dominy, 2009 ; L. Cavalli-Sforza, 1986 ; J. M. Diamond, 1991 ; Migliano et al., 2007). Les études de génétique des populations ont détecté des signatures moléculaires de sélection positive pour un grand nombre de gènes dans plusieurs populations de chasseurs-cueilleurs Pygmées (Hsieh et al., 2016a ; Lachance et al., 2012 ; Jarvis et al., 2012 ; Perry et

al., 2014 ; Mendizabal et al., 2012 ; Amorim et al., 2015 ; Migliano et al., 2013 ; Lopez Herraez et al., 2009). Cependant, peu de signaux reportés dans ces différentes études sont communs, ce qui peut être le signe de nombreux faux positifs inhérents aux approches de détection par valeurs extrêmes (*outlier approach*). La pertinence de certains loci candidats de sélection positive est cependant soutenue par des études fonctionnelles et épidémiologiques comme les gènes *FLNB* et *EPHB1* qui présentent à la fois des signatures robustes de sélection positive et une association avec la taille chez l'Homme ou dans des organismes modèles (Hsieh et al., 2016a ; Lachance et al., 2012 ; Jarvis et al., 2012). Ces travaux qui détectent différents gènes sous sélection positive dans différents groupes de Pygmées peuvent aussi être le signe d'une adaptation convergente des populations, comme le suggère une étude basée sur les taux de croissance dans des groupes de Pygmées à l'ouest et à l'est de l'Afrique centrale (Rozzi et al., 2015).

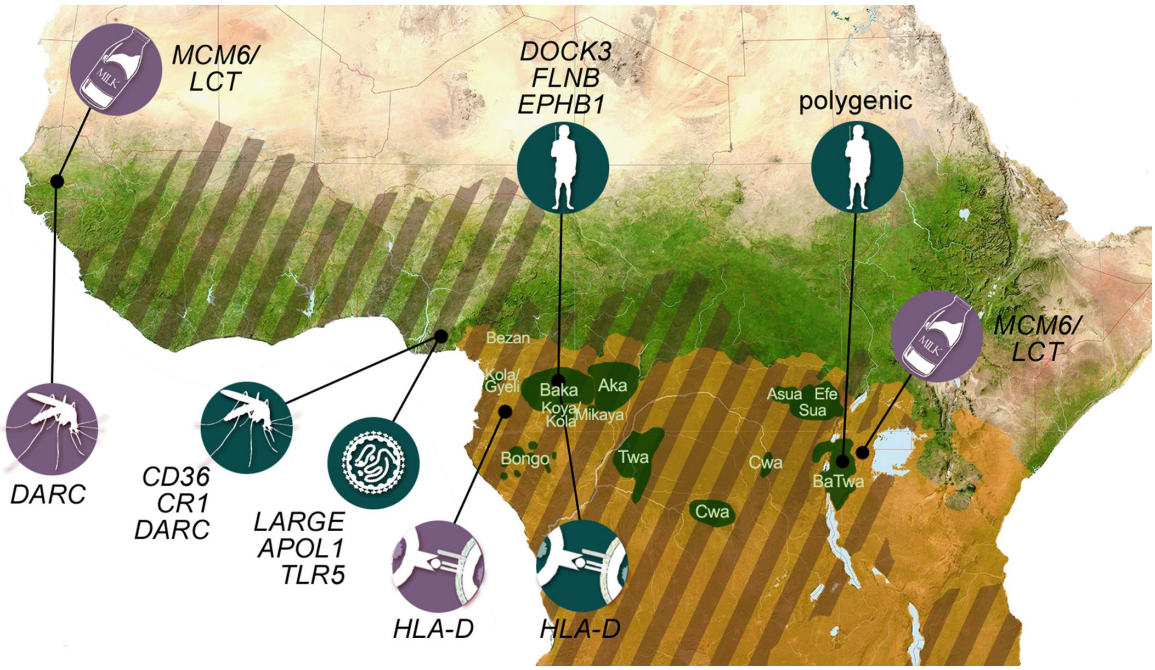


Fig. 3.6 ***Adaptation locale des populations de Pygmées et non-Pygmées d'Afrique centrale (Patin & Quintana-Murci, 2018))*** *Les cercles bleus représentent les phénotypes et les gènes candidats de sélection positive dans un modèle de balayage sélectif. Les cercles violets indiquent des évènements de métissage adaptatif où les loci sous sélection sont acquis par mélange génétique. Les hachures indiquent la présence endémique de Plasmodium falciparum responsable de la malaria.*

Les gènes détectés comme étant sous sélection positive chez les Pygmées appartiennent à différent groupes fonctionnels tels que la reproduction, la signalisation cellulaire, le développement neural et les fonctions immunitaires (Hsieh et al., 2016a ; Lachance et al., 2012 ; Jarvis et al., 2012). De plus, des signaux de sélection polygénique liés au phénotype Pygmées et aux processus immunitaires ont été détectés (Perry et al., 2014 ; Hsieh et al., 2016a).

Dans les populations d'agriculteurs non-Pygmées, des tests de sélection positive ont identifié de nombreux gènes candidats impliqués dans la défense contre les pathogènes. Les

pressions sélectives imposées par l'agent causal du paludisme *Plasmodium* ont menés des évènements de selection positive en réponse au pathogènes comme *HBB*, *DARC* (*ACKR1*), *G6PD*, *CR1* et *CD36* (Gurdasani et al., 2015 ; Sabeti et al., 2002 ; Hamblin & Di Rienzo, 2000 ; Allison, 1954 ; Ruwende et al., 1995 ; Deschamps et al., 2016 ; Barreiro et al., 2008) dont les variations génétiques protègent contre le paludisme provoqué par *Plasmodium falciparum* et *Plasmodium vivax*. D'autres cas d'adaptations génétiques ont détecté le gène *APOL1* qui confère une résistance à *Trypanosoma brucei* (Ko et al., 2013), *LARGE* qui est impliqué dans l'infection par le virus de Lassa (Andersen et al., 2015), et *TLR5* qui est associé avec une diminution de l'activité de NF-κB (Grossman et al., 2013), ainsi que des loci impliqués dans l'adaptation à la température et à l'osmorégulation (Gurdasani et al., 2015).

Chapitre 4
Objectifs de la thèse

L'histoire évolutive des populations façonne leur diversité génétique et influence les mécanismes de sélection naturelle qui s'appliquent sur leurs génomes. Au cours son histoire récente, l'homme a colonisé de nouveaux environnements, adopté différents modes de vie et connu des fluctuations démographiques considérables. Ainsi, la diversification des populations humaines depuis la fin du Pléistocène s'est accompagnée d'importantes modifications des contraintes sélectives qui ont modelé le génome des individus, bien que les mécanismes d'action de la sélection naturelle associés à ces changements restent en partie à déterminer.

En premier lieu, le processus de sélection négative joue un rôle fondamental dans l'élimination des mutations délétères dans les populations, et la vitesse à laquelle ces mutations sont éliminées dépend de la taille effective des populations. Ainsi, l'efficacité de la sélection négative est conditionnée par les événements démographiques qui font varier le nombre d'individus au cours du temps. Cependant, l'efficacité de la sélection reste difficile à évaluer dans les populations humaines dont l'histoire démographique a été marquée par de nombreux évènements à la fois rapides et complexes.

De plus, aux variations du nombre de mutations délétères dans le génome s'ajoutent d'autres mécanismes de sélection qui, à l'inverse, conservent les mutations qui améliorent la survie des individus. Ces mutations, qui confèrent un avantage adaptatif aux individus dans un environnement donné, se propagent dans les populations et modifient localement les profils de diversité génétique. Cependant, le temps nécessaire à cette adaptation dans le modèle canonique de sélection naturelle n'est que peu compatible avec l'observation des nombreux phénotypes adaptatifs dans les populations humaines et d'autres mécanismes d'adaptation sont à considérer.

Dans ce contexte, l'étude des populations de Pygmées et de non-Pygmées d'Afrique centrale, qui présentent des modes de subsistance (chasse et cueillette vs agriculture), des environnements (forêt équatoriale vs savane) ainsi que des histoires démographiques différentes, constitue une excellente opportunité d'évaluer plusieurs questions fondamentales relatives à l'action de la sélection naturelle sur le génome humain. A partir de données de séquençage d'exomes et de données de génotypage obtenues pour plus de 600 individus Pygmées et non-Pygmées provenant de 14 populations différentes, les objectifs de ce travail de thèse se sont articulés autour de deux problématiques principales qui sont (1) de déterminer l'impact des changements démographiques récents de ces populations sur l'efficacité de la sélection néga-

tive dans leur génome et (2) de comprendre les mécanismes de sélection positive à l'origine de l'adaptation des Pygmées aux contraintes environnementales propres à leur habitat et d'identifier les gènes et les fonctions biologiques impliqués.

Pour cela, mon travail de thèse a d'abord consisté à la mise en place d'une suite d'outils permettant le traitement des données de séquençage de 600 exomes des individus tout en s'assurant de leur qualité. Puis, à partir d'un large échantillon de 200 individus à l'ouest (100 Pygmées et 100 non-Pygmées) et 100 individus à l'est (50 Pygmées et 50 non-Pygmées), nous avons réévalué le modèle démographique de ces populations et suivi l'impact des fluctuations des tailles effectives sur leurs fardeaux de mutations délétères. Ce travail qui associe l'analyse de données empiriques et de données de simulations nous a permis de comparer l'efficacité de la sélection négative dans les populations de Pygmées et de non-Pygmées d'Afrique centrale. Ensuite, l'utilisation des données de 566 individus répartis dans 12 populations à l'ouest et 2 populations à l'est nous a permis de scanner leur génome afin d'identifier les gènes porteurs de signaux de sélection positive. Plusieurs scans de sélection nous ont permis d'identifier les gènes candidats sous balayage sélectif, communs ou spécifiques, aux différents groupes de Pygmées et nous nous sommes intéressés aux signatures de sélection polygénique et de métissage adaptatif propres à ces populations.

Chapitre 5

Résultats 1 : Histoire démographique et fardeaux de mutations délétères des chasseurs-cueilleurs et agriculteurs en Afrique

5.1 Contexte

Comme nous l'avons abordé dans les chapitres précédents, les régimes démographiques des populations humaines ont considérablement varié au cours des 100 000 dernières années et ces fluctuations de tailles effectives de populations se reflètent dans leurs niveaux de diversité génétique (Nielsen et al., 2017 ; Henn et al., 2016). Ainsi, le SFS des populations européennes et asiatiques se caractérise par la présence d'un nombre accru de mutations dérivées à forte ou à très faible fréquence résultant d'épisodes successifs de réduction et d'expansion de leurs tailles de populations lors de la sortie d'Afrique (Henn et al., 2015 ; Keinan & Clark, 2012 ; Tennessen et al., 2012 ; Coventry et al., 2010 ; Fu et al., 2013). De plus, le niveau de diversité génétique des populations humaines a été largement influencé par d'autres facteurs tels que le mélange génétique entre les populations modernes (Hellenthal et al., 2014 ; Pickrell et al., 2014) ou avec des populations archaïques (Green et al., 2010 ; Sankararaman et al., 2014 ; Nielsen et al., 2017), le niveau de consanguinité (Narasimhan et al., 2016) et les effets fondateurs (Casals et al., 2013 ; Henn et al., 2016 ; Peischl et al., 2018). Les conséquences de ces variations rapides de N_e sur l'efficacité de la sélection négative dans l'espèce humaine sont longtemps restées méconnues.

Le développement de techniques de séquencage à haut débit a permis de quantifier le nombre et la distribution des mutations délétères qui ségrègent dans les différentes populations humaines, et d'estimer l'impact des fluctuations de N_e sur l'efficacité de la sélection purificatrice (Lohmueller et al., 2008 ; Do et al., 2015 ; Simons et al., 2014 ; Casals et al., 2013 ; Henn et al., 2016). Cependant, l'estimation du fardeau de mutations délétères dans les populations humaines reste controversée puisque plusieurs travaux ont abouti à des conclusions différentes affirmant à la fois qu'il existait ou n'existait pas de différences d'efficacité

de la sélection purificatrice entre les populations humaines (Lohmueller et al., 2008 ; Simons et al., 2014 ; Do et al., 2015 ; Casals et al., 2013 ; Pedersen et al., 2017 ; Peischl et al., 2018). Ces différences sont attribuables en grande partie à la variété des statistiques utilisées pour estimer le fardeau de mutations délétères, aux méthodes statistiques employées pour tester les différences entre populations ainsi qu'aux différents algorithmes de prédiction de la sévérité des mutations dans le génome (Simons & Sella, 2016). Par ailleurs, la plupart de ces travaux comparent les effets de la démographie sur le fardeau de mutations délétères au sein de populations aux histoires démographiques complexes combinant les effets du goulot d'étranglement associé à la sortie d'Afrique et d'expansions explosives récentes, probablement liées à la transition Néolithique (Henn et al., 2016 ; Lohmueller et al., 2008 ; Simons et al., 2014).

Afin de répondre à cette question, nos travaux se sont concentrés sur l'impact de la démographie sur l'efficacité de la sélection purificatrice au sein de populations africaines de Pygmées et de non-Pygmées qui se différencient par leurs modes de subsistance (chasseurs-cueilleurs et agriculteurs) associés à des régimes démographiques opposés (réduction et augmentation des tailles de population chez les Pygmées et non-Pygmées respectivement). Pour cela, nous avons analysé les séquences exoniques de 300 individus chasseurs-cueilleurs Pygmées (Baka et BaTwa) et agriculteurs non-Pygmées (Nzebi, Bapunu et BaKiga) à l'ouest et à l'est de l'Afrique que nous avons comparé à 100 individus d'origine européenne (Quach et al 2016). En premier lieu, nous avons réévalué le modèle démographique de ces populations en estimant leurs temps de divergence, leurs variations de taille de population effective et leurs taux de migrations à partir de données de séquençage. Nous avons ensuite estimé la probabilité de fixation des allèles délétères dans ces populations et suivi par simulation l'évolution du fardeau de mutations délétères au cours de leurs histoires démographiques respectives. Enfin, nous avons mesuré empiriquement le fardeau de mutations délétères dans chacune des populations en considérant un modèle de dominance additif (semi-dominant) et récessif des mutations.

5.2 Article 1

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0496-4

The demographic history and mutational load of African hunter-gatherers and farmers

Marie Lopez[1,2,3,10], Athanasios Kousathanas[1,2,3,10*], Hélène Quach[1,2,3], Christine Harmant[1,2,3], Patrick Mouguiama-Daouda[4,5], Jean-Marie Hombert[5], Alain Froment[6], George H. Perry[7], Luis B. Barreiro[8], Paul Verdu[9], Etienne Patin[1,2,3] and Lluís Quintana-Murci[1,2,3*]

Understanding how deleterious genetic variation is distributed across human populations is of key importance in evolutionary biology and medical genetics. However, the impact of population size changes and gene flow on the corresponding mutational load remains a controversial topic. Here, we report high-coverage exomes from 300 rainforest hunter-gatherers and farmers of central Africa, whose distinct subsistence strategies are expected to have impacted their demographic pasts. Detailed demographic inference indicates that hunter-gatherers and farmers recently experienced population collapses and expansions, respectively, accompanied by increased gene flow. We show that the distribution of deleterious alleles across these populations is compatible with a similar efficacy of selection to remove deleterious variants with additive effects, and predict with simulations that their present-day additive mutation load is almost identical. For recessive mutations, although an increased load is predicted for hunter-gatherers, this increase has probably been partially counteracted by strong gene flow from expanding farmers. Collectively, our predicted and empirical observations suggest that the impact of the recent population decline of African hunter-gatherers on their mutation load has been modest and more restrained than would be expected under a fully recessive model of dominance.

Human populations have undergone radical changes in size over the past 100,000 years due to range expansions, bottlenecks and periods of rapid growth[1–3]. Such demographic fluctuations may have differently affected the efficacy of natural selection, relative to drift, to remove deleterious genetic variation from populations[4]. Genomic studies have indeed reported differences in the number, frequency and distribution of putatively deleterious variants across populations and it has been suggested that these differences result from their various demographic histories[5–13]. Understanding the potential importance of recent demographic events on how selection operates is thus of tremendous interest, as deleterious mutations may increase the risks of common disease[8,14,15] or may have contributed to past population extinctions, as suggested for Neanderthals[16].

The burden of deleterious mutations of a population has traditionally been quantified as the mutation load[17]; that is, the reduction in the average fitness of a population due to deleterious mutations compared with the theoretical optimal fitness, which depends only on the mutation rate and the model of dominance for a population at mutation-selection balance equilibrium[18]. However, demographic history can also alter the load for non-equilibrium populations; for example, during a prolonged bottleneck, a fraction of deleterious mutations may shift between being weakly selected to being effectively neutral and drift to fixation, thus permanently increasing the load[19,20]. Furthermore, under a recessive model of dominance, the load can fluctuate severely during demographic changes before reaching a new equilibrium[21,22].

Population genetic studies have examined the dynamics of the mutation load in the context of different, non-equilibrium human populations[7,22,23]. These studies have typically approximated the load at the present time using a variety of statistics, such as the number of deleterious alleles per individual, and have mostly contrasted African with non-African populations. The general consensus derived from theoretical work is that under an additive model of dominance only small differences in load are expected between populations, whereas under a recessive model expectations vary according to their specific demography[22]. However, the expectations of only a limited number of population demographic scenarios have been examined and tested using empirical data.

While the majority of human populations have undergone substantial recent growth associated with the advent of agriculture in the past 10,000 years[24,25], about 5% of human groups are expected to have maintained low population sizes because they have continued to subsist primarily by hunting and gathering[26]. Africa harbours the largest group of hunter-gatherers—the rainforest hunter-gatherers (historically referred to as 'pygmies')—who have traditionally lived in small, mobile groups scattered across the central African forest[27]. Conversely, the neighbouring agriculturalists are sedentary and descend from early farming communities that recently expanded across sub-Saharan Africa[28]. While there is increasing evidence to

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France. [2]Centre National de la Recherche Scientifique UMR 2000, Paris, France. [3]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. [4]Laboratoire de Langue, Culture et Cognition, Université Omar Bongo, Libreville, Gabon. [5]Centre National de la Recherche Scientifique UMR 5596, Dynamique du Langage, Université Lumière-Lyon 2, Lyon, France. [6]Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, Paris, France. [7]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, USA. [8]Université de Montréal, Centre de Recherche du Centre Hospitalier Universitaire Sainte-Justine, Montréal, Canada. [9]Centre National de la Recherche Scientifique UMR 7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Sorbonne Paris Cité, Paris, France. [10]These authors contributed equally: Marie Lopez and Athanasios Kousathanas. *e-mail: kousathanas2@gmail.com; quintana@pasteur.fr

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

suggest that these hunter-gatherer and farmer populations have experienced contractions and growth, respectively[29–33], important aspects of their demographic history remain unclear.

Here, we aimed to understand how the demographic history of African hunter-gatherer and farmer populations has affected the efficacy of selection to purge deleterious alleles (defined as the ratio of the fixation probability of new deleterious mutations relative to neutral mutations)[4] and shaped their past and present mutation load (defined as the average fitness reduction of a population due to deleterious mutations)[34]. The different subsistence strategies of these groups are expected to be associated with unique demographic regimes[29–33], thus providing an excellent model for exploring the temporal trajectory of mutation load across populations and examining the model of dominance that is most compatible with the observed patterns of deleterious variation.

Results

Population exome sequencing dataset. We analysed 100 Baka rainforest hunter-gatherers (wRHG) and 100 Nzebi and Bapunu farmers (wAGR) from Gabon and Cameroon in western central Africa and 50 BaTwa rainforest hunter-gatherers (eRHG) and 50 BaKiga farmers (eAGR) from Uganda in eastern central Africa (Fig. 1a and Supplementary Table 1). On the basis of genome-wide single nucleotide polymorphism (SNP) data, populations separated by mode of subsistence (RHG versus AGR), before RHG split into western and eastern groups[29,31,35] (Fig. 1b, Supplementary Note 1 and Supplementary Figs. 1 and 2), and presented no evidence of internal substructure (Supplementary Figs. 3 and 4). We performed whole-exome sequencing for a selection of 300 unrelated individuals at high coverage (mean depth 68×) and identified 406,270 quality-filtered variants, including 67,037 newly identified variants (Supplementary Table 1 and Supplementary Figs. 5 and 6). This dataset was supplemented with high-coverage exome sequences from 100 European-descent Belgians (EUR)[36], yielding a final dataset of 488,653 SNPs.

To obtain a broad view of the genetic diversity of RHG and AGR populations, we calculated summary statistics for synonymous variants (fourfold degenerate variants). Watterson's θ (θ_W) was higher in the wAGR and eAGR populations than in the RHG populations ($P < 10^{-3}$; Fig. 1c and Supplementary Table 2) due to a larger proportion of low-frequency variants, as demonstrated by the more negative Tajima's D values obtained ($P < 10^{-3}$ for both comparisons; Fig. 1e). However, wRHG had the highest pairwise nucleotide diversity θ_π ($P < 10^{-3}$ for all comparisons; Fig. 1d), suggesting a large effective population size (N_e). These results indicate that the African populations studied have similar levels of genetic diversity, regardless of their mode of subsistence, but their different allele frequency distributions suggest a contrasting demographic past.

Model-based inference of population divergence times, size changes and gene flow. Demographic parameters characterizing RHG and AGR populations were estimated by fitting models incorporating all the populations studied, including EUR, into non-cytosine-phosphate-guanine (non-CpG) synonymous pairwise (two-dimensional) site frequency spectra (SFS), using the coalescent-based composite likelihood approach fastsimcoal2 (Supplementary Note 2). We assumed unlinked sites (Supplementary Fig. 7), a mutation rate of 1.36×10^{-8} per site per generation and a generation time of 29 years (see Methods). All models assumed an early population size change for the ancestors of all populations, as previously proposed[8,37], followed by size changes coinciding with population splits, and an additional population size change for EUR (Supplementary Fig. 8). We formulated three branching models, each assuming that a different population (EUR, RHG or AGR) was the first to split off from the remaining groups (that is, EUR-first, RHG-first and AGR-first; Fig. 2). Furthermore, because admixture between wRHG and wAGR, eRHG and eAGR, and eAGR and EUR has been documented (Fig. 1b)[33,38,39], we estimated parameters by considering two epochs of continuous migration between population pairs, allowing for asymmetric gene flow (Supplementary Note 2).

The three branching models produced non-significant differences in likelihood (non-adjusted $P > 0.05$ for all comparisons; Supplementary Table 3) and presented an excellent fit to both observed marginal one-dimensional SFS and fixation index (F_{ST}) values (Supplementary Figs. 9 and 10). Importantly, the three models consistently provided similar estimates for key demographic

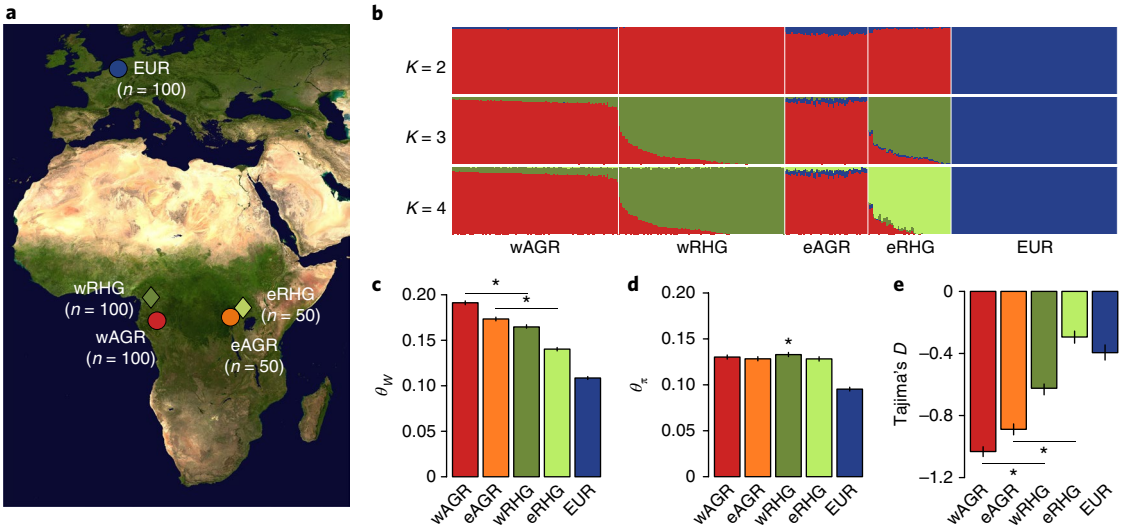


Fig. 1 | Genetic structure and diversity of African rainforest hunter-gatherers and farmers. a, Locations of the sampled populations. Map credit: NASA. **b**, Estimation of ancestry proportions with the clustering algorithm ADMIXTURE[77] using the SNP array data. Cross-validation values were the lowest at K (number of clusters) = 4 (Supplementary Fig. 2). **c**, Watterson's estimator θ_W. **d**, Pairwise nucleotide diversity θ_π. **e**, Tajima's D. In **c**–**e**, all neutrality statistics were calculated with exome sequencing data at fourfold degenerate synonymous sites. Means, 95% confidence intervals and significance were obtained by bootstrapping by site 1,000 times. Significance was assessed between wAGR and wRHG and between eAGR and eRHG in **c** and **e**, and for all pairwise comparisons involving wRHG in **d**. Non-adjusted P values are shown: $^*P < 10^{-3}$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

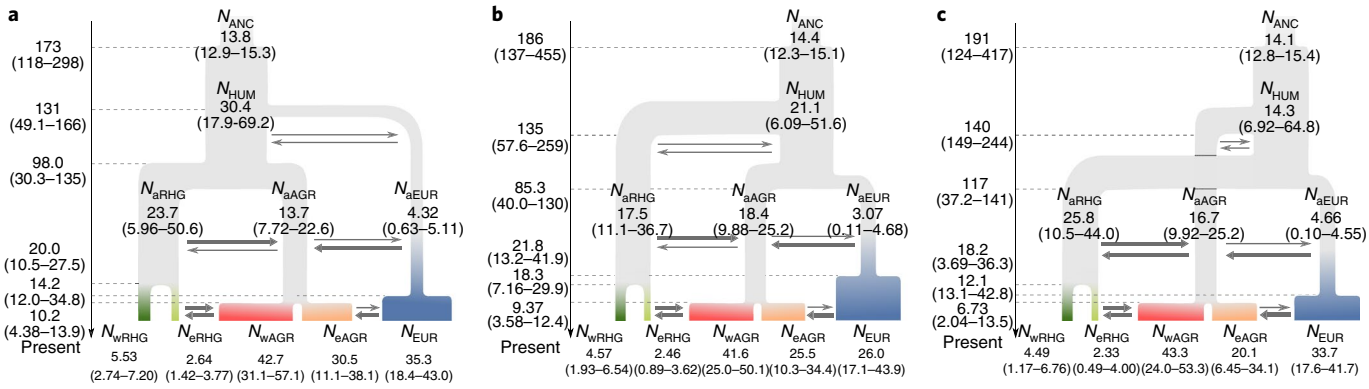


Fig. 2 | Inferred demographic models of the studied populations. a, EUR-first branching model, in which ancestors of EUR (aEUR) diverged from African populations before the divergence of the ancestors of RHG (aRHG) and AGR (aAGR). **b**, RHG-first branching model, in which aRHG were the first to diverge from the other groups. **c**, AGR-first branching model, in which aAGR were the first to diverge from the other groups. We assumed an ancient change in the size of the ancestral population of all humans (ANC). We assumed that each subsequent divergence of populations was followed by an instantaneous change in the effective population size (N_e). We also assumed that there were two epochs of migration between the following population pairs: wAGR/aAGR and wRHG/aRHG, eAGR/aAGR and eRHG/aRHG, and EUR and eAGR/aAGR. The figure labels correspond to the parameters of the model estimated by maximum likelihood and the 95% confidence intervals assessed by bootstrapping by site 100 times (Supplementary Table 4). Vertical arrow corresponds to the direction of time, from past to present, with divergence times given on the left and expressed in thousand years ago (ka). Effective population sizes (N) are given within the diagram and expressed in thousands of individuals. Bold horizontal arrows indicate an estimated parameter for the effective strength of migration $2Nm > 1$, while thin horizontal arrows indicate $2Nm \leq 1$.

parameters (Fig. 2, Supplementary Table 4 and Supplementary Note 3). Specifically, we obtained significant support, based on ratios of ancestral to current N_e, for a recent bottleneck in both wRHG and eRHG populations, and for a recent expansion of wAGR and EUR populations (that is, $N_{aRHG}/N_{RHG} > 1$, whereas N_{aAGR}/N_{wAGR} and $N_{aEUR}/N_{EUR} < 1$; $P < 0.05$; Supplementary Table 5). The estimated parameters for the effective strength of migration ($2Nm$) were also mostly similar between the branching models (Supplementary Tables 4 and 6). We inferred that the average migration between the ancestors of RHG and AGR was limited, but significantly higher than zero ($P < 0.05$). However, over the past 10,000–20,000 years, migration between RHG and AGR increased markedly (for all models, $2Nm > 17$ for western groups and $2Nm > 8$ for eastern groups; $P < 0.05$) (Fig. 2, Supplementary Table 4 and Supplementary Note 3).

Differences in the efficacy of purifying selection across populations. We explored whether the different demographic histories of RHG and AGR affected the efficacy with which deleterious alleles were purged by natural selection. First, we assessed the deleteriousness of variants using a method based on sequence conservation (genomic evolutionary rate profiling-rejected substitution (GERP RS))[40], which avoids reference-bias effects[7]. We compared, across populations, the SFS of non-synonymous-derived mutations assigned to different GERP RS score classes (Fig. 3a) and their proportion in different frequency bins (Fig. 3b). We observed that singletons in EUR are enriched in mutations with GERP RS > 4 (that is, predicted slightly-to-moderately deleterious mutations) relative to African populations, while singletons in RHG are slightly depleted in mutations with GERP RS > 4 relative to AGR (Fig. 3b).

To test whether the observed population differences in deleterious SFS result from a difference in the efficacy of purifying selection among populations, we used model-based approaches to infer the distribution of fitness effects (DFE) of new non-synonymous mutations, as implemented in ∂a∂i/Fit∂a∂i and *DFE-α*[41–43] (Supplementary Table 7). We explicitly incorporated a three-epoch model of non-equilibrium demography and a gamma distribution of deleterious mutations with additive effects (see Methods and Supplementary Table 8). Because the results obtained with ∂a∂i/Fit∂a∂i and *DFE-α* were very similar (Supplementary Table 7) and the fit of the demographic and selection models to the data was excellent for ∂a∂i/Fit∂a∂i (Supplementary Figs. 11 and 12), we present the results using ∂a∂i/Fit∂a∂i only.

First, we summarized the inferred DFEs by computing the proportion of mutations assigned into four selection strength ($N_e s$) ranges (0–1, 1–10, 10–100 and >100, corresponding to neutral, weakly, moderately and strongly deleterious mutations, respectively). The inferred DFE histogram showed that the fraction of mutations in the neutral range of $N_e s \sim 0$–1 was almost identical across populations, whereas a larger fraction of new mutations was expected to be strongly deleterious to lethal ($N_e s > 100$) in Africans than Europeans (Fig. 4a). We then calculated the ratio of the fixation probability (u) for a new deleterious mutation versus a neutral mutation (u_{del}/u_{neu}; Supplementary Notes 4 and 5) to quantify the relative strength of selection versus drift (that is, a small u_{del}/u_{neu} ratio means greater efficacy of selection)[4]. We found that the inferred u_{del}/u_{neu} was almost identical across populations (Fig. 4b). More generally, the strong leptokurtic shape of the inferred DFEs suggests that differences between populations in u_{del}/u_{neu} are limited given their modest differences in average N_e (see average effective population size over the three-epoch demographic history (N_w) estimates; Supplementary Note 4 and Supplementary Table 7), even when assuming the same underlying distribution for s across populations (Fig. 4c).

Together, our results show that the limited population differences in the allele frequency distribution of deleterious mutations are compatible with a similar efficacy of selection for purging new additive deleterious mutations across populations.

Expected temporal trajectories of the mutation load. We performed simulations to explore how mutation load has changed through time as a function of the recent demography of RHG, AGR and EUR populations and for the additive (dominance coefficient, $h = 0.5$) and recessive ($h = 0$) models of dominance. Our simulations suggested that under an additive model the load is fairly insensitive to the demographic changes experienced through time by the studied populations (Fig. 5a,b and Supplementary Figs. 13 and 14). When stratifying mutations in different ranges of selection coefficients, we nevertheless observed a slight increase in the additive

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

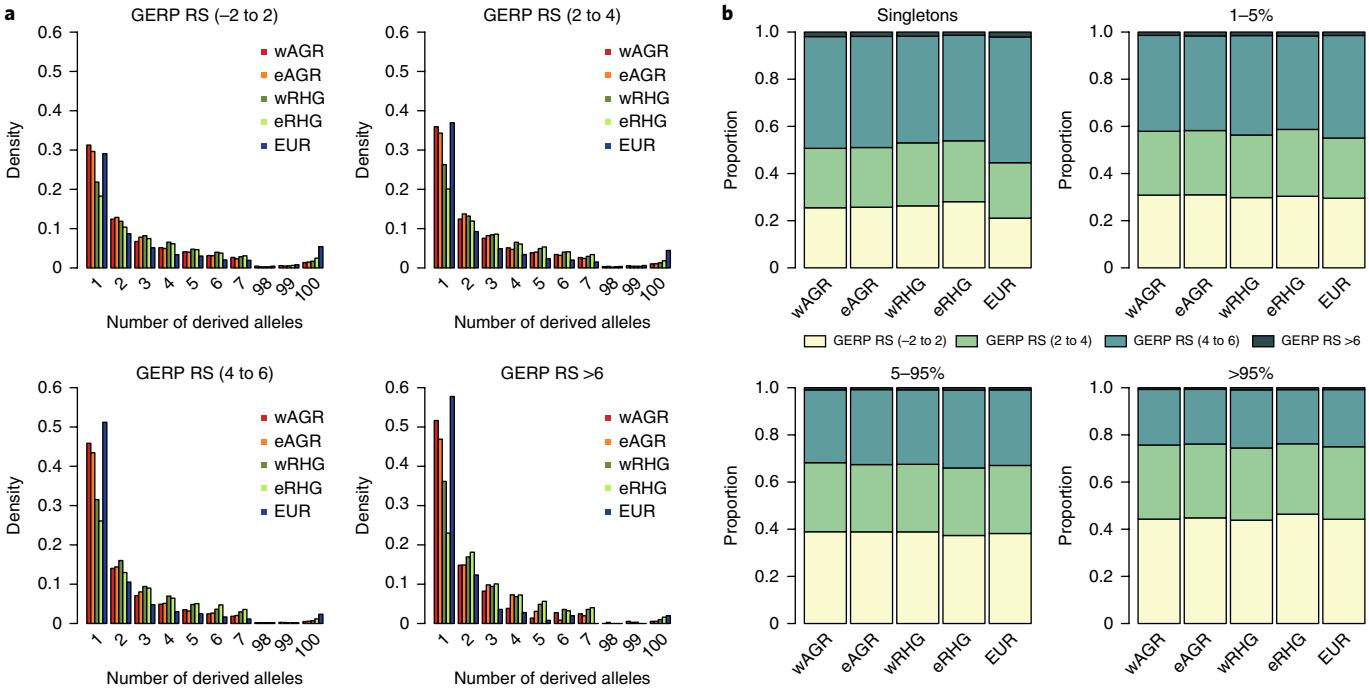


Fig. 3 | Population frequencies and predicted selective effects of non-synonymous mutations in African rainforest hunter-gatherers and farmers, compared with Europeans. a, Derived allele frequency spectra of non-synonymous variants segregating in each studied population for different GERP RS score categories. The proportion of singletons increases with increasing GERP RS scores in all populations. **b**, Proportion of derived non-synonymous variants assigned to different GERP RS score categories that segregate at different allele frequencies (singletons, 1–5%, 5–95% and >95%) in the populations studied. In **a** and **b**, populations were downsampled to the same sample size of 50 individuals to allow visual comparison.

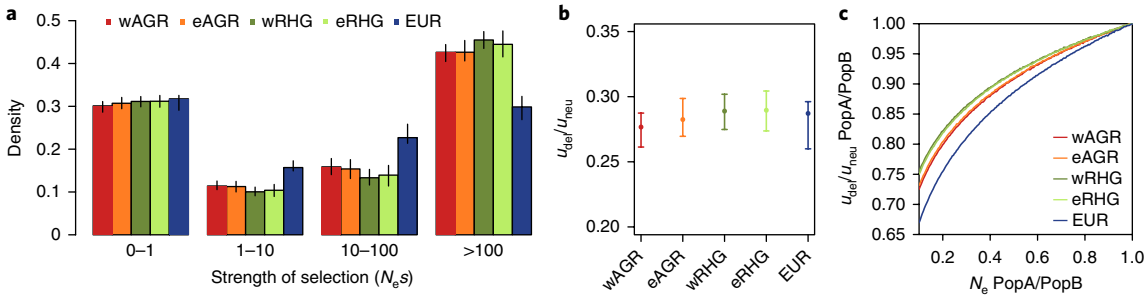


Fig. 4 | DFE of new non-synonymous mutations. a, Fraction of new mutations in different bins of selection strength ($N_e s = 0$–1, 1–10, 10–100 or >100), assuming a three-epoch demography fitted with ∂a∂i to the unfolded synonymous SFS and a gamma distribution model for the DFE fitted with Fit∂a∂i to the unfolded non-synonymous SFS for each population separately. **b**, Ratio of the fixation probability of a new deleterious mutation relative to a neutral mutation (u_{del}/u_{neu}) inferred for each population. **c**, Relative u_{del}/u_{neu} for a given ratio in N_e between two populations (PopA and PopB), assuming the DFE inferred for wAGR, eAGR, wRHG, eRHG and EUR. We used non-CpG sites, and 95% confidence intervals, represented by the error bars in **a** and **b**, were calculated by bootstrapping by site 100 times.

mutation load in EUR, contributed by weakly-to-moderately deleterious mutations ($10^{-4} < s < 10^{-3}$; Supplementary Fig. 15).

By contrast, the impact of demography on mutation load is more severe under a recessive model of dominance (Fig. 5c and Supplementary Figs. 13 and 14), although its extent depends on the selection coefficients of the mutations considered (Supplementary Fig. 15). Specifically, after a bottleneck, strongly deleterious-to-lethal recessive mutations can be found at the homozygous state, leading to a higher yet transient mean mutation load, as these homozygous mutations are eliminated at a higher rate per generation. Indeed, our simulations predicted that the duration of the bottleneck for EUR, together with their recent expansion, was sufficient to overcome the transient surge in load contributed by strongly deleterious-to-lethal recessive mutations (Supplementary Fig. 15). Conversely, the recent collapse experienced by RHG, which has not been accompanied by recovery, results in a present recessive load that is expected to be higher than in AGR and EUR (Fig. 5c and Supplementary Figs. 13 and 14). However, the magnitude of the increase in the load of RHG compared with EUR would depend on the fraction of new deleterious recessive mutations that are strongly deleterious to lethal relative to the fraction of weakly-to-moderately deleterious mutations (Supplementary Fig. 15c). Weakly-to-moderately deleterious mutations contribute more to the mutation load of EUR than to the mutation load of RHG, due to the prolonged bottleneck of EUR that allowed these mutations to reach higher frequencies through drift (Supplementary Fig. 15c).

Finally, we investigated whether gene flow may have attenuated population differences in mutation load by performing simulations

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

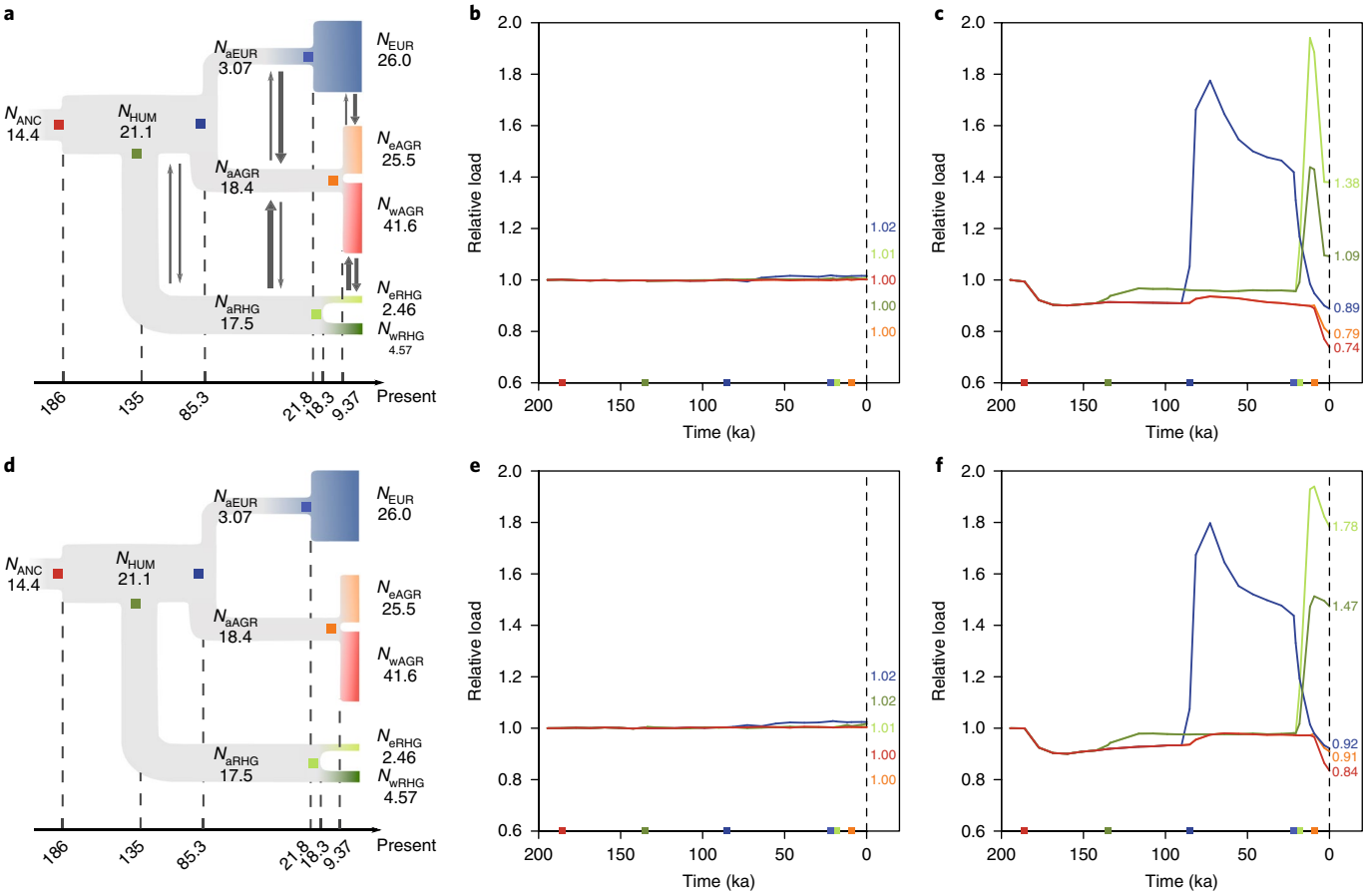


Fig. 5 | Trajectory of mutation load through time obtained with simulations. a–f, The mutation load (*L*) relative to the ancestral population at equilibrium was calculated as a function of time expressed in thousands of years ago (ka) during the recent history of wAGR (dark red), eAGR (orange), wRHG (dark green), eRHG (light green) and EUR (blue), assuming the full RHG-first demographic model (**a**) for additive ($h=0.5$; **b**) and recessive ($h=0$; **c**) mutations, and assuming the RHG-first demographic model without migration (**d**) for additive ($h=0.5$; **e**) and recessive ($h=0$; **f**) mutations. For **a** and **d**, the horizontal arrows correspond to the direction of time from past to present. For **b**, **c**, **e** and **f**, dashed vertical lines indicate the present time and coloured numbers indicate the relative mutation load at the present time. Coloured boxes indicate events in the demographic history of the populations also depicted in **a** and **d**. We assumed the DFE inferred for wAGR and the RHG-first demographic model. Similar trajectories of mutation load through time were obtained with simulations assuming the EUR-first demographic model (Supplementary Fig. 13) and the DFE inferred for EUR (Supplementary Fig. 14).

with migration set to zero (Fig. 5d and Supplementary Figs. 13–15). Interestingly, removing migration had no impact on the additive mutation load, but increased the recessive mutation load, particularly for RHG (Fig. 5e,f and Supplementary Figs. 13–15). The effect of gene flow on reducing the recessive load of RHG was observed for all ranges of selection coefficients (Supplementary Fig. 15).

Exploring present-day mutation load and the dominance of deleterious variation. Because our simulations indicated that population differences in mutation load at the present time are dependent on the dominance model assumed, next, we sought to gain insight into the most likely model of dominance of deleterious mutations segregating in our dataset. We thus compared observed summary statistics for different classes of sites with expectations of these statistics generated with simulations under different dominance models. Specifically, we compared the per-individual number of derived alleles ($N_{alleles}$) and homozygous genotypes (N_{hom}) between populations—two statistics previously used to quantify mutation load under additive and recessive dominance models[7,9,22] and reported to be informative for the dominance of deleterious mutations[44]—for variants stratified into different classes of GERP RS scores. Of note, $N_{alleles}$ and N_{hom} were not affected by sequencing quality (Supplementary Fig. 16) and were examined for all sites and non-CpG sites separately.

Under an additive model, non-synonymous $N_{alleles}$ is expected to be similar across populations, while under a fully recessive model non-synonymous $N_{alleles}$ is expected to be lower in EUR than in Africans—a pattern that is further accentuated with increasing selective coefficients (Fig. 6a). In agreement with expectations under the additive model, observed population differences in non-synonymous $N_{alleles}$ did not exceed 2% and were not significant for any population comparison or GERP RS categories (Fig. 6b). Similar results were obtained using independent annotation software to predict the fitness effects of mutations[45], or when considering variants in the genes responsible for dominant and recessive diseases separately (Supplementary Figs. 17 and 18 and Supplementary Notes 6 and 7). The ratio of $N_{alleles}$ was not significantly different from 1 and did not exceed 8% for loss-of-function (LOF) mutations, which are expected to be enriched in deleterious variants (Supplementary Fig. 19).

In contrast with $N_{alleles}$, expectations for N_{hom} were generally very similar under different models of dominance and were not distinguishable from the observed data (Supplementary Fig. 20). Observed N_{hom} was more than 20% higher in Europeans than Africans, but much smaller differences in N_{hom} were observed between RHG and AGR: non-synonymous N_{hom} was about 2–4% lower in wRHG than AGR and ~1–4% higher in eRHG than other African populations (Supplementary Fig. 20). Very similar results were obtained for

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

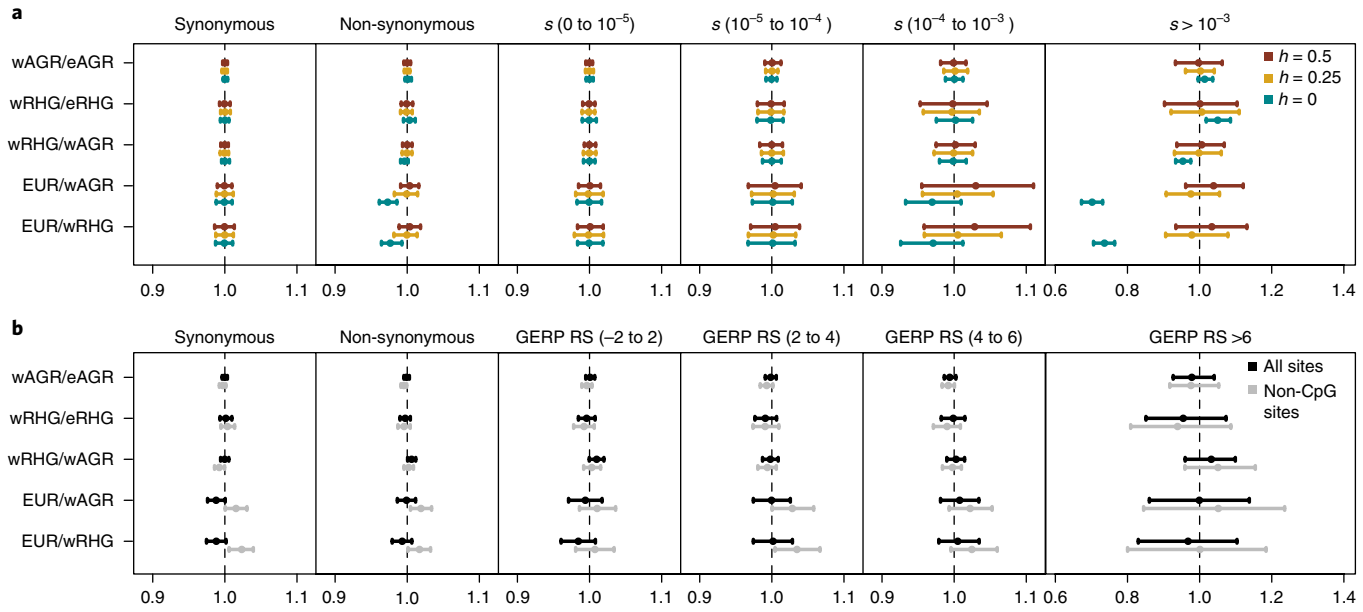


Fig. 6 | Comparison of observed $N_{alleles}$ with simulated predictions between the studied populations. a, Between-population ratios of the mean per-individual number of derived alleles ($N_{alleles}$) that were simulated under additive ($h = 0.5$, red), partially recessive ($h = 0.25$, yellow) and fully recessive ($h = 0$, blue) models of dominance for synonymous and non-synonymous mutations stratified in bins of increasing selection coefficient, s. For these simulations, we assumed the DFE inferred for wAGR and the RHG-first demographic model. **b**, Between-population ratios of the mean per-individual number of derived alleles ($N_{alleles}$) for observed synonymous and non-synonymous mutations involving all sites (black) and non-CpG sites (grey) and stratified in several classes of GERP RS scores reflecting the severity of mutational effects ('neutral' (−2 to 2), 'moderate' (2 to 4), 'large' (4 to 6) and 'extreme' (>6)). For the simulated ratios (**a**), the error bars represent the 0.025 and 0.975 quantiles of ratios across 100 simulations. For the observed ratios (**b**), the error bars represent the 95% confidence intervals that were calculated by dividing the exome data into 1,000 blocks and carrying out bootstrap resampling of blocks 1,000 times. The non-adjusted P value obtained by testing for observed ratios different from 1 was never equal to or lower than 10^{-3}, which was our required threshold for significance due to the large number of tests performed.

$N_{alleles}$ and N_{hom} when considering deleterious non-coding variants present in our extended exome dataset (Supplementary Fig. 21).

We also searched for an effect of gene flow between AGR and RHG on mutation load, as expected from simulations under the recessive model, by examining $N_{alleles}$ and N_{hom} as a function of AGR ancestry in RHG individuals. We attempted to enrich for recessive mutations by focusing on variants located in genes previously associated with recessive diseases (Supplementary Note 7). We did not detect a significant correlation between the non-synonymous to synonymous ratio of $N_{alleles}$ and N_{hom} and AGR ancestry (Supplementary Figs. 22–24), possibly reflecting the difficulty in isolating recessive deleterious mutations. Finally, although the studied populations displayed variable levels of parental relatedness, we found no evidence for population differences in the distribution of deleterious variants across runs of homozygosity (ROH) (Supplementary Note 8 and Supplementary Figs. 25–27).

Collectively, the observed limited differences in $N_{alleles}$ and N_{hom} across populations suggest that the majority of deleterious alleles segregating in RHG, AGR and EUR present a model of dominance that is not fully recessive, resulting in no detectable differences in mutation load among these human groups.

Discussion

Several demographic models have been previously proposed to explain the history of rainforest hunter-gatherers and farmers from central Africa[29–33,35,46]. Our estimation that the ancestors of RHG and AGR diverged 98 to 140 thousand years ago (ka) is compatible with previous estimations[29–31,35] once the models are recalibrated using recent estimates of mutation rate and generation time (Supplementary Table 9). These results support a deep divergence time between the ancestors of contemporary rainforest hunter-gatherers and agricultural groups. Furthermore, the similar likelihoods obtained for the three branching models suggest that the ancestors of RHG, AGR and EUR diverged from each other at around the same time (~85–140 ka). This coincides with a period of major climate change (for example, megadroughts dated between 75 and 135 ka)[47], which probably promoted population isolation and ancient structure on the African continent.

Our demographic inference indicates that the effective population size of RHG was at least as large as that of the ancestors of AGR for most of their evolutionary past. This suggests that RHG groups were more demographically successful, or more interconnected by gene flow, in the distant past than in the most recent millennia, as has also been suggested for the KhoeSan hunter-gatherers of southern Africa[48,49]. More recently, RHG have experienced major size reductions while AGR have experienced mild expansions, accompanied by a marked increase in gene flow between them. Overall, our results extend earlier findings by providing parameter estimates for more complex demographic models than previously investigated[29–32], thus characterizing the history of African RHG and AGR over the past 150,000 years more realistically.

The impact of population growth and decline on the efficacy of purifying selection and the burden of deleterious mutations in humans has been the subject of intense research over the past few years[5–13,22,23,25]. Our estimates for the ratio of the average fixation probability of new deleterious mutations over neutral mutations (u_{del}/u_{neu}) suggest that the efficacy of selection for removing deleterious mutations has been very similar across populations, when assuming an additive model of dominance. Although we analysed non-synonymous variants for the estimation of u_{del}/u_{neu}, we would expect our results to generalize to deleterious variants in non-coding DNA if their underlying DFE is as leptokurtic as that of

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

non-synonymous sites. Furthermore, our simulations of the temporal trajectory of mutational load predict almost identical trajectories for all populations under the additive model.

Under the recessive model of dominance, the trajectory of mutational load indicates that the population decline experienced by Europeans and African hunter-gatherers led to a surge in load. Although the European bottleneck was sufficiently long-lasting to overcome the transient surge in load contributed by strongly deleterious mutations, its long duration led to an increase in load contributed by weakly deleterious mutations. In contrast, the more recent collapse of African hunter-gatherers has led to an increased load at present, relative to African farmers, which is mainly contributed by strongly deleterious mutations. This prediction would imply an increased genetic risk for diseases for African hunter-gatherers (Supplementary Note 9 and Supplementary Table 10) and could theoretically further exacerbate their population size decline. Yet, as we demonstrate with simulations, increased gene flow from neighbouring farmers is possibly counteracting the effect of the recent collapse of hunter-gatherers on their present and future mutation load.

We also examined whether an additive or recessive model of dominance was more compatible with empirical data by comparing the number of deleterious alleles per individual ($N_{alleles}$) for observed versus simulated data[44]. A fully recessive model predicts patterns of $N_{alleles}$ that are incompatible with our observed data for deleterious coding and non-coding variants, especially for comparisons involving Europeans. This clearly suggests that a partially-recessive-to-additive model is more realistic for our dataset. In the light of this, we expect differences in load between hunter-gatherer and farmer populations to be less pronounced than predicted by the fully recessive model. However, a quantitative conclusion for the mutation load of African hunter-gatherers at the present time would strongly depend on the precise estimation of the average or even the full distribution of dominance coefficients, for which we have little a priori knowledge[21,50].

In conclusion, our study highlights the unique and under-studied demography of a population group that had a historically large effective population size, but has recently experienced a strong decline. Because such a demographic history probably characterizes many other small-sized populations, we expect our conclusions on mutation load to apply to other African hunter-gatherers, as well as to non-African groups who have experienced recent bottlenecks. These populations are predicted to have an increased recessive load at present times, particularly if they are isolated and have not experienced substantial gene flow from expanding populations. By jointly modelling the genomic diversity of ancient and modern populations with contrasting demographic histories and various levels of admixture, future studies on the dynamics of the mutation load will provide new insights into the probable dominance model of deleterious mutations and their impact on population differences in disease risk.

Methods

Population and individual selection. In total, we included 317 individuals from the AGR and RHG populations of western and eastern central Africa and 101 individuals of European ancestry[36] who were chosen based on previous analyses of SNP array data[33,36,51–53] (Supplementary Note 1 and Supplementary Table 1). Informed consent was obtained from all participants, which was overseen by the institutional review board of Institut Pasteur, France (2011-54/IRB/7), the Comité National d'Ethique du Gabon (0016/2016/SG/CNE), the University of Chicago (IRB 16986 A) and Makerere University, Uganda (IRB 2009-137).

Exome sequencing and quality controls. We sequenced the exome of 314 African samples and processed these data together with 101 European individuals[36]. All samples were sequenced with the Nextera Rapid Capture Expanded Exome Kit, which delivers 62 megabases of genomic content per individual, including exons, untranslated regions and microRNAs. Using the GATK Best Practices recommendations[54], read pairs were first mapped onto the human reference genome (GRCh37) with Burrows–Wheeler Aligner version 0.7.7 (ref. [55]) and reads duplicating the start position of another read were marked as duplicates with Picard Tools version 1.94 (http://broadinstitute.github.io/picard/). We used GATK version 3.5 (ref. [56]) for base quality score recalibration ('BaseRecalibrator'), 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'.

As criteria to remove low-quality samples, we required at least 40× mean depth of coverage (3 excluded samples), 85% of the positions in the BAM file to be covered at 5× minimum (8 excluded samples) and a total genotype missingness lower than 5% (1 excluded sample) (Supplementary Fig. 6). In addition, we checked for unexpectedly high or low heterozygosity values suggesting high levels of inbreeding or DNA contamination, and excluded 3 additional individuals presenting heterozygosity levels 4 s.d. higher than their population average (Supplementary Fig. 6). We thus retained exome data from 400 individuals, with an average depth of coverage of 68×, ranging from 40× to 168×, and an individual breadth of coverage above 5 × for 93% of the exome target on average, ranging from 85 to 97%. Finally, we removed indels and discarded from the 768,143 SNPs obtained those that: (1) were not biallelic, (2) presented missingness above 5%, (3) were monomorphic in our sample, (4) were located on the sex chromosomes, (5) presented a Hardy–Weinberg test P value $< 10^{-3}$ in at least one of the populations and (6) had an unknown ancestral state using the 6-EPO multi-alignment from Ensembl Compara version 59, which was used to obtain the ancestral and derived allele of each variant. The application of these quality-control filters resulted in a final dataset of 488,653 SNPs (406,270 SNPs segregating in African populations and 82,383 European-specific SNPs). We intersected the variants with SNP database (dbSNP) build 149 and identified 67,037 previously unreported SNPs segregating in African populations.

We also examined whether variation in individual missingness affected the per-individual number of homozygotes (N_{hom}) or alleles ($N_{alleles} = N_{het} + 2 \times N_{hom}$) and found a significant correlation between individual missingness and these parameters. We thus allowed no missingness and required a depth of coverage of at least 5× for each variant across all individuals, leading to a filtered dataset of 382,786 SNPs. To test the influence of depth of coverage on the detection of SNPs, we explored the correlation between the per-individual mean and variance of coverage and the number of SNPs identified per individual and observed no significant correlation (Supplementary Fig. 16). This filtered dataset was then employed for all analyses that used per-individual genotypes and allele counts.

Functional annotation of variants. To annotate synonymous and non-synonymous variants in our dataset, we first obtained a bed file with genomic coordinates of exons for each canonical transcript from the University of California, Santa Cruz genome browser (http://genome.ucsc.edu/) with the 'Table' browser feature track from the 'KnownCanonical' University of California, Santa Cruz genes table. We then used a custom script to parse the concatenated exons of each transcript, codon by codon, and annotate sites within each codon as zerofold or fourfold degenerate according to the genetic code. The annotation of LOF variants was performed using the Ensembl Variant Effect Predictor (VEP version 84). Only stop-gained and splice-disrupting variants were considered. Frameshift variants were not detected because of preliminary filters removing indels. We applied the Loss-of-Function Transcript Effect Estimator plugin (LOFTEE; available at https://github.com/konradjk/loftee) to attribute high or low confidence to the LOF variants. With LOFTEE, we filtered out LOF mutations known to be false positives, such as variants near the end of transcripts and in non-canonical splice sites, and kept only functional annotations based on the transcript having the highest confidence label.

Additionally, to assess the biological impact of each variant, we used the conservation-based scores computed by the algorithm GERP on an alignment of 35 mammals[40]. Positive GERP RS scores indicate a high degree of conservation and, therefore, a high probability for variants at conserved sites to be deleterious. We stratified non-synonymous variants in 4 different classes of GERP RS scores, −2 to 2, 2 to 4, 4 to 6 and >6, which have previously been used to assign variants as neutral, weakly, moderately and strongly deleterious, respectively[7]. We also assessed variant deleteriousness using independent functional annotation software that estimates 'fitness consequences' (fitCons) scores by integrating information from both evolutionary data and cell-type-specific functional genomic data[45] (Supplementary Note 6). Finally, we identified variants located in genes associated with known dominant or recessive genetic diseases using the Online Mendelian Inheritance in Man catalogue[57] (Supplementary Note 7).

Genetic diversity of hunter-gatherer and agriculturalist populations. We computed several summary statistics on synonymous sites. Watterson expected nucleotide diversity (θ_W), pairwise nucleotide diversity (θ_π) and Tajima's D were calculated based on SFS[58] with custom R scripts. Confidence intervals and P values for comparisons of these summary statistics between populations were computed by bootstrapping 1,000 times and resampling sites with replacement.

Linkage disequilibrium between sites. We calculated the levels of linkage disequilibrium in wAGR, wRHG and EUR (each with $n = 100$) using PLINK[59],

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

between all possible pairs of SNPs in 1-megabase-pair windows sliding along the genome. Only SNPs with a minor allele frequency >5% were considered. We used the square correlation coefficient between SNP pairs (r^2). For each SNP, we retrieved its maximum r^2 value with other SNPs located in the same gene and its maximum r^2 value with SNPs located in the surrounding genes and contrasted both distributions. Linkage disequilibrium decay with physical distance was assessed in the same population groups by calculating the average r^2 value of SNP pairs separated by a given distance, binned into 2,000 intervals. We discarded bins containing fewer than 50 SNP pairs.

SFS. Our demographic and selection inferences were based on fitting models to unfolded and folded SFS data. To obtain the SFS for non-synonymous and synonymous site classes, we created a variant-call format file containing variant and invariant sites and kept loci with at least 5× coverage, and allowed no missingness across individuals. We then used PGDspider[60] to transform our variant-call format (VCF) files to arlequin project (ARP) files, and used Arlequin version 2.5 (ref. [61]) to calculate one-dimensional SFS per population, as well as pairwise two-dimensional SFS.

Demographic inference. To estimate parameters of demographic models, we used the programme fastsimcoal2 version 2.5.2.21 (http://cmpg.unibe.ch/software/fastsimcoal2/)[62]. fastsimcoal2 performs coalescent simulations to approximate the likelihood of the data given a certain demographic model and specific parameter values. Maximization of the likelihood is achieved through several cycles of an expectation maximization algorithm. Therefore, it is critical to perform several simulations to approximate with high precision the likelihood, enough cycles of the expectation maximization algorithm to ensure the maximum was reached and several independent replicate estimations to ensure the global maximum likelihood was found. For all our point estimates, we performed 500,000 simulations, 30 cycles of the expectation maximization and 100 replicate runs from different random starting values. We recorded the maximum likelihood parameter estimates that were obtained across replicate runs. For calculation of confidence intervals, we fitted the demographic models to resampled SFS data obtained by bootstrapping 100 times by site using Arlequin[61]. Parameter inference for each bootstrap replicate was achieved by performing 500,000 simulations, 20 cycles of the expectation maximization and 20 replicate runs from different starting values.

We used non-CpG fourfold degenerate synonymous sites from 21,782 genes for our demographic inferences, and a total of 2,383,014 invariant and 223,356 variant sites passed quality filters. We used these data to generate one-dimensional folded SFS for each population and two-dimensional folded SFS for each pair of populations with Arlequin[61]. We fitted two types of demographic models to the SFS data, assuming a mutation rate of 1.36×10^{-8} per site per generation (that is, the mutation rate inferred for non-CpG sites)[63] and a generation time of 29 years[64]. First, three-epoch models of demographic change were fitted to the one-dimensional SFS of each population to obtain rough approximations of the size changes that each population had experienced. Second, we fitted complex models including splits, gene flow and size changes to the pairwise SFS of all the sampled populations. For these computations, we assumed that pairwise SFS are independent and, therefore, fastsimcoal2 computed a composite likelihood. The maximum likelihood method employed by fastsimcoal2 assumes that sites are unlinked; therefore, it calculates the full likelihood of the data by multiplying the likelihood for each site. Since we used only a few SNPs per gene and, assuming that on average there is no substantial linkage between genes (Supplementary Fig. 7), we expect that the assumption of fastsimcoal2 that sites are unlinked is reasonable for our dataset.

We used fourfold degenerate synonymous sites for demographic inference because they are, among all sites in our data, the least likely site class to be directly under natural selection. We expect, however, that linked selection (that is, background or positive) on the tightly linked non-synonymous sites might have impacted the diversity and shape of the synonymous SFS and, particularly, the variance of these statistics among sites[65,66]. Using the average folded SFS across sites for demographic inference, and not variance statistics, we reduced the potential impact of linked selection on our inference. Moreover, the effects of background selection on introducing bias for inference of population size changes are most severe when the strength of linked selection is weak to moderate ($N_e s = 2$–20)[65]. The strongly leptokurtic DFE expected for non-synonymous sites would predict that the proportion of non-synonymous mutations in this $N_e s$ range is rather small (10–20%)[67,68]; thus, we would not expect a substantial bias on the population size change estimates. For migration rates, strong background selection can improve estimation[65]. Furthermore, in humans, who have a large recombining genome, we would expect positive selection to impact demographic inference only if it is highly pervasive and mostly occurring through de novo mutations[66]—a scenario that is highly unrealistic[69]. Finally, we expect that the confidence intervals on the demographic parameters, which were computed by bootstrapping by site, should reflect some of the uncertainty over the parameter estimates introduced by background selection or selective sweeps.

DFEs of new mutations. We used methods implemented in ∂a∂i/Fit∂a∂i[41,42] and *DFE-α* version 2.15 (ref. [43]) to infer the DFE of new non-synonymous mutations. Both methods fit a demographic model to a class of sites that is assumed to be neutral and, conditional to the demographic model inferred, fit a gamma DFE model to the SFS of the focal class of sites. We used unfolded SFS for the analysis of ∂a∂i/Fit∂a∂I, accounting for ancestral misspecification, and folded SFS for the analysis of *DFE-α*. DFE estimation with *DFE-α* and analogous methods has been shown to be generally robust to linked selection when using SFS of neutral and focal classes that are interdigitated; that is, synonymous and non-synonymous sites[70,71]. Therefore, we used the synonymous and non-synonymous SFS as neutral and focal classes, respectively. We fit a three-epoch demographic model to synonymous SFS per population (Supplementary Table 7), yielding broadly consistent results with those generated by fastsimcoal2 based on one-dimensional SFS (Supplementary Note 2 and Supplementary Table 8). Both methods infer the mean (E(s)) and shape (β) of a gamma distribution DFE model fit on the non-synonymous SFS, accounting for demography. We assumed that new mutations reduce the fitness of the heterozygotes by $s/2$ and the homozygotes by s. Because ∂a∂i/Fit∂a∂i assumes that the fitness of homozygotes is reduced by $2s$, we multiplied ∂a∂i/Fit∂a∂i estimates for E(s) by 2. We calculated a weighted N_e over the inferred demographic changes through time (Supplementary Note 4) and interpolated the proportion of mutations that are assigned to four $N_e s$ ranges (0–1, 1–10, 10–100 and >100) corresponding to neutral, weakly selected, strongly selected and lethal mutations, respectively. We computed the average fixation probability of a new mutation (u) by integrating over the DFE inferred for each population separately and weighting by N_w inferred by ∂a∂i (Supplementary Note 4). We computed the fixation probability of a new deleterious mutation (u_{del}) and calculated the ratio of u_{del} over the fixation probability of a neutral mutation (u_{neu}) as a way to quantify the relative strength of selection versus drift at removing deleterious mutations (Supplementary Note 5). The estimation u_{del}/u_{neu} has been shown to be rather robust to misspecification of the DFE model when assuming a gamma distribution[70]. We calculated confidence intervals for estimated parameters by bootstrapping by site 100 times.

Genetic and mutation load. Genetic load is defined as the difference in fitness between the average genotype in the population and the genotype with maximum fitness, $L = \frac{w_{max} - \overline{w}}{w_{max}}$. We define maximum fitness as equal to 1; thus, genetic load is $L = 1 - \overline{w}$. Here, we measured the genetic load due to deleterious mutations; that is, the mutation load. We assumed that each mutation contributes to an average population fitness reduction of $l = s(2hq + (1-2h)q^2)$, where q is the derived frequency of the mutation, s is the selection coefficient against the mutation and h is the dominance coefficient. We assumed that effects across loci combine multiplicatively; thus, the total mutation load can be obtained by multiplying the reduction in fitness contributed by each locus ($1-l$). This product can be well approximated[34] by the exponential of the sum of the loads across m loci: $L = 1 - \exp\left(\sum_{i=1}^{m} l_i\right)$.

Simulations of trajectories of the mutation load. We performed forward simulations with the software SLiM version 2.2.1 (ref. [72]). We simulated all five populations jointly under the best-fitting demographic model inferred with fastsimcoal2 (RHG-first), but examined how the results would change under another branching model (EUR-first). We assumed a mutation rate of 1.36×10^{-8} mutations per base pair and a recombination rate of 10^{-8} crossovers per base pair per generation. We assumed a realistic genome structure of 20 unlinked chromosomes, each composed of 1,000 genes, separated by 50 kilobase pairs of intergenic regions, with each gene containing 8 exons of 100 base pairs separated by 5 kilobase pairs of intronic sequence. We assumed that each exon was composed of three-base-pair codons, whose first two sites were under selection (non-synonymous) and third evolved neutrally (synonymous). Intergenic and intronic regions were assumed to evolve neutrally. We also assumed that the DFE for non-synonymous sites was the one inferred for wAGR with ∂a∂i/Fit∂a∂i—a gamma distribution with E(s) = 0.083 and β = 0.14—since this population had the simplest demography and was therefore the best to use for extrapolating the distribution of s from the inferred distribution of $N_e s$. However, we also examined how our results would change under a different DFE (E(s) = 0.018 and β = 0.175). In the simulations, we ran an initial burn-in phase of $8N$ generations to reach equilibrium. After the initial burn-in phase, the simulations were run for an additional 6,710 generations comprising the recent history of the five modelled populations. We sampled the mutations segregating in the whole populations every 300 generations and also sampled individuals from the 5 populations at the present time, with the sample sizes matching those of our data (n = 100 for wAGR, wRHG and EUR, and n = 50 for eAGR and eRHG). In some cases, when we had rapid fluctuations in the inferred mutation load, we also took samples at smaller generation intervals (that is, every 50 or 100 generations) to ensure that the trajectory was accurate. To speed up the simulations, we used a rescaling procedure, where the population sizes and generation times were divided by ten, whereas the recombination rate, mutation rate, migration rates and selection coefficients were multiplied by ten. We do not expect this procedure to impact the accuracy of the simulations, since we calculated only ratios between populations for quantities such as mutation load and summary statistics.

Summary statistics for approximating the mutation load. We used summary statistics based on individual genotypic data to approximate the mutation

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

load: the number of deleterious alleles in each individual as $N_{alleles} = N_{het} + 2 \times N_{hom}$, with N_{het} and N_{hom} corresponding to the numbers of heterozygous and homozygous genotypes, respectively[7,22]. We also stratified $N_{alleles}$ and N_{hom} for putatively deleterious variants into several functional categories, including non-synonymous variants, variants with GERP RS in different classes[40] and LOF mutations. The significance of population differences in per-individual observed and simulated genotype counts was assessed by estimating the confidence interval of ratios between population pairs. Confidence intervals for observed data were calculated by paired bootstrapping: we split the SNP data into 1,000 blocks and resampled, with replacement, 1,000 times. This approach takes into account the variance introduced by demographic processes[21,73]. The predictions for $N_{alleles}$ and N_{hom} for deleterious mutations stratified in ranges of selection coefficients were generated with simulations under additive ($h = 0.5$), recessive ($h = 0$) and partially recessive models ($h = 0.25$) for sample sizes matching the observed data. For these simulations, we assumed the RHG-first demographic model and a single underlying DFE for s across populations (that is, the DFE of wAGR). Given the multiple between-population comparisons performed for the observed data, we required a non-adjusted P value $\leq 10^{-3}$ to declare significance.

ROH. We searched for ROH along the genome of all individuals using the sliding window approach implemented in PLINK[59] on the SNP genotyping data. The whole genome of each sample was explored using sliding windows of 50 SNPs, and ROH were detected if the 50 SNPs were homozygous, with the possible exception of two heterozygous and five missing genotypes. Various minimum lengths were tested to define ROH regions. Finally, we allowed any length of ROH regions and discerned inbreeding from strong linkage disequilibrium on homozygous segments by classifying ROH in two size classes, adapted from ref. [74]. Class A ROH were 0–0.5 megabases in length and are attributable to high levels of local linkage disequilibrium likely to be generated by small population sizes. Class B ROH were >0.5 megabases in length, resulting from background relatedness due to limited population size or recent inbreeding.

Hotspots and coldspots of recombination. We defined regions of high recombination rates and coldspots of recombination using the full list of autosomal regions provided in ref. [75]. We obtained the intersections of the positions of each SNP in our dataset with the coordinates of high recombination rates and coldspots of recombination regions, and assigned them to one of the two categories.

Clinically relevant variants. To detect clinical variants with validated pathogenicity in the exomes of the African populations, we intersected the set of 406,270 SNPs segregating in the RHG and AGR groups with the curated ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/docs/maintenance_use/)[76]. We considered ClinVar entries with the most supported evidence for clinical significance, recorded as '5—Pathogenic'. We therefore identified a total of 334 pathogenic variants distributed in 251 genes and mostly segregating at a derived allele frequency lower than 3% (Supplementary Note 9).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. A complete description of the programmes and models used in this study is provided in the Methods and Supplementary Information. Custom scripts used to parse and analyse the data are available upon request from the corresponding authors.

Data availability. The newly generated exome sequencing data for the central African rainforest hunter-gatherers and agriculturalists ($n = 300$) have been deposited in the European Genome-phenome Archive under accession code EGAS00001002457. Exome sequencing data for the European population ($n = 100$) are available under accession code EGAS00001001895.

Received: 2 May 2017; Accepted: 8 February 2018;
Published online: 12 March 2018

References

1. Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* **15**, 149–162 (2014).
2. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
3. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc. Natl Acad. Sci. USA* **109**, 17758–17764 (2012).
4. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
5. Lohmueller, K. E. et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).
6. Casals, F. et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet.* **9**, e1003815 (2013).
7. Henn, B. M. et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl Acad. Sci. USA* **113**, E440–E449 (2016).
8. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
9. Pedersen, C. T. et al. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: insights from the Greenlandic Inuit. *Genetics* **205**, 787–801 (2017).
10. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
11. Fu, W., Gittelman, R. M., Bamshad, M. J. & Akey, J. M. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet.* **95**, 421–436 (2014).
12. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).
13. Peischl, S. et al. Relaxed selection during a recent human expansion. *Genetics* **208**, 763–777 (2018).
14. Agarwala, V., Flannick, J., Sunyaev, S., Go, T. D. C. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* **45**, 1418–1427 (2013).
15. Maher, M. C., Uricchio, L. H., Torgerson, D. G. & Hernandez, R. D. Population genetics of rare variants and complex diseases. *Hum. Hered.* **74**, 118–128 (2012).
16. Harris, K. & Nielsen, R. The genetic cost of Neanderthal introgression. *Genetics* **203**, 881–891 (2016).
17. Muller, H. J. Our load of mutations. *Am. J. Hum. Genet.* **2**, 111–176 (1950).
18. Wright, S. The distribution of gene frequencies in populations. *Science* **85**, 504 (1937).
19. Kimura, M., Maruyama, T. & Crow, J. F. The mutation load in small populations. *Genetics* **48**, 1303–1312 (1963).
20. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1984).
21. Simons, Y. B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Curr. Opin. Genet. Dev.* **41**, 150–158 (2016).
22. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
23. Do, R. et al. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* **47**, 126–131 (2015).
24. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).
25. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
26. Bocquet-Appel, J. P. When the world's population took off: the springboard of the Neolithic Demographic Transition. *Science* **333**, 560–561 (2011).
27. Cavalli-Sforza, L. L. African Pygmies (Academic, New York, 1986).
28. Bostoen, K. et al. Middle to Late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of western central Africa. *Curr. Anthropol.* **56**, 354–384 (2015).
29. Patin, E. et al. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* **5**, e1000448 (2009).
30. Verdu, P. et al. Origins and genetic diversity of pygmy hunter-gatherers from western central Africa. *Curr. Biol.* **19**, 312–318 (2009).
31. Veeramah, K. R. et al. 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 (2012).
32. Hsieh, P. et al. 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 (2016).
33. Patin, E. et al. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat. Commun.* **5**, 3163 (2014).
34. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics* (Roberts & Company, Greenwood Village, 2010).
35. Batini, C. et al. Insights into the demographic history of African pygmies from complete mitochondrial genomes. *Mol. Biol. Evol.* **28**, 1099–1110 (2011).
36. Quach, H. et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656.e17 (2016).
37. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

38. Pagani, L. et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* **91**, 83–96 (2012).
39. Hodgson, J. A., Mulligan, C. J., Al-Meeri, A. & Raaum, R. L. Early back-to-Africa migration into the Horn of Africa. *PLoS Genet.* **10**, e1004393 (2014).
40. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
41. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
42. Kim, B. Y., Huber, C. D. & Lohmueller, K. E. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics* **206**, 345–361 (2017).
43. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
44. Balick, D. J., Do, R., Cassa, C. A., Reich, D. & Sunyaev, S. R. Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS Genet.* **11**, e1005436 (2015).
45. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
46. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
47. Scholz, C. A. et al. East African megadroughts between 135 and 75 thousand years ago and bearing on early-modern human origins. *Proc. Natl Acad. Sci. USA* **104**, 16416–16421 (2007).
48. Kim, H. L. et al. Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nat. Commun.* **5**, 5692 (2014).
49. Skoglund, P. et al. Reconstructing prehistoric African population structure. *Cell* **171**, 59–71.e21 (2017).
50. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
51. Fagny, M. et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat. Commun.* **6**, 10047 (2015).
52. Patin, E. et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546 (2017).
53. Perry, G. H. et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc. Natl Acad. Sci. USA* **111**, E3596–E3603 (2014).
54. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
56. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
57. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
58. Kousathanas, A., Oliver, F., Halligan, D. L. & Keightley, P. D. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol. Biol. Evol.* **28**, 1183–1191 (2011).
59. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
60. Lischer, H. E. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).
61. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
62. Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
63. Lipson, M. et al. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLoS Genet.* **11**, e1005550 (2015).
64. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
65. Ewing, G. & Jensen, J. The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* **25**, 135–141 (2016).
66. Schrider, D. R., Shanku, A. G. & Kern, A. D. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223 (2016).
67. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
68. Boyko, A. R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
69. Hernandez, R. D. et al. Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
70. Kousathanas, A. & Keightley, P. D. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* **193**, 1197–1208 (2013).
71. Messer, P. W. & Petrov, D. A. Frequent adaptation and the McDonald–Kreitman test. *Proc. Natl Acad. Sci. USA* **110**, 8615–8620 (2013).
72. Haller, B. C. & Messer, P. W. SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).
73. Gravel, S. When is selection effective? *Genetics* **203**, 451–462 (2016).
74. Pemberton, T. J. et al. Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
75. Hussin, J. G. et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat. Genet.* **47**, 400–404 (2015).
76. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
77. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

Acknowledgements

We thank all the participants for providing the DNA samples used in this study. We thank the Paleogenomics and Molecular Genetics Platform of the Musée de l'Homme-Muséum National d'Histoire Naturelle for technical assistance with DNA sample preparation. We thank G. Laval, L. Excoffier, M. Rotival and S. Ait Kaci Azzou for helpful discussions, and N. Joly for help with computational resources. This work was supported by the Institut Pasteur, Centre National de la Recherche Scientifique and Agence Nationale de la Recherche grant 'AGRHUM' (ANR-14-CE02-0003-01). M.L. was supported by the Fondation pour la Recherche Médicale (FDT20170436932) and A.K. by a Pasteur-Roux fellowship.

Author contributions

A.K. and M.L. designed the analytical approach and performed the analyses with input from E.P. and L.Q.-M. H.Q. and C.H. performed the experiments. P.M.-D., J.-M.H., A.F., G.H.P., L.B.B. and P.V. collected the samples. L.Q.-M. conceived the study with input from E.P. and obtained funding. The manuscript was written by A.K., M.L., E.P. and L.Q.-M. with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0496-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.K. or L.Q.-M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Corresponding author(s): Lluis Quintana-Murci
Athanasios Kousathanas

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

See Supplementary Note 1 and Supplementary Table 1. We used a total of 400 samples from populations with different historical lifestyles: 100 Baka mobile rainforest hunter gatherers (wRHG) and 100 Nzebi and Bapunu sedentary farmers (wAGR) from Gabon and Cameroon in western central Africa, and 50 BaTwa rainforest hunter-gatherers (eRHG) and 50 BaKiga farmers (eAGR) from Uganda in eastern central Africa and 100 Belgians of European ancestry (EUR). Signatures for recent demographic events can only be detected if the full site frequency spectrum is obtained from sequencing data for relatively large sample sizes (>>10). Moreover, we performed analyses to characterize the efficacy of purifying selection. For these analyses accurate estimation of parameters depends on detection of low frequency (1-5%) deleterious variants which is possible only with rather large sample sizes (Keightley and Eyre-Walker 2010).

2. Data exclusions

Describe any data exclusions.

See Supplementary Note 1. Overall we sequenced the exome of 314 African samples, and processed these data together with 101 European individuals. As a criterion to remove low-quality samples, we required at least 40x of mean depth of coverage (3 excluded samples), 85% of the positions in the BAM file to be covered at 5x minimum (8 excluded samples) and a total genotype missingness lower than 5% (1 excluded sample) (Supplementary Fig. 6). In addition, we checked for unexpectedly high or low heterozygosity values suggesting high levels of inbreeding or DNA contamination, and excluded 3 additional individuals presenting heterozygosity levels 4 SD higher than their population average (Supplementary Fig. 6). We thus retained exome data from 400 individuals, with an average depth of coverage of 68x, ranging from 40x to 168x, and an individual breadth of coverage above 5x for 93% of the exome target on average, ranging from 85% to 97%.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

The sampling strategy of this study allowed the comparison of a western RHG population (n=100) with a neighboring population of western AGR (n=100). All findings were replicated in a similar setting of populations on the eastern part of Central Africa, with the comparisons of an eastern RHG population (n=50) with a neighboring population of eastern AGR (n=50).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

In the context of this study, individuals did not belong to experimental groups and have been grouped based on their historical modes of subsistence (hunter-gatherers vs agriculturalists).

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	A complete description of the programs and models used in this study is provided in the Methods and Supplementary Information. All programs used are already published (GATK, bwa, fastsimcoal2, DFE-alpha, Arlequin, dadi/fitdadi), except some basic, home-made scripts for preparing input and parsing the output for these programs. All custom scripts are available upon request.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	Not applicable

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable
b. Describe the method of cell line authentication used.	Not applicable
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	Not applicable

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Phenotypic information about participants was not collected in this study. Genotypic information about participants is available in Methods, "Data availibility".

nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0496-4

In the format provided by the authors and unedited.

The demographic history and mutational load of African hunter-gatherers and farmers

Marie Lopez[1,2,3,10], Athanasios Kousathanas[1,2,3,10]*, Hélène Quach[1,2,3], Christine Harmant[1,2,3], Patrick Mouguiama-Daouda[4,5], Jean-Marie Hombert[5], Alain Froment[6], George H. Perry[7], Luis B. Barreiro[8], Paul Verdu[9], Etienne Patin[1,2,3] and Lluís Quintana-Murci[1,2,3]*

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France. [2]Centre National de la Recherche Scientifique UMR 2000, Paris, France. [3]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. [4]Laboratoire de Langue, Culture et Cognition, Université Omar Bongo, Libreville, Gabon. [5]Centre National de la Recherche Scientifique UMR 5596, Dynamique du Langage, Université Lumière-Lyon 2, Lyon, France. [6]Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, Paris, France. [7]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, USA. [8]Université de Montréal, Centre de Recherche du Centre Hospitalier Universitaire Sainte-Justine, Montréal, Canada. [9]Centre National de la Recherche Scientifique UMR 7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Sorbonne Paris Cité, Paris, France. [10]These authors contributed equally: Marie Lopez and Athanasios Kousathanas. *e-mail: kousathanas2@gmail.com; quintana@pasteur.fr

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Supplementary Notes

Note 1 - Population and individual selection

Four populations of rainforest hunter-gatherers (RHG) and neighbouring farmers (AGR) were selected for exome sequencing: an AGR and a RHG population from western central Africa, and an AGR and a RHG population from eastern central Africa (Supplementary Table 1). As to western RHG, the Baka of southern/eastern Cameroon and northern Gabon were selected because they show lower levels of admixture with AGR compared with other RHG populations, and because of the large number of individuals who have been sampled in this ethnic group. As to western AGR, the Bantu-speaking Nzebi and Bapunu sedentary agriculturalists of Gabon were selected because they showed limited genetic differentiation (data not shown). When merging Nzebi and Bapunu populations, ADMIXTURE analysis supported a single ancestry (K=1; Supplementary Fig. 3) and the inbreeding coefficient F_{IS} distribution across synonymous SNPs was not significantly different from zero (Supplementary Fig. 4). As to eastern central Africa, we chose the BaTwa RHG and BaKiga AGR from Uganda[1]. More generally, ADMIXTURE analyses and F_{IS} distributions across all populations provided no evidence of internal substructure within the groups studied (Supplementary Figs. 3 and 4).

In each population, individuals were selected based on previous genome-wide SNP data, obtained with the Illumina HumanOmniExpress BeadChip array for 730,525 SNP markers, to avoid cryptically-related pairs of individuals; a summary of the different datasets used and their accession numbers is shown in Supplementary Table 1. In total, 317 African samples were analysed, including 111 Baka RHG of Cameroon and Gabon, 46 Bapunu AGR and 56 Nzebi AGR from Gabon, 52 BaKiga AGR and 52 BaTwa RHG from Uganda. We removed a total of 92,968 SNPs because of their physical location (SNPs on the Y-chromosome and SNPs unmapped on dbSNP build 137), problematic genotype clustering profiles (*i.e.,* Illumina GenTrain score > 0.35) or a SNP call rate <95%. We merged this filtered dataset with 101 individuals of European ancestry sampled in Belgium, which were genotyped with the Illumina Omni5 array[2], and applied the same SNP quality filters. We kept only 599,559 SNPs common to both datasets. We removed a total of 53 C/G or A/T SNPs to correct for misaligned SNPs, and excluded a total 3,395 additional SNPs that were under Hardy-Weinberg disequilibrium in at least one of the five populations (P-value < 0.001), leading to a final SNP array dataset of 596,111 SNPs. All individuals had heterozygosity levels within 4 standard deviations (SD) higher or lower than the population mean, and individual missingness values did not exceed 2.4% (Supplementary Fig. 6, Supplementary Table 1).

Based on this data, we searched for 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[3] (Supplementary Fig. 5). Three individuals with first-degree relationships with other samples were thus removed, leading to a total of 314 African individuals that were retained for exome sequencing.

The ADMIXTURE model-based approach[4] was also used to estimate, for each individual, the proportions of their genome originating from K ancestral populations, K being specified *a priori.* To obtain the most supported results and test for their stability, all ADMIXTURE analyses were run ten times with different random seeds, for each K value. We show in Fig. 1b results for K=4 providing the lowest cross-validation (CV) among iterations (Supplementary Fig. 2).

Note 2 - Demographic inference: initial checks and model-building

Prior to fitting complex demographic models, we began our demographic inference by considering three-epoch models of population size change, which were fitted to the one-dimension (1D) site frequency spectra (SFS) calculated for each population separately. The parameter estimates for these three-epoch models suggested that the studied populations have experienced fairly different size changes. We found that an ancient expansion in both RHG and AGR populations was followed by a more recent population size reduction in RHG and an expansion in AGR (Supplementary Table 8). With respect to EUR, we inferred a strong bottleneck of 5-times the initial N_e followed by a massive 25-times population growth (Supplementary Table 8).

We then modelled all the populations together and fitted models into non-CpG synonymous pairwise (2D) SFS. Because the chronology of divergences between these populations remains unknown, we considered three possible branching models (EUR-first, RHG-first, AGR-first, Fig. 2), to evaluate different scenarios regarding the order of population splits. To capture the possible complexity of the size changes for the different populations considered, we constructed models that included 4 epochs of population size changes (Supplementary Fig. 8). We assumed that an ancient epoch of constant population size (epoch 1 for all populations) was followed by a size change for the ancestor of all human populations (epoch 2 for all populations). Subsequently, population size changes coincided with population splits: the split of Europeans from African populations (epoch 3 for Europeans), the split of African populations into the ancestors of RHG and AGR (epoch 3 for Africans), and the subsequent split of RHG and AGR into western and eastern populations (epoch 4 for Africans). We also included a recent population size change for Europeans (epoch 4 for Europeans). The size changes could be expansions or reductions in size and were not constrained in any way.

To choose the most likely migration scenario, we evaluated how different migration scenarios affect the fit of the branching models to the data. We considered three scenarios: the first scenario assumed no migration between populations, the second scenario assumed continuous constant migration between populations across their history, and the third scenario assumed that migration occurred in two epochs, an ancient migration epoch that extended from the initial population split (T_{DIV1}) to the split between western and eastern AGR (T_{AGR}), followed by a recent migration epoch that lasted until present (Fig. 2). We evaluated the likelihood differences between these three migration scenarios (Supplementary Table 3), although we could not formally test for significance due to the fact that we calculated composite likelihoods. When no migration or a single continuous migration rate over time was assumed, we obtained a much lower likelihood than when two epochs of migration were considered for all branching models (Supplementary Table 3). We thus decided to use the complex two epoch migration scenario for our main analyses. Having a complex migration scenario might limit our power to discriminate between branching models or estimate parameters with high precision due to the many possible ways that this scenario can fit to the data. However, it allows us to integrate over the migration parameter space and thus be conservative when considering other parameters such as ratios of population sizes: a significant population size change in our inference framework (*i.e.*, as judged by the 95% confidence intervals of ratios of population sizes) will be robust to either weak or strong migration in the ancient past, although we cannot estimate the precise strength or direction of migration.

Note 3 - Demographic inference: detailed demography results

Regardless of the population branching model considered, our results suggested that the ancestors of the contemporary RHG, AGR and EUR diverged between 85 and 140 thousand years ago (kya), from an ancestral population that underwent demographic growth between 173 and 191 kya (T_{ANC}) (Fig. 2), supporting an ancient expansion of modern humans during the Pleistocene[5]. This expansion signal, which was mostly apparent in the EUR-first model, could also result from ancient structure within Africa, corresponding to a model that would include a "ghost" unsampled population[6]. After the initial population splits, the N_e of AGR and RHG (N_{aAGR} and N_{aRHG}) remained within a range extending from 0.55 to 2.2 times the ancestral African N_e (N_{HUM}), whereas EUR (N_{aEUR}) experienced a decrease in N_e by a factor of 3 to 7 (Supplementary Tables 4 and 5). The ancestors of the wRHG and eRHG populations diverged 18 to 20 kya (T_{RHG}), and underwent a decrease in N_e by a factor of 3.8 to 5.7 for the wRHG (N_{wRHG}) and 7.1 to 11 for the eRHG (N_{eRHG}), regardless of the branching model considered. The ancestors of the AGR (N_{aAGR}) split into western and eastern populations 6.7 to 11 kya (T_{AGR}), and underwent a mild expansion, by a factor of 2.3 to 3.1 for the wAGR (N_{wAGR}) and 1.2 to 2.2 for the eAGR (N_{eAGR}). The EUR population experienced a 7.1- to 8.3-fold expansion (N_{EUR}) 12 to 22 kya (T_{EUR}).

The estimated migration parameters were also mostly similar across models (Supplementary Table 6). To account for the differences in N_e of the recipient populations, we compared the effective strength of migration ($2Nm$) between populations exchanging migrants (Supplementary Table 4). For the ancient migration epoch, we found that asymmetric migration was stronger in the direction RHG to AGR than the inverse (3.5-fold stronger) for the EUR-first model, but not for the RHG-first and the AGR-first models (from 1.6- to 1.8-times stronger from AGR to RHG than the inverse; Supplementary Table 4). This discrepancy between models likely reflects the moderate power to estimate the rate and direction of ancient migration events, which is also reflected in the wide confidence intervals for these parameters (Supplementary Table 4). With respect to the recent migration epoch, migration between RHG and AGR was inferred to be very strong and mostly symmetric across models ($2Nm$ consistently higher than 17 and 8 in western and eastern groups, respectively; Supplementary Table 4), providing evidence that admixture between these populations has mostly occurred in recent times. Finally, we inferred that $2Nm$ was larger for migration from EUR to AGR than the opposite direction, for both migration epochs (from 7- to 120-fold stronger across models and epochs; Supplementary Table 4), consistent with back-to-Africa migrations[7-9]. More generally, the 95% confidence intervals of estimated parameters were wide (Supplementary Table 4), owing to the complexity of the models and the limited number of synonymous sites used.

Note 4 - Average effective population size for a 3-epoch demographic history

For averaging N across the demographic history inferred for the 3-epoch model, we used a weighting scheme, following Eyre-Walker & Keightley (2009) (ref.[10]). Assuming that a diploid population of size N is sufficiently large, the probability that a pair of alleles has coalesced by time $T \leq t$, where t is the duration of an epoch of constant size can be approximated by:

$$\mathrm{P}(T \leq t) = 1 - \exp(-\frac{t}{2\mathrm{N}})$$

Thus, for a demographic model of 3 epochs, where population size is assumed to be piecewise constant, we can devise weights to account for the average time spent in a particular epoch before coalescing as follows:

for the most recent epoch 3:

$$w_3 = \mathrm{P}(T \leq t_3) = 1 - \exp(-\frac{t_3}{2N_3})$$

for preceding epoch 2:

$$w_2 = \mathrm{P}(t_3 < T \leq t_2 + t_3) = \exp(-\frac{t_3}{2N_3}) - \exp\left(-\left(\frac{t_2}{2N_2} + \frac{t_3}{2N_3}\right)\right)$$

and for the oldest epoch 1:

$$w_1 = 1 - \mathrm{P}(T \leq t_2 + t_3) = \exp\left(-(\frac{t_2}{2N_2} + \frac{t_3}{2N_3})\right)$$

And calculate a weighted average for the effective population size as:

$$N_w = N_1 w_1 + N_2 w_2 + N_3 w_3$$

Note 5 - Fixation probability of new deleterious mutations

We calculated the fixation probability of a new mutation for selection coefficient *s* and population size *N* as:

$$u(N, s) = \frac{1 - \exp(-s)}{1 - \exp(-2Ns)}$$

Kimura, 1962 (ref.[11])

We used the weighting scheme described in Note 4 to calculate *N*. We then integrated $u(N,s)$ over a given distribution of fitness effects with parameters scale (α) and shape (β) to calculate the average fixation probability of a new deleterious mutation (u_{del}):

$$\overline{u_{del}} = \int_0^\infty u\,(N, s) f(s \mid \alpha, \beta) ds$$

Keightley and Eyre-Walker, 2007 (ref.[12])

Finally, we computed the ratio u_{del}/u_{neu} as:

$$\frac{\overline{u_{del}}}{u_{neu}} = \frac{\overline{u_{del}}}{1/2N} = 2N\overline{u_{del}}$$

where, u_{neu} is the fixation probability of a new neutral mutation.

Note 6 - fitCons analyses for quantifying the fitness effects of mutations

In order to validate the results obtained for $N_{alleles}$ and N_{hom} ratios when partitioning the variants in different bins of GERP RS (Genomic Evolutionary Rate Profiling-Rejected Substitution)[13], we chose to partition the variants according to an additional function annotation software, fitCons[14]. fitCons scores estimate the probability of a variant to influence fitness. Importantly, fitCons integrates information from both evolutionary data and cell type-specific functional genomic data, and does not rely on the human reference, therefore avoiding the reference-bias effect. We stratified ratios of $N_{alleles}$ and N_{hom} across populations in bins of scores corresponding to different quantiles of the distribution of fitCons scores, obtained for nonsynonymous variants segregating in our dataset: "<25%" (fitCons scores ≤ 0.53), "25% - 50%" (fitCons scores (0.53; 0.63]), "50% - 75%" (fitCons scores (0.63; 0.70]), "75% - 95%" (fitCons scores (0.70; 0.723]) and ">95%" (fitCons scores > 0.723) (Supplementary Fig. 17). These bins of scores are likely to reflect the severity of the mutational impact, and we calculated confidence intervals by bootstrapping by site 1,000 times as recommended by previous studies[15,16]. Observed population differences in $N_{alleles}$ did not exceed 4% in any of the fitCons score bins, and were not significant for any of the population pairs examined. Observed N_{hom} ratios were 17% to 35% higher in Europeans than in Africans, but differences in N_{hom} ratios among African populations did not exceed 5.3% in any of the fitCons bins (Supplementary Fig. 17).

Note 7 - Analysis of variants in genes associated with recessive and dominant diseases

Very little is known about the dominance coefficients of variants segregating in natural populations. To get further insight on the distribution of potentially deleterious variants likely to be dominant or recessive, we considered genes listed in OMIM (Online Mendelian Inheritance in Man)[17]. The OMIM database reports genes as being responsible for known recessive or dominant diseases. We found a total of 63,169 nonsynonymous mutations distributed in 12,453 OMIM genes in our dataset. We stratified nonsynonymous variants into recessive and dominant, by assuming that variants within a gene implicated in recessive or dominant diseases have also recessive or dominant effects on fitness. Considering loss-of-function (LOF) mutations could have been a better proxy of dominant and recessive mutations, but this analysis would have been underpowered since we find only 1,172 LOF segregating in OMIM genes in our dataset. We calculated $N_{alleles}$ and N_{hom} ratios for nonsynonymous variants located in OMIM genes. We found that the population differences in $N_{alleles}$ did not exceed 1-5%, and were not significantly different for any of the population comparisons or between putatively dominant or recessive variants (Supplementary Fig. 18). Regarding $N_{hom,}$ we found 23-30% higher N_{hom} in Europeans compared to Africans for both putatively dominant and recessive variants, whereas population differences in N_{hom} did not exceed 5% between African populations (Supplementary Fig. 18).

Note 8 - Differences in the Genomic Distribution of Deleterious Variants

Despite the observed similarity between populations in terms of genomic estimates of mutation load, we explored possible differences between populations in the distribution of putatively deleterious alleles across the genome, due to variable levels of parental relatedness that lead to different patterns of extended runs of homozygosity (ROH) between populations Recent parental relatedness results in unexpectedly long segments of ROH, while parental relatedness due to a historically low N_e results in more numerous and short ROH segments[18-21]. We thus assigned segments of ROH to two length classes: short (0-0.5 Mb) and long (>

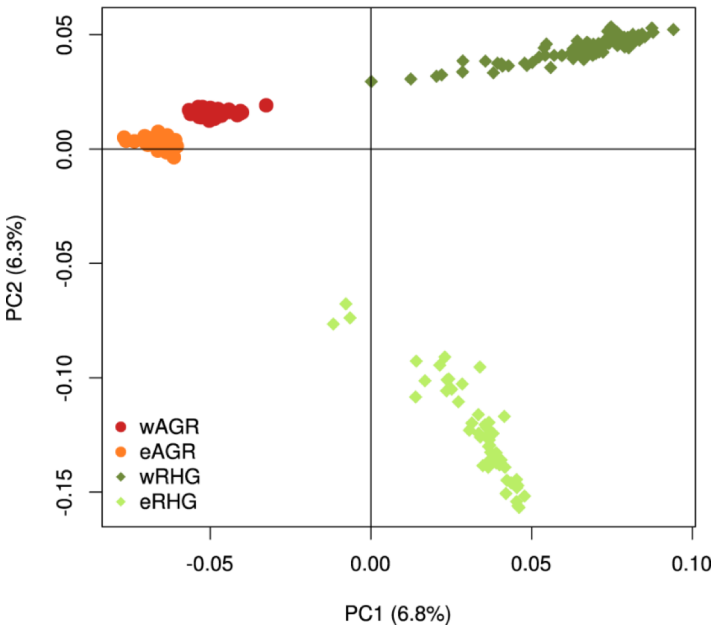
0.5 Mb) ROH[19,20]. We observed higher proportions of long ROH segments in RHG populations (wRHG: 24% ± 2.7%, eRHG: 26% ± 2.8%) than in AGR (wAGR: 20% ± 2.0%, eAGR: 21% ± 2.6%) and EUR (22% ± 1.1%). Longer ROH segments in RHG groups suggest that parental relatedness, together with the recent bottleneck they experienced, has increased further their levels of homozygosity (Supplementary Fig. 25).

We tested whether putatively deleterious variation accumulated differentially inside and outside ROH, by comparing between these genomic compartments the ratio of nonsynonymous to synonymous N_{hom} and $N_{alleles}$. We observed an overrepresentation of nonsynonymous variants within both short (Class A) and long (Class B) ROH for all populations (Supplementary Fig. 26). This excess was slightly, yet not significantly, higher for long ROH. These results suggest that ROH, including long ROH resulting from recent parental relatedness, have preferentially accumulated non-lethal deleterious variants. We then tested whether ROH segments were also associated with regions of particular recombination rates, *i.e.*, highly recombining regions (HRR) and coldspots (CS) of recombination[22]. We found that nonsynonymous alleles in ROH (either long or short) are more likely to be in a coldspot of recombination (Supplementary Fig. 27). We also found significantly larger ratios of nonsynonymous to synonymous N_{hom} and $N_{alleles}$ in CS compared to HRR, in all populations (Supplementary Fig. 26). Together, our results suggest that regions of low recombination accumulate deleterious variants, likely due to a decreased selection efficacy.

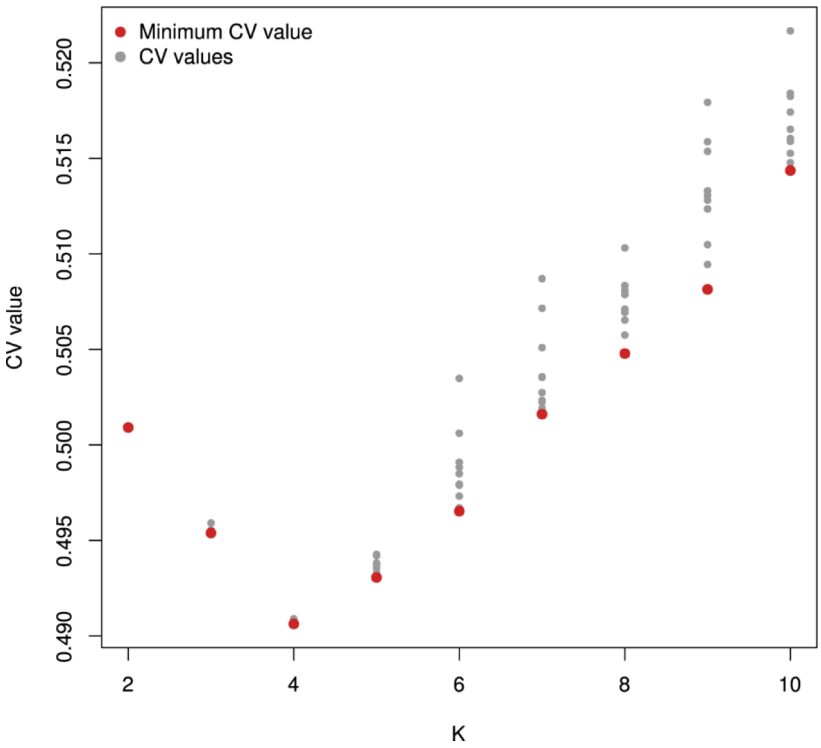
Note 9 - Population distribution of disease-related variants

We searched for clinically-relevant variants that could be differentially distributed between populations, by annotating RHG and AGR exome variation with the ClinVar database[23]. We found a higher number of disease-related variants specific to AGR than to RHG (98 *vs.* 39, respectively; Fisher's exact test *P*-value=3.2×10^{-3}) — probably reflecting the bias associated with the discovery of African pathogenic variants in African Americans who share ancestry with AGR populations[24]. However, clinically-relevant, disease-related variants were observed at intermediate frequencies in either RHG or AGR populations (Supplementary Table 10).

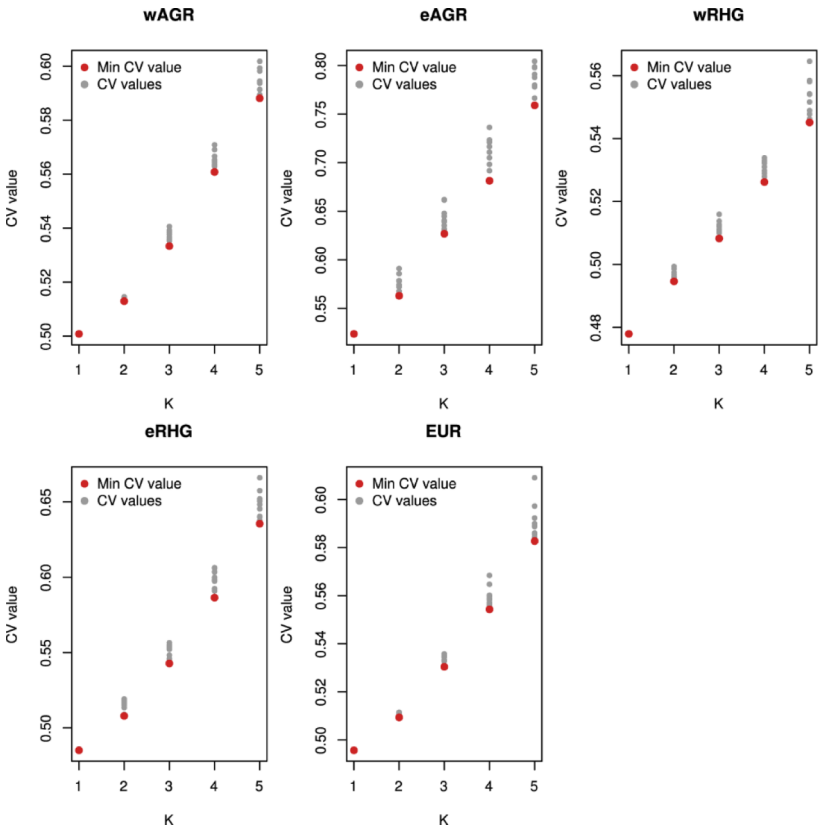
Highlighting some relevant examples, we found the regulatory variant -59356-T in the *CCR5* gene, known to increase the rate of HIV-1 mother-to-child transmission in African Americans[25] and mother's risk of death due to HIV infection[26], at low frequency in RHG (1-6%), intermediate frequency in AGR (15-20%) and absent in Europeans. Likewise, the p.Thr1209Ala variant in *CDH23*, which causes a recessive form of the Usher syndrome type I characterized by hearing loss and retinitis pigmentosa[27], was observed at 1-5% in RHG, 13-18% in AGR and absent in Europeans. These results indicate that populations of rainforest hunter-gatherers and farmers can carry deleterious, most likely recessive, standing variation at appreciable frequencies. Under the assumption of weak gene-environment interactions, this suggests that the genetic susceptibility to specific diseases can vary across populations, despite their similar mutation load.



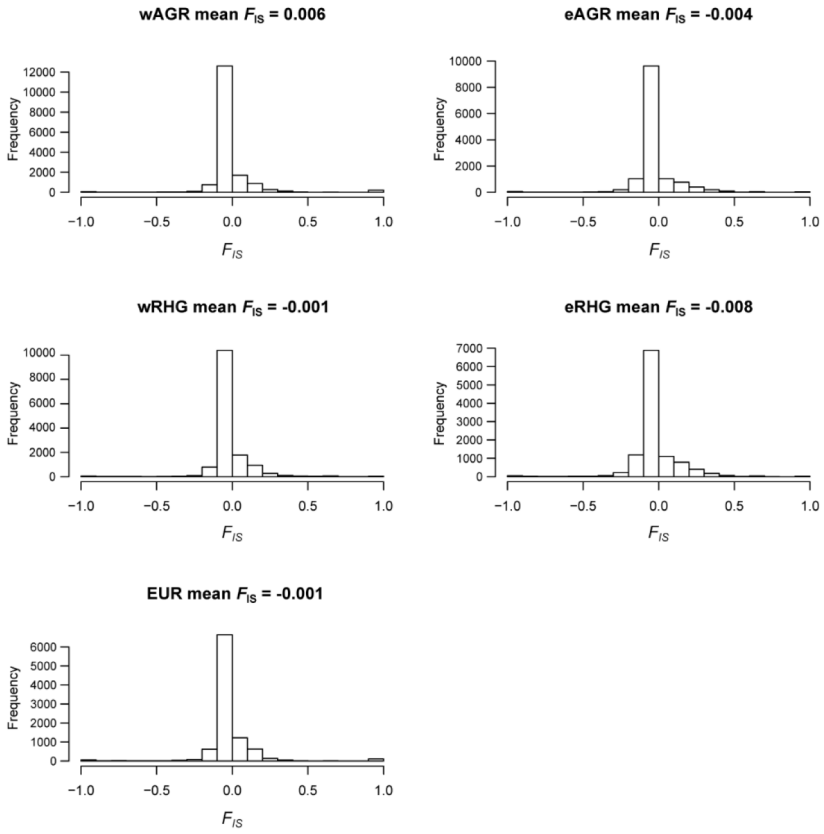
Supplementary Fig. 1 | Principal Component Analysis (PCA) implemented in EIGENSTRAT[28]. The African populations studied include western Bantu-speaking agriculturalists (wAGR; i.e., the Bapunu and Nzebi of Gabon; *n*=100), eastern Bantu-speaking agriculturalists (eAGR; i.e., the BaKiga of Uganda; *n*=50), western rainforest hunter-gatherers (wRHG; i.e., the Baka of Cameroon and Gabon; *n*=100) and eastern rainforest hunter-gatherers (eRHG; the BaTwa of Uganda; *n*=50).



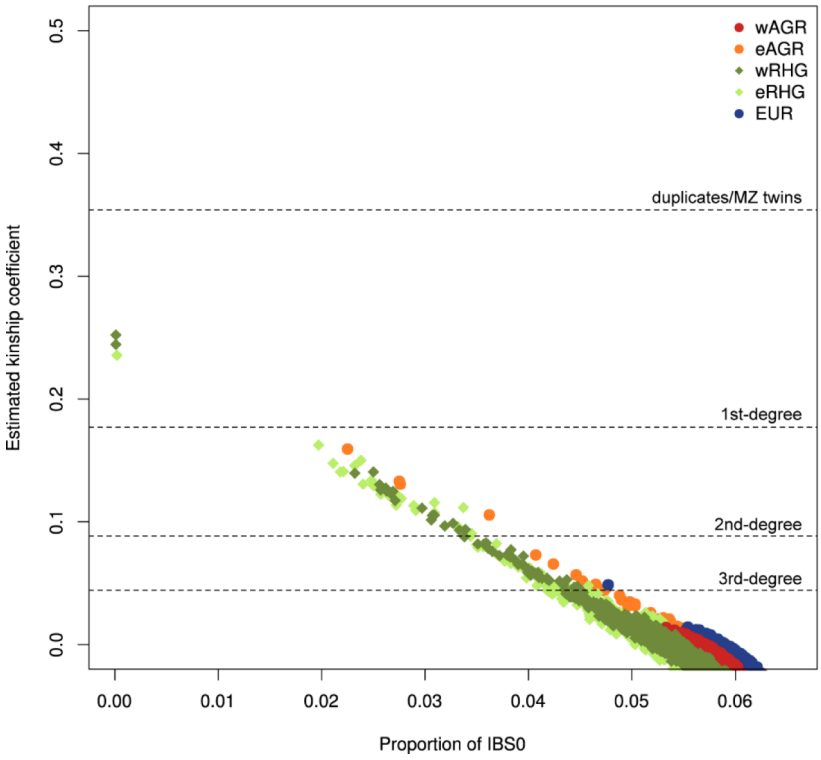
Supplementary Fig. 2 | Cross-validation (CV) values. CV values of different *K* values from the ADMIXTURE analysis (Fig. 1b). The minimum CV value obtained across ten independent runs of ADMIXTURE is represented in red, other values are in gray.



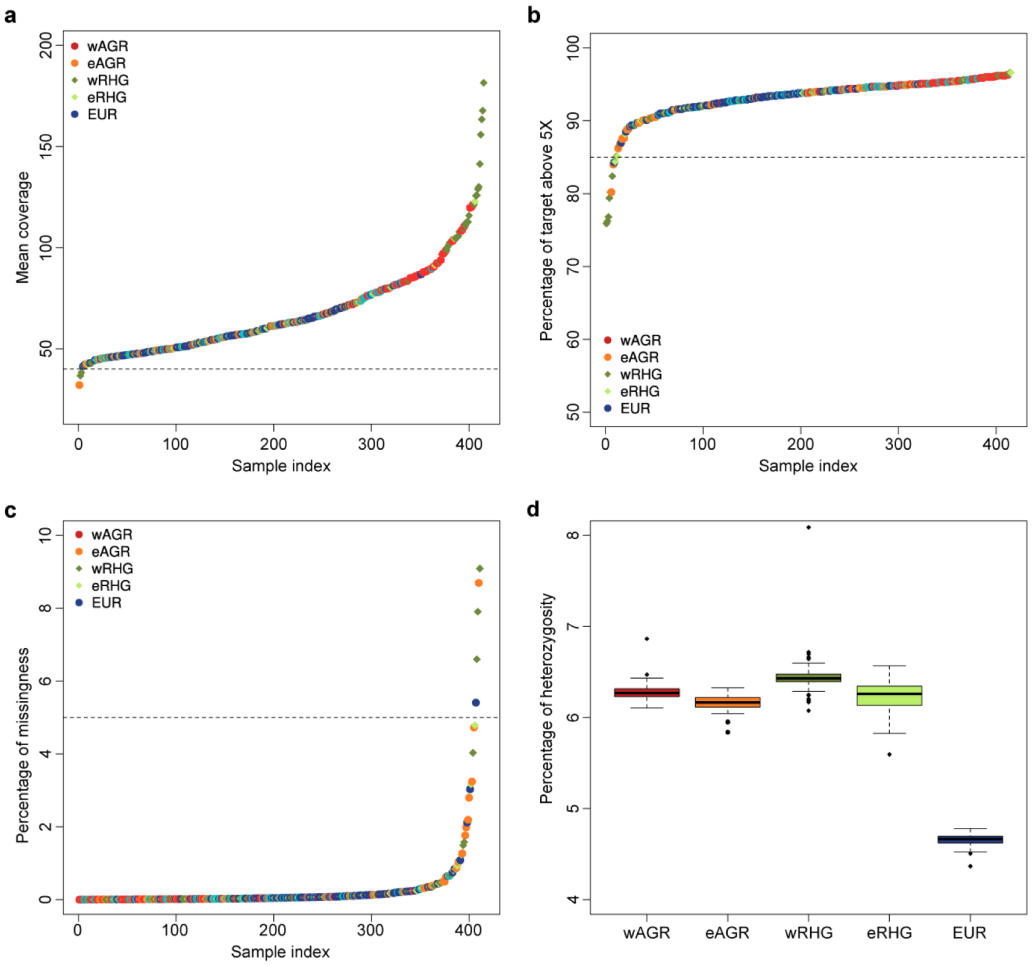
Supplementary Fig. 3 | Cross-validation (CV) values from the ADMIXTURE analysis of population groups. We performed ADMIXTURE analyses for each population separately, with ten independent runs for *K*=1,2,3,4,5. The minimum CV value is represented in red, other values are in gray.



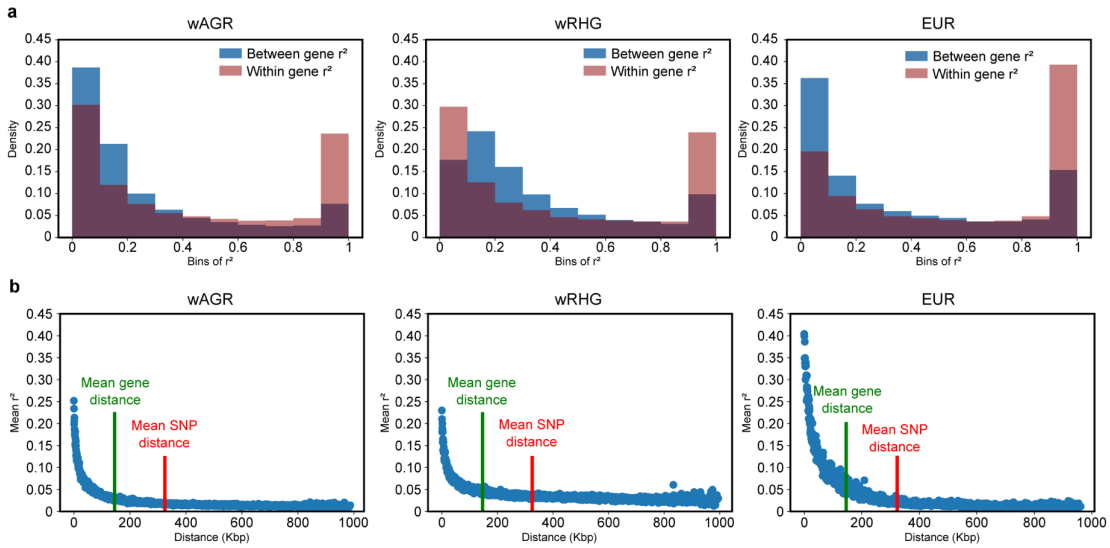
Supplementary Fig. 4 | Distribution of F_{IS}. We computed the distribution of the inbreeding coefficient (F_{IS}) for 4-fold synonymous SNPs in each population separately.



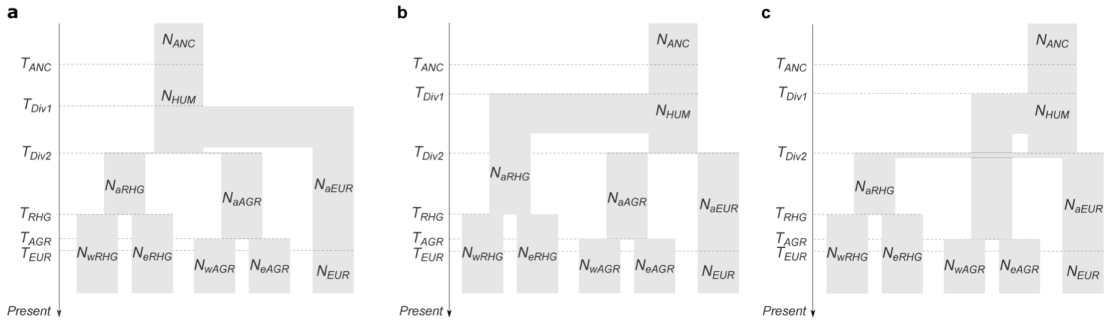
Supplementary Fig. 5 | Cryptic relatedness within populations. Estimated kinship coefficients and genomic proportion of non-Identical-By-State (IBS0) SNPs were estimated for each pair of individuals with KING. Kinship thresholds were defined as in ref.[3], and are shown with dashed lines.



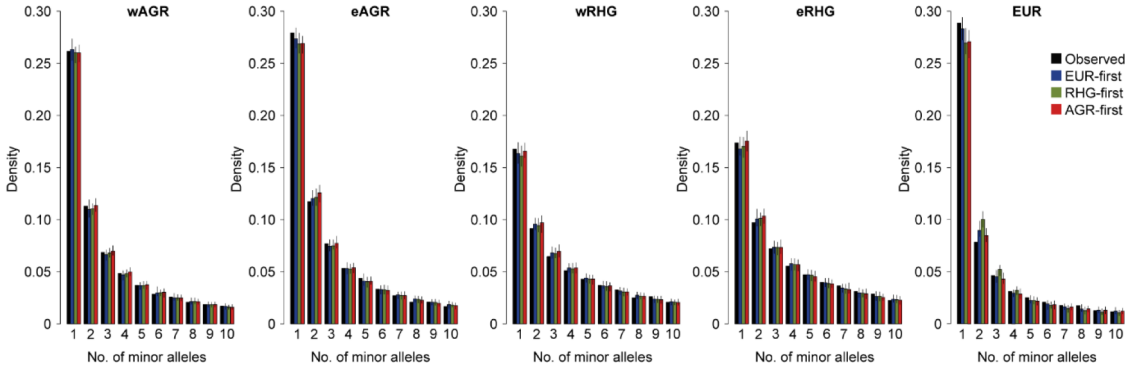
Supplementary Fig. 6 | Variant quality metrics. **a**, Individual mean depth of coverage including duplicated reads. **b**, Individual proportion of sites in the BAM file covered by at least five reads including duplicated reads. **c**, Individual proportion of missing data. **d**, Individual heterozygosity. The bold black line indicates the median, and box limits the 25th and 75th percentiles, of distributions.



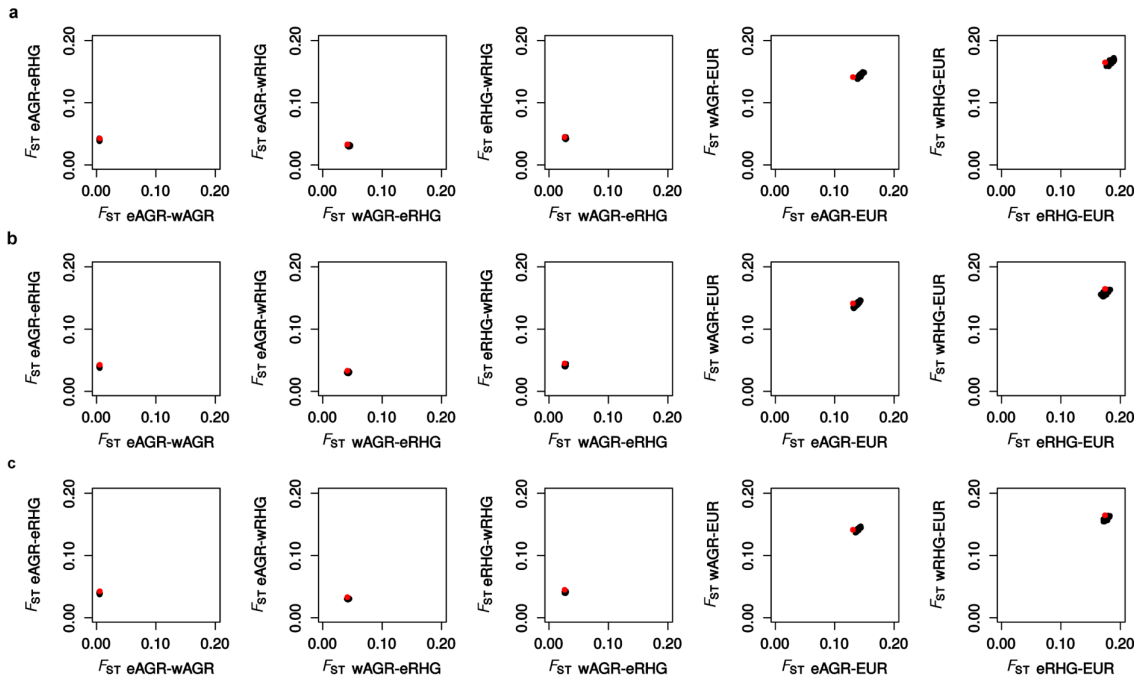
Supplementary Fig. 7| Linkage disequilibrium in the exome dataset. **a**, Population distribution of the maximum r^2 values per SNP considering pairs of SNPs located either within the same gene (red) or in different genes (blue). **b**, Mean r^2 values for 2,000 bins of genomic distances between SNP pairs. The green and red lines indicate, respectively, the mean distance between the center of two genes bodies in the genome and the mean distance between two SNPs located in different genes in 1Mb windows. The average r^2 between SNPs separated by ~146Kbp (*i.e.*, the average distance between two gene bodies) is r^2=0.03, 0.04, 0.05, for wAGR, wRHG and EUR, respectively.



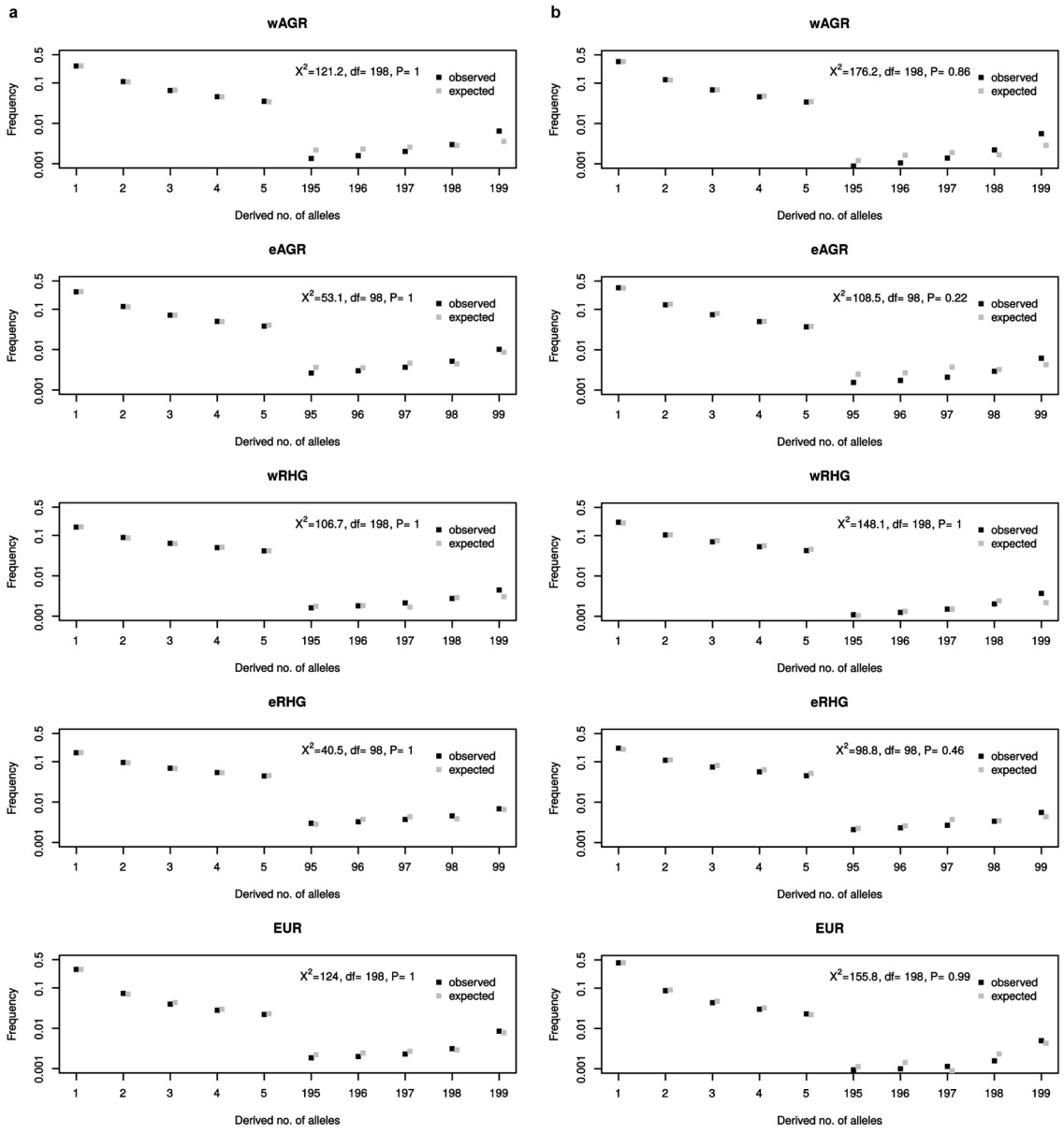
Supplementary Fig. 8 | Branching models tested. Schematic representation of the estimated parameters (times and effective population sizes) for the three branching models **a**, EUR-first, **b**, RHG-first and **c**, AGR-first, tested with *fastsimcoal2*.



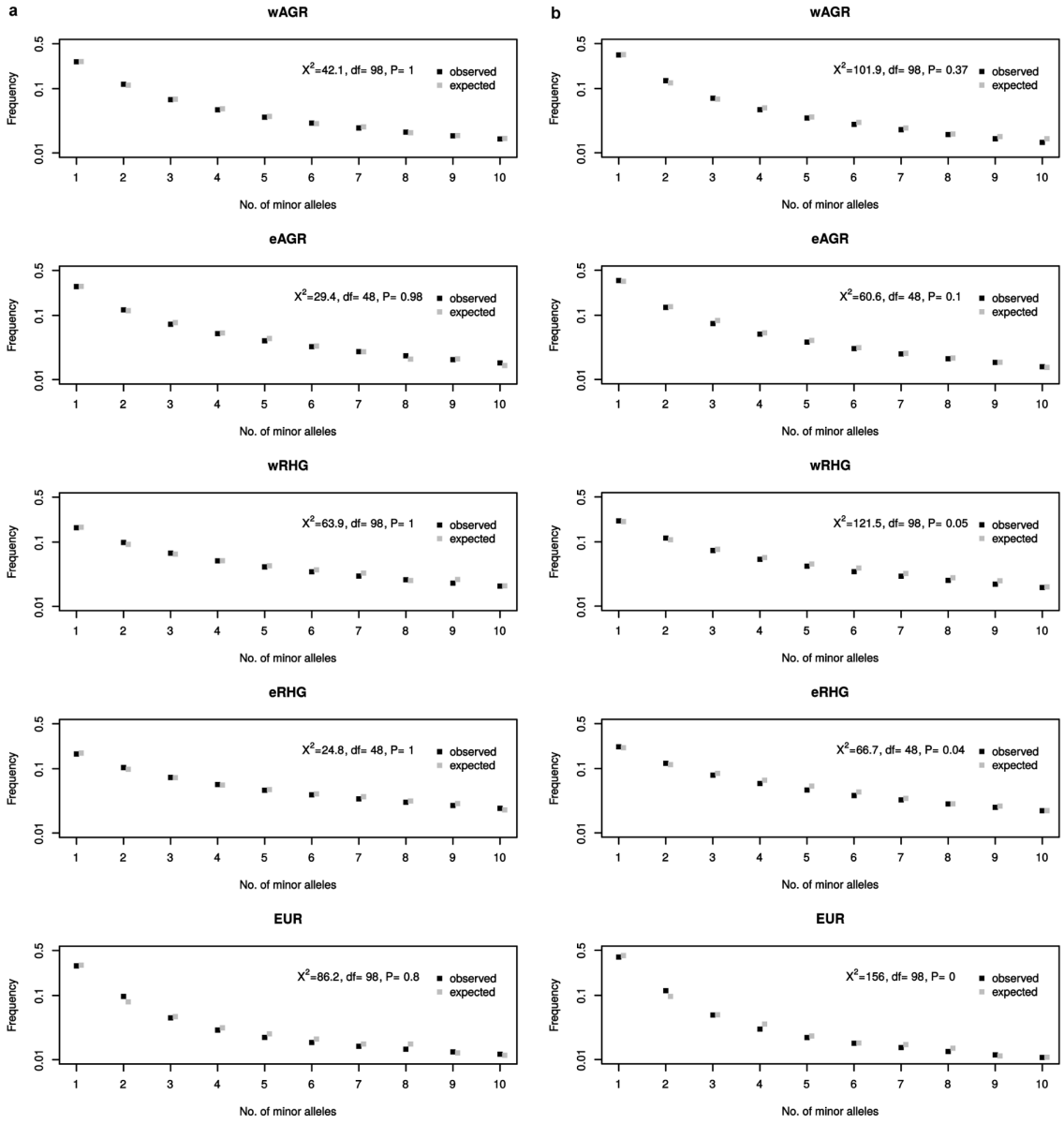
Supplementary Fig. 9 | Observed and expected synonymous site frequency spectra (SFS) from several best-fitted demographic models using *fastsimcoal2*. The bins for the number of minor alleles greater than 10 were omitted for reasons of presentation. Error bars on expectations correspond to the 2.5 and 97.5 percentiles of distributions of SFSs, generated by performing 100 simulations for each of the best-fitted models.



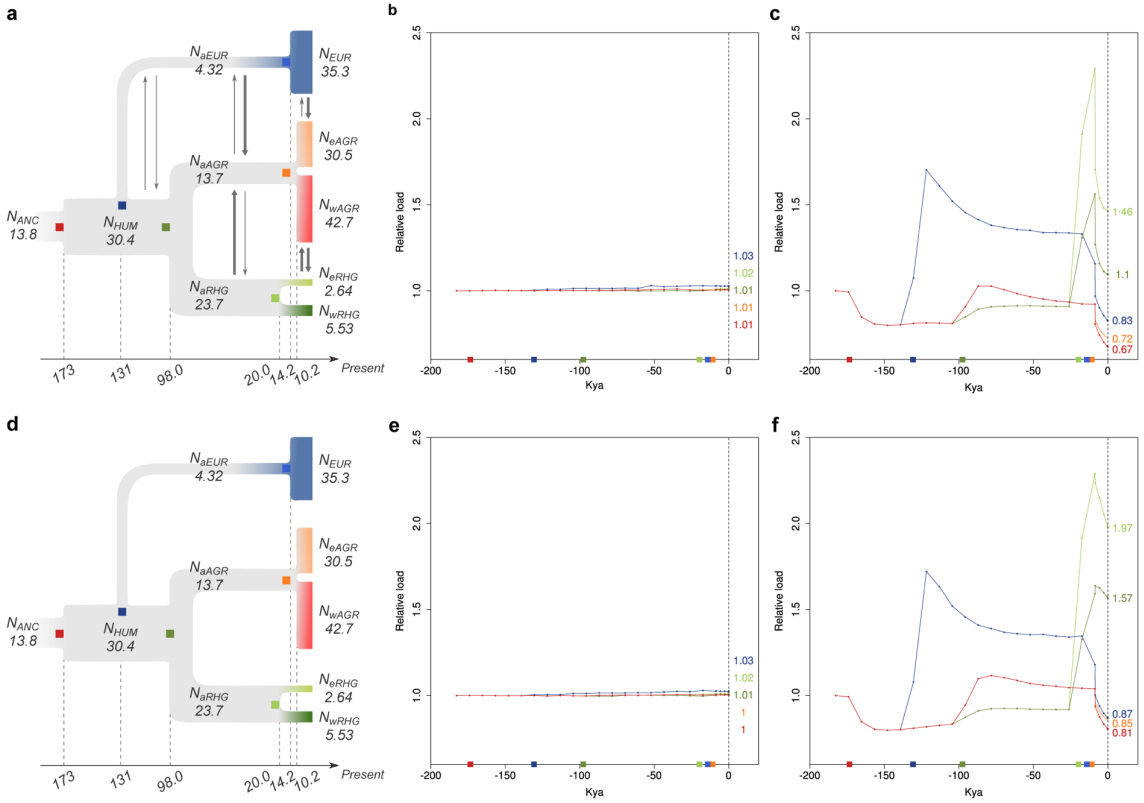
Supplementary Fig. 10 | Observed versus predicted F_{ST} for demographic models fitted to synonymous sites with *fastsimcoal2*. Predicted distributions of F_{ST} between all pairs of populations (black dots) were obtained with 100 simulations under the combinations of parameters of each tested model: **a**, EUR-first, **b**, RHG-first, **c**, AGR-first. Observed F_{ST} values are indicated in red.



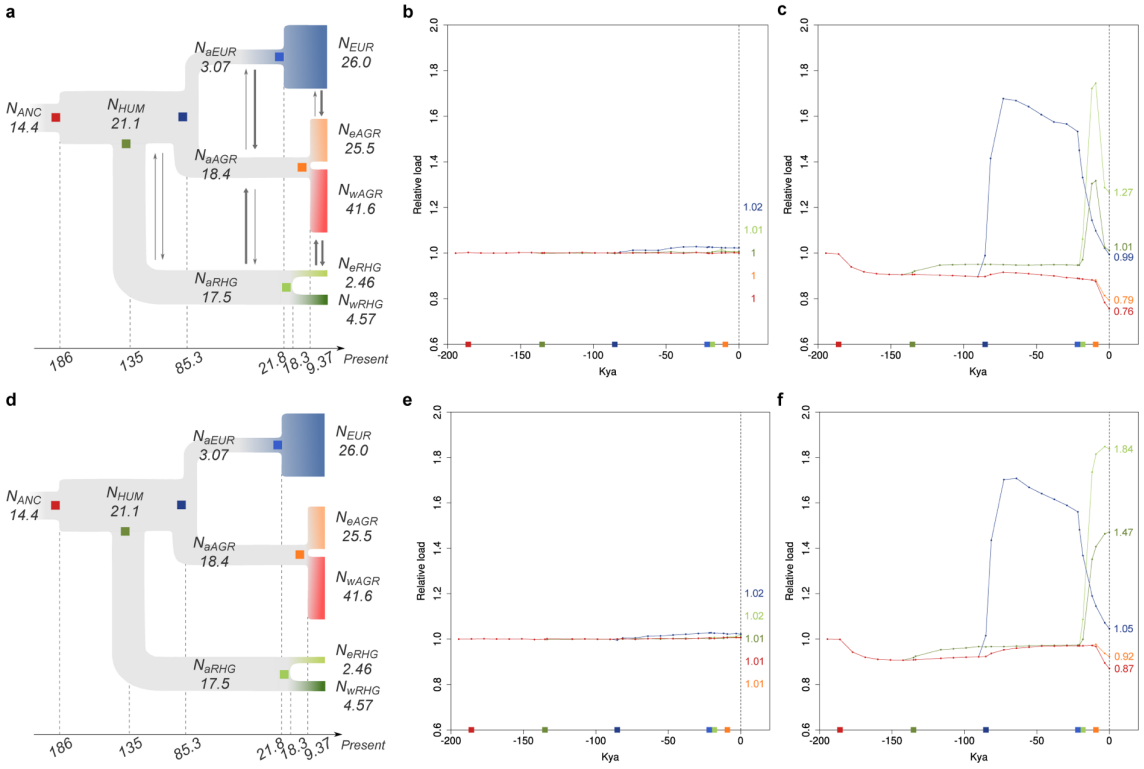
Supplementary Fig. 11 | Fitted site frequency spectra obtained with ∂a∂i/Fit∂a∂i. Observed (black) and expected (gray) unfolded site frequency spectra from ∂a∂i/Fit∂a∂i for **a**, synonymous and **b**, nonsynonymous sites. The bins for the number of derived alleles between 6-94 and between 6-194 for population sample sizes equal to 50 and 100, respectively, were omitted for reasons of presentation. Goodness of fit of expected to observed SFSs was tested using the non-parametric χ^2 test of non-independence.



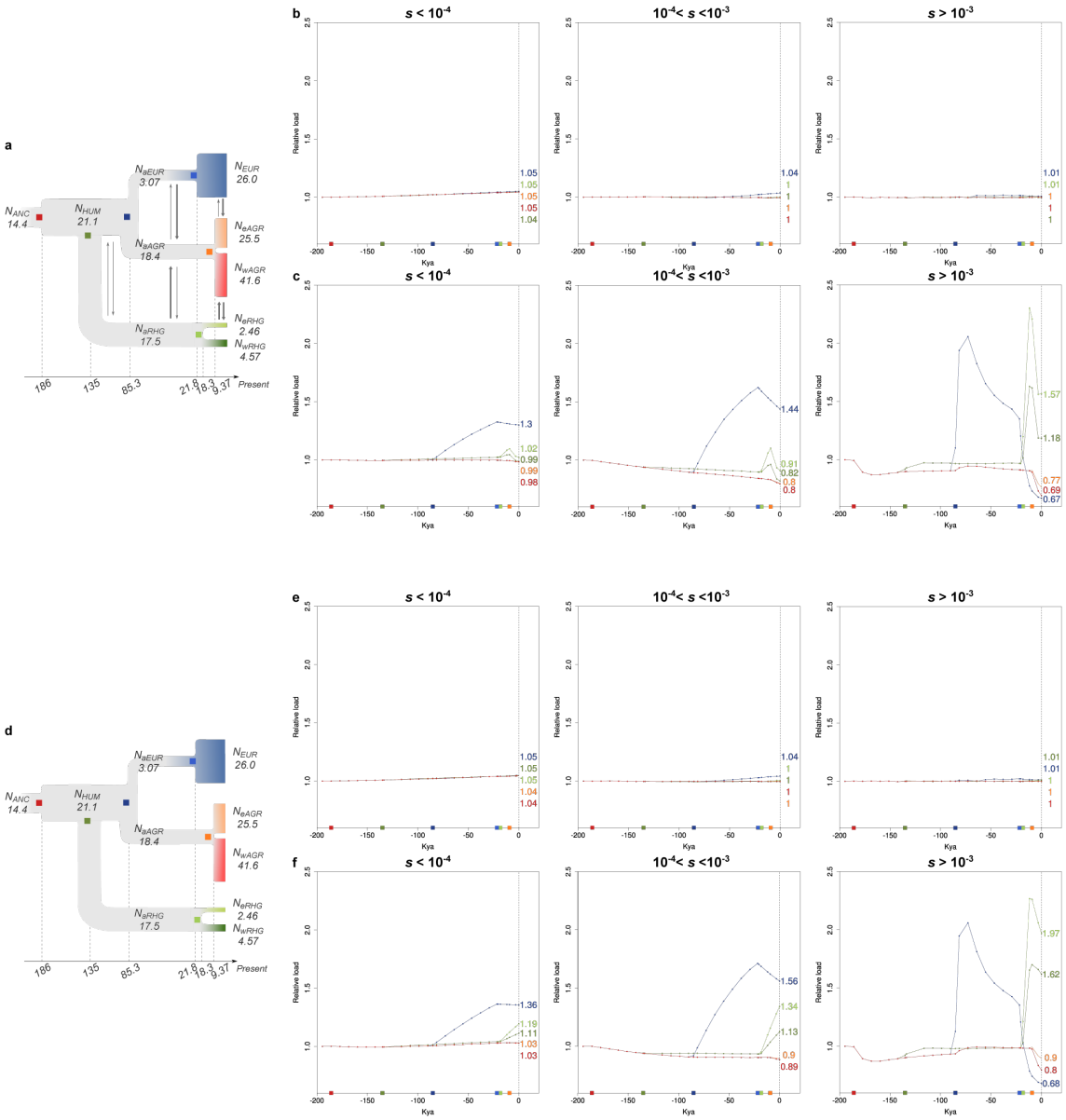
Supplementary Fig. 12 | Fitted site frequency spectra obtained with *DFE-α*. Observed (black) and expected (gray) unfolded site frequency spectra from *DFE-α* for **a**, synonymous and **b**, nonsynonymous sites. The bins for the number of minor alleles over 10 were omitted for reasons of presentation. Goodness of fit of expected to observed SFSs was tested using the non-parametric χ^2 test of non-independence.



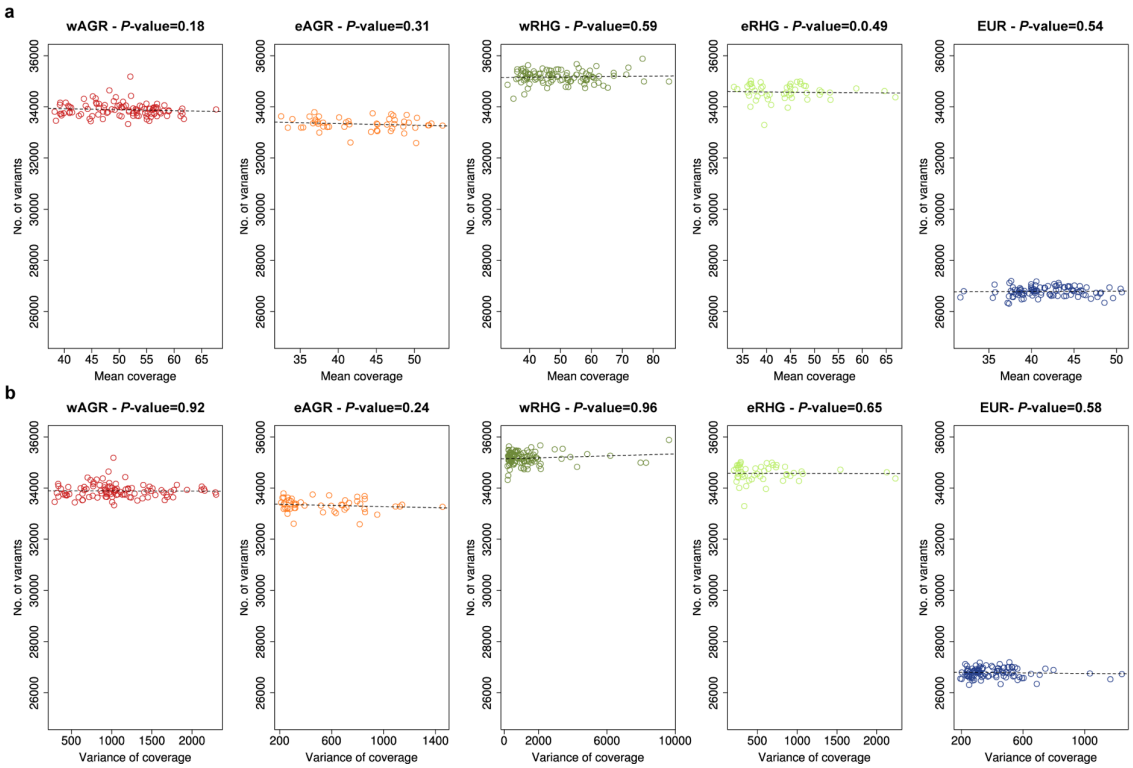
Supplementary Fig. 13 | Trajectory of mutation load through time obtained with simulations assuming the distribution of fitness effects inferred for wAGR and the EUR-first demographic model. Mutation load (L) relative to the ancestral population at equilibrium has been calculated as a function of time during the recent history of wAGR (dark red), eAGR (orange), wRHG (dark green), eRHG (light green) and EUR (blue), assuming **a,** the full EUR-first demographic model for **b,** additive and **c,** recessive mutations, and assuming **d,** the EUR-first demographic model without migration for **e,** additive and **f,** recessive mutations. For panels **b**, **c**, **e** and **f**, dashed vertical lines indicate the present time, and colored numbers indicate the relative mutation load at present time. Colored boxes indicate events in the demographic history of populations also depicted in panels **a** and **d**. Solid points in the trajectory indicate the time-points at which mutations were sampled in the simulations.



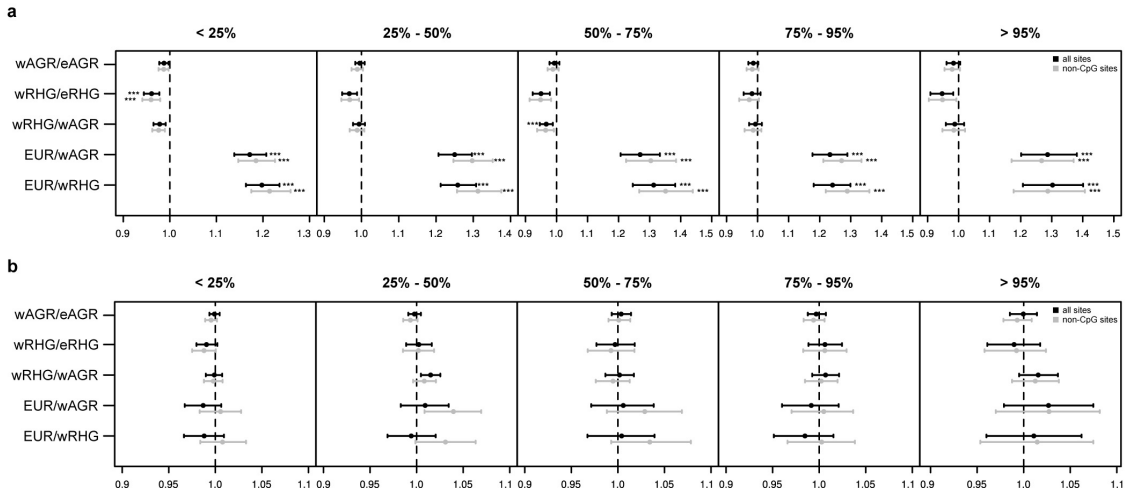
Supplementary Fig. 14 | Trajectory of mutation load through time obtained with simulations assuming the distribution of fitness effects inferred for EUR and the RHG-first demographic model. Mutation load (L) relative to the ancestral population at equilibrium has been calculated as a function of time during the recent history of wAGR (dark red), eAGR (orange), wRHG (dark green), eRHG (light green) and EUR (blue), assuming **a,** the full RHG-first demographic model for **b,** additive and **c,** recessive mutations, and assuming **d,** the RHG-first demographic model without migration for **e,** additive and **f,** recessive mutations. For panels **b**, **c**, **e** and **f**, dashed vertical lines indicate the present time, and colored numbers indicate the relative mutation load at present time. Colored boxes indicate events in the demographic history of populations also depicted in panels **a** and **d**. Solid points in the trajectory indicate the time-points at which mutations were sampled in the simulations.



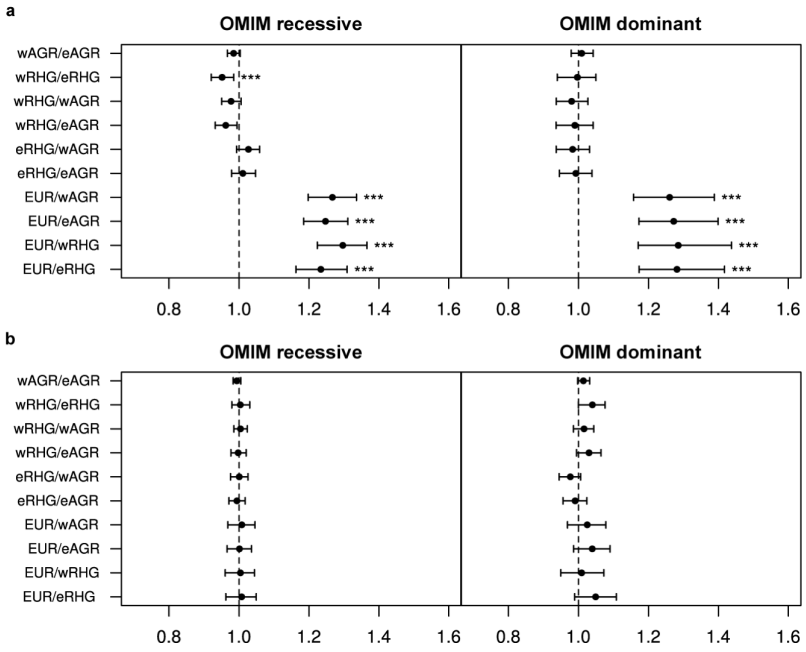
Supplementary Fig. 15 | Trajectory of mutation load through time obtained with simulations assuming the distribution of fitness effects inferred for wAGR and the RHG-first demographic model, and stratifying mutations according to their selection coefficient in three ranges: $s < 10^{-4}$, $10^{-4} < s < 10^{-3}$, $s > 10^{-3}$. Mutation load (L) relative to the ancestral population at equilibrium has been calculated as a function of time during the recent history of wAGR (dark red), eAGR (orange), wRHG (dark green), eRHG (light green) and EUR (blue), and assuming **a,** the full RHG-first demographic model for **b,** additive and **c,** recessive mutations, and assuming **d,** the RHG-first demographic model without migration for **e,** additive and **f,** recessive mutations. For panels **b**, **c**, **e** and **f**, dashed vertical lines indicate the present time, and colored numbers indicate the relative mutation load at present time. Colored boxes indicate events in the demographic history of populations also depicted in panels **a** and **d**. Solid points in the trajectory indicate the time-points at which mutations were sampled in the simulations.



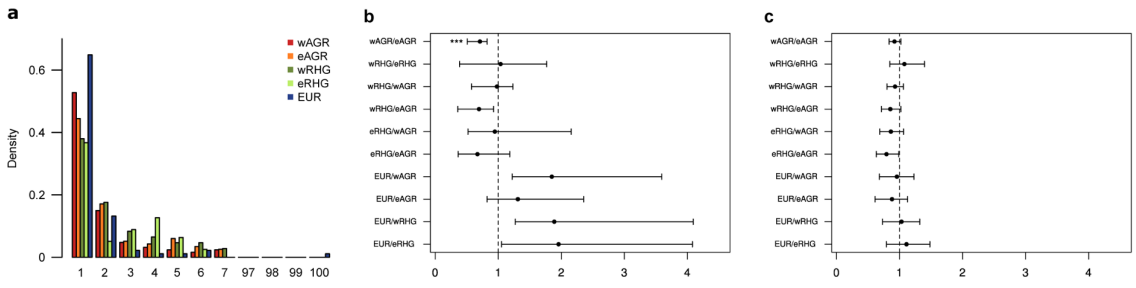
Supplementary Fig. 16 | Effect of the mean coverage depth on the number of variant sites per-individual. The **a**, mean and **b**, variance of coverage for each individual were calculated from the VCF file after removing duplicated reads. All variants considered had no missingness and at least 5x coverage. Dashed lines indicate the fit of a linear regression to the data, and association was tested using the non-parametric Spearman's rank correlation.



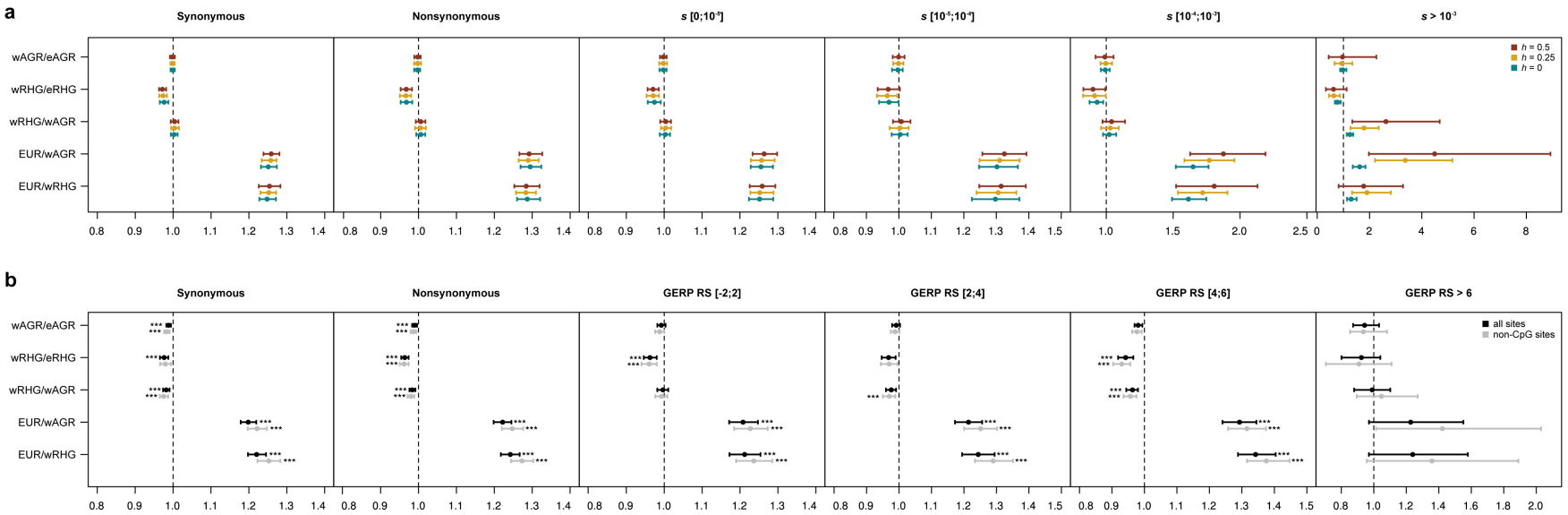
Supplementary Fig. 17 | Comparison of population differences in N_{hom} and $N_{alleles}$ across bins of fitCons scores. Between-population ratios of the mean per-individual counts of nonsynonymous (0-fold) **a**, homozygous genotypes (N_{hom}) and **b**, numbers of alleles ($N_{alleles}$). These ratios were computed with the observed data for several ranges of fitCons scores, corresponding to several quantiles of the distribution of fitCons scores for nonsynonymous variants: "<25%" (fitCons scores ≤ 0.53), "25% - 50%" (fitCons scores (0.53; 0.63]), "50% - 75%" (fitCons scores (0.63; 0.70]), "75% - 95%" (fitCons scores (0.70; 0.723]) and ">95%" (fitCons scores > 0.723). Confidence intervals were calculated by bootstrapping 1,000 times by site. ***: non-adjusted *P*-value < 10^{-3}.



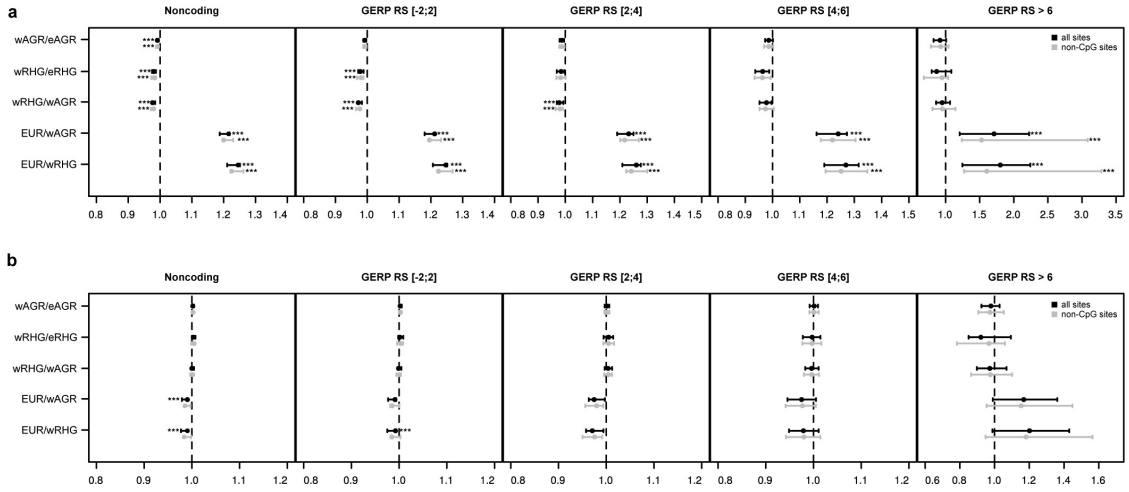
Supplementary Fig. 18 | Comparison of population differences in N_{hom} and $N_{alleles}$ across variants in OMIM genes associated to dominant and recessive diseases. Between-population ratios of the mean per-individual counts of **a**, homozygous genotypes (N_{hom}) and **b**, numbers of alleles ($N_{alleles}$). These ratios were computed with the observed data for 63,169 nonsynonymous (0-fold) mutations located in 12,453 genes reported by OMIM catalog as being responsible for either dominant or recessive diseases. Confidence intervals were calculated by bootstrapping 1,000 times by site. ***: non-adjusted *P*-value < 10^{-3}.



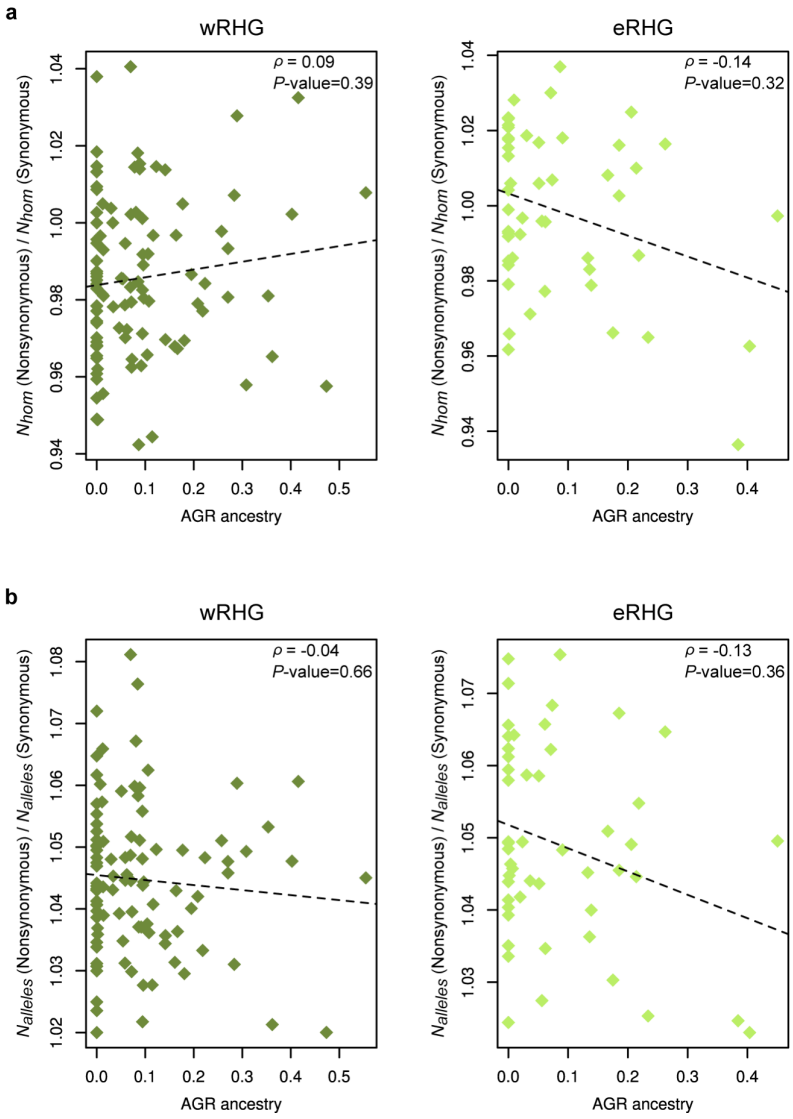
Supplementary Fig. 19 | Distribution of allele frequencies and comparison of observed N_{hom} and $N_{alleles}$ for loss-of-function (LOF) variants. **a,** Derived allele frequency spectra for each population for nonsynonymous LOF mutations. Between-population ratios of the mean per-individual counts of **b,** homozygous genotypes (N_{hom}) and **c,** numbers of alleles ($N_{alleles}$) of nonsynonymous LOF mutations. Confidence intervals were calculated by dividing the SNP data into 1,000 blocks and carrying out bootstrap resampling of sites 1,000 times. ***: non-adjusted *P*-value < 10^{-3}. We present here data including CpG sites because we found no homozygous genotypes for LOF non-CpG.



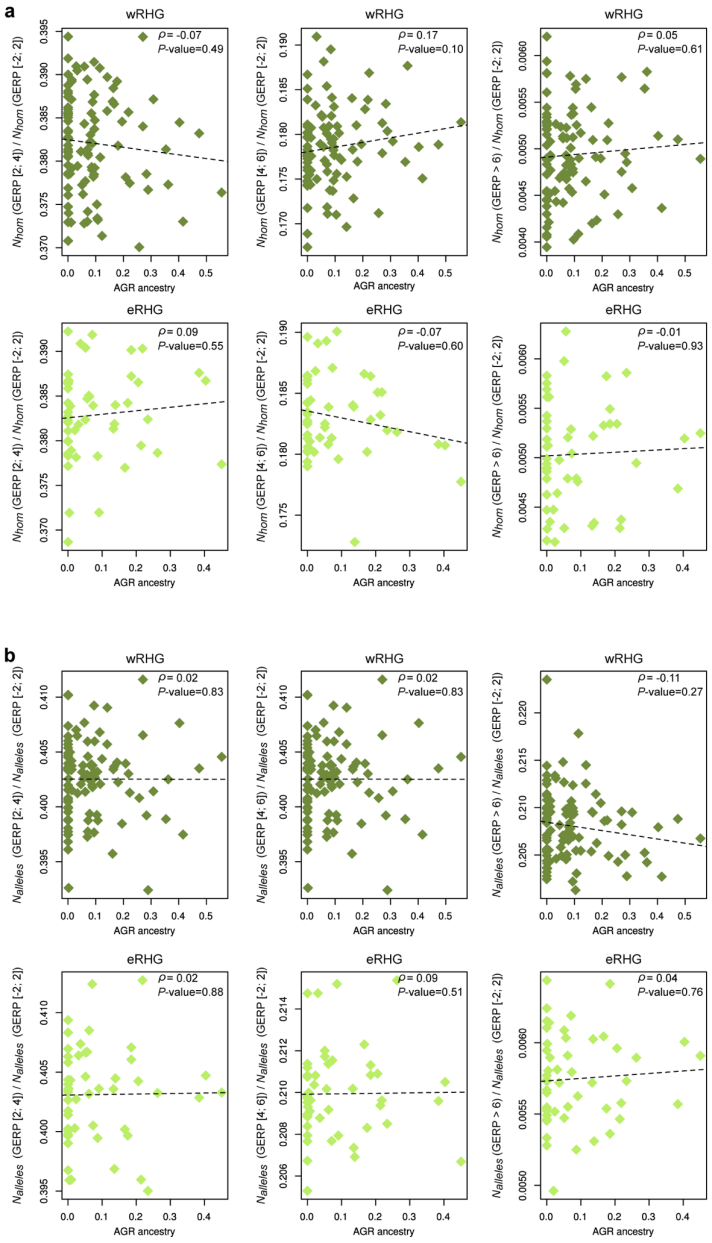
Supplementary Fig. 20 | Comparison of observed N_{hom} with simulated predictions between the studied populations. **a**, Between-population ratios of the mean per-individual counts of numbers of homozygotes (N_{hom}) that were simulated under additive (h=0.5, red), partially recessive (h=0.25, yellow) and fully recessive (h=0, blue) models of dominance for synonymous and nonsynonymous mutations, stratified in bins of increasing selection coefficients s. **b,** Between-population ratios of the mean per-individual counts of numbers of homozygotes (N_{hom}) for observed synonymous and nonsynonymous mutations involving all sites (black) and non-CpG sites (gray), and stratified in several GERP RS classes reflecting the severity of mutational effects: "neutral" [-2; 2], "moderate" [2; 4], "large" [4; 6], and "extreme" [> 6]. For simulated ratios, confidence intervals are the 0.025 and 0.975 quantiles of ratios across 100 simulations. For observed ratios, confidence intervals were calculated by dividing the SNP data into 1,000 blocks and carrying out bootstrap resampling of sites 1,000 times. The non-adjusted P-value testing for ratios different from one was never lower than 10^{-3}.



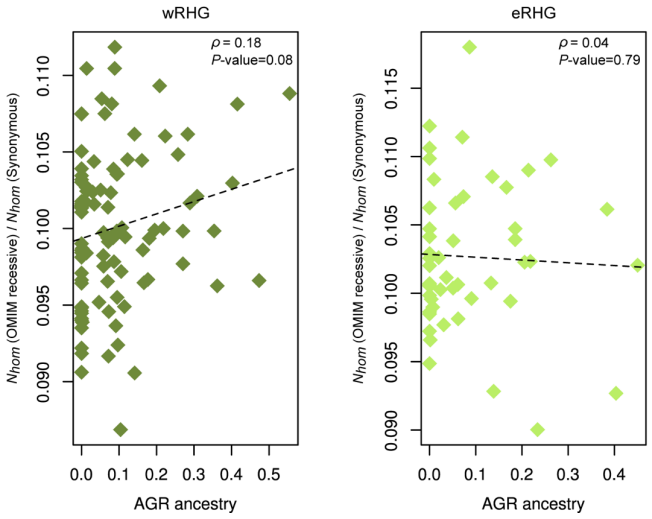
Supplementary Fig. 21 | Comparison of observed N_{hom} and $N_{alleles}$ for noncoding variants. Between-population ratios of the mean per-individual counts of noncoding **a,** homozygous genotypes (N_{hom}) and **b,** numbers of alleles ($N_{alleles}$). Counts have been performed for mutations in several GERP RS classes reflecting the severity of mutational effects: "neutral" [-2; 2], "moderate" [2; 4], "large" [4; 6], and "extreme" [> 6]. Confidence intervals were calculated by dividing the SNP data into 1,000 blocks and carrying out bootstrap resampling of sites 1,000 times. ***: non-adjusted *P*-value < 10^{-3}.



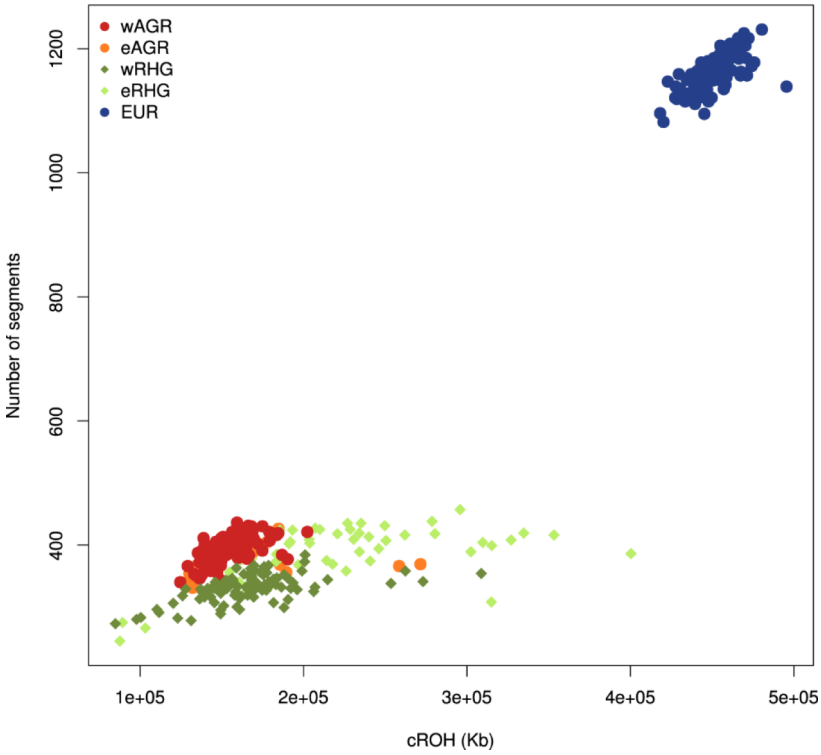
Supplementary Fig. 22 | Effects of AGR admixture on the nonsynonymous to synonymous ratio of N_{hom} and $N_{alleles}$ for RHG populations. The nonsynonymous to synonymous ratio of **a**, homozygous genotypes (N_{hom}) and **b**, number of alleles ($N_{alleles}$) was obtained. All variants considered exhibit no missingness and at least 5x coverage. Per-individual proportions of AGR ancestry were obtained from the ADMIXTURE results at K=4, producing the lowest cross-validation value. Dashed lines indicate the fit of a linear regression to the data, and P-values correspond to an association test using the non-parametric Spearman's rank correlation.



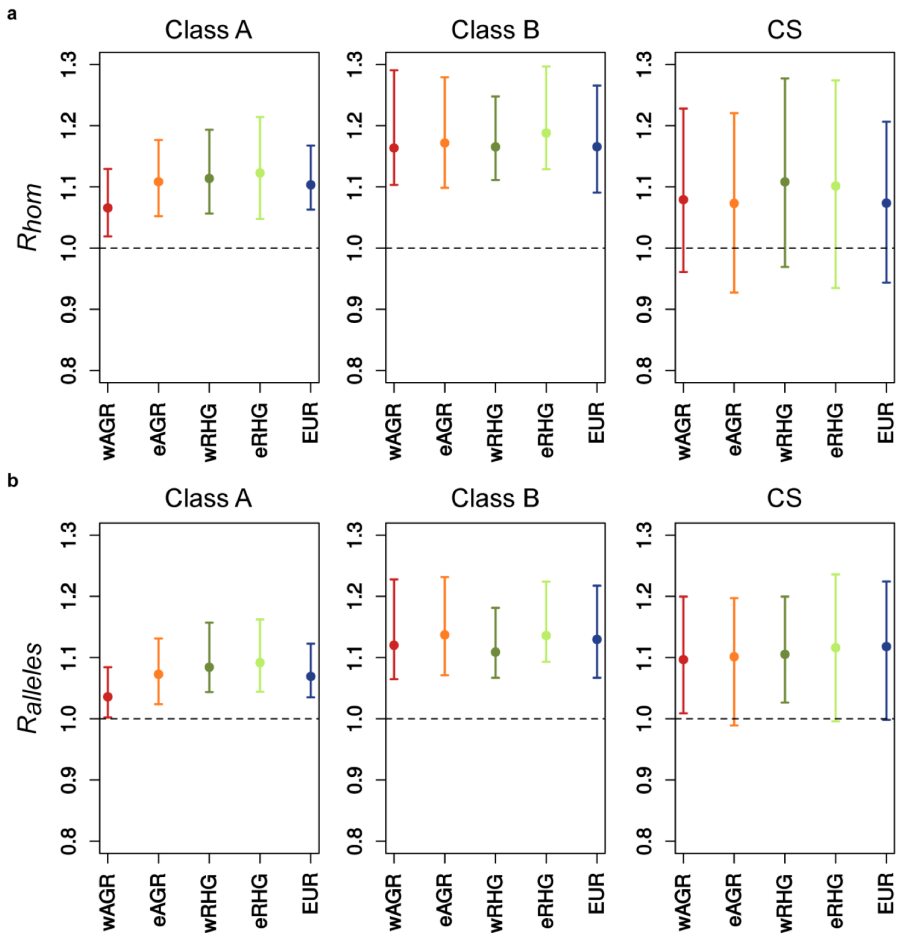
Supplementary Fig. 23 | Effects of AGR admixture in RHG populations on N_{hom} and $N_{alleles}$ ratios across bins of different GERP RS scores. The per-individual counts of **a**, homozygous genotypes (N_{hom}) and **b**, number of alleles ($N_{alleles}$) were obtained for several GERP RS classes reflecting the severity of mutational effects: "neutral" [-2;2], "moderate" [2;4], "large" [4;6], "extreme" [>6]. We present the ratios of N_{hom} and $N_{alleles}$ in "moderate" [2;4], "large" [4;6], "extreme" [>6] bins with counts obtained for the "neutral" [-2;2] bin. All variants considered exhibit no missingness and at least 5x coverage. Per-individual proportions of AGR ancestry were obtained from the ADMIXTURE results at K=4, producing the lowest cross-validation value. Dashed lines indicate the fit of a linear regression to the data, and P-values correspond to an association test using the non-parametric Spearman's rank correlation.



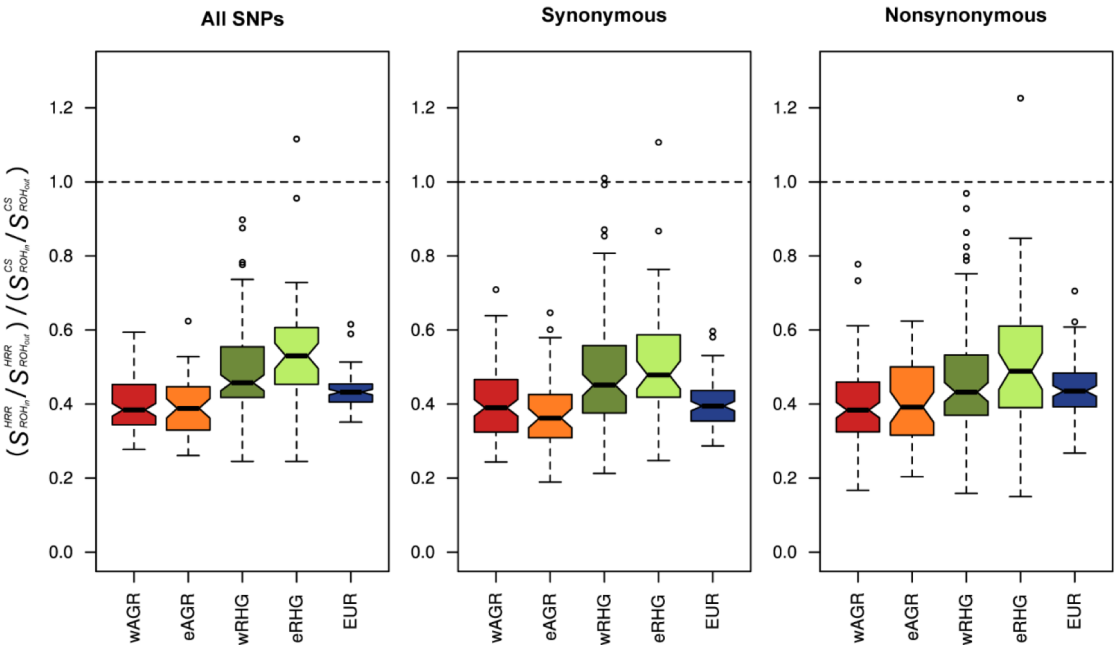
Supplementary Fig. 24 | Effects of AGR admixture in RHG populations on N_{hom} nonsynonymous to synonymous ratios in OMIM genes associated to recessive diseases. The per-individual counts of homozygous genotypes (N_{hom}) were obtained for synonymous (4-fold) and nonsynonymous (0-fold) sites located in genes reported by OMIM catalog as being responsible for recessive diseases. All variants considered exhibit no missingness and at least 5x coverage. Per-individual proportions of AGR ancestry were obtained from the ADMIXTURE results at K=4, producing the lowest cross-validation value. Dashed lines indicate the fit of a linear regression to the data, and P-values correspond to an association test using the non-parametric Spearman's rank correlation.



Supplementary Fig. 25 | Runs of homozygosity (ROH). ROH in wRHG, eRHG, wAGR, eAGR and EUR populations. Cumulative ROH (cROH) is reported per individual against the total number of observed ROH segments.



Supplementary Fig. 26 | Effects of ROH and recombination on the per-individual nonsynonymous to synonymous ratio of alleles. Ratios of nonsynonymous to synonymous counts of **a**, homozygous genotypes (R_{hom}) and **b**, number of alleles ($R_{alleles}$) inside ROH class A (panel "Class A"), inside ROH class B (panel "Class B"), and inside coldspots of recombination (panel "CS"), over corresponding ratios of nonsynonymous to synonymous counts of homozygous genotypes and number of alleles outside ROH class A, outside ROH class B and inside hotspots of recombination. Confidence intervals were calculated by bootstrapping by site 100 times.



Supplementary Fig. 27 | ROH segments and recombination rates. Ratios of the per-individual number of variant sites inside both hotspots of recombination (HRR) and ROH segments (classes A and B together) (S^{HRR}_{ROHin}) and the number of variant sites inside hotspots of recombination but outside ROH segments (S^{HRR}_{ROHout}), over the ratios of the per-individual number of variant sites inside both coldspots of recombination (CS) and ROH segments (S^{CS}_{ROHin}) and the number of variant sites inside coldspots of recombination but outside ROH segments (S^{CS}_{ROHout}), calculated for different functional categories. The bold black line indicates the median, and box limits the 25th and 75th percentiles, of distributions. The dashed line indicates the expected ratio, under the assumption of no enrichment in HRR or CS in ROH regions.

Supplementary Table 1 | Population description, samples and exome quality metrics. Per-population mean values and standard deviation (SD) are given for several metrics assessing exome sequencing quality.

Group	Population	Country	Language	N	Source genotyping	Breadth of coverage 5x (%)	Depth of coverage (BAM)	Depth of coverage (VCF)	Missingness (%)	Heterozygosity (%)	Heterozygosity (%) ± 4 SD
wAGR	Bapunu	Gabon	Niger-Congo, Narrow Bantu, Northwest, B43	45	Patin et al., 2017 (EGAS00001002078)	94 ± 1.3	62.4 ± 13.3	45.5 ± 6.9	0.06 ± 0.1	6.3 ± 0.09	3.9 - 8.9
wAGR	Nzebi	Gabon	Niger-Congo, Narrow Bantu, Northwest, B52	55	Patin et al., 2017 (EGAS00001002078)	95.2 ± 0.8	86.8 ± 14.3	51.4 ± 5.1	0.03 ± 0.04	6.3 ± 0.09	3.9 - 8.9
eAGR	BaKiga	Uganda	Niger-Congo, Narrow Bantu, Central, J.10	50	Perry et al., 2014 (EGAS00001000908)	92.4 ± 2.6	58.2 ± 9.0	40.7 ± 6.7	0.4 ± 0.7	6.2 ± 0.05	5.8 - 6.5
wRHG	Baka	Cameroon/ Gabon	Niger-Congo, Ubangi, Ngbaka, Baka-Gundi	70/ 30	Fagny et al., 2015 (EGAS00001001066) Patin et al., 2014 (EGAS00001000605)	94 ± 2.0	78.1 ± 29.6	47.4 ± 10.4	0.2 ± 0.5	6.4 ± 0.2	5.7 - 7.2
eRHG	BaTwa	Uganda	Niger-Congo, Narrow Bantu, Central, J.10	50	Perry et al., 2014 (EGAS00001000908)	92.7 ± 2.3	60.8 ± 15.5	41.4 ± 7.8	0.3 ± 0.5	6.2 ± 0.06	5.5 - 7
EUR	Belgian	Belgium	Indo-European, Italic/Germanic, French/Dutch	100	Quach et al., 2016 (EGAS00001001895)	92.7 ± 1.4	57.6 ± 10.0	39.7 ± 4.1	0.2 ± 0.4	4.7 ± 0.1	4.4 - 4.9

Supplementary Table 2 | Summary statistics of genetic diversity. The 95% confidence intervals are reported in brackets and were obtained by bootstrapping by site 1,000 times.

CpG status	Site Class	Population	θ_W (%)	θ_π (%)	Tajima's *D*	$\theta_{\pi(0\text{-fold})}/\ \theta_{\pi(4\text{-fold})}$
All sites	0-fold	wAGR	0.0727 [0.0721,0.0734]	0.0395 [0.0389,0.0401]	-1.478 [-1.499,-1.456]	0.303 [0.297, 0.311]
		eAGR	0.0622 [0.0616,0.0629]	0.0391 [0.0384,0.0397]	-1.272 [-1.297,-1.247]	0.304 [0.297, 0.312]
		wRHG	0.0594 [0.0590,0.0602]	0.0405 [0.0399,0.0411]	-1.029 [-1.060,-1.010]	0.305 [0.298, 0.312]
		eRHG	0.0487 [0.0481,0.0493]	0.0390 [0.0384,0.0397]	-0.679 [-0.708,-0.647]	0.304 [0.297, 0.313]
		EUR	0.0473 [0.0467,0.0478]	0.0300 [0.0295,0.0307]	-1.178 [-1.210,-1.143]	0.315 [0.306, 0.325]
	4-fold	wAGR	0.191 [0.189,0.193]	0.130 [0.128,0.132]	-1.032 [1.064,1.001]	-
		eAGR	0.173 [0.171,0.175]	0.128 [0.126,0.131]	-0.889 [-0.924,-0.854]	-
		wRHG	0.1645 [0.163,0.167]	0.133 [0.131,0.135]	-0.624 [-0.667,-0.597]	-
		eRHG	0.140 [0.138,0.142]	0.128 [0.126,0.131]	-0.294 [-0.334,-0.256]	-
		EUR	0.109 [0.107,0.110]	0.0952 [0.0932,0.0974]	-0.395 [-0.442,-0.345]	-
Non-CpG	0-fold	wAGR	0.0523 [0.0518,0.0530]	0.0309 [0.0306,0.0314]	-1.330 [-1.360,-1.299]	0.349 [0.339,0.360]
		eAGR	0.0458 [0.0452,0.0464]	0.0306 [0.0300,0.0312]	-1.131 [-1.161,-1.096]	0.350 [0.341,0.362]
		wRHG	0.0433 [0.0429,0.0440]	0.0316 [0.0310,0.0322]	-0.877 [-0.916,-0.848]	0.352 [0.342,0.363]
		eRHG	0.0364 [0.0359,0.0369]	0.0304 [0.0298,0.0310]	-0.562 [-0.601,-521]	0.351 [0.340,0.363]
		EUR	0.0342 [0.0337,0.0347]	0.0237 [0.0231,0.0242	-0.999 [-1.039,-0.957]	0.359 [0.346,0.374]
	4-fold	wAGR	0.119 [0.117,0.121]	0.0884 [0.0864, 0.0904]	-0.832 [-0.874,-0.790]	-
		eAGR	0.109 [0.107,0.111]	0.0873 [0.0852, 0.0894]	-0.678 [-0.726,-0.631]	-
		wRHG	0.104 [0.102,0.106]	0.0897 [0.0877, 0.0917]	-0.433 [-0.488,-0.393]	-
		eRHG	0.0890 [0.0874, 0.0907]	0.0864 [0.0844, 0.0884]	-0.0976 [-0.152,-0.0451]	-
		EUR	0.0683 [0.0669, 0.0696]	0.0659 [0.0639, 0.0677]	-0.113 [-0.181,-0.0462]	-

Supplementary Table 3 | Difference in log likelihood between branching models under different assumptions for migration.

Submodel	No. Parameters	EUR-first	RHG-first	AGR-first
No migration	16	-2336.3*	-2777	-2847.9
1 migration epoch	20	-316.1*	-371.8	-353.7
2 migration epochs	24	-4	0*	-28.8

We modelled migration as continuous and occurring at different rates between AGR and RHG and between AGR and EUR populations. We assumed that migration can be asymmetric (*i.e.*, having separate parameters for either direction of migration) and that migration rate can change at a specific time. We fitted the three different branching models (EUR-first, RHG-first, AGR-first) under the assumption of no migration (No migration), migration that does not change in time (1 migration epoch), and migration that changes at the time of split between wAGR and eAGR (2 migration epochs). The difference in likelihood between each model and the best-fitting model are reported. For each submodel category, we highlight with a star (*) the best branching model.

Supplementary Table 4 | Inferred parameters under demographic models fitted to the five populations jointly. The demographic models were fitted to the ten pairwise (2D) SFSs of eAGR, wAGR, eRHG, wRHG and EUR. The 95% confidence intervals and significance for differences in likelihood between models was assessed by bootstrapping by site 100 times. Non-CpG sites were used for the estimations.

		Point estimates			95% Confidence Intervals		
Parameter type	**Parameter**	**EUR-first**	**RHG-first**	**AGR-first**	**EUR-first**	**RHG-first**	**AGR-first**
Times	T_{ANC}	173420	186006	190501	[118074,297992]	[137117,454858]	[123865,416585]
	T_{DIV1}	130761	135140	139809	[49086.6,165978]	[57594.0,259243]	[49475.5,244198.9]
	T_{DIV2}	97585	85318	116638	[30323,134807]	[40061,130188]	[37227,141043]
	T_{EUR}	14152	21750	12470	[12035,34789]	[13192,41894]	[13102,42759]
	T_{RHG}	19749	18270	18154	[10499,27454]	[7156.5,29889]	[3693.2,36273]
	T_{AGR}	10846	9367	6728	[4377.6,13882]	[3579.3,12447]	[2038.7,13462]
Population sizes	N_{ANC}	13822	14427	14093	[12940,15258]	[12797,15099]	[12775,15429]
	N_{HUM}	30371	21128	14301	[17915,69216]	[6085.1,51574]	[6918.2,64757]
	N_{aEUR}	4315	3065	4655	[632.9,5106]	[111.0,4680]	[101.7,4553]
	N_{aRHG}	23711	17543	25770	[5962.8,50618]	[11070,36659]	[10526,43979]
	N_{aAGR}	13695	18423	16670	[7715.7,22614]	[9881.9,25167]	[9915.6,25196]
	N_{EUR}	35323.5	25989	33750	[18380,43016]	[17054,43946]	[17557,41670]
	N_{wRHG}	5259	4570	4495	[2742,7199]	[1930,6544]	[1167,6760]
	N_{eRHG}	2642	2461	2335	[1422,3774]	[892.3,3617]	[493.8,3997]
	N_{wAGR}	42715	41570	43296	[31146,57149]	[24996,50840]	[23981,53294]
	N_{eAGR}	30467	25461	20064	[11107,38106]	[10270,34412]	[6448.3,34121]
Migration rate (ancient epoch)	$2\times Nm_{aAGR\rightarrow aEUR}$ $_{HUM\rightarrow aEUR}$ $_{HUM\rightarrow aRHG}$ $_{HUM\rightarrow aAGR}$	0.38 0.38 - -	0.21 - 1.2 -	0.24 - - 0.88	[0.014,0.50] [0.014,0.50] - -	[0.0040,0.36] - [0.040,12] -	[0.0024,0.48] - - [0.081,8.9]
	$2\times Nm_{aEUR\rightarrow aAGR}$ $_{aEUR\rightarrow HUM}$ $_{aRHG\rightarrow HUM}$ $_{aAGR\rightarrow HUM}$	2.7 6.0 - -	2.1 - 2.5 -	2.6 - - 2.3	[0.85,4.2] [1.4,16] - -	[0.041,3.3] - [0.043,9.0] -	[0.39,4.6] - - [0.014,13]
	$2\times Nm_{aRHG\rightarrow aAGR}$	2.9	2.5	3.3	[0.037,9.5]	[0.072,6.9]	[0.081,8.9]
	$2\times Nm_{aAGR\rightarrow aRHG}$	0.84	3.9	5.8	[0.016,9.2]	[0.040,12]	[0.026,17]
Migration rate (recent epoch)	$2\times Nm_{eAGR\rightarrow EUR}$	0.36	0.054	0.33	[0.034,1.0]	[0.19,0.88]	[0.014,0.94]
	$2\times Nm_{EUR\rightarrow eAGR}$	5.6	6.5	12.0	[0.17,15]	[1.8,19]	[0.90,21]
	$2\times Nm_{wRHG\rightarrow wAGR}$ $_{eRHG\rightarrow eAGR}$	20 17	18 11	34 16	[8.3,33] [5.2,17]	[8.1,37] [4.1,16]	[6.5,46] [4.5,15]
	$2\times Nm_{wAGR\rightarrow wRHG}$ $_{eAGR\rightarrow eRHG}$	17 8.3	18 9.5	20 10	[13,21] [6.3,12]	[15,22] [7.4,12]	[14,21] [6.5,12]
	LogL	-917440.8	-917436.8	-917465.6			

Supplementary Table 5 | Estimated magnitude of population size changes for the three branching models. The 95% confidence intervals are shown in brackets and were obtained by bootstrapping by site 100 times.

Population size epoch	Population size ratio	EUR-first	RHG-first	AGR-first
Epoch 2	N_{ANC}/N_{HUM}	0.46 [0.20,0.80]	0.68 [0.27,2.3]	0.99 [0.23,2.2]
Epoch 3	N_{HUM}/N_{aAGR}	2.2 [1.0,7.6]	1.2 [0.32,3.2]	0.86 [0.36,4.8]
	N_{HUM}/N_{aRHG}	1.3 [0.46,11]	1.2 [0.29,3.7]	0.55 [0.24,4.1]
	N_{HUM}/N_{aEUR}	7.0 [3.5,75]	6.9 [2.2,379]	3.1 [1.7,319]
Epoch 4	N_{aAGR}/N_{wAGR}	0.32 [0.14,0.65]	0.44 [0.21,0.88]	0.39 [0.19,1.1]
	N_{aAGR}/N_{eAGR}	0.45 [0.22,1.9]	0.72 [0.35,2.2]	0.83 [0.33,3.7]
	N_{aRHG}/N_{wRHG}	4.5 [0.96,13]	3.8 [2.1,14]	5.7 [1.8,33]
	N_{aRHG}/N_{eRHG}	9.0 [2.0,24]	7.1 [4.2,32]	11 [3.4,72]
	N_{aEUR}/N_{EUR}	0.12 [0.035,0.18]	0.12 [0.0058,0.15]	0.14 [0.0048,0.18]

Supplementary Table 6 | Unscaled parameter estimates for migration rates. The 95% confidence intervals were obtained by bootstrapping by site 100 times.

Parameter type	Parameter	EUR-first	RHG-first	AGR-first
Migration rates	$m_{aAGR\rightarrow aEUR}$ $m_{HUM\rightarrow aEUR}$ $m_{HUM\rightarrow aRHG}$ $m_{HUM\rightarrow aAGR}$	4.36×10^{-5} $[8.24\times10^{-6}, 6.66\times10^{-5}]$	3.48×10^{-5} $[5.67\times10^{-6}, 5.87\times10^{-5}]$	2.63×10^{-5} $[6.10\times10^{-6}, 5.97\times10^{-5}]$
	$m_{aEUR\rightarrow aAGR}$ $m_{aEUR\rightarrow HUM}$ $m_{aRHG\rightarrow HUM}$ $m_{aAGR\rightarrow HUM}$	9.91×10^{-5} $[2.40\times10^{-5}, 1.77\times10^{-4}]$	5.79×10^{-5} $[1.13\times10^{-6}, 1.10\times10^{-4}]$	7.88×10^{-5} $[8.33\times10^{-6}, 1.45\times10^{-4}]$
	$m_{eAGR\rightarrow EUR}$	5.10×10^{-6} $[5.78\times10^{-7}, 1.87\times10^{-5}]$	1.03×10^{-6} $[3.32\times10^{-7}, 2.22\times10^{-5}]$	4.82×10^{-6} $[3.37\times10^{-7}, 2.27\times10^{-5}]$
	$m_{EUR\rightarrow eAGR}$	9.22×10^{-5} $[2.92\times10^{-6}, 5.14\times10^{-4}]$	1.27×10^{-4} $[3.69\times10^{-5}, 8.33\times10^{-4}]$	3.00×10^{-4} $[1.55\times10^{-5}, 1.74\times10^{-3}]$
	$m_{aRHG\rightarrow aAGR}$	1.05×10^{-4} $[1.29\times10^{-6}, 5.87\times10^{-4}]$	6.70×10^{-5} $[1.53\times10^{-6}, 3.31\times10^{-4}]$	9.75×10^{-5} $[1.95\times10^{-6}, 4.11\times10^{-4}]$
	$m_{aAGR\rightarrow aRHG}$	1.77×10^{-5} $[3.48\times10^{-7}, 3.56\times10^{-4}]$	1.10×10^{-4} $[1.04\times10^{-6}, 4.14\times10^{-4}]$	1.13×10^{-4} $[4.98\times10^{-7}, 4.50\times10^{-4}]$
	$m_{wRHG\rightarrow wAGR}$ $m_{eRHG\rightarrow eAGR}$	2.28×10^{-4} $[9.08\times10^{-5}, 4.31\times10^{-4}]$	2.12×10^{-4} $[8.69\times10^{-5}, 5.39\times10^{-4}]$	3.88×10^{-4} $[7.35\times10^{-5}, 7.03\times10^{-4}]$
	$m_{wAGR\rightarrow wRHG}$ $m_{eAGR\rightarrow eRHG}$	1.57×10^{-3} $[9.66\times10^{-4}, 3.21\times10^{-3}]$	1.93×10^{-3} $[1.21\times10^{-3}, 4.94\times10^{-3}]$	2.22×10^{-3} $[1.17\times10^{-3}, 8.23\times10^{-3}]$

Supplementary Table 7 | Parameter estimates obtained with *DFE-α* and *∂a∂i*/Fit*∂a∂i*. We analysed the folded SFS with *DFE-α*, and the unfolded SFS with *∂a∂i*/ Fit*∂a∂i*. The demographic parameters were estimated for a three-epoch model of population size changes for both methods. The parameter P_{mis} corresponds to a scaling parameter used by *∂a∂i*/Fit*∂a∂i* to account for the effect of ancestral misidentification on the unfolded SFS. The N_w parameter corresponds to the weighted *N* across the 3-epoch model inferred by each method, calculated as described in Note 4. The selection parameters correspond to a gamma distribution of selection coefficients. The ratio u_{del}/u_{neu} corresponds to the ratio of the average fixation probability of a new deleterious mutation over the fixation probability of a neutral mutation, calculated as described in Note 5. The confidence intervals for *DFE-α* and *∂a∂i*/Fit*∂a∂i* were obtained by bootstrapping by site 100 times and are shown in brackets. Non-CpG sites were used.

	Demographic parameters							Selection parameters								
Method	**Population**	***N1***	***N2/N1***	***N3/N2***	***t2/N1***	***t3/N1***	***pmis***	***Nw***	***β***	***E(s)***	***E(Nes)***	***Nes* 0-1**	***Nes* 1-10**	***Nes* 10-100**	***Nes* >100**	***udel/uneu***
DFE-α	wAGR	100	1.7 [1.7,2]	2.35 [2,2.647]	0.53 [0.397,0.688]	0.062 [0.05,0.081]	-	112.4 [110,115.5]	0.133 [0.12,0.146]	8.3 [5.5,15.5]	934.3 [623.7,1741.2]	0.324 [0.313,0.336]	0.118 [0.107,0.126]	0.159 [0.141,0.175]	0.398 [0.384,0.42]	0.298 [0.288,0.312]
	eAGR	100	1.7 [1.7,2]	2.06 [1.5,2.353]	0.45 [0.304,0.607]	0.062 [0.05,0.084]	-	110.8 [108.8,114.5]	0.128 [0.105,0.144]	11.4 [5.8, 32.5]	1260.4 [631.9,3703.2]	0.33 [0.319,0.347]	0.112 [0.096,0.126]	0.15 [0.123,0.173]	0.408 [0.38,0.436]	0.306 [0.294,0.324]
	wRHG	100	3.5 [2.5,4.5]	0.86 [0.378,0.889]	5.48 [0.335,8.107]	0.16 [0.104,0.417]	-	237.4 [118.7,310.2]	0.106 [0.093,0.125]	19.8 [6.7,79.4]	4689 [1569.9,12197.1]	0.342 [0.327,0.353]	0.093 [0.084,0.108]	0.119 [0.104,0.145]	0.446 [0.422,0.465]	0.321 [0.304,0.332]
	eRHG	100	2 [2,4.5]	0.7 [0.267,0.8]	2.61 [0.262,5.932]	0.16 [0.099,0.351]	-	147.5 [111.2,214.9]	0.119 [0.101,0.139]	10.9 [4.2,46.8]	1614.7 [748.8,5752.1]	0.335 [0.325,0.353]	0.108 [0.091,0.122]	0.142 [0.115,0.166]	0.415 [0.387,0.454]	0.311 [0.299,0.331]
	EUR	100	0.5 [0.5,0.6]	8 [7.5,9]	0.39 [0.244,0.47]	0.05 [0.05,0.05]	-	85.8 [84.3,92.4]	0.172 [0.16,0.187]	1.9 [1.4,2.7]	163.5 [121.6,228.4]	0.328 [0.315,0.341]	0.161 [0.15,0.172]	0.231 [0.21,0.249]	0.28 [0.254,0.308]	0.296 [0.284,0.31]
∂a∂i/ Fit*∂a∂i*	wAGR	12462 [11914,12946]	1.76 [1.59,1.95]	2.04 [1.73,3.31]	0.552 [0.426,0.788]	0.054 [0.018,0.104]	0.022 [0.019,0.025]	14072 [13729,14411]	0.14 [0.127,0.156]	0.083 [0.048,0.153]	1166 [683,2114]	0.301 [0.286,0.311]	0.114 [0.105,0.126]	0.159 [0.14,0.178]	0.426 [0.405,0.444]	0.277 [0.261,0.287]
	eAGR	12248 [11551,12785]	1.73 [1.29,2.04]	1.53 [1.43,32.33]	0.512 [0.374,1.05]	0.09 [0.008,0.272]	0.031 [0.026,0.036]	13801 [13470,14117]	0.136 [0.118,0.152]	0.091 [0.05,0.217]	1261 [699,2990]	0.307 [0.294,0.321]	0.113 [0.1,0.125]	0.154 [0.131,0.176]	0.426 [0.406,0.454]	0.282 [0.27,0.299]
	wRHG	11980 [11354,12530]	2.91 [2.29,41.23]	0.39 [0.03,0.48]	0.616 [0.338,0.84]	0.114 [0.058,0.216]	0.023 [0.02,0.027]	14253 [13906,14591]	0.122 [0.108,0.138]	0.202 [0.099,0.457]	2882 [1431,6430]	0.311 [0.299,0.323]	0.1 [0.091,0.111]	0.133 [0.116,0.153]	0.455 [0.435,0.475]	0.289 [0.275,0.302]
	eRHG	11279 [10085,12146]	1.88 [1.67,7.26]	0.47 [0.12,0.52]	1.134 [0.448,2.01]	0.092 [0.038,0.226]	0.031 [0.026,0.036]	13664 [13327,13999]	0.125 [0.105,0.143]	0.162 [0.073,0.56]	2217 [1003,7548]	0.312 [0.298,0.325]	0.104 [0.09,0.118]	0.139 [0.114,0.162]	0.445 [0.415,0.476]	0.29 [0.274,0.304]
	EUR	83668 [15708,123465]	0.08 [0.02,0.41]	13.78 [9.86,18.78]	0.514 [0.404,1.006]	0.004 [0.002,0.024]	0.024 [0.02,0.033]	10345 [10058,10663]	0.175 [0.167,0.2]	0.018 [0.012,0.024]	191 [129,246]	0.318 [0.291,0.326]	0.157 [0.149,0.173]	0.227 [0.214,0.258]	0.299 [0.269,0.323]	0.287 [0.26,0.296]

Supplementary Table 8 | Three-epoch size change parameter estimates based on 1D SFS. The three-epoch demographic parameters were inferred by fitting to the 1D SFS of each population separately with *fastsimcoal2*.

Population	N_3	N_3/N_2	N_2/N_1	t_2/N_3	t_3/N_1
wAGR	36854	1.77	1.60	0.04	0.42
eAGR	37269	1.54	1.94	0.02	0.62
wRHG	11732	0.45	2.55	0.03	1.62
eRHG	12654	0.42	2.87	0.20	1.22
EUR	46612	25.67	0.19	0.01	0.03

Supplementary Table 9 | Comparison of the divergence times between AGR and RHG estimated from previous studies

Study reference	Nature of the genetic data	Number of polymorphic sites	Sample size	Populations	Method	Estimated time (years)	95%CI	Mutation rate: per site per generation	Generation time	Recalibrated time*	95%CI
Quintana-Murci (ref.[29])	Sanger sequencing of complete mtDNA genomes	~120	22	**wRHG:** Baka, Bakola; **wAGR:** Akele, Bateke, Bapunu, Eshira, Fang, Obamba, Shake	Coalescence time of L1c1 lineages	73,811	66,668–80,954	5.00×10^{-7}	29	-	-
Verdu (ref.[30])	Microsatellites	28	544	**wRHG:** Bezan, Baka, Kola, Koya; **wAGR:** Tikar, Nzime, Bangando, Fang, Akele, Teke, Nzebi, Kota, Tsogho, Ewondo	ABC	89,675	23,025–123,275	1.60×10^{-4}	25	-	-
Patin (ref.[31])	Sanger sequencing of autosomal DNA	340	124	**wRHG:** Baka, Biaka; **eRHG:** Mbuti, Twa; **wAGR:** Yoruba, Ngumba, Akele; **seAGR:** Chagga, Mozambicans	ABC	56,049	25,814–130,548	2.50×10^{-8}	25	119,516	55,045–278,374
Veeramah (ref.[32])	Sanger sequencing of autosomal DNA	~500	85	**wRHG:** Bakola, Biaka; **eRHG:** Mbuti; **wAGR:** Ngoumba; **seAGR:** Luhya, Shona	ABC	48,927	10,000–105,909	2.50×10^{-8}	25	104,330	21,324–225,835
Batini (ref.[33])	Sanger sequencing of complete mtDNA genomes	402	121	**wRHG:** Baka, Bakola, Babinga, Biaka, Mbenzele; **eRHG:** Mbuti; **wAGR:** Fang, Nzebi	ABC	70,866	51,789–106,388	2.80×10^{-7}	25	-	-
Hsieh (ref.[34])	Next-generation sequencing of whole genomes	1.6×10^{6}	16	**wRHG:** Baka, Biaka; **wAGR:** Yoruba	∂a∂i	*Model 1:* 155,671 *Model 2:* 89,645	139,661–164,280 85,503–91,725	2.35×10^{-8}	25	*Model 1:* 312,029 *Model 2:* 179,685	279,938–329,285 171,383–183,855
Malick (ref.[35])	Next-generation sequencing of whole genomes	~1.3×10^{6} / sample	2	**eRHG:** Mbuti; **wAGR:** Yoruba	MSMC	56,000	32,000–84,000**	1.25×10^{-8}	29	51,471	29,412–77,206
This study	Next-generation sequencing of whole exomes	24,794	400	**wRHG:** Baka; **eRHG:** BaTwa; **wAGR:** Nzebi, Bapunu; **seAGR:** BaKiga; **EUR:** Belgians	*fastsimcoal2*	135,140	57,594–259,243	1.36×10^{-8}	29	-	-

*Recalibrated times were obtained assuming a mutation rate $\mu=1.36\times10^{-8}$ and a generation time g=29

**Estimated times at which 25% and 75% of lineages of the two populations of interest are descended from the same ancestral population

Supplementary Table 10 | Clinically-relevant variants identified in the exome of African hunter-gathering and farming populations. List of variants annotated with the highest clinical significance in ClinVar database and their frequencies in each of the examined populations.

(provided as an excel file)

References

1. Perry, G.H. et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc Natl Acad Sci U S A* **111**, E3596-3603 (2014).
2. Quach, H. et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167**, 643-656 e617 (2016).
3. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
4. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009).
5. Excoffier, L. Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev* **12**, 675-682 (2002).
6. Malaspinas, A.S. et al. A genomic history of Aboriginal Australia. *Nature* **538**, 207-214 (2016).
7. Henn, B.M., Cavalli-Sforza, L.L. & Feldman, M.W. The great human expansion. *Proc Natl Acad Sci U S A* **109**, 17758-17764 (2012).
8. Pagani, L. et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am J Hum Genet* **91**, 83-96 (2012).
9. Hodgson, J.A., Mulligan, C.J., Al-Meeri, A. & Raaum, R.L. Early back-to-Africa migration into the Horn of Africa. *PLoS Genet* **10**, e1004393 (2014).
10. Eyre-Walker, A. & Keightley, P.D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* **26**, 2097-2108 (2009).
11. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713-719 (1962).
12. Keightley, P.D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251-2261 (2007).
13. Davydov, E.V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
14. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* **47**, 276-283 (2015).
15. Simons, Y.B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Curr Opin Genet Dev* **41**, 150-158 (2016).
16. Simons, Y.B., Turchin, M.C., Pritchard, J.K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220-224 (2014).
17. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).
18. Kirin, M. et al. Genomic runs of homozygosity record population history and consanguinity. *PLoS One* **5**, e13996 (2010).
19. Szpiech, Z.A. et al. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* **93**, 90-102 (2013).
20. Pemberton, T.J. et al. Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* **91**, 275-292 (2012).
21. Narasimhan, V.M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474-477 (2016).

22. Hussin, J.G. et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* **47**, 400-404 (2015).
23. Landrum, M.J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
24. Tishkoff, S.A. et al. The Genetic Structure and History of Africans and African Americans. *Science* **324**, 1035-1044 (2009).
25. Kostrikis, L.G. et al. A polymorphism in the regulatory region of the CC-chemokine receptor 5 gene influences perinatal transmission of human immunodeficiency virus type 1 to African-American infants. *J Virol* **73**, 10264-10271 (1999).
26. John, G.C. et al. Correlates of mother-to-child human immunodeficiency virus type 1 (HIV-1) transmission: association with maternal plasma HIV-1 RNA load, genital HIV-1 DNA shedding, and breast infections. *J Infect Dis* **183**, 206-212 (2001).
27. Astuto, L.M. et al. Searching for evidence of DFNB2. *Am J Med Genet* **109**, 291-297 (2002).
28. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
29. Quintana-Murci, L. et al. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A* **105**, 1596-1601 (2008).
30. Verdu, P. et al. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* **19**, 312-318 (2009).
31. Patin, E. et al. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* **5**, e1000448 (2009).
32. Veeramah, K.R. et al. 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 (2012).
33. Batini, C. et al. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Mol Biol Evol* **28**, 1099-1110 (2011).
34. Hsieh, P. et al. 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 (2016).
35. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).

5.3 Résumé des résultats et nouveautés

Afin de comprendre l'impact de la démographie sur l'efficacité de la sélection purificatrice chez les populations de Pygmées et de non-Pygmées, nous avons tout d'abord réévalué leurs modèles démographiques à partir des données de séquencage d'exome. Pour cela, nous avons testé trois modèles possibles pour la chronologie des divergences entre les populations de Pygmées, de non-Pygmées et d'européens (respectivement RHG-first, AGR-first et EUR-first). Ces estimations réalisées grâce à des méthodes de maximum de vraisemblance indiquent que la divergence des ces trois groupes de populations pourraient avoir eu lieu simultanément entre 85 000 et 140 000 ans ou peuvent indiquer l'existence à cette époque de populations ancestrales structurées sur le continent africain.

L'estimation de paramètres démographiques tels que les dates de divergence, les variations de tailles effectives de populations ainsi que les taux de migrations entre les groupes nous ont permis de montrer que les populations de Pygmées sont issues de populations ancestrales prospères dont la taille génétique était comparable à celles des ancêtres des agriculteurs (17 000 à 26 000 individus) et qu'un métissage existait déjà entre ces deux groupes il y a plus de 20 000 ans. Nos résultats indiquent également que quelque soit le modèle démographique considéré, les groupes de Pygmées ont vu leurs tailles de population diminuer d'environ 80% alors que les tailles effectives des populations d'agriculteurs ont triplé à la même époque. Cette période de variations démographiques considérables entre Pygmées et non-Pygmées a été accompagnée par une intensification des événements des migrations entre les deux groupes.

Afin d'évaluer l'efficacité de la sélection purificatrice dans les groupes de Pygmées et de non-Pygmées, nous avons tout d'abord estimé le DFE (Distribution of Fitness Effects) de chacune des populations qui permet d'évaluer la fraction de mutations dans différents intervalles de $N_e s$ parmi les nouvelles mutations qui apparaissent dans la population. Ces résultats ont ensuite été utilisés pour calculer la probabilité de fixation des allèles délétères dans les populations en comparaison aux allèles neutres. Ces résultats indiquent que les forces de sélection purificatrice et de dérive ont eu la même intensité dans toutes les populations, malgré leurs histoires démographiques récentes opposées.

Bien que le ratio des probabilités de fixation des allèles délétères et neutres ne soit pas différents dans les populations considérées à l'échelle de leur histoire évolutive globale, ce résultat ne nous renseigne pas sur les éventuelles variations de fardeau de mutations au cours du temps. Afin de répondre à cette question, nous avons obtenu par simulation l'évolution du fardeau de mutations délétères au cours du temps en intégrant l'histoire démographique des populations et leur DFE selon deux hypothèses de dominance : l'additivité (h=0.5) et la complète récessivité (h=0) des mutations délétères. Alors que le fardeau de mutations délétères est identique pour toutes les populations dans le cas de l'additivité des mutations délétères, on observe une forte augmentation du fardeau de mutations délétères récessives lors d'événements de réduction de tailles de population (Simons & Sella, 2016). Lorsqu'on ne considère pas d'événements de migration entre les populations, nos résultats indiquent une

augmentation substantielle du fardeau de mutations délétères récessives chez les Pygmées, suggérant que le métissage a compensé le fardeau de mutations délétères récessives accru chez ces populations.

Ces observations prédisent alors que l'hypothèse du modèle de dominance des mutations délétères constitue l'élément majeur permettant de conclure sur l'existence d'une différence de l'efficacité de sélection entre les populations. Afin de déterminer si les données empiriques soutiennent ou non l'existence d'une différence d'efficacité de sélection purificatrice entre les populations, nous avons mesuré empiriquement le fardeau de mutations délétères entre les groupes de Pygmées et de non-Pygmées. Pour cela nous avons estimé les fardeaux de mutations délétères de chacune des populations en calculant le nombre de mutations non-synonymes prédites comme délétères par individu. Ces estimations empiriques n'ont montré aucune différence de fardeaux de mutations délétères et donc d'efficacité de la sélection entre les populations de Pygmées et de non-Pygmées. De plus, un travail de simulations nous a permis de montrer que le nombre de mutations non-synonymes par individu n'est pas compatible avec un modèle strictement récessif des mutations délétères dans le génome humain. Ces travaux mettent également en évidence la difficulté d'isoler des mutations fortement délétères dans le génome d'individus adultes sains.

L'ensemble de ces résultats a permis de mettre en évidence l'absence de différences dans l'efficacité de la sélection purificatrice de populations de Pygmées et non-Pygmées aux histoires démographiques différentes, en considérant un modèle de dominance additif ou récessif, ainsi que le rôle bénéfique que le métissage peut jouer sur la diminution du fardeau de mutations délétères récessives.

Chapitre 6

Résultats 2 : Contribution des mécanismes de sélection naturelle à l'adaptation génétique des chasseurs-cueilleurs

6.1 Contexte

Le phénotype "Pygmée" se caractérise par une taille moyenne de moins de 150 cm (Perry & Dominy, 2009 ; Perry et al., 2014 ; Migliano et al., 2007 ; Rozzi et al., 2015) et a été observée dans plusieurs populations occupant les forêts tropicales d'Afrique, d'Amérique du sud ou d'Asie du sud-est (Perry & Dominy, 2009) qui vivent de la chasse et de la cueillette. Ce phénotype particulier résulterait d'une adaptation convergente de ces populations aux conditions de vie dans les forêts tropicales où la faible stature pourrait conférer de nombreux avantages en terme de thermorégulation (L. Cavalli-Sforza, 1986), de déplacement dans un milieu dense (J. M. Diamond, 1991), une adaptation aux ressources alimentaires limitées (Shea & Bailey, 1996) et un âge de reproduction avancé pour compenser une espérance de vie réduite (Migliano et al., 2007).

Chez les populations de Pygmées d'Afrique, des perturbations de la voie métabolique liée à l'hormone de croissance GH1-IGF1 ont été montrées comme potentiellement impliquées dans ce phénotype adaptatif bien qu'aucun variant génétique n'ai été identifié dans la séquence codante de ces gènes (Hattori et al., 1996 ; Merimee et al., 1989 ; Baumann et al., 1989 ; Bozzola et al., 2009). Une étude récente suggère cependant que des variants introniques dans *GHR* et *IGF1* seraient sous sélection positive, et associés avec la taille chez les Pygmées (Becker et al., 2013). De nombreux travaux se sont concentrés sur les signatures moléculaires de sélection naturelle dans différents groupes de Pygmées, à l'ouest et à l'est de l'Afrique, afin d'identifier les bases génétiques de ce phénotype adaptatif (Jarvis et al., 2012 ; Lachance et al., 2012 ; Perry et al., 2014 ; Hsieh et al., 2016a ; Pickrell et al., 2009). Ces travaux ont pour la plupart utilisé des méthodes de détection de balayages sélectifs, et deux d'enter elles ont identifié une large région de 15 Mb sur le chromosome 3 qui présente des signatures de sélection positive récente (Jarvis et al., 2012 ; Lachance et al., 2012). Ces études ont corrélé plusieurs gènes de cette région génomique avec la faible stature des populations de Pygmées comme *DOCK3*, qui est associé avec la taille dans les populations non-africaines,

et *CISH*, qui régule à la fois l'activité du récepteur à l'hormone de croissance et certaines fonctions immunitaires(Jarvis et al., 2012). D'autres études ont également mis en évidence un haplotype à forte fréquence chez les populations de Pygmées mais à moindre fréquence dans d'autres populations africaines autour du gène *HESX1* impliqué dans le développement de l'antéhypophyse et qui régule la production d'hormone de croissance (Lachance et al., 2012). Et d'autres travaux ont mis en évidence des gènes impliqués dans l'ostéogénèse et l'homéostasie osseuse comme *EPHB1* et *FLNB* et qui expliqueraient la faible stature des individus Pygmées (Hsieh et al., 2016a).
Ces résultats suggèrent que la faible stature des populations de Pygmées d'Afrique centrale serait due à des événements de balayages sélectifs forts affectant un faible nombre de loci. Par ailleurs, il a été établi que la taille adulte des Pygmées est fortement corrélée à leur niveau de métissage avec les populations non-Pygmées où les individus les plus métissés présentent les statures les moins faibles (Becker et al., 2011) et 16 régions génomiques ont été identifiées chez les populations de l'est comme associées à la taille (Perry et al., 2014), suggérant alors que ce phénotype est sous sélection polygénique, correspondant à une faible augmentation en fréquence d'un grand nombre de loci ayant un faible impact sur le phénotype.

Afin d'évaluer en détails les mécanismes de sélection naturelle ainsi que les gènes et les fonctions biologiques impliqués dans le phénotype des populations de Pygmées, notre étude s'est concentrée sur l'analyse de séquences exoniques et de données de génotypage provenant de 566 individus Pygmées et non-Pygmées répartis dans 14 populations à l'ouest et à l'est de l'Afrique centrale. En premier lieu, nous avons recherché les signaux de sélection positive par balayage sélectif partagé entre les groupes de Pygmées puis nous avons réalisé un second scan afin d'identifier les signaux spécifiques à chacune des populations. Ensuite, afin d'évaluer la contribution d'autres mécanismes de sélection positive dans l'adaptation des populations de Pygmées, nous avons cherché à identifier des signaux de sélection polygénique et des événements de métissage adaptatif.

6.2 Article 2

Exploring the history of genetic adaptation of African rainforest hunter-gatherers using whole-exome sequencing data

Marie Lopez[a,b,c], Alexandre Gouy[d], Hélène Quach[a,b,c], Christine Harmant[a,b,c], Patrick Mouguiama-Daouda[e,f], Jean-Marie Hombert[f], Alain Froment[g], George H. Perry[h], Luis B. Barreiro[i], Paul Verdu[j], Laurent Excoffier[d], Etienne Patin[a,b,c], Lluís Quintana-Murci[a,b,c]

[a]Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France; [b]Centre National de la Recherche Scientifique (CNRS) UMR2000, 75015 Paris, France; [c]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, 75015 Paris, France; [d]Computational and Molecular Population Genetics lab, Institute of Ecology and Evolution, University of Bern, CH-3012 Bern, Switzerland; [e]Laboratoire Langue, Culture et Cognition (LCC), Université Omar Bongo, BP 13131 Libreville, Gabon; [f]CNRS UMR 5596, Dynamique du Langage, Université Lumière-Lyon 2, 69007 Lyon, France; [g]Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, 75005 Paris, France; [h]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA; [i]Université de Montréal, Centre de Recherche CHU Sainte-Justine, H3T 1C5 Montréal, Canada; [j]CNRS UMR7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Sorbonne Paris Cité, Paris 75016, France

Corresponding author: quintana@pasteur.fr (L.Q.-M.)

Keywords: population genetics; positive selection; immunity; pygmies; farmers; Africa

Abstract

The study of the occurrence and mechanisms of genetic adaptation can be highly informative for understanding the nature of the selective pressures that have acted upon human populations occupying different habitats. Foraging lifestyle in tropical rainforests has exposed humans to specific ecological, nutritional and pathogenic constraints, which have likely driven their adaptive history and distinctive phenotypes. To identify shared and population-specific signatures of genetic adaptation to rainforest environments and lifestyles, we report an analysis of high-coverage exome sequences of 566 African rainforest hunter-gatherers and neighboring farmers from 14 different populations across central African, combined with new data from 40 whole genome sequences. Using this dataset, we evaluated the relative contribution of classic sweeps, polygenic adaptation and adaptive admixture characterizing the evolutionary history of these populations. Although we found limited evidence for shared recent selective sweeps among rainforest hunter-gatherer groups, most signals detected involved functions related to immunity and hormone related pathways. Furthermore, our analyses of gene subnetworks detected signals of polygenic adaptation in populations of rainforest hunter-gatherers targeting primarily FoxO and Jak-STAT biological pathways. Collectively, our analyses suggest that both immune functions and transduction of growth factors or insulin-like growth factor signaling are functions that have most likely participated in the adaptation of rainforest hunter-gatherers to their unique environments, through mechanisms that involved polygenic adaptation.

Introduction

African rainforest hunter-gatherers, which have historically been grouped under the derogative term “pygmies” (Hewlett, 2014; Perry and Dominy, 2009), represent the largest living group of active hunter-gatherers worldwide, with an estimated total number ranging from 300,000 to 500,000 people (Jackson, 2005). They live all along the equatorial belt of Africa, which extends more than 4,000 kilometers from the Congo basin to Lake Victoria and encompasses a large variety of biomes vastly dominated by dense tropical rainforests bordered by woodlands, grasslands and savannas. Today, rainforest hunter-gatherers consist in at least 15 different ethnolinguistic groups, live in ten central African countries (Angola, Gabon, Cameroon, Equatorial Guinea, Central African Republic, Republic of the Congo, Democratic Republic of the Congo, Uganda, Rwanda, and Burundi), and are broadly subdivided into two groups that reflect their geographic location (Perry and Verdu, 2016). Western rainforest hunter-gatherers, which inhabit the Congo Basin, include populations such as the Baka, Aka, Koya or Bongo, whereas Eastern rainforest hunter-gatherers live nearby the Ituri rainforest and comprise groups such as the Asua, Sua, Efe or Twa. Rainforest hunter-gatherers are traditionally mobile, live in huts in the rainforest, and maintain socio-economic relationships with neighboring Bantu-speaking farmers, which have adopted a sedentary lifestyle in the last 5,000 years (Diamond and Bellwood, 2003; Patin et al., 2017; Philipson, 2005). These two population groups differ not only in their subsistence patterns, but also in their ecologies and exposure to environmental pressures and diseases (Hewlett, 2014; Ohenjo, 2006).

The iconic feature of rainforest hunter-gatherers is the so-called ‘pygmy’ phenotype, which is characterized by a small body size and an average adult height below 150 cm (Migliano et al., 2007; Perry and Dominy, 2009; Perry et al., 2014; Rozzi et al., 2015), and has been observed in several populations traditionally hunting and gathering in tropical rainforests

(Perry and Verdu, 2016; Perry and Dominy, 2009). These traits are unlikely to result from sexual selection (Becker et al., 2012), which suggests an adaptive role of reduced body proportions to humid forest conditions and probably extensive convergent evolution (Bergey et al., 2018; Perry and Dominy, 2009; Perry et al., 2014). Many hypotheses have been proposed regarding the adaptive role of short stature to the hunter-gatherer lifestyle in tropical forests such as improved thermoregulation (Cavalli-Sforza, 1986), improved mobility (Diamond, 1991), adaptation to low caloric alimentation (Shea and Bailey, 1996), and earlier reproduction compensating shorter lifespans (Migliano et al., 2007). Metabolic and physiological studies in African rainforest hunter-gatherers (Baumann et al., 1989; Capute et al., 1969; Hattori et al., 1996; Merimee et al., 1989; Merimee et al., 1987; Rimoin et al., 1969) have reported no differences in serum levels of human growth hormone (HGH) but low serum levels of insulin-like growth hormone 1 (IGF-1) whose production is induced by HGH (Baumann et al., 1989; Merimee et al., 1989; Rimoin et al., 1969). In addition, reduced level of IGF-1 in African hunter-gatherers is accompanied by a reduced expression of the IGF-1 receptor (*IGF1R*) as well as a severe reduction of the expression of the HGH receptor (*HGHR*) although no causal variants have been identified in the sequence of these genes (Baumann et al., 1989; Becker et al., 2013; Bozzola et al., 2009; Hattori et al., 1996).

Over the last decade, genomic studies of rainforest hunter-gatherers and farmers have greatly increased our knowledge of their past demography, in terms of population splits, size changes and gene flow. The split between these two population groups is one of the most ancient divergence times estimated in humans, dating back to the Late Pleistocene between 90,000 and 135,000 years ago (Hsieh et al., 2016; Lopez et al., 2018). A later split between rainforest hunter-gatherer populations according to their geographic locations (west and east) occurred during the Holocene, around 20,000 years ago, followed by the separation of western groups of hunter-gatherers around 3,000 years ago (Batini et al., 2011; Lopez et al.,

2018; Patin et al., 2009; Verdu et al., 2009; Verdu et al., 2013). The distinctive subsistence patterns of these populations have had a marked impact on their demographic history; rainforest hunter-gatherers and farmers have recently experienced population collapses and expansions, respectively, accompanied by increased gene flow (Lopez et al., 2018; Patin et al., 2009; Patin et al., 2014; Veeramah et al., 2012).

While the demographic history of rainforest hunter-gatherers and farmers is increasingly well characterized, their history of natural selection has received less attention. Recent data has shown that despite marked differences in past demographic regimes between these two groups, negative selection has been equally efficient to remove deleterious variants from these populations, who present an almost identical additive mutation load (Do et al., 2015; Lopez et al., 2018)). With respect to their history of positive selection, most studies have focused on height, and identified genes and functions that have likely participated to the adaptive nature of the 'pygmy' phenotype including insulin-signaling pathway (Jarvis et al., 2012; Migliano et al., 2013; Pickrell et al., 2009), thyroid pathway and pituitary development (Lachance et al., 2012; Lopez Herraez et al., 2009), and skeletal development and bone homeostasis (Hsieh et al., 2016; Mendizabal et al., 2012). Despite the major challenge imposed by the high microbial burden of the rainforest (Ohenjo, 2006), only a few genes involved in immune functions have been identified as candidates for positive selection (Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012). Furthermore, all studies have used so far ascertained genotyping data (Amorim et al., 2015; Jarvis et al., 2012; Lopez Herraez et al., 2009; Mendizabal et al., 2012; Migliano et al., 2013; Perry et al., 2014) or few individuals and/or populations (Hsieh et al., 2016; Lachance et al., 2012), so the history of genetic adaptation of African rainforest hunter-gatherers remains to be explored in further detail.

Here, we aimed to obtain new insights into how rainforest hunter-gatherers from central Africa have genetically adapted to their specific and challenging habitat, and to decipher

which evolutionary mechanisms — including classic sweeps, polygenic adaptation and admixture — have participated in such selective processes. To do so, we generated and analyzed whole-exome sequencing data for a total of 566 individuals from an extensive set of 14 populations of rainforest hunter-gatherers and neighboring farmers from both western and eastern central Africa, as well as 40 whole-genome sequences for a subset of these individuals. This dataset was used to detect signatures of selective sweeps characterizing both rainforest hunter-gatherers as a whole as well as populations-specific cases of genetic adaptation. We also explored the genes and biological functions that have been most likely targeted by polygenic selection as well as investigated how adaptive admixture has participated in the adaptation of rainforest hunter-gatherers to their specific environments.

Results

Population structure and admixture. We generated and analyzed whole-exome sequencing data from a collection of ethnologically well-defined populations of rainforest hunter-gatherers (RHG) and agriculturalists (AGR), to obtain a comprehensive view of how RHG in general, or individual RHG populations, have genetically adapted to their unique ecologies. These populations include the RHG Bezan from Cameroon, Baka from Cameroon and Gabon, BaKoya, BaBongo from the center, the east, and the south of Gabon, as well as BaTwa from Uganda. These RHG were compared to an extensive set of traditional AGR Bantu-speaking populations, which include the Fang, Galoa, Tsogo, Shake, Bapunu, and Nzebi from Gabon as well as the BaKiga from Uganda (Fig. 1A; Supplementary Table 1). We generated new whole-exome sequencing data for 266 individuals, which were analyzed together with 300 individuals originally reported in Lopez et al., 2018, yielding a final dataset of 566 individuals from seven RHG populations and seven AGR populations distributed across central Africa (Fig. S1; Table S1).

We detected substantial levels of genetic differentiation (F_{ST}) among the different RHG groups, estimated from a total of 728,154 exome variants, which were equivalent to those observed between RHG and AGR groups (mean F_{ST} between RHG = 0.02, RHG-AGR F_{ST} = 0.02, and F_{ST} between AGR = 0.008 ; Fig. 2B). To increase the genome-wide breath of polymorphic sites available, we combined the exome variants with genome-wide SNP genotyping data obtained for the same individuals (Fig. S1; Table S1) (Fagny et al., 2015; Patin et al., 2017; Patin et al., 2014). Using a subset of 412,869 independent SNPs, which were pruned for linkage disequilibrium ($r^2 < 0.5$), we then estimated ancestry proportions with the clustering algorithm ADMIXTURE (Alexander et al., 2009). At K=5, which exhibited the lowest cross-validation error (Fig. S2), RHG populations separated into four main ancestry clusters (Fig. 1C), although all groups displayed similar proportions of ancestry from AGR

populations (~8% in Baka [s.d.=11%], ~9% in Bezan [s.d.=13%], ~4% BaKoya [s.d.=5%], ~9%, BaBongo Center [s.d.=9%] and ~9% in BaTwa [s.d.=12%]), with the exception of two groups of BaBongo, who presented very high AGR ancestry (~43% in BaBongo East [s.d.=11%] and ~24% in BaBongo South, [s.d.=17%]). AGR populations presented similar proportions of overall RHG ancestry (RHG ancestry estimated at *K*=2), varying from ~2% to ~9%, except the Shake who presented higher levels of RHG ancestry (mean ~22%, [s.d.=2%]). These results indicate that RHG are a subdivided population whose genetic diversity can be accounted for by four different ancestry components, and have received varying degrees of gene flow from neighboring AGR populations.

Shared signatures of local adaptation in African hunter-gatherers

For all subsequent analyses, we increased further marker density by imputing SNPs in the dataset combining exome sequencing and genotyping, using two different reference panels (Fig. S1). To maximize the imputation efficacy in our African dataset, we generated new whole-genome sequences from 20 RHG and 20 AGR individuals of Gabon (Baka and Nzebi, respectively) at ~6× coverage, providing with a total of 17,687,206 variants along the genome. We combined this new dataset with a second ethnically diverse reference panel (1000G Phase 3;(Auton et al., 2015)) and imputed our data, yielding a final dataset of 9,129,128 high quality variants with MAF > 1% in all 566 individuals, after imputation (Fig. S1, S3 and S4). To identify genomic signatures of classic sweeps shared by all RHG (i.e., shared events of adaptation across RHG), we analyzed the RHG Bezan, Baka, BaKoya, BaBongo Center and BaTwa separately, and compared them to western or eastern AGR populations (i.e., western Fang, Nzebi, Bapunu, Shake, Tsogo, and Galoa and eastern BaKiga) using a population of European ancestry as an outgroup (EUR: Table S1 and S2;(Quach et al., 2016)). Note that the two RHG populations presenting a high degree of AGR ancestry (i.e.,

BaBongo East and BaBongo South) were not included in the RHG selection analyses unless otherwise stated.

We considered two intra-population tests (|ΔiHH|(Grossman et al., 2010) and |iHS|(Voight et al., 2006)) and three inter-population statistics (PBS(Yi et al., 2010), XP-EHH and |ΔiHHD|(Sabeti et al., 2007)), which were combined into a Fischer score (Fsc, see Methods). Composite scores have been shown to increase power and minimize the detection of false positives (Grossman et al., 2013; Grossman et al., 2010; Pybus et al., 2015). Furthermore, we previously showed that Fsc is not affected by the inclusion of correlated statistics and displays similar power than the composite of multiple signals (CMS,(Grossman et al., 2013)) (Deschamps et al., 2016; Patin et al., 2017). We calculated the Fsc of each SNP and defined candidate regions for positive selection as 100kb windows presenting an enrichment in SNPs with outlier Fsc scores (i.e., top 1% of the empirical genome-wide distribution). As higher proportions of outlier SNPs are more likely to be found in windows containing a reduced number of sites, our top hits were defined as windows presenting an enrichment in outlier SNPs within the top 0.5% of the empirical distribution in five different bins of SNPs density ([50;96]; [97;124]; [125;153]; [154;193]; [194;880]).

In order to validate our approach, we searched for canonical examples of selection in AGR populations. Candidate regions for positive selection in wAGR include a locus that encompasses the *TLR5* gene (chr1:223,214,464-223,314,464; P_{value} = 0.0003), which was previously detected under selection in the Yoruba from Nigeria (Grossman et al., 2013) and is involved in the response to bacterial flagellin. Another well-established positively-selected gene was detected in eAGR, in a genomic region that includes the *DARC* gene, variation in which is known to impart almost complete resistance to *Plasmodium vivax* malaria (McManus et al., 2017) (chr1:159,164,464-159,264,464, P_{value} = 0.004). To restrict our analysis to loci that included variants with a likely functional impact, we then required candidate genes to

include a high scoring SNP (Fsc > 10), in addition to being found in a window enriched for outlier SNPs (top 0.5% of the empirical distribution; Fig. S5 and S6). Using these stringent filters, we confirmed the strong signal of positive selection found in the *LCT*/*MCM6* region, involved in lactase persistency in adulthood, in eAGR (chr2:136,510,797-136,610,797, P_{value} = 0.0006) (Patin et al., 2017). We also found signals in AGR populations at *FOLH1*, shown to regulate folate intake in the intestine and presumably linked to the adaptation to high-UV environment (Cummings et al., 2017; Jablonski, 2012; Jones et al., 2018) (Fig. 2B).

As to signals of adaptation in RHG, we first observed that the number of shared candidate regions (windows enriched in outlier SNPs) between RHG and AGR is higher than expected ($P_{value} < 10^{-4}$, resampling 10,000 times), in agreement with a shared history of positive selection for these African populations, both in western (wAGR and wRHG) and eastern (eAGR and eRHG) central Africa ($P_{value} < 10^{-4}$, for both comparisons) (Fig. 2A). We found a total of 36 candidate regions (windows or contiguous windows) that were shared by at least two RHG populations (Fig. 2C; Table S3). Remarkably, we did not find signals of selection shared by all five populations of RHG, adding to an emerging body of evidence suggesting that the pygmy phenotype result from parallel evolution (Hsieh et al., 2016; Jarvis et al., 2012; Perry and Dominy, 2009). Nevertheless, we identified a large region of 31 Mb under selection on the chromosome 3, previously reported as a "pygmy" specific signal (Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012), in western RHG (Fig. 2D). This selection signal was particularly extended in the Bezan from North Cameroon, while shorter in the Baka from Cameroun and Gabon, and in the BaBongo Center from Gabon (Fig. S5 and S6). Several genes involved in growth hormone stimulation, association with height, pituitary development and bone homeostasis lie in this cluster, including *CISH*, *MAPKAPK3*, *DOCK3* (Jarvis et al., 2012), *HESX1* (Lachance et al., 2012), and *ARHGEF3* (Fig. 2D). Stringent filters to define candidate genes (windows enriched in outlier SNPs and Fsc > 10 in the gene body) detected

MAPKAPK3, a mitogen activated protein kinase central to the regulation of several signaling pathways mediating inflammation, and *ARHGEF3*, encoding a Rho guanine exchange factor, as containing a high scoring SNP in the Baka and BaBongo Center, respectively (Fig. 2D; Fig. S5). Importantly, *MAPKAPK3* is shown to interact directly with the hepatitis C virus (HCV) and its silencing is associated with HCV infectivity levels in cells (Ngo et al., 2013). Furthermore, its interplay with *MAPKAPK2* is crucial for *STAT3* activation in fine-tuned response to bacterial infection (Ehlting et al., 2011; Moens et al., 2013). With respect to *ARGHEF3,* it is associated with platelet function in humans and knock-out mice for this gene present increased mean platelet volume (Zou et al., 2017).

Furthermore, it is interesting to note that the chromosome 3 candidate region in RHG harbors multiple SNPs identified by GWAS as significant hits for several immune-related traits ("Macrophage inflammatory protein 1b levels", "Inflammatory Bowel disease"," Ulcerative colitis"), metabolism ("HDL cholesterol levels", "Blood protein levels") and life history traits ("Menarche") (Table S4). These observations are collectively consistent with a history of positive selection acting in the ancestor of western RHG, with subsequent drift or additional selection having acted upon each population after their divergence (Fig. S7).

When applying stringent filters to define additional genome-wide candidate genes for selection (windows enriched in outlier SNPs and Fsc > 10 in the gene body), we highlighted several genes shared by RHG populations and absent from AGR (Fig. 2B; Fig. S5 and S6). Among them, we identified *POLR3E* (RNA Polymerase III Subunit E) involved in virus sensing, as well as *EEF2K* (eukaryotic elongation factor 2 kinase) shared by BaBongo Center and Baka from Gabon (Carter-Timofte et al., 2018). Interestingly, the *EEF2K* kinase plays a major role in cell resistance to caloric restriction, especially glucose restriction (Leprivier et al., 2013), and was previously found in the vicinity of a top hit associated with height in Baka, Bezan and BaKola RHG (Jarvis et al., 2012) (Fig. 2B).

Finally, several genes previously suggested to be linked to the "pygmy" phenotype did not pass our threshold of significance (Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012). For example, we found suggestive signals of positive selection in the genomic regions containing the *OBSCN* gene (Hsieh et al., 2016) in both the Bezan and BaTwa, but corresponding windows were not significantly enriched in outlier SNPs (top 5% of the distribution, with fractions of outlier SNPs of 0.2 and 0.16 respectively). Likewise, *FLNB* (Hsieh et al., 2016; Lachance et al., 2012), which was detected as a candidate in the Bezan RHG, also showed suggestive but non-significant signals among both wAGR and eAGR (top 5% of the distribution, with fractions of outlier SNPs of 0.08, 0.11 and 0.07 respectively), thus suggesting that this signal is not RHG-specific. Collectively, our results provide with a comprehensive view of adaptive signatures in an extensive set of RHG populations, and highlight a limited sharing of candidate genes among RHG, suggesting that RHG adaptation is likely to rely on complex signals of positive selection involving multiple genes and parallel evolution.

Specific signatures of local adaptation in African hunter-gatherers

To explore recent population-specific signatures of positive selection in RHG groups, we next compared each group of western RHG to all other groups of RHG and used western AGR as outgroup (Fig. 3; Table S2). Remarkably, using our stringent criteria (top 0.5% windows enriched in outliers and containing at least one high scoring SNP Fsc > 10), we found a large number of candidate genes per population that are involved in pathways regulating the activation of immunity, insulin resistance, thyroid hormone or central pathways regulating several of these functions, suggesting an interplay between pathways participating to the adaptive phenotype of RHG (Fig. 3; Table S5). For example, we detected *PIAS1* among the Baka; *PIAS1* is a key regulator in inflammation responses of innate immunity protein through

the repression of *STAT1* (Zhang et al., 2012). Animal knock-out models of this gene have shown that the absence of *PIAS1* amplifies the inflammatory cascade and macrophage migration in adipocytes which in turn leads to insulin resistance (Liu et al., 2004; Liu et al., 2015). In this line, we found additional immunity genes with pleiotropic effects on insulin resistance among the Baka, including *PPARD*, which regulates insulin sensitivity and intestinal immunity (Lee et al., 2006; Tanaka et al., 2014) (Fig. 3; Table S5). Likewise, in the BaBongo Center, we detected *PTGDS*, shown to regulate mast cell regulation (Taketomi et al., 2013) as well as insulin resistance and early hypothalamic pituitary-adreanal axis in mice (Evans et al., 2013), and *SERPINA12*, which is involved in insulin resistance and inflammation (Heiker, 2014) (Fig. 3; Table S5).

We also found several genes involved in thyroid functions such as *THRB*, the thyroid hormone receptor beta, in the Baka; *IGSF10*, involved in gonadotrophin-releasing hormone resulting in delayed puberty (Howard and Dunkel, 2018) in the BaBongo Center, and genes involved in growth-hormone receptors such as the *IGF2BP3* insulin-like growth factor mRNA-binding protein 3 repressing the translation of *IGF2* (Panebianco et al., 2017) in BaKoya (Fig. 3; Table S5). Moreover, we identified an extensive set of candidate genes involved in immunity and inflammation pathways alone. For example, we found *IFNGR2*, the interferon gamma receptor 2 responsible for signal transduction and activation of the Jak-STAT pathway (Rosenzweig et al., 2004; Soh et al., 1994) in the Baka, *SERPINA5* a complement element regulating inflammation and coagulation (Rajeevan et al., 2015) and *MAP3K1* involved in the progression of inflammatory cytokine in type 2 diabetes (Torkamandi et al., 2016) in the BaBongo Center. In addition, we found *LRBA* involved in inflammation (Kostel Bal et al., 2017), and *TXNRD3* which is involved in parasitic helminthiasis and increased parasitic clearance capacity of myeloid cells (Williams et al., 2013) in the BaKoya (Fig. 3; Table S5). Finally, we found genes linked to muscle formation

such as *MYO1D* involved in myoblast differentiation in BaBongo Center (Tapscott, 2005) (Fig. 3; Table S5).

Collectively, these results provide evidence that positive selection has targeted an abundant set of pleiotropic genes in western groups of RHG all involved in similar functions, thus suggesting the contribution of multiple biological pathways to the adaptation of RHG populations.

Signals of polygenic adaptation in rainforest hunter-gatherer populations

While our scans for selection provide with complementary information on the adaptive history of RHG populations, classic selective sweeps are known to be rare in human populations (Hernandez et al., 2011) and other processes of adaptation might be more likely to explain how human populations have adapted to the rainforest (Jarvis et al., 2012; Perry et al., 2014). In light of this, we next aimed to detect events of polygenic adaptation, resulting from subtle changes in allele frequencies across many loci, all contributing to variation in an adaptive selected trait or pathway (Pritchard et al., 2010). To do so, we applied a method detecting enrichments in selection scores for gene subnetworks involved in several biological pathways (Gouy et al., 2017) (Methods). We examined the selection scores of 4,527 genes related to 301 pathways from the KEGG database (Kyoto Encyclopedia for Genes and Genomes;(Kanehisa et al., 2017)), and assigned selection scores to genes based on the $-\log10(P_{value})$ of outlier enrichment in the genomic window they overlap. To obtain an overview of polygenic signals shared across RHG populations, we computed gene scores from our first selection scan comparing RHG to AGR populations (Table S2; Table S6). Although only subnetwork enrichments in "Olfactory transduction" passed multiple testing correction (Methods), we found surprisingly high levels of sharing in pathways enriched in selected signals ($P_{value} < 0.05$) between RHG populations (Fig. 4; Table S6). Indeed, the

western Bezan, BaBongo Center and eastern BaTwa all exhibited signals of polygenic selection in "Jak-STAT signaling pathway", which is crucial for transducing signal from cytokines, hormones and growth factor (Dodington et al., 2018) (Table S6). The Jak-STAT pathway has a well-established role in immune cell activation and modulates a multitude of metabolic processes including adiposity, energy expenditure, glucose tolerance, growth and insulin sensitivity. Remarkably, we found that *IL4* is part of the subnetwork enrichment in Jak-STAT for all three populations, and *IL13* is found in western Bezan and eastern BaTwa, both cytokines playing a central role in the protection against intestinal parasites by inducing a T-helper 2 cell response (Finkelman et al., 2004).

We also found signals of polygenic selection related to the "FoxO signaling pathway" in western RHG Bezan, Baka and BaKoya (Table S6). This pathway, which acts downstream of the IGF-1 signaling pathway, integrates insulin signaling with glucose and lipid metabolism. Specifically, we found that *FOXO6*, playing a pivotal role in regulating glucose and lipid homeostasis in response to fasting, is present in the FoxO subnetworks enriched for selection signals in all three populations (Kim et al., 2011; Lee and Dong, 2017). Moreover, *STAT3*, which appears in the enriched FoxO subnetwork found in the Baka and BaKoya, regulates insulin sensitivity by acting on lipolysis, gluconeogenesis, growth and inflammation (Dodington et al., 2018). Additional subnetwork enrichments related to development and immunity were detected in some RHG populations including "Wnt-signaling pathway" in the BaTwa, "HTLV-1 infection" in the BaKoya, and "NOD-like receptor signaling pathway" and "Natural killer cell mediated toxicity" in the Bezan.

Collectively, our results provide evidence that polygenic selection has participated in the adaptive history of RHG populations, targeting primarily pathways involved in signal transduction of growth factors and immune homeostasis, suggesting a crosstalk between these pathways participating to their adaptive phenotype.

Adaptive admixture in admixed rainforest hunter-gatherers

To evaluate whether adaptive variation in RHG might have been introduced through admixture with neighboring AGR populations, we scanned the genome of two groups of highly admixed RHG, the BaBongo South and BaBongo East (Fig 1C), for regions presenting both selection signals and unusually high levels of AGR ancestry (Fig. 5A). We performed local ancestry inference with RFMix (Maples et al., 2013) (Methods) in the two BaBongo populations, using as parental populations wAGR and wRHG individuals with the lowest RHG and AGR ancestry proportions, respectively (Methods; Fig. S8). We found a very a strong correlation between genome-wide ancestry proportions obtained with RFMix (Maples et al., 2013) and ADMIXTURE (Alexander et al., 2009) (Fig. S9). We did not find regions of significant excesses (> 3 s.d. from genome-wide average) of either AGR of RHG ancestry in admixed RHG, probably owing to our relatively small sample sizes, limiting the power to detect ancestry blocks (Bhatia et al., 2014) or possibly because of undetectable AGR ancestry in RHG resulting from >5,000-old admixture events (Hsieh et al., 2016; Lopez et al., 2018) (Fig. 1C). However, we detected one region, which presents an excess of AGR ancestry equal or higher than the genome-wide mean + 2 s.d (chr20:30,311,795-44,111,795), that is centered on the *GHRH* gene, playing a major role in stimulating growth hormone release from the pituitary gland (Ranke and Wit, 2018) although this gene presents no molecular signature of selection. (Fig. 5B). We also found a slight increase in AGR ancestry at the *FOLH1* locus in admixed RHG, which we identified as a positively-selected candidate gene in AGR populations (Fig. 5A: Fig. S5).

Based on local ancestry patterns, we next searched for enrichments of AGR or RHG ancestry in subnetworks of genes in biological pathways, using the same method applied for the search of polygenic adaptation (Gouy et al., 2017) (Methods). Although few of these

enrichments were significant after multiple testing correction, we consistently found that several pathways previously identified as being under polygenic selection in RHG populations were also enriched in RHG ancestry in the two admixed RHG populations ($P_{value} < 0.05$; Fig 5C; Fig. 4). These pathways include the "FoxO signaling pathway" (subnetwork of 39 genes $P_{value} = 0.009$ for RHG ancestry and 10 genes $P_{value} = 0.04$ in for AGR ancestry) and "Jak-STAT signaling pathway" (subnetwork of 47 genes, $P_{value} = 0.007$). We also found a pathway related to "Salmonella infection" enriched in RHG ancestry in our set of admixed RHG (subnetwork of 22 genes, $P_{value} = 0.01$) (Fig. 5C). Altogether, these results indicate that pathways previously found as targets of polygenic selection across multiple RHG populations are also enriched in RHG ancestry in the admixed RHG groups, strongly supporting the notion that these pathways have contributed to the adaptation of RHG populations to rainforest environments.

Discussion

Although several genome-wide scans of selection have been previously performed in African rainforest hunter-gatherers, most have focused on investigating the genetic basis of the 'pygmy' phenotype in Africa (Amorim et al., 2015; Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012; Lopez Herraez et al., 2009; Mendizabal et al., 2012; Migliano et al., 2013; Perry et al., 2014) and have used a few individuals and/or populations as representing African "pygmies". In this study, we have studied a large sample of different populations of rainforest hunter-gatherers located all along the central African belt and utilized an integrative approach incorporating scans for classic sweeps, detection of polygenic selection and local ancestry estimation. In doing so, we have identified several functions that may have contributed to the adaptive phenotypes of RHG populations in central Africa.

Genome scans for classic sweeps show little sharing across RHG populations, consistent with their marked levels of genetic differentiation (Batini et al., 2011; Lopez et al., 2018; Patin et al., 2009; Verdu et al., 2009; Verdu et al., 2013). Nevertheless, in agreement with previous analyses (Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012), we identified a large region of 31 Mb exhibiting strong signatures of selection on the chromosome 3 that is associated with GWAS traits related to innate immunity and lipid levels. Our analyses indicate that when considering populations separately, this region displays the strongest signatures of selection in the Bezan RHG, with some more genomically-restricted signals in the Baka and BaBongo Center. Although fine mapping of the candidate target for positive selection is challenging, particularly for this large gene-rich genomic region, the selection signal is maximal in two genes, *MAPKAPK3*, whose expression is linked to the regulation of both viral and bacterial infection, and *ARGHEF3*, associated to mean platelet volume. These results extend previous findings about the phenotypic importance of this selected region and its role in the adaptation of various RHG populations. Indeed, the substantial level of sharing

of this region supports at least one ancient event of selection in the ancestral population of western RHG, probably linked to infection and wound healing functions, a signal that may have been later amplified, or lost, depending on the demographic history of each of these RHG populations(Lopez et al., 2018).

Although candidate genes for positive selection show little overlap among RHG, owing to different demographic pasts and statistical power, our results indicate that they consistently relate to similar functions. Genes involved in immunity, growth factor transduction signal, insulin in particular, as well as thyroid hormone are consistently found under polygenic selection across RHG populations. Many of these pathways have independently been proposed as underlying the genetic basis of the 'pygmy' phenotype (Hsieh et al., 2016; Jarvis et al., 2012; Lachance et al., 2012). However, our results extend previous findings by identifying candidate genes that are key regulators of several pathways. Selection signals on genes regulating both immunity and insulin sensitivity, such as *PIAS1*, *PPARD*, *PTGDS* and *SERPINA12*, raise the possibility that highly pleiotropic functions have been selected in RHG, which might be or not directly linked to stature. Emerging evidence reveals that variation in linear growth depends on the balance between proliferation and senescence of chondrocytes and this process can be affected by multiple processes including inflammation and serum level of thyroid hormone (Sederquist et al., 2014). Moreover, several studies have highlighted the reciprocal relationship between proinflammatory cytokines, which govern the quality and the amplitude of immune responses, and the regulation of the growth hormone and IGF-1 axis (Smith, 2010). In this line, multiple studies showed that the activation of the immune system reduces IGF-1 sensitivity (Maggio et al., 2013; Wolters et al., 2017; Wong et al., 2010) in multiple tissues and ultimately could have more important effects on health than direct effects of the growth hormone and IGF-1, even though the regulation of IGF-1 expression during pathogenesis remains complex and relatively unexplored.

We examined the occurrence of polygenic selection in all RHG populations by searching for subnetwork enrichment in signatures of selection among biological pathways. Our results clearly indicate similar pathways enriched for selection across RHG populations, specifically in “Jak-STAT signaling pathway” and “FoxO signaling pathway” raising the possibility of parallel polygenic adaptation acting on these functions. In agreement with signatures of classic sweeps, both pathways appear to regulate immune functions, as well as core metabolic processes such as signal transduction. More specifically, enriched Jak-STAT subnetworks include two interleukines playing a role in parasitic clearances (Finkelman et al., 2004) and similar functions are found under selection in the BaKoya. Notably, among all infectious diseases affecting central African populations, parasitic diseases are those with the highest prevalence in RHG, with respect to AGR populations (Ohenjo 2006). These findings clearly suggest that this high load imposed by parasites has participated in the evolutionary history of RHG populations, and informs us about past exposures of RHG to pathogens. Additionnally, the FoxO pathways participates directly to the signal tranduction of IGF-1 and is involved in response to stress, cell growth, glycometabolism and regulation of lifespan (Becker et al., 2010; Eijkelenboom and Burgering, 2013; Lee and Dong, 2017). Enrichment in selection signals in this pathway involved *STAT3* that is also part of the Jak-STAT pathway and controls lipolysis, gluconeogenesis, growth and inflammation, and *FOXO6* playing a role in the homeostasis of glucose levels during fasting periods. These results are consistent with the signal of selection on *EEF2K* shared by several RHG populations, previously found near one of the strongest associations with height in an RHG sample composed of Baka, BaKola and Bezan RHG (Jarvis et al., 2012), and involved in cell resistance during nutrient deprivation. Altogether, these results highlight the importance of caloric restriction as a cause of adaptation of RHG, through different selective mechanisms.

Altogether, our study provides evidence that changes in immunity and insulin related pathways have been under polygenic selection in several groups of hunter-gatherers, possibly because of high pathogenic exposure and the limited availability of food resources in the rainforest, thus causing pleiotropic effects that could involve reduced height. Dissecting the genetic architecture of these traits in African populations will help characterize the molecular basis of human adaptation to the rainforest.

Material and Methods

Population and individual selection. A total of 695 individuals were analyzed, including 277 newly generated and 317 already published exome sequences from rainforest hunter-gatherer (RHG) and agriculturalist (AGR) populations of western and eastern central Africa, together with 101 individuals of European ancestry (Fig. S1; Table S1). Informed consent was obtained from all participants, which was overseen by the institutional review board of Institut Pasteur, France (2011-54/IRB/7), the Comité National d'Ethique du Gabon (0016/2016/SG/CNE), the University of Chicago (IRB 16986A) and Makerere University Uganda (IRB 2009-137).

Exome sequencing and quality. We generated new exome data for 277 African samples (Table S1) and processed these data together with 101 European individuals from Quach et al 2016 and 317 African samples from Lopez et al 2018. All samples were sequenced with the Nextera Rapid Capture Expanded Exome Kit, which delivers 62Mb of genomic content per individuals, including exons, untranslated regions and micros RNAs. Using the GATK Best Practices recommendations (Van der Auwera et al., 2013), read pairs were first mapped onto the human reference genome (GRCh37) with Burrows-Wheeler Aligner version 0.7.7 (Li and Durbin, 2009) and reads duplicating the start position of another read were marked as

duplicates with Picard Tools version 1.94 (http://broadinstitute.github.io/picard/). We used GATK version 3.5 (DePristo et al., 2011) 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'. From the 1,005,574 sites detected we removed indels and removed SNPs that (1) were located on the sex chromosomes, (2) were not biallelic, (3) were monomorphic in our samples, (4) had a depth of coverage lower than 5×, (5) had a genotype quality score (GQ) lower than 20, (6) presented missingness above 15%,(7) presented a Hardy-Weinberg test $P_{value} < 10^{-6}$ in at least one of the populations. As criteria to remove low-quality samples, we required a total genotype missingness lower than 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 additional individuals presenting heterozygosity levels 4 s.d higher than their population average. The application of these quality-control filters resulted in a final exome dataset consisting of 671 individuals, among which 268 were new, and 728,154 SNPs (Fig. S1).

Genotyping and quality-controls. In addition to exome sequencing, we downloaded the genotyping data of the same 671 individuals from Quach et al 2016, Patin et al 2014, Patin et al 2017 and Fagny et al 2015 and genotyped 21 additional BaBongo from the center of Gabon (BaBongo Center) with the Illumina HumanOmniExpress array for 730,525 SNP markers (Fig. 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 a SNP call rate <95%. We kept only 599,559 SNPs common to different genotyping chips. We removed a total of 53 C/G or A/T SNPs to correct for misaligned SNPs, and excluded a total 5 additional SNPs that

were under Hardy-Weinberg disequilibrium in at least one of the populations ($P_{value} < 10^{-6}$), leading to a final SNP array dataset of 559,501 SNPs. We applied additional filters on the genotyping dataset of the 671 individuals retained in exome. We removed individuals with heterozygosity levels higher or lower than the population mean +/- 4 s.d (2 individuals excluded). Based on this data, we searched for 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 (Manichaikul et al., 2010). One individual with first-degree relationships with other samples was thus removed. Additionally, we removed one individual that did not present any first-degree relatedness but was related in second-degree to many others. The application of these quality-controls filters resulted in a final genotyping dataset of 667 individuals and 599,501 SNPs (Fig. S1).

Merging exome and genotyping datasets. Before merging the genotyping array and exome data from the 667 high-quality individuals in common, we flipped alleles for 8400 SNPs with incompatible allelic states, and removed 11 SNPs with alleles that remained incompatible after allele flipping from the genotyping dataset. The total concordance rate has been evaluated on 28,418 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 (599,490 and 728,154 SNPs, respectively) were then merged, yielding a final dataset of 1,299,226 SNPs for 667 individuals, 566 of which are African farmers or hunter-gatherers (Fig. S1).

Whole-genome sequencing of Baka and Nzebi populations. We generated whole genome sequences data for 20 RHG Baka and 20 AGR Nzebi also included in the exome and genotyping dataset. The quality of the variants was assessed using the 'VariantQualityScoreRecalibration' tool (VQSR) from GATK (DePristo et al., 2011). Only SNPs with a VQSR tranche > 99.9 were considered as confidently called yielding a final dataset of 17,687,206 variants (Fig. 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 Genome-wide SNP and Exome Data. Before imputation, we phased the data with SHAPEIT2 (Delaneau et al., 2013). SNPs and allelic states were then aligned with the 1,000 Genomes Project imputation reference panel (Phase 3, The 1000 Genomes project Consortium), referred to as 'reference panel 1', as well as 40 whole genome sequences of Nzebi AGR and Baka RHG referred to as 'reference panel 2' (Fig. 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 a 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 (Howie et al., 2009) considering 1-Mb windows and both reference panel using the -merge_ref_panels option. Of the 13,137,985 SNPs obtained after imputation, we removed SNPs that: (i) presented an information metric below 0.8 (ii) had a duplicate, (iii) presented a call rate < 95%, and (iv) were monomorphic. The final imputed dataset included 10,306,084 SNPs and 9,129,128 after filtering SNPs with MAF < 1%. To evaluate imputation accuracy, we estimated correlation coefficients R^2 between true genotypes (*i.e.*, obtained by Illumina array

genotyping 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) (Fig. S3 and S4). The average correlation coefficient across all genotyped SNPs with information metric > 0.8 were 0.60 and 0.56 for reference panels 1 and 2 respectively showing that our quality filters ensure to keep accurately imputed SNPs to be analyzed further.

Genome-wide scans for recent positive selection. Candidate genomic regions of recent positive selection were detected in seven populations of RHG: Bezan, Baka, BaBongo Center (BaBongoC), BaKoya, BaBongo South and East (BaBongoS and BaBongoE), BaTwa and two populations of AGR, wAGR and eAGR. We used an outlier approach that considers five neutrality statistics:two intrapopulation statistics (|ΔiHH| (Grossman et al., 2010) and |iHS| (Voight et al., 2006)) and three interpopulation statistics (PBS (Population Branch Score,(Yi et al., 2010)), |ΔiHHD| and |XP-EHH| (Sabeti et al., 2007)). We combined these scores in a Fischer score (Fsc,(Deschamps et al., 2016)) equal to the sum, over the five statistics, of –log10(rank of the statistic/number of SNPs). Interpopulation statistics require a reference population and PBS statistics an outgroup population. We performed two separate scans of positive selection in RHG using: (1) western/eastern AGR as reference and EUR as outgroup (Fig. 2 and Figs. S4 and S5) and (2) pooled populations of wRHG as reference and wAGR as outgroup (Fig. 3; Table S2). The derived allele of each SNP was defined based on the 6-EPO alignment and statistics based on extended haplotype homozygosity were computed in 100kb windows, with home-made scripts (available upon request). Only SNPs with a derived allele frequency (DAF) between 10% and 90% were analyzed further. All statistics except PBS were normalized in 40 separate bins of derived allele frequency. An outlier SNP was defined as a SNP with an Fsc among the highest 1% of the genome. Putatively selected genomic

regions were defined as 100kb windows presenting a proportion of outlier SNPs (top 1% Fsc) among the highest 0.5% of windows in five bins of numbers of SNPs. Windows containing less than 50 SNPs were discarded as well as regions of 100kb around empty windows to avoid biases on the outlier enrichment scores.

Subnetwork enrichments. To detect enrichment of selection scores in biological pathways we applied the subnetwork enrichment analysis as described in Gouy et al 2017 and implemented in the R package *signet*. In order to limit the effects of gene length on the selection score, we used the P_{value} of enrichments in outlier SNPs per window to assign scores to the genes they overlap. Thus, we assigned selection to scores to 27,623 genes with Entrez IDs present in our dataset by considering the $-\log10(P_{value})$ of the enrichment in outliers (top 1% SNPs) of the window in which they are located. When genes overlapped several windows, we considered only the lowest P_{value} for the enrichment in outlier SNPs to compute the gene selection score. To compute subnetwork enrichments for all RHG populations we used the results from the selection scan comparing RHG and AGR populations and using EUR as an outgroup. We considered a total of 301 pathways from the KEGG (Kyoto Encyclopaedia of Genes and Genomes) database (Kanehisa et al., 2017) using the "graphite" R package and analyzed a total of 4,527 genes for subnetwork enrichment. We then performed a correction for multiple testing by calculating qvalues using the *qvalue* R package.

Inference of local ancestry in admixed BaBongo (BaBongo South and BaBongo East). To infer local ancestry in BaBongo South and East individuals, we first constituted parental populations representative of AGR and RHG ancestry. We considered 163 individuals with at least 80% of their ancestry assigned to the agriculturalist component based on the

ADMIXTURE (Alexander et al., 2009) analysis at *K*=5 as the parental wAGR population. Similarly, we considered 101 individuals with less than 5% AGR ancestry as representative of the wRHG population. The genomes of BaBongo South and Babongo East were decomposed into segments of different ancestries with RFMix v.1.5.4 (Maples et al., 2013), including two EM steps. Genomic regions within the 2 Mb from the telomeres of each chromosome were excluded. The genome-wide AGR ancestry estimated by RFMix was 94% in the parental AGR population (ranging from 77% to 99%), 62% in the BaBongo South and East (ranging from 43% to 82%) and 27% in the parental RHG population (ranging from 15% to 41%) (Fig. S8).

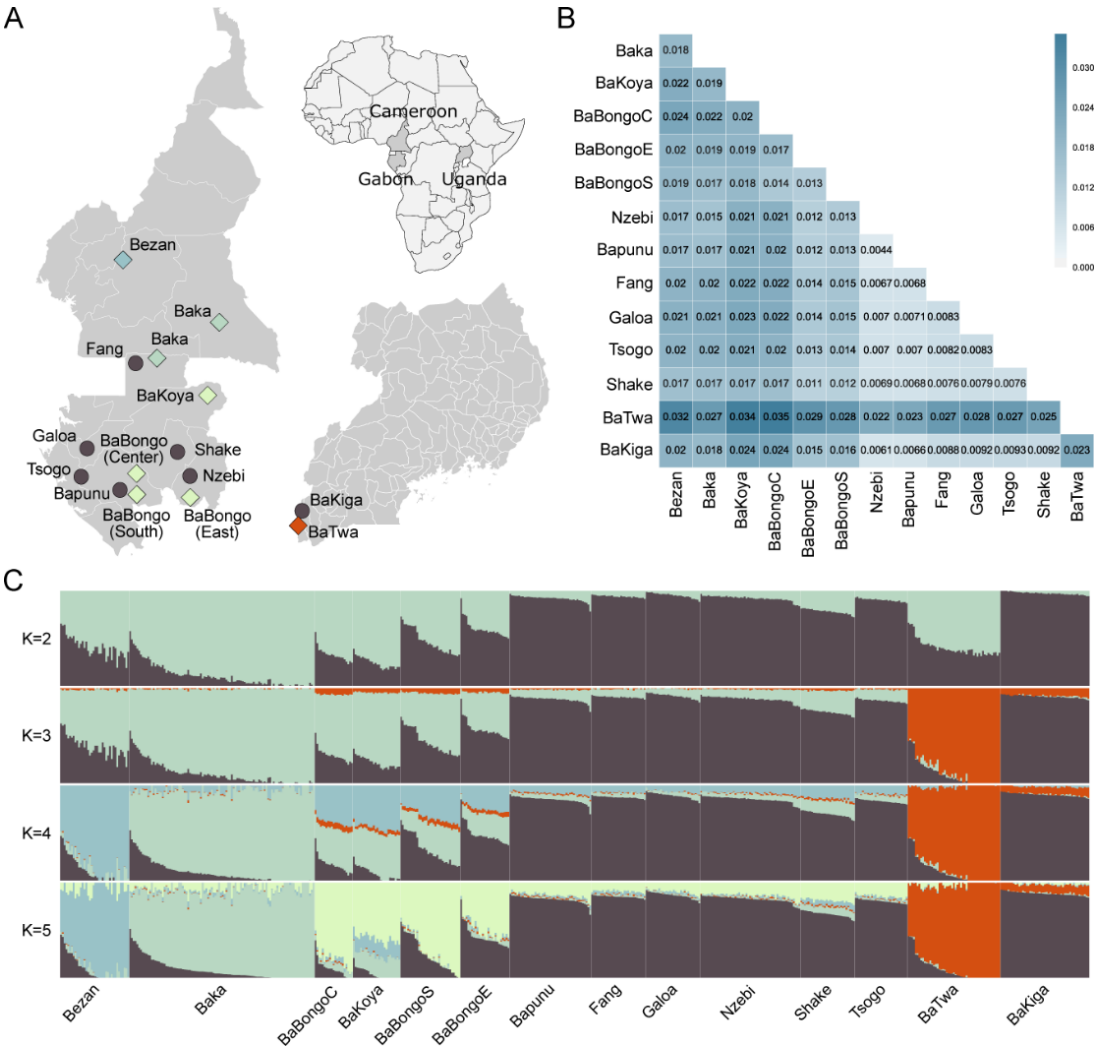


Figure 1. Genetic differentiation and structure of African rainforest hunter-gatherer populations. (A) Geographic location of the sampled populations. Populations of rainforest hunter–gatherers (RHG, diamonds) and neighboring farmers (AGR, circles) have been sampled from three different African countries shown in the continental map. Colors indicate the major genetic ancestry in each population, as determined by ADMIXTURE (Fig. 1C). (B) Genetic differentiation between studied populations, measured by the mean pairwise F_{ST} calculated on the exome data. (C) Estimation of ancestry proportions with the clustering algorithm ADMIXTURE using 412,869 independent SNPs from the exome and SNP array data. Cross-validation values were lowest at K (the number of clusters) = 5 (Fig. S2). BaBongo populations from the Center, South and East Gabon are abbreviated by BaBongoC, BaBongoS and BaBongoE respectively in panels B and C.

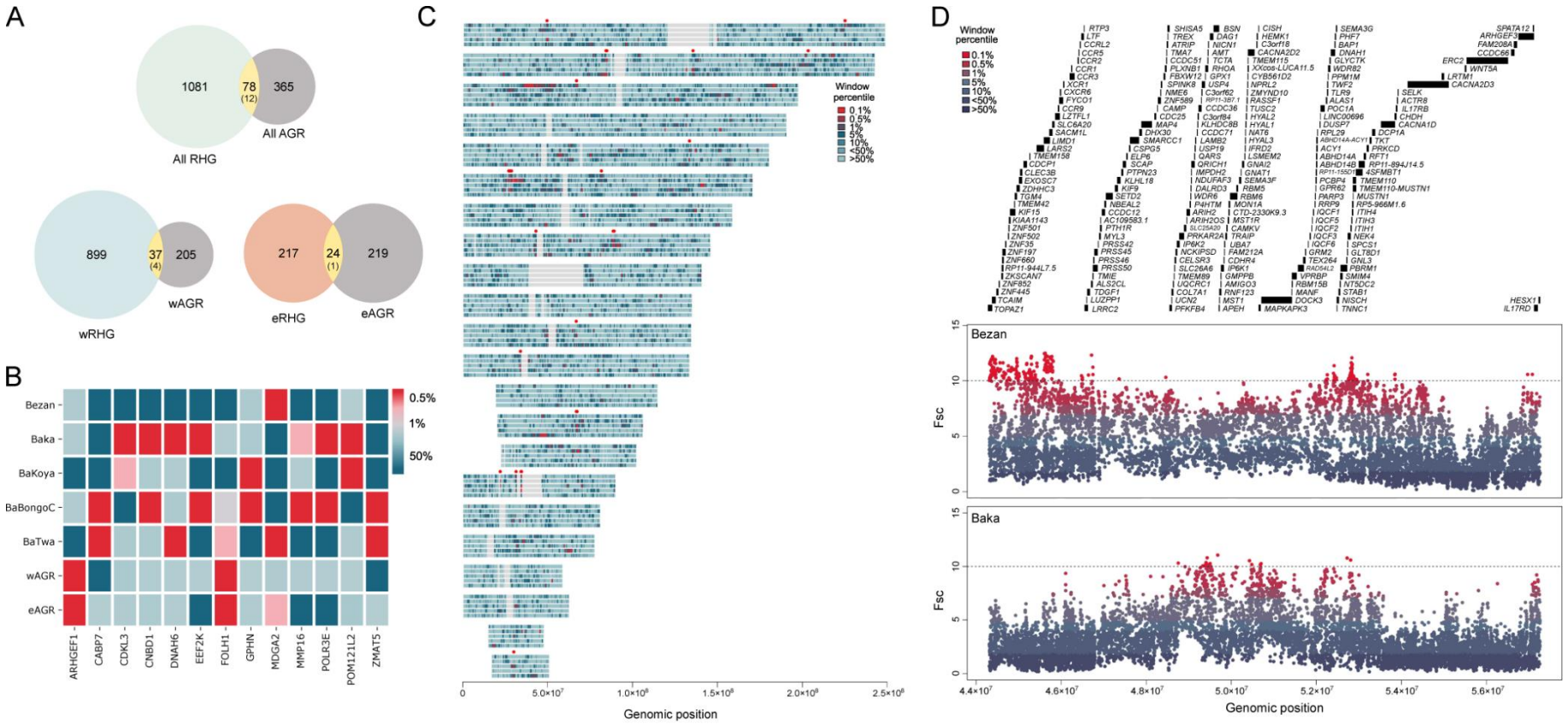


Figure 2. Shared signals of positive selection in African rainforest hunter-gatherer populations. (A) Number of candidate genomic windows shared between RHG and AGR populations. The expected number of overlapping windows was obtained by resampling. (B) Genes located in candidate windows with at least a SNP Fsc score > 10 and specific to AGR or RHG populations. Genes shared by both groups are not shown. (C) Genome-wide clusters of signals of natural selection. The chromosomes of each of the five RHG populations (from top to bottom: Bezan, Baka, BaBongo Center, BaKoya, BaTwa) are shown, and red dots indicate windows or contiguous regions shared by at least two RHG populations (details found in Table S3) (D) Local positive selection signals on chromosome 3 between positions 49,310,197 and 52,260,197 for in the Bezan and Baka populations.

Figure 3. Genome-wide signals of positive selection in wRHG populations. Proportions of outlier SNPs (Fsc in top 1% of the empirical distribution) in 100kb genomic windows in the Bezan, Baka, BaBongo Center and BaKoya. Only candidate genes in windows within the top 0.5% of the enrichment in SNP outliers and with at least one high-scoring SNP (Fsc > 10) are shown. Combined selection signals were computed using wAGR as an outgroup population.

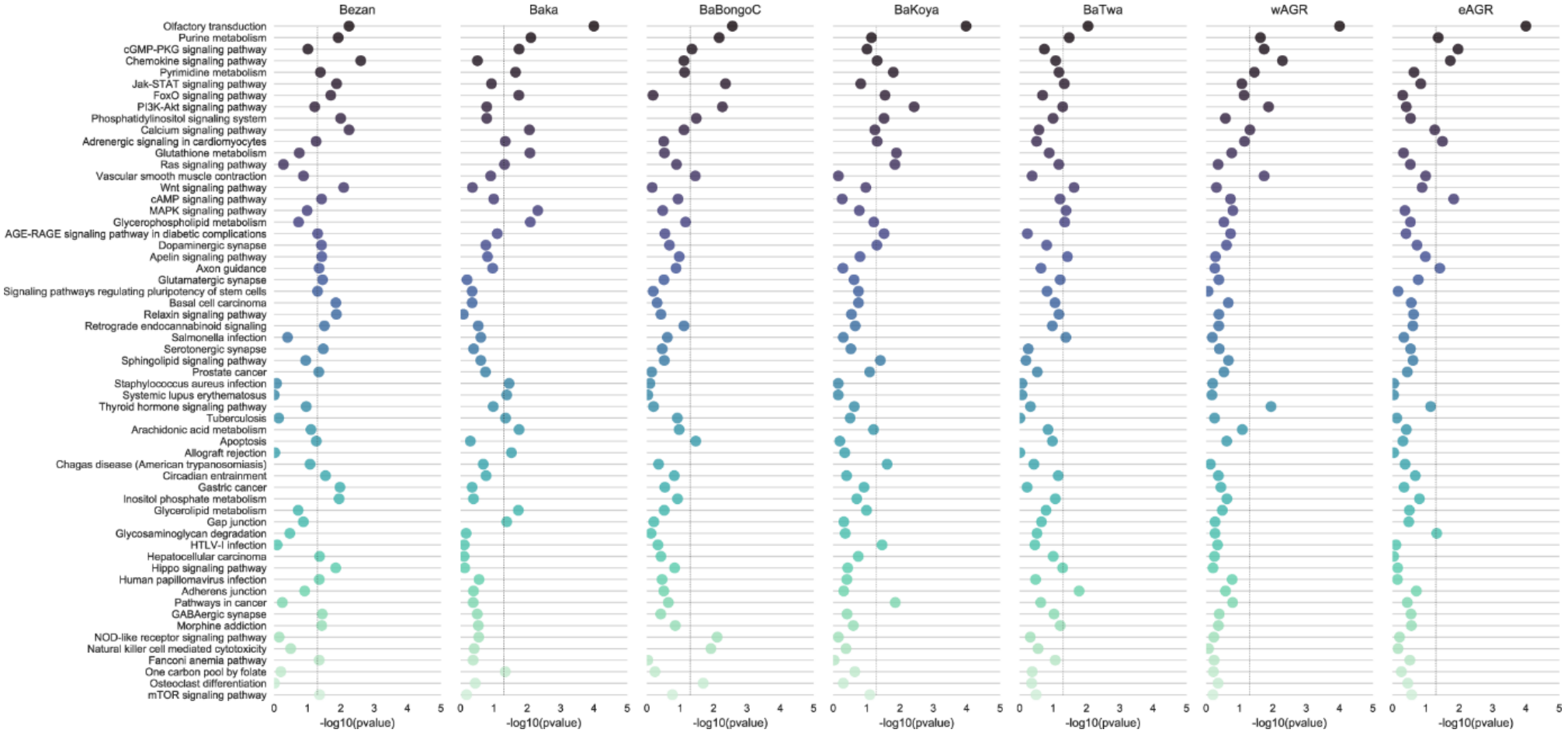


Figure 4. Signals of polygenic selection in African rainforest hunter-gatherers and agriculturalists. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways presenting an enrichment of selection scores in gene subnetworks with a $P_{value} < 0.05$ ($-\log_{10}(0.05)$ indicating significance is represented with a vertical bar). Pathways are shown from the highest to the lowest level of sharing between populations.

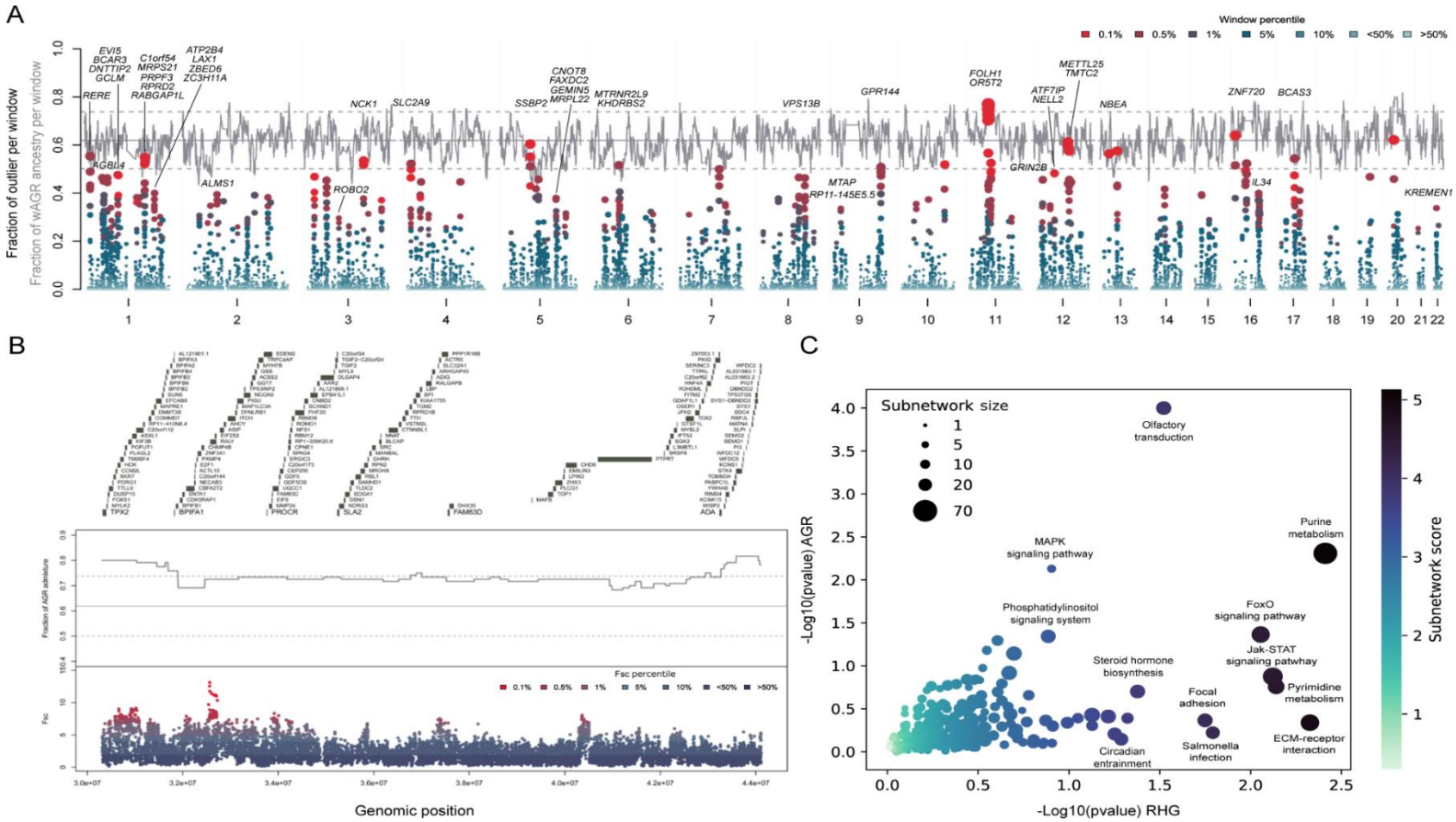


Figure 5. Signals of adaptive admixture in admixed rainforest hunter-gatherers. (A) Proportion of outlier SNPs (Fsc in top 1% of the empirical distribution) in 100kb windows in merged populations of Babongo South and Babongo East. Only candidate genes in windows within the top 0.5% of the enrichment in outliers and with at least one high-scoring SNP (Fsc > 10) are shown. The gray line indicates the proportion of wAGR ancestry along the genome. Solid and dashed line represent the genome-wide mean wAGR ancestry and the mean +/- 2 s.d respectively. (B) Signals of Fsc on chromosome 20 between positions 30,311,795 and 44,111,795 presenting a mean wAGR ancestry higher than the genome average+2s.d.in BaBongo South and East. (C) KEGG pathways presenting gene subnetworks enriched in wRHG or wAGR ancestry (P_{value} < 0.05) in BaBongo South and East. Colors and dot sizes indicate the scores and sizes of subnetworks enriched in wRHG ancestry

References

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res *19*, 1655-1664.

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.

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.

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.

Baumann, G., Shaw, M.A., and Merimee, T.J. (1989). Low levels of high-affinity growth hormone-binding protein in African pygmies. N Engl J Med *320*, 1705-1709.

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.

Becker, N.S.A., Touraille P, Froment, A., Heyer, E., and Courtiol, A. (2012). Short stature in African pygmies is not explained by sexual selection. Evolution and Human Behavior *33*, 615-622.

Becker, T., Loch, G., Beyer, M., Zinke, I., Aschenbrenner, A.C., Carrera, P., Inhester, T., Schultze, J.L., and Hoch, M. (2010). FOXO-dependent regulation of innate immune homeostasis. Nature *463*, 369-373.

Bergey, C.M., Lopez, M., Harrison, G.F., Patin, E., Cohen, J., Quintana-Murci, L., Barreiro, L.B., and Perry, G.H. (2018). Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. bioRxiv.

Bhatia, G., Tandon, A., Patterson, N., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J.*, et al.* (2014). Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. Am J Hum Genet *95*, 437-444.

Bozzola, M., Travaglino, P., Marziliano, N., Meazza, C., Pagani, S., Grasso, M., Tauber, M., Diegoli, M., Pilotto, A., Disabella, E.*, et al.* (2009). The shortness of Pygmies is associated with severe under-expression of the growth hormone receptor. Mol Genet Metab *98*, 310-313.

Capute, A.J., Rimoin, D.L., Konigsmark, B.W., Esterly, N.B., and Richardson, F. (1969). Congenital deafness and multiple lentigines. A report of cases in a mother and daughter. Arch Dermatol *100*, 207-213.

Carter-Timofte, M.E., Hansen, A.F., Christiansen, M., Paludan, S.R., and Mogensen, T.H. (2018). Mutations in RNA Polymerase III genes and defective DNA sensing in adults with varicella-zoster virus CNS infection. Genes Immun.

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

Cummings, D., Dowling, K.F., Silverstein, N.J., Tanner, A.S., Eryilmaz, H., Smoller, J.W., and Roffman, J.L. (2017). A Cross-Sectional Study of Dietary and Genetic Predictors of Blood Folate Levels in Healthy Young Adults. Nutrients *9*.

Delaneau, O., Howie, B., Cox, A.J., Zagury, J.F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. Am J Hum Genet *93*, 687-696.

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.

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.

Diamond, J., and Bellwood, P. (2003). Farmers and their languages: the first expansions. Science *300*, 597-603.

Diamond, J.M. (1991). Anthropology. Why are pygmies small? Nature *354*, 111-112.
Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S., and Reich, D. (2015). No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet *47*, 126-131.

Dodington, D.W., Desai, H.R., and Woo, M. (2018). JAK/STAT - Emerging Players in Metabolism. Trends Endocrinol Metab *29*, 55-65.

Ehlting, C., Ronkina, N., Bohmer, O., Albrecht, U., Bode, K.A., Lang, K.S., Kotlyarov, A., Radzioch, D., Gaestel, M., Haussinger, D.*, et al.* (2011). Distinct functions of the mitogen-activated protein kinase-activated protein (MAPKAP) kinases MK2 and MK3: MK2 mediates lipopolysaccharide-induced signal transducers and activators of transcription 3 (STAT3) activation by preventing negative regulatory effects of MK3. J Biol Chem *286*, 24113-24124.

Eijkelenboom, A., and Burgering, B.M. (2013). FOXOs: signalling integrators for homeostasis maintenance. Nat Rev Mol Cell Biol *14*, 83-97.

Evans, J.F., Islam, S., Urade, Y., Eguchi, N., and Ragolia, L. (2013). The lipocalin-type prostaglandin D2 synthase knockout mouse model of insulin resistance and obesity demonstrates early hypothalamic-pituitary-adrenal axis hyperactivity. J Endocrinol *216*, 169-180.

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.

Finkelman, F.D., Shea-Donohue, T., Morris, S.C., Gildea, L., Strait, R., Madden, K.B., Schopf, L., and Urban, J.F., Jr. (2004). Interleukin-4- and interleukin-13-mediated host protection against intestinal nematode parasites. Immunol Rev *201*, 139-155.

Gouy, A., Daub, J.T., and Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. Nucleic Acids Res *45*, e149.

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.* (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703-713.

Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O.*, et al.* (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883-886.

Hattori, Y., Vera, J.C., Rivas, C.I., Bersch, N., Bailey, R.C., Geffner, M.E., and Golde, D.W. (1996). Decreased insulin-like growth factor I receptor expression and function in immortalized African Pygmy T cells. J Clin Endocrinol Metab *81*, 2257-2263.

Heiker, J.T. (2014). Vaspin (serpinA12) in obesity, insulin resistance, and inflammation. J Pept Sci *20*, 299-306.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920-924.

Hewlett, B.S. (2014). Hunter-gatherers of the Congo Basin: Cultures, Histories and Biology of African Pygmies (Transaction Publishers).

Howard, S.R., and Dunkel, L. (2018). The Genetic Basis of Delayed Puberty. Neuroendocrinology *106*, 283-291.

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.
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.

Jablonski, N.G. (2012). Human skin pigmentation as an example of adaptive evolution. Proc Am Philos Soc *156*, 45-57.

Jackson, D. (2005). Implementation of international commitmentson traditional forest-related knowledge; indigenous peoples' experiences in Central Africa. (H Newing, ed. Our knowledge for our survivial, Vol 1. Regional case studies in traditional forest- related knowledge and implementation of related international commitments. International Alliance of the Indigenous and Tribal Peoples of the Tropical Forests, Forest Peoples Programme, and Centre for International Forestry Research.), pp. 150–303.

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.

Jones, P., Lucock, M., Veysey, M., Jablonski, N., Chaplin, G., and Beckett, E. (2018). Frequency of folate-related polymorphisms varies by skin pigmentation. Am J Hum Biol *30*.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res *45*, D353-D361.

Kim, D.H., Perdomo, G., Zhang, T., Slusher, S., Lee, S., Phillips, B.E., Fan, Y., Giannoukakis, N., Gramignoli, R., Strom, S.*, et al.* (2011). FoxO6 integrates insulin signaling with gluconeogenesis in the liver. Diabetes *60*, 2763-2774.

Kostel Bal, S., Haskologlu, S., Serwas, N.K., Islamoglu, C., Aytekin, C., Kendirli, T., Kuloglu, Z., Yavuz, G., Dalgic, B., Siklar, Z.*, et al.* (2017). Multiple Presentations of LRBA Deficiency: a Single-Center Experience. J Clin Immunol *37*, 790-800.

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.

Lee, C.H., Olson, P., Hevener, A., Mehl, I., Chong, L.W., Olefsky, J.M., Gonzalez, F.J., Ham, J., Kang, H., Peters, J.M.*, et al.* (2006). PPARdelta regulates glucose metabolism and insulin sensitivity. Proc Natl Acad Sci U S A *103*, 3444-3449.

Lee, S., and Dong, H.H. (2017). FoxO integration of insulin signaling with glucose and lipid metabolism. J Endocrinol *233*, R67-R79.

Leprivier, G., Remke, M., Rotblat, B., Dubuc, A., Mateo, A.R., Kool, M., Agnihotri, S., El-Naggar, A., Yu, B., Somasekharan, S.P.*, et al.* (2013). The eEF2 kinase confers resistance to nutrient deprivation by blocking translation elongation. Cell *153*, 1064-1079.

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

Liu, B., Mink, S., Wong, K.A., Stein, N., Getman, C., Dempsey, P.W., Wu, H., and Shuai, K. (2004). PIAS1 selectively inhibits interferon-inducible genes and is important in innate immunity. Nat Immunol *5*, 891-898.

Liu, Y., Ge, X., Dou, X., Guo, L., Zhou, S.R., Wei, X.B., Qian, S.W., Huang, H.Y., Xu, C.J., Jia, W.P.*, et al.* (2015). Protein Inhibitor of Activated STAT 1 (PIAS1) Protects Against Obesity-Induced Insulin Resistance by Inhibiting Inflammation Cascade in Adipose Tissue. Diabetes *64*, 4061-4074.

Lopez Herraez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One *4*, e7888.

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.

Maggio, M., De Vita, F., Lauretani, F., Butto, V., Bondi, G., Cattabiani, C., Nouvenne, A., Meschi, T., Dall'Aglio, E., and Ceda, G.P. (2013). IGF-1, the cross road of the nutritional, inflammatory and hormonal pathways to frailty. Nutrients *5*, 4184-4205.

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.
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.

McManus, K.F., Taravella, A.M., Henn, B.M., Bustamante, C.D., Sikora, M., and Cornejo, O.E. (2017). Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS Genet *13*, e1006560.

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.

Merimee, T.J., Hewlett, B.S., Wood, W., Bowcock, A.M., and Cavalli-Sforza, L.L. (1989). The growth hormone receptor gene in the African pygmy. Trans Assoc Am Physicians *102*, 163-169.

Merimee, T.J., Zapf, J., Hewlett, B., and Cavalli-Sforza, L.L. (1987). Insulin-like growth factors in pygmies. The role of puberty in determining final stature. N Engl J Med *316*, 906-911.

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.

Migliano, A.B., Vinicius, L., and Lahr, M.M. (2007). Life history trade-offs explain the evolution of human pygmies. Proc Natl Acad Sci U S A *104*, 20216-20219.

Moens, U., Kostenko, S., and Sveinbjornsson, B. (2013). The Role of Mitogen-Activated Protein Kinase-Activated Protein Kinases (MAPKAPKs) in Inflammation. Genes (Basel) *4*, 101-133.

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.

Ohenjo, N., R. Willis, D. Jackson, C. Nettleton, K. Good, and B. Mugarura (2006). Health of indigenous people in Africa. The Lancet *367:1937–46*.

Panebianco, F., Kelly, L.M., Liu, P., Zhong, S., Dacic, S., Wang, X., Singhi, A.D., Dhir, R., Chiosea, S.I., Kuan, S.F.*, et al.* (2017). THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. Proc Natl Acad Sci U S A *114*, 2307-2312.

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.

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.

Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Regnault, B., Lemee, 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.

Perry, G., and Verdu, P. (2016). Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. Quaternary International.
Perry, G.H., and Dominy, N.J. (2009). Evolution of the human pygmy phenotype. Trends Ecol Evol *24*, 218-225.

Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nedelec, 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 U S A *111*, E3596-3603.

Philipson, D. (2005). African Archaeology (Cambridge, Cambridge University Press).

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*., et al.* (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res *19*, 826-837.

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*, R208-215.

Pybus, M., Luisi, P., Dall'Olio, G.M., Uzkudun, M., Laayouni, H., Bertranpetit, J., and Engelken, J. (2015). Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics *31*, 3946-3952.

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 e617.

Rajeevan, M.S., Dimulescu, I., Murray, J., Falkenberg, V.R., and Unger, E.R. (2015). Pathway-focused genetic evaluation of immune and inflammation related genes with chronic fatigue syndrome. Hum Immunol *76*, 553-560.

Ranke, M.B., and Wit, J.M. (2018). Growth hormone - past, present and future. Nat Rev Endocrinol *14*, 285-300.

Rimoin, D.L., Merimee, T.J., Rabinowitz, D., Cavalli-Sforza, L.L., and McKusick, V.A. (1969). Peripheral subresponsiveness to human growth hormone in the African pygmies. N Engl J Med *281*, 1383-1388.

Rosenzweig, S.D., Dorman, S.E., Uzel, G., Shaw, S., Scurlock, A., Brown, M.R., Buckley, R.H., and Holland, S.M. (2004). A novel mutation in IFN-gamma receptor 2 with dominant negative activity: biological consequences of homozygous and heterozygous states. J Immunol *173*, 4000-4008.

Rozzi, F.V., Koudou, Y., Froment, A., Le Bouc, Y., and Botton, J. (2015). Growth pattern from birth to adulthood in African pygmies of known age. Nat Commun *6*, 7672.

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.* (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913-918.

Sederquist, B., Fernandez-Vojvodich, P., Zaman, F., and Savendahl, L. (2014). Recent research on the growth plate: Impact of inflammatory cytokines on longitudinal bone growth. J Mol Endocrinol *53*, T35-44.

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.
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.

Soh, J., Donnelly, R.J., Kotenko, S., Mariano, T.M., Cook, J.R., Wang, N., Emanuel, S., Schwartz, B., Miki, T., and Pestka, S. (1994). Identification and sequence of an accessory factor required for activation of the human interferon gamma receptor. Cell *76*, 793-802.

Taketomi, Y., Ueno, N., Kojima, T., Sato, H., Murase, R., Yamamoto, K., Tanaka, S., Sakanaka, M., Nakamura, M., Nishito, Y.*, et al.* (2013). Mast cell maturation is driven via a group III phospholipase A2-prostaglandin D2-DP1 receptor paracrine axis. Nat Immunol *14*, 554-563.

Tanaka, T., Tahara-Hanaoka, S., Nabekura, T., Ikeda, K., Jiang, S., Tsutsumi, S., Inagaki, T., Magoori, K., Higurashi, T., Takahashi, H.*, et al.* (2014). PPARbeta/delta activation of CD300a controls intestinal immunity. Sci Rep *4*, 5412.

Tapscott, S.J. (2005). The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development *132*, 2685-2695.

Torkamandi, S., Bastami, M., Ghaedi, H., Moghadam, F., Mirfakhraie, R., and Omrani, M.D. (2016). MAP3K1 May be a Promising Susceptibility Gene for Type 2 Diabetes Mellitus in an Iranian Population. Int J Mol Cell Med *5*, 134-140.

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*, 11 10 11-33.

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.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Thery, 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.

Verdu, P., Becker, N.S., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., Hombert, J.M., Van der Veen, L., Le Bomin, S., Bahuchet, S.*, et al.* (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. Mol Biol Evol *30*, 918-937.

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

Williams, D.L., Bonilla, M., Gladyshev, V.N., and Salinas, G. (2013). Thioredoxin glutathione reductase-dependent redox networks in platyhelminth parasites. Antioxid Redox Signal *19*, 735-745.
Wolters, T.L.C., Netea, M.G., Hermus, A., Smit, J.W.A., and Netea-Maier, R.T. (2017). IGF1 potentiates the pro-inflammatory response in human peripheral blood mononuclear cells via MAPK. J Mol Endocrinol *59*, 129-139.

Wong, S.C., Smyth, A., McNeill, E., Galloway, P.J., Hassan, K., McGrogan, P., and Ahmed, S.F. (2010). The growth hormone insulin-like growth factor 1 axis in children and adolescents with inflammatory bowel disease and growth retardation. Clin Endocrinol (Oxf) *73*, 220-228.
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.

Zhang, Y., Gan, Z., Huang, P., Zhou, L., Mao, T., Shao, M., Jiang, X., Chen, Y., Ying, H., Cao, M.*, et al.* (2012). A role for protein inhibitor of activated STAT1 (PIAS1) in lipogenic regulation through SUMOylation-independent suppression of liver X receptors. J Biol Chem *287*, 37973-37985.

Zou, S., Teixeira, A.M., Kostadima, M., Astle, W.J., Radhakrishnan, A., Simon, L.M., Truman, L., Fang, J.S., Hwa, J., Zhang, P.X.*, et al.* (2017). SNP in human ARHGEF3 promoter is associated with DNase hypersensitivity, transcript level and platelet function, and Arhgef3 KO mice have increased mean platelet volume. PLoS One *12*, e0178095.

Supplementary Information

Exploring the history of genetic adaptation of African rainforest hunter-gatherers using whole-exome sequencing data.

Marie Lopez, Alexandre Gouy, Hélène Quach, Christine Harmant, Patrick Mouguiama-Daouda, Jean-Marie Hombert, Alain Froment, George H. Perry, Luis B. Barreiro, Paul Verdu, Laurent Excoffier, Etienne Patin, Lluís Quintana-Murci

This file contains:

Supplementary Tables 1-6
Supplementary Figures 1-9

Supplementary Table 1. Population description and sample sizes of the final dataset of 667 individuals.

Group	Population	Country	N	Source genotyping	Source of exome sequencing
wAGR	Tsogo	Gabon	29	Patin et al., 2017 (EGAS00001002078)	This study
wAGR	Galoa	Gabon	30	Patin et al., 2017 (EGAS00001002078)	This study
wAGR	Shake	Gabon	30	Patin et al., 2017 (EGAS00001002078)	This study
wAGR	Fang	Gabon	31	Patin et al., 2017 (EGAS00001002078)	This study
wAGR	Bapunu	Gabon	44	Patin et al., 2017 (EGAS00001002078)	Lopez et al., (EGAS00001002457)
wAGR	Nzebi	Gabon	55	Patin et al., 2017 (EGAS00001002078)	Lopez et al., (EGAS00001002457)
eAGR	BaKiga	Uganda	49	Perry et al., 2014 (EGAS00001000908)	Lopez et al., (EGAS00001002457)
wRHG	Bezan	Cameroon	38	Patin et al., 2014 (EGAS00001000605)	This study
wRHG	BaBongo Center	Gabon	21	This study	This study
wRHG	BaBongo East	Gabon	27	Patin et al., 2014 (EGAS00001000605)	This study
wRHG	BaBongo South	Gabon	33	Patin et al., 2014 (EGAS00001000605)	This study
wRHG	BaKoya	Gabon	26	Patin et al., 2014 (EGAS00001000605)	This study
wRHG	Baka	Cameroon/ Gabon	72/ 30	Fagny et al., 2015 (EGAS00001001066) Patin et al., 2014 (EGAS00001000605)	Lopez et al., (EGAS00001002457) This study
eRHG	BaTwa	Uganda	51	Perry et al., 2014 (EGAS00001000908)	Lopez et al., (EGAS00001002457)
EUR	Belgian	Belgium	101	Quach et al., 2016 (EGAS00001001895)	Quach et al., 2016 (EGAS00001001895)

Supplementary Table 2. Test, reference and outgroup populations used in genome-wide scans of positive selection

Scans of positive selection	Test population	Reference population	Outgroup population
Scan comparing RHG and AGR	Bezan	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	EUR
	Baka	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	EUR
	BaKoya	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	EUR
	BaBongo Center	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	EUR
	Babongo East and South	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	EUR
	BaTwa	eAGR (BaKiga)	EUR
	wAGR (Fang, Galoa, Tsogo, Shake, Bapunu, Nzebi)	wRHG (Bezan, Baka, BaKoya, BaBongo Center)	EUR
	eAGR (BaKiga)	BaTwa	EUR
Scan comparing western RHG	Bezan	wRHG except Bezan and Babongo East and South	wAGR
	Baka	wRHG except Baka and Babongo East and South	wAGR
	BaKoya	wRHG except BaKoya and Babongo East and South	wAGR
	BaBongo Center	wRHG except BaBongo Center and Babongo East and South	wAGR
	Babongo East and South	wRHG except Babongo East and South	wAGR

Supplementary Table 3. Windows or contiguous regions in the top 0.5% of the enrichment in outlier SNPs and shared by at least two RHG populations.

Genomic region	Size (kb)	Populations	Genes
chr1:225164464-225314464	150	Bezan, BaTwa	DNAH14
chr1:49714464-49814464	100	Bezan, BaKoya	AGBL4
chr2:84410797-84860797	450	Baka, BaTwa	SUCLG1, DNAH6
chr2:84960797-85060797	100	Baka, BaTwa	DNAH6, TRABD2A
chr2:135710797-135860797	150	BaTwa, BaKoya	MAP3K19, CCNT2, RAB3GAP1
chr2:203510797-203610797	100	BaTwa, BaBongoC	FAM117B
chr3:67060197-67160197	100	Bezan, BaBongoC	KBTBD8
chr5:133611856-133711856	100	Bezan, BaBongoC	CDKL3, UBE2B
chr6:27202452-27402452	200	Baka, BaKoya	ZNF391, POM121L2, PRSS16
chr6:27352452-27502452	150	Bezan, Baka, BaKoya	ZNF391, ZNF184
chr6:28402452-28502452	100	Baka, BaKoya	ZSCAN23, GPX6, GPX5
chr6:28752452-28852452	100	Baka, BaKoya	No genes in the region
chr6:81702452-81802452	100	Bezan, Baka	No genes in the region
chr8:43152422-43302422	150	BaTwa, BaBongoC	No genes in the region
chr8:88652422-88902422	250	Baka, BaBongoC	DCAF4L2
chr8:89202422-89302422	100	Baka, BaBongoC	MMP16
chr11:66877091-66977091	100	BaTwa, BaBongoC	KDM2A, AP001885.1
chr12:34038009-34138009	100	BaBongoC, BaKoya	No genes in the region
chr14:66973030-67223030	250	BaBongoC, BaKoya	GPHN
chr14:67423030-67573030	150	BaBongoC, BaKoya	GPHN
chr16:22168464-22368464	200	Baka, BaBongoC	EEF2K, SDR42E2, POLR3E, CDR2
chr16:31468464-31618464	150	Baka, BaKoya	TGFB1I1, SLC5A2, ARMC5, C16orf58, AHSP
chr16:34418464-34618464	200	Bezan, BaKoya	No genes in the region
chr16:34618464-34718464	100	Bezan, BaBongoC, BaKoya	No genes in the region
chr16:34718464-34818464	100	Bezan, BaBongoC	No genes in the region
chr22:30106379-30206379	100	BaTwa, BaBongoC	CABP7, ZMAT5, UQCR10, ASCC2

Supplementary Table 4. Association with GWAS traits in chr3:49,310,197-52,260,197.

Chr	Position	rs#	Bezan Fsc (DAF)	Baka Fsc (DAF)	BaBongo C Fsc (DAF)	BaKoya Fsc (DAF)	BaTwa Fsc (DAF)	wAGR Fsc (DAF)	eAGR Fsc (DAF)	Trait	Reference
3	49391082	rs13090388	7.16 (0.5)	6.76 (0.319)	1.34 (0.332)	3.11 (0.173)	5.35 (0.588)	5.61 (0.336)	2.35 (0.265)	Educational attainment;	Okbay et al., Nature 2016;
3	49510931	rs7647973	X	X	X	X	X	X	2.19 (0.11)	Menarche;	Perry et al., Nature 2014;
3	49572140	rs4625	6.60 (0.5)	6.89 (0.27)	1.33 (0.332)	2.79 (0.17)	6.08 (0.588)	5.90 (0.33)	2.48 (0.276)	Pediatric autoimmune diseases;	Li et al., Nat Med 2015;
3	49697459	rs9836291	5.67 (0.382)	X	1.22 (0.214)	2.27 (0.154)	6.62 (0.588)	2.25 (0.34)	2.36 (0.29)	Inflammatory bowel disease;	Liu et al., Nat Genet 2015;
3	49701983	rs9858542	5.067 (0.342)	X	X	1.28 (0.115)	3.14 (0.284)	2.54 (0.265)	2.91 (0.204)	Crohn's disease; Crohn's disease; Ulcerative colitis; Blood protein levels;	WTCCC et al., Nature 2007; Parkes et al., Nat Genet 2007; Barrett et al., Nat Genet 2009; Suhre et al., Nat Commun 2017;
3	49719729	rs9822268	5.65 (0.342)	X	1.41 (0.19)	1.80 (0.135)	5.92 (0.598)	2.71 (0.308)	1.70 (0.265)	Ulcerative colitis;	Anderson et al., Nat Genet 2011;
3	49721532	rs3197999	5.09 (0.342)	X	X	1.25 (0.115)	3.09 (0.284)	2.78 (0.265)	3.20 (0.204)	Crohn's disease; Primary sclerosing cholangitis; Crohn's disease; Inflammatory bowel disease; Ulcerative colitis; Crohn's disease; Ulcerative colitis; Primary sclerosing cholangitis; Crohn's disease; Ulcerative colitis; Inflammatory bowel disease;	Barrett et al., Nat Genet 2008; Melum et al., Nat Genet 2010; Franke et al., Nat Genet 2010; Jostins et al., Nature 2012; Liu et al., Nat Genet 2015; Liu et al., Nat Genet 2015; McGovern et al., Nat Genet 2010; Ji et al., Nat Genet 2016; de et al., Nat Genet 2017; de et al., Nat Genet 2017; de et al., Nat Genet 2017;
3	49731861	rs9858213	5.49 (0.342)	X	X	0.94 (0.115)	3.16 (0.284)	3.10 (0.265)	3.35 (0.204)	Educational attainment;	Rietveld et al., Proc Natl Acad Sci U S A 2014;
3	49770032	rs3749237	4.10 (0.105)	X	X	X	X	5.71 (0.153)	X	Resting heart rate;	Eppinga et al., Nat Genet 2016;
3	49898000	rs2777888	3.36 (0.11)	3.52 (0.127)	1.44 (0.19)	1.83 (0.28)	5.41(0.48)	3.61 (0.365)	3.16 (0.28)	Age at first birth; Age at first birth;	Barban et al., Nat Genet 2016; Barban et al., Nat Genet 2016;
3	49914397	rs11712056	6.40 (0.526)	1.20 (0.319)	0.85 (0.31)	1.57 (0.327)	7.55 (0.529)	1.43 (0.42)	3.71 (0.418)	Educational attainment;	Okbay et al., Nature 2016;
3	50093209	rs6762477	6.46 (0.289)	3.15 (0.147)	2.07 (0.119)	X	X	2.65 (0.249)	X	Menarche; Menarche;	Elks et al., Nat Genet 2010; Perry et al., Nature 2014;
3	50129399	rs2013208	4.03 (0.513)	4.87 (0.76)	2.22 (0.762)	2.77 (0.75)	6.13 (0.675)	4.07 (0.61)	3.76 (0.59)	HDL cholesterol; HDL cholesterol levels;	Willer et al., Nat Genet 2013; Spracklen et al., Hum Mol Genet 2017;
3	50998816	rs9869826	X	X	X	X	X	3.15 (0.135)	X	Blood protein levels;	Suhre et al., Nat Commun 2017;
3	52115752	rs77861329	4.61 (0.184)	1.58 (0.309)	3.38 (0.357)	3.77 (0.25)	X	1.70 (0.201)	X	Macrophage inflammatory protein 1b levels;	Ahola-Olli et al., Am J Hum Genet 2016;

Supplementary Table 5. Candidate genes in western RHG involved in immunity, thyroid and insulin-resistance functions.

Gene	Population	Functions	Window	Fraction of outlier	P_{value}	Position of the best Fsc	Fsc	P_{value}
PIAS1	Baka	Immunity and insuline sensitivity	chr15:68250075-68350075	0.32	0.002942	68348514	12.1	0.00022
PPARD	Baka	Immunity and insuline sensitivity	chr6:35302452-35402452	0.31	0.002548	35354976	13.0	0.00010
PTGDS	BaBongo Center	Immunity, insuline sensitivity and thyroid function	chr9:139814588-139914588	0.46	0.00245	139876483	10.2	0.00068
SERPINA12	BaBongo Center	Immunity and insuline sensitivity	chr14:94973030-95073030	0.32	0.001961	94980575	10.2	0.00064
THRB	Baka	Thyroid function	chr3:24460197-24560197	0.34	0.002739	24536309	10.2	0.00077
IGSF10	BaBongo Center	Thyroid function	chr3:151160197-151260197	0.32	0.002871	151167762	13.9	3.92E-05
IGF2BP3	BaKoya	Insulin-like growth-factor	chr7:23335326-23435326	0.35	0.004563	23363118	10.2	0.00064
IFNGR2	Baka	Immunity	chr21:34794305-34894305	0.36	0.004145	34808122	11.9	0.00024
SERPINA5	BaBongo Center	Immunity	chr14:94973030-95073030	0.32	0.001961	95045774	10.2	0.00065
MAP3K1	BaBongo Center	Immunity	chr5:56061856-56161856	0.31	0.00352	56160689	11.2	0.00027
LRBA	BaKoya	Immunity	chr4:151203233-151303233	0.46	0.001521	151293258	11.9	0.00011
TXNRD3	BaKoya	Immunity	chr3:126210197-126310197	0.41	0.000607	126297864	11.3	0.00022
MYO1D	BaBongo Center	Muscle	chr17:30950828-31050828	0.42	0.00347	30974937	13.3	5.96E-05

Supplementary Table 6. Subnetworks enriched in scores of selection ($P_{value} < 0.05$).

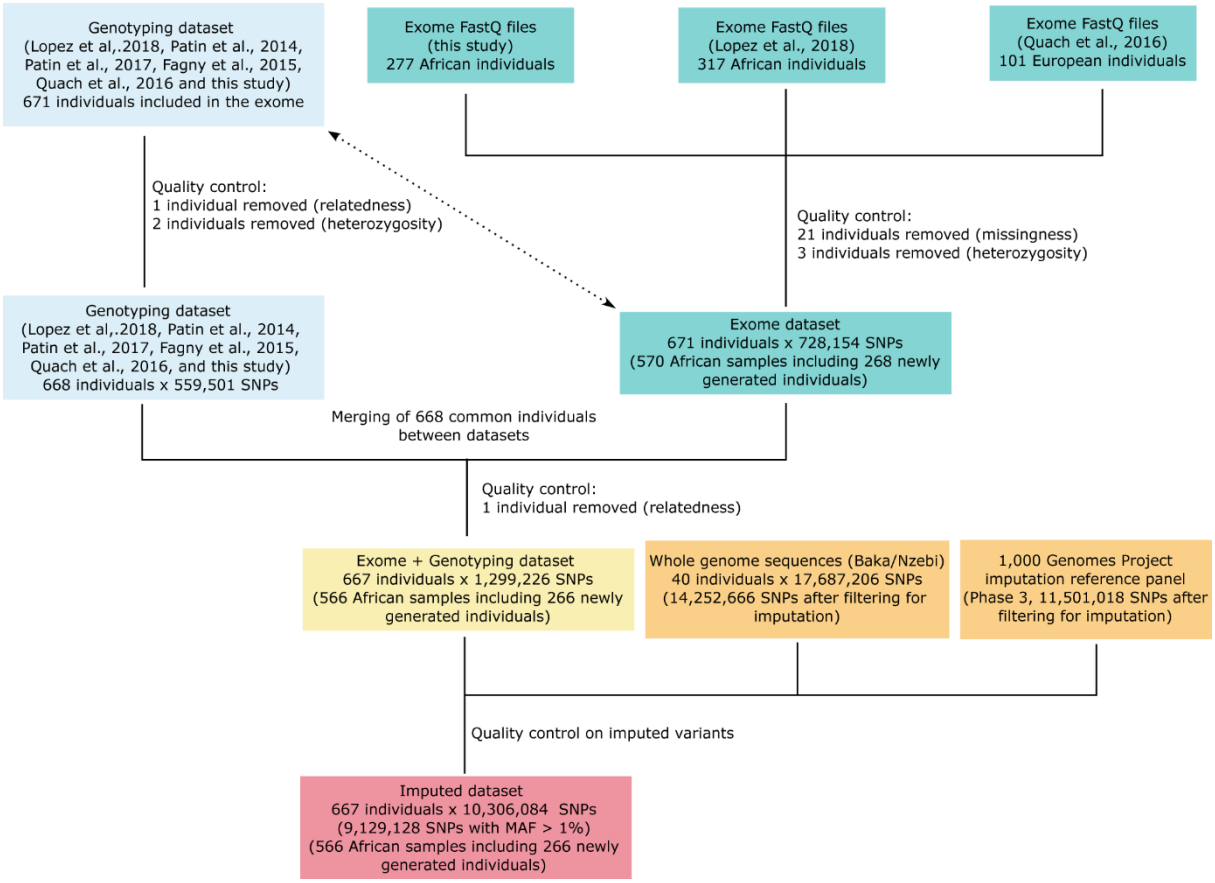
Population	Pathway	Network size	Subnetwork score	P_{value}	qvalue	Genes in the subnetwork
eRHG	Adherens junction	69	7.5831782	0.0165671	0.7146465	CTNNA1, CTNNA2, RHOA, RAC3, ACTB, ACTG1, VCL, ACTN1, IQGAP1
eAGR	Adrenergic signaling in cardiomyocytes	144	7.0556827	0.0317027	0.991392	PPP1CA, PPP2R3C, PPP2R2A, PPP2R3A, PPP2R5B, PPP2R5C, PPP2R5D, CACNB1, CACNA2D2
Baka	Adrenergic signaling in cardiomyocytes	144	7.0607646	0.045445	0.6965135	CACNG2, PPP1CA, PPP2CA, PPP2CB, PPP2R2B, PPP2R5B, PPP2R5D, PPP2R5E, PRKACA, SCN4B, CAMK2A, CACNA2D2
BaKoya	Adrenergic signaling in cardiomyocytes	144	6.9826176	0.0482824	0.7909093	PPP2CA, PPP2R1A, PPP2R2B, PPP2R2C, CACNA2D3, CACNA2D2
BaKoya	AGE-RAGE signaling pathway in diabetic complications	89	7.3883394	0.0300587	0.7435157	MAPK14, DIAPH1, AGER, NRAS, NOX4, PLCD1, PRKCA, PRKCB, PRKCE, PRKCZ, MAPK10, MAPK13, CDC42
Bezan	AGE-RAGE signaling pathway in diabetic complications	89	6.7771658	0.0492766	0.3748992	PLCD3, PLCB1, NOX4, PLCB4, PLCD1, PRKCD, PRKCZ, MAPK8, MAPK10
Baka	Allograft rejection	28	7.5565854	0.0296335	0.6829236	NA, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-E
eRHG	Apelin signaling pathway	133	6.8688213	0.0369089	0.7146465	KLF2, ADCY1, ADCY2, ADCY3, ADCY5, ADCY7, GNAI1, GNAI2, PRKAG2, GNG13, PIK3CG, PLCB3, PRKAG3, GNG2, SPHK1
Bezan	Apelin signaling pathway	133	7.0227134	0.0370099	0.3748992	GNB5, ADCY6, ADCY8, ADCY9, MRAS, PLCB1, GNAI2, GNB2, GNG5, GNG13, PLCB4, PRKAA2
BaBongoC	Apoptosis	131	7.2631172	0.03418	0.9927668	ERN1, GZMB, BIRC5, LMNA, NFKB1, TRAF2, CASP8, CASP10, CFLAR
Baka	Arachidonic acid metabolism	62	7.9844859	0.0174869	0.5096589	CYP2J2, GPX6, PLA2G2D, GPX1, GPX3, GPX5, GPX7, PLA2G2C, PLA2G5, PLA2G2F
eAGR	Axon guidance	166	6.8909083	0.0382112	0.991392	FYN, RHOA, PLXNB1, SSH3, PPP3CA, PPP3CB, RAC1, SMO, TRPC1
Bezan	Axon guidance	166	6.8581752	0.0444538	0.3748992	FYN, NA, SEMA3G, RAC1, SEMA3F, SEMA3B
Bezan	Basal cell carcinoma	59	7.9797454	0.0140491	0.2897598	DVL2, DVL3, FZD2, GLI3, WNT9A, WNT3A
Bezan	Calcium signaling pathway	179	9.0046359	0.0056616	0.2897598	PTK2B, ITPKB, P2RX5, PLCD1, TNNC1, VDAC3, CACNA1D, ORAI3
Baka	Calcium signaling pathway	179	8.6178224	0.0085864	0.3957579	PPIF, GRIN2A, ITPR3, P2RX3, SLC8A2, SLC8A3, TNNC2, VDAC1, PLCD4, CALML4, ORAI3
wAGR	Calcium signaling pathway	179	6.8151215	0.0483905	0.9872492	ITPR3, ATP2A2, P2RX5, PLCD1, RYR1, VDAC1, VDAC2, ITPKC, CALML4, ORAI3
eAGR	cAMP signaling pathway	196	7.7444383	0.0145916	0.991392	ADCY3, EP300, GNAI2, RHOA, ATP2B2, PDE3B, PPP1CA, PRKACA, ADCY10, PPP1R1B
Bezan	cAMP signaling pathway	196	6.9996133	0.0375341	0.3748992	ADCY9, CNGA1, GNAI2, GNAS, HTR1A, PRKACB
eAGR	cGMP-PKG signaling pathway	153	8.1073708	0.0108125	0.991392	KCNMB2, NPR1, ATP2B2, PDE2A, PDE3B, PPP1CA, PRKCE, PRKG1, SLC8A2
Baka	cGMP-PKG signaling pathway	153	7.9827304	0.0174869	0.5096589	ATF6B, KCNU1, PLCB1, ATP1B3, ATP2B4, PLN, PRKG1, RAF1, CREB3L2, SLC8A2, SLC8A3, CREB5
wAGR	cGMP-PKG signaling pathway	153	7.7263684	0.0181856	0.8516931	KCNMB2, CREB3L4, PPP1R12A, ATP1A3, ATP2B2, ATP2B4, PRKG1, SLC8A3, CREB5
BaBongoC	cGMP-PKG signaling pathway	153	7.0060947	0.0444235	0.9938696	KCNMB2, ATF2, CREB3L4, KCNMA1, ATP1A2, PPP1CA, PRKG1, RGS2, SLC8A3, GTF2IRD1
BaKoya	Chagas disease (American trypanosomiasis)	87	7.5847902	0.0242983	0.7435157	MAPK14, IKBKB, PPP2CA, PPP2R1A, PPP2R1B, PPP2R2B, PPP2R2C, MAPK13, TRAF6
Bezan	Chemokine signaling pathway	185	10.98794	0.0025163	0.2897598	CXCR6, CCR9, CCR3, CCR8, CX3CR1, GNAI2, XCR1, CCL19, CCR2
wAGR	Chemokine signaling pathway	185	9.1225168	0.0051212	0.7195339	CXCR6, CCR9, CCR1, CCR3, CCR5, XCR1, CCL28, CCR2
eAGR	Chemokine signaling pathway	185	7.5657744	0.0184758	0.991392	CXCR6, CCR9, CCR5, ADRBK1, ADRBK2, GNAI2, XCR1, CCR2, CXCR4

BaKoya	Chemokine signaling pathway	185	6.9818796	0.0483871	0.7909093	GNB5, AKT2, GNGT1, ITK, NRAS, PIK3CB, PTK2, PREX1, SOS1
Bezan	Circadian entrainment	96	7.3237558	0.0287272	0.3748992	GNB5, ADCY6, ADCY8, ADCY9, PLCB1, GNAI2, GNAS, GNB2, GNG3, GNG5, GNG10, GNGT2, KCNJ3, KCNJ6, GNG13, PLCB4, ADCY10, GNB4
Bezan	Dopaminergic synapse	124	6.9942172	0.0377438	0.3748992	GNB5, DRD2, PLCB1, GNAI2, GNB2, GNG5, GNG10, GNGT2, KCNJ3, KCNJ6, GNG13, PLCB4, GNB4
BaKoya	Dopaminergic synapse	124	6.9563008	0.0493297	0.7909093	DRD2, PPP2CA, PPP2R1A, PPP2R1B, PPP2R2B, PPP2R2C, PPP2R5D
Bezan	Fanconi anemia pathway	40	6.8609694	0.0443489	0.3748992	APITD1-CORT, FANCD2, FANCG, APITD1, MLH1, FANCI, BRCA2, BRIP1, ATRIP
Baka	FoxO signaling pathway	124	7.9684064	0.017801	0.5096589	AKT3, FOXO6, C8orf44-SGK3, CDKN1A, CDKN1B, CDKN2B, PLK2, GADD45G, CHUK, PLK3, MAPK14, SGK3, PRMT1, PRKAG2, PRKAG3, MAPK10, MAPK13, PTEN, RAG1, RAG2, BCL6, SOD2, STAT3, STK4, TNFSF10
Bezan	FoxO signaling pathway	124	7.6800279	0.0199203	0.3748992	FOXO6, AKT2, IKBKB, PDPK1, PIK3CB, PIK3CD, PTEN
BaKoya	FoxO signaling pathway	124	7.4661021	0.0280687	0.7435157	FOXO6, CDKN1A, PLK4, GABARAP, GABARAPL2, CHUK, MAPK14, IKBKB, ATM, MAPK10, MAPK13, BNIP3, STAT3, STK4, CAT
Bezan	GABAergic synapse	66	7.0495569	0.0359614	0.3748992	GNB5, ADCY6, ADCY8, ADCY9, GNAI2, GNB2, GNG3, GNG5, GNG10, KCNJ6, GNG13, GNB4, CACNA1D
Baka	Gap junction	88	7.2041878	0.0409424	0.6965135	CSNK1D, TUBB, GJA1, MAPK7, MAP2K5, TUBA4A
Bezan	Gastric cancer	148	8.2750093	0.0105892	0.2897598	CTNNB1, DVL2, DVL3, AKT2, FZD2, GSK3B, PIK3CA, PIK3CB, PIK3CD
Bezan	Glutamatergic synapse	89	7.0651756	0.0352275	0.3748992	GNB5, ADCY6, ADCY8, ADCY9, GNAI2, GNAS, GNB2, GNG3, GNG5, GNG10, GRM7, KCNJ3, GNB4
Baka	Glutathione metabolism	53	8.6925579	0.0082723	0.3957579	GSTA5, GPX6, GGT7, HPGDS, GPX1, GPX3, GPX5, GPX7, GSTA3, GSTA4, MGST3
BaKoya	Glutathione metabolism	53	8.2365349	0.0127775	0.6519795	GGT6, GSTA5, GPX6, GGT7, OPLAH, GCLC, GCLM, GPX2, GPX5, ANPEP, GSTA1, GSTA2, GSTA3, GSTA4, GSTM3, GSTM5, CHAC1
Baka	Glycerolipid metabolism	59	7.938616	0.0184293	0.5096589	AGPAT1, MOGAT1, AGK, AGPAT3, DGKI
Baka	Glycerophospholipid metabolism	96	8.7575808	0.0080628	0.3957579	PLA2G15, PLA2G2D, PLA2G2E, PNPLA7, PLA2G2C, LCAT, PLA2G5, PLD2
eRHG	Glycerophospholipid metabolism	96	6.729087	0.0445633	0.7146465	LPCAT3, PEMT, LCAT, PLA2G1B, PLA2G6, PLA2G12B, EPT1, PLA2G4C
eAGR	Glycosaminoglycan degradation	19	6.6851832	0.0470292	0.991392	GUSB, HYAL1, HYAL3, HYAL2
Bezan	Hepatocellular carcinoma	168	6.8883557	0.0430908	0.3748992	DVL2, DVL3, FZD2, LRP6, WNT9A, WNT3A
Bezan	Hippo signaling pathway	151	7.9768888	0.0141539	0.2897598	CSNK1E, DVL2, DVL3, FZD2, WNT5B, WNT3A
BaKoya	HTLV-I infection	192	7.2694511	0.0344575	0.774839	DVL2, AKT2, GSK3B, IKBKB, WNT16, PIK3CB, WNT2, FZD3
Bezan	Human papillomavirus infection	299	6.8636069	0.0440344	0.3748992	DVL2, DVL3, FZD2, WNT9A, WNT3A
Bezan	Inositol phosphate metabolism	71	8.1774113	0.0112183	0.2897598	PLCD3, PLCB1, ITPKB, PIK3CA, PIK3CB, PLCD1, PIP4K2C, MTMR2
BaBongoC	Jak-STAT signaling pathway	162	9.7377514	0.0043901	0.440577	IFNA1, IFNA2, IFNA4, IFNA7, IFNA8, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNB1, IFNW1, IL4, IL6ST, JAK1, JAK2, JAK3, LEPR, IL21R, IL21, STAT4, THPO, STAM
Bezan	Jak-STAT signaling pathway	162	8.0244398	0.0133152	0.2897598	CNTFR, CSF2RB, CTF1, EGFR, EPO, IFNAR1, IL4, IL4R, IL5, IL6R, IL10RA, IL10RB, IL11RA, IL12RB1, IL13, IL21R, IL20RB, IL21, THPO
eRHG	Jak-STAT signaling pathway	162	6.7018433	0.0459264	0.7146465	CNTF, CSF2, IL3, IL4, IL5, IL12RB2, IL13, JAK2, LIF, IL20
Baka	MAPK signaling pathway	295	9.4930099	0.0048168	0.3957579	MAPK14, DUSP4, DUSP7, MAPK10, MAPK13, MAP2K6, MAPKAPK3, CDC25B

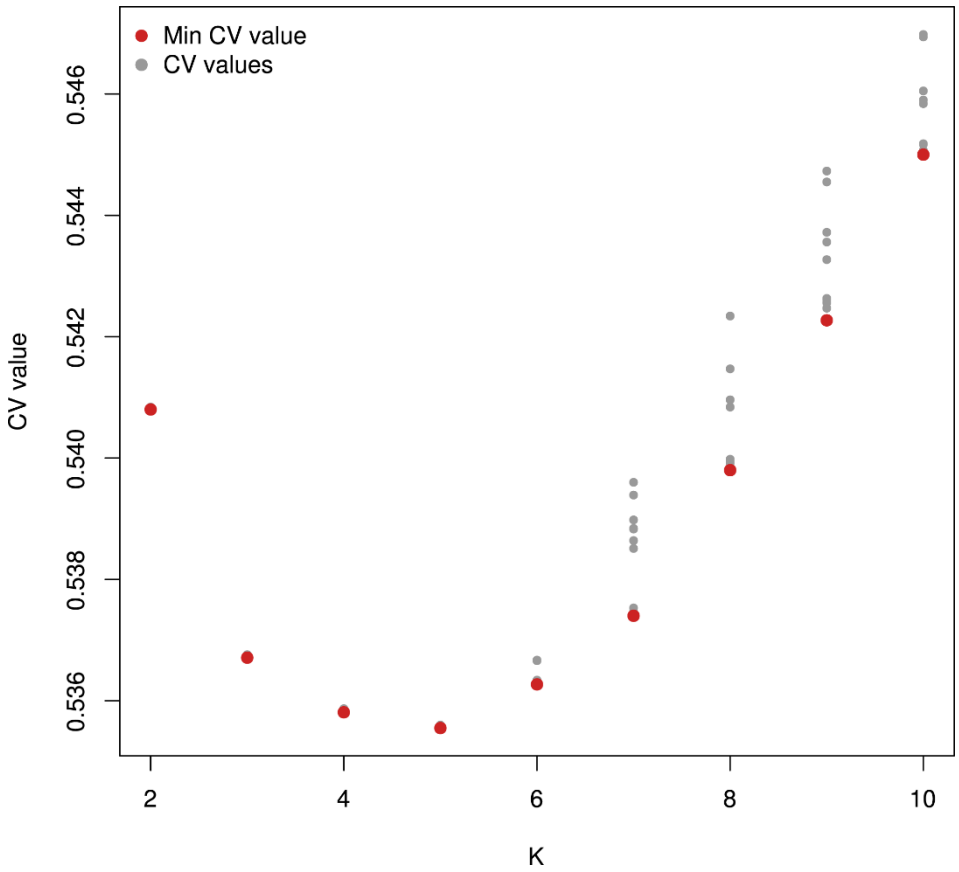
eRHG	MAPK signaling pathway	295	6.7813352	0.0407885	0.7146465	EFNA1, EFNA3, EFNA4, EGFR, ERBB2, ERBB4, FGF3, FGF4, FGF6, FGF7, FGFR3, FLT3, FLT4, FGF21, ANGPT2, KITLG, NGF, NTF4, NTRK1, PDGFB, PGF, BDNF, VEGFB, FGF23, FGF19
Bezan	Morphine addiction	54	7.0110914	0.0372195	0.3748992	GNB5, ADCY8, ADCY9, GNAI2, GNAS, GNB2, GNG5, KCNJ3, KCNJ6, OPRM1, GNG13, GNB4
Bezan	mTOR signaling pathway	142	6.8883557	0.0430908	0.3748992	DVL2, DVL3, FZD2, LRP6, WNT9A, WNT3A
BaBongoC	Natural killer cell mediated cytotoxicity	132	8.3004557	0.0120205	0.5629595	IFNA1, IFNA2, IFNA4, IFNA6, IFNA7, IFNA8, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA21, IFNAR2, IFNB1, IFNG
BaBongoC	NOD-like receptor signaling pathway	148	8.7763036	0.0078394	0.440577	IFNA1, IFNA2, IFNA4, IFNA6, IFNA7, IFNA8, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA21, IFNB1, NFKB1, TNF
Baka	Olfactory transduction	414	15.4894	9.9E-05	0	OR4X2, OR4B1, ADRBK1, OR4S2, OR4C3, OR2B6, OR12D2, OR2W1, OR2J2, OR2H1, OR4X1, OR5D13, OR4A47, OR2B3, OR14J1, OR10C1, OR4P4, OR4C15, OR5V1, OR12D3
wAGR	Olfactory transduction	414	18.884268	9.9E-05	0	OR5I1, OR4X2, OR4B1, OR9A2, OR4C16, OR8K5, OR5T2, OR8J1, OR5R1, OR5M3, OR5M11, OR9Q1, OR9Q2, OR1S1, OR10Q1, OR4C3, OR4S1, OR12D2, OR11A1, OR2W1, OR5F1, OR5AP2, OR6V1, OR4X1, OR5D16, OR5W2, OR8H2, OR8H3, OR5T3, OR5T1, OR8K1, OR5M9, OR5M10, OR5M1, OR9G1, OR5AK2, OR5B12, OR5C1, OR5B3, OR2B3, OR14J1, OR10C1, PRKACG, OR51B2, OR8J3, OR10W1, OR5V1, OR12D3
eAGR	Olfactory transduction	414	17.23376	9.9E-05	0	OR2D2, OR9A2, ADRBK1, ADRBK2, OR4C16, OR4C11, OR4S2, OR4C6, OR5L1, OR5D18, OR5AS1, OR5T2, OR8H1, OR8K3, OR5R1, OR5M3, OR5M8, OR5AR1, OR1S1, OR10G3, OR5J2, OR10A2, OR4X1, OR5W2, OR8H2, OR8H3, OR5T3, OR5T1, OR5M10, OR10J3, OR8J3, OR4P4
BaKoya	Olfactory transduction	414	14.142177	0.0001047	0.0259065	OR5I1, OR8I2, OR7G1, OR1M1, OR13C3, OR7D2, OR8U1, OR4C16, OR4S2, OR4C6, OR5L1, OR8H1, OR8K3, OR5R1, OR10G9, OR9I1, OR9Q2, OR5B17, OR5A2, OR4D11, OR2B6, OR5L2, OR5K1, OR10G2, OR4E2, OR4E1, OR2J2, OR2H1, OR5F1, OR5AP2, OR2AT4, OR6F1, OR5D16, OR5W2, OR8H2, OR8H3, OR5M9, OR5M10, OR5M1, OR9G1, OR6X1, OR10G4, OR10G7, OR5AU1, OR7G2, OR10K2, OR5K4, ARRB2, OR2B3, OR14J1, OR10C1, OR8J3, OR4P4, OR10W1, OR2G2, OR2C3, OR5V1, OR12D3
BaBongoC	Olfactory transduction	414	10.841366	0.0027177	0.440577	OR5I1, OR4X2, OR10J5, ADRBK1, OR5AS1, OR8K5, OR5T2, OR5R1, OR4D11, OR4C3, OR4S1, OR10AG1, OR5J2, OR8S1, OR8H3, OR5T1, OR5AN1, OR11G2, OR11H4, OR4A47, OR1F1, OR2C1, OR8J3, OR4P4
Bezan	Olfactory transduction	414	9.0637044	0.0056616	0.2897598	OR4D2, OR10H4, OR10T2, OR6P1, ADRBK2, OR4C6, OR5M8, OR5M11, OR9I1, OR9Q2, OR10Q1, OR6C74, OR6C3, OR2F1, OR2B6, OR2K2, OR7C2, OR4D1, OR2J2, OR6C6, OR6C4, OR10R2, OR5M1, OR10K1, OR6K3, OR2A25, OR2A5, OR6C68, OR2C1, GNG13, OR4C15, OR2AE1
eRHG	Olfactory transduction	414	8.2849511	0.0089127	0.7146465	OR4X2, ADRBK1, OR4C16, OR4C11, OR4S2, OR5D18, OR9Q1, OR4S1, OR5L2, OR10J1, OR6V1, OR52D1, GNG13, OR4P4
Baka	One carbon pool by folate	20	7.060446	0.045445	0.6965135	MTFMT, GART, AMT
BaBongoC	Osteoclast differentiation	119	7.7070658	0.0204871	0.8224104	FCGR3B, TAB2, RAC1, SYK, TEC, TRAF2, TNFRSF11A
BaKoya	Pathways in cancer	454	8.149802	0.0138249	0.6519795	IFNA1, IFNA2, IFNA4, IFNA6, IFNA7, IFNA8, IFNA10, IFNA21, IFNG, IL3, IL4R, IL6, IL7, IL12RB1, JAK1, JAK2, JAK3, PIM1, STAT3, STAT4, STAT5A, STAT5B, VEGFA
Bezan	Phosphatidylinositol signaling system	86	8.3413059	0.010065	0.2897598	PLCD3, DGKB, SACM1L, PLCB1, ITPKB, PIK3C3, PIK3CB, PIK3CD, PLCD1, INPP4B

BaKoya	Phosphatidylinositol signaling system	86	7.3834715	0.0300587	0.7435157	DGKB, IMPA1, PIK3C2G, PIK3CB, PIP4K2A, PLCZ1, DGKI
BaBongoC	Phosphatidylinositol signaling system	86	7.3008556	0.0330302	0.9927668	PLCE1, PI4KA, TMEM55A, PIP4K2B, SYNJ2, MTMR3, TMEM55B
BaKoya	PI3K-Akt signaling pathway	351	10.137443	0.0037704	0.4663165	CREB1, AKT2, MCL1, ATF4, CREB3L2, CREB5
BaBongoC	PI3K-Akt signaling pathway	351	9.2974629	0.0054354	0.440577	CSF3, CSH1, CSH2, GH1, GH2, IFNA1, IFNA2, IFNA4, IFNA6, IFNA7, IFNA8, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA21, IFNB1, IL2, IL2RA, IL3, IL4, IL4R, JAK1, JAK2, JAK3, OSM
wAGR	PI3K-Akt signaling pathway	351	7.9995378	0.0133779	0.8516931	CHUK, THEM4, CREB3L4, CRTC2, AKT2, PDPK1, PPP2R3C, PPP2R3A, PPP2R5C, PPP2R5E
Bezan	Prostate cancer	85	6.8430682	0.0455022	0.3748992	EGFR, AKT2, IGF1R, INSRR, PDPK1, PIK3CA, PIK3CB, PIK3CD, PTEN
BaBongoC	Purine metabolism	171	8.9765013	0.0067942	0.440577	ENTPD8, PDE11A, GMPR2, PDE1A, PDE4D, PDE6B, PDE8B
Baka	Purine metabolism	171	8.8069718	0.0077487	0.3957579	POLD3, PDE10A, NUDT16, POLA2, NT5C, ITPA, ENTPD8, NT5C3A, PDE1C, PDE6D, PKM, POLR3E
Bezan	Purine metabolism	171	8.1106812	0.0119522	0.2897598	NME6, PDE10A, ADCY6, ADCY9, FHIT, GUCY1A2, GUK1, NT5C, ENTPD8, PDE11A, PDE2A, POLR2D, POLR2F, POLR2H, NTPCR, POLR1C, ENTPD3
wAGR	Purine metabolism	171	7.4980823	0.0234114	0.9397993	NME6, ADCY3, FHIT, POLR1A, NUDT2, PKM, POLR2F, PNPT1
eRHG	Purine metabolism	171	6.9928432	0.0327147	0.7146465	ADCY3, NT5C2, AK5, GUK1, PDE11A, PDE2A, PDE6A
eAGR	Purine metabolism	171	6.7965544	0.0425152	0.991392	ADK, ADSS, AK4, TWISTNB, NT5C2, AK5, GUK1, NPR1, GMPR2, PDE3B, PDE6D, POLR2D, POLR2F, POLR2K, ADCY10, NME1-NME2, PNPT1
BaKoya	Pyrimidine metabolism	102	7.9950251	0.0158148	0.6519795	POLR3G, POLA2, ENTPD8, NME1, NME2, NT5C3A, ENPP1, POLR2A, POLR2I, PRIM2, NME1-NME2, UMPS, UCK1, ENTPD6, CDA
Baka	Pyrimidine metabolism	102	7.7800551	0.0223037	0.5607301	POLD3, POLA2, ENTPD8, NT5C3A, ENPP1, CMPK1, POLR3E, POLR1C
wAGR	Pyrimidine metabolism	102	7.1165476	0.0356396	0.9872492	POLR3C, POLR1A, NUDT2, POLR2C, POLR2F, POLR3E, RRM1, PNPT1, ENTPD6
Bezan	Pyrimidine metabolism	102	6.9341702	0.0408891	0.3748992	NME6, POLD3, NUDT2, ENTPD8, POLR2D, POLR2F, POLR2H, POLR2K, POLR3D, ENTPD3
BaKoya	Ras signaling pathway	229	8.1332951	0.0141391	0.6519795	GNB5, MRAS, NF1, NRAS, PIK3CB, RASGRF1, RASGRF2, SOS1, RASAL2
Baka	Ras signaling pathway	229	6.9993926	0.0478534	0.6965135	RASSF1, PAK1, PIK3R1, RAC3, RRAS, TIAM1, PIK3R3
Bezan	Relaxin signaling pathway	130	8.0091878	0.0135248	0.2897598	GNB5, AKT2, PLCB1, GNAI2, GNB2, GNG3, GNG5, GNG10, GNGT2, GNG13, PIK3CA, PIK3CB, PIK3CD, PLCB4, GNB4, SRC
Bezan	Retrograde endocannabinoid signaling	106	7.2170093	0.0309289	0.3748992	GNB5, ADCY6, ADCY9, GNAI2, GNB2, GNG13, PRKACB, MAPK10, GNB4
eRHG	Salmonella infection	72	6.7741563	0.0413128	0.7146465	ARPC1B, ARPC1A, FLNB, PFN4, ACTB, ACTG1
Bezan	Serotonergic synapse	80	7.1062552	0.03355	0.3748992	GNB5, PLCB1, GNAI2, GNB2, GNG3, GNG10, HTR1A, KCNJ3, KCNJ6, GNG13, PLCB4, GNB4
Bezan	Signaling pathways regulating pluripotency of stem cells	109	6.7689296	0.0495911	0.3748992	DVL3, FZD2, WNT9A, WNT3A
BaKoya	Sphingolipid signaling pathway	96	7.183898	0.0387516	0.7909093	MAPK14, AKT2, PPP2CA, PPP2R1A, PPP2R1B, PPP2R2B, PPP2R2C, PPP2R3A, PPP2R2D, PRKCZ, MAPK13
Baka	Staphylococcus aureus infection	36	7.3931768	0.0347644	0.6965135	CFB, C2, C3, C4A, C4B
Baka	Systemic lupus erythematosus	19	7.2058039	0.0409424	0.6965135	C2, C4A, C4B
wAGR	Thyroid hormone signaling pathway	110	8.224422	0.0112876	0.8516931	NCOA2, SIN3A, MED4, MED1, RXRB, THRA, MED24
Baka	Tuberculosis	173	7.1039758	0.0445026	0.6965135	CREB1, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1

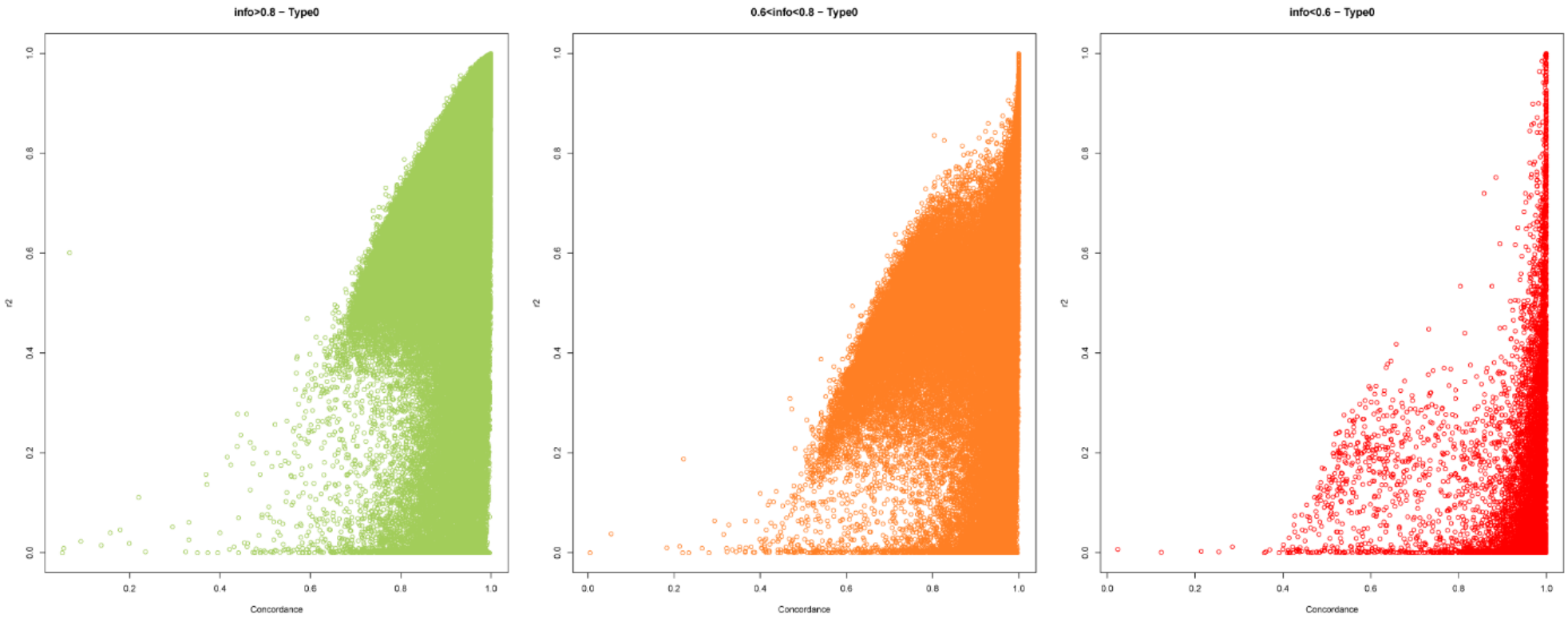
wAGR	Vascular smooth muscle contraction	114	7.7256238	0.0181856	0.8516931	AGTR1, GNA12, RHOA, PPP1R12A, ROCK1, ARHGEF1, PPP1R14A
BaBongoC	Vascular smooth muscle contraction	114	7.2283827	0.0353298	0.9927668	KCNMB2, MYL9, KCNMA1, MYH11, MYLK, CALML5, PPP1CA, PRKG1, ACTG2
Bezan	Wnt signaling pathway	144	8.5089989	0.0081778	0.2897598	FZD2, DKK2, LRP6, SFRP4, WNT9A, WNT3A
eRHG	Wnt signaling pathway	144	7.282413	0.0234875	0.7146465	LRP5, PLCB3, SFRP4, WNT5A, WNT6, WNT8A, WNT11, FZD6, WNT3A



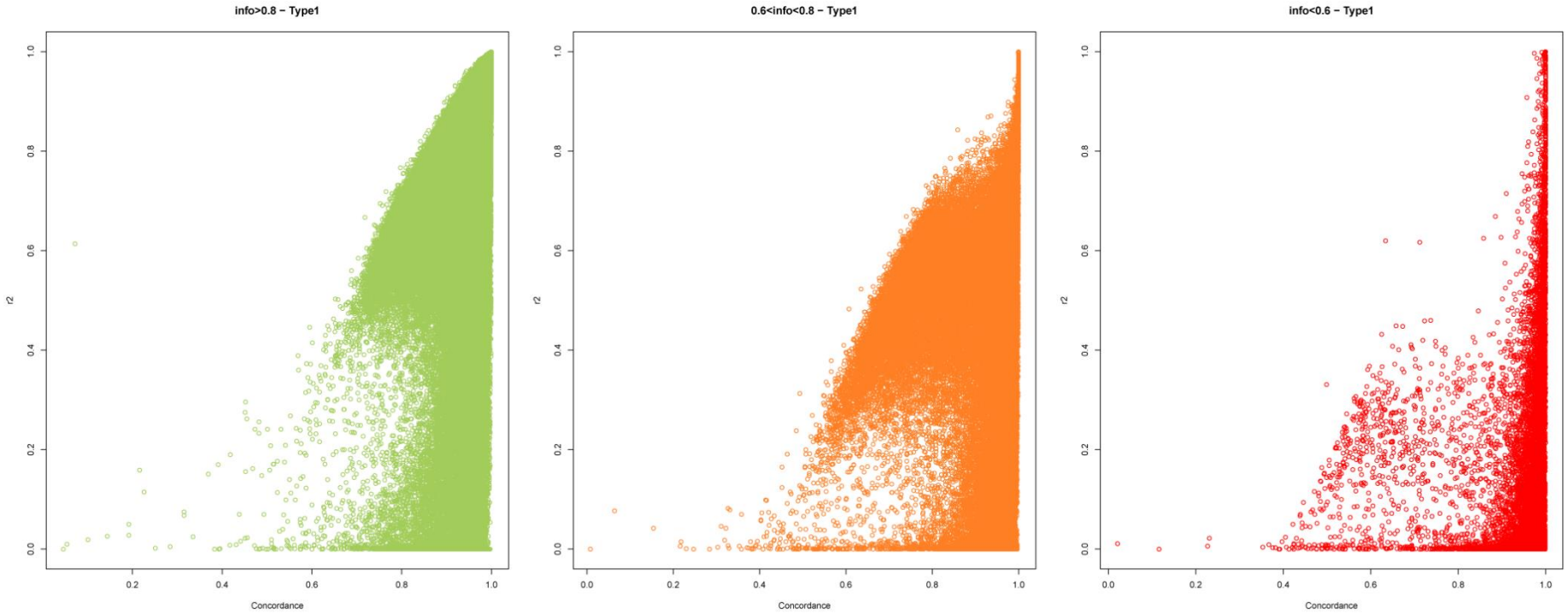
Supplementary Figure 1. Datasets used in this study. Summary of the data processing performed in this study, colors represent the nature of the data and the arrow indicates the correspondence between individuals analyzed in both datasets.



Supplementary Figure 2. Cross-validation (CV) values. CV values of different *K* values from the ADMIXTUE analysis (Fig. 1C). The minimum CV value obtained across ten independent runs of ADMIXTURE is represented in red, other values in gray.

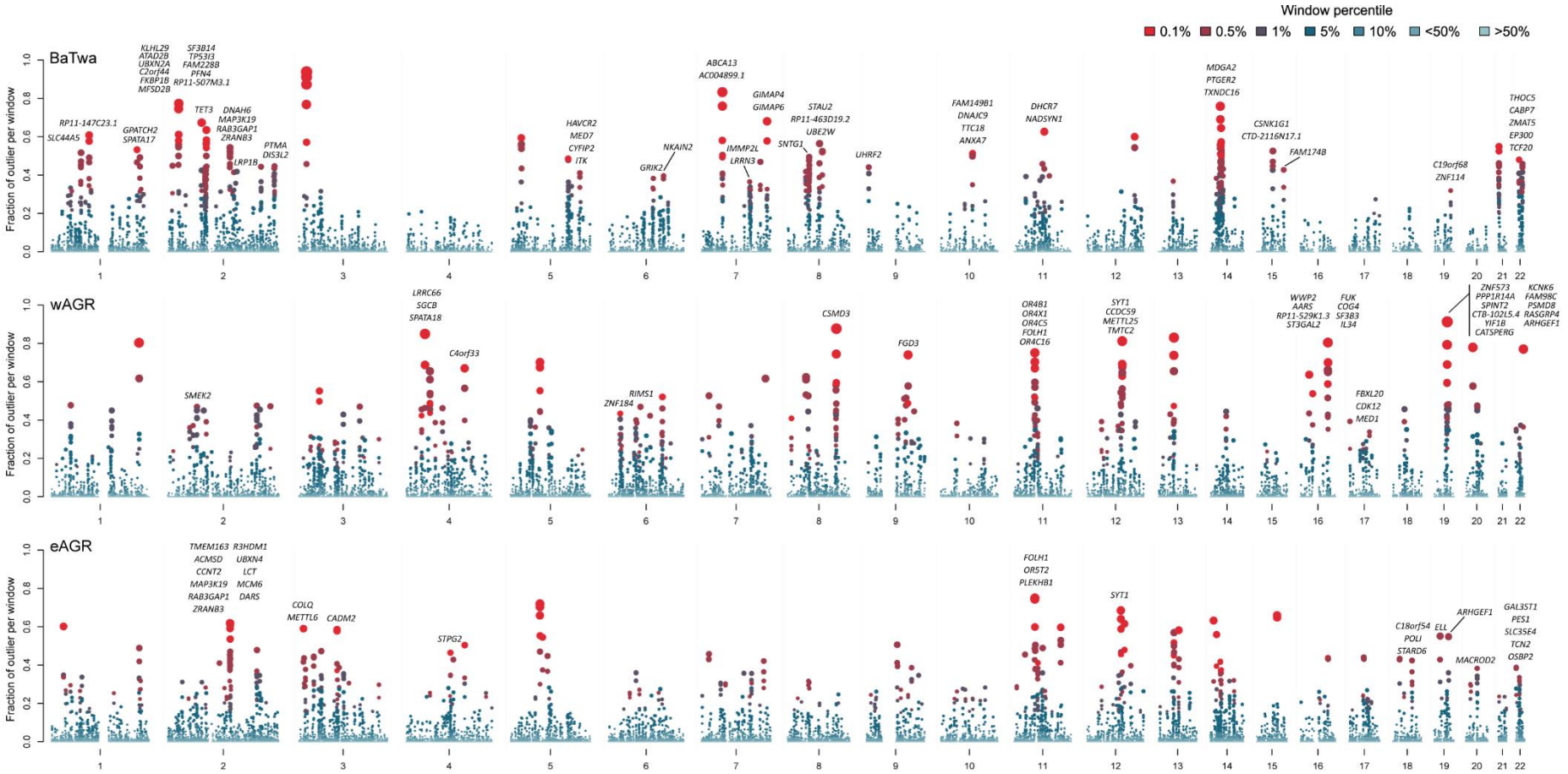


Supplementary Figure 3. Imputation accuracy with 1,000 Genomes dataset. Estimated correlation coefficient r^2 and allele concordance between true genotypes (obtained by genotyping or exome sequencing) and imputed genotypes for the same SNPs (obtained by artificially removing genotyped SNPs from the data and then imputing them). Info indicate the information metrics of the imputed SNPs.

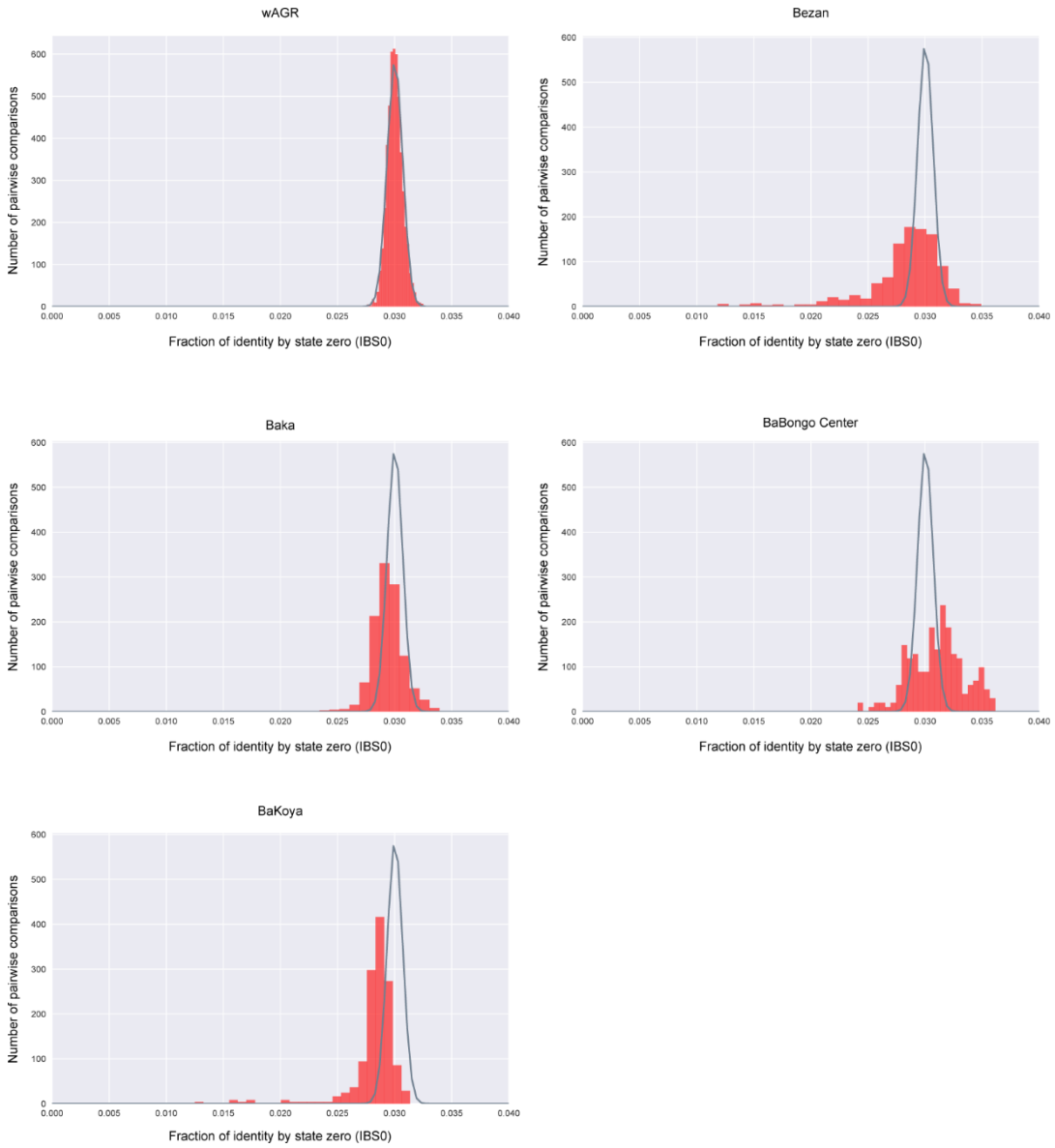


Supplementary Figure 4. Imputation accuracy with Nzebi/Baka WGS sequences. Estimated correlation coefficient r^2 and allele concordance between true genotypes (obtained by genotyping or exome sequencing) and imputed genotypes for the same SNPs (obtained by artificially removing genotyped SNPs from the data and then imputing them). Info indicate the information metrics of the imputed SNPs.

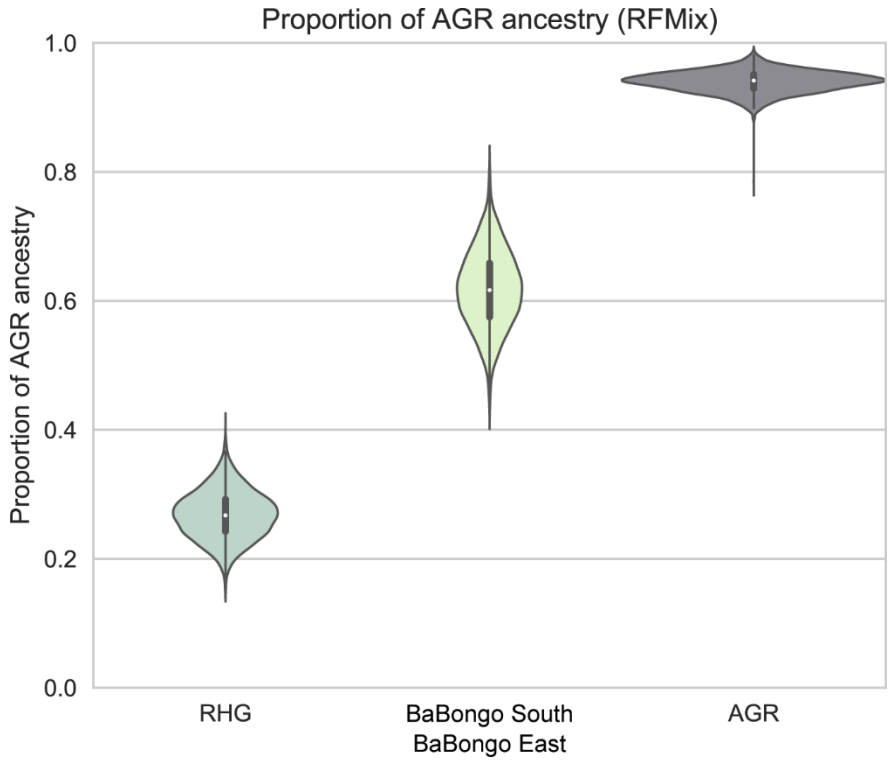
Supplementary Figure 5. Genome-wide signals of natural selection in wRHG populations. Proportions of outlier SNPs (Fsc in top 1% of the empirical distribution) in 100kb windows in populations of Bezan, Baka, BaBongo Center and BaKoya. Only candidate genes in windows within the top 0.5% of the enrichment of outlier and presenting at least one high scoring SNPs (Fsc > 10) are presented. Combined selection signals have been computing using EUR as an outgroup population.



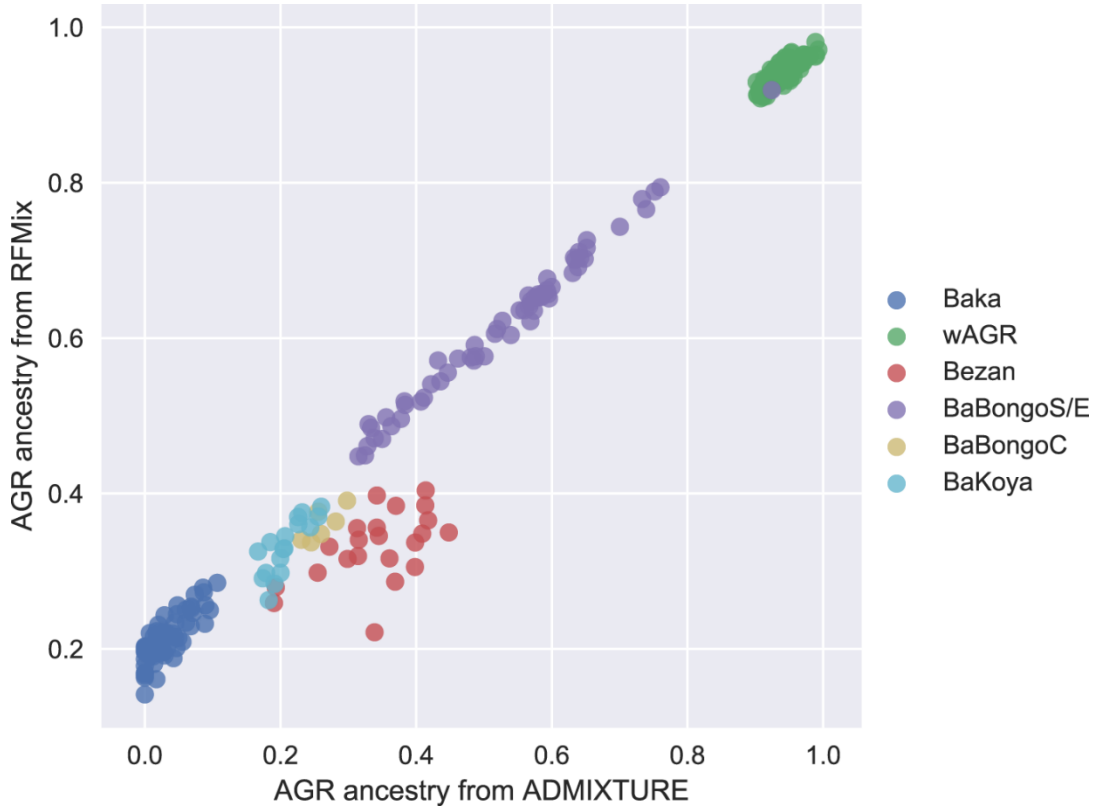
Supplementary Figure 6. Genome-wide signals of natural selection in eRHG, wAGR and eAGR populations. Proportions of outlier SNPs (Fsc in top 1% of the empirical distribution) in 100kb windows in populations of eRHG, wAGR and eAGR. Only candidate genes in windows within the top 0.5% of the enrichment of outlier and presenting at least one high scoring SNPs (Fsc > 10) are presented. Combined selection signals have been computing using EUR as an outgroup population.



Supplementary Figure 7. Distribution of IBS0 in western populations. Distribution (red) of the fraction of sites with no identity by state (IBS0) in pairwise comparisons of individuals in western populations of AGR (wAGR) and RHG (Bezan, Baka, BaBongo Center, BaKoya). The distribution of IBS0 in wAGR (fitted line in gray) is reported in all plots to facilitate comparisons.



Supplementary Figure 8. Mean AGR ancestry estimated with RFMix. Distribution of the mean AGR ancestry estimated by RFMix in each individual from parental populations (AGR and RHG) and admixed RHG (BaBongo South and BaBongo East).



Supplementary Figure 9. Correlation of the mean AGR ancestry estimated with RFMix and ADMIXTURE. Correlation of the mean AGR ancestry estimated in western AGR and RHG individuals with RFMix and ADMIXTURE at *K*=2.

6.3 Résumé des résultats et nouveautés

Afin d'évaluer la contribution des différents mécanismes de sélection positive ayant eu lieu dans les différentes populations de Pygmées, nous avons d'abord réalisé un scan identifiant les signaux de balayages sélectifs classiques. Pour cela, nous avons utilisé un score de sélection combinant cinq statistiques de sélection positive, à la fois intra-populationnelles (|iHS|, |ΔiHH|), et inter-populationnelles (PBS, XP-EHH et $|\Delta iHH_D|$). En premier lieu, nous avons réalisé un scan permettant d'identifier les signaux de sélection potentiellement partagés par toutes les populations de Pygmées en les comparant aux populations de non-Pygmées, et en utilisant une population européenne comme population fortement différenciée (outgroup). En règle générale, les approches de sélection positive par valeur extrême (outlier approach) identifient de nombreux faux positifs attribuables aux variations locales des profils de diversité le long du génome. Pour limiter ces effets, nous avons considéré des critères très stringents pour identifier les gènes candidats, en ne considérant que les fenêtres génomiques les plus enrichies en SNPs ayant des scores extrêmes de sélection (top 0.5% de la distribution empirique de l'enrichissement en SNPs dont le score est dans le top 1% du génome), et au moins un SNP avec un score combiné supérieur à 10 localisé dans la séquence du gène.

Nos résultats mettent tout d'abord en évidence une région de 31 Mb sur le chromosome 3 présentant des signatures de sélection positive dans plusieurs groupes de Pygmées et préalablement proposée comme candidate par plusieurs études (Lachance et al., 2012 ; Jarvis et al., 2012). En considérant les populations séparément, cette région apparaît sous très forte sélection positive chez les Bedzan du Cameroun et quelques fenêtres génomiques sont également sous sélection chez les populations de Baka du Cameroun et du Gabon et chez les BaBongo du Centre du Gabon. Parmi les gènes de cette région qui présentent un SNP dont le score est particulièrement élevé, nous avons identifié le gène *MAPKAPK3* impliqué dans la régulation des certaines infections virales et bactériennes ainsi que *ARGHEF3*, qui est associé au volume sanguin de plaquettes. Ces résultats pourraient indiquer un événement de sélection ancien dans le groupe ancestral des populations de Pygmées de l'ouest et dont les signatures ont probablement été exacerbées chez les Bedzan, qui présentent un niveau élevé d'homozygotie dû à la consanguinité. A l'exception des gènes *EEF2K*, qui permet la résistance cellulaire à l'absence de nutriments et *POLR3E* impliqué dans la détection d'ADN viral, dont les signaux sont partagés par au moins deux populations de Pygmées, la recherche de gènes candidats sur le reste du génome a mis en évidence un faible niveau de partage des signatures de sélection entre ces populations.

Nous avons réalisé un second scan de sélection comparant les groupes de Pygmées entre eux afin d'identifier des signaux spécifiques à chaque population. Ces résultats mettent en évidence un grand nombre de gènes candidats impliqués dans des voies métaboliques liées à l'immunité, à la résistance à l'insuline ou aux hormones thyroidiennes et dont l'association au phénotype pygmée avait été suggéré par d'autres études. Cependant nos résultats indiquent

que plusieurs gènes candidats sont impliqués dans plusieurs de ces voies métaboliques comme c'est le cas de *PIAS1*, *PPARD*, *PTGDS* et *SERPINA12* qui régulent à la fois des fonctions immunitaires et la résistance à l'insuline. L'interaction des fonctions inflammatoires et de la sensibilité à l'insuline a fait l'objet de nombreux travaux et nos résultats suggèrent un effet pléiotropique des signaux de sélection chez les Pygmées qui impliquerait ces deux voies métaboliques. Ces résultats soulèvent également la possibilité que la sélection aurait pu agir sur des fonctions qui ne seraient pas directement liées à la faible stature des Pygmées mais aurait pu agir sur l'immunité et la régulation de l'inflammation.

Nous avons évalué la contribution des mécanismes de sélection polygénique aux phénotypes adaptatifs chez les Pygmées en recherchant des réseaux de gènes enrichis en signaux de sélection au sein de voies métaboliques. Nos résultats indiquent que les populations de Pygmées présenteraient un enrichissement de signaux de sélection principalement dans deux voies métaboliques : "Jak-STAT signaling pathway" et "FoxO signaling pathway". Ces deux voies métaboliques sont impliquées dans le régulation des fonctions immunitaires et FoxO permet la transduction du signal de IGF-1, impliqué dans un grand nombre de processus cellulaires tels que la croissance, la régulation des niveaux de lipides et de glucose dans le sang et les processus immunitaires. De plus, dans les populations de Pygmées les plus métissées avec les agriculteurs, ces mêmes voies métaboliques présentent également un enrichissement en segments génomiques d'origine Pygmées, ce qui indique que ces voies métaboliques sont particulièrement associées aux populations Pygmées.

Ces résultats suggèrent que des gènes impliqués à la fois dans la régulation des processus immunitaires et les voies métaboliques reliées à l'insuline seraient sous sélection polygénique convergente chez les populations de Pygmées d'Afrique centrale.

Chapitre 7
Discussion générale de la thèse

7.1 Vers un tableau plus complet de l'histoire démographique des populations d'Afrique centrale ?

7.1.1 Des populations ancestrales structurées de chasseurs-cueilleurs

La disponibilité croissante des données génétiques pour un grand nombre de populations humaines a permis de caractériser en détail la structure des populations qui ont colonisé l'Afrique sub-Saharienne au corus des derniers millénaires (Busby et al., 2016 ; Gurdasani et al., 2015 ; Patin et al., 2017). Ces résultats apportent de nouveaux éléments sur l'histoire des migrations bantoues depuis l'ouest de l'Afrique centrale vers l'est et le sud du continent, il y a environ 4 000 à 5 000 ans (Patin et al., 2017). Cependant, les événements démographiques antérieurs à cette époque restent largement méconnus. Nos travaux sur la réévaluation du modèle démographique des Pygmées et non-Pygmées d'Afrique centrale nous ont permis d'estimer avec précision les temps de divergence entre les groupes ainsi que leurs fluctuations de tailles de populations au cours des 200 000 dernières années. Ces résultats ont mis en évidence une divergence ancienne de la population ancestrale des Pygmées de l'ouest et de l'est datant de la fin du Pléistocène. Il est important de noter que les différences entre les premières estimations faites à partir de données autosomales limitées, datant leur divergence à 60-70 000 ans (Verdu et al., 2009 ; Patin et al., 2009 ; Batini et al., 2011 ; Veeramah et al., 2012), sont compatibles avec nos dernières estimations, une fois que les mêmes taux de mutation et temps générationnel sont utilisés. De plus, ces analyses ont montré que la taille effective de la population ancestrale des Pygmées était quatre à neuf fois supérieure à celle des populations actuelles, et au moins égale à la taille de la population ancestrale des agriculteurs. L'origine commune des Pygmées, leur phénotype partagé et la forte structure génétique observée entre les groupes soulèvent de nombreuses questions sur l'occupation préhistorique des forêts équatoriales africaines, et sur la répartition géographique des populations ancestrales et leurs interactions avec d'autres groupes de chasseurs-cueilleurs.

Le processus de diversification des populations humaines en Afrique au cours du Pléistocène est difficile à adresser sur le plan archéologique en raison de la quasi absence de restes

fossiles en Afrique de l'ouest et en Afrique centrale. En effet, le climat chaud et humide de cette région favorise la dégradation rapide des restes fossiles et ce phénomène est accentué par les sols acides de la forêt équatoriale. Quelques études génétiques ont montré que les temps de divergence les plus anciens observés en Afrique sont rapportés entre les différentes populations de chasseurs-cueilleurs organisés en communautés de petites tailles, tels que les Pygmées, les chasseurs-cueilleurs Khoe-San en Afrique du sud et les Hadza de Tanzanie (Figure 1.1) (Gronau et al., 2011 ; Schlebusch et al., 2012 ; Veeramah et al., 2012 ; Hsieh et al., 2016a ; Schlebusch et al., 2017). Ces observations, associées aux données linguistiques, suggèrent l'existence de ces groupes sur le continent africain depuis plusieurs dizaines de milliers d'années.

Une étude récente réalisée à partir de 16 génomes préhistoriques africains a caractérisé pour la première fois la structure génétique ancienne et les interactions entre les populations africaines sub-Sahariennes au cours de 8 000 dernières années (Skoglund et al., 2017). Ces travaux ont mis en évidence le remplacement de plusieurs populations de chasseurs-cueilleurs au Kenya, en Tanzanie et au Malawi par des groupes d'agriculteurs à la suite des expansions bantoues (Skoglund et al., 2017). Ces analyses ont également relevé qu'au cours de ces remplacements de populations, les chasseurs-cueilleurs préhistoriques n'ont quasiment pas contribué à la diversité génétique des populations agricultrices nouvellement arrivées (Skoglund et al., 2017). Ces résultats obtenus pour des populations de chasseurs-cueilleurs non forestiers contrastent avec les observations réalisées dans notre étude chez les populations de Pygmées qui présentent des signatures de métissage ancien avec les ancêtres des agriculteurs et qui s'est intensifié à une époque plus récente. Il semblerait donc que l'histoire des interactions entre Pygmées et non-Pygmées en Afrique centrale soit différente de celle des chasseurs-cueilleurs non forestiers, probablement en lien avec l'exploitation des ressources de la forêt ayant favorisé la mise en place de relations économiques. Ces résultats soutiennent donc l'hypothèse d'une structure ancienne des populations africaines mais la question de la structure des populations ancestrales de Pygmées ainsi que leur répartition géographique à l'intérieur ou en lisière des forêts avant l'arrivée des agriculteurs reste ouverte.

7.1.2 Populations préhistoriques africaines et homininés archaïques

Les résultats obtenus à partir d'ADN ancien (Skoglund et al., 2017) apportent un nouvel éclairage à l'histoire complexe des populations préhistoriques sub-Sahariennes. Cependant, ces données sont relativement "récentes" (environ 8 000 ans) et n'incluent aucun fossile de chasseur-cueilleur forestier, probablement en raison des mauvaises conditions de conservation mais aussi de la difficulté d'accès aux sites archéologiques dans des zones de conflit armé. La découverte récente d'un site archéologique en RDC a révélé la présence d'une centaine de fragments d'os d'homme moderne datant du Pléistocène Supérieur soit environ -25 000 à -20 000 ans (Crevecoeur et al., 2016). Ces restes appartiennent à une communauté de chasseurs-cueilleurs-pêcheurs dont la diversité phénotypique était plus grande que celle des populations

africaines actuelles, et présentent plus de similitudes morphologiques avec d'autres fossiles du Pléistocène Moyen et Inférieur qu'avec les populations locales modernes (Crevecoeur et al., 2016). L'obtention de données génétiques à partir de ces échantillons et leur intégration dans un modèle démographique incluant un plus grand nombre de populations de Pygmées, en particulier les populations de RDC très peu représentées, serait une avancée considérable dans l'étude de l'histoire de la préhistoire africaine. En effet, l'estimation d'un tel modèle apporterait de nombreux éléments à la compréhension des phénomènes de diversification et de dispersion de l'homme moderne en Afrique au cours du Pléistocène, marqué par des événements climatiques majeurs comme le Dernier Maximum Glaciaire.

Nos résultats ont mis en évidence l'existence d'un métissage ancien entre les populations ancestrales de Pygmées et de non Pygmées, cependant notre modèle ne prend pas en compte l'estimation du métissage des populations d'Afrique centrale avec des homininés archaïques. La cohabitation et le métissage de populations modernes avec des populations archaïques ont été démontrés à plusieurs reprises dans les populations européennes et asiatiques (Green et al., 2010 ; Prufer et al., 2014 ; Reich et al., 2010, 2011 ; Sankararaman et al., 2014) et ces événements ont parfois permis l'introgression dans les populations modernes de loci adaptatifs (Huerta-Sanchez et al., 2014 ; Vernot & Akey, 2014 ; Racimo et al., 2015, 2017 ; Quach et al., 2016 ; Deschamps et al., 2016). Dans les populations Africaines, l'absence d'ADN provenant d'homininés archaïques connus empêchent la comparaison directe de leurs séquences avec les populations modernes. Cependant l'utilisation d'approches basées sur la divergence des séquences et le déséquilibre de liaison a permis d'identifier des événements putatifs d'introgression entre les populations africaines et un homininé inconnu (Hammer et al., 2011 ; Lachance et al., 2012 ; Hsieh et al., 2016b). Ces études indiquent une introgression de matériel génétique archaïque datant de 35 000 ans chez les populations de chasseurs-cueilleurs Biaka, Mbuti et Khoe-San et constituant près de 2% de leur génome, avec un homininé qui aurait divergé des populations modernes il y a 700 000 ans (Hammer et al., 2011). Comme en Europe et en Asie, le mélange génétique avec des populations archaïques a pu être adaptatif. Un exemple de cette adaptation a été montré pour le gène *MUC7* qui code une protéine de la salive et qui est porté par un haplotype archaïque chez certaines populations africaines (Xu et al., 2017).

7.2 Efficacité de la sélection purificatrice : Quelles conclusions pour les populations humaines ?

7.2.1 Intérêts et confusion des statistiques mesurant le fardeau de mutations délétères

Nos travaux sur l'inférence démographique des populations d'Afrique centrale ont montré une réduction considérable de la taille des populations de Pygmées de l'ouest et de l'est, ainsi qu'une expansion démographique des groupes de non-Pygmées au cours des 20 000 dernières

années. Cependant, ces variations récentes de N_e n'ont pas eu d'impact sur l'efficacité de la sélection purificatrice et donc le fardeau de mutations délétères porté par ces populations. Ces résultats s'ajoutent à plusieurs travaux qui ont évalué l'impact des différentes histoires démographiques humaines sur leurs fardeaux de mutations délétères. Cependant, plusieurs de ces travaux ont abouti à des conclusions opposées qui rendent difficile l'interprétation des résultats dans un contexte général (Lohmueller et al., 2008 ; Henn et al., 2016 ; Casals et al., 2013 ; Do et al., 2015 ; Simons et al., 2014 ; Pedersen et al., 2017 ; Peischl et al., 2018).

En pratique, il n'est pas possible de mesurer directement la valeur sélective d'un individu car la distribution des coefficients de sélection et de dominance, ainsi que les effets délétères combinés des différents loci, ne sont pas connus. Cependant, la disponibilité croissante de données de séquençage pour un grand nombre d'individus a permis de quantifier empiriquement le nombre et la distribution des mutations délétères portées par les populations humaines. Les études menées sur des données empiriques ont utilisé différentes statistiques comme approximation du fardeau de mutations délétères dans les populations humaines (Lohmueller et al., 2008 ; Henn et al., 2016 ; Casals et al., 2013 ; Do et al., 2015 ; Simons et al., 2014 ; Pedersen et al., 2017 ; Peischl et al., 2018). Certaines études ont observé une diminution du nombre de mutations non-synonymes hétérozygotes et une augmentation d'allèles dérivés homozygotes dans les populations européennes en comparaison aux populations africaines (Lohmueller et al., 2008) et ces effets augmentent avec la distance des populations au continent africain (Henn et al., 2015, 2016). D'autres travaux ont montré en revanche qu'il n'existait pas de différence dans le nombre de mutations potentiellement délétère portées par chaque individu, quelque soit la population considérée (Simons et al., 2014 ; Simons & Sella, 2016 ; Do et al., 2015), concluant alors que l'histoire démographique des populations humaines n'a pas de conséquence sur leurs fardeaux de mutations délétères. D'autres études ont détecté des différences dans le nombre de mutations non-synonymes par individu mais n'ont en revanche pas réalisé de tests statistiques prenant en compte la variabilité des processus évolutifs en jeu (Henn et al., 2016, 2015 ; Simons & Sella, 2016).

La raison majeure pour laquelle différentes études ont conclu à l'existence ou non de différences de fardeaux de mutations délétères et donc d'efficacité de la sélection purificatrice entre les populations humaines réside dans l'emploi de différents estimateurs et de tests statistiques plus ou moins appropriés (Simons & Sella, 2016). Des travaux de simulations ont montré que le nombre de mutations non-synonymes dérivées portées par un individu est directement lié à son fardeau de mutations délétères en faisant l'hypothèse d'un modèle additif (Simons & Sella, 2016). De plus, cette étude suggère de prendre en compte la stochasticité des processus mutationnels et généalogiques afin de tester la significativité des différences entre populations. En d'autres termes, en répétant le même scénario démographique plusieurs fois, le nombre de mutations non-synonymes dérivées par individu varie considérablement entre ces deux expériences et il est préférable d'utiliser des approches de bootstrap par bloc génomique afin de prendre en compte tous les facteurs qui peuvent varier

le long du génome (Simons & Sella, 2016 ; Simons et al., 2014 ; Do et al., 2015).

Dans un modèle additif (sans d'interactions entre les loci), bien que les événements démographiques puissent modifier de façon considérable le nombre et la fréquence des mutations délétères dans une population, ces effets ont tendance à se compenser aboutissant à un fardeau de mutations délétères identique pour tous les individus (Simons et al., 2014 ; Simons & Sella, 2016 ; Do et al., 2015). De plus, les événements démographiques ayant eu lieu dans l'espèce humaine sont trop récents pour avoir entraîné la fixation d'allèles délétères augmentant le fardeau de mutations de façon définitive dans les populations modernes (Simons et al., 2014). En revanche, certains travaux ont évalué le fardeau de mutations délétères des populations archaïques de Néandertal et de Denisova, qui présentent des tailles effectives réduites (Meyer et al., 2012) et mis en évidence une efficacité de la sélection purificatrice réduite chez l'homme de Denisova (Do et al., 2015). Ces observations suggèrent donc que de nouvelles données sur l'étude de l'impact de la démographie sur le fardeau de mutations délétères pourraient être obtenues en considérant des variations démographiques extrêmes ayant perduré dans le temps, comme c'est le cas de certaines populations archaïques ou d'autres espèces.

7.2.2 L'impact du coefficient de dominance

Une partie de la confusion engendrée par les différents résultats réside également dans la formulation de différentes hypothèses concernant le coefficient de dominance des mutations délétères. Par définition, les mutations dont les effets sont additifs (h=0.5) présentent une certaine pénétrance et auront un impact direct sur la valeur sélective des individus alors que les mutations récessives (h=0) n'auront un impact qu'à l'état homozygote. L'hypothèse de dominance considérée peut donc aboutir à l'observation de différences de fardeaux de mutations délétères entre populations car le nombre de mutations à l'état homozygote augmente fortement au cours d'événements de réduction des tailles de populations effectives (i.e. goulot d'étranglement, effets fondateurs successifs). En faisant l'hypothèse d'une complète récessivité des mutations délétères, on observe un fardeau de mutations délétères plus élevé pour les populations européennes que pour les populations africaines et des observations similaires ont été faites pour des populations ayant subi des événements démographiques similaires tels que les Québécois (Lohmueller et al., 2008 ; Henn et al., 2016 ; Peischl et al., 2018). Nos travaux ont montré que même en ne considérant que les mutations délétères homozygotes comme estimateur du fardeau de mutations délétères, les populations de Pygmées et non-Pygmées présentent des fardeaux similaires malgré des histoires démographiques opposées. L'absence d'un excès de mutations homozygotes chez les Pygmées suite à la réduction de leur population, et donc l'action limitée de la dérive génétique, peut s'expliquer par leur grande taille effective de population passée, ainsi que par l'échange de migrants avec les non-Pygmées. Les effets de la migration n'ont été que peu étudiés dans le cadre de l'estimation du fardeau de mutations délétères chez l'homme, mais pourrait être un paramètre majeur à prendre en compte pour les mutations délétères les plus récessives.

En plus de calculer le fardeau de mutations délétères à partir des données empiriques, nos analyses ont permis de suivre l'évolution du fardeau génétique des individus Pygmées et non-Pygmées au cours de leur histoire démographique et en particulier en présence ou non de flux géniques. Ce travail de simulation montre une augmentation du fardeau de mutations délétères récessives dans les populations de Pygmées de l'ouest et de l'est lors de la réduction de leur taille de population. Cette augmentation transitoire du fardeau récessif, attribuable à l'augmentation du nombre de mutations homozygotes dans la population, diminue ensuite. Nos travaux ont montré que cette diminution est accélérée par les événements de migration avec les populations de non-Pygmées. Le fait que les traces de cette augmentation prédite du fardeau de mutations délétères récessives ne soit pas observée dans les données empiriques peut s'expliquer par la difficulté à identifier ces mutations chez des individus sains. En effet, il est probable que les mutations dont le coefficient de sélection correspond à des valeurs de $N_e s$ élevées ne soient jamais présentes dans le génome des individus non malades à l'état homozygote. Une autre explication pourrait également être que ces mutations existent dans les populations mais sont difficiles à isoler empiriquement en raison du manque de spécificité des algorithmes de prédiction fonctionnelle. De plus, une large fraction des mutations dont les effets sont potentiellement délétères se situent dans des régions pour lesquelles la prédiction fonctionnelle est difficile à réaliser, telles que les régions régulatrices.

Ces observations suggèrent donc que c'est en grande partie le coefficient de dominance qui contrôle l'impact de la démographie sur le fardeau de mutations délétères dans les populations, en particulier suite à une réduction de la taille de population (Simons et al., 2014 ; Balick et al., 2015). Nos travaux montrent qu'un modèle de dominance complètement récessif des mutations délétères n'est pas compatible avec les observations empiriques du fardeau de mutations délétères dans les populations humaines. En accord avec cette observation, les travaux de Wrigth et Haldane prédisent une corrélation négative entre le coefficient de dominance et l'effet délétère des mutations suggérant que les mutations faiblement délétères ont tendance à avoir des effets additifs alors que les mutations ayant des effets délétères important seront récessives. Cependant, il n'existe pas de méthode permettant d'estimer précisément la relation entre les coefficients de dominance et de sélection des mutations dans les populations humaines. Des travaux menés dans d'autres espèces disposant de modes de reproduction basés sur l'autofécondation comme *Arabidopsis* proposent de co-estimer le DFE et le coefficient de dominance (Huber et al., 2018). Ces études ont montré que les mutations le plus délétères ont plus de chances d'être récessives que les mutations ayant moins d'impact sur le phénotype et que ces effets sont modulés par la connectivité des gènes et leurs profils d'expression (Huber et al., 2018). L'exploration de la dominance des mutations chez l'homme reste cependant à déterminer.

7.3 Contribution des modèles alternatifs de sélection naturelle à l'histoire adaptative humaine

7.3.1 Méthodes de détection et modèle de balayage sélectif classique

Au cours des 100 000 dernières années, les populations humaines ont colonisé des régions du monde qui présentent une extraordinaire variété de conditions climatiques et de contraintes pathogéniques. Ces populations ont également exploité différentes ressources nutritionnelles et majoritairement adopté des régimes alimentaires basés sur l'agriculture et l'élevage, depuis son émergence il y a environ 10 000 ans. L'étude de l'adaptation génétique des populations humaines est particulièrement informative sur les mécanismes de sélection naturelle, les gènes et les fonctions biologiques ayant permis l'occupation de ces différents habitats. Le phénotype pygmée est un exemple de l'adaptation des populations humaines à un mode de vie de chasseur-cueilleur forestier dans différentes régions du monde et plusieurs hypothèses ont été avancées pour expliquer les causes de cette adaptation (Perry et al., 2014). En Afrique centrale, les populations de Pygmées ont fait l'objet de plusieurs scans visant à identifier les signatures moléculaires de sélection positive présentes dans leurs génomes (Lachance et al., 2012 ; Jarvis et al., 2012 ; Hsieh et al., 2016a ; Amorim et al., 2015 ; Migliano et al., 2013 ; Lopez Herraez et al., 2009 ; Mendizabal et al., 2012 ; Perry et al., 2014). Les méthodes de détection de balayages sélectifs employées identifient les profils de variations qui se caractérisent par une réduction de la diversité haplotypique, l'augmentation de la fraction d'allèles rares et l'augmentation des différences de fréquences alléliques.

Nos travaux sur la détection des signatures d'événements de sélection positive dans sept populations de Pygmées d'Afrique centrale a mis en évidence un très faible partage des signaux de sélection entre les populations, suggérant l'absence de balayages sélectifs classiques au cours de l'histoire commune à ces populations. Bien que de nombreuses méthodes aient été développées pour détecter les événements de sélection positive au niveau du génome entier, il existe un manque de concordance entre les méthodes (Akey, 2009) en partie due aux différentes échelles de temps que ces statistiques sont capables de détecter (Sabeti et al., 2006) ou au grand nombre de faux positifs et négatifs détectés dans les approches de valeurs extrêmes, dont les signaux de "sélection" peuvent être créés par d'autres effets génomiques (Hernandez et al., 2011). La prévalence du modèle de balayage sélectif dans les événements adaptatifs humains a été débattu, en particulier dans le cadre de la sélection d'arrière plan (i.e. background selection) qui élimine les mutations à proximité des mutations délétères (Hernandez et al., 2011) et peut mener à l'apparition dans le génome de régions dont les signatures moléculaires ressemblent à celles des régions sélectionnées positivement. De plus, une étude récente a montré que les effets de la sélection d'arrière plan sont amplifiés par les événements de réduction de taille des populations, en particulier dans les régions fortement contraintes (Torres et al., 2018) et met en évidence les interactions entre démographie et sélection naturelle qui peuvent rendre complexe la détection d'événements de sélection (Wilson et al., 2014).

7.3.2 Prévalence des modèles adaptatifs ?

Une évaluation complète de l'importance de la sélection positive dans les phénotypes adaptatifs chez l'homme nécessite de s'intéresser à d'autres modes de sélection positive dont le point de départ ne serait l'augmentation en fréquence d'une mutation *de novo* mais d'autres mécanismes comme la sélection sur variant pré-existant, la sélection polygénique et le métissage adaptatif (Pritchard et al., 2010 ; Schrider et al., 2016). Bien que ces modes de sélection soient difficiles à détecter à l'aide d'outils classiques car ils ne produisent pas les mêmes signatures moléculaires que les balayage sélectifs (Pritchard et al., 2010 ; Przeworski et al., 2005), ils pourraient représenter le mode d'adaptation le plus prévalent dans le génome humain (Pritchard et al., 2010 ; Schrider et al., 2016 ; Messer & Petrov, 2013). De nombreux traits comme la taille ou la résistance aux pathogènes sont déterminés non pas par un seul mais plusieurs gènes et la sélection polygénique permet une modulation fine de ces phénotypes dans le cadre d'une adaptation à un changement environnemental.

Nos travaux évoquent la possibilité d'une sélection polygénique chez les différents groupes de Pygmées ayant agi sur des voies métaboliques impliquées à la fois dans l'immunité et la régulation de certains mécanismes comme la sensibilité à l'insuline. Cependant, la détection simultanée de plusieurs loci adaptatifs ayant un faible impact sur le phénotype, comme c'est le cas des événements de sélection polygénique, reste difficile (Messer & Petrov, 2013). Plusieurs méthodes ont été proposées afin de détecter de tels événements sélectifs comme la recherche d'enrichissements en signaux de sélection positive dans des gènes appartenant à une même voie métabolique ou régulatrice (Gouy et al., 2017 ; Daub et al., 2013 ; Fraser, 2013) ou bien de SNPs associés au même phénotype (Turchin et al., 2012 ; Berg & Coop, 2014 ; Racimo et al., 2018). Les méthodes qui évaluent les signatures de sélection portées par des SNPs associés à un même phénotype et identifiés par des études d'association en génome entier (Genome Wide Association Studies, GWAS) permettent de tester les variations de fréquence alléliques entre les populations, cependant ces résultats doivent être interprétés avec précaution en raison des biais induits par la structure génétique des populations étudiées (Novembre & Barton, 2018 ; Sohail et al., 2018 ; Berg et al., 2018). Le développement de méthodes basées sur l'apprentissage automatique (Schrider et al., 2016) pourraient représenter une avancée méthodologique dans la détection des signaux alternatifs de sélection positive en intégrant des modèles démographiques qui ne sont pas à l'équilibre, en utilisant une approche non biaisée par l'identification au préalable de gènes dans une même voie métabolique ou de SNPs associés à un trait et en détectant les régions génomiques à la marge de balayages classiques qui peuvent être très similaires aux signatures de sélection générées par d'autres mécanismes de sélection.

7.3.3 Proposition de validation fonctionnelle des candidats sous sélection chez les Pygmées

L'une des limitations majeures de l'étude de l'adaptation génétique de l'homme par une approche génomique est lié au rôle des régions régulatrices et du phénotype en résultant. De nombreux efforts ont été réalisés pour décrire et cartographier les régions régulatrices. L'analyse de données d'expression de gènes et épigénétiques ont permis d'identifier un certain nombre de promoteurs et d'enhancers (Encode Project Consortium, 2012 ; GTEx Consortium, 2015 ; GTEx Consortium et al., 2017). De plus, des travaux ont estimé que les régions régulatrices pourraient être la cible de la sélection positive environ dix fois plus fréquemment que les mutations non-synonymes (Fraser, 2013 ; Grossman et al., 2013).

Les données d'expression de gènes au sein de groupes de cellules ou de tissus ont permis d'identifier des variants génétiques qui contribuent à la variation dans l'intensité de l'expression de certains gènes : les eQTLs (expression quantitative trait loci). L'identification de ces variants fonctionnels sont d'un importance majeure et permettent de faire le lien entre la variation génétique des individus et leur phénotype. De récentes études ont identifié des eQTLs importants dans la réponse immunitaire à partir des profil d'expression de gènes de monocytes infectés par différents ligands imitant l'infection par des pathogènes dans des populations africaines et européennes (Figure 7.1) (Quach et al., 2016). La présence de signaux de sélection naturelle portés par des eQTL a permis d'identifier les variants génétiques ayant conféré un avantage sélectif aux populations exposées à des environnements pathogéniques distincts (Quach et al., 2016 ; Nédélec et al., 2016).

Bien que ces études soient pertinentes pour identifier les variants génétiques ayant un effet direct sur le phénotype des populations, il n'existe pas d'étude comparative des profils d'expression cellulaires pour les populations de Pygmées et de non-Pygmées. L'étude de l'expression des gènes dans un contexte d'infection dans des lignées cellulaires de Pygmées et de non-Pygmées permettrait de tester empiriquement si les gènes qui présentent des signaux de sélection naturelle participent différemment à la régulation des phénotypes immunitaires chez les Pygmées. Cette étude permettrait également de suivre conjointement les mécanismes d'inflammation et de sensibilité à l'insuline apportant alors de nouveaux éléments de réponse sur un effet pléiotropique entre immunité et taille dans les populations de Pygmées d'Afrique centrale.

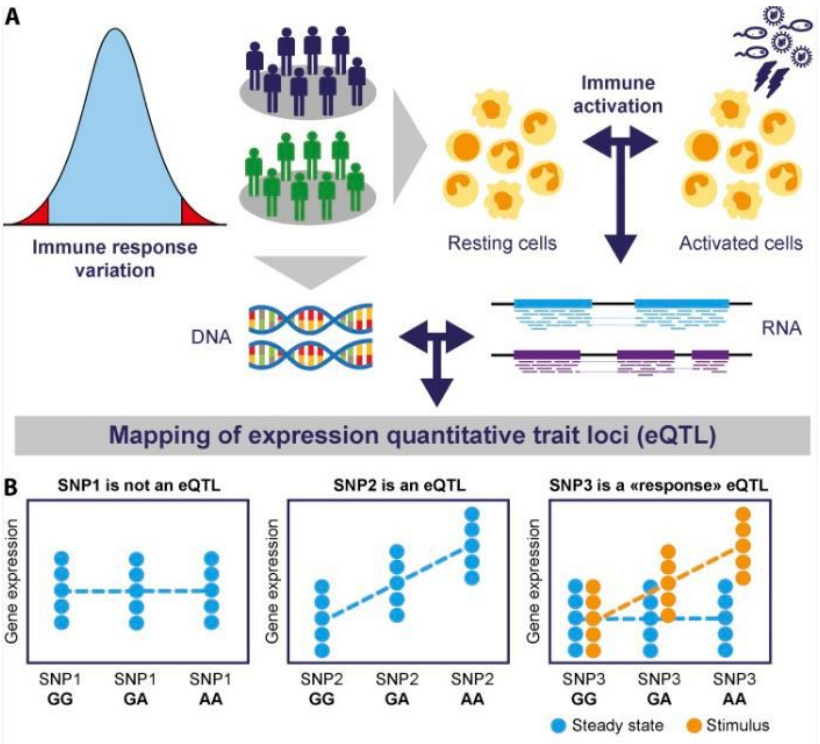


Fig. 7.1 ***Principe de la détection d'eQTL associés aux phénotypes immunitaires (Quach & Quintana-Murci, 2017).*** ***(A)*** *Protocole permettant de rechercher les bases génétiques de la variation des phénotypes immunitaires à partir des données génétiques et des profils d'expression de cellules saines ou infectées par des pathogènes.* ***(B)*** *Représentation schématique de l'action d'un eQTL contrôlant l'expression génétique. Contrairement au SNP2, le SNP1 ne contrôle pas l'intensité de l'expression, ce n'est pas un eQTL. Le génotype du SNP3 n'est corrélé à l'expression que lors de l'infection, c'est donc un eQTL spécifique de la réponse immunitaire.*

Références

Agarwala, V., Flannick, J., Sunyaev, S., Go, T. D. C., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet*, *45*(12), 1418-27.

Aghokeng, A. F., Ayouba, A., Mpoudi-Ngole, E., Loul, S., Liegeois, F., Delaporte, E., & Peeters, M. (2010). Extensive survey on the prevalence and genetic diversity of sivs in primate bushmeat provides insights into risks for potential new cross-species transmissions. *Infect Genet Evol*, *10*(3), 386-96.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans : where do we go from here ? *Genome Res*, *19*(5), 711-22.

Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J*, *1*(4857), 290-4.

Amorim, C. E., Daub, J. T., Salzano, F. M., Foll, M., & Excoffier, L. (2015). Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One*, *10*(4), e0121557.

Andersen, K. G., Shapiro, B. J., Matranga, C. B., Sealfon, R., Lin, A. E., Moses, L. M., Folarin, O. A., Goba, A., Odia, I., Ehiane, P. E., Momoh, M., England, E. M., Winnicki, S., Branco, L. M., Gire, S. K., Phelan, E., Tariyal, R., Tewhey, R., Omoniwa, O., Fullah, M., Fonnie, R., Fonnie, M., Kanneh, L., Jalloh, S., Gbakie, M., Saffa, S., Karbo, K., Gladden, A. D., Qu, J., Stremlau, M., Nekoui, M., Finucane, H. K., Tabrizi, S., Vitti, J. J., Birren, B., Fitzgerald, M., McCowan, C., Ireland, A., Berlin, A. M., Bochicchio, J., Tazon-Vega, B., Lennon, N. J., Ryan, E. M., Bjornson, Z., Milner, J., D. A., Lukens, A. K., Broodie, N., Rowland, M., Heinrich, M., Akdag, M., Schieffelin, J. S., Levy, D., Akpan, H., Bausch, D. G., Rubins, K., McCormick, J. B., Lander, E. S., Gunther, S., Hensley, L., Okogbenin, S., Viral Hemorrhagic Fever, C., Schaffner, S. F., Okokhere, P. O., Khan, S. H., Grant, D. S., Akpede, G. O., Asogun, D. A., Gnirke, A., Levin, J. Z., Happi, C. T., Garry, R. F., & Sabeti, P. C. (2015). Clinical sequencing uncovers origins and evolution of lassa virus. *Cell*, *162*(4), 738-50.

Arom, S. (1987). *Anthologie de la musique des pygmée aka (centrafrique).* Paris : OCORA.

Arom, S., & Fürniss, S. (1992). *The pentatonic system of the aka pygmies of the central african republic. in european studies in ethnomusicology.* Berlin : International Institute for Traditional Music.

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., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.

Bahuchet, D. M., S., & de Garine, I. (1991). Wild yams revisited : Is independence from agriculture possible for rain forest hunter-gatherers ? *Human Ecology and evolution*, *19(2) :213–43.*.

Bahuchet, S. (1987). *Le filet de chasse des pygmées aka (r.c.a.). in de la voûte céleste au terroir, du jardin au foyer : Mosaïque sociographique.* Paris : Ed. de l'ehess.

Bahuchet, S. (1992). *Spatial mobility and access to the resources among the african pygmies.", in mobility and territoriality : Social and spatial boundaries among foragers, fishers,*

pastoralists and peripatetics. New York/Oxford : Berg.

Bahuchet, S. (1993). L'invention des pygmées. *Cahiers d'études africaines*, *33*(153-181).

Bahuchet, S., & Guillaume, H. (1982). *Aka-farmer relations in the northwest congo basin. in politics and history in band societies.* Cambridge/Paris : Cambridge University Press.

Bailey, G. H. M. J. B. O. R. R., R. C., & Zechenter., E. (1989). The tropical rainforest : Is it a productive environment for human foragers ? *American Anthropologist*, *91 :59–82*.

Balick, D. J., Do, R., Cassa, C. A., Reich, D., & Sunyaev, S. R. (2015). Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS genetics*, *11*(8), e1005436.

Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nat Rev Genet*, *4*(2), 99-111.

Barham, L. S. (2001). *Central africa and the emergence of regional identity in the middle pleistocene.* Bristol : Western Academic and Specialist Press,.

Barker, G., Barton, H., Bird, M., Daly, P., Datan, I., Dykes, A., Farr, L., Gilbertson, D., Harrisson, B., Hunt, C., Higham, T., Kealhofer, L., Krigbaum, J., Lewis, H., McLaren, S., Paz, V., Pike, A., Piper, P., Pyatt, B., Rabett, R., Reynolds, T., Rose, J., Rushworth, G., Stephens, M., Stringer, C., Thompson, J., & Turney, C. (2007). The 'human revolution' in lowland tropical southeast asia : the antiquity and behavior of anatomically modern humans at niah cave (sarawak, borneo). *J Hum Evol*, *52*(3), 243-61.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nat Genet*, *40*(3), 340-5.

Barreiro, L. B., & Quintana-Murci, L. (2009). From evolutionary genetics to human immunology : how selection shapes host defence genes. *Nature Reviews Genetics*, *11*, 17.

Barreiro, L. B., & Quintana-Murci, L. (2010). From evolutionary genetics to human immunology : how selection shapes host defence genes. *Nat Rev Genet*, *11*(1), 17-30.

Barton, T. D. K. N., H., & Arroyo-Kalin., M. (2012). Long term perspectives on human occupation of tropical rainforests : An introductory overview. *Quaternary International*, *249 :1–3*.

Batini, C., Coia, V., Battaggia, C., Rocha, J., Pilkington, M. M., Spedini, G., Comas, D., Destro-Bisol, G., & Calafell, F. (2007). Phylogeography of the human mitochondrial l1c haplogroup : genetic signatures of the prehistory of central africa. *Mol Phylogenet Evol*, *43*(2), 635-44.

Batini, C., Ferri, G., Destro-Bisol, G., Brisighelli, F., Luiselli, D., Sanchez-Diz, P., Rocha, J., Simonson, T., Brehm, A., Montano, V., Elwali, N. E., Spedini, G., D'Amato, M. E., Myres, N., Ebbesen, P., Comas, D., & Capelli, C. (2011). Signatures of the preagricultural peopling processes in sub-saharan africa as revealed by the phylogeography of early y chromosome lineages. *Mol Biol Evol*, *28*(9), 2603-13.

Baumann, G., Shaw, M. A., & Merimee, T. J. (1989). Low levels of high-affinity growth hormone-binding protein in african pygmies. *The New England journal of medicine*, *320*(26), 1705-9.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P. A., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., Weale, M. E., Xu, Y., Xu, Z., Yang, L., Zaman, M. J., Zeng, C., Zhang, L., Zhang, X., Zhaxi, P., & Zheng, Y. T. (2010). Natural selection on epas1 (hif2alpha) associated with low hemoglobin concentration in tibetan highlanders. *Proc Natl Acad Sci U S A*, *107*(25), 11459-64.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, *162*(4), 2025-35.

Becker, N. S. A., Verdu, P., Froment, A., Le Bomin, S., Pagezy, H., Bahuchet, S., & Heyer, E. (2011, Jul). Indirect evidence for the genetic determination of short stature in african pygmies. *American journal of physical anthropology*, *145*(3), 390–401. doi: 10.1002/ajpa.21512

Becker, N. S. A., Verdu, P., Georges, M., Duquesnoy, P., Froment, A., Amselem, S., Le Bouc, Y., & Heyer, E. (2013, Jun). The role of ghr and igf1 genes in the genetic determination of african pygmies' short stature. *European journal of human genetics : EJHG*, *21*(6), 653–8. doi: 10.1038/ejhg.2012.223

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

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., & Coop, G. (2018). Reduced signal for polygenic adaptation of height in uk biobank. *bioRxiv*.

Bherer, C., Campbell, C. L., & Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun*, *8*, 14994.

Bozzola, M., Travaglino, P., Marziliano, N., Meazza, C., Pagani, S., Grasso, M., Tauber, M., Diegoli, M., Pilotto, A., Disabella, E., Tarantino, P., Brega, A., & Arbustini, E. (2009). The shortness of pygmies is associated with severe under-expression of the growth hormone receptor. *Molecular genetics and metabolism*, *98*(3), 310-3.

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A., Kwiatkowski, D. P., Spencer, C. C., & Malaria Genomic Epidemiology, N. (2016). Admixture into and within sub-saharan africa. *Elife*, *5*.

Calattini, S., Betsem, E. B., Froment, A., Mauclere, P., Tortevoye, P., Schmitt, C., Njouom, R., Saib, A., & Gessain, A. (2007). Simian foamy virus transmission from apes to humans, rural cameroon. *Emerg Infect Dis*, *13*(9), 1314-20.

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. *Nat Genet*, *44*(11), 1277-81.

Campbell, M. C., Hirbo, J. B., Townsend, J. P., & Tishkoff, S. A. (2014). The peopling of the african continent and the diaspora into the new world. *Curr Opin Genet Dev*, *29*, 120-32.

Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J. C., Gbeha, E., Hamdan, F. F., Girard, S., Spinella, J. F., Lariviere, M., Saillour, V., Healy, J., Fernandez, I., Sinnett, D., Michaud, J. L., Rouleau, G. A., Haddad, E., Le Deist, F., & Awadalla, P. (2013). Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS genetics*, *9*(9), e1003815.

Cavalli-Sforza, L. (1986). *African pygmies*. New York Academic Press.

Cavalli-Sforza, L. L., Zonta, L. A., Nuzzo, F., Bernini, L., de Jong, W. W., Meera Khan, P., Ray, A. K., Went, L. N., Siniscalco, M., Nijenhuis, L. E., van Loghem, E., & Modiano, G. (1969). Studies on african pygmies. i. a pilot investigation of babinga pygmies in the central african republic (with an analysis of genetic distances). *Am J Hum Genet*, *21*(3), 252-74.

Charlesworth, B. (2009). Fundamental concepts in genetics : effective population size and patterns of molecular evolution and variation. *Nature reviews. Genetics*, *10*(3), 195-

205.

Charlesworth, B. (2010). *Elements of evolutionary genetics*. Greenwood Village, Colo. : Roberts and Co.

Compton, J. S. (2011). Pleistocene sea-level fluctuations and human evolution on the southern coastal plain of south africa. , *30*(5-6), 506-527.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Wellcome Trust Case Control, C., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., & Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704-12.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., & Pritchard, J. K. (2009). The role of geography in human adaptation. *PLoS Genet*, *5*(6), e1000500.

Cooper, G. M., Nickerson, D. A., & Eichler, E. E. (2007). Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, *39*(7 Suppl), S22-9.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Iv, W. H., Templeton, A. R., Boerwinkle, E., Gibbs, R., & Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, *1*, 131.

Crevecoeur, I., Brooks, A., Ribot, I., Cornelissen, E., & Semal, P. (2016). Late stone age human remains from ishango (democratic republic of congo) : New insights on late pleistocene modern human diversity in africa. *J Hum Evol*, *96*, 35-57.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., & Excoffier, L. (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol*, *30*(7), 1544-58.

Delegue, M. A., Fuhr, M., Schwartz, D., Mariotti, A., & Nasi, R. (2001). Recent origin of a large part of the forest cover in the gabon coastal area based on stable carbon isotope data. *Oecologia*, *129*(1), 106-113.

Delobeau, J. M. (1989). *Yandenga et yamonzombo. les relations entre les villages monzombo et les campements pygmées aka dans la sous-préfecture de mongoumba (centrafrique)*. Paris : Peeters-SELAF.

deMenocal, P., Ortiz, J., Guilderson, T., & Sarnthein, M. (2000). Coherent high- and low-latitude climate variability during the holocene warm period. *Science*, *288*(5474), 2198-202.

Demolin, D. (1990). *Chants de l'orée de la forêt. polyphonies des pygmées efe* (Vol. 185). Fonti Musicali Fmd.

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. *American journal of human genetics*, *98*(1), 5-21.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G., & Capelli, C. (2004). Variation of female and male lineages in sub-saharan populations : the importance of sociocultural factors. *Mol Biol Evol*, *21*(9), 1673-82.

Diamond, J., & Bellwood, P. (2003). Farmers and their languages : the first expansions. *Science*, *300*(5619), 597-603.

Diamond, J. M. (1991). Anthropology. why are pygmies small ? *Nature*, *354*(6349), 111-2.

Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S., & Reich, D. (2015). No evidence that selection has been less effective at removing deleterious mutations in europeans than in africans. *Nature genetics*, *47*(2), 126-31.

Dounias, E. (2001). The management of wild yam tubers by the baka pygmies in southern cameroon. *African Study Monographs*, *supplement 26 :135–56*.

Du Chaillu, P. B., & Owen., R. (1867). *A journey to ashango-land : And further penetration into equatorial africa.* New York : D. Appleton And Company.

Dunn, F. L. (1977). *Health and disease in hunter-gatherers : epidemiological factors in : D. landy (ed.), culture, disease, and healing,.* New York : Macmillan.

Encode Project Consortium. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, *489*(7414), 57-74.

Eory, L., Halligan, D. L., & Keightley, P. D. (2010). Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol*, *27*(1), 177-92.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and snp data. *PLoS genetics*, *9*(10), e1003905.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in ecology & evolution*, *21*(10), 569-75.

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution*, *26*(9), 2097-108.

Fan, S., Hansen, M. E., Lo, Y., & Tishkoff, S. A. (2016). Going global by adapting local : A review of recent human adaptation. *Science*, *354*(6308), 54-59.

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive darwinian selection. *Genetics*, *155*(3), 1405-13.

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

Froment, A., & Koppert., G. (1999). *Malnutrition chronique et gradient climatique en milieu tropical. in : S. bahuchet, d. bley, h. pagezy and n. vernazza- licht (eds), l'homme et la forêt tropicale.* Société d'Ecologie Humaine-APFT, Marseille, Editions de Bergier,.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., & Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, *493*(7431), 216-20.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*, *7*(11), e1002355.

Fürniss, S., & Bahuchet, S. (1995). *Existe-t-il des instruments de musique pygmées ?" in ndroje balendro, musiques, terrains et disciplines.* Paris : Peters-Selaf.

Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *Am J Hum Genet*, *98*(3), 456-472.

Garine, I. (1990). Adaptation biologique et bien-être psycho-culturel. *Bulletins et Mémoires de la Société d'Anthropologie de Paris 2 : 151–74.*.

Gazave, E., Chang, D., Clark, A. G., & Keinan, A. (2013). Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics*, *195*(3), 969-78.

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome

variation from population-scale sequencing. *Nature*, *467*(7319), 1061-73.

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56-65.

Gomez, A., Petrzelkova, K. J., Burns, M. B., Yeoman, C. J., Amato, K. R., Vlckova, K., Modry, D., Todd, A., Jost Robinson, C. A., Remis, M. J., Torralba, M. G., Morton, E., Umana, J. D., Carbonero, F., Gaskins, H. R., Nelson, K. E., Wilson, B. A., Stumpf, R. M., White, B. A., Leigh, S. R., & Blekhman, R. (2016). Gut microbiome of coexisting baaka pygmies and bantu reflects gradients of traditional subsistence patterns. *Cell Rep*, *14*(9), 2142-2153.

Gouy, A., Daub, J. T., & Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. *Nucleic acids research*, *45*(16), e149.

Granka, J. M., Henn, B. M., Gignoux, C. R., Kidd, J. M., Bustamante, C. D., & Feldman, M. W. (2012). Limited evidence for classic selective sweeps in african populations. *Genetics*, *192*(3), 1049-64.

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.

Grinker, R. R. (1994). *Houses in the rainforest : Ethnicity and inequality among farmers and foragers in central africa* (Vol. 226). Berkeley : University of California Press.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet*, *43*(10), 1031-4.

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., Cabili, M., Adegbola, R. A., Bamezai, R. N., Hill, A. V., Vannberg, F. O., Rinn, J. L., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, *152*(4), 703-13.

Grossman, S. R., Shlyakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, *327*(5967), 883-6.

GTEx Consortium. (2015). Human genomics. the genotype-tissue expression (gtex) pilot analysis : multitissue gene regulation in humans. *Science*, *348*(6235), 648-60.

GTEx Consortium, D. A., and Laboratory, 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.

Guernier, V., Hochberg, M. E., & Guegan, J. F. (2004). Ecology drives the worldwide distribution of human diseases. *PLoS Biol*, *2*(6), e141.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S. A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., & Sandhu, M. S. (2015). The african genome variation project shapes medical genetics in africa. *Nature*, *517*(7534), 327-32.

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the signature of natural selection in humans : evidence from the duffy blood group locus. *Am J Hum Genet*, *66*(5), 1669-79.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., & Wall, J. D. (2011). Genetic evidence for archaic admixture in africa. *Proc Natl Acad Sci U S A*, *108*(37), 15123-8.

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2011). Adaptations to climate-mediated selective pressures in humans. *PLoS Genet*, *7*(4), e1001375.

Harako, R. (1976). *The mbuti as hunters : A study of ecological anthropology of the mbuti pygmies (i)*. Kyoto University African Studies.

Hattori, Y., Vera, J. C., Rivas, C. I., Bersch, N., Bailey, R. C., Geffner, M. E., & Golde, D. W. (1996). Decreased insulin-like growth factor i receptor expression and function in immortalized african pygmy t cells. *The Journal of clinical endocrinology and metabolism*, *81*(6), 2257-63.

Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747-51.

Henn, B. M., Botigue, L. R., Bustamante, C. D., Clark, A. G., & Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature reviews. Genetics*, *16*(6), 333-43.

Henn, B. M., Botigue, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., & Comas, D. (2012). Genomic ancestry of north africans supports back-to-africa migrations. *PLoS genetics*, *8*(1), e1002397.

Henn, B. M., Botigue, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., & Bustamante, C. D. (2016). Distance from sub-saharan africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(4), E440-9.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Sella, G., & Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, *331*(6019), 920-4.

Hewlett, B. (1996). *Cultural diversity among african pygmies, in cultural diversity among twentieth-century foragers : An african perspective* (Vol. 1). Cambridge : Cambridge University Press.

Hewlett, B. (2014). *Hunter-gatherers of the congo basin : Cultures, histories and biology of african pygmies*. Transaction Publishers.

Hladik, A., & Dounias, E. (1993). Wild yams of the african forest as potential food resources. *In Tropical Forests, People and Food. A. H. C. M. Hladik, O. F. Linares, H. Pagezy,*

A. Semple, and M. Hadley, Ed. Pp. 163–76. Man and the Biosphere, Vol. 13. Paris : UNESCO..

Hsieh, P., Veeramah, K. R., Lachance, J., Tishkoff, S. A., Wall, J. D., Hammer, M. F., & Gutenkunst, R. N. (2016a). 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 research*, *26*(3), 279-90.

Hsieh, P., Woerner, A. E., Wall, J. D., Lachance, J., Tishkoff, S. A., Gutenkunst, R. N., & Hammer, M. F. (2016b). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in central african pygmies. *Genome research*, *26*(3), 291-300.

Huber, C. D., Durvasula, A., Hancock, A. M., & Lohmueller, K. E. (2018). Gene expression drives the evolution of dominance. *Nat Commun*, *9*(1), 2750.

Hudson, R. R., Kreitman, M., & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153-9.

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.

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends Genet*, *24*(5), 238-45.

Hussin, J. G., Hodgkinson, A., Idaghdour, Y., Grenier, J. C., Goulet, J. P., Gbeha, E., Hip-Ki, E., & Awadalla, P. (2015). Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature genetics*, *47*(4), 400-4.

Hutchison, D. W., & Templeton, A. R. (1999). Correlation of pairwise genetic and geographic distance measures : Inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution*, *53*(6), 1898-1914.

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., Tsui, S. K., Xue, H., Wong, J. T., 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., Sallee, 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., Song, Y. Q., Tam, P. K., 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*(7164), 851-61.

Izagirre, N., Garcia, I., Junquera, C., de la Rua, C., & Alonso, S. (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol Biol Evol*, *23*(9), 1697-706.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., & Singleton,

A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, *451*(7181), 998-1003.

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J. M., Beggs, W., Hoffman, G., Mezey, J., & Tishkoff, S. A. (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in western african pygmies. *PLoS genetics*, *8*(4), e1002641.

Joiris, D. V. (2003). *The framework of central african hunter-gatherers and neighbouring societies.* African Study Monographs,.

Joiris, D. V., & Bahuchet, S. (1994). *"afrique Équatoriale." in situation des populations indigènes des forêts denses et humides.* Bruxelles : Comission Européenne.

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nat Rev Genet*, *15*(6), 379-93.

Keightley, P. D., & Lynch, M. (2003). Toward a realistic model of mutations affecting fitness. *Evolution*, *57*(3), 683-5 ; discussion 686-9.

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, *336*(6082), 740-3.

Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W., & Akey, J. M. (2006, Aug). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome research*, *16*(8), 980–9.

Kelso, J., & Prufer, K. (2014). Ancient humans and the origin of modern humans. *Curr Opin Genet Dev*, *29*, 133-8.

Kiezun, A., Pulit, S. L., Francioli, L. C., van Dijk, F., Swertz, M., Boomsma, D. I., van Duijn, C. M., Slagboom, P. E., van Ommen, G. J., Wijmenga, C., Genome of the Netherlands, C., de Bakker, P. I., & Sunyaev, S. R. (2013). Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS Genet*, *9*(2), e1003301.

Kitanishi, K. (1995). easonal changes in the subsistence activities and food intake of the hunter-gatherers in northeastern congo. *African Study Monographs*, *16(2) :73–118*.

Klieman, K. A. (2003). *"the pygmies were our compass :" bantu and batwa in the history of west central africa, early times to c. 1900 c.e.* ortsmouth, New Hampshire : Heinemann.

Ko, W. Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C. A., Froment, A., Nyambo, T. B., Omar, S. A., Wambebe, C., Ranciaro, A., Hirbo, J. B., & Tishkoff, S. A. (2013). Identifying darwinian selection acting on different human apol1 variants among diverse african populations. *Am J Hum Genet*, *93*(1), 54-66.

Koppert, E. D. A. F., G., & Pasquet., P. (1993). *Food consumption in three forest populations of the southern coastal cameroon. in : Hladik c. m., hladik a., linares o., pagezy h., semple a., and hadley m. editors.* Tropical Forests : People and Food, Man and the Biosphere.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J. M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M., & Tishkoff, S. A. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell*, *150*(3), 457-69.

Lapierre, M., Blin, C., Lambert, A., Achaz, G., & Rocha, E. P. (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol Biol Evol*, *33*(7), 1711-25.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). Dbvar and dgva : public archives for genomic structural variation. *Nucleic Acids Res*, *41*(Database issue), D936-41.

Lohmueller, K. E. (2014a). The distribution of deleterious genetic variation in human

populations. *Current opinion in genetics & development*, *29*, 139-46.

Lohmueller, K. E. (2014b). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS genetics*, *10*(5), e1004379.

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A. F., Grarup, N., Jorgensen, T., Jiang, T., Witte, D. R., Sandbaek, 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 Genet*, *7*(10), e1002326.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in european than in african populations. *Nature*, *451*(7181), 994-7.

Lopez Herraez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M. R., Gross, A., Scholz, M., & Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations : evidence from nearly 1 million snps. *PloS one*, *4*(11), e7888.

Louicharoen, C., Patin, E., Paul, R., Nuchprayoon, I., Witoonpanich, B., Peerapittayamongkol, C., Casademont, I., Sura, T., Laird, N. M., Singhasivanon, P., Quintana-Murci, L., & Sakuntabhai, A. (2009). Positively selected g6pd-mahidol mutation reduces plasmodium vivax density in southeast asians. *Science*, *326*(5959), 1546-9.

Luca, F., Perry, G. H., & Di Rienzo, A. (2010). Evolutionary adaptations to dietary changes. *Annu Rev Nutr*, *30*, 291-314.

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat Genet*, *39*(7 Suppl), S43-7.

Maher, M. C., Uricchio, L. H., Torgerson, D. G., & Hernandez, R. D. (2012). Population genetics of rare variants and complex diseases. *Hum Hered*, *74*(3-4), 118-28.

Malaspinas, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., Dupanloup, I., Eriksson, A., Margaryan, A., Moltke, I., Pugach, I., Korneliussen, T. S., Levkivskyi, I. P., Moreno-Mayar, J. V., Ni, S., Racimo, F., Sikora, M., Xue, Y., Aghakhanian, F. A., Brucato, N., Brunak, S., Campos, P. F., Clark, W., Ellingvag, S., Fourmile, G., Gerbault, P., Injie, D., Koki, G., Leavesley, M., Logan, B., Lynch, A., Matisoo-Smith, E. A., McAllister, P. J., Mentzer, A. J., Metspalu, M., Migliano, A. B., Murgha, L., Phipps, M. E., Pomat, W., Reynolds, D., Ricaut, F. X., Siba, P., Thomas, M. G., Wales, T., Wall, C. M., Oppenheimer, S. J., Tyler-Smith, C., Durbin, R., Dortch, J., Manica, A., Schierup, M. H., Foley, R. A., Lahr, M. M., Bowern, C., Wall, J. D., Mailund, T., Stoneking, M., Nielsen, R., Sandhu, M. S., Excoffier, L., Lambert, D. M., & Willerslev, E. (2016). A genomic history of aboriginal australia. *Nature*, *538*(7624), 207-214.

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., 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., Paabo, S., Kelso, J., Patterson, N., & Reich, D. (2016). The simons genome diversity project : 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201-206.

Mann, A. R. D. L. P., G. V., & Merrill., J. M. (1962). Cardiovascular disease in african pygmies : a survey of the health status, serum lipids, and diet of pygmies in congo. *Journal of Chronic Diseases 15 :341–71*.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the adh locus in drosophila. *Nature*, *351*(6328), 652-4.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*(5670), 581-4.

Mendizabal, I., Marigorta, U. M., Lao, O., & Comas, D. (2012). Adaptive evolution of loci covarying with the human african pygmy phenotype. *Human genetics*, *131*(8), 1305-17.

Mercader, J. (2002). orest people : The role of african rainforests in human evolution and dispersal. *Evolutionary Anthropology*, *11(117–24)*.

Merimee, T. J., Hewlett, B. S., Wood, W., Bowcock, A. M., & Cavalli-Sforza, L. L. (1989). The growth hormone receptor gene in the african pygmy. *Transactions of the Association of American Physicians*, *102*, 163-9.

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol*, *28*(11), 659-69.

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, 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., Andres, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., & Paabo, S. (2012). A high-coverage genome sequence from an archaic denisovan individual. *Science*, *338*(6104), 222-6.

Migliano, A. B., Romero, I. G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., Huang, D. W., Sherman, B. T., Siddle, K., Scholes, C., Hudjashov, G., Kaitokai, E., Babalu, A., Belatti, M., Cagan, A., Hopkinshaw, B., Shaw, C., Nelis, M., Metspalu, E., Magi, R., Lempicki, R. A., Villems, R., Lahr, M. M., & Kivisild, T. (2013). Evolution of the pygmy phenotype : evidence of positive selection fro genome-wide scans in african, asian, and melanesian pygmies. *Human biology*, *85*(1-3), 251-84.

Migliano, A. B., Vinicius, L., & Lahr, M. M. (2007). Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20216-9.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321-4.

Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., Barnett, A. H., Bates, C., Bellary, S., Bockett, N. A., Giorda, K., Griffiths, C. J., Hemingway, H., Jia, Z., Kelly, M. A., Khawaja, H. A., Lek, M., McCarthy, S., McEachan, R., O'Donnell-Luria, A., Paigen, K., Parisinos, C. A., Sheridan, E., Southgate, L., Tee, L., Thomas, M., Xue, Y., Schnall-Levin, M., Petkov, P. M., Tyler-Smith, C., Maher, E. R., Trembath, R. C., MacArthur, D. G., Wright, J., Durbin, R., & van Heel, D. A. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, *352*(6284), 474-7.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, *154*(2), 931-42.

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

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302-310.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One*, *7*(7), e37558.

Novembre, J., & Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics*, *208*(4), 1351-1355.

Novembre, J., & Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet*, *10*(11), 745-55.

Novembre, J., & Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. *Annu Rev Genomics Hum Genet*, *12*, 245-74.

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., et al. (2016). Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, *167*(3), 657–669.

Ohenjo, R. W. D. J. C. N. K. G., N., & Mugarura, B. (2006). Health of indigenous people in africa. *The Lancet*, *367 :1937–46*.

Pagani, L., Lawson, D. J., Jagoda, E., Morseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., Wall, J. D., Cardona, A., Magi, R., Wilson Sayres, M. A., Kaewert, S., Inchley, C., Scheib, C. L., Jarve, M., Karmin, M., Jacobs, G. S., Antao, T., Iliescu, F. M., Kushniarevich, A., Ayub, Q., Tyler-Smith, C., Xue, Y., Yunusbayev, B., Tambets, K., Mallick, C. B., Saag, L., Pocheshkhova, E., Andriadze, G., Muller, C., Westaway, M. C., Lambert, D. M., Zoraqi, G., Turdikulova, S., Dalimova, D., Sabitov, Z., Sultana, G. N. N., Lachance, J., Tishkoff, S., Momynaliev, K., Isakova, J., Damba, L. D., Gubina, M., Nymadawa, P., Evseeva, I., Atramentova, L., Utevska, O., Ricaut, F. X., Brucato, N., Sudoyo, H., Letellier, T., Cox, M. P., Barashkov, N. A., Skaro, V., Mulahasanovic, L., Primorac, D., Sahakyan, H., Mormina, M., Eichstaedt, C. A., Lichman, D. V., Abdullah, S., Chaubey, G., Wee, J. T. S., Mihailov, E., Karunas, A., Litvinov, S., Khusainova, R., Ekomasova, N., Akhmetova, V., Khidiyatova, I., Marjanovic, D., Yepiskoposyan, L., Behar, D. M., Balanovska, E., Metspalu, A., Derenko, M., Malyarchuk, B., Voevoda, M., Fedorova, S. A., Osipova, L. P., Lahr, M. M., Gerbault, P., Leavesley, M., Migliano, A. B., Petraglia, M., Balanovsky, O., Khusnutdinova, E. K., Metspalu, E., Thomas, M. G., Manica, A., Nielsen, R., Villems, R., Willerslev, E., Kivisild, T., & Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of eurasia. *Nature*, *538*(7624), 238-242.

Pagezy, H., & Hauspie, R. (1989). Longitudinal study of growth in weight of african babies : an analysis of seasonal variations in the average growth rate and the effects of infectious diseases on individual and average growth patterns. *Acta Paediatrica Scandinavia suppl. 350 : 37–43.*.

Paolucci, M. A. S., A. M., & Pennetti., V. (1973). Modification of serum-free amino acid patterns of babinga adult pygmies after short-term feeding on a balanced diet. *American Journal of Clinical Nutrition 26 :429–34*.

Patin, E., Harmant, C., Kidd, K. K., Kidd, J., Froment, A., Mehdi, S. Q., Sica, L., Heyer, E., & Quintana-Murci, L. (2006). Sub-saharan african coding sequence variation and haplotype diversity at the nat2 gene. *Hum Mutat*, *27*(7), 720.

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., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., & Quintana-Murci, L. (2009). Inferring the demographic history of african farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics*, *5*(4), e1000448.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J. M., Pereira, J. B., Fernandes, V., Pereira, L., Van der Veen, L., Mouguiama-Daouda, P., Bustamante, C. D., Hombert, J. M., & Quintana-Murci, L. (2017). Dispersals and genetic adaptation of bantu-speaking populations in africa and north america. *Science*, *356*(6337), 543-546.

Patin, E., Siddle, K. J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Regnault, B., Lemee, L., Gravel, S., Hombert, J. M., Van der Veen, L., Dominy, N. J., Perry, G. H., Barreiro, L. B., Verdu, P., Heyer, E., & Quintana-Murci, L. (2014). The impact of agricultural emergence on the genetic history of african rainforest hunter-gatherers and agriculturalists. *Nature communications*, *5*, 3163.

Pedersen, C. T., Lohmueller, K. E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H. R., Moltke, I., & Albrechtsen, A. (2017). The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation : Insights from the greenlandic inuit. *Genetics*, *205*(2), 787-801.

Peischl, S., Dupanloup, I., Foucal, A., Jomphe, M., Bruat, V., Grenier, J.-C., Gouy, A., Gbeha, E., Bosshard, L., Hip-Ki, E., Agbessi, M., Hodgkinson, A., Vézina, H., Awadalla, P., & Excoffier, L. (2016). Relaxed selection during a recent human expansion. *bioRxiv*.

Peischl, S., Dupanloup, I., Foucal, A., Jomphe, M., Bruat, V., Grenier, J. C., Gouy, A., Gilbert, K. J., Gbeha, E., Bosshard, L., Hip-Ki, E., Agbessi, M., Hodgkinson, A., Vezina, H., Awadalla, P., & Excoffier, L. (2018). Relaxed selection during a recent human expansion. *Genetics*, *208*(2), 763-777.

Perry, G. H., & Dominy, N. J. (2009). Evolution of the human pygmy phenotype. *Trends in ecology & evolution*, *24*(4), 218-25.

Perry, G. H., Foll, M., Grenier, J. C., Patin, E., Nedelec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S. L., Excoffier, L., Quintana-Murci, L., Dominy, N. J., & Barreiro, L. B. (2014). Adaptive, convergent origins of the pygmy phenotype in african rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(35), E3596-603.

Philipson, D. (2005). *African archaeology*. Cambridge : Cambridge University Press.

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., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, *19*(5), 826-37.

Pickrell, J. K., Patterson, N., Loh, P. R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., & Reich, D. (2014). Ancient west eurasian ancestry in southern and eastern africa. *Proc Natl Acad Sci U S A*, *111*(7), 2632-7.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Dna recombination. recombination initiation maps of individual human genomes. *Science*, *346*(6211), 1256442.

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.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human y chromosomes : a study of y chromosome microsatellites. *Mol Biol Evol*, *16*(12), 1791-8.

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.

Przeworski, M., Coop, G., & Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, *59*(11), 2312-23.

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, *155*(3), 1429-37.

Quach, H., & Quintana-Murci, L. (2017). Living in an adaptive world : Genomic dissection of the genus homo and its immune response. *J Exp Med*, *214*(4), 877-894.

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.

Quintana-Murci, L., & Clark, A. G. (2013). Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol*, *13*(4), 280-93.

Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S., Balanovsky, O., Kidd, K. K., Kidd, J. R., van der Veen, L., Hombert, J. M., Gessain, A., Verdu, P., Froment, A., Bahuchet, S., Heyer, E., Dausset, J., Salas, A., & Behar, D. M. (2008). Maternal traces of deep common ancestry and asymmetric gene flow between pygmy hunter-gatherers and bantu-speaking farmers. *Proc Natl Acad Sci U S A*, *105*(5), 1596-601.

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.

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sanchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat Rev Genet*, *16*(6), 359-71.

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.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., Ko, A. M., Ko, Y. C., Jinam, T. A., Phipps, M. E., Saitou, N., Wollstein, A., Kayser, M., Paabo, S., & Stoneking, M. (2011). Denisova admixture and the first modern human dispersals into southeast asia and oceania. *Am J Hum Genet*, *89*(4), 516-28.

Robillard, M. (2012). De la nécessité d'étudier les relations interethniques pour appréhender la dynamique du changement : Le cas des baka et des fang-mvè de minvoul (gabon). *Journal des Africanistes*, 82(1–2) :137–165.

Rozzi, F. V., Koudou, Y., Froment, A., Le Bouc, Y., & Botton, J. (2015). Growth pattern from birth to adulthood in african pygmies of known age. *Nature communications*, *6*,

7672.

Ruwende, C., Khoo, S. C., Snow, R. W., Yates, S. N., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C. E., Gilbert, S. C., Peschu, N., & et al. (1995). Natural selection of hemi- and heterozygotes for g6pd deficiency in africa by resistance to severe malaria. *Nature*, *376*(6537), 246-9.

Sabbagh, A., Darlu, P., Crouau-Roy, B., & Poloni, E. S. (2011). Arylamine n-acetyltransferase 2 (nat2) genetic diversity and traditional subsistence : a worldwide population survey. *PloS one*, *6*(4), e18507. doi: 10.1371/journal.pone.0018507

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832-7.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, *312*(5780), 1614-20.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., 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., 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., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., 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., Sallee, 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., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913-8.

Salas, A., Richards, M., De la Fe, T., Lareu, M. V., Sobrino, B., Sanchez-Diz, P., Macaulay, V., & Carracedo, A. (2002). The making of the african mtdna landscape. *Am J Hum Genet*, *71*(5), 1082-111.

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.

Santachiara-Benerecetti, A. S., Beretta, M., Negri, M., Ranzani, G., Antonini, G., Barberio, C., Modiano, G., & Cavalli-Sforza, L. L. (1980). Population genetics of red cell enzymes in pygmies : a conclusive account. *Am J Hum Genet*, *32*(6), 934-54.

Sanz, J., Randolph, H. E., & Barreiro, L. B. (2018). Genetic and evolutionary determinants of human population variation in immune responses. *Curr Opin Genet Dev*, *53*, 28-35.

Schebesta, P. (1938). *Die bambuti pygmäen vom ituri* (Vol. 1). Falk, Bruxelles : Mémoire de l'Institut Royal Colonial Belge.

Schebesta, P. (1941). *Die bambuti-pygmäen von ituri, bd ii. ethnographie der ituri-bambuti, 1 teil : Die wirtschaft der ituri-bambuti (belgisch kongo)* (Vol. 284). Bruxelles.

Schebesta, P. (1948). *Die bambuti-pygmäen von ituri, bd ii– ethnographie der ituri- bambuti, teil ii : Das soziale leben* (Vol. 551). Bruxelles.

Schebesta, P. (1952). *Les pygmées du congo belge* (Vol. 432). Bruxelles.

Scheinfeldt, L. B., & Tishkoff, S. A. (2013). Recent human adaptation : genomic approaches, interpretation and insights. *Nat Rev Genet*, *14*(10), 692-702.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., & Jakobsson, M. (2017). Southern african ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*. doi: 10.1126/science.aao6266

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H., & Jakobsson, M. (2012). Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science*, *338*(6105), 374-9.

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*(3), 1207-1223.

Schwartz, H. D. F. R. D., D., & Lanfranchi, R. (1990). Découverte d'unpremier site de l'age de fer ancien (2,100 b.p.) dans le mayumbe congolais. implications paléobotaniques et pédologiques. *Comptes Rendus De L'academie Des Sciences. Série II, Sciences De La Terre Et Des Plants*, *310(2) :1293–8*.

Seielstad, M. T., Minch, E., & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nat Genet*, *20*(3), 278-80.

Shea, B. T., & Bailey, R. C. (1996). Allometry and adaptation of body proportions and stature in african pygmies. *American journal of physical anthropology*, *100*(3), 311-40.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbsnp : the ncbi database of genetic variation. *Nucleic Acids Res*, *29*(1), 308-11.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M., & Jones, K. W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal snps. *Hum Genomics*, *1*(4), 274-86.

Siddle, K. J., & Quintana-Murci, L. (2014). The red queen's long race : human adaptation to pathogen pressure. *Curr Opin Genet Dev*, *29*, 31-8.

Simons, Y. B., & Sella, G. (2016). The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current opinion in genetics & development*, *41*, 150-158.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., & Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature genetics*, *46*(3), 220-4.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., & Ge, R. (2010). Genetic evidence for high-altitude adaptation in tibet. *Science*, *329*(5987), 72-5.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Roman, J. I., Chiumia, C., Crowther, A., Gomani-Chindebvu, E., Gidna, A. O., Grillo, K. M., Helenius, I. T., Hellenthal, G., Helm, R., Horton, M., Lopez, S., Mabulla, A. Z. P., Parkington, J., Shipton, C., Thomas, M. G., Tibesasa, R., Welling, M., Hayes, V. M., Kennett, D. J., Ramesar, R., Meyer, M., Paabo, S., Patterson, N., Morris, A. G., Boivin, N., Pinhasi, R., Krause, J., & Reich, D. (2017). Reconstructing prehistoric african population structure. *Cell*, *171*(1), 59-71 e21.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W. K., Hirschhorn, J. N., Daly, M., Patterson, N., Neale, B., Mathieson, I., Reich, D., & Sunyaev, S. R. (2018). Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*.

Sowunmi, M. A. (1999). The significance of the oil palm (elaeis guineenis jacq.) in the late holocene environments of west and west central africa : A further consideration. *Vegetation History and Archaeobotany*, *8 :199–210*.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, *61*, 437-55.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., Palsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Vermeulen, S. H., Goldstein, A. M., Tucker, M. A., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., & Stefansson, K. (2008). Two newly identified genetic determinants of pigmentation in europeans. *Nat Genet*, *40*(7), 835-7.

Tanno, K. (1976). *The mbuti net-hunters in the ituri forest, eastern zaire : Their hunting activities and band composition*. Kyoto University African Studies.

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*, *3*, 92.

Tavare, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, *145*(2), 505-18.

Taylor, N. (2011). The origins of hunting & gathering in the congo basin : A perspective on the middle stone age lupemban industry. *Before Farming*.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., & Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*(6090), 64-9.

Terashima, H. (1980). *Mota and other hunting activities of the mbuti archers*. African Study Monographs.

Teshima, K. M., Coop, G., & Przeworski, M. (2006, Jun). How reliable are empirical genomic scans for selective sweeps ? *Genome research*, *16*(6), 702–12.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., & Williams, S. M. (2009). The genetic structure and history of africans and african americans. *Science*, *324*(5930), 1035-1044.

Torres, R., Szpiech, Z. A., & Hernandez, R. D. (2018). Human demographic history has amplified the effects of background selection across the genome. *PLoS Genet*, *14*(6), e1007387.

Turchin, M. C., Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D., Genetic Investigation of, A. T. C., & Hirschhorn, J. N. (2012). Evidence of widespread selection on standing variation in europe at height-associated snps. *Nat Genet*, *44*(9), 1015-9.

Turnbull, C. M. (1965). *Wayward servants, the two worlds of the african pygmies*. New York : The Natural History Press.

Vansina, J. (1984). Western bantu expansion. *Journal of African History*, *25 :129–145*.

Vansina, J. (1990). *Paths in the rainforest*. Milwaukee : University of Wisconsin Press.

Vattathil, S., & Akey, J. M. (2015). Small amounts of archaic admixture provide big insights into human history. *Cell*, *163*(2), 281-4.

Veeramah, K. R., & Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nat Rev Genet*, *15*(3), 149-62.

Veeramah, K. R., Wegmann, D., Woerner, A., Mendez, F. L., Watkins, J. C., Destro-Bisol,

G., Soodyall, H., Louie, L., & 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. *Molecular biology and evolution*, *29*(2), 617-30.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Thery, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J. M., Van der Veen, L., Quintana-Murci, L., Bahuchet, S., & Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from western central africa. *Current biology : CB*, *19*(4), 312-8.

Verdu, P., Becker, N. S., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., Hombert, J. M., Van der Veen, L., Le Bomin, S., Bahuchet, S., Heyer, E., & Austerlitz, F. (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in central african pygmies and non-pygmies. *Molecular biology and evolution*, *30*(4), 918-37.

Vernot, B., & Akey, J. M. (2014). Human evolution : genomic gifts from archaic hominins. *Curr Biol*, *24*(18), R845-R848.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu Rev Genet*, *47*, 97-120.

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.

Wall, J. D. (1999). Recombination and the power of statistical tests of neutrality. *Genetical Research,*, *74(1)*, 65-79.

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterlander, M., Hollfelder, N., Potekhina, I. D., Schier, W., Thomas, M. G., & Burger, J. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in europeans during the last 5,000 y. *Proc Natl Acad Sci U S A*, *111*(13), 4832-7.

Wilson, B. A., Petrov, D. A., & Messer, P. W. (2014). Soft selective sweeps in complex demographic scenarios. *Genetics*, *198*(2), 669-84.

Wlasiuk, G., Khan, S., Switzer, W. M., & Nachman, M. W. (2009). A history of recurrent positive selection at the toll-like receptor 5 in primates. *Mol Biol Evol*, *26*(4), 937-49.

Wollstein, A., & Stephan, W. (2015). Inferring positive selection in humans from genomic data. *Investig Genet*, *6*, 5.

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

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114-38.

Wright, S. (1965). The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution*, *19*(3), 395-420.

Xu, D., Pavlidis, P., Taskent, R. O., Alachiotis, N., Flanagan, C., DeGiorgio, M., Blekhman, R., Ruhl, S., & Gokcumen, O. (2017). Archaic hominin introgression in africa contributes to functional salivary muc7 genetic variation. *Mol Biol Evol*, *34*(10), 2704-2715.

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, *15*(12), 496-503.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Yang, H., Nielsen, R., & Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, *329*(5987), 75-8.

Zerner, C. (2003). *The viral forest in motion. in : Candace wade, ed., in search of the rainforest.* Durham : Duke University Press.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, *22*(12), 2472-9.

Annexe 1

HUMAN GENETICS

Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America

Etienne Patin,[1,2,3]* Marie Lopez,[1,2,3] Rebecca Grollemund,[4,5] Paul Verdu,[6] Christine Harmant,[1,2,3] Hélène Quach,[1,2,3] Guillaume Laval,[1,2,3] George H. Perry,[7] Luis B. Barreiro,[8] Alain Froment,[9] Evelyne Heyer,[6] Achille Massougbodji,[10,11] Cesar Fortes-Lima,[6,12] Florence Migot-Nabias,[13,14] Gil Bellis,[15] Jean-Michel Dugoujon,[12] Joana B. Pereira,[16,17] Verónica Fernandes,[16,17] Luisa Pereira,[16,17,18] Lolke Van der Veen,[19] Patrick Mouguiama-Daouda,[19,20] Carlos D. Bustamante,[21,22] Jean-Marie Hombert,[19] Lluís Quintana-Murci[1,2,3]*

Bantu languages are spoken by about 310 million Africans, yet the genetic history of Bantu-speaking populations remains largely unexplored. We generated genomic data for 1318 individuals from 35 populations in western central Africa, where Bantu languages originated. We found that early Bantu speakers first moved southward, through the equatorial rainforest, before spreading toward eastern and southern Africa. We also found that genetic adaptation of Bantu speakers was facilitated by admixture with local populations, particularly for the *HLA* and *LCT* loci. Finally, we identified a major contribution of western central African Bantu speakers to the ancestry of African Americans, whose genomes present no strong signals of natural selection. Together, these results highlight the contribution of Bantu-speaking peoples to the complex genetic history of Africans and African Americans.

Linguistic and archaeological records indicate that Bantu languages, together with agriculture, expanded ~4000 to 5000 years ago from western central Africa to eastern and southern Africa (*1*). Population genetics studies have informed us about the genetic structure of African populations and demonstrated that the expansion of Bantu languages was accompanied by a diffusion of people (*2*–*5*). However, most genomic studies have focused on comparisons between farming and hunter-gathering populations (*6*, *7*) rather than on patterns of diversity among Bantu-speaking populations across the continent. Thus, although Bantu speakers today account for one-third of sub-Saharan Africans, many aspects of their genetic history remain unknown.

One debated question concerns the routes followed by Bantu speakers during their dispersal across sub-Saharan Africa, owing to the poor population coverage in the Bantu heartland (i.e., the Nigeria/Cameroon frontier) or the limited genetic resolution of previous studies (*2*–*5*). Furthermore, documentation of how Bantu speakers adapted to the new environments they encountered—from the grasslands of Cameroon to the African rainforest, the East African plateau, and the Kalahari desert—is unknown. Their rapid adaptation may have been facilitated by the acquisition, via admixture, of adaptive alleles from local populations; the impact of this process on recent human evolution remains largely unexplored (*7*, *8*). Finally, large-scale movements of Bantu peoples have not been limited to Africa, as historical records indicate that people from western central Africa were massively deported to North America during the transatlantic slave trade (*9*).

We dissected the genetic and adaptive history of Bantu-speaking populations (BSPs, which refers here to traditional farming groups and does not include Bantu-speaking rainforest hunter-gatherers) by generating genome-wide single-nucleotide polymorphism (SNP) data for 1318 individuals from 35 linguistically and anthropologically well-defined populations of western and western central Africa, including the Bantu homeland (Fig. 1A and table S1). After quality control (fig. S1) (*10*), we combined these data with data sets for other BSPs and non-BSPs from sub-Saharan Africa (table S2). We obtained a total of 548,055 high-quality SNPs in 2055 individuals from 57 populations.

Genetic cluster analyses (*11*) showed that BSPs from western central (wBSP), eastern (eBSP), southwestern (swBSP), and southeastern (seBSP) Africa clustered together (Fig. 1B and figs. S2 to S4), echoing their modest levels of genetic differentiation (analysis of molecular variance–based F_{ST} < 0.01) (table S3). This relative homogeneity reflects the recent separation of BSPs after their expansions throughout sub-Saharan Africa (*5*, *12*). Furthermore, wBSPs, eBSPs, and seBSPs displayed moderate proportions of ancestry from western rainforest hunter-gatherers (~16%), Afroasiatic-speaking farmers (~17%), and San hunter-gatherers (~23%), respectively, suggesting admixture with local populations.

Two hypotheses have been proposed concerning the dispersal of Bantu-speaking populations across sub-Saharan Africa (*2*–*4*). According to the "early-split" hypothesis, the western and eastern branches split early, within the Bantu heartland, into separate migration routes. By contrast, the "late-split" model suggests an initial spread southward from the Bantu homeland into the equatorial rainforest (i.e., Gabon/Angola), followed by expansions toward the rest of the subcontinent. We tested these hypotheses by determining whether eBSPs and seBSPs were genetically closer to wBSPs from the southern part, relative to wBSPs from the northern part, of western central Africa. The populations from this core region can be distinguished along the first axis of the haplotype-based principal component analysis (PCA) (Fig. 1C and figs. S5 to S7) (*13*), mirroring genetic isolation due to both geography and linguistic barriers (fig. S8) (*10*). We overcame problems due to the levels of non-BSP ancestry detected in eBSPs and seBSPs (Fig. 1B) by using haplotype-based admixture inference with GLOBETROTTER (*14*) to account for potential admixture.

The GLOBETROTTER method estimated that eBSPs resulted from two consecutive admixture events (P < 0.05) occurring 1000 to 1500 years ago and 150 to 400 years ago between a wBSP (~75% contribution) and an Afroasiatic-speaking population from Ethiopia (~10% contribution) (table S4). For both events, the best-matching parental wBSP was located in Angola and support for a northern central African origin was weak (Fig. 2A and figs. S9 and S10). In southern Africa, seBSPs displayed

[1]Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France. [2]Centre National de la Recherche Scientifique URA3012, 75015 Paris, France. [3]Center of Bioinformatics, Biostatistics, and Integrative Biology, Institut Pasteur, 75015 Paris, France. [4]Evolutionary Biology Group, School of Biological Sciences, University of Reading, Reading RG6 6BX, England. [5]Departments of English and Anthropology, University of Missouri, Columbia, Missouri, MO 65211, USA. [6]Centre National de la Recherche Scientifique UMR7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Sorbonne Paris Cité, 75016 Paris, France. [7]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA. [8]Université de Montréal, Centre de Recherche CHU Sainte-Justine, Montréal, Québec H3T 1C5, Canada. [9]Institut de Recherche pour le Développement, UMR 208, Muséum National d'Histoire Naturelle, 75005 Paris, France. [10]Centre d'Etude et de Recherche sur le Paludisme Associé à la Grossesse et l'Enfance (CERPAGE), Cotonou, Bénin. [11]Institut de Recherche Clinique du Bénin (IRCB), 01 BP 188 Cotonou, Bénin. [12]Anthropologie Moléculaire et Imagerie de Synthèse, Centre National de la Recherche Scientifique UMR 5288/Université Paul Sabatier Toulouse 3, 31073 Toulouse Cedex 3, France. [13]Institut de Recherche pour le Développement, UMR 216, 75006 Paris, France. [14]Communautés d'Universités et Etablissements (COMUE) Sorbonne Paris Cité, Faculté de Pharmacie, Université Paris Descartes, 75006 Paris, France. [15]Institut National d'Etudes Démographiques, 75020 Paris, France. [16]Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto 4200-135, Portugal. [17]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto 4200-465, Portugal. [18]Faculdade de Medicina da Universidade do Porto, Porto 4200-319, Portugal. [19]Centre National de la Recherche Scientifique UMR 5596, Dynamique du Langage, Université Lumière-Lyon 2, 69007 Lyon, France. [20]Laboratoire Langue, Culture et Cognition (LCC), Université Omar Bongo, 13131 Libreville, Gabon. [21]Department of Genetics, Stanford University, Stanford, CA 94305, USA. [22]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: epatin@pasteur.fr (E.P.); quintana@pasteur.fr (L.Q.-M.)

Downloaded from http://science.sciencemag.org/ on August 4, 2018

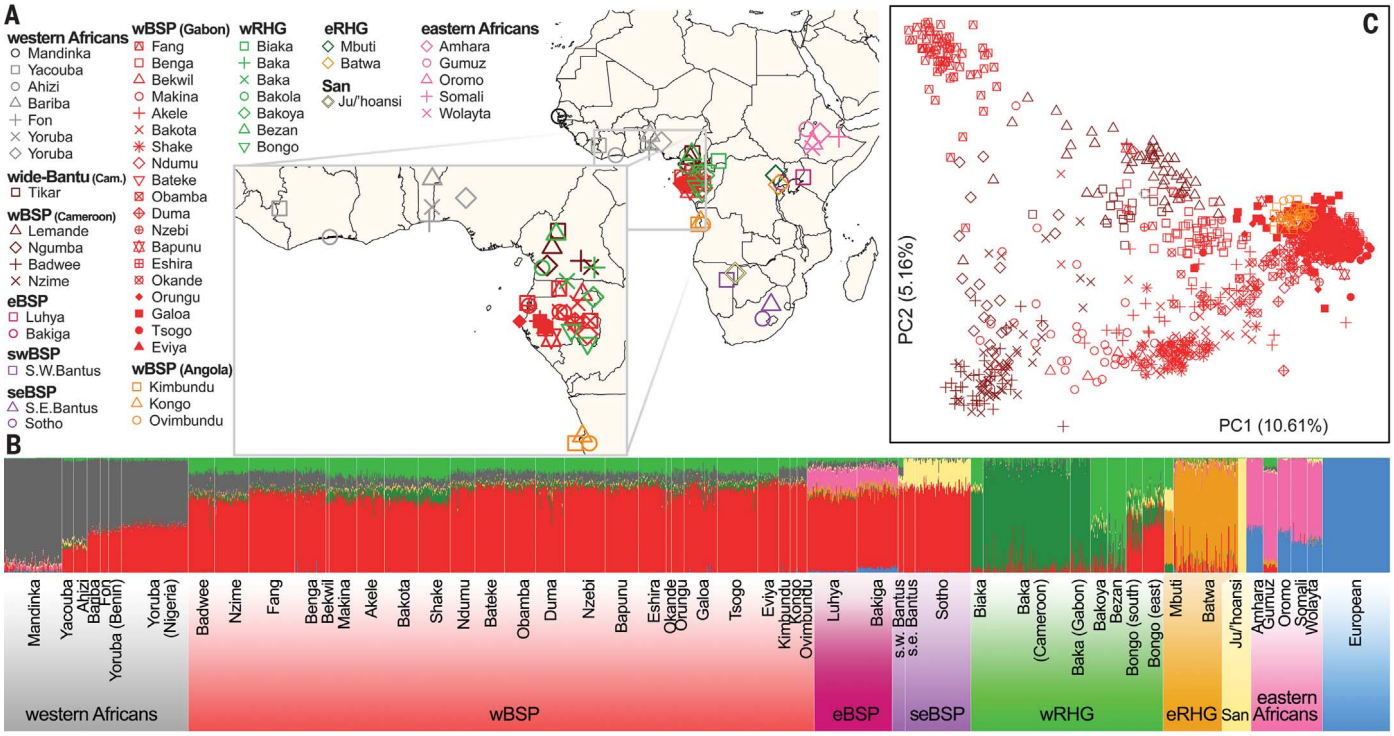


Fig. 1. Genetic structure of African populations. (**A**) Geographic locations of sampled populations. The inset shows the homeland of Bantu expansions. wRHG and eRHG correspond to western and eastern rainforest hunter-gatherers, respectively. (**B**) Clustering analysis was performed on 2055 individuals and 406,798 independent SNPs with ADMIXTURE (*11*). Results for varying numbers of postulated ancestral populations (*K*) are shown in fig. S2. (**C**) Haplotype-based PCA of wide-Bantu-speaking and narrow-Bantu-speaking populations from western central Africa, on 1015 individuals and 429,972 SNPs, with the software fineSTRUCTURE (*13*). The proportions of variance explained, expectedly larger than for unlinked SNP data, are shown in brackets.

signals of a unique admixture event ($P < 0.01$) occurring ~700 years ago between a parental BSP (~70% contribution) and the Ju/'hoansi San from Namibia (~23% contribution). The best parental BSP was located in Angola, with some contribution from eBSPs (Fig. 2B and figs. S9 and S10). Furthermore, eastern and southeastern Bantu speakers shared more identical-by-descent segments with Angolans, relative to northern wBSPs (Mann-Whitney test; $P < 10^{-16}$) (table S5). Although additional sampling of African populations may further refine these patterns, our results, together with previous genetic data supporting the late-split model (*2*, *3*), indicate that BSPs first moved southward through the rainforest before migrating toward eastern and southern Africa, where they admixed with local populations. This model is further supported by linguistics (*15*) and archaeoclimate data (*16*), suggesting that a climatic crisis ~2500 years ago fragmented the rainforest into patches and facilitated the early movements of BSPs farther southward from their original homeland.

As they dispersed through the rainforest, Bantu speakers encountered local populations of rainforest hunter-gatherers (RHGs). We found that the RHG ancestry detected in wBSPs (Fig. 1B and figs. S2 and S5) resulted from an admixture event occurring ~800 years ago, using admixture linkage disequilibrium decay with ALDER ($P < 10^{-8}$) (table S6) (*17*) and GLOBETROTTER ($P < 0.01$) (table S4) (*14*). These results, together

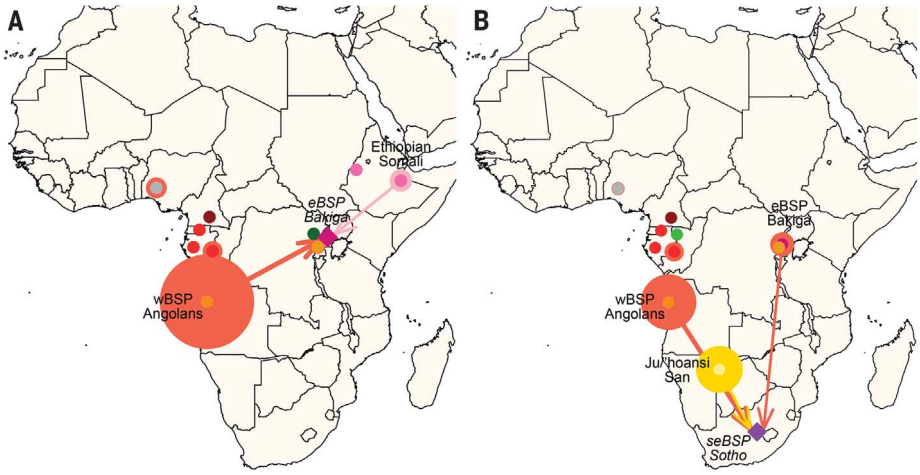


Fig. 2. Reconstructing the dispersal of Bantu-speaking populations. Haplotype-based inference of the genetic origins of (**A**) eBSPs and (**B**) seBSPs. The names of the tested admixed populations are shown in italics. Circle sizes are proportional to the relative genetic contribution of parental populations to admixed populations. Only the oldest admixture event in eBSPs is represented; the most recent admixture event and other examples are shown in fig. S9.

with the low western RHG ancestry detected among BSPs from eastern and southeastern Africa (<5%), indicate that admixture between wBSPs and RHGs occurred mostly after BSPs had expanded throughout sub-Saharan Africa.

The adaptive history of farming BSPs, which were rapidly exposed and had to adapt to new ecosystems, remains largely unknown. We scanned their genomes for signatures of strong, recent positive selection—i.e., regions showing a high proportion of SNPs presenting both greater extended haplotype homozygosity and population differentiation, relative to a closely related reference population (*10*). We detected eight, five, and

Downloaded from http://science.sciencemag.org/ on August 4, 2018

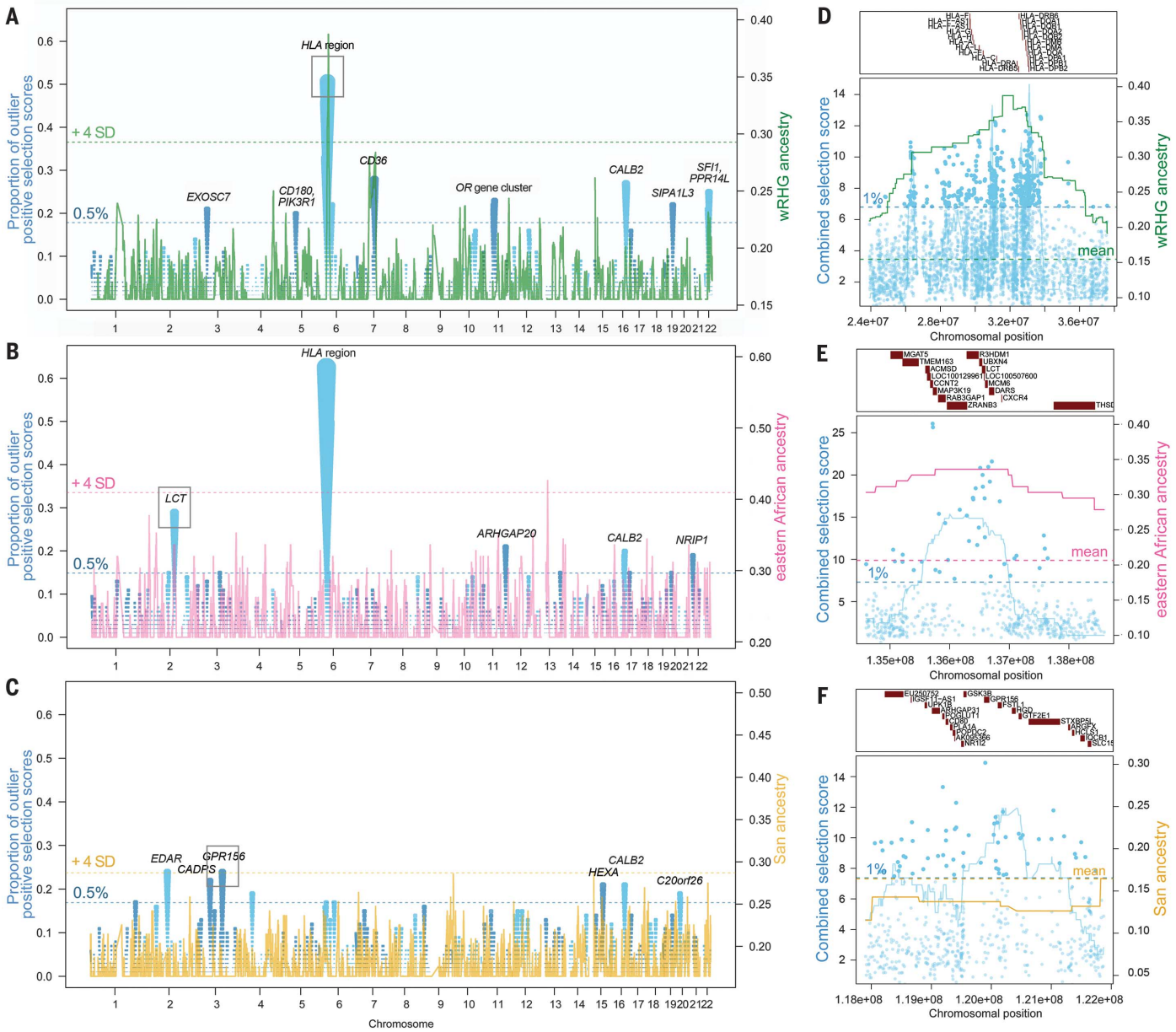


Fig. 3. Genomic signatures of recent positive selection. (**A** to **C**) Genomic signatures of recent positive selection in (A) wBSPs, (B) eBSPs, and (C) seBSPs. Blue points, and their sizes, indicate the proportion, in 100-SNP windows, of SNPs showing outlier neutrality statistics (*10*). (**D** to **F**) Local selection signatures for (D) the *HLA* region in wBSPs, (E) the *LCT* region in eBSPs, and (F) the *GPR156* region in seBSPs. Blue points indicate selection scores for individual SNPs (*10*). The blue line indicates the proportion, in 100-SNP windows, of SNPs showing outlier neutrality statistics. Other candidate loci are shown in figs. S11, S13, and S15. [(A) to (F)] The green, pink, and gold solid lines indicate the local ancestry in BSPs from western RHG, Eastern African, and San populations, respectively.

seven genomic regions presenting strong signatures of recent positive selection in wBSPs, eBSPs, and seBSPs, respectively (tables S7 to S9) (*10*).

The *HLA* locus, which mediates immune response, presented the genome-wide highest proportion of selection signals in both wBSPs and eBSPs (50.5 and 62.4%, respectively) (Fig. 3, A and B, and tables S7 and S8). The most prominent peaks for individual SNP scores were observed in the vicinity of *HLA-D* genes [rs3129302, empirical P (P_{emp}) = 2.9×10^{-5} and rs6907291, $P_{emp} = 6.9 \times 10^{-5}$, respectively] (Fig. 3D and figs. S11 to S14). In wBSPs, the second-strongest hit encompassed *CD36* (Fig. 3A; figs. S11 and S12; and table S7), associated with susceptibility to *Plasmodium falciparum* malaria (*18*). The putatively selected SNP in *CD36* was observed at 25% frequency in wBSPs, yet was essentially absent from non-BSPs from western Africa (rs3211881, $P_{emp} = 5.8 \times 10^{-6}$) (fig. S11F). Adaptive evolution has been demonstrated for a different, unlinked variant at *CD36* in the western African Yoruba of Nigeria (rs3211938) (*19*), suggesting convergent adaptation.

In eBSPs, the next-strongest selection signal overlapped the *LCT* gene region, which encodes the lactase enzyme (28.7%) (Fig. 3, B and E; figs. S13 and S14; and table S8). The derived allele of the best candidate SNP at this locus (rs4954204, $P_{emp} = 5.7 \times 10^{-6}$) displayed high levels of both haplotype homozygosity and genetic differentiation and was linked to the lactase persistence allele C-14010 (*20*). In seBSPs, the proportions of selection signals were lower (<24%) (Fig. 3, C and F; fig. S15; and table S9), possibly reflecting a different demographic and adaptive history.

We scanned the genomes of BSPs for the presence of regions with unusually high levels of non-BSP ancestry (*10*). Again, the *HLA* region in wBSPs showed a strong excess of ancestry from rainforest hunter-gatherers, at 38%, 6.74 SD higher than the genome-wide average of 16% (Fig. 3A). Similar results were obtained when excluding the classical *HLA* region and restricting the analysis to data from a single SNP array

Downloaded from http://science.sciencemag.org/ on August 4, 2018

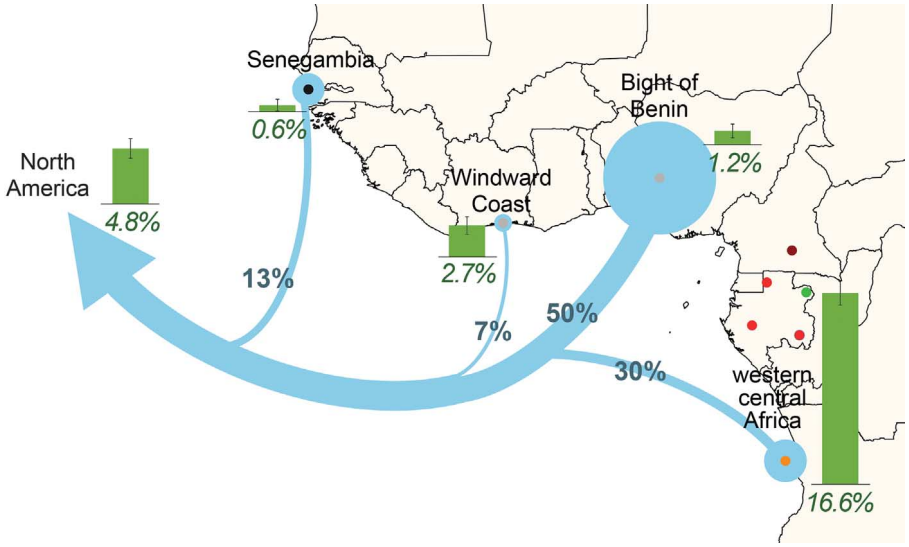


Fig. 4. Dissecting the African origins of African Americans. Estimated genetic contribution, indicated by blue circles, of diverse African populations to African Americans of North America (table S11). African populations were chosen to represent the historical ports from which slaves were embarked during the transatlantic slave trade (*9*). Green bars indicate the western RHG ancestry of African populations and of the African genome of African Americans (fig. S19).

(fig. S11, A and B), indicating that our findings are unlikely to result from the incorrect modeling of the complex *HLA* haplotype structure or misalignments of alleles between SNP arrays. Simulations under realistic demographic models showed that drift or continuous gene flow from RHGs could not account for the high frequency of introgressed *HLA* variants in wBSPs ($P < 0.0001$) (fig. S16 and table S10). Given that these introgressed variants are independent from those presenting the strongest selection signals (*10*), our results indicate that the *HLA* locus has been a hotspot of recent adaptation in BSPs.

We found a local excess of eastern African ancestry in the *LCT* region of eBSPs, and the introgressed variants were those that also showed the strongest positive selection scores of the region (Fig. 3, B and E) (*10*). Simulations indicated that the high frequency of these variants in eBSPs (up to 30% in the Bakiga eBSP and <1% in wBSPs) (fig. S13D and table S8) could not be explained by strong drift or continuous gene flow from eastern Africans ($P < 0.0001$) (fig. S17 and table S10). These observations support a model in which eBSPs acquired the lactase persistence trait from eastern Africans (*20*) and illustrate that the rapid adaptation of human populations migrating to new environments can be facilitated by admixture with local populations.

Last, we estimated the genetic contribution of Bantu-speaking populations to African Americans by analyzing the African ancestry of 5244 African Americans from various locations in North America (table S2). Consistent with previous analyses (*5*, *21*–*23*), the program ADMIXTURE estimated that African Americans had 73% and 78% African ancestry in the northern and southern United States, respectively (fig. S18 and table S11). GLOBETROTTER partitioned their African ancestry into different contributions: 13% from Senegambia, 7% from the Windward Coast, 50% from the Bight of Benin, and up to 30% from western central Africa, mostly from Angola (Fig. 4 and table S11). The estimated contribution of BSPs from western central Africa is consistent with historical records reporting that 23% of slaves transported to North America between 1619 and 1860 originated from this region (*9*). Furthermore, ADMIXTURE estimated that western RHG ancestry accounted for ~4.8% of the African ancestry of African Americans (Fig. 4 and fig. S19). Given that a direct RHG contribution to the slave trade is unlikely (table S12) (*10*), this result further supports that a large fraction of the genome of African Americans derives from wBSPs, who themselves have ~16% western RHG ancestry (Fig. 4). Our results indicate that the ultimate African origins of African Americans are more diverse than previously suggested (*5*, *21*, *23*).

Relaxed selective pressure at the malaria-associated *HBB* and *CD36* genes has been suggested in African Americans, based on large allele frequency differences between African Americans and their assumed, unique African parental population, the non-BSP Yoruba from western Africa (*24*). We replicated this result for *CD36* when considering western Africans only (rs3211938; χ^2-test $P = 2.7 \times 10^{-10}$) (fig. S20A), but it was entirely lost when a more diverse, and realistic, set of African parental sources was used (χ^2-test $P = 0.42$) (fig. S20B). Thus, the *CD36* signal (*24*) is due to the use of the Yoruba as the sole source of African ancestry in African Americans. Furthermore, our analyses did not detect any excess of African ancestry in African American genomes (*25*), using either set of parental populations (fig. S21) (*10*), collectively suggesting that no major changes in selective pressure have occurred in the history of African Americans.

Our study reconstructs the genetic history of Bantu-speaking farming communities, from their initial expansions within Africa to the most recent forced migrations of a subset of these populations to North America. Additional large-scale resequencing studies of geographically and linguistically diverse populations from Africa are needed to provide insight into the evolutionary forces acting on genome diversity at a fine geographic and temporal scale, ultimately facilitating the unbiased identification of variants contributing to diseases in the Southern Hemisphere.

REFERENCES AND NOTES

1. D. W. Phillipson, *African Archaeology* (Cambridge Univ. Press, ed. 2, 1993).
2. G. B. Busby *et al.*, *eLife* **5**, e15266 (2016).
3. C. de Filippo, K. Bostoen, M. Stoneking, B. Pakendorf, *Proc. Biol. Sci.* **279**, 3256–3263 (2012).
4. S. Li, C. Schlebusch, M. Jakobsson, *Proc. Biol. Sci.* **281**, 20141448 (2014).
5. S. A. Tishkoff *et al.*, *Science* **324**, 1035–1044 (2009).
6. J. K. Pickrell *et al.*, *Nat. Commun.* **3**, 1143 (2012).
7. C. M. Schlebusch *et al.*, *Science* **338**, 374–379 (2012).
8. C. Jeong *et al.*, *Nat. Commun.* **5**, 3281 (2014).
9. D. Eltis, D. Richardson, *Atlas of the Transatlantic Slave Trade* (Yale Univ. Press, 2010).
10. See the supplementary materials.
11. D. H. Alexander, J. Novembre, K. Lange, *Genome Res.* **19**, 1655–1664 (2009).
12. D. Gurdasani *et al.*, *Nature* **517**, 327–332 (2015).
13. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, *PLOS Genet.* **8**, e1002453 (2012).
14. G. Hellenthal *et al.*, *Science* **343**, 747–751 (2014).
15. R. Grollemund *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13296–13301 (2015).
16. K. Bostoen *et al.*, *Curr. Anthropol.* **56**, 354–384 (2015).
17. P. R. Loh *et al.*, *Genetics* **193**, 1233–1254 (2013).
18. D. P. Kwiatkowski, *Am. J. Hum. Genet.* **77**, 171–192 (2005).
19. M. Deschamps *et al.*, *Am. J. Hum. Genet.* **98**, 5–21 (2016).
20. S. A. Tishkoff *et al.*, *Nat. Genet.* **39**, 31–40 (2007).
21. K. Bryc *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 786–791 (2010).
22. K. Bryc, E. Y. Durand, J. M. Macpherson, D. Reich, J. L. Mountain, *Am. J. Hum. Genet.* **96**, 37–53 (2015).
23. F. Montinaro *et al.*, *Nat. Commun.* **6**, 6596 (2015).
24. W. Jin *et al.*, *Genome Res.* **22**, 519–527 (2012).
25. G. Bhatia *et al.*, *Am. J. Hum. Genet.* **95**, 437–444 (2014).

ACKNOWLEDGMENTS

We thank all participants who donated samples and participated in this study. We thank C. Schlebusch and G. Hellenthal for helpful discussions. We thank E. Soumonni, a historian whose advice guided the recruitment of Beninese individuals, and J.-P. Chippaux (CERPAGE, Cotonou, Benin) for his help with local authorities. We thank the African Variation Genome Project, the Data Access Committee Chair for the National Human Genome Research Institute (particularly V. Ota Wang), the Electronic Medical Records and Genomics (eMERGE) *Genome-Wide Association* Study, the Multiethnic Cohort Study, the Gene, Environment Association Studies consortium (GENEVA), and the Health Aging and Body Composition (Health ABC) Study for kindly providing access to their data. Detailed acknowledgments can be found elsewhere (*10*). This work was funded by the Institut Pasteur, the Centre National de la Recherche Scientifique (CNRS), Agence Nationale de la Recherche (ANR) grant AGRHUM (ANR-14-CE02-0003-01), and the "Histoire du Génome des Populations Humaines Gabonaises" project (Institut Pasteur/Republic of Gabon). The newly generated SNP genotype data have been deposited in the European Genome-Phenome Archive under accession code EGAS00001002078.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/356/6337/543/suppl/DC1
Materials and Methods
Figs. S1 to S22
Tables S1 to S12
References (*26*–*54*)

12 October 2016; accepted 11 April 2017
10.1126/science.aal1988

Downloaded from http://science.sciencemag.org/ on August 4, 2018

Science

Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America

Etienne Patin, Marie Lopez, Rebecca Grollemund, Paul Verdu, Christine Harmant, Hélène Quach, Guillaume Laval, George H. Perry, Luis B. Barreiro, Alain Froment, Evelyne Heyer, Achille Massougbodji, Cesar Fortes-Lima, Florence Migot-Nabias, Gil Bellis, Jean-Michel Dugoujon, Joana B. Pereira, Verónica Fernandes, Luisa Pereira, Lolke Van der Veen, Patrick Mouguiama-Daouda, Carlos D. Bustamante, Jean-Marie Hombert and Lluís Quintana-Murci

Science **356** (6337), 543-546.
DOI: 10.1126/science.aal1988

On the history of Bantu speakers

Africans are underrepresented in many surveys of genetic diversity, which hinders our ability to study human evolution and the health of modern populations. Patin *et al.* examined the genetic diversity of Bantu speakers, who account for one-third of sub-Saharan Africans. They then modeled the timing of migration and admixture during the Bantu expansion. The analysis revealed adaptive introgression of genes that likely originated in other African populations, including specific immune-related genes. Applying this information to African Americans suggests that gene flow from Africa into the Americas was more complex than previously thought.

Science, this issue p. 543

ARTICLE TOOLS http://science.sciencemag.org/content/356/6337/543

SUPPLEMENTARY MATERIALS http://science.sciencemag.org/content/suppl/2017/05/03/356.6337.543.DC1

REFERENCES This article cites 51 articles, 17 of which you can access for free
http://science.sciencemag.org/content/356/6337/543#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science* is a registered trademark of AAAS.

Downloaded from http://science.sciencemag.org/ on August 4, 2018

Annexe 2

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers

Christina M. Bergey[1,2], Marie Lopez[3,4,5], Genelle F. Harrison[6], Etienne Patin[3,4,5], Jacob Cohen[2], Lluis Quintana-Murci[3,4,5,*], Luis B. Barreiro[6,*], and George H. Perry[1,2,7,*]

[1]Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[2]Department of Biology, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[3]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France.

[4]Centre National de la Recherche Scientifique UMR 2000, Paris, France.

[5]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France.

[6]Université de Montréal, Centre de Recherche CHU Sainte-Justine, Montréal, Canada.

[7]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[*] *Co-senior authors*

June 14, 2018

Corresponding authors: C.M.B. (cxb585@psu.edu) and G.H.P. (ghp3@psu.edu)

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Abstract:

Different human populations facing similar environmental challenges have sometimes evolved convergent biological adaptations, for example hypoxia resistance at high altitudes and depigmented skin in northern latitudes on separate continents. The pygmy phenotype (small adult body size), a characteristic of hunter-gatherer populations inhabiting both African and Asian tropical rainforests, is often highlighted as another case of convergent adaptation in humans. However, the degree to which phenotypic convergence in this polygenic trait is due to convergent vs. population-specific genetic changes is unknown. To address this question, we analyzed high-coverage sequence data from the protein-coding portion of the genomes (exomes) of two pairs of populations, Batwa rainforest hunter-gatherers and neighboring Bakiga agriculturalists from Uganda, and Andamanese rainforest hunter-gatherers (Jarawa and Onge) and Brahmin agriculturalists from India. We observed signatures of convergent positive selection between the Batwa and Andamanese rainforest hunter-gatherers across the set of genes with annotated 'growth factor binding' functions ($p < 0.001$). Unexpectedly, for the rainforest groups we also observed convergent and population-specific signatures of positive selection in pathways related to cardiac development (e.g. 'cardiac muscle tissue development'; $p = 0.001$). We hypothesize that the growth hormone sub-responsiveness likely underlying the pygmy phenotype may have led to compensatory changes in cardiac pathways, in which this hormone also plays an essential role. Importantly, in the agriculturalist populations we did not observe similar patterns of positive selection on sets of genes associated with either growth or cardiac development, indicating that our results most likely reflect

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

a history of convergent adaptation to the similar ecology of rainforest hunter-gatherers rather than a more common or general evolutionary pattern for human populations.

Introduction

Similar ecological challenges may repeatedly result in similar evolutionary outcomes, and many instances of phenotypic convergence arising from parallel changes in the same genetic loci have been uncovered (reviewed in [1–3]). Many examples of convergent genetic evolution reported to date are for simple monogenic traits, for example depigmentation in independent populations of Mexican cave fish living in lightless habitats [4, 5] and persistence of the ability to digest lactose in adulthood in both European and African agriculturalist/pastoralist humans [6]. Most biological traits, however, are highly polygenic. Since the reliable detection of positive selection in aggregate on multiple loci of individually small effect (i.e., polygenic adaptation) is relatively difficult [7–11], the extent to which convergent genetic changes at the same loci and functional pathways or changes affecting distinct genetic pathways may underlie these complex traits is less clear.

Human height is a classic example of a polygenic trait with approximately 800 known loci significantly associated with stature in Europeans collectively accounting for 27.4% of the heritable portion of height variation in this population [12]. A stature phenotype also represents one of most striking examples of convergent evolution in humans. Small body size (or the "pygmy" phenotype, e.g. average adult male stature <155 cm) appears to have evolved independently in rainforest hunter-gatherer populations from Africa, Asia, and South America [13], as groups on different continents do not share common ancestry to the exclusion of nearby agriculturalists [14, 15]. Positive correlations between stature and the degree of admixture with neighboring agriculturalists have confirmed that the pygmy phenotype is,

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

at least in part, genetically mediated and therefore potentially subject to natural selection [16–20].

Indeed, previous population genetic studies have identified signatures of strong positive natural selection across the genomes of various worldwide rainforest hunter-gatherer groups [15, 19, 21, 22]. In some cases, the candidate positive selection regions were significantly enriched for genes involved in growth processes and pathways [15, 19]. However, in one rainforest hunter-gatherer population, the Batwa from Uganda, an admixture mapping approach was used to identify 16 genetic loci specifically associated with the pygmy phenotype [17]. While these genomic regions were enriched for genes involved in the growth hormone pathway and for variants associated with stature in Europeans, there was no significant overlap between the pygmy phenotype-associated regions and the strongest signals of positive selection in the Batwa genome. Rather, subtle shifts in allele frequencies were observed across these regions in aggregate, consistent with a history of polygenic adaptation for the Batwa pygmy phenotype [17] and underscoring the importance of using different types of population genetic approaches to study the evolutionary history of this trait. Similar studies focused on other rainforest hunter-gatherer groups have found enrichment for signatures of selection on genes involved in growth [15] and various growth factor signaling pathways [19], immunity [19, 21, 22], metabolism [19, 21, 22], development [15, 22], and reproduction [19, 21, 22].

Here, we investigate population-specific and convergent patterns of positive selection in African and Asian hunter-gatherer populations using genome-wide sequence data from two sets of populations: the Batwa rainforest hunter-gatherers of Uganda in East Africa and the nearby Bakiga agriculturalists [23], and the Jarawa and Onge rainforest hunter-gatherers of the Andaman Islands in South Asia and the Uttar Pradesh Brahmin agriculturalists from mainland India [24, 25]. We specifically test whether convergent or population-specific signatures of positive selection, as detected both with 'outlier' tests designed to identify strong signatures of positive selection and tests designed to identify signatures of polygenic adaptation,

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

are enriched for genes with growth-related functions. After studying patterns of convergent- and population-specific evolution in the Batwa and Andamanese hunter-gatherers, we then repeat these analyses in the paired Bakiga and Brahmin agriculturalists to evaluate whether the evolutionary patterns most likely relate to adaptation to hunter-gatherer subsistence in rainforest habitats, rather than being more generalized evolutionary patterns for human populations.

Results

We sequenced the protein coding portions of the genomes (exomes) of 50 Batwa rainforest hunter-gatherers and 50 Bakiga agriculturalists (dataset originally reported in [23]), identified single nucleotide polymorphisms (SNPs), and analyzed the resultant data alongside those derived from published whole genome sequence data for 10 Andamanese rainforest hunter-gatherers and 10 Brahmin agriculturalists (dataset from [25]). We restricted our analysis to exonic SNPs, for comparable analysis of the Asian whole genome sequence data with the African exome sequence data. To polarize allele frequency differences observed between each pair of hunter-gatherer and agriculturalist populations, we merged these data with those from outgroup comparison populations from the 1000 Genomes Project [26]: exome sequences of 30 unrelated British individuals from England and Scotland (GBR) for comparison with the Batwa/Bakiga data, and exome sequences of 30 Luhya individuals from Webuye, Kenya (LWK) for comparison with the Andamanese/Brahmin data. Outgroup populations were selected for genetic equidistance from the test populations. While minor levels of introgression from a population with European have been observed for the Batwa and Bakiga [23, 27], PBS is relatively robust to low levels of admixture [28].

To identify regions of the genome that may have been affected by positive selection in each of our test populations, we computed the population branch statistic (PBS; [29]) for

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

each exonic SNP identified among or between the Batwa and Bakiga, and Andamanese and Brahmin populations (Fig. S1, S2; Table S15). PBS is an estimate of the magnitude of allele frequency change that occurred along each population lineage following divergence of the most closely related populations, with the allele frequency information from the outgroup population used to polarize frequency changes to one or both branches. Larger PBS values for a population reflect greater allele frequency change on that branch, which in some cases could reflect a history of positive selection [29].

For each analyzed population, we computed a PBS selection index for each gene by comparing the mean PBS for all SNPs located within that gene to a distribution of values estimated by shuffling SNP-gene associations (without replacement) and re-computing the mean PBS value for that gene 100,000 times (Table S17). The PBS selection index is the percentage of permuted values that is higher than the actual (observed) mean PBS value for that gene. Per-gene PBS selection index values were not significantly correlated with gene size (linear regression of log adjusted selection indices against gene length: adjusted $R^2 = -2.74 \times 10^{-5}$, F-statistic $p = 0.81$; Fig. S3), suggesting that this metric is not overtly biased by gene size.

Convergent evolution can operate at different scales, including on the same mutation or amino acid change, different genetic variants between populations but within the same genes, or across a set of genes involved in the same biomolecular pathway or functional annotation. Given that our motivating phenotype is a complex trait and signatures of polygenic adaptation are expected to be relatively subtle and especially difficult to detect at the individual mutation and gene levels, in this study we principally consider patterns of convergence versus population specificity at the functional pathway/annotation level. We do note that when we applied the same approaches described in this study to individual SNPs, we identified several individual alleles with patterns of convergent allele frequency evolution between the Batwa and Andamanese that may warrant further study (Table S16),

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

including a nonsynonymous SNP in the gene *FIG4*, which when disrupted in mice results in a phenotype of small but proportional body size [30]. However, likely related to the above-discussed challenges of identifying signatures of polygenic adaptation at the locus-specific level, the results of our individual SNP and gene analyses were otherwise largely unremarkable, and thus the remainder of our report and discussion focuses on pathway-level analyses.

Outlier signatures of strong convergent and population-specific selection

The set of genes with the lowest (outlier) PBS index values for each population may be enriched for genes with histories of relatively strong positive natural selection. We used a permutation-based analysis to test whether curated sets of genome-wide growth-associated genes (4 lists tested separately ranging from 266-3,996 genes; 4,888 total genes; Suppl. Text) or individual Gene Ontology (GO) annotated functional categories of genes (GO categories with fewer than 50 genes were excluded) have significant convergent excesses of genes with low PBS selection index values (< 0.01) in both of two cross-continental populations, for example the Batwa and Andamanese. Specifically, we first used Fisher's exact tests to estimate the probability that the number of genes with PBS selection index values < 0.01 was greater than that expected by chance, for each functional category set of genes and population. We then reshuffled the PBS selection indices across all genes 1,000 different times for each population to generate distributions of permuted enrichment p-values for each functional category set of genes. We compared our observed Batwa and Andamanese Fisher's exact test p-values to those from the randomly generated distributions as follows. We computed the joint probability of the null hypotheses for both the Andamanese and Batwa being false as $(1 - p_{Batwa})(1 - p_{Andamanese})$, where p_{Batwa} and $p_{Andamanese}$ are the p-values of

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

the Fisher's exact test, and we compared this joint probability estimate to the same statistic computed for the p-values from the random iterations. We then defined the p-value of our empirical test for convergent evolution as the probability that this statistic was more extreme (lower) for the observed values than for the randomly generated values. The resultant p-value summarizes the test of the null hypothesis that both results could have been jointly generated under random chance. While each individual population's outlier-based test results are not significant after multiple test correction, this joint approach provides increased power to identify potential signatures of convergent selection by assessing the probability of obtaining two false positives in these independent samples.

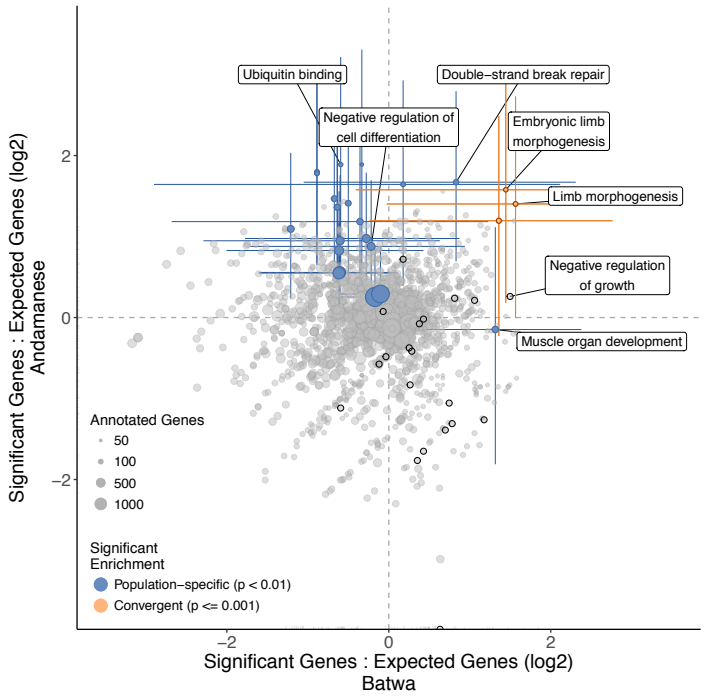
Several GO biological processes were significantly overrepresented—even when accounting for the number of tests performed—among the sets of genes with outlier signatures of positive selection in both the Batwa and Andamanese hunter-gatherer populations (empirical test for convergence $p < 0.005$; Table S1; Fig. 1A). These GO categories include 'limb morphogenesis' (GO:0035108; empirical test for convergence $p < 0.001$; $q < 0.001$; Batwa: genes observed = 5, expected = 1.69, Fisher's exact $p = 0.027$; Andamanese: observed = 6, expected = 2.27, Fisher's exact $p = 0.025$).

Other functional categories of genes were overrepresented in the sets of outlier loci for one of these hunter-gatherer populations but not the other (Fig. 1A; Table S2, S24). The top population-specific enrichments for genes with outlier PBS selection index values for the Batwa were associated with growth and development: 'muscle organ development' (GO:0007517; observed genes: 10; expected genes: 4.02; $p = 0.007$) and 'negative regulation of growth' (GO:0045926; observed = 7; expected = 2.48; $p = 0.012$). Significantly overrepresented GO biological processes for the Andamanese included 'negative regulation of cell differentiation' (GO:0045596; observed genes: 18; expected genes: 9.79; $p = 0.009$). However, these population-specific enrichments were not significant following multiple test correction (false discovery rate $q = 0.71$ for both Batwa terms and $q = 0.22$ for the An-

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

GO categories with significant enrichment for signatures of strong positive selection

A. Rainforest hunter-gatherers



B. Agriculturalists

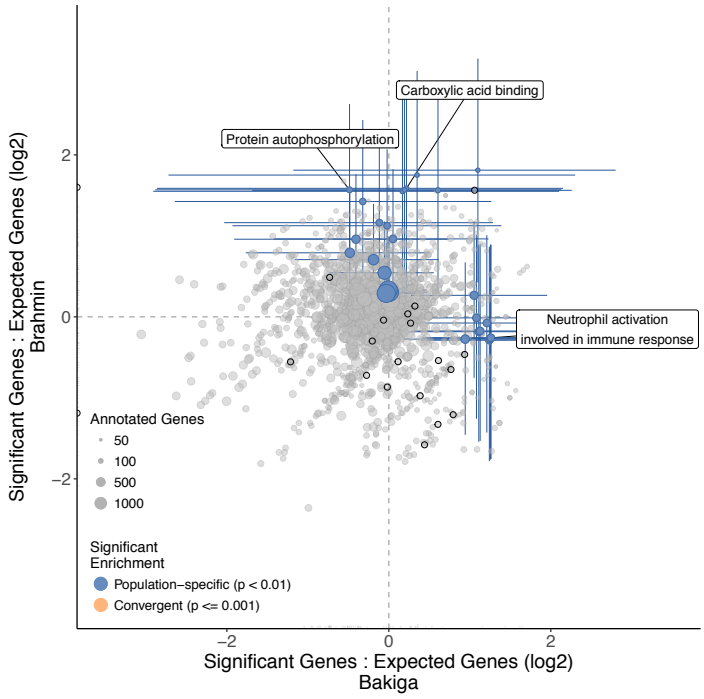


Figure 1: (Continued on the following page.)

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 1: Gene Ontology (GO) functional categories' ratios of expected to observed counts of outlier genes (with PBS selection index < 0.01) in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control comparison (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly overrepresented for genes under positive selection (Fisher $p < 0.01$) in either population are shown in blue and for both populations convergently (empirical permutation-based $p \leq 0.001$) are shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories.

damanese result).

In contrast, no GO functional categories were observed to have similarly significant convergent excesses of 'outlier' genes with signatures of positive selection across the two agriculturalist populations as that observed for the rainforest hunter-gatherer populations (Fig. 1B; Table S19), and the top ranked GO categories from both the convergent evolution analysis and the population-specific analyses were absent any obvious connections to skeletal growth. The top-ranked functional categories with enrichments for genes with outlier PBS selection index values for the individual agriculturalist populations included 'neutrophil activation involved in immune response' for the Bakiga (GO:0002283; observed = 13; expected = 5.43; $p = 0.003$; $q = 0.41$) and 'protein autophosphorylation' for the Brahmin (GO:0046777; observed = 11; expected = 3.71; $p = 0.0012$; $q = 0.16$; Table S24).

Signatures of convergent and population-specific polygenic adaptation

Outlier-based approaches such as that presented above are expected to have limited power to identify signatures of polygenic adaptation [7–11], which is our expectation for the pygmy phenotype [17]. Unlike the previous analyses in which we identified functional categories

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

with an enriched number of genes with outlier PBS selection index values, for our polygenic evolution analysis we computed a "distribution shift-based" statistic to instead identify functionally-grouped sets of loci with relative shifts in their distributions of PBS selection indices. Specifically, we used the Kolmogorov-Smirnov (KS) test to quantify the distance between the distribution of PBS selection indices for the genes within a functional category to that of the genome-wide distribution. Significantly positive shifts in the PBS selection index distribution for a particular functional category may reflect individually subtle but consistent allele frequency shifts across genes within the category, which could result from either a relaxation of functional constraint or a history of polygenic adaptation. Our approach is similar to another recent method that was used to detect polygenic signatures of pathogen-mediated adaptation in humans [31]. As above, we identified functional categories with convergently high KS values between cross-continental groups by repeating these tests 1,000 times on permuted gene-PBS values and computing the joint probability of both null hypotheses being false for the two populations. We then compared this value from the random iterations to the same statistic computed with the observed KS p-values for each functional category. For example, for the Batwa and Andamanese, we tallied the number of random iterations for which the joint probability of both null hypotheses being false was more extreme (lower) than those of the random iterations. In this way we tested the null hypothesis that both of our observed p-values could have been jointly generated by random chance.

The GO molecular function with the strongest signature of a convergent polygenic shift in PBS selection indices across the Batwa and Andamanese populations was 'growth factor binding' (Table S3; Fig. 2A; GO:0019838; Batwa $p = 0.021$; Andamanese $p = 0.027$; Fisher's combined $p = 0.0048$; empirical test for convergence $p < 0.001$; $q < 0.001$), and the top GO biological process was 'organ growth' (GO:0035265; Batwa $p = 0.028$; Andamanese $p = 0.045$; Fisher's combined $p = 0.0095$; empirical test for convergence $p = 0.001$; $q = 1$).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

The other top Batwa-Andamanese convergent GO biological processes are not as obviously related to growth, but instead involve muscles, particularly heart muscles. A significant convergent shift in PBS selection indices across both hunter-gatherer populations was observed for 'cardiac muscle tissue development' (GO:0048738; Batwa $p = 0.046$; Andamanese $p = 0.003$; Fisher's combined $p = 0.001$; empirical test for convergence $p = 0.001$; $q = 1$).

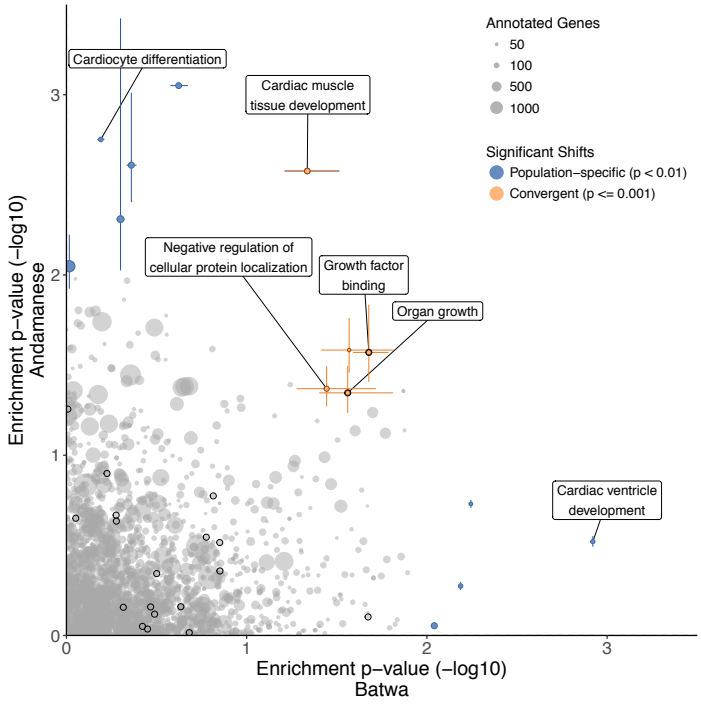
In contrast, when this analysis was repeated on the agriculturalist populations, no growth- or muscle-related functional annotations were observed with significantly convergent shifts in both populations (Fig. 2B; Table S26). The GO categories with evidence of potential convergent evolution between the agriculturalists were the biological processes 'leukocyte differentiation' (GO:0002521; Bakiga $p = 0.0086$; Brahmin $p = 0.0149$; Fisher's combined $p = 0.00128$; convergence empirical $p < 0.001$; $q < 0.001$) and 'protein autophosphorylation' (GO:0046777; Bakiga $p = 0.033$; Brahmin $p = 0.0099$; Fisher's combined $p = 0.003$; convergence empirical $p = 0.001$; $q = 1$).

We also used Bayenv, a Bayesian linear modeling method for identifying loci with allele frequencies that covary with an ecological variable [9, 32], to assess the level of consistency with our convergent polygenic PBS shift results. Specifically, we used Bayenv to test whether the inclusion of a binary variable indicating subsistence strategy would increase the power to explain patterns of genetic diversity for a given functional category of loci over a model that only considered population history (as inferred from the covariance of genome-wide allele frequencies in the dataset.) We converted Bayes factors into per-gene index values via permutation of SNP-gene associations (Table S21) and identified GO terms with significant shifts in the Bayenv Bayes factor index distribution [9, 32] (Table S27). The top results from this analysis included 'growth factor activity' (GO:0008083; $p = 0.006$; $q = 0.11$), categories related to enzyme regulation (e.g. 'enzyme regulator activity'; GO:0030234; $p = 0.003$; $q = 0.01$), and categories related to muscle cell function (e.g. 'microtubule binding'; GO:0008017; $p = 0.003$; $q = 0.10$). There were more GO terms that were highly ranked ($p < 0.05$) in both

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

GO categories with significant enrichment for signatures of polygenic selection

A. Rainforest hunter-gatherers



B. Agriculturalists

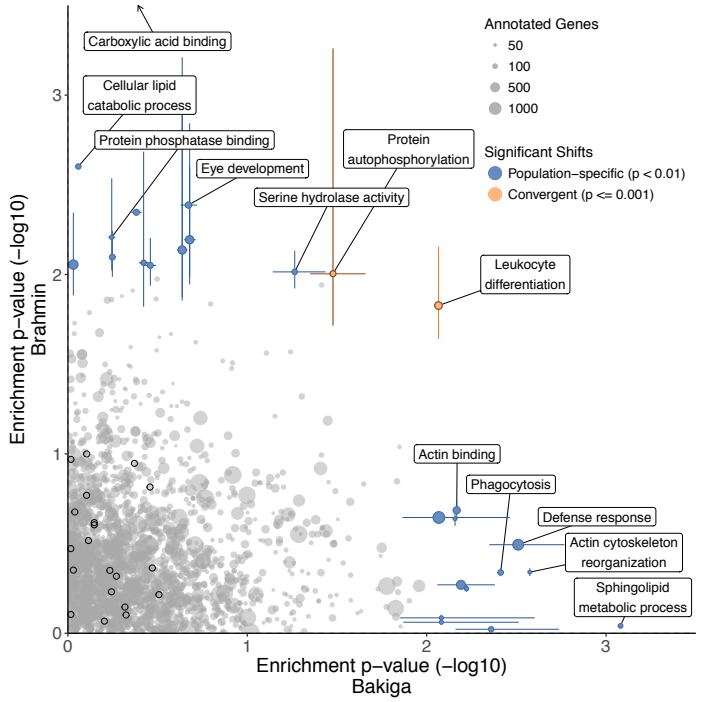


Figure 2: (Continued on the following page.)

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 2: Gene Ontology (GO) functional categories' distribution shift test p-values, indicating a shift in the PBS selection index values for these genes, in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control comparison (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly enriched for genes under positive selection (Kolmogorov-Smirnov $p < 0.01$) in either population are shown in blue and for both populations convergently (empirical permutation-based $p \leq 0.001$) are shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories. One GO molecular function, "carboxylic acid binding" (GO:0031406; Brahmin $p = 7.3 \times 10^{-5}$; $q = 0.0050$) not shown, but indicated with arrow.

the hunter-gatherer PBS shift-based empirical test of convergence and the Bayenv analysis than expected by chance (for biological processes GO terms: observed categories in common = 13, expected = 8.03, Fisher's exact test $p = 9.67 \times 10^{-5}$; for molecular function GO terms: observed categories in common = 4, expected = 1.45, Fisher's exact test $p = 0.045$).

While we did not observe any significant population-specific shifts in PBS selection index values for growth-associated GO functional categories in any of our studied populations (Table S4; Suppl. Text), for each individual rainforest hunter-gatherer population we did observe nominal shifts in separate biological process categories involving the heart (Fig. 2A). For the Batwa, 'cardiac ventricle development' (GO:0003231) was the top population-specific result (median PBS index = 0.272 vs. genome-wide median PBS index = 0.528; $p = 0.001$; $q = 0.302$). For the Andamanese, 'cardiocyte differentiation' (GO:0035051) was also ranked highly (median PBS index = 0.353 vs. genome-wide median PBS index = 0.552; $p = 0.002$; $q = 0.232$). We note that while these are separate population-specific signatures, 17 genes are shared between the above two cardiac-related pathways (of 61 total 'cardiocyte differentiation' genes total, 28%; of 71 total 'cardiac ventricle development' genes, 24%; Table S28).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

In contrast, cardiac development-related GO categories were not observed among those with highly-ranked population-specific polygenic shifts in selection index values for either the Bakiga or Brahmin agriculturalists (Fig. 2B; Table S29). The only GO term with a significant population-specific shift in the agriculturalists after multiple test correction was molecular function 'carboxylic acid binding' in the Brahmins (GO:0031406; $p = 7.30 \times 10^{-5}$; $q = 0.005$).

To ensure that our results were robust to several possible biases, we repeated the above analyses with several modifications. First, to control for potential biases related to variation in gene length and SNP minor allele frequency (MAF), we repeated all analyses after computing the PBS selection index with binning of genes by length and SNPs by MAF, respectively. Our results were not materially different (Tables S5-S12; Figs. S4-S8; Suppl. Text). Second, to account for the effect of linkage disequilibrium among SNPs within a gene, we re-computed the empirical test for convergence p-values by permuting gene-GO relationships when generating the random null distributions for the PBS selection index values instead of gene-PBS relationships as in our original analysis. Again, downstream results were largely unchanged (Table S13-S14; Suppl. Text). These additional analyses increase our confidence that our results are not artifactual.

Discussion

The independent evolution of small adult body size in multiple different tropical rainforest environments worldwide presents a natural human model for comparative study of the genetic and evolutionary bases of growth and body size. Through an evolutionary genomic comparison of African and Asian rainforest hunter-gatherer populations to one another and with nearby agriculturalists, we have gained additional, indirect insight into the genetic structure of body size, a fundamental biological trait. Specifically, we identified a signa-

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

ture of potential convergent positive selection on the growth factor binding pathway that could partially underlie the independent evolution of small body size in African and Asian rainforest hunter-gatherers.

Unexpectedly, we also observed signatures of potential polygenic selection across functional categories of genes related to heart development in the rainforest hunter-gatherer populations, both convergently and on a population-specific basis. To a minor extent, the growth factor- and heart-related functional categories highlighted in our study do overlap: of the 123 total genes annotated across the three heart-related categories ('cardiac muscle tissue development' GO:0048738, 'cardiac ventricle development' GO:0003231, and 'cardiocyte differentiation' GO:0035051), nine (7.3%) are also included among the 66 annotated genes in the 'growth factor binding' category (GO:0019838). However, even after excluding these nine genes from our dataset, we still observed similar polygenic PBS shifts in the Batwa and Andamanese for both growth factor- and heart-related functional categories (Suppl. Text), demonstrating that our observations are not driven solely by cross-annotated genes.

We hypothesize that the evolution of growth hormone sub-responsiveness, which appears to at least partly underlie short stature in some rainforest hunter-gatherer populations [33–37] may in turn have also resulted in strong selection pressure for compensatory adaptations in cardiac pathways. The important roles of growth hormone (GH1) in the heart are evident from studies of patients deficient in the hormone. For example, patients with growth hormone deficiency are known to be at an increased risk of atherosclerosis and mortality from cardiovascular disease [38] and have worse cardiac function [39]. More broadly, shorter people have elevated risk of coronary artery disease [40], likely due to the pleiotropic effects of variants affecting height and atherosclerosis development [41]. Such health outcomes may relate to the important roles that growth hormone plays in the development and function in the myocardium [42, 43], which contains a relatively high concentration of receptors for growth hormone [44]. We hypothesize that the adaptive evolution of growth hormone sub-

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

328 responsiveness underlying short stature in rainforest hunter-gatherers may have necessitated
329 compensatory adaptations in the cardiac pathways reliant on growth hormone.

330 An alternative explanation for our finding of potential convergent positive selection on
331 cardiac-related pathways relates to the nutritional stress of full-time human rainforest habi-
332 tation. Especially prior to the ability to trade forest products for cultivated goods with
333 agriculturalists, the diets of full-time rainforest hunter-gatherers may have been calorically
334 and nutritionally restricted on at least a seasonal basis [13]. Caloric restriction has a direct
335 functional impact on cardiac metabolism and function, with modest fasting in mice leading to
336 the depletion of myocardial phospholipids, which potentially act as a metabolic reserve to en-
337 sure energy to essential heart functions [45]. In human rainforest hunter-gatherers, selection
338 may have favored variants conferring cardiac phenotypes optimized to maintain myocardial
339 homeostasis during the nutritional stress that these populations may have experienced in
340 the past.

341 An important caveat to our study is the lack of statistical significance for our population-
342 specific analyses after controlling for the multiplicity of tests resulting from hierarchically
343 nested GO terms. The absence of strong signals of positive selection that are robust to
344 the multiple testing burden likely reflects both the expected subtlety of evolutionary sig-
345 nals of selection on polygenic traits and the restriction of our dataset to gene coding region
346 sequences. However, our comparative approach to identify signatures of convergent evo-
347 lution is more robust. Therefore, while we cannot yet accurately estimate the extent to
348 which signatures of positive selection that potentially underlie the evolution of the pygmy
349 phenotype occurred in the same versus distinct genetic pathways between the Batwa and
350 Andamanese, we do feel confident in our findings of convergent growth-related and cardiac-
351 related pathways evolution. The concurrent signatures of convergent evolution across these
352 two pathways in both African and Asian rainforest hunter-gatherers is an example of the in-
353 sight into a biomedically-relevant phenotype that can be gained from the comparative study

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

of human populations with non-pathological natural variation.

Materials and Methods

Sample collection and dataset generation

Sample collection, processing, and sequencing have been previously described [17, 23]. Briefly, sampling of biomaterials (blood or saliva) from Batwa rainforest hunter-gatherers and Bakiga agriculturalists of southwestern Uganda took place in 2010 [17]. The study was approved by the Institutional Review Boards (IRBs) of both the University of Chicago (#16986A) and Makerere University, Kampala, Uganda (#2009-137), and local community approval and individual informed consent were obtained before collection. DNA samples of 50 Batwa and 50 Bakiga adults were included in the present study. Exome capture, sequencing, and variant calling were described previously [23]. Briefly, sequence reads were aligned to the hg19/GRCh37 genome with BWA v.0.7.7 mem with default settings [46], PCR duplicates were detected with Picard Tools v.1.94 (http://broadinstitute.github.io/picard), and re-alignment around indels and base quality recalibration was done with GATK v3.5 [47] using the known indel sites from the 1000 Genomes Project [26]. Variants were called individually with GATK HaplotypeCaller [47], and variants were pooled together with GATK GenotypeGVCF and filtered using VQSR. Only biallelic SNPs with a minimum depth of 5x and less than 85% missingness that were polymorphic in the entire dataset were retained for analyses.

Variant data for the Andamanese individuals (Jarawa and Onge) and an outgroup mainland Indian population (Uttar Pradesh Brahmins) from [25] were downloaded in VCF file format from a public website. To ensure the exome capture-derived African and whole genome shotgun sequencing-derived Asian datasets were comparable, we restricted our analyses of these data to exonic SNPs only.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Merging with 1000 Genomes data

We chose outgroup comparison populations from the 1000 Genomes Project [26] to be equally distantly related to the ingroup populations: Reads from a random sample of 30 unrelated individuals from British in England and Scotland (GBR) and Luhya in Webuye, Kenya (LWK) were chosen for the Batwa/Bakiga and Andamanese/Brahmin datasets, respectively. We re-called variants in each 1000 Genomes comparison population at loci that were variable in the ingroup populations using GATK UnifiedGenotyper [47]. Variants were filtered to exclude those with QD $<$ 2.0, MQ $<$ 40.0, FS $>$ 60.0, HaplotypeScore $>$ 13.0, MQRankSum $<$ -12.5, or ReadPosRankSum $<$ -8.0. We removed SNPs for which fewer than 10 of the 30 individuals from the 1000 Genomes datasets had genotypes.

Computation of the Population Branch Statistic (PBS) and the per-gene PBS index

Using these merged datasets, we computed F_{ST} between population pairs using the unbiased estimator of Weir and Cockerham [48], transformed it to a measure of population divergence $[T = -log(1 - F_{ST})]$, and then calculated the Population Branch Statistic (PBS), after [29]. PBS was computed on a per-SNP basis. We computed an empirical p-value for each SNP, simply the proportion of coding SNPs with PBS greater than the value for this SNP, which we adjusted for FDR.

SNPs were annotated with gene-based information using ANNOVAR [49] with refGene (Release 76) [50] and PolyPhen [51] data. As the Andamanese/Brahmin dataset spanned the genome and the Batwa/Bakiga exome dataset included off target intronic sequences as well as untranslated regions (UTRs), and microRNAs, we restricted our analysis to only exonic SNPs. For both the Batwa/Bakiga and Andamanese/Brahmin datasets, we computed a "PBS selection index" for each gene as follows. We compared the mean PBS for all

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

SNPs located within that gene to a distribution of values estimated by shuffling SNP-gene associations (without replacement) and re-computing the mean PBS value for that gene 10,000 times. We defined the PBS selection index of the gene as the percentage of these empirical mean values that is higher than its observed mean PBS value. When identifying outlier genes, gene-based indices were adjusted for FDR.

In order to assess potential biases related to variation in gene length and SNP minor allele frequencies (MAF), we repeated all analyses after computing the PBS selection index with binning of genes by length or SNPs by MAF. Complete details of these methods are included in the Supplemental Text.

To identify SNPs with allele frequencies correlated with subsistence strategy (hunter-gatherer: Andamanese and Batwa; agriculturalists: Bakiga and Brahmin), we used Bayenv2.0 [32] to assess whether the addition of a binary variable denoting subsistence strategy improved the Bayesian model that already took into account covariance between samples due to ancestry. As with the PBS results, we computed an index for each gene by sampling new values for each SNP from the distribution of all Bayes factors and comparing the actual average for this gene to those of the bootstrapped replicates.

Creation of *a priori* lists of growth-related genes

To test the hypothesis that genes with known influence on growth would show increased positive selection in rainforest hunter-gatherer populations, we curated *a priori* lists of growth-related genes as described fully in the Supplemental Text. Briefly, we obtained the following gene lists: i) 3,996 genes that affect growth or size in mice (MP:0005378) from the Mouse/Human Orthology with Phenotype Annotations database [52]; ii) 266 genes associated with abnormal skeletal growth syndromes in the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org), as assembled by [53]; iii) 427 genes expressed substantially more highly in the mouse growth plate, the cartilaginous region on the end of

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

long bones where bone elongation occurs, than in soft tissues [lung, kidney, heart; >= 2.0 fold change; [54]]; and iv) 955 genes annotated with the Gene Ontology "growth" biological process (GO:0040007). As the GH/IGF1 pathway is a major regulator of growth and disruptions to the pathway have been implicated in the pygmy phenotype, we also collected lists of genes associated with GH1 and IGF1 respectively from the OPHID database of proteinprotein interaction (PPI) networks [55]. Separately, we also used a list of genes found to be associated with the pygmy phenotype in the Batwa [17].

Statistical overrepresentation and distribution shift tests

Using the PBS and Bayenv indices, we next tested for a statistical over-representation of extreme values ($p < 0.01$) for the above *a priori* gene lists as well as all Gene Ontology (GO) terms using the topGO package of Bioconductor [56], gene-to-GO mapping from the org.Hs.eg.db package [57], and Fisher's exact test in "classic" mode (i.e., without adjustment for GO hierarchy). We similarly performed a statistical enrichment test using the Kolmogorov-Smirnov test again in "classic" mode, which tested for a shift in the distribution of the PBS or Bayenv statistic, rather than an excess of extreme values. In all cases, we pruned the GO hierarchy to exclude GO terms with fewer than 50 annotated genes to reduce the number of tests, leaving 1,742 and 1,816 GO biological processes and 266 and 285 GO molecular functions tested for the African and Asian datasets, respectively. To further reduce the number of redundant tests, we also computed the semantic similarity between GO terms to remove very similar terms. We computed the similarity metric of [58] as implemented in the GoSemSim R package [59], a measure of the overlapping information content in each term using the annotation statistics of their common ancestor terms, and then clustered based on these pairwise distances between GO terms using Ward Hierarchical Clustering. We then pruned GO terms by cutting the tree at a height of 0.5 and retaining the term in each cluster with the lowest p-value. With this reduced set of GO overrepresentation and

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

distribution shift results, we adjusted the p-value for FDR.

Identification of signatures of convergent evolution

We used two methods to identify convergent evolution: i.) computation of simple combined p-values for SNPs, genes, and GO overrepresentation and distribution shift tests using Fisher's and Edgington's methods, and ii.) a permutation based approach to identify GO pathways for which both the Batwa and Andamanese overrepresentation or distribution shift test results are more extreme than is to be expected by chance (the "empirical test for convergence"). These two approaches are summarized below.

We searched for convergence between Batwa and Andamanese individuals by computing the joint p-value for PBS on a per-SNP, per-gene, and per-GO term basis. We calculated all joint p-values using Fisher's method (as the sum of the natural logarithms of the uncorrected p-values for the Batwa and Andamanese tests [60]) as well as via Edgington's method (based on the sum of all p-values [61]). Meta-analysis of p-values was done via custom script and the metap R package [62].

We also assessed the probability of getting two false positives in the Batwa and Andamanese selection results by shuffling the genes' PBS indices 1,000 times and performing GO overrepresentation and distribution shift tests on these permuted values. We compared the observed Batwa and Andamanese p-values to this generated distribution of p-values, as described above. We computed the joint probability of both null hypotheses being false for the Andamanese and Batwa as $(1-p_{Batwa})(1-p_{Andamanese})$, where p_{Batwa} and $p_{Andamanese}$ are the p-values of the Fisher's exact test or of the Kolmogorov-Smirnov test for the outlier- and shift-based tests, respectively, and we compared the joint probability to the same statistic computed for the p-values from the random iterations. The empirical test for convergence p-value was simply the number of iterations for which this statistic was more extreme (lower) for the observed values than for the randomly generated values.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

We also performed a variation of this analysis, but to preserve patterns of linkage disequilibrium among SNPs within a gene in the null distribution, instead of permuting gene-PBS relationships to generate the random null distributions for the PBS selection index values of the two populations considered jointly, we instead permuted the gene-GO relationships. That is, to compute the PBS selection index, the one-to-many relationships between genes and GO terms were shuffled when generating the null distribution, maintaining the groupings of GO terms that were assigned together to an original gene. Full details of this analysis are available in the Supplemental Text.

Script and data availability

All scripts used in the analysis are available at https://github.com/bergeycm/rhg-convergence-analysis and released under the GNU General Public License v3. Exome data for the Batwa and Bakiga populations have previously been deposited in the European Genome-phenome Archive under accession code EGAS00001002457. Extended data tables are available at https://doi.org/10.18113/S1N63M.

References

[1] Stern, D. L. The genetic causes of convergent evolution. *Nature Reviews Genetics* **14**, 751–764 (2013).

[2] Elmer, K. R. & Meyer, A. Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends in Ecology and Evolution* **26**, 298–306 (2011).

[3] Christin, P. A., Weinreich, D. M. & Besnard, G. Causes and evolutionary significance of genetic convergence. *Trends in Genetics* **26**, 400–405 (2010).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

[4] Protas, M. E. *et al.* Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics* **38**, 107–111 (2006).

[5] Gross, J. B., Borowsky, R. & Tabin, C. J. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish Astyanax mexicanus. *PLoS Genetics* **5** (2009).

[6] Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**, 31–40 (2007).

[7] Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nature Reviews Genetics* **11**, 665–667 (2010).

[8] Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**, R208–R215 (2010).

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

[10] Stephan, W. Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology* **25**, 79–88 (2016).

[11] Wellenreuther, M. & Hansson, B. Detecting polygenic evolution: Problems, pitfalls, and promises. *Trends in Genetics* **32**, 155–164 (2016).

[12] Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).

[13] Perry, G. H. & Dominy, N. J. Evolution of the human pygmy phenotype. *Trends in Ecology and Evolution* **24**, 218–225 (2009).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

[14] Rasmussen, M., Guo, X. & Wang, Y. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).

[15] Migliano, A. B. *et al.* Evolution of the pygmy phenotype: Evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Human Biology* **85**, 251–284 (2013).

[16] Perry, G. H. & Verdu, P. Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. *Quaternary International* **448**, 150–157 (2016).

[17] Perry, G. H. *et al.* Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences* **111**, E3596–E3603 (2014).

[18] Becker, N. S. A. *et al.* Indirect evidence for the genetic determination of short stature in African pygmies. *American Journal of Physical Anthropology* **145**, 390–401 (2011).

[19] Jarvis, J. P. *et al.* Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genetics* **8**, e1002641 (2012).

[20] Pemberton, T. J., Verdu, P., Becker, N. S., Willer, C. J. & Hewlett, B. S. A genome scan for genes underlying adult body size differences between Central African pygmies and their non-pygmy neighbors. *bioRxiv* 1–35 (2017).

[21] Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).

[22] Hsieh, P. *et al.* 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 Research* **26**, 279–290 (2015).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

[23] Lopez, M. *et al.* The demographic history and mutational load of African hunter-gatherers and farmers. *Nature Ecology & Evolution* **2**, 721–730 (2018).

[24] Mondal, M. *et al.* Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics* **48**, 1066–1070 (2016).

[25] Mondal, M., Casals, F., Majumder, P. P. & Bertranpetit, J. Further confirmation for unknown archaic ancestry in Andaman and South Asia. *bioRxiv* (2016).

[26] Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[27] Patin, E. *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications* **5**, 3163 (2014).

[28] Huerta-Sánchez, E. *et al.* Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Molecular Biology and Evolution* **30**, 1877–1888 (2013).

[29] Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).

[30] Campeau, P. M. *et al.* Yunis-Varón syndrome is caused by mutations in FIG4, encoding a phosphoinositide phosphatase. *American Journal of Human Genetics* **92**, 781–791 (2013).

[31] Daub, J. T. *et al.* Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution* **30**, 1544–1558 (2013).

[32] Günther, T. & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195**, 205–220 (2013).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

564 [33] Rimoin, D. L., Merimee, T. J., Rabinowitz, D., Cavalli-Sforza, L. L. & McKusick, V. A.
565 Peripheral subresponsiveness to human growth hormone in the African pygmies. *The*
566 *New England Journal of Medicine* **281**, 1383–1388 (1969).

567 [34] Merimee, T. J., Rimoin, D. L., Cavalli-Sforza, L. C., Rabinowitz, D. & McKusick, V. A.
568 Metabolic effects of human growth hormone in the African pygmy. *The Lancet* **292**,
569 194–195 (1968).

570 [35] Merimee, T. J., Rimoin, D. L. & Cavalli-Sforza, L. L. Metabolic studies in the African
571 pygmy. *The Journal of Clinical Investigation* **51**, 395–401 (1972).

572 [36] Geffner, M. E., Bailey, R. C., Bersch, N., Vera, J. C. & Golde, D. W. Insulin-like growth
573 factor-I unresponsiveness in an Efe Pygmy. *Biochemical and Biophysical Research Com-*
574 *munications* **193**, 1216–1223 (1993).

575 [37] Geffner, M. E., Bersch, N., Bailey, R. C. & Golde, D. W. Insulin-like growth factor I
576 resistance in immortalized T cell lines from African Efe Pygmies. *Journal of Clinical*
577 *Endocrinology and Metabolism* **80**, 3732–3738 (1995).

578 [38] Carroll, P. V. *et al.* Growth hormone deficiency in adulthood and the effects of growth
579 hormone replacement: A Review. *The Journal of Clinical Endocrinology & Metabolism*
580 **83**, 382–395 (1998).

581 [39] Arcopinto, M. *et al.* Growth hormone deficiency is associated with worse cardiac func-
582 tion, physical performance, and outcome in chronic heart failure: Insights from the
583 T.O.S.CA. GHD study. *PLoS ONE* **12**, e0170058 (2017).

584 [40] Paajanen, T. A., Oksala, N. K., Kuukasjärvi, P. & Karhunen, P. J. Short stature is
585 associated with coronary heart disease: A systematic review of the literature and a
586 meta-analysis. *European Heart Journal* **31**, 1802–1809 (2010).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

587 [41] Nelson, C. P. *et al.* Genetically determined height and coronary artery disease. *New*
588 *England Journal of Medicine* **372**, 1608–1618 (2015).

589 [42] Devesa, J., Almengló, C. & Devesa, P. Multiple effects of growth hormone in the body:
590 Is it really the hormone for growth? *Clinical Medicine Insights: Endocrinology and*
591 *Diabetes* **9**, 47–71 (2016).

592 [43] Meyers, D. E. & Cuneo, R. C. Controversies regarding the effects of growth hormone
593 on the heart. *Mayo Clinic Proceedings* **78**, 1521–1526 (2003).

594 [44] Mathews, L. S., Enberg, B. & Norstedt, G. Regulation of rat growth hormone receptor
595 gene expression. *The Journal of Biological Chemistry* **264**, 9905–9910 (1989).

596 [45] Han, X., Cheng, H., Mancuso, D. J. & Gross, R. W. Caloric restriction results in phos-
597 pholipid depletion, membrane remodeling, and triacylglycerol accumulation in murine
598 myocardium. *Biochemistry* **43**, 15584–15594 (2004).

599 [46] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
600 transform. *Bioinformatics* **25**, 1754–1760 (2009).

601 [47] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
602 generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

603 [48] Weir, B. & Cockerham, C. Estimating F-statistics for the analysis of population struc-
604 ture. *Evolution* **38**, 1358–1370 (1984).

605 [49] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
606 variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164 (2010).

607 [50] O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status,
608 taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–
609 D745 (2016).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

[51] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).

[52] Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723–D729 (2017).

[53] Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).

[54] Lui, J. C. *et al.* Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Human Molecular Genetics* **21**, 5193–5201 (2012).

[55] Brown, K. R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082 (2005).

[56] Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version* **2** (2016).

[57] Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human* (2017).

[58] Jiang, J. J. & Conrath, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)* (1997).

[59] Yu, G. *et al.* GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

[60] Mosteller, F. & Fisher, R. Questions and answers. *The American Statistician* **2**, 30–31 (1948).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

631 [61] Edgington, E. S. An additive method for combining probability values from independent
632 experiments. *The Journal of Psychology* **80**, 351–363 (1972).

633 [62] Dewey, M. *metap: meta-analysis of significance values* (2017).

634 Acknowledgments

635 The authors would like to thank the Batwa and Bakiga communities and all individuals
636 who participated in this study, and J.A. Hodgson and E.C. Reeves for helpful discussions.
637 This work was supported by NIH R01-GM115656 (to G.H.P and L.B.B.), 1 F32 GM125228-
638 01A1 (to C.M.B), and ANR AGRHUM ANR-14-CE02-0003-01 (to L.Q.-M.). M.L. was
639 supported by the Fondation pour la Recherche Médicale (FDT20170436932). This research
640 was conducted with Advanced CyberInfrastructure computational resources provided by The
641 Institute for CyberScience at The Pennsylvania State University.

642 Competing interests statement

643 The authors declare no competing interests.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Supplemental Material for: Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers

Christina M. Bergey[1,2], Marie Lopez[3,4,5], Genelle F. Harrison[6], Etienne Patin[3,4,5], Jacob Cohen[2], Lluis Quintana-Murci[3,4,5,*], Luis B. Barreiro[6,*], and George H. Perry[1,2,7,*]

[1]Department of Anthropology, Pennsylvania State University, Pennsylvania, U.S.A.

[2]Department of Biology, Pennsylvania State University, Pennsylvania, U.S.A.

[3]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France.

[4]Centre National de la Recherche Scientifique UMR 2000, Paris, France.

[5]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France.

[6]Université de Montréal, Centre de Recherche CHU Sainte-Justine, H3T 1C5 Montréal, Canada.

[7]Huck Institutes of the Life Sciences, Pennsylvania State University, Pennsylvania, U.S.A.

[*] *Co-senior authors*

June 14, 2018

Corresponding authors: C.M.B. (cxb585@psu.edu) and G.H.P. (ghp3@psu.edu)

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Contents

List of Figures

List of Tables

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

1 Supplemental Text

Positive selection signatures on growth-associated genes We examined whether gene-specific signatures of strong positive selection (using an "outlier-based"" designation of genes with PBS index values < 0.01) in the rainforest populations were enriched for known functional associations with growth using *a priori* lists of 4,888 total growth-related genes, consisting of (with some redundancy among individual categories, as expected): i) 3,996 genes that affect growth or size in mice (MP:0005378) from the Mouse/Human Orthology with Phenotype Annotations database [1]; ii) 266 genes associated with abnormal skeletal growth syndromes in the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org/), as assembled by [2]; iii) 427 genes expressed substantially more highly in the mouse growth plate, the cartilaginous region on the end of long bones where bone elongation occurs, than in soft tissues (lung, kidney, heart; >= 2.0 fold change; [3]; and iv) 955 genes annotated with the Gene Ontology "growth" biological process (GO:0040007). Separately, we also considered in our analyses the set of 166 genes located within the 16 genomic regions previously associated with the pygmy phenotype in the Batwa, using an admixture mapping approach [4], as well as GH1- and IGF1-associated genes using data from OPHID database of proteinprotein interaction (PPI) networks [5].

We used each of the curated *a priori* growth-related gene lists for testing the hypothesis that such loci are enriched for genes with signatures of strong positive selection (outlier PBS selection index values) or have a shift in the distribution of PBS selection index values consistent with subtle polygenic adaptation in the Batwa and Andamanese rainforest hunter-gatherer but not the Bakiga and Brahmin agriculturalist populations. We identified 202, 188, 291, and 252 outlier strong selection candidate genes (with PBS index values < 0.01) in each of the Batwa, Bakiga, Andamanese, and Brahmin populations, respectively. Genes in the *a priori* growth-related gene lists were not significantly overrepresented among

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

PBS outliers in any populations, except for those associated with mouse growth phenotype in the Brahmin (68 observed, 47.7 expected; Fisher $p = 0.0179$) (Table S23). Though the lack of over-representation of growth-related gene lists among loci with outlier signatures of strong positive natural selection related to growth is perhaps unsurprising considering the polygenic phenotype, our distribution shift-based test also showed no significant shifts in the distribution of PBS indices for any population (Table S25). Genes in genomic regions previously associated with the pygmy phenotype in the Batwa [4] were enriched for genes with outlier PBS selection index values in the Batwa (outlier-based test: 5 observed, 1.39 expected; Fisher $p = 0.017$; Table S23) and the PBS distribution for the phenotype-associated genes was shifted relative to the genome-wide distribution (distribution shift-based test: Kolmogorov-Smirnov test $p = 0.056$; Table S25). We found no evidence that genes associated with GH1 and IGF1 were enriched for outlier or polygenic selection.

Impact of cross-annotated genes between growth factor- and cardiac-related pathways To assess whether shared genes in GO categories relating to the heart and growth factor binding were responsible for the significant shift in PBS selection index values for genes in these annotations, we compared the distributions of PBS selection indices before and after removing 9 genes common to heart pathways and growth factor binding. The heart GO terms assessed were: 'cardiocyte differentiation' (GO:0035051), with a shift in the Andamanese hunter-gatherers; 'cardiac ventricle development' (GO:0003231), with a shift in the Batwa hunter-gatherers; and 'cardiac muscle tissue development' (GO:0048738) with a convergent shift in the Batwa and Andamanese. Of the 123 heart related genes contained in these pathways, 9 were also annotated to the GO molecular function 'growth factor binding' (GO:0019838): *ACVR1*, *EGFR*, *ENG*, *FGFR2*, *FGFRL1*, *LTBP1*, *SCN5A*, *TGFBR1*, and *TGFBR3*.

After removing the 9 shared genes, the mean PBS selection index for the Andamanese

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

among genes annotated to 'cardiocyte differentiation' decreased slightly from 0.444 to 0.443 and the pre- and post-filtration distributions were not significantly different (Kolmogorov-Smirnov $D = 0.023$, $p = 1$). Similarly, the mean PBS selection index for the Batwa for genes in 'cardiac ventricle development' decreased slightly from 0.654 to 0.652, and the distributions were not significantly different ($D = 0.044$, $p = 1$). Finally, for 'cardiac muscle tissue development', the mean PBS selection index for the Andamanese increased from 0.450 to 0.453, and for the Batwa increased from 0.474 to 0.486. Again the pre- and post-filtering distributions were not significantly different for the Andamanese ($D = 0.015$, $p = 1$) or Batwa ($D = 0.015$, $p = 1$).

Similarly, after removing 9 shared genes, the mean PBS selection index for genes annotated to 'growth factor binding' (GO:0019838) for the Batwa increased slightly from 0.437 to 0.440 and for the Andamanese decreased from 0.455 to 0.437. Again, the pairs of distributions were not significantly different (Batwa: $D = 0.030$, $p = 1$; Andamanese: $D = 0.036$, $p = 1$).

Correcting for potential bias from differing gene size or global minor allele frequency (MAF) In order to assess the potential biases related to differences in gene length (e.g. number of SNPs) or in SNP global minor allele frequencies (MAF), we repeated the analysis after modifying how the PBS selection index was computed. As in the uncorrected analysis, these corrected PBS selection index values were computed using 1,000 iterations.

First, to control for gene size, we sampled the PBS values for each SNP from only genes with the same number of SNPs during the computation of the selection index. For larger genes, gene sizes were binned to ensure sufficient SNPs from which to sample, using sets $[11, 15]$, $[16, 20]$, and $[21, \infty)$.

Second, to control for differing MAF values for SNPs, we did the permutation-based computation of the PBS selection index while matching SNPs on global MAF (computed

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

using the African or Asian datasets for within-continent analyses.) SNPs were grouped by MAF into bins of size 0.01, and for each SNP in a gene, SNPs were sampled from only the set in the MAF bin.

Neither modification to the PBS selection index computation algorithm majorly affected the PBS selection index values nor the GO-based downstream analyses. Corrected and uncorrected PBS selection index values were highly correlated ($R^2 = 0.993$ to 0.997 and 0.953 to 0.985 for the gene size- and MAF-corrections respectively; Fig. S4).

The GO biological processes and molecular functions with the strongest evidence of enrichment for strong selection were similar for the convergent (Tables S5 and S6; Figs. S5 and S6) and population-specific selection analyses (Tables S7 and S8; Figs. S5 and S6). The only mentioned growth- or heart-associated pathway that was no longer significant after correction was the biological process "negative regulation of growth," which was significantly enriched for genes with evidence of strong selection in the Batwa in the original analysis, but its p-value rose to 0.0448 after correction for gene size. In contrast, "cardiac muscle tissue development" (GO:0048738) which originally had a convergent empirical p-value of 0.025, was significantly enriched for strong positive selection convergently in the Batwa and Andamanese after MAF-based filtration ($p = 0.001$).

Similarly, the top GO categories with evidence of polygenic selection were largely unchanged for the convergent (Tables S9 and S10; Figs. S7 and S8) and population-specific selection analyses (Tables S11 and S12; Figs. S7 and S8). Minor changes include "growth factor binding" (GO:0019838) which rose to be no longer significant with the MAF-based correction (original convergent empirical $p < 0.001$; MAF corrected $p = 0.005$).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Modification of significance testing in empirical test for convergent evolution We also modified and repeated the analysis that computes the significance of the convergence GO tests using a permutation-based approach. Whereas we originally permuted gene-PBS relationships to generate the random null distributions of PBS selection index values for two populations considered jointly, we instead permuted the gene-GO relationships to preserve LD patterns. The one-to-many relationships between genes and GO terms were shuffled, maintaining the groupings of GO terms that were assigned together to an original gene. We repeated the GO-based analyses for enrichment of strong selection or polygenic selection 1,000 times with these randomized gene-GO annotations, and compared our actual observed values to this randomly-generated null distribution. As before, we then defined the p-value of our empirical test for convergent evolution as the probability that this statistic was more extreme (lower) for the observed values than for the randomly generated values. The resultant p-value summarizes the test of the null hypothesis that both results could have been jointly generated under random chance. The results of the modified test were only slightly different than the original for both convergence in strong outlier selection (Table S13) and in a shifted PBS selection index (Table S14).

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

2 Figures

Fig. S1: Population Branch Statistic (PBS) schematic.

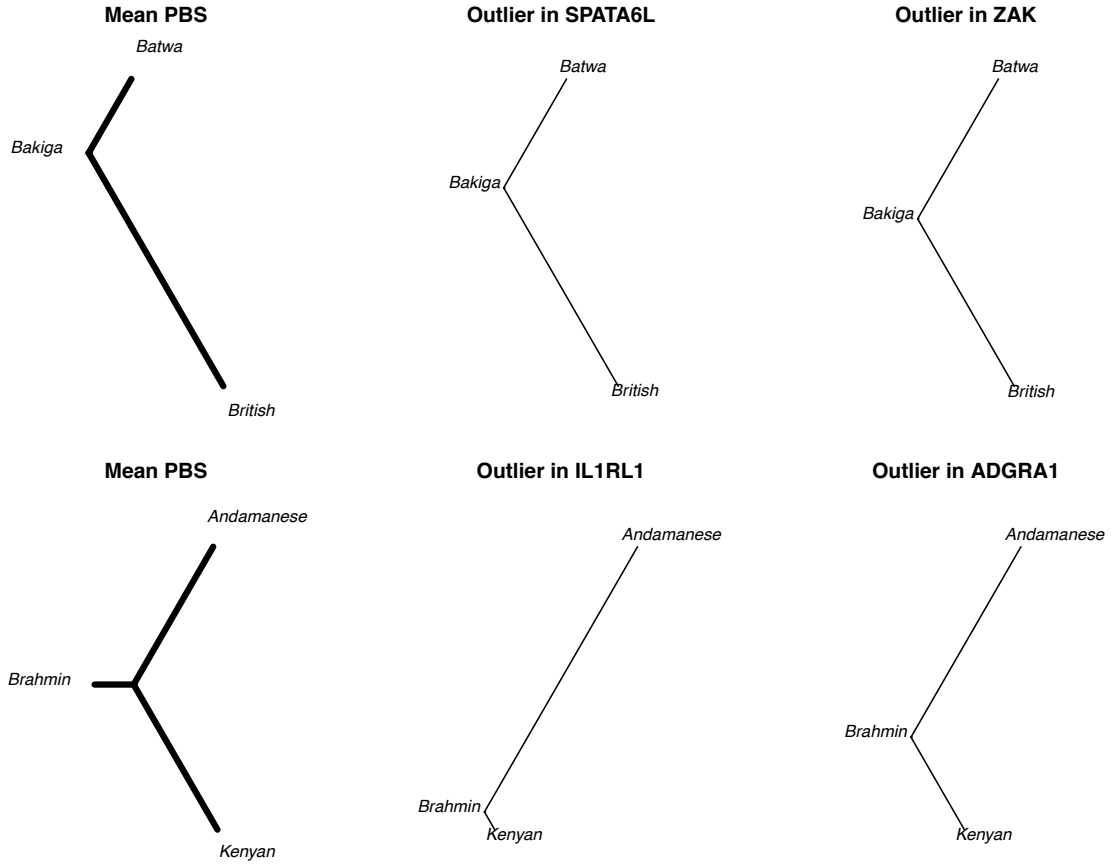


Figure S1: Mean values of the Population Branch Statistic (PBS; left) for the African dataset (Batwa, Bakiga, and outgroup British populations; upper row) and Asian dataset (Andamanese, Brahmin, and outgroup Kenyan populations; lower row). Middle and right columns contain PBS values for two outlier SNPs in each population.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S2: Population Branch Statistic (PBS) by SNP.

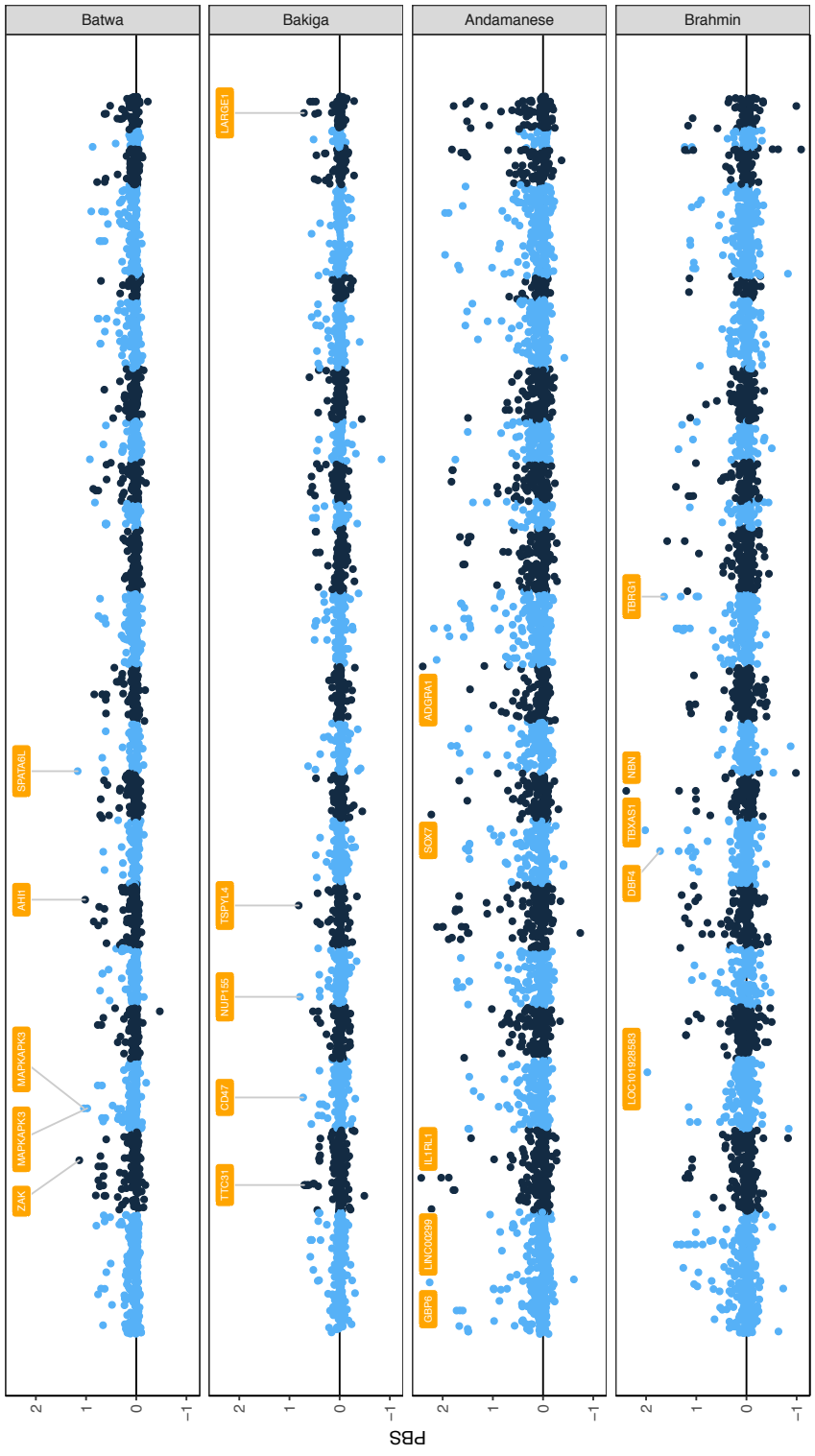


Figure S2: Population Branch Statistic (PBS) values plotted across the genome for the four focal populations. The genes containing the SNPs with the 5 highest PBS values in each population are labeled.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S3: Population Branch Statistic (PBS) by gene SNP count.

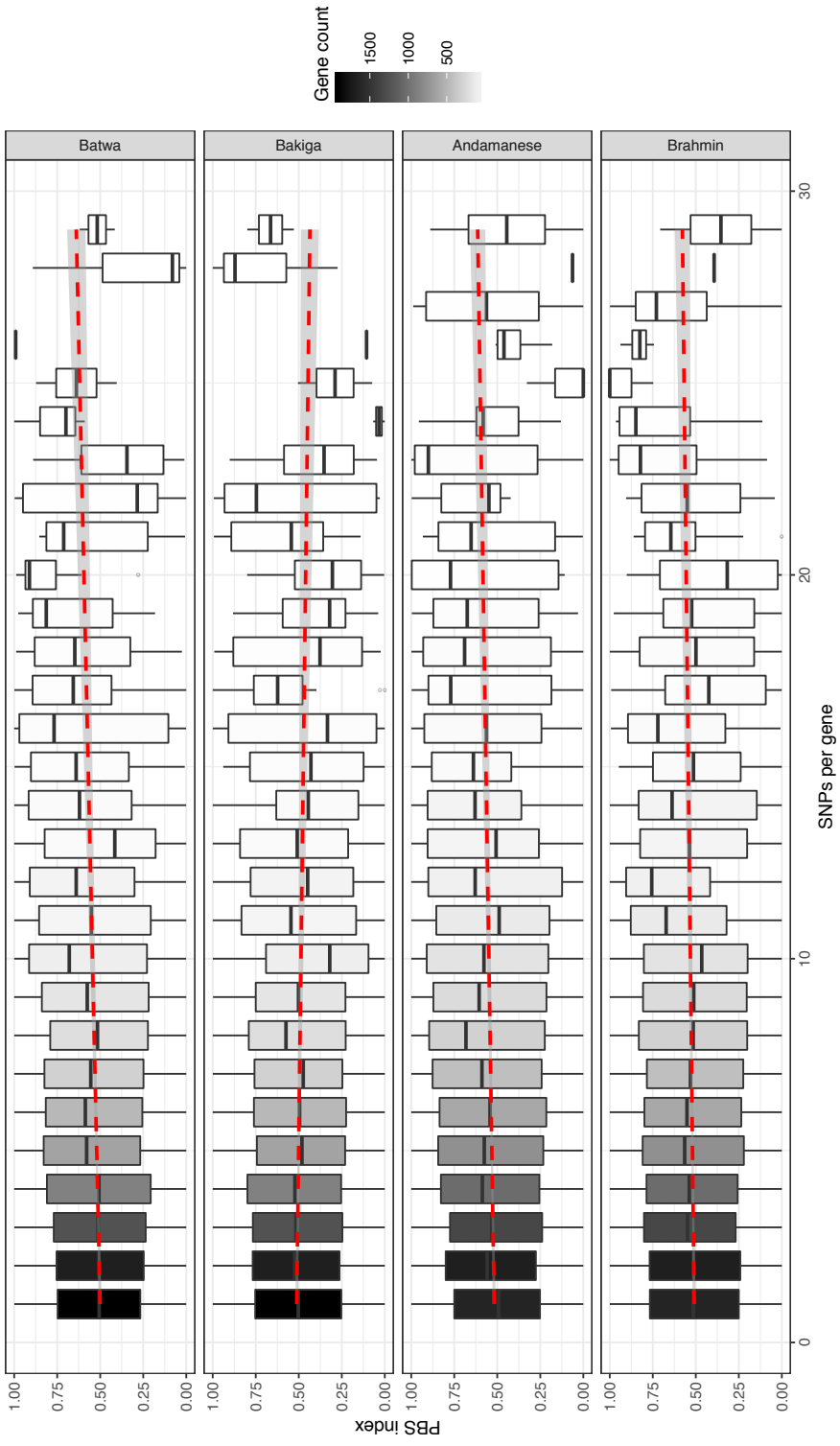


Figure S3: Population Branch Statistic (PBS) selection index values plotted by number of SNPs in gene. Color indicates number of genes with that SNP count. Only SNP counts from 1 to 30 shown.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

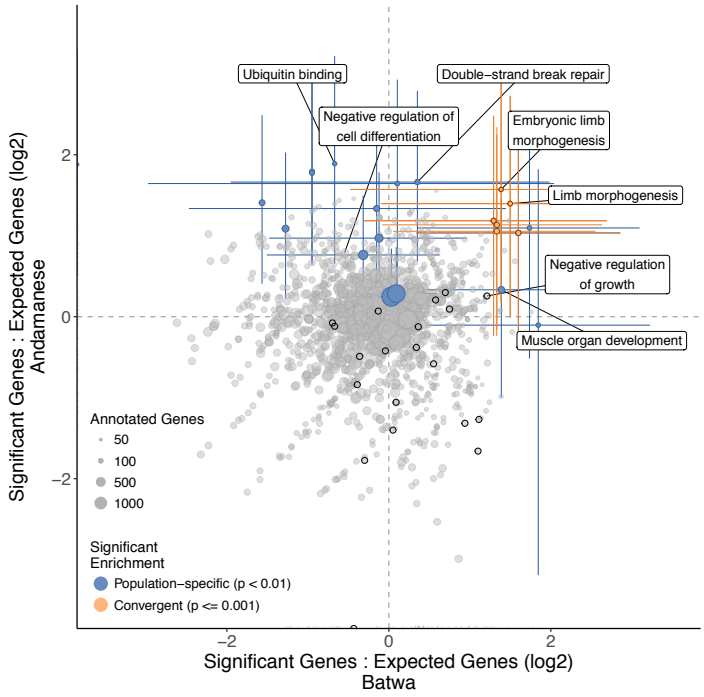
Fig. S4: Gene size- and MAF-based corrections' impact on p-value.

Figure S4: Plots of PBS selection index values for genes corrected for gene size and MAF shown compared to the original uncorrected values (with both plotted on a logarithmic scale. Red shading indicates higher percent difference from original value.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S5: Gene size-corrected strong positive selection enrichment results.

A. Rainforest hunter-gatherers



B. Agriculturalists

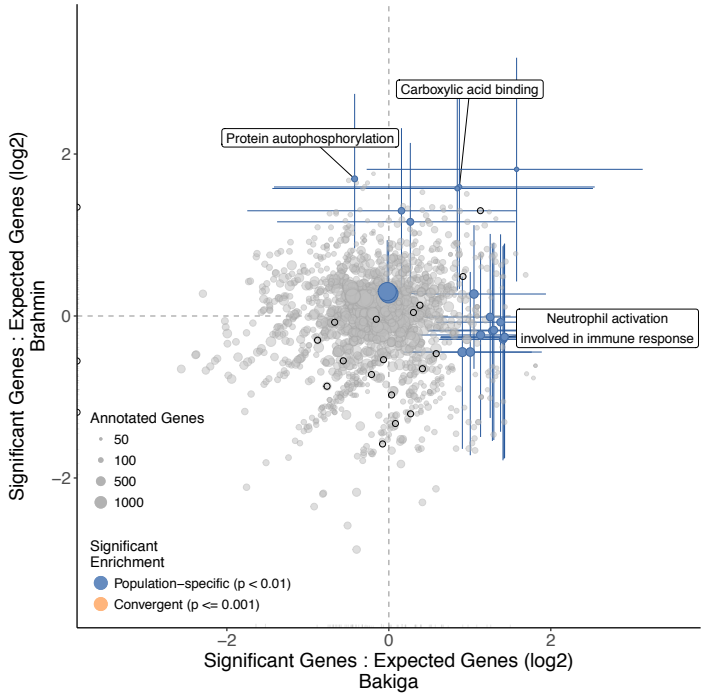


Figure S5: (Continued on the following page.)

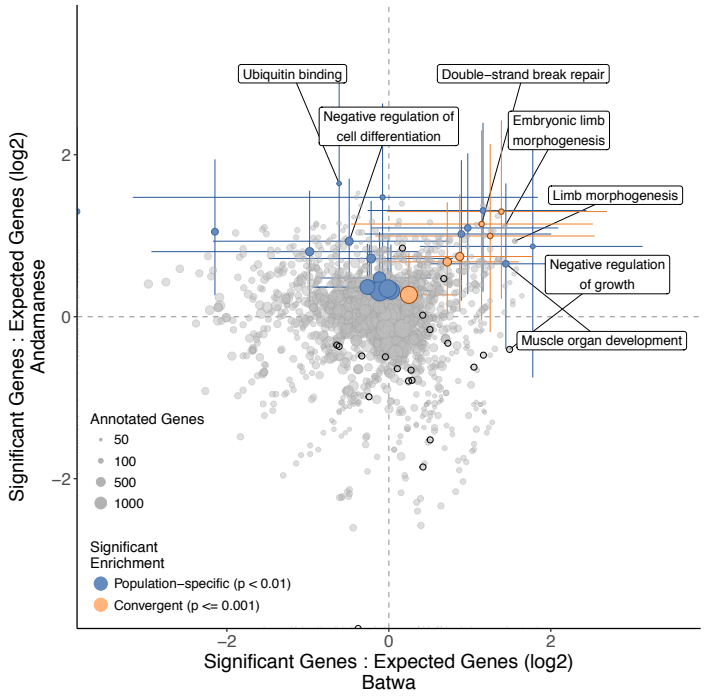
bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure S5: After gene size-based correction, Gene Ontology (GO) functional categories' ratios of expected to observed counts of outlier genes (with PBS selection index < 0.01) in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly overrepresented for genes under positive selection (Fisher $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S6: MAF-corrected strong positive selection enrichment results.

A. Rainforest hunter-gatherers



B. Agriculturalists

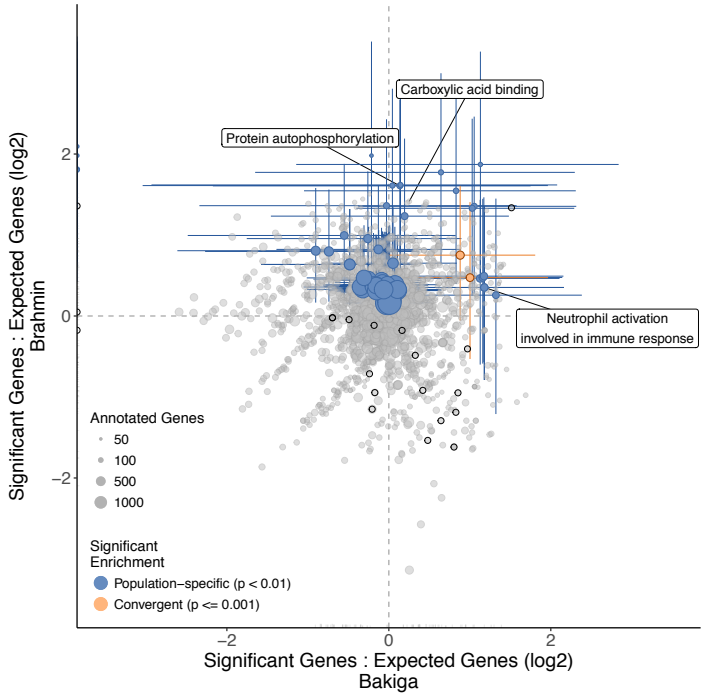


Figure S6: (Continued on the following page.)

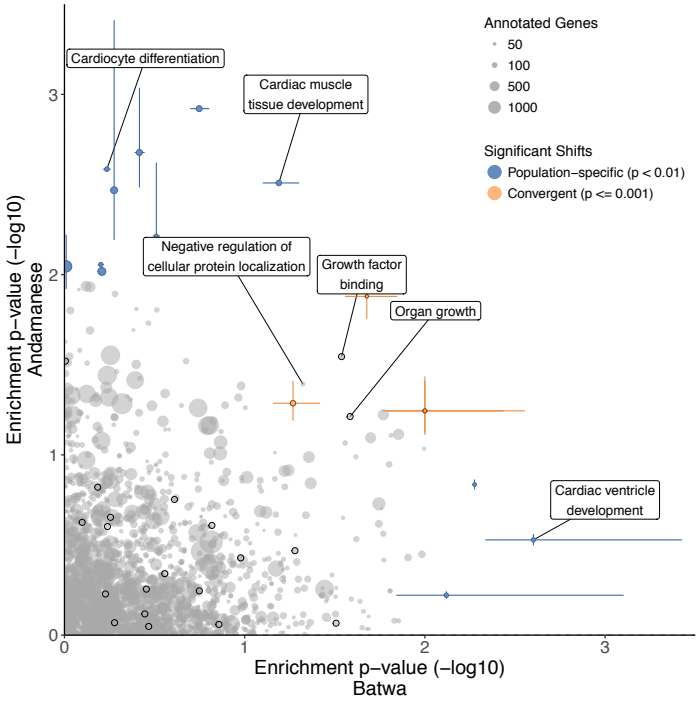
bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure S6: After MAF-based correction, Gene Ontology (GO) functional categories' ratios of expected to observed counts of outlier genes (with PBS selection index < 0.01) in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly overrepresented for genes under positive selection (Fisher $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S7: Gene size-corrected polygenic distribution shift test results.

A. Rainforest hunter-gatherers



B. Agriculturalists

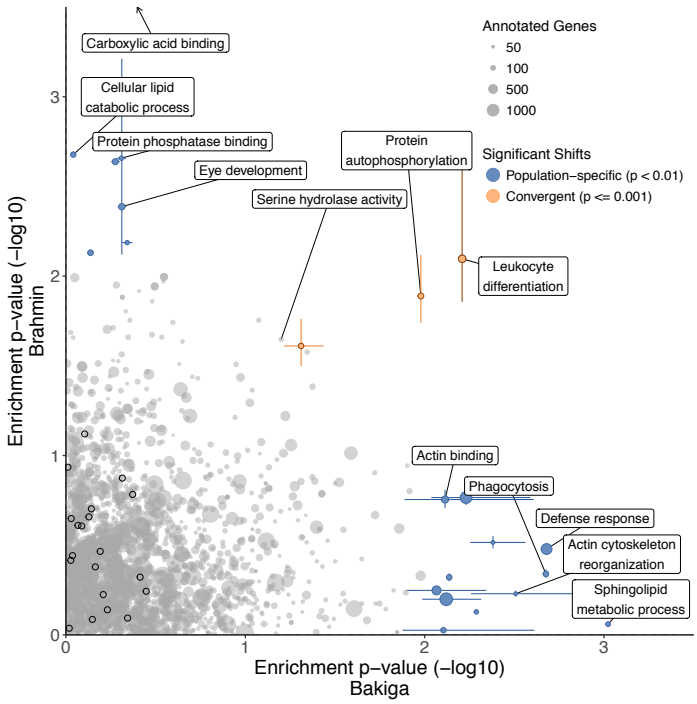


Figure S7: (Continued on the following page.)

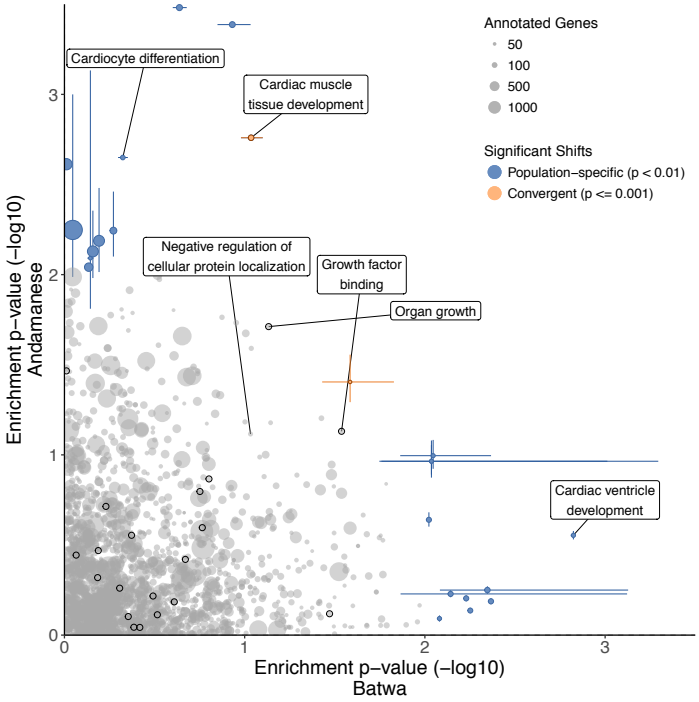
bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure S7: After gene size-based correction, Gene Ontology (GO) functional categories' distribution shift test p-values, indicating a shift in the PBS selection index values for genes, in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly enriched for genes under positive selection (Kolmogorov-Smirnov $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories. One GO molecular function, "carboxylic acid binding" (GO:0031406; Brahmin $p = 7.3 \times 10^{-5}$; $q = 0.0157$) not shown.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Fig. S8: Gene size-corrected polygenic distribution shift test results.

A. Rainforest hunter-gatherers



B. Agriculturalists

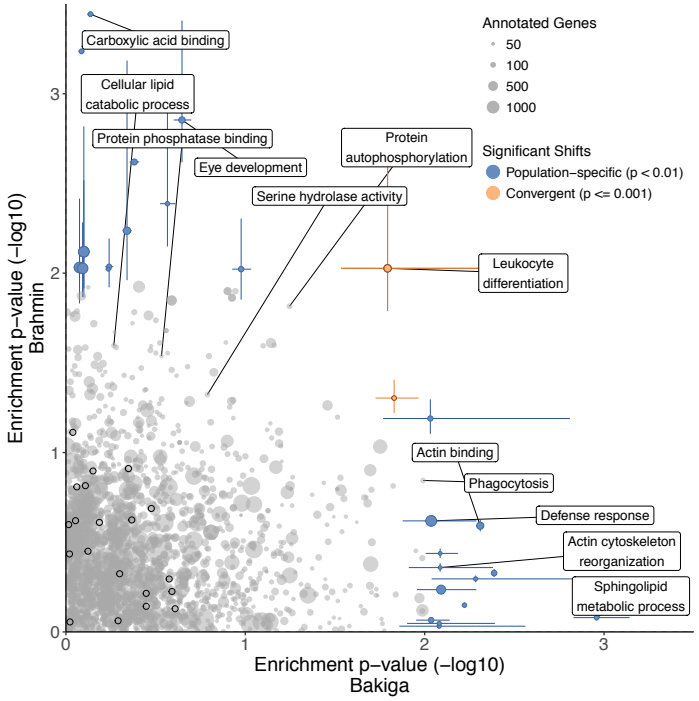


Figure S8: (Continued on the following page.)

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure S8: After MAF-based correction, Gene Ontology (GO) functional categories' distribution shift test p-values, indicating a shift in the PBS selection index values for genes, in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly enriched for genes under positive selection (Kolmogorov-Smirnov $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their sub-categories.

3 Tables

Table S1: Gene Ontology (GO) biological processes with evidence of convergent enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. No molecular functions were found to be convergently enriched. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

			Batwa:				*Andamanese:*			
	GO Biological Process	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0030326	embryonic limb morphogenesis	0.000	1.47	4	0.0584	1	2.01	6	0.0147	0.901
GO:0035107	appendage morphogenesis	0.000	1.69	5	0.0267	1	2.27	6	0.0254	0.901
GO:0035108	limb morphogenesis	0.000	1.69	5	0.0267	1	2.27	6	0.0254	0.901
GO:0035113	embryonic appendage morphogenesis	0.000	1.47	4	0.0584	1	2.01	6	0.0147	0.901
GO:0048736	appendage development	0.001	1.95	5	0.0448	1	2.62	6	0.047	1.000
GO:0060173	limb development	0.001	1.95	5	0.0448	1	2.62	6	0.047	1.000
GO:0030048	actin filament-based movement	0.002	1.72	4	0.0927	1	2.33	5	0.0834	1.000
GO:0048705	skeletal system morphogenesis	0.002	2.20	5	0.0688	1	2.95	6	0.0743	1.000
GO:0007034	vacuolar transport	0.003	1.37	4	0.0470	1	1.75	4	0.0968	1.000

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S2: Gene Ontology (GO) biological processes with evidence of population-specific enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. Results with $p < 0.01$ shown.

GO		Exp.	Obs.	p	adj. p
Batwa RHG - Biological Processes:					
GO:0007517	muscle organ development	10	4.02	0.0069	0.708
GO:0045926	negative regulation of growth	7	2.48	0.0118	0.708
Andamanese RHG - Biological Processes:					
GO:0006302	double-strand break repair	10	3.14	0.0011	0.171
GO:0070085	glycosylation	12	4.34	0.0013	0.171
GO:0000723	telomere maintenance	8	2.30	0.0020	0.175
GO:0033365	protein localization to organelle	25	14.09	0.0036	0.189
GO:1903827	regulation of cellular protein localization	19	9.66	0.0036	0.189
GO:0007569	cell aging	6	1.62	0.0052	0.208
GO:0009101	glycoprotein biosynthetic process	12	5.28	0.0065	0.208
GO:0034613	cellular protein localization	41	27.87	0.0067	0.208
GO:0051179	localization	116	97.30	0.0079	0.208
GO:0060249	anatomical structure homeostasis	13	6.09	0.0079	0.208
GO:0045596	negative regulation of cell differentiation	18	9.79	0.0090	0.215
Batwa RHG - Molecular Functions:					
GO:0003723	RNA binding	26	17.66	0.028	0.732
GO:0043167	ion binding	83	70.68	0.034	0.732
Andamanese RHG - Molecular Functions:					
GO:0043130	ubiquitin binding	7	1.89	0.0026	0.177
GO:0008233	peptidase activity	17	9.58	0.0153	0.383

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S3: Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of convergent distribution shifts in PBS selection index values in the hunter-gatherer populations. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

				Batwa:		*Andamanese:*	
		GO	Joint p	p	adj. p	p	adj. p
BP	GO:0035265	organ growth	0.001	0.0275	0.997	0.04509	1.000
	GO:0048738	cardiac muscle tissue development	0.001	0.0461	0.997	0.00265	1.000
	GO:1903828	negative regulation of cellular protein localization	0.001	0.0360	0.997	0.04275	1.000
	GO:0016202	regulation of striated muscle tissue development	0.002	0.0135	0.997	0.04406	1.000
	GO:1901861	regulation of muscle tissue development	0.002	0.0135	0.997	0.04406	1.000
	GO:0045444	fat cell differentiation	0.004	0.0573	0.997	0.04058	1.000
MF	GO:0019199	transmembrane receptor protein kinase activity	0.000	0.027	0.817	0.0261	0.784
	GO:0019838	growth factor binding	0.000	0.021	0.817	0.0269	0.784
	GO:0032559	adenyl ribonucleotide binding	0.003	0.020	0.817	0.0579	0.877
	GO:0030554	adenyl nucleotide binding	0.004	0.017	0.817	0.0755	0.877

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S4: Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of population-specific distribution shifts in PBS selection index values in the hunter-gatherer populations. No molecular functions were found to be significantly shifted for the Batwa. Results with $p < 0.01$ are shown.

GO		p	adj. p
Batwa RHG - Biological Processes:			
GO:0003231	cardiac ventricle development	0.001	0.302
GO:0061351	neural precursor cell proliferation	0.007	0.348
GO:0034976	response to endoplasmic reticulum stress	0.009	0.348
Andamanese RHG - Biological Processes:			
GO:0016579	protein deubiquitination	0.001	0.232
GO:0035051	cardiocyte differentiation	0.002	0.232
GO:0048738	cardiac muscle tissue development	0.003	0.232
GO:1901800	positive regulation of proteasomal protein catabolic process	0.004	0.262
GO:0006508	proteolysis	0.009	0.453
Andamanese RHG - Molecular Functions:			
GO:0005085	guanyl-nucleotide exchange factor activity	0.005	0.278

Table S5: After gene size-based correction, Gene Ontology (GO) biological processes and molecular functions with evidence of convergent enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

			Batwa:				*Andamanese:*			
	GO Biological Process	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0035107	appendage morphogenesis	0.000	1.77	5	0.0315	1	2.28	6	0.0258	0.966
GO:0035108	limb morphogenesis	0.000	1.77	5	0.0315	1	2.28	6	0.0258	0.966
GO:0048736	appendage development	0.000	2.04	5	0.0524	1	2.64	6	0.0478	1.000
GO:0060173	limb development	0.000	2.04	5	0.0524	1	2.64	6	0.0478	1.000
GO:1901617	organic hydroxy compound biosynthetic process	0.000	2.38	6	0.0314	1	3.19	7	0.0401	0.980
GO:0030326	embryonic limb morphogenesis	0.001	1.53	4	0.0665	1	2.02	6	0.0150	0.966
GO:0035113	embryonic appendage morphogenesis	0.001	1.53	4	0.0665	1	2.02	6	0.0150	0.966
GO:0030048	actin filament-based movement	0.002	1.80	6	0.0089	1	2.34	5	0.0845	1.000
GO:0007034	vacuolar transport	0.004	1.43	4	0.0537	1	1.76	4	0.0979	1.000
			Batwa:				*Andamanese:*			
	GO Molecular Function	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0008514	organic anion transmembrane transporter activity	0.000	2.78	7	0.0212	0.6853	3.37	7	0.0515	0.902
GO:0015081	sodium ion transmembrane transporter activity	0.000	2.31	7	0.0081	0.6853	2.93	6	0.0726	0.902

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S6: After MAF-based correction, Gene Ontology (GO) biological processes and molecular functions with evidence of convergent enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

			Batwa:				*Andamanese:*			
	GO Biological Process	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0048522	positive regulation of cellular process	0.000	55.57	66	0.0534	1	87.93	106	0.01153	0.946
GO:1901617	organic hydroxy compound biosynthetic process	0.000	2.29	6	0.0267	1	3.66	9	0.01069	0.946
GO:0006302	double-strand break repair	0.001	2.26	5	0.0757	1	3.62	8	0.02808	0.946
GO:0006897	endocytosis	0.001	8.49	14	0.0447	1	13.76	22	0.01948	0.946
GO:0048738	cardiac muscle tissue development	0.001	2.52	6	0.0399	1	4.01	8	0.04694	0.946
GO:0080135	regulation of cellular response to stress	0.001	7.63	14	0.0203	1	11.95	20	0.01624	0.946
GO:0014706	striated muscle tissue development	0.002	4.07	8	0.0509	1	6.55	14	0.00585	0.946
GO:0070302	regulation of stress-activated protein kinase signaling cascade	0.002	2.52	6	0.0399	1	3.93	7	0.09815	0.946
GO:1901615	organic hydroxy compound metabolic process	0.002	5.70	9	0.1168	1	8.87	14	0.06046	0.946
GO:2001020	regulation of response to DNA damage stimulus	0.002	2.09	5	0.0572	1	3.35	7	0.04997	0.946
GO:0051592	response to calcium ion	0.003	1.75	6	0.0079	1	2.74	5	0.13789	0.946
GO:0015718	monocarboxylic acid transport	0.004	1.63	4	0.0793	1	2.39	5	0.08980	0.946
GO:0030048	actin filament-based movement	0.004	1.73	4	0.0942	1	2.78	7	0.02039	0.946
GO:0030326	embryonic limb morphogenesis	0.004	1.48	4	0.0594	1	2.31	5	0.08052	0.946
GO:0031098	stress-activated protein kinase signaling	0.004	3.15	7	0.0383	1	4.97	8	0.12454	0.946
GO:0035113	embryonic appendage morphogenesis	0.004	1.48	4	0.0594	1	2.31	5	0.08052	0.946
GO:0060537	muscle tissue development	0.004	4.30	8	0.0660	1	6.90	14	0.00910	0.946
			Batwa:				*Andamanese:*			
	GO Molecular Function	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0008514	organic anion transmembrane transporter activity	0.002	2.68	6	0.051	0.798	4.03	10	0.0068	0.959
GO:0005342	organic acid transmembrane transporter activity	0.003	1.97	5	0.046	0.798	2.96	7	0.0282	1.000
GO:0015081	sodium ion transmembrane transporter activity	0.003	2.22	5	0.071	0.821	3.50	7	0.0600	1.000
GO:0046943	carboxylic acid transmembrane transporter activity	0.004	1.79	4	0.103	0.964	2.70	7	0.0177	1.000

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S7: After gene size-based correction, Gene Ontology (GO) biological processes with evidence of population-specific enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. Results with $p < 0.01$ shown.

GO		Exp.	Obs.	p	adj. p
Batwa RHG - Biological Processes:					
GO:0007517	muscle organ development	11	4.20	0.0032	0.5650
GO:1903825	organic acid transmembrane transport	6	1.67	0.0061	0.5652
GO:0030048	actin filament-based movement	6	1.80	0.0061	0.5652
Andamanese RHG - Biological Processes:					
GO:0006302	double-strand break repair	10	3.16	0.0012	0.231
GO:0000723	telomere maintenance	8	2.31	0.0020	0.231
GO:1903827	regulation of cellular protein localization	19	9.70	0.0038	0.231
GO:0070085	glycosylation	11	4.36	0.0042	0.231
GO:1900180	regulation of protein localization to nucleus	10	3.77	0.0044	0.231
GO:0007569	cell aging	6	1.63	0.0053	0.232
GO:0060249	anatomical structure homeostasis	13	6.12	0.0082	0.299
GO:0051234	establishment of localization	99	81.24	0.0091	0.299
Batwa RHG - Molecular Functions:					
GO:0015081	sodium ion transmembrane transporter activity	7	2.31	0.0081	0.33245
Andamanese RHG - Molecular Functions:					
GO:0043130	ubiquitin binding	7	1.89	0.0026	0.177

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S8: After MAF-based correction, Gene Ontology (GO) biological processes with evidence of population-specific enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. No molecular functions were found to be significantly shifted for the Batwa. Results with $p < 0.01$ shown.

GO		Exp.	Obs.	p	adj. p
Batwa RHG - Biological Processes:					
GO:0007517	muscle organ development	11	4.04	0.0023	0.492
GO:0051592	response to calcium ion	6	1.75	0.0079	0.492
Andamanese RHG - Biological Processes:					
GO:0051179	localization	142	114.34	0.00044	0.115
GO:0045596	negative regulation of cell differentiation	22	11.53	0.0026	0.266
GO:0071229	cellular response to acid chemical	9	3.24	0.0048	0.266
GO:0002460	adaptive immune response based on somatic recombination...	11	4.47	0.0050	0.266
GO:0014706	striated muscle tissue development	14	6.55	0.0059	0.266
GO:0048584	positive regulation of response to stimulus	53	38.05	0.0067	0.266
GO:0016337	single organismal cell-cell adhesion	22	12.61	0.0076	0.266
Andamanese RHG - Molecular Functions:					
GO:0043130	ubiquitin binding	7	2.24	0.0067	0.221
GO:0008514	organic anion transmembrane transporter activity	10	4.03	0.0068	0.221

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S9: After gene size-based correction, Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of convergent distribution shifts in PBS selection index values in the hunter-gatherer populations. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

				Batwa:		*Andamanese:*	
		GO	Joint p	p	adj. p	p	adj. p
BP	GO:0016202	regulation of striated muscle tissue development	0.000	0.0100	0.994	0.0570	1.000
	GO:1901861	regulation of muscle tissue development	0.000	0.0100	0.994	0.0570	1.000
	GO:0045444	fat cell differentiation	0.001	0.0539	0.994	0.0517	1.000
	GO:0048634	regulation of muscle organ development	0.002	0.0101	0.994	0.0924	1.000
	GO:0035265	organ growth	0.003	0.0260	0.994	0.0613	1.000
	GO:0048738	cardiac muscle tissue development	0.003	0.0646	0.994	0.0031	1.000
	GO:0051147	regulation of muscle cell differentiation	0.003	0.0242	0.994	0.1026	1.000
	GO:1903828	negative regulation of cellular protein localization	0.003	0.0475	0.994	0.0405	1.000
	GO:0046434	organophosphate catabolic process	0.004	0.0154	0.994	0.0780	1.000
MF	GO:0019199	transmembrane receptor protein kinase activity	0.000	0.021	0.698	0.0132	0.736
	GO:0019838	growth factor binding	0.002	0.029	0.750	0.0285	0.736
	GO:0030554	adenyl nucleotide binding	0.002	0.014	0.698	0.0769	0.877
	GO:0032559	adenyl ribonucleotide binding	0.002	0.017	0.698	0.0599	0.877

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S10: After MAF-based correction, Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of convergent distribution shifts in PBS selection index values in the hunter-gatherer populations. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

				Batwa:		*Andamanese:*	
		GO	Joint p	p	adj. p	p	adj. p
BP	GO:0048738	cardiac muscle tissue development	0.000	0.0921	0.993	0.00174	0.878
	GO:0033002	muscle cell proliferation	0.002	0.0919	0.993	0.02565	1.000
	GO:0035265	organ growth	0.003	0.0736	0.993	0.01943	1.000
	GO:0003007	heart morphogenesis	0.004	0.0375	0.993	0.06262	1.000
	GO:0016579	protein deubiquitination	0.004	0.1171	0.993	0.00041	0.369
MF	GO:0019199	transmembrane receptor protein kinase activity	0.001	0.026	0.675	0.039	0.874

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S11: After gene size-based correction, Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of population-specific distribution shifts in PBS selection index values in the hunter-gatherer populations. No molecular functions were found to be significantly shifted for the Batwa. Results with $p < 0.01$ are shown.

GO		p	adj. p
Batwa RHG - Biological Processes:			
GO:0003231	cardiac ventricle development	0.0025	0.371
GO:0061351	neural precursor cell proliferation	0.0080	0.371
Andamanese RHG - Biological Processes:			
GO:0016579	protein deubiquitination	0.001	0.273
GO:0035051	cardiocyte differentiation	0.003	0.273
GO:0048738	cardiac muscle tissue development	0.003	0.273
GO:1901800	positive regulation of proteasomal protein catabolic process	0.006	0.322
GO:0001936	regulation of endothelial cell proliferation	0.006	0.322
GO:0006508	proteolysis	0.009	0.396
Andamanese RHG - Molecular Functions:			
GO:0005085	guanyl-nucleotide exchange factor activity	0.0034	0.224
GO:0008134	transcription factor binding	0.0096	0.224

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S12: After MAF-based correction, Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of population-specific distribution shifts in PBS selection index values in the hunter-gatherer populations. No molecular functions were found to be significantly shifted for the Batwa. Results with $p < 0.01$ are shown.

	GO	p	adj. p
Batwa RHG - Biological Processes:			
GO:0003231	cardiac ventricle development	0.0015	0.346
GO:0034284	response to monosaccharide	0.0043	0.346
GO:0008217	regulation of blood pressure	0.0056	0.346
GO:0050864	regulation of B cell activation	0.0083	0.346
GO:0048634	regulation of muscle organ development	0.0090	0.346
GO:1901861	regulation of muscle tissue development	0.0092	0.346
Andamanese RHG - Biological Processes:			
GO:0070646	protein modification by small protein removal	0.0003	0.087
GO:0048738	cardiac muscle tissue development	0.0017	0.160
GO:0035051	cardiocyte differentiation	0.0022	0.160
GO:0006508	proteolysis	0.0024	0.156
GO:0071840	cellular component organization or biogenesis	0.0057	0.283
GO:0007155	cell adhesion	0.0065	0.283
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	0.0091	0.298
Andamanese RHG - Molecular Functions:			
GO:0005085	guanyl-nucleotide exchange factor activity	0.0057	0.198
GO:0019783	ubiquitin-like protein-specific protease activity	0.0081	0.198

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S13: Comparison of results of two methods for computing empirical test for convergence in strong outlier selection in both the Batwa and Andamanese RHGs. In the original method, genes and PBS selection index values are permuted to create an empirical null distribution. In the modified case, genes and their Gene Ontology (GO) annotations are instead permuted to create the null distribution. Biological processes (BP) with empirical test for convergence $p < 0.005$ in either method shown. No molecular functions were found to be significantly convergently enriched in both RHG populations.

	GO Biological Process	Original convergence p	Modified convergence p
GO:0035107	appendage morphogenesis	0.000	0.000
GO:0035108	limb morphogenesis	0.000	0.000
GO:0030326	embryonic limb morphogenesis	0.000	0.002
GO:0035113	embryonic appendage morphogenesis	0.000	0.002
GO:0048736	appendage development	0.001	0.003
GO:0060173	limb development	0.001	0.003
GO:0048705	skeletal system morphogenesis	0.002	0.003
GO:0048522	positive regulation of cellular process	0.006	0.003
GO:0080135	regulation of cellular response to stress	0.018	0.004
GO:0030048	actin filament-based movement	0.002	0.006
GO:0007034	vacuolar transport	0.003	-

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table S14: Comparison of results of two methods for computing empirical test for convergence in PBS selection index shift in both the Batwa and Andamanese RHGs. In the original method, genes and PBS selection index values are permuted to create an empirical null distribution. In the modified case, genes and their Gene Ontology (GO) annotations are instead permuted to create the null distribution. Biological processes (BP) and molecular functions (MF) with empirical test for convergence $p < 0.005$ in either method shown.

	GO Biological Process	Original convergence p	Modified convergence p
GO:0016202	regulation of striated muscle tissue development	0.002	0.000
GO:1901861	regulation of muscle tissue development	0.002	0.000
GO:0048738	cardiac muscle tissue development	0.001	0.001
GO:0048634	regulation of muscle organ development	0.005	0.001
GO:0045444	fat cell differentiation	0.004	0.002
GO:0003007	heart morphogenesis	0.006	0.002
GO:0035265	organ growth	0.001	0.004
GO:0046434	organophosphate catabolic process	0.007	0.004
GO:1903828	negative regulation of cellular protein localization	0.001	0.006
	GO Molecular Function	Original convergence p	Modified convergence p
GO:0019838	growth factor binding	0.000	0.001
GO:0019199	transmembrane receptor protein kinase activity	0.000	0.002
GO:0032559	adenyl ribonucleotide binding	0.003	0.004
GO:0030554	adenyl nucleotide binding	0.004	0.004
GO:0005524	ATP binding	0.005	0.004

bioRxiv preprint first posted online Apr. 13, 2018; doi: http://dx.doi.org/10.1101/300574. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

References

[1] Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723–D729 (2017).

[2] Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).

[3] Lui, J. C. *et al.* Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Human Molecular Genetics* **21**, 5193–5201 (2012).

[4] Perry, G. H. *et al.* Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences* **111**, E3596–E3603 (2014).

[5] Brown, K. R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082 (2005).