Université de Montréal

Impact des variants génétiques sur la réponse immunitaire des populations humaines

par Yohann Nédélec

Bio-informatique

Faculté de médecine

Thèse présentée

en vue de l'obtention du grade de docteur

en bio-informatique

Juin 2017

Résumé

Les mécanismes impliqués dans les variations de susceptibilité aux infections bactériennes entre individus et populations demeurent méconnus. Durant mon doctorat, j'ai mis à jour l'existence de profondes différences dans la réponse immunitaire entre individus d'origine Africaine et Européenne. Nous avons observé que 34% des gènes exprimés par les macrophages montrent au moins un type de différence transcriptionnelle entre populations : au niveau de l'expression totale du gène (30%), à sa réponse lors de l'infection (9.3%) ou, plus rarement, au niveau des proportions d'isoformes (1%). Les individus d'origine Africaine semblent disposer d'une réponse inflammatoire exacerbée ainsi que d'une capacité plus importante à détruire les bactéries intracellulaires. Une large proportion de ces différences sont sous le contrôle de variants génétiques : pour 804 gènes, plus de 75% des différences d'expression entre populations peuvent être expliquées par un unique variant génétique en cis ou en trans. Parmi ces variants, j'ai identifié un enrichissement en signatures de sélection naturelle suggérant que des pathogènes ont exercé des pressions sur le génome de notre espèce et ont ainsi participé à l'adaptation de notre espèce à son environnement. Notre métissage avec l'homme de Néandertal a également joué un rôle important dans l'adaptation des populations non-Africaines aux nouveaux environnements pathogéniques auxquels l'homme moderne a été confronté lorsqu'il a quitté l'Afrique pour coloniser le reste du monde.

Mots-clés : immunité, infection, génétique des populations, génomique, séquençage, variants génétiques, médecine de précision

Abstract

The mechanisms that explain inter-individual and population differences in susceptibility to infectious diseases remain poorly understood. During my PhD, I uncovered important variabilities in the immune response to pathogens between individuals of African and European descent. We observed that 34% of genes expressed in macrophages show at least one type of ancestry-related transcriptional divergence, whether in the form of differences in gene expression (30%), the transcriptional response to infection (9.3%), or, less commonly, differences in isoform usage (1%). African ancestry also specifically predicts a stronger inflammatory response and reduced intracellular bacterial growth. A large proportion of these differences are under genetic control: for 804 genes, more than 75% of ancestry effects on the immune response can be explained by a single cis- or trans-acting expression quantitative trait locus (eQTL). Among those variants, I identified several signatures of natural selection suggesting that pathogens were involved in the adaptation of our species to fluctuating pathogenic realities, contributing therefore to the recent evolution of our immune system. I also showed that admixture between our species and Neandertals played a significant role in the adaptation of non-African populations to the new pathogenic environments modern humans had to face when leaving the Africa to colonize the globe.

Keywords : immunity, infection, population genetics, genomics, sequencing, genetic variants, precision medicine

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1

Chapitre 2

Article

Chapitre 3

Chapitre 4

Chapitre 5

Liste des sigles

ADN : Acide Désoxyribonucléique

API : Application Programming Interface

ARN : Acide Ribonucléique

ASE : Alternative Splicing Event

asQTL : alternative splicing Quantitative Trait Loci

CRP : C-Reactive Protein

eQTL : expression Quantitative Trait Loci

FST : F-statistic (Fixation index)

GWAS : Genome Wide Association Study

HTML5 : HyperText Markup Langage 5

iHS : integrated Haplotype Score

IL6 : Interleukine 6

IL6R : Interleukine 6 Receptor

IPSC : Induced Pluripotent Stem Cell

LPS : Lipopolysaccharide

OAS : 2',5'-oligoadénylate synthase

OMS : Organisation Mondiale de la Santé

PAMP : Pathogen Associated Molecular Pattern

PBMC : Peripheral Blood Mononuclear Cell

PCR : Polymerase Chain Reaction

reQTL : response Quantitative Trait Loci

REST : REpresentational State Transfer

SNP : Single Nucleotide Polymorphism

SVG : Scalable Vector Graphics

TLR : Toll-Like Receptor

VIH : Virus de l'Immunodéficience Humaine

Remerciements

Je tiens à remercier chaleureusement toutes les personnes qui m'ont accompagné durant cette aventure que représente le doctorat, et en particulier :

- Luis Barreiro, mon directeur, qui m'a soutenu tout au long de cette thèse et éclairé de ses précieux conseils ;
- mon équipe, qui a fait preuve de convivialité, d'excellentes compétences scientifiques, mais également d'une résistance accrue à mon sens de l'humour ;
- Jean-Christophe Grenier, Vanessa & Philippe Bruat et Manon & Thibault de Malliard, pour leur amitié et leur présence ;
- Julie, mon épouse, pour m'avoir encouragé dans l'effort et apporté la sérénité ;
- que mes deux fils, Awen et Victor, n'hésitent pas à briser par leurs cris aigus ;
- mes parents, beaux-parents, ma sœur et mes nièces pour leur soutien dans ce périple ;
- mes plus proches amis: Anna, Audrey, Claire, Elise, Laetitia, Mary, Marie-Eve, Valérie, Alexis, Dadoo, David, Jérem, Yoann qui m'accompagnent depuis de longues années ; cher Docteur Alexis, félicitations pour ta thèse !
- Elaine Meunier, pour avoir gardé son calme face à ma phobie administrative ;
- les membres du jury pour avoir accepté d'évaluer mon travail ;
- le CHU Sainte-Justine, pour leur confiance et leur soutien, notamment en m'accordant la bourse de la fondation ; la Faculté des études supérieures et

postdoctorales, à travers une bourse d'excellence ; le Réseau de médecine génétique appliquée, en m'attribuant une bourse de formation ;

- Calcul Canada, en mettant à disposition des ressources informatiques indispensables à mes projets de recherche. Sans cela, mon travail aurait nécessité 75 années de calcul !

Chapitre 1 : Introduction

Les maladies infectieuses sont responsables de près de 15 millions de décès chaque année et représentent un challenge de santé publique à l'échelle mondiale. Notre système immunitaire est le fruit de longs processus d'évolution où la survie dépend de la capacité de l'organisme à se défendre face à des pathogènes eux-mêmes en constante adaptation.

Grâce à l'émergence des technologies de séquençage, de génotypage et la mise au point de méthodes analytiques novatrices, nous disposons de nouveaux moyens pour tenter d'élucider l'évolution du système immunitaire. La compréhension de ces mécanismes constitue un espoir de recherche décisif dans le cadre de notre combat contre les maladies infectieuses.

Enjeux de la réponse immunitaire

Le fonctionnement du système immunitaire repose essentiellement sur sa capacité à discerner les molécules de l'organisme de celles portées par un pathogène : virus, bactérie, champignon ou parasite.

La salmonellose et la listériose sont deux zoonoses d'origine bactérienne causées respectivement par les bactéries du genre *Salmonella et Listeria*.

La salmonellose se caractérise par des fièvres typhoïdes ou paratyphoïdes associées à des gastro-entérites à la suite de contaminations alimentaires. *Salmonella enterica* est l'une des deux espèces du genre *Salmonella* (*enterica* et *bongori*), elle-même subdivisée en six sous-espèces, dont la sous-espèce *enterica*. Au sein de cette sous-espèce, on distingue de nombreux sérotypes dont *Typhi*, *Paratyphi* et *Typhimurium*. *Salmonella Typhi* est responsable de la fièvre typhoïde et constitue un problème de santé public

majeur dans les pays en développement. L'OMS estime en effet que 21 millions de personnes chaque année présentent une fièvre typhoïde et que 220 000 d'entre elles décèderont de complications liées à la maladie. Les infections par *Salmonella Typhimurium* sont impliquées dans des septicémies chez les enfants de moins de 5 ans, avec des taux de mortalité supérieurs à 25% lorsque ces individus sont conjointement atteints du VIH (Gordon et al. 2008). Cette infection se révèle particulièrement inquiétante depuis la découverte de la souche ST313 en Afrique sub-saharienne. Comparativement à la souche ST19, plus commune et répandue, la souche ST313 est plus facilement phagocytée par les macrophages et montre une résistance accrue à la destruction qui lui permet de se répliquer activement au sein de la cellule (Ramachandran et al. 2015; Feasey et al. 2012; Kingsley et al. 2009).

La listériose est une infection bactérienne relativement rare mais redoutable qui se développe à la suite de la consommation de nourriture contaminée par *Listeria monocytogenes*. Aux États-Unis, 1600 personnes sont atteintes chaque année et environ 260 d'entre elles en mourront. La listériose est particulièrement préoccupante chez les femmes enceintes, nouveaux nés et personnes âgées ou avec un système immunitaire compromis.

Depuis la découverte de la pénicilline en 1928 par Alexander Fleming, les antibiotiques se révèlent être des alliés particulièrement utiles dans le traitement des infections bactériennes. Cependant, dans la perpétuelle course à l'adaptation qui nous oppose aux pathogènes, la recrudescence d'infections bactériennes multi-résistantes aux antibiotiques est un phénomène particulièrement préoccupant qui rend crucial la découverte de nouveaux traitements.

Classe d'antibiotique	Année d'introduction	Premières résistances

Sulfamides	1936	1940
Pénicilline G	1943	1946
Streptomycine	1943	1959
Chloramphénicol	1947	1959
Tétracycline	1948	1953
Erythromycine	1952	1988
Ampicilline	1961	1973
Ciprofloxacine	1987	1997

TABLEAU 1. GRANDES CLASSES D'ANTIBIOTIQUES ET PHENOMENES DE RESISTANCES

Dates issues de la littérature (Palumbi 2001; Jacoby 2005).

Au-delà des pathologies infectieuses, les maladies auto-immunes et certains cancers sont étroitement liés à des dysfonctionnements du système immunitaire : soit à travers une réaction disproportionnée ou, à l'inverse, un défaut d'activation. Le diabète de type 1, la sclérose en plaque, le lupus ou la maladie de Crohn sont quelques exemples de maladies attribuées au système immunitaire.

Dans le monde occidental, où les maladies infectieuses ont une importance plus limitée, les maladies inflammatoires et auto-immunes sont la troisième cause de mortalité, derrière le cancer et les affections cardio-vasculaires.

Il est judicieux d'étudier le système immunitaire à travers son histoire évolutive, pour tenter de comprendre les mécanismes que la nature a mis en place pour assurer la survie de l'homme face aux pathogènes.

Evolution de l'homme moderne

Ancêtres

Malgré les avancées dans le domaine de la paléontologie et paléogénétique, la reconstruction de l'histoire précise de l'émergence et de la progression de notre espèce demeure un challenge.

La divergence entre la lignée conduisant aux chimpanzés et celle de l'homme remonterai à 6 Ma (Patterson et al. 2006). Il y a 1.8 Ma, à l'occasion de la première sortie de l'Afrique, Homo erectus se serait établit à la frontière de l'Asie et de l'Europe, comme en témoignent cinq crânes mis à jour entre 1991 et 2005 dans la région de Dmanisi, en Géorgie (Gabunia et al. 2001). Ces découvertes furent d'abord associées à Homo ergaster, un des plus ancien représentant du genre Homo qui vivait en Afrique du sud et partagerait probablement un ancêtre commun avec Homo erectus.

Homo erectus est le premier représentant du genre homo disposant des mêmes caractéristiques que celles de l'homme moderne. Ainsi, Homo erectus était parfaitement adapté à la bipédie, pouvait probablement courir sur de longues distances et montrait des comportements sociaux particulièrement avancés. Homo erectus se distingue des autres représentants du genre homo à travers sa survie remarquablement étendue dans le temps puisqu'on estime qu'elle s'étend de 1.3 million à 200 milles années avant notre ère (Figure 1).

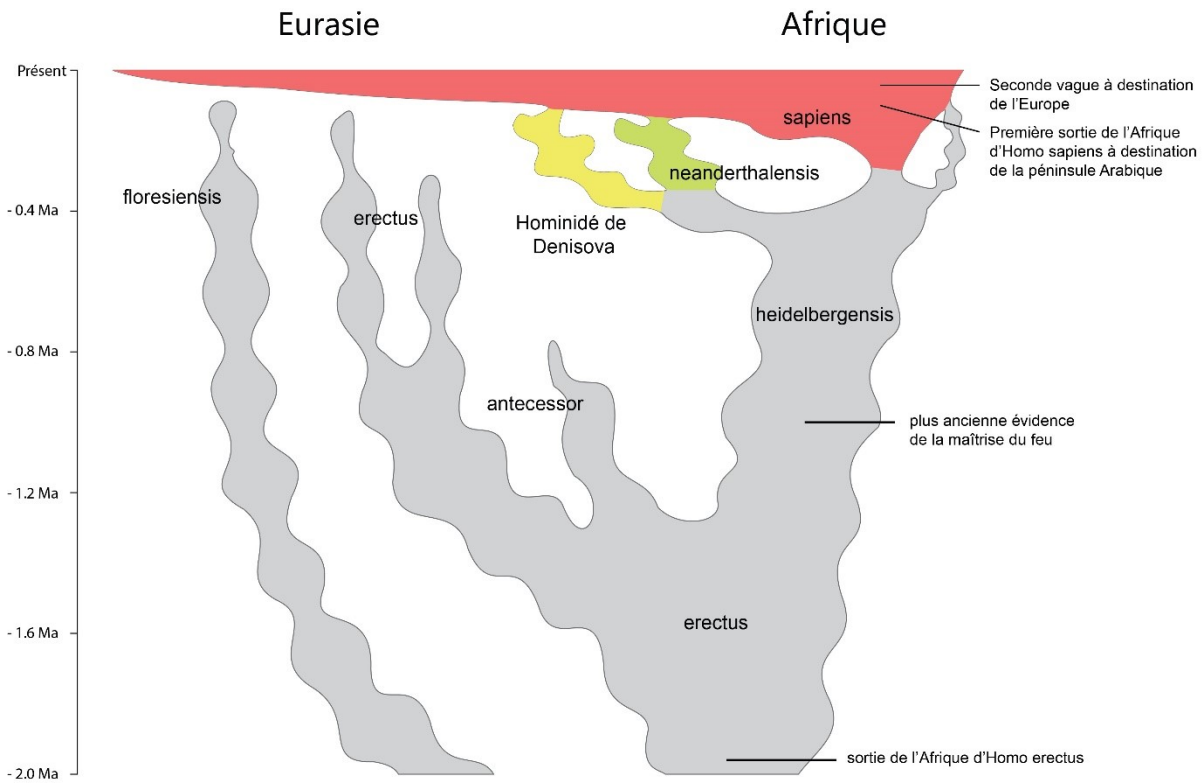

Figure 1. **REPARTITION DES REPRESENTANTS DU GENRE HOMO**

Reproduit et complété à partir de la littérature (Stringer 2012).

Des représentants d'Homo erectus en Afrique auraient évolué pour donner une nouvelle espèce : Homo heidelbergensis. Celle-ci semble être l'ancêtre commun entre Homo sapiens et deux espèces archaïques : Homo neanderthalensis et l'hominidé de Denisova.

Dispersions humaines

La reconstitution des migrations des populations hors de l'Afrique demeure un challenge important. Celle-ci s'appuie à la fois sur des études paléontologiques (à travers les interprétations d'échantillons fossiles), paléoclimatiques, mais également génomiques (où l'on cherche à retracer l'ascendance du prélèvement par rapport aux autres populations humaines).

La paléontologie s'intéresse à l'étude de fossiles des êtres vivants pour en comprendre l'évolution. Ainsi, deux crânes découverts en Ethiopie en 1967 permirent d'établir que les premières populations de l'homme moderne existaient déjà en Afrique il y a près de 200 milles ans (McDougall, Brown, et Fleagle 2005).

En ce qui concerne la sortie de l'Afrique, deux modèles s'opposent (Tucci et Akey 2016). Le premier modèle repose sur un seul et unique évènement qui aurait eu lieu entre 40 000 et 80 000 ans et aurai donné naissance à l'ensemble des populations non-Africaines (Mallick et al. 2016; Malaspinas et al. 2016). A l'opposé, un second scénario est proposé où une première migration, à l'origine des populations du sud-est de l'Asie et d'Océanie, se serait déroulée il y a environ 130 000 ans. Une deuxième migration aurait suivi pour entreprendre le peuplement de l'Eurasie.

Des modifications du climat ont probablement joué un rôle crucial dans la migration des populations humaines hors de l'Afrique. Ainsi, de récentes études en paléoclimatologie ont mis au point des simulations complexes dans le but de retracer le climat qui régnait en Afrique et dans la péninsule arabique au cours des 125 000 dernières années. L'objectif principal était de comprendre si la situation était propice à une unique ou à plusieurs vagues de migrations (deMenocal et Stringer 2016; Tucci et Akey 2016).

De nos jours, les déserts du Sahara et Arabique constituent des barrières infranchissables pour des groupes d'individus souhaitant s'établir en dehors de l'Afrique. Les résultats étayent l'hypothèse selon laquelle plusieurs évènements de sortie d'Homo sapiens auraient existé puisque plusieurs fenêtres de conditions favorables ont été identifiées entre -100 000 et -50 000 ans. A ces périodes, le climat dans l'ouest de l'Afrique et dans la péninsule arabique était plus humide et permis l'apparition de

véritables couloirs propices aux migrations où la végétation et les ressources étaient abondantes (Figure 2).

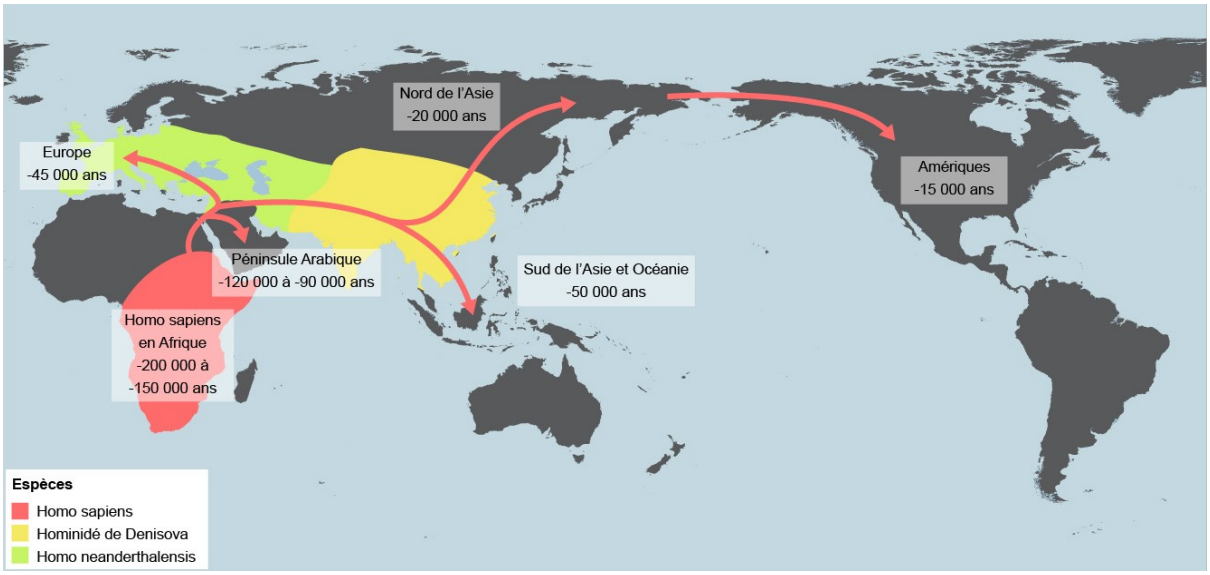

Figure 2. MIGRATIONS D'HOMO SAPIENS ET REPARTITIONS GEOGRAPHIQUES D'ESPECES ARCHAÏQUES DU GENRE HOMO.

Les récentes améliorations dans les processus de collecte, traitement et séquençage des traces d'ADN à partir de fossiles ont participé à la reconstruction de l'histoire de l'homme moderne (Figure 3). Ces méthodes reposent principalement sur les études de l'ADN mitochondrial, celui du chromosome Y ou nucléaire et permettent d'élucider la parenté des populations humaines actuelles, mais également de celles d'échantillons fossiles.

Introgression avec l'homme de Néandertal et l'hominidé de Denisova

Homo neanderthalensis et l'hominidé de Denisova sont deux représentants archaïques du genre homo qui résidaient respectivement en Europe et en Asie plusieurs centaines de milliers d'années avant l'arrivée d'Homo sapiens.

En 2010, l'analyse d'un génome partiel d'Homo neanderthalensis confirme l'existence d'un métissage avec Homo sapiens (Green et al. 2010). Mais l'obtention d'échantillons dans une grotte au sud de la Sibérie, puis leur séquençage a permis d'explorer sans précédents l'origine et l'apparenté de l'homme moderne avec deux

espèces archaïques : Homo neanderthalensis (Prüfer et al. 2014) et l'hominidé de Denisova (qui ne dispose pas encore de nom scientifique) (M. Meyer et al. 2012).

Ce métissage entre Homo sapiens et Homo neanderthalensis aurait eu lieu peu de temps après la sortie de l'Afrique de notre espèce et concernerait ainsi uniquement les populations non-africaines à hauteur de 1.5 à 2.1% de leur génome. Les populations d'Asie seraient en outre le fruit d'un métissage particulièrement important avec l'hominidé de Denisova puisqu'on estime que près de 5% de leur génome proviendrait de cet ancêtre (Sankararaman et al. 2016).

La Figure 3 résume les hypothèses actuelles quant aux flux de gènes de plusieurs espèces au sein du genre Homo.

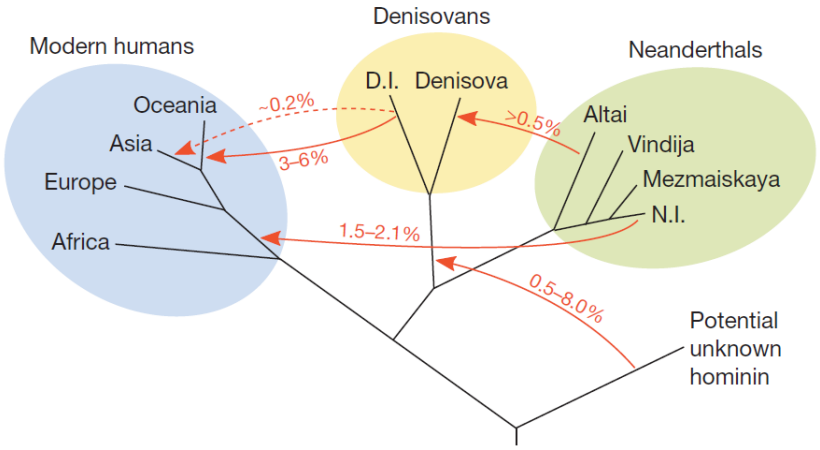

Figure 3. **UN MODELE POSSIBLE DES FLUX DE GENES A LA FIN DU PLEISTOCENE**

"The direction and estimated magnitude of inferred gene flow events are shown. Branch lengths and timing of gene flows are not drawn to scale. The dashed line indicates that it is uncertain if Denisovan gene flow into modern humans in mainland Asia occurred directly or via Oceania. D.I. denotes the introgressing Denisovan, N.I. the introgressing Neanderthal. Note that the age of the archaic genomes precludes detection of gene-flow from modern humans into the archaic hominins."

Figure et légende issues de la littérature (Prüfer et al. 2014)

Variants génétiques

Les variants génétiques sont des modifications de l'ADN qui apparaissent aléatoirement à l'occasion de mutations.

A l'heure actuelle, environ 88 millions de variants génétiques ont été caractérisés dans le génome humain (Auton et al. 2015). Parmi ceux-ci, on distingue : 84.7M de polymorphismes nucléotidiques (affectant un seul nucléotide); 3.6M de courtes insertions ou délétions; et 60 000 variants structuraux (évènements d'insertions, délétions, inversions, translocations ou duplications s'étendant sur 1Kb à 3Mb) (Feuk, Carson, et Scherer 2006).

Dans le cadre de mes travaux, je me suis intéressé plus particulièrement aux polymorphismes nucléotidiques (SNP) car : ceux-ci représentent la majorité des variants génétiques (>95%) du génome humain ; leurs rôles dans de nombreuses pathologies a été établi à travers des études d'associations (Figure 4); et nous disposons de technologies adaptées pour les caractériser.

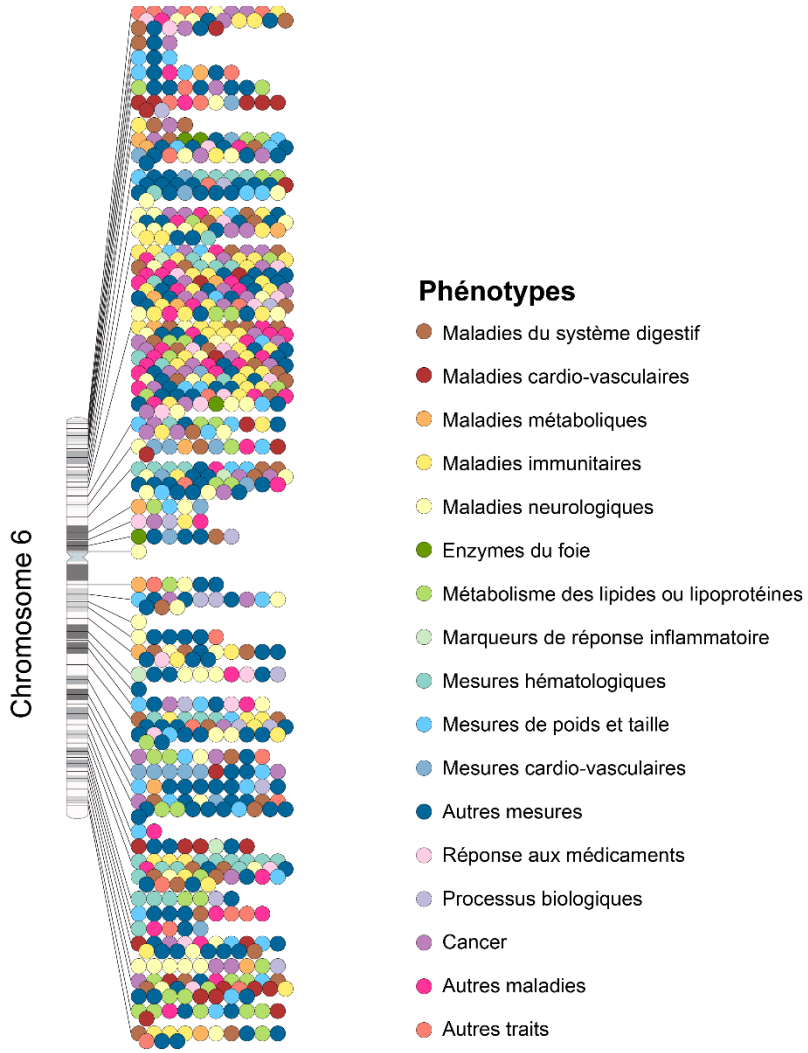

Figure 4. **TRAITS ASSOCIES A DES VARIANTS GENETIQUES SUR LE CHROMOSOME 6.**

Version interactive de l'ensemble des chromosomes disponible sur https://www.ebi.ac.uk/gwas/diagram

Les puces de génotypages permettent de déterminer rapidement le génotype de plusieurs millions de marqueurs prédéfinis (SNPs principalement). A ce titre, il est nécessaire de sélectionner parmi les variants ceux qui présentent un intérêt avéré ou potentiel, c'est-à-dire qui sont connus comme étant impliqués dans des maladies ou susceptibles d'affecter l'expression de gènes. On cherche également à constituer une liste de variants répartis sur l'ensemble du génome pour disposer d'une vision globale des polymorphismes.

Le fonctionnement des puces de génotypage repose sur l'hybridation de petits fragments d'ADN avec des séquences complémentaires fixées sur la puce. Une ADN polymérase complète le brin opposé en utilisant un nucléotide fluorescent que le scanner est en mesure de détecter.

Dans le cadre des analyses d'association, où l'on cherche à mettre en évidence un lien entre un polymorphisme et une maladie ou l'expression d'un gène, il peut être intéressant de disposer d'un plus grand nombre de variants génotypés. En effet, à partir d'un set plus conséquent de SNPs, il est possible d'identifier de nouvelles associations, mais également d'affiner nos analyses pour nous rapprocher du variant causal. On a souvent recourt à des méthodes d'imputations reposant sur l'inférence statistique qui permettent de déterminer l'état de certains SNPs non observés à partir de données de références, telles que celles collectées par « The 1000 Genomes Project » (Durbin et al. 2010).

Dérive génétique

La fréquence d'un variant génétique au sein d'une population fluctue aléatoirement sous l'effet du hasard de la rencontre des gamètes lors de la reproduction sexuée. Au bout d'un certain nombre de générations, il est susceptible d'être "fixé" : seul cet allèle subsiste ; ou est éradiqué : plus aucun individu n'est porteur de la mutation. Ses effets sur la diversité génétique d'une population sont particulièrement remarqués à l'occasion d'un goulet d'étranglement (bottleneck), d'une différenciation (split) ou d'une expansion.

Lors d'un goulet d'étranglement, seul un groupe d'individu est échantillonné à partir de la population d'origine. Cette réduction de la taille de la population est susceptible d'entrainer une diminution de la variabilité génétique : les variants

génétiques préservés ne reflètent pas nécessairement la structure génétique de la population originelle. A la suite de cet évènement, les effets de la dérive génétique sont plus importants : un variant, même délétère à la survie, peut subsister et être fixé au sein de la population. Les migrations, catastrophes naturelles, épidémies massives ou génocides sont des exemples de goulets d'étranglement qui, en sélectionnant une partie de la population, en diminuent sa variabilité génétique. La population québécoise porte les stigmates de cet effet fondateur car celle-ci s'est développée depuis le XIV^{ème} siècle à partir d'un nombre restreint d'individus provenant de France. Une étude de l'Université de Montréal (Casals et al. 2013) a ainsi montré une réduction significative de la diversité génétique au sein de la population québécoise comparativement à la population Française dont elle est issue depuis seulement 20 générations. De plus, cette population comporterait une plus grande proportion de variants fonctionnels ayant un rôle potentiellement délétère, ce phénomène expliquerait en partie la prévalence plus importante de certaines maladies génétiques dans cette province. A titre d'exemple, nous pouvons citer le syndrome d'Andermann qui est presqu'exclusif de la population franco-canadienne du Québec, et plus particulièrement de celle du Saguenay–Lac-Saint-Jean. Cette maladie neurodégénérative récessive est associée à des mutations du gène codant pour le transporteur transmembranaire SLC12A6 (Dupré et al. 2003). On estime qu'1/23 de la population du Saguenay–Lac-Saint-Jean serait porteur d'un allèle incriminé et qu'1 enfant sur 2117 serait concerné par cette maladie extrêmement rare dans le reste du monde.

On parle de différenciation lorsque des individus sont isolés et cessent de se reproduire avec ceux de la population d'origine. Ils tendent ainsi à accumuler des différences génétiques dont le degré conditionnera l'apparition de deux populations ou de deux nouvelles espèces.

Lorsqu'une population est en expansion, c'est-à-dire que le nombre d'individu augmente, les effets de la dérive génétique sur la fréquence des variants sont retardés.

ll faut noter que le phénomène de dérive génétique s'applique à l'ensemble des variants et est indépendant de la sélection naturelle. Seule la sélection naturelle est susceptible d'accélérer la fixation ou la disparition d'un variant lorsque celui-ci a un effet sur le succès reproductif (fitness) de l'individu dans son environnement. Ainsi, l'apparition d'un variant conférant un avantage reproductif aura plus de chance de se fixer rapidement au sein de la population.

Sélection naturelle

Plusieurs types de sélection naturelle sont à distinguer. La sélection positive favorise les variants génétiques qui confèrent un avantage reproductif alors que la sélection négative s'attelle à purger ceux qui sont délétères. Il existe également la sélection balancée qui maintient la diversité génétique où les allèles tendent à être conservés dans le cas où l'état hétérozygote favorise la survie de l'individu, comparativement aux homozygotes.

La sélection naturelle laisse des traces à certaines positions de notre génome. Pour détecter ces évènements de sélection, nous disposons de plusieurs approches, parmi celles-ci, F_{ST} et iHS sont deux statistiques couramment utilisées.

L'indice de fixation (F_{ST}) (Hudson, Slatkin, et Maddison 1992; Holsinger et Weir 2009) est une méthode qui permet d'apprécier le niveau de différenciation de deux populations, en se basant sur la variance des fréquences alléliques. Pour un variant donné, la valeur F_{ST} est élevée lorsque les fréquences alléliques sont différentes entre les deux populations ; à l'inverse, lorsqu'elle est petite, cela signifie que les fréquences alléliques sont similaires. Dans le cas où la sélection naturelle favorise un allèle par

rapport à un autre dans une des deux populations, alors la valeur de F_{ST} pour ce variant sera importante. Des valeurs élevées de F_{ST} peuvent être le fruit d'un évènement de sélection naturelle ou apparaître simplement sous l'effet de la dérive génétique. En effet, lorsque deux populations sont séparées et cessent de se reproduire entre elles, les fréquences alléliques d'un même variant sont susceptibles de varier aléatoirement sans que ce variant n'ai été ciblé par la sélection naturelle.

Integrated Haplotype Score (iHS) (Voight et al. 2006) est un second test destiné à mettre en évidence des évènements de sélection positive récents (moins de 60 000 ans). Cette méthode repose sur l'identification de différences dans le déséquilibre de liaison à proximité d'un allèle sélectionné positivement, comparativement à l'autre allèle. En d'autres termes, iHS permet de détecter des variants qui ont été sélectionnés positivement car ceux-ci, en augmentant rapidement en fréquence au sein de la population, échappent aux mécanismes de recombinaison et sont présents sur des haplotypes anormalement longs. On estime qu'une valeur absolue d'iHS supérieure à 2 est un signe de sélection positive récente.

L'ensemble de ces évènements de sélection, en affectant la fréquence des haplotypes, a participé à l'évolution des populations humaines mais également à celle des autres représentants archaïques du genre Homo, tels que l'hominidé de Denisova et l'homme de Néandertal.

Différences entre populations, un héritage de l'adaptation

Au cours de l'évolution, nous savons que la compétition entre les pathogènes et notre organisme a exercé des pressions extrêmement importantes sur notre génome

(Barreiro et Quintana-Murci 2010; Fumagalli et al. 2011; Karlsson, Kwiatkowski, et Sabeti 2014). Les migrations récentes de nos ancêtres, en colonisant de nouveaux environnements, ont exposé les populations à des pressions pathogéniques différentes.

Ainsi, lorsque les populations ont quitté l'Afrique, l'organisme s'est adapté pour assurer sa survie face à de nouvelles conditions. Nous pensons que la pression pathogénique en dehors de l'Afrique était différente à la fois en ce qui concerne la quantité mais également les espèces de pathogènes. Tout d'abord, la quantité de pathogènes était probablement moins importante dans l'environnement en Europe, étant donné que son abondance semble inversement corrélée à la latitude (Guernier, Hochberg, et Guégan 2004). En outre, le développement de l'agriculture dans le croissant fertile (Egypte et nord de la péninsule Arabique) il y a près de 15 000 ans a renforcé les contacts entre l'homme et les animaux domestiqués et probablement entraîné une recrudescence des zoonoses qui représentaient de nouvelles infections auxquelles l'homme n'avait pas été confronté.

Même si nous savons qu'il existe des différences de réponses immunitaires entre les populations humaines, le rôle joué par l'évolution et son impact au niveau moléculaire restent méconnus. Par exemple, des études épidémiologiques ont montré que les individus d'origine Africaine étaient trois fois plus à risque de développer des maladies en rapport avec le système immunitaire, telles que la tuberculose, le lupus, le psoriasis ou des septicémies (Brinkworth et Barreiro 2014; Richardus et Kunst 2001). Il semblerait que ces différences de susceptibilités soient en partie expliquées par une inflammation particulièrement prononcée chez les populations d'origine africaine (Pennington et al. 2009).

Impact des variants génétiques sur l'expression des gènes

L'expression génétique est le processus selon lequel l'information contenue dans un gène est utilisée pour aboutir à la synthèse de molécules responsables de fonctions au sein de la cellule. Parmi ces effecteurs, on distingue principalement les protéines et les ARN non codants.

Plusieurs étapes sont nécessaires à la production d'une protéine fonctionnelle : transcription, maturation et épissage de l'ARN, traduction et modifications post-traductionnelles de la protéine. Les protéines sont les effecteurs principaux de la cellule et peuvent remplir des fonctions diverses, par exemple en assurant l'intégrité structurelle, la signalisation, les fonctions métaboliques ou les mécanismes de défense face aux pathogènes.

Les variants génétiques peuvent avoir des répercussions sur l'expression génétique à la fois en altérant la fonction du transcrit ou de la protéine produite mais également en modifiant la quantité de ces molécules.

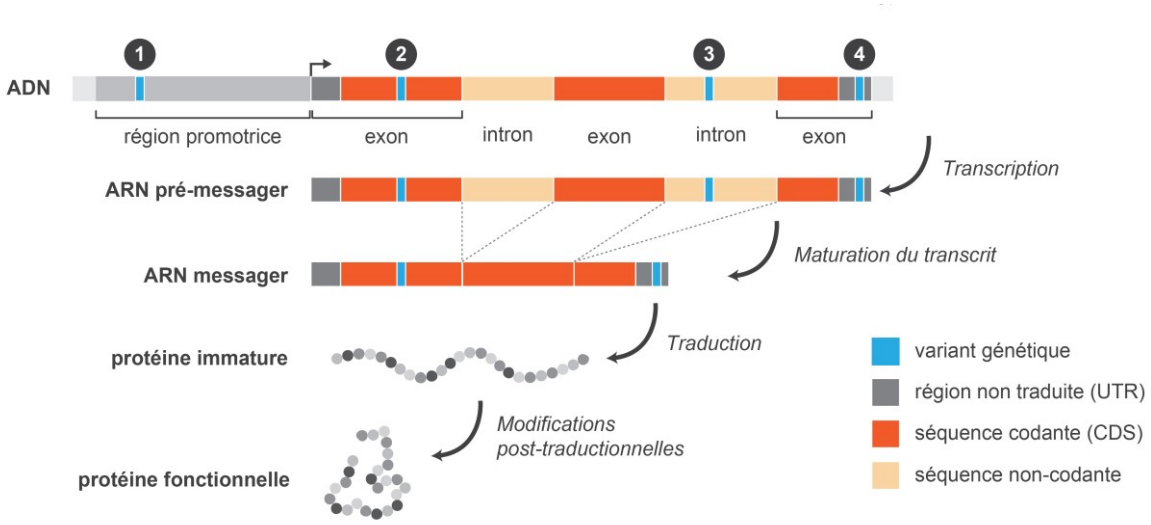

Figure 5. **VARIANTS ET EXPRESSION GENETIQUE.**

Il est important de distinguer les effets que peuvent avoir les mutations en fonction de leur localisation et de leur nature (Figure 5). Un variant génétique dans la région promotrice (1) est susceptible d'affecter le site de liaison pour la machinerie cellulaire de transcription et donc de faire varier la quantité de transcrits produits. Lorsqu'un variant génétique est en région codante du gène (2), celui-ci peut entrainer une modification de l'acide aminé de la protéine associé à ce codon (mutation faux-sens) ou l'apparition d'un signal de fin de traduction prématuré (mutation non-sens). Le code génétique étant redondant, c'est-à-dire que plusieurs codons peuvent être associés au même acide aminé, il est possible qu'une mutation en séquence codante soit silencieuse. Un variant génétique dans un intron (3) peut perturber des sites de liaisons pour des molécules chargées de l'épissage des exons et favoriser l'orientation vers un isoforme à l'occasion de la maturation de l'ARN messager. Ce dernier phénomène pourra donner lieu à la modification des proportions des isoformes d'un gène dans le cas où celui-ci est capable de produire plusieurs effecteurs protéiques. Enfin, un variant génétique dans une région non-traduite du gène (UTR) (4), bien que n'ayant pas de retentissement sur la séquence de la protéine, peut affecter l'interaction de l'ARN avec la machinerie

cellulaire responsable de la maturation, du transport et de la survie. Ce variant peut également compromettre le repliement du messager et mettre en péril sa stabilité avant le début de la traduction.

Etant donné que seul 1.5% de notre génome est impliqué dans la production de protéines, les séquences non-codantes ont longtemps été caractérisées « d'ADN poubelle ». Le projet ENCODE a identifié que 80% des séquences de notre génome seraient fonctionnelles, c'est-à-dire qu'elles auraient un rôle biochimique et que leur altération par des variants génétiques serait susceptible d'affecter les mécanismes cellulaires (ENCODE Project et al. 2012). Même s'il semble difficile de valider l'utilité[1] de l'ensemble de ces séquences, nous savons qu'une partie d'entre-elles ont un rôle fonctionnel évident ; par exemple en permettant la fixation de facteurs de transcription, la modulation de l'accessibilité des séquences d'ADN au sein de la chromatine ou la production de microARNs capables de réguler la survie d'ARNs messagers cibles.

Si l'on considère l'expression des gènes comme un trait quantitatif (QT), nous pouvons mesurer son héritabilité en estimant la contribution des facteurs génétiques à la variance du phénotype (Cui et al. 2006). Le phénomène est appelé *expression quantitative trait loci* (eQTL) et repose sur l'établissement d'associations entre des régions du génome et des variations de l'expression génétique (Cheung et al. 2005; Gibson et Weir 2005; Morley et al. 2004).

Il est important de noter qu'au-delà du niveau global d'expression du gène, c'est-à-dire la quantité totale d'expression de tous les transcrits d'un gène, il est possible de

[1] Parmi les critiques formulées à l'encontre de ces déclarations du projet d'ENCODE, on peut citer la curieuse disparité dans la proportion d'ADN non-codant au sein des espèces (Palazzo et Gregory 2014). Ainsi, le génome d'un oignon comporte 5 fois plus d'ADN non-codant que celui de l'homme.

s'intéresser aux différentes proportions de transcrits produits. En effet, l'épissage alternatif est le mécanisme responsable de la sélectionner des exons destinés à faire partie de l'ARN messager mature qui sera ensuite traduit en protéine. De ce fait, un variant génétique impliqué dans un *alternative splicing quantitative trait loci* (asQTL) est susceptible de perturber les proportions de chaque transcrit produit et donc d'affecter la quantité et la proportion des protéines issues d'un même gène. Pour identifier des eQTL et asQTL, il est nécessaire de connaître les variants génétiques des individus et de mesurer l'expression de leurs transcrits.

Plusieurs technologies sont disponibles pour mesurer en parallèle l'expression d'un grand nombre de gènes. Alors que les puces à ADN peuvent interroger un set prédéfini de transcrits, le séquençage de l'ARN permet d'examiner l'ensemble du génome et il est ainsi possible de mettre en évidence de nouveaux transcrits propres au type cellulaire que l'on étudie. Bien que cette technologie devienne de plus en plus abordable, elle demeure dispendieuse et nécessite un traitement bio-informatique avancé, notamment à travers l'alignement, le contrôle de qualité et l'estimation des valeurs d'expression.

La technique de séquençage d'ARN se découpe en trois parties : la préparation de l'échantillon, son séquençage et l'analyse des données. Lors de la préparation de l'échantillon, les ARN messagers comportant une queue poly-adénylée sont capturés, fragmentés puis des séquences adaptatrices sont ajoutées à leurs extrémités. Ils sont ensuite placés sur le support de séquençage (flowcell) où ils se répliqueront pour former des clusters. Les nucléotides marqués sont enfin ajoutés et leur incorporation au sein des brins complémentaires est suivie grâce à la fluorescence. A l'issue du séquençage,

on dispose de plusieurs millions de lectures d'environ 100 nucléotides appelés « reads »[2]. Ils seront ensuite alignés contre le génome pour identifier les régions d'où ils proviennent et ainsi estimer l'expression des gènes.

Les études portant sur les eQTL réalisées sur l'ensemble du génome ont montré que la variation de l'expression génétique est répandue et hautement héritable, que les loci identifiés expliquent une grande proportion de la variance observée et que les associations les plus fortes reposent sur des polymorphismes à proximité des gènes qu'ils régulent. Lorsqu'un polymorphisme génétique est proche d'un gène, il a plus de chance d'être associé à la variation de l'expression du transcrit que s'il est éloigné ou sur un autre chromosome (Zhang et al. 2014).

Lorsque l'on cherche à caractériser les eQTL, ceux-ci peuvent reposer sur une association en *cis*, où le polymorphisme affecte l'expression d'un allèle ou en *trans*, où le polymorphisme affecte l'expression des deux allèles. Une association en *cis* implique un mécanisme moléculaire local qui n'affecte que l'allèle où il siège tel que la fixation d'un facteur de transcription, répresseur ou d'autres régulateurs opérant à grande distance (Sanyal et al. 2012). A l'opposé, une association en *trans* fait intervenir un élément régulateur qui affecte sans distinction les deux copies du gène à travers un facteur de transcription ou un micro-ARN (Selbach et al. 2008). D'usage, on parle d'une association en *cis* lorsque le variant est proche du gène affecté et en *trans* lorsque celui-ci en est éloigné ou est porté par autre chromosome[3].

[2] La taille des reads dépend de la technologie utilisée et du paramétrage de la machine de séquençage.

[3] Il est important de noter qu'un effet en *trans* pour un SNP à proximité du gène est possible.

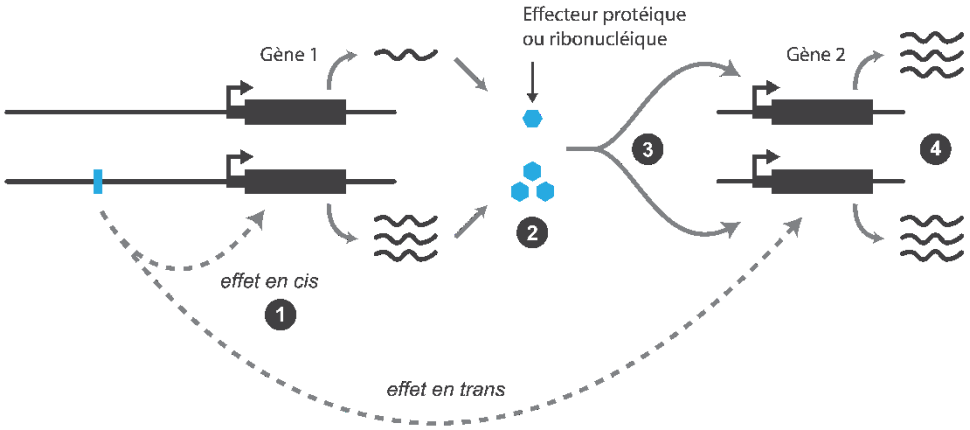

Figure 6. **ILLUSTRATION D'EFFETS EN CIS ET TRANS**

1. Un effet en cis est présent lorsqu'un variant génétique est susceptible d'affecter l'expression d'un gène. **2.** Ici, un allèle favorise l'expression du gène porté par cette même copie du chromosome, aboutissant à une quantité plus importante d'effecteur. **3.** Ces molécules, par exemple des facteurs de transcriptions, sont ensuite capable d'altérer les deux copies d'un gène cible porté sur le même ou un autre chromosome, **4.** entrainant une augmentation de la transcription. Nous sommes en mesure de détecter un effet en trans en cherchant une corrélation entre les états alléliques du variant génétique et le niveau d'expression du gène 2.

Le principal avantage des eQTL est qu'ils permettent de comprendre, d'un point de vue fonctionnel, comment des variants génétiques affectent l'expression génétique. Ainsi, si l'on identifie un eQTL dont le polymorphisme génétique est juste en amont du site d'initiation de la transcription du gène, ce variant siège dans la région promotrice et est susceptible d'affecter l'affinité de la fixation d'un facteur de transcription. Prenons l'exemple (Figure 7) de trois individus respectivement : homozygote à l'allèle ancestral (`CC`), hétérozygote (`CT`) et homozygote à l'allèle alternatif (`TT`). Si l'allèle alternatif `T` siège dans une région promotrice du gène et dégrade l'affinité de fixation d'un facteur de transcription, les individus `CT` et `TT` montreront des niveaux d'expression du gène plus faible.

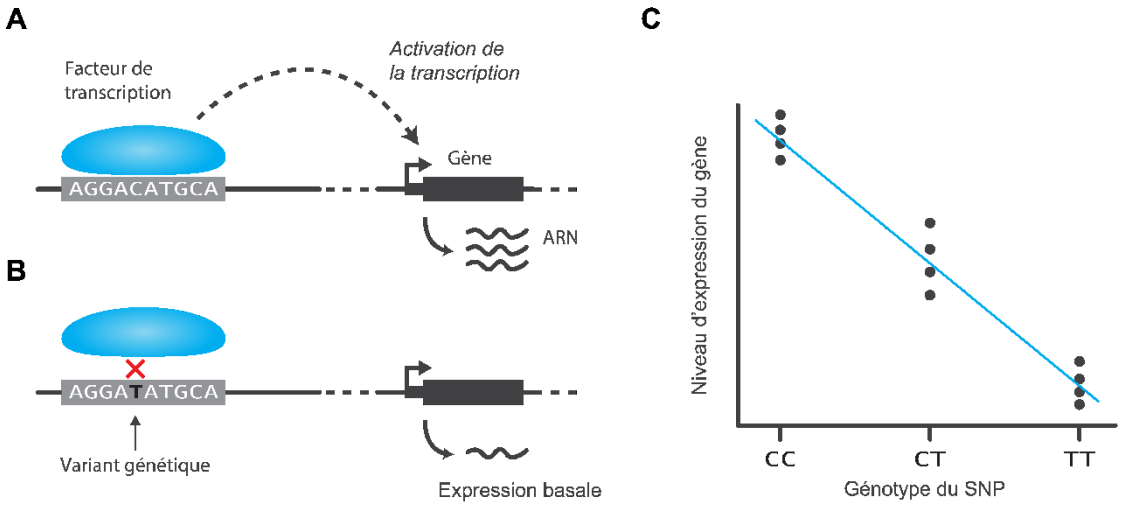

Figure 7. EFFET D'UN VARIANT GENETIQUE EN REGION REGULATRICE SUR L'EXPRESSION D'UN GENE

A. Un facteur de transcription activateur est capable de reconnaitre et de se lier sur une séquence nucléotidique spécifique pour stimuler l'expression du gène. **B**. Dans le cas d'un allèle T qui perturbe la séquence du motif, le facteur de transcription n'est plus capable de se lier à l'ADN et ne peut pas activer la transcription. **C**. Lorsque l'on compare l'expression des individus en fonction de leur génotype, on distingue un effet du variant génétique sur le niveau d'expression du gène, c'est un eQTL.

Plusieurs méthodes existent pour mettre en évidence les eQTL. La plus répandue est basée sur la régression linéaire et cherche à identifier une relation entre le génotype de l'individu et le niveau d'expression du gène. Cette approche repose sur le postulat selon lequel les effets rencontrés sont additifs, c'est-à-dire que les individus de génotype `CC` montreront un niveau d'expression égal à 2 fois celui des individus `CT`, puisque qu'ils disposent de deux copies de l'allèle `C`. L'avantage principal de la régression linéaire est sa simplicité de calcul puisqu'elle permet de cribler rapidement un nombre conséquent d'associations entre un SNP et un gène, même sur l'ensemble du génome. A titre d'exemple, si l'on dispose de 13 millions de génotypes et de l'expression de 13 000 gènes, cela représente 169 milliards d'associations possibles qu'il est nécessaire d'évaluer individuellement. Entreprendre une telle analyse demeure un challenge technique et bien que les algorithmes et systèmes informatiques les plus rapides permettent d'évaluer ces associations en une cinquantaine d'heures, la gestion des fichiers et leur interrogation

reste complexe. En outre, la réalisation d'un tel nombre de tests nécessite des approches plus agressives pour contrer le fardeau des tests multiples et discerner les associations faussement significatives de celles qui dépendent d'un véritable mécanisme biologique.

L'étude des eQTL représente un axe de recherche important dans notre lutte contre les maladies immunitaires puisqu'elle permet de mettre en évidence des marqueurs génétiques associés aux phénomènes de susceptibilité aux infections. Dans une précédente étude de notre laboratoire (Barreiro et al. 2011), on s'est intéressé aux eQTL dans un contexte d'infection des cellules dendritiques par *Mycobacterium tuberculosis*, la bactérie responsable de la tuberculose. Les individus ont été génotypés et l'expression des gènes a été quantifiée avant et après infection. L'analyse a montré que les gènes qui présentent une variation de leur expression entre les états non-infectés et infectés ont plus de chance d'être impliqués dans un eQTL. Cette remarque confirme que les eQTL permettent d'expliquer les variations d'expression génétiques.

Plus récemment, une publication (Fairfax et al. 2014) s'est intéressée aux eQTL dans des contextes de stimulations de monocytes humains à l'aide de *LPS* et *IFN-γ*. Un autre projet (Lee et al. 2014) a été plus loin en étudiant de telles associations entre les polymorphismes et l'expression des gènes à travers l'infection de cellules dendritiques par Influenzavirus A de sous-type H1N1. Ces deux travaux sont proches de mon propre projet sur le plan mécanistique, c'est-à-dire la mise en évidence d'eQTL lors de la réponse immunitaire, mais ne situent pas ces découvertes dans un contexte de génétique des populations humaines. La mise en évidence de susceptibilités aux infections bactériennes et la compréhension des mécanismes génétiques sous-jacentes au niveau populationnel sont des axes de recherche cruciaux pour renforcer nos connaissances en médecine.

Mise en contexte du projet

Nous savons qu'il existe des différences de prévalences de maladies liées au système immunitaire entre populations humaines et supposons que celles-ci sont déterminées par une combinaison de facteurs environnementaux et génétiques. La contribution des variants génétiques vis-à-vis de ce phénomène demeure mal comprise, en dépit de l'intérêt que représente ce type d'information pour améliorer notre compréhension du système immunitaire. Pour mener à bien ce projet de doctorat, je propose 3 axes principaux.

Axe 1 : différences de réponses immunitaires entre populations humaines

Hypothèse : Il existe des différences de réponse immunitaire entre populations humaines que nous sommes en mesure de mettre en évidence à travers des expériences in vitro.

Objectifs :

- Obtenir des macrophages à partir de donneurs sains Afro- et Euro-américains puis les infecter par Listeria *monocytogenes* et Salmonella *typhimurium.*
- Collecter des phénotypes en rapport avec l'infection bactérienne : capacité du macrophage à lyser les bactéries intracellulaires (clairance bactérienne) et expression des gènes (séquençage de l'ARN)
- Rechercher des différences de clairance bactérienne entre les populations.
- Identifier des gènes et isoformes différentiellement exprimés entre Afro- et Euro-américains.

Axe 2 : impact des variants génétiques dans la réponse immunitaire

Hypothèse : Une partie des différences de réponse immunitaire observées entre Afro- et Euro-américains sont conditionnées par des variants génétiques.

Objectifs :

- Identifier les variants génétiques impliqués dans des variations d'expression de gènes et isoformes entre individus.
- Evaluer la contribution de ces variants vis-à-vis des gènes et isoformes différentiellement exprimés entre Afro- et Euro-américains.

Axe 3 : système immunitaire et histoire évolutive des populations humaines

Hypothèse : certains des variants génétiques associés à des modifications transcriptionnelles ont été sélectionnés au cours de l'adaptation des populations à leur environnement pathogénique respectif.

Objectifs :

- Mettre en évidence des signatures de sélection naturelle récente au sein de ces variants, suggérant des pressions évolutives.
- Placer ces résultats en perspective avec nos connaissances sur l'origine de l'homme moderne et plus particulièrement le métissage entre notre espèce et l'Homme de Néandertal.

Chapitre 2 : Genetic ancestry and natural selection drive population differences in immune responses to pathogens.

Contexte de l'article

Cet article est le fruit d'une large étude inédite qui a débuté en 2012 qui repose sur des techniques avancées en laboratoire humide : culture cellulaire, différenciation et infection de macrophages humains (Vania Yotova, Ariane Pagé Sabourin, Anne Dumaine); et fait appel à des analyses bio-informatiques particulièrement innovantes : séquençage de l'ARN, génotypage, identification de gènes et transcrits différentiellement exprimés et mise en évidence des variants génétiques impliqués (Yohann Nédélec, Joaquin Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Jean-Christophe Grenier).

Cet article incarne l'aboutissement du travail que j'ai effectué depuis le début de mon doctorat, en janvier 2013. J'ai été en charge des tâches suivantes :

- supervision du volet bio-informatique de l'étude et coordination avec l'équipe menant les expérimentations dans notre laboratoire humide ;
- création et déploiement des approches nécessaires au contrôle de qualité, à l'analyse et à l'interprétation des résultats ;
- production de l'ensemble des figures présentes dans l'article.

Notre article a été soumis le 13 avril 2016 à la revue Cell, accepté à l'issue de la première révision le 15 septembre puis publié le 20 octobre de la même année.

Nous avons également eu la chance de réaliser la couverture du volume 169, numéro 3 du journal Cell, en partenariat avec deux artistes graphistes.

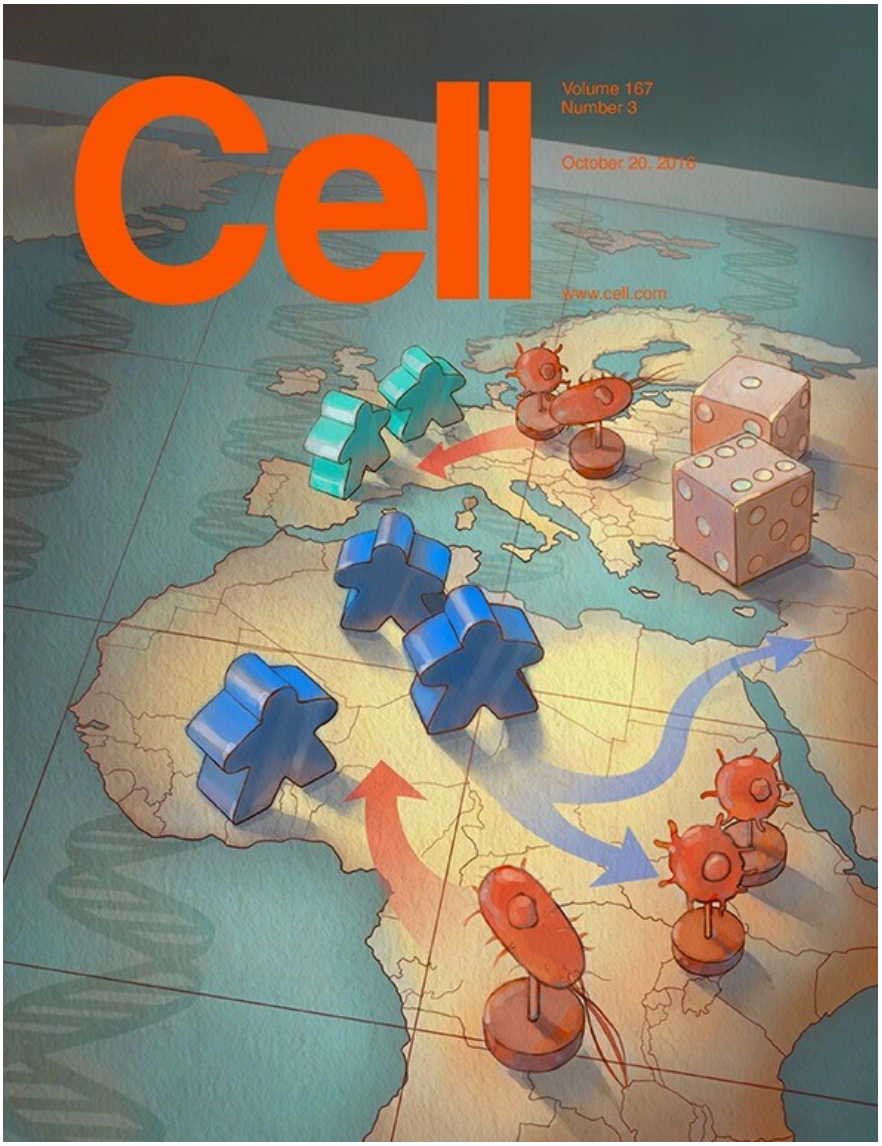

Figure 1. **COUVERTURE DE CELL, VOLUME 169, NUMERO 3**

Cette couverture montre un jeu de plateau qui illustre la bataille perpétuelle entre les humains et les pathogènes ainsi que les pressions de sélection qui en résultent et ont participé à l'adaptation des populations à leur environnement. Illustration réalisée par Edwin Choi et Sigrid Knemeyer selon une idée de Luis Barreiro.

Les tableaux supplémentaires sont disponibles sur le site de l'éditeur : http://www.cell.com/cell/fulltext/S0092-8674(16)31307-1

Authors

Yohann Nédélec[1,2,9], Joaquín Sanz[1,2,9], Golshid Baharian[1,2,9], Zachary A. Szpiech[3], Alain Pacis[1,2], Anne Dumaine[2], Jean-Christophe Grenier[2], Andrew Freiman[4], Aaron J. Sams[5], Steven Hebert[2], Ariane Pagé Sabourin[2], Francesca Luca[4], Ran Blekhman[6], Ryan D. Hernandez[3], Roger Pique-Regi[4] , Jenny Tung[7], Vania Yotova[2], Luis B. Barreiro [2,8,*]

Affiliations

[1] Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montreal, H3T1J4, Canada

[2] Department of Genetics, CHU Sainte-Justine Research Center, Montreal, H3T1C5, Canada

[3] Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, and Quantitative Biosciences Institute, University of California, San Francisco, CA 94143, USA

[4] Center for Molecular Medicine and Genetics and Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan 48202, USA

[5] Department of Biological Statistics & Computational Biology, Cornell University, USA

[6] Department of Genetics, Cell Biology, and Development, and Department of Ecology, University of Minnesota, Twin Cities, MN 55108, USA

[7] Departments of Evolutionary Anthropology and Biology and Duke Population Research Institute, Duke University, Durham, North Carolina 27708, USA

[8] Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[9] Co-first author.

Abstract

Individuals from different populations vary considerably in their susceptibility to immune-related diseases. To understand how genetic variation and natural selection contribute to these differences, we tested for the effects of African versus European ancestry on the transcriptional response of primary macrophages to live bacterial pathogens. 9.3% of macrophage-expressed genes show ancestry-associated differences in the gene regulatory response to infection, and African ancestry specifically predicts a stronger inflammatory response and reduced intracellular bacterial growth. A large proportion of these differences are under genetic control: for 804 genes, more than 75% of ancestry effects on the immune response can be explained by a single *cis*- or *trans*-acting eQTL. Finally, we show that genetic effects on the immune response are strongly enriched for recent, population-specific signatures of adaptation. Together, our results demonstrate how historical selective events continue to shape human phenotypic diversity today, including for traits that are key to controlling infection.

Introduction

As our primary interface with the environment, the immune system is thought to have evolved under strong selective pressure from pathogens (Barreiro et Quintana-Murci 2010; Fumagalli et al. 2011; Karlsson, Kwiatkowski, et Sabeti 2014). When human populations migrated out of Africa, they encountered markedly different pathogenic environments, likely resulting in population-specific selection on the immune response (Barreiro et Quintana-Murci 2010; Fumagalli et al. 2011; Karlsson, Kwiatkowski, et Sabeti 2014). Substantial evidence supports this hypothesis at the genetic level. However, we still know little about the extent to which neutral or adaptive inter-population genetic differences affect the actual immune response to pathogens.

Addressing this gap is not only important for understanding recent human evolution, but may also help reveal the molecular basis of ancestry-related differences in disease susceptibility. Individuals from different populations vary considerably in their susceptibility to many infectious diseases, chronic inflammatory disorders, and autoimmune disorders. For tuberculosis, systemic lupus erythematosus, systemic sclerosis, psoriasis, and septicemia, African American (AA) and European American (EA) individuals exhibit an up to 3-fold difference in prevalence (reviewed in: (Brinkworth et Barreiro 2014; Pennington et al. 2009; Richardus et Kunst 2001)). These observations argue in favor of significant ancestry-related differences in immune response, especially in susceptibility to inflammation (Pennington et al. 2009; Richardus et Kunst 2001).

Such differences almost certainly involve major contributions from the environment. However, genome-wide association studies (GWAS) also support a key role for genetic factors, as many of the GWAS-variants associated with infectious, autoimmune, and inflammatory diseases present extreme differences in allele frequency ($F_{\mathrm{st}} > 0.4$) between human populations, again supporting a possible history of population-specific selection (Brinkworth et Barreiro 2014).

GWAS results also indicate that susceptibility to many common immune-related diseases is primarily controlled by non-coding variants (Gusev et al. 2014; Hindorff et al. 2009; Schaub et al. 2012). Thus, many ancestry-related differences in disease susceptibility may result from genetically controlled transcriptional differences in immune responses to inflammatory signals. This idea is consistent with recent expression quantitative trait locus (eQTL) mapping studies in innate immune cells exposed to immune antigens or live infectious agents (Barreiro et al. 2011; Çalışkan et al. 2015; Lee et al. 2014). Such immune "response eQTL" studies have identified

hundreds of genetic variants that both explain variation in the host immune response and are significantly enriched among GWAS-associated loci. However, because studies to date have mostly focused on individuals of European ancestry, the degree to which such variants contribute to population differences in the immune response remains unclear.

Here, we report an RNA-sequencing-based immune response eQTL study to test for the effects of African versus European ancestry on the transcriptional response to several live bacterial pathogens. We integrate statistical and evolutionary genetic analyses with primary macrophage gene expression levels, before and after infection, to characterize ancestry-related differences in the immune response. Our analyses address three fundamental questions about recent evolution in the human immune system: (*i*) the degree to which innate immune responses are differentiated by European versus African ancestry; (*ii*) the genetic variants that account for such differences; and (*iii*) the evolutionary mechanisms (neutral genetic drift vs positive selection) that led to their establishment in modern human populations. Finally, to facilitate the use of our data by the research community, we have developed an accessible, publicly available browser for exploring our results: the ImmunPop QTL browser (http://www.immunpop.com).

Results

Transcriptional response of macrophages to Listeria and Salmonella

We infected monocyte-derived macrophages – phagocytic cells that are essential for fighting foreign invaders, tissue development, and homeostasis (Okabe et Medzhitov 2016) – derived from 80 AA and 95 EA individuals (**Table S1**) with either *Listeria monocytogenes* (a Gram-positive bacterium) or *Salmonella typhimurium* (a Gram-negative bacterium). Following 2 hours of infection, we collected RNA-seq data from matched non-infected and infected samples, for a total of 525 RNA-seq profiles across individual-treatment combinations (mean = 36 million reads per sample; see Methods; **Figure S1A**). Each individual was genotyped for over 4.6 million single nucleotide polymorphisms (SNPs), with additional imputation to ~13 million SNP genotypes (see Methods). After quality control (**Figure S1A**), we were able to study 171 individuals with high-quality RNA-seq data, among which 168 were also successfully genotyped.

The first principal component of the resulting gene expression data accounted for 85% of the variance in our dataset and separated non-infected macrophages from macrophages infected with either *Listeria* or *Salmonella* (**Figure 1A**). We found extensive differences in gene expression levels between infected and non-infected cells, with 5,201 (44%) and 6,701 (56%) differentially expressed genes after infection with *Listeria* and *Salmonella*, respectively (see Methods, False Discovery Rate (FDR)<0.01 and $|\log_2(\text{fold change})| > 0.5$; **Table S2A**). As expected, the sets of genes that responded to either infection were strongly enriched (FDR < 0.01) for gene sets involved in immune function, including the regulation of inflammatory responses, cytokine production, T-cell activation, and apoptosis (**Table S3**).

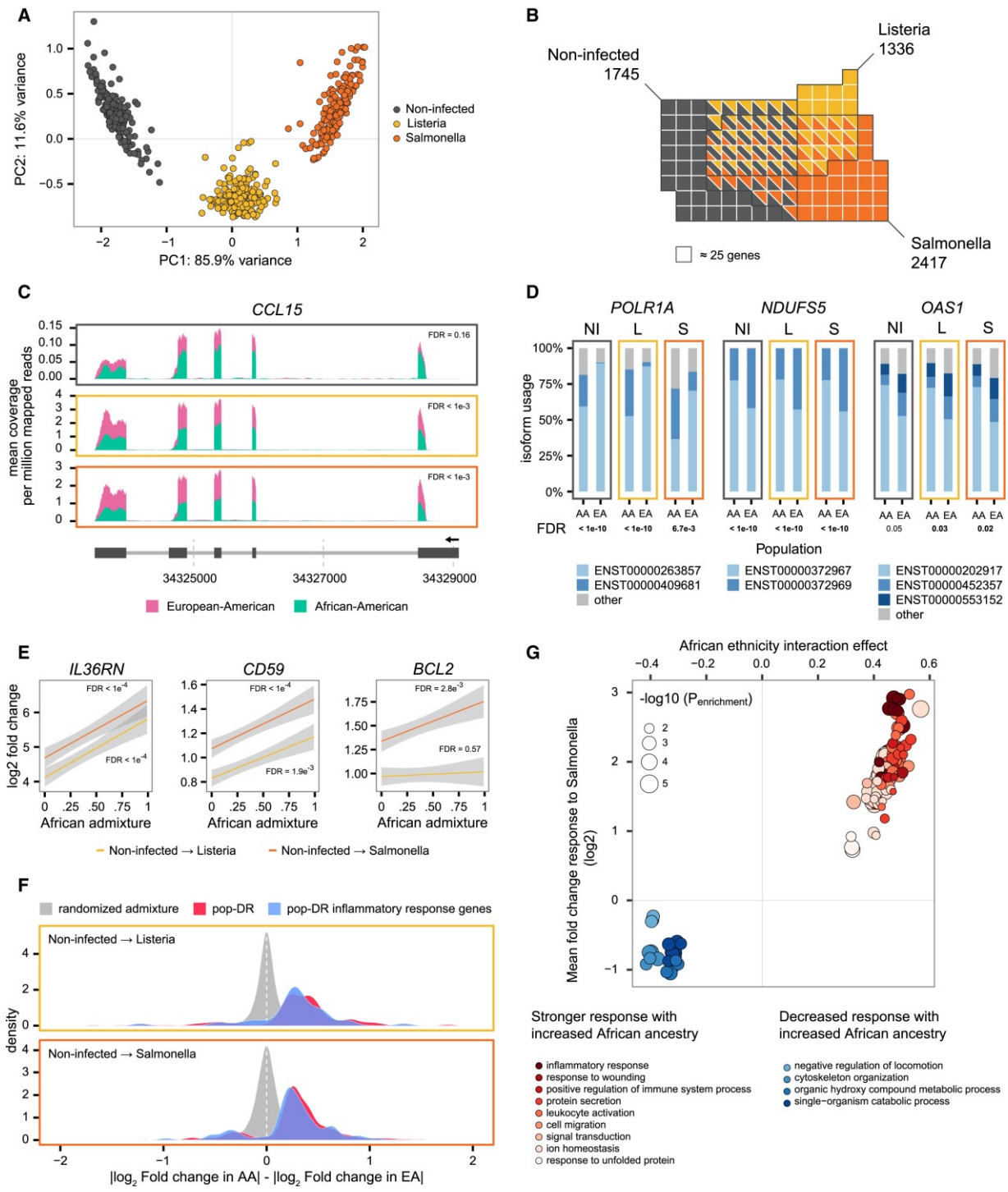

Figure 1. EUROPEAN AND AFRICAN ANCESTRY-ASSOCIATED DIFFERENCES IN IMMUNE RESPONSE

A. Principal component analysis of gene expression data from all samples. PC1 (x axis) and PC2 (y axis) clearly separate non-infected macrophages from macrophages infected with either *Listeria* or *Salmonella*. **B.** Venn diagram illustrating the number of pop-DE genes (FDR < 0.05) in non-infected (black), *Listeria*-infected (yellow), and *Salmonella*-infected (orange) macrophages. **C**. Example of a gene, the chemokine *CCL15*, for which expression levels in all conditions are significantly

associated with levels of African versus European ancestry. The average sequencing depth for each base (normalized per million mapped reads) is shown on the y axis. **D.** Example of three genes (*POLR1A*, *NDUFS5*, and *OAS1*) for which ancestry predicts differences in isoform usage. **E.** Example of three immune-related pop-DR genes. The y axis shows the log2 fold changes in gene expression levels in response to *Listeria* and *Salmonella*, as a function of continuous differences in African ancestry (x axis). **F.** Absolute difference in the log2 fold change response to *Salmonella* (top panel) and *Listeria* infection (bottom panel) between European and African individuals (x axis), among all pop-DR genes (red) and pop-DR genes associated with the inflammatory response (blue). The null expectation from permuting admixture levels across individuals is shown in light gray for comparison. A shift in the distribution to the right reflects a stronger response to infection in African-ancestry individuals. **G.** GO enrichment analysis for genes showing a significant interaction between ancestry and the response to *Salmonella*. Only GO terms with an enrichment at FDR < 0.1 are displayed, and GO terms are color-coded into functionally related terms based on the overlap among gene sets (Bindea et al. 2009). For each GO term, the average interaction effect is plotted on the x axis and the mean log2 fold change in gene expression levels in response to infection is plotted on the y axis. See also Tables S2 and S3.

Ancestry-related differences in the innate immune response to infection

We first aimed to characterize European versus African ancestry-related transcriptional differences in non-infected and infected macrophages. Because self-identified ethnicity is an imprecise proxy for the actual genetic ancestry of an individual, we used the genotype data to estimate genome-wide levels of European and African ancestry in each sample using the program ADMIXTURE (Alexander, Novembre, et Lange 2009). Consistent with previous reports (Bryc et al. 2010; Tishkoff et al. 2009), we found that many self-identified AA individuals have a high proportion of European ancestry (mean= 30%, range 0.9-100%; **Figure S1B**). In contrast, self-identified EA showed more limited levels of African admixture (mean= 0.4%, range 0-18%; **Figure S1B**). Thus, we used these continuous estimates (as opposed to a binary classification of individuals into African or European ancestry) to identify ancestry-associated differentially expressed genes (*i.e.*, pop-DE genes: genes for which gene expression levels are linearly correlated with ancestry levels; see Methods for details on the nested linear model used for this analysis).

Of the 11,914 genes we tested, we identified 3,563 pop-DE genes (30%) in at least one of the experimental conditions, explaining a mean 8.2% of expression variance (range 1.8-44%) (FDR<0.05: 1,745 in non-infected (NI), 1,336 in *Listeria*-infected (L),

and 2,417 in *Salmonella*-infected (S) macrophages; **Figure 1B** and **C**; **Table S2B**). These differences primarily influence mean gene expression levels across transcript isoforms, as opposed to the proportion of isoform usage within genes. Specifically, among genes with at least two annotated isoforms (N = 10,223), only 62, 39, and 48 genes exhibited evidence for ancestry-associated differential isoform usage, in the non-infected, *Listeria*-infected, and *Salmonella*-infected conditions, respectively (multivariate generalization of the Welch's t-test; FDR<0.05; **Figure 1D**; **Figure S2A**, **Table S2D**). These results were unaltered by using an alternative identification approach (Wilcoxon rank sum test, as in (Lappalainen et al. 2013); see Methods for details) or when relaxing the FDR threshold used to define significance (**Figure S2B**). Despite the low number of genes showing ancestry associated differences in isoform usage, many of these genes are key regulators of innate immunity, including *OAS1* that encodes isoforms with varying enzymatic activity against viral infections (Bonnevie-Nielsen et al. 2005).

Next we sought to identify genes for which the *response* to infection (*i.e.*, fold change in gene expression in infected versus non-infected macrophages, cultured in parallel) significantly correlates with ancestry (see Methods). We term these genes "population differentially responsive" (pop-DR) genes. We detected 1,005 and 206 pop-DR genes (FDR<0.05) in response to *Salmonella* and *Listeria*, respectively (**Figure 1E**, **Table S2C**; the increased power for *Salmonella* likely results from the stronger transcriptional response induced by *Salmonella* relative to *Listeria*: see **Figure 1A**). These genes explain a mean 7.4% (range 2.6-24%) of variance in transcriptional response to infection. Overall, we found that macrophages from individuals of African ancestry produced a markedly stronger transcriptional response to both bacterial infections (Mann–Whitney test, $P < 1\text{x}10^{-15}$, **Figure 1F**). GO term enrichment analyses further

revealed that genes related to inflammatory processes were the most enriched among pop-DR genes showing a stronger response to infection in African-descent individuals (**Figure 1G**, **Figure S2C**). Together, these results indicate that increased African ancestry predicts a stronger inflammatory response to infection.

We hypothesized that ancestry-associated differences in the transcriptional response to infection could translate into ancestry-associated differences in the ability of macrophages to clear the infection. We tested this hypothesis in a subset of 89 individuals by quantifying the number of bacteria remaining inside the macrophages right after the infection step (T0), 2 hours (T2), and 24 hours (T24) post-infection. For both bacteria, increased African ancestry predicted improved control of intracellular bacterial growth. This effect was particularly noticeable in our infection experiments with *Listeria*. Despite no significant difference in the initial number of bacteria infecting macrophages (**Figure 2A**, P=0.95), the number of bacteria inside the macrophages of individuals with high levels of African ancestry at T24 was 3.2-fold lower than that of Europeans (**Figure 2A**, P=2.0x10^{-4}).

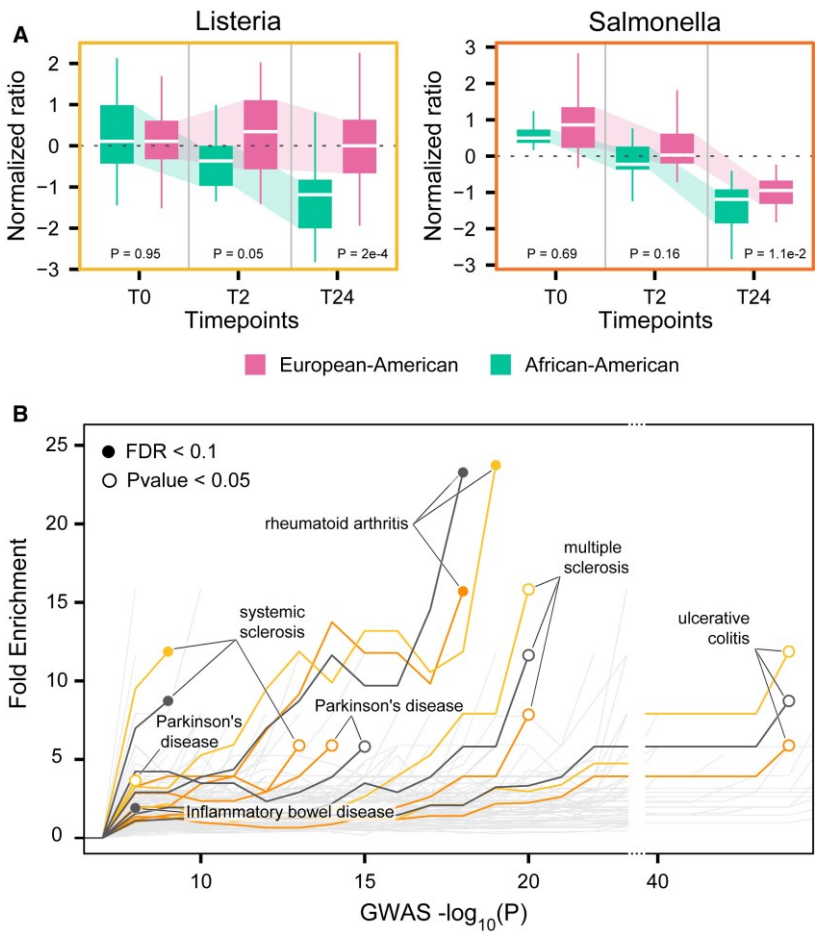

Figure 2. INCREASED AFRICAN ANCESTRY PREDICTS IMPROVED CONTROL OF BACTERIAL GROWTH INSIDE MACROPHAGES

A. Boxplots showing the quantile normalized number of bacteria inside infected macrophages (y axis) immediately after infection (T0), 2 hr post-infection (T2), and 24 hr post-infection (T24) (x axis). We quantile normalized data across individuals and time points. Analyses were conducted using a continuous measure of ancestry; however, for visualization purposes, African and European samples were defined as those with an estimated African ancestry >75% (green) and <25% (pink), respectively. **B.** Fold enrichment of pop-DE genes (y axis) among genes identified in disease susceptibility GWAS, at progressively stringent p value thresholds (x axis). Grey, yellow and orange lines show significant enrichments (filled circles at an FDR < 0.1 and open circles at a nominal p < 0.05) for pop-DE genes identified in non-infected and Listeria- and Salmonella-infected macrophages, respectively. Light gray lines show non-significant diseases/traits. See also Figure S2.

Finally, we tested if pop-DE genes were enriched among GWAS-associated genes. We found seven diseases for which susceptibility genes reported by GWAS were significantly enriched among genes classified as pop-DE, in at least one experimental condition (**Figure 2B**). Contributing to these enrichments are several HLA genes (*HLA-DQA1*, *HLA-DPA1*, *HLA-DRB1*, *HLA-DPB1*, *HLA-DRA*), known to be the

main genetic risk factors for several immune-disorders. Strikingly, six of these seven diseases (all but Parkinson's Disease) are immune-related and tightly connected to a dysregulated inflammatory response. Further, among the diseases most significantly enriched for pop-DE genes were rheumatoid arthritis, systemic sclerosis, and ulcerative colitis, all of which have been reported to differ in incidence or disease severity between AA and EA individuals (Brinkworth et Barreiro 2014; Pennington et al. 2009). Thus, ancestry-associated gene regulatory differences likely contribute to known ethnic disparities in inflammatory and autoimmune disease susceptibility, in part through affecting the ability of macrophages to control bacterial infections.

Gene expression QTL in non-infected and infected macrophages.

To identify whether pop-DE and pop-DR genes are explained by genetic differences between European and African populations, we first mapped genetic variants that are associated with gene expression levels (*i.e.*, eQTL) or transcript isoform usage (alternative splicing QTL: asQTL) in the complete sample. To do so, we used a linear regression model that accounts for population structure and principal components of the expression data, thus limiting the effect of unknown confounding factors (see Methods for details). Given that our sample size is too small to robustly detect *trans*-acting eQTL, we focused our analyses on local associations which, for simplicity, we refer to as *cis*-eQTL. We define *cis*-eQTL and *cis*-asQTL here as SNPs located in the gene body or in the 100 kb flanking the gene of interest.

We identified *cis*-eQTL for 1,647 genes (14% of all genes tested; FDR<0.01) in at least one of the experimental conditions (875 in non-infected macrophages, 1,087 in the *Listeria*-infected condition, and 983 in the *Salmonella*-infected condition: **Figure 3A**, **Table S4A,** and **Figure S3A** for number of eQTL found at more relaxed cutoffs).

Similarly, we detected a large number of *cis*-asQTL affecting the ratio of alternative isoforms used for the same gene (1,120 genes, 10% of all genes tested; FDR<0.01, **Figure 3A**, **Table S4C**: 886 in non-infected macrophages, 746 in *Listeria*-infected samples, and 615 in *Salmonella*-infected samples).

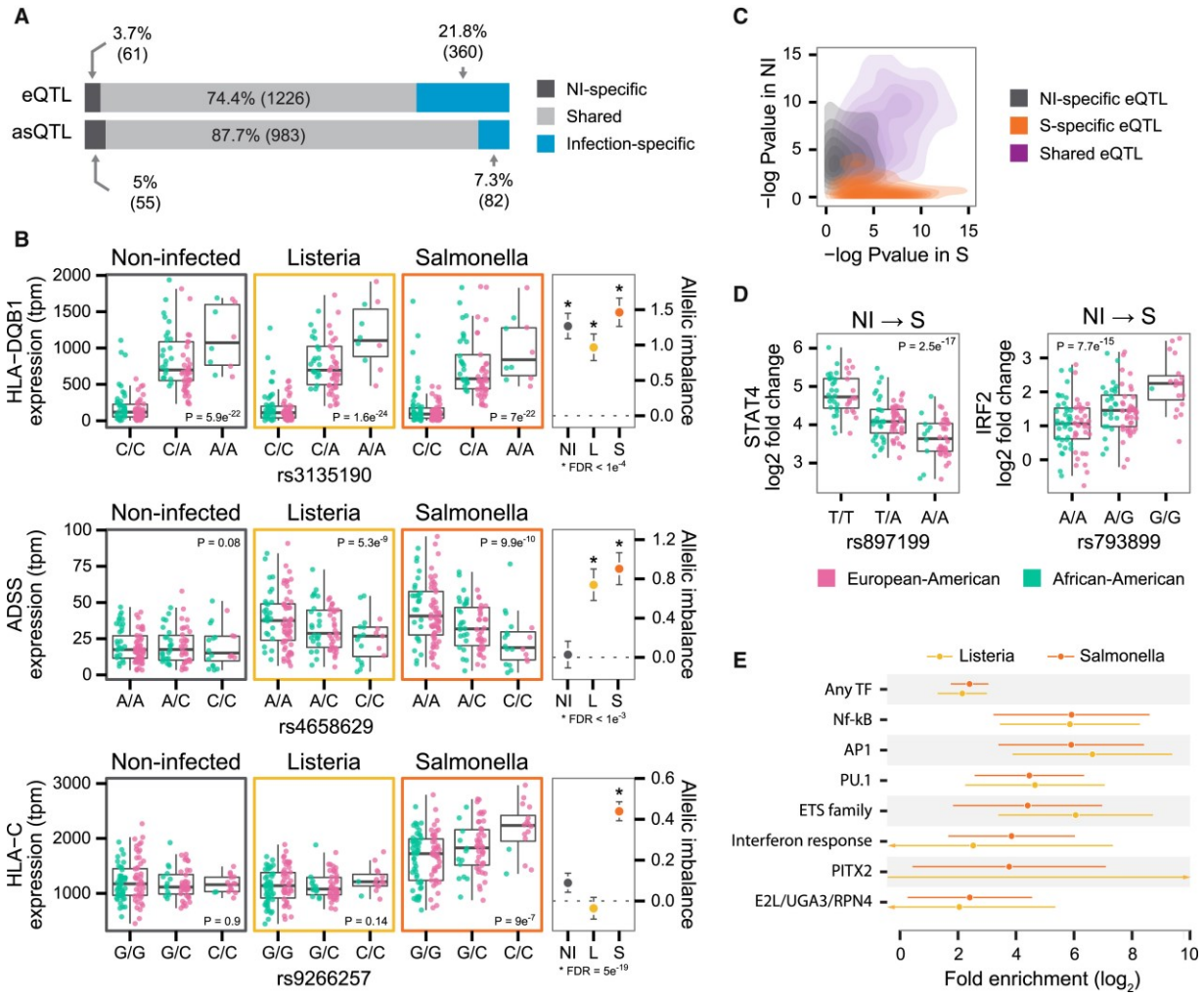

Figure 3. EQTL AND ASE ANALYSES REVEAL EXTENSIVE *CIS*-REGULATION OF GENE EXPRESSION RESPONSES TO PATHOGENS IN MACROPHAGES

A. Schematic representation of the number of *cis*-eQTL and *cis*-asQTL shared across all conditions, or only found in non-infected macrophages or *Listeria* and/or *Salmonella* infected macrophages. Infection-specific eQTL were defined as those showing very strong evidence of eQTL in the infected state (FDR always lower than 0.01), and very limited in the non-infected state (FDR always higher than 0.3). **B.** Examples of a *cis*-eQTL observed in all conditions (*HLA-DQB1*), a *cis*-eQTL observed only in infected macrophages (*ADSS*), and a *cis*-eQTL observed only in *Salmonella*-infected macrophages (*HLA-C*). Pink and green dots inside the boxplots distinguish African (>75% African ancestry) and European (<25% African ancestry) samples, respectively. **C.** Plot contrasting the evidence for ASE (-log10 p values) in non-infected macrophages (y axis) and in macrophages infected with *Salmonella* (x axis), for genes where we identified *cis*-eQTL in both conditions (purple), genes for which *cis*-eQTL were only found in non-infected macrophages (gray), and genes for which *cis*-eQTL were only found in

Salmonella-infected macrophages (orange). Qualitatively similar results were obtained when contrasting non-infected and *Listeria*-infected cells (Figure S3C). **D.** Examples of two *cis*-reQTL where genotype (x axis) has a significant effect on the response of *STAT4* (left) and *IRF2* (right) to *Salmonella* infection. **E.** reQTL enrichments (x axis) in actively regulated TF binding sites annotated by ATAC-seq footprinting. Error bars show 95% confidence intervals. Binding sites were grouped into functionally overlapping "TF clusters" using sequence similarity and co-localization in the genome (Table S6; STAR Methods). See also Table S4

Out of all genes with *cis*-eQTL, a large fraction (21.8%) were associated with an eQTL *only* in infected macrophages. In contrast, only 7.3% of genes showed an infection-specific *cis*-asQTL (**Figure 3A** and **B)**. Infection-specific *cis*-eQTL were further supported by analysis of allele-specific expression (ASE) levels, which provides independent but complementary evidence for functional *cis*-regulatory variation. As expected, genes with *cis*-eQTL were significantly enriched for genes with ASE, compared to the background of all 9,588 genes tested (**Figure S3B**, Fisher's exact test, $P < 1\text{x}10^{-15}$ for all conditions). Further, genes harboring infection-specific eQTL also tended to exhibit infection-specific ASE in the same condition (*Listeria* or *Salmonella*) in which the eQTL was identified (**Figure 3C,** ~27 fold-enrichment of infection-specific ASE among infection-specific eQTL, relative to shared eQTL; $P < 1\text{x}10^{-15}$). Thus, in agreement with previous studies (Fairfax et Knight 2014; Lee et al. 2014), genotype-environment (G x E) interactions are common in the context of immune responses to infection, albeit more so for mean expression levels than alternative isoform usage.

A complementary approach to identifying G x E interactions for expression levels is to directly map response eQTL (reQTL): QTL associated with the magnitude of *change* in expression levels after infection (Barreiro et al. 2011; Çalışkan et al. 2015; Lee et al. 2014). In contrast to condition-specific eQTL (an extreme case of G x E interaction), reQTL can capture more subtle interaction effects: eQTL can be shared across conditions as long as their effect size differs between infected and non-infected samples. We detected 244 and 503 genes with a *cis*-reQTL (FDR<0.01, **Table S4B**) for the response to *Listeria* and *Salmonella*, respectively. Interestingly, among genes

associated with a *cis*-reQTL, we found several key regulators of the immune response, including the transcription factors *STAT4* and *IRF2* (**Figure 3D**). We also found *cis*-reQTL for known susceptibility loci for ulcerative colitis (e.g., *HLA-A*, *HLA-DQA2*, *PMPCA*), systemic lupus erythematosus (*ITGAX*, *HLA-DQA1*), and the infectious diseases hepatitis B and leprosy (e.g., *HLA-C*, *NOD2*).

To investigate the regulatory mechanisms that account for immune reQTL, we next profiled the genome-wide chromatin accessibility landscape of non-infected and *Listeria* and *Salmonella*-infected cells using ATAC-seq (Buenrostro et al. 2013). This approach allowed us to identify transcription factor (TF) binding motifs likely to be occupied by their respective TFs, in both conditions (see Methods). We found that SNPs within accessible TF binding sites were >4 times more likely to be identified as reQTL (**Figure 3E**). Further, reQTL in our analyses were strongly enriched (>20-fold) for PU.1 binding sites (a pioneer TF involved in regulating enhancer activity in macrophages (Garber et al. 2012)), and for virtually all TFs that orchestrate innate immune responses to infection (**Figure 3E**) (e.g., NF-kB: >50-fold; AP1: >55-fold; and IRFs: 14-fold for *Salmonella* only). In striking contrast, we found no such enrichment for eQTL identified in non-infected macrophages ($P > 0.05$ for NF-kB, AP1, and IRFs, **Figure S3D**). These results show that reQTL variants are often conditionally silent in resting macrophages but become functionally relevant post-infection, and that this transition is explained by disruption of binding sites for immune response-activated TFs.

Genetic basis of ancestry-associated differences in the immune response to pathogens

We hypothesized that differences in allele frequencies for some of the eQTL identified above could explain the observed ancestry-associated differences in the

transcriptional response to infection. In support of this hypothesis, we found that pop-DE genes were enriched up to 3.3-fold for genes with *cis*-eQTL (P <1x10^{-10}), and that pop-DR genes were enriched up to 5.8-fold for genes with *cis*-reQTL (P <1x10^{-10}) (**Figure 4A**, **Figure S4A**). Additionally, ~60% of genes that exhibited ancestry-associated isoform usage were associated with an asQTL (up to 24-fold enrichment, P <1x10^{-10}). Thus, although rare, ancestry-associated changes in isoform usage are largely genetically driven.

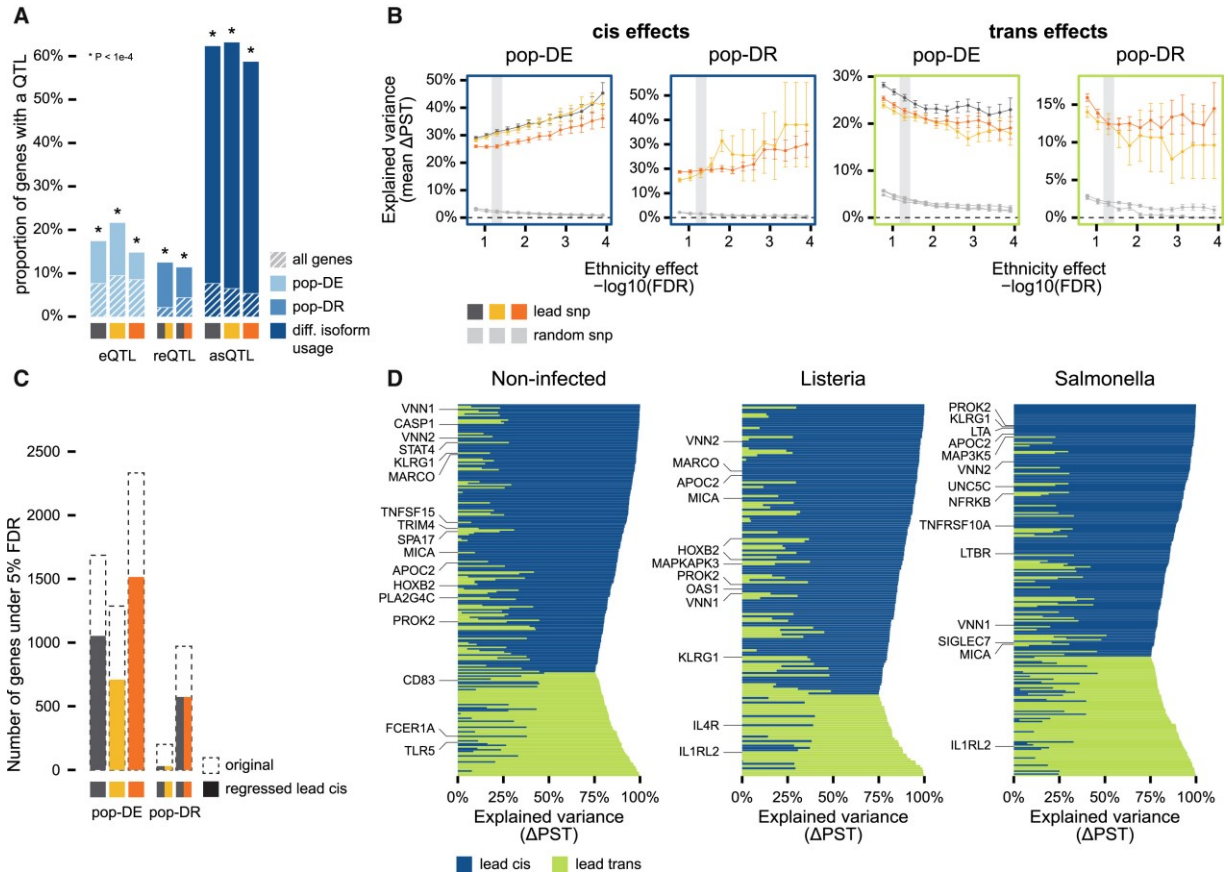

Figure 4. CONTRIBUTION OF CIS AND TRANS GENETIC VARIATION TO POP-DE AND POP-DR GENES

A. The proportion of pop-DE, pop-DR, and genes that exhibit ancestry-associated isoform usage that are associated with a *cis*-eQTL, *cis*-reQTL, or *cis*-asQTL, respectively (FDR < 0.01). Null expectations (based on the genome-wide proportion of genes associated with each QTL class) are shown in gray. Similar results are obtained when focusing on transcriptional QTL identified at an FDR of 0.05 (Figure S4B). **B.** Average ΔP_{ST} obtained (±SE) when regressing out the genotype effect of the lead *cis*- or the lead *trans*-SNP for pop-DE and pop-DR genes (y axis), defined using progressively stringent FDR cutoffs (x axis). Colored lines show average ΔP_{ST} values based on the real data; gray lines show the same values when regressing out the genotype effect of the lead SNP identified based on permuted genotypes. **C.** Number of genes identified as pop-DE and

pop-DR at an FDR < 0.05 (y axis) before (dashed bars) and after (filled bars) regressing out the effect of the lead *cis*-SNP associated with these genes. **D.** Examples of genes for which the lead *cis*- (blue) or the lead *trans*-SNP (green), explains at least 75% of the differences in gene expression associated with African versus European ancestry.

To explicitly quantify the contribution of our eQTL set to transcriptional differences detected between populations, we devised a new score based on P_{ST} estimates (Leinonen et al. 2013; Pujol et al. 2008). P_{ST} is the phenotypic analog of the population genetic parameter F_{ST}, providing a measure of the proportion of overall gene expression variance explained by between-population phenotypic divergence (as opposed to within-population diversity). P_{ST} values range from 0 to 1, with values close to 1 implying that the majority of a gene's expression variance is due to differences between populations. Our score, deltaP$_{ST}$ (ΔP$_{ST}$), is defined as the difference between P$_{ST}$ values before and after regressing out the effect of the *cis*-SNP that was most strongly associated with the target gene's expression level (regardless of significance level), divided by the P$_{ST}$ value observed before removing the genotype effect. ΔP$_{ST}$ therefore quantifies the proportion of ancestry-associated expression level differences that stem from the strongest *cis*-associated variant.

Among all pop-DE genes, we found that *cis*-regulatory variants explained an average of 31%, 31% and 26% of ancestry-related differences in expression observed in non-infected, *Listeria*-infected and *Salmonella*-infected samples, respectively (**Figure 4B)**. Further, the larger the effect of ancestry in the original pop-DE analysis, the larger the contribution of *cis*-regulatory variation to these differences: for pop-DE genes identified at a stringent FDR of $1x10^{-4}$, *cis*-regulatory variation explained close to 50% (on average) of ancestry effects (**Figure 4B)**. We observed a similar pattern for pop-DR genes after regressing out the genotype effect of the lead cis-reQTL SNP (**Figure 4B**). In support of the substantial role of *cis*-regulatory variation in explaining pop-DE and pop-DR genes, gene expression values for 30% and 45% of pop-DE and pop-DR

genes, respectively, were no longer significantly associated with ancestry once we regressed out *cis*-genetic effects (**Figure 4C**). Importantly, ΔP_{st} values never exceeded 5% when we regressed out either *(i)* the genotype effect of randomly selected SNPs matched for the allele frequency of the lead *cis*-SNP, or *(ii)* lead *cis*-SNPs identified after permuting the genotype data. Thus, our results cannot be simply explained by population structure (**Figure 4B)**.

Based on their known importance in the genetic control of gene regulation and because of power limitations, our main analysis of ancestry-associated gene expression patterns focused on the role of *cis*-eQTL. However, in a separate analysis, we recalculated ΔP_{ST} using the lead *trans*-SNP for each gene in place of the lead *cis*-SNP (although only 51, 21 and 22 *trans*-eQTL genes survived genome-wide multiple testing correction (FDR<0.1) in non-infected, *Listeria*-infected and *Salmonella*-infected samples, respectively). Intriguingly, we found that lead *trans*-SNPs accounted for an average of ~23% and ~20% of ancestry effects on gene expression levels for pop-DE and pop-DR genes, respectively (**Figure 4B;** at least twofold more than estimates based on permuted data, $P < 1x10^{-10}$). These results suggest that lead *trans*-SNPs, although difficult to detect at a genome-wide significance level, are enriched for true *trans*-associations that could be resolved with larger sample sizes. Together, a single *cis* or *trans*-acting variant was sufficient to explain almost all ancestry effects ($\Delta P_{ST} > 75\%$) on gene expression levels for 804 pop-DE genes and pop-DR genes (**Figure 4D)**, including for master regulators of the immune response such as *CASP1*, *STAT4* and *MICA*. Our results thus provide a comprehensive genome-wide map of *cis*- and *trans*-genetic variants associated with African and European ancestry-related differences in the immune response to infection.

Natural selection and genetic ancestry effects on gene expression divergence

Finally, we sought to determine the impact of recent local positive selection in either African or European populations on ancestry-related divergence in gene expression levels. To do so, we first calculated F_{ST} values between the Yoruba African (YRI) and the western European population (CEU) in Phase 3 data from the 1000 Genomes Project (Auton et al. 2015). To generate gene-specific estimates, we averaged F_{ST} values for variants within a window of 10 kb around the transcription start site (TSS) of each gene we analyzed (11,914 genes). As a complementary approach, we also calculated Integrated Haplotype Scores (iHS) for all SNPs with a minor allele frequency (MAF) >5% in the CEU and YRI samples. In contrast to F_{ST}, iHS is a *within*-population measure of recent positive selection that is not affected by the levels of population differentiation (Voight et al. 2006).

Our analyses identified significantly higher mean F_{ST} values among genes that were pop-DE, pop-DR, or showing differences in isoform usage between populations ($P \leq 1\text{x}10^{-3}$; **Figure 5A** and **Figure S5A** for similar results when using alternative window sizes). Further, variants identified as *cis*-eQTL were significantly enriched (~2-fold) for high iHS values (*i.e.*, iHS > 99$^{\text{th}}$ percentile of genome-wide distribution, **Figure 5B,** $P<1\text{x}10^{-8}$), consistent with the importance of regulatory genetic variation in recent human evolution (Fraser 2013). *cis*-reQTL and *cis*-asQTL were even more strongly enriched among high iHS values (up to 3.6-fold; **Figure 5B**, $P<1\text{x}10^{-5}$).

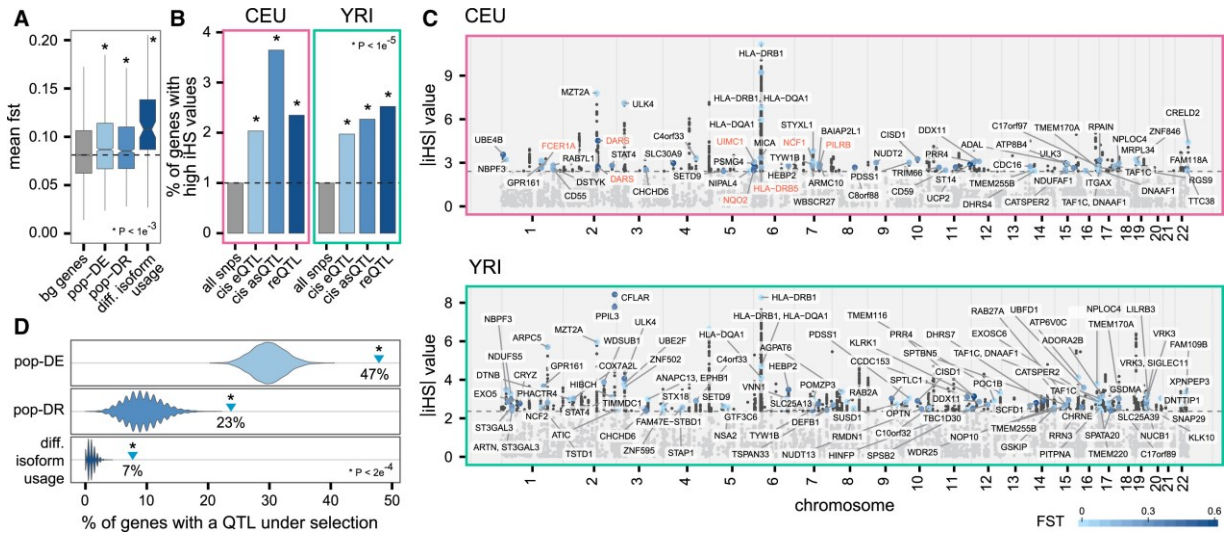

Figure 5. NATURAL SELECTION ON EQTL AND ITS CONTRIBUTION TO ANCESTRY-ASSOCIATED REGULATORY DIFFERENCES

A. Mean F_{ST} values in a window of ±10 kb around the TSS of all genes, pop-DE genes, pop-DR genes, and genes showing differences in isoform usage between populations (top panel). **B.** Proportion of all SNPs, *cis*-eQTL, *cis*-reQTL, and *cis*-asQTL with an iHS value above the 99th percentile of the genome-wide distribution in the CEU (|iHS| > 2.70) and the YRI (|iHS| > 2.68) populations. See Figure S5B for similar results considering QTL identified at an FDR < 0.05 (instead of 0.01). **C.** Manhattan plot of a genome-wide scan for selection in CEU (top) and YRI (bottom) for SNPs identified as regulatory QTL in macrophages. The dashed line represents the 99th percentile of the genome-wide distribution. Darker shades of blue represent larger F_{ST} values for SNPs with elevated |iHS| values; blue circled dots highlight genes that show one or more transcriptional associations with African versus European ancestry. Genes in red are regulated by NLS with elevated |iHS| values in CEU (|iHS| > 2.7), supporting adaptive introgression from Neanderthals into the ancestors of modern Europeans. **D.** Proportion of genes regulated by eQTL and targeted by recent positive selection (among the 258 represented by the blue circles in C) that are pop-DE, pop-DR, or show population differences in isoform usage (blue triangles), compared to random expectations when sampling the same total number of genes 10,000 times from all genes tested (violin plots). See also Table S5.

Overall, within the set of *cis*-eQTL, *cis*-reQTL, or *cis*-asQTL-associated genes, 258 carried a signature of recent positive selection in either CEU or YRI samples (|iHS| ≥ 99th percentile of the genome-wide distribution: **Figure 5C, Table S5A**). These variants were also significantly enriched for high XP-EHH values (Sabeti et al. 2007) (~6-fold, $P<1x10^{-10}$, **Figure S5C**), further supporting that these variants have been important in recent, population-specific human adaptation. However, because outlier methods for detecting selection can be susceptible to false positives (Kelley et al. 2006), we complemented our iHS analysis with a model-based approach. Specifically, we compared the observed iHS value for each putatively selected allele to those observed

under neutral coalescent simulations matched to the known demographic history of African and European populations (Gutenkunst et al. 2009), the candidate allele's observed frequency, and the local recombination rate. The vast majority (92%) of all sites tested exhibited significantly larger observed iHS statistics than expected under a neutral model ($P<0.01$, **Table S5B**), providing strong convergent support for recent positive selection at these loci. Far more of these genes are pop-DE or pop-DR than expected by chance (47% and 23%, respectively: **Figure 5D**, $P < 0.001$), showing that natural selection has contributed to present-day inter-population differences in innate immune responses to infection.

Neanderthal ancestry makes up approximately 2% of the ancestry of living humans found outside of Africa (Kelso et Prüfer 2014). It is therefore plausible that interbreeding between Neanderthal and modern human populations could also contribute to some of the ancestry-related differences in gene expression we observed, especially if it enabled the ancestors of modern Europeans to more rapidly adapt to a new pathogen environment (Ségurel et Quintana-Murci 2014). To test this hypothesis, we identified sites where the derived allele is shared between Neanderthals and non-African populations, but is absent in sub-Saharan Africans samples considered. This class of sites, which we call "Neanderthal-like sites" (NLS), is a conservative indicator of Neanderthal introgression (Sankararaman et al. 2014). Among the 18,862 NLS tested in our *cis*-QTL analyses, 297 were significantly associated with transcriptional variation of 145 genes (NLS-QTL). Among these 145 genes, 46% (FDR<0.05) were differentially regulated in at least one experimental condition (non-infected, *Listeria*-infected, *Salmonella*-infected, or in the response to either type of infection) between Europeans and Africans (63% at a more relaxed FDR < 0.1). Thus, a non-negligible proportion of ancestry-related gene expression divergence probably results from introgression of

functional Neanderthal variants into the ancestors of modern Europeans. Interestingly, some of these variants (n=16) also have elevated iHS values (|iHS|≥2, **Figure 5C**, **Table S5A**) and therefore represent new candidates for adaptive introgression in humans.

Discussion

Together, our results provide a comprehensive characterization of genes for which the transcriptional responses of primary cells to live pathogenic bacteria differs depending on European versus African ancestry. We show that 34% of genes expressed in macrophages show at least one type of ancestry-related transcriptional divergence, whether in the form of differences in gene expression (30%), the transcriptional response to infection (9.3%), or, less commonly, differences in isoform usage (1%). Notably, the modest contribution of differences in isoform usage to ancestry-related expression levels differs from previous results in lymphoblastoid cell lines (LCLs), where they were found to be quite common (Lappalainen et al. 2013). The discrepancy between our results and those reported for LCLs may be related to differences in the experimental procedures used to produce the two sets of LCL lines, which were generated more than 20 years apart (Dausset et al. 1990).

One of the most striking observations from our study was the markedly stronger response to infection induced in macrophages from individuals of African descent, particularly among inflammatory response genes. This result agrees with previous reports showing that AAs have higher frequencies of alleles associated with an increased pro-inflammatory response (Ness et al. 2004), increased levels of circulating C-reactive protein (Kelley-Hedgepeth et al. 2008), and a much higher rate of inflammatory diseases than EA individuals (Pennington et al. 2009). Although the exact causal link between

ancestry and the pro-inflammatory response has yet to be established, we speculate that the stronger inflammatory response associated with African ancestry accounts for the increased ability of macrophages in African ancestry individuals to control bacterial growth post-infection.

Nevertheless, the evolutionary pressures that explain these differences remain an open question. One possibility is that, after human populations migrated out of Africa, they were exposed to lower pathogen levels (Guernier, Hochberg, et Guégan 2004), which reduced the need for strong, costly pro-inflammatory signals. Change in this direction may have been favored due to the detrimental consequences of acute or chronic inflammation, which are key contributors to the development of autoinflammatory and autoimmune diseases (Okin et Medzhitov 2012). This hypothesis is consistent with previous reports showing a signature of positive selection in Europeans on a high-frequency non-synonymous variant in the Toll-like receptor 1 gene, which is also associated with impaired NF-κB-mediated signaling (Barreiro et al. 2009). Alternatively, the weaker inflammatory response detected in Europeans could have resulted from relaxation of selective constraint in an environment where the pathogen burden was reduced, or at least different in nature, from that found in Africa.

Because our samples were derived from individuals with their own, unknown life histories and environmental exposures, the ancestry-related differences we observed could be explained by both environmental and genetic factors. However, our eQTL analyses suggest that genetic contributions are probably substantial. We estimate that, on average, about 30% and 20% of ancestry-associated expression differences in pop-DE genes are accounted for by *cis*- and *trans*-regulatory variants, respectively. Further, among the genes with the most robust association with genetic ancestry (pop-DE genes with FDR $< 1\text{x}10^{-4}$), putatively *cis*-acting variants explain up to ~50% of ancestry

effects. Notably, these numbers probably underestimate the true genetic contribution to ancestry-related differences in gene expression, given our low power to detect *trans* associations; our exclusion of non-SNP regulatory variants, which may also influence gene expression (Gymrek et al. 2016); our conservative assumption that genes have only one major *cis*-eQTL (many genes have at least two independent *cis*-eQTL: (Lappalainen et al. 2013)); and the fact that we limited our *cis*-eQTL mapping to an 100kb window around the targeted gene.

The extent to which positive selection has contributed to recent human evolution remains a matter of intense debate (Enard, Messer, et Petrov 2014; Fagny et al. 2014; Hernandez et al. 2011). Here we show that variants associated with regulatory QTL are strongly enriched for signatures of recent selection, supporting an important role of adaptive regulatory variation in recent human evolution. More specifically, our results suggest that a significant fraction of population differences in transcriptional responses to infection are a direct consequence of local adaptation driven by regulatory variants. Notably, several positively selected regulatory QTL (or SNPs in strong LD with them (r^2>0.8)) have been associated with common diseases by GWAS, further reinforcing the link between past selection and present-day susceptibility to disease (Barreiro et Quintana-Murci 2010; Brinkworth et Barreiro 2014). Some examples include positively selected variants affecting the expression of *HLA-DQA1*, the major genetic susceptibility factor for celiac disease (Abadie et al. 2011); *ERAP2*, a susceptibility factors for Crohn's disease (Jostins et al. 2012); and the transcription factor *IRF5*, which is associated with systemic lupus erythematosus, rheumatoid arthritis, ulcerative colitis and systemic sclerosis (reviewed in: (Eames, Corbin, et Udalova 2016)).

Finally, our results provide new insight into the contribution of adaptive introgression from admixture with Neanderthals to the diversification of the immune

system among modern human populations. We found 20 positively selected NLS regulatory-QTL (associated with 17 genes) that are candidates for adaptive introgression in humans. These genes include previously identified candidates such as *TLR1* (Dannemann, Andrés, et Kelso 2016; Deschamps et al. 2016), but also a large set of loci that have not previously been associated with adaptive introgression. For example, one of the strongest signatures of selection was found for eQTL for *DARS*, a gene associated with neuroinflammatory and white matter disorders (Wolf et al. 2015). However, in agreement with evidence that most introgressed variation from Neanderthals was probably deleterious (Sankararaman et al., 2014; Vernot and Akey, 2014) (Sankararaman et al. 2014; Vernot et al. 2014), as putative cases of adaptive introgression remain relatively rare.

All data generated in this study are freely accessible via a custom web-based browser that enables easy querying and visualization of ancestry-related transcriptional differences and associated QTL. This resource, the ImmunPop QTL browser (http://www.immunpop.com), should serve as a useful tool for fine mapping of genetic association signals and for the continued quest to understand how pathogens have shaped global human population diversity today.

Author contributions

L.B.B. conceived and directed the study. A.D., A.P.S, V.Y. and F.L., performed experimental work, Y.N., J.S., and G.B. led the computational analyses with contributions from Z.A.S, A.F., A.J.A., R.B., R.D.H, R.P.R, and L.B.B. Y.N. developed and implemented the ImmunPop QTL browser with help from S.H. L.B.B., J.T., G.B., and J.S. wrote the paper, with input from all authors.

Acknowledgments

We thank members of the Barreiro lab for helpful discussions and comments on the manuscript; Danielle Malo for a gift of the *Listeria* and *Salmonella* strains used in this study; and L. Tailleux for advice on the infection experiments. This study was funded by grants to L.B.B. from the Canadian Institutes of Health Research (301538 and 232519), the Human Frontiers Science Program (CDA-00025/2012) and the Canada Research Chairs Program (950-228993). We thank Calcul Québec and Compute Canada for providing access to the supercomputer Briaree from the University of Montreal. Y.N. was supported by a fellowship from the *Réseau de Médecine Génétique Appliquée* (RMGA); J.S by a fellowship from the *Fonds de recherche du Québec - Nature et technologies* (FRQNT), and G.B. and J.S by a fellowship from the *Fonds de recherche du Québec –Santé* (FRQS).

Data and Software Availability

Software

All the Software packages and methods used in this study have been properly detailed and referenced under "QUANTIFICATION AND STATISTICAL ANALYSIS."

Data Resources

All data generated in this study are freely accessible via the ImmunPop QTL browser (http://www.immunpop.com). RNA sequencing data reported in this paper is available in GEO: GSE81046.

Methods

Voir annexe

Chapitre 3 : Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans

Contexte de l'article

Cet article explore l'impact du métissage des populations européennes avec Néandertal sur l'expression de gènes antiviraux. Nous avons recherché des signaux de sélection positive suggérant que cette région ait participé à l'adaptation des populations européennes à leur environnement.

L'essentiel du travail fut réparti entre les 3 premiers co-auteurs :

- Aaron J. Sams a réalisé le volet évolutif de l'article à travers l'identification des haplotypes de Néandertal, les modèles démographiques et les simulations de coalescence ;
- Anne Dumaine s'est occupée des expérimentations de laboratoire pour valider les niveaux d'expression des gènes et transcrits d'intérêt ;
- J'ai réalisé les analyses eQTL et asQTL et mis en évidence les effets de l'haplotype issu de Néandertal sur l'expression des gènes et transcrits d'intérêt ;

L'article a été soumis à *Genome Biology* le 28 mai 2016, accepté le 4 novembre et publié le 29 novembre 2016.

Les tableaux et données associées sont disponibles librement sur le site de l'éditeur : http://doi.org/10.1186/s13059-016-1098-6

Authors

Aaron J. Sams[1,* †], Anne Dumaine[2†,] Yohann Nédélec[2,3†], Vania Yotova[2], Carolina Alfieri[2,4], Jerome E. Tanner[2], Philipp W. Messer[1,*‡], Luis B. Barreiro[2,5,* ‡]

Affiliations

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY

[2]Department of Genetics, Sainte-Justine Hospital Research Centre, University of Montreal, Montreal, Qc, Canada

[3]Department of Biochemistry, University of Montreal, Montreal, Qc, Canada

[4]Department of Microbiology and Immunology, University of Montreal, Montreal, Qc, Canada

[5]Department of Pediatrics, University of Montreal, Montreal, Qc, Canada

[†]These authors contributed equally to this work

[‡]These authors jointly supervised this work

[*]Corresponding author

Abstract

Background

The 2'-5' oligoadenylate synthetase (OAS) locus encodes for three OAS enzymes (OAS1-3) involved in innate immune response. This region harbors high amounts of Neandertal ancestry in non-African populations; yet, strong evidence of positive selection in the OAS region is still lacking.

Results

Here we used a broad array of selection tests in concert with neutral coalescent simulations to demonstrate a signal of adaptive introgression at the OAS locus.

Furthermore, we characterized the functional consequences of the Neandertal haplotype in the transcriptional regulation of OAS genes at baseline and infected conditions. We found that cells from people with the Neandertal-like haplotype express lower levels of OAS3 upon infection, as well as distinct isoforms of OAS1 and OAS2.

Conclusions

We present evidence that a Neandertal haplotype at the OAS locus was subjected to positive selection in the human population. This haplotype is significantly associated with functional consequences at the level of transcriptional regulation of innate immune responses. Notably, we suggest that the Neandertal-introgressed haplotype likely reintroduced an ancestral splice variant of OAS1 encoding a more active protein, suggesting that adaptive introgression occurred as a means to resurrect adaptive variation that had been lost outside Africa.

Background

Whole genome sequencing of several archaic human genomes (Green et al. 2010; M. Meyer et al. 2012; Prüfer et al. 2014; Sawyer et al. 2015) representing Neandertals and an as yet geographically and paleontologically unknown population referred to as Denisovans has revealed gene flow between these populations and the ancestors of present-day humans. Neandertal ancestry makes up approximately 0.5-2 percent of the ancestry of most living humans, with higher amounts of Neandertal ancestry found outside of Africa (Sankararaman et al. 2014; Llorente et al. 2015; Vernot et al. 2014). While it seems that there may have been widespread purifying selection against Neandertal ancestry in humans (Prüfer et al. 2014; Vernot et al. 2014; Harris et Nielsen 2016) , some positive selection on Neandertal genes (adaptive introgression) has also been observed (Racimo et al. 2015; Racimo, Marnetto, et Huerta-Sanchez 2016).

Neandertals and other archaic populations inhabited Eurasia for several hundred thousand years (Hublin 2009) and were likely well adapted to their environments. Therefore, some genetic variation inherited from these archaic humans may have been adaptive in modern humans, particularly across phenotypes that are strongly influenced by direct interactions with the surrounding environment (Racimo et al. 2015), such as our immune response to infectious agents (Ségurel et Quintana-Murci 2014).

The OAS locus on chromosome 12, which harbors three genes (*OAS1, OAS2, OAS3*) encoding the 2'-5' oligoadenylate synthetase enzymes has received considerable attention due to its clear signatures of multiple archaic haplotypes in populations outside of Africa (Fernando L Mendez, Watkins, et Hammer 2012; Fernando L. Mendez, Watkins, et Hammer 2013), and the critical role of OAS genes in the innate immune response to viruses (Player et Torrence 1998). Mendez and colleagues (F L Mendez, Watkins, et Hammer 2012) first identified an introgressed haplotype of *OAS1* from Denisovans that is restricted to individuals in Indonesia and Melanesia. Later, these authors (Fernando L. Mendez, Watkins, et Hammer 2013) identified a Neandertal haplotype at the OAS locus that spans a ~190 kilobase region between two surrounding recombination hotspots.

The elevated frequency of the Neandertal derived alleles in the OAS locus (Figure 1) relative to average levels of Neandertal ancestry in Europeans (Sankararaman et al. 2014), along with the key role OAS genes play in protective immunity against viral infections raises the possibility that introgressed Neandertal haplotypes at OAS may have been adaptive in modern humans. While some studies provide suggestive evidence of adaptive introgression at the OAS locus (Sankararaman et al. 2014; Racimo, Marnetto, et Huerta-Sanchez 2016), strong evidence of positive selection in the OAS region is still lacking. Indeed, several studies failed to reject a model of neutral evolution

for the Neandertal haplotype when using standard neutrality tests (Fernando L. Mendez, Watkins, et Hammer 2013; Deschamps et al. 2016).

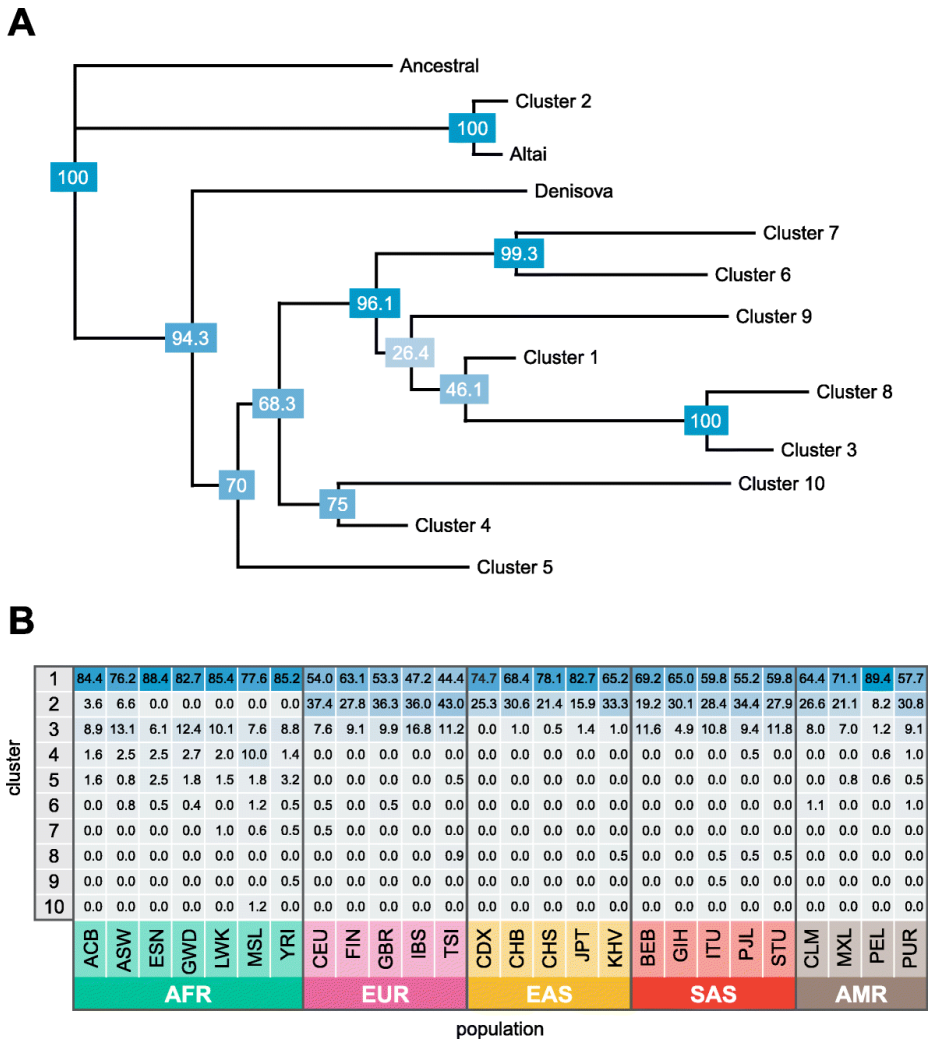

Figure 1. NEANDERTAL INTROGRESSSED HAPLOTYPES IN THE OAS REGION

a *Neighbor-joining tree* of 5008 phased haplotypes spanning chr12: 113344739–113449528 (hg19) from phase 3 of the 1000 Genomes Project. Haplotypes were condensed into 10 core haplotypes based on the majority allele after clustering into groups of haplotypes with pairwise differences of 60 or less. The *figure* illustrates that the Altai haplotype is very similar to "cluster 2" haplotypes found in several human populations, while the Denisovan haplotype is not more closely related to any of the remaining (non-cluster 2) sequences in this dataset. Bootstrap values (1000 replicates) are provided in *blue boxes* at each node. b Frequencies of the 10 core haplotypes within each 1000 Genomes Project population sample. The most common Neandertal-like haplotype, cluster 2, is found only outside of sub-Saharan African samples, with the exception of recently admixed populations. Population codes can be found in Additional file 2: Table S7

We hypothesize that the overall lack of signals of selection in the OAS region stems from the low power of standard neutrality tests to detect adaptive introgression

(Sankararaman et al. 2016). Here we circumvent this issue by testing the hypothesis of adaptive introgression using extensive neutral coalescent simulations specifically tailored to match the genomic features of the OAS region in combination with several empirical observations of Eurasian genetic variation, and ancient DNA data from Eurasia. We firmly demonstrate a population genetic signal of adaptive introgression at the OAS locus and characterize the functional consequences of the Neandertal haplotype in the transcriptional regulation of OAS genes in macrophages and peripheral blood mononuclear cells (PBMCs) at baseline and infected conditions. We note that the term 'adaptive introgression' will be used in a broad sense, as our tests do not allow us to determine the exact timing when selection started to act on the Neanderthal alleles.

Results

Mendez et al. (Fernando L. Mendez, Watkins, et Hammer 2013) previously reported evidence of archaic introgression at the OAS locus. We first verified these results, using data from phase 3 of the 1000 Genomes Project (Auton et al. 2015), by examining the relationships of all modern human sequences at the OAS locus with the Altai Neandertal, Denisovan, and an inferred ancestral sequence. Clustering the human haplotypes resulted in 10 consensus sequences representing human haplotype clusters, which we combined with the archaic sequences in a neighbor-joining tree (Methods, Figure 1A, Additional file 1: Figure S1). This tree confirms that the Denisovan haplotype is not present in the population samples represented in the 1000 Genomes Project dataset. In contrast, the Neandertal haplotype is found at relatively high frequencies outside of Africa, reaching highest frequencies in European population samples (up to 43%, Figure 1B). Additionally, we corroborate the finding of Mendez

and colleagues (Fernando L. Mendez, Watkins, et Hammer 2013) that Neandertal-like haplotypes in the OAS region are too long to have resulted from incomplete lineage sorting (ILS) ($P \leq 2x10^{-3}$, Methods).

To investigate the hypothesis of non-neutral evolution at the OAS locus we first tested whether the observed frequencies of Neandertal-like sites (NLS) in the OAS region are higher than expected under neutrality. We defined NLS as bi-allelic SNPs with derived alleles that are shared between Neandertals and a non-African population sample, but absent in a sub-Saharan African sample (Sankararaman et al. 2014; Fu et al. 2014; Yang et al. 2012). Specifically, we simulated the expected allele frequency of NLS under neutrality, using a demographic model based on previously inferred parameters of human demographic history (Gravel et al. 2011; Tennessen et al. 2012; Vernot et Akey 2015) (Additional file 1: Figure S1, Additional file 2: Table S1). We considered both a model with a single pulse of Neandertal introgression occurring over a span of 500 years into the ancestral Eurasian population after their population split from Africa, and two additional two-pulse models (Additional file 1: Figure S1, Additional file 2: Table S1). We found that in all European populations (with the exception of Finnish (FIN)), the highest frequency NLS fall in the extreme 1% of all simulations and are significantly elevated in frequency (all five European samples pass FDR < 0.05, see Additional file 2: Table S2), regardless of whether we assume a single-pulse introgression model (Figure 2A) or two-pulse introgression models (Additional file 1: Figure S3), indicating that the high frequencies of NLS found in most European populations likely reflects a history of adaptation.

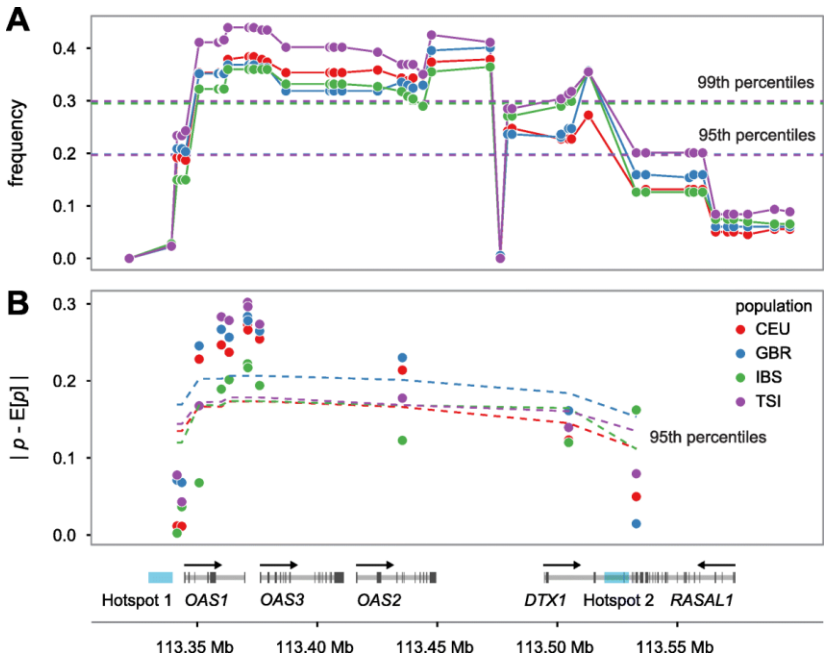

Figure 2. OAS-INTROGRESSED HAPLOTYPES ARE FOUND AT HIGHER FREQUENCIES IN EUROPEAN POPULATIONS THAN EXPECTED UNDER NEUTRALITY

a Comparison of frequency (*y-axis*) of NLS in the OAS locus in the CEU, GBR, IBS, and TSI European population samples with respect to neutral expectations (*dashed lines*) based on coalescent simulations. **b** Absolute difference between observed and expected allele frequency in the same four present-day European samples (*y-axis*) based on ancient DNA data. *Dashed lines* represent the 95th percentile of the expected distribution based on similar deviations calculated on a dataset of approximately 1 million SNPs scattered around the genome and with comparable present-day frequencies to those found for NLS in the OAS region

We next sought to examine the consistency of the simulation results described above using genetic data from a dataset of 230 ancient Eurasian individuals (Mathieson et al. 2015). Assuming neutrality, the expected frequency of an allele in contemporary European populations can be predicted as a linear combination of allele frequencies sampled from representative ancient populations that have contributed ancestry to present-day European populations in different proportions (Mathieson et al. 2015; Lazaridis et al. 2014) (Methods). Using this approach, we calculated the expected allele frequency in four present day European samples from the 1,000 Genomes Project (Auton et al. 2015) at the 11 NLS falling within the bounds of the three OAS genes, based on the ancient allele frequencies estimated by Mathieson and colleagues. To set

up our null expectations we performed a similar analysis on a dataset of approximately one million SNPs scattered around the genome, generated by Mathieson and colleagues (by merging 213 ancient samples dated between 6,500 and 300 BCE with sequencing data from four European samples from the 1,000 Genomes Project). We found that the NLS at OAS are outliers in the genome with respect to deviations from ancient frequencies. More specifically, we found that the allele frequencies of 6 out of the 11 OAS SNPs tested in the OAS1-OAS3 region (those 6 SNPs fall within the block of SNPs with NLS frequency greater than 99% of neutral simulations) have increased above the frequency predicted by ancient Eurasian samples by more than 20%, significantly more than what we observed for other SNPs genome-wide with comparable present-day frequencies (lowest $P = 0.00476$, Figure 2B, Additional file 2: Table S3). Our findings at this single locus are consistent with results from the genome-wide selection scan performed by Mathieson and colleagues (Mathieson et al. 2015) where the OAS region also showed evidence of selection ($P < 10^{-7}$, Additional file 2: Table S3), even if it did not reach genome-wide significance after multiple-test correction.

We next searched for additional evidence of recent selection, as measured by the iHS (Voight et al. 2006) and DIND (Barreiro et al. 2009) statistics. These tests share a similar rationale: an allele that has recently been driven to high population frequency by positive selection should be associated with unusually long-range LD (iHS) and reduced intra-allelic nucleotide diversity (DIND) (Voight et al. 2006; Barreiro et al. 2009; Sabeti, P.C. et al. 2002). We found that several NLS in the OAS region show significantly high iHS and DIND values with respect to genome-wide expectations (Figure 3A and Additional file 1: Figure S4), further supporting that they have been targeted by positive selection.

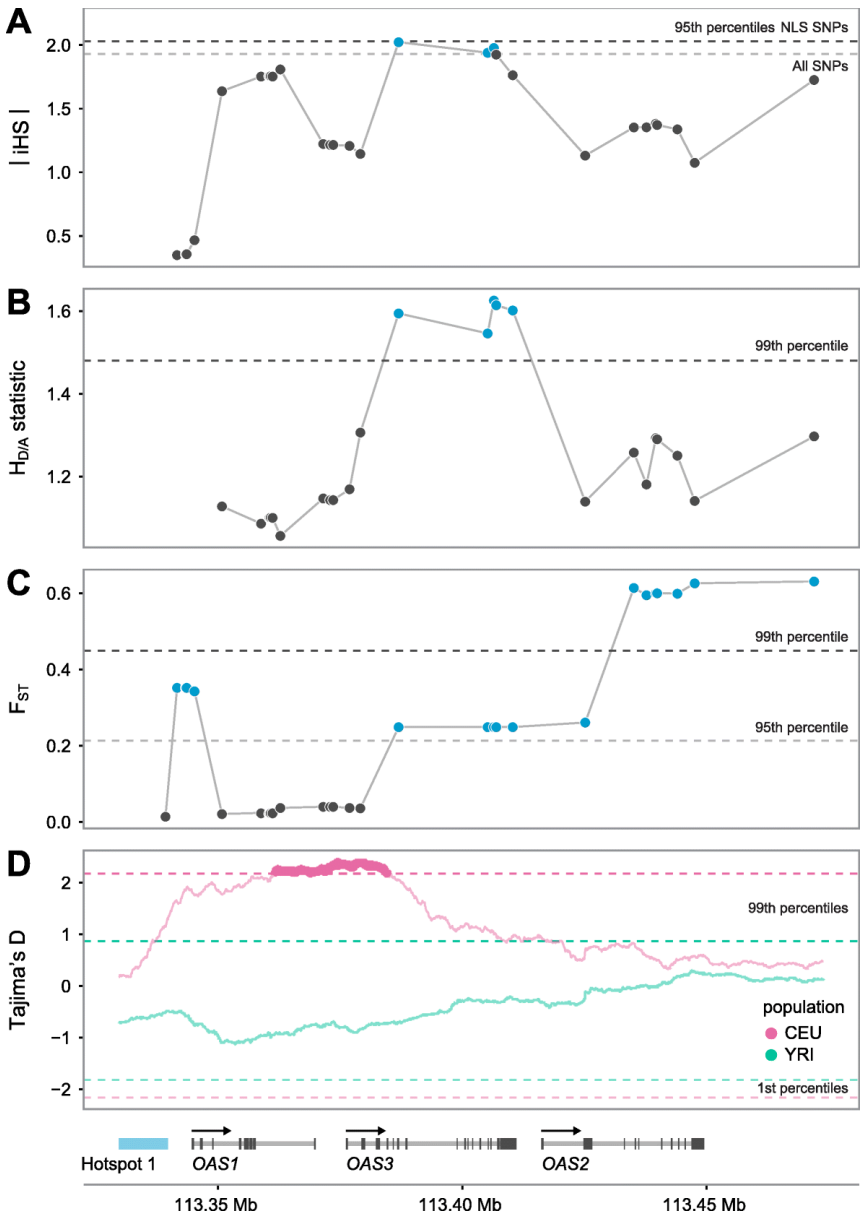

Figure 3. OAS-INTROGRESSED HAPLOTYPES SHOW MULTIPLE SIGNATURES OF POSITIVE SELECTION

a Normalized absolute iHS scores (*y-axis*) across SNPs in the CEU sample. *Dashed lines* represent the 95th percentile of iHS calculated genome-wide across all SNPs (*gray*) and all NLS SNPs (*black*). **b** The $H_{D/A}$ statistic in CEU (*y-axis*). *Dashed line* indicates the 99th percentile of $H_{D/A}$ calculated across 1000 simulations. **c** Fst calculated between CEU and all East Asian populations from the 1000 Genomes projects (*y-axis*). *Dashed lines* indicate the 95th (*gray*) and 99th (*black*) genome-wide percentiles of Fst. Similar results are obtained when comparing any other European population against East Asians (Additional file 2: Table S5). **d** Tajima's D (*y-axis*) calculated for CEU and YRI samples. *Dashed lines* indicate the 1st and 99th genome-wide percentiles of Tajima's D

As an additional means of understanding whether or not Neandertal haplotypes at OAS are longer than would be expected if they had evolved neutrally, we followed

up these empirical haplotype-based tests with a simulation approach using a simple test statistic, $H_{D/A}$. This statistic (fully defined in Methods) compares pairwise haplotype homozygosity lengths within the set of haplotypes carrying a derived (Neandertal) allele and within those carrying the ancestral allele. We utilized the same demographic models described above for our frequency-based test to understand whether the haplotypes surrounding derived alleles at NLS are longer than expected under a neutral model, conditional on the map of recombination in the OAS region. Again, we found that several derived Neandertal-like alleles are associated with a haplotype that is significantly longer than observed in simulated loci, and that these results were robust to a series of alternative simulation models tested, including the one- and two-pulse models described above and two additional one-pulse models which varied the mutation and recombination rates (Figure 3B, Additional file 2: Table S4). A caveat to this simulation-based approach is that the parameters of the true demographic history of Neandertal-modern human admixture remain uncertain. Therefore, comparisons of neutral demographic models of Neandertal introgression to tests of selection utilizing haplotype lengths, such as HD/A, are limited by the current state of demographic parameter estimates.

Finally, we calculated the levels of population differentiation between European and East Asian populations for all NLS in the OAS region. Interestingly, we found extreme levels of differentiation for most NLS (Fst as high as 0.6, $P_{empirical} < 0.01$, Figure 3C. Additional file 2: Table S5), except within the genomic region covering OAS1 and part of OAS3. These results suggest that NLS surrounding OAS1 have been positively selected in both European and East Asian populations but that distinct haplotypes, possibly recombinant forms of an original introgressed Neandertal haplotype, have been selected in Europe and Asia.

Our population genetic results provide evidence that Neandertal alleles at the OAS locus have likely experienced positive selection during one or several phases after their introduction into the human population, suggesting a possible functional role of these alleles in human innate immune responses. To study this possibility, we analyzed RNA-sequencing data collected on primary macrophages from 96 European-descent individuals, before and after in-vitro infection with Salmonella typhimurium. After 2 hours of infection, we found that all OAS genes were strongly up-regulated (up to 19-fold, $P < 1$ x 10-10, Additional file 1: Figure S5), confirming the ability of Salmonella to activate the interferon (IFN) production pathway (Nauciel et Espinasse-Maes 1992; Gulig et al. 1997; LaRock, Chaudhary, et Miller 2015). Using genotype data available for the same individuals (673 SNPs spanning the OAS region, see methods) we tested if NLS were associated with variation in the expression levels of OAS1, OAS2 or OAS3, in either infected or non-infected macrophages. We found that among NLS, individuals that are heterozygous or homozygous for the Neandertal allele show reduced expression levels of OAS3 (i.e., they were expression quantitative trait loci, or cis eQTL for OAS3) (Figure 4A, false discovery rate (FDR) $< 10\%$, Additional file 2: Table S6). Interestingly, these cis eQTL showed a much stronger effect in infected macrophages (best $P_{salmonella} = 3.5$x10-3 vs best $P_{non\text{-}infected} = 0.027$), supporting an interaction between the Neandertal haplotype and the OAS3 response to Salmonella infection.

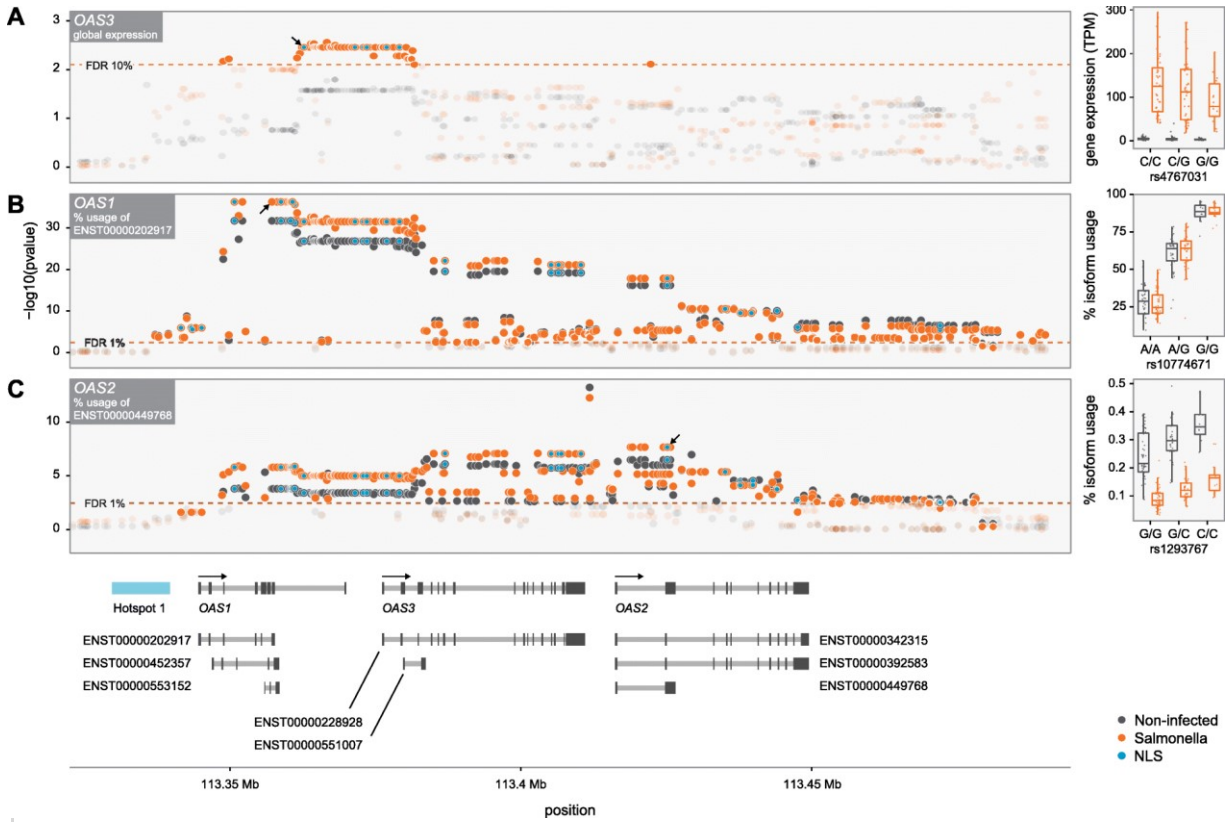

Figure 4. PERVASIVE IMPACT OF THE NEANDERTAL HAPLOTYPE ON THE REGULATION OF OAS GENES IN PRIMARY MACROPHAGES

a –log 10 Ps (*y-axis*) for the association between genotypes for SNPs with a MAF > 10% in the OAS region and expression levels of *OAS3* in non-infected (*black*) and Salmonella-infected macrophages (*orange*). The *dashed line* shows the *P* cutoff corresponding to an FDR of 10%. The *right panel* shows a *boxplot* for the association between genotypes at the NLS rs4767031 (*x-axis*) and the expression levels of *OAS3* in TPM (transcripts per kilobase million) (*y-axis*). The lower expression level of *OAS3* in individuals harboring the Neandertal haplotype were confirmed by real-time polymerase chain reaction (Additional file 1: Figure S11). **b** –log 10 Ps (*y-axis*) for the association between genotypes in the OAS regions and the percentage usage of isoform ENST00000202917 (i.e. p46 in the text) in non-infected (*black*) and Salmonella-infected macrophages (*orange*). The *dashed line* shows the *P* cutoff corresponding to an FDR of 1%. The *right panel* shows a *boxplot* for the association between genotypes at the splicing variant rs10774671 (*x-axis*) and the percentage usage of isoform p46 (*y-axis*). **c** Similar to (**b**) but for the percentage usage of isoform ENST00000449768 of OAS2. In all the panels NLS are highlighted by *blue dots*. The *arrows* on (**a**)–(**c**) highlight the location of the SNPs for which the *boxplots* are shown on the *right*

In addition to overall changes in expression, we took advantage of the power of RNA-sequencing data to test if NLS in the OAS regions influenced the ratio of alternative isoforms used for each of the OAS genes (i.e., alternative splicing QTL: asQTL). We found that SNPs associated with the Neandertal haplotype are significant asQTL for OAS1 and OAS2 in both infected and non-infected macrophages (FDR << 1%; Figure 4B-C). The effect of the splice site variant rs10774671 at determining what

isoform is primarily encoded by OAS1 was particularly strong ($P \leq 2x10^{-32}$). This SNP is also a strong asQTL and protein QTL in lymphoblastoid cell lines (Wu et al. 2013; Pickrell et al. 2010). The ancestral G allele at this SNP (A<u>G</u> at acceptor site) retains the splice site whereas the derived allele, A, (A<u>A</u> at acceptor site) disrupts the splice site leading to the usage of a distinct isoform (Additional file 1: Figure S6). The Neandertal haplotype harbors the ancestral allele (encoding the p46 isoform), which is associated with high enzyme activity (Bonnevie-Nielsen et al. 2005). Interestingly, despite the fact that this ancestral allele is found at ~ 60-70% in most Sub-Saharan African populations, outside of Africa this allele is only found on individuals with the Neanderthal haplotype (with rare exceptions; ~2% of all haplotypes, Additional file 1: Figure S2 suggesting that the Neandertal segment is effectively reintroducing an ancestral variant that was already present in other human groups. To validate that this SNP had the same impact at determining what OAS1 isoforms are expressed in African-descent individuals, we analyzed additional RNA-sequencing data collected from 41 African-American individuals. As among Europeans, rs10774671 is a strong asQTL for OAS1 in both non-infected and infected macrophages ($P < 1x10^{-15}$, Additional file 1: Figure S7), despite laying in a non-Neandertal haplotype.

Because OAS genes are primarily involved in the control of viral infections we decided to validate our functional findings on peripheral blood mononuclear cells (PBMCs) from 40 individuals stimulated/infected with viral-ligands (polyI:C and gardiquimod), and live viruses (Influenza, Herpes simplex virus (HSV) 1 and HSV2). The individuals were chosen based on their genotype for the NLS rs1557866, a SNP that is a strong proxy for the presence or absence of the Neandertal haplotype in the OAS region (9 were homozygous for the Neandertal haplotype, 15 were heterozygous, and 16 homozygous for the modern human sequence).

As expected, we found that all viral-associated immune triggers led to a marked increase in OAS1-3 gene expression levels, as measured by real-time PCR (up to 27-fold, $P \leq 1.2x10^{-8}$, Figure 5A), concomitantly with the up-regulation of type-I and type-II interferon genes (Additional file 1: Figure S8). Confirming the QTL results obtained in macrophages, we found that rs10774671 was a strong asQTL for OAS1 in both non-infected and infected PBMCs ($P \leq 4.9x10^{-5}$, Figure 5B). Likewise, we found that the presence of the Neandertal haplotype was associated with reduced expression levels of OAS3, particularly in PBMCs infected with influenza ($P = 6.1x10^{-3}$) and the synthetic ligand gardiquimod ($P = 2.0x10^{-4}$), which mimics a single strand RNA infection (Figure 5B). Interestingly, the Neandertal haplotype harbors additional regulatory variants that only impact expression levels in a cell-type and immune stimuli specific fashion. For example, we found that the Neandertal haplotype is associated with increased expression levels of OAS2 in non-infected ($P=1.2x10^{-3}$), and gardiquimod-stimulated PBMCs ($P=4.9x10^{-3}$), but neither in macrophages nor in PBMCs treated with other viral agents. Collectively, our functional data provide evidence for a pervasive impact of the Neandertal haplotype on the regulation of OAS genes that varies depending on the cell type and the immune stimuli to which the cells are responding.

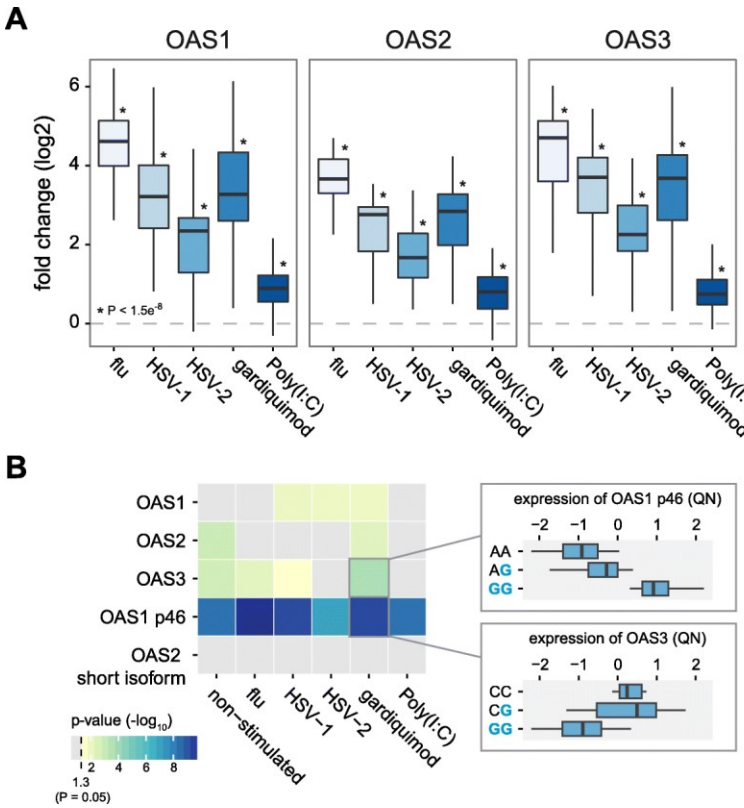

Figure 5. **THE NEANDERTAL HAPLOTYPE IN THE OAS REGIONS HAS A DIFFERENT IMPACT ON THE REGULATION OF OAS GENES DEPENDING ON THE VIRAL AGENTS PBMCS ARE EXPOSED TO**

a Log 2 fold induction (*y-axis*) of *OAS1*, *OAS2*, and *OAS3* in response to different viral agents or viral-associated immune stimuli, relative to non-infected PBMCs. b –log10 P for the association between genotype status for the Neandertal haplotype and overall expression levels of OAS genes and the expression of specific isoforms of *OAS1* and *OAS2* (those identified in Fig. 3 as associated with NLS). *Boxplots* show the association between genotype status (*blue* referring to Neanderthal alleles) and expression levels of the p46 isoform of *OAS1* and *OAS3* in Gardiquimod-stimulated PBMCs

Discussion

The Neandertal lineage was present in Eurasia for at least 400,000 years (Matthias Meyer et al. 2016), providing ample time for Neandertals to adapt to local disease environments. The admixture process, which likely fostered the transmission of pathogens between Neandertals and humans migrating out of Africa, could have also led to the exchange of genes useful in responding to local pathogens. Here, we have demonstrated that a previously reported case of Neandertal introgression at the OAS

locus (Fernando L. Mendez, Watkins, et Hammer 2013) displays population genetic signatures that are characteristic of positive selection in the European population. We note, however, that our results are limited by an incomplete knowledge of the demographic history of admixture between Neandertals and modern humans and the fact that the neutrality tests we used are not perfectly suited to detect adaptive introgression. Yet, we have strengthened the case for adaptive introgression by providing direct functional evidence of a role for the Neandertal OAS haplotype in the regulatory responses in innate immune cells to infectious agents.

Our results show that the Neandertal haplotype at OAS is associated with several regulatory variants that reduce expression of *OAS3* in response to infection, as well as encode alternate isoforms of *OAS1* and *OAS2*. These dramatic functional implications of the Neandertal OAS haplotype support our case for adaptive introgression at OAS. Yet, because distinct functional polymorphisms segregate together in the same haplotype, inferring the exact variant(s) targeted by positive selection remains a daunting task. We speculate, however, that one of the strongest direct targets of selection is likely to have been the splice variant identified in *OAS1*.

The Neandertal haplotype carries the ancestral allele (G) of the *OAS1* splice variant (rs10774671), which is common both inside and outside of Africa. However, outside of Africa, the only haplotypes carrying this ancestral splice site are closely related to the Neandertal haplotype, with a few exceptions being rare recombined haplotypes (~2% of all haplotypes with the ancestral allele). This pattern reflects the possibility that Neandertal introgression, in effect, served as a means to resurrect the ancestral splice site from local extinction outside of Africa, probably following the out-of-Africa exodus. Interestingly, the same ancestral splice site is found in the Denosivan genome, and may similarly have impacted adaptive introgression into the ancestors of

Melanesians. While we cannot at present confirm that the ancestral splice site was missing from the ancestral Eurasian population, the presence of this allele only on the Neandertal haplotype hints at the possibility that this splice site was lost to drift and subsequently re-introduced by Neandertals, providing beneficial genetic variation at the OAS locus (Additional file 1: Figure S2). The Neandertal-introgressed allele encodes a protein variant (p46) that is associated with higher enzymatic activity (Bonnevie-Nielsen et al. 2005). The adaptive potential of this variant is supported by the observation that this variant (or other variants in strong LD with it) was shown to be associated with: *(i)* reduced infection and replication rates of West Nile virus ((Lim et al. 2009), but see (Bigham et al. 2011)), *(ii)* improved resistance to hepatitis C virus (HCV) infection (Kwon et al. 2013; El Awady et al. 2011), and *(iii)* variable symptomology of Tick-Borne Encephalitis (TBE) Virus-Induced Disease (homozygous individuals for the Neandertal haplotype show the most severe symptoms of TBE). Strikingly, West Nile, hepatitis C and TBE are all members of the *Flaviviridae* family, suggesting that Flaviviruses might have been the main drivers of selection in *OAS1*.

The differential responses of homozygous carriers of the Neandertal OAS haplotype to the different viruses described above suggest that the Neandertal haplotype is not uniformly beneficial in humans. Thus, it is plausible that both spatial and temporal fluctuations in virus populations could have led to fluctuating selection pressure on the Neandertal OAS haplotype, consistent with findings that genes in the immune system have been disproportionately targeted by positive selection since the dawn of agriculture (Barreiro et Quintana-Murci 2010).

Alternatively, and not mutually exclusively, alleles at OAS, and particularly the *OAS1* splice variant, might be evolving under balancing selection. This hypothesis is supported by the observation that the *OAS1* splice variant (rs10774671) is found at

high frequency worldwide (0.11-0.7), and the significantly higher Tajima's D values observed around *OAS1*, as compared to genome-wide expectations (Figure 3D). Moreover, *OAS1* is among the most diverse genes in both humans and non-human primates. Indeed, a recent analysis of genome-wide sequence data from a total of 55 individuals from four non-human ape species, chimpanzee (*Pan troglodytes ellioti*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla gorilla*), and orangutan (*Pongo abelii*), identified *OAS1* as in the top 1% of genes showing the largest levels of nucleotide diversity among ape species, consistent with a scenario of long-term balancing selection (*OAS2* and *OAS3* are ranked in the 60th and 36th percentile of the genome-wide distribution, respectively) or, as previous research has suggested, rapid evolution across the primate order (Hancks et al. 2015). Further supporting the idea of balancing selection on the introgressed haplotype, our functional data suggest that the Neandertal haplotype contributes a range of gene expression responses in a cell-type and stimulus-specific manner.

Estimates of historical allele frequencies at the OAS locus, from ancient DNA samples dated between roughly 8,500 and 2,500 years ago, support the notion that the Neandertal haplotype did not follow a classic selective sweep model with constant directional selection. Under such a model, the observed present-day frequency of the Neandertal haplotype at the OAS locus of roughly 38% would suggest a codominant fitness effect of $s \sim 0.0014$ - 0.0017, when assuming an initial frequency of 0.02 at introgression approximately 2,000-2,400 generations ago (Methods). However, the observed allele frequency shift of 0.26 over the past 200-340 generations (maximum shift in CEU from ancient samples; see Figure 2B) suggests that the selection coefficient associated with the Neandertal haplotype during this recent human evolution would have been $s \sim 0.0044$ - 0.0075: 2.6-5.4 times larger than the above estimate. Therefore,

both temporally varying selection and balancing selection could explain why the Neandertal haplotype at OAS did not show clear signatures of adaptive introgression in previous studies (Fernando L. Mendez, Watkins, et Hammer 2013; Deschamps et al. 2016).

Our study illustrates the difficulty of identifying strong candidates for adaptive introgression. In a genome-wide statistical framework, our use of neutral coalescent simulations as a null distribution for adaptive introgression likely would not have provided significant results after multiple test correction across loci. This suggests that other instances of less obvious adaptive introgression, especially those that did not follow a classic selective sweep model, may remain to be identified. Moving forward, novel methods must be developed to identify such cases. Some progress is now being made on this front. For example, Racimo and colleagues (Racimo, Marnetto, et Huerta-Sanchez 2016) recently developed a genome-wide statistical framework which relies on distributions of uniquely shared derived alleles between humans and Neandertals (as does our study) to identify candidate regions for adaptive introgression. This study also highlighted the OAS locus as a candidate for adaptive introgression. Additionally, estimates of historical allele frequencies with increased spatial and temporal resolution provided by the sequencing of ancient human genomes are likely to play an important role in illuminating candidates for adaptive introgression which do not conform to classical selective sweep models.

Conclusions

In conclusion, our study demonstrates that the frequency and haplotype distribution of Neandertal-like sites can be used in a neutral simulation framework that accounts for local genomic context to investigate the history of selection at a candidate

locus for which genome-wide tests of selection provide ambiguous results. When combined with functional data, our results provide the strongest evidence to date in support of adaptive introgression in the OAS region. More generally, our study raises the possibility that adaptive introgression might not necessarily occur to select newly introduced variants but rather as a means to resurrect adaptive variation in modern human populations that had been lost due to demographic events.

Methods

Détails dans les annexes.

Notes

Philipp W. Messer and Luis B. Barreiro are Co-senior author.

Declarations

Acknowledgements

We thank all members from the Barreiro and Messer labs for helpful discussions and comments on the manuscript. We thank Dr. Silvia Vidal for the gift of the Influenza PR8 WT used in this study, Dr. Hugo Soudeyns and Doris Ransy for advice with the viral infections, and Marc Montero and George (PJ) Perry for sharing their analysis on nucleotide diversity levels of OAS genes in non-human primate species.

Funding

This study was funded by grants from the Canadian Institutes of Health Research (301538 and 232519), the Human Frontiers Science Program (CDA-00025/2012), and the Canada Research Chairs Program (950–228993) (to LBB). YN

was supported by a fellowship from the Réseau de Médecine Génétique Appliquée (RMGA).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and in Additional file *2* : Tables S8 and S9. The raw data were deposited under GEO accession numbers GSE73765 and GSE81046.

Authors' contributions

Conception and design: AJS, PWM, LBB; acquisition of data: AJS, AD, YN, VY, LBB; analysis and interpretation of data: AJS, AD, YN, PWM, LBB; contributed unpublished, essential data, or reagents: CA, JET; drafting or revising the article: AJS, PWM, LBB. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Institutional ethics (IRB) approval was obtained by the CHU Sainte-Justine Institutional Review Board (approved project #3557) and all subjects gave written informed consent prior to participation. Experimental methods comply with the Declaration of Helsinki.

Chapitre 4 : Discussion

C'est la première fois qu'une étude s'attelle à mettre en évidence des différences de réponses immunitaires entre populations humaines à différents niveaux. Au niveau

cellulaire nous avons montré que les macrophages des individus d'origine Africaine semblent capables de détruire les bactéries intracellulaires plus rapidement.

D'un point de vue génétique, nous avons à la fois mis en évidence des différences entre populations dans l'expression des gènes mais également des proportions d'isoformes lors de la réponse immunitaire. En accord avec les études épidémiologiques qui font état d'une prévalence plus importante de maladies inflammatoires chez les Afro-Américain, nous avons montré que cette population dispose d'une réponse pro-inflammatoire plus soutenue.

Parmi les gènes et isoformes différemment exprimés entre les populations, nous avons remarqué qu'un grand nombre (804) sont sous le contrôle de variants génétiques qui expliquent au moins 75% des différences observées entre Afro- et Euro-Américains.

Nous avons ensuite mis en perspective ces découvertes avec l'évolution de l'homme moderne et l'adaptation des populations humaines à leurs environnements. En effet, nous avons confirmé que des évènements récents de sélection naturelle ont participé à l'émergence et au maintien de variants génétiques nécessaires à la défense contre les pathogènes et ce, différemment entre les Africains et Européens.

Enfin, nous avons montré qu'une partie de variants génétiques sélectionnés durant l'évolution et responsables de différences dans la réponse immunitaire entre populations trouvent leur origine dans notre métissage avec l'Homme de Néandertal.

Il est intéressant de noter que nos résultats s'accordent avec ceux d'une récente étude bien que ceux-ci découlent d'expérimentations différentes (Quach et al. 2016). En effet, là où nous avons choisi d'étudier l'infection de macrophages par des bactéries intracellulaires vivantes, Quach et al. ont basé leur travail sur des monocytes stimulés par des antigènes bactériens ou infectés par un virus grippal. Nous pensons que ces

différences expérimentales sont responsables d'une partie des écarts que nous observons dans les listes de gènes différentiellement exprimés entre les populations ou dont l'expression est sous l'effet de variants génétiques.

Ce travail effectué durant mon doctorat, bien qu'ayant été scrupuleusement préparé et soutenu par des technologies et protocoles performants (collecte des échantillons, culture cellulaire, infection, génotypage, séquençage de l'ARN, analyses) n'en demeure pas moins soumis à plusieurs réserves.

Bien que nous ayons détecté un grand nombre d'associations eQTL, reQTL ou asQTL en cis, nous n'avons pas été en mesure de découvrir de tels signaux en trans, c'est-à-dire de discerner les associations significatives qui ont une réalité biologique, de celles qui proviennent de la quantité de tests statistiques que nous avons dû réaliser. Ce fardeau des tests multiples est difficile à maîtriser et notre approche avec un FDR basé sur la méthode de Storey-Tibshirani n'a pas suffi à révéler des associations en trans. Notre analyse suggère toutefois que 20% des différences d'expression observées entre Afro- et Euro-Américains sont sous le contrôle de variants génétiques en trans. Des méthodes basées sur la reconstruction de réseaux de co-expression puis leur interrogation, en limitant le nombre de tests à effectuer, pourraient améliorer notre capacité à détecter ces associations en trans (Rakitsch et Stegle 2016) (Figure 1).

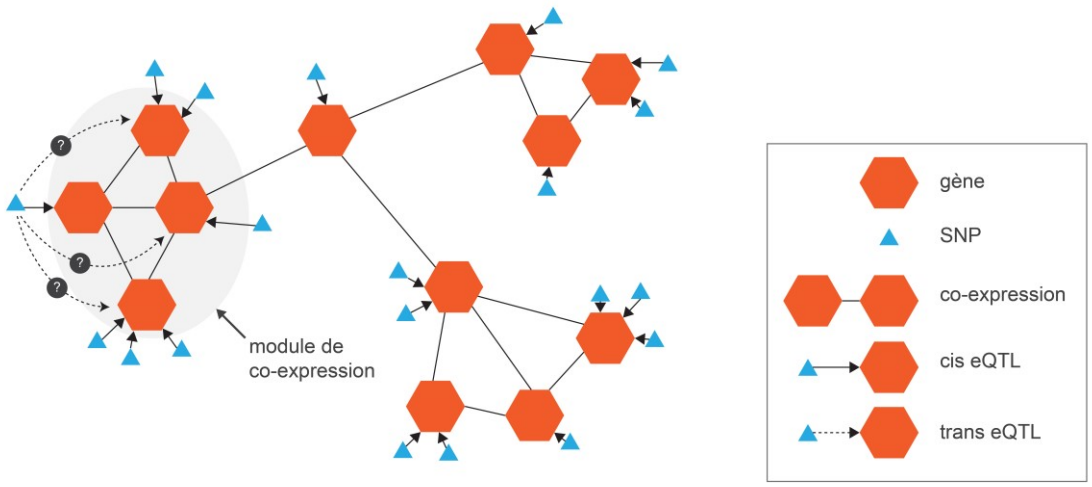

Figure 1. EXEMPLE DE L'UTILISATION D'UN RESEAU DE CO-EXPRESSION POUR LA DECOUVERTE DE TRANS EQTL

Dans un premier temps, un réseau de co-expression est reconstruit pour regrouper des gènes dont l'expression est corrélée au sein de modules. Ensuite, chaque SNP impliqué dans un cis eQTL pourra être testé contre l'expression des autres gènes faisant partie du même module.

Il faut garder à l'esprit qu'une fraction des différences que nous observons dans l'expression des gènes et isoformes entre Afro- et Euro-Américains est probablement le fruit d'évènements épigénétiques. En d'autres termes, même si nous pensons que la sélection d'individus Nord-Américains permet de s'affranchir d'une partie des variations dues à l'environnement, les conditions de vie inégales de ces deux populations (situation socio-professionnelle, alimentation, accès aux soins, hygiène de vie) sont susceptibles d'engendrer des différences mesurables dans l'expression de leurs gènes.

Notre étude suggère que les individus d'origine Africaine disposent d'une clairance bactérienne plus efficace que les Européens. Cependant, nous n'avons pas réalisé d'expérimentations plus avancées nous permettant d'élucider les mécanismes mis en jeu et de répondre à plusieurs questions en suspens :

- *Quels sont les gènes responsables de ces différences de clairance bactérienne entre Afro- et Euro-Américains ?*
- *Quels sont les variants génétiques susceptibles d'affecter l'expression de ces gènes et ainsi de moduler la clairance bactérienne ?*

En utilisant la cytométrie de flux, nous pourrions suivre plus précisément la quantité de bactéries à l'intérieur des macrophages dans l'objectif d'utiliser la clairance bactérienne comme un phénotype qui reflète la réalité des infections bactériennes au sein de l'organisme humain. Ainsi, nous serions en mesure d'identifier des profils d'expression de gènes reliés à ce phénotype et, au final, d'établir un set de variants génétiques associés à des variations dans l'efficacité de la clairance bactérienne. Pour se faire, nous pourrions rechercher des groupes de gènes dont la variation d'expression est corrélée à la clairance bactérienne car ceux-ci représentent des candidats pour expliquer le phénotype observé.

Pour tenter d'améliorer notre compréhension de l'évolution du système immunitaire et les effets du métissage avec Néandertal sur celui-ci, nous avons choisi de nous concentrer sur l'analyse d'une région du génome impliquée dans la protection contre les infections virales : les gènes OAS. Dans cette portion de notre génome, l'haplotype provenant de Néandertal est à la fois associé à une diminution de l'expression d'OAS3 en réponse à une infection mais également relié à l'expression de transcrits alternatifs pour OAS1 et OAS2. Etant donné que cet haplotype renferme un grand nombre de variants génétiques qui ségrégent entre eux, c'est-à-dire dont les allèles varient conjointement, il demeure difficile d'identifier les polymorphismes causaux. Toutefois, le SNP rs10774671, en perturbant un site d'épissage d'OAS1, est associé à une expression augmentée du transcrit codant pour l'isoforme protéique p46 dont l'activité enzymatique est plus importante (Bonnevie-Nielsen et al. 2005). En accord

avec plusieurs études établissant un lien entre ce SNP et des susceptibilités aux infections par des virus de la famille des *Flaviviridae* (El Awady et al. 2011; Lin et al. 2009; Kwon et al. 2013; Bigham et al. 2011; Lim et al. 2009), nous pensons que ces virus sont à l'origine de pressions de sélection naturelle sur ce polymorphisme depuis son introgression dans notre génome.

Lorsque l'homme moderne est arrivé en Europe, il a pu se métisser et bénéficier de certains des variants génétiques de l'homme de Néandertal qui était déjà adapté à son environnement. Parmi ceux-ci, nous avons pu identifier des variants impliqués dans le contrôle de l'inflammation et dans la réponse aux infections virales. Ceux-ci ont probablement participé à la survie de notre espèce en permettant à l'organisme de se défendre contre des pressions pathogéniques différentes de celles rencontrées en Afrique tout en limitant l'apparition de pathologies auto-immunes.

Chapitre 5 : Perspectives

Notre combat contre les infections bactériennes demeure un challenge de santé public à l'échelle mondiale. Les traitements actuels reposent sur une approche en deux grandes étapes : identification de la bactérie responsable de l'infection puis mise en application de la conduite à tenir sans prendre en compte les susceptibilités génétiques de l'individu.

De ce fait, la nature du traitement, sa posologie et son administration ne sont pas adaptées au malade et l'équilibre entre l'efficacité du traitement et l'exposition aux effets secondaires n'est pas atteint.

La médecine de précision s'écarte de la pratique traditionnelle en ciblant les particularités de l'individu plutôt qu'en se basant sur une approche thérapeutique

identique pour tous les patients. Depuis l'avènement de la génomique, notre domaine de recherche croule sous la quantité de données maintenant disponibles. Malheureusement, nous éprouvons des difficultés à en faire bon usage dans le but d'améliorer notre compréhension des mécanismes biologiques qui s'avèrent particulièrement complexes. L'apprentissage automatique représente un espoir dans l'analyse de données à grande échelle où l'informatique permettrait d'élucider les acteurs et les mécanismes impliqués dans le fonctionnement de notre organisme, de la cellule jusqu'au phénotype (Deo 2015; Leung et al. 2016). Cette approche pourra nous permettre de détecter les facteurs génétiques ou épigénétiques susceptibles d'entrainer une pathologie, ou qui se révèlent utiles pour adapter une prise en charge thérapeutique.

Pour atteindre ce but, je propose deux axes principaux :

- identification des mécanismes responsables des variations interindividuelles dans la clairance bactérienne ;
- utilisation de l'apprentissage automatique pour guider le praticien dans une prise en charge des maladies infectieuses adaptée au patient.

Identification des mécanismes responsables de variations inter-individuelles de la clairance bactérienne

La clairance bactérienne est un phénotype cellulaire qui traduit la capacité du macrophage à lyser les bactéries intracellulaires. Pour la première fois, nous avons montré que ce phénomène diffère entre Afro- et Euro-américains. Toutefois, il semble important de collecter d'autres observations sur ce phénomène dans le but d'en comprendre les mécanismes : *les différences que l'on observe entre populations sont-elles associées à la capacité du macrophage à lyser les bactéries ou à internaliser l'agent*

pathogène ? Même si la qualité des mesures que nous avons collectées ne permettent pas de l'affirmer, nous supposons que ces différences sont en partie la conséquence de variations dans l'expression des gènes et sous le contrôle de variants génétiques. Cette observation reflète probablement assez fidèlement le combat du système immunitaire inné contre les bactéries au sein de l'organisme et représente ainsi un intérêt de recherche significatif.

Comme je l'ai déjà précisé dans la discussion de mon premier article, l'obtention de mesures de clairance plus précises constitue la première étape nécessaire à la découverte et à l'identification des gènes et de leurs variants génétiques impliqués.

Ainsi, dans la continuité de mon travail, je pense qu'il serait intéressant de recueillir à la fois des phénotypes cellulaires mais également sécrétoires. Cette étude pourrait utiliser les mêmes individus que mon propre projet grâce aux cellules mononucléaires périphériques sanguines (PBMCs) qui ont été conservées. Les expérimentations de laboratoire s'appuieraient sur des méthodes innovantes de culture cellulaire[4] pour, dans un premier temps, « dé-différencier » les PBMCs en cellules souches pluripotentes induites (IPSCs) puis en les orientant vers la lignée monocytaire dans le but d'obtenir des macrophages. Ces macrophages seraient enfin soumis au même protocole que celui de mon étude, à travers une infection par *Listeria monocytogenes* et *Salmonella typhimurium*. Plusieurs phénotypes cellulaires et sécrétoires en rapport avec l'infection pourraient être recueillis pour permettre, dans un premier temps, de mettre en évidence les étapes responsables des différences de clairance entre populations humaines que nous avons déjà observées, et dans un second temps de tenter de corréler

[4] Cette technique (Takahashi et Yamanaka 2006) a été développée par Shinya Yamanaka dont le travail a été récompensé par le prix Nobel de médecine de 2012.

ces traits avec des variants génétiques. Nous devrions ainsi aboutir à une liste de variants génétiques dont les allèles sont corrélés avec certains traits, tels que le délai d'internalisation et de destruction de la bactérie, le nombre de bactéries à l'intérieur de la cellule, la proportion de macrophages détruits par l'infection ou le profil de cytokines produites.

Les attributs sont les informations qui serviront à l'algorithme d'apprentissage automatique pour en prédire l'issue, je propose d'utiliser des mesures génétiques (génotypes), épigénétiques (empreintes ATAC-seq) et les niveaux d'expression des gènes (Burga et Lehner 2013). Ces derniers, en étant un phénotype intermédiaire entre le génome et le trait observé, devraient améliorer les performances de la prédiction (Burga et Lehner 2013). Des algorithmes d'apprentissage automatique se basant sur le transcriptome ont déjà été utilisés pour prédire le pronostic et la réponse au traitement de patients atteints d'un cancer (van't Veer et al. 2002; Onken et al. 2004; Golub et al. 1999).

Il semble judicieux de choisir un algorithme capable de prédire la variable cible sous la forme d'une valeur quantitative (proportion de bactéries lysées à 24h, par exemple) tout en conservant une certaine visibilité pour en comprendre la classification. Ce cas semble adapté à l'utilisation d'un arbre de régression (Waldmann 2016) et implique l'usage d'un apprentissage supervisé où l'entrainement de l'algorithme repose sur l'utilisation d'ensembles de données pour lesquels la variable-cible est connue.

L'objectif de cette approche est de mettre en évidence les attributs qui contribuent de façon importante à la variabilité de la clairance bactérienne. Par exemple, des gènes associés à une diminution de la clairance représentent des cibles potentielles pour des traitements pharmacologiques. Dans le cas de variants génétiques importants pour la prédiction de la clairance, ceux-ci ont un intérêt diagnostic pour

déceler des troubles de la clairance chez un patient et adapter sa surveillance thérapeutique à l'occasion d'une infection bactérienne.

Apprentissage automatique et traitements de précision des maladies infectieuses

L'objectif à long terme est d'utiliser une approche d'apprentissage automatique pour prédire l'évolution de l'infection chez le patient et s'orienter vers une approche thérapeutique adaptée. Cette démarche s'appuiera sur plusieurs types de données, telles que le pathogène mis en cause, les caractéristiques génétiques du patient ou ses antécédents de maladies en rapport avec le système immunitaire.

Dans un premier temps, on cherchera à constituer une liste aussi complète que possible des variants génétiques ayant potentiellement un effet sur la réponse immunitaire. Lors de l'entrainement de l'algorithme d'apprentissage automatique, ces variants seront génotypés chez les patients et utilisés conjointement avec d'autres attributs pouvant améliorer notre capacité à effectuer la classification. Les attributs ayant des potentiels prédictifs seront retenus pour être dépistés à l'occasion du déploiement de cette approche en milieu hospitalier. Enfin, cet outil permettra de guider le praticien et d'étayer ses choix thérapeutiques dans le but d'assurer une efficacité optimale du traitement, une minimisation des effets secondaires et une anticipation des complications.

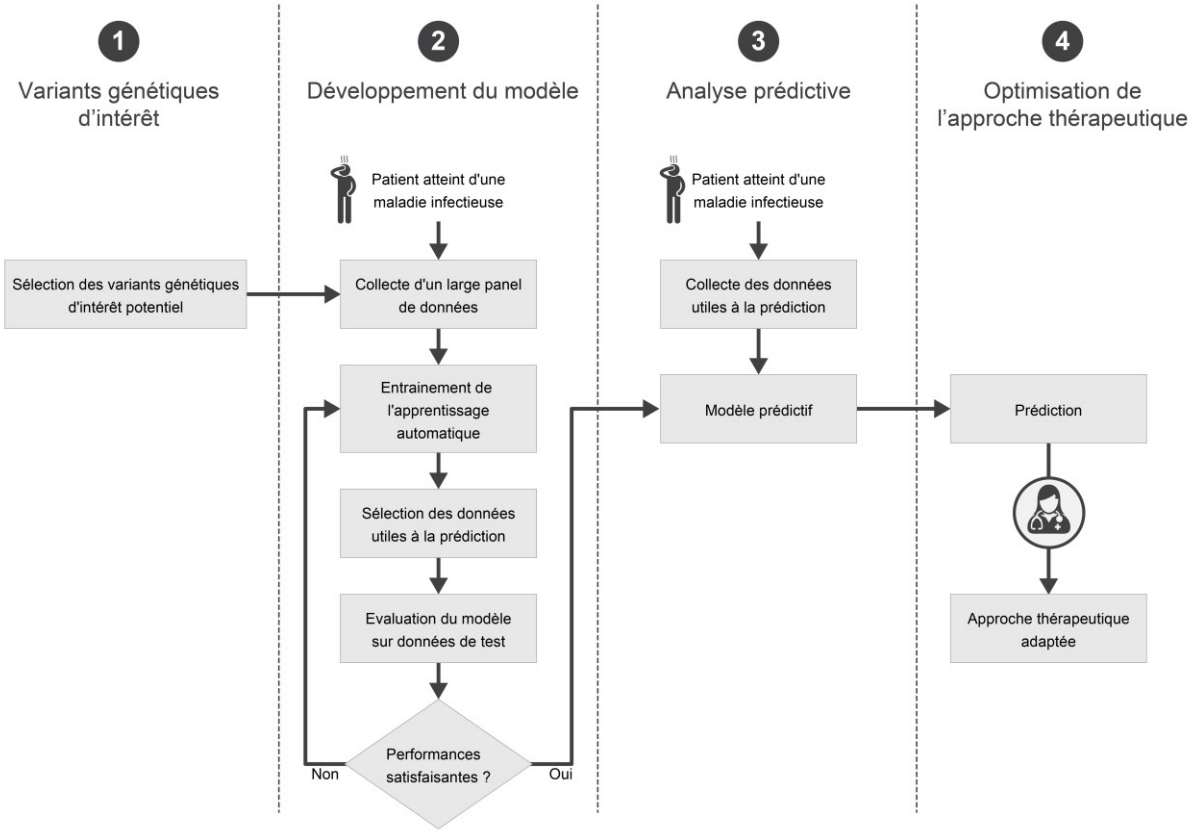

Figure 1. PROPOSITION D'UNE DEMARCHE DE MEDECINE DE PRECISION ADAPTEE AUX MALADIES INFECTIEUSES

Identification des variants génétiques d'intérêt

L'avènement du séquençage de l'ADN et du génotypage nous permet d'identifier les facteurs de risques d'origine génétique d'un individu. Bien entendu, il est nécessaire, de connaître au préalable les variants génétiques associés à des gènes, des pathologies ou à des phénotypes en rapport avec la réponse aux infections.

Je propose de sélectionner un ensemble de variants ayant potentiellement un retentissement sur les fonctions de l'immunité et susceptibles d'être utiles dans le cadre d'un dépistage. Ces SNPs sont issus de 3 sources :

1. A partir d'analyses eQTL, reQTL ou asQTL pour des gènes dont l'ontologie est en rapport avec l'immunité ;

2. Détectés à travers des GWAS des pathologies du système immunitaire ;
3. Pour lesquels un lien avec des phénotypes cellulaires ou sécrétoires a été établi expérimentalement.

SNPs issus d'analyses -QTL

Les variants génétiques ayant un effet sur l'expression d'un gène constituent des candidats potentiels dans le cadre d'une approche de médecine de précision. En effet, en interrogeant ces variants chez un patient, nous sommes en mesure d'estimer leur retentissement sur l'expression des gènes et donc sur les effecteurs de la cellule. Ces effecteurs peuvent ensuite être responsables de variations dans des phénotypes reliés à l'infection, tels que la capacité du macrophage à produire des cytokines pro-inflammatoire, à détruire la bactérie ou sa capacité à initier la réponse immunitaire.

A titre d'exemple, nous savons que l'interleukine 6 (IL6) est une cytokine pro-inflammatoire extrêmement importante dans la réponse immunitaire. Celle-ci est produite par les macrophages en réponse à la reconnaissance de certains antigènes microbiens (PAMPs) par les récepteurs de type Toll (TLRs). Ces récepteurs, présents à la surface de la cellule et dans les compartiments intracellulaires, initient des cascades de signalisation cellulaire qui aboutissent à la production de cytokines inflammatoires. IL6 est notamment un médiateur important dans la mise en place de la fièvre et la production des protéines de phase aigüe par le foie en réponse à une inflammation[5]. Lors de l'étude présentée dans le chapitre 2, nous avons détecté des variants génétiques associés à des variations dans l'expression du récepteur de cette cytokine : IL6R. Nous

[5] Parmi les protéines de phase aigüe, la plus connue est la protéine C-réactive (CRP) qui est dosée fréquemment en milieu hospitalier pour évaluer l'inflammation.

pouvons émettre l'hypothèse que les individus disposant de l'allèle responsable d'une diminution de l'expression de ce gène ont une réponse inflammatoire plus limitée à l'occasion d'une infection bactérienne. Ce premier exemple montre le potentiel que représente le dépistage de ce variant génétique dans le cadre d'une approche de médecine de précision. Il est toutefois important de rester prudent quant à la contribution réelle de celui-ci vis-à-vis de la réponse immunitaire étant donné que son effet n'a pas été validé expérimentalement.

Pour identifier de tels variants avec un intérêt diagnostic potentiel, nous avons développé une plateforme pour simplifier la présentation des résultats et le partage des données à destination de la communauté scientifique.

Je suis convaincu que des améliorations dans le partage des données sont nécessaires afin d'accélérer nos recherches en médecine. En ce sens, le site internet ImmunPop permet de remplir plusieurs fonctions :

- vérifier si un gène dispose d'un QTL d'expression (eQTL), de réponse (reQTL) ou associé à un épissage alternatif (asQTL) ;
- vérifier si un SNP est associé à la variation d'expression d'un gène ou d'un isoforme ;
- vérifier si un gène est différentiellement exprimé entre les populations ou lors d'une infection ;
- exporter les figures sous un format vectoriel permettant leur modification ;
- récupérer programmatiquement les résultats en interrogeant le serveur à travers l'interface de programmation (API REST).

A partir des résultats d'une analyse QTL, les données spécifiques au projet (génotypes, valeurs d'expressions et associations testées) sont chargées dans la base de données SQL. A titre d'information, cela correspond à plus d'une centaine de millions d'entrées pour mon projet. La base de données est interrogée à travers des requêtes générée par le moteur développé en Python. Le moteur est à la fois responsable de la production des pages web en HTML5 lorsque le client utilise un navigateur, et du renvoi des données sous format JSON lorsque des requêtes sont formulées à travers l'API REST (Annexe).

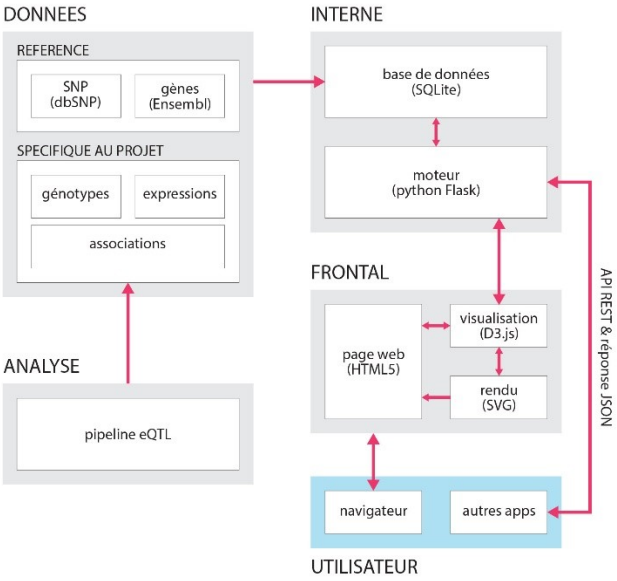

Figure 2. **STRUCTURE D'IMMUNPOP**

ImmunPop a été conçu pour pouvoir s'adapter à différents projets et permettre le chargement de leurs données. Ainsi, nous avons déjà entrepris l'ajout de deux autres projets qui concernent respectivement des études sur *Mycobacterium tuberculosis* (Barreiro et al. 2011) et *Mycobacterium leprae* (manuscrit soumis).

SNPs identifiés en GWAS

Une GWAS est une étude d'association pangénomique qui analyse, au sein d'un large groupe d'individu, la corrélation entre un grand nombre de variants génétiques et un phénotype. Certaines GWAS se sont concentrées sur des pathologies du système immunitaire, des marqueurs d'inflammation ou des susceptibilités à des infections bactériennes. L'objectif est d'identifier des SNPs qui ont un effet significatif sur le phénotype mesuré. Les données peuvent être facilement récupérées sur GWAS Central (www.gwascentral.org) ou GWAS Catalog (https://www.ebi.ac.uk/gwas). Ce dernier recense respectivement 705 et 42 SNPs significativement associés à des maladies du système immunitaire et à des marqueurs de l'inflammation. En analysant ces phénotypes macroscopiques, les GWAS cherchent à mettre en cause des variants génétiques qui ont en réalité un effet sur un grand nombre de processus moléculaires sous-jacents. De ce fait, les variants génétiques associés au phénotype sont difficiles à mettre en évidence car leurs effets n'expliquent généralement qu'une petite partie du phénotype d'intérêt. Les GWAS nécessitent donc sur le recrutement d'un grand nombre de participants pour tenter d'élucider l'ensemble des variants génétiques impliqués.

SNPs associés à des phénotypes de l'infection bactérienne

Je propose d'utiliser les variants génétiques qui auront été associés à des différences de clairance bactérienne, comme je l'ai proposé dans le premier axe de perspectives. En effet, à l'heure actuelle, nous ne disposons pas de données nous permettant de comprendre les effets des variants génétiques sur la dynamique de l'infection bactérienne.

Développement du modèle

Un algorithme d'apprentissage automatique (*machine learning*) repose sur l'utilisation d'un grand nombre d'attributs du patient pour tenter de prédire l'évolution de l'infection, il représente ainsi un outil décisionnel pour classifier les patients et adapter leur prise en charge.

La première étape consiste à entrainer l'algorithme d'apprentissage automatique en utilisant à la fois les données prédictives mais également celles associées à l'évolution de l'infection et au résultat de la prise en charge du patient.

Lorsqu'un patient se présentera à l'hôpital avec une maladie infectieuse, je propose que deux grands types de données soient collectées : les attributs observables qui peuvent aider à la classification, et les données qui sont associées à l'évolution de l'infection.

Pour ce premier groupe d'attributs prédictifs :

- l'agent pathogène et de ses particularités : recherche de résistances par PCR ou antibiogramme, séquençage ;
- les caractéristiques du patient : age, sexe, antécédents de maladies en rapport avec le système immunitaire (infections ou maladies auto-immunes, immunodépression), co-morbidités, traitements associés (anti-inflammatoires, immuno-suppresseurs, chimiothérapie, radiothérapie) ;
- ses paramètres cliniques : manifestations de l'infection, température corporelle à l'admission, pouls, saturation ;
- résultats d'explorations biologiques : numération formule sanguine, vitesse de sédimentation, dosage de marqueurs inflammatoires (CRP, procalcitonine, fibrinogène), bactériémie ;

- l'expression des gènes impliqués dans la réponse immunitaire ;
- les génotypes de l'individu pour les SNPs qui ont été sélectionnés préalablement et qui représentent un potentiel prédictif ;

Parmi les attributs que l'on cherche à prédire, c'est-à-dire ceux qui reflètent l'évolution de l'infection et de sa prise en charge, je suggère de recueillir :

- la bactériémie et la température ;
- la durée d'hospitalisation ;
- les complications rencontrées : propagation de l'infection, choc septique (Mayr, Yende, et Angus 2014) ;
- les effets secondaires du traitement : réaction allergique, perturbation importante de la flore intestinale, réactions cutanées, atteinte hépatique (Yee et al. 2003; Westphal, Vetter, et Brogard 1994) ;

Pour faciliter l'entrainement de l'apprentissage et ses performances, il est important de sélectionner les attributs qui ont un réel potentiel prédictif : c'est-à-dire qui contribuent significativement à l'amélioration de la classification.

Plusieurs types d'algorithmes d'apprentissage existent avec des caractéristiques différentes (Deo 2015). Pour mener à bien notre objectif, les arbres décisionnels ou la régression logistique sont deux choix envisageables qu'il est nécessaire toutefois d'évaluer en conditions réelles.

Une fois celui-ci suffisamment entraîné, c'est-à-dire que le modèle permet une prédiction dont les performances sont satisfaisantes, il est possible de classifier les patients en fonction de leur susceptibilité aux infections bactériennes ou leur probabilité à souffrir de complications ou d'effets secondaires.

Analyse prédictive

L'objectif du système d'apprentissage automatique est de prédire les phénotypes en rapport avec l'infection bactérienne à partir d'un minimum de données :

- les génotypes et l'expression des gènes ayant été déclarés informatifs par le système d'apprentissage ;
- les paramètres cliniques ou caractéristiques du patient utiles à la classification ;
- la nature de l'agent infectieux responsable de l'infection.

Notons qu'il est nécessaire de disposer d'une technologie capable de génotyper un grand nombre de variants et de mesure l'expression des gènes dans un délai en adéquation avec l'urgence que peut représenter une infection bactérienne. Actuellement, le séquençage du génome peut être réalisé en 26h (Miller et al. 2015) mais il est toutefois possible de s'orienter vers un génotypage de première intention en sélectionnant un nombre réduit de variants génétiques dont on détermine les allèles à l'aide d'une technique de réaction en chaîne par polymérase (PCR) multiplexée.

Le système d'apprentissage automatique sera en mesure de classifier le patient pour évaluer, notamment, les facteurs de risques auxquels il est soumis, la probabilité d'apparition de complications ou la réponse de l'organisme à un agent pharmacologique.

Optimisation de l'approche thérapeutique

Le praticien disposera d'informations permettant d'orienter la prise en charge thérapeutique du patient dans l'objectif d'optimiser la balance bénéfice-risque.

Cette approche s'appuiera sur notamment sur l'adaptation du traitement pharmacologique aux particularités du patient, tant sur la nature de la molécule

antibiotique mais également sur sa posologie dans le but de limiter les effets secondaires (Van Driest et al. 2015).

Lorsque le patient est classifié par l'algorithme comme potentiellement à risque de développer des complications, le médecin pourra prescrire des examens complémentaires et ajuster la surveillance lors de l'hospitalisation.

L'utilisation de l'apprentissage automatique en médecine est récente et n'a, à ce jour, pas donné lieu à une profonde refonte de la pratique clinique (Deo 2015; Leung et al. 2016). Plusieurs obstacles s'opposent à la découverte et à l'adoption de ce type d'approches.

Nous sommes en effet réticents à faire confiance à un système de classification automatique dont le fonctionnement est difficile à appréhender. A ce titre, il semble judicieux d'opter pour un apprentissage supervisé où l'humain est en mesure de comprendre et de contrôler la classification. En ce sens, nous pourrions envisager une application en infectiologie où l'on combinerait un algorithme ayant une grande sensibilité avec la spécificité du diagnostic d'un praticien.

L'apprentissage automatique se heurte également à des défis techniques qui retarde son adoption en médecine. En premier lieu, il est nécessaire d'acquérir un grand nombre de mesures, souvent coûteuses, pour obtenir un modèle avec une précision suffisante. Dans le cas de modèles particulièrement complexes, c'est-à-dire qui cherchent à prédire des phénotypes qui sont contrôlés par un grand nombre d'attributs, il est difficile d'échapper à des phénomènes de surapprentissage (overfitting) où l'algorithme montre un taux d'erreur de classification très faible sur le set de données d'entrainement mais est incapable de généraliser face à de nouvelles observations. Notons également que le choix des attributs à utiliser s'avère crucial et l'utilisation d'un trop grand

nombre d'entre eux se fera au détriment des performances de la classification. Pour cela, on peut tout d'abord sélectionner les attributs informatifs grâce à un algorithme d'apprentissage non-supervisé comme par exemple une forêt d'arbres décisionnels (X. Chen et Ishwaran 2012).

Chapitre 6 : Conclusion

La bio-informatique se révèlent être une alliée importante de la recherche médicale en permettant d'analyser de grandes quantités de données dans le but d'améliorer notre compréhension des systèmes biologiques. Au-delà de l'importance que représente les nouvelles découvertes en biologie, je suis convaincu que celles-ci doivent s'inclure dans un effort collectif où l'on cherche à améliorer le partage des résultats, des techniques, mais également la reproductibilité d'une étude.

Ce projet de doctorat repose sur l'utilisation de techniques avancées de bio-informatique, notamment à travers le séquençage et l'estimation des niveaux d'expression du transcriptome de cellules humaines en réponse à une infection bactérienne. L'estimation de l'abondance de chaque isoforme demeure un défi computationnel qui nécessite l'utilisation d'algorithmes avancés permettant d'affronter la complexité combinatoire de cette tâche. Dans ce projet, nous avons fait le choix d'utiliser le programme RSEM qui repose sur l'estimation de l'abondance d'isoformes à l'aide d'un algorithme espérance-maximisation capable de retracer le transcrit de provenance d'un read lorsque celui-ci peut provenir de plusieurs isoformes (Li et Dewey 2011). L'émergence de ce type d'approches d'estimation au niveau des isoformes (RSEM (Li et Dewey 2011), Kallisto (Bray et al. 2016), Salmon (Patro et al. 2017)) est un des exemples d'avancées en bio-informatique qui permettent d'identifier des variations dans les profils transcriptomiques et se présentent comme des alternatives intéressantes face

aux solutions plus classiques de quantification d'expression des gènes (featureCounts (Liao, Smyth, et Shi 2014), HTseq (Anders, Pyl, et Huber 2015)). Si la recherche d'eQTL n'a pas nécessité l'utilisation d'approches particulièrement complexes, il est important de rappeler l'impressionnante quantité de tests statistiques (169 milliards) que représente le test de chaque SNP contre l'expression de chaque gène. Ainsi, sans l'utilisation de Matrix eQTL qui repose sur des calculs matriciels, cette analyse aurait nécessité plusieurs années de temps de calcul par un unique processeur (Shabalin 2012). Pour améliorer notre capacité à discerner les signaux biologiques des faux positifs inhérents aux nombreux tests que nous effectuons, nous avons adapté un algorithme innovant basé sur le travail de Storey and Tibshirani (Storey et Tibshirani 2003). Enfin, nous avons mis en place un site internet, accessible sur www.immunpop.com, permettant de présenter et de partager facilement les données d'autres projets similaires avec l'ensemble de la communauté.

Notre génome porte les stigmates de longs processus évolutifs qui ont façonné notre espèce. La reconstruction de ces évènements demeure un défi qui implique d'aborder la situation à travers la génétique des populations et son pendant évolutif. Durant ce doctorat, j'ai montré qu'il existe des différences de réponse immunitaire entre Afro- et Euro-américains, à la fois au niveau de l'expression des macrophages, mais également en ce qui concerne leur capacité à lyser les bactéries intracellulaires. A ce titre, nous avons identifié qu'environ 30% des gènes exprimés dans ces cellules montrent des différences de niveau d'expression entre populations à l'état basal ou lorsqu'elles sont infectées par Listeria ou Salmonella. L'observation la plus intéressante de ce projet, et probablement celle qui me tient le plus à cœur, concerne la différence de clairance bactérienne entre macrophages issus d'individus Afro- et Euro-américains. Pour la première fois, nous montrons qu'une telle différence existe entre populations humaines

et que les macrophages d'individus d'ascendance Africaine disposent d'une meilleure capacité à lyser les bactéries intracellulaires. Ce phénotype s'avère fascinant car il est probablement l'observation *in vitro* la plus à même de nous renseigner sur le déroulement et le retentissement de l'infection sur l'organisme. Ce phénomène est en accord avec des études épidémiologiques qui font état de prévalences plus importantes de maladies inflammatoires chez les Afro-Américains (Pennington et al. 2009).

Même si nous pensons qu'une partie des différences que nous observons entre les populations peuvent provenir de facteurs environnementaux, nous avons validé l'importance des polymorphismes génétiques sur la différence d'expression des gènes entre populations. En effet, nous avons observé que 30% des différences d'expression entre populations pour les gènes différentiellement exprimés, peut être attribué à l'effet d'un unique variant génétique en cis.

Nous avons également montré un enrichissement marqué de signatures de sélection naturelle parmi les variants impliqués dans des eQTL, suggérant un rôle adaptatif de ceux-ci dans la récente évolution de l'homme moderne. Nous pensons que ces variants ont été sélectionnés au cours de l'adaptation des populations humaines dans le but de protéger l'organisme des pathogènes présents dans son environnement. Les variants génétiques associés à un système immunitaire plus inflammatoire, autrefois nécessaires à la survie de l'individu face aux considérables pressions de sélection qu'exerçaient les pathogènes en Afrique, révèlent maintenant leur effet délétère dans des pathologies du système immunitaire (maladies auto-immunes, inflammatoires ou certains cancers). Ainsi, il est plausible que l'adaptation des populations non-africaines se soit accompagnée d'une sélection négative ou d'une relaxation des pressions de sélection à l'encontre de certains de ces variants.

Au sein des populations non-Africaines, certains des polymorphismes génétiques sont issus de l'introgression avec l'Homme de Néandertal, déjà présent en Eurasie avant la sortie de l'Afrique de l'homme moderne. Nous avons décrit le cas des gènes antiviraux OAS où un haplotype issu de Néandertal est associé à une réduction de l'expression d'OAS3 en réponse à une infection ainsi qu'à la surexpression de certains isoformes des gènes OAS1 et OAS2. Notons que dans ce cas, cette introgression adaptative aurait servi à restaurer un allèle perdu à la sortie de l'Afrique responsable de la production d'une version plus active de la protéine OAS1, conférant probablement un avantage pour la survie des populations européennes.

L'objectif à long terme de ce doctorat a toujours été de repousser les limites de nos connaissances sur le système dans l'espoir de faire progresser nos approches thérapeutiques. Les données collectées lors de ce doctorat sont probablement utiles à la mise en place de méthodes d'apprentissage automatique pour identifier des variants génétiques ayant un potentiel diagnostic et prédictif en médecine. Ainsi, nous serions en mesure d'adapter nos soins aux spécificités de chaque patient, dans l'objectif perpétuel d'améliorer leur efficacité tout en limitant les risques.

Bibliographie

Abadie, Valérie, Ludvig M Sollid, Luis B Barreiro, et Bana Jabri. 2011. « Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis ». *Annual Review of Immunology* 29 (1). Annual Reviews:493-525. https://doi.org/doi:10.1146/annurev-immunol-040210-092915.

Alexander, David H, John Novembre, et Kenneth Lange. 2009. « Fast model-based estimation of ancestry in unrelated individuals ». *Genome Research* 19 (9):1655-64.

Anders, Simon, Paul Theodor Pyl, et Wolfgang Huber. 2015. « HTSeq-A Python framework to work with high-throughput sequencing data ». *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu638.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. « A global reference for human genetic variation ». *Nature* 526 (7571). Nature Research:68-74. https://doi.org/10.1038/nature15393.

Awady, Mostafa K. El, Mohamed A. Anany, Gamal Esmat, Naglaa Zayed, Ashraf A. Tabll, Amr Helmy, Abdel Rahman El Zayady, et al. 2011. « Single nucleotide polymorphism at exon 7 splice acceptor site of OAS1 gene determines response of hepatitis C virus patients to interferon therapy ». *Journal of Gastroenterology and Hepatology (Australia)*. https://doi.org/10.1111/j.1440-1746.2010.06605.x.

Barreiro, Luis B, Meriem Ben-Ali, Hélène Quach, Guillaume Laval, Etienne Patin, Joseph K Pickrell, Christiane Bouchier, et al. 2009. « Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. » *PLoS genetics* 5 (7):e1000562.

Barreiro, Luis B, et Lluís Quintana-Murci. 2010. « From evolutionary genetics to human immunology: how selection shapes host defence genes. » *Nature Reviews Genetics* 11 (1):17-30. https://doi.org/10.1038/nrg2698.

Barreiro, Luis B, Ludovic Tailleux, Athma A Pai, Brigitte Gicquel, John C Marioni, et Yoav Gilad. 2011. « Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. » *Proceedings of the National Academy of Sciences of the United States of America* 109 (4):1204-9. https://doi.org/10.1073/pnas.1115761109/-/DCSupplemental.www.pnas.org/cgi/.

Bigham, Abigail W., Kati J. Buckingham, Sofia Husain, Mary J. Emond, Kathryn M. Bofferding, Heidi

Gildersleeve, Ann Rutherford, et al. 2011. « Host genetic risk factors for West Nile virus infection and disease progression ». *PLoS ONE*. https://doi.org/10.1371/journal.pone.0024745.

Bindea, Gabriela, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf Herman Fridman, Franck Pagès, Zlatko Trajanoski, et Jérôme Galon. 2009. « ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks ». *Bioinformatics* 25 (8). Oxford University Press ({OUP}):1091-93. https://doi.org/10.1093/bioinformatics/btp101.

Bonnevie-Nielsen, V, L L Field, S Lu, D J Zheng, M Li, P M Martensen, T B Nielsen, H Beck-Nielsen, Y L Lau, et F Pociot. 2005. « Variation in antiviral 2',5'-oligoadenylate synthetase (2'5'AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene ». *Am J Hum Genet*. https://doi.org/10.1086/429391.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, et Lior Pachter. 2016. « Near-optimal probabilistic RNA-seq quantification ». *Nature Biotechnology*. https://doi.org/10.1038/nbt.3519.

Brinkworth, Jessica F., et Luis B. Barreiro. 2014. « The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease ». *Current Opinion in Immunology* 31. Elsevier Ltd:66-78. https://doi.org/10.1016/j.coi.2014.09.008.

Bryc, Katarzyna, Adam Auton, Matthew R Nelson, Jorge R Oksenberg, Stephen L Hauser, Scott Williams, Alain Froment, et al. 2010. « Genome-wide patterns of population structure and admixture in West Africans and African Americans. » *Proceedings of the National Academy of Sciences of the United States of America* 107 (2). Proceedings of the National Academy of Sciences:786-91. https://doi.org/10.1073/pnas.0909559107.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, et William J Greenleaf. 2013. « Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. » *Nature methods* 10 (12). Springer Nature:1213-18. https://doi.org/10.1038/nmeth.2688.

Burga, Alejandro, et Ben Lehner. 2013. « Predicting phenotypic variation from genotypes, phenotypes and a combination of the two ». *Current Opinion in Biotechnology*. https://doi.org/10.1016/j.copbio.2013.03.004.

Çalışkan, Minal, Samuel W Baker, Yoav Gilad, et Carole Ober. 2015. « Host genetic variation influences gene expression response to rhinovirus infection. » Édité par Greg Gibson. *PLoS genetics* 11 (4).

Public Library of Science ({PLoS}):e1005111. https://doi.org/10.1371/journal.pgen.1005111.

Casals, Ferran, Alan Hodgkinson, Julie Hussin, Youssef Idaghdour, Vanessa Bruat, Thibault de Maillard, Jean Cristophe Grenier, et al. 2013. « Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans ». *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1003815.

Chen, Gary K, Paul Marjoram, et Jeffrey D Wall. 2009. « Fast and flexible simulation of DNA sequence data. » *Genome Research* 19 (1):136-42.

Chen, X, et H Ishwaran. 2012. « Random forests for genomic data analysis ». *Genomics*. https://doi.org/10.1016/j.ygeno.2012.04.003.

Cheung, V G, R S Spielman, K G Ewens, T M Weber, M Morley, et J T Burdick. 2005. « Mapping determinants of human gene expression by regional and genome-wide association ». *Nature* 437 (7063):1365-69. https://doi.org/10.1038/nature04244.

Cui, Xiangqin, Jason Affourtit, Keith R Shockley, Yong Woo, et Gary A Churchill. 2006. « Inheritance patterns of transcript levels in F1 hybrid mice. » *Genetics* 174 (2):627-37.

Dannemann, Michael, Aida M. Andrés, et Janet Kelso. 2016. « Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors ». *American Journal of Human Genetics* 98 (1). Elsevier {BV}:22-33. https://doi.org/10.1016/j.ajhg.2015.11.015.

Dausset, Jean, Howard Cann, Daniel Cohen, Mark Lathrop, Jean Marc Lalouel, et Ray White. 1990. « Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome ». *Genomics* 6 (3). Elsevier {BV}:575-77. https://doi.org/10.1016/0888-7543(90)90491-C.

deMenocal, Peter B., et Chris Stringer. 2016. « Human migration: Climate and the peopling of the world ». *Nature* 538 (7623):49. https://doi.org/10.1038/nature19471.

Deo, Rahul C. 2015. « Machine learning in medicine ». *Circulation* 132 (20):1920-30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.

Deschamps, Matthieu, Guillaume Laval, Maud Fagny, Yuval Itan, Laurent Abel, Jean-Laurent Casanova, Etienne Patin, et Llu'\is Quintana-Murci. 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.

Driest, Sara L. Van, Tracy L. McGregor, Digna R. Velez Edwards, Ben R. Saville, Terrie E. Kitchner,

Scott J. Hebbring, Murray Brilliant, et al. 2015. « Genome-wide association study of serum creatinine levels during vancomycin therapy ». *PLoS ONE* 10 (6):1-14. https://doi.org/10.1371/journal.pone.0127791.

Dupré, Nicolas, Heidi C. Howard, Jean Mathieu, George Karpati, Michel Vanasse, Jean Pierre Bouchard, Stirling Carpenter, et Gu A. Rouleau. 2003. « Hereditary motor and sensory neuropathy with agenesis of the corpus callosum ». *Annals of Neurology*. https://doi.org/10.1002/ana.77777.

Durbin, Richard M., David L. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, et al. 2010. « A map of human genome variation from population-scale sequencing. » *Nature* 467 (7319):1061-73. https://doi.org/10.1038/nature09534.

Eames, Hayley L., Alastair L. Corbin, et Irina A. Udalova. 2016. « Interferon regulatory factor 5 in human autoimmunity and murine models of autoimmune disease ». *Translational Research* 167 (1). Elsevier {BV}:167-82. https://doi.org/10.1016/j.trsl.2015.06.018.

Enard, David, Philipp W. Messer, et Dmitri A. Petrov. 2014. « Genome-wide signals of positive selection in human evolution ». *Genome Research* 24 (6). Cold Spring Harbor Laboratory Press:885-95. https://doi.org/10.1101/gr.164822.113.

ENCODE Project, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, et Michael Snyder. 2012. « An integrated encyclopedia of DNA elements in the human genome. » *Nature*. https://doi.org/nature11247 [pii]\n10.1038/nature11247.

Excoffier, Laurent, et Heidi E L Lischer. 2010. « Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows ». *Molecular Ecology Resources* 10 (3):564-67.

Fagny, Maud, Etienne Patin, David Enard, Luis B. Barreiro, Lluis Quintana-Murci, et Guillaume Laval. 2014. « Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets ». *Molecular Biology and Evolution* 31 (7). Oxford University Press ({OUP}):1850-68. https://doi.org/10.1093/molbev/msu118.

Fairfax, Benjamin P., et Julian C. Knight. 2014. « Genetics of gene expression in immunity to infection ». *Current Opinion in Immunology* 30. Elsevier Ltd:63-71. https://doi.org/10.1016/j.coi.2014.07.001.

Fairfax, Benjamin P, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, et al. 2014. « Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. » *Science (New York, N.Y.)* 343 (March):1246949.

https://doi.org/10.1126/science.1246949.

Feasey, Nicholas A., Gordon Dougan, Robert A. Kingsley, Robert S. Heyderman, et Melita A. Gordon. 2012. « Invasive non-typhoidal salmonella disease: An emerging and neglected tropical disease in Africa ». *The Lancet* 379 (9835). Elsevier Ltd:2489-99. https://doi.org/10.1016/S0140-6736(11)61752-2.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, et Rasmus Nielsen. 2014. « On detecting incomplete soft or hard selective sweeps using haplotype structure. » *Molecular Biology and Evolution* 31 (5):1275-91.

Feuk, Lars, Andrew R. Carson, et Stephen W. Scherer. 2006. « Structural variation in the human genome ». *Nature Reviews Genetics* 7 (2). Nature Publishing Group:85-97. https://doi.org/10.1038/nrg1767.

Fraser, Hunter B. 2013. « Gene expression drives local adaptation in humans ». *Genome Research* 23 (7). Cold Spring Harbor Laboratory Press:1089-96. https://doi.org/10.1101/gr.152710.112.

Fu, Qiaomei, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip L F Johnson, et al. 2014. « Genome sequence of a 45,000-year-old modern human from western Siberia. » *Nature* 514 (7523):445-49. https://doi.org/10.1038/nature13810.

Fumagalli, Matteo, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admettla, Linda Pattini, et Rasmus Nielsen. 2011. « Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution ». *PLoS Genetics* 7 (11). https://doi.org/10.1371/journal.pgen.1002355.

Gabunia, Leo, Susan C. Antón, David Lordkipanidze, Abesalom Vekua, Antje Justus, et Carl C. Swisher. 2001. « Dmanisi and dispersal ». *Evolutionary Anthropology* 10:158-70. https://doi.org/10.1002/evan.1030.

Garber, Manuel, Nir Yosef, Alon Goren, Raktima Raychowdhury, Anne Thielke, Mitchell Guttman, James Robinson, et al. 2012. « A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals ». *Molecular Cell* 47 (5). Elsevier {BV}:810-22. https://doi.org/10.1016/j.molcel.2012.07.030.

Gibson, Greg, et Bruce Weir. 2005. « The quantitative genetics of transcription. » *Trends in genetics : TIG* 21 (11):616-23. https://doi.org/10.1016/j.tig.2005.08.010.

Golub, T R, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, et al. 1999.

« Molecular classification of cancer: class discovery and class prediction by gene expression monitoring ». *Science*. https://doi.org/10.1126/science.286.5439.531.

Gordon, M A, S M Graham, A L Walsh, L Wilson, A Phiri, E Molyneux, E E Zijlstra, R S Heyderman, C A Hart, et M E Molyneux. 2008. « Epidemics of invasive Salmonella enterica serovar enteritidis and S. enterica Serovar typhimurium infection associated with multidrug resistance among adults and children in Malawi ». *Clin Infect Dis* 46 (7):963-69. https://doi.org/10.1086/529146.

Gravel, Simon, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, et 1000 Genomes Project. 2011. « Demographic history and rare allele sharing among human populations ». *Proceedings of the National Academy of Sciences of the United States of America* 108 (29):11983-88.

Green, Richard E, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. « A draft sequence of the Neandertal genome. » *Science* 328 (May):710-22. https://doi.org/10.1126/science.1188021.

Guernier, Vanina, Michael E. Hochberg, et J. F. Guégan. 2004. « Ecology drives the worldwide distribution of human diseases ». *PLoS Biology* 2 (6):740-46. https://doi.org/10.1371/journal.pbio.0020141.

Gulig, P A, T J Doyle, M J Clare-Salzler, R L Maiese, et H Matsui. 1997. « Systemic infection of mice by wild-type but not Spv- Salmonella typhimurium is enhanced by neutralization of gamma interferon and tumor necrosis factor alpha. » *Infection and Immunity* 65 (12):5191-97.

Gusev, Alexander, S Hong Lee, Benjamin M Neale, Gosia Trynka, B. J. Vilhjalmsson, H. Finucane, H. Xu, et al. 2014. « Regulatory variants explain much more heritability than coding variants across 11 common diseases ». *Genomics*. https://doi.org/10.1101/004309.

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson, et Carlos D. Bustamante. 2009. « Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data ». Édité par Gil McVean. *PLoS Genetics* 5 (10). Public Library of Science ({PLoS}):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Gymrek, Melissa, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J Daly, et al. 2016. « Abundant contribution of short tandem repeats to gene expression variation in humans. » *Nature genetics* 48 (1). Springer Nature:22-29. https://doi.org/10.1038/ng.3461.

Hancks, Dustin C, Melissa K Hartley, Celia Hagan, Nathan L Clark, et Nels C Elde. 2015. « Overlapping Patterns of Rapid Evolution in the Nucleic Acid Sensors cGAS and OAS1 Suggest a Common Mechanism of Pathogen Antagonism and Escape ». *PLoS Genetics* 11 (5):e1005203.

Harris, Kelley, et Rasmus Nielsen. 2016. « The Genetic Cost of Neanderthal Introgression ». *Genetics*, janvier, genetics.116.186890.

Hernandez, Ryan D, Joanna L Kelley, Eyal Elyashiv, S Cord Melton, Adam Auton, Gilean McVean, 1000 Genomes Project, Guy Sella, et Molly Przeworski. 2011. « Classic selective sweeps were rare in recent human evolution. » *Science (New York, N.Y.)* 331 (6019):920-24. https://doi.org/10.1126/science.1198878.

Hindorff, Lucia a, Praveen Sethupathy, Heather a Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, et Teri a Manolio. 2009. « Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. » *Proceedings of the National Academy of Sciences of the United States of America* 106 (23):9362-67. https://doi.org/10.1073/pnas.0903103106.

Holsinger, Kent E, et Bruce S Weir. 2009. « Genetics in geographically structured populations: defining, estimating and interpreting F(ST). » *Nature reviews. Genetics* 10 (September):639-50. https://doi.org/10.1038/nrg2611.

Howie, Bryan N, Peter Donnelly, et Jonathan Marchini. 2009. « A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies ». *PLoS Genetics* 5 (6):e1000529.

Hublin, J J. 2009. « The origin of Neandertals ». *Proceedings of the National Academy of Sciences* 106 (38):16022-27.

Hudson, R R. 2002. « Generating samples under a Wright-Fisher neutral model of genetic variation ». *Bioinformatics*.

Hudson, R R, M Slatkin, et W P Maddison. 1992. « Estimation of levels of gene flow from DNA-sequence data. » *Genetics* 132 (2):583-89.

Huerta-Sanchez, Emilia, Xin Jin, Asan, Zhuoma Bianba, Benjamin M Peter, Nicolas Vinckenbosch, Yu Liang, et al. 2014. « Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA ». *Nature* 512 (7513):194-97.

Jacoby, George A. 2005. « Mechanisms of resistance to quinolones. » *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 41 Suppl 2 (Supplement_2):S120-

6. https://doi.org/10.1086/428052.

Jostins, Luke, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, et al. 2012. « Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. » *Nature* 491 (7422):119-24. https://doi.org/10.1038/nature11582.

Karlsson, Elinor K, Dominic P Kwiatkowski, et Pardis C Sabeti. 2014. « Natural selection and infectious disease in human populations ». *Nature Reviews Genetics* 15 (6). Nature Publishing Group:379-93. https://doi.org/10.1038/nrg3734.

Kelley-Hedgepeth, Alyson, Donald M. Lloyd-Jones, Alicia Colvin, Karen A. Matthews, Janet Johnston, MaryFran R. Sowers, Barbara Sternfeld, Richard C. Pasternak, et Claudia U. Chae. 2008. « Ethnic differences in C-reactive protein concentrations ». *Clinical Chemistry* 54 (6). American Association for Clinical Chemistry ({AACC}):1027-37. https://doi.org/10.1373/clinchem.2007.098996.

Kelley, Joanna L., Jennifer Madeoy, John C. Calhoun, Willie Swanson, et Joshua M. Akey. 2006. « Genomic signatures of positive selection in humans and the limits of outlier approaches ». *Genome Research* 16 (8). Cold Spring Harbor Laboratory Press:980-89. https://doi.org/10.1101/gr.5157306.

Kelso, Janet, et Kay Prüfer. 2014. « Ancient humans and the origin of modern humans ». *Current Opinion in Genetics and Development* 29 (décembre). Elsevier {BV}:133-38. https://doi.org/10.1016/j.gde.2014.09.004.

Kingsley, Robert A, Chisomo L Msefula, Nicholas R Thomson, Robert A Kingsley, Chisomo L Msefula, Nicholas R Thomson, Samuel Kariuki, et al. 2009. « Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype », 2279-87. https://doi.org/10.1101/gr.091017.109.

Kwon, Young Chan, Ju Il Kang, Soon B. Hwang, et Byung Yoon Ahn. 2013. « The ribonuclease l-dependent antiviral roles of human 2,5-oligoadenylate synthetase family members against hepatitis C virus ». *FEBS Letters*. https://doi.org/10.1016/j.febslet.2012.11.010.

Lappalainen, Tuuli, Michael Sammeth, Marc R Friedländer, Peter a C 't Hoen, Jean Monlong, Manuel a Rivas, Mar Gonzàlez-Porta, et al. 2013. « Transcriptome and genome sequencing uncovers functional variation in humans. » *Nature* 501:506-11. https://doi.org/10.1038/nature12531.

LaRock, Doris L, Anu Chaudhary, et Samuel I Miller. 2015. « Salmonellae interactions with host

processes ». *Nature Reviews Microbiology* 13 (4):191-205.

Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, et al. 2014. « Ancient human genomes suggest three ancestral populations for present-day Europeans ». *Nature* 513 (7518):409-13.

Lee, Mark N., Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, et al. 2014. « Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells ». *Science* 343:1246980. https://doi.org/10.1126/science.1246980.

Leinonen, Tuomas, R J Scott McCairns, Robert B O'Hara, et Juha Merilä. 2013. « Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. » *Nature reviews. Genetics* 14 (March). Springer Nature:179-90. https://doi.org/10.1038/nrg3395.

Leung, Michael K K, Andrew Delong, Babak Alipanahi, et Brendan J. Frey. 2016. « Machine learning in genomic medicine: A review of computational problems and data sets ». *Proceedings of the IEEE*. https://doi.org/10.1109/JPROC.2015.2494198.

Li, Bo, et Colin N Dewey. 2011. « RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. » *BMC bioinformatics* 12 (1). Springer Nature:323. https://doi.org/10.1186/1471-2105-12-323.

Liao, Yang, Gordon K. Smyth, et Wei Shi. 2014. « FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features ». *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt656.

Lim, Jean K., Andrea Lisco, David H. McDermott, Linda Huynh, Jerrold M. Ward, Bernard Johnson, Hope Johnson, et al. 2009. « Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man ». *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1000321.

Lin, R J, H P Yu, B L Chang, W C Tang, C L Liao, et Y L Lin. 2009. « Distinct antiviral roles for human 2',5'-oligoadenylate synthetase family members against dengue virus infection ». *J Immunol*. https://doi.org/10.4049/jimmunol.0902728.

Llorente, M. G., E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, et al. 2015. « Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent ». *Science* 350 (6262):820-22. https://doi.org/10.1126/science.aad2879.

Maclean, Colin A, Neil P Chue Hong, et James G D Prendergast. 2015. « hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. »

Molecular Biology and Evolution 32 (11):3027-29.

Malaspinas, Anna-Sapfo, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, et al. 2016. « A genomic history of Aboriginal Australia ». *Nature* 538 (7624). Nature Publishing Group:207-14. http://dx.doi.org/10.1038/nature18299.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. « The Simons Genome Diversity Project: 300 genomes from 142 diverse populations ». *Nature* 538 (7624). Macmillan Publishers Limited, part of Springer Nature. All rights reserved.:201-6. http://dx.doi.org/10.1038/nature18964%5Cnhttp://10.1038/nature18964%5Cnhttp://www.nature.com/nature/journal/v538/n7624/abs/nature18964.html#supplementary-information%5Cnhttp://www.nature.com/doifinder/10.1038/nature18964.

Mathieson, Iain, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, et al. 2015. « Genome-wide patterns of selection in 230 ancient Eurasians ». *Nature*, novembre.

Mayr, Florian B, Sachin Yende, et Derek C Angus. 2014. « Epidemiology of severe sepsis. » *Virulence*. https://doi.org/10.4161/viru.27372.

McDougall, Ian, Francis H Brown, et John G Fleagle. 2005. « Stratigraphic placement and age of modern humans from Kibish, Ethiopia ». *Nature* 433 (7027):733-36. https://doi.org/10.1038/nature03258.

Mendez, F L, J C Watkins, et M F Hammer. 2012. « Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations ». *Molecular Biology and Evolution* 29 (6):1513-20.

Mendez, Fernando L., Joseph C. Watkins, et Michael F. Hammer. 2013. « Neandertal origin of genetic variation at the cluster of OAS immunity genes ». *Molecular Biology and Evolution* 30 (4):798-801. https://doi.org/10.1093/molbev/mst004.

Mendez, Fernando L, Joseph C Watkins, et Michael F Hammer. 2012. « A Haplotype at STAT2 Introgressed from Neanderthals and Serves as a Candidate of Positive Selection in Papua New Guinea ». *The American Journal of Human Genetics* 91 (2):265-74.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, et al. 2012. « A High-Coverage Genome Sequence from an Archaic Denisovan Individual ». *Science* 338 (October):222-26. https://doi.org/10.1126/science.1224344.

Meyer, Matthias, Juan Luis Arsuaga, Cesare de Filippo, Sarah Nagel, Ayinuer Aximu-Petri, Birgit

Nickel, Ignacio Mart'\inez, et al. 2016. « Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins ». *Nature*, mars.

Miller, Neil A., Emily G. Farrow, Margaret Gibson, Laurel K. Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, et al. 2015. « A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases ». *Genome Medicine* 7 (1). Genome Medicine:100. https://doi.org/10.1186/s13073-015-0221-8.

Morley, Michael, Cliona M Molony, Teresa M Weber, James L Devlin, Kathryn G Ewens, Richard S Spielman, et Vivian G Cheung. 2004. « Genetic analysis of genome-wide variation in human gene expression. » *Nature* 430 (7001):743-47.

Nauciel, C, et F Espinasse-Maes. 1992. « Role of Gamma Interferon and Tumor-Necrosis-Factor-Alpha in Resistance to Salmonella-Typhimurium Infection ». *Infection and Immunity* 60 (2):450-54.

Nédélec, Yohann, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, et al. 2016. « Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens ». *Cell* 167 (3). Elsevier:657-669.e21. https://doi.org/10.1016/j.cell.2016.09.025.

Ness, Roberta B, Catherine L Haggerty, Gail Harger, et Robert Ferrell. 2004. « Differential distribution of allelic variants in cytokine genes among African Americans and White Americans. » *American journal of epidemiology* 160 (11):1033-38. https://doi.org/10.1093/aje/kwh325.

Okabe, Yasutaka, et Ruslan Medzhitov. 2016. « Tissue biology perspective on macrophages. » *Nature immunology* 17 (1). Springer Nature:9-17. https://doi.org/10.1038/ni.3320.

Okin, Daniel, et Ruslan Medzhitov. 2012. « Evolution of inflammatory diseases ». *Current Biology* 22 (17). Elsevier {BV}:R733--R740. https://doi.org/10.1016/j.cub.2012.07.029.

Onken, Michael D., Lori A. Worley, Justis P. Ehlers, et J. William Harbour. 2004. « Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death ». *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-04-1750.

Palazzo, Alexander F., et T. Ryan Gregory. 2014. « The Case for Junk DNA ». *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004351.

Palumbi, S R. 2001. « Evolution - Humans as the world's greatest evolutionary force ». *Science* 293 (5536):1786-90. https://doi.org/10.1017/CBO9781107415324.004.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, et Carl Kingsford. 2017. « Salmon provides

fast and bias-aware quantification of transcript expression ». *Nature Methods*. https://doi.org/10.1038/nmeth.4197.

Patterson, Nick, Daniel J Richter, Sante Gnerre, Eric S Lander, et David Reich. 2006. « Genetic evidence for complex speciation of humans and chimpanzees. » *Nature* 441 (June):1103-8. https://doi.org/10.1038/nature04789.

Pennington, Renee, Chandler Gatenbee, Brett Kennedy, Henry Harpending, et Gregory Cochran. 2009. « Group differences in proneness to inflammation. » *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases.* 9 (6):1371-80. https://doi.org/10.1016/j.meegid.2009.09.017.

Pickrell, J K, J C Marioni, A A Pai, et J F Degner. 2010. « Understanding mechanisms underlying human gene expression variation with RNA sequencing ». *Nature* 464 (7289):768-72.

Player, Mark R, et Paul F Torrence. 1998. « The 2{\textendash}5 A system: Modulation of viral and cellular processes through acceleration of RNA degradation ». *Pharmacology {&} Therapeutics* 78 (2):55-113.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. « The complete genome sequence of a Neanderthal from the Altai Mountains. » *Nature* 505:43-49. https://doi.org/10.1038/nature12886.

Pujol, B., A. J. Wilson, R. I C Ross, et J. R. Pannell. 2008. « Are QST-FST comparisons for natural populations meaningful? » *Molecular Ecology* 17 (22). Wiley-Blackwell:4782-85. https://doi.org/10.1111/j.1365-294X.2008.03958.x.

Quach, Hélène, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, et al. 2016. « Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations ». *Cell* 167 (3):643-656.e17. https://doi.org/10.1016/j.cell.2016.09.024.

Racimo, Fernando, Davide Marnetto, et Emilia Huerta-Sanchez. 2016. « The landscape of uniquely shared archaic alleles in present-day human populations ». *bioRxiv*, mars, 45237.

Racimo, Fernando, Sriram Sankararaman, Rasmus Nielsen, et Emilia Huerta-Sánchez. 2015. « Evidence for archaic adaptive introgression in humans. » *Nature reviews. Genetics* 16 (6). https://doi.org/10.1038/nrg3936.

Rakitsch, Barbara, et Oliver Stegle. 2016. « Modelling local gene networks increases power to detect

trans-acting genetic effects on gene expression ». *Genome Biology* 17 (1). Genome Biology:33. https://doi.org/10.1186/s13059-016-0895-2.

Ramachandran, Girish, Darren J. Perkins, Patrick J. Schmidlein, Mohan E. Tulapurkar, et Sharon M. Tennant. 2015. « Invasive Salmonella Typhimurium ST313 with Naturally Attenuated Flagellin Elicits Reduced Inflammation and Replicates within Macrophages ». *PLoS Neglected Tropical Diseases* 9 (1):1-12. https://doi.org/10.1371/journal.pntd.0003394.

Richardus, Jan H, et Anton E Kunst. 2001. « Black–White Differences in Infectious Disease Mortality in the United States. » *American journal of public health* 91 (8). © American Journal of Public Health 2001:1251-53.

Sabeti, P.C., D E Reich, J M Higgins, H Z P Levine, D J Richter, S F Schaffner, S B Gabriel, J V Platko, N J Patterson, et G J McDonald. 2002. « Detecting recent positive selection in the human genome from haplotype structure ». *Nature* 419 (6909):832-37.

Sabeti, P C, B Fry, J Lohmueller, E Hostetter, C Cotsapas, X Xie, E H Byrne, et al. 2007. « Genome-wide detection and characterization of positive selection in human populations ». *Nature* 449 (7164):913-18. https://doi.org/10.1038/nature06250.

Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, et David Reich. 2014. « The genomic landscape of Neanderthal ancestry in present-day humans. » *Nature* 507:354-57. https://doi.org/10.1038/nature12961.

Sankararaman, Sriram, Swapan Mallick, Nick Patterson, et David Reich. 2016. « The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans ». *Current Biology* 26 (9). Elsevier Ltd:1241-47. https://doi.org/10.1016/j.cub.2016.03.037.

Sanyal, Amartya, Job Dekker, Gaurav Jain, et Bryan R Lajoie. 2012. « The long-range interaction landscape of gene promoters. » *Nature* 489 (7414):109-13. https://doi.org/10.1038/nature11279.

Sawyer, Susanna, Gabriel Renaud, Bence Viola, J.-J. Jean-jacques Hublin, Marie-theres M.-T. Gansauge, M. V. Shunkov, A. P. Derevianko, K. Pru fer, J. Kelso, et S. Pa a bo. 2015. « Nuclear and mitochondrial DNA sequences from two Denisovan individuals ». *Proceedings of the National Academy of Sciences* 112 (51):2-6. https://doi.org/10.1073/pnas.1519905112.

Schaub, Marc A., Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, et Michael Snyder. 2012. « Linking disease associations with regulatory information in the human genome ». *Genome Research* 22 (9):1748-59. https://doi.org/10.1101/gr.136127.111.

Schlamp, Florencia, Julian Made, Rebecca Stambler, Lewis Chesebrough, Adam R Boyko, et Philipp W Messer. 2016. « Evaluating the performance of selection scans to detect selective sweeps in domestic dogs ». *Molecular Ecology* 25 (1):342-56.

Ségurel, Laure, et Lluis Quintana-Murci. 2014. « Preserving immune diversity through ancient inheritance and admixture ». *Current Opinion in Immunology* 30:79-84. https://doi.org/10.1016/j.coi.2014.08.002.

Selbach, Matthias, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, et Nikolaus Rajewsky. 2008. « Widespread changes in protein synthesis induced by microRNAs. » *Nature* 455 (7209):58-63.

Shabalin, Andrey A. 2012. « Matrix eQTL: Ultra fast eQTL analysis via large matrix operations ». *Bioinformatics* 28 (10). Oxford University Press ({OUP}):1353-58. https://doi.org/10.1093/bioinformatics/bts163.

Storey, John D, et Robert Tibshirani. 2003. « Statistical significance for genomewide studies. » *Proceedings of the National Academy of Sciences of the United States of America* 100 (16):9440-45.

Stringer, Chris. 2012. « What makes a modern human ». *Nature* 485 (7396):4-6. https://doi.org/10.1038/485033a.

Takahashi, Kazutoshi, et Shinya Yamanaka. 2006. « Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors ». *Cell* 126 (4):663-76. https://doi.org/10.1016/j.cell.2006.07.024.

Tennessen, J. a., a. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, et al. 2012. « Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes ». *Science* 337 (July):64-69. https://doi.org/10.1126/science.1219240.

Tishkoff, Sarah a, Floyd a Reed, Françoise R Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B Hirbo, et al. 2009. « The genetic structure and history of Africans and African Americans. » *Science (New York, N.Y.)* 324 (May):1035-44. https://doi.org/10.1126/science.1172257.

Tucci, Serena, et Joshua M. Akey. 2016. « Population genetics: A map of human wanderlust ». *Nature*. http://www.nature.com/doifinder/10.1038/nature19472.

van't Veer, L J, H Dai, M J van de Vijver, Y D He, A A Hart, M Mao, H L Peterse, et al. 2002. « Gene expression profiling predicts clinical outcome of breast cancer ». *Nature*.

https://doi.org/10.1038/415530a.

Vernot, Benjamin, et Joshua M Akey. 2015. « Complex History of Admixture between Modern Humans and Neandertals ». *The American Journal of Human Genetics*, février.

Vernot, Benjamin, Joshua M Akey, A. D. Twyford, R. A. Ennos, D. Zinner, M. L. Arnold, C. Roos, et al. 2014. « Resurrecting surviving Neandertal lineages from modern human genomes. » *Science* 343 (6174):1017-21. https://doi.org/10.1126/science.1245938.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, et Jonathan K Pritchard. 2006. « A map of recent positive selection in the human genome. » *PLoS biology* 4 (3):e72.

Waldmann, Patrik. 2016. « Genome-wide prediction using Bayesian additive regression trees ». *Genetics Selection Evolution*. https://doi.org/10.1186/s12711-016-0219-8.

Westphal, J. F., D. Vetter, et J. M. Brogard. 1994. « Hepatic side-effects of antibiotics ». *Journal of Antimicrobial Chemotherapy*. https://doi.org/10.1093/jac/33.3.387.

Wolf, Nicole I., Camilo Toro, Ilya Kister, Kartikasalwah Abd Latif, Richard Leventer, Amy Pizzino, Cas Simons, et al. 2015. « DARS-associated leukoencephalopathy can mimic a steroid-responsive neuroinflammatory disorder ». *Neurology* 84 (3). Ovid Technologies (Wolters Kluwer Health):226-30. https://doi.org/10.1212/WNL.0000000000001157.

Wu, L, S I Candille, Y Choi, D Xie, et L Jiang. 2013. « Variation and genetic control of protein abundance in humans ». *Nature* 499 (7456):79-82.

Yang, M A, A S Malaspinas, E Y Durand, et M Slatkin. 2012. « Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity ». *Molecular Biology and Evolution* 29 (10):2987-95.

Yee, Daphne, Chantal Valiquette, Marthe Pelletier, Isabelle Parisien, Isabelle Rocher, et Dick Menzies. 2003. « Incidence of Serious Side Effects from First-Line Antituberculosis Drugs among Patients Treated for Active Tuberculosis ». *American Journal of Respiratory and Critical Care Medicine*. https://doi.org/10.1164/rccm.200206-626OC.

Zhang, Xiaoling, Hinco J Gierman, Daniel Levy, Andrew Plump, Radu Dobrin, Harald HH Goring, Joanne E Curran, et al. 2014. « Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs ». *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-532.

Annexe 1 : Contenu supplémentaire du chapitre 2

Figures

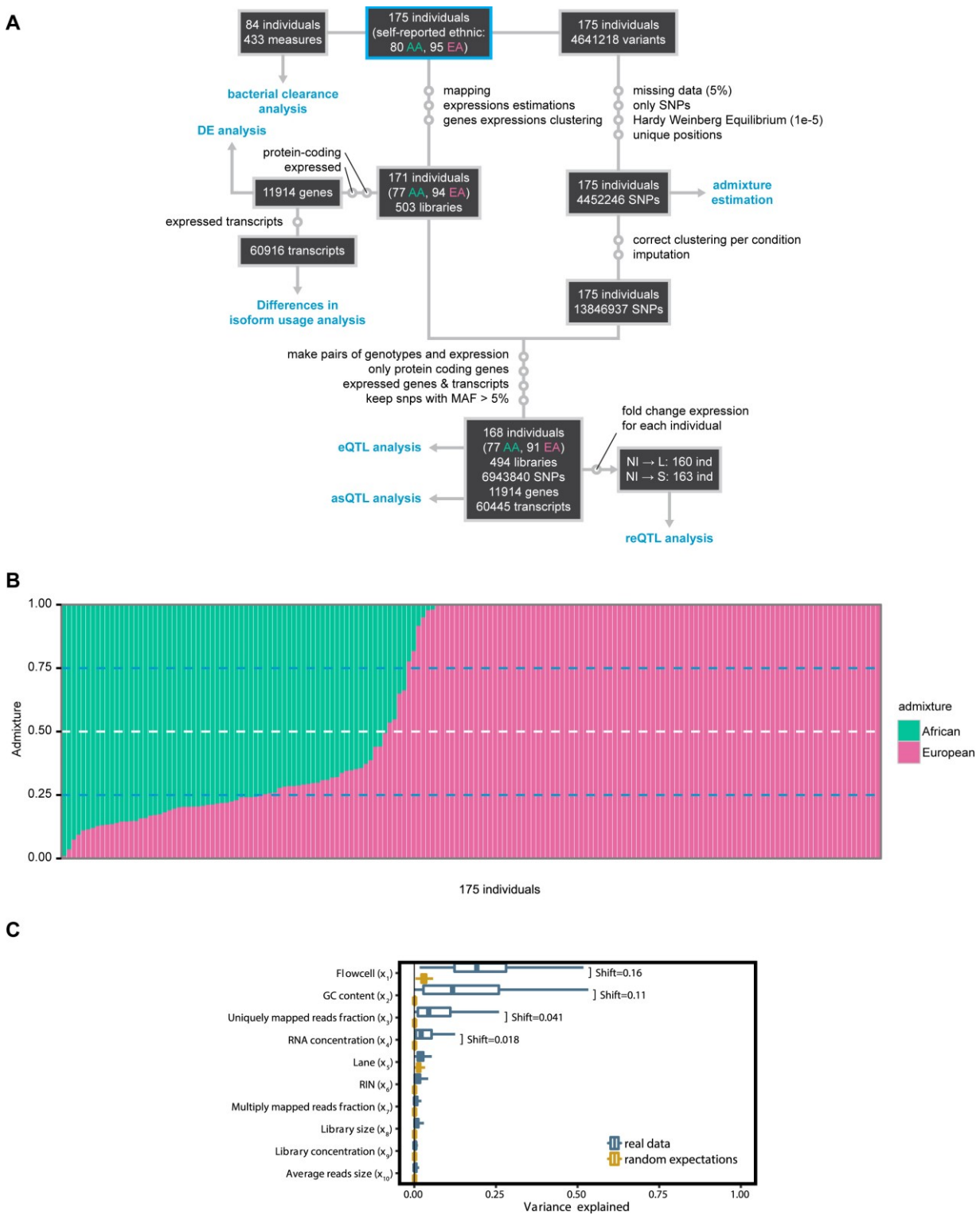

FIGURE S1: STUDY DESIGN AND EVALUATION OF TECHNICAL CONFOUNDERS, RELATED TO STAR METHODS

A. Schematic representation of the different steps used to process the RNA-sequencing and the genotyping data. The figure also depicts the number of samples and SNPs included in each of the analyses reported in the manuscript. **B.** Population structure analysis of all samples based on autosomal SNPs. Each individual is represented as a vertical line, with population origins indicated below the lines. Cluster membership proportions are depicted in green (inferred proportion of African ancestry) and pink (inferred proportion of European ancestry). **C.** Fraction of variance explained in the RNA-seq data by potential technical confounders. For each confounder, the shift between the distribution of variance explained by the real data and by random (when shuffling the real data) was estimated using a non-parametric Mann-Whitney U test. Confounders with shifts larger than 1% variance (shown in the figure) were corrected for in downstream analysis.

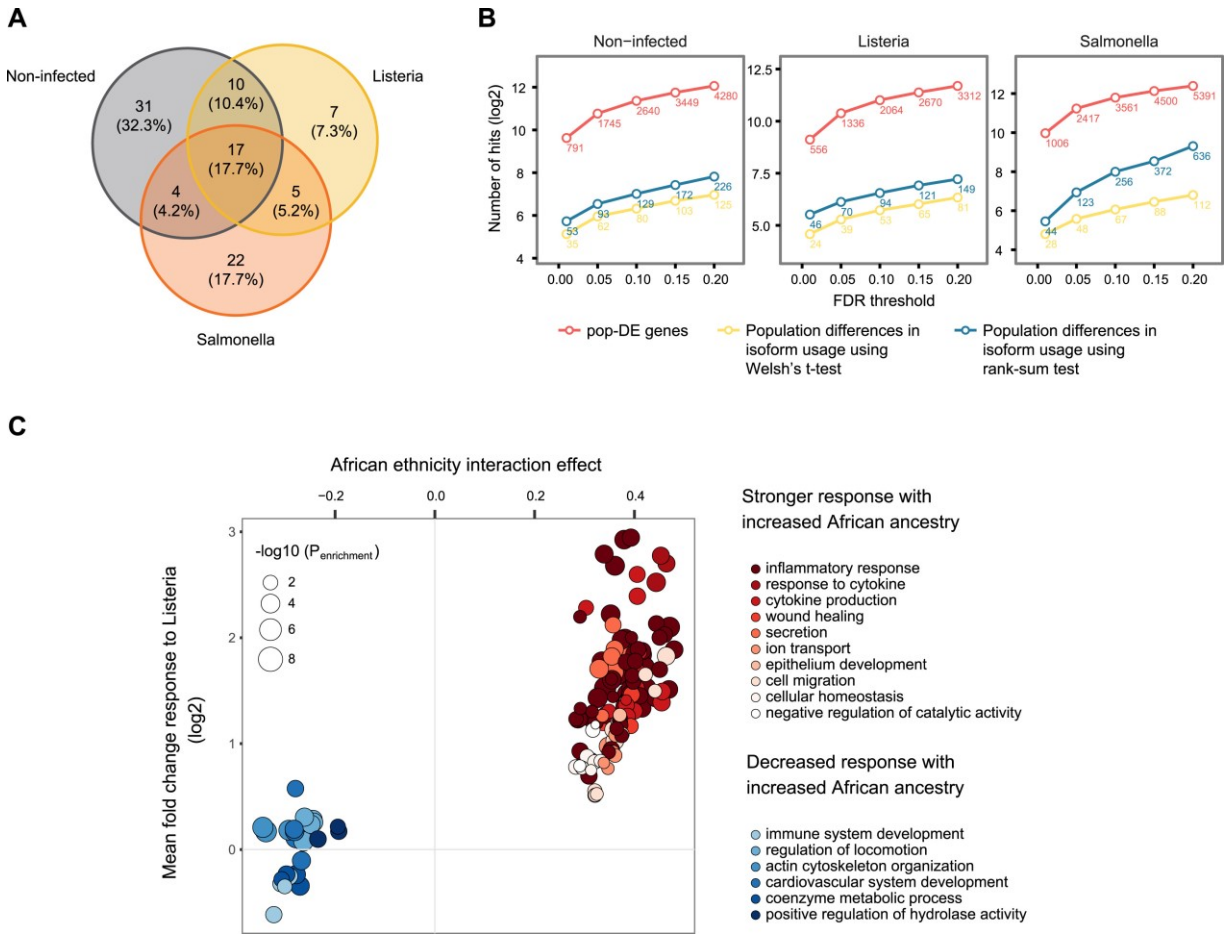

FIGURE S2: POPULATION DIFFERENCES IN GENE EXPRESSION AND ISOFORM USAGE, RELATED TO FIGURE 1

A. Venn diagram for the overlap of genes showing significant differences in isoform usage in the different experimental conditions. Non-infected, *Listeria*-infected, and *Salmonella*-infected genes with ancestry-associated differential isoform usage at FDR < 0.05 are illustrated in gray, yellow, and orange, respectively. **B.** Number of genes identified as pop-DE and with ancestry-associated changes in isoform usage at different FDR cutoffs. The number of significant genes at the different cutoffs (x axis) is reported in log2 scale (y axis). For differences in isoform usage, we report results obtained using the Welsh's t test (yellow) and the rank-sum test (blue). **C.** Gene ontology enrichment analysis for genes showing a significant interaction between ancestry and response to *Listeria*. Enrichments were performed separately for genes showing a significantly stronger and a significantly weaker response to Listeria (pop-DR) with increasing African ancestry (i.e., positive and negative interaction effects, respectively, as illustrated on the x axis). Only GO-terms with an enrichment at FDR < 0.1 are displayed,

where GO terms are grouped into clusters and colored accordingly based on the overlap among gene sets (also obtained from ClueGO's clustering functionality). For each GO-term (each circle), the average interaction effect is plotted on the x axis, against the mean log2 fold change in gene expression levels in response to infection for that term (y axis).

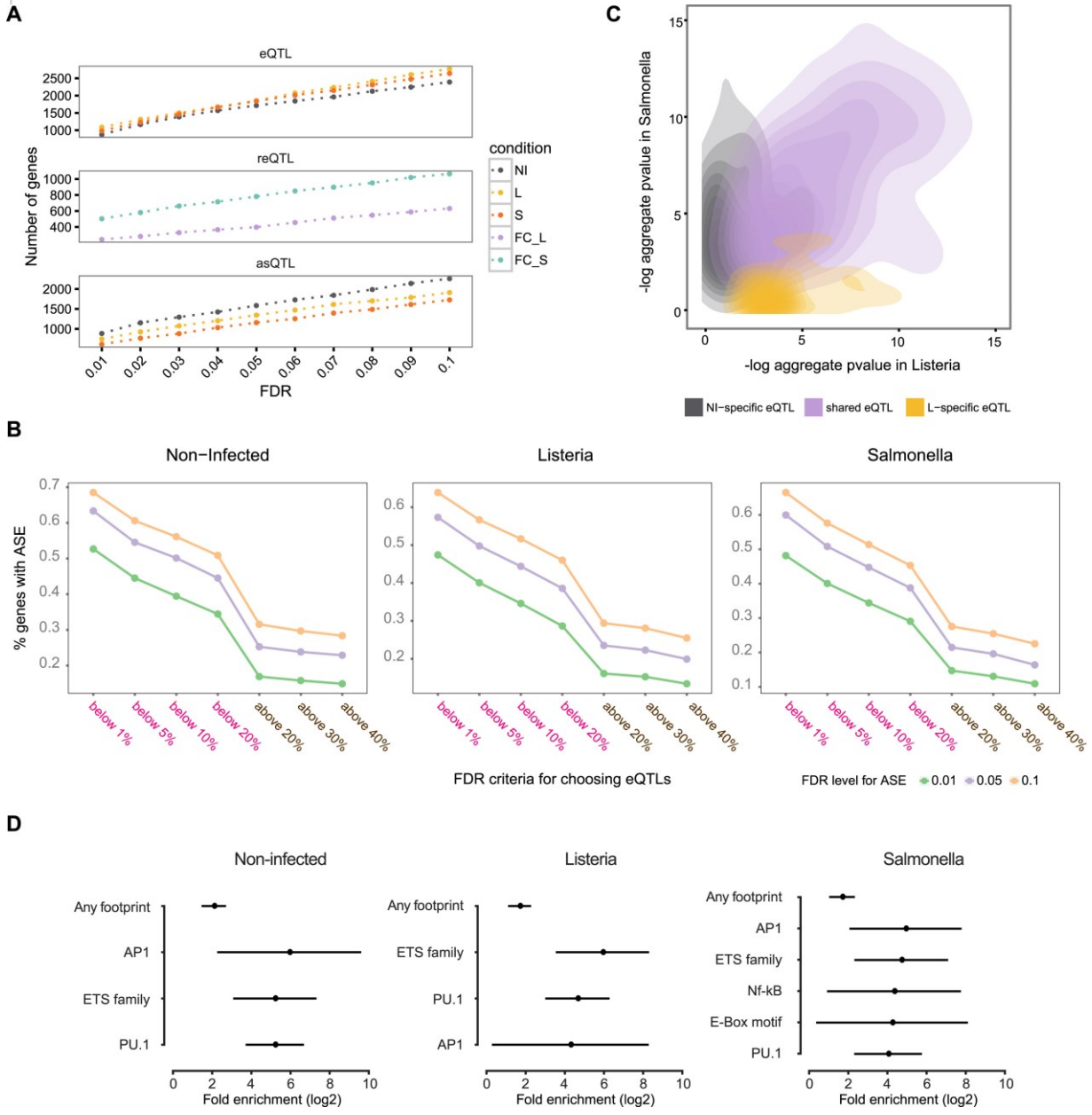

FIGURE S3: EQTL ARE ENRICHED FOR BINDING SITES OF ACTIVELY REGULATED TF BINDING SITES, RELATED TO FIGURE 3

A. Comparison of the number of genes associated with an eQTL (top), reQTL (middle), and asQTL (bottom) at increasingly higher FDR cutoffs (x axis). **B.** Percentage of genes showing significant ASE based on different FDR cutoffs adopted for ASE and cis-eQTL mapping. The plots depict the percentage of genes showing significant ASE (y axis) out of the total number of genes with cis-eQTL that pass FDR thresholds shown on the x axis. Three different FDR cutoffs are studied for ASE statistical significance, while the eQTL FDR thresholds on the x axis cover a wide range from extremely significant (to the left of the x axis) to extremely non-significant (to the right of the x axis). In particular, FDR criteria for selecting significant

and non-significant *cis* regulatory variants are illustrated on the x axis in pink and brown, respectively. **C.** Plot contrasting the evidence for ASE (-log10 *P values*) in non-infected macrophages (y axis) and in macrophages infected with *Listeria* (x axis), for genes where we identified cis-eQTL in both conditions (purple), genes for which cis-eQTL were only found in non-infected macrophages (gray), and genes for which cis-eQTL were only found in Listeria-infected macrophages (yellow). **D.** ATAC-seq eQTL enrichments (x axis) in actively-regulated TF binding sites annotated by ATAC-seq footprinting. Error bars show 95% confidence intervals. Only significant enrichments are shown. Binding sites were grouped into functionally-overlapping "TF clusters" using sequence similarity and co-localization in the genome.

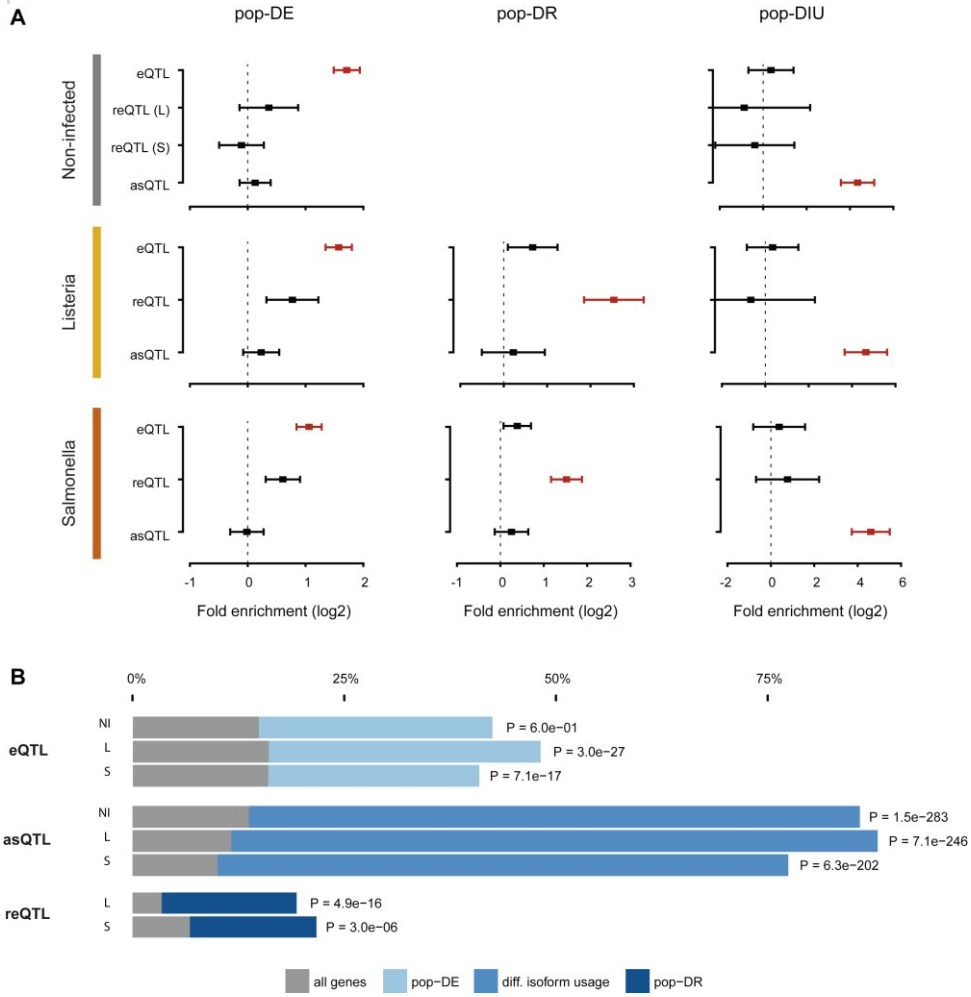

FIGURE S4: CONTRIBUTION OF *CIS* GENETIC VARIATION TO ANCESTRY-ASSOCIATED TRANSCRIPTIONAL VARIATION, RELATED TO FIGURE 4

A. Enrichment of regulatory variants among pop-DE, pop-DR and genes that exhibit ancestry-associated differential isoform usage (pop-DIU). The enrichment factors are shown on the x axis in a log2 scale. The bars around the estimated enrichments reflect the 95% confidence intervals around the estimates. For pop-DE genes, enrichments were obtained from a logistic regression model aimed at testing if pop-DE (FDR < 0.05) (as compared to non-pop-DE genes; FDR > 0.05) were enriched among cis-eQTL, cis-reQTL or cis-asQTL. The same was done for pop-DR and genes showing differences in isoform usage between populations. **B.** Proportion of pop-DE, pop-DR and genes that exhibit ancestry-associated differential isoform usage that are associated with a cis-eQTL, cis-reQTL or cis-asQTL, respectively (FDR < 0.05). Null expectations (based on the genome-wide proportion of genes associated with each QTL class) are shown in gray.

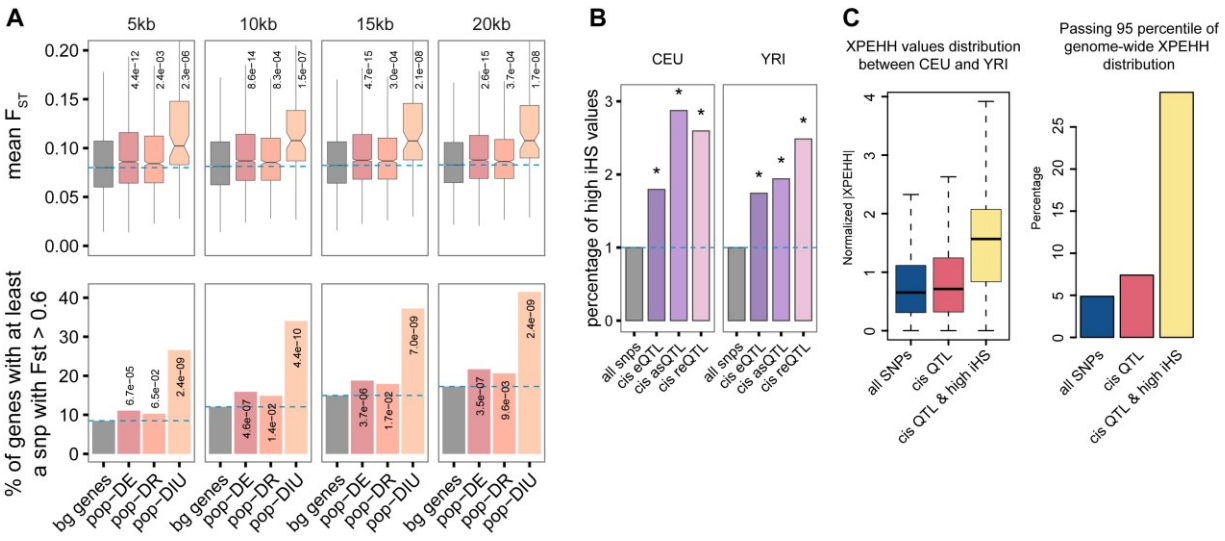

FIGURE S5: NATURAL SELECTION CONTRIBUTES TO ANCESTRY-ASSOCIATED REGULATORY DIFFERENCES, RELATED TO FIGURE 5

A. The top panel shows boxplots of mean Fst values in a window of different sizes (mentioned above the plots) around the TSS of all genes, pop-DE, pop-DR and genes showing differences in isoform usage between populations. The bottom panel shows that proportion of genes in each of the above-mentioned categories that have at least one SNP in the window with an Fst value above 0.6 (the 99^{th} percentile of the genome-wide distribution). **B.** Proportion of all SNPs, cis-eQTL, cis-reQTL, and cis-asQTL identified at an FDR < 0.05 with an iHS value above the 99th percentile of the genome-wide distribution in the CEU ($|iHS| > 2.70$) and the YRI ($|iHS| > 2.68$) populations. **C**. Boxplot showing the distribution of absolute XP-EHH values (x axis) among all *cis* SNPs tested (blue), the group of SNPs impacting one or more transcriptional phenotypes (i.e., cis-eQTL, cis-reQTL or cis-asQTL; pink), and SNPs impacting one or more transcriptional phenotypes that show an elevated iHS values (yellow). The right panel shows the proportion of SNPs (y axis) belonging to each of the groups described above that have an XP-EHH value above the 95th percentile of the genome-wide distribution. For all comparisons, QTL with an elevated iHS values show significantly higher XP-EHH values ($p < 1x10^{-10}$) as compared to all SNPs or cis-regulatory variants.

Méthodes

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Antibody against CD14	BD Biosciences	Cat#:555398; RRID: AB_395799
Antibody against CD1a	BD Biosciences	Cat#:555807; RRID: AB_396141
Antibody against CD83	BD Biosciences	Cat#:556855; RRID: AB_396526
Antibody against HLA-DR	BD Biosciences	Cat#:555561; RRID: AB_395943
Chemicals, Peptides, and Recombinant Proteins		
FICOLL-PAQUE PREMIUM	GE Healthcare	Cat#:17-5446-52
GENTAMICIN REAGENT SOLUTION LIQUID 10ML	Thermo Fisher Scientific	Cat#:15710-054
IGEPAL CA-630	Sigma Aldrich	Cat#:I3021-50ML
SYBR Green I Nucleic Acid Gel Stain - 10,000X concentrate in DMSO (500ul)	Thermo Fisher Scientific	Cat#:S7563
Triton X-100	Sigma Aldrich	Cat#:X100-500ML
FBS premium, US origin	WISENT	Cat#:80150
Human CD14 microbeads for Macs	Miltenyi Biotec	Cat#:130-050-201

L-glutamine	WISENT	Cat#:609-065-EL
NEBNext High-Fidelity 2X PCR Master Mix	New England Biolabs	Cat#:M0541S
Recombinant Human M-CSF, Animal-Free	R&D Systems	Cat#:AFL216
RPMI-1640	HyClone	Cat#:SH30096.01
Tryptic Soy Broth (TSB) (BBL Trypticase Soy Broth)	BD Biosciences	Cat#:211768
Trypticase Soy agar	BD Biosciences	Cat#:221283
Critical Commercial Assays		
Nextera DNA Sample Preparation Kit (24 samples)	Illumina	Cat#:FC-121-1030
Nextera index kit (24 indexes, 96 samples)	Illumina	Cat#:FC-121-1011
TRUSEQ RNA SAMPLE PREP KIT V2, SET A	Illumina	Cat#:RS-122-2001
DNA extraction kit (Gentra Systems)	QIAGEN	Cat#:1042606
MinElute PCR Purification Kit (50)	QIAGEN	Cat#:28004
miRNeasy Mini kit	QIAGEN	Cat#:217004

RNA Nano Chips	Agilent Technologies	Cat#:5067-1521
Deposited Data		
Raw and analyzed data	This study	GEO: GSE81046
Experimental Models: Organisms/Strains		
Listeria monocytogenes	This study	N/A
Salmonella typhimurium	This study	N/A
Software and Algorithms		
Impute2 v 2.3.0	Howie et al. (2012))	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
shapeIT v2.r790	Delaneau et al. (2013))	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
PLINK 1.9	Chang et al. (2015))	https://www.cog-genomics.org/plink2
RseQC	Wang et al. (2012))	http://rseqc.sourceforge.net
R		https://www.r-project.org/
edgeR	Robinson et al. (2010))	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Limma	Ritchie et al. (2015))	https://bioconductor.org/packages/release/bioc/html/limma.html

Matrix eQTL	Shabalin (2012))	http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/
Sva	Leek et al. (2012))	https://www.bioconductor.org/packages/release/bioc/html/sva.html
cbcbSEQ	Okrah n. d.	https://github.com/kokrah/cbcbSEQ
WASP	van de Geijn et al. (2015))	https://www.encodeproject.org/software/wasp/
SAMtools	Li et al. (2009))	http://samtools.sourceforge.net
QuASAR	Harvey et al. (2015))	https://github.com/piquelab/QuASAR
ADMIXTURE	Alexander et al. (2009))	https://www.genetics.ucla.edu/software/admixture/
Trim Galore!	Krueger n.d.	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Picard-tools	Broad Institute	http://broadinstitute.github.io/picard/
Bowtie 2	Langmead and Salzberg (2012))	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Centipede	Pique-Regi et al. (2011))	http://centipede.uchicago.edu/

STAR	Dobin et al. (2013))	https://github.com/alexdobin/STAR
RSEM	Li and Dewey (2011))	http://deweylab.github.io/RSEM/
Vcftools	Danecek et al. (2011))	http://vcftools.sourceforge.net/
CLUEGO	Bindea et al. (2009))	http://apps.cytoscape.org/apps/cluego
TORUS	Wen et al. (2016))	https://github.com/xqwen/dap/tree/master/torus_src
PGDSpider	Lischer and Excoffier (2012))	http://www.cmpg.unibe.ch/software/PGDSpider/
Arlequin	Excoffier and Lischer (2010))	http://cmpg.unibe.ch/software/arlequin35/
Selscan v1.1.0b	Szpiech and Hernandez, 2014	https://github.com/szpiech/selscan
Other		
eQTL visualization web browser		http://www.immunpop.com

Contact for Reagent and Resource Sharing

Reagent and resource requests should be addressed and will be fulfilled by the Lead Contact, Luis Barreiro (luis.barreiro@umontreal.ca).

Experimental Model and Subject Details

Sample Collection

Buffy coats from 175 healthy donors were obtained from the Indiana Blood Center (Indianapolis, IN, USA). A signed written consent was obtained from each participant and the project was approved by the ethics committee at the CHU Sainte-Justine (protocol #4022). All individuals recruited in this study were males, self-identified as African-American (AA) (n = 76) or European-American (EA) (n = 99) between the age of 18 and 55 years old. The average age across AA and EU samples was similar (34.2 years (AA) versus 35 years (EA), t test, p = 0.7). We decided to only focus on males to avoid the potentially confounding effects of sex-specific differences in immune responses to infection. Only individuals self-reported as currently healthy and not under medication were included in the study. In addition, each donor's blood was tested for Hepatitis B, Hepatitis C, Human Immunodeficiency Virus (HIV), and West Nile Virus, and only samples negative for all of the tested pathogens were used.

Method Details

Isolation of Monocytes and Differentiation of Macrophages

Blood mononuclear cells were isolated by Ficoll-Paque centrifugation. Monocytes were purified from peripheral blood mononuclear cells by positive selection with magnetic CD14 MicroBeads (Miltenyi Biotech) using the autoMACS Pro Separator. The purity of the isolated monocytes was verified using an antibody against CD14 (BD

Biosciences) and only samples showing $>$ 90% purity were used to differentiate into macrophages. Monocytes were then cultured for 7 days in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent), L-glutamine (Fisher) and M-CSF (20ng/mL; R&D systems). Cell cultures were fed every 2 days with complete medium supplemented with the cytokines previously mentioned. Before infection, we checked the differentiation/activation status of the monocyte-derived macrophages by flow cytometry, using antibodies against CD1a, CD14, CD83, and HLA-DR (BD Biosciences). Only samples presenting the expected phenotype for non-activated macrophages (CD1a+, CD14+, CD83, and HLA-DRlow) were used in downstream experiments.

Bacterial Preparation and Infection of Macrophages

We infected macrophages with two bacteria, *Salmonella typhimurium* and *Listeria monocytogenes*. The day prior to infection, aliquots of *Salmonella typhimurium* and *Listeria monocytogenes* were thawed and bacteria were grown overnight in Tryptic Soy Broth (TSB) media. Bacterial culture was diluted to mid-log phase prior to infection and supernatants density was checked at OD_{600}.

Monocyte-derived macrophages were infected at a multiplicity of infection (MOI) of 10:1 for *Salmonella typhimurium* and an MOI of 5:1 for *Listeria monocytogenes* for 2h at 37°C. A control group of non-infected macrophages was treated the same way but with only medium without bacteria. After 2 hr in contact with the bacteria, macrophages were washed and cultured for another hour in the presence of $50\mu g/ml$ gentamycin in order to kill all extracellular bacteria present in the medium. The cells were then washed a second time and cultured in complete medium with $3\mu g/ml$ gentamycin for an additional 2h, the time point we refer to in the main text. A control

group of non-infected macrophages was treated the same way but with only medium without bacteria. We note that we did not run technical replicates for the infections because we could not derive sufficient macrophages from one individual to perform multiple infections with both bacteria. However, the impact of technical confounds are reduced by our large set of biological replicates (and are probably overall small, given our power to detect so many eQTL and ancestry-associated responses).

Estimation of the Number of Infected Macrophages

To determine bacterial counts in infected cells, monolayers of $2 \cdot 10^6$ infected macrophages in 6-well plates were used. Culture medium was removed and replaced with 1ml of 1% Triton X-100 in distilled water. Serial 10-fold dilutions were made, in duplicates, in Trypticase Soy broth and plated on Trypticase Soy agar plates. Plates were kept at 37°C and counted after 24h. Enumeration of intracellular bacteria was performed at T0, corresponding to the percentage of infected macrophages, and T2 and T24, corresponding to the number of bacteria inside the macrophages 2- and 24 hr post-infection, respectively. Data was collected for T0 for all the samples (to control for variation in the number of infected macrophages among individuals), and for T2 and T24 for a subset of 89 individuals for which enough macrophages were available to perform the experiment.

RNA Extraction, Library Preparation, and Sequencing

Total RNA was extracted from the non-infected and infected macrophages using the miRNeasy kit (QIAGEN). RNA quantity was evaluated spectrophotometrically, and the quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Only samples with no evidence for RNA degradation (RNA integrity number > 8) were kept for further experiments. RNA-sequencing libraries were prepared using the

Illumina TruSeq protocol. Once prepared, indexed cDNA libraries were pooled (6 libraries per pool) in equimolar amounts and were sequenced with single-end 100bp reads on an Illumina HiSeq2500. Samples were carefully balanced across flow cells and sequencing lanes. Specifically, we multiplexed infected and non-infected samples from the same individual in the same lane, and balanced the number of African Americans and European Americans in each of the flowcells. Additionally, we multiplexed non-infected and infected macrophages (*Salmonella* and *Listeria*) from one European American and one African American in each lane. Because we had a larger number of European ancestry samples than African ancestry samples, the ideal 50-50 ratio was significantly violated for samples sequenced in two of 16 total flowcells. Yet, these samples account for only 5% of all the RNA-seq libraries sequenced. Sequencing libraries from both infected and non-infected conditions were always prepared in parallel with a balanced amount of samples derived from EA and AA individuals.

ATAC-Seq Library Preparation and Sequencing

ATAC-seq libraries were generated from 100,000 cells, as previously described in (Buenrostro et al., 2013) and sequenced on an Illumina HiSeq 2500 using 100-bp paired-end reads. We found high concordance between the ATAC-seq signals for the two biological replicates sequenced for each of the conditions (Spearman $r > 0.80$), which allowed us to merge them for downstream footprint analyses.

DNA Extraction and Genome-wide Genotyping

DNA was extracted from each of the blood samples using the PureGene DNA extraction kit (Gentra Systems). Each individual was genotyped for over 4.6 million single nucleotide polymorphisms (SNPs), using the Illumina HumanOmni5Exome BeadChip, which interrogates $>$ 4.3 million whole-genome variants, plus the content of

the Illumina exome BeadChip. Genotypes were called in all samples together using Genome Studio v2010. All samples had genotype call rates (CR) above 98%, with the exception of 2 samples that were excluded from further analysis. SNPs with $> 5\%$ of missing data or deviating from Hardy–Weinberg equilibrium in at least one of the studied populations (at a $p < 10^{-5}$) were excluded. In total, 4,452,246 SNPs passed our quality-control filters. Since samples were collected anonymously, we systematically tested for relatedness in our samples by estimating the pair-wise genome-wide identity by state (IBS) between all possible pairs of individuals using PLINK (Chang et al., 2015). We found 2 pairs of individuals that appeared to be genetically identical, suggesting that these pairs of sample are from the same individual that donated blood twice during our recruitment process. Therefore, we randomly excluded the data of one individual from each of these pairs. All other samples were unrelated as defined by an estimated proportion of IBS < 0.2. Finally, all samples were confirmed to be males on the basis of the genotype data from the X chromosome. After various quality control checks, we ended up with 171 individuals for which genotype data was available for eQTL analyses.

Quantification and Statistical Analysis

Imputation

Imputation was done using Impute2 (ver. 2.3.0) (Howie et al., 2012), on the pre-filtered genotype data and using as reference panels phased genotype data from phase 3 of the 1000 Genomes project (downloaded from: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). Our genotype data was phased (per chromosome) using shapeIT (version 2.r790). Post-imputation, we removed genotype calls with likelihood lower than 0.9. In addition, we removed

positions with an information metric lower than 0.5, more than 5% of missing genotype calls or deviating from Hardy–Weinberg equilibrium in at least one of the studied populations (at $P < 10^{-5}$). After all filters, we kept 13,846,937 SNPs.

Estimation of Genome-wide Admixture Levels

Self-reported EA and AA have variable degrees of African and European ancestry. In particular, the genome-wide levels of European genetic ancestry among self-reported AAs average 30% and can attain close to 100% in some individuals (Bryc et al., 2010). Thus, instead of relying on self-reported ancestry labels, we calculated the actual proportion of European and African ancestry for each of the samples included in the study using the unsupervised clustering algorithm ADMIXTURE (Alexander et al., 2009). We included 56 Yoruban samples in our analyses to have a group of African individuals that are arguably not admixed. A total of 86,329 unlinked SNPs (*i.e.*, r^2 between all pairs < 0.1) were used for ancestry assignments, assuming $K = 2$ ancestral clusters. The estimated ancestry proportions were used to assess differences in immune responses between populations, unless mentioned otherwise.

Estimation of Gene- and Isoform-Level Expressions

Adaptor sequences and low quality score bases (Phred score < 20) were first trimmed using Trim Galore (version 0.2.7). The resulting reads were then mapped to the human genome reference sequence (Ensembl GRCh37 release 75) using STAR (2.4.1d) (Dobin et al., 2013) with a hg19 transcript annotation GTF file downloaded from Ensembl (date: 2014-02-07).

The following parameters were used for STAR index generation (other than default):

- `genomeSAindexNbases 2`
- `genomeChrBinNbits 14`
- `sjdbOverhang 99`

In order to obtain aligned reads in transcriptome coordinates, we used the following options specifically recommended for downstream analysis with RSEM:

- `outSAMattributes NH HI`
- `outFilterMultimapNmax 20`
- `outFilterMismatchNmax 999`
- `outFilterMismatchNoverLmax 0.05`
- `alignIntronMin 20`
- `alignIntronMax 1000000`
- `alignSJoverhangMin 8`
- `alignSJDBoverhangMin 1`
- `quantMode TranscriptomeSAM`

Transcript- and gene-level expression estimates were calculated using RSEM (version 1.2.21) (Li and Dewey, 2011), with default parameters considering a mean and standard deviation of 178bp and 58bp, respectively, for insert sizes across our RNA-seq libraries.

Differences in Expression between Populations and in Response to Infection

Quality Control

A total of 22 RNA-seq libraries (out of 525 in total) were removed from downstream analyses because the genotype calls made on the RNA-seq data did not match those obtained from the genotyping array (n = 12), the non-infected samples were clustering close to infected samples in a principal component analysis (n = 4) or the *Listeria*-infected samples were clustering together with *Salmonella*-infected samples (n = 6).

We subset our phenotype data by keeping protein-coding genes that were sufficiently expressed: median TPM value above 0.5 in at least one of the three conditions.

Identification of Relevant Technical Confounders

As a preliminary step for the differential expression analysis, we aimed at identifying confounders that amounted to unwanted technical sources of variability in the expression data. To do this in a systematic way, we began by considering the following pool of putatively relevant technical confounders:

- x_1: sequencing flowcell
- x_2: mean GC content estimated per sample (using RseQC (Wang et al., 2012))
- x_3: fraction of uniquely mapped reads
- x_4: RNA concentration post-extraction
- x_5: sequencing lane
- x_6: RNA integrity numbers (RIN)
- x_7: fraction of multiply mapped reads
- x_8: Total number of sequenced reads
- x_9: RNA library concentration used for sequencing
- x_{10}: library insert size (based on Bioanalyzer)

two of which (*i.e.*, sequencing flowcell and lane) are categorical variables, while the rest are continuous variables that were standardized before the analysis (*i.e.*, rescaled to have $mean = 0$ and $sd = 1$). In order to identify the confounding variables, among the above-mentioned list, that explain a significant amount of the variance in the data, we implemented the following iterative procedure:

- *Step 1:* Let $M_{ref} : E \sim 1$ denote the reference model with no covariates, where only an intercept is estimated for the gene expression data E. In addition, assume that $M_i : E \sim x_i$ models the gene expression data by only considering the ith technical confounder as the covariate, for $i \in \{1, \ldots, 10\}$. The fraction of variance

in the expression data explained by the ith technical covariate was then estimated by $v_i = (SS_{M_{ref}} - SS_{M_i})/SS_{M_{ref}}$ for each gene, where $SS_{M_{ref}}$ and SS_{M_i} represent the residual sum of squares in M_{ref} and M_i, respectively.

- *Step 2:* For each technical confounder listed above, the following procedure was repeated for $N_{iter} = 200$ iterations per gene: The entries of the original confounder vector x_i were permuted and the permuted vector was denoted by $\tilde{x}_i$. Afterward, the randomized model $M_{rand(i)} : E \sim x_i$ was set up. The expected amount of variance explained by the randomized confounding variable was then estimated by $v_{rand(i)} = (SS_{M_{ref}} - SS_{M_{rand(i)}})/SS_{M_{ref}}$, where $SS_{M_{rand(i)}}$ denotes the residual sum of squares for $M_{rand(i)}$.

- *Step 3:* For each confounder, the distribution of v_i (*i.e.*, the observed or true fraction of variance explained by the confounder) across all genes was compared to the corresponding distribution of randomized values, $v_{rand(i)}$, through the non-parametric Mann–Whitney U test. The shift between these two distributions at a significance level of p = 0.05 is denoted by δ_i for the ith confounder.

- *Step 4:* We compared the δ_i values across the ten confounders and chose the technical confounder with the maximum shift $\delta_{i^*} := \max_i \delta_i$. If $\delta_{i^*} > 0.01$ (*i.e.*, the contribution of this confounder in explaining the variability in the data is least 1% more than that of an arbitrary random variable), then the confounder was selected and added as a covariate to the reference model M_{ref}.

- *Step 5:* We repeated Steps 1 to 3, using the updated reference model. After re-evaluating the distribution shifts δ_i in Step 3, we proceeded as follows: (i) Among the set of confounders currently present in M_{ref}, the one with the lowest amount of shift was removed from M_{ref}, given that the shift was below 0.01. (ii) Among

the set of confounders currently absent in M_{ref}, the one which satisfied the selection procedure described in Step 4 was added to M_{ref}.

- *Step 6:* Step 5 was repeated until we obtained a reference model where, out of the ten studied confounders, only the covariates present in M_{ref} satisfied the condition mentioned in Step 4 (*i.e.*, their contribution in explaining the variability in the data is least 1% more than that of an arbitrary random variable).

It turned out that only five iterations of the above procedure were sufficient, leading to the following reference model:

$$(1)\; M_{ref} : E \sim x_1 + x_2 + x_3 + x_4$$

containing four technical confounders that are controlled for in the downstream analysis (see Figure S1C).

Data Pre-processing

To account for differences in read counts at the tails of the distribution, we normalized the samples using the weighted trimmed mean of M-values algorithm (TMM), as implemented in the R package edgeR (Robinson et al., 2010). Afterward, we log-transformed the data using the voom function in the limma package (Ritchie et al., 2015) and removed the flowcell batch effect using the ComBat function in sva Bioconductor package (Leek et al., 2012). We then applied the voomMod function from package cbcbSEQ (https://github.com/kokrah/cbcbSEQ/blob/master/README.md), specifically devised to work on log-transformed data as opposed to voom which works on count data, to recover new sample weights for the batch-corrected data. Following this pre-processing of the data, we fitted the log-transformed expression estimates to linear

models (with design details explained in the subsequent paragraphs), using the lmFit function from the limma package (Ritchie et al., 2015). This function uses the sample weights previously estimated, from the overall mean-variance trend by voomMod, to rescale model residuals and improve the quality of the fit. In these models, the three numerical confounders shown in Equation 1 (*i.e.*, x_2, x_3 and x_4; GC means, RNA concentrations, and fractions of uniquely mapped reads, respectively) are introduced as model covariates. Note that the categorical confounder x_1, or flowcell, has already been corrected for using ComBat. Finally, differential expression effects across conditions (DE) and across populations (pop-DE), along with Ethnicity Condition interaction effects resulting in differential response across populations (pop-DR), were estimated using these linear models. In what follows, each model is elaborately explained.

Ancestry-Related Differential Expression

The following nested linear model was used to identify genes for which expression levels correlated with the African-ancestry levels estimated for each of our samples:

(2)

$$M_1 : E(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & if \quad Condition = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & if \quad Condition = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & if \quad Condition = S \end{cases}$$

Here, $E(i,j)$ *shows the expression level of gene* i for individual j, $\beta_{Af}^{NI}(i)$, $\beta_{Af}^{L}(i)$, and $\beta_{Af}^{S}(i)$ indicate the effects of African admixture *(Af)* on gene i within each condition, $\beta_L(i)$ and $\beta_S(i)$ represent the intrinsic infection effects of each pathogen, and $\beta_c^L(i)$ and $\beta_c^S(i)$ are the effects of the standardized bacterial counts (denoted by c_L and c_L)

registered in the samples immediately after infection with each pathogen. Furthermore, $\{x_k, k=2,3,4\}$ represents the three numerical covariates previously detected as significant technical confounders; i.e., mean GC content per sample, RNA concentration, and fractions of uniquely mapped reads, with β_{x_k} being their corresponding effects on gene expression. Finally, $e^C(i,j)$ represents the residuals at condition C (NI, L or S) for the gene-i individual-j pair, and $\beta_o(i)$ is the global intercept accounting for the expected expression of gene i in a 100% European non-infected sample (i.e., $Af = 0$). Note that for each individual, we assessed only one sample per condition. In other words, no technical replicates were used in the design.

Fitting the model using the Bioconductor's limma pipeline (Ritchie et al., 2015), we extract the estimates $\beta^{NI}_{Af}(i)$, $\beta^{L}_{Af}(i)$, and $\beta^{S}_{Af}(i)$ of the ethnicity effects across all genes, along with their corresponding p values. Each of these estimates represents the ancestry-related differential expression effects within each condition (pop-DE). Afterward, we control for false discovery rates using an approach analogous to that of Storey and Tibshirani (Storey and Tibshirani, 2003), which makes no explicit distributional assumption for the null-model but instead derives it empirically from 200 permutation tests, where African admixture values are permuted across individuals (see section "Estimation of false discovery rates" below for details). Before proceeding to the table below, which includes results and further details on these effects, some notation is introduced. Let $\langle E_C \rangle_{Af=x}$ denote the expected expression value for a gene in condition $C \in \{NI, L, S\}$ for an individual with African admixture $Af = x$, under M_1. According to this definition, $\langle E_C \rangle_{Af=1} - \langle E_C \rangle_{Af=0}$ represents the expected African ancestry effect within condition C. This effect is denoted by pop-DE:C in the table below and by the β^C_{Af} coefficient in model M_1.

Within Condition Ancestry-Associated Differences in Gene Expression				
Pop-DE Effect	**Linear Model Coefficient M_1**	**No. Genes under 0.05 FDR**	**Permuted Variable at Null Model**	**Estimated Fraction of True Negatives π_o**
Pop-DE:NI $\langle E_{\mathrm{NI}}\rangle_{\mathrm{Af}=1} - \langle E_{\mathrm{NI}}\rangle_{\mathrm{Af}=0}$	β_{Af}^{NI}	1,745	African ancestry across individuals	0.601
Pop-DE:L $\langle E_{\mathrm{L}}\rangle_{\mathrm{Af}=1} - \langle E_{\mathrm{L}}\rangle_{\mathrm{Af}=0}$	β_{Af}^{L}	1,336	African ancestry across individuals	0.658
Pop-DE:S $\langle E_{\mathrm{S}}\rangle_{\mathrm{Af}=1} - \langle E_{\mathrm{S}}\rangle_{\mathrm{Af}=0}$	β_{Af}^{S}	2,417	African ancestry across individuals	0.528

Infection Effects: Condition-Related Differential Expression

Contrary to the case of pop-DE analysis, expression levels of samples corresponding to the same individuals are compared in order to test for global infection effects (condition-related DE). To this end, a paired design is used, in which individuals are introduced as additional covariates:

(3)

$$M_2 : E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if } Condition = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{L}(j) + \varepsilon^{L}(i,j) & \text{if } Condition = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{S}(j) + \varepsilon^{S}(i,j) & \text{if } Condition = S \end{cases}$$

Specifically, $\beta_o(i,j)$ represents the intercept corresponding to gene i and individual j; *i.e.*, the model's expectation for the expression level of gene i at the non-infected sample of individual j. Analyzing model M_2 results in the global (condition-related or ethnicity-independent) estimates of *Salmonella* and *Listeria* infection effects, β_L and β_S, approximated using the within-individual variation in gene expression across conditions.

Similar to the previous model, M_2 is fit using limma; however, the 200 permutation tests implemented here to estimate FDRs are based on random reshuffling of condition labels within each individual (see the table below); moreover, considering the large effect of infection on gene expression, FDRs are obtained from Benjamini-Hochberg's more conservative approach in order to avoid false positives. In the table below, $\langle E_C - E_{NI} \rangle$ shows the expected response upon infection with pathogen C ∈ {L, S} (or equivalently, C infection effect), which is denoted by DE:C in the table and by the β_C coefficient in model M_2.

Condition-Related Differential Expression Effects			
Condition DE Effect	**Linear Model Coefficient M_2**	**No. Genes under 0.05 FDR (BH)**	**Permuted Variable at Null Model**
DE:L $\langle E_{\mathrm{L}} - E_{\mathrm{NI}} \rangle$	β_{L}	10,663	conditions within individual
DE:S $\langle E_{\mathrm{S}} - E_{\mathrm{NI}} \rangle$	β_{S}	10,751	condition within individual

Infection-Ethnicity Interactions: Ancestry-Associated Differential Response to Infection (pop-DR genes)

After obtaining global infection effects, we explored for genes whose response to infection significantly depend on ethnic ancestry. Specifically, we fit the following linear model:

(4)

$$M_3: E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if } Condition = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & \text{if } Condition = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & \text{if } Condition = S \end{cases}$$

which is quite similar to M_2 with the difference that the infection effect of, say, *Listeria* is no longer built in an ethnicity-independent fashion as in model M_2 (*i.e.*, $\langle E_L - E_{NI} \rangle_{M_2} = \beta_L$), since it is in fact dependent on ethnic ancestry as follows: $\langle E_L - E_{NI} \rangle_{M_3} = \beta_L + \beta_{Af}^L(i) \cdot Af$. In this framework, β_{Af}^L and β_{Af}^S denote ethnicity-infection interactions, which represent variations in response to infection observed across ethnic groups (pop-DR). Similar to previous models, 200 permutations were implemented here to estimate FDRs (see the table below for details). According to the notation introduced for models M_1 and M_2, $\langle E_C - E_{NI} \rangle_{Af=x}$ denotes the expected response upon infection with pathogen $C \in \{L, S\}$ for an individual of African admixture $Af = x$. It follows that $\langle E_C - E_{NI} \rangle_{Af=1} - \langle E_C - E_{NI} \rangle_{Af=0}$ represents the infection-ethnicity interaction induced by pathogen C (or, C infection-ethnicity interaction). This interaction term is denoted by pop-DR:C for pathogen C in the table below and by the β_{Af}^C coefficient in model M_3.

Ancestry-Associated Differential Response to Infection

Interaction Effect pop-DR	**Linear Model Coefficient M_3**	**No. Genes under 0.05 FDR (BH)**	**Permuted Variable at Null Model**	**Estimated Fraction of True Negatives π_o**
Pop-DR:L $\langle E_{\mathrm{L}} - E_{\mathrm{NI}} \rangle_{\mathrm{Af}=1} - \langle E_{\mathrm{L}} - E_{\mathrm{NI}} \rangle_{\mathrm{Af}=0}$	β_{L}	206	African admixture across individuals	0.683
Pop-DR:S $\langle E_{\mathrm{S}} - E_{\mathrm{NI}} \rangle_{\mathrm{Af}=1} - \langle E_{\mathrm{S}} - E_{\mathrm{NI}} \rangle_{\mathrm{Af}=0}$	β_{S}	1,005	African admixture across individuals	0.631

Considering that, and taking *Listeria* as an example, from M_3 we can build the mentioned expected response upon infection for African-Americans: $\langle E_L - E_{NI} \rangle_{Af=1} = \beta_L + \beta_{Af}^L$, and compare it against the corresponding effect in Europeans: $\langle E_L - E_{NI} \rangle_{Af=0} = \beta_L$, as it is done in Figure 1F, after extracting absolute values.

Applying models M_2 and M_3, instead of M_1, allows us to obtain estimates that are solely based on the within-individual variability. The upside to considering only within-individual variability is that despite the many degrees of freedom consumed by the individual-specific offsets $\beta_o(i,j)$, it augments the statistical power for detecting both global infection effects and ethnicity-infection interaction effects.

False Discovery Rate Estimations

Throughout the paper, (unless stated otherwise), FDRs were calculated separately for each dataset, following a procedure analogous to that proposed by Storey and Tibshirani (Storey and Tibshirani, 2003), which can be described as the following two-component model:

(5)

$$F(p) = \pi_o F_o(p) + (1 - \pi_o) F_A(p)$$

where $F_o(p)$ represents the cumulative density of p values for tests truly fulfilling the null hypothesis (*i.e.*, true negatives) and $F_A(p)$ is the equivalent cumulative distribution for tests truly verifying the alternative hypothesis (*i.e.*, true positives). In addition, π_o refers to the fraction of true negatives of the experiment. If the null cumulative distribution $F_o(p)$ is approximately linear (or equivalently, the p values are uniformly distributed under the null hypothesis), the above-mentioned model reduces to Storey and Tibshirani's model, corresponding to the case with $F_o(p) = p$. However, when null distributions deviate from uniform (for example, when the most strongly associated variant is assigned as a single eQTL for a gene), comparisons to empirical, permutation-based null distributions are more appropriate. Indeed, this approach, which requires a minor modification to the method in Storey and Tibshirani, is also appealing because it avoids any assumptions about uniformity. We thus elected to use it here, despite the fact that our empirical nulls are consistently uniform or close to uniform.

Here, we use the empirical cumulative distribution functions (ECDFs) $\widetilde{F_o}(p)$ and $\tilde{F}(p)$ as estimates of $F_o(p)$ and $F(p)$, respectively. To be more specific, $\tilde{F}(p)$ is the ECDF of the actual p values of any effect of interest (either pop-DE, pop-DR or response to infection, for example), whereas $\widetilde{F_o}(p)$ denotes the ECDF obtained from a suitable permutation test performed on that effect.

From Equation 5, the fraction of true negatives under a give p value can be derived as $\pi_o F_o(p)/F(p)$; or $\pi_o \tilde{F}_o(p)/\tilde{F}(p)$, once we accept the ECDF as pertinent estimators of the underlying distributions. Correcting that fraction to ensure monotonicity yields the definition of the tail-area-based false discovery rate FDR(p):

(6)

$$FDR(p) = \min_{p' \geq p} \left(\frac{\pi_o \tilde{F}_o(p')}{\tilde{F}(p')} \right)$$

In order to compute $FDR(p)$, we must first estimate π_o. As proposed in (Storey and Tibshirani, 2003), this is achieved by :

(7)

$$\hat{\pi}_o(p) = \frac{1 - F(p)}{1 - F_o(p)} \approx \frac{1 - \tilde{F}(p)}{1 - \tilde{F}_o(p)}$$

yielding a biased estimator of π_o, where the amount of bias declines as p approaches the maximum p value registered in the experiment, p_{max}. Therefore, to obtain a better estimation of π_o, the estimator $\hat{\pi}_o(p)$ is fitted to a suitable smooth function f(p) - typically a decreasing cubic spline- evaluated at p_{max} : $\pi_o \simeq f(p_{max})$.

Differential Isoform Usage between Populations

Prior to performing the DIU analysis, we removed the lowly expressed isoforms and only kept those with median TPM value (strictly) above zero in at least one of the three experimental conditions, using isoform-level TPM values estimated by RSEM. In the next step of data pre-processing, the ComBat function in the sva Bioconductor package (Leek et al., 2012) was applied to the log-transformed isoform-level TPM data to remove the flowcell batch effect. Then, the following linear model was designed using limma (Ritchie et al., 2015):

(8)

$$M_4 : log_2\left(TPM + 10^{-5}\right)(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + e^{NI}(i,j) & if \quad Condition = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + e^L(i,j) & if \quad Condition = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + e^S(i,j) & if \quad Condition = S \end{cases}$$

where 0.00001 (*i.e.*, $0.001 \times \min_{TPM > 0} TPM$) was added to all the TPM values to avoid instances of $\log(0)$. Note that this model has the same design as model $M_1$for pop-DE effects. After obtaining covariate estimates for this model, the effect of the technical confounders (*i.e.*, mean GC content, fraction of uniquely mapped reads, and RNA concentration) were regressed out from the log-transformed isoform-level TPM values. Finally, the obtained values were transformed back, from log_2-scale, to TPM-scale. These values were then used in the downstream DIU analysis. To detect differences in isoform usage between African-Americans and European-Americans, we applied a multivariate generalization of the Welch's t test to the set of 10223 genes, out of the total number of genes, with at least two an-notated isoforms (which remained after the elimination of lowly expressed isoforms in the pre-processing step). To implement the method, we started by calculating the proportional abundance of the different isoforms for each tested gene using the isoform-level TPM values estimated by RSEM. The proportional isoform abundance (or relative isoform usage) for a target gene g with D isoforms is denoted by a vector of size D, where its elements sum to one and the ith element denotes the proportional abundance of isoform i. Next, we tested whether the means of the two multivariate distributions, associated with African-American and European-American populations, were equal. Specifically, suppose that group i consists of $n_i$samples, and let $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijD})$ be the vector of proportional isoform abundance for sample j of group I, with $i \in \{1, 2\}$ and $j \in \{1, \ldots, n_i\}$. Clearly, $\sum_{d=1}^{D} \pi_{ijd} = 1$, with π_{ijd}

denoting the relative isoform usage associated with isoform d. Following this notation, the relative isoform usage data falls into the category of compositional data (Aitchison, 1982), where components (or vector elements) are proportions of total isoform abundance that sum to one. Mathematically, the state space of such compositional data is defined as an open simplex (*i.e.*, a generalization of the notion of a two-dimensional triangle to higher dimensions) as follows (Egozcue et al., 2003):

(9)

$$S^D = \left\{ (x_1, \ldots, x_D) \,|\, x_d > 0 \;\forall\; d \in \{1, \ldots, D\}, \sum_{d=1}^{D} x_d = 1 \right\}$$

The fact that the proportions have a fixed sum implies that (1) there is dependency between relative isoform usage values within each sample and (2) S^D is not a vector space. This results in specific numerical characteristics, which interfere substantially with the approaches taken in the statistical analysis of compositional data. For one thing, the familiar Euclidean geometry cannot be applied when dealing with compositional data; specifically, although the distance between two real vectors can be easily computed with the standard Euclidean metric, it is not the proper metric to use for compositional data. To illustrate, consider the following pairs of compositions: {(0.25, 0.05, 0.7), (0.25, 0.1, 0.65)} and {(0.25, 0.5, 0.25), (0.25, 0.55, 0.2)}. The Euclidean distance between the compositions in the first pair equals that of the second pair, as the element-wise difference between the compositions is (0, 0.05, 0.05) for both pairs. However, the second component has doubled in the first pair, while it has only increased by ten percent in the second pair. The fold changes associated with the third components are more comparable between the pairs (0.9 and 0.8 for the first and the second pairs, respectively). In other words, while the Euclidean distances between compositions of both pairs are equal, fold changes imply that the actual distance is

larger for the first pair. Therefore, the relative variation of components, rather than their absolute differences, provide the basis to the statistical analysis of compositional data. Lending a linear vector space structure to the open simplex, the Aitchison geometry (Aitchison, 1982) provides us with a way to work with compositional data that is analogous to the real space. Any statistical analysis on compositional data can be performed using this vector space structure; however, it is easier to use alternative methods, which transform compositional data to the familiar Euclidean space. Egozcue et al. (Egozcue et al., 2003) proposes the isometric log-ratio transformation (*ilr*), which is obtained with orthogonal coordinates. Using a set of D-1 orthonormal vectors as the basis for S^D, the *ilr* transformation maps the log of a given composition to a vector of size D-1 in the Euclidean space. The advantage of applying this distance-preserving mapping is that it allows for the familiar Euclidean geometry to be applied to the obtained vectors in $\mathbb{R}^{D-1}$, since the vector elements are no longer dependent on one another after the transformation.

As one can come up with more than one orthonormal basis for the open simplex, the *ilr* transformation is not unique. In this paper, we employed a specific one defined by (Egozcue et al., 2003). For any $\boldsymbol{x} = (x_1, \ldots, x_D) \in S^D$,

(10)

$$ilr(x) : = \log(x) \cdot U$$

where $U = [U_1, \ldots, U_{D-1}]$is the $D \times (D-1)$orthonormal basis, with $U_i \in \mathbb{R}^D$denoting its ith column:

(11)

$$U_{ji} = \begin{cases} \frac{1}{i}\sqrt{\frac{i}{1+i}}, & if \quad j \le i \\ -\sqrt{\frac{i}{1+i}}, & if \quad j = i+1 \\ 0, & otherwise \end{cases}$$

for $j \in \{1, \ldots, D\}$. Initially, and before applying the *ilr*-transformation on the relative isoform usage data, the statistical hypothesis test for differential isoform usage between African-American and European-American groups within each condition was set up as:

(12)

$$H_o : \mu_{\pi_1} = \mu_{\pi_2}$$
$$H_1 : \mu_{\pi_1} \neq \mu_{\pi_2}$$

where μ_{π_i} is the mean relative isoform usage for group i. To prepare the data for the *ilr* transformation and statistical analysis, two preliminary steps were undertaken as follows. To make sure that the statistical test yielded biologically meaningful results, if the average relative abundance of an isoform across samples was less than 0.05 in both African-American and European-American groups, that isoform was eliminated from statistical testing analysis. Any relative isoform usage value that remained after the isoform-elimination step and was estimated as zero was replaced by a small strictly positive value of 0.0005 to make sure that all samples belong to the open simplex S^D. Note that if, for a specific isoform and a specific sample, the relative abundance is below 0.05 and strictly greater than 0, and if that isoform is not removed in the isoform-elimination step, then the relative abundance value is retained. After performing the *ilr* transformation on each sample, a multivariate normal distribution on $\mathbb{R}^{D-1}$ was assumed for the *ilr*-transformed relative isoform usage data:

(13)

$$ilr(\pi_{i1}), \ldots, ilr(\pi_{in_i}) \sim \mathcal{N}_{D-1}\left(\mu_{ilr(\pi_i)}, \Sigma_i\right) \quad for \quad i = 1, 2,$$

where Σ_i is the covariance matrix for group i, with $\Sigma_1 \neq \Sigma_2$. Consequently, differential isoform usage boils down to testing the equality of means of two multivariate normal populations, with distinct covariance matrices. This is mathematically shown by:

(14)

$$H_o : \mu_{ilr(\pi_1)} = \mu_{ilr(\pi_2)}$$
$$H_1 : \mu_{ilr(\pi_1)} \neq \mu_{ilr(\pi_2)}$$

where $\mu_{ilr(\pi_i)}$ is the mean of *ilr*-transformed relative isoform usage vectors of group i. This problem is referred to as the multivariate Behrens-Fisher problem, and different approaches have been proposed to tackle the multi-dimensional case. In this paper, we adopted the method proposed by (Krishnamoorthy and Yu, 2004), which reduces to the well-known Welch's t test for one-dimensional data (or equivalently, when D = 2). This test, referred to as T_{KY} herein, cannot be employed when $D - 1 \geq \min\{n_1, n_2\}$ (a case that results in either of the estimated covariance matrices to be singular and non-invertible). This is not a concern in our analysis, since we have a large number of samples per group. The result of differential isoform usage test is reported in Table S2D, where estimation of isoform expression values was done using the RSEM software package. Out of the 10223 genes tested, 62, 39, and 48 genes showed statistically significant DIU between African-American and European-American populations, in the non-infected, *Listeria*-infected, and *Salmonella*-infected samples, respectively (FDR < 0.05).

To verify the robustness of our results, we re-conducted our DIU analysis with the approach adopted by (Lappalainen et al., 2013). Specifically, we used the Mann-Whitney U test to compare the distributions of relative abundances, for each isoform, between African-American and European-American populations so as to detect transcripts with significantly different ratios between populations. Afterward, the Benjamini–Hochberg FDR method was applied to adjust the p-values obtained from

these individual comparisons (*i.e.*, between-population comparisons per isoform). Subsequently, a gene was labeled DIU provided that at least one of its isoforms showed significant evidence of differential usage between African-Americans and European-Americans. In particular, using this approach, 93, 70, and 123 genes were detected with significant DIU (FDR <0.05) between African-American and European-American populations, in the non-infected, *Listeria*-infected, and *Salmonella*-infected samples, respectively. Figure S2B compares the number of DIU genes detected using the multivariate Welch's t test (our original approach) with that obtained by the rank sum test at different FDR thresholds.

Enrichment of GWAS-Associated Genes among pop-DE Genes

To identify enrichment of pop-DE genes among genes that were previously found to be associated with complex human disease and traits, we used data from the GWAS catalog (Hindorff et al., 2009). Since each GWAS has a different distribution of P-values and significance cutoffs, we chose to use a set of $-log_{10}(p)$ cutoffs in the range of 8-60 (plotted along the x axis in Figure 2B). For a given disease, we identified the overlap between the genes significantly associated with the disease at each cutoff and pop-DE genes, and calculated a fold enrichment (plotted along the y axis in Figure 2B), defined as the ratio of observed/expected overlap between the two gene sets. We used a Fisher Exact Test to calculate a P-value for each cutoff, and corrected these P-values for multiple tests using the FDR approach within each disease.

Genotype-Phenotype Association Analysis

eQTL, asQTL and reQTL mappings were performed against a set of 11,927 protein coding genes. We examined associations between SNP genotypes and the phenotype of interest using a linear regression model, in which phenotype was regressed

against genotype. In particular, expression levels were considered as the phenotype when searching for eQTL and asQTL, while to identify reQTL, fold changes in response to infection were treated as the quantitative trait to be mapped. In all cases, we assumed that alleles affected the phenotype in an additive manner. For the eQTL and asQTL analyses, we mapped *Salmonella*-infected, *Listeria*-infected, and non-infected macrophages, separately. All regressions were performed using the R package Matrix eQTL (Shabalin 2012). To avoid low power caused by rare variants, only SNPs with a minor allele frequency of 5% across all individuals were tested. Local associations (*i.e.*, putative *cis* QTL) were tested against all SNPs located within the gene body or 100Kb upstream and downstream of the transcript start site (TSS) and transcript end site (TES) of each tested gene. We recorded the minimum P-value (*i.e.*, the strongest association) observed for each gene, which we used as statistical evidence for the presence of at least one eQTL for that gene. *Trans*-eQTL were defined as SNPs located $>$ 500kb of the gene they regulate and could be on the same or different chromosomes. To estimate an FDR, we permuted the phenotypes (expression levels, fold changes or percent of isoform usage) ten times, re-performed the linear regressions, and recorded the minimum P-values for the gene for each permutation. These sets of minimum P-values were used as our empirical null distribution and FDRs were calculated using the method described in the section "Estimation of FDRs."

Consistent with previous reports (Barreiro et al., 2012), we found that we could increase the power to detect *cis*-eQTL by accounting for unmeasured-surrogate—confounders. To identify such confounders, we first performed principal component analysis (PCA) on a correlation matrix based on genes expressions, for non-infected, *Salmonella*- or *Listeria*-infected samples. Subsequently, we regressed out up to 15 principal components before performing the association analysis for each gene. This

specific number of PCs was chosen since it empirically led to the identification of the largest number of eQTL in each condition. The exact number of PCs regressed in each of the analyses can be found in the table below. Note that for the *trans* analysis we did not regress any PCs to avoid inadvertently removing the effect of true *trans* signals.

Principal Components Regressed				
Analysis	**Condition**	**Regressed PCs**	**No. Genes under 0.01 FDR**	**Estimated Fraction of True Negatives π_o**
Cis eQTL	non-infected	1 to 3	875	0.56
	Listeria	1 to 7	1,087	0.47
	Salmonella	1 to 5	983	0.47
Cis reQTL	fold change *Listeria*	1 to 10	244	0.69
	fold change *Salmonella*	1 to 7	503	0.66
Cis asQTL	non-infected	1 to 3	886	0.67
	Listeria	1 to 3	746	0.66
	Salmonella	1 to 3	615	0.65

Importantly, although the PC corrections clearly increase power to detect eQTL, they do not affect the underlying structure of the expression data. Indeed, over 80% of the eQTL observed before any PC correction are also observed after PC correction at the same FDR cutoff. A similar approach was used for asQTL and reQTL mapping, with the only difference being that for those analyses the PCA were performed in a matrix of isoform proportional abundance or fold-change responses, respectively.

Mapping was performed combining AA and EA samples to increase power. To avoid spurious outcomes resulting from population structure, the first five eigenvectors obtained from a PCA on the genotype data were included in the linear model as well. For each library, we also took into account the potential biases and significant technical confounders identified before the DE analyses; *i.e.*, bacteria counts used when infecting the macrophages (c), sequencing flowcell (x_1), mean GC content estimated per sample (x_2), proportion of uniquely mapped reads (x_3), and RNA concentration (x_4). The covariate subscripts or superscripts corresponding to the experimental condition, in which they were measured, are dropped in the following models for simplicity:

- eQTL models, non-infected condition:

(15)

$$Gene\ expression \sim Genotype + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5$$

- eQTL models, *Listeria* and *Salmonella* conditions:

(16)

$$Gene\ expression \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5$$

- reQTL models, response to *Listeria* and *Salmonella* infections:

(17)

$$Gene\ fold\ change \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5$$

- asQTL models, non-infected condition:

(18)

$$Transcript\ proportion \sim Genotype + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5$$

- asQTL models, *Listeria* and *Salmonella* conditions:

(19)

$$Transcript\ proportion \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5$$

Identification of Condition-Specific eQTL and asQTL

We classified condition-specific *cis*-QTL using a conservative criterion aimed at minimizing the risk that true eQTL in both resting and infected macrophages are only identified in one condition because of incomplete power. Specifically, we defined condition-specific QTL when we found strong evidence (FDR < 0.01) of a *cis*-QTL in one condition and no statistical evidence, using a relaxed FDR threshold (0.3), supporting a *cis*-QTL for the same gene in the other conditions.

Multiple Testing Correction

To estimate a FDR, we permuted the phenotypes ten times and used the distribution of the acquired minimum p value per gene to calculate the FDR associated with the p value obtained from the real data, as described above.

Allele-Specific Expression Detection

The sequenced samples were preprocessed using WASP (van de Geijn et al., 2015) program in order to correct for mapping biases toward the reference sequence. To this end, we removed all the monomorphic sites and hence only the positions showing polymorphism in at least one of the 171 samples were included in the analysis for correction. The resulting fastq files from WASP were mapped the same way the original

alignment files were obtained (*i.e.*, using the STAR pipeline). Allele counts per sample for positions that overlap with the Omni5Exome-4v1-1 genotyping array were obtained using SAMtools mpileup (Li et al., 2009) v0.1.19-44428cd, with minimum base quality of 13.

Genotype calls obtained from these steps were then used as the input files for ASE identification with the QuASAR software (Harvey et al., 2015), which can jointly genotype and detect allelic imbalances for each SNP. Starting with three samples from each individual, corresponding to the two bacterial infections and the non-infected control, QuASAR can simultaneously identify heterozygous SNPs and ASE by taking into account base-calling errors and over-dispersion in the ASE ratio. The prior genotype probabilities in QuASAR are obtained from the 1000 Genomes Project minor allele frequencies assuming Hardy–Weinberg equilibrium; however, as we had the genotype information available, we manually input the prior genotype probabilities. Specifically, the prior genotype probabilities in QuASAR are indicated in a matrix of three columns, where the columns denote homozygous reference, heterozygous, and homozygous alternate probabilities, respectively, and each row corresponds to an exonic location. As an illustration, to input our genotype information for a heterozygous exonic location, we set the corresponding row equal to $(\gamma, 1-2\gamma, \gamma)$where $\gamma = 0.001$accounts for genotyping error. This is done through changing the gmat argument of the *fitAseNullMulti()* function. Manually setting up the genotype probabilities, as mentioned above, assures that the prior genotype information will not change drastically as QuASAR iterates through the EM algorithm steps; moreover, it enables us to estimate the base-calling error rate. In the subsequent step of inferring ASE, we set the min.cov argument of the *aseInference()* function equal to 10 to only assess the sites represented in at least 10 reads across all the samples.

QuASAR outputs the allelic imbalance estimate for each exonic location as $\log(p/1-p)$, where p denotes the proportion of the number of reference reads over the total number of reads, with no allelic imbalance (*i.e.*, $p=0.5$) resulting in an effect size estimate of zero. After obtaining estimates of allelic imbalance and the corresponding standard errors from QuASAR for each (heterozygous) exonic SNP in each sample, we used the CRG Alignability tracks in the framework of the GEM (GEnome Multitool) project to only keep the exonic SNPs with a mapability score of S = 1; *i.e.*, with only one match in the genome. We then performed meta-analysis on this output to aggregate the effect sizes across samples for each exonic SNP. Specifically, we only retained the exonic SNPs with a frequency of at least three samples and adopted the inverse-variance weighting method for each of these exonic locations to combine the effect sizes across samples, where each effect size was weighted by its inverse variance. A Z-score is then calculated for each weighted mean effect size to allow for a Z-test of significant deviation from zero, to test the null hypothesis of no allelic imbalance at each exonic SNP.

ATAC-Seq Data Processing and Footprinting Analysis

ATAC-seq paired-end reads were mapped to the human reference genome (GRCh37/hg19) using Bowtie 2 (Langmead and Salzberg, 2012) with the following parameters: -N 1 -X 2000–no-mixed–no-discordant–no-unal. Only reads that had a paired and unique alignment were retained. PCR duplicates were removed using Picard's MarkDuplicates tool.

To detect TF binding footprints in the ATAC-seq data we used the program Centipede (Pique-Regi et al., 2011). We ran Centipede separately for each of the three conditions. We started by defining the set of transcription factors that were active (*i.e.*, had motif instances with footprints) before and after infection with *Listeria* and

Salmonella using a reduced set of high-confidence motif instances for each TF. Using these reduced set of motifs, we calculated a Z-score corresponding to the PWM effect in the prior probability in Centipede's logistic model. The Z-score corresponds to the parameter in:

(20)

$$log\left(\pi_l/1-\pi_l\right)=\alpha+\beta\cdot PMW_{score_l}$$

where $\pi_l$represents the prior probability of binding in Centipede model in motif location l. We considered a TF binding site as active if the estimate was supported at Bonferroni-corrected $P < 10^{-5}$. In total, 369, 420 and, 422 motifs were identified as active in non-infected, *Listeria*-infected and *Salmonella*-infected macrophages, respectively. We then scanned the entire genome for motif instances matching the original PWM for these active motifs, separately for each of the three conditions.

Footprints were grouped into clusters using sequence similarity. The positional overlap of predicted bound regions of active motifs was first determined using bedtools multiinter. Using the overlap scores, footprints were then divided into clusters using R function hclust with a distance cutoff of 0.9. Well-supported footprints (posterior Pr > 0.9) were used for the enrichment analysis.

Enrichment of TF Binding Sites among eQTL and reQTL

To estimate the enrichment level of particular transcription factor binding locations among eQTL and reQTL, we used the method described in (Wen et al., 2016) and implemented in TORUS (https://github.com/xqwen/dap/tree/master/torus_src).

This method uses a hierarchical model that aggregates eQTL signals across all genetic variants to model the characteristics shared among those most likely to be causal. This is an iterative process starting from eQTL summary statistics calculated

with matrix-eQTL using a comprehensive set of imputed genotypes. Using the deterministic approximation of posteriors (DAP) approach, we then learn a prior for each genetic variant using a logistic that can use different types annotations that are informative for determining SNPs that are more likely to disrupt transcription. In our case, we seek to determine if SNPs in binding sites for certain TFs are more likely to be eQTL or reQTL, and therefore determine the likely molecular mechanisms underlying our QTL signals.

The putative binding sites (*i.e.*, ATAC-seq footprints) were determined using ATAC-seq, as described in the section above. To analyze eQTL in non-infected, *Salmonella*-infected, and *Listeria*-infected macrophages, we used the footprints derived in each condition. To analyze reQTLs to *Salmonella* and *Listeria* infection, we used the footprints collected in *Salmonella*-infected and *Listeria*-infected macrophages, respectively. To avoid spurious enrichments resulting from the fact that several TF binding sites are non-randomly distributed with respect to the TSS, we used a background model that captures the effects of distance to the TSS.

Footprints corresponding to different types of transcription factors were analyzed separately. For each annotation, we also calculated a 95% confidence interval. The "enrichment" parameter represents the log-odds that genetic variants in a particular annotation are more likely to harbor causal SNPs for an eQTL compared to a baseline background model that takes into account distance to TSS. In our application, higher enrichment for genetic variants in a specific transcription factor binding sites provide evidence for a likely causal mechanism underlying many of the measured eQTLs and reQTLs.

Genetic Control of Ancestry Effects on Gene Expression

To determine the extent to which a set of genetic variants control the signal associated with ethnic admixture in gene expression variation, a comparison should be made between the models M_1, M_2, and M_3-previously introduced to produce estimates of pop-DE, condition-DE and pop-DR effects, respectively, and their corresponding extensions, which take the effect of SNPs into consideration.

Control of Ancestry-Related Differential Expression (pop-DE genes) by Individual SNPs

Extending M_1 to account for *cis*-genetic variation effects as mentioned above, we obtain the following model, $M_1^{G_{cis}}$, that has the effects of the best-associated *cis*-SNPs of each gene (regardless of significance level) in each condition as additional covariates:

(21)

$$M_1^{G_{cis}}: E(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \beta_G^{NI}(i) \cdot G^{NI}(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if } Condition = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \beta_G^L(i) \cdot G^L(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & \text{if } Condition = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \beta_G^S(i) \cdot G^S(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & \text{if } Condition = S \end{cases}$$

In this model, $G^{NI}(i,j)$, $G^L(i,j)$, and $G^S(i,j)$ represent the genotypes of the *cis*-SNPs with the highest association to the expression of gene i in individual j and take values in the set $\{0,1,2\}$ accounting for the copies of the less abundant allele. It is worth mentioning that $G^{NI}(i,j)$, $G^L(i,j)$, and $G^S(i,j)$ generally differ across genes and conditions. Furthermore, these SNPs are not necessarily the true eQTLs, as being the top association for a given gene-condition pair does not automatically imply that the SNP satisfies the FDR-threshold criteria of statistical significance to be an eQTL. The reason why we do not require SNPs to pass any FDR threshold for significance is that smaller

effect eQTL may still contribute to population differences, with the most strongly associated variant still being the best candidate eQTL. Moreover, we note that most of these "best variants" actually do have reasonably strong evidence for eQTL association (e.g., 40% of best variants are identified at an FDR < 0.1, and 61% are identified at an FDR < 0.2), even if they do not pass our more stringent primary threshold (FDR $<$ 0.01).

In order to compare the role of ethnic admixture in shaping gene expression levels before and after regulatory variants are introduced to the model, for each gene, let us define its reduced expression vector associated to a given condition C as the set of expression values within such condition from which the effects of all model covariates except ethnic admixture have been removed:

(22)

$$e^C_{M_1}(j) = \beta^C_{Af} \cdot Af(j) + \varepsilon^C(j)$$

where β^C_{Af} coefficients and $\varepsilon^C(j)$ come from M_1 fit. If we denote the gene mean of the reduced expression values across individuals for condition C as $\langle e \rangle^C_{M_1}$, then the total sum of square deviations from $\langle e \rangle^C_{M_1}$ for that gene is:

$$SS^C_{tot(M_1)} = \sum_j (e^C_{M_1}(j) - \langle e \rangle^C_{M_1})^2 \quad ;$$

and can be expressed as the sum of two components: a between-population component, explained by the admixture effect in the regression model:

$$SS^C_{reg(M_1)} = \sum_j (\beta^C_{Af} \cdot Af(j) - \langle e \rangle^C_{M_1})^2$$

and a within-population, or unexplained component corresponding to the residuals:

$$SS^C_{res(M_1)} = \sum_j \varepsilon^C(j)^2 \quad .$$

From these magnitudes, the P_{ST} statistics is build for each gene as follows:

(23)

$$Pst_{M_1}(C) = \frac{SS^{C}_{reg(M_1)}}{SS^{C}_{reg(M_1)} + SS^{C}_{res(M_1)}}$$

which measures the fraction of variance of the reduced expression that is explained by ethnicity. Defined this way, the P_{ST} indexes constitutes a phenotypic analog of the population genetics parameter F_{ST} (Leinonen et al., 2013), when the population's structure is not defined in a binary fashion but according to a continuous trait as the ethnic admixture. From a merely formal point of view, the P_{ST} statistics is just the coefficient of determination R^2 associated to the regression model that defines the reduced expression vectors.

Likewise, we can obtain the reduced expression vectors from $M_1^{G_{cis}}$, which we denote by $e^{C}_{M_1^{G_{cis}}}$; and from these, the respective $Pst_{M_1^{G_{cis}}}(C)$. In order to assess the contribution of genetic information to reduce the fraction of variance that ethnicity explains with respect to the residuals, we can compare $Pst_{M_1}(C)$ against $Pst_{M_1^{G_{cis}}}(C)$ through the following relative variation index:

(24)

$$\Delta Pst^{cis}_{M_1}(C) = \frac{Pst_{M_1}(C) - Pst_{M_1^{G_{cis}}}(C)}{Pst_{M_1}(C)}$$

Analogously, we also build a version of M_1 including, instead of *cis*-SNPs, the best *trans*-SNP of each gene as an additional covariate, denoted as $M_1^{G_{trans}}$; from whose comparison to M_1 we ultimately derive $\Delta Pst^{trans}_{M_1}(C)$ in the same way.

Control of Ancestry-Associated Differential Response to Infection by Individual SNPs

Similar to what proceeded, the following extension of $M_3$is set up to incorporate the genetic variants that affect the gene-expression responses to infection:

(25)

$$M_3^{G_{cis}}: E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & if \quad Condition = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \beta_G^L(i) \cdot \widehat{G}^L(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & if \quad Condition = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \beta_G^S(i) \cdot \widehat{G}^S(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & if \quad Condition = S \end{cases}$$

For the gene-*i* individual-*j* pair, $\widehat{G}^L(i,j)$ and $\widehat{G}^L(i,j)$ represent the genotypes of the *cis*-SNPs with the highest association to the gene-expression fold changes after infection with *Listeria* and *Salmonella*, respectively. When statistically significant, these top associations constitute response eQTLs, as their effects on gene expression differ across conditions.

In a similar fashion to the comparative analysis of M_1 and $M_1^{G_{cis}}$, so as to compare $M_3$and $M_3^{G_{cis}}$, we start by defining the reduced response vectors as the expected fold change after infection with each pathogen, from which all the covariates but ethnicity have been removed using model M_3 estimates:

(26)

$$fc_{M_3}^L(j) = \beta_{Af}^L \cdot Af(j) + \varepsilon_{fc}^L(j)$$
$$fc_{M_3}^S(j) = \beta_{Af}^S \cdot Af(j) + \varepsilon_{fc}^S(j)$$

where the fold change residuals are derived from the differences between infected and non-infected samples residuals for each individual j from which a valid sample of each condition was collected: $\varepsilon_{fc}^L(j) = (\varepsilon^L - \varepsilon^{NI})(j)$ and $\varepsilon_{fc}^S(j) = (\varepsilon^S - \varepsilon^{NI})(j)$.

From this point, the computation of P_{ST} statistics provides us with a measure of the proportion of variance that is explained by the interaction terms β_L^{Af} and β_S^{Af} at the reduced fold changes obtained from M_3 estimates. More precisely, for the infected condition C (*Listeria* or *Salmonella*), the total sum of square deviations from the average $\langle fc \rangle_{M_3}^C$ reads.

As $SS_{tot(M_3)}^C = \sum_j (fc_{M_3}^C(j) - \langle fc \rangle_{M_3}^C)^2$; while the regression and residuals components are $SS_{reg(M_3)}^C = \sum_j (\beta_{Af}^C \cdot Af(j) - \langle fc \rangle_{M_3}^C)^2$ and $SS_{res(M_3)}^C = \sum_j \varepsilon^C(j)^2$, respectively. Therefore, the corresponding P_{ST} index reads as follows:

(27)

$$Pst_{M_3}(C) = \frac{SS_{reg(M_3)}^C}{SS_{reg(M_3)}^C + SS_{res(M_3)}^C}$$

Moreover, the *Pst* coefficients obtained from model M*3*, can be also obtained from $M_3^{G_{cis}}$ within each infected condition to get the corresponding $Pst_{M_3^{G_{cis}}}(C)$ statistics. Finally, these values are compared to $Pst_{M_3}(C)$ through the following relative variation index:

(28)

$$\Delta Pst_{M_3}^{cis}(C) = \frac{Pst_{M_3}(C) - Pst_{M_3^{G_{cis}}}(C)}{Pst_{M_3}(C)}$$

Finally, the equivalent analysis is performed using *trans*-SNPs: from $M_3^{G_{trans}}$, we calculate the corresponding P_{ST} statistics $Pst_{M_3^{G_{trans}}}(C)$, and compare those against $Pst_{M_3}(C)$ through the corresponding relative variation indexes $\Delta Pst_{M_3^{G_{trans}}}(C)$.

Null Model

As a means of validating the significance of the comparisons conducted before and after introducing genetic information in our linear models, we consider a null model in which the top associated SNPs in $M_1^{G_{cis}}$ and $M_3^{G_{cis}}$ (or $M_1^{G_{trans}}$ and $M_3^{G_{trans}}$) are substituted by (*i*) randomly selected SNPs matched for the allele frequency of the lead *cis*-SNP, or *(ii)* lead *cis*-SNPs identified after permuting the genotype data.

Gene Ontology Enrichment Analysis

Using ClueGO cytoscape module (Bindea et al., 2009), we interrogated for enrichments of ontology terms related to Biological processes in different target sets of DE genes with respect to a background composed by all genes analyzed. For this particular, for pop-DR, genes within the target sets were required to present an absolute fold change larger than 0.2 -positive or negative, depending on the direction of the effect- and a false discovery rate lower than 0.2 for the DE effect considered. For infection DE enrichments, characterized by much larger size effects, we conducted one analysis per pathogen, regardless direction of effects, using a more stringent cutoff (absolute fold change larger than 0.5 and FDR $<$ 0.01). Regarding the program's parameters defining the test to consider, we configured them as follow:

- Statistical Test Used = Enrichment (Right-sided hypergeometric test).
- Fusion of related Parent-child terms activated.
- Correction Method Used = Benjamini-Hochberg.
- Min GO Level = 3.
- Max GO Level = 8.
- Minimum Number of Genes = 20.
- Min Percentage = 5.0.

For the graphical representation of the enrichment analysis among pop-DR genes, ClueGO clustering functionality was used (kappa threshold score for considering

or rejecting term-to-term links set to 0.4 for *Salmonella* and 0.5 for *Listeria*, fraction for groups merging = "25%" in both cases). Only clusters with at least three GO terms were plotted. The detailed results of this analysis are presented in Table S3, where terms enriched at FDR < 0.1 are registered.

Signatures of Selection

F_{ST} values between the YRI and CEU individuals were obtained using data coming from 1000 Genomes data (phase 3 20130502). The phased data were downloaded from Impute reference data panel (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html) and were filtered for biallelic SNPs found in either the CEU (n = 99) or the YRI (n = 108) samples. The phased genotypes were obtained using ShapeIT v2.r790 (Delaneau et al., 2013) with the '-output-vcf' option. The vcf files (one for each chromosome) were then converted to the PLINK format using vcftools v0.1.12b (https://vcftools.github.io/man_0112b.html) (Danecek et al., 2011). PLINK files were afterward converted to Arlequin format using the PGDSpider program (Lischer and Excoffier, 2012). Arlequin version 3.5.1.3 (http://cmpg.unibe.ch/software/arlequin35/) (Excoffier and Lischer, 2010) was then used to calculate F_{st} estimates derived from ANOVA. Integrated Haplotype Scores (iHS) (Voight et al., 2006) and cross-population Extended Haplotype Homozygosity (XP-EHH) (Sabeti et al., 2007) scores were calculated with the program selscan v1.1.0b (Szpiech and Hernandez, 2014) with default parameters. We defined high iHS values as those above the 99th percentile of genome-wide distribution in the CEU ($|\text{iHS}| > 2.70$) and the YRI ($|\text{iHS}| > 2.68$) populations. For XP-EHH, we used YRI as the reference set of haplotypes. Therefore, negative XP-EHH values correspond to longer haplotypes in the YRI population, and positive XP-EHH values correspond to longer haplotypes in the CEU population.

Coalescence Neutral Simulations for Evaluating Putatively Selected eQTL Sites

We identified 258 genes carrying a signature of recent positive selection (|iHS| > 99th percentile of the genome-wide distribution) in either CEU or YRI samples. In order to provide additional support for adaptive evolution for each of these genes, we performed 500 replicates of neutral simulations matched to the known demographic histories of CEU and YRI populations (Gutenkunst et al., 2009), the observed allele frequency of the putatively selected variant (if several were present because of strong linkage disequilibrium, we chose the one with the highest iHS value), and the local recombination rate around that candidate eQTL. Simulations were performed with the program mssel, a modified version of ms (Hudson, 2002) that simulates the coalescent process conditional on a frequency trajectory.

For each site, and separately for each population, we simulated 500 allele frequency trajectories. These were simulated backward in time with the appropriate demographic history (Gutenkunst et al., 2009) from a fixed present-day allele frequency, which we take to be the observed allele frequency in the population of interest.

To determine the local recombination rate surrounding the putatively selected eQTL site, we took a 1Mbp window (500kb on each side) around the site and calculate the number of centimorgans based on the genetic map. For a region of length d centimorgans, we use the Haldane's Map Function to estimate the probability of recombination, $r = (1/2)(1 - e^{-2d/100})$. We then compute the population scaled recombination parameter as $\rho = 4N_e r$, where $N_e = 10000$ is taken to be the ancestral effective population size.

The population scaled mutation rate in mssel/ms is parameterized as $\theta = 4N_e L\mu$, where L = 1,000,000 is the length in bp of the region, $N_e = 10000$ is the effective population size, and $\mu = 10^{-9}$ is the mutation rate per site per generation.

Thirteen of our putatively selected variants were within 500kb of the edge of our genetic map, and hence, we could not estimate recombination rates, and 2 had such high recombination rates that coalescent simulations did not finish. These were excluded from the analyses.

Determining Significance of iHS Scores

We calculate iHS scores for all neutral simulations using selscan v1.1.0b, as we did for the real data. For each eQTL site in the real data, we took the unstandardized iHS score that was observed for a given population and normalize it (subtracting the mean and dividing by the standard deviation) with the iHS score of the frequency matched neutral simulation, giving 500+1 total scores in the normalization. As normalized iHS scores have a standard normal distribution (Voight et al., 2006), the scores can be treated as Z-scores. From these Z-scores, we calculated a p value for the observed scores based on the standard normal distribution.

Identification of Neanderthal-like Sites

Bi-allelic SNPs across five European population samples (CEU, FIN, GBR, IBS, TSI), three African population samples with low levels of Eurasian ancestry (ESN, MSL, YRI), and ancestral allele were extracted from the phase 3 release of the 1,000 Genomes Project. SNPs with alleles segregating in any of the three African samples were removed from the analysis.

Genotypes at the remaining SNPs were extracted from the high-coverage Altai Neanderthal genome after applying the minimum set of quality filters (map35_50 downloaded from https://bioinf.eva.mpg.de/altai_minimal_filters/). To summarize, Neanderthal-like sites were called as all bi-allelic SNPs for which the Altai genome carries a derived allele, the derived allele is segregating in the European sample, and the African samples are fixed for the ancestral allele.

Annexe 2 : contenu complémentaire du chapitre 3

Figures

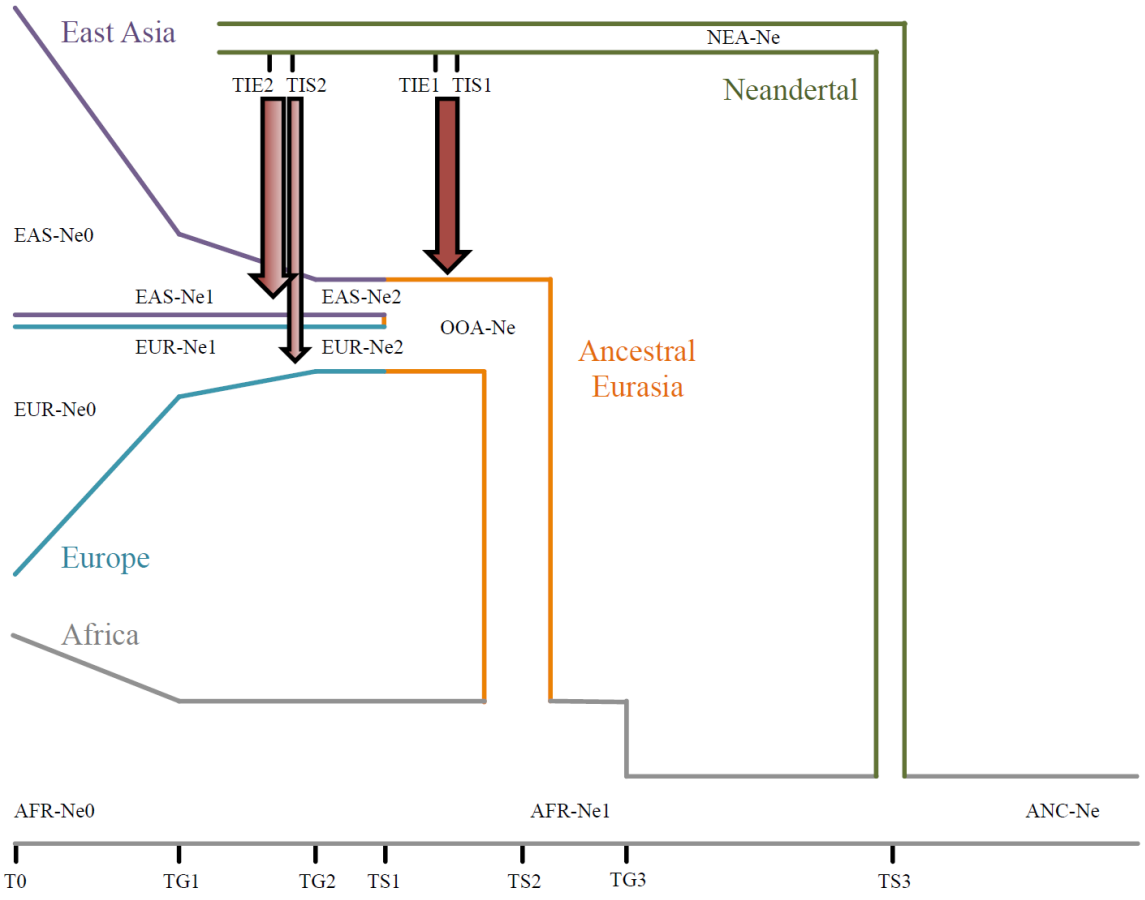

FIGURE S1: SCHEMATIC REPRESENTATION OF DEMOGRAPHIC MODEL USED IN SIMULATIONS.

Red arrows indicate points of introgression. The full range of demographic parameters used for the simulation can be found in Table S1.

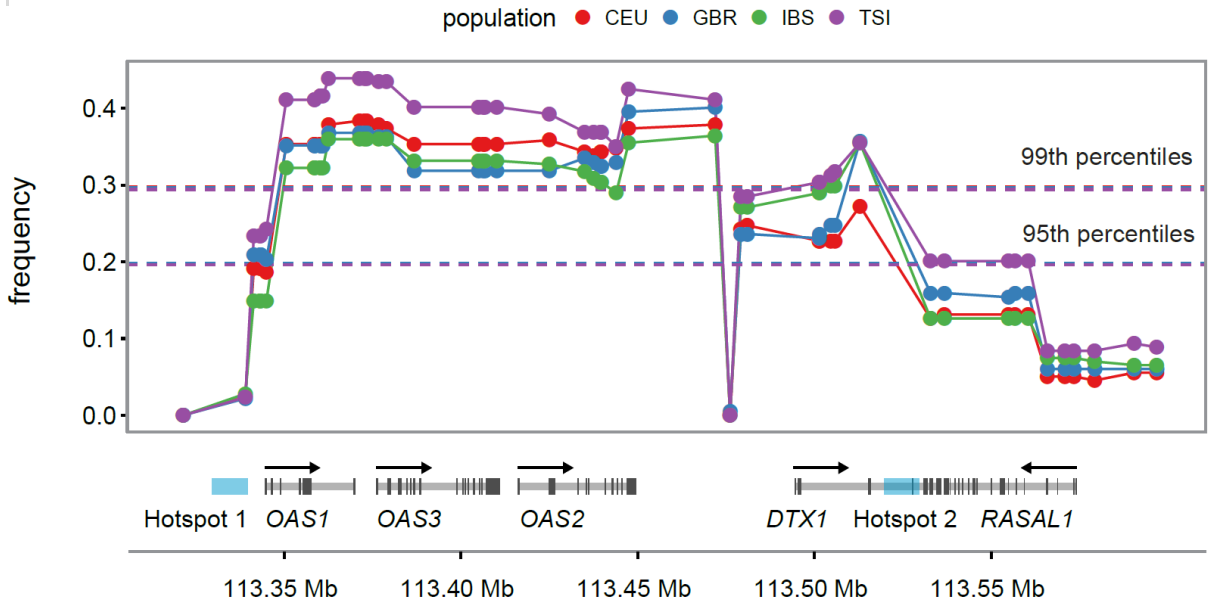

FIGURE S2: OAS-INTROGRESSED HAPLOTYPES ARE FOUND AT HIGHER FREQUENCIES IN EUROPEAN POPULATIONS THAN EXPECTED UNDER NEUTRALITY WHEN USING A TWO-PULSE INTROGRESSION MODEL

Comparison of frequency (y- axis) of NLS in the OAS locus in the CEU, GBR, IBS, and TSI European population samples with respect to neutral expectations (dashed lines) based on coalescent simulations considering two-pulse introgression in both Europeans and Asians, or a second pulse of introgression only in Europeans. Dashed lines overlap almost completely, reflecting little variation between a two-pulse introgression in both European and Asians, or a second pulse only in Europeans.

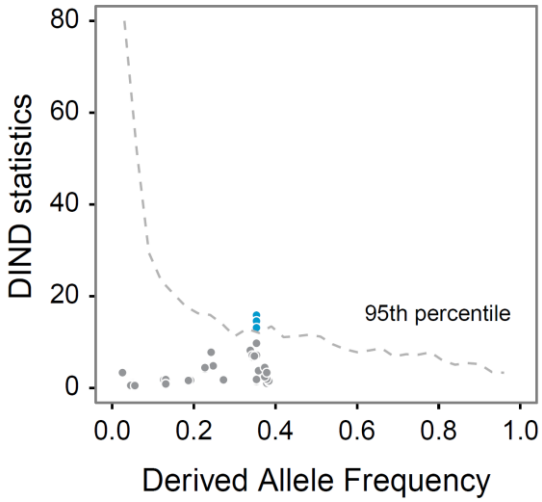

FIGURE S3: THE DIND STATISTIC INDICATES REDUCED DIVERSITY SURROUNDING NLS.

DIND (y-axis) is plotted for SNPs in OAS locus by derived (Neandertal) allele frequency. Dashed line indicates genome-wide 95th percentile by frequency.

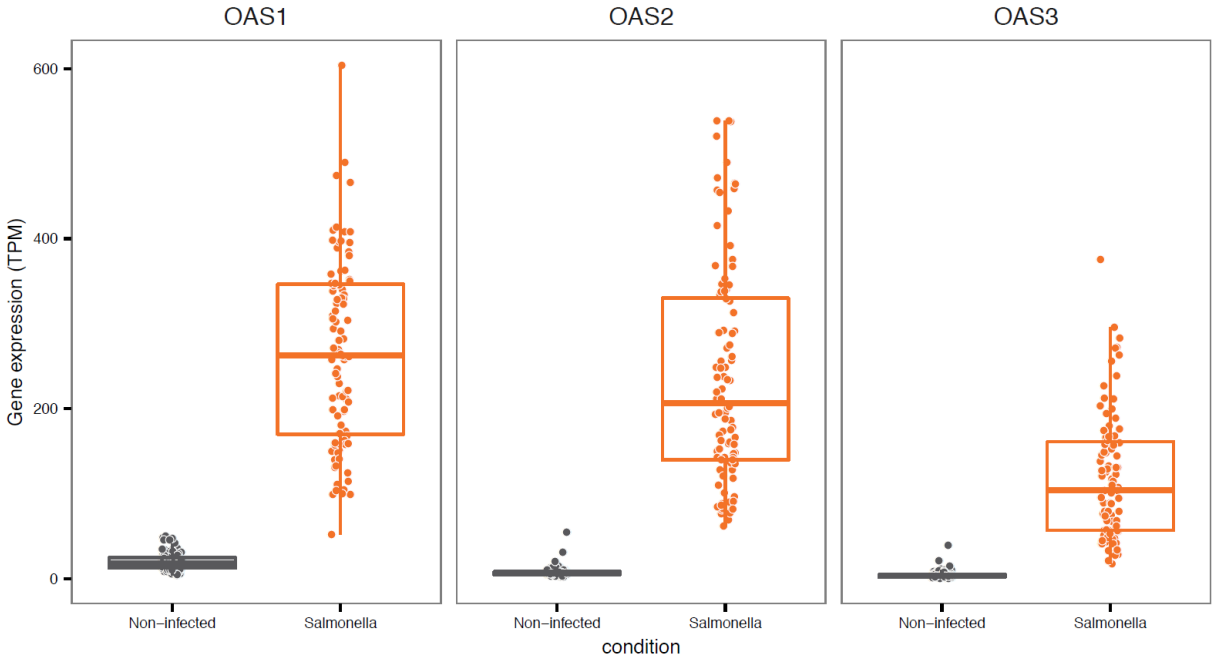

FIGURE S4: EXPRESSION LEVELS OF OAS GENES IN PRIMARY MACROPHAGES (EUROPEAN) BEFORE AND AFTER INFECTION WITH SALMONELLA.

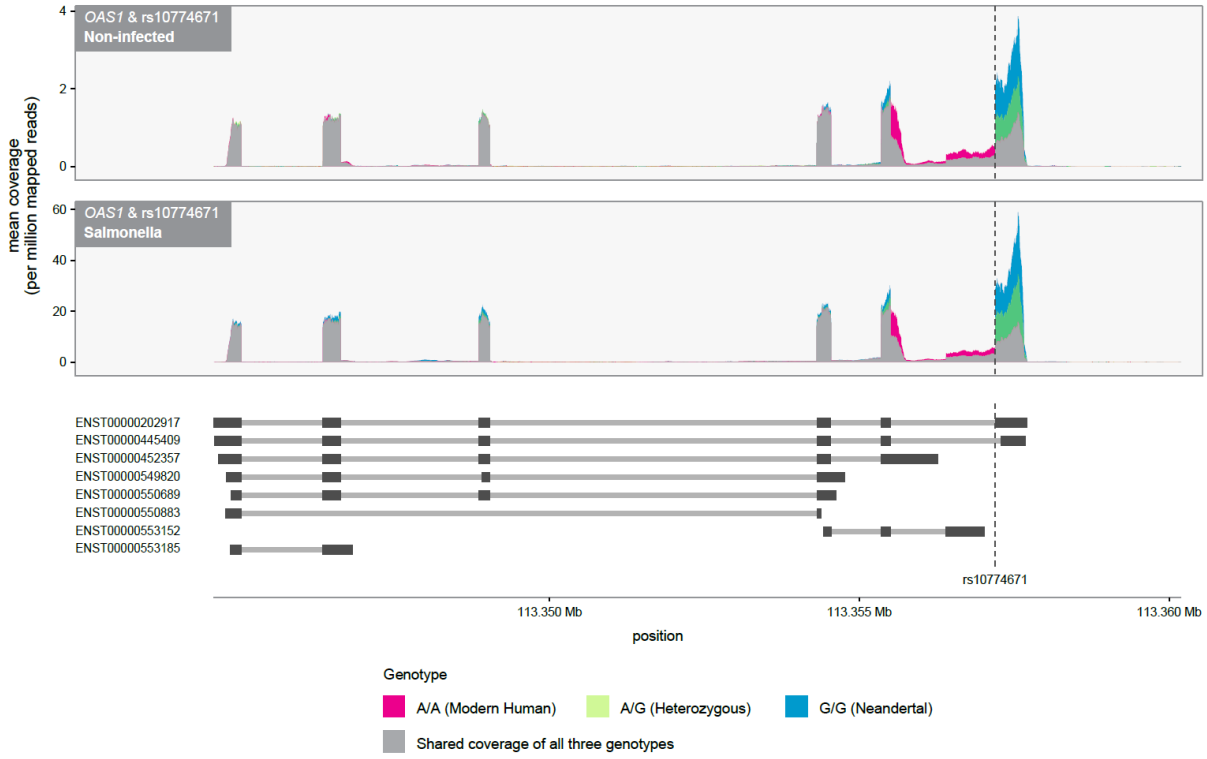

FIGURE S5: THE SPLICING VARIANT RS10774671 IS A STRONG ASQTL FOR OAS1

Plotted is the normalized average coverage at which each base was sequenced along the genomic regions encoding the gene OAS1. Individuals were stratified according to their genotype at rs10774671. Below the figure are gene models from the Ensembl database. Individuals carrying the G allele at rs10774671 (i.e. the Neandertal allele) primarily express the transcript ENST00000202917 (referred to as p46 in the text) whereas individuals carrying the A derived allele lose the splice site, which leads to the usage of a distinct isoform.

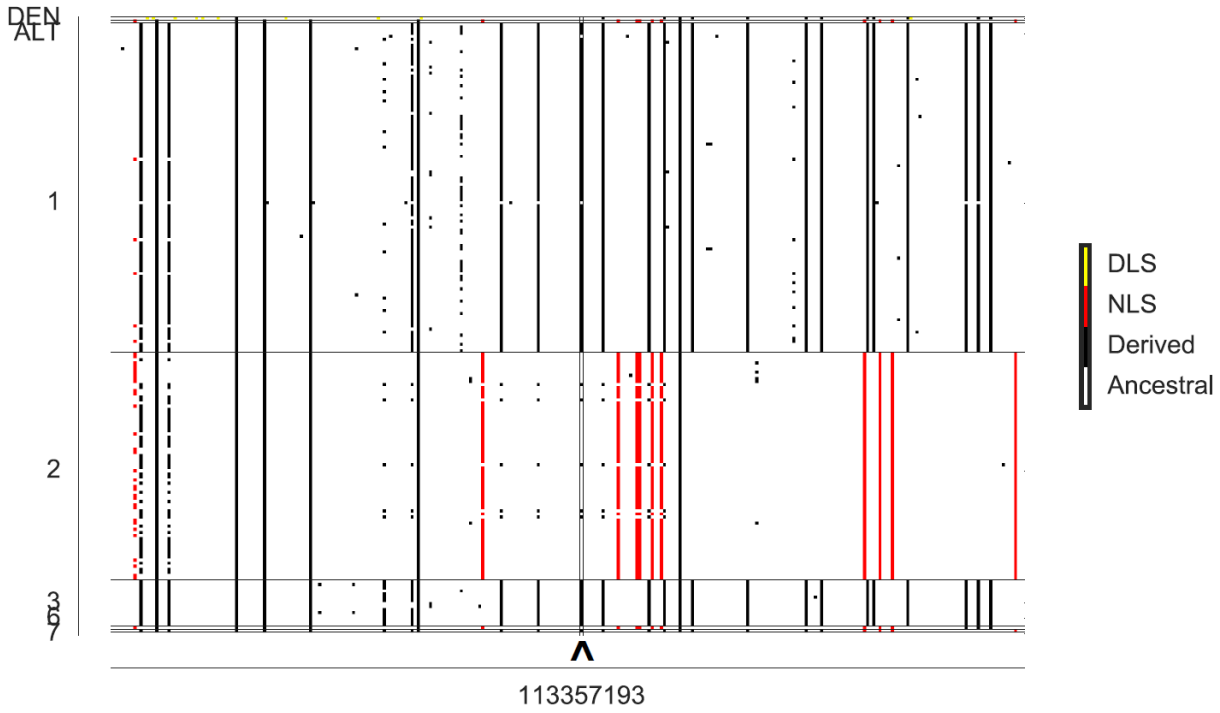

FIGURE S6: HAPLOGRAM OF REGION SURROUNDING OAS1 SPLICE SITE (RS10774671). 198 PHASED HAPLOTYPES FROM THE CEU POPULATION SAMPLE ARE ILLUSTRATED ALONG WITH THE ALTAI AND DENISOVAN HAPLOTYPES

Ancestral alleles are colored white, derived alleles are (1) yellow if derived in Denisovan and absent from Africans (YRI) (DLS), (2) red if derived in Neandertals and absent from Africans (YRI) (NLS), (3) black otherwise. With the exception of two rare haplotypes in cluster 1, all occurrences of the ancestral variant at rs10774671 are surrounded by derived Neandertal alleles, indicating that these were introduced in a Neandertal haplotype. Only sites where the Neandertal and Denisovan genomes are homozygous were used in this haplogram. Cluster labels (y-axis) correspond to clusters in Figure 1.

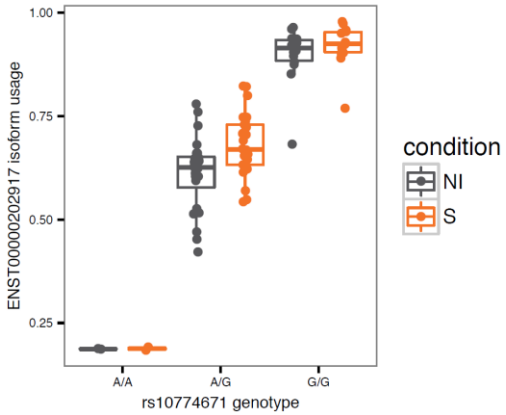

FIGURE S7: ISOFORM USAGE FOR TRANSCRIPT ENST00000202917 STRATIFIED BY GENOTYPE AT RS10774671 IN CELLS DERIVED FROM INDIVIDUALS WITH AFRICAN ANCESTRY

The derived (Neandertal) allele leads to increased expression of ENST00000202917 in both non-infected and Salmonella infected conditions.

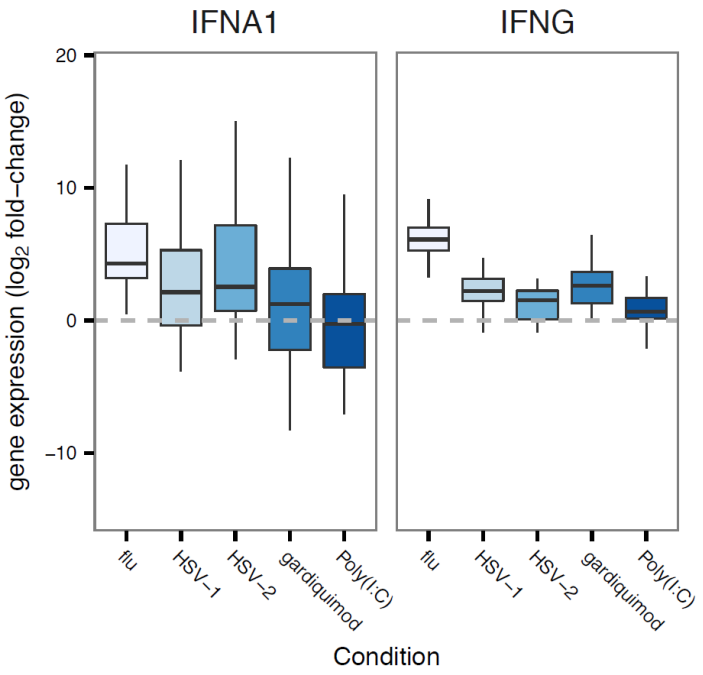

FIGURE S8: LOG 2 FOLD INDUCTION (Y-AXIS) OF IFNA1 (TYPE 1 INTERFERON) AND IFNG (TYPE II INTERFERON) IN PBMCS UPON STIMULATION WITH SEVERAL VIRAL AGENTS

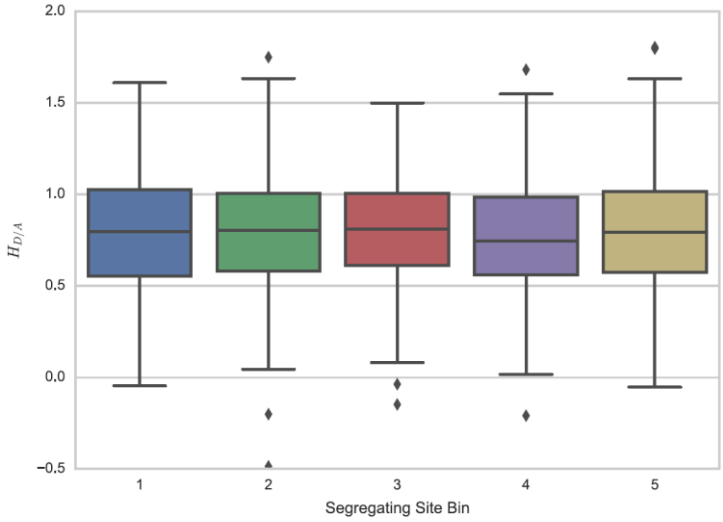

FIGURE S9: $H_{D/A}$ PLOTTED AGAINST QUINTILES OF SEGREGATING SITES IN 1,000 SIMULATIONS. THE $H_{D/A}$ STATISTIC DOES VARY WITH NUMBER OF SEGREGATING SITES ACROSS SIMULATED DATA.

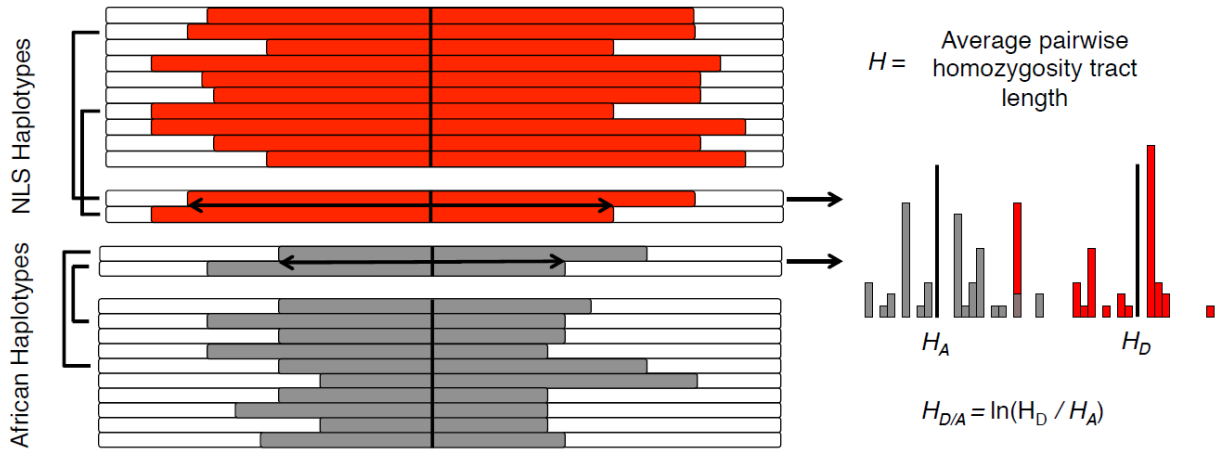

FIGURE S10: SCHEMATIC ILLUSTRATION OF THE $H_{D/A}$ STATISTIC

At each NLS, haplotypes are divided into two groups, based on whether they carry the Neandertal (derived) or African (ancestral) state. Within each haplotype subset, haplotype homozygosity is calculated across all pairs of haplotypes and then averaged (H), resulting in two values; H_A and H_D (A-ancestral, D-derived). Finally, $H_{D/A}$ is calculated as the natural log of the ratio of derived to ancestral H values. Excessively high values of this statistic reflect particularly long haplotypes carrying Neandertal-derived alleles.

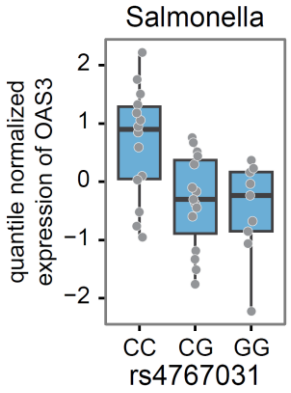

FIGURE S11: BOXPLOT FOR THE ASSOCIATION BETWEEN GENOTYPES AT THE NLS RS4767031 (X-AXIS) AND THE EXPRESSION LEVELS OF OAS3 IN SALMONELLA-INFECTED MACROPHAGES USING REAL-TIME PCR.

Méthodes

Genome alignments and identification of Neandertal-like sites

Human/chimpanzee ancestral states were computed by parsimony using alignments from the UCSC Genome Browser for the human reference (hg19) and three

outgroups chimpanzee (panTro2), orangutan (ponAbe2), and rhesus macaque (rheMac2) (Hudson 2002). Ancestral state was assumed to be the chimpanzee allele (if available) if its state was confirmed by matching either orangutan or macaque. All sites with no inferred ancestral state were removed from our analysis.

We filtered the Altai Neandertal genome (Prüfer et al. 2014) and Denisovan genome (M. Meyer et al. 2012) using the map35_50 set of minimum filters provided at (https://bioinf.eva.mpg.de/altai_minimal_filters/). For the frequency analysis, we combined these filtered datasets with 10 samples from outside of Africa (5-Eur European: CEU, FIN, GBR, IBS, TSI; 5-East Asian: CDX, CHB, CHS, JPT, KHB; see Additional file 2: Table S7 for population codes) and one sub-Saharan African sample YRI (Yoruba in Ibadan, Nigeria) from the 1000 Genomes Project Phase 3 (Auton et al. 2015), which we downloaded from :

https://mathgen.stats.ox.ac.uk/impute/1000GP%20Phase%203%20haplotypes%206%20October%202014.html.

We first extracted all bi-allelic variants in our human, Neandertal, and chimpanzee alignments. We considered as NLS only those variants where the African sample (YRI) had a derived allele frequency of zero and both the non-African sample and Neandertal carried the derived allele. We restricted the follow-up haplotype-based analysis to only the CEU sample. In that analysis, we required the derived allele to be present in CEU in at least two copies in order to calculate our haplotype-based test statistic ($H_{D/A}$).

Haplotype clustering in the 1000 Genomes Project data

To assess the relationships and frequencies of Neandertal-like haplotypes in the OAS region, we identified haplotype clusters based on sequence similarity. First, we

filtered all 5008 haplotypes in the 1000 Genomes Project phase 3 dataset to include only SNPs within the bounds of the three OAS genes (hg19 chr12: 113344739–113449528) and at which the Altai Neandertal (Prüfer et al. 2014) and Denisovan (M. Meyer et al. 2012) genomes carry homozygous genotypes. Next, we clustered the human haplotypes such that all haplotypes in a cluster had no more than 60 pairwise differences, which resulted in 10 clusters. We inferred a consensus haplotype for each cluster based on the majority allele at each position in the cluster. Finally, we produced a neighbor-joining tree of the human cluster consensus sequences and the Altai, Denisovan, and ancestral sequences using the R package "ape." We estimated confidence values for nodes with 1000 bootstraps.

We then calculated haplotype frequencies for each of the 1000 Genomes Project population samples based on the assignment of each haplotype to the clusters described above.

Demographic model and neutral coalescent simulations

We performed coalescent simulations of the demographic history of the European, East Asian, African, and Neandertal populations applied by Vernot and Akey (Vernot et Akey 2015) based on previously inferred demographic models (Gravel et al. 2011; Tennessen et al. 2012). In our first set of simulations, which were performed with ms (G. K. Chen, Marjoram, et Wall 2009), we extracted only allele frequencies at NLS. We gathered a total of 1 million simulation results. In each simulation run, a result was collected if a NLS was present in either the European or the East Asian population sample (or both). Therefore, null distributions specific to Europeans or East Asians included less than 1 million results (but typically $> 800{,}000$). Haplotype simulations were performed with macs (Schlamp et al. 2016) in order to explicitly

simulate the genetic map (downloaded with the 1000 Genomes samples at the link above) of the 600 kb region centered on OAS (chr12:113100000–113700000). The general shape of the demographic model is illustrated in Additional file 1: Figure S1. Our baseline simulations were performed with the parameters specified in Additional file 2: Table S1, assuming 25 years per generation and a mutation rate of 2.5×10^{-8} per bp per generation. Additional file 2: Table S1 also provides all parameters used in one-introgression and two-introgression pulse simulations. Additionally, we examined the one-pulse model with a slower mutation rate (1.25×10^{-8}) and a one-pulse model with a uniform recombination rate (1.3×10^{-8}; a value conservatively lower than the average for the OAS region, 1.7×10^{-8}). Sample ms and macs commands are given at the end of this section.

For haplotype-based simulations, data were thinned in a manner similar to Sankararaman et al. (Sankararaman et al. 2014) to account for imperfect SNP ascertainment in the 1000 Genomes dataset, such that SNPs with minor allele count of 1, 2, 3, 4, 5, 6, 7, 8, 9, and $>=$10 were accepted with probabilities 0.25, 0.5, 0.75, 0.8, 0.9, 0.95, 0.96, 0.97, 0.98, and 0.99, respectively. Additionally, we only kept SNPs that were polymorphic in the simulated CEU sample. Finally, we performed an additional thinning of SNPs with uniform probability of 0.05 of removal to account for slightly elevated SNP density in the simulated data. The resulting simulated datasets had an average SNP density of 3.8 SNPs per kb compared to 2.5 in the real data. This is a slightly larger than ideal difference in SNP density, but we note that neither derived allele frequency nor our primary haplotype-based test statistic (described below) should be particularly sensitive to SNP density. In fact, Additional file 1: Figure S9 illustrates that our statistic is conservative with respect to SNP density. Additionally, our results hold in a simulation with half the mutation rate of our baseline simulation. This

simulation resulted in an average SNP density of 1.9 SNPs per kb. So our observation of long haplotypes surrounding NLS is robust to SNP density across our simulations.

Sample ms command:

```
ms 752 1 -s 1 -I 4 250 250 250 2 0 -n 1 58.0027359781 -n 2 70.0410396717
-n 3 187.549931601 -n 4 0.205198358413 -eg 0 1 482.67144247 -eg 0 2
505.592963281 -eg 0 3 720.224280773 -em 0 1 2 0.358619661426 -em 0 2 1
0.358619661426 -em 0 1 3 0.111889334365 -em 0 3 1 0.111889334365 -em 0 2
3 0.446122858814 -em 0 3 2 0.446122858814 -en 0.00699726402189 1
1.98002735978 -eg 0.00699726402189 2 0.0 -eg 0.00699726402189 3
17.5076517287 -en 0.03146374829 2 2.03666547538 -en 0.03146374829 3
0.700185007205 -ej 0.0641347992705 3 2 -em 0.0641347992705 1 2 0.00015 -
em 0.0641347992705 2 1 0.00015 -em 0.0839811779742 2 4 0.00075 -em
0.0846651725023 2 4 0 -ej 0.0957592339261 2 1 -en 0.202462380301 1 1.0 -
ej 0.957592339261 4 1
```

Sample macs command:

```
macs 416 600000 -t 0.000731 -R oas_recrates.txt -I 4 216 198 0 2 0 -n 1
58.0027359781 -n 2 70.0410396717 -n 3 187.549931601 -n 4 0.205198358413 -
eg 0 1 482.67144247 -eg 1e-12 2 460.409556336 -eg 2e-12 3 720.224280773 -
em 3e-12 1 2 0.4441436762 -em 4e-12 2 1 0.4441436762 -em 5e-12 1 3
0.138572826974 -em 6e-12 3 1 0.138572826974 -em 7e-12 2 3 0.552514733193
-em 8e-12 3 2 0.552514733193 -en 0.00699726402189 1 1.98002735978 -eg
0.00699727402189 2 0.0 -eg 0.00699728402189 3 18.896348561 -en
0.03146374829 2 2.79399962717 -en 0.03146375829 3 0.670250141711 -ej
0.051785044418 3 2 -em 0.051785054418 1 2 0.00015 -em 0.051785064418 2 1
0.00015 -em 0.0555806720117 2 4 0.00075 -em 0.0562646665398 2 4 0 -ej
0.0957592339261 2 1 -en 0.202462380301 1 1.0 -ej 0.957592339261 4 1 -h
1e3
```

Frequency and haplotype-based tests of neutrality using simulations

We examined the consistency of genetic variation with our neutral model using several approaches. First, we examined the likelihood of observing (Neandertal) allele frequencies as high as the OAS locus. Under neutrality, allele frequency is not dependent upon recombination rate; therefore, we can estimate the likelihood of our observed NLS

frequency in the OAS region. To create a distribution of expected NLS frequency in each introgression model, we first extracted derived allele counts (DAC) at simulated NLS for samples of 250 individuals each for Europeans and East Asians in the simulated model. Next, for each non-African population sample from the 1000 Genomes Project, we binomially sampled the DACs from the appropriate simulated population (European for CEU, GBR, IBS, TSI; East Asian for CDX, CHB, CHS, JPT, KHV) using the true population sample size to create a distribution of NLS frequencies that is specific to each population sample (Fig. 2a, Additional file 1: Figure S3). Minimum *P* values across the OAS locus for each population are listed in Additional file 2: Table S2.

Additionally, we wanted to examine if the haplotypes carrying NLS at the OAS locus are longer than expected under neutrality when conditioning on the observed frequencies and the underlying genetic map, which would provide an additional signature of selection on introgressed haplotypes beyond empirical haplotype signatures. For this purpose, we modified a simple haplotype statistic *H* (Ferrer-Admetlla et al. 2014), which measures the average length of pairwise homozygosity tracts in base pairs—a quantity that is very straightforward to interpret. As selective sweeps are expected to create long haplotypes around the selected site, the *H* statistic should be higher in samples containing positively selected haplotypes compared to samples containing neutrally evolving haplotypes, when frequency and recombination are properly controlled, similar to other statistics based on haplotype lengths, such as *EHH*, *iHS*, and *nSL* (Voight et al. 2006; Sabeti, P.C. et al. 2002; Huerta-Sanchez et al. 2014). However, in contrast to these other statistics, *H* does not require specification of analysis parameters such as minimum haplotype homozygosity levels below which haplotypes are no longer extended.

Under adaptive introgression, we specifically expect the introgressed (derived allele carrying) haplotypes to be longer than the ancestral haplotypes, due to the fact that selected Neandertal haplotypes will, on average, spend less time at low frequency, where they have a greater opportunity for recombination with non-introgressed haplotypes than neutral Neandertal haplotypes. We therefore defined our test statistic, $H_{D/A}$, as:

$$H_{D/A} = ln(H_D/H_A);$$

where H_D and H_A are H calculated across haplotypes carrying derived NLS allele versus the ancestral allele (see Additional file 1: Figure S10).

For each model we generated 1000 simulated OAS regions containing NLS that reach a frequency greater than the 99th percentile for CEU in our frequency simulations. We calculated $H_{D/A}$ for every simulated NLS and for every NLS in our true sample. We then compared our simulation results to the observed data in two ways. First, to visually examine the distribution of true $H_{D/A}$ values to neutral simulations, we took the mean $H_{D/A}$ score across all NLS in each of 1000 simulations and calculated the 95th and 99th percentiles of those 1000 simulations (Fig. 3b). Second, to examine the probability under neutrality of the observed $H_{D/A}$ values at the peak of $H_{D/A}$ between chr12:113380000–113420000, we compared the mean $H_{D/A}$ score for SNPs in this window to the mean of SNPs in this window for each simulation that produced a NLS in that window. These results for all tested models are listed in Additional file 2: Table S4.

Analysis of ancient Eurasian data

We utilized supplementary data Table 3 from Mathieson et al. (Mathieson et al. 2015). This table includes maximum likelihood allele frequency estimates for three ancient population samples (HG: hunter-gatherer, EF: early farmer, SA: Steppe

ancestry) and four present-day European samples from the 1000 Genomes Project (see the "Genome-wide scan for selection" section of methods in (Mathieson et al. 2015)). We intersect this table with allele frequencies for 1000 Genomes Yorubans (YRI) and the Altai Neandertal genotypes and only analyze sites for which we have data for all samples (1,004,612 SNPs).

To calculate the expected allele frequency in modern samples under drift, we used the estimated proportions (m) of (HG, EF, SA) in each of the four present-day samples estimated by Mathieson et al. (Mathieson et al. 2015): CEU = (0.196, 0.257, 0.547), GBR = (0.362, 0.229, 0.409), IBS = (0, 0.686, 0.314), and TSI = (0, 0.645, 0.355). We calculated the expected frequency E[p] of site as:

$$E[p] = \sum_{i}^{\{HG,EF,SA\}} (p_i \times m_i)$$

Next, we calculated the absolute difference between observed and expected allele frequency in all four present-day European samples at all available sites. To test the null hypothesis that OAS NLS have not changed in frequency more than expected under neutrality at 11 SNPs in the central OAS region (chr12:113200000–113600000), we first calculated the fraction of all autosomal SNPs in the dataset at similar present-day frequency (within 1% in the folded frequency spectrum) of each OAS NLS. Finally, we calculated the fraction of autosomal SNPs with an absolute observed minus expected frequency difference greater than or equal to the OAS NLS. These results are given in Additional file 2: Table S3 and illustrated in Fig. 2b.

This test does not explicitly incorporate variance in estimated ancient allele frequency. However, any bias in ancient allele frequency estimation should be distributed randomly across the genome. Therefore, our comparison to a genome-wide

distribution of SNPs at similar present-day frequency should incorporate most of this error. Nonetheless, the selection test performed by Mathieson et al. (Mathieson et al. 2015) does incorporate such error, so we can also look to the Ps from that test to ensure consistency with our results.

Estimation of selection coefficients

To estimate the selection coefficient s under constant positive selection for a given starting frequency (x_0), final frequency (x_1), and number of generations between these estimates (Δt), we assumed a model of standard logistic growth of a co-dominant allele:

$$x_1 = \frac{x_0}{x_0 + (1 - x_0)e^{-s\Delta t}}$$

This equation can be easily solved to obtain s, given x_0, x_1, and Δt.

Probabilities that Neandertal haplotypes are shared due to incomplete lineage sorting

We estimated the probability that Neandertal haplotypes observed in the OAS region could be a product of ILS using the method outlined by Huerta-Sanchez et al. (Excoffier et Lischer 2010). This method estimates the probability that a haplotype of length (H) is shared due to incomplete lineage sorting as:

$$P(H) = 1 - gammaCDF\left(H, shape = 2, rate = \frac{1}{L}\right);$$

where L is the expected shared length, $\mathrm{L} = 1/(\mathrm{r} \times \mathrm{t})$ where r is recombination rate and t is time in generations.

We first estimated the maximum haplotype length observed in each of the 1000 Genomes Project population samples by identifying tracts of identity-by-state (IBS)

that include NLS between the Altai Neandertal haplotype (considering only homozygous positions) and individual human haplotypes. We identified all tracts of IBS bounded by NLS (using YRI as the African reference) and took the maximum of these per-population sample IBS distributions.

We identified the maximum tract length (198,441 bp) in European (and some admixed Native American) samples. Maximum tract length for East Asian samples (54,385 bp) is notably shorter. As expected, no IBS tracts were identified in African population samples.

Assuming a conservatively low recombination rate for the OAS region (1.3×10^{-8}) and a conservatively short divergence between Neandertals and humans of 300,000 years (and 25 years per generation), the probability of a length of at least 54,385 is $P(54{,}385) = 0.002$. The probability of the longest tracts is much smaller, $P(198{,}441) = 1.15 \times 10^{-12}$. Therefore, it is unlikely that Neandertal haplotypes are shared with non-Africans due to shared ancestral variation.

We repeated this analysis using the Denisovan genome and found no IBS tracts defined by Denisovan-like sites, suggesting that introgressed Denisovan haplotypes described by Mendez et al. (F L Mendez, Watkins, et Hammer 2012) are not present in the 1000 Genomes Project Samples, consistent with our findings in Fig. 1a.

Neutrality tests in phase 3 of the 1000 Genomes data

We calculated iHS, DIND, Fst, and Tajima's D on different populations from phase 3 of the 1000 Genomes project. The phased data were downloaded from Impute reference data panel (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html) and were filtered for bi-allelic SNPs. Arlequin version 3.5.1.3 (Maclean, Chue Hong, et Prendergast 2015) was then used to calculate Fst estimates derived from ANOVA

among all pairs of European and Asian populations. DIND was calculated as previously described (Barreiro et al. 2009) and Tajima's D was calculated for all SNPs using a window of 25 kb either side of each SNP (i.e. in window of 50 kb in total). iHS values were recovered from a genomic scan performed using hapbin (Howie, Donnelly, et Marchini 2009) on the same phase 3 release of the 1000 Genomes dataset.

Sample collection

Buffy coats from 96 healthy European American donors and 41 African Americans were obtained from Indiana Blood Center (Indianapolis, IN, USA). Only individuals self-reported as currently healthy and not under medication were included in the study. The individuals recruited in this study were men aged 18–55 years old. For 80% of our samples, we have information on their exact age (for the remaining 20% of donors we only know that they were adults aged less than 55 years) and we found no association between the presence of the Neandertal haplotype (as measure by rs1557866) and age ($P > 0.5$). Thus, variation in age is not likely impact any of our conclusions. In addition to self-identified ancestry labels, we used the genome-wide genotype data (see the "DNA extraction and genotyping" section) to estimate genome-wide levels of European and African ancestry in each sample using the program ADMIXTURE (Howie, Donnelly, et Marchini 2009). Consistent with previous reports, we found that self-identified European Americans showed limited levels of African admixture (mean = 0.4%, range 0–13%). The African Americans used in our study were selected on having at least 75% of African ancestry.

DNA extraction and genotyping

DNA from each of the blood donors was extracted using the Gentra Pure Gene blood kit (Qiagen). Genotyping of each individual was then performed by Illumina's

HumanOmni5Exome bead Chip array and complemented with imputed data from the 1000 Genomes data using Impute2 (Storey et Tibshirani 2003). Here, we only focused on genetic diversity surrounding the OAS region – chr12:113229549–113574044 (~344 kb) spanning from the beginning of *RPH3A* to the end of *RASAL1*—for a total of 673 SNPs with a MAF above 10%. Genotypes can be found in Additional file 2: Table S8.

Isolation of monocytes and differentiation of macrophages

Blood mononuclear cells were isolated by Ficoll-Paque centrifugation. Monocytes were purified from PBMCs by positive selection with magnetic CD14 MicroBeads (Miltenyi Biotech) using the autoMACS Pro Separator. All samples had purity levels above 90%, as measured by flow cytometry using an antibody against CD14 (BD Biosciences). Monocytes were then cultured for seven days in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent), L-glutamine (Fisher), and M-CSF (20 ng/mL; R&D systems). Cell cultures were fed every two days with complete medium supplemented with the cytokines previously mentioned. Before infection, we systematically verified that the differentiated macrophages presented the expected phenotype for non-activated macrophages (CD1a+, CD14+, CD83−, and HLA-DRlow (BD Biosciences)).

Bacterial preparation and infection of macrophages

The day prior to infection, aliquots of *Salmonella typhimurium* (Keller strain) were thawed and bacteria were grown overnight in Tryptic Soy Broth (TSB) medium. Bacterial culture was diluted to mid-log phase prior to infection and supernatant density was checked at OD600. Monocyte-derived macrophages were then infected with *S. typhimurium* at a multiplicity of infection (MOI) of 10:1. A control group of

non-infected macrophages was treated the same way but using medium without bacteria. After 2 h in contact with the bacteria, macrophages were washed and cultured for another hour in the presence of 50 mg/mL of gentamycin in order to kill all extracellular bacteria present in the medium. The cells were then washed a second time and cultured in complete medium with 3 mg/mL gentamycin for an additional 2 h, the time point to which we refer in the main text.

Infection/stimulation of PBMC

PBMCs from a subset of 40 individuals used to derive macrophages were cultured in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent) and 1% L-glutamine (Fisher). The 30 individuals were chosen based on their genotype for rs1557866, a SNP which derived allele is of Neandertal origin and that we used as a proxy to identify individuals harboring the Neandertal haplotype in the OAS region. From the 40 individuals, and based on this SNP, nine individuals were homozygous for the Neandertal haplotype, 15 were heterozygous, and 16 homozygous for the modern human sequence.

For each of the tested individuals, PBMCs (1 million per condition) were stimulated/infected with one of the following viral-associated immune challenges: polyI:C (10 μg/mL, TLR3 agonist), Gardiquimod (0.5 μg/mL, TLR7 and TLR8 agonist), Influenza PR8 WT (multiplicity of infection (MOI) of 0.05:1), HSV 1 (1.55×10^2 CPE), and HSV2 (19.5×10^4 CPE). PBMCs were stimulated/infected for 4 h with TLR ligands and Influenza and 6 h with HSV1 and HSV2. A control group of non-infected PBMC was treated the same way but with only medium.

RNA extraction, RNA-seq library preparation, and sequencing

Total RNA was extracted from the non-infected and infected/stimulated cells using the miRNeasy kit (Qiagen). RNA quantity was evaluated spectrophotometrically and the quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Only samples with no evidence of RNA degradation (RNA integrity number > 8) were kept for further experiments. RNA-seq libraries were prepared using the Illumina TruSeq protocol. Once prepared, indexed complementary DNA (cDNA) libraries were pooled (6 libraries per pool) in equimolar amounts and sequenced with single-end 100 bp reads on an Illumina HiSeq2500. Results based on the entire dataset are described elsewhere (Nédélec et al. 2016). Here, we only studied transcript-level and gene-level expression estimates for *OAS1*, *OAS2*, and *OAS3*. The transcript-level and gene-level expression data for these genes can be found in Additional file 2: Table S9.

Quantifying gene expression values from RNA-seq data

Adaptor sequences and low quality score bases (Phred score < 20) were first trimmed using Trim Galore (version 0.2.7). The resulting reads were then mapped to the human genome reference sequence (Ensembl GRCh37 release 65) using TopHat (version 2.0.6) and using a hg19 transcript annotation database downloaded from UCSC. Gene-level expression estimates were calculated using featureCounts (version 1.4.6-p3) and transcript-level expression values were obtained using RSEM under default parameters.

Quantitative real-time PCR

For the PBMC samples, we measured the expression levels of OAS and interferon genes using real-time PCR. A total of 100 ng of high-quality RNA was reverse-transcribed into cDNA using the qScript cDNA SuperMix (Quanta Biosciences).

Quantitative real-time PCR was performed using 96.96 Dynamic Array™ IFCs and the BioMark™ HD System from Fluidigm. For the TaqMan gene assays, we used the following TaqMan Gene Expression Assay (Applied BioSystems) to quantify the expression levels of interferon genes: *IFNA1* (Hs03044218), *IFNA6* (Hs00819627), and *IFNG* (Hs00989291). To quantify the overall expression levels of OAS genes, we used probes that capture all common isoforms of *OAS1* (Hs00973635), *OAS2* (Hs00942643), and *OAS3* (Hs00196324).

Custom-made probes were designed to specifically target the short-isoform of *OAS2* (Forward Primer Sequence CTGCAGGAACCCGAACAGTT; Reverse Primer Sequence ACTCATGGCCTAGAGGTTGCA; Reporter Sequence AGAGAAAAGCCAAAGAA).

As housekeeping genes, we used: *GAPDH* (Hs02758991), *GUSB* (Hs99999908), *HPRT1* (Hs99999909) and *POLR2A* (Hs00172187). The results reported in the manuscript used *POLR2A* as a reference but all conclusions remain unchanged when using any of the other housekeeping genes.

We start by doing a preamplification of the cDNA using the PreAmp Master Mix (Fluidigm). Preamplified cDNA was then diluted 2× on a solution of 10 mM Tris–HCl (pH 8.0) and 0.1 mM EDTA. In order to prepare samples for loading into the integrated fluid circuit (IFC), a mix was prepared consisting of 360 μL TaqMan Fast Advanced Master Mix (Applied BioSystems) and 36 μL 20× GE Sample Loading Reagent (Fluidigm). A total of 2.75 μL of this mix was dispensed to each well of a 96-well assay plate and mixed with 2.25 μL of preamplified cDNA. Following priming of the IFC in the IFC Controller HX, 5 μL of the mixture of cDNA and loading reagent were dispensed in each of the sample inlet of the 96.96 IFC. For the TaqMan gene

assays, 5 μL of mixes consisting of 2.5 μL 20× TaqMan Gene Expression Assay (Applied BioSystems) and 2.5 μL 2× Assay Loading Reagent (Fluidigm) were dispensed to each detector inlet of the 96.96 IFC. After loading the assays and samples into the IFC in the IFC Controller HX, the IFC was transferred to the BioMark HD and PCR was performed using the thermal protocol GE 96 × 96 Fast v1.pcl. This protocol consists of a Thermal Mix of 70 °C, 30 min; 25 °C, 10 min, Hot Start at 95 °C, 1 min, PCR Cycle of 35 cycles of (96 °C, 5 s; 60 °C, 20 s). Data were analyzed using Fluidigm Real-Time PCR Analysis software using the Linear (Derivative) Baseline Correction Method and the Auto (Detectors) Ct Threshold Method.

To quantify the expression levels of the *OAS1* isoform associated with the derived allele at the splicing variant rs10774671, we used SybrGreen and the following forward (GCTGAGGCCTGGCTGAATTA) and reverse (CCACTTGTTAGCTGATGTCCTTGA) primers. PCR was performed using the thermal protocol 50 °C, 2 min; 95 °C, 10 min, PCR cycle of 40 cycles of (95 °C, 15 s; 60 °C, 1 min). A melting curve was also performed to check for non-specific amplification.

Genotype–phenotype association analysis

eQTL and asQTL were performed against *OAS1*, *OAS2*, and *OAS3*. We examined associations between SNP genotypes and the phenotype of interest using a linear regression model, in which phenotype was regressed against genotype. In particular, expression levels were considered as the phenotype when searching for eQTL and the percentage usage of each isoform in each gene when mapping asQTL. To avoid low power caused by rare variants, only SNPs in the OAS region with a minor allele frequency of 10% across all individuals were tested (i.e. 673 SNPs within the region

chr12:113229549–113574044). In all cases, we assumed that alleles affected the phenotype in an additive manner. For the eQTL and asQTL analyses on macrophages, we mapped Salmonella-infected and non-infected samples separately. We controlled for FDRs using an approach analogous to that of Storey and Tibshirani (Storey et Tibshirani 2003), which makes no explicit distributional assumption for the null model but instead derives it empirically null from permutation tests, where expression levels were permuted 1000 times across individuals. For the non-infected and infected/stimulated PBMCs we only tested expression levels against the SNPs identified as eQTL or asQTL in the macrophage data (specifically, the SNPs for which boxplots are shown in Fig. 4).

Contenu supplémentaire du chapitre 5

Exemples de l'utilisation de l'API ImmunPop

Exemple d'une requête lorsqu'on désire rechercher les associations en cis entre un SNP et le gène TLR2 (identifiant *Ensembl* : ENSG00000137462) qui sont significatives :

http://immunpop.com/nedelec/eQTL/association/gene/ENSG00000137462/local/1pc

Variable	Commentaires
`projet`	nom du projet
`analyse`	type d'analyse : **`eQTL`**, **`reQTL`** ou **`asQTL`**
`élément`	type d'élément à rechercher : **`gene`** ou **`snp`**

`ID`	identifiant de l'élément à rechercher. snp_id ou ensembl_gene_id
type_asso	type d'association recherchée : **`local`** (cis) ou **`distant`** (trans)

TABLEAU 1 - EXEMPLE DE SYNTAXE A DESTINATION DE L'API

URL : **immunpop.com/<projet>/<analyse>/association/<élément>/<ID>/<type_asso>**

Le serveur répond à travers un contenu JSON :


```
{
  "data": {
    "associations": [
      {
        "ensembl_gene_id": "ENSG00000137462",
        "external_gene_id": "TLR2",
        "gene_chr": "4",
        "gene_end": 154626851,
        "gene_start": 154622652,
        "with": [
          {
            "interaction": "local",
            "snp_chr": "4",
            "snp_id": "rs62323831",
            "snp_position": 154610018,
            "conditions": [
              {
                "condition": "Salmonella",
                "estimate": -15.3194715780919,
                "pvalue": 1.52403400632591e-7
              },
              {
                "condition": "Listeria",
                "estimate": -22.399409979735,
                "pvalue": 2.64705682105746e-10
              }
            ]
          }
        ]
      }
    ],
    "filterFdr": "1pc",
    "filterInteraction": "local"
  }
```