Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :

Génétique Evolutive Humaine

Présentée et soutenue par :
Estelle VASSEUR

le : mardi 20 décembre 2011

Titre :

Contraintes sélectives et adaptation chez l'homme :
histoire évolutive des senseurs microbiens

JURY

Howard Cann Rapporteur, Christine Dillmann Rapportrice,
Philippe Sansonetti Examinateur, Jérôme Clain Examinateur,
Brigitte Crouau-Roy Directrice de thèse, Lluis Quintana-Murci Directeur de thèse

---

Ecole doctorale :
Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :
Evolution et Diversité Biologique et Unité Génétique Evolutive Humaine, Institut Pasteur

Directeur(s) de Thèse :

Brigitte Crouau-Roy et Lluis Quintana-Murci

Rapporteurs :

Howard Cann et Christine Dillmann

A Cédric,

Remerciements

Je souhaiterais tout d'abord remercier Christine Dillmann et Howard Cann de m'avoir fait l'honneur d'accepter d'être les rapporteurs de ma thèse et pour le temps consacré à la lire et à la commenter.

Je remercie également Jérôme Clain d'avoir accepté avec enthousiasme de faire partie de mon jury de thèse.

Je remercie Philippe Sansonetti, qui, malgré un emploi du temps chargé, s'est montré disponible pour mes comités de thèse annuels et a également accepté de faire partie de ce jury, mais aussi pour nos discussions très enrichissantes sur les NLRs.

Un grand merci à Kenneth Vernick, pour sa disponibilité et pour la pertinence de ses questions lors de nos échanges.

Je remercie profondément Brigitte Crouau-Roy, ma co-directrice de thèse, qui a eu l'idée de cette collaboration et qui s'est toujours montrée disponible et enthousiaste.

Je tiens bien évidemment à exprimer mes sincères remerciements à Lluìs Quintana-Murci, pour avoir été mon co-directeur de thèse. Sans toi, bien sûr, ce projet n'existerait pas, mais bien au-delà de cela, j'ai été très heureuse de travailler dans ton équipe. Comme je te l'ai répété, tu es un chef exceptionnel, compréhensif et soucieux de l'épanouissement de chaque membre de ton unité. Et en tant que « super manager », tu as su t'assurer de l'avancement de mon projet, tout en me laissant de la liberté et en me faisant confiance. Plus que tout, merci pour ta disponibilité : malgré tes responsabilités, tu as toujours trouvé du temps pour discuter du projet, aussi bien que de mon avenir en général et pour me conseiller.

Aux autres membres, passés et présents du laboratoire. Je remercie Christine Harmant pour nos discussions, en particulier de m'avoir fait connaître Groupon… Et plus sérieusement, merci de ton soutien quotidien. Ton rôle de bonne fée au laboratoire va bien au-delà de ta responsabilité hygiène et sécurité, en veillant véritablement sur chacun d'entre nous.

A Hélène Quach et Jérémy Manry pour nos discussions scientifiques et non-scientifiques et pour avoir résisté au mouvement des gens qui font la grève de la faim le midi. Vous êtes probablement les deux personnes que j'aurai sollicitées le plus durant ma thèse. Merci pour tous vos conseils avisés d'« anciens ».

A Etienne Patin, pour tes explications claires, rapides et précises, pour les discussions toujours très productives que nous avons eu ; ta présence au laboratoire durant ma troisième année aura très certainement fortement contribué à la qualité des articles inclus dans ma thèse et à la publication de l'un d'eux dans HMG, face à un Lluis dubitatif...

Un merci également à Guillaume Laval pour son travail de bioinformaticien. Bien que tu aies ruiné mes espoirs plus d'une fois en m'annonçant que mes résultats n'étaient plus significatifs avec ton modèle, tu m'auras permis de considérer le moindre résultat comme une perle rare.

A Simona Fornarino, pour ton extrême rigueur scientifique, mais aussi pour notre passion commune pour l'Inde et ses grands principes sur le respect de la vie.

A Katie Siddle et Maud Fagny pour leur présence et leur bonne humeur. Je leur souhaite bonne chance, ainsi qu'à Eddie Loh.

A Michele Boniotto, merci pour ta collaboration, lors du regroupement NALPs/NODs, qui n'aura pas été simple, mais va bientôt s'achever avec, espérons-le, un bon article à la clé.

A Sandra Pajon, pour toutes les manips que tu as réalisées avec rigueur et clarté ; quelle facilité de reprendre le travail où tu l'avais laissé ! J'espère au départ t'avoir guidée au mieux pour que ton stage de master démarre rapidement et te permette d'entrevoir de véritables conclusions à la fin de ces quelques mois. Je te souhaite bonne chance à Marseille ou ailleurs.

A Elise Merlo et Cécile Roux, pour votre aide tout au long de ces 3 ans, toujours avec bonne humeur, facilitant grandement l'ensemble des démarches administratives.

De Toulouse, je remercie également Claude Maranges, directeur actuel de SEVAB et ancien professeur à l'INSA, pour son soutien ; depuis le tout début de cette thèse, vos décisions ont joué un rôle majeur dans le bon déroulement de ma thèse.

Merci à Dominique Pantalacci, qui de Toulouse, s'est montrée particulièremeent rapide et efficace, afin de faciliter les démarches et l'organisation de ma soutenance de thèse. Celle-ci ne se serait certainement pas passée dans d'aussi bonnes conditions si vous n'aviez pas été là.

A Lucie, Caro, Stéphane, Capucine, Minh et Karim, vous pour qui je suis partie dans le grand Nord, loin de Toulouse, merci d'avoir accepté de mettre votre équipement de ski pour venir nous rendre visite régulièrement.

A Karim et Caro en particulier, bon courage, votre tour se rapproche pour la soutenance!

A Manu, pour ton humour, tes réflexions incroyablement réalistes et pour ta vision absolument unique du monde.

Pour finir, je remercie ma famille, ma mère, Ludi, Alex : malgré la distance, vous restez toujours présents pour moi.

A ma mère, tu m'as toujours soutenue, rassurée, aidée, encouragée (et la liste serait encore longue) depuis le moment où je clamais, petite, « plus tard, je veux être chercheur scientifique en génétique !» jusqu'à aujourd'hui ; disons tout simplement que c'est grâce à toi que j'en suis arrivée là ; tu es une mère comme il n'en existe aucune autre.

A ma grande sœur Ludi, dans mon cœur, peu importe le temps qui passe.

A Alex, mon grand frère, déjà mari et maintenant papa : je vous souhaite beaucoup de bonheur, à Marie, Camille et toi. Mais s'il vous plaît, rapprochez-vous de quelques milliers de km, j'en ai assez de skype…

A mes boules de poils passées et présentes, Automne, Attila, Fritz, Shadow et Souricette, pour le bonheur de vous avoir trouvées, pour le réconfort que vous m'apportez. Comme l'a dit le docteur Albert Schweitzer « Il y a deux moyens d'oublier les tracas de la vie : la musique et les chats ». J'ai choisi les deux. Et j'espère vous donner ce même réconfort en retour.

Et bien sûr à toi, Cédric, ma moitié, à qui je dédie cette thèse, pour ton soutien, ta gentillesse et ta patience. Pour avoir été auprès de moi dans les instants de doutes autant que dans les bons moments; pour t'être intéressé à la biologie comme aucun informaticien ne l'avait jamais fait avant et pour avoir été heureux pour moi (pour ne pas me vexer) quand je t'annonçais avec le sourire « Plus que 70 plaques de PCR à faire !». Merci pour les délicieux plats que tu m'a préparés en attendant mon retour du labo… Et enfin, merci pour ta franchise (qui a dit manque de tact ?) qui est l'une des nombreuses qualités que j'aime chez toi.

Résumé en français

Les maladies infectieuses ont joué un rôle considérable dans la diversité de notre génome. La détection de sélection naturelle s'est avérée très utile pour mieux comprendre les mécanismes qui ont gouverné notre évolution. Elle permet également de distinguer les gènes qui sont essentiels à notre survie de ceux qui partagent des fonctions plus redondantes. Nous nous intéressons ici aux pressions de sélection exercées par les pathogènes sur notre génome et donc aux molécules chez l'homme qui sont impliquées dans leur reconnaissance.

Nous avons pour cela étudié deux familles de récepteurs microbiens de l'immunité innée ou *Pattern Recognition Receptors* (PRRs) se trouvant dans le cytosol, les 21 NLRs (*Nod-like receptors*) qui reconnaissent *a priori* les bactéries et les signaux de danger cellulaire et les 3 RLRs (*Rig-like receptors*) détectant les ARN provenant majoritairement de virus. Pour cela, nous avons reséquencé ces gènes dans un panel d'individus sains représentatif de la population mondiale. A partir de cette diversité, nous avons détecté des signatures de sélection naturelle en utilisant différents tests inter- et intra-spécifiques.

Nous avons d'une part mis en évidence que la majorité des *NALPs*, l'une des sous-familles de *NLRs*, était très contrainte par la sélection, montrant un déficit en mutations non-synonymes. Cela suggère qu'ils ont joué un rôle essentiel à notre survie et pourraient être impliqués dans certaines maladies conférant des phénotypes sévères. Au contraire, la plupart des *NOD/IPAF*, autre sous-famille des *NLRs*, ainsi que les 3 *RLRs* ne présentent pas de fortes contraintes. Il semble donc ils aient joué un rôle relativement moins important, accumulant un nombre conséquent de mutations altérant la protéine. Ainsi, l'intégration de ces données à celles d'autres groupes de senseurs microbiens nous a permis de proposer un modèle hiérarchique général, traduisant les contributions relatives des différentes familles de senseurs microbiens humains à notre survie. Parmi les gènes codant les récepteurs impliqués dans la reconnaissance de virus, ceux des *Toll-like-receptors* (TLRs) endosomaux se trouvent très contraints car ils doivent assurer des fonctions essentielles, laissant les *RLRs* muter « plus librement ». Quant aux gènes des récepteurs de produits bactériens ou de signaux de danger cellulaire, les *NALPs* semblent très contraints et doivent jouer des rôles importants, comparés aux *NOD/IPAF*, aux *RLRs* et aux gènes codant les TLRs situés à la surface des cellules. Les informations fonctionnelles dont nous disposons actuellement semblent indiquer des rôles peu classiques pour certains NLRs, en particulier pour les NALPs : leur essentialité pourrait par conséquent également résulter de fonctions telles que le développement embryonnaire ou l'homéostasie cellulaire.

D'autre part, nous avons détecté des signaux de sélection positive sur certains des gènes étudiés ; les plus forts ont été trouvés pour *NLRP1*, dont un haplotype semble avoir augmenté en fréquence dans l'ensemble des populations étudiées (Afrique, Europe et Asie), probablement parce qu'il a conféré dans le passé une plus grande résistance face à certaines infections.

De manière générale, ce projet aura d'une part permis de mieux cerner quels sont les senseurs microbiens humains ayant joué un rôle prépondérant et qu'il faudrait donc étudier en priorité dans une perspective médicale. D'autre part, les gènes et variants particuliers que nous avons trouvés comme étant sous sélection positive pourraient expliquer les différences de résistance qui existent actuellement face à certaines maladies infectieuses.

English summary

Infectious diseases have played a major role in shaping the patterns of human genetic diversity. Detecting signatures of natural selection in the human genome represents a useful tool for improving our understanding of the mechanisms governing our evolution. It also allows us to distinguish genes that are essential to survival from those that share redundant functions. Here we focused on the selective pressures exerted by pathogens in our genomes and therefore on human molecules that are specialised in the sensing of pathogens.

To this end, we studied the levels of naturally-occurring variation of two major gene families determining innate immunity microbial sensors, or Pattern Recognition Receptors (PRRs), localized to the cytosol. These included the 21 NLRs (Nod-like receptors), which *a priori* detect bacteria and signals of cell danger, and the 3 RLRs (Rig-like receptors), which are specialized in the recognition of RNA, mainly from viruses. We resequenced these genes in a panel of individuals, not selected for disease, from different populations worldwide. We used the resulting dataset to detect signatures of natural selection, using various inter- and intra-specific tests.

First, we showed that most of the genes encoding the NALPs, a sub-family of the NLRs, were under the action of strong purifying selection, exhibiting a deficit in non-synonymous mutations. This suggests that they have played an essential role in host survival and could be involved in several diseases associated with severe phenotypes. Conversely, the majority of NOD/IPAF genes, another *NLR* subfamily, together with the *RLRs* were found to evolve under weaker selective constraints. It seems therefore that they have played a relatively lesser essential role, accumulating a high number of protein-altering mutations. The integration of these data with those from other genes of microbial sensors allowed us to propose a general hierarchical model, indicating differences in their relative contributions of the human microbial sensors to our survival. Among the genes encoding receptors specialised in the sensing of nucleic acids particularly from virus, the endosomal *Toll-like receptor* (TLR) genes are those under the strongest constraints, highlighting their essential role, whereas such constraints are more relaxed among the *RLRs*. With respect to the genes encoding sensors detecting products from bacteria and stress signals, *NALPs* appeared to be the most strongly constrained, with respect to those encoding most NOD/IPAF members, RLRs and cell-surface TLRs. Functional information from these families of microbial sensors highlighted « non-classical roles » for several NLR receptors, especially among the NALPs: the strong selective constraints observed for their genes could result from functions that go beyond immunity to infection, such as embryonic development and cellular homeostasis.

Moreover, we detected signatures of positive selection for some of the genes studied; the strongest signatures for a *NLRP1*-containing haplotype that seems to have increased in frequency in populations from each of the three continental regions studied (*i.e.* Africa, Europe and Asia), probably because it conferred a higher resistance against some infectious diseases in the past.

Overall, this study has allowed us to assess the relative biological relevance of different microbial sensors in humans and highlights those that have played an essential role in host survival; their genes should be now studied from a medical perspective. In addition, the genes and variants found here to be targeted by positive selection could explain the actual differences observed between individuals and populations in the susceptibility to, and pathogenesis of, infectious diseases.

Table des matières

INTRODUCTION

I) La diversité génétique, source de diversité phénotypique

Depuis la première espèce vivante nommée LUCA (pour *Last Universal Common Ancestor* et par analogie avec l'australopithèque Lucy) apparue il y a au moins 3,8 milliards d'années, l'ensemble des êtres vivants semble partager une caractéristique commune, les acides nucléiques (Forterre, Gribaldo, Brochier 2005); qu'ils s'agissent d'acides ribonucléiques (ARN) ou d'acides désoxyribonucléiques (ADN), ceux-ci contiennent l'ensemble des informations nécessaires au développement, au fonctionnement d'un être vivant. De plus, étant transmis lors de la reproduction, ils vont servir de support de l'hérédité. Ces longues chaînes d'acides nucléiques sont constituées de simples petites molécules, les nucléotides (A pour Adénine, T pour Thymine — ou U pour Uracile dans l'ARN — G pour Guanine et C pour Cytosine). Et c'est l'enchaînement particulier de ces « briques » élémentaires qui va établir un code constituant un message génétique précis, distinguant chaque espèce des autres espèces, chaque individu des autres membres de son espèce.

1) La diversité génétique humaine

Le génome humain totalise 3,2 milliards de nucléotides. Cependant, plus de 99,9% de ceux-ci sont identiques d'un individu à l'autre. C'est donc uniquement le reste de ce code qui différencie nos génomes et rend chacun d'entre nous unique. L'espèce humaine — *Homo sapiens* — est celle qui présente le moins de diversité parmi les hominidés actuels, incluant notamment le chimpanzé possédant pourtant une taille effective bien plus faible (Lander et al. 2001; Venter et al. 2001; Fischer et al. 2004; Thalmann et al. 2007). Cette faible différenciation de l'espèce humaine peut s'expliquer par son apparition récente, il y a seulement 200 000 ans environ (Li, Sadler 1991; Chen et al. 1995; Harpending, Rogers 2000; Ingman et al. 2000; McDougall, Brown, Fleagle 2005; Santos-Lopes et al. 2007). Ainsi, notre diversité génétique repose sur les 0.1% de différences restantes, soit quelques 3 millions de nucléotides qui distinguent en moyenne deux humains (Lander et al. 2001; Venter et al. 2001) ; cela correspond environ à une paire de base tous les 100 à 500 nucléotides entre deux séquences uniques prises au hasard chez deux individus non apparentés. Cependant, la majeure partie de ces différences génétiques n'aura pas d'effet connu à l'échelle de l'individu ni au niveau moléculaire, *i.e.* d'effet phénotypique. En effet, un grand nombre d'entre elles touchera des régions dites « neutres », qui peuvent s'avérer en majeure partie être des régions inter-géniques ou introniques, mais aussi certains sites exoniques changeant dans la protéine un acide aminé en un autre acide aminé aux propriétés chimiques proches. Et ainsi, seule une

faible part de notre diversité génétique se traduira par des différences phénotypiques : ces mutations peuvent être responsables de différences morphologiques ou physiologiques ; elles peuvent en partie expliquer nos spécificités en matière de digestion de certains aliments, ou encore de sensibilité à des environnements climatiques extrêmes (Tishkoff et al. 2007; Enattah et al. 2008; Beall et al. 2010; Itan et al. 2010; Yi et al. 2010) ; plus important encore, elles peuvent nous permettre de mieux comprendre nos disparités en matière de résistance ou de susceptibilité à certaines maladies, ou encore nos différentes réactions aux vaccins et traitements utilisés (Allison 1961; Dean et al. 1996; Cooke, Hill 2001; Ma et al. 2007; Poland et al. 2007).

Un point particulièrement intéressant, il a récemment été estimé que chacun d'entre nous porte un nombre notable de mutations conduisant à des gènes non-fonctionnels : on trouverait pour chaque individu 190 à 210 insertions/délétions dans le cadre de lecture, entre 80 et 100 mutations non-sens précoces et 40 à 50 mutations qui rompent un site d'épissage (Durbin 2010). De manière générale, chacun d'entre nous possèderait 250 à 300 gènes portant des mutations responsables d'une perte de fonction, parmi les 20 000 à 25 000 gènes codants identifiés au sein de l'espèce humaine (Lander et al. 2001; Venter et al. 2001; Durbin 2010). Un résultat plus frappant encore, chaque génome comporterait entre 50 et 100 mutations classées dans la *Human Gene Mutation Database* comme responsables de maladies héréditaires.

2) Où en sommes-nous dans l'étude de cette diversité?

Les variations les plus répandues et donc les plus étudiées de notre génome sont les SNPs, ou *single nucleotide polymorphisms* : ils consistent en un remplacement d'un nucléotide par un autre. En 2001, le séquençage de notre génome a apporté une base solide pour l'étude de ces SNPs (Lander et al. 2001; Venter et al. 2001). En 2002, le projet international HapMap (http://www.hapmap.org) a ensuite fourni une base de données de plus de 3 millions de mutations sur 270 échantillons d'ADN (Consortium 2003). Le lancement en 2008 du projet *1000 Genomes*, qui exploite les technologies de séquençage d'ADN de nouvelle génération, a permis de développer une base de données publique contenant les informations des génomes de 2500 personnes issues de 27 populations dans le monde ; il contribue ainsi fortement à la découverte de nouveaux SNPs (Durbin 2010). Aujourd'hui, on trouve plus de 15 millions de SNPs validés dans dbSNP, Build135 (Sherry et al. 2001). Ces SNPs sont couramment utilisés en tant que marqueurs moléculaires dans des associations génotypes/phénotypes. Le projet

HapMap a d'ailleurs réalisé une première génération d'études d'association du génome (GWAS, *Genome Wide Association Studies*) qui ont permis de localiser plus de 600 facteurs de risques génétiques pour diverses maladies telles que le cancer du sein, le diabète et la schizophrénie par exemple (Consortium 2007; Figueroa et al. 2011; Ripke et al. 2011). Ces informations s'avèrent très précieuses, dans la mesure où elles peuvent permettre de mieux comprendre de nombreuses pathologies et à terme, de développer des traitements contre celles-ci ; la thérapie génique, qui vise à remplacer les gènes défectueux, où des mutations délétères apparaissent, constitue un exemple évident des applications possibles (Cavazzana-Calvo et al. 2010).

Cependant, d'autres approches utilisant également les données de SNPs peuvent s'avérer utiles : c'est le cas de la génétique des populations, sur laquelle est basée ce projet de thèse. La génétique des populations pourrait être définie comme l'étude de la distribution et des modifications de fréquence des mutations observées au sein de différents groupes d'individus donnés. Un des principaux intérêts de cette approche est qu'elle peut notamment permettre d'identifier des signatures de sélection naturelle sur certains SNPs, montrant qu'ils ont joué un rôle assez important dans notre évolution pour être ciblés (Sabeti et al. 2006; Quintana-Murci et al. 2007; Barreiro, Quintana-Murci 2010). Néanmoins, avant de développer davantage les méthodes de détection de sélection, il convient de rappeler les différents facteurs qui affectent notre diversité génétique.

II) Quelles sont les forces modifiant notre diversité génétique ?

Ainsi, l'existence de ces 3,2 milliards de nucléotides que nous avons en commun, ainsi que des quelques 3 millions de SNPs qui différencient en moyenne deux humains résultent de l'action et l'interaction de différents facteurs (**Figure 1**). Si les forces génomiques (mutations et recombinaison) ont tendance à favoriser l'augmentation de diversité génétique, la démographie (modification de taille ou migration d'une population), la dérive génétique et la sélection naturelle (adaptation à un environnement pathogénique, climatique, etc.) vont affecter les fréquences de ces mutations. Les facteurs sociaux et culturels, qui peuvent s'avérer très complexes chez l'homme, peuvent également jouer un rôle non négligeable dans sa variabilité génétique (Salem et al. 1996; Seielstad, Minch, Cavalli-Sforza 1998; Perez-Lezaun et al. 1999; Chaix et al. 2004; Wilder, Mobasher, Hammer 2004; Chaix et al. 2007). L'intensité et le poids relatif de ces facteurs peuvent varier entre individus, entre populations, et dépendent pour la plupart de la région génomique étudiée.

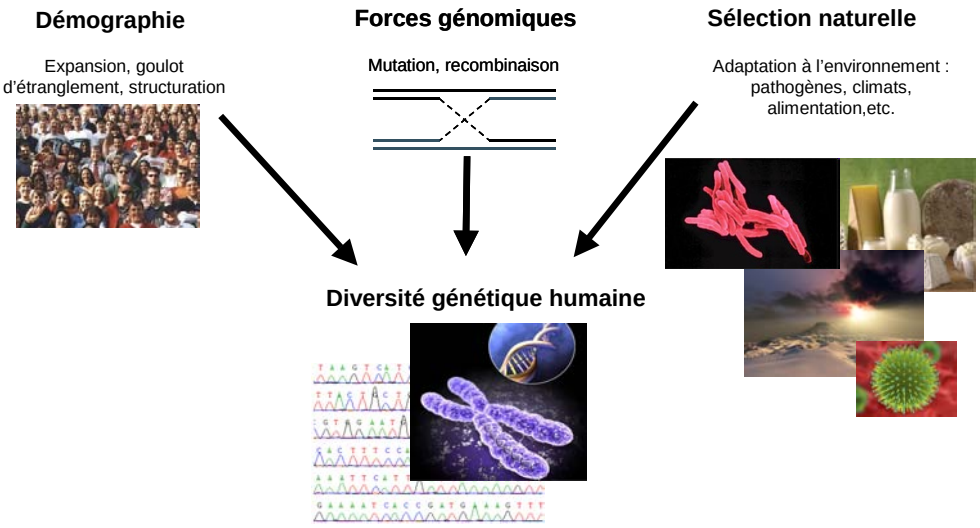


Figure 1. Forces influençant notre diversité génétique

1) Les forces génomiques, créatrices de diversité

Tout d'abord, l'existence même de diversité génétique repose sur ce que nous pouvons appeler de manière générale les forces génomiques.

a) Les mutations

Tout au long de notre vie, se produisent des modifications spontanées de notre ADN, les mutations. Celles-ci peuvent résulter d'erreurs lors de la réplication (amplification de l'ADN), mais aussi lors de la réparation de ces erreurs. Notamment lors de cassure de l'ADN, les remaniements réalisés pour réparer le chromosome peuvent mener à des reconstitutions chromosomiques qui ne correspondent plus à la configuration de départ. Ainsi, que ce soit pendant la réplication ou la réparation, différents types de mutations peuvent survenir dans notre ADN. Nous l'avons abordé dans le paragraphe précédent, une mutation peut affecter un seul nucléotide, en le remplaçant par un autre. Une telle mutation sera dite polymorphe si elle atteint une certaine fréquence dans la population (arbitrairement choisie à 1%). Mais d'autres types de mutations peuvent également apparaître ; certains motifs de plus d'un nucléotide répétés peuvent être amplifiés en tandem : on distingue les microsatellites (motifs de 1 à 6 paires de bases, pb), les minisatellites (motifs de 6 à 100 pb) et les CNVs (*Copy Number Variants*, allant de 1 000 pb à plusieurs millions de pb) (Sebat et al. 2004; Tuzun et al. 2005; Redon et al. 2006; Cooper, Nickerson, Eichler 2007; Kidd et al. 2008). On peut aussi observer des insertions, des délétions, des translocations (échanges d'une fraction d'ADN entre deux chromosomes non-homologues) ou encore des inversions pouvant atteindre plusieurs milliers de paires de bases.

Certains éléments extérieurs, tels que les rayonnements ultraviolets et certains agents chimiques dits « mutagènes » peuvent augmenter la probabilité de ces erreurs (Altenburg 1928; Auerbach, Robson, Carr 1947). De manière générale, les mutations vont créer de la diversité et si elles apparaissent dans une cellule germinale, c'est-à-dire qui peut être transmise à la descendance, alors elles seront potentiellement « héréditaires » et pourront se propager dans la population.

b) Les mécanismes de recombinaison

Outre les mutations, la recombinaison crée également de la diversité. En effet, en l'absence de recombinaison, les gènes situés sur le même chromosome seraient toujours transmis ensemble des parents à l'enfant. Cependant, la recombinaison intra-chromosomique va briser les associations de gènes existantes : au cours de la méiose, la division cellulaire menant à la production des gamètes, un brassage génétique est réalisé ; les deux chromosomes d'une même paire (portant des allèles différents à un certain nombre de loci) peuvent échanger certaines de leurs portions par un mécanisme d'enjambement (« crossing-over ») (Creighton, McClintock 1931). Il peut ainsi y avoir des modifications dans l'association des

allèles portés par chacun des chromosomes, c'est-à-dire dans ce que l'on appelle le déséquilibre de liaison (DL). La recombinaison n'est pas constante le long des chromosomes : elle varie en fonction de la distance au centromère, de la proportion en G+C, ou de la présence de certains motifs cibles (Frazer et al. 2007). On distingue en particulier des régions montrant un taux très fort, les *hotspots*, observés en moyenne tous les 200 kilobases (kb) (McVean et al. 2004; Myers et al. 2005; Przeworski 2005) et d'autres au contraire qui arborent un taux de recombinaison particulièrement bas, les *coldspots*. De manière générale, le taux de recombinaison reste faible dans les régions géniques, probablement en raison du risque d'inactivation du gène. Les régions régulatrices du gène (en 5') sont par contre des cibles privilégiées de la recombinaison, du fait du fort pourcentage en G+C qu'elles arborent généralement (Myers et al. 2005; Lindsay et al. 2006). Au sein de l'espèce humaine, il a été estimé qu'en 30 000 ans, un chromosome subit une recombinaison toutes les 100 kb.

D'autre part, toujours lors de la méiose, il se produit une séparation aléatoire des chromosomes, c'est la recombinaison inter-chromosomique : chaque futur gamète reçoit une combinaison particulière de chromosomes. Nous possédons 23 paires de chromosomes, il y a donc 2^{23}, soient plus de 8 millions de combinaisons possibles, fournissant ainsi un large nombre de gamètes possibles et donc une grande diversité. Mutation et recombinaison seront les seules forces véritablement génératrices de diversité.

2) Comment notre histoire démographique affecte-t-elle notre diversité ?

En premier lieu, il paraît évident que les modifications de taille d'une population ont un impact fort sur sa diversité génétique : ainsi par exemple, une population subissant une réduction drastique présentera une diversité génétique bien moindre que celle qu'elle avait au départ. Inversement, une population en expansion verra sa diversité génétique augmenter. En outre, la migration, qui consiste le plus souvent en un faible échantillon de population se déplaçant dans un nouvel environnement géographique, peut aussi influencer sa diversité génétique (Wright 1943; Kimura, Weiss 1964; Bodmer, Cavalli-Sforza 1971). Les migrations assurent en fait un brassage des populations, en confrontant des allèles à de nouveaux contextes. Et lorsque les migrants participent à la diversité génétique de la population « receveuse » à la génération suivante, on parlera de flux génique (Wright 1943; Kimura, Weiss 1964; Bodmer, Cavalli-Sforza 1971; Cavalli-Sforza, Feldman 2003) : celui-ci aura tendance à augmenter la diversité génétique de la population « receveuse » uniquement.

Voyons maintenant où en sont nos connaissances en ce qui concerne ces forces chez l'homme. L'archéologie et l'anthropologie nous ont apporté de précieuses informations concernant l'histoire démographique de l'espèce humaine : en 2001, Michel Brunet et son équipe découvrent au Tchad plusieurs fossiles dont un crâne presque complet datant d'environ 7 millions d'années, qu'ils appelleront Toumaï, définissant une nouvelle espèce, *Sahelanthropus tchadensis*, aux caractéristiques morphologiques considérées comme plus proches de l'humain que des grands singes : ce fossile est aujourd'hui admis par de nombreux paléoanthropologues comme le plus vieux de la lignée humaine, se rapprochant ainsi de l'ancêtre commun de l'homme et du chimpanzé (**Figure 2, gauche**) (Brunet et al. 2002; Lebatard et al. 2008). Quant à l'homme moderne (*Homo sapiens*), le fossile le plus ancien, découvert en 1967 en Ethiopie, est daté d'approximativement 200 000 ans (McDougall, Brown, Fleagle 2005) (**Figure 2, droite**); les datations génétiques confirment que notre ancêtre commun le plus récent vivait il y a 150 000 à 200 000 ans en Afrique (Chen et al. 1995; Ingman et al. 2000; Santos-Lopes et al. 2007).

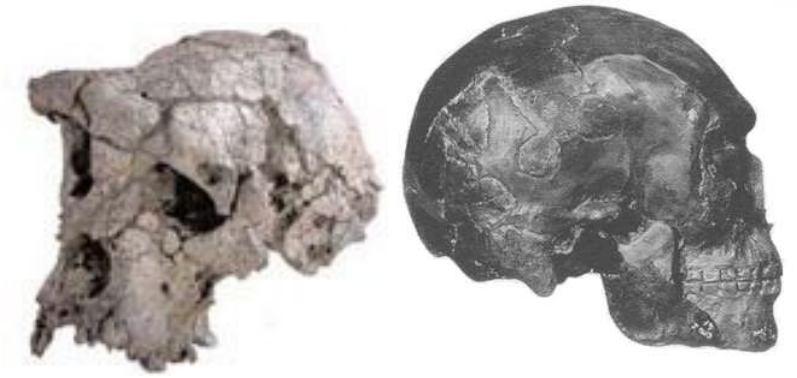
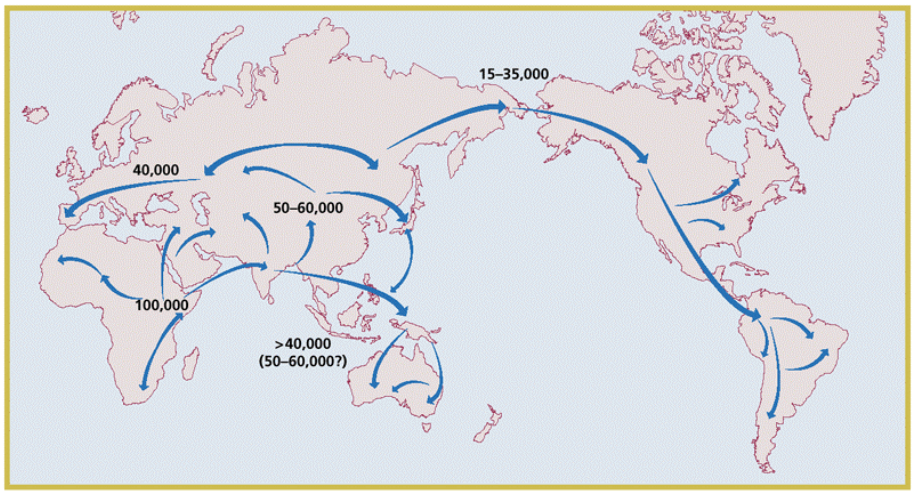


Figure 2. Reconstitution des fossiles crâniens de Toumaï à gauche (Brunet et al. 2002) **et d'Omo I à droite** (McDougall, Brown, Fleagle 2005) (à noter que la proportion des tailles n'est pas respectée).

De même, les datations génétiques des premières dispersions humaines hors d'Afrique (Quintana-Murci et al. 1999; Macaulay et al. 2005; Fagundes et al. 2007) confirment celles des fossiles les plus anciens d'homme moderne retrouvés en Asie et en Europe (entre 50 000 et 75 000 ans) (Mellars 2006). Enfin, diverses analyses de génétique des populations dans des régions neutres ou supposées neutres du chromosome Y, de l'ADN mitochondrial, des autosomes ou encore des insertions *Alu* nous montrent que les populations non-africaines possèdent une diversité génétique plus faible que les populations africaines ; de plus, elles indiquent que la majeure partie de la variabilité trouvée hors d'Afrique est un sous-ensemble de celle observée en Afrique (Wallace, Brown, Lott 1999; Ingman et al. 2000; Underhill et al.

2000; Hammer et al. 2001; Stephens et al. 2001; Watkins et al. 2001; Jobling, Tyler-Smith 2003; Mishmar et al. 2003; Akey et al. 2004; Pakendorf, Stoneking 2005; Underhill, Kivisild



2007). Ces observations sont à l'origine du modèle « *Out-of-Africa* » (Lewin 1987), proposant l'existence d'un groupe fondateur d'humains ayant quitté l'Afrique de l'Est il y a environ 100 000 ans pour coloniser l'Asie, l'Océanie, l'Europe et l'Amérique aux cours de différents événements de migration successifs (**Figure 3**).

Figure 3. Schéma obtenu à partir d'informations génétiques représentant la migration de l'espèce humaine. Extrait de (Cavalli-Sforza, Feldman 2003)

3) Un facteur aléatoire : la dérive génétique

Lors de la fécondation, seuls quelques-uns des nombreux gamètes créés vont être transmis. En dehors de toute pression de sélection, le hasard des croisements des gamètes mâles et femelles au sein d'une population conduit à ce que l'on appelle la dérive génétique. Ce phénomène purement aléatoire va modifier au cours des générations les fréquences alléliques existant au sein d'une population (Wright 1931; Nei 1987). Et ainsi, lorsqu'un allèle neutre (pas d'avantage en matière de survie ou de reproduction) apparaît, la dérive génétique va déterminer si son destin est de disparaître, de se fixer, ou de fluctuer sans jamais se fixer ni disparaître. La probabilité de fixation d'un allèle neutre dans une population de taille finie étant égale à la fréquence de cet allèle dans la population de départ, un nouvel allèle n'aura donc que peu de chances de se fixer (**Figure 4**).

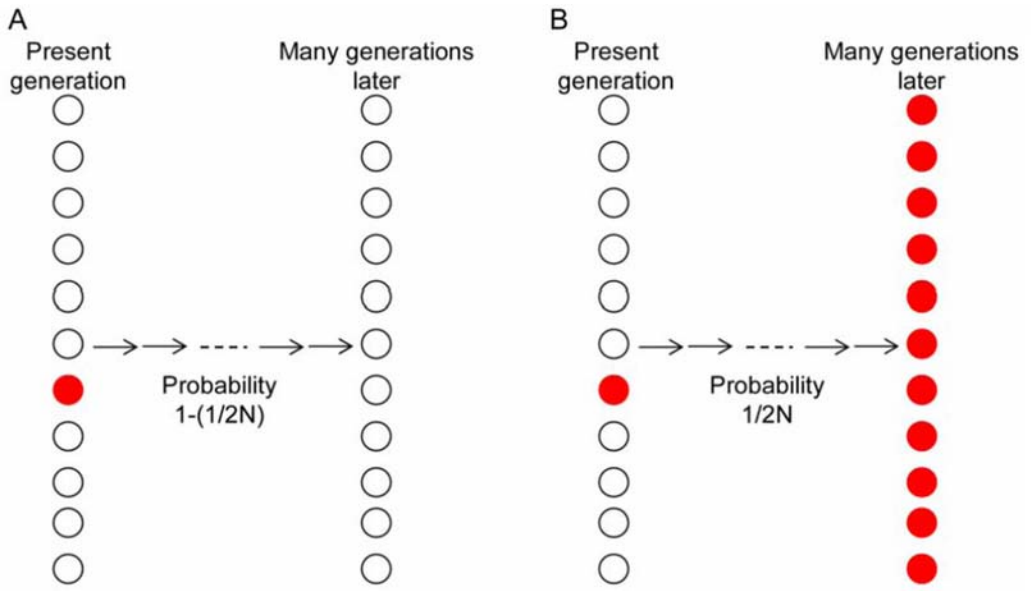


Figure 4. La dérive génétique. Ici, les ronds rouges représentent un allèle nouvellement apparu dans une population diploïde de taille efficace N. **A) La probabilité de disparition de ce nouvel allèle est de 1-(1/2N).** Pour des populations de grande taille, la probabilité est très élevée. **B) La probabilité de fixation de ce nouvel allèle est de 1/2N.** Ainsi, la probabilité de fixation dans une population d'un nouvel allèle est extrêmement faible dans de grandes populations. Adapté de *(Hartl, Clark 2007).*

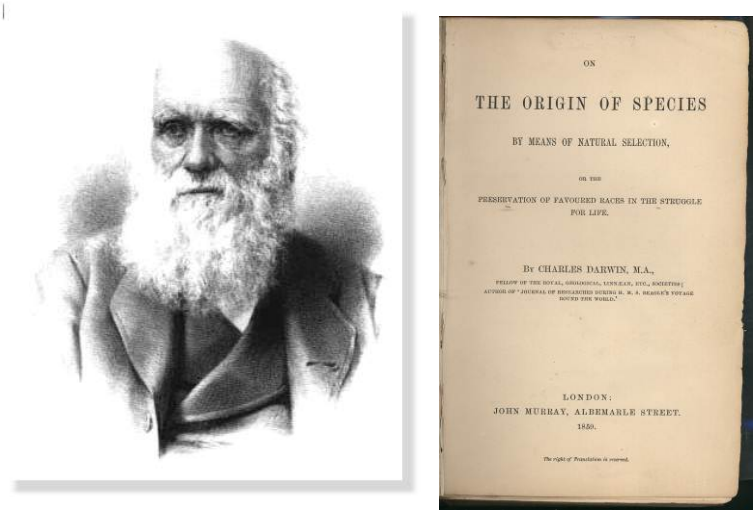
Comme décrit dans la figure 4, la dérive génétique dépend fortement de la taille efficace de la population, c'est-à-dire du nombre d'individus pouvant se reproduire dans une population dont les rencontres sont aléatoires (Wright 1931; Nei 1987; Clark et al. 2007). Plus l'effectif d'une population est restreint, plus un allèle peut disparaître rapidement (au profit d'un autre allèle qui, réciproquement, se fixera rapidement). La dérive génétique aura donc un impact considérable dans des populations de faible taille, réduisant leur diversité génétique. En particulier, comme nous l'avons introduit dans le paragraphe consacré à la démographie, toutes les populations non-africaines semblent avoir subi ce que l'on appelle un « goulot d'étranglement » : leur diversité génétique a été réduite drastiquement, puis elles ont subi une réexpansion à partir de cette faible diversité. En effet, lorsque dans une population originelle, un faible nombre d'individus migre dans un nouvel environnement, la « nouvelle » population n'emporte avec elle qu'une faible part de ses variants : cela réduit donc considérablement sa diversité génétique initiale, c'est l'effet fondateur. En quittant l'Afrique, les populations humaines migrantes ont ainsi affronté deux « goulots d'étranglement » majeurs, l'un entre 50 000 et 60 000 ans de l'Afrique au Moyen-Orient, l'autre en traversant le détroit de Béring pour coloniser l'Amérique. L'hypothèse est que seule une quantité réduite d'individus a pu franchir les obstacles naturels (montagnes, mers) et résister aux différences de climats. Et

ainsi le faible nombre de reproducteurs restants n'ayant transmis aux générations suivantes qu'une partie de la diversité génétique initiale, c'est ce qui pourrait avoir contribué à l'appauvrissement génétique des populations hors de l'Afrique (Amos, Hoffman 2010).

La dérive constitue donc un processus aléatoire. La sélection naturelle, dont je vais maintenant aborder les principes généraux, cible quant à elle les variants particuliers qui vont avoir un impact en matière de survie ou de reproduction.

4) Adaptation et sélection naturelle

a) *« On the origin of species »*, C. Darwin, 1859


ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

Figure 5. Charles Darwin et *« On the origin of species »*, 1859

En 1859, dans « *On the origin of species* », sans conteste le plus connu de ses ouvrages, Charles Darwin (1809-1882) décrit les observations des nombreuses espèces vivantes qu'il a étudiées en tant que naturaliste et propose un mécanisme expliquant la transformation et la diversification adaptative des espèces dans leur milieu (**Figure 5**). C'est la première fois qu'apparaît le principe qui sera défini ensuite comme la sélection naturelle. Il s'agit du processus par lequel les individus les plus aptes à survivre et à se reproduire dans un environnement donné vont être favorisés. Les exemples d'adaptation présents dans ce livre sont bien trop nombreux pour être abordés ici. Sa réflexion s'appuya tout d'abord sur le mécanisme de sélection artificielle, pratiquée par les créateurs de nouvelles races d'animaux domestiques ou de plantes cultivées. Puis, la lecture du livre de l'économiste Thomas Robert Malthus (1766-1834) "*Essai sur la population*" paru en 1798, l'inspira dans l'application de

ces principes à la sélection naturelle. D'après T.R. Malthus, plusieurs causes limitaient la prolifération des plantes et des animaux, "*le défaut de place et de nourriture*" ou encore le fait que "*les animaux sont réciproquement la proie des uns des autres*". Ces remarques confirmaient les intuitions de Darwin : la sélection naturelle était pour lui le résultat de la "*lutte pour l'existence*" (ces termes étant employés au sens large, en incluant l'ensemble des rapports écologiques qui participent à l'équilibre naturel).

En particulier, la visite de l'archipel des Galapagos fut pour lui l'occasion d'assister en grandeur nature au processus de l'évolution :

« *Cet archipel avec ses innombrables cratères et ses ruisseaux de lave dénudée, paraît être d'origine récente ; et je me figurais presque d'assister à l'acte même de la création* »

C. Darwin, *L'origine des espèces*, 1859

Je citerai notamment l'exemple des pinsons des Galapagos (aussi appelés pinsons de Darwin), qui illustre de manière évidente le principe de sélection : 14 espèces de pinsons ont évolué à partir d'une seule espèce originelle ; en colonisant les différentes îles, elles ont dû s'adapter à des conditions d'habitat et des possibilités alimentaires très différentes, acquérant des becs de taille et de forme très variées qui ont pu être directement corrélées à leur nouveaux environnements (**Figure 6**).

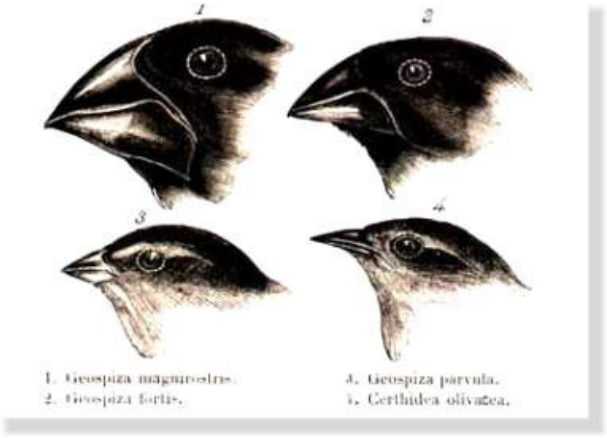


Figure 6. Dessin mettant en évidence les caractéristiques variées de plusieurs espèces de pinsons des Galapagos, sous l'effet de la sélection naturelle. Extrait de *« On the origin of species »*, C. Darwin, 1859.

Et c’est en 1863 que H. W. Bates apporte la première confirmation observationnelle de la théorie évolutionniste : il traite de l’évolution d’une espèce de papillon amazonien qui a adopté une couleur semblable à une autre espèce voisine que les oiseaux prédateurs ne mangent pas. La théorie de la sélection naturelle constituera donc la base de nos connaissances actuelles de l’évolution. Cependant, les mécanismes de transmission des caractères restent inexpliqués pour C. Darwin. Les travaux de son contemporain, le botaniste Johann Gregor Mendel (1822-1884), qui permettent de définir la manière dont les gènes sont transmis de génération, auraient pu répondre à ces questions. Ce n'est qu’au XXème siècle, avec la redécouverte des lois de J. G. Mendel (Hugo de Vries, Carl Erich Correns et Erich von Tschermak) et grâce aux progrès de la génétique, que le darwinisme devient véritablement une théorie de l'évolution : celle-ci prend en compte les mécanismes de l'hérédité de la génétique et met en évidence la possibilité de mutations ou de «sauts» en plus de la sélection des caractères héréditaires. Les individus les plus adaptés à un milieu donné sont favorisés et ce sont leurs variants génétiques qui vont augmenter en fréquence dans la population ; les variants désavantageux vont au contraire avoir tendance à disparaître.

b) Sélection naturelle chez l’homme

Comme les autres espèces, l’homme a dû, au cours de son histoire, faire face à de nombreux changements d’environnement, notamment lors de sa colonisation du monde. Il a en effet dû s’adapter à différents climats, différentes sources d’alimentation et également à différents pathogènes. On peut citer le cas de la persistance de la lactase, qui constitue probablement l’un des exemples les plus connus de sélection naturelle chez l’homme : la lactase, enzyme permettant la digestion du lactose (et donc du lait), est fortement produite pendant les premières semaines suivant la naissance ; plus tard, le gène codant l’enzyme étant de plus en plus réprimé, la production de lactase (et donc la digestion du lait) diminue jusqu’à atteindre un taux résiduel de 5 à 10% à l’âge adulte. Cependant, dans certaines populations, des mutations de ce gène permettent la persistance de cette enzyme (et donc la digestion de lait) à l’âge adulte : or, il a été montré que, sous l’effet de la sélection naturelle, ces mutations ont augmenté en fréquence dans des populations dont l’agriculture a été un mode de subsistance important (en Afrique et Europe) (Tishkoff et al. 2007; Enattah et al. 2008; Itan et al. 2010). Ceci explique en partie la forte proportion d’individus digérant le lait dans ces régions géographiques (**Figure 7**).

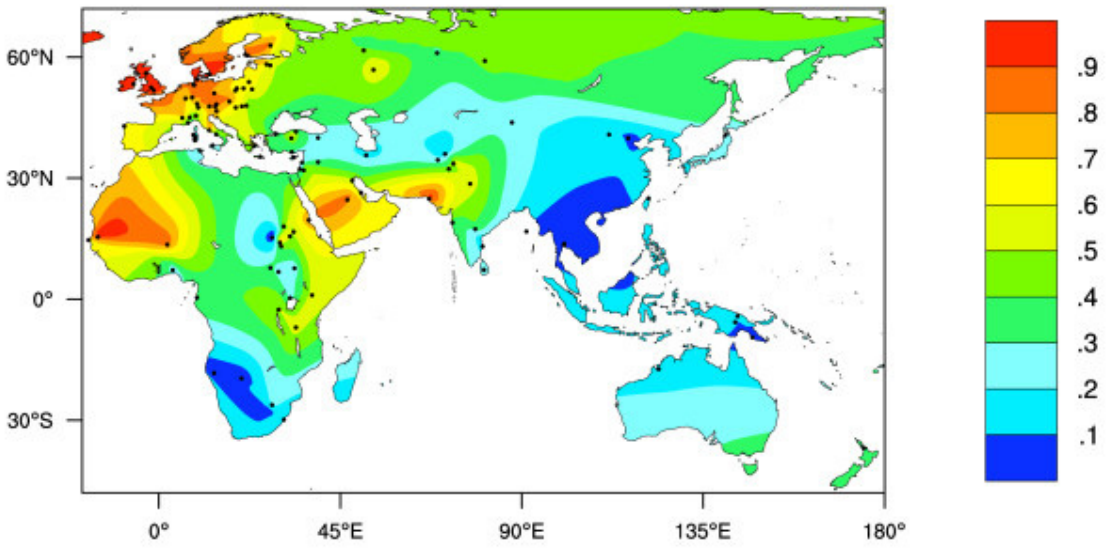


Figure 7. Carte des fréquences phénotypiques de la persistance de la lactase. Les points représentent les lieux d'échantillonnage. Les couleurs montrent les fréquences du phénotype de persistance de la lactase par interpolation de surface. Extrait de (Itan et al. 2010).

Ceci constitue un cas de sélection naturelle particulièrement fort. Cependant, il existe de nombreux autres exemples d'adaptation de l'homme, notamment face à des environnements pathogéniques variés (Barreiro, Quintana-Murci 2010). Certains d'entre eux sont cités dans la partie suivante, dans laquelle je précise les différentes formes de sélection.

Ainsi, si l'on résume l'ensemble des forces affectant notre variabilité génétique, on a vu que les forces génomiques créent de la diversité en faisant apparaître de nouveaux allèles ; les migrations assurent un brassage des populations confrontant ces allèles à de nouveaux contextes ; la dérive génétique influence les fréquences de ces allèles de manière aléatoire ; et la sélection naturelle affecte quant à elle de façon très ciblée certains variants particulièrement avantageux ou au contraire délétères. Notre diversité résulte donc de l'ensemble de ces facteurs et de leurs interactions. En particulier, si un allèle apparaît et qu'il est particulièrement avantageux ou délétère, la sélection naturelle agira sur lui (et sur la région génomique qui l'entoure) plus rapidement que la dérive génétique ; et dans ce cas, les variations des fréquences alléliques de cette région seront davantage dictées par les pressions de sélection que par le hasard.

III) Les différentes formes de sélection naturelle

La sélection naturelle peut agir de différentes manières. Nous allons voir les trois grands types de sélection, ainsi que leurs effets sur la diversité génétique (**Figure 8**).

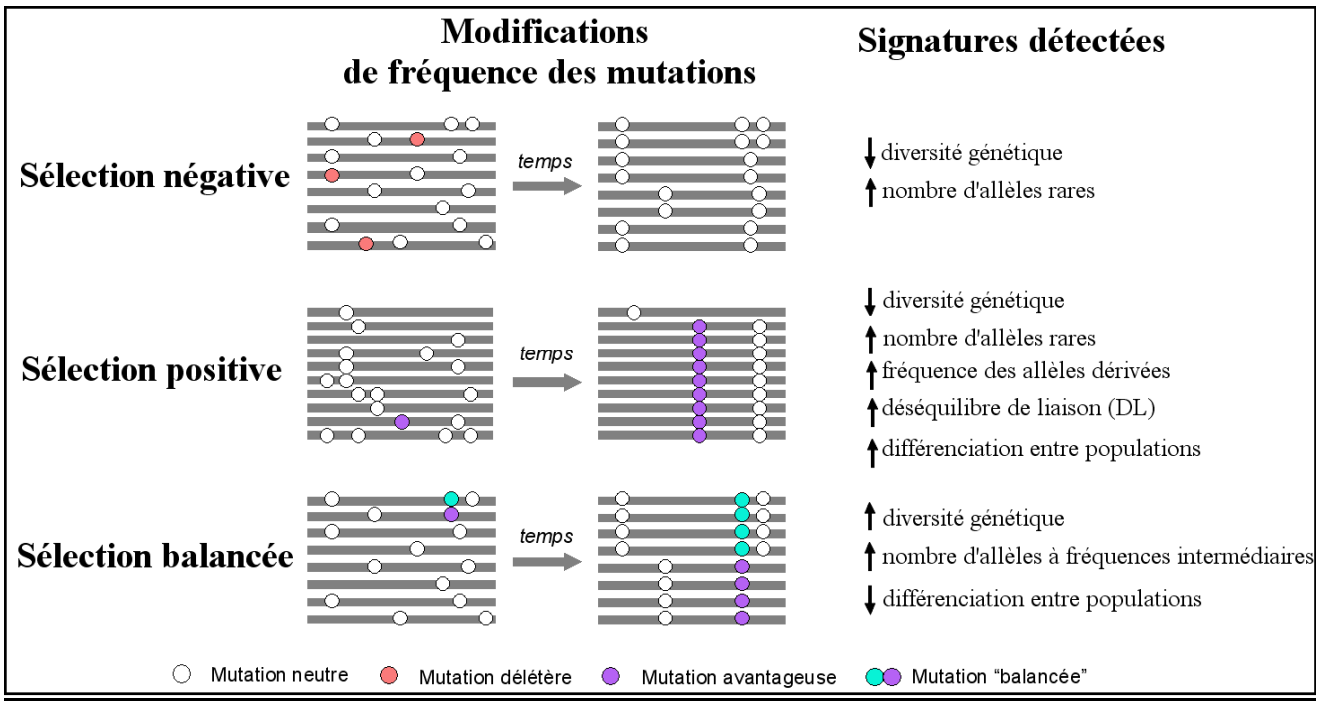


Figure 8. Les différentes formes de sélection naturelle. Pour chaque type de sélection, sont représentés ses effets sur la diversité génétique : on voit la distribution des mutations (ronds de couleur) sur des chromosomes (barres grises) avant et après sélection. Lorsque des mutations délétères apparaissent, la sélection négative va avoir tendance à les éliminer de la population, parce qu'elles sont trop désavantageuses pour les individus qui les portent. La sélection positive va au contraire augmenter la fréquence d'une mutation qui confère un avantage. Enfin, la sélection balancée maintient à fréquences intermédiaires plusieurs allèles différents (ici deux) à un même locus. Il faut noter que le devenir des mutations neutres sera influencé par la dérive génétique uniquement, sauf si elles se trouvent liées génétiquement (en fort déséquilibre de liaison, DL) avec les mutations sélectionnées et que leurs fréquences sont donc modifiées par « entraînement ». Sont précisés à droite les effets correspondants détectables en génétique des populations (le chapitre suivant détaille les tests utilisés pour détecter ces effets).

1) La sélection négative

Tout d'abord, le cas le plus connu et initialement décrit comme celui touchant la plupart des gènes est celui de la sélection négative (Bamshad, Wooding 2003; Bustamante et al. 2005; Nielsen 2005). Si une mutation apparaît et que, dans un environnement donné, elle est désavantageuse pour l'individu qui la porte en matière de survie ou de reproduction, celle-ci va avoir tendance à diminuer en fréquence, voire disparaître totalement de la population. Dans des cas extrêmes, c'est-à-dire lorsque des mutations ont un effet fortement délétère, on parlera de sélection purificatrice : celles-ci seront alors rapidement éliminées de la population

sous l'effet de la sélection naturelle. On s'attend à ce que la sélection exerce des pressions extrêmement fortes sur les gènes remplissant des fonctions essentielles, tels que les gènes de ménage ; mais cela peut concerner bien d'autres gènes impliqués dans le métabolisme ou encore l'immunité. Dans certains cas plus modérés, des mutations désavantageuses peuvent apparaître et perdurer dans une population, mais seulement en se maintenant à de faibles fréquences : on parlera de sélection négative faible. Il est ainsi possible de distinguer les gènes qui se trouvent sous sélection purificatrice de ceux qui sont neutres ou montrent des signes de sélection négative faible. On peut alors avoir une meilleure idée des gènes qui remplissent *a priori* des fonctions plus importantes que les autres.

2) La sélection positive

D'autre part, la sélection positive va augmenter en fréquence les variants qui confèrent un avantage dans un environnement donné (Sabeti et al. 2006; Nielsen et al. 2007). Ainsi par exemple, différentes mutations permettant une plus grande résistance au paludisme ont été sélectionnées dans des régions où le parasite responsable de cette maladie était très présent ; une mutation affectant la production de la G6PD (*Glucose-6-Phosphate Dehydrogenase*), molécule essentielle dans le métabolisme de glucose, a augmenté très fortement en fréquence dans les régions où le paludisme sévit, car cette mutation confère une plus grande résistance au parasite *Plasmodium falciparum* et à *Plasmodium vivax* en Asie du Sud-Est, provoquant le paludisme (Beutler 1994; Tishkoff et al. 2001; Saunders, Hammer, Nachman 2002; Louicharoen et al. 2009). Une mutation affectant le promoteur du gène *DARC* (*Duffy Antigen Receptor for Chemokines*) fournit un exemple encore plus extrême : celle-ci est responsable de l'absence totale de la protéine dans 90% de la population africaine et dans certaines autres régions touchées par le paludisme ; c'est la protection contre *Plasmodium vivax* qui semble avoir fait augmenter en fréquence cette mutation, qui reste par ailleurs très rare dans les autres régions du monde (Chitnis, Miller 1994; Tournamille et al. 1995; Hamblin, Di Rienzo 2000; Hamblin, Thompson, Di Rienzo 2002; Sabeti et al. 2006). De manière générale, les gènes qui sont sous sélection positive présenteront localement ou tout le long de leur séquence des mutations qui altèrent la protéine et qui peuvent se trouver à de fortes fréquences dans la population étudiée.

3) La sélection balancée

Enfin, on distingue une autre forme de sélection, la sélection balancée, qui consiste à maintenir dans une population à un même locus plusieurs variants à fréquence intermédiaire. Divers cas peuvent se présenter. L'un des exemples les plus connus chez l'homme est le système HLA (pour H*uman Leucocyte Antigen)* qui reconnaît de nombreux antigènes pouvant provenir de pathogènes et les présentent aux lymphocytes T pour déclencher la réponse immunitaire. Ces gènes se caractérisent par un degré de polymorphisme élevé (le plus élevé des gènes humains) maintenu par la sélection balancée. Etant donné la grande diversité des motifs pathogéniques, on comprend bien l'intérêt de conserver une forte diversité dans ce système de reconnaissance (Hughes, Nei 1988; Takahata 1990; Hedrick, Whittam, Parham 1991; Ohta 1991; Takahata, Satta, Klein 1992; Hughes et al. 1994; Hughes, Yeager 1998; Walsh et al. 2003; Akey et al. 2004). Ceci est encore plus évident, lorsque l'on sait que la diversité du système HLA a été corrélée positivement avec l'enrichissement d'une zone géographique en pathogènes (Prugnolle et al. 2005). Un autre cas de sélection balancée résulte cette fois-ci de ce que l'on appelle l'avantage à l'hétérozygote ; en d'autres termes, les individus qui portent deux allèles différents à un même locus (par opposition à homozygote) sont avantagés. Un exemple bien connu de cette situation est celui de l'allèle Hémoglobine S (HbS). Cet allèle est responsable, à l'état homozygote, de la drépanocytose, une maladie génétique au phénotype très sévère, provoquant des anémies (baisse de concentration en hémoglobine dans le sang responsable d'un mauvais transport de l'oxygène). Cependant, cet allèle est trouvé à une fréquence relativement forte, en particulier dans les régions du monde touchées par le paludisme (**Figure 9**). Il a en effet été montré, qu'à l'état hétérozygote, cet allèle permettait de mieux résister aux infections par *Plasmodium falciparum* provoquant le paludisme, tout en ayant peu d'effet sur le transport de l'oxygène par le sang (Allison 1954; Allison 1961; Cooke, Hill 2001; Allison 2004).

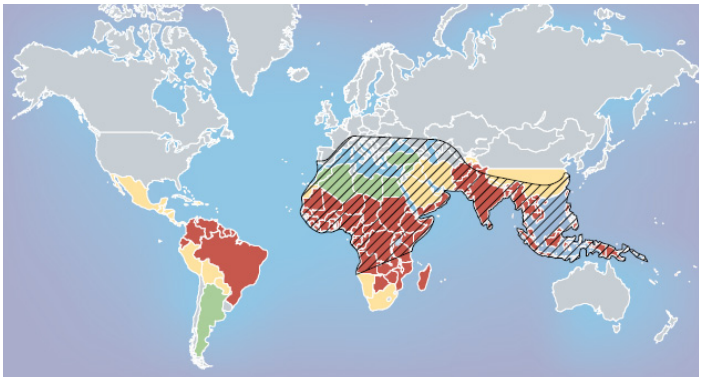


Figure 9. Distributions globales du paludisme et des maladies associées aux globules rouges. En vert, sont indiquées les régions où le paludisme n'est trouvé que dans de rares cas isolés ; la couleur jaune indique les zones où le risque de paludisme est intermédiaire et en rouge, les régions où le risque est élevé. Quant à l'aire hachurée, elle montre la distribution des maladies des globules rouges. Extrait de (Cooke, Hill 2001).

La sélection naturelle peut donc affecter notre diversité de manière différente, favorisant les individus les mieux adaptés, assurant la pérennité des espèces les plus « fortes ».

IV) Comment détecter la sélection naturelle?

Nous avons vu auparavant que la sélection naturelle a un impact fort sur notre diversité. C'est en 1968 que Motoo Kimura a proposé sa théorie de la neutralité : elle suggère que la plupart des polymorphismes observés au niveau moléculaire sont neutres ou quasi-neutres du point de vue de la sélection (Kimura 1968a; Kimura 1968b); c'est donc la dérive génétique qui gouverne de manière totalement aléatoire leurs variations de fréquence. Selon Motoo Kimura, les mutations fortement délétères seraient donc rapidement éliminées, celles-ci ayant peu d'effets sur les niveaux de polymorphisme et de divergence observés. Et dans de rares cas, lorsqu'un allèle apparaît et qu'il est particulièrement avantageux, la sélection naturelle va modifier sa fréquence plus rapidement que ne le fait la dérive génétique au hasard. Cette théorie gouverne aujourd'hui l'ensemble des tests de génétique des populations : la neutralité en matière de sélection constitue « l'hypothèse nulle » qui devra être rejetée lorsque la région testée a été ciblée par la sélection naturelle. Il est ainsi possible de détecter les « empreintes » laissées par cette sélection dans notre génome, en mesurant ces « écarts à la neutralité ».

Il existe pour cela différents types de tests (**Tableau 1**). On distingue d'une part les tests inter-spécifiques, pouvant mettre en évidence un excès ou un défaut de divergence entre deux espèces. Nous avons comparé ici l'homme et le chimpanzé. On peut de cette façon détecter des événements de sélection très anciens, dont certains peuvent avoir participé au phénomène de spéciation. D'autre part, les tests intra-spécifiques considèrent les niveaux de polymorphisme au sein d'une espèce ou d'une population particulière. On peut ainsi mettre en évidence des événements de sélection plus récents, c'est-à-dire dans notre cas qui se sont produits après l'apparition de l'homme moderne, et locaux, *i.e.* qui sont spécifiques d'une population particulière. On peut trouver dans la littérature un grand nombre de tests, ainsi que d'exemples de leur application pour détecter des signatures de sélection chez l'homme (Bamshad, Wooding 2003; Sabeti et al. 2006; Nielsen et al. 2007; Barreiro et al. 2008).

Tableau 1. Caractéristiques des différents tests de génétique des populations humaines utilisés dans cette étude

Aspect des données	*Données nécessaires*	*Evénements détectés*	*Tests statistiques*
Comparaison de polymorphisme et divergence aux sites silencieux et non-synonymes	Séquençage d'un ou plusieurs loci dans une espèce de primate et dans plusieurs individus humains	Sélection diversificatrice, purificatrice, ou pseudogénisation au sein de la lignée humaine	**Test MKPRF**
Spectre des fréquences alléliques dérivées chez l'homme	Séquençage d'un ou plusieurs loci dans une ou plusieurs espèces de primates et dans plusieurs individus humains	Excès d'allèles dérivés fréquents : goulot d'étranglement, sélection positive (balayage sélectif complet ou quasi-complet)	***H* de Fay & Wu significativement négatif**
Spectre des fréquences alléliques chez l'homme	Séquençage d'un ou plusieurs loci dans plusieurs individus humains	Excès d'allèles rares : expansion de population, sélection positive (balayage sélectif complet ou quasi-complet), sélection négative Excès d'allèles fréquents : stratification des populations, goulot d'étranglement, sélection balancée, sélection positive (balayage sélectif partiel)	***D* de Tajima, *F* et *D* de Fu & Li significativement négatifs (allèles rares) ou positifs (allèles à fréquences intermédiaires)**
Comparaison de la fréquence d'un allèle entre populations humaines	Génotypage ou séquençage de plusieurs populations humaines	Niveaux d'isolement et de flux génique entre populations ; si excès de différenciation à un SNP, sélection positive locale de ce SNP dans une des populations	**F_{ST} comparé à une distribution empirique**
Comparaison de la diversité interne et de la fréquence d'un allèle/haplotype	Séquençage ou génotypage d'une région recombinante autour d'un locus donné	Déficit de diversité interne et fréquence élevée : goulot d'étranglement, sélection positive récente (balayages sélectifs incomplets et complets)	**Tests LRH, iHS, XP-EHH**
Comparaison de la diversité interne d'un allèle dérivé par rapport à l'allèle ancestral en fonction de sa fréquence	Séquençage d'un locus dans plusieurs individus humains	Diversité de l'allèle dérivé faible et fréquence élevée : sélection positive récente (balayages sélectifs incomplets)	**Test DIND**

1) Les tests inter-spécifiques

De manière générale, l'utilisation de séquences d'ADN orthologues (homologues entre espèces) peut s'avérer utile pour inférer une fonction biologique connue dans une espèce à une autre où cette fonction reste à déterminer; c'est la génomique comparative. Le principe est que les régions qui ont une importance fonctionnelle vont avoir tendance à être très conservées entre espèces génétiquement proches, alors que les autres vont muter davantage. Ainsi, à l'aide de cette technique, on peut mettre en lumière une importance biologique jusqu'alors insoupçonnée de certaines régions codantes ou non-codantes d'une espèce d'intérêt. Dans notre cas, nous souhaitons savoir si l'espèce humaine a accumulé plus ou moins de mutations fonctionnelles que ce qui est attendu, en comparant sa divergence avec le chimpanzé. Les ratios de divergences non-synonyme et synonyme (dN/dS) ont notamment été estimés. Néanmoins, ils ne permettent pas d'avoir une quelconque preuve de significativité. Le test McDonald-Kreitman (MK) (McDonald, Kreitman 1991), très utilisé aujourd'hui, donne une précision du niveau de significativité obtenu : de manière intéressante, il prend en compte le niveau de polymorphisme de l'espèce testée, ainsi que le niveau de divergence avec une autre espèce proche, et ce en séparant les sites non-synonymes et synonymes ou non-synonymes et silencieux (non-codants + synonymes). On obtient alors un tableau de contingence, indiquant les nombres de polymorphismes au sein d'une espèce et les différences fixées entre espèces. La non-indépendance entre les lignes et les colonnes est évaluée en utilisant le test exact de Fisher (**Figure 10 et Tableau 1**).

A

	Test MK	
	Synonymes	Nonsynonymes
Polymorphisme	20	30
Divergence	10	15

Le locus est neutre

$20/10 = 30/15$

B

	Test MK	
	Synonymes	Nonsynonymes
Polymorphisme	20	50
Divergence	10	15

Le locus n'est pas neutre

$20/10 \neq 50/15$

Figure 10. Exemple de tableaux de contingence du test McDonaldKreitman. (A) Sous neutralité, le ratio entre polymorphismes et divergences est identique pour les SNPs synonymes et non-synonymes. (B) Le locus présente un excès de polymorphismes non-synonymes par rapport à la divergence. Les résultats de ce deuxième cas pourraient résulter de sélection négative faible.

L'un des principaux avantages de ce test est qu'il est peu sensible à la démographie car si les variants synonymes et non-synonymes sont affectés par elle, ils le sont de la même

manière. Une *P-value* significativement inférieure au seuil considéré (0,05 ici) constituera une signature de sélection naturelle. Néanmoins, il est important de préciser que ce test considère qu'on a de la sélection si on obtient un niveau inattendu de mutations non-synonymes, les mutations synonymes étant considérées comme neutres. On peut donc se poser la question de la validité de ce test, dans la mesure où certaines mutations non-synonymes n'ont aucun effet fonctionnel et inversement, des mutations non-codantes peuvent avoir un impact phénotypique important. Toutefois, c'est l'excès ou le défaut de l'ensemble de la classe des non-synonymes d'une région génomique donnée qu'on observe ; et on peut donc s'attendre à ce que les signatures trouvées reflètent bien le comportement de la majeure partie des mutations. Ainsi par exemple, un excès de différences non-synonymes fixées (comparées aux silencieuses fixées) résulte vraisemblablement d'un effet de sélection adaptative : les mutations non-synonymes avantageuses se sont fixées rapidement dans l'espèce considérée. Et dans le cas où l'on observe un déficit en mutations non-synonymes, à la fois en matière de polymorphisme et de divergence, on peut suspecter la présence de sélection purificatrice, l'accumulation de mutations non-synonymes en général n'étant pas tolérée. Enfin, un dernier exemple pourrait être le cas d'un excès de mutations polymorphes non-synonymes par rapport aux mutations polymorphes silencieuses ; dans ce cas, on peut l'interpréter par de la sélection positive ou négative faible. Si les fréquences des mutations non-synonymes dans la population restent faibles, on conclura davantage en faveur de sélection négative faible, ces mutations étant vraisemblablement maintenues à faibles fréquences parce qu'elles ont un effet délétère non négligeable ; et si on contraire, les mutations non-synonymes se trouvent à des fréquences élevées, on privilégiera l'hypothèse de sélection positive.

Pour ce projet de thèse, j'ai utilisé une extension du test de McDonald-Kreitman (McDonald, Kreitman 1991), la méthode MKPRF (McDonald and Kreitman Poisson Random Field) (Sawyer, Hartl 1992; Bustamante et al. 2002; Bustamante et al. 2005), qui avait déjà été utilisée sur approximativement 11 000 gènes. Ce test est basé sur l'estimation de deux paramètres : ω compare les niveaux de divergence et polymorphisme à des sites non-synonymes et synonymes ; γ permet d'estimer le logarithme du ratio comparant divergence et polymorphisme aux sites non-synonymes. Ainsi, un ω significativement inférieur à 1 montrera un déficit en mutations non-synonymes (à la fois en matière de divergence et de polymorphisme), suggérant de la sélection purificatrice contraignant très fortement la région étudiée. Un γ négatif mettra en évidence un déficit de divergence (comparée au polymorphisme) aux sites non-synonymes ; celui-ci peut notamment traduire un effet de

sélection négative faible : les mutations désavantageuses étant maintenues à faibles fréquences, on observe un polymorphisme non-négligeable au sein des populations humaines, mais qui n'est jamais fixé chez l'homme et réduit donc la divergence avec le chimpanzé. Cependant, comme pour le test simple de McDonald Kreitman, d'autres explications sont possibles dans ce cas : la diversification d'une population humaine (sélection positive locale) ou la sélection balancée peuvent également conduire à un déficit de divergence aux sites non-synonymes, comparé au polymorphisme ; c'est la distribution des fréquences des mutations de la région étudiée qui pourra permettre de conclure parmi les différentes interprétations possibles (une forte proportion d'allèles à fréquences intermédiaires permettra par exemple de conclure en faveur de sélection balancée).

2) Les tests intra-spécifiques

On peut également utiliser des tests intra-spécifiques, c'est-à-dire des tests basés sur le polymorphisme observé au sein des populations humaines. Ils permettent donc de détecter de la sélection plus récente que les tests inter-spécifiques.

a) Les écarts au spectre de fréquences alléliques

L'un des principaux moyens de détecter de la sélection dans un gène consiste à analyser la distribution de ses fréquences alléliques (**Figure 11 et Tableau 1**). Les différentes formes de sélection modifieront le spectre des fréquences alléliques de manière variée, par rapport à celui attendu sous neutralité (Bamshad, Wooding 2003; Nielsen 2005; Sabeti et al. 2006; Nielsen et al. 2007). Différents tests de neutralité permettent de mesurer les écarts du spectre de fréquences alléliques observés par rapport à la neutralité (**Annexe 1**). Nous avons utilisé le *D* de Tajima, les *D** et *F** de Fu et Li, le *H* de Fay et Wu. Toutes ces statistiques seront proches de 0 sous neutralité. Bien que l'ensemble de ces tests soit basé sur le même principe, ils vont se focaliser sur des parties différentes du spectre des fréquences alléliques. Le *D* de Tajima (Tajima 1989) prend en compte l'ensemble du spectre : il mesure la différence entre la moyenne du nombre de différences observées entre paires de séquences (θ_π) et le nombre total de sites polymorphes observés (θ_W) ; sous neutralité, ces deux valeurs sont proches. Les *D* et *F* de Fu & Li (Fu, Li 1993) mesurent quant à eux l'excès ou le déficit de mutations présentes une seule fois dans l'échantillon (singletons), en comparant le nombre de singletons aux valeurs de θ_W et θ_π, respectivement. Si la région testée se trouve sous sélection positive ou négative, le *D* de Tajima et les *D* et *F* de Fu & Li seront négatifs,

indiquant un excès d'allèles rares. Si elle est sous sélection balancée, ces statistiques seront positives, révélant un excès d'allèles à fréquences intermédiaires.

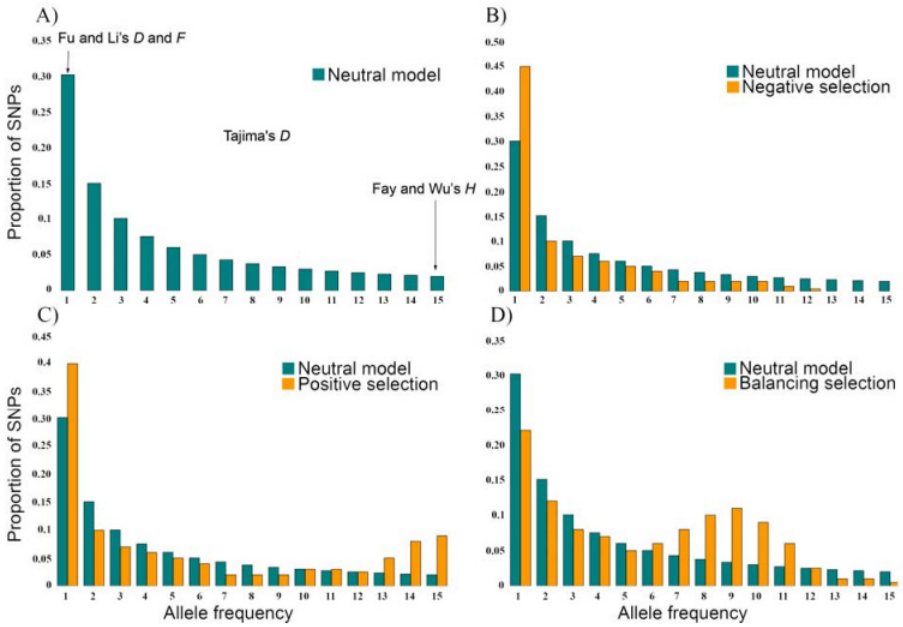


Figure 11. Exemples de formes que peuvent adopter le spectre des fréquences alléliques. Les parties du spectre qui sont évaluées par les différents tests de neutralité sont indiquées. Le *D* de Tajima considère la totalité du spectre. Les *D* et *F* de Fu et Li se concentrent sur les singletons (allèles retrouvés une seule fois dans l'échantillon) à l'état dérivé. Le *H* de Fay et Wu détecte l'excès d'allèles dérivés à fortes fréquences. (A) Neutralité. (B) Sélection négative. On observe un excès d'allèles rares. (C) Sélection positive. On observe à la fois un excès d'allèles rares et à fortes fréquences. (D) Sélection balancée. On observe un excès d'allèles à fréquences intermédiaires.

Enfin, le *H* de Fay & Wu (Fay, Wu 2000) indique un excès d'allèles dérivées à forte fréquence, caractéristique de sélection positive. L'allèle dérivé est défini par opposition à l'allèle ancestral porté par l'ancêtre commun à l'homme et au chimpanzé ; en pratique, il est déterminé par parcimonie à partir d'espèces proches comme le chimpanzé, le gorille, l'orang-outan et le macaque rhésus. Et ainsi, sous sélection positive, tous les allèles dérivés en déséquilibre de liaison avec la mutation avantageuse subiront l'augmentation en fréquence de cette dernière (balayage sélectif), provoquant un excès général d'allèles dérivés à fortes fréquences.

b) <u>Les indices de différenciation de populations</u>

On peut également détecter de la sélection positive relativement moins ancienne en analysant les variations de fréquences alléliques de SNPs entre populations, c'est-à-dire la différenciation de populations, que l'on peut quantifier par la statistique F_{ST} (Cavalli-Sforza 1966; Lewontin, Krakauer 1973; Excoffier, Smouse, Quattro 1992; Weir, Hill 2002). Les

valeurs de cette statistique sont comprises entre 0 (aucune différenciation) et 1 (différenciation totale). En l'absence de sélection, la valeur de F_{ST} ne sera influencée que par la dérive génétique (et bien sûr aussi par la démographie) et ce de manière homogène sur l'ensemble du génome. En revanche, la sélection va cibler spécifiquement un locus donné et peut créer des différences génétiques très importantes entre populations. Ainsi, une population dont une région génomique particulière est sous sélection positive va accumuler des mutations et combinaisons de mutations (ou haplotypes) qui vont lui être propres et la différencier des autres populations : on observera alors des valeurs de F_{ST} fortes entre la population sous sélection et les autres populations étudiées. Cette statistique pouvant être évaluée pour tout un gène, mais aussi SNP par SNP, elle permet potentiellement d'identifier le variant particulier sous sélection positive. Au contraire, des gènes sous sélection balancée, négative ou positive convergente (cas dans lequel une même mutation est sélectionnée dans les différentes régions du monde, car elle confère un avantage dans l'ensemble de ces régions) montreront des F_{ST} réduites (**Tableau 1**) (Cavalli-Sforza 1966; Bamshad, Wooding 2003; Barreiro et al. 2008).

c) Le déséquilibre de liaison

Enfin, pour identifier des signatures de sélection positive encore plus récente, on utilise les tests de déséquilibre de liaison, basés sur l'association génétique entre variants. Tous ces tests s'appuient sur le même principe (**Figure 12 et Tableau 1**) : lorsqu'une mutation apparaît, elle possède une forte association génétique avec les régions voisines, qui va diminuer au cours du temps, du fait de la recombinaison génétique qui a tendance à « casser » ces blocs d'association. Même si la mutation augmente en fréquence par hasard (du fait de la dérive génétique), on observera une mutation à forte fréquence, mais montrant une association génétique faible avec ses régions voisines. Bien au contraire, si une mutation est sélectionnée positivement, elle va augmenter en fréquence dans la population à une telle vitesse que la recombinaison n'aura pas le temps de briser totalement le « déséquilibre de liaison » ; on aura donc une mutation à forte fréquence associée à un DL fort, ce qui constitue un signe de sélection positive. Cette signature aura cependant tendance à s'atténuer au cours du temps, c'est pourquoi ces tests ne restent puissants que pour détecter des événements récents de sélection. Les EHH, XP-EHH et iHS (integrated Haplotype Scores) (Sabeti et al. 2002; Sabeti et al. 2006; Voight et al. 2006; Sabeti et al. 2007) mesurent la diminution d'homozygotie d'haplotypes depuis la mutation d'intérêt en s'éloignant latéralement le long du chromosome. Le test DIND (Derived Intra-allelic Nucleotide Diversity) mesure le ratio des diversités associées aux allèles ancestral et dérivé (π_A/π_D) (Barreiro et al. 2009). Il permet

ainsi de mettre en évidence, pour une fréquence donnée, une diminution de diversité associée à l'allèle dérivé comparée à celle de l'allèle ancestral. En effet, la sélection positive va avoir un effet non pas sur la mutation uniquement, mais sur l'ensemble de la région qui la contient ; et ainsi, si une mutation est sélectionnée, c'est l'ensemble de sa région qui est conservée et la variation à son voisinage est donc largement réduite par rapport à ce qu'elle aurait été sans sélection (valeur estimée par la diversité associée à l'allèle ancestral).

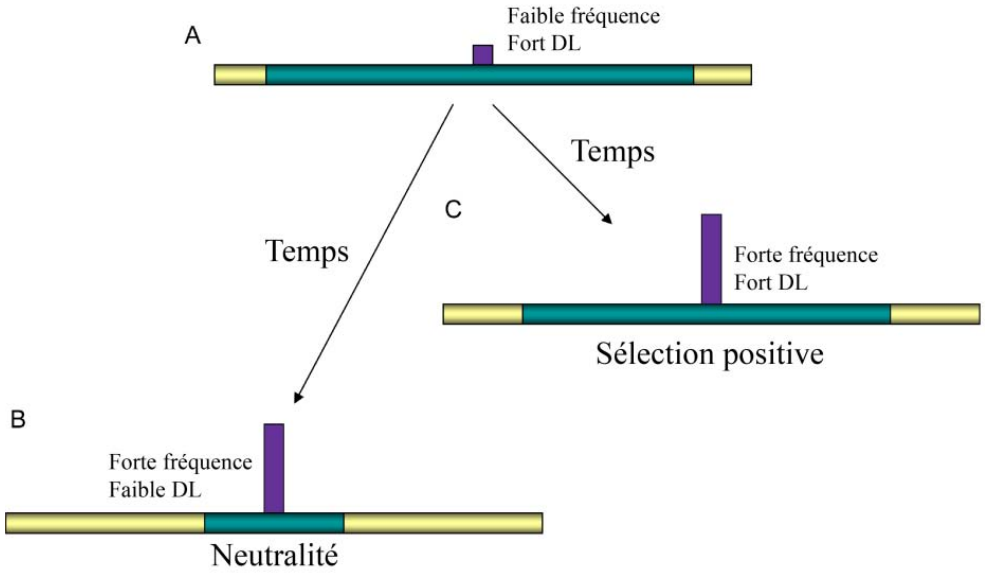


Figure 12. Tests basés sur le déséquilibre de liaison (DL). (A) Un nouvel allèle (barre violette) se trouve à faible fréquence (représentée par la longueur de la barre violette) sur un haplotype donné (barre jaune) caractérisé par un long DL (barre verte) entre l'allèle d'intérêt et les autres allèles liés. (B) Sous neutralité, l'allèle d'intérêt peut augmenter en fréquence sous l'effet de la dérive génétique, mais assez lentement pour que la recombinaison ait le temps de casser le DL. On aura donc une mutation à forte fréquence associée à un faible DL. (C) Si l'allèle d'intérêt est sélectionné, il va augmenter en fréquence à une vitesse telle que la recombinaison n'aura pas le temps de casser le DL. On aura donc une mutation à forte fréquence associée à un fort DL. Adapté de (Bamshad, Wooding 2003).

3) Corriger les effets démographiques mimant la sélection naturelle

Comme précisé auparavant, la démographie joue également un rôle majeur dans notre diversité génétique. Celle-ci peut donc mimer les effets de la sélection ; ainsi par exemple, un excès d'allèles rares peut tout autant résulter d'une expansion de population que de sélection positive. C'est pourquoi, nous avons pris en compte ces effets pour les tests dits « sensibles à la démographie ». Or, la principale distinction entre un évènement démographique et la sélection naturelle réside dans le fait que la sélection naturelle va cibler un ou plusieurs loci particuliers alors que la démographie va affecter l'ensemble du génome. Ainsi, on peut différencier ces deux facteurs en comparant la région potentiellement sous sélection avec le

profil global de la variabilité génétique estimée à partir de régions supposées neutres (non-codantes) réparties dans le génome et séquencées sur les mêmes individus (Bamshad, Wooding 2003; Voight et al. 2006; Barreiro et al. 2009; Quach et al. 2009). Et ainsi, la part de diversité observée dans les régions neutres pourra être imputée à la démographie. Dans le cas des F_{ST} par exemple, on compare les valeurs des SNPs étudiés à la distribution des 659 000 SNPs génotypés dans des régions considérées comme neutres et dans le même panel d'individus (Li et al. 2008). Pour les autres tests sensibles à la démographie, on utilise des simulations qui prennent en compte les données admises concernant l'histoire démographique humaine : on peut ainsi redéfinir une « nouvelle neutralité », *i.e.* une nouvelle hypothèse nulle ; les écarts à cette hypothèse nulle pourront ainsi être imputés à la sélection naturelle uniquement. Les modèles démographiques utilisés (Voight et al. 2005; Laval et al. 2010) considèrent tous les deux une expansion de population en Afrique et un goulot d'étranglement dans les populations eurasiennes (c'est-à-dire une diminution rapide de taille de population, due à la sortie de petits groupes d'individus hors d'Afrique, suivie d'une réexpansion à partir de ces populations de taille réduite). Il est à noter que le modèle de Laval prend également en compte les migrations au sein de ces différentes populations. En utilisant parallèlement ces deux modèles, on peut ainsi avoir une idée plus juste de la significativité de nos résultats, étant donnée l'absence de certitudes quant au modèle démographique réel. Ainsi pour chacun de ces tests et pour chaque modèle, une valeur de significativité (*P-value*) sera attribuée aux différences observées. On considèrera comme significatifs les écarts à la neutralité montrant des *P-values* inférieures à 0.05 (seuil d'erreur à 5%).

Tous ces tests permettent donc de détecter différents types de sélection naturelle et ce à différentes échelles de temps. La nature des pressions de sélection exercées, nous l'avons vu, est très variable. Nous nous intéressons en particulier dans ce projet aux signatures de sélection chez l'homme résultant de son adaptation à divers environnements pathogéniques. C'est pourquoi dans la partie suivante, je vais aborder les différents mécanismes développés par l'hôte pour lutter contre les pathogènes.

V) Maladies infectieuses et mécanismes de défense

Depuis toujours, les espèces vivantes ont dû développer des mécanismes de défense contre les pathogènes ; depuis les microbes eux-mêmes, jusqu'aux espèces végétales et animales, tous possèdent un système de défense. Ces mécanismes sont en perpétuelle évolution, les pathogènes développant également en permanence de nouveaux moyens d'échapper à la surveillance immunitaire de l'hôte. Et ainsi, hôte et pathogènes doivent co-évoluer continuellement pour survivre (Van Valen 1973).

1) Immunités innée et adaptative

On distingue deux types de mécanisme immunitaire, d'une part l'immunité innée, codée dans les cellules germinales et d'autre part l'immunité adaptative codée dans les cellules somatiques et se diversifiant par recombinaison et variation de segments de gènes (Medzhitov, Janeway 1998; Janeway et al. 2001). L'immunité innée n'est pas spécifique d'antigènes particuliers et est incapable de mémoire. Elle constitue par contre une réponse immédiate, assurant les premiers mécanismes de défense contre les pathogènes. En effet, lors d'une infection de notre organisme, c'est-à-dire lorsque des structures physiques telles que la peau, les muqueuses ou encore le suc digestif ne suffisent à empêcher l'intrusion d'un pathogène, ce sont les cellules de l'immunité innée telles que les cellules dendritiques et les macrophages qui vont entrer en jeu : elles vont déclencher des mécanismes permettant à la fois l'élimination directe du pathogène, mais aussi stimuler la réponse adaptative. Nous en arrivons donc à l'immunité adaptative, dont la mise en œuvre est donc retardée, mais qui est spécifique et capable de mémoire. L'immunité adaptative permettra en effet d'acquérir des récepteurs spécifiques d'un pathogène particulier et de les garder en mémoire pour pouvoir les produire plus rapidement lors d'un futur contact avec ce pathogène ; les lymphocytes sont les principales cellules représentantes de ce système.

L'étude de l'immunité adaptative a longtemps été privilégiée, celle-ci permettant comme je l'ai précisé l'établissement d'un large répertoire de récepteurs d'antigènes et une mémoire immunitaire. Cependant, l'intérêt pour l'immunité innée a été relancé récemment lorsqu'il a été mis en évidence que celle-ci, partagée par l'ensemble des espèces vivantes, constituait un système évolutivement très ancien (Hoffmann et al. 1999; Janeway, Medzhitov 2002; Litman, Cannon, Dishaw 2005) ; l'immunité adaptative n'est quant à elle présente que chez les vertébrés. De plus, il a été montré que l'induction de certains types de réponse adaptative dépend de la reconnaissance préalable de pathogènes par les cellules de l'immunité

innée (Medzhitov, Janeway 1998). En résumé, l'immunité innée intervient en premier, jouant un rôle essentiel dans les primo-infections. Et ce sont ensuite les deux systèmes qui vont agir en synergie pour défendre l'organisme.

2) Des simples senseurs microbiens aux acteurs clés d'une surveillance constante

En 1989, Charles Janeway propose un modèle selon lequel l'immunité innée serait la véritable gardienne des clés du déclenchement d'une réponse immunitaire. Nous sommes en permanence en contact avec de nombreuses espèces microbiennes, dont certaines peuvent avoir un impact négatif sur la santé, d'autres étant devenues au contraire indispensables au bon fonctionnement de notre organisme. L'une des forces de l'immunité est donc de savoir distinguer les unes des autres. Le fait de réagir ou non face à un agent étranger reposerait sur la reconnaissance de motifs par des récepteurs qu'il nomme les PRRs (pour *Pattern Recognition Receptors*) (Kimbrell, Beutler 2001; Medzhitov 2001; Medzhitov, Janeway 2002). Ces récepteurs reconnaissent diverses molécules qui sont absentes des organismes multicellulaires, mais qui sont caractéristiques d'un microbe et de son intrusion : celles-ci seront appelées PAMPs (*pour Pathogen-Associated Molecular Patterns*) ; en tant qu'exemples de PAMPs on pourrait citer les fortes concentrations de mannose dans les sucres, le lipopolysaccharide (LPS) et le peptidoglycane (composants respectifs des parois des bactéries à Gram négatif et positif), la méthylation d'ADN ou encore les ARN double-brin (provenant de virus) (Medzhitov 2001; Medzhitov, Janeway 2002).

A partir de 1994, Polly Matzinger développe la théorie du danger. L'immunité innée n'est plus définie comme un ensemble de senseurs du « soi » et du « non-soi », mais plutôt comme des senseurs de « ce qui est normal » et « ce qui est anormal » (Matzinger 1994; Matzinger 2002). Depuis, ce modèle a été validé expérimentalement par l'identification de récepteurs de signaux de danger et de certains de leurs ligands (Shi, Evans, Rock 2003; Mariathasan et al. 2006; Petrilli et al. 2007; Fink, Bergsbaken, Cookson 2008; Wickliffe, Leppla, Moayeri 2008; Schroder, Tschopp 2010) et il a conduit à une nouvelle définition des PRRs : ils sont désormais considérés comme des capteurs de signaux cellulaires de détresse : ces signaux peuvent être des composants bactériens, viraux ou provenant d'autres parasites, mais aussi des changements de concentration de différentes molécules comme le potassium ou l'ATP. Ces déséquilibres en molécules chimiques peuvent résulter de l'intrusion de pathogènes, de toxines ou plus généralement d'un mauvais fonctionnement de la cellule.

Ainsi, même si les PRRs ont initialement été décrits comme des senseurs microbiens détectant les PAMPs, il faut donc garder à l'esprit qu'ils peuvent également détecter des signaux de danger cellulaire, les DAMPs (pour Damage associated molecular pattern molecules). Les PRRs pourraient même plus généralement jouer un rôle dans la surveillance du niveau de différents paramètres indiquant le bon fonctionnement de la cellule. Nous en arrivons donc à la notion d'homéostasie (du grec μοιος, homoios, « similaire » et στημι, histēmi, « immobile »), initialement définie par le médecin et physiologiste français du XIXème Claude Bernard (1813-1878) : il s'agit de la capacité que peut avoir un système quelconque (ouvert ou fermé) à conserver son équilibre de fonctionnement en dépit des contraintes qui lui sont extérieures. Le physiologiste américain Walter Bradford Cannon (1871-1945), définit lui l'homéostasie comme « l'équilibre dynamique qui nous maintient en vie. » Les différents rôles des PRRs sont davantage développés dans la description qui va suivre, ainsi que dans la discussion.

Actuellement, plusieurs familles de PRRs, définies par leurs structures et leurs fonctions particulières ont été décrites chez l'homme. Nous allons aborder dans cette partie les principales familles connues à l'heure actuelle, tout en gardant à l'esprit qu'il existe également certains senseurs isolés qui ne seront pas décrits ici (**Figure 13**). Nous parlerons des *Toll-like receptors* (TLRs) les premiers et donc les plus étudiés des PRRs. Nous nous attacherons ensuite à la présentation de deux groupes de senseurs cytosoliques, les *RIG-I-like receptors* (RLRs) et les *Nod-Like receptors* (NLRs). Les *C-type lectins* (CLRs), récepteurs situés à la surface des cellules seront présentés succinctement à la fin de cette partie.

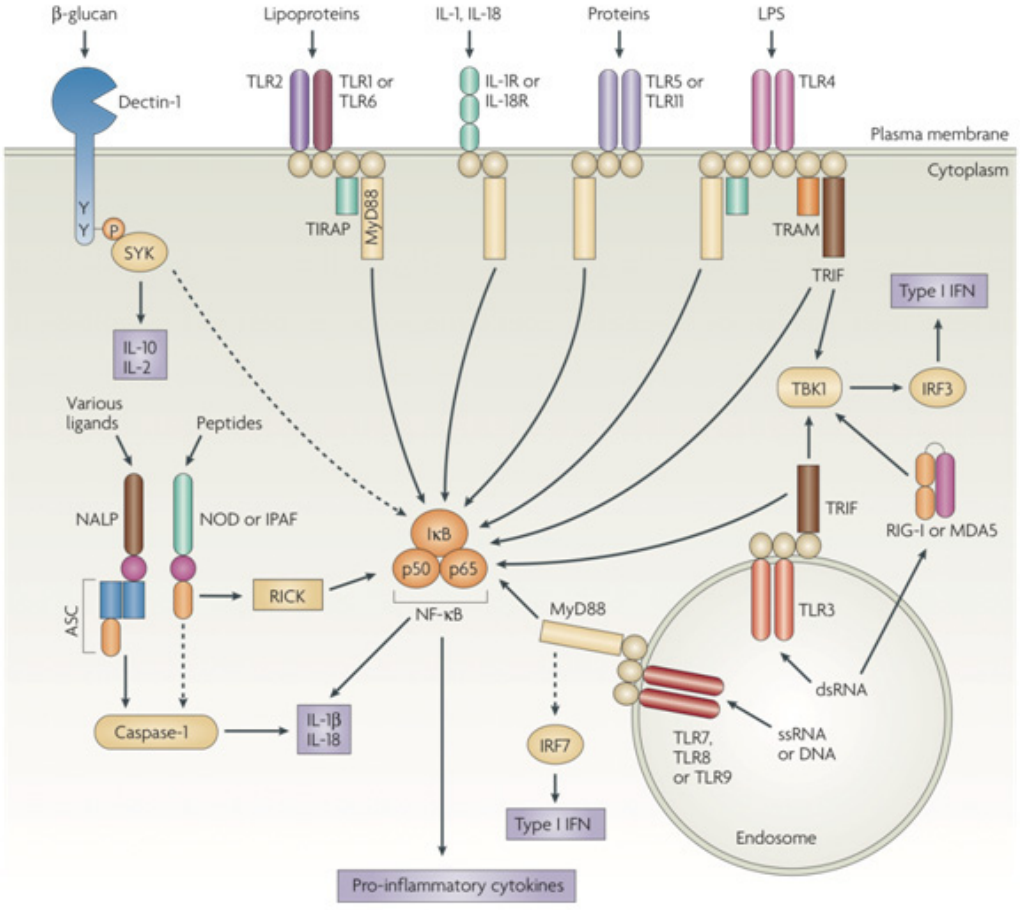


Figure 13. Activation des principales voies de signalisation des *Pattern Recognition Receptors (PRRs)* par leurs ligands. Extrait de (Trinchieri, Sher 2007).

3) Les principales familles de PRRs

Les plus largement étudiés restent sans aucun doute les *Toll-like receptors* (TLRs), et ce sous tous les angles (études génétiques, immunologiques, cliniques, etc) (Casanova, Abel, Quintana-Murci 2011). On sait aujourd'hui qu'ils sont présents dans la plupart des vertébrés et invertébrés, faisant de ces récepteurs une famille évolutivement très ancienne.

a) Les *Toll-like receptors*

Le nom de ces récepteurs provient de leur similarité avec la protéine codée par le gène *Toll*, identifié chez la drosophile (Anderson, Jurgens, Nusslein-Volhard 1985; Hashimoto, Hudson, Anderson 1988). Quand ce gène est muté, la drosophile adopte une manière inhabituelle de voler, révélant ainsi le rôle phare de ce gène dans son développement. On raconte que les chercheurs surpris de constater cet effet s'exclamèrent « Das ist ja toll ! », ce qui signifie en allemand « C'est super! », donnant ainsi le nom de *Toll* à ce gène. C'est en 1996 qu'il fut montré à partir de drosophiles déficientes pour le gène *Toll* que la protéine avait

un rôle-clé dans l'immunité antifongique et la défense contre les bactéries à Gram positif, déclenchant la synthèse de peptides antimicrobiens (Lemaitre et al. 1996; Hoffmann 2003; Lemaitre, Hoffmann 2007). Tous ces travaux ont amené les scientifiques à revoir la place et le rôle de l'immunité innée chez les vertébrés. En effet, la présence, chez les vertébrés uniquement, d'un processus de défense plus spécifique et plus complexe, l'immunité adaptative, avait jusque-là placé l'immunité innée au second plan.

Un an après, se basant sur l'identification de Toll comme un récepteur de l'immunité innée chez la drosophile, une recherche d'homologie dans les bases de données mena à la découverte d'un homologue au gène *Toll* chez l'homme (Medzhitov, Preston-Hurlburt, Janeway 1997). Il s'agissait en fait du récepteur que l'on nomme aujourd'hui TLR4, impliqué dans la production de cytokines inflammatoires et de molécules de co-stimulation (Medzhitov, Preston-Hurlburt, Janeway 1997). Les analyses qui ont suivi ont permis la découverte d'autres protéines structurellement proches de TLR4, constituant aujourd'hui chez l'homme une famille de 10 TLRs fonctionnels (Medzhitov, Preston-Hurlburt, Janeway 1997; Rock et al. 1998; Takeuchi et al. 1999; Chuang, Ulevitch 2000; Du et al. 2000; Chuang, Ulevitch 2001). Une partie d'entre eux (TLR3, 7, 8 et 9) sont intracellulaires, se situant plus précisément dans l'endosome, un sous-compartiment responsable du « tri cellulaire » des molécules provenant du milieu extra-cellulaire et qui ont été internalisées dans des vésicules d'endocytose : ces récepteurs sont principalement impliqués dans la reconnaissance d'acides nucléiques, provenant en particulier de virus. L'autre partie des TLRs est située à la surface des cellules, sur la membrane plasmique (TLR1, TLR2, TLR4, TLR5, TLR6 et TLR10) : mis à part TLR10, pour lequel on ne connaît toujours aucun ligand, ce groupe détecte préférentiellement des motifs bactériens (comme le LPS pour TLR4 ou la flagelline pour TLR5) ou de certains parasites, ainsi que des particules virales (Akira, Uematsu, Takeuchi 2006; Kawai, Akira 2006). Les TLRs participent à diverses voies de signalisation, recrutant différentes combinaisons d'adaptateurs, MyD88, TIRAP (MAL), TRIF (TICAM1) et TRAM (TICAM2). La cascade de signalisation en résultant peut mener à la production de cytokines inflammatoires telles que l'interleukine-1-β, *via* l'activation des facteurs de transcription NFκB et AP-1 (*Activator Protein-1*) (Bochud et al. 2007; Lee, Kim 2007) ; ces cytokines vont ainsi initier et amplifier la réponse inflammatoire en activant et recrutant diverses cellules de l'immunité innée, telles que les monocytes, les neutrophiles ou encore les cellules *Natural Killer* (NK) (Akira, Takeda 2004; Lee, Kim 2007). D'autre part, l'activation de TLR3, TLR4, TLR7, TLR8 peut conduire, *via* l'activation du facteur de régulation IRF3 (*Interferon Regulatory Factor 3*) ou IRF7, à la production d'interférons (IFN) de type I tels que les IFN-α

et IFN-*β*, jouant un rôle antiviral (Isaacs, Lindenmann 1957; Hallum, Younger 1966; Repik, Flamand, Bishop 1974).

En outre, certaines familles de senseurs se trouvant directement dans le cytoplasme vont également participer à notre immunité innée.

b) Les *Rig-I-like receptors*

Le premier des RLRs, RIG-I (*Retinoic acid-inducible gene I*), fut identifié en 2004 (Yoneyama et al. 2004), donnant ensuite son nom à la famille d'hélicases à ARN ayant des structures proches. On compte ainsi aujourd'hui trois membres dans cette famille, RIG-I, IFIH1 (*Interferon-induced helicase C domain-containing protein 1*, aussi connu sous le nom de MDA5 pour *Melanoma Differentiation-Associated gene 5*) et LGP2 (*Laboratory of Genetics and Physiology 2*) : ceux-ci détectent les ARN viraux présents dans le cytosol cellulaire et participent à l'immunité antivirale (Yoneyama et al. 2005; Yoneyama, Fujita 2007; Baum, Garcia-Sastre 2010; Loo, Gale 2011).

Ces trois molécules ont en commun (i) un domaine C-Terminal (CTD) principalement impliqué dans la reconnaissance d'ARN ; ce domaine est appelé domaine répresseur (RD) dans le cas de RIG-I et LGP2, pour sa capacité supplémentaire à bloquer la signalisation induite par RIG-I , (ii) un domaine hélicase central qui permet d'hydroliser l'ATP et potentiellement de dérouler l'ARN, (iii) enfin, en N-terminal, deux CARDs (*CAspase Recruitment Domains*) que seuls RIG-I et IFIH1 possèdent et qui sont associés au déclenchement de la signalisation de la réponse inflammatoire et dans certains cas à une mort cellulaire (Kang et al. 2002; Kovacsovics et al. 2002; Yoneyama et al. 2005; Saito et al. 2007). Plus précisément, ces CARDs vont recruter un adaptateur appelé IPS-1 (*Interferon-beta Promoter stimulator 1*, aussi dénommé MAVS, Cardif ou VISA) (Kawai et al. 2005; Meylan et al. 2005; Seth et al. 2005; Xu et al. 2005). Deux complexes pourront ainsi être formés : le complexe TBK1 (*TANK-binding kinase-1*) mènera à la phosphorylation des facteurs de régulation IRF3 et IRF7, induisant l'expression des interférons de type I. Parallèlement, le complexe IKK (*I Kappa B Kinase*) pourra également être formé, activant le facteur de transcription NF-*κ*B menant à la production de cytokines proinflammatoires (Kato et al. 2005; Kawai et al. 2005). Malgré cette fonction commune, RIG-I et IFIH1 ne semblent pas avoir les mêmes ligands. RIG-I reconnaît un large éventail de virus : on peut notamment citer le virus de l'hépatite C, ou encore ceux de l'*Influenza* A, de la rage et de la rougeole (Melchjorsen et al. 2005; Kato et al. 2006; Pichlmair et al. 2006; Liu et al. 2007; Loo et al. 2008; Mikkelsen et al. 2009). Le spectre de virus détectés par IFIH1 se limite quant à lui aux

familles des picornavirus (incluant notamment le virus responsable de l'encéphalomyocardite), des rhinovirus et des entérovirus (comprenant les poliovirus) (Gitlin et al. 2006; Wang et al. 2009). Certains flavivirus, tels que le virus de la Dengue de type 2, et certains réovirus semblent être reconnus à la fois par RIG-I et IFIH1 (Loo et al. 2008). Les différences entre RIG-I et IFIH1 s'expriment également en matière de types d'ARN : RIG-I a une forte affinité pour les ARN simple-brin portant trois groupements phosphates en 5' et pour les ARN double-brin de petite taille (~1kb) ; de son côté, IFIH1 se lie davantage aux ARN double-brin de plus grande taille (~1-5kb) (Kato et al. 2008; Takahasi et al. 2008; Schlee et al. 2009; Loo, Gale 2011). Il est intéressant de noter que dans tous ces cas, l'ARN reconnu (double-brin ou simple-brin portant un groupement triphosphate) n'est pas présent dans une cellule humaine dans des conditions « normales » et sa présence constitue donc un signal du non-soi, déclenchant la réponse inflammatoire (Nallagatla, Toroney, Bevilacqua 2008).

Enfin, le troisième et dernier membre des RLRs, LGP2, ne peut initier la réponse antivirale ; il s'apparente davantage à un régulateur de la réponse inflammatoire. En effet, celui-ci ne dispose pas de domaines CARD essentiels au déclenchement de la cascade de signalisation : il semble qu'il régule négativement RIG-I ; plusieurs mécanismes ont été proposés : (i) LGP2 possédant tout de même une forte affinité pour les ARN double-brin, la formation de complexes LGP2-ARN pourrait diminuer la quantité d'ARN disponibles, réduisant ainsi la réponse *via* RIG-I ; (ii) le domaine répresseur (RD) de LGP2 pourrait réprimer directement RIG-I en se fixant sur son domaine de signalisation CARD ; (iii) il a également été suggéré que LGP2 pourrait se lier à l'adaptateur IPS-1 nécessaire lors de la signalisation, entrant ainsi en compétition avec RIG-I, et diminuant son activité. Cependant, les véritables mécanismes restent peu clairs ; il semble que LGP2 ait au contraire un rôle de co-activateur d'IFIH1 (Venkataraman et al. 2007).

De nombreuses études génétiques ont permis ces dernières années d'identifier divers polymorphismes fonctionnels dans RIG-I et IFIH1, dont certains semblent modifier la réponse aux infections virales (Pothlichet et al. 2009; Shigemoto et al. 2009; Hu et al. 2010). D'autre part, des mutations d'*IFIH1* ont été associées à plusieurs maladies autoimmunes, dont le diabète de type I (Smyth et al. 2006; Barrett et al. 2009; Nejentsev et al. 2009; Jermendy et al. 2010; Reddy et al. 2011), la maladie de Grave (Sutherland et al. 2007), le lupus (Gateva et al. 2009) ou encore le psoriasis (Strange et al. 2010). Le lien qui peut exister entre maladies infectieuses, sélection naturelle et maladies auto-immunes sera discuté dans la partie « Résultats et discussion » de ce mémoire.

Nous avons donc vu dans cette partie que les RLRs et 3 des 4 TLRs endosomaux (TLR3, 7 et 8) détectaient les ARN viraux, conduisant à la production d'interférons de type I et les cytokines pro-inflammatoires (Akira, Takeda 2004; Yoneyama et al. 2004; Kato et al. 2005; Kawai et al. 2005). Cependant, il semble qu'ils utilisent des adaptateurs distincts pour initier la cascade de signalisation : ils agissent ainsi davantage en parallèle qu'en interaction (Akira, Takeda 2004; Kato et al. 2005; Kawai, Akira 2006). Il pourrait donc y avoir une redondance dans leurs fonctions.

c) Les *Nod-like receptors*

Avec ses 22 membres, la famille des *NOD-like receptors* (NLRs) est probablement la plus mal connue des PRRs. Il s'agit pourtant d'une famille très ancienne. Pour preuve, ils ont été trouvés chez différents vertébrés et invertébrés (Rast et al. 2006; Tian, Pascal, Monget 2009; Lange et al. 2011) ; il existe également chez les plantes des senseurs de structure similaire, les NBS-LRRs (*Nucleotide-Binding Site Leucine-Rich Repeats*), formant une famille qui s'est particulièrement élargie au cours de l'évolution et constitue l'un des piliers de leur système immunitaire (Bent, Mackey 2007; Caplan, Padmanabhan, Dinesh-Kumar 2008) ; il semble d'ailleurs que ces récepteurs, et plus précisément leurs domaines LRRs, soient hypervariables et particulièrement sujets à la sélection positive chez les espèces végétales (Parniske et al. 1997; Meyers et al. 1998; Noel et al. 1999; Ellis, Dodds, Pryor 2000; Mondragon-Palomino et al. 2002; Wang et al. 2006); cela leur permet sans doute d'acquérir de nouvelles spécificités de résistance (Endo, Ikeo, Gojobori 1996). Chez l'homme, on dispose tout de même de nombreuses informations sur certains des NLRs, dont beaucoup ont été extrapolées (à tort ou à raison) à l'ensemble d'entre eux. Les NLRs sont pour la plupart localisés dans le cytoplasme et reconnaissent *a priori* davantage des motifs bactériens ou des signaux de danger cellulaire (Kufer, Sansonetti 2010). La majorité d'entre eux semble impliquée dans la régulation de l'inflammation et la mort cellulaire. Les NLRs sont constitués de trois parties : (i) un domaine LRR qui peut jouer un rôle dans la liaison au ligand bactérien, (ii) un domaine central NACHT qui permet l'oligomérisation de la protéine et potentiellement l'utilisation d'NTP (*Nucleoside-TriPhosphate*) pour son repliement et (iii) un domaine effecteur en N-terminal qui semble intervenir la plupart du temps dans la signalisation ou la mort cellulaire. Ce dernier domaine est très variable entre NLRs et permet de distinguer deux principales sous-familles, les NODs (Nucleotide Oligomerization Domain) et les NALPs (NACHT, leucine-rich repeat, pyrin domain, ou NLRPs).

Parmi les NODs, ce sont principalement NOD1 et NOD2 qui ont été étudiés. Activés par des composants bactériens, l'acide meso-diaminopimélique (meso-DAP, composant des bactéries à Gram négatif) pour NOD1 et le dipeptide-muramyl (MDP, constitutif de la paroi bactérienne) pour NOD2, ils mènent tous deux à l'activation du facteur de transcription NFκB *via* les molécules adaptatrices RICK (ou RIP2) et CARD9 ; ils permettent ainsi la production de cytokines inflammatoires (Inohara et al. 2000; Girardin et al. 2001; Ogura et al. 2001a; Kobayashi et al. 2002; Chamaillard et al. 2003; Inohara et al. 2003; Girardin et al. 2003a; Girardin et al. 2003b; Hsu et al. 2007). Il a de plus été montré que NOD1 et NOD2 jouaient un rôle important dans la réponse immunitaire face à des infections gastro-intestinales, indépendamment des TLRs (Kim, Lee, Kagnoff 2004; Viala et al. 2004; Kobayashi et al. 2005). Toutefois, *NOD2* reste principalement connu pour son association avec la maladie de Crohn, une maladie chronique inflammatoire rare, à composante autoimmune (Hugot et al. 2001; Ogura et al. 2001b; Lesage et al. 2002). Trois mutations principales et une trentaine de mutations mineures de *NOD2* ont été observées chez les malades. Et il a été estimé que le risque d'être malade est 1,5 à 3 fois plus élevé chez les hétérozygotes simples que chez les sujets n'ayant pas de mutation et environ 40 fois plus fort chez les homozygotes, ce qui représente le facteur de risque le plus important connu à ce jour de développer la maladie de Crohn (Desreumaux 2005).

En ce qui concerne les membres de la sous-famille des NALPs (NALP1-14), ils ont à l'origine été identifiés comme des éléments clés faisant partie intégrante de grands complexes protéiques appelés inflammasomes : ceux-ci, qui incluent notamment un adaptateur nommé ASC (pour *Apoptosis-associated speck-like*), interviennent dans le déclenchement de la réponse inflammatoire et la mort cellulaire (Srinivasula et al. 2002; Tschopp, Martinon, Burns 2003). C'est l'inflammasome formé avec NALP1 qui a été identifié en premier (Martinon, Burns, Tschopp 2002). Suite à l'assemblage de l'inflammasome, une caspase-1 est activée, processus nécessaire à l'activation des pro-interleukines IL-1β et IL-18 menant à la réponse inflammatoire. On sait maintenant que ce complexe est formé avec d'autres NLRs, tels que NALP3 et NLRC4 (IPAF) (Mariathasan et al. 2006; Schroder, Tschopp 2010). En ce qui concerne les ligands impliqués, NALP3 semble activé par le MDP provenant de *Listeria monocytogenes* et *Staphylococcus aureus*, par différents signaux de danger cellulaire (ATP, cristaux d'acide urique, flux de potassium, etc) ou encore par les ARN de certains virus et bactéries (Martinon et al. 2004; Mariathasan et al. 2006; Martinon et al. 2006; Meylan, Tschopp, Karin 2006; Kanneganti et al. 2006a; Kanneganti et al. 2006b; Petrilli et al. 2007). L'activation de NALP1 semble quant à elle dépendre de motifs bactériens, tels que le MDP et

le LPS (Martinon, Burns, Tschopp 2002; Mariathasan et al. 2006; Faustin et al. 2007). Certaines mutations de l'homologue murin Nalp1b semblent affecter la suceptibilité à la toxine de l'anthrax (Boyden, Dietrich 2006).

Même s'ils sont étudiés de manière beaucoup plus précise ces dernières années, on dispose pour les autres NLRs d'informations assez éparses et parfois difficiles à concilier ; il reste en particulier de nombreuses zones d'ombre concernant leurs ligands. De manière générale, on les trouve impliqués dans des fonctions de régulation de la réponse inflammatoire, dans la mort cellulaire, l'homéostasie, la production de gamètes, ou encore le développement embryonnaire (**Tableau 2**). Et ainsi, si les NLRs sont souvent considérés comme des PRRs cytosoliques, reconnaissant majoritairement les bactéries et déclenchant la réponse inflammatoire, ils restent une famille très hétérogène en matière de localisation, de ligands et de fonctions. On peut citer quelques exemples : NLRC5 détecte *a priori* des acides nucléiques viraux (Neerincx et al. 2010). CIITA, facteur de transcription, active les molécules du complexe majeur d'histocompatibilité de classe II (CMH II) et se trouve donc en majeure partie dans le noyau cellulaire (Steimle et al. 1993). Par ailleurs, NLRX1, qui joue un rôle dans la régulation antivirale, est principalement localisé dans la mitochondrie (Moore et al. 2008). NOD2 est impliqué dans l'autophagie, c'est-à-dire la digestion par une cellule de ses propres organites pour la présentation des antigènes aux cellules de l'immunité adaptative (Cooney et al. 2010). En présence des bactéries *Salmonella* et *Pseudomonas*, l'activation de NLRC4 (IPAF) conduit à une forme particulière de mort cellulaire, la pyroptose (Bortoluci, Medzhitov 2010). NAIP semble quant à lui inhiber l'apoptose (Liston et al. 1996; Davoodi et al. 2010).

Tablau 2. Tableau récapitulant les fonctions cellulaires des NLRs, ainsi que les maladies qui leur ont été associées chez l'homme

NLR	Signaling pathway	Acts as PRR	Function(s)	Associated disease(s)
CIITA	MHC class II transcriptional regulator	?	Regulates MHC class II gene expression	Bare lymphocyte syndrome, rheumatoid arthritis, multiple sclerosis
NAIP	JNK activation by TAK1 and others	Yes	Flagellin sensing, pyroptosis, inhibition of apoptosis	*Legionella* susceptibility (mice), spinal muscle atrophy
NOD1	RIP2-dependent NF-κB and MAPK activation	Yes	PRR for bacterial PGN, inducer of apoptosis and of autophagy	Asthma (?), inflammatory bowel disease (?)
NOD2	RIP2-dependent NF-κB and MAPK activation	Yes	PRR for bacterial PGN, attenuates TLR2 responses, induces autophagy	Blau syndrome, Crohn's disease, early-onset sarcoidosis, graft-versus-host disease
NLRC3	Unknown	?	Negative regulator of T cells	Unknown
NLRC4	Inflammasome formation	Yes	PRR for flagellin and bacterial secretion systems, inducer of pyroptosis	Unknown
NLRC5	Type I interferon responses, MHC I regulator	?	Regulates innate immune responses, influences MHC class I presentation	Unknown
NLRP1	Inflammasome formation	Yes	Activated by MDP and DAMPs	Vitiligo, Addison's disease, type 1 diabetes
NLRP2	Inflammasome formation	?	Negative regulator of NF-κB	Beckwith-Wiedemann syndrome, graft-versus-host disease
NLRP3	Inflammasome formation	Yes	Responds to DAMPs	Cryopyrin-associated periodic syndromes (CAPS)[a], type 2 diabetes
NLRP4	Unknown	?	Suppressor of NF-κB (?)	Unknown
NLRP5	Unknown	?	Maternal effect gene	Sterility in female (mice)
NLRP6	Unknown	?	Negative regulator of NF-κB and IL-1β (?)	Unknown
NLRP7	Unknown	?	Negative regulator of IL-1β (?)	Familial biparental hydatidiform moles
NLRP8	Unknown	?	Unknown	Unknown
NLRP9	Unknown	?	Unknown	Unknown
NLRP10	Caspase-1 activation	?	Negative regulator of caspase-1 activation	Unknown
NLRP11	Unknown	?	Unknown	Unknown
NLRP12	Unknown	?	Negative regulator of classical and non-canonical NF-κB pathways involved in MHC-I gene expression	Hereditary periodic fevers with skin urticaria
NLRP13	Unknown	?	Unknown	Unknown
NLRP14	Unknown	?	Spermatogenesis	Spermatogenic failure, embryogenesis failure in mice
NLRX1	Mitochondrially located	?	ROS induction, viral sensing	Unknown
Apaf-1	Apoptosome formation	?	Inducer of apoptosis by cytochrome *c* sensing	Unknown

Extrait de (Kufer, Sansonetti 2010)

Ainsi, il sera particulièrement intéressant dans notre étude de voir si certains membres de cette famille évoluent sous des contraintes spécifiques (*i.e.* montrent des signatures de sélection particulières) et de déterminer si groupes fonctionnels et signatures de sélection peuvent être reliés.

Enfin, j'ajouterai qu'il existe bien sûr d'autres récepteurs, dont en particulier les CLRs (*C-type lectins*), qui se trouvent à la surface des cellules, enchâssés dans la membrane. Ces senseurs sont caractérisés par la reconnaissance de glucides, mécanisme requérant la présence de calcium. Ils peuvent détecter un large spectre de pathogènes (virus, bactéries, …) et semblent jouer un rôle dans l'immunité (Zelensky, Gready 2005). Ainsi par exemple, DC-SIGN reconnaît des virus tels que HIV-1, Ebola, ceux de l'hépatite C et de la Dengue, ainsi que les bactéries *Mycobacterium tuberculosis*, *Mycobacterium leprae* et *Helicobacter pylori*, ou encore le parasite *Leishmania pifanoi* (Geijtenbeek et al. 2000; Alvarez et al. 2002; Colmenares et al. 2002; Appelmelk et al. 2003; Geijtenbeek et al. 2003; Lozach et al. 2003; Tailleux et al. 2003; Tassaneetrithep et al. 2003; Bergman et al. 2004; Barreiro et al. 2006).

VI) Maladies infectieuses, évolution humaine et sélection naturelle : exemples et objectifs de thèse

Ainsi, l'homme, au cours de sa sortie d'Afrique, puis de sa colonisation du reste du monde a dû faire face à plusieurs reprises à de nouvelles conditions environnementales. Les pathogènes et à travers eux les maladies infectieuses ont exercé des pressions particulièrement fortes ; avant l'arrivée des vaccins et des antibiotiques, les populations humaines étaient régulièrement décimées, seuls les individus les plus résistants pouvant survivre (Casanova, Abel 2005). John Burdon Sanderson Haldane (1860-1936) est l'un de premiers à avoir souligné le rôle prépondérant des maladies infectieuses comme agent de sélection naturelle. Et en effet, de nombreuses études détectant de la sélection à l'échelle du génome confirment l'importance des gènes qui sont impliqués dans l'immunité et qui sont en interaction avec les pathogènes : ceux-ci constituent clairement des cibles privilégiées de sélection chez l'homme (Clark et al. 2003; Vallender, Lahn 2004; Bustamante et al. 2005; Nielsen 2005; Voight et al. 2006; Wang et al. 2006; Quintana-Murci et al. 2007; Barreiro et al. 2008; Barreiro, Quintana-Murci 2010). Rappelons donc quelques un des nombreux exemples qui peuvent attester de leur importance.

1) Exemples de signatures de sélection chez l'homme dues aux pathogènes

Lors de la définition des différents types de sélection, j'ai notamment cité le cas du système HLA, montrant la plus forte diversité des gènes humains du fait de l'action de la sélection balancée, permettant ainsi une reconnaissance d'une grande variété de pathogènes. Nous avons aussi vu que le paludisme, *via Plasmodium falciparum* et *Plasmodium vivax,* avait exercé de fortes pressions sur nos génomes, probablement les plus fortes des forces évolutives de l'histoire humaine récente (Kwiatkowski 2005). On dénote ainsi la présence dans notre génome de différentes signatures de sélection dues à l'existence de ces pathogènes : celles du gène codant pour la G6PD, de l'allèle HbS ou encore de l'allèle nulle du gène *DARC,* confèrent différentes formes de résistance au paludisme. On pourrait d'autre part citer l'exemple de CCR5 (*Chemokine CC motif receptor 5*), dont je n'ai pas encore parlé jusqu'ici. Ce récepteur s'avère être un composant indispensable au virus HIV-1 (*Human immunodeficiency virus type 1*) pour entrer dans les cellules (Alkhatib et al. 1996). Originellement, le rôle de CCR5 dans le SIDA a été démontré à partir de l'observation d'individus homozygotes pour une délétion de 32 paires de bases dans *CCR5* qui se trouvaient

plus résistants à l'infection par HIV-1 que les individus n'ayant pas cette délétion (Dean et al. 1996). Appuyées par la présence de nombreuses mutations non-synonymes dans *CCR5*, signe potentiel de sélection positive, de nombreuses études ont été réalisées dans ce sens (Carrington et al. 1997). Certains chercheurs concluent en faveur d'une sélection positive ciblant la délétion *CCR5*-Δ32, celle-ci se trouvant à des fréquences notables en Europe (jusqu'à 16%) alors qu'elle est absente des populations africaines, asiatiques et natives américaines (Stephens et al. 1998; Galvani, Novembre 2005). Sabeti a au contraire montré que la présence de la délétion pouvait être expliquée par une évolution neutre (Sabeti et al. 2005). L'attention a alors été attirée sur des polymorphismes apparaissant en 5' de *CCR5*, dans sa région régulatrice, et qui semblaient affecter la progression de la maladie (Gonzalez et al. 1999; Gonzalez et al. 2001). On sait maintenant que ces polymorphismes ont été ciblés par la sélection balancée (Bamshad et al. 2002) et sont associés à une résistance à la fois à l'infection par HIV-1 et à la progression de la maladie. Toutefois, étant donnée l'apparition récente du virus HIV-1, il est peu probable qu'il soit la cause réelle de sélection ; mais ce sont certainement des virus utilisant le même mode d'invasion qu'HIV-1, c'est-à-dire *via* CCR5 (Lalani et al. 1999; Bamshad et al. 2002).

Enfin, une étude réalisée au laboratoire a montré que différents types de contraintes étaient exercées sur les gènes codant deux principaux groupes de senseurs microbiens de la famille des TLRs (Barreiro et al. 2009). Les gènes codant les TLRs endosomaux (TLR3, 7, 8 et 9), récepteurs davantage spécialisés dans la reconnaissance d'acides nucléiques, provenant majoritairement de virus, se trouvaient sous sélection purificatrice. Arborant une très faible proportion de mutations non-synonymes, il avait été conclu qu'ils doivent jouer un rôle essentiel dans notre survie. A l'inverse, les gènes codant les TLRs situés à la surface des cellules (TLR1, 2, 4, 5, 6 et 10) et qui détectent principalement des motifs provenant de bactéries et de certains parasites, accumulaient un grand nombre de mutations altérant la protéine, suggérant que leurs rôles sont plus redondants dans notre système immunitaire. D'autre part, il a été montré que *TLR1* était sous sélection positive en Europe, la mutation I602S constituant une cible probable : l'allèle sélectionné semble conférer une réponse inflammatoire plus faible à certaines infections, ce qui a pu s'avérer très avantageux contre des maladies dont une réponse inflammatoire excessive peut être fatale (Hawn et al. 2007; Ma et al. 2007; Barreiro et al. 2009). L'absence de récepteur ou tout du moins son expression réduite peut également constituer un avantage lorsque le pathogène l'utilise afin d'échapper au système immunitaire (Hamblin, Di Rienzo 2000; Johnson et al. 2007; Misch et al. 2008).

2) Objectifs de thèse

Ainsi, il semble clair que les pathogènes ont joué un rôle prépondérant dans l'évolution de notre génome et les cas présentés ci-dessus ne sont que quelques exemples parmi ceux existants et ceux qui seront découverts dans les années à venir. Nous, et par là j'entends les génomes que nous portons, sommes en quelque sorte aujourd'hui les survivants du passé. Des variations génétiques, même infimes, ont pu dans notre histoire (et probablement encore aujourd'hui) être responsables de réponses immunitaires particulières qui ont permis une plus grande résistance à certaines maladies infectieuses. Si nous nous intéressons aux régions de notre génome qui ont permis cette résistance, il paraît judicieux d'étudier les molécules de notre organisme qui participent à la détection de ces pathogènes. Nous l'avons vu, les PRRs, récepteurs de l'immunité innée, peuvent interagir directement avec les pathogènes et se trouvent donc au cœur du mécanisme de coévolution hôte-pathogène. Ils paraissent donc constituer de très bonnes cibles potentielles de sélection naturelle. D'autre part, nous avons pu voir que la génétique des populations *via* la détection de sélection pouvait être une bonne alternative aux études cliniques et épidémiologiques dans la découverte de gènes/variants impliqués dans des maladies humaines.

Ainsi, nous cherchons ici à déterminer s'il existe sur les gènes codant deux familles de PRRs cytosoliques, les *Nod-like receptors* (NLRs) et les *RIG-I-like receptors* (RLRs), des pressions de sélection particulières. Et ainsi, à partir de la diversité et des signatures de sélection naturelle observées au sein de ces gènes, nous cherchons :

- à identifier les *NLRs* et *RLRs* qui montrent des signatures de sélection purificatrice et sont donc très contraints : ils doivent avoir joué (et jouent probablement encore aujourd'hui) un rôle essentiel à notre survie : ce sont ceux qui devraient être étudiés en priorité dans une perspective clinique ou épidémiologique, car ils pourraient être impliqués dans des maladies mendéliennes ou complexes présentant des symptômes sévères. On pourra alors distinguer ces gènes essentiels de ceux qui ont évolué sous des contraintes plus relâchées, dont certains partagent certainement des fonctions plus redondantes.

- à identifier, parmi les *NLRs* et *RLRs*, les gènes qui montrent des signes de sélection positive et évoluent donc de manière plus adaptative. Ils accumulent généralement des mutations qui altèrent la protéine, certaines pouvant atteindre de fortes fréquences. On

peut également détecter pour ces gènes certains variants particuliers qui constituent des cibles probables de sélection positive : ils ont pu augmenter en fréquence parce qu'ils ont conféré à un moment donné un avantage en matière de survie. Ceci présente un grand intérêt dans la mesure où cela permet de prédire les mutations qui ont potentiellement un effet fonctionnel ; celles-ci pourraient en particulier expliquer bon nombre de nos différences en matière de résistance/susceptibilité à certaines maladies infectieuses.

- à établir un modèle hiérarchique général, prédisant les contributions relatives des différentes familles de PRRs à notre survie. Les TLRs avaient déjà été étudiés dans ce contexte ; les résultats de l'analyse présente permettent d'établir une comparaison de la variabilité génétique et des signatures de sélection détectées sur les différents senseurs microbiens, mettant en évidence ceux qui ont joué un rôle prépondérant chez l'homme. Cela pourrait ainsi permettre de mieux comprendre quels sont les voies et acteurs particuliers de l'immunité innée qui ont été mis en place et conservés chez l'homme pour lutter contre les pathogènes (et ceux qui semblent au contraire avoir adopté un rôle plus secondaire).

RESULTATS ET DISCUSSION

I) Contexte

Comme je l'ai largement décrit dans l'introduction, les maladies infectieuses ont joué un rôle essentiel dans la survie des différentes espèces, y compris des pathogènes eux-mêmes. Elles constituent des pressions de sélection majeures, nous « forçant » à nous adapter et modelant nos génomes. Les molécules impliquées dans les premiers mécanismes mis en place par notre organisme pour lutter contre ces infections doivent donc être particulièrement concernés par cette évolution. Ainsi, au laboratoire, différentes études ont été menées sur les principaux acteurs de l'immunité innée, depuis les senseurs, tels que les Toll-like receptors (TLRs) (Barreiro et al. 2009), qui reconnaissent directement les pathogènes, en passant par les adaptateurs participant aux mécanismes de signalisation (Fornarino et al. 2011) (**Annexe 2**), et ce jusqu'aux interférons, produits à la fin de cette chaîne (Manry et al. 2011a; Manry et al. 2011b). Le but est ici de savoir s'il existe des contraintes particulières sur certaines voies de l'immunité innée ou si quelques acteurs seulement jouent des rôles prépondérants. Les cinq adaptateurs contenant un domaine TIR (Toll/interleukine-1), MyD88, MAL, TRIF, TRAM and SARM, utilisés par les TLRs pour déclencher la cascade de signalisation, ont été étudiés (Fornarino et al. 2011) (**Annexe 4**); il a été montré que les gènes codant ces molécules, situées à un véritable carrefour de récepteurs et de complexes de signalisation, ont été particulièrement ciblées par la sélection, comparées aux *TLRs* eux-mêmes. Deux d'entre eux (*MyD88* et *TRIF*) se trouvent sous sélection purificatrice, suggérant qu'ils ont été contraints dans une période ancienne ; par ailleurs, les cinq adaptateurs étudiés montrent des signatures de sélection positive dans l'ensemble des populations humaines ou dans certaines spécifiquement. Ces résultats indiquent ainsi que les contraintes exercées sont spécifiques de certains acteurs et non de voies entières, dans le cas des gènes codant les TLRs/adaptateurs tout du moins. Cependant, ces analyses sont loin d'être exhaustives.

Nous tentons dans l'étude présente de compléter la partie « haute » de ce système, c'est-à-dire les senseurs chargés de la détection microbienne. Comme précisé dans l'introduction, il avait été montré que les gènes codant les TLRs endosomaux, davantage spécialisés dans la reconnaissance d'acides nucléiques, provenant majoritairement de virus, se trouvaient sous sélection purificatrice et qu'ils devaient donc jouer un rôle essentiel dans notre survie. A l'inverse, les gènes codant les TLRs situés à la surface des cellules et qui détectent principalement des motifs provenant de bactéries et de certains parasites semblaient être sous des contraintes plus relâchées, jouant probablement des rôles plus secondaires dans notre défense (Barreiro et al. 2009). Nous complétons ici ces résultats en réalisant le même type

d'étude sur les gènes codant les deux principales familles de senseurs cytosoliques connues, les *NLRs* et les *RLRs*. En d'autres termes, nous avons étudié la diversité génétique observée actuellement chez l'homme dans un panel d'individus n'ayant pas été choisis pour des maladies particulières ; pour cela, nous avons reséquencé les exons, une longueur comparable d'introns et les régions potentiellement régulatrices de ces gènes, et ce dans un panel d'individus représentatif de la population mondiale (HGDP-CEPH) (Cann et al. 2002). Nous avons utilisé des tests inter-spécifiques (comparant l'homme et le chimpanzé) et des tests intra-spécifiques (comparant des populations de tailles similaires provenant d'Afrique sub-saharienne, d'Europe et d'Asie de l'est) pour détecter des signatures de sélection naturelle. Ceci devrait nous permettre de savoir si la sélection a agi sur certains d'entre eux de manière spécifique et dans certaines populations particulières.

Dans un premier article, sont présentés les résultats obtenus pour une première famille de senseurs microbiens situés dans le cytoplasme cellulaire, les *Rig-I-like receptors* (RLRs), reconnaissant principalement les ARN viraux. Sont mis en évidence les différences de diversité (par gène et par domaine) et de pressions de sélection observées entre les trois *RLRs*.

Dans un second article, nous avons tout d'abord étudié une autre famille de senseurs cytosoliques, les *Nod-like receptors* (NLRs), *a priori* davantage spécialisés dans la réponse antibactérienne. Nous avons distingué la sous-famille des NALPs (NALP1-14, qui partagent un domaine pyrine (PYR) en N-terminal) et les NLRs restants (NOD1, NOD2, NLRC3, NLRC4, NLRC5, NLRX1 et CIITA) que nous avons regroupés sous la dénomination « NOD/IPAF » ; ce deuxième groupe inclut les protéines appelées classiquement « NODs » (NOD1, NOD2, NLRC3, NLRC5 et NLRX1), ainsi que NLRC4 (IPAF) et CIITA qui ont en N-terminal un domaine CARD, les rapprochant davantage des NODs (les NALPs sont quant à eux définis par leur domaine PYR). De plus, nous avons exclu de toutes ces analyses le gène *NAIP*. En effet, celui-ci a la particularité d'être constitué de nombreux CNVs (*Copy Number Variants*) ; en d'autres termes, il existe plusieurs exemplaires très similaires de ce même gène sur notre génome. De ce fait, des analyses comparables (séquençage Sanger, identification précise de SNPs) n'ont pu être réalisées pour ce gène et ont fait l'objet d'une étude séparée au laboratoire (Boniotto et al. 2011).

D’autre part, dans le second article présenté ici, nous avons intégré l’ensemble des résultats des senseurs étudiés au laboratoire, les NLRs, les RLRs et les *Toll-like receptor*s (TLRs) et proposons un modèle hiérarchique général. Il est à noter que le panel que j’ai utilisé est comparable (en matière de structure et de représentativité) à celui des TLRs, permettant ainsi l’établissement de ce schéma général. A la fin de cet article, nous abordons l’existence de fonctions moins classiques pour ces PRRs et discutons de l’implication de ces rôles atypiques pour l’interprétation de nos résultats.

II) Article 1 : *« The selective footprints of viral pressures at the human RIG-I-like receptor family »*

(pour les *supplementary material*, voir **Annexe 2**)

Human Molecular Genetics, 2011, Vol. 20, No. 22 **4462–4474**
doi:10.1093/hmg/ddr377
Advance Access published on August 24, 2011

The selective footprints of viral pressures at the human RIG-I-like receptor family

Estelle Vasseur[1,2,3], Etienne Patin[1,2], Guillaume Laval[1,2], Sandra Pajon[1,2], Simona Fornarino[1,2], Brigitte Crouau-Roy[3] and Lluis Quintana-Murci[1,2,*]

[1]Unit of Human Evolutionary Genetics and [2]Centre National de la Recherche Scientifique, URA3012, Institut Pasteur, 25–28 Rue du Dr Roux, F-75015 Paris, France and [3]CNRS, Université de Toulouse, EDB (Laboratoire Evolution et Diversité Biologique), UMR 5174, 118 route de Narbonne, F-31062 Toulouse, France

Received June 28, 2011; Revised and Accepted August 21, 2011

The RIG-I-like receptors (RLRs)—RIG-I, IFIH1 (or MDA5) and LGP2—are thought to be key actors in the innate immune system, as they play a major role in sensing RNA viruses in the cytosol of host cells. Despite the increasingly recognized importance of the RLR family in antiviral immunity, no population genetic studies have yet attempted to compare the evolutionary history of its different members in humans. Here, we characterized the levels of naturally occurring genetic variation in the RLRs in a panel of individuals of different ethnic origins, to assess to what extent natural selection has acted on this family of microbial sensors. Our results show that amino acid-altering variation at *RIG-I*, particularly in the helicase domain, has been under stronger evolutionary constraint than that at *IFIH1* and *LGP2*, reflecting an important role for *RIG-I* in sensing numerous RNA viruses and/or functional constraints related to the binding of viral substrates. Such evolutionary constraints have been much more relaxed at *IFIH1* and *LGP2*, which appear to have evolved adaptively in specific human populations. Notably, we identified several mutations showing signatures of positive selection, including two non-synonymous polymorphisms in *IFIH1* (R460H and R843H) and one in *LGP2* (Q425R), suggesting a selective advantage related to the sensing of RNA viruses by IFIH and to the regulatory functions of LGP2. In light of the fact that some of these mutations have been associated with altered risks of developing autoimmune disorders, our study provides an additional example of the evolutionary conflict between infection and autoimmunity.

INTRODUCTION

Viral recognition by host cells and the subsequent induction of interferons (IFNs) are mediated by pattern-recognition receptors, which sense pathogen-associated molecular patterns (PAMPs) from diverse groups of microbes (1–3). Two systems of RNA virus detection and IFN induction have been characterized so far; the Toll-like receptor (TLR) (4–6) and the RIG-I-like receptor (RLR) systems (7,8). Although the endosomally located TLRs, specialized in the sensing of viral nucleic acids, have been studied from different angles—immunological, genetic and evolutionary (9,10), much less attention has been paid to the RLRs. Identified in 2004, RLRs are RNA helicases that function as cytosolic sensors of viral RNA infection to initiate and modulate antiviral immunity (7,8,11–13). To date, three RLR members have been identified: *RIG-I* (retinoic acid-inducible gene I, or *DDX58*), *IFIH1* (IFN-induced helicase C domain-containing protein 1, or *MDA5*) and *LGP2* (laboratory of genetics and physiology 2, or *DHX58*) (11). They share a number of structural similarities including their organization into different protein domains: (i) a C-terminal domain (CTD) mostly involved in RNA recognition, with a particular conformation in the repressor domain (RD) for RIG-I and LGP2; (ii) a central DExD/H box RNA helicase domain with the capacity to hydrolyze ATP and to bind and possibly unwind RNA; and (iii) two caspase recruitment domains (CARDs), mostly associated with cell death and inflammatory signalling pathways for RIG-I and IFIH1 (12,14–16). RLRs and TLRs use distinct and specific adaptors to initiate their

*To whom correspondence should be addressed at: Unit of Human Evolutionary Genetics, CNRS URA3012, Institut Pasteur, 25 rue du Dr. Roux, 75724 Paris, Cedex 15, France. Tel: +33 140613443; Fax: +33 145688639; Email: quintana@pasteur.fr

respective signalling cascades, which ultimately converge to the expression of type-I IFNs and proinflammatory cytokines (8,17–19).

The three RLRs, although activating the same signalling cascade, differ from each other in both their specific functions and the ligands which they sense. RIG-I and IFIH1 are both involved in the direct recognition of viral PAMPs, but they differ in their viral substrates and mechanisms of recognition [see (11) for an extensive review]. RIG-I detects a large range of RNA viruses including Newcastle disease virus, vesicular stomatitis virus, Sendai Virus, hepatitis C virus, influenza A virus, rabies virus, measles virus and respiratory syncytial virus (20–25). In turn, the viral substrates of IFIH1 are essentially restricted to picornaviruses, such as the encephalomyocarditis virus, rhinovirus and enterovirus (in particular Coxsackie B virus) (26–28). Some flaviviruses, such as Dengue type-2, or reoviruses have been shown to be sensed by both RIG-I and IFIH1, illustrating the occasionally overlapping functions of these two receptors (22). The different viral substrates of RIG-I and IFIH1 are mainly accounted for by the distinct structures of viral RNA the two sensors recognize; RIG-I is specialized in the sensing of ssRNAs bearing a 5′-triphosphate moiety as well as short dsRNAs (∼1 kb) (29–31), whereas IFIH1 binds long dsRNA (1–5 kb) (11,30). In both cases, these specificities allow them to distinguish viral ligands from self RNA (32). In contrast to RIG-I and IFIH1, LGP2 is not able to initiate antiviral signalling owing to its complete lack of a CARD domain, and has been shown instead to negatively regulate RIG-I signalling and positively regulate IFIH1 signalling (33).

In the past few years, genetic association studies and functional analyses have allowed a better understanding of the role of RLRs in human disease, suggesting a critical relationship between RLR polymorphisms, viral infections and autoimmunity (11). Genetic screens have led to the identification of a number of functional polymorphisms in both *RIG-I* and *IF1H1*, which are in some cases associated with altered cell responses to viral infection (34–36). The direct involvement of RLR polymorphisms in susceptibility to disease is well supported in the case of autoimmune disorders. Specifically, *IFIH1* has been replicably identified as a strong susceptibility gene for type-I diabetes (37–41). The non-synonymous mutation identified as the most strongly associated (rs1990760, A946T) has also been associated with other autoimmune diseases, including Graves' disease (42) systemic lupus erythematosus (43) and psoriasis (44). The A946T variant, which confers susceptibility to these diseases, also increases susceptibility to immunoglobin A deficiency, the most common heritable immunodeficiency (45), which is diagnosed more often than expected in patients with autoimmune disorders. Altogether, these findings point to an important role of the RLR system in human autoimmunity. Interestingly, the observation that type-I diabetes susceptibility alleles in *IFIH1* are found at high population frequencies (39) may attest to an advantage conferred by these variants against past infections (46,47). Indeed, the maintenance of risk alleles for autoimmune or inflammatory disorders, such as Crohn's disease, celiac disease or multiple sclerosis, as the by-product of past adaptation to pathogen exposure is supported by a growing body of data (reviewed in 46,48). Perhaps the best documented example is that of celiac disease, where evolutionary and functional analyses have shown that several celiac disease-risk alleles have been positively selected and that individuals carrying these alleles benefit from increased protection against certain infectious agents (46,49,50).

Despite the increasingly recognized importance of the RLR family in host defence, no population genetic study has so far attempted to compare the evolutionary history of its different members. In this context, the evolutionary genetics approach has been shown to provide an indispensable complement to clinical and epidemiological genetics studies in helping to delineate the essential and redundant functions of host defence genes in the natural setting (10,46,51). Here, we characterized the levels of naturally occurring sequence-based variation of the three members of the RLR family in a panel of healthy individuals, representative of various populations worldwide. We used these data to search for selection signatures, by means of allele frequency spectrum- and linkage disequilibrium (LD)-based tests, together with population differentiation indexes. Our analyses reveal that natural selection has targeted to different extents the members of the RLR family and identify various genetic variants as putative targets of positive selection, providing a number of good candidates that may affect the susceptibility to infectious and autoimmune diseases.

RESULTS

To gain insight into the levels of naturally occurring genetic variation at the three members of the RLR family, we resequenced them in a panel of 186 healthy individuals from sub-Saharan Africa, Europe and East-Asia (Supplementary Material, Table S1). A total of 21.3 kb was sequenced per individual (7.5, 8.0 and 5.8 kb for *RIG-I*, *IFIH1* and *LGP2*, respectively), 37% of which corresponded to exonic regions, the rest accounting for intronic and putative *cis*-regulatory regions (Supplementary Material, Table S2 and Fig. S1). Overall, we identified 206 single nucleotide substitutions, 8 insertions and 3 deletions (Supplementary Material, Table S3). Specifically, we detected 69 mutations at *RIG-I*, 78 at *IFIH1* and 59 at *LGP2*, of which 14, 16 and 11 corresponded to non-synonymous variants, respectively (Table 1 and Fig. 1).

Varying levels of nucleotide and functional diversity among RLRs

Globally, *RIG-I* presented lower levels of nucleotide diversity per site (π) ($4.5*10^{-4}$) than *IFIH1* and *LGP2*, which displayed similar values ($6.7*10^{-4}$ and $7.4*10^{-4}$, respectively) (Table 2). To compare these levels of nucleotide diversity with background genic expectations, we used the SeattleSNPs database, which reports the population sequence diversity of 327 genes associated with inflammatory pathways. As the SeattleSNPs database does not include Asian populations, we excluded our Asian sample from this analysis to have comparable panels. As a consequence, the overall levels of diversity at each gene were slightly increased, because the Asian population is known to present the lowest genetic diversity, as a result of greater genetic drift (52). *RIG-I* ($5.1*10^{-4}$)

Table 1. Non-synonymous changes identified in RLRs in this study

Gene	SNP position[a]	AA change[b]	Domain	Polyphen 2[c]	dbSNP	Allele frequency[d]		
						Africa	Europe	Asia
RIG-I	19 C>T	R7C	CARD1	P damaging	rs10813831	18.55	26.61	6.45
RIG-I	23 G>A	S8N	CARD1	Benign	–	0.81		
RIG-I	88 G>A	A30T[e]	CARD1	Benign	–	1.61		
RIG-I	89 C>T	A30V[e]	CARD1	Benign	–	1.61		
RIG-I	25272 A>G	M51V	CARD1	Benign	–			0.81
RIG-I	25333 G>A	R71H	CARD1	P damaging	rs72710678		0.81	
RIG-I	32398 A>C	I139L	CARD2	P damaging	rs78786043	4.84	1.61	
RIG-I	33636 C>T	S144F	CARD2	Benign	rs55789327	6.45	0.81	4.03
RIG-I	36758 A>G	N245S	Helicase	Benign	–	4.03		
RIG-I	44826 G>A	R546Q	Helicase	P damaging	rs61752945		2.42	
RIG-I	45914 T>A	D580E	Helicase	Ps damaging	rs17217280		11.29	4.84
RIG-I	59857 G>A	E773K	RD	Benign	–	4.03		
RIG-I	68933 T>C	I889T	RD	P damaging	–			0.81
RIG-I	69014 T>C	I916T	RD	Benign	rs77788008	4.03		
IFIH1	413 G>T	C138F	CARD2	P damaging	–			0.81
IFIH1	30124 A>G	K349R	Helicase	Benign	rs72650664		0.81	
IFIH1	35733 T>A	V366E	Helicase	P damaging	–	0.81		
IFIH1	35817 C>A	T394N	Helicase	Benign	–	0.81		
IFIH1	35931 T>C	V432A	Helicase	Ps damaging	–	0.81		
IFIH1	36835 G>A	R460H	Helicase	P damaging	rs10930046	52.42	100.00	92.74
IFIH1	36937 G>T	G494V	Helicase	P damaging	–	3.23		
IFIH1	40061 A>G	T575A	Helicase	Ps damaging	–	0.81		
IFIH1	40814 T>A	D655E	Helicase	Benign	–	0.81		
IFIH1	41432 A>C	R705S	Helicase	P damaging	–			0.81
IFIH1	41549 T>A	F744L	Helicase	P damaging	–			0.81
IFIH1	45994 G>A	R843H	CTD	P damaging	rs3747517	41.13	34.68	69.35
IFIH1	50181 A>G	I923V	CTD	Ps damaging	rs35667974		0.81	
IFIH1	50199 G>A	V929I	CTD	P damaging	–	1.61	0.00	
IFIH1	50767 G>A	A946T	CTD	Benign	rs1990760	11.29	53.23	21.77
IFIH1	50930 C>A	A967D	CTD	Benign	–	0.81		
LGP2	125 T>C	L42P	Helicase	P damaging	–	0.81		
LGP2	453 A>G	T76A	Helicase	Benign	rs34891485		2.42	
LGP2	511 G>A	R95Q	Helicase	Ps damaging	rs35118457	2.42	4.84	
LGP2	994 G>A	V129M	Helicase	P damaging	–	0.81		
LGP2	5906 C>T	R334C	Helicase	Ps damaging	rs76998797			1.61
LGP2	6748 A>G	Q425R	Helicase	Benign	rs2074158	84.68	22.58	16.94
LGP2	6856 A>G	N461S	Helicase	P damaging	rs34016093	2.42	13.71	1.61
LGP2	7048 T>C	I495T	RD	Benign	–			1.61
LGP2	8099 G>A	R523Q	RD	Benign	rs2074160	21.77	3.23	13.71
LGP2	8168 A>G	Q546R	RD	Benign	–		2.42	
LGP2	10056 G>A	R654H	RD	Benign	–		0.81	

A full description of all SNPs identified in this study is available in Supplementary Material, Table S3.
[a]SNP numbering is related to the ATG position, the first nucleotide corresponding to the ancestral allele.
[b]'AA change' stands for 'amino acid change', the first amino acid corresponding to the ancestral state.
[c]'Ps damaging' stands for 'possibly damaging', and 'P damaging' for 'probably damaging'.
[d]The frequencies (in %) of each SNP in the different populations refer to the derived allele frequency.
[e]Nucleotide substitutions at positions 88 and 89 are always observed together; consequently, the actual amino acid substitution is A30I, which is predicted as benign by Polyphen v2.

was found to be in the 17th percentile of genes presenting the lowest nucleotide diversity, whereas *IFIH1* ($7.9^{*}10^{-4}$) and *LGP2* ($8.2^{*}10^{-4}$) felt in the 54th and 58th, respectively. At the continental level, we observed fluctuations in the levels of nucleotide diversity across populations, the most extreme case being that of *IFIH1*, which displays very high levels of nucleotide diversity in Africa while very low in European and Asian populations (Supplementary Material, Fig. S2). We next focused on the comparison of non-synonymous/synonymous ratios between species (divergence, d_N/d_S) and within humans (polymorphism, π_N/π_S). Strong differences among genes emerged from our data: *RIG-I* exhibited a dramatic decay in its d_N/d_S ratio with respect to the other RLRs (Fig. 2), consistent with a strong deficit of fixed non-synonymous mutations (i.e. no non-synonymous mutations fixed between the chimpanzee and human lineages were detected). Furthermore, when comparing the π_N/π_S ratios between RLRs, we observed both a smaller proportion and a lower frequency of segregating non-synonymous polymorphisms for *RIG-I* (Fig. 2).

Finally, we predicted the functional effects at the protein level of the 41 non-synonymous mutations at the RLRs found to segregate in the population (Table 1), using the Polyphen v2 HumDiv algorithm (53). This method, which considers protein structure and/or sequence conservation for each gene, has been shown to be the best available predictor of

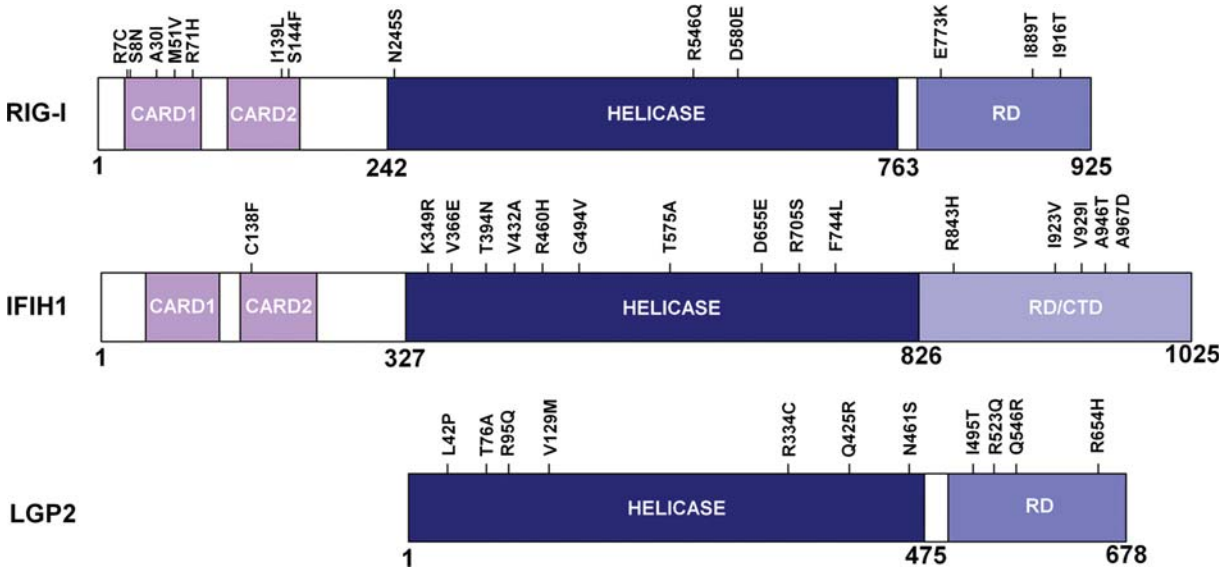


Figure 1. Distribution of the non-synonymous variants identified in this study across the RLRs. The location of each non-synonymous variant within the different protein domains is shown.

the fitness effects of non-synonymous variants (53,54). *IFIH1* exhibited the highest proportion of 'probably damaging' mutations and these mutations were found at high population frequencies. In contrast, *RIG-I* and *LGP2* displayed a relatively low proportion of probably damaging mutations, which were observed at low population frequencies (Fig. 3). In light of this, *RIG-I* and *LGP2* seem to be much more constrained with respect to non-synonymous mutations that are likely to have an impact on protein function.

Low amino acid-altering variation in the RIG-I helicase domain

We next aimed to understand whether constraints of amino acid sequence variation are evenly distributed among the different protein domains. To this end, we compared the levels of non-synonymous and synonymous polymorphisms, considering both the number of nucleotide substitutions (p_N/p_S) and the average number of differences between pairs of individuals (π_N/π_S), the latter of which also takes into account population frequency data. With respect to the CARD domain, involved in cell death and inflammation signalling, *IFIH1* exhibits a lower occurrence of non-synonymous mutations than *RIG-I*, as attested to by the p_N/p_S ratio (Fig. 4A). Moreover, non-synonymous mutations in the CARD domain for *IFIH1* were generally at low frequency; the π_N/π_S of *IFIH1* was in the lowest 25% of the resampled values while that of *RIG-I* was among the 14% highest, although it did not reach statistical significance. With respect to the Helicase domain, *RIG-I* exhibits a lower p_N/p_S ratio than *IFIH1* and *LGP2* (Fig. 4B), as well as a significant deficit in high-frequency non-synonymous mutations (resampling $P < 0.0001$). Finally, with respect to the CTD/RD domain, primarily responsible for RNA recognition, *RIG-I* displayed lower p_N/p_S and π_N/π_S ratios than the other two RLRs, although this did not reach statistical significance (Fig. 4C). Overall, the comparisons of p_N/p_S and π_N/π_S ratios per domain illustrated lower occurrences and frequencies of non-synonymous mutations in the helicase and CTD domains at *RIG-I* than at *IFIH1* and *LGP2*, a deficit that was significant in the case of the helicase domain.

Significant deviations from neutrality at *IFIH1*

We tested for deviations of the allele frequency spectrum from neutral expectations over the genomic regions encompassing the three RLRs. We used summary statistics based on within-population allele frequency distribution, such as Tajima's *D*, Fu & Li's D^* & F^* and normalized Fay & Wu's *H* (reviewed in 55). As these tests are known to be sensitive to demographic effects, we considered two previously validated demographic models based on a set of unlinked noncoding regions sequenced in a panel of populations from Africa, Europe and Asia (56,57). The main difference between the two models is that the one of Voight *et al*. (57) does not consider inter-continental population migration, while the one of Laval *et al*. (56) does. Our analyses unmasked an excess of low-frequency variants for the three genes, indicative of either negative or positive selection. This was illustrated by the significantly negative values of the various neutrality tests, which were virtually restricted to the European population (Table 2). To be conservative, we focused mainly on those signatures of selection that turned out to be significant after considering both demographic models. The strongest signals were clearly detected at *IFIH1*, as illustrated by the significantly negative Tajima's *D*, Fu & Li's D^* and F^* in Europe. With respect to the Asian population, we also detected a signature, although weaker, of positive selection at *IFIH1,* as attested to by the significantly negative value of the Fay and Wu's *H* statistics, which detects an excess of high-frequency-derived alleles. This signature was supported by the Tajima's *D* values, which were significantly negative (or close to significance) for the two models in Asia (Table 2). Overall, the strongest

Table 2. Mean diversity indices and neutrality tests across *RIG-I*, *IFIH1* and *LGP2* genomic regions

	RIG-I Africa ($N = 124$)	Europe ($N = 124$)	Asia ($N = 124$)	Global ($N = 372$)	*IFIH1* Africa ($N = 124$)	Europe ($N = 124$)	Asia ($N = 124$)	Global ($N = 372$)	*LGP2* Africa ($N = 124$)	Europe ($N = 124$)	Asia ($N = 124$)	Global ($N = 372$)
H	36	24	14	60	29	19	20	59	51	24	20	79
Hd	0.93	0.87	0.80	0.89	0.87	0.80	0.74	0.85	0.95	0.81	0.72	0.89
Syn	8	4	2	10	7	2	2	10	4	2	0	5
Non-syn	8	6	4	12	11	4	6	16	6	7	5	11
S	45	27	16	66	48	20	25	77	36	28	18	59
Singletons	6	10	3	17	18	12	6	34	16	12	4	27
INDELS	3	1	0	4	3	0	1	3	1	3	2	4
π (10^{-4})	6.1	3.9	3.1^{+}	4.6	11.0	$2.0^{*/++}$	3.6	6.7	6.2	5.3	3.9	7.4
θ_W (10^{-4})	11.6	7.0	4.1	14.1	11.2	4.7	5.8	14.9	11.5	8.9	5.7	15.6
TD	−1.45	-1.28^{+}	−0.66	–	−0.05	$-1.61^{*/+}$	-1.12^{+}	–	−1.40	-1.21^{+}	−0.90	–
D*	0.62	-1.88^{+}	−0.01	–	−2.22	$-3.86^{**/+}$	−0.53	–	−2.85	-2.50^{+}	−0.32	–
F*	−0.28	-1.97^{+}	−0.30	–	−1.58	$-3.59^{**/++}$	−0.91	–	−2.71	$-2.38^{*/+}$	−0.65	–
Hn	−0.37	−0.03	−0.19	–	0.48	−0.46	-2.21^{*}	–	−0.04	−0.42	−1.10	–

N, number of chromosomes sequenced; H, number of haplotypes; Hd, haplotype diversity; Syn, number of synonymous mutations; Non-syn, number of non-synonymous mutations; S, number of segregating sites; π, nucleotide diversity per site from average pairwise differences; θ_W, nucleotide diversity per site from number of segregating sites; TD, Tajima's *D*; D, Fu & Li's *D**; F, Fu & Li's *F**; Hn, Normalized Fay & Wu's *H*.
$^{**/*}P$-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Voight *et al.* (57); $^{++/+}P$-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Laval *et al.* (56). *P*-values were obtained from coalescent simulations according to the former model that considers each continental population separately, and to the latter model that considers inter-continental population migration.

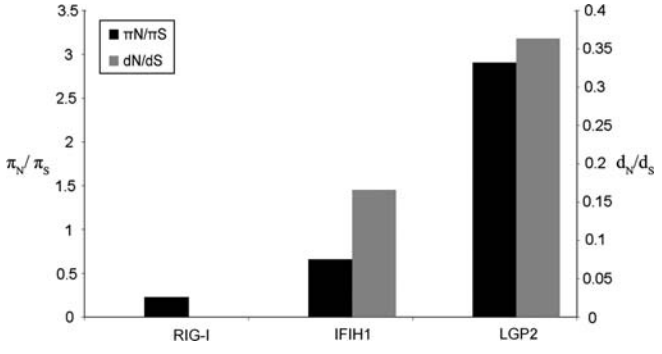


Figure 2. Degree of non-synonymous diversity observed at *RIG-I*, *IFIH1* and *LGP2*. Non-synonymous/synonymous diversity (π_N/π_S) based on the mean number of pairwise differences between human sequences (π), and non-synonymous/synonymous divergence (d_N/d_S) based on the number of fixed differences between human and chimpanzee for *RIG-I*, *IFIH1* and *LGP2*. π_N/π_S ratios are represented by black bars; grey bars correspond to d_N/d_S ratios.

deviations from neutrality were observed in *IFIH1*, especially in Europe.

Strong population differentiation at *IFIH1* and *LGP2*

To gain further insight into the signatures of selection detected by the previous sequence-based neutrality tests, we estimated the degree of population differentiation by means of F_{ST}. Local positive selection is known to increase the levels of population differentiation with respect to neutrally evolving loci (55,58–61). When calculating F_{ST} averaged across each gene, both worldwide and population-pairwise, *IFIH1* and *LGP2* were strongly differentiated between populations ($F_{ST} = 0.25$ and 0.40, respectively) with respect to *RIG-I* ($F_{ST} = 0.05$). The very high F_{ST} displayed by *LGP2* ($F_{ST} = 0.40$) with respect to the mean across the genome ($F_{ST} = 0.15$) suggests that this gene has been targeted by population-specific positive selection. The analysis of pairwise F_{ST} revealed that the strong differentiation observed at *IFIH1* and *LGP2* is mainly accounted for by differences between Africans and non-Africans (*IFIH1* $F_{ST} = 0.30$, 0.23 and 0.12 and *LGP2* $F_{ST} = 0.46$, 0.51 and 0.05 for Africa/Europe, Africa/Asia and Europe/Asia, respectively), whereas all pairwise F_{ST} gave similarly low values for *RIG-I* ($F_{ST} = 0.05$, 0.06 and 0.06).

At the level of individual single nucleotide polymorphisms (SNPs) (i.e. 206 SNPs), strongly differentiated SNPs appeared to be restricted to *IFIH1* and *LGP2,* when estimating F_{ST} as a function of expected heterozygosity (Fig. 5). *IFIH1* presented 11 highly differentiated SNPs in the Africa/Europe comparison, with respect to the rest of the genome: SNPs 29773 ($P = 0.03$), 36033 ($P = 0.009$), 36835 ($P = 0.007$), 36947 ($P = 0.009$), 38047 ($P = 0.009$), 38091 ($P = 0.013$), 40330 ($P = 0.009$), 45935 ($P = 0.009$), 46093 ($P = 0.008$), 49917 ($P = 0.008$) and 51114 ($P = 0.009$), including one non-synonymous SNP in the helicase domain (SNP 36835, R460H). Three of these highly differentiated SNPs were also significant outliers in the Africa/Asia comparison (SNPs 40330, 45935 and 51114, $P = 0.013$), while the non-synonymous SNP 36835 showed a very high F_{ST} but did not reach statistical significance. The

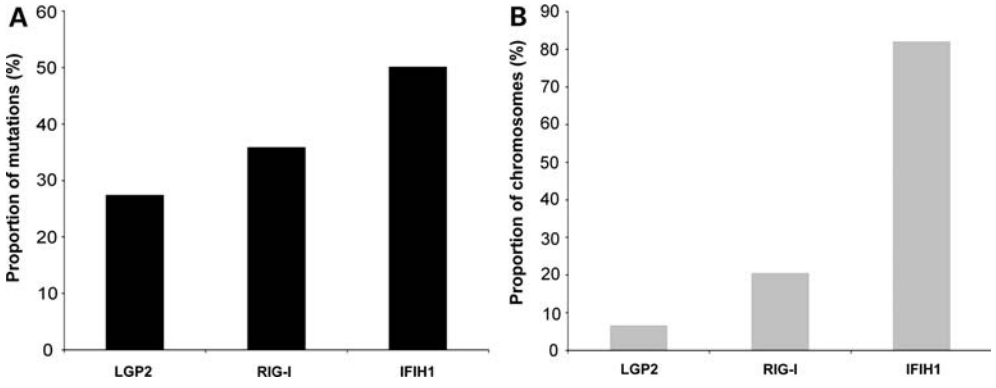


Figure 3. Distribution of functional diversity among the different RLRs. The functional impact of each non-synonymous mutation observed at *RIG-I*, *IFIH1* and *LGP2* was predicted by means of the Polyphen algorithm v2 HumDiv. (**A**) Proportion of 'Probably damaging' mutations among non-synonymous mutations and (**B**) proportion of chromosomes carrying at least one 'probably damaging' mutation.

derived state of SNP 36835 was found at a frequency of ~52% in Africa, at fixation in Europe and at ~93% frequency in Asia (Supplementary Material, Table S3). Concerning LD levels among highly differentiated SNPs at *IFIH1*, we found a group of seven SNPs (SNPs 36033, 36947, 38047, 38091, 40330, 45935 and 51114) and a group of four (29773, 36835, 46093 and 49917) in high LD ($r^2 > 0.8$) in Africa, and a group of seven SNPs (SNPs 36033, 36835, 36947, 38047, 38091, 46093 and 49917) in complete LD in Asia. No SNPs with a MAF > 0.05 were found in high LD ($r^2 > 0.8$) in Europe (Supplementary Material, Fig. S3).

With respect to *LGP2*, we detected five SNPs (SNPs −751, −215, 3931, 5784 and 6748) that were outliers for the Africa/Europe ($P = 0.009$, 0.01, 0.01, 0.016 and 0.042, respectively) and the Africa/Asia ($P = 0.001$, 0.023, 0.023, 0.054 and 0.043, respectively) comparisons (Fig. 5). Three of these five SNPs are in LD in all populations, namely SNP −215 with SNPs 3931 and 5784. SNP −751 is in strong LD with these SNPs in Africa and Europe only (Supplementary Material, Fig. S3). Interestingly, SNPs −751 and −215 are located in the potential *cis*-regulatory region of the gene. The SNP 6748, which is in low LD with the 'potential regulatory' block of LD (Supplementary Material, Fig. S3), is a non-synonymous mutation (Q425R) located, as for *IFIH1*, in the helicase domain.

The footprints of recent positive selection targeting *IFIH1* and *LGP2*

Finally, we searched for signatures of population-specific positive selection by means of several LD-based tests. Specifically, we used statistics based on specific signatures of recent positive selection including low levels of derived intra-allelic nucleotide diversity (DIND) (9) and extended haplotype homozygosity [integrated haplotype scores (iHS), cross population extended haplotype homozygosity (XP-EHH) and extended haplotype homozygosity (EHH)] (61–63). The strongest signatures of positive selection were again observed at *IFIH1* and *LGP2*, identifying potential targets within both of them. For *IFIH1*, SNP 36835, which is found to be highly differentiated in Europe and to a lesser extent in Asia based on F_{ST}, was not assessed for EHH in Europeans because it reaches fixation. In Asia, the levels of EHH around the derived allele of SNP 36835 were similar to those observed in Africa (Supplementary Material, Fig. S4). In Africa, SNP 36835 (together with SNPs 46093 and 49917 which are in complete LD, see Supplementary Material, Fig. S3) appeared to be an outlier in the DIND test when using the recombination rates of UCSC ($P = 0.006$ for Laval's model and $P = 0.039$ for Voight's model) but not when using HapMap ($P = 0.164$ for Laval's model and $P = 0.181$ for Voight's model). In this view, LD-based tests did not support the notion that the signature of positive selection at SNP 36835 revealed by F_{ST} was of recent nature. In contrast, the DIND test identified another non-synonymous mutation (SNP 45994, R843H) as a clearer outlier in Africa ($P = 0.000$ Laval/UCSC; $P = 0.002$ Voight/UCSC; $P = 0.023$ Laval/HapMap; $P = 0.027$ Voight/HapMap) and, to a lesser extent, in Asia ($P = 0.0128$ Laval/UCSC; $P = 0.048$ Voight/UCSC; $P = 0.080$ Laval/HapMap; $P = 0.116$ Voight/HapMap) (Supplementary Material, Fig. S5). Likewise, EHH analysis for SNP 45994 revealed a much higher homozygosity for the derived haplotype in Africa and Asia (Supplementary Material, Fig. S4). Note that SNPs 36835 and 45994 display intermediate LD levels in Africa and low in Asia (Supplementary Material, Fig. S3) (LD was not assessed in Europe because of the derived allele fixation of SNP 36835). Finally, the DIND test provided strongly significant values for two additional non-coding SNPs in complete LD in Africa (Supplementary Material, Fig. S3), SNPs 36531/41712 ($P = 4 \times 10^{-4}$ Laval/UCSC; $P = 0.004$ Voight/UCSC; $P = 0.018$ Laval/HapMap; $P = 0.022$ Voight/HapMap), but they present a frequency lower than 0.20, at which the DIND test has been shown to have a higher false discovery rate (9).

With respect to *LGP2*, the non-synonymous SNP 6748, which was found to be highly differentiated between Africans and non-Africans based on F_{ST}, exhibited an iHS value in Europe of 2.3, suggesting the action of positive selection on the ancestral allele. In Asia, the iHS value was of 1.401, which although not higher than the significance threshold of 2, represents a high value with respect to the observed iHS of 0.391 in Africa (Supplementary Material, Table S3). The ancestral state of this SNP is found at 15% frequency in Africa, 77% in Europe and 83% in Asia. In further support of this, EHH plots around SNP 6748 revealed higher haplotype

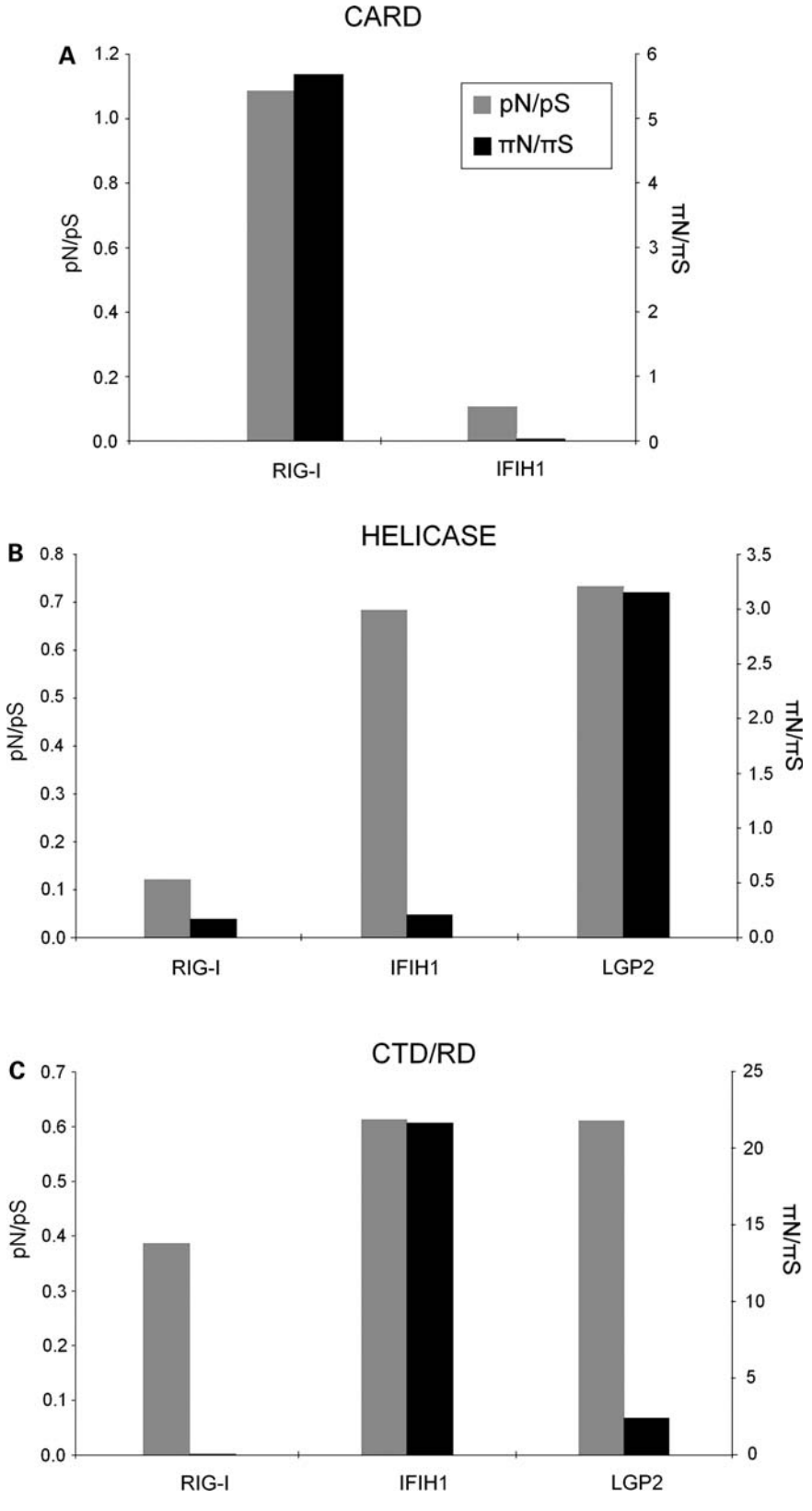


Figure 4. Occurrence and frequency of non-synonymous mutations across the different RLR protein domains. Non-synonymous/synonymous polymorphisms in humans (p_N/p_S), and non-synonymous/synonymous diversity (π_N/π_S) assessed by the mean numbers of pairwise differences between human sequences (π), for *RIG-I*, *IFIH1* and *LGP2* in (**A**) the CARD domain, (**B**) the helicase domain and (**C**) the CTD/RD domain. Grey and black bars represent p_N/p_S and π_N/π_S, respectively.

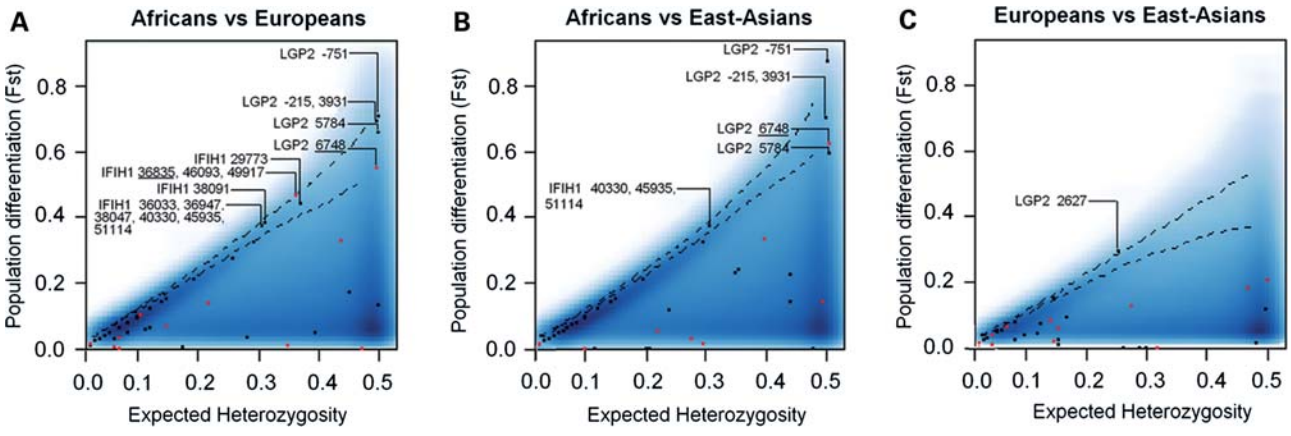


Figure 5. Levels of population differentiation displayed by the three RLR members. F_{ST} as a function of heterozygosity per SNP is shown between (**A**) Africans versus Europeans, (**B**) Africans versus East-Asians and (**C**) Europeans versus East-Asians, for *RIG-I*, *IFIH1* and *LGP2*. The 95th and 99th percentiles of the HGDP–CEPH genotyping data set using the same individuals are represented with dashed lines, while the density blue area corresponds to the 99.9th percentile. Black and red points represent silent and non-synonymous SNPs, respectively. For each outlier SNP, the gene name is indicated, followed by position to ATG. Groups of outlying SNPs separated by a comma correspond to SNPs in complete LD. Outlying non-synonymous SNP positions are underlined.

homozygosity associated with the ancestral allele among non-African populations (Supplementary Material, Fig. S6). With respect to the other SNPs that were found to be highly differentiated in the F_{ST} analysis, SNPs −751, −215 and 3931 showed higher haplotype homozygosity associated with the derived allele in Europe and Asia (Supplementary Material, Fig. S6). Note that although in some cases the levels of homozygosity extended over neighbouring genes, recombination rate variation and additional tests of selection peaked around the RLR genes (Supplementary Material, Figs S4 and S6). For example, a number of genes are located in the vicinity of *LGP2*, but the high recombination rate variation of the region reduced to four the number of putative genes explaining the signature of selection observed. Furthermore, within this genomic region, the highest iHS and F_{ST} values were observed in *LGP2*, making this gene the most likely target of the signature of selection detected here.

DISCUSSION

Detecting to what extent natural selection has targeted host microbial sensors, such as the RLRs, can provide insight into their respective biological roles in host defence as well as into the genetic basis of their phenotypic variation. Because RLRs constitute key players in the initial sensing of RNA viruses, genetic variation at these genes may impact downstream immune responses and therefore contribute to the susceptibility or pathogenesis of various infectious diseases. This makes these genes putative substrates of natural selection. Results presented herein showed indeed that natural selection has driven the evolution of RLRs, in different directions and with different intensities.

First, we showed that the degree of gene and protein diversity vary substantially among the three RLR members. *RIG-I* is clearly the most constrained RLR, as attested by its low levels of nucleotide diversity and population differentiation as well as by its low tolerance of amino acid-altering variation. Indeed, *RIG-I* presents fewer, and lower frequencies of, non-synonymous mutations than *IFIH1* and *LGP2*. Such a constraint may reflect the pressure imposed by the specific viruses recognized by RIG-I with respect to IFIH1; RIG-I senses a much larger variety of viruses, including several families such as paramyxoviridae, rhabdoviridae, filoviridae, arenaviridae and coronaviridae, whereas IFIH1 is mainly involved in the recognition of picornaviruses [see (11) and references therein]. In addition, the structural nature of the viral substrates of RIG-I may also account for the stronger constraint observed. Indeed, RIG-I and IFIH1 detect short and long dsRNA, respectively, but RIG-I further senses 5′PPPssRNA (20,30,31). Short dsRNA and 5′PPPssRNA binding could require more strict binding sites, thus allowing for less protein flexibility, compared with long dsRNA. In this context, it is interesting to note that the C-terminal and helicase domains of RIG-I, notably involved in RNA recognition, are more constrained with respect to amino acid-altering variation than in the other RLRs. Our results suggest that, in the context of sensing RNA viruses, the function fulfilled by RIG-I may allow for less protein variation than that of IFIH1. However, viral presence may not be the only force driving the evolution of these microbial sensors, as suggested by experimental data in mice. Indeed, both *RIG-I*$^{-/-}$ and *IFIH1*$^{-/-}$ mice are highly susceptible, to varying extents, to specific RNA virus infections. However, while *IFIH1*$^{-/-}$ mice grow healthily without any developmental abnormalities before 24 weeks of age, most *RIG-I*$^{-/-}$ mice have been shown to be embryonic lethal due to developmental defects (18,20). These observations suggest a generally more important role for RIG-I compared with IFIH1, which may go well beyond immunity to infection. In light of this, our evolutionary findings in humans could also reflect different levels of involvement for these two genes in other functions, such as reproduction and development. This would not be surprising, considering that mutations in various NOD-like receptors—another major family of microbial sensors—have been shown to lead to reproductive wastage or abnormalities in early development both in mice and humans (64–67).

Secondly, the patterns of diversity displayed by *IFIH1* and *LGP2* indicate that these two genes, in contrast to *RIG-I*, have

evolved under more relaxed selective constraints. Most importantly, different lines of evidence based on the allele frequency spectrum, the levels of population differentiation and LD-based tests clearly show that *IFIH1* and *LGP2* appear to evolve adaptively. The action of positive selection at *IFIH1* has been previously proposed, with the non-synonymous SNP 36835 (R460H) suggested as a plausible target (47). Our study extends previous observations and clarifies the nature of the selective event targeting *IFIH1*. Our F_{ST} analysis clearly indicates that the R460H variant is the most differentiated non-synonymous SNP in Eurasian populations, pointing to this variant as the genuine target of selection. Further support for the action of positive selection comes from the median-joining network, where the branch carrying the derived 460H allele includes virtually all Europeans and Asians (Supplementary Material, Fig. S7). In addition, the absence of signals of recent positive selection at this SNP based on LD-based tests (iHS test) (61), which is not surprising given the very high frequencies of the derived allele (100 and 92.7% in Europe and Asia, respectively), suggests that the selective event targeting R460H is of a sufficiently ancient nature to have allowed recombination to reduce haplotype lengths. In support of this, the results of the XP-EHH test over the entire gene were close to the threshold of 2 in Europe and East-Asia, but were not significant (data not shown). Interestingly, the ancestral allele at this position (460R) has been associated with higher resistance against psoriasis, a chronic autoimmune disease caused by an abnormal speeding up of growth cycle for skin cells (68). It is thus tempting to hypothesize that the derived 460H variant has increased in frequency among non-Africans owing to an increased advantage in host defence, despite its deleterious, rather minor effects associated with higher risk of psoriasis. This provides an additional example that supports the notion whereby the present-day increased incidence of autoimmune, inflammatory or allergic diseases may result from past adaptation to infectious agents (46,48).

In addition, our study has identified another non-synonymous variant at *IFIH1*, the SNP 45994 (R843H), which exhibit signatures of positive selection in Africa and, to a lesser extent, in Asia, based on LD tests. The R843H variant has been suggested to behave neutrally or almost-neutrally and to have hitch-hiked with the genuine selective variant, the SNP 36835-R460H variant (47). Our data challenge this view, because the strongest signatures on SNP 45994 were found in Africa, while those detected at SNP 36835 are restricted to Europe and Asia. Moreover, the signature observed at SNP 45994 is provided by LD-based tests, while that at SNP 36835 is based on F_{ST}. These two signatures—low diversity associated with the derived allele in Africans and Asians for SNP 45994 versus strong differentiation between African and non-African populations for SNP 36835—cannot convincingly be explained by each other. It is thus likely that the selection signatures we detected at SNP 45994 reflect a selective event of different and more recent nature than that observed at SNP 36835. Interestingly, SNP 45994 has been associated with altered risk for type-1 diabetes in a European cohort (39), together with another non-synonymous SNP in strong LD ($r^2 = 0.6$), i.e. SNP 50767-A946T (37–41). SNP 50767 has been, in turn, associated with systemic lupus erythematosus and IgA deficiency through genome-wide association studies (43,45). However, we did not detect any selection signal on SNP 50767, and functional studies have failed to find a loss of function phenotype when focusing on SNP 50767 alone (34,69). Instead, it seems that the effect could be due to SNP 45994 (R843H) or at least that this variant tags a haplotype conferring both differential expression of *IFIH1* and an altered risk for auto-immune diseases (34,69). Our results indicate that the positively selected allele corresponds to the haplotype associated with a lower *IFIH1* expression and a higher protection against autoimmune diseases. Note that all these association studies have been conducted in European cohorts only and still have to be replicated in populations with different ethnic backgrounds. Our results are nevertheless consistent with two events of positive selection targeting the haplotype carrying the derived state of SNP 45994, which differ in time and space; a more ancient event in Asia (frequency of 69%) and a more recent one in Africa (41% frequency) (LD-based tests, which are powerful to detect recent events, provided stronger signatures in Africa with respect to Asia). With respect to Europeans, where the derived allele presents a 35% frequency and no selective signature was detected, SNP 45994 appears to have been evolving neutrally or under negative selection.

Thirdly, our analyses pinpointed two independent regions in *LGP2* which are strongly differentiated between African and non-African populations, based on F_{ST}. The first encompasses four SNPs in LD (−751, −215, 3931 and 5784), including the two SNPs −751 and −215 in potential regulatory regions, while the second involves a non-synonymous variant (SNP 6748, Q425R, rs2074158) (Supplementary Material, Fig. S3). As neutrality tests have indicated an excess of singletons in Europe, this could reflect either positive or background selection in this region (or more generally in non-Africans). Nevertheless, LD-based tests showed an excess of homozygosity associated with the derived alleles at the four SNPs in LD (see EHH results, Supplementary Material, Fig. S6, especially SNP −215) in Europe and Asia, supporting the action of positive selection among non-Africans. Furthermore, as illustrated in the network, SNPs −215 and 3931 in particular ($r^2 = 1$; $D' = 1$ in all three populations) define the branch where virtually all European and Asian lineages are observed, pointing to these mutations as the potential targets of positive selection (Supplementary Material, Fig. S8). With respect to the non-synonymous SNP 6748, we found a higher homozygosity for the ancestral haplotype in Europe and Asia (Supplementary Material, Fig. S6), together with an iHS value of 2.3 in Europe (1.4 in Asia and 0.4 in Africa) (Supplementary Material, Table S3), indicating that the ancestral allele of SNP 6748 (425Q) is the most likely target of positive selection among non-Africans. This variant, which is located in the RNA-binding helicase domain, could have a functional impact on the RNA sequestration by LGP2 (12) and compete with RIG-I and IFIH1 in ligand recognition. Under this hypothesis, the derived 425R allele would have risen in frequency by genetic drift and then, after the out-of-Africa exodus, the ancestral allele would have become advantageous in Europe and Asia. Taken together, our results suggest that two independent events of positive selection have targeted

LGP2, highly differentiating Africans and non-Africans, and support the notion that some genes appear to be more subject to recurrent events of positive selection (70).

More generally, in the context of pathogen sensing of RNA viruses, RLRs are not the only family that induce antiviral immunity but are part of a concerted antiviral program mediated by a variety of microbial sensors including the TLRs (11,13). Specifically, among the 10 functional members of the TLR family in humans, the endosomally located TLR3, TLR7 and TLR8 have been characterized as principal sensors of RNA viruses, while the other TLRs are responsible for detecting bacteria, fungi and DNA viruses (4,6,71). Earlier work from our laboratory has shown that endosomally located TLRs evolve under strong purifying selection, indicating their essential role in host survival, while the remaining TLRs display higher levels of immunological redundancy (9). When comparing the two groups of microbial sensors specialized in the detection of viral RNA (i.e. RLRs versus TLRs 3/7/8), we found that RLRs and TLRs display similar levels of nucleotide diversity ($\pi = 6.2^{*}10^{-4}$ and $6.4^{*}10^{-4}$, respectively) and d_N/d_S ratios (0.18 and 0.23, respectively). However, we observed both a higher proportion and a higher frequency of segregating non-synonymous polymorphisms at RLRs than at TLRs, as attested by the 10-fold higher π_N/π_S ratio of RLRs with respect to TLRs (1.26 and 0.12, respectively). It should be noted that the very high π_N/π_S ratio of RLRs is mainly accounted by the values of *LGP2*; however, the combined π_N/π_S ratio of *RIG-I* and *IFIH1* is 0.44, a value which is still three times higher than that observed at the TLRs. Furthermore, among all non-synonymous mutations identified, RLRs displayed a higher proportion of mutations predicted as having an impact on protein function ['probably damaging' mutations as predicted by Polyphen v2 (53)] with respect to the TLRs (39% for RLRs versus 15% for TLRs 3/7/8). Overall, our data suggest that the TLRs specialized in the detection of viral RNA have been under stronger evolutionary constraints than their RLR counterparts. Nonetheless, RLRs have been under stronger selective constraints than the other group of TLRs, the cell-surface TLRs. This group not only displays generally high levels of genetic and functional diversity but, in contrast to RLRs and TLR3/7/8, can accumulate stop mutations at high population frequencies (9). As TLRs 3/7/8 and RLRs share partially overlapping functions (8,17–19), the generally lower constraints observed at RLRs taken as a group could reflect some level of redundancy in their innate immunity functions.

In summary, our study has shown that natural selection has targeted differently the three members of the RLR family of microbial sensors. Evolutionary constraint in amino acid-altering variation at *RIG-I*, especially in its RNA recognition domains, suggests an important role for this gene in fighting against numerous and severe RNA viruses, structurally strict mechanisms of recognition or, more generally, an essential role in host survival. In turn, such evolutionary constraints have been much more relaxed at *IFIH1* and *LGP2*, which appear to have evolved adaptively. We identified three non-synonymous mutations, two in *IFIH1* (R460H and R843H) and one in *LGP2* (Q425R), as well as several SNPs located in the regulatory region of *LGP2* showing signatures of positive selection. Our study provides evolutionary evidence that these mutations have conferred a selective advantage to specific human populations, which may have impacted the sensing of RNA viruses by IFIH and the regulatory effect of LGP2 on RIG-I and/or IFIH1. Delineation of the functional impact of these mutations on immune responses to viral infections is now needed to understand the extent to which these variants have conferred a phenotypic advantage on the human host. Given the role of the RLRs in immunity to infection, particularly to RNA viruses, the variants identified as potential targets of positive selection should be further studied in the context of disease, as they may play a role in the susceptibility to, or pathogenesis of, infectious and autoimmune diseases.

MATERIAL AND METHODS

DNA samples

We resequenced *RIG-I, IFIH1* and *LGP2* in a total of 372 chromosomes from the human genome diversity panel (HGDP)–centre d'etude du polymorphisme humain (CEPH) panel (72). Specifically, this subpanel includes 62 sub-Saharan Africans, 62 Europeans and 62 East-Asians. Sub-Saharan African populations were composed of 19 Bantu from Kenya, 21 Mandenka from Senegal and 22 Yoruba from Nigeria; European populations include 20 French, 14 Italians, 6 Orcadians and 22 Russians; and East-Asian populations were composed of 10 Japanese, 4 Cambodians, 15 Han Chinese and 33 individuals from Chinese minorities. For a complete description of this HGDP–CEPH subpanel, see Supplementary Material, Table S1. This study was approved by the Institut Pasteur Institutional Review Board (no. RBM 2008.06).

Resequencing

For each gene, we resequenced the whole of the exonic region and at least an equivalent amount of non-exonic sequence, including intronic, 5′ and 3′ regions (Supplementary Material, Table S2). We used NM_014314.3, NM_022168.2 and NM_024119.2 as reference sequences for *RIG-I*, *IFIH1* and *LGP2*, respectively. Sequence files and chromatograms were inspected using the GENALYS software (73). All sequences were analysed by two different operators to avoid SNP discovery errors, and ambiguous polymorphisms were systematically reamplified and resequenced. Concerning nucleotide numbering, SNP positions correspond to locations in genomic DNA, taking the A of the ATG translation initiation codon in the reference sequence as the +1 position). For various neutrality tests that need the definition of the ancestral/derived state at each SNP, we used the orthologous regions of each of the three RLRs from various species. To determine ancestral states at each SNP by parsimony, we used the UCSC database to retrieve the orthologous sequences of chimpanzee, gorilla, orangutan and rhesus macaque.

Statistical analyses

Haplotype reconstruction was performed by means of the Bayesian statistical method implemented in Phase (v.2.1.1) (74). We applied the algorithm five times, using different

randomly generated seeds, and consistent results were obtained across runs. LD levels between SNPs were determined for the three RLRs in Africa, Europe and Asia separately, using Haploview 4.0 (75). We assessed the d_N/d_S (divergent non-synonymous sites/divergent synonymous sites) ratio for the three genes using DnaSP package v. 5.1 (76); divergent sites refer to positions that are different between the human and chimpanzee lineages, whereas polymorphic sites refer to alleles that are segregating within humans. The different summary statistics, such as the number of segregating sites (S), haplotype diversity (Hd), the average number of pairwise differences (π) and the sequence-based neutrality tests, such as Tajima's *D*, Fu and Li's D^*, F^* and normalized Fay and Wu's *H* tests were also performed using DnaSP. *P*-values for the various neutrality tests were estimated from 10^4 coalescent simulations, performed using SIMCOAL 2.0 (77), under a finite-site neutral model and considering the recombination rate of the concerned region as reported in UCSC (78) and HapMap Phase II (79). Each of the 10^4 coalescent simulations was conditional on the sample size and the number of segregating sites observed in each gene. To correct for the mimicking effects of demography on the patterns of diversity, we considered two previously validated demographic models based on resequencing data from noncoding autosomal regions in a set of populations similar to ours (i.e. African, European and Asian) (56,57). The main difference between these two demographic models is that the model of Laval *et al*. (56) considers inter-continental population migration.

To detect events of positive selection, we used a number of haplotype-based tests as well as levels of population differentiation. Specifically, we used the DIND test based on the ratio π_A/π_D, where π_A and π_D are the levels of nucleotide diversity associated with the haplotypes carrying the ancestral and the derived allele, respectively (9). The rationale behind this test is that a derived allele under positive selection that is at high population frequency should present lower levels of nucleotide diversity at linked sites than expected. We also used tests based on the levels of haplotype homozygosity, such as the extended haplotype homozygosity (62) using the EHH web calculator (http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl) (80) and the XP-EHH test (63). When available, we also used the integrated iHS obtained from HapMap Phase II (61). To assess the levels of population differentiation for the entire SNP panel, we used the F_{ST} statistics derived from the analysis of variance (81). To identify SNPs presenting extreme levels of population differentiation, we compared the observed F_{ST} values at each SNP within RLRs against the F_{ST} distribution of 659 000 SNPs genotyped in the same panel of individuals (82). F_{ST} comparisons were conditioned to SNPs presenting similar expected heterozygosity. Median-joining networks were built using NETWORK 4.6 (83).

Functional impact of non-synonymous mutations and protein domain analysis

The fitness status of all amino acid-altering mutations (i.e. benign, possibly damaging and probably damaging) was predicted using the Polyphen algorithm v2 HumDiv (53). Nucleotide diversity indexes such as p_N/p_S and π_N/π_S were assessed for each protein domain (CARD1, CARD2, helicase and CTD/RD) and compared among RLRs. The significance of differences in π_N/π_S ratios observed between domains was assessed by resampling. For each domain, we resampled 10 000 times a region of the same length as the domain of interest at a given RLR from the entire coding region of the two other RLRs put end to end and alternatively ordered first or second. The π_N/π_S value observed in a specific domain at one RLR was then compared with that obtained in the resampling along the coding regions of the other two RLRs. We then determined where each specific domain falls in the resampling in terms of π_N/π_S and deduced the corresponding *P*-value. The domain division was done according to Baum and Garcia-Sastre (13). Details for the CARD1/CARD2 division were retrieved from the Uniprot database (84).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We would like to thank Katherine Siddle for critical reading of the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This work has been supported by the Institut Pasteur, the ANR (ANR-08-MIEN-009-01), the Fondation pour la Recherche Médicale, the CNRS, Merck-Serono and a EPFL-Debiopharm Life Sciences Award to L.Q.-M.

REFERENCES

1. Beutler, B. and Rietschel, E.T. (2003) Innate immune sensing and its roots: the story of endotoxin. *Nat. Rev. Immunol.*, **3**, 169–176.
2. Janeway, C.A. Jr. and Medzhitov, R. (2002) Innate immune recognition. *Annu. Rev. Immunol.*, **20**, 197–216.
3. Medzhitov, R. and Janeway, C.A. Jr. (1997) Innate immunity: the virtues of a nonclonal system of recognition. *Cell*, **91**, 295–298.
4. Akira, S. and Hemmi, H. (2003) Recognition of pathogen-associated molecular patterns by TLR family. *Immunol. Lett.*, **85**, 85–95.
5. Kawai, T. and Akira, S. (2007) TLR signaling. *Semin. Immunol.*, **19**, 24–32.
6. Medzhitov, R. (2001) Toll-like receptors and innate immunity. *Nat. Rev. Immunol.*, **1**, 135–145.
7. Yoneyama, M. and Fujita, T. (2007) Function of RIG-I-like receptors in antiviral innate immunity. *J. Biol. Chem.*, **282**, 15315–15318.
8. Yoneyama, M., Kikuchi, M., Natsukawa, T., Shinobu, N., Imaizumi, T., Miyagishi, M., Taira, K., Akira, S. and Fujita, T. (2004) The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat. Immunol.*, **5**, 730–737.
9. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B. *et al.* (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.*, **5**, e1000562.
10. Casanova, J.L., Abel, L. and Quintana-Murci, L. (2011) Human TLRs and IL-1Rs in host defense: natural insights from evolutionary, epidemiological, and clinical genetics. *Annu. Rev. Immunol.*, **29**, 447–491.
11. Loo, Y.M. and Gale, M. Jr. (2011) Immune signaling by RIG-I-like receptors. *Immunity*, **34**, 680–692.

12. Yoneyama, M., Kikuchi, M., Matsumoto, K., Imaizumi, T., Miyagishi, M., Taira, K., Foy, E., Loo, Y.M., Gale, M. Jr., Akira, S. *et al.* (2005) Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J. Immunol.*, **175**, 2851–2858.
13. Baum, A. and Garcia-Sastre, A. (2010) Induction of type I interferon by RNA viruses: cellular receptors and their substrates. *Amino Acids*, **38**, 1283–1299.
14. Kang, D.C., Gopalkrishnan, R.V., Wu, Q., Jankowsky, E., Pyle, A.M. and Fisher, P.B. (2002) mda-5: an interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. *Proc. Natl Acad. Sci. USA*, **99**, 637–642.
15. Kovacsovics, M., Martinon, F., Micheau, O., Bodmer, J.L., Hofmann, K. and Tschopp, J. (2002) Overexpression of Helicard, a CARD-containing helicase cleaved during apoptosis, accelerates DNA degradation. *Curr. Biol.*, **12**, 838–843.
16. Saito, T., Hirai, R., Loo, Y.M., Owen, D., Johnson, C.L., Sinha, S.C., Akira, S., Fujita, T. and Gale, M. Jr. (2007) Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. *Proc. Natl Acad. Sci. USA*, **104**, 582–587.
17. Akira, S. and Takeda, K. (2004) Functions of toll-like receptors: lessons from KO mice. *C. R. Biol.*, **327**, 581–589.
18. Kato, H., Sato, S., Yoneyama, M., Yamamoto, M., Uematsu, S., Matsui, K., Tsujimura, T., Takeda, K., Fujita, T., Takeuchi, O. *et al.* (2005) Cell type-specific involvement of RIG-I in antiviral response. *Immunity*, **23**, 19–28.
19. Kawai, T., Takahashi, K., Sato, S., Coban, C., Kumar, H., Kato, H., Ishii, K.J., Takeuchi, O. and Akira, S. (2005) IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. *Nat. Immunol.*, **6**, 981–988.
20. Kato, H., Takeuchi, O., Sato, S., Yoneyama, M., Yamamoto, M., Matsui, K., Uematsu, S., Jung, A., Kawai, T., Ishii, K.J. *et al.* (2006) Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature*, **441**, 101–105.
21. Liu, P., Jamaluddin, M., Li, K., Garofalo, R.P., Casola, A. and Brasier, A.R. (2007) Retinoic acid-inducible gene I mediates early antiviral response and Toll-like receptor 3 expression in respiratory syncytial virus-infected airway epithelial cells. *J. Virol.*, **81**, 1401–1411.
22. Loo, Y.M., Fornek, J., Crochet, N., Bajwa, G., Perwitasari, O., Martinez-Sobrido, L., Akira, S., Gill, M.A., Garcia-Sastre, A., Katze, M.G. *et al.* (2008) Distinct RIG-I and MDA5 signaling by RNA viruses in innate immunity. *J. Virol.*, **82**, 335–345.
23. Melchjorsen, J., Jensen, S.B., Malmgaard, L., Rasmussen, S.B., Weber, F., Bowie, A.G., Matikainen, S. and Paludan, S.R. (2005) Activation of innate defense against a paramyxovirus is mediated by RIG-I and TLR7 and TLR8 in a cell-type-specific manner. *J. Virol.*, **79**, 12944–12951.
24. Mikkelsen, S.S., Jensen, S.B., Chiliveru, S., Melchjorsen, J., Julkunen, I., Gaestel, M., Arthur, J.S., Flavell, R.A., Ghosh, S. and Paludan, S.R. (2009) RIG-I-mediated activation of p38 MAPK is essential for viral induction of interferon and activation of dendritic cells: dependence on TRAF2 and TAK1. *J. Biol. Chem.*, **284**, 10774–10782.
25. Pichlmair, A., Schulz, O., Tan, C.P., Naslund, T.I., Liljestrom, P., Weber, F. and Reis e Sousa, C. (2006) RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′-phosphates. *Science*, **314**, 997–1001.
26. Gitlin, L., Barchet, W., Gilfillan, S., Cella, M., Beutler, B., Flavell, R.A., Diamond, M.S. and Colonna, M. (2006) Essential role of mda-5 in type I IFN responses to polyriboinosinic: polyribocytidylic acid and encephalomyocarditis picornavirus. *Proc. Natl Acad. Sci. USA*, **103**, 8459–8464.
27. Wang, J.P., Cerny, A., Asher, D.R., Kurt-Jones, E.A., Bronson, R.T. and Finberg, R.W. (2010) MDA5 and MAVS mediate type I interferon responses to coxsackie B virus. *J. Virol.*, **84**, 254–260.
28. Wang, Q., Nagarkar, D.R., Bowman, E.R., Schneider, D., Gosangi, B., Lei, J., Zhao, Y., McHenry, C.L., Burgens, R.V., Miller, D.J. *et al.* (2009) Role of double-stranded RNA pattern recognition receptors in rhinovirus-induced airway epithelial cell responses. *J. Immunol.*, **183**, 6989–6997.
29. Schlee, M., Roth, A., Hornung, V., Hagmann, C.A., Wimmenauer, V., Barchet, W., Coch, C., Janke, M., Mihailovic, A., Wardle, G. *et al.* (2009) Recognition of 5′ triphosphate by RIG-I helicase requires short blunt double-stranded RNA as contained in panhandle of negative-strand virus. *Immunity*, **31**, 25–34.
30. Kato, H., Takeuchi, O., Mikamo-Satoh, E., Hirai, R., Kawai, T., Matsushita, K., Hiiragi, A., Dermody, T.S., Fujita, T. and Akira, S. (2008) Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid-inducible gene-I and melanoma differentiation-associated gene 5. *J. Exp. Med.*, **205**, 1601–1610.
31. Takahasi, K., Yoneyama, M., Nishihori, T., Hirai, R., Kumeta, H., Narita, R., Gale, M. Jr, Inagaki, F. and Fujita, T. (2008) Nonself RNA-sensing mechanism of RIG-I helicase and activation of antiviral immune responses. *Mol. Cell*, **29**, 428–440.
32. Nallagatla, S.R., Toroney, R. and Bevilacqua, P.C. (2008) A brilliant disguise for self RNA: 5′-end and internal modifications of primary transcripts suppress elements of innate immunity. *RNA Biol.*, **5**, 140–144.
33. Venkataraman, T., Valdes, M., Elsby, R., Kakuta, S., Caceres, G., Saijo, S., Iwakura, Y. and Barber, G.N. (2007) Loss of DExD/H box RNA helicase LGP2 manifests disparate antiviral responses. *J. Immunol.*, **178**, 6444–6455.
34. Shigemoto, T., Kageyama, M., Hirai, R., Zheng, J., Yoneyama, M. and Fujita, T. (2009) Identification of loss of function mutations in human genes encoding RIG-I and MDA5: implications for resistance to type I diabetes. *J. Biol. Chem.*, **284**, 13348–13354.
35. Hu, J., Nistal-Villan, E., Voho, A., Ganee, A., Kumar, M., Ding, Y., Garcia-Sastre, A. and Wetmur, J.G. (2010) A common polymorphism in the caspase recruitment domain of RIG-I modifies the innate immune response of human dendritic cells. *J. Immunol.*, **185**, 424–432.
36. Pothlichet, J., Burtey, A., Kubarenko, A.V., Caignard, G., Solhonne, B., Tangy, F., Ben-Ali, M., Quintana-Murci, L., Heinzmann, A., Chiche, J.D. *et al.* (2009) Study of human RIG-I polymorphisms identifies two variants with an opposite impact on the antiviral immune response. *PLoS ONE*, **4**, e7582.
37. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703–707.
38. Jermendy, A., Szatmari, I., Laine, A.P., Lukacs, K., Horvath, K.H., Korner, A., Madacsy, L., Veijola, R., Simell, O., Knip, M. *et al.* (2010) The interferon-induced helicase IFIH1 Ala946Thr polymorphism is associated with type 1 diabetes in both the high-incidence Finnish and the medium-incidence Hungarian populations. *Diabetologia*, **53**, 98–102.
39. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
40. Reddy, M.V., Wang, H., Liu, S., Bode, B., Reed, J.C., Steed, R.D., Anderson, S.W., Steed, L., Hopkins, D. and She, J.X. (2011) Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. *Genes Immun.*, **12**, 208–212.
41. Smyth, D.J., Cooper, J.D., Bailey, R., Field, S., Burren, O., Smink, L.J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D.B. *et al.* (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.*, **38**, 617–619.
42. Sutherland, A., Davies, J., Owen, C.J., Vaikkakara, S., Walker, C., Cheetham, T.D., James, R.A., Perros, P., Donaldson, P.T., Cordell, H.J. *et al.* (2007) Genomic polymorphism at the interferon-induced helicase (IFIH1) locus contributes to Graves' disease susceptibility. *J. Clin. Endocrinol. Metab.*, **92**, 3338–3341.
43. Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., Nordmark, G. *et al.* (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.*, **41**, 1228–1233.
44. Strange, A., Capon, F., Spencer, C.C., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., Bellenguez, C., Bergboer, J.G. *et al.* (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, **42**, 985–990.
45. Ferreira, R.C., Pan-Hammarstrom, Q., Graham, R.R., Gateva, V., Fontan, G., Lee, A.T., Ortmann, W., Urcelay, E., Fernandez-Arquero, M., Nunez, C. *et al.* (2010) Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat. Genet.*, **42**, 777–780.
46. Barreiro, L.B. and Quintana-Murci, L. (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.*, **11**, 17–30.
47. Fumagalli, M., Cagliani, R., Riva, S., Pozzoli, U., Biasin, M., Piacentini, L., Comi, G.P., Bresolin, N., Clerici, M. and Sironi, M. (2010) Population genetics of IFIH1: ancient population structure, local selection, and

implications for susceptibility to type 1 diabetes. *Mol. Biol. Evol.*, **27**, 2555–2566.
48. Sironi, M. and Clerici, M. (2010) The hygiene hypothesis: an evolutionary perspective. *Microbes Infect.*, **12**, 421–427.
49. Abadie, V., Sollid, L.M., Barreiro, L.B. and Jabri, B. (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annu. Rev. Immunol.*, **29**, 493–525.
50. Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G., Franke, L., Oosting, M., Barisani, D. *et al.* (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.*, **86**, 970–977.
51. Quintana-Murci, L., Alcais, A., Abel, L. and Casanova, J.L. (2007) Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nat. Immunol.*, **8**, 1165–1171.
52. Keinan, A., Mullikin, J.C., Patterson, N. and Reich, D. (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.*, **39**, 1251–1255.
53. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
54. Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R. and Bustamante, C.D. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci. USA*, **102**, 7882–7887.
55. Kreitman, M. (2000) Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.*, **1**, 539–559.
56. Laval, G., Patin, E., Barreiro, L.B. and Quintana-Murci, L. (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*, **5**, e10284.
57. Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R. and Di Rienzo, A. (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA*, **102**, 18508–18513.
58. Barreiro, L.B., Laval, G., Quach, H., Patin, E. and Quintana-Murci, L. (2008) Natural selection has driven population differentiation in modern humans. *Nat. Genet.*, **40**, 340–345.
59. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A.G. (2007) Recent and ongoing selection in the human genome. *Nat. Rev. Genet.*, **8**, 857–868.
60. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
61. Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
62. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
63. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
64. Murdoch, S., Djuric, U., Mazhar, B., Seoud, M., Khan, R., Kuick, R., Bagga, R., Kircheisen, R., Ao, A., Ratti, B. *et al.* (2006) Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. *Nat. Genet.*, **38**, 300–302.
65. Qian, J., Deveault, C., Bagga, R., Xie, X. and Slim, R. (2007) Women heterozygous for NALP7/NLRP7 mutations are at risk for reproductive wastage: report of two novel mutations. *Hum. Mutat.*, **28**, 741.
66. Tian, X., Pascal, G. and Monget, P. (2009) Evolution and functional divergence of NLRP genes in mammalian reproductive systems. *BMC Evol. Biol.*, **9**, 202.
67. Tong, Z.B., Gold, L., Pfeifer, K.E., Dorward, H., Lee, E., Bondy, C.A., Dean, J. and Nelson, L.M. (2000) Mater, a maternal effect gene required for early embryonic development in mice. *Nat. Genet.*, **26**, 267–268.
68. Li, Y., Liao, W., Cargill, M., Chang, M., Matsunami, N., Feng, B.J., Poon, A., Callis-Duffin, K.P., Catanese, J.J., Bowcock, A.M. *et al.* (2010) Carriers of rare missense variants in IFIH1 are protected from psoriasis. *J. Invest. Dermatol.*, **130**, 2768–2772.
69. Downes, K., Pekalski, M., Angus, K.L., Hardy, M., Nutland, S., Smyth, D.J., Walker, N.M., Wallace, C. and Todd, J.A. (2010) Reduced expression of IFIH1 is protective for type 1 diabetes. *PLoS ONE*, **5**, e12646.
70. Gompel, N. and Prud'homme, B. (2009) The causes of repeated genetic evolution. *Dev. Biol.*, **332**, 36–47.
71. Kawai, T. and Akira, S. (2006) Innate immune recognition of viral infection. *Nat. Immunol.*, **7**, 131–137.
72. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A. *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
73. Takahashi, M., Matsuda, F., Margetic, N. and Lathrop, M. (2003) Automated identification of single nucleotide polymorphisms from sequencing data. *J. Bioinform. Comput. Biol.*, **1**, 253–265.
74. Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.
75. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
76. Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
77. Laval, G. and Excoffier, L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.
78. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A. *et al.* (2010) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
79. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
80. Mueller, J.C. and Andreoli, C. (2004) Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. *Bioinformatics*, **20**, 786–787.
81. Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
82. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.
83. Bandelt, H.J., Forster, P. and Rohl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, **16**, 37–48.
84. Consortium, U. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*, **37**, D169–D174.

III) Article 2 : *« The evolutionary landscape of cytosolic microbial sensors in humans»*

(pour les *supplementary material*, voir **Annexe 3**)

The evolutionary landscape of cytosolic microbial sensors in humans

Estelle Vasseur[a,b,c,1], Michele Boniotto[a,b,1], Etienne Patin[a,b], Guillaume Laval[a,b], Hélène Quach[a,b], Jeremy Manry[a,b], Brigitte Crouau-Roy[c], Lluis Quintana-Murci[a,b,2]

[a]Institut Pasteur, Unit of Human Evolutionary Genetics, Paris 75015, France

[b]Centre National de la Recherche Scientifique, URA3012, Paris 75015, France

[c]CNRS, Université de Toulouse, EDB, UMR 5174, Toulouse 31062, France

[1]These authors contributed equally to this work

[2]Correspondence should be addressed to L.Q.-M. (quintana@pasteur.fr)

Classification: BIOLOGICAL SCIENCES - Genetics

Abstract

Host-pathogen interactions are generally initiated by host recognition of microbial components or of danger signals triggered by microbial invasion. This recognition involves germline-encoded microbial sensors or pattern recognition receptors (PRRs). By studying the way in which natural selection has driven the evolution of these microbial sensors, we can identify genes playing an essential role and distinguish them from other, redundant genes. We characterized the sequence diversity of the NOD-like receptor family, including the NALP and NOD/IPAF subfamilies, in various populations worldwide, and compared this diversity with that of other PRR families, such as Toll-like receptors (TLRs) and RIG-I-like receptors (RLRs). We found that most NALPs had evolved under strong selective constraints, suggesting that their functions are essential, and possibly much broader than previously thought. Conversely, most NOD/IPAF subfamily members were subject to more relaxed selective constraints, suggesting greater redundancy. Furthermore, some NALP genes, including *NLRP1* and *NLRP14*, and *CIITA* were found to have evolved adaptively. We identified the most likely targets of positive selection, corresponding to variants conferring a selective advantage on some human populations. More generally, the strength of selection differed considerably between the major families of microbial sensors. Endosomal TLRs and most NALPs were found to evolve under stronger purifying selection than most NOD/IPAF subfamily members, cell-surface TLRs and RLRs, suggesting some degree of redundancy in the signaling pathways triggered by these molecules. This study provides novel perspectives and experimentally testable hypotheses concerning the relative biological relevance of the various families of microbial sensors in humans.

\body

Introduction

The innate immune system is responsible for the immediate response of the host to infectious or noxious assaults (1). It is based on germline-encoded microbial sensors, known as pattern-recognition receptors (PRRs) (2), which recognize conserved pathogen-associated molecular patterns (PAMPs) from microorganisms or damage-associated molecular patterns (DAMPs) resulting from tissue damage or cellular stress (3). The effective sensing of PAMPs and DAMPs induces the activation of signaling pathways that culminate in the induction of inflammatory responses, which facilitate the eradication of pathogens and infected, injured or dying cells (2, 3). Several families of PRRs have been identified, including the Toll-like receptors (TLRs), the C-type lectin receptors (CLRs), the RIG-I-like receptors (RLRs) and the NOD-like receptors (NLRs). The members of these families can be distinguished on the basis of ligand specificity, cellular distribution and downstream signaling pathways (2, 4-7). The use of different families of PRRs provides the host with a high degree of functional redundancy and multiple mechanisms for responding to a diverse range of pathogens (2).

PRRs may be membrane-bound, cytoplasmic or secreted. The membrane-bound TLRs — the most studied group of PRRs (4, 8) — and CLRs survey the extracellular milieu and endosomal compartments, whereas the more recently discovered RLRs and NLRs scan the cytosol for signs of infection or danger (2, 3, 5-7, 9). The importance of the role of NLRs, in particular, has been increasingly recognized in the last few years (6, 7). In humans, the NLR proteins are encoded by a family of 22 genes and are characterized by three distinct domains: the ligand-sensing leucine-rich repeats (LRRs); the NACHT domain, which mediates oligomerization, and an effector domain, which may be a pyrin domain (PYD), a CARD (caspase recruitment domain) or a BIR (baculovirus IAP repeat) domain. Different NLRs have different N-terminal domains, and this

domain defines different subfamilies, including the large NALP subfamily, the NODs and other proteins, such as CIITA, NLRC4 (or IPAF) and NAIP (10). Substantial advances in the functional characterization of some NLRs, such as NOD1 and NOD2 and the two NALP-members NLRP1 and NLRP3 in particular, have been made, highlighting the roles of these proteins in the sensing of microbial and nonmicrobial danger signals. Upon ligand recognition, these sensors activate either NF-κB or MAP kinases to induce inflammatory responses, or large cytoplasmic complexes called inflammasomes, which link the sensing of microbial or danger signals to both the proteolytic activation of proinflammatory cytokines and the initiation of cell death (6, 7, 11, 12). There is also increasing evidence to suggest that NLRs have a diverse range of biological functions, extending well beyond pathogen detection (6). These include roles in autoimmunity (13-15), autophagy (16), development and reproduction (17, 18) and tissue homeostasis (19). However, the precise functions, ligands and sensing mechanisms of most NLRs remain largely unknown.

We used an evolutionary genetics approach to assess the biological relevance of the NLR family and, more generally, of the various families of microbial sensors in humans. This approach has proved indispensable and complementary to immunological, clinical and epidemiological genetics studies, making it possible to distinguish between essential and redundant functions of host defense genes in the natural setting (8, 20, 21). We assessed the extent to which natural selection had affected the evolution of NLR family members, and compared our findings with those for other PRR families, including cell-surface TLRs, endosomal TLRs and cytosolic RLRs. We thus provide here the first comprehensive view of the evolutionary landscape characterizing the major families of microbial sensors, providing insight into the biological relevance of these sensors in humans.

Results and Discussion

Full resequencing of the NLR family members in various human populations. We sequenced the 21 NLR genes in a panel of 185 individuals of African, European and East-Asian descent, which were apparently healthy individuals who were not selected for any disease status (Table S1). The NLRs can be broadly divided into two large subfamilies (10). The first, the "NALP subfamily", is comprised of the PYD-containing NALPs, which are encoded by the 14 *NLRP1-14* genes. The second, which is here collectively referred as to the "NOD/IPAF subfamily", includes the five NODs – *NOD1* (*CARD4*), *NOD2* (*CARD15*), *NLRC3* (*NOD3*), *NLRC5* (*NOD4*) and *NLRX1* (*NOD9*) – together with *CIITA* (*NLRA*) and *NLRC4* (*IPAF*), which have the N-terminal CARD domain common to most NODs. The NLR gene *NAIP* was not resequenced here, owing to its highly repeated genomic organization (22). We generated 171.2 kb of resequenced data per individual, 67.9 kb of which corresponded to coding regions, the rest corresponding to the 5'and 3'-UTRs and introns (Table S2). We identified 2,084 SNPs (49% of which are newly described here), 396 of which were nonsynonymous, four of which were nonsense and 12 of which were coding-region indels (Fig. S1, Table S3). NALPs generally displayed higher or similar levels of nucleotide diversity (π) than expected under neutrality (0.1-0.4 times higher in Africa and Asia), whereas most NOD/IPAF members had levels of nucleotide diversity similar to or lower than neutral expectations (0.4-0.6 times lower in all populations) (Fig. S2, Table S4). This dataset was used to explore the effects of natural selection on the NLRs, and other major families of microbial sensors, since the divergence of the human and chimpanzee lineages and within different human populations.

The NALP and NOD/IPAF subfamilies have evolved under different functional constraints. We investigated whether and how natural selection had driven the heterogeneous patterns of diversity observed, by first estimating the direction and strength of selection within the human species as a whole. We used the McDonald-Kreitman Poisson Random Field (MKPRF) (23, 24) method to estimate ω, which compares divergence and polymorphism at nonsynonymous and silent sites, and γ, which relies on the ratio of divergence and polymorphism at nonsynonymous sites. Ten of the 14 NALPs *(NLRP3, 4, 5, 7, 8, 9, 10, 11, 12* and *13*) and *NLRC5* had a ω value significantly lower than 1, indicating a deficit of nonsynonymous variants (Fig. 1), which reflects the action of purifying selection. Conversely, most NOD/IPAF subfamily members (*NOD1, NLRC3, NLRC4, NLRX1* and *CIITA*), and *NLRP2* had significantly negative γ values, attesting to an excess of nonsynonymous SNPs segregating in the human population with respect to divergence (Fig. S3). Inspection of the allele frequency spectra showed that the excess of nonsynonymous polymorphism observed at these genes could be accounted for mostly by low-frequency variants, suggesting that negative selection prevented these alleles from increasing to high frequency and eventually becoming fixed (Fig. S4). Calculation of the π_N/π_S ratio per domain for each gene showed that the differences in selective constraints between the NALP and NOD/IPAF subfamilies could not be attributed to a particular protein domain (Fig. S5).

Our results indicate that the subfamilies of NLRs have followed very different evolutionary pathways. NALPs, the least known group of NLRs, have mostly evolved under strong purifying selection and are characterized by an overall deficit of functional diversity. By contrast, most members of the NOD/IPAF subfamily have been subject to more relaxed selection constraints, reflecting a higher degree of redundancy. Interestingly, the NALP subfamily shows a significant enrichment in genes under the action of purifying selection (71% observed with respect to 20%

expected genomewide on the basis of ~11,600 genes (23), P<0.01). These observations are consistent with most NALPs having acquired a function that is essential and non redundant in host survival, and with the rapid elimination of mutations of these genes from the population, due to their highly deleterious effects. Our observations are supported by medical genetic studies for *NLRP3* (also known as *NALP3*), subject to the highest degree of evolutionary constraint, in which mutations have been associated with rare severe inflammatory diseases (25-27). *NLRP3* expression is essentially limited to immune and non keratinizing epithelial cells (28), and the protein encoded by this gene is known to activate caspase-1 in the sensing of bacteria or DAMPs (6). The purifying selection regime under which this gene has evolved suggests an important role of this sensor in caspase-1-mediated immunity signaling and in processes that are caspase-1-independent (29) or unrelated to pathogen recognition (6).

NALPs are encoded by a multigene family (30), most of the members of which are under strong selective constraints. This situation contrasts with other multigene families, many of the members of which have become pseudogenes or are subject to relaxed selective constraints (31, 32). This observation highlights the important role that most NALPs may have, probably in a much more diverse range of functions than the mere sensing of microbial and danger signals (6). For example, there is increasing evidence to suggest that NALPs are involved in the maintenance of intestinal homeostasis, as shown for the murine *Nalp3*, the absence of which is associated with greater tissue damage and colitis (19, 33-35). It is also becoming evident that many NALP genes are expressed specifically in gametes and embryos (36), consistent with an important role in early development and reproduction. Mutations of the human *NLRP7* gene, for example, are associated with abnormal human pregnancies, spontaneous abortions and intrauterine growth retardation (17, 37). These reproductive and developmental functions may be more closely related to immune functions than previously anticipated (10). Indeed, activation of the proinflammatory

cytokine IL-1β after inflammasome formation plays an essential role in ovulation and oocyte maturation (38). However, the genuine functions of most NALPs remain poorly documented and little studied. In this context, our evolutionary data are particularly informative, as they are consistent with most NALPs playing key roles in host survival, highlighting the need for unbiased functional data on this major family of ATPases in humans, for the assessment of their physiological relevance in health and disease.

Some NLRs have been subject to positive selection in specific populations. We then investigated whether some NLRs had evolved adaptively, with a view to identifying functional variation that may have conferred a selective advantage on the host. We thus performed various intra-species neutrality tests on various aspects of the data, including the allele frequency spectrum (i.e., Tajima's *D*, Fu and Li's *D** and *F**, and Fay and Wu's *H* tests), levels of population differentiation (i.e., F_{ST}) and haplotype-based tests (i.e., DIND and iHS tests) (for extensive reviews, see (39, 40)). At a first glance, allele frequency spectrum tests showed that most NOD/IPAF members were characterized by an excess of singletons, as attested by the significantly negative values obtained for these sequence-based neutrality tests (Table S5). However, because no significant signals of positive selection were detected based on independent tests, the excess of singletons observed most likely reflects the effects of weak negative selection alone (as attested by the γ estimate, see MKPRF results, Fig. S3). Furthermore, and more generally, we detected some signatures suggesting the action of positive selection at some genes based on single tests, but these were not confirmed by various, independent neutrality statistics. In light of this, we discuss only genes for which significant results were obtained at least in two independent tests for selection, after demographic correction, in a given population (see Materials

and Methods for details). With these conservative criteria, most NLRs showed no significant deviation from neutral expectations, although *NLRP1*, *NLRP14* and *CIITA* represented notable exceptions (Tables S3 and S5, Figs. 2 and S6).

The strongest signature of positive selection was detected for *NLRP1* (NALP1). Allele frequency spectrum-based tests (particularly Fay & Wu's *H* and *DH* tests) showed an excess of high-frequency derived alleles in Africa and Europe (Table S5, Fig. S7). We identified a group of 15 SNPs for which derived alleles with a frequency >90% were found in Africa and Europe. In Europe, these high-frequency derived alleles were in strong linkage disequilibrium (LD) with 29 high-frequency ancestral alleles in a haplotype block of ~45 kb (Table S6, Fig. S8*B*). The LD block was shorter in Africa, but most of the SNPs characterizing the European haplotype were present at similar frequencies (Table S6, Fig. S8*A*). In Asia, most of the corresponding derived alleles were fixed, accounting for the lack of significance of Fay & Wu' *H* values (Table S6, Fig. S8*C*). These observations suggest the occurrence of a selective sweep that is still underway in Africa and Europe but has been completed in Asia. In LD-based tests, 11 of the mutations characterizing the European haplotype gave iHS values > 2 in Europe, and one mutation gave an iHS value > 2 in Africa (Table S3). These significant iHS values were observed at highly-frequent ancestral alleles only because derived alleles characterizing the selected haplotype were not available from HapMap (41). Overall, our results are thus consistent with a worldwide selective sweep targeting a long haplotype comprised of seven nonsynonymous SNPs (Table S6). *NLRP1* is one of the few NALPs for which a PRR function has been well documented (6, 42). These amino acid changes may therefore have conferred a selective advantage related to microbial sensing and may underlie differences in susceptibility to infectious and immunity-related disorders.

We detected an independent, positive selection event that has more recently targeted *NLRP1* in Europe. The DIND test identified the nonsynonymous SNP 51015 (V1059M) and four linked intronic variants (Fig. 3*A*), which were not in LD (r^2= 0.04) with the long haplotype described above. Two of these SNPs (including the V1059M variant) had iHS values < -2, providing evidence of the action of positive selection on the derived alleles. Furthermore, the frequencies of the derived alleles were higher in Europe than elsewhere (33% vs. 4-8% elsewhere). F_{ST} analyses identified another nonsynonymous variant, SNP 1911 (L155H), as being strongly differentiated in Europe (F_{ST} =0.40 and 0.35 for Africa/Europe and Europe/Asia, respectively) (Fig. 2). L155H was found to be in intermediate LD (r^2=0.52) with the V1059M variant and the four intronic variants (Fig. S8*D*). Interestingly, the L155H variant has been shown to be associated with various autoimmune diseases, including Addison's disease, type I diabetes and vitiligo (15, 43, 44). Most of our analyses localized however the signature of selection to V1059M, rather than L155H, suggesting that V1059M is the actual target of selection, increasing the frequency of the linked L155H mutation in Europe, which is nowadays responsible for several autoimmune diseases. These findings provide support for the hypothesis that the current high incidence of autoimmune or inflammatory disorders results from past adaptation to infectious agents (20, 45).

Another positive selection signal was identified for *NLRP14* (NALP14). This gene displayed an excess of low-frequency alleles in Asia (Table S5) and LD-based tests confirmed this signature, with XP-EHH values greater than 2 recorded for Asia (Fig. S9). Interestingly, the highest XP-EHH value in this genomic region was observed for the nonsynonymous SNP 19221 (E808K) and two intronic SNPs, in most HapMap Asian populations (41). Furthermore, DIND identified the E808K variant as an outlier in the three populations and particularly in Asia (Figs. 3*B* and S6), suggesting that E808K is the most likely target of either stronger or earlier selection

in Asian populations, consistent with the clear star-like shape of the *NLRP14* network, particularly in Asia (Fig. S10).

Finally, we detected a signature of positive selection for the NOD/IPAF member *CIITA* based on the F_{ST} analysis, which identified SNP -286 as highly differentiated in Europe and SNPs -730 and -561 in Asia (Fig. 2). The signal in Europe was confirmed by the iHS value > 2 obtained for SNP –712 (Table S3) and other HapMap SNPs (rs8052975, rs6498114, rs12922863, rs6498116, rs12928665, rs11074934) in high LD (r^2>0.6) with SNP -286. These results are therefore consistent with a positive selection event targeting the *CIITA* promoter region, in Europe. Interestingly, the positively selected SNP -286 has been associated with different susceptibilities to inflammatory diseases, such as rheumatoid arthritis, multiple sclerosis and myocardial infarction (46). The overlap between the positive selection signatures observed for some *NLRP1* and *CIITA* variants and previous associations with disease states provides proof-of-concept for the use of this evolutionary approach to predict the functional impact of other, as yet uncharacterized variants of *NLRP1* and *NLRP14* in human disease.

Global patterns of selection differ between the major families of microbial sensors. Pathogens harbor multiple ligands that are sensed by multiple families of PRRs, through crosstalk between the corresponding signaling pathways (2, 47), which may display various degrees of redundancy. Our results for NLRs therefore cannot be interpreted in isolation. We thus evaluated the strength of selection observed for the NALP and NOD/IPAF subfamilies, in the context of the results obtained for other major families of microbial sensors (48, 49). These families included the TLRs, both those located in the endosome, (TLR3, TLR7, TLR8 and TLR9) and those expressed on the cell surface (TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10), and the cytosolic RLRs – DDX58 (RIG-I), IFIH1 (MDA5) and DHX58 (LGP2). The MKPRF test was performed

on the 34 genes encoding these receptors. Based on the results of this test, we identified a group of genes subject to strong selective constraint, including most NALPs and the endosomal TLRs, and a group of genes subject to weaker evolutionary constraints, most NOD/IPAF subfamily members, the cell-surface TLRs and the cytosolic RLRs (Fig. 1). We next evaluated the contributions of divergence and polymorphism to the patterns of purifying selection observed at some genes, by comparing the p_N/p_S and d_N/d_S values obtained for the microbial sensors presenting ω values < 1 with a genomewide distribution (23) (Fig. S11). None of the PRR genes showed unexpected values. In light of this, the most parsimonious explanation is that the strong evolutionary constraints observed have been exerted with the same strength in the human/chimpanzee ancestor and in the human lineage.

Our study revealed important differences in the intensity of selection driving the evolution of the major families of microbial sensors and provided information about the biological relevance of the mechanisms triggered by these molecules (Fig. 4). For PRRs specialized in the sensing of nucleic acids, particularly from viruses, we found that endosomal TLRs were under stronger evolutionary constraints than cytosolic RLRs, suggesting a non redundant, essential role for endosomal TLRs. Indeed, TLRs and RLRs use specific adaptors to initiate their respective signaling cascades, but these pathways ultimate converge in the production of type I IFNs and proinflammatory cytokines (4, 5). The more relaxed evolutionary constraints on RLRs than on endosomal TLRs may therefore reflect some redundancy of this system in antiviral immunity. For PRRs involved in the sensing of non-nucleic acid products, mostly from bacteria, and stress signals, we found that the members of the NALP family were generally subject to stronger purifying selection than NOD/IPAF subfamily members and cell-surface TLRs. This supports a higher degree of redundancy of these latter two groups.

Conclusions

Overall, our analyses allowed us to distinguish three groups of innate immunity genes that differ in their evolutionary patterns: genes under strong selective constraints, genes under weaker constraints, and genes for which no deviation from neutrality was detected. Our analyses also revealed that some NALPs and NOD/IPAF subfamily members have evolved adaptively, attesting to the presence of functional variation that may confer an advantage on specific human populations. These data open up new research perspectives and facilitate the formulation of experimentally testable hypotheses, by providing a general hierarchical model for the biological relevance of the various microbial sensors, some of which are essential, whereas others are more expendable. These findings should stimulate future functional studies aiming to determine whether the strong constraints on the genes for some of these sensors, including the little studied NALPs, provide evidence of the importance of the PRR functions putatively mediated by these sensors or, more generally, for broader processes extending to basic, early developmental mechanisms and the maintenance of body homeostasis.

Materials and Methods

Full details of the experimental and statistical procedures can be found in SI Materials and Methods. Sequence variation for the 21 NLR genes was determined for a total of 370 chromosomes from the Human Genome Diversity Panel (HGDP)–CEPH panel (50), including 62 from Sub-Saharan Africans, 62 from Europeans and 61 from East Asians (Table S1). This study was approved by the Institut Pasteur Ethics Committee (no. RBM 2008.06). We resequenced all the exonic regions and an equivalent amount of non-exonic material, including intronic, 5' and 3' regions (Table S2). The resulting sequence-based dataset was used to estimate a number of general diversity indices and population genetic parameters.

Haplotype reconstruction was carried out by the Bayesian statistical method implemented in Phase (v.2.1.1) (51). Haploview software (52) was used to evaluate and visualize LD levels. Diversity indices were calculated and allele frequency spectrum-based neutrality tests were performed with the DnaSP package v. 5.1 (53). *P*-values for the various neutrality tests were estimated from coalescent simulations with SIMCOAL 2.0 (54). Demographic corrections were applied to neutrality tests by integrating two validated demographic models in a set of populations similar to ours (55, 56). The number of silent and nonsynonymous fixed differences between humans and chimpanzees, and polymorphic within humans, together with the d_N, d_S, p_N and p_S values were determined with the DnaSP package v. 5.1 (53). We used the McDonald-Kreitman Poisson Random Field (MKPRF) method (23, 24) to estimate ω and γ. To assess the time depth of events of purifying selection exerted on some microbial sensors, we compared the p_N/p_S and d_N/d_S ratios of these genes to a genomewide distribution of genes (23). The division of the protein into domains was determined on the basis of information in the Uniprot database (57) and the π_N/π_S ratio per domain was calculated with the DnaSP package v. 5.1 (53). Various

haplotype-based tests, including the derived intra-allelic nucleotide diversity (DIND) test (48), the extended haplotype homozygosity (EHH) test (58), the cross population extended haplotype homozygosity (XP-EHH) test (59) and, when available, the integrated haplotype scores (iHS) (60) obtained from HapMap Phase II (41), were used to detect positive selection signatures. Population differentiation was assessed by calculating F_{ST} statistics and comparing them with the F_{ST} distribution of 650,000 SNPs genotyped for the same panel of individuals (61). We used NETWORK 4.6 to construct median-joining networks for the inference of haplotype genealogy (62).

Acknowledgments

This work was supported by Institut Pasteur, the ANR (ANR-08-MIEN-009-01), the *Fondation pour la Recherche Médicale*, the CNRS, Merck-Serono, and an EPFL-Debiopharm Life Sciences Award to L.Q.-M

Author contributions

EV and MB performed the experiments and the statistical analyses, with contributions from EP, GL, HQ and JM. BCR contributed reagents and materials. LQM conceived and designed the study. EV and LQM wrote the paper, with contributions from all authors.

References

1. Janeway CA, Jr. & Medzhitov R (2002) Innate immune recognition. *Annu Rev Immunol* 20:197-216.
2. Takeuchi O & Akira S (2010) Pattern recognition receptors and inflammation. *Cell* 140(6):805-820.
3. Matzinger P (2002) The danger model: a renewed sense of self. *Science* 296(5566):301-305.
4. Kawai T & Akira S (2007) TLR signaling. *Semin Immunol* 19(1):24-32.
5. Loo YM & Gale M, Jr. (2011) Immune signaling by RIG-I-like receptors. *Immunity* 34(5):680-692.
6. Kufer TA & Sansonetti PJ (2011) NLR functions beyond pathogen recognition. *Nat Immunol* 12(2):121-128.
7. Schroder K & Tschopp J (2010) The inflammasomes. *Cell* 140(6):821-832.
8. Casanova JL, Abel L, & Quintana-Murci L (2011) Human TLRs and IL-1Rs in host defense: natural insights from evolutionary, epidemiological, and clinical genetics. *Annu Rev Immunol* 29:447-491.
9. Geijtenbeek TB, van Vliet SJ, Engering A, t Hart BA, & van Kooyk Y (2004) Self- and nonself-recognition by C-type lectins on dendritic cells. *Annu Rev Immunol* 22:33-54.
10. Martinon F, Mayor A, & Tschopp J (2009) The inflammasomes: guardians of the body. *Annu Rev Immunol* 27:229-265.
11. Tschopp J, Martinon F, & Burns K (2003) NALPs: a novel protein family involved in inflammation. *Nat Rev Mol Cell Biol* 4(2):95-104.
12. Martinon F, Burns K, & Tschopp J (2002) The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. *Mol Cell* 10(2):417-426.
13. Hugot JP, *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411(6837):599-603.
14. Ogura Y, *et al.* (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 411(6837):603-606.
15. Jin Y, *et al.* (2007) NALP1 in vitiligo-associated multiple autoimmune disease. *N Engl J Med* 356(12):1216-1225.
16. Cooney R, *et al.* (2010) NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat Med* 16(1):90-97.
17. Murdoch S, *et al.* (2006) Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. *Nat Genet* 38(3):300-302.
18. Westerveld GH, *et al.* (2006) Mutations in the testis-specific NALP14 gene in men suffering from spermatogenic failure. *Hum Reprod* 21(12):3178-3184.
19. Zaki MH, Lamkanfi M, & Kanneganti TD (2011) The Nlrp3 inflammasome: contributions to intestinal homeostasis. *Trends Immunol* 32(4):171-179.
20. Barreiro LB & Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet* 11(1):17-30.
21. Quintana-Murci L, Alcais A, Abel L, & Casanova JL (2007) Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nat Immunol* 8(11):1165-1171.
22. Boniotto M, *et al.* (2011) Population variation in NAIP functional copy number

confers increased cell death upon /Legionella pneumophila/ infection. *Hum Immunol (in press).*
23. Bustamante CD, *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153-1157.
24. Sawyer SA & Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161-1176.
25. Aganna E, *et al.* (2002) Association of mutations in the NALP3/CIAS1/PYPAF1 gene with a broad phenotype including recurrent fever, cold sensitivity, sensorineural deafness, and AA amyloidosis. *Arthritis Rheum* 46(9):2445-2452.
26. Dode C, *et al.* (2002) New mutations of CIAS1 that are responsible for Muckle-Wells syndrome and familial cold urticaria: a novel mutation underlies both syndromes. *Am J Hum Genet* 70(6):1498-1506.
27. Hoffman HM, Mueller JL, Broide DH, Wanderer AA, & Kolodner RD (2001) Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. *Nat Genet* 29(3):301-305.
28. Kummer JA, *et al.* (2007) Inflammasome components NALP 1 and 3 show distinct but separate expression profiles in human tissues suggesting a site-specific role in the inflammatory response. *J Histochem Cytochem* 55(5):443-452.
29. Allen IC, *et al.* (2009) The NLRP3 inflammasome mediates in vivo innate immunity to influenza A virus through recognition of viral RNA. *Immunity* 30(4):556-565.
30. Tian X, Pascal G, & Monget P (2009) Evolution and functional divergence of NLRP genes in mammalian reproductive systems. *BMC Evol Biol* 9:202.
31. Force A, *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531-1545.
32. Ohno S (1970) *Evolution by gene duplication*. *Springer, New York.*
33. Allen IC, *et al.* (2010) The NLRP3 inflammasome functions as a negative regulator of tumorigenesis during colitis-associated cancer. *J Exp Med* 207(5):1045-1056.
34. Hirota SA, *et al.* (2010) NLRP3 inflammasome plays a key role in the regulation of intestinal homeostasis. *Inflamm Bowel Dis* 17(6):1359-1372.
35. Zaki MH, *et al.* (2010) The NLRP3 inflammasome protects against loss of epithelial integrity and mortality during experimental colitis. *Immunity* 32(3):379-391.
36. Zhang P, *et al.* (2008) Expression analysis of the NLRP gene family suggests a role in human preimplantation development. *PLoS One* 3(7):e2755.
37. Qian J, Deveault C, Bagga R, Xie X, & Slim R (2007) Women heterozygous for NALP7/NLRP7 mutations are at risk for reproductive wastage: report of two novel mutations. *Hum Mutat* 28(7):741.
38. Gerard N, Caillaud M, Martoriati A, Goudet G, & Lalmanach AC (2004) The interleukin-1 system and female reproduction. *J Endocrinol* 180(2):203-212.
39. Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* 1:539-559.
40. Nielsen R, Hellmann I, Hubisz M, Bustamante C, & Clark AG (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet* 8(11):857-868.
41. Frazer KA, *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851-861.
42. Hsu LC, *et al.* (2008) A NOD2-NALP1 complex mediates caspase-1-dependent IL-1beta secretion in response to Bacillus anthracis infection and muramyl dipeptide. *Proc Natl Acad Sci U S A* 105(22):7803-7808.

43. Magitta NF, *et al.* (2009) A coding polymorphism in NALP1 confers risk for autoimmune Addison's disease and type 1 diabetes. *Genes Immun* 10(2):120-124.
44. Zurawek M, *et al.* (2010) A coding variant in NLRP1 is associated with autoimmune Addison's disease. *Hum Immunol* 71(5):530-534.
45. Sironi M & Clerici M (2010) The hygiene hypothesis: an evolutionary perspective. *Microbes and infection / Institut Pasteur* 12(6):421-427.
46. Swanberg M, *et al.* (2005) MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. *Nat Genet* 37(5):486-494.
47. van Vliet SJ, den Dunnen J, Gringhuis SI, Geijtenbeek TB, & van Kooyk Y (2007) Innate signaling and regulation of Dendritic cell immunity. *Curr Opin Immunol* 19(4):435-440.
48. Barreiro LB, *et al.* (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet* 5(7):e1000562.
49. Vasseur E, *et al.* (2011) The selective footprints of viral pressures at the human RIG-I-like receptor family. *Hum Mol Genet* 20(22):4462-4474.
50. Cann HM, *et al.* (2002) A human genome diversity cell line panel. *Science* 296(5566):261-262.
51. Stephens M & Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73(5):1162-1169.
52. Barrett JC, Fry B, Maller J, & Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263-265.
53. Rozas J, Sanchez-DelBarrio JC, Messeguer X, & Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19(18):2496-2497.
54. Laval G & Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20(15):2485-2487.
55. Laval G, Patin E, Barreiro LB, & Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5(4):e10284.
56. Voight BF, *et al.* (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* 102(51):18508-18513.
57. Consortium TU (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37:D169-174.
58. Sabeti PC, *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832-837.
59. Sabeti PC, *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913-918.
60. Voight BF, Kudaravalli S, Wen X, & Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3):e72.
61. Li JZ, *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100-1104.
62. Bandelt HJ, Forster P, & Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16(1):37-48.

Figure Legends

Fig. 1. Purifying selection acting on individual *NLR* genes and genes from other major families of microbial sensors. The strength of purifying selection was assessed by calculating ω in the MKPRF test. Bars indicate 95% CIs and red diamonds indicate genes with ω estimates significantly lower than 1. The results for the population selection parameter γ, for the identification of genes subject to selection operating on nonsynonymous mutations that are polymorphic in humans, are presented in Fig. S3. The MKPRF test does not consider nonsense mutations. However, nonsense mutations were either absent or present at very low frequency (<1%) in all the families of PRRs considered other than the cell-surface molecules TLR10 and TLR5, for which 5% and up to 23% of individuals from the general population, respectively, carried a nonsense mutation (Table S3) (48, 49).

Fig. 2. Levels of population differentiation at the NLRs. The F_{ST} statistics is presented as a function of heterozygosity for each SNP in (*A*) Africans vs. Europeans, (*B*) Africans vs. East-Asians, and (*C*) Europeans vs. East Asians. The 95^{th} and 99^{th} percentiles of the HGDP-CEPH genotyping dataset for the same individuals are shown as dashed lines, whereas the blue area corresponds to the 99.9^{th} percentile. Black and red points represent silent and nonsynonymous SNPs, respectively. For each outlier SNP, the gene name is indicated, followed by its position respective to the ATG. Outlier SNPs separated by a comma correspond to SNPs in complete LD, and nonsynonymous SNPs are underlined. Note that a group of SNPs in *NALP6*, including the nonsynonymous SNP 1652 (L163M), displayed some of the highest levels of differentiation of any of the SNPs studied, in Asian populations.

Fig. 3. Detection of recent positive selection acting on the *NLRP1* gene in Europe and *NLRP14* gene in Asia. We plotted $i\pi_A/i\pi_D$ values against derived allele frequencies (DAFs) for (*A*) *NLRP1* in Europe and (*B*) *NLRP14* in Asia. The rationale underlying the DIND test is that a derived allele under positive selection with a high population frequency should present lower levels of nucleotide diversity at linked sites than would be expected under neutrality (48). *P*-values were obtained by comparing the $i\pi_A/i\pi_D$ values for the *NLRP1* and *NLRP14* genes with the expected values obtained from 10^4 simulations, taking into account the most conservative demographic model (55). The upper dashed line on the graph corresponds to the 99th percentile, and the lower, to the 95th percentile. Black and red points represent silent and nonsynonymous SNPs, respectively. Outlier SNPs separated by a comma correspond to SNPs in complete LD, and nonsynonymous SNPs are underlined. As to *NLRP1*, in addition to the selected SNP 51015 (V1059M), our analyses also identified another nonsynonymous SNP (SNP 62201, V1184M) linked to an intronic variant (SNP 63236). This signal may be a complex repercussion on the worldwide selective sweep. For the DIND analyses of all genes in all populations, see Fig. S6.

Fig. 4. Hierarchical model outlining the evolutionary dynamics and biological relevance of the various subfamilies of PRRs. This representation is based on the intensity of the selective constraints (based on the MKPRF results) detected for the 34 PRRs. These analyses allowed us to distinguish three groups of genes: genes under purifying selection ($\omega<1$, in red), genes under weaker selective constraints ($\gamma<0$, in yellow), and genes for which no deviation from neutrality was detected (in grey). The color intensity is proportional to the log P-value. Cellular sub-localization, functions and ligands are given as an indication but they are not exhaustive.

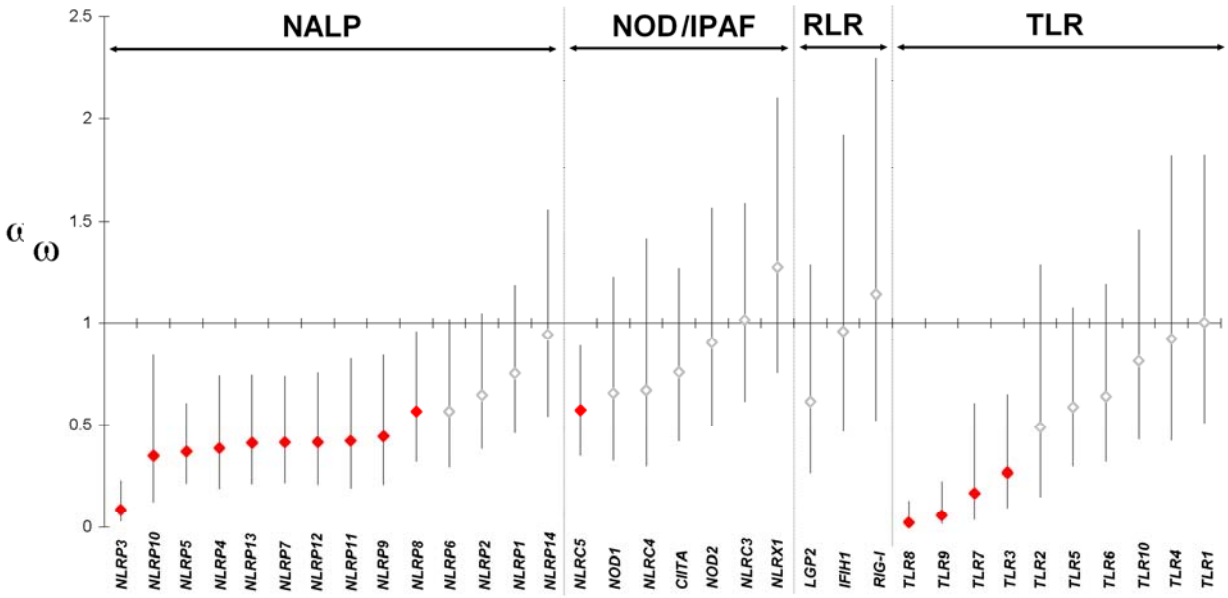

NALP
NOD/IPAF
RLR
TLR
ω
2.5
2
1.5
1
0.5
0
NLRP3
NLRP10
NLRP5
NLRP4
NLRP13
NLRP7
NLRP12
NLRP11
NLRP9
NLRP8
NLRP6
NLRP2
NLRP1
NLRP14
NLRC5
NOD1
NLRC4
CIITA
NOD2
NLRC3
NLRX1
LGP2
IFIH1
RIG-I
TLR8
TLR9
TLR7
TLR3
TLR2
TLR5
TLR6
TLR10
TLR4
TLR1

Figure 1

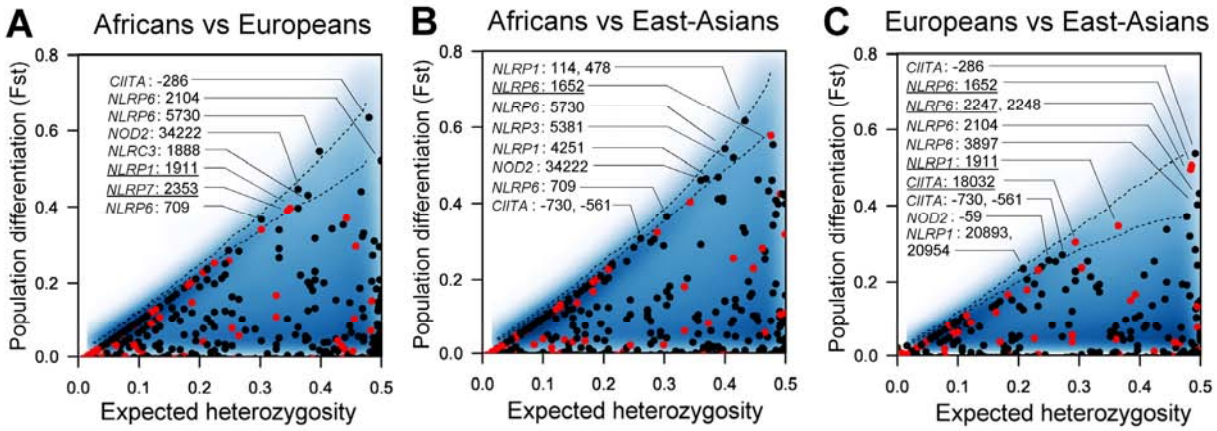

A
Africans vs Europeans
B
Africans vs East-Asians
C
Europeans vs East-Asians
Population differentiation (Fst)
Expected heterozygosity
0.0
0.2
0.4
0.6
0.8
0.1
0.3
0.5
CIITA: -286
NLRP6: 2104
NLRP6: 5730
NOD2: 34222
NLRC3: 1888
NLRP1: 1911
NLRP7: 2353
NLRP6: 709
NLRP1: 114, 478
NLRP6: 1652
NLRP6: 5730
NLRP3: 5381
NLRP1: 4251
NOD2: 34222
NLRP6: 709
CIITA: -730, -561
CIITA: -286
NLRP6: 1652
NLRP6: 2247, 2248
NLRP6: 2104
NLRP6: 3897
NLRP1: 1911
CIITA: 18032
CIITA: -730, -561
NOD2: -59
NLRP1: 20893, 20954

Figure 2

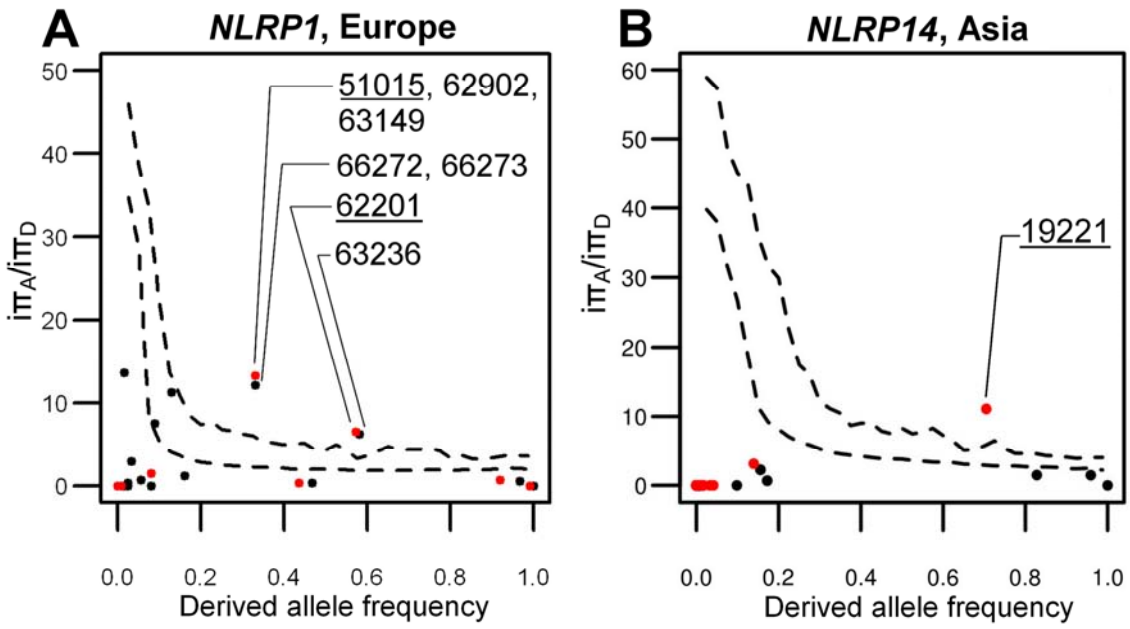

A
NLRP1, Europe
51015, 62902, 63149
66272, 66273
62201
63236
iπA/iπD
Derived allele frequency
B
NLRP14, Asia
19221
iπA/iπD
Derived allele frequency

Figure 3

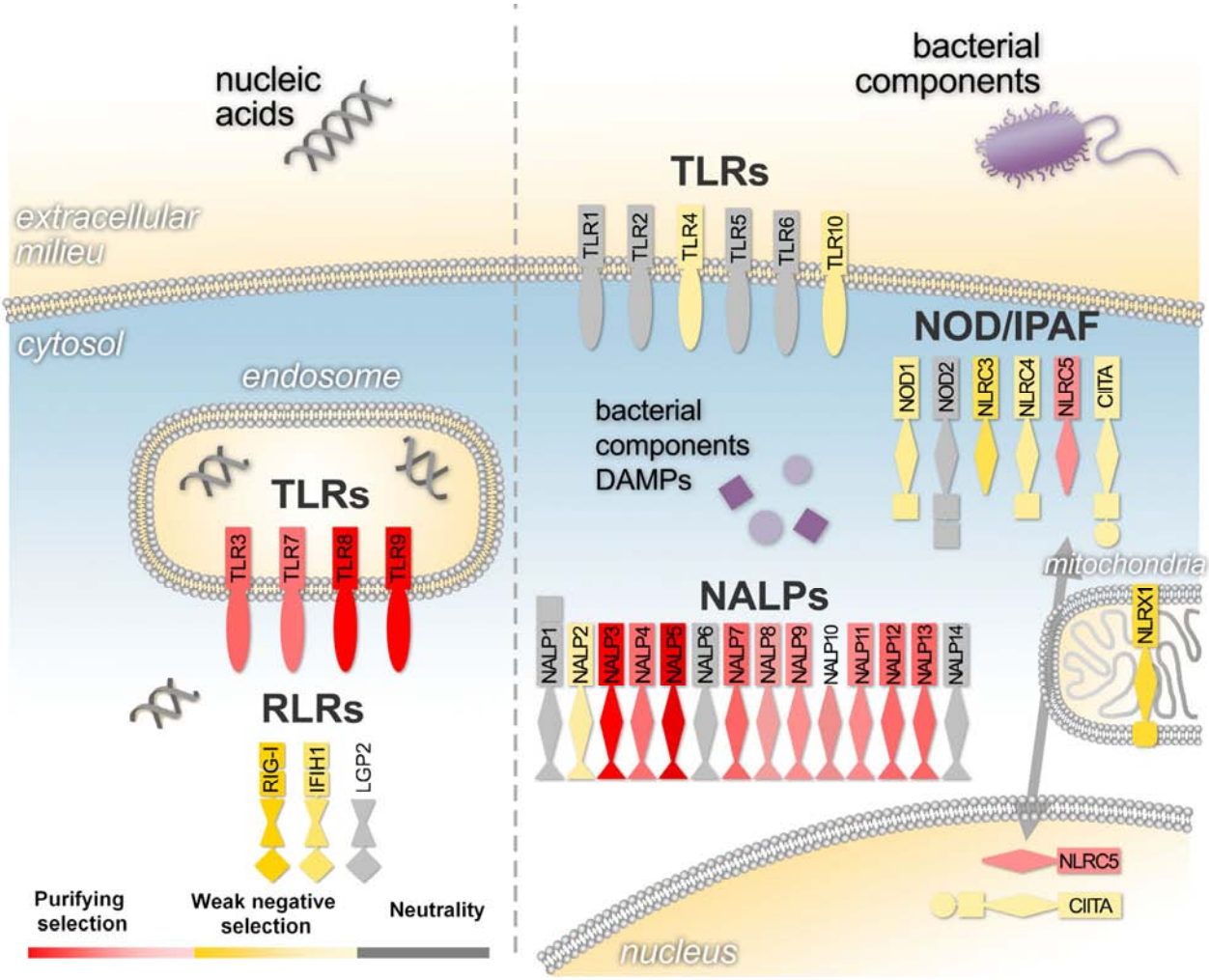

nucleic acids
bacterial components
TLRs
extracellular milieu
TLR1
TLR2
TLR4
TLR5
TLR6
TLR10
cytosol
NOD/IPAF
endosome
NOD1
NOD2
NLRC3
NLRC4
NLRC5
CIITA
bacterial components DAMPs
TLRs
TLR3
TLR7
TLR8
TLR9
mitochondria
NALPs
NALP1
NALP2
NALP3
NALP4
NALP5
NALP6
NALP7
NALP8
NALP9
NALP10
NALP11
NALP12
NALP13
NALP14
NLRX1
RLRs
RIG-I
IFIH1
LGP2
NLRC5
CIITA
Purifying selection
Weak negative selection
Neutrality
nucleus

Figure 4

IV) Discussion

1) Précisions sur l'interprétation des contraintes sélectives

Afin de mieux comprendre les interprétations que nous avons faites dans ces articles, il est important de clarifier les termes employés en matière de différence d'importance. Au cours de ce type d'étude, nous sommes souvent enclins à opposer les gènes portant des signatures de sélection purificatrice (les gènes essentiels) aux gènes qui accumulent des mutations non-synonymes et qui peuvent se trouver sous sélection positive (les gènes évoluant de manière adaptative). Cependant, ces cas ne sont pas totalement opposables et les gènes sous sélection positive, même s'ils accumulent passagèrement des mutations non-synonymes, ne doivent pas être considérés comme des gènes ayant un rôle moins important. Ainsi, je redéfinirais chacun des cas possibles pour une espèce donnée, à un instant t, par (i) les gènes sous sélection purificatrice qui ont déjà acquis une fonction essentielle qui ne peut être modifiée, (ii) les gènes qui accumulent beaucoup de mutations, fonctionnelles ou non, parce qu'ils ne remplissent pas une fonction essentielle (et non redondante) et ne semblent pas être en train d'acquérir de fonction nouvelle et particulière pendant la période étudiée ; ces gènes sont trouvés sous neutralité, (iii) les gènes qui accumulent localement ou plus largement des mutations non-synonymes, dont certaines à fortes fréquences qui ont conféré un avantage dans le passé : ces gènes montrent des signatures de sélection positive. Ils sont « en cours » de modification et ont déjà varié par rapport aux versions ancestrales. Dans ce cas, la diversité qu'ils accumulent n'est que « passagère » et si la nouvelle fonction conférée est très avantageuse, ces gènes se verront ensuite très contraints pour garder cette nouvelle propriété. Il ne s'agit donc que d'une question de temps. De manière intuitive, il est évident que la sélection positive n'agit que sur des gènes qui sont susceptibles d'avoir un effet avantageux ; s'ils n'avaient qu'un rôle mineur, nous les trouverions plutôt sous neutralité. De manière générale, les oppositions que nous faisons lors de cette analyse entre gènes essentiels et gènes moins essentiels devront donc être comprises dans le sens « gènes essentiels et conservés à la période étudiée » opposés à « gènes mutant plus librement, en train ou en attente d'acquérir une fonction assez importante pour être conservée ».

Un autre point qui me semble fondamental est que, comme je l'ai décrit dans l'introduction lors de la présentation des différents tests utilisés, nous détectons des signatures à différentes échelles de temps. Les traces de sélection purificatrice dont nous parlons ici résultent toutes du test MKPRF, basé sur la comparaison homme-chimpanzé et mettent en lumière des contraintes anciennes ; celles-ci sont partagées par l'ensemble des populations

humaines et ne dépendent donc pas de l'environnement. En revanche, les signatures de sélection positive que nous avons trouvées ici avec les tests comparant Afrique, Europe et Asie détectent des événements plus récents et qui dépendent de l'environnement (Novembre, DI Rienzo 2009). Il n'est donc pas du tout contradictoire de trouver certains gènes sous sélection négative lors d'une analyse comparant polymorphisme chez l'homme et divergence avec le chimpanzé (montrant qu'à l'époque de la spéciation de l'homme, ces gènes étaient plutôt contraints) et de trouver des signatures de sélection positive récente chez l'homme (indiquant qu'un ou plusieurs polymorphismes avantageux ont depuis augmenté en fréquence).

2) Contraintes sélectives au sein des deux familles de senseurs étudiées

Ainsi, au cours cette thèse, j'ai pu étudier la diversité génétique et détecter des signes de sélection au sein des gènes codant des senseurs cytosoliques humains, les NLRs (contenant les sous-familles des NALPs et des NOD/IPAF) et les RLRs. J'ai montré que la majorité des gènes codant les NALPs, probablement les moins connus des PRRs, ont été particulièrement contraints au cours de l'évolution. En effet, malgré leur grande diversité génétique, probablement due à des taux de mutations plus élevés dans leurs régions chromosomiques respectives, leur diversité protéique semble au contraire particulièrement contrainte. En effet, nos analyses inter-espèces (comparant l'homme et le chimpanzé) montrent que la majorité (10 parmi 14) des gènes codant les NALPs est soumise à une sélection purificatrice. Cela signifie que l'accumulation de mutations altérant la protéine n'est pas tolérée. Les valeurs d'ω du test MKPRF significativement inférieures à 1 mettent en effet en évidence ce déficit en mutations non-synonymes pour la majorité des gènes codant les NALPs. La plupart d'entre eux pourrait donc avoir joué un rôle essentiel à notre survie dans le passé, et probablement encore aujourd'hui. Par ailleurs, il est intéressant de constater l'absence apparente de pseudogénisation au sein des gènes codant les NALPs. Cette famille résulte en réalité de la duplication d'un gène unique, datant probablement d'avant la divergence des mammifères (Tian, Pascal, Monget 2009). La théorie de Susumu Ohno (1928-2000), considérant deux copies d'un gène obtenues par duplication de celui-ci, prévoit dans la majorité des cas l'évolution suivante (**Figure 14)** : une copie va être très conservée, afin de maintenir la fonction ancestrale essentielle du gène ; l'autre exemplaire va être moins contraint et peut potentiellement accumuler des mutations : dans la plupart des cas, ces mutations vont

inactiver le gène qui sera pseudogénisé (la copie reste sur le génome, mais aucune protéine n'est exprimée). Dans de rares cas, les mutations accumulées confèrent une nouvelle fonction avantageuse, qui pourra ensuite elle-même être très conservée ; et ainsi, deux fonctions différentes seront obtenues à partir d'un même gène de départ (néo-fonctionnalisation) (Ohno 1970; Force et al. 1999). On peut également distinguer un autre cas, celui de la sous-fonctionnalisation : les deux copies peuvent tour à tour accumuler des mutations de façon complémentaire, c'est-à-dire qu'un site muté dans une copie sera conservé dans l'autre ; et ainsi, les deux exemplaires du gène seront tous les deux indispensables pour maintenir la fonction active. Ceci a été proposé comme un mécanisme d'évolution « plus sûr », permettant de garder longtemps des copies intactes même si elles n'apportent pas immédiatement d'avantage supplémentaire, tout en maintenant une certaine flexibilité adaptative (Force et al. 1999) ; en effet, une copie pourra par hasard dans le temps acquérir des mutations conférant une nouvelle fonction (néo-sous-fonctionnalisation) (Dittmar, Liberles 2010).

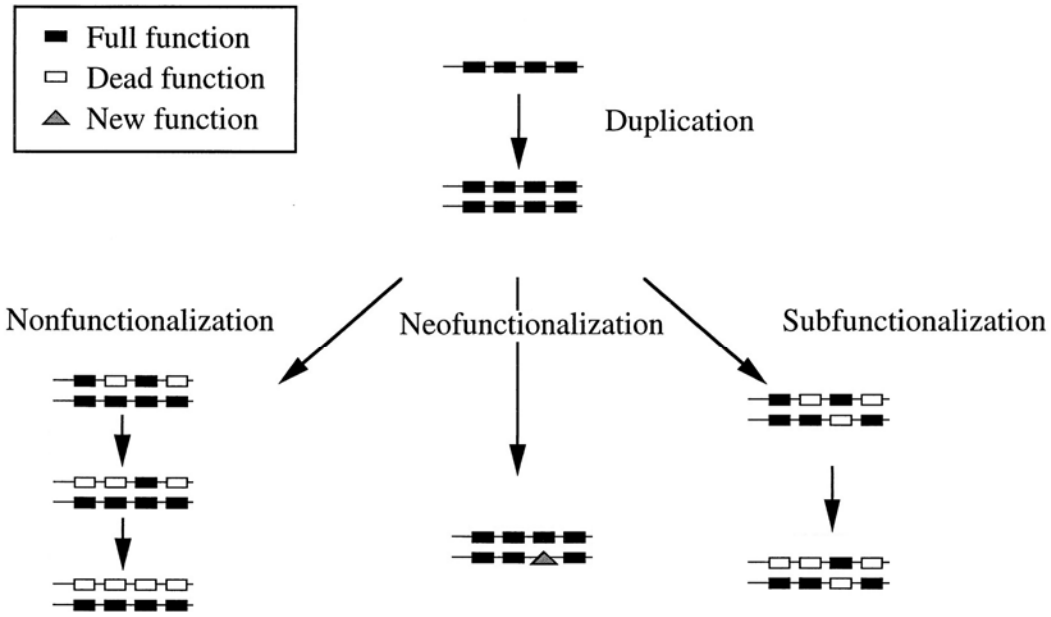


Figure 14. Evolutions possibles de deux copies résultant de la duplication d'un gène. Un gène est représenté par 4 rectangles (mutations) reliés entre eux. En cas de pseudogénisation (à gauche), l'une des deux copies devient non-fonctionnelle. D'autre part, l'une des copies peut acquérir une nouvelle fonction (au milieu). Enfin, les deux copies peuvent être complémentaires et indispensables pour assurer la fonction initiale (à droite). Il faut cependant noter qu'il est également possible que les deux copies aient la même fonction : c'est par exemple le cas de copies qui donnent les mêmes protéines, mais en quantité différentes. Adapté de (Force et al. 1999).

Le cas des gènes *DC-SIGN* et *L-SIGN*, codant deux récepteurs de type lectine (CLRs) et qui résultent d'une duplication, illustre très bien l'une des évolutions possibles de duplicats. *DC-SIGN* semble avoir été très contraint par la sélection, alors que *L-SIGN* accumule de nombreuses mutations non-synonymes et serait vraisemblablement sous sélection balancée dans les populations non-africaines; la cible de cette sélection serait probablement une région impliquée dans la reconnaissance des pathogènes, permettant ainsi la reconnaissance d'un répertoire plus large de pathogènes (Barreiro et al. 2005). Ce modèle peut s'appliquer de la même manière dans le cas de duplications successives. Les interférons (IFN), également étudiés au laboratoire, constituent un exemple intéressant des différentes possibilités quant au devenir possible des membres d'une famille multigénique résultant de duplications. Parmi les 21 gènes codant les IFN chez l'homme, seuls ceux de l'*IFN-γ* et des *IFN-α6, IFN-α8, IFN-α13* et *IFN-α14* sont sous sélection purificatrice, ceux-ci-ci assurant donc une fonction essentielle. *IFN-α10,* qui montre en particulier de nombreuses mutations stop à fortes fréquences, constitue probablement un exemple de gène en cours de pseudogénisation. Enfin, il semble que certains IFN de type III sont davantage sujets à des néo-fonctionnalisations, ceux-ci montrant de fortes signatures de sélection positive en Europe et en Asie (Manry et al. 2011b). Concernant les gènes codant les NALPs, nous avons trouvé 10 parmi les 14 identifiés chez l'homme sous fortes contraintes, 1 sous contraintes plus faibles et 3 qui semblent évoluer de manière plus adaptative, en accumulant des mutations non-synonymes dont certaines à forte fréquence dans la population. Il s'agit donc d'un cas peu classique pour une famille de gènes, nombre de ses membres ayant apparemment acquis une fonction assez importante pour être conservée (néo-fonctionnalisations). Aucun cas clair de gène en cours de pseudogénisation n'a pu être montré. Etant donné les fonctions proches tout en étant différentes des NALPs, on pourrait envisager que leur configuration actuelle chez l'homme résulte de néo-sous-fonctionnalisations successives.

Au contraire, la majorité des gènes codant les NOD/IPAF (6 parmi 7) et les RLRs ne montrent pas de valeur d'ω significative. Il semble donc que des contraintes moins fortes aient été exercées sur ces récepteurs, les laissant muter « plus librement ». Ils doivent par conséquent jouer un rôle plus secondaire dans notre survie, ou n'ont tout du moins pas acquis actuellement chez l'homme une fonction assez importante pour que la protéine soit totalement conservée. Seul NLRC5 se distingue des autres NOD/IPAF, le gène qui le code arborant des signatures de sélection purificatrice. On peut faire l'hypothèse que ses caractéristiques particulières (la reconnaissance de virus et son rôle dans l'activation du complexe du CMH I)

expliquent les contraintes importantes qui s'exercent sur lui. Il joue certainement un rôle essentiel dans la reconnaissance virale, les virus qu'il reconnaît ne lui permettant pas l'accumulation de mutations dans sa protéine. Par ailleurs, il est particulièrement surprenant de constater que *NOD2* n'est pas trouvé sous sélection négative (ni purificatrice, ni négative faible), suggérant qu'il aurait donc joué un rôle plus secondaire que la majorité des *NLRs* ou aurait évolué de manière adaptative. Or, il a été associé à plusieurs reprises à des maladies sévères : en particulier, plusieurs mutations dans *NOD2* s'avèrent être la composante génétique majoritairement responsable de la maladie de Crohn, une maladie inflammatoire chronique intestinale rare de nature auto-immune (Hampe et al. 2001; Hugot et al. 2001; Ogura et al. 2001b). Il faut toutefois souligner que les tests de neutralité montrent pour *NOD2* un excès très significatif d'allèles de faibles fréquences dans les populations non-africaines (et également un excès d'allèles de fortes fréquences en Asie) : étant données les valeurs extrêmes obtenues pour les D* et F* de Fu & Li (les plus fortes observées en Asie au sein des *PRRs* étudiés), cela suggère qu'il est certainement sous sélection positive dans ces populations ; cependant les tests que nous avons employés n'ont probablement pas permis de détecter de sélection positive dans *NOD2* de manière assez soutenue pour être abordée dans l'article (résultats significatifs pour deux tests de types différents au minimum). Si l'on revient aux définitions des trois types de gènes que j'ai faites au début de cette partie, *NOD2* se placerait donc dans le troisième cas : les gènes qui peuvent être essentiels, mais qui montrent actuellement des signatures de sélection positive : ils sont donc « en cours » de modification, acquérant potentiellement une nouvelle fonction qui s'avèrera éventuellement très conservée dans des analyses réalisées dans le futur.

Concernant l'étude détaillée que j'ai faite des *RLRs*, (voir le premier article), je ne discuterai ici que de l'intérêt que de telles comparaisons peuvent apporter. Nous avons en effet montré que *RIG-I* et *IFIH1*, codant deux RLRs qui possèdent des structures et fonctions très similaires, arborent des diversités génétiques très différentes : *RIG-I* est beaucoup plus contraint qu'*IFIH1*, en particulier en matière de mutations fonctionnelles. RIG-I a donc probablement déjà acquis une fonction plus importante qui est relativement plus conservée ; *IFIH1*, pour lequel nous trouvons des signes de sélection positive, accumule encore des mutations non-synonymes, éventuellement en vue de gagner de nouvelles propriétés. Plus intéressant encore, les changements d'acides aminés que l'on trouve dans les deux protéines ne sont pas distribués de la même manière : RIG-I en accumule très peu dans sa région LRR, impliquée dans la reconnaissance de l'ARN, alors que les changements d'acides aminés sont beaucoup plus nombreux et à plus fortes fréquences dans IFIH1. Ceci pourrait également être

expliqué par la localisation de la cible potentielle de sélection positive d'*IFIH1* dans cette région. Cependant, cela n'explique pas pourquoi *RIG-I* par ailleurs si conservé accumule de nombreuses mutations dans le domaine CARD, impliqué dans la signalisation menant à la production d'interférons, domaine indispensable également. Nos résultats suggèrent donc que ces protéines doivent subir des contraintes différentes en matière de ligands ou n'utilisent pas leurs domaines de la même manière. Ceci ouvre en particulier la voie à des analyses structurales plus précises des complexes ligands-récepteurs.

3) Intégration de ces résultats dans le contexte général des PRRs humains

Afin d'intégrer ces résultats dans un contexte plus général, nous les avons comparés à ceux précédemment obtenus pour d'autres récepteurs microbiens, les *Toll-like receptors* (TLRs). Nous proposons un modèle hiérarchique général mettant en évidence les contraintes trouvées et par déduction les contributions relatives des principales familles de senseurs microbiens à la survie de l'homme (**Figure 15**).

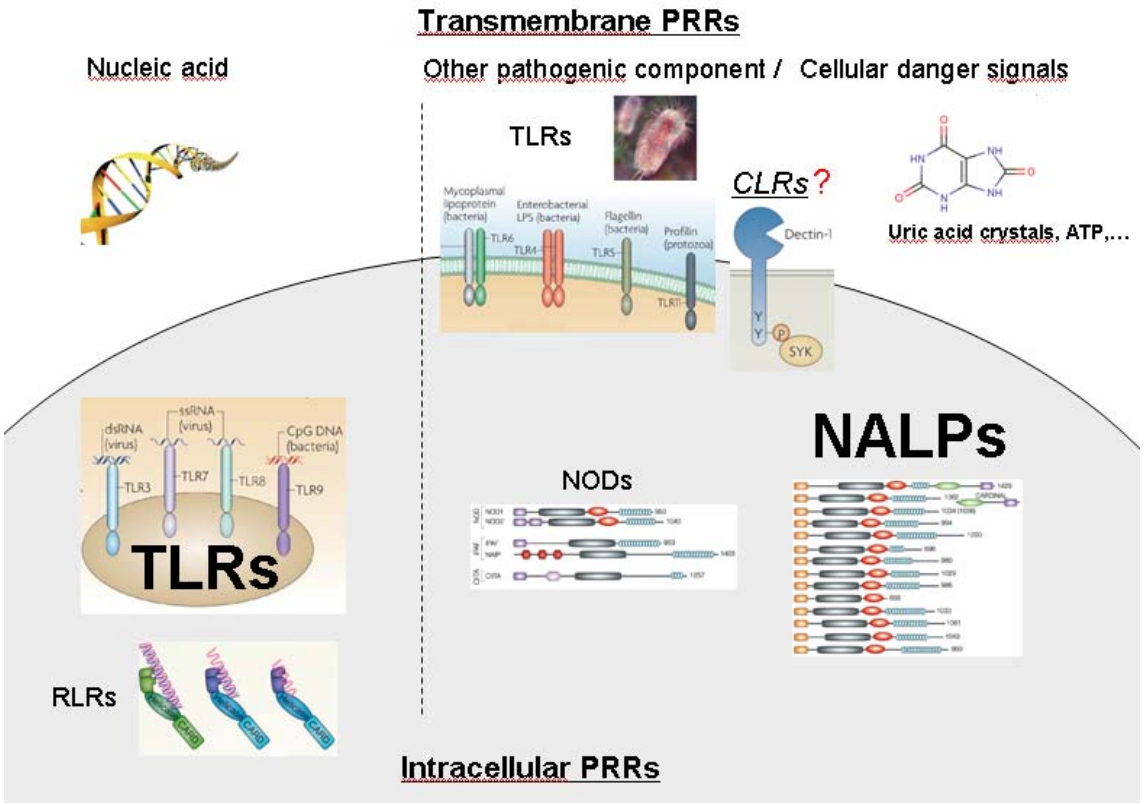


Figure 15. Modèle général hypothétique basé sur les contraintes sélectives observées et proposant les contributions relatives des différentes familles de PRRs humains à notre survie. On distingue ici les différents types de ligands et les diverses localisations cellulaires de ces familles (malgré l'existence d'exceptions). Sont précisées en gras et avec une police de taille supérieure les

familles dont la majorité des membres sont sous sélection purificatrice, et assurent probablement des fonctions essentielles. Les CLRs sont indiqués à titre informatif sur ce schéma, mais les résultats pour cette famille ne sont toujours pas disponibles à l'heure actuelle.

Parmi les senseurs d'acides nucléiques, principalement d'origine virale, les TLRs endosomaux semblent remplir des fonctions biologiques plus importantes que les RLRs. En effet, les gènes codant les TLRs endosomaux (*TLR3, 7, 8* et *9*) s'étaient avérés être très contraints, montrant des signes de sélection purificatrice (Barreiro et al. 2009). Dans l'analyse présente (premier article), j'ai pu mettre en évidence les contraintes beaucoup plus faibles qui étaient exercées sur les *RLRs* : aucun d'entre eux ne se trouve sous sélection purificatrice et ils ont tendance à accumuler des mutations non-synonymes dont certaines à fortes fréquences dans les différentes populations humaines. Ainsi, étant donné que 3 des TLRs endosomaux reconnaissent, comme les RLRs, des ARN provenant principalement de virus et que les spectres des virus reconnus par ces deux groupes semblent comparables, on peut faire l'hypothèse suivante : les gènes codant les TLRs endosomaux ont été particulièrement contraints du fait du rôle essentiel de ces récepteurs, très certainement dans la détection virale ; ainsi par exemple, il a été montré qu'un allèle du gène de *TLR3*, qui est dominant-négatif (la protéine codée par un chromosome muté suffit à altérer le fonctionnement de celle du chromosome homologue « normal »), est associé au développement rare mais sévère de l'encéphalite herpétique (due au virus herpès simplex de type 1, HSV-1) (Zhang et al. 2007b). Quant aux gènes codant les RLRs, qui remplissent des fonctions similaires à celles des TLRs, ils peuvent donc muter plus « librement ». On aurait donc un cas de redondance des senseurs viraux.

En ce qui concerne les récepteurs spécialisés dans la reconnaissance d'autres pathogènes (bactéries et parasites) et de signaux de danger cellulaire, 10 des 14 *NALPs* sont sous sélection purificatrice, contre 1 des 7 *NOD/IPAF* et aucun des gènes codant les TLRs se trouvant à la surface cellulaire. On peut de la même manière conclure que la plupart des NALPs (et NLRC5) pourraient jouer un rôle plus important dans ce type de reconnaissance, comparés aux autres NLRs et aux TLRs à la surface cellulaire. Les contraintes sont cependant beaucoup moins claires pour les *NLRs* sous contraintes que pour les *TLRs* endosomaux. En effet, les *TLRs* montrent de manière générale un très faible nombre de mutations non-synonymes, ce qui rend évidente leur essentialité. En revanche, les *NALPs* sous sélection purificatrice et *NLRC5*, arborent dans la plupart des cas un nombre non-négligeable de mutations non-synonymes. En particulier, pour *NLRC5*, qui est sous sélection purificatrice, on trouve un nombre incroyablement élevé de ce type de mutations ; mais c'est aussi une

protéine particulièrement longue, comparée aux autres NLRs (voir figure 1, article 2) et aux TLRs (1866 acides aminés pour NLRC5 contre 1049 pour TLR7, la protéine la plus longue des TLRs, Uniprot, http://www.uniprot.org/), ce qui pourrait expliquer la différence en nombres de mutations non-synonymes. Il n'en reste pas moins que la proportion de mutations non-synonymes comparée à celle des synonymes doit être assez faible pour que le test MKPRF indique de la sélection purificatrice pour ces *NLRs*.

D'autre part, nos résultats n'ont pas mis en évidence de variations significatives au cours du temps de l'intensité des contraintes exercées sur les gènes sous sélection purificatrice (gènes codant les NALPs, NLRC5 et les TLRs endosomaux) (voir figure S11 de l'article 2). L'hypothèse la plus probable reste donc qu'ils étaient contraints chez l'ancêtre commun à l'homme et au chimpanzé et le sont toujours dans la lignée humaine. En ce qui concerne les gènes qui ont un γ significatif, nous n'avons pas pu évaluer les variations de ces contraintes de cette manière, l'odds ratio ne réflétant que la valeur du γ elle-même ; en d'autres termes, l'odds ratio tout comme le γ montrent un excès de polymorphisme comparé à la divergence aux sites non-synonymes, caractéristique de la sélection négative faible identifiée pour la plupart de ces gènes. Il n'en reste pas moins que ces contraintes peuvent être exercées à différents degrés d'intensité au cours du temps, certaines pouvant par exemple s'avérer spécifiques à l'homme. Je soulignerai à cette occasion la possibilité de biais lors de l'interprétation chez l'homme de fonctions établies dans d'autres espèces, en particulier pour les fonctions de l'immunité. En effet, l'homme, comme toutes les espèces vivantes, a à sa disposition un certain nombre de molécules pour assurer sa défense contre les intrusions de pathogènes ou de substances toxiques : au cours de l'évolution, en particulier après chaque spéciation, certaines de ces molécules peuvent devenir moins essentielles, voire inutiles et d'autres peuvent au contraire permettre l'acquisition de nouvelles fonctions ou de mécanismes plus efficaces (Ohno 1970; Force et al. 1999). Il est donc attendu de trouver chez l'homme un répertoire comprenant des récepteurs dont la fonction est partagée par plusieurs espèces et des récepteurs qui lui sont donc propres (Schroder, Tschopp 2010). Ceci résulte de la notion-même de sélection naturelle, c'est-à-dire d'adaptation d'une espèce donnée à un environnement donné. Deux espèces distinctes peuvent avoir différentes résistances aux températures, les pathogènes qui affectent une espèce peuvent ne pas affecter une autre, etc. Comme souligné par l'histoire humoristique imaginée par Richard Dawkins (1941), la sélection n'est pas une compétition inter-espèce mais intra-espèce.

Deux brontosaures voient un tyrannosaure royal avancer dans leur direction et se mettent à courir aussi vite qu'ils le peuvent. Puis l'un des deux dit à l'autre : « Pourquoi nous fatiguons-nous au juste ? Nous n'avons de toute façon pas la moindre chance d'arriver à courir plus vite qu'un tyrannosaure ! »

Et l'autre lui répond cyniquement : « Je ne cherche pas à courir plus vite que le tyrannosaure. Je cherche juste à courir plus vite que toi ! »

Et ceci est d'autant plus important lorsque l'on étudie des récepteurs de l'immunité innée : il vaut mieux être très précautionneux lorsque l'on infère des fonctions d'une espèce à l'autre. Ainsi, cette remarque semble être particulièrement vraie chez les NLRs, pour lesquels certains résultats semblent contradictoires, notamment parce que certaines études comparent des fonctions chez la souris avec celles des protéines humaines. Il s'avère d'ailleurs par exemple que NOD1 ne possède pas les mêmes types de ligands chez la souris et chez l'homme (Magalhaes et al. 2005). On peut également observer des différences évolutives majeures dans des espèces phylogénétiquement proches telles que l'homme et le chimpanzé : ainsi par exemple, on constate un relâchement important des contraintes sélectives au sein des *TLRs* humains, comparé au chimpanzé (Wlasiuk et al. 2010). Or, l'un des intérêts ici de l'approche de génétique des populations est qu'elle nous permet de proposer des cibles qui ont probablement un effet fonctionnel, et ce dans l'espèce considérée.

4) Des rôles peu classiques pour des PRRs

L'établissement de la comparaison entre les différents senseurs de produits bactériens ou de signaux de danger cellulaire (NALPs, NODs, TLRs à la surface cellulaire) est plus complexe que pour les senseurs d'acides nucléiques, étant donné l'hétérogénéité des fonctions de ces familles. Si les rôles démontrés pour les TLRs sont clairement ceux de senseurs microbiens, nombreux sont les NLRs dont on ne connaît pas les ligands ou qui ont été impliqués dans des rôles moins classiques pour des PRRs.

De manière générale, la présence de nombreux NALPs dans les gamètes et les cellules reproductives uniquement suscite tout d'abord quelques interrogations sur ces molécules, qui montrent tout de même une structure classique de PRRs (en particulier un domaine LRR, généralement impliqué dans la reconnaissance de pathogènes) (Kobe, Deisenhofer 1995; Rock et al. 1998; Zhang et al. 2008). On peut ajouter à cela le fait que pour bon nombre de NALPs (tous exceptés NALP1, 2 et 3), aucun ligand direct n'a été identifié précisément. Enfin, certains d'entre eux semblent clairement impliqués dans l'homéostasie cellulaire et le développement embryonnaire (Kufer, Sansonetti 2010). L'analyse

phylogénétique des NALPs chez les mammifères avait conclu à deux groupes, l'un spécialisé dans l'immunité, l'autre dans le développement (Tian, Pascal, Monget 2009). Néanmoins, les informations fonctionnelles dont nous disposons sont loin d'être exhaustives et de plus en plus d'études mettent en évidence un aspect crucial : ces différentes fonctions peuvent être associées à une même protéine. Ainsi par exemple NALP6, qui active la production d'IL-1β et participe donc à la réponse inflammatoire, est surexprimée dans les cellules intestinales du fœtus chez la souris (Kempster et al. 2010); d'autre part, le promoteur du gène qui le code s'avère être régulé par PPAR-γ (Peroxisome Proliferator-Activated Receptor- γ), un facteur de transcription impliqué dans l'homéostasie intestinale (Dubuquoy et al. 2006). Quant à NALP7, il active la voie NF*κ*B, mais certaines des mutations du gène qui le code chez l'homme semblent responsables de grossesses anormales, pouvant être provoquées par un mauvais développement de l'embryon (Murdoch et al. 2006; Qian et al. 2007). Cela souligne certainement un fonctionnement beaucoup plus complexe que supposé pour ces molécules. Un point important, il est communément admis que les différentes fonctions de notre organisme évoluent en forme et en intensité tout au long de notre vie : on n'aura ainsi pas le même métabolisme, ni les mêmes défenses immunitaires à la naissance, l'adolescence, l'âge adulte, ou à un âge plus avancé (Wedderburn et al. 2001; Troelsen 2005; Salvioli et al. 2006; Reik 2007; Kempster et al. 2010). Paradoxalement, de nombreuses études cherchent à établir une fonction unique pour une même protéine au cours du temps et dans différentes conditions. Si le gain de connaissances n'en est pas moins grand, il est nécessaire de garder à l'esprit la possibilité d'autres rôles et de ne pas chercher à tout prix à classer chaque protéine dans une catégorie bien distincte. Dans ce cadre, il est intéressant de noter que NALP3, dont le gène s'avère être le plus contraint des *NLRs*, semble jouer d'une part un rôle essentiel en tant que senseur (sa localisation est limitée à des cellules se trouvant à des positions stratégiques de risques d'infections et il est impliqué dans des maladies rares et sévères) (Hoffman et al. 2001; Aganna et al. 2002; Aksentijevich et al. 2002; Dode et al. 2002; Feldmann et al. 2002; Kummer et al. 2007) ; cependant, il pourrait d'autre part jouer un rôle plus général que celui de senseur microbien : le taux de survie chez les souris dont le gène *NALP3* a été éteint est très faible, comparé à celles dont l'expression de caspase-1 (molécule activée par NALP3) a été abolie (Allen et al. 2009). Cela laisse donc supposer chez l'homme la possibilité pour NALP3 (et potentiellement pour les autres NALPs dont les gènes sont sous contraintes fortes) de rôles plus généraux que l'immunité face aux infections et qui affectent directement la survie. Les NALPs pourraient ainsi se montrer essentiels, car ils exercent une fonction à la fois très générale et fondamentale, mais aussi variable au cours de l'existence, fonction que je

définirais comme «la préparation du système immunitaire durant le développement embryonnaire, et le maintien de l'homéostasie cellulaire plus tard dans notre vie ». Au centre de ce système, il semblerait que l'activation de l'inflammasome et la production d'interleukines jouent des rôles prépondérants (Martinon, Mayor, Tschopp 2009). L'IL1-β en particulier, qui résulte de la formation de l'inflammasome, joue un rôle essentiel dans l'ovulation et la maturation des oocytes (Gerard et al. 2004). On pourrait par ailleurs faire l'hypothèse de rôles semblables pour les autres PRRs déclenchant la production d'interleukines, en particulier chez les TLRs. Il ne faut pas oublier l'origine de la découverte de cette famille : chez la drosophile, *Toll-1* s'est avéré jouer un rôle dans son développement dorso-ventral (Anderson, Jurgens, Nusslein-Volhard 1985). Cependant, aucune fonction affectant clairement le développement embryonnaire n'a pour l'instant été mise évidence pour les TLRs des mammifères (« *The role of Toll receptors in innate immunity* » de Zlatco Dembic, extrait de (Rich 2004)). En ce qui concerne l'implication des TLRs dans l'homéostasie, on dispose par contre d'exemples assez soutenus (Rakoff-Nahoum et al. 2004; Zhang et al. 2006).

Si nous revenons aux conclusions de cette étude, il est donc possible que ces fonctions peu classiques pour des PRRs aient contribué aux contraintes observées. Et ainsi, même si notre étude ne permet pas d'expliquer en définitive pourquoi les *NALPs* ont été plus contraints sélectivement, nous avons tout de même mis en évidence la forte probabilité qu'ils jouent un rôle essentiel et donc leur implication potentielle dans des maladies ayant des phénotypes sévères. Nous avons pu de plus faire des hypothèses sur les raisons possibles de ces contraintes, en s'appuyant sur les associations cliniques trouvées dans la littérature.

5) Sélection positive et adaptation

D'autre part, nous avons également identifié quelques-uns des gènes étudiés sous sélection positive. En d'autres mots, ces gènes évoluent de manière adaptative, arborant localement ou sur de larges régions un fort polymorphisme au sein des populations humaines : certaines de leurs mutations ont fortement augmenté en fréquence dans une ou plusieurs populations humaines car elles ont dû conférer dans le passé un avantage en matière de survie. Elles ont pu éventuellement permettre aux récepteurs qu'ils codent d'acquérir une nouvelle fonction ou une plus grande efficacité par exemple. Ces mutations nous différencient encore aujourd'hui et pourraient notamment être responsables de nos différences actuelles de résistances/susceptibilité face à certaines infections. Nos analyses intra-populationnelles ont en particulier montré que *NLRP1* était sous sélection positive. Plus précisément, nous avons

clairement mis en évidence qu'un haplotype a été sélectionné positivement dans les trois populations, cet haplotype étant apparemment déjà fixé en Asie. Il est probable qu'il soit apparu il y a longtemps (avant la migration des populations humaines hors d'Afrique), puisque les 3 populations le portent. Il a dû conféré très tôt un avantage, probablement en matière de résistance à certaines maladies infectieuses, étant donné le rôle de *NLRP1* (Martinon, Burns, Tschopp 2002; Faustin et al. 2007). Et ainsi, le goulot d'étranglement en Asie a pu conduire à la fixation de cet haplotype ; quant à l'Europe, les individus de la population fondatrice devaient contenir une faible proportion d'haplotype ancestral, de même que la population africaine dont ils ont émergé. Une deuxième solution, toutefois moins plausible, serait que la sélection ait agi après migration, conférant un avantage commun aux trois régions.

Nous avons également trouvé d'autres signaux moins forts de sélection sur *NLRP1* et sur d'autres gènes. Je ne discuterai que peu de ces signaux (voir article 2) qui devront être confirmés par des études fonctionnelles. Reste que le cas d'un événement potentiel de sélection positive sur la mutation V1059M de *NLRP1* en Europe peut constituer un exemple intéressant. Il semble que cette mutation ou une mutation génétiquement associée (étant donné la proximité des natures chimiques de ces deux acides aminés) ait été sélectionnée positivement, car elle a conféré un avantage dans le passé, probablement une résistance à certaines infections. Or, l'analyse du déséquilibre de liaison (associations génétiques) de ce gène suggère que cet événement, par entraînement, a pu augmenter en fréquence une autre mutation L155H associée de nos jours à plusieurs maladies auto-immunes (Jin et al. 2007; Magitta et al. 2009; Zurawek et al. 2010). Nos résultats fournissent un exemple supplémentaire des conséquences actuelles des effets de la sélection sur nos génomes d'autrefois. En effet, de nombreuses études confirment le maintien d'allèles conférant un risque plus grand de développer des maladies auto-immunes ou inflammatoires, telles que la maladie de Crohn, la maladie cœliaque ou encore la sclérose en plaque, résultant d'événements de sélection passées en réaction aux pathogènes (Barreiro, Quintana-Murci 2010; Sironi, Clerici 2010). Le cas le plus documenté est certainement celui de la maladie cœliaque, dans lequel il est montré que les individus atteints bénéficient d'une protection supérieure contre certains agents infectieux ; or, les allèles qui augmentent le risque de développer la maladie cœliaque semblent avoir été sélectionnés positivement (Barreiro, Quintana-Murci 2010; Zhernakova et al. 2010; Abadie et al. 2011).

Cependant, la sélection n'agit pas toujours de la même manière vis-à-vis des maladies. Le cas de sélection positive sur *IFIH1* (l'un des trois RLRs) en est un exemple : l'un des

variants associés à une résistance à diverses maladies auto-immunes pourrait au contraire avoir été augmenté en fréquence sous l'effet de la sélection : ce variant est associé à une plus faible réponse inflammatoire et donc un plus faible risque (Nejentsev et al. 2009; Vasseur et al. 2011). Un cas encore différent est celui de *CIITA*, qui montre que des mutations de son promoteur conférant une résistance à des maladies auto-immunes ont vraisemblablement été sélectionnées positivement en Europe ; or, l'une de ces mutations est associée à la fois à une plus forte expression des molécules du CMH II (et donc une plus forte réponse immunitaire) et à une plus faible susceptibilité aux maladies auto-immunes (Swanberg et al. 2005). Ainsi, le fonctionnement des récepteurs de l'immunité et des voies associées n'est pas toujours facile à élucider. Si notre survie repose sur la capacité du système immunitaire à réagir, il peut dans certains cas être avantageux de montrer une réponse diminuée, notamment lorsqu'une réaction trop forte (réponse inflammatoire trop importante) peut être fatale (Johnson et al. 2007; Ma et al. 2007; Barreiro et al. 2009). Les maladies auto-immunes constituent l'exemple le plus extrême, dans lequel les mécanismes eux-mêmes de l'immunité déclenchent une réponse inflammatoire en l'absence de véritables signaux de danger (Witebsky et al. 1957; Steiner, Volpe 1961; Rose, Bona 1993).

Il n'en reste pas moins que tous les variants que nous avons identifiés comme cibles potentielles de sélection positive pourraient avoir un impact fonctionnel chez l'homme. Et ainsi, cette étude a un rôle prédictif, nos résultats ouvrant la voie à de futures études biochimiques, immunologiques et cliniques.

CONCLUSIONS ET PERSPECTIVES

I) Les apports de la génétique des populations

L'approche de génétique des populations peut s'avérer utile, et ce de différentes manières. Pour les protéines dont la fonction est déjà connue, ce type d'étude peut permettre de mieux comprendre leur évolution ; on peut par exemple comprendre pourquoi une mutation conférant un plus grand risque de développer un diabète peut se trouver à forte fréquence. Pour les protéines peu connues, comme les NLRs, on peut ainsi prédire quelles sont celles qui ont *a priori* plus de chances d'être impliquées dans des maladies graves ou des différences de résistance à certaines pathologies.

On pourrait cependant reprocher à ce type d'étude l'absence *in fine* de conclusion définitive. Et même si nous pouvons apporter de précieuses informations, la plupart de nos hypothèses devront être validées par des études fonctionnelles. Toutefois, c'est là même le principe de cette approche qui a un aspect prédictif, permettant à des chercheurs des domaines cliniques et médicaux de cibler plus facilement les gènes ou variants à étudier dans une perspective fonctionnelle ; cela peut ainsi éviter que le choix des gènes ou variants à tester soit fait de manière arbitraire ou qu'il se base sur des ressemblances structurales avec d'autres protéines connues, ou sur des connaissances de ces molécules dans d'autres espèces uniquement. En particulier dans cette étude, on pourrait aussi souligner la différence entre le cadre de travail de départ et l'interprétation que nous avons finalement pu faire. En effet, nous avons étudié ces familles de gènes, en ayant pour but d'interpréter leurs différences de diversité génétique par des différences d'importance des senseurs microbiens qu'ils codent. Nous avons proposé un modèle précisant les différentes contraintes sélectives exercées sur chacune de ces familles, en fonction de leurs ligands potentiels ; ainsi par exemple, notre modèle suggère que la majorité des NALPs a joué un rôle important dans la reconnaissance des bactéries et des signaux de danger cellulaire, comparés à la plupart des NOD/IPAF. Or, si les PRRs étudiés ont tous des structures en relation avec cette fonction de senseur microbien et qu'il est clair que la plupart interviennent dans l'inflammation en général, nombre d'entre eux n'ont toujours pas été véritablement identifiés en tant que véritables senseurs. Il est fortement possible qu'ils ne détectent en réalité pas directement ces molécules et l'on s'aperçoit de plus en plus de leur intervention dans des fonctions très variées, telles que la reproduction et le développement embryonnaire. Néanmoins, cette étude nous aura permis de proposer des hypothèses concernant les contributions relatives de ces molécules à notre survie, et donc celles à étudier en priorité.

II) Remarques méthodologiques

D'un point de vue méthodologique, plusieurs remarques peuvent être faites. Un premier point qui pourrait venir à l'esprit concerne la technique de biologie moléculaire utilisée. Le long et fastidieux reséquençage de type Sanger qui a été réalisé pour obtenir ces données pourrait paraître obsolète et peu utile, étant donnés les techniques de séquençage de dernière génération. En effet, des projets comme le projet *1000 genomes* permettent d'obtenir en très peu de temps des jeux de données comparables en matière de représentation de populations, mais de taille largement supérieure. Cependant, le séquençage de type Sanger fournit des données beaucoup plus robustes que les techniques de dernière génération : en particulier, de nombreuses régions séquencées par le le projet *1000 genomes* ont une faible couverture, diminuant la probabilité de détecter correctement les allèles rares et à très fortes fréquences ; or, pour notre approche, comparant différentes régions génomiques entre elles pour des populations variées, il est essentiel d'obtenir des données de qualité homogène, les variants à faibles et fortes fréquences étant particulièrement importants pour la détection de sélection naturelle. De plus, la présence pour certains des gènes étudiés de longues régions codantes riches en GC ou totalement identiques rend inutilisable les techniques « récentes », qui permettent pour le moment d'obtenir une centaine de paires de bases. Le séquençage Sanger permet généralement d'obtenir des séquences d'environ 400 à 500 paires de bases, ce qui était suffisamment long pour différencier les régions de forte identité des gènes étudiés.

Si on s'intéresse maintenant à l'analyse, il est important de préciser que la significativité de nos données dépend hautement des modèles démographiques qui ont été simulés. Or, ceux-ci sont très difficiles à estimer de manière exacte, étant donnés l'ancienneté des événements et la quantité de paramètres variables. Néanmoins, les événements intégrés dans les modèles démographiques sont directement basés sur les résultats d'études archéologiques, anthropologiques appuyés par des données génétiques. Et si les modèles utilisés ici ne prennent pas en compte toutes les variations, elles considèrent tout de même les événements les plus soutenus et donc probablement ceux qui ont été assez forts pour influencer notre diversité de manière significative.

D'autre part, le *multiple testing*, c'est-à-dire l'accumulation de différents tests pour un même jeu de données peut conduire à une augmentation de faux positifs. En d'autres termes, dans notre étude, plus on va réaliser de tests de détection de sélection pour un gène, plus on va trouver de signaux qui ne sont qu'en fait des résultats du hasard. Certains tests permettent d'éviter ce problème, en comparant les niveaux de significativité pour des comparaisons

simples ou multiples ; cependant, notre quantité de données n'est généralement pas assez élevée pour atteindre la significativité. Pour compenser cela, nous avons uniquement considéré comme sous sélection les gènes dont plusieurs signaux de natures différentes (par exemple un test basé sur le déséquilibre de liaison et un basé sur les fréquences alléliques) confirmaient l'existence d'un signal de sélection naturelle.

Enfin, dans toutes ces analyses, nous avons considéré les SNPs uniquement. Il est clair qu'il s'agit des mutations les plus courantes, en particulier dans les régions géniques, et qu'ils apportent donc une grande quantité de données. Cependant, les insertions et délétions peuvent également constituer de précieuses informations ; dans notre cas, concernant les *RLRs*, aucun indel n'a été trouvé dans les régions codantes ; quant aux *TLRs* et *NLRs*, certains indels ont été trouvés dans les exons codants, mais uniquement à très faible fréquence et jamais à l'état homozygote dans le panel étudié ; de plus, dans la majorité des cas, ces indels affectent la fin de la protéine, suggérant qu'ils auraient un effet faible sur celle-ci, ou tout du moins potentiellement plus faible que lorsqu'un indel affecte la protéine dès son début ou au milieu. Ainsi, on peut en déduire que les *PRRs* étudiés ne peuvent accumuler ce genre de mutations souvent responsables de l'apparition d'un codon stop prématuré menant à un raccourcissement de la séquence protéique. Cela suggère que les PRRs sont impliqués dans un système qui n'est pas très flexible. On peut supposer que même les gènes sous contraintes plus relâchées (comme les *RLRs* par exemple) ne peuvent accumuler ce type de mutation, car l'expression d'une protéine raccourcie pourrait avoir de lourdes conséquences pour la cellule, si elle interagit avec les autres molécules intracytosoliques. Ainsi, les indels n'apportent pas ici de conclusions majeures permettant de distinguer les PRRs les uns des autres.

III) Conclusions sur les PRRs et perspectives possibles

Ainsi, nous avons montré que certains PRRs, avaient d'une manière générale davantage contribué à la survie de l'homme. Il serait donc pertinent de les étudier en priorité dans une perspective clinique ou épidémiologique. La compréhension plus précise des mécanismes impliqués dans la formation des inflammasomes et de leur champ d'action pourrait s'avérer essentielle. A cette fin, il semble indispensable d'éviter les barrières disciplinaires pour étudier ces complexes, au domaine d'action potentiellement très large. La résolution de ces mécanismes nécessiterait une prise en compte précise dans chaque étude des paramètres expérimentaux, du type cellulaire étudié, du répertoire de récepteurs présents, etc. Ainsi par exemple, suivant les études considérées, certains NLRs montrent alternativement des rôles d'activateur ou d'inhibiteur de la réponse inflammatoire (Inohara et al. 2001; Ogura et al. 2001a; Watanabe et al. 2004; Kufer, Sansonetti 2010) ; on sait d'autre part qu'il peut exister des interactions en hétérodimères telles que certaines combinaisons de TLRs ou celle de NALP1/NOD2 (Akira, Uematsu, Takeuchi 2006; Hsu et al. 2008). Ainsi, outre les conditions expérimentales, on peut penser que le répertoire de PRRs présents et interagissant avec le récepteur étudié va influencer son comportement (Zhang et al. 2002; Triantafilou et al. 2004). Les gènes et variants qui semblent avoir été particulièrement ciblés par la sélection pourraient être étudiés de la manière la plus exhaustive possible. Tout d'abord, il conviendrait de localiser clairement les types cellulaires dans lesquels on peut trouver ces molécules, et en quelle quantité. Enfin, une analyse des propriétés chimiques des NLRs, conjointement à celle de leur localisation cellulaire pourrait fournir une idée précise de ceux qui peuvent interagir avec celle testée. On aurait ainsi une meilleure idée de son domaine d'action. Ces études nécessiteraient bien sûr du temps et des fonds importants. Cependant, elles paraissent indispensables si l'on veut comprendre les mécanismes qui gouvernent des molécules se trouvant à l'intersection de plusieurs fonctions majeures de notre organisme.

Par ailleurs, les variants que nous avons identifiés comme cibles de sélection devraient être considérés dans les études d'association avec diverses pathologies. Les GWAS (*Genome-Wide Association Studies*) s'avèrent essentielles dans la découverte de mutations impliquées dans la susceptibilité à des maladies : ces études à l'échelle du génome qui comparent des individus sains et malades détectent donc les variants ayant un effet relativement fort sur l'ensemble du phénotype complexe qu'est « la maladie ». Or, certains variants peuvent jouer un rôle important dans une pathologie, sans être repérés par ces analyses souvent très conservatives ; la recherche d'eQTL (*expression Quantitative Trait Loci)* vise à identifier les

variants qui modifient un trait phénotypique beaucoup plus ciblé, le niveau d'expression des gènes, c'est-à-dire leur abondance de transcrit. Ainsi, elle vient compléter les GWAS dans la compréhension des mécanismes impliqués dans des maladies et dans la détermination des mutations impliquées dans celles-ci. Il semble d'ailleurs que les eQTL constituent de très bonnes cibles de sélection naturelle et sont donc potentiellement impliqués dans de nombreuses maladies complexes. En effet, les variations du niveau d'expression de certains gènes entre populations et entre individus se trouvent souvent contraintes par la sélection (Drake et al. 2006). D'autres au contraire montrent des signatures de sélection positive, comme dans l'exemple de la lactase, précédemment cité (Tishkoff et al. 2007; Enattah et al. 2008; Itan et al. 2010). De manière générale, il a même été suggéré que les eQTL étaient des cibles privilégiées de sélection positive récente et que certains variants associés à des traits phénotypiques étaient plus probablement des eQTL (Nicolae et al. 2010). Dans le cadre de notre étude, nous avons vu le cas de *CIITA*, dont la sélection positive semble avoir cibléle promoteur en Europe ; l'une des cibles proposées semble modifier le niveau d'expression de sa protéine (Swanberg et al. 2005). Et ainsi, l'intégration des données de génétique de populations aux études d'associations cliniques et analyses du transcriptome nous permettra de mieux comprendre la complexité phénotypique humaine.

IV) Variations phénotypiques et épigénome

Pour finir, j'ai parlé jusqu'ici des différents types de mutations pouvant affecter notre génome. Or, il existe d'autres mécanismes qui peuvent altérer l'expression des gènes, sans pour autant modifier l'enchaînement de nucléotides de notre ADN, c'est l'épigénétique.

Ainsi tout d'abord un exemple évident, chacun d'entre nous possède un grand nombre de cellules, toutes codées par la même séquence d'ADN, mais qui s'avèrent pourtant très différenciées. Il est donc évident que des variations indépendantes de la séquence nucléotidique interviennent. Dans le cas ici de la différenciation cellulaire, on sait qu'il y a un rapport avec la structuration de l'ADN qui va affecter l'expression de certains gènes : la différenciation se met en place par « extinction » progressive du répertoire des gènes pouvant être exprimés (Reik 2007). Les mécanismes épigénétiques incluent notamment les modifications covalentes de l'ADN et des histones (protéines permettant la compaction de l'ADN), ou encore la régulation par de petits ARN non-codants, les micro-ARN (Klose, Bird 2006; Rassoulzadegan et al. 2006). L'un des mieux connus est probablement la méthylation de l'ADN (Riggs 1975). On sait qu'elle joue un rôle prépondérant dans l'inactivation de gènes, en ciblant particulièrement les sites CpG chez les mammifères (Ehrlich et al. 1982; Clark, Harrison, Frommer 1995) ; il n'est donc pas étonnant de constater que 60% des gènes humains présentent des sites CpG dans leur promoteur. La méthylation participe également à l'inactivation du chromosome X (chez la femme, seul l'un des deux X est exprimé), ou encore à la reprogrammation génétique au cours du développement (Riggs 1975; Clark, Harrison, Frommer 1995; Chow et al. 2005; Reik 2007); on peut également citer l'empreinte parentale : il existe des gènes dont l'activité est soumise à une empreinte : les deux copies de ce gène portées par un même individu montrent des activités différentes, selon leur origine (maternelle ou paternelle dans le cas de l'empreinte parentale). Ces gènes nécessitent en fait que l'une des copies parentales soit réprimée, l'autre pouvant être exprimée. Actuellement, plus d'une centaine de gènes soumis à l'empreinte parentale ont été identifiés chez l'homme et la souris (Hirasawa, Feil 2010). Ceux-ci sont rarement isolés sur le génome, 80% d'entre eux étant organisés en clusters de gènes, et la plupart jouent un rôle dans la croissance fœtale et le développement (Reik, Walter 2001). La méthylation de sites particuliers de l'ADN joue un rôle majeur dans la détermination des copies à activer (Morison, Ramsay, Spencer 2005). Ces mécanismes semblent être héritables d'une génération à l'autre (Kaminsky et al. 2009) et il a été démontré que des profils de méthylation déréglés conduisent à des troubles sévères du développement (Oberle et al. 1991; Nan et al. 1998) ; il semble également clair que ce

processus est impliqué dans certains cancers (Jones, Baylin 2002; Suter, Martin, Ward 2004). Malgré une héritabilité évidente, la méthylation de l'ADN est très instable et dépend hautement des stimuli environnementaux et de l'histoire individuelle (contrairement aux mutations altérant la séquence d'ADN) (Mehler 2008). Ainsi par exemple, il a été montré que des jumeaux monozygotes ayant grandi dans des environnements divers et se nourrissant de manières différentes peuvent présenter des différences biologiques et physiques (Fraga et al. 2005). Les variations épigénétiques peuvent donc jouer un rôle tout aussi essentiel que les modifications de la séquence d'ADN. Il pourrait donc être très intéressant de les étudier dans le but de mieux comprendre certaines différences phénotypiques observées dans les populations humaines ; en particulier dans le cas des maladies infectieuses ou reliées au systeme immunitaire. Le développement actuel des techniques à haut-débit utilisées pour détecter des profils de méthylation à l'échelle du génome devrait aider à répondre à ces questions (Beck, Rakyan 2008).

D'autre part, les gènes peuvent être régulés par de petits ARN simple-brin non codants, les microARN. Ceux-ci peuvent inactiver un gène en interagissant directement avec l'ARNm correspondant. Ils se fixent en effet par complémentarité de séquence à la partie 3' (non-traduite) des ARNm, ce qui provoque le clivage de l'ARNm cible ou le blocage de sa traduction (Lagos-Quintana et al. 2001; Lim et al. 2003; Ambros 2004; Bartel 2004). Ils peuvent jouer un rôle dans la croissance, la différenciation et la prolifération cellulaire, ou encore l'immunité innée (Ambros 2004; Plasterk 2006). Il a été montré récemment que les microARN pourraient également agir en tant qu'oncogènes ou suppresseurs de tumeurs (Zhang et al. 2007a). Il a d'ailleurs été montré que les régions contenant des microARN se trouvaient sous sélection purificatrice, soulignant l'importance de leur rôle et leur implication probable dans des maladies sévères (Quach et al. 2009).

V) Conclusion générale

Les analyses de génétique des populations viennent compléter les approches cliniques et épidémiologiques dans la recherche de gènes et de mutations impliqués dans des pathologies mendéliennes ou complexes. En prenant en compte les mutations observées dans les séquences nucléiques ainsi que les modifications épigénétiques, on pourra mieux comprendre une grande partie de nos différences phénotypiques. On pourrait notamment élucider les mécanismes responsables de disparités actuelles en matière de résistance à certaines maladies infectieuses, et plus généralement à certaines pathologies encore inexpliquées ; et ainsi, cela pourrait contribuer à terme à accélérer le développement de nouvelles thérapies. De manière générale, les résultats obtenus à partir de ce type d'approche pourraient aider à améliorer notre compréhension des mécanismes évolutifs qui gouvernent la diversité humaine, ainsi que celle des autres espèces vivantes.

BIBLIOGRAPHIE

Abadie, V, LM Sollid, LB Barreiro, B Jabri. 2011. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu Rev Immunol 29:493-525.

Aganna, E, F Martinon, PN Hawkins, et al. 2002. Association of mutations in the NALP3/CIAS1/PYPAF1 gene with a broad phenotype including recurrent fever, cold sensitivity, sensorineural deafness, and AA amyloidosis. Arthritis Rheum 46:2445-2452.

Akey, JM, MA Eberle, MJ Rieder, CS Carlson, MD Shriver, DA Nickerson, L Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286.

Akira, S, K Takeda. 2004. Toll-like receptor signalling. Nat Rev Immunol 4:499-511.

Akira, S, S Uematsu, O Takeuchi. 2006. Pathogen recognition and innate immunity. Cell 124:783-801.

Aksentijevich, I, M Nowak, M Mallah, et al. 2002. De novo CIAS1 mutations, cytokine activation, and evidence for genetic heterogeneity in patients with neonatal-onset multisystem inflammatory disease (NOMID): a new member of the expanding family of pyrin-associated autoinflammatory diseases. Arthritis Rheum 46:3340-3348.

Alkhatib, G, C Combadiere, CC Broder, Y Feng, PE Kennedy, PM Murphy, EA Berger. 1996. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science 272:1955-1958.

Allen, IC, MA Scull, CB Moore, EK Holl, E McElvania-TeKippe, DJ Taxman, EH Guthrie, RJ Pickles, JP Ting. 2009. The NLRP3 inflammasome mediates in vivo innate immunity to influenza A virus through recognition of viral RNA. Immunity 30:556-565.

Allison, AC. 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J 1:290-294.

Allison, AC. 1961. Genetic factors in resistance to malaria. Ann N Y Acad Sci 91:710-729.

Allison, AC. 2004. Two lessons from the interface of genetics and medicine. Genetics 166:1591-1599.

Altenburg, E. 1928. The limit of radiation frequency effective in producing mutations. The American Naturalist, 62, 540-545.

Alvarez, CP, F Lasala, J Carrillo, O Muniz, AL Corbi, R Delgado. 2002. C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in cis and in trans. J Virol 76:6841-6844.

Ambros, V. 2004. The functions of animal microRNAs. Nature 431:350-355.

Amos, W, JI Hoffman. 2010. Evidence that two main bottleneck events shaped modern human genetic diversity. Proc Biol Sci 277:131-137.

Anderson, KV, G Jurgens, C Nusslein-Volhard. 1985. Establishment of dorsal-ventral polarity in the Drosophila embryo: genetic studies on the role of the Toll gene product. Cell 42:779-789.

Appelmelk, BJ, I van Die, SJ van Vliet, CM Vandenbroucke-Grauls, TB Geijtenbeek, Y van Kooyk. 2003. Cutting edge: carbohydrate profiling identifies new pathogens that interact with dendritic cell-specific ICAM-3-grabbing nonintegrin on dendritic cells. J Immunol 170:1635-1639.

Auerbach, C, JM Robson, JG Carr. 1947. The Chemical Production of Mutations. Science 105:243-247.

Bamshad, M, SP Wooding. 2003. Signatures of natural selection in the human genome. Nat Rev Genet 4:99-111.

Bamshad, MJ, S Mummidi, E Gonzalez, et al. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A 99:10539-10544.

Barreiro, LB, M Ben-Ali, H Quach, et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet 5:e1000562.
Barreiro, LB, G Laval, H Quach, E Patin, L Quintana-Murci. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet 40:340-345.
Barreiro, LB, E Patin, O Neyrolles, HM Cann, B Gicquel, L Quintana-Murci. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am J Hum Genet 77:869-886.
Barreiro, LB, H Quach, J Krahenbuhl, S Khaliq, A Mohyuddin, SQ Mehdi, B Gicquel, O Neyrolles, L Quintana-Murci. 2006. DC-SIGN interacts with Mycobacterium leprae but sequence variation in this lectin is not associated with leprosy in the Pakistani population. Hum Immunol 67:102-107.
Barreiro, LB, L Quintana-Murci. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17-30.
Barrett, JC, DG Clayton, P Concannon, et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703-707.
Bartel, DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281-297.
Baum, A, A Garcia-Sastre. 2010. Induction of type I interferon by RNA viruses: cellular receptors and their substrates. Amino Acids 38:1283-1299.
Beall, CM, GL Cavalleri, L Deng, et al. 2010. Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107:11459-11464.
Beck, S, VK Rakyan. 2008. The methylome: approaches for global DNA methylation profiling. Trends Genet 24:231-237.
Bent, AF, D Mackey. 2007. Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. Annu Rev Phytopathol 45:399-436.
Bergman, MP, A Engering, HH Smits, SJ van Vliet, AA van Bodegraven, HP Wirth, ML Kapsenberg, CM Vandenbroucke-Grauls, Y van Kooyk, BJ Appelmelk. 2004. Helicobacter pylori modulates the T helper cell 1/T helper cell 2 balance through phase-variable interaction between lipopolysaccharide and DC-SIGN. J Exp Med 200:979-990.
Beutler, E. 1994. G6PD deficiency. Blood 84:3613-3636.
Bochud, PY, M Bochud, A Telenti, T Calandra. 2007. Innate immunogenetics: a tool for exploring new frontiers of host defence. Lancet Infect Dis 7:531-542.
Bodmer, WF, LL Cavalli-Sforza. 1971. Fear of enlightenment. Nature 229:71-72; discussion 72.
Boniotto, M, L Tailleux, M Lomma, B Gicquel, C Buchrieser, S Garcia, L Quintana-Murci. 2011. Population variation in NAIP functional copy number confers increased cell death upon /Legionella pneumophila/ infection. Hum Immunol (in press).
Bortoluci, KR, R Medzhitov. 2010. Control of infection by pyroptosis and autophagy: role of TLR and NLR. Cell Mol Life Sci 67:1643-1651.
Boyden, ED, WF Dietrich. 2006. Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin. Nat Genet 38:240-244.
Brunet, M, F Guy, D Pilbeam, et al. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418:145-151.
Bustamante, CD, A Fledel-Alon, S Williamson, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157.
Bustamante, CD, R Nielsen, SA Sawyer, KM Olsen, MD Purugganan, DL Hartl. 2002. The cost of inbreeding in Arabidopsis. Nature 416:531-534.

Cann, HM, C de Toma, L Cazes, et al. 2002. A human genome diversity cell line panel. Science 296:261-262.
Caplan, J, M Padmanabhan, SP Dinesh-Kumar. 2008. Plant NB-LRR immune receptors: from recognition to transcriptional reprogramming. Cell Host Microbe 3:126-135.
Carrington, M, T Kissner, B Gerrard, S Ivanov, SJ O'Brien, M Dean. 1997. Novel alleles of the chemokine-receptor gene CCR5. Am J Hum Genet 61:1261-1267.
Casanova, JL, L Abel. 2005. Inborn errors of immunity to infection: the rule rather than the exception. J Exp Med 202:197-201.
Casanova, JL, L Abel, L Quintana-Murci. 2011. Human TLRs and IL-1Rs in host defense: natural insights from evolutionary, epidemiological, and clinical genetics. Annu Rev Immunol 29:447-491.
Cavalli-Sforza, LL. 1966. Population structure and human evolution. Proc R Soc Lond B Biol Sci 164:362-379.
Cavalli-Sforza, LL, MW Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl:266-275.
Cavazzana-Calvo, M, E Payen, O Negre, et al. 2010. Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. Nature 467:318-322.
Chaix, R, F Austerlitz, T Khegay, S Jacquesson, MF Hammer, E Heyer, L Quintana-Murci. 2004. The genetic or mythical ancestry of descent groups: lessons from the Y chromosome. Am J Hum Genet 75:1113-1116.
Chaix, R, L Quintana-Murci, T Hegay, MF Hammer, Z Mobasher, F Austerlitz, E Heyer. 2007. From social to genetic structures in central Asia. Curr Biol 17:43-48.
Chamaillard, M, M Hashimoto, Y Horie, et al. 2003. An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. Nat Immunol 4:702-707.
Chen, YS, A Torroni, L Excoffier, AS Santachiara-Benerecetti, DC Wallace. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133-149.
Chitnis, CE, LH Miller. 1994. Identification of the erythrocyte binding domains of Plasmodium vivax and Plasmodium knowlesi proteins involved in erythrocyte invasion. J Exp Med 180:497-506.
Chow, JC, Z Yen, SM Ziesche, CJ Brown. 2005. Silencing of the mammalian X chromosome. Annu Rev Genomics Hum Genet 6:69-92.
Chuang, T, RJ Ulevitch. 2001. Identification of hTLR10: a novel human Toll-like receptor preferentially expressed in immune cells. Biochim Biophys Acta 1518:157-161.
Chuang, TH, RJ Ulevitch. 2000. Cloning and characterization of a sub-family of human toll-like receptors: hTLR7, hTLR8 and hTLR9. Eur Cytokine Netw 11:372-378.
Clark, AG, S Glanowski, R Nielsen, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963.
Clark, AM, L Hartling, B Vandermeer, SL Lissel, FA McAlister. 2007. Secondary prevention programmes for coronary heart disease: a meta-regression showing the merits of shorter, generalist, primary care-based interventions. Eur J Cardiovasc Prev Rehabil 14:538-546.
Clark, SJ, J Harrison, M Frommer. 1995. CpNpG methylation in mammalian cells. Nat Genet 10:20-27.
Colmenares, M, A Puig-Kroger, OM Pello, AL Corbi, L Rivas. 2002. Dendritic cell (DC)-specific intercellular adhesion molecule 3 (ICAM-3)-grabbing nonintegrin (DC-SIGN, CD209), a C-type surface lectin in human DCs, is a receptor for Leishmania amastigotes. J Biol Chem 277:36766-36769.
Consortium, TIH. 2003. The International HapMap Project. Nature 426:789-796.

Consortium, TWTCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.
Cooke, GS, AV Hill. 2001. Genetics of susceptibility to human infectious disease. Nat Rev Genet 2:967-977.
Cooney, R, J Baker, O Brain, B Danis, T Pichulik, P Allan, DJ Ferguson, BJ Campbell, D Jewell, A Simmons. 2010. NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. Nat Med 16:90-97.
Cooper, GM, DA Nickerson, EE Eichler. 2007. Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39:S22-29.
Creighton, HB, B McClintock. 1931. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. Proc Natl Acad Sci U S A 17:492-497.
Davoodi, J, MH Ghahremani, A Es-Haghi, A Mohammad-Gholi, A Mackenzie. 2010. Neuronal apoptosis inhibitory protein, NAIP, is an inhibitor of procaspase-9. Int J Biochem Cell Biol 42:958-964.
Dean, M, M Carrington, C Winkler, et al. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science 273:1856-1862.
Desreumaux, P. 2005. [NOD2/CARD15 and Crohn's disease]. Gastroenterol Clin Biol 29:696-700.
Dittmar, K, D Liberles. 2010. *Evolution after gene duplication*. Wiley.
Dode, C, N Le Du, L Cuisset, et al. 2002. New mutations of CIAS1 that are responsible for Muckle-Wells syndrome and familial cold urticaria: a novel mutation underlies both syndromes. Am J Hum Genet 70:1498-1506.
Drake, JA, C Bird, J Nemesh, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223-227.
Du, X, A Poltorak, Y Wei, B Beutler. 2000. Three novel mammalian toll-like receptors: gene structure, expression, and evolution. Eur Cytokine Netw 11:362-371.
Dubuquoy, L, C Rousseaux, X Thuru, L Peyrin-Biroulet, O Romano, P Chavatte, M Chamaillard, P Desreumaux. 2006. PPARgamma as a new therapeutic target in inflammatory bowel diseases. Gut 55:1341-1349.
Durbin, R. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.
Ehrlich, M, MA Gama-Sosa, LH Huang, RM Midgett, KC Kuo, RA McCune, C Gehrke. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res 10:2709-2721.
Ellis, J, P Dodds, T Pryor. 2000. The generation of plant disease resistance gene specificities. Trends Plant Sci 5:373-379.
Enattah, NS, TG Jensen, M Nielsen, et al. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am J Hum Genet 82:57-72.
Endo, T, K Ikeo, T Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685-690.
Excoffier, L, PE Smouse, JM Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.
Fagundes, NJ, N Ray, M Beaumont, S Neuenschwander, FM Salzano, SL Bonatto, L Excoffier. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A 104:17614-17619.

Faustin, B, L Lartigue, JM Bruey, F Luciano, E Sergienko, B Bailly-Maitre, N Volkmann, D Hanein, I Rouiller, JC Reed. 2007. Reconstituted NALP1 inflammasome reveals two-step mechanism of caspase-1 activation. Mol Cell 25:713-724.

Fay, JC, CI Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Feldmann, J, AM Prieur, P Quartier, P Berquin, S Certain, E Cortis, D Teillac-Hamel, A Fischer, G de Saint Basile. 2002. Chronic infantile neurological cutaneous and articular syndrome is caused by mutations in CIAS1, a gene highly expressed in polymorphonuclear cells and chondrocytes. Am J Hum Genet 71:198-203.

Figueroa, JDM Garcia-ClosasM Humphreys, et al. 2011. Associations of common variants at 1p11.2 and 14q24.1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: findings from the Breast Cancer Association Consortium. Hum Mol Genet.

Fink, SL, T Bergsbaken, BT Cookson. 2008. Anthrax lethal toxin and Salmonella elicit the common cell death pathway of caspase-1-dependent pyroptosis via distinct mechanisms. Proc Natl Acad Sci U S A 105:4312-4317.

Force, A, M Lynch, FB Pickett, A Amores, YL Yan, J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Fornarino, S, G Laval, LB Barreiro, J Manry, E Vasseur, L Quintana-Murci. 2011. Evolution of the TIR Domain-Containing Adaptors in Humans: Swinging between Constraint and Adaptation. Mol Biol Evol 28:3087-3097.

Forterre, P, S Gribaldo, C Brochier. 2005. [Luca: the last universal common ancestor]. Med Sci (Paris) 21:860-865.

Fraga, MF, E Ballestar, MF Paz, et al. 2005. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A 102:10604-10609.

Frazer, KADG BallingerDR Cox, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Fu, YX, WH Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Galvani, AP, J Novembre. 2005. The evolutionary history of the CCR5-Delta32 HIV-resistance mutation. Microbes Infect 7:302-309.

Gateva, V, JK Sandling, G Hom, et al. 2009. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat Genet 41:1228-1233.

Geijtenbeek, TB, DS Kwon, R Torensma, et al. 2000. DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. Cell 100:587-597.

Geijtenbeek, TB, SJ Van Vliet, EA Koppel, M Sanchez-Hernandez, CM Vandenbroucke-Grauls, B Appelmelk, Y Van Kooyk. 2003. Mycobacteria target DC-SIGN to suppress dendritic cell function. J Exp Med 197:7-17.

Gerard, N, M Caillaud, A Martoriati, G Goudet, AC Lalmanach. 2004. The interleukin-1 system and female reproduction. J Endocrinol 180:203-212.

Girardin, SE, IG Boneca, LA Carneiro, et al. 2003a. Nod1 detects a unique muropeptide from gram-negative bacterial peptidoglycan. Science 300:1584-1587.

Girardin, SE, IG Boneca, J Viala, M Chamaillard, A Labigne, G Thomas, DJ Philpott, PJ Sansonetti. 2003b. Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. J Biol Chem 278:8869-8872.

Girardin, SE, R Tournebize, M Mavris, et al. 2001. CARD4/Nod1 mediates NF-kappaB and JNK activation by invasive Shigella flexneri. EMBO Rep 2:736-742.

Gitlin, L, W Barchet, S Gilfillan, M Cella, B Beutler, RA Flavell, MS Diamond, M Colonna. 2006. Essential role of mda-5 in type I IFN responses to polyriboinosinic:polyribocytidylic acid and encephalomyocarditis picornavirus. Proc Natl Acad Sci U S A 103:8459-8464.

Gonzalez, E, M Bamshad, N Sato, et al. 1999. Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. Proc Natl Acad Sci U S A 96:12004-12009.
Gonzalez, E, R Dhanda, M Bamshad, et al. 2001. Global survey of genetic variation in CCR5, RANTES, and MIP-1alpha: impact on the epidemiology of the HIV-1 pandemic. Proc Natl Acad Sci U S A 98:5199-5204.
Hallum, JV, JS Younger. 1966. Quantitative aspects of inhibition of virus replication by interferon in chick embryo cell cultures. J Bacteriol 92:1047-1050.
Hamblin, MT, A Di Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669-1679.
Hamblin, MT, EE Thompson, A Di Rienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369-383.
Hammer, MF, TM Karafet, AJ Redd, H Jarjanazi, S Santachiara-Benerecetti, H Soodyall, SL Zegura. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18:1189-1203.
Hampe, J, A Cuthbert, PJ Croucher, et al. 2001. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet 357:1925-1928.
Harpending, H, A Rogers. 2000. Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361-385.
Hartl, D, A Clark. 2007. Principles of Population Genetics (Sunderlan: Sinauer Associates, Inc.)
Hashimoto, C, KL Hudson, KV Anderson. 1988. The Toll gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. Cell 52:269-279.
Hawn, TR, EA Misch, SJ Dunstan, et al. 2007. A common human TLR1 polymorphism regulates the innate immune response to lipopeptides. Eur J Immunol 37:2280-2289.
Hedrick, PW, TS Whittam, P Parham. 1991. Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. Proc Natl Acad Sci U S A 88:5897-5901.
Hirasawa, R, R Feil. 2010. Genomic imprinting and human disease. Essays Biochem 48:187-200.
Hoffman, HM, JL Mueller, DH Broide, AA Wanderer, RD Kolodner. 2001. Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. Nat Genet 29:301-305.
Hoffmann, JA. 2003. The immune response of Drosophila. Nature 426:33-38.
Hoffmann, JA, FC Kafatos, CA Janeway, RA Ezekowitz. 1999. Phylogenetic perspectives in innate immunity. Science 284:1313-1318.
Hsu, LC, SR Ali, S McGillivray, et al. 2008. A NOD2-NALP1 complex mediates caspase-1-dependent IL-1beta secretion in response to Bacillus anthracis infection and muramyl dipeptide. Proc Natl Acad Sci U S A 105:7803-7808.
Hsu, YM, Y Zhang, Y You, D Wang, H Li, O Duramad, XF Qin, C Dong, X Lin. 2007. The adaptor protein CARD9 is required for innate immune responses to intracellular pathogens. Nat Immunol 8:198-205.
Hu, J, E Nistal-Villan, A Voho, A Ganee, M Kumar, Y Ding, A Garcia-Sastre, JG Wetmur. 2010. A common polymorphism in the caspase recruitment domain of RIG-I modifies the innate immune response of human dendritic cells. J Immunol 185:424-432.
Hughes, AL, MK Hughes, CY Howell, M Nei. 1994. Natural selection at the class II major histocompatibility complex loci of mammals. Philos Trans R Soc Lond B Biol Sci 346:359-366; discussion 366-357.

Hughes, AL, M Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.
Hughes, AL, M Yeager. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet 32:415-435.
Hugot, JP, M Chamaillard, H Zouali, et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599-603.
Ingman, M, H Kaessmann, S Paabo, U Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.
Inohara, N, T Koseki, J Lin, L del Peso, PC Lucas, FF Chen, Y Ogura, G Nunez. 2000. An induced proximity model for NF-kappa B activation in the Nod1/RICK and RIP signaling pathways. J Biol Chem 275:27823-27831.
Inohara, N, Y Ogura, FF Chen, A Muto, G Nunez. 2001. Human Nod1 confers responsiveness to bacterial lipopolysaccharides. J Biol Chem 276:2551-2554.
Inohara, N, Y Ogura, A Fontalba, et al. 2003. Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. J Biol Chem 278:5509-5512.
Isaacs, A, J Lindenmann. 1957. Virus interference. I. The interferon. Proc R Soc Lond B Biol Sci 147:258-267.
Itan, Y, BL Jones, CJ Ingram, DM Swallow, MG Thomas. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evol Biol 10:36.
Janeway, C, P Travers, M Walport, M Shlomchik. 2001. Immunobiology: The Immune System in Health and Disease, 5th Edition, New York and London: Garland Science.
Janeway, CA, Jr., R Medzhitov. 2002. Innate immune recognition. Annu Rev Immunol 20:197-216.
Jermendy, A, I Szatmari, AP Laine, et al. 2010. The interferon-induced helicase IFIH1 Ala946Thr polymorphism is associated with type 1 diabetes in both the high-incidence Finnish and the medium-incidence Hungarian populations. Diabetologia 53:98-102.
Jin, Y, CM Mailloux, K Gowan, SL Riccardi, G LaBerge, DC Bennett, PR Fain, RA Spritz. 2007. NALP1 in vitiligo-associated multiple autoimmune disease. N Engl J Med 356:1216-1225.
Jobling, MA, C Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598-612.
Johnson, CM, EA Lyle, KO Omueti, VA Stepensky, O Yegin, E Alpsoy, L Hamann, RR Schumann, RI Tapping. 2007. Cutting edge: A common polymorphism impairs cell surface trafficking and functional responses of TLR1 but protects against leprosy. J Immunol 178:7520-7524.
Jones, PA, SB Baylin. 2002. The fundamental role of epigenetic events in cancer. Nat Rev Genet 3:415-428.
Kaminsky, ZA, T Tang, SC Wang, et al. 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41:240-245.
Kang, DC, RV Gopalkrishnan, Q Wu, E Jankowsky, AM Pyle, PB Fisher. 2002. mda-5: An interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. Proc Natl Acad Sci U S A 99:637-642.
Kanneganti, TD, M Body-Malapel, A Amer, et al. 2006a. Critical role for Cryopyrin/Nalp3 in activation of caspase-1 in response to viral infection and double-stranded RNA. J Biol Chem 281:36560-36568.
Kanneganti, TD, N Ozoren, M Body-Malapel, et al. 2006b. Bacterial RNA and small antiviral compounds activate caspase-1 through cryopyrin/Nalp3. Nature 440:233-236.

Kato, H, S Sato, M Yoneyama, et al. 2005. Cell type-specific involvement of RIG-I in antiviral response. Immunity 23:19-28.
Kato, H, O Takeuchi, E Mikamo-Satoh, R Hirai, T Kawai, K Matsushita, A Hiiragi, TS Dermody, T Fujita, S Akira. 2008. Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid-inducible gene-I and melanoma differentiation-associated gene 5. J Exp Med 205:1601-1610.
Kato, H, O Takeuchi, S Sato, et al. 2006. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. Nature 441:101-105.
Kawai, T, S Akira. 2006. Innate immune recognition of viral infection. Nat Immunol 7:131-137.
Kawai, T, K Takahashi, S Sato, C Coban, H Kumar, H Kato, KJ Ishii, O Takeuchi, S Akira. 2005. IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. Nat Immunol 6:981-988.
Kempster, SL, G Belteki, AJ Forhead, AL Fowden, RD Catalano, BY Lam, I McFarlane, DS Charnock-Jones, GC Smith. 2010. Developmental control of the Nlrp6 inflammasome and a substrate, IL-18, in mammalian intestine. Am J Physiol Gastrointest Liver Physiol 300:G253-263.
Kidd, JM, GM Cooper, WF Donahue, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56-64.
Kim, JG, SJ Lee, MF Kagnoff. 2004. Nod1 is an essential signal transducer in intestinal epithelial cells infected with bacteria that avoid recognition by toll-like receptors. Infect Immun 72:1487-1495.
Kimbrell, DA, B Beutler. 2001. The evolution and genetics of innate immunity. Nat Rev Genet 2:256-267.
Kimura, M. 1968a. Evolutionary rate at the molecular level. Nature 217:624-626.
Kimura, M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res 11:247-269.
Kimura, M, GH Weiss. 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics 49:561-576.
Klose, RJ, AP Bird. 2006. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31:89-97.
Kobayashi, K, N Inohara, LD Hernandez, JE Galan, G Nunez, CA Janeway, R Medzhitov, RA Flavell. 2002. RICK/Rip2/CARDIAK mediates signalling for receptors of the innate and adaptive immune systems. Nature 416:194-199.
Kobayashi, KS, M Chamaillard, Y Ogura, O Henegariu, N Inohara, G Nunez, RA Flavell. 2005. Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. Science 307:731-734.
Kobe, B, J Deisenhofer. 1995. Proteins with leucine-rich repeats. Curr Opin Struct Biol 5:409-416.
Kovacsovics, M, F Martinon, O Micheau, JL Bodmer, K Hofmann, J Tschopp. 2002. Overexpression of Helicard, a CARD-containing helicase cleaved during apoptosis, accelerates DNA degradation. Curr Biol 12:838-843.
Kufer, TA, PJ Sansonetti. 2010. NLR functions beyond pathogen recognition. Nat Immunol 12:121-128.
Kummer, JA, R Broekhuizen, H Everett, L Agostini, L Kuijk, F Martinon, R van Bruggen, J Tschopp. 2007. Inflammasome components NALP 1 and 3 show distinct but separate expression profiles in human tissues suggesting a site-specific role in the inflammatory response. J Histochem Cytochem 55:443-452.
Kwiatkowski, DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77:171-192.

Lagos-Quintana, M, R Rauhut, W Lendeckel, T Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. Science 294:853-858.
Lalani, AS, J Masters, W Zeng, J Barrett, R Pannu, H Everett, CW Arendt, G McFadden. 1999. Use of chemokine receptors by poxviruses. Science 286:1968-1971.
Lander, ESLM LintonB Birren, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.
Lange, C, G Hemmrich, UC Klostermeier, JA Lopez-Quintero, DJ Miller, T Rahn, Y Weiss, TC Bosch, P Rosenstiel. 2011. Defining the origins of the NOD-like receptor system at the base of animal evolution. Mol Biol Evol 28:1687-1702.
Laval, G, E Patin, LB Barreiro, L Quintana-Murci. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5:e10284.
Lebatard, AE, DL Bourles, P Duringer, et al. 2008. Cosmogenic nuclide dating of Sahelanthropus tchadensis and Australopithecus bahrelghazali: Mio-Pliocene hominids from Chad. Proc Natl Acad Sci U S A 105:3226-3231.
Lee, MS, YJ Kim. 2007. Signaling pathways downstream of pattern-recognition receptors and their cross talk. Annu Rev Biochem 76:447-480.
Lemaitre, B, J Hoffmann. 2007. The host defense of Drosophila melanogaster. Annu Rev Immunol 25:697-743.
Lemaitre, B, E Nicolas, L Michaut, JM Reichhart, JA Hoffmann. 1996. The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. Cell 86:973-983.
Lesage, S, H Zouali, JP Cezard, et al. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet 70:845-857.
Lewin, R. 1987. Africa: cradle of modern humans. Science 237:1292-1295.
Lewontin, RC, J Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195.
Li, JZ, DM Absher, H Tang, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.
Li, WH, LA Sadler. 1991. Low nucleotide diversity in man. Genetics 129:513-523.
Lim, LP, ME Glasner, S Yekta, CB Burge, DP Bartel. 2003. Vertebrate microRNA genes. Science 299:1540.
Lindsay, SJ, M Khajavi, JR Lupski, ME Hurles. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. Am J Hum Genet 79:890-902.
Liston, P, N Roy, K Tamai, et al. 1996. Suppression of apoptosis in mammalian cells by NAIP and a related family of IAP genes. Nature 379:349-353.
Litman, GW, JP Cannon, LJ Dishaw. 2005. Reconstructing immune phylogeny: new perspectives. Nat Rev Immunol 5:866-879.
Liu, P, M Jamaluddin, K Li, RP Garofalo, A Casola, AR Brasier. 2007. Retinoic acid-inducible gene I mediates early antiviral response and Toll-like receptor 3 expression in respiratory syncytial virus-infected airway epithelial cells. J Virol 81:1401-1411.
Loo, YM, J Fornek, N Crochet, et al. 2008. Distinct RIG-I and MDA5 signaling by RNA viruses in innate immunity. J Virol 82:335-345.
Loo, YM, M Gale, Jr. 2011. Immune signaling by RIG-I-like receptors. Immunity 34:680-692.
Louicharoen, C, E Patin, R Paul, et al. 2009. Positively selected G6PD-Mahidol mutation reduces Plasmodium vivax density in Southeast Asians. Science 326:1546-1549.

Lozach, PY, H Lortat-Jacob, A de Lacroix de Lavalette, et al. 2003. DC-SIGN and L-SIGN are high affinity binding receptors for hepatitis C virus glycoprotein E2. J Biol Chem 278:20358-20366.

Ma, X, Y Liu, BB Gowen, EA Graviss, AG Clark, JM Musser. 2007. Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PLoS One 2:e1318.

Macaulay, V, C Hill, A Achilli, et al. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034-1036.

Magalhaes, JG, DJ Philpott, MA Nahori, et al. 2005. Murine Nod1 but not its human orthologue mediates innate immune detection of tracheal cytotoxin. EMBO Rep 6:1201-1207.

Magitta, NF, AS Boe Wolff, S Johansson, et al. 2009. A coding polymorphism in NALP1 confers risk for autoimmune Addison's disease and type 1 diabetes. Genes Immun 10:120-124.

Manry, J, G Laval, E Patin, et al. 2011b. Evolutionary genetic dissection of human interferons. J Exp Med (in press).

Manry, J, G Laval, E Patin, S Fornarino, M Tichit, C Bouchier, LB Barreiro, L Quintana-Murci. 2011a. Evolutionary genetics evidence of an essential, nonredundant role of the IFN-gamma pathway in protective immunity. Hum Mutat 32:633-642.

Mariathasan, S, DS Weiss, K Newton, J McBride, K O'Rourke, M Roose-Girma, WP Lee, Y Weinrauch, DM Monack, VM Dixit. 2006. Cryopyrin activates the inflammasome in response to toxins and ATP. Nature 440:228-232.

Martinon, F, L Agostini, E Meylan, J Tschopp. 2004. Identification of bacterial muramyl dipeptide as activator of the NALP3/cryopyrin inflammasome. Curr Biol 14:1929-1934.

Martinon, F, K Burns, J Tschopp. 2002. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. Mol Cell 10:417-426.

Martinon, F, A Mayor, J Tschopp. 2009. The inflammasomes: guardians of the body. Annu Rev Immunol 27:229-265.

Martinon, F, V Petrilli, A Mayor, A Tardivel, J Tschopp. 2006. Gout-associated uric acid crystals activate the NALP3 inflammasome. Nature 440:237-241.

Matzinger, P. 1994. Tolerance, danger, and the extended family. Annu Rev Immunol 12:991-1045.

Matzinger, P. 2002. The danger model: a renewed sense of self. Science 296:301-305.

McDonald, JH, M Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McDougall, I, FH Brown, JG Fleagle. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733-736.

McVean, GA, SR Myers, S Hunt, P Deloukas, DR Bentley, P Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584.

Medzhitov, R. 2001. Toll-like receptors and innate immunity. Nat Rev Immunol 1:135-145.

Medzhitov, R, CA Janeway, Jr. 1998. Innate immune recognition and control of adaptive immune responses. Semin Immunol 10:351-353.

Medzhitov, R, CA Janeway, Jr. 2002. Decoding the patterns of self and nonself by the innate immune system. Science 296:298-300.

Medzhitov, R, P Preston-Hurlburt, CA Janeway, Jr. 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388:394-397.

Mehler, MF. 2008. Epigenetic principles and mechanisms underlying nervous system functions in health and disease. Prog Neurobiol 86:305-341.

Melchjorsen, J, SB Jensen, L Malmgaard, SB Rasmussen, F Weber, AG Bowie, S Matikainen, SR Paludan. 2005. Activation of innate defense against a paramyxovirus is mediated by RIG-I and TLR7 and TLR8 in a cell-type-specific manner. J Virol 79:12944-12951.

Mellars, P. 2006. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. Nature 439:931-935.

Meyers, BC, KA Shen, P Rohani, BS Gaut, RW Michelmore. 1998. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell 10:1833-1846.

Meylan, E, J Curran, K Hofmann, D Moradpour, M Binder, R Bartenschlager, J Tschopp. 2005. Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. Nature 437:1167-1172.

Meylan, E, J Tschopp, M Karin. 2006. Intracellular pattern recognition receptors in the host response. Nature 442:39-44.

Mikkelsen, SS, SB Jensen, S Chiliveru, J Melchjorsen, I Julkunen, M Gaestel, JS Arthur, RA Flavell, S Ghosh, SR Paludan. 2009. RIG-I-mediated activation of p38 MAPK is essential for viral induction of interferon and activation of dendritic cells: dependence on TRAF2 and TAK1. J Biol Chem 284:10774-10782.

Misch, EA, M Macdonald, C Ranjit, BR Sapkota, RD Wells, MR Siddiqui, G Kaplan, TR Hawn. 2008. Human TLR1 deficiency is associated with impaired mycobacterial signaling and protection from leprosy reversal reaction. PLoS Negl Trop Dis 2:e231.

Mishmar, D, E Ruiz-Pesini, P Golik, et al. 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100:171-176.

Mondragon-Palomino, M, BC Meyers, RW Michelmore, BS Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res 12:1305-1315.

Moore, CB, DT Bergstralh, JA Duncan, et al. 2008. NLRX1 is a regulator of mitochondrial antiviral immunity. Nature 451:573-577.

Morison, IM, JP Ramsay, HG Spencer. 2005. A census of mammalian imprinting. Trends Genet 21:457-465.

Murdoch, S, U Djuric, B Mazhar, et al. 2006. Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. Nat Genet 38:300-302.

Myers, S, L Bottolo, C Freeman, G McVean, P Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321-324.

Nallagatla, SR, R Toroney, PC Bevilacqua. 2008. A brilliant disguise for self RNA: 5'-end and internal modifications of primary transcripts suppress elements of innate immunity. RNA Biol 5:140-144.

Nan, X, HH Ng, CA Johnson, CD Laherty, BM Turner, RN Eisenman, A Bird. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393:386-389.

Neerincx, A, K Lautz, M Menning, E Kremmer, P Zigrino, M Hosel, H Buning, R Schwarzenbacher, TA Kufer. 2010. A role for the human nucleotide-binding domain, leucine-rich repeat-containing family member NLRC5 in antiviral responses. J Biol Chem 285:26223-26232.

Nei, M. 1987. Molecular evolutionary genetics.

Nejentsev, S, N Walker, D Riches, M Egholm, JA Todd. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387-389.

Nicolae, DL, E Gamazon, W Zhang, S Duan, ME Dolan, NJ Cox. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6:e1000888.

Nielsen, R. 2005. Molecular signatures of natural selection. Annu Rev Genet 39:197-218.
Nielsen, R, I Hellmann, M Hubisz, C Bustamante, AG Clark. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet 8:857-868.
Noel, L, TL Moores, EA van Der Biezen, M Parniske, MJ Daniels, JE Parker, JD Jones. 1999. Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell 11:2099-2112.
Novembre, J, A DI Rienzo. 2009. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet 10(11):745-55.
Oberle, I, F Rousseau, D Heitz, C Kretz, D Devys, A Hanauer, J Boue, MF Bertheas, JL Mandel. 1991. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science 252:1097-1102.
Ogura, Y, DK Bonen, N Inohara, et al. 2001b. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603-606.
Ogura, Y, N Inohara, A Benito, FF Chen, S Yamaoka, G Nunez. 2001a. Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB. J Biol Chem 276:4812-4818.
Ohno, S. 1970. *Evolution by gene duplication*. Springer, New York.
Ohta, T. 1991. Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. Proc Natl Acad Sci U S A 88:6716-6720.
Pakendorf, B, M Stoneking. 2005. Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6:165-183.
Parniske, M, KE Hammond-Kosack, C Golstein, CM Thomas, DA Jones, K Harrison, BB Wulff, JD Jones. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell 91:821-832.
Perez-Lezaun, A, F Calafell, D Comas, et al. 1999. Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208-219.
Petrilli, V, S Papin, C Dostert, A Mayor, F Martinon, J Tschopp. 2007. Activation of the NALP3 inflammasome is triggered by low intracellular potassium concentration. Cell Death Differ 14:1583-1589.
Pichlmair, A, O Schulz, CP Tan, TI Naslund, P Liljestrom, F Weber, C Reis e Sousa. 2006. RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. Science 314:997-1001.
Plasterk, RH. 2006. Micro RNAs in animal development. Cell 124:877-881.
Poland, GA, IG Ovsyannikova, RM Jacobson, DI Smith. 2007. Heterogeneity in vaccine immune response: the role of immunogenetics and the emerging field of vaccinomics. Clin Pharmacol Ther 82:653-664.
Pothlichet, J, A Burtey, AV Kubarenko, et al. 2009. Study of human RIG-I polymorphisms identifies two variants with an opposite impact on the antiviral immune response. PLoS One 4:e7582.
Prugnolle, F, A Manica, M Charpentier, JF Guegan, V Guernier, F Balloux. 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15:1022-1027.
Przeworski, M. 2005. Genetics. Motivating hotspots. Science 310:247-248.
Qian, J, C Deveault, R Bagga, X Xie, R Slim. 2007. Women heterozygous for NALP7/NLRP7 mutations are at risk for reproductive wastage: report of two novel mutations. Hum Mutat 28:741.
Quach, H, LB Barreiro, G Laval, et al. 2009. Signatures of purifying and local positive selection in human miRNAs. Am J Hum Genet 84:316-327.

Quintana-Murci, L, A Alcais, L Abel, JL Casanova. 2007. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. Nat Immunol 8:1165-1171.

Quintana-Murci, L, O Semino, HJ Bandelt, G Passarino, K McElreavey, AS Santachiara-Benerecetti. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23:437-441.

Rakoff-Nahoum, S, J Paglino, F Eslami-Varzaneh, S Edberg, R Medzhitov. 2004. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell 118:229-241.

Rassoulzadegan, M, V Grandjean, P Gounon, S Vincent, I Gillot, F Cuzin. 2006. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature 441:469-474.

Rast, JP, LC Smith, M Loza-Coll, T Hibino, GW Litman. 2006. Genomic insights into the immune system of the sea urchin. Science 314:952-956.

Reddy, MV, H Wang, S Liu, B Bode, JC Reed, RD Steed, SW Anderson, L Steed, D Hopkins, JX She. 2011. Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. Genes Immun 12:208-212.

Redon, R, S Ishikawa, KR Fitch, et al. 2006. Global variation in copy number in the human genome. Nature 444:444-454.

Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447:425-432.

Reik, W, J Walter. 2001. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2:21-32.

Repik, P, A Flamand, DH Bishop. 1974. Effect of interferon upon the primary and secondary transcription of vesicular stomatitis and influenza viruses. J Virol 14:1169-1178.

Rich, T. 2004. Toll and Toll-like receptors: an immunologic perspective. University of Cambridge.

Riggs, AD. 1975. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14:9-25.

Ripke, SAR SandersKS Kendler, et al. 2011. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43:969-976.

Rock, FL, G Hardiman, JC Timans, RA Kastelein, JF Bazan. 1998. A family of human receptors structurally related to Drosophila Toll. Proc Natl Acad Sci U S A 95:588-593.

Rose, NR, C Bona. 1993. Defining criteria for autoimmune diseases (Witebsky's postulates revisited). Immunol Today 14:426-430.

Sabeti, PC, DE Reich, JM Higgins, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sabeti, PC, SF Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, TS Mikkelsen, D Altshuler, ES Lander. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sabeti, PCP VarillyB Fry, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Sabeti, PC, E Walsh, SF Schaffner, et al. 2005. The case for selection at CCR5-Delta32. PLoS Biol 3:e378.

Saito, T, R Hirai, YM Loo, D Owen, CL Johnson, SC Sinha, S Akira, T Fujita, M Gale, Jr. 2007. Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. Proc Natl Acad Sci U S A 104:582-587.

Salem, AH, FM Badr, MF Gaballah, S Paabo. 1996. The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. Am J Hum Genet 59:741-743.

Salvioli, S, M Capri, S Valensin, P Tieri, D Monti, E Ottaviani, C Franceschi. 2006. Inflamm-aging, cytokines and aging: state of the art, new hypotheses on the role of mitochondria and new perspectives from systems biology. Curr Pharm Des 12:3161-3171.

Santos-Lopes, SS, RW Pereira, IJ Wilson, SD Pena. 2007. A worldwide phylogeography for the human X chromosome. PLoS One 2:e557.

Saunders, MA, MF Hammer, MW Nachman. 2002. Nucleotide variability at G6pd and the signature of malarial selection in humans. Genetics 162:1849-1861.

Sawyer, SA, DL Hartl. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161-1176.

Schlee, M, A Roth, V Hornung, et al. 2009. Recognition of 5' triphosphate by RIG-I helicase requires short blunt double-stranded RNA as contained in panhandle of negative-strand virus. Immunity 31:25-34.

Schroder, K, J Tschopp. 2010. The inflammasomes. Cell 140:821-832.

Sebat, J, B Lakshmi, J Troge, et al. 2004. Large-scale copy number polymorphism in the human genome. Science 305:525-528.

Seielstad, MT, E Minch, LL Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278-280.

Seth, RB, L Sun, CK Ea, ZJ Chen. 2005. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. Cell 122:669-682.

Sherry, ST, MH Ward, M Kholodov, J Baker, L Phan, EM Smigielski, K Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308-311.

Shi, Y, JE Evans, KL Rock. 2003. Molecular identification of a danger signal that alerts the immune system to dying cells. Nature 425:516-521.

Shigemoto, T, M Kageyama, R Hirai, J Zheng, M Yoneyama, T Fujita. 2009. Identification of loss of function mutations in human genes encoding RIG-I and MDA5: implications for resistance to type I diabetes. J Biol Chem 284:13348-13354.

Sironi, M, M Clerici. 2010. The hygiene hypothesis: an evolutionary perspective. Microbes Infect 12:421-427.

Smyth, DJ, JD Cooper, R Bailey, et al. 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38:617-619.

Srinivasula, SM, JL Poyet, M Razmara, P Datta, Z Zhang, ES Alnemri. 2002. The PYRIN-CARD protein ASC is an activating adaptor for caspase-1. J Biol Chem 277:21119-21122.

Steimle, V, LA Otten, M Zufferey, B Mach. 1993. Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). Cell 75:135-146.

Steiner, JW, R Volpe. 1961. Autoimmunization - A Possible Mechanism of Tissue Injury I-Theoretical and Experimental Aspects. Can Med Assoc J 84:1165-1172.

Stephens, JC, DE Reich, DB Goldstein, et al. 1998. Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507-1515.

Stephens, JC, JA Schneider, DA Tanguay, et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-493.

Strange, A, F Capon, CC Spencer, et al. 2010. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet 42:985-990.
Suter, CM, DI Martin, RL Ward. 2004. Germline epimutation of MLH1 in individuals with multiple cancers. Nat Genet 36:497-501.
Sutherland, A, J Davies, CJ Owen, et al. 2007. Genomic polymorphism at the interferon-induced helicase (IFIH1) locus contributes to Graves' disease susceptibility. J Clin Endocrinol Metab 92:3338-3341.
Swanberg, M, O Lidman, L Padyukov, et al. 2005. MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nat Genet 37:486-494.
Tailleux, L, O Schwartz, JL Herrmann, et al. 2003. DC-SIGN is the major Mycobacterium tuberculosis receptor on human dendritic cells. J Exp Med 197:121-127.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.
Takahasi, K, M Yoneyama, T Nishihori, R Hirai, H Kumeta, R Narita, M Gale, Jr., F Inagaki, T Fujita. 2008. Nonself RNA-sensing mechanism of RIG-I helicase and activation of antiviral immune responses. Mol Cell 29:428-440.
Takahata, N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc Natl Acad Sci U S A 87:2419-2423.
Takahata, N, Y Satta, J Klein. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130:925-938.
Takeuchi, O, T Kawai, H Sanjo, NG Copeland, DJ Gilbert, NA Jenkins, K Takeda, S Akira. 1999. TLR6: A novel member of an expanding toll-like receptor family. Gene 231:59-65.
Tassaneetrithep, B, TH Burgess, A Granelli-Piperno, et al. 2003. DC-SIGN (CD209) mediates dengue virus infection of human dendritic cells. J Exp Med 197:823-829.
Tian, X, G Pascal, P Monget. 2009. Evolution and functional divergence of NLRP genes in mammalian reproductive systems. BMC Evol Biol 9:202.
Tishkoff, SA, FA Reed, A Ranciaro, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31-40.
Tishkoff, SA, R Varkonyi, N Cahinhinan, et al. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455-462.
Tournamille, C, Y Colin, JP Cartron, C Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10:224-228.
Triantafilou, M, K Brandenburg, S Kusumoto, K Fukase, A Mackie, U Seydel, K Triantafilou. 2004. Combinational clustering of receptors following stimulation by bacterial products determines lipopolysaccharide responses. Biochem J 381:527-536.
Trinchieri, G, A Sher. 2007. Cooperation of Toll-like receptor signals in innate immune defence. Nat Rev Immunol 7:179-190.
Troelsen, JT. 2005. Adult-type hypolactasia and regulation of lactase expression. Biochim Biophys Acta 1723:19-32.
Tschopp, J, F Martinon, K Burns. 2003. NALPs: a novel protein family involved in inflammation. Nat Rev Mol Cell Biol 4:95-104.
Tuzun, E, AJ Sharp, JA Bailey, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet 37:727-732.
Underhill, PA, T Kivisild. 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet 41:539-564.

Underhill, PA, P Shen, AA Lin, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361.
Vallender, EJ, BT Lahn. 2004. Positive selection on the human genome. Hum Mol Genet 13 Spec No 2:R245-254.
Van Valen, L. 1973. A new evolutionary law. . Evol Theory 1:1-30.
Vasseur, E, E Patin, G Laval, S Pajon, S Fornarino, B Crouau-Roy, L Quintana-Murci. 2011. The selective footprints of viral pressures at the human RIG-I-like receptor family. Hum Mol Genet 20:4462-4474.
Venkataraman, T, M Valdes, R Elsby, S Kakuta, G Caceres, S Saijo, Y Iwakura, GN Barber. 2007. Loss of DExD/H box RNA helicase LGP2 manifests disparate antiviral responses. J Immunol 178:6444-6455.
Venter, JCMD AdamsEW Myers, et al. 2001. The sequence of the human genome. Science 291:1304-1351.
Viala, J, C Chaput, IG Boneca, et al. 2004. Nod1 responds to peptidoglycan delivered by the Helicobacter pylori cag pathogenicity island. Nat Immunol 5:1166-1174.
Voight, BF, AM Adams, LA Frisse, Y Qian, RR Hudson, A Di Rienzo. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102:18508-18513.
Voight, BF, S Kudaravalli, X Wen, JK Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol 4:e72.
Wallace, DC, MD Brown, MT Lott. 1999. Mitochondrial DNA variation in human evolution and disease. Gene 238:211-230.
Walsh, EC, KA Mather, SF Schaffner, et al. 2003. An integrated haplotype map of the human major histocompatibility complex. Am J Hum Genet 73:580-590.
Wang, ET, G Kodama, P Baldi, RK Moyzis. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103:135-140.
Wang, Q, DR Nagarkar, ER Bowman, et al. 2009. Role of double-stranded RNA pattern recognition receptors in rhinovirus-induced airway epithelial cell responses. J Immunol 183:6989-6997.
Watanabe, T, A Kitani, PJ Murray, W Strober. 2004. NOD2 is a negative regulator of Toll-like receptor 2-mediated T helper type 1 responses. Nat Immunol 5:800-808.
Watkins, WS, CE Ricker, MJ Bamshad, ML Carroll, SV Nguyen, MA Batzer, HC Harpending, AR Rogers, LB Jorde. 2001. Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet 68:738-752.
Wedderburn, LR, A Patel, H Varsani, P Woo. 2001. The developing human immune system: T-cell receptor repertoire of children and young adults shows a wide discrepancy in the frequency of persistent oligoclonal T-cell expansions. Immunology 102:301-309.
Weir, BS, WG Hill. 2002. Estimating F-statistics. Annu Rev Genet 36:721-750.
Wickliffe, KE, SH Leppla, M Moayeri. 2008. Anthrax lethal toxin-induced inflammasome formation and caspase-1 activation are late events dependent on ion fluxes and the proteasome. Cell Microbiol 10:332-343.
Wilder, JA, Z Mobasher, MF Hammer. 2004. Genetic evidence for unequal effective population sizes of human females and males. Mol Biol Evol 21:2047-2057.
Witebsky, E, NR Rose, K Terplan, JR Paine, RW Egan. 1957. Chronic thyroiditis and autoimmunization. J Am Med Assoc 164:1439-1447.
Wlasiuk G & Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. Mol Biol Evol 27(9):2172-2186.
Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.
Wright, S. 1943. Isolation by Distance. Genetics 28:114-138.

Xu, LG, YY Wang, KJ Han, LY Li, Z Zhai, HB Shu. 2005. VISA is an adapter protein required for virus-triggered IFN-beta signaling. Mol Cell 19:727-740.
Yi, X, Y Liang, E Huerta-Sanchez, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75-78.
Yoneyama, M, T Fujita. 2007. Function of RIG-I-like receptors in antiviral innate immunity. J Biol Chem 282:15315-15318.
Yoneyama, M, M Kikuchi, K Matsumoto, et al. 2005. Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. J Immunol 175:2851-2858.
Yoneyama, M, M Kikuchi, T Natsukawa, N Shinobu, T Imaizumi, M Miyagishi, K Taira, S Akira, T Fujita. 2004. The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. Nat Immunol 5:730-737.
Zelensky, AN, JE Gready. 2005. The C-type lectin-like domain superfamily. FEBS J 272:6179-6217.
Zhang, B, X Pan, GP Cobb, TA Anderson. 2007a. microRNAs as oncogenes and tumor suppressors. Dev Biol 302:1-12.
Zhang, H, PN Tay, W Cao, W Li, J Lu. 2002. Integrin-nucleated Toll-like receptor (TLR) dimerization reveals subcellular targeting of TLRs and distinct mechanisms of TLR4 activation and signaling. FEBS Lett 532:171-176.
Zhang, P, M Dixon, M Zucchelli, F Hambiliki, L Levkov, O Hovatta, J Kere. 2008. Expression analysis of the NLRP gene family suggests a role in human preimplantation development. PLoS One 3:e2755.
Zhang, SY, E Jouanguy, S Ugolini, et al. 2007b. TLR3 deficiency in patients with herpes simplex encephalitis. Science 317:1522-1527.
Zhang, X, P Shan, G Jiang, L Cohn, PJ Lee. 2006. Toll-like receptor 4 deficiency causes pulmonary emphysema. J Clin Invest 116:3050-3059.
Zhernakova, A, CC Elbers, B Ferwerda, et al. 2010. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am J Hum Genet 86:970-977.
Zurawek, M, M Fichna, D Januszkiewicz-Lewandowska, M Gryczynska, P Fichna, J Nowak. 2010. A coding variant in NLRP1 is associated with autoimmune Addison's disease. Hum Immunol 71:530-534.

ANNEXE 1 : Tests de neutralité

Tajima *D* statistic

Tajima (1989) developed a statistical test of neutrality that uses only polymorphism data within a population. The test statistic, *D*, uses information on the total number of polymorphic nucleotide sites observed (θ_W) and the average number of differences between all pairs of sequences sampled ($\theta\pi$). θ_W it is given by

$$\theta_W = S / a_n, \text{ where, } a_n = \sum_{i=1}^{n-1} 1/i$$

where (S) represents the number of segregating sites observed and (n) the sample size.
The average number of differences between all pairs of sequences sampled ($\theta\pi$) it is given by,

$$\theta_\pi = \frac{\sum_{i=1}^{n_A-1} \sum_{j=i+1}^{n_A} d_{ij}}{{}_{n_A}C_2}$$

with d_{ij} the number of observed differences between the i^{th} and j^{th} haplotypes and ${}_{nA}C_2$ the number of pairwise comparisons.

Under the standard neutral model of a randomly mating population of constant size, θ_W and $\theta\pi$ are unbiased estimators of the population mutation rate, θ. Thus, under neutrality, both estimators are expected to produce the same value, and Tajima's *D* statistic is designed to test whether the difference between them is unlikely to be observed by chance alone

$$D = \frac{\theta\pi - \theta w}{\sqrt{Var(\theta\pi - \theta w)}}$$

Watterson's estimator (θ_W) is only influenced by the number of segregating sites whereas $\theta\pi$ is sensitive to allele frequencies at segregating sites, such that alleles at intermediate frequencies contribute more to this estimator than alleles at low frequencies. Consequently, if a sample has an excess of rare variants, Tajima *D* will be negative (i.e. $\theta\pi << \theta_W$). In contrast, if there is an excess of intermediate frequency variants, Tajima *D* will be positive (i.e. $\theta\pi >> \theta_W$). Thus, Tajima's *D* statistic will tend to be negative under positive selection due to an excess of rare alleles and positive under balancing selection from an excess of intermediate frequency alleles.

Fu and Li's *D* statistic

In 1993 Fu and Li showed that the expected number of derived mutations that are present only once in a sample (i.e. singletons), ηe, is also equal to θ. Consequently, it is possible to construct a test statistic in a similar manner to Tajima's.

Fu and Li's $D = \frac{S - an\eta e}{\sqrt{Var(S - an\eta e)}}$

Like for Tajima's *D*, a negative value indicates an excess of singletons (which would also give a negative Tajima), and a positive value indicates a lack of singletons (which would typically, though not necessarily, give a positive Tajima's *D*). However, certain population genetic scenarios, particularly selective sweeps, tend to generate an excess of singletons, to which this test can be more sensitive than Tajima's *D*.

Fu and Li's *F* statistic

This test is identical to the Fu and Li's *D* statistic, except that instead of comparing the number of singletons against the number of polymorphic sites (S) it will do it against the average number of differences between all pairs of sequences sampled ($\theta\pi$)

Fu and Li's $F = \frac{\theta\pi - \eta e}{\sqrt{Var(\theta\pi - \eta e)}}$

Fay and Wu's *H* statistic

Fu (1995) showed that the expected number of mutations at which the derived allele is represented *i* times in a sample, ξ_i, is given by

$$\mathrm{E}[\xi i] = \theta / i$$

In 2000 Fay and Wu proposed a new test which uses an estimator of θ, $\theta_{H,,}$ which is heavily influenced by high frequency derived mutations (that is, nonancestral).

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

By comparing π with the new estimator, the test statistic *H* provides a way of identifying samples in which there is an excess of high-frequency derived mutations, which is a hallmark of positive selection. Unlike the previous tests, the variance of the test has to be estimated by stochastic simulation.

ANNEXE 2 : *Supplementary material for article 1*

This file contains three supplementary tables (S1-S3) and eight supplementary figures (S1 to S8)

Table S1. Populations included in the HGDP-CEPH resequencing sub-panel.

Population	Geographical origin	Region	Number of individuals
Bantu	South Africa	sub-Saharan Africa	8
Bantu	Kenya	sub-Saharan Africa	11
Yoruba	Nigeria	sub-Saharan Africa	22
Mandenka	Senegal	sub-Saharan Africa	21
sub-Saharan African		***sub-Saharan Africa***	***62***
Adygei	Russia Caucasus	Europe	7
Russian	Russia	Europe	15
French	France	Europe	8
French Basque	France	Europe	12
Orcadian	Orkney Islands	Europe	6
North Italian	Italy (Bergamo)	Europe	3
Sardinian	Italy	Europe	11
European		***Europe***	***62***
Han	China	Asia	15
Dai	China	Asia	2
Lahu	China	Asia	2
Naxi	China	Asia	3
She	China	Asia	3
Yizu	China	Asia	2
Miaozu	China	Asia	4
Tujia	China	Asia	1
Tu	China	Asia	2
Xibo	China	Asia	4
Hezhen	China	Asia	3
Mongola	China	Asia	4
Daur	China	Asia	2
Oroqen	China	Asia	1
Cambodian	Cambodia	Asia	4
Japanese	Japan	Asia	10
Asian		***Asia***	***62***

Table S2. Genomic regions resequenced for the RLR genes. Chromosome location is given according to the hg19 (GRCh37) human assembly coordinates. Positions are given as relative to ATG position. Lengths are given in base pairs.

Gene	Chromosome location	Fragments sequenced	Gene location	Total length	Exonic length	Non-exonic length
RIG_I NM_014314.3	chr9:32,455,300-32,526,322	-567 : -159	upstream	7485	2775	4710
		-158 : -1	5'UTR			
		1 : 106	exon 1			
		107 : 120	intronic			
		24930 : 25227	intronic			
		25228 : 25362	exon 2			
		25363 : 25426	intronic			
		31830 : 32224	intronic			
		32225 : 32406	exon 3			
		32407 : 32410	intronic			
		33599 : 33628	intronic			
		33629 : 33776	exon 4			
		33777 : 34290	intronic			
		34414 : 34746	intronic			
		34747 : 34866	exon 5			
		34867 : 34887	intronic			
		36591 : 36715	intronic			
		36716 : 36823	exon 6			
		36824 : 37279	intronic			
		37280 : 37433	exon 7			
		37434 : 37440	intronic			
		37951 : 37963	intronic			
		37964 : 38222	exon 8			
		38223 : 38533	intronic			
		38534 : 38696	exon 9			
		38697 : 38722	intronic			
		40877 : 40887	intronic			
		40888 : 40992	exon 10			
		40993 : 41016	intronic			
		44650 : 44669	intronic			
		44670 : 44827	exon 11			
		44828 : 44842	intronic			
		45312 : 45812	intronic			
		45813 : 45948	exon 12			
		45949 : 46039	intronic			
		48999 : 49035	intronic			
		49036 : 49184	exon 13			
		49185 : 49519	intronic			
		52957 : 53101	intronic			
		53102: 53192	exon 14			
		53193 : 53269	intronic			
		58229 : 58234	intronic			
		58235 : 58405	exon 15			

		58406 : 58460	intronic			
		59692 : 59725	intronic			
		59726 : 59877	exon 16			
		59878 : 59951	intronic			
		66628 : 66652	intronic			
		66653 : 66796	exon 17			
		66797 : 66949	intronic			
		68749 : 69045	exon 18			
IFIH1 NM_022168.2	chr2:163,123,589-163,175,039	-117 : -1	5'UTR	7955	3075	4880
		1 : 453	exon 1			
		454 : 480	intronic			
		7227 : 7374	intronic			
		7375 : 7543	exon 2			
		7544 : 7670	intronic			
		11441 : 11452	intronic			
		11453 : 11599	exon 3			
		11600 : 11666	intronic			
		24538 : 24639	intronic			
		24640 : 24744	exon 4			
		24745 : 24840	intronic			
		29757 : 29952	intronic			
		29953 : 30173	exon 5			
		30174 : 30180	intronic			
		35645 : 35731	intronic			
		35732 : 35942	exon 6			
		35943 : 36762	intronic			
		36763 : 36980	exon 7			
		36981 : 37060	intronic			
		37996 : 38195	intronic			
		38196 : 38312	exon 8			
		38313 : 38342	intronic			
		39838 : 39979	intronic			
		39980 : 40103	exon 9			
		40104 : 40125	intronic			
		40191 : 40614	intronic			
		40615 : 40893	exon 10			
		40894 : 40900	intronic			
		41335 : 41361	intronic			
		41362 : 41621	exon 11			
		41622 : 41755	intronic			
		44221 : 44363	intronic			
		44364 : 44513	exon 12			
		44514 : 44820	intronic			
		45561 : 45920	intronic			
		45921 : 46082	exon 13			
		46083 : 46250	intronic			
		49812 : 50030	intronic			
		50031 : 50221	exon 14			
		50222 : 50738	intronic			
		50739 : 50829	exon 15			

		50830 : 50928	intronic			
		50929 : 51108	exon 16			
		51109 : 51228	3'UTR			
		51229 : 51300	downstream			
LGP2 NM_024119.2	chr17:40,253,422-40,264,751	-914 : -842	upstream	5805	2034	3771
		-841 : -713	5'UTR			
		-712 : -687	intronic			
		-621 : -314	intronic			
		-313 : -221	5'UTR			
		-220 : -2	intronic			
		-1	5'UTR			
		1 : 168	exon 1			
		169 : 259	intronic			
		340 : 395	intronic			
		396 : 597	exon 2			
		598 : 979	intronic			
		980 : 1170	exon 3			
		1171 : 1199	intronic			
		2461 : 2506	intronic			
		2507 : 2623	exon 4			
		2624 : 2960	intronic			
		3603 : 3784	intronic			
		3785 : 3911	exon 5			
		3912 : 4097	intronic			
		4098 : 4289	exon 6			
		4290 : 4523	intronic			
		5722 : 5903	intronic			
		5904 : 6157	exon 7			
		6158 : 6181	intronic			
		6714 : 6725	intronic			
		6726 : 6875	exon 8			
		6876 : 6965	intronic			
		6966 : 7127	exon 9			
		7128 : 7469	intronic			
		8070 : 8094	intronic			
		8095 : 8285	exon 10			
		8286 : 8549	intronic			
		9551 : 9580	intronic			
		9581 : 9677	exon 11			
		9678 : 9946	intronic			
		9947 : 10132	exon 12			
		10133 : 10270	3'UTR			

Table S3. Nucleotide substitutions at RLRs identified in this study.

Gene	Position to ATG	Mutation	Gene location	Mutation type	Protein domain	AA change	PolyPhen[a]	Allele Frequency			dbSNP	AFR iHS	EUR iHS	ASI iHS
								AFR	EUR	ASI				
RIG-I	-494	G>A	upstream	non-coding	-	-	-	10.48	0.00	0.00	rs2777730	-	NA	NA
RIG-I	-207	A>T	upstream	non-coding	-	-	-	1.61	0.00	0.00	-			
RIG-I	-153	A>T	5'UTR	non-coding	-	-	-	0.00	0.00	1.61	rs12235719	NA	NA	NA
RIG-I	-113	T>G	5'UTR	non-coding	-	-	-	0.00	2.42	0.00	-			
RIG-I	-69	C>G	5'UTR	non-coding	-	-	-	38.71	66.13	40.32	rs3739674	1.506	1.327	2.916
RIG-I	-8	A>C	5'UTR	non-coding	-	-	-	0.00	1.61	0.00	-			
RIG-I	19	C>T	exon 1	non-synonymous	CARD1	R7C	P damaging	18.55	26.61	6.45	rs10813831	0.126	-1.014	0.009
RIG-I	23	G>A	exon 1	non-synonymous	CARD1	S8N	Benign	0.81	0.00	0.00	-			
RIG-I	88	G>A	exon 1	non-synonymous	CARD1	A30T[b]	Benign	1.61	0.00	0.00	-			
RIG-I	89	C>T	exon 1	non-synonymous	CARD1	A30V[b]	Benign	1.61	0.00	0.00	-			
RIG-I	25272	A>G	exon 2	non-synonymous	CARD1	M51V	Benign	0.00	0.00	0.81	-			
RIG-I	25313	C>G	exon 2	synonymous	CARD1	-	-	0.00	0.81	0.00	-			
RIG-I	25333	G>A	exon 2	non-synonymous	CARD1	R71H	P damaging	0.00	0.81	0.00	rs72710678	NA	NA	NA
RIG-I	31875	C>A	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
RIG-I	31926	G>A	intron	non-coding	-	-	-	2.42	0.00	0.00	-			
RIG-I	31937	G>T	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
RIG-I	31960	T>C	intron	non-coding	-	-	-	0.00	2.42	0.00	-			
RIG-I	31990	G>A	intron	non-coding	-	-	-	7.26	0.00	0.00	-			
RIG-I	32136	T>A	intron	non-coding	-	-	-	0.00	1.61	0.00	-			
RIG-I	32277	A>G	exon 3	synonymous	CARD2	-	-	2.42	0.00	0.00	-			
RIG-I	32398	A>C	exon 3	non-synonymous	CARD2	I139L	P damaging	4.84	1.61	0.00	rs78786043	NA	NA	NA
RIG-I	33610	C>G	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
RIG-I	33636	C>T	exon 4	non-synonymous	CARD2	S144F	Benign	6.45	0.81	4.03	rs55789327	NA	NA	NA
RIG-I	33813	G>A	intron	non-coding	-	-	-	92.74	87.90	95.16	rs17289927	1.043	0.25	-
RIG-I	34460	T>C	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
RIG-I	34473	G>A	intron	non-coding	-	-	-	6.45	0.81	4.03	rs75507670	NA	NA	NA
RIG-I	34527	T>G	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
RIG-I	34562	C>T	intron	non-coding	-	-	-	4.84	0.00	0.00	-			
RIG-I	34693	C>G	intron	non-coding	-	-	-	1.61	0.00	0.00	-			
RIG-I	34712	T>C	intron	non-coding	-	-	-	1.61	0.00	0.00	-			

RIG-I	36613	C>A	intron	non-coding	-	-	-	98.39	100.00	100.00	-		
RIG-I	36758	A>G	exon 6	non-synonymous	Helicase	N245S	Benign	4.03	0.00	0.00	-		
RIG-I	36924	G>A	intron	non-coding	-	-	-	0.00	0.00	0.81	-		
RIG-I	36926	T>A	intron	non-coding	-	-	-	1.61	0.00	0.00	-		
RIG-I	37017	A>C	intron	non-coding	-	-	-	4.03	0.00	0.00	-		
RIG-I	37059	T>C	intron	non-coding	-	-	-	0.00	0.81	0.00	-		
RIG-I	37123	T>C	intron	non-coding	-	-	-	1.61	0.00	0.00	-		
RIG-I	37206	A>C	intron	non-coding	-	-	-	1.61	0.00	0.00	-		
RIG-I	37258	T>C	intron	non-coding	-	-	-	5.65	0.00	0.00	-		
RIG-I	37371	T>G	exon 7	synonymous	Helicase	-	-	98.39	100.00	100.00	-		
RIG-I	38069	C>T	exon 8	synonymous	Helicase	-	-	0.00	0.81	0.00	rs45471397	NA	NA	NA
RIG-I	38253	A>G	intron	non-coding	-	-	-	12.90	0.00	0.00	rs12001044	-1.189	NA	NA
RIG-I	38268	C>G	intron	non-coding	-	-	-	1.61	0.00	0.00	-			
RIG-I	38338	C>G	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
RIG-I	38442	C>T	intron	non-coding	-	-	-	1.61	0.00	0.00	-			
RIG-I	38644	A>G	exon 9	synonymous	Helicase	-	-	11.29	1.61	12.10	rs2274863	-0.442	NA	1.655
RIG-I	40885	T>C	intron	non-coding	-	-	-	10.48	1.61	12.10	rs2274862	-0.67	NA	1.769
RIG-I	44800	C>T	exon 11	synonymous	Helicase	-	-	2.42	0.00	0.00	-			
RIG-I	44826	G>A	exon 11	non-synonymous	Helicase	R546Q	P damaging	0.00	2.42	0.00	rs61752945	NA	NA	NA
RIG-I	44838	C>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
RIG-I	45573	A>G	intron	non-coding	-	-	-	22.58	11.29	4.84	rs34804468	NA	NA	NA
RIG-I	45581	A>G	intron	non-coding	-	-	-	2.42	0.00	0.00	rs55895980	NA	NA	NA
RIG-I	45736	A>C	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
RIG-I	45851	A>G	exon 12	synonymous	Helicase	-	-	0.81	0.00	0.00	-			
RIG-I	45911	C>T	exon 12	synonymous	Helicase	-	-	1.61	0.00	0.00	-			
RIG-I	45914	T>A	exon 12	non-synonymous	Helicase	D580E	Ps damaging	0.00	11.29	4.84	rs17217280	NA	0.211	-
RIG-I	49272	A>G	intron	non-coding	-	-	-	10.48	0.00	0.00	-			
RIG-I	49414	G>A	intron	non-coding	-	-	-	1.61	0.00	0.00	rs7341899	NA	NA	NA
RIG-I	52997	C>T	intron	non-coding	-	-	-	2.42	0.00	0.00	rs78930212	NA	NA	NA
RIG-I	53077	A>G	intron	non-coding	-	-	-	0.00	11.29	4.84	rs13295938	NA	NA	NA
RIG-I	58300	C>T	exon 15	synonymous	Helicase	-	-	7.26	0.00	0.00	rs61757209	NA	NA	NA
RIG-I	59857	G>A	exon 16	non-synonymous	RD	E773K	Benign	4.03	0.00	0.00	-			
RIG-I	59885	T>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
RIG-I	59888	G>A	intron	non-coding	-	-	-	0.00	11.29	5.65	rs45589431	NA	NA	NA
RIG-I	66646	G>A	intron	non-coding	-	-	-	0.00	1.61	0.00	-			

RIG-I	66715	A>C	exon 17	synonymous	RD	-	-	19.35	34.68	45.97	rs3205166	1.556	-0.602	0.301
RIG-I	68933	T>C	exon 18	non-synonymous	RD	I889T	P damaging	0.00	0.00	0.81	-			
RIG-I	68976	A>G	exon 18	synonymous	RD	-	-	0.00	5.65	0.00	rs10970987	NA	-	NA
RIG-I	69014	T>C	exon 18	non-synonymous	RD	I916T	Benign	4.03	0.00	0.00	rs77788008	NA	NA	NA
IFIH1	-109	C>T	5'UTR	non-coding	-	-	-	1.61	0.00	0.00	-			
IFIH1	-89	A>T	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	-13	C>G	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	103	C>T	exon 1	synonymous	CARD1	-	-	0.00	0.00	0.81	rs75342243	NA	NA	NA
IFIH1	177	T>C	exon 1	synonymous	CARD1	-	-	2.42	0.00	0.00	-			
IFIH1	258	C>G	exon 1	synonymous	CARD1	-	-	1.61	0.00	0.00	-			
IFIH1	413	G>T	exon 1	non-synonymous	CARD2	C138F	P damaging	0.00	0.00	0.81	-			
IFIH1	427	C>T	exon 1	synonymous	CARD2	-	-	2.42	0.00	0.00	-			
IFIH1	7350	T>G	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
IFIH1	7664	A>G	intron	non-coding	-	-	-	0.00	4.03	0.00	rs17715295	NA	NA	NA
IFIH1	24769	C>G	intron	non-coding	-	-	-	0.00	0.81	0.00	rs74162077	NA	NA	NA
IFIH1	29773	G>T	intron	non-coding	-	-	-	52.42	98.39	92.74	rs6746073	-0.595	-	-0.521
IFIH1	29919	C>T	intron	non-coding	-	-	-	0.00	1.61	0.00	rs35502110	NA	NA	NA
IFIH1	29942	A>C	intron	non-coding	-	-	-	0.00	0.00	1.61	-			
IFIH1	30124	A>G	exon 5	non-synonymous	Helicase	K349R	Benign	0.00	0.81	0.00	rs72650664	NA	NA	NA
IFIH1	35671	C>A	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
IFIH1	35733	T>A	exon 6	non-synonymous	Helicase	V366E	P damaging	0.81	0.00	0.00	-			
IFIH1	35817	C>A	exon 6	non-synonymous	Helicase	T394N	Benign	0.81	0.00	0.00	-			
IFIH1	35931	T>C	exon 6	non-synonymous	Helicase	V432A	Ps damaging	0.81	0.00	0.00	-			
IFIH1	36033	T>C	intron	non-coding	-	-	-	62.10	100.00	92.74	rs6734769	0.077	-	-0.59
IFIH1	36131	A>G	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	36319	T>C	intron	non-coding	-	-	-	5.65	0.81	0.00	rs73973095	NA	NA	NA
IFIH1	36335	T>G	intron	non-coding	-	-	-	15.32	0.00	0.00	rs6759894	NA	NA	NA
IFIH1	36508	T>C	intron	non-coding	-	-	-	3.23	0.00	0.00	-			
IFIH1	36531	C>G	intron	non-coding	-	-	-	16.13	0.00	0.00	rs6731458	NA	NA	NA
IFIH1	36666	A>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	36701	A>T	intron	non-coding	-	-	-	21.77	0.00	0.00	rs58625397	NA	NA	NA
IFIH1	36835	G>A	exon 7	non-synonymous	Helicase	R460H	P damaging	52.42	100.00	92.74	rs10930046	-0.773	-	-0.59
IFIH1	36937	G>T	exon 7	non-synonymous	Helicase	G494V	P damaging	3.23	0.00	0.00	-			
IFIH1	36947	G>C	exon 7	synonymous	Helicase	-	-	37.90	0.00	7.26	rs12479043	NA	NA	NA

IFIH1	38047	A>G	intron	non-coding	-	-	-	37.90	0.00	7.26	rs7590692	NA	NA	NA
IFIH1	38091	C>G	intron	non-coding	-	-	-	38.71	0.00	7.26	rs2287293	0.821	-	1.394
IFIH1	40061	A>G	exon 9	non-synonymous	Helicase	T575A	Ps damaging	0.81	0.00	0.00	-			
IFIH1	40248	A>T	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
IFIH1	40330	A>T	intron	non-coding	-	-	-	62.10	100.00	100.00	rs888284	-0.071	NA	NA
IFIH1	40340	C>T	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
IFIH1	40390	G>A	intron	non-coding	-	-	-	0.00	8.87	1.61	rs34977319	NA	NA	NA
IFIH1	40439	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	40496	C>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	40715	G>A	exon 10	synonymous	Helicase	-	-	0.81	0.00	0.00	-			
IFIH1	40760	G>A	exon 10	synonymous	Helicase	-	-	0.00	0.81	0.00	-			
IFIH1	40814	T>A	exon 10	non-synonymous	Helicase	D655E	Benign	0.81	0.00	0.00	-			
IFIH1	41432	A>C	exon 11	non-synonymous	Helicase	R705S	P damaging	0.00	0.00	0.81	-			
IFIH1	41549	T>A	exon 11	non-synonymous	Helicase	F744L	P damaging	0.00	0.00	0.81	-			
IFIH1	41662	T>C	intron	non-coding	-	-	-	0.00	0.00	16.13	rs74269124	NA	NA	NA
IFIH1	41708	C>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	41712	G>C	intron	non-coding	-	-	-	16.13	0.00	0.00	rs76054377	NA	NA	NA
IFIH1	44229	G>A	intron	non-coding	-	-	-	0.00	0.00	8.87	rs2287290	NA	NA	NA
IFIH1	44269	C>T	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
IFIH1	44345	T>C	intron	non-coding	-	-	-	8.87	0.00	0.00	rs74162085	NA	NA	NA
IFIH1	44519	T>C	intron	non-coding	-	-	-	8.87	0.00	0.00	rs6748554	NA	NA	NA
IFIH1	44522	A>G	intron	non-coding	-	-	-	0.00	0.00	1.61	-			
IFIH1	44761	G>A	intron	non-coding	-	-	-	0.00	1.61	0.00	rs13422273			
IFIH1	44777	G>A	intron	non-coding	-	-	-	0.00	0.00	3.23	-			
IFIH1	44788	C>A	intron	non-coding	-	-	-	0.00	0.00	7.26	rs41463049			
IFIH1	45610	G>A	intron	non-coding	-	-	-	21.77	0.00	0.00	rs13418892			
IFIH1	45884	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	45884	G>C	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
IFIH1	*45914*	T>A	intron	non-coding	-	-	-	8.87	0.00	0.00	rs41399348			
IFIH1	45935	C>T	exon 13	synonymous	Helicase			37.90	0.00	0.00	rs13418718	NA	-	-
IFIH1	45994	G>A	exon 13	non-synonymous	RD	R843H	P damaging	41.13	34.68	69.35	rs3747517	-1.079	0.626	-1.5
IFIH1	46093	C>T	intron	non-coding	-	-	-	52.42	100.00	92.74	rs3747518	-0.479	-	-0.331
IFIH1	49917	G>A	intron	non-coding	-	-	-	52.42	100.00	92.74	rs12474565	-0.471	-	-0.331
IFIH1	50181	A>G	exon 14	non-synonymous	RD	I923V	Ps damaging	0.00	0.81	0.00	rs35667974	NA	NA	NA
IFIH1	50199	G>A	exon 14	non-synonymous	RD	V929I	P damaging	1.61	0.00	0.00	-			

IFIH1	50222	G>A	intron	non-coding	-	-	-	0.00	0.00	1.61	rs35732034	NA	NA	NA
IFIH1	50312	T>A	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
IFIH1	50508	A>G	intron	non-coding	-	-	-	0.00	5.65	0.00	rs17764770			
IFIH1	50535	A>T	intron	non-coding	-	-	-	1.61	0.00	0.00	-			
IFIH1	50627	C>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
IFIH1	50664	G>T	intron	non-coding	-	-	-	0.00	0.00	7.26	rs3761652			
IFIH1	50714	G>A	intron	non-coding	-	-	-	0.00	0.81	0.00	rs36070180			
IFIH1	50767	G>A	exon 15	non-synonymous	RD	A946T	Benign	11.29	53.23	21.77	rs1990760	0	-0.243	1.079
IFIH1	50930	C>A	exon 16	non-synonymous	RD	A967D	Benign	0.81	0.00	0.00	-			
IFIH1	50949	G>C	exon 16	synonymous	RD	-	-	0.00	0.81	0.00	-			
IFIH1	50976	C>T	exon 16	synonymous	RD	-	-	0.81	0.00	0.00	rs74162089	NA	NA	NA
IFIH1	51114	G>A	3'UTR	non-coding	-	-	-	37.90	0.00	0.00	rs11891191	0.709	NA	NA
IFIH1	51237	C>T	downstream	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-871	C>G	upstream	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-861	C>A	upstream	non-coding	-	-	-	4.84	0.00	0.00	-			
LGP2	-808	G>A	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-755	C>A	5'UTR	non-coding	-	-	-	0.00	3.23	0.00	rs1557743	NA	NA	NA
LGP2	-751	C>G	5'UTR	non-coding	-	-	-	9.68	83.87	97.58	rs2277621	0.768	-0.956	-0.483
LGP2	-716	G>A	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-689	C>T	intron	non-coding	-	-	-	34.68	0.81	0.81	-			
LGP2	-461	C>T	intron	non-coding	-	-	-	0.00	2.42	10.48	rs12601619	NA	-	-0.413
LGP2	-302	A>G	5'UTR	non-coding	-	-	-	0.00	1.61	0.00	-			
LGP2	-283	A>G	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-275	G>A	5'UTR	non-coding	-	-	-	0.00	0.81	0.00	-			
LGP2	-242	G>A	5'UTR	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	-215	A>T	intron	non-coding	-	-	-	8.87	81.45	82.26	rs3744484	0.768	-1.227	-0.658
LGP2	-166	C>A	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
LGP2	-91	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	99	G>A	exon 1	synonymous	Helicase	-	-	0.81	0.00	0.00	-			
LGP2	125	T>C	exon 1	non-synonymous	Helicase	L42P	P damaging	0.81	0.00	0.00	-			
LGP2	434	C>T	exon 2	synonymous	Helicase	-	-	8.06	0.81	0.00	rs16967493	0.16	NA	NA
LGP2	453	A>G	exon 2	non-synonymous	Helicase	T76A	Benign	0.00	2.42	0.00	rs34891485	NA	NA	NA
LGP2	511	G>A	exon 2	non-synonymous	Helicase	R95Q	Ps damaging	2.42	4.84	0.00	rs35118457	NA	NA	NA
LGP2	811	A>T	intron	non-coding	-	-	-	0.00	0.81	0.00	-			

LGP2	876	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	933	A>C	intron	non-coding	-	-	-	10.48	0.00	0.00	rs78026875	NA	NA	NA
LGP2	994	G>A	exon 3	non-synonymous	Helicase	V129M	P damaging	0.81	0.00	0.00	-			
LGP2	2566	G>C	exon 3	synonymous	Helicase	-	-	0.00	0.81	0.00	-			
LGP2	2627	C>T	intron	non-coding	-	-	-	0.81	29.84	0.00	rs1557744	NA	NA	NA
LGP2	2641	C>T	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
LGP2	2659	C>T	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	2661	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	2820	G>A	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
LGP2	2883	G>A	intron	non-coding	-	-	-	2.42	0.00	0.00	-			
LGP2	2951	G>A	intron	non-coding	-	-	-	1.61	0.00	0.00	-			
LGP2	3607	C>T	intron	non-coding	-	-	-	0.00	0.00	1.61	-			
LGP2	3709	G>A	intron	non-coding	-	-	-	0.81	0.00	0.00	-			
LGP2	3724	C>T	intron	non-coding	-	-	-	0.00	0.00	6.45	-			
LGP2	3768	A>C	intron	non-coding	-	-	-	3.23	0.00	0.00	-			
LGP2	3931	A>T	intron	non-coding	-	-	-	8.87	81.45	82.26	rs2074154	0.575	-1.436	-0.66
LGP2	3972	G>T	intron	non-coding	-	-	-	0.00	0.81	1.61	-			
LGP2	4308	C>A	intron	non-coding	-	-	-	2.42	0.00	0.00	-			
LGP2	4423	C>T	intron	non-coding	-	-	-	5.65	0.00	0.00	-			
LGP2	5784	A>G	intron	non-coding	-	-	-	83.06	12.90	17.74	rs35370188	NA	NA	NA
LGP2	5789	T>C	intron	non-coding	-	-	-	0.00	0.00	0.81	-			
LGP2	5901	C>T	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
LGP2	5906	C>T	exon 7	non-synonymous	Helicase	R334C	Ps damaging	0.00	0.00	1.61	rs76998797	NA	NA	NA
LGP2	6748	A>G	exon 8	non-synonymous	Helicase	Q425R	Benign	84.68	22.58	16.94	rs2074158	0.391	2.292	1.401
LGP2	6856	A>G	exon 8	non-synonymous	Helicase	N461S	P damaging	2.42	13.71	1.61	rs34016093	NA	NA	NA
LGP2	7032	C>T	exon 9	synonymous	RD	-	-	4.03	0.00	0.00	-			
LGP2	7048	T>C	exon 9	non-synonymous	RD	I495T	Benign	0.00	0.00	1.61	-			
LGP2	7237	G>A	intron	non-coding	-	-	-	4.03	0.00	0.00	rs77051434	NA	NA	NA
LGP2	7413	C>G	intron	non-coding	-	-	-	49.19	19.35	15.32	rs2074159	-0.414	1.996	1.4
LGP2	8099	G>A	exon 10	non-synonymous	RD	R523Q	Benign	21.77	3.23	13.71	rs2074160	0.831	0.155	0.706
LGP2	8168	A>G	exon 10	non-synonymous	RD	Q546R	Benign	0.00	2.42	0.00	-			
LGP2	9732	G>C	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
LGP2	9846	G>A	intron	non-coding	-	-	-	0.00	1.61	0.00	rs77876788	NA	NA	NA
LGP2	9919	G>A	intron	non-coding	-	-	-	0.00	0.81	0.00	-			
LGP2	9937	T>C	intron	non-coding	-	-	-	0.00	0.81	0.00	-			

LGP2	10056	G>A	exon 12	non-synonymous	RD	R654H	Benign	0.00	0.81	0.00	-
LGP2	10126	G>C	exon 12	synonymous	RD	-	-	0.81	0.00	0.00	-
LGP2	10150	T>C	3'UTR	non-coding	-	-	-	0.81	0.00	0.00	-

[a]For the Polyphen analyses, "Ps damaging" stands for "possibly damaging", and "P damaging" for "probably damaging".

[b]The nucleotide substitutions G88A and C89T are always observed together in our dataset; consequently, the actual amino acid substitution is A30I, which is predicted as benign by Polyphen v2.

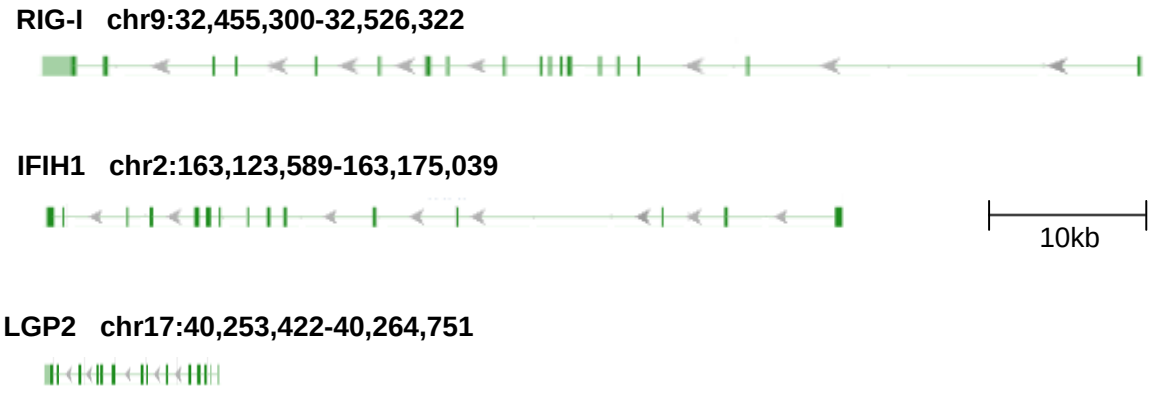


Fig. S1. Genomic organization and structure of *RIG-I*, *IFIH1* and *LGP2*.

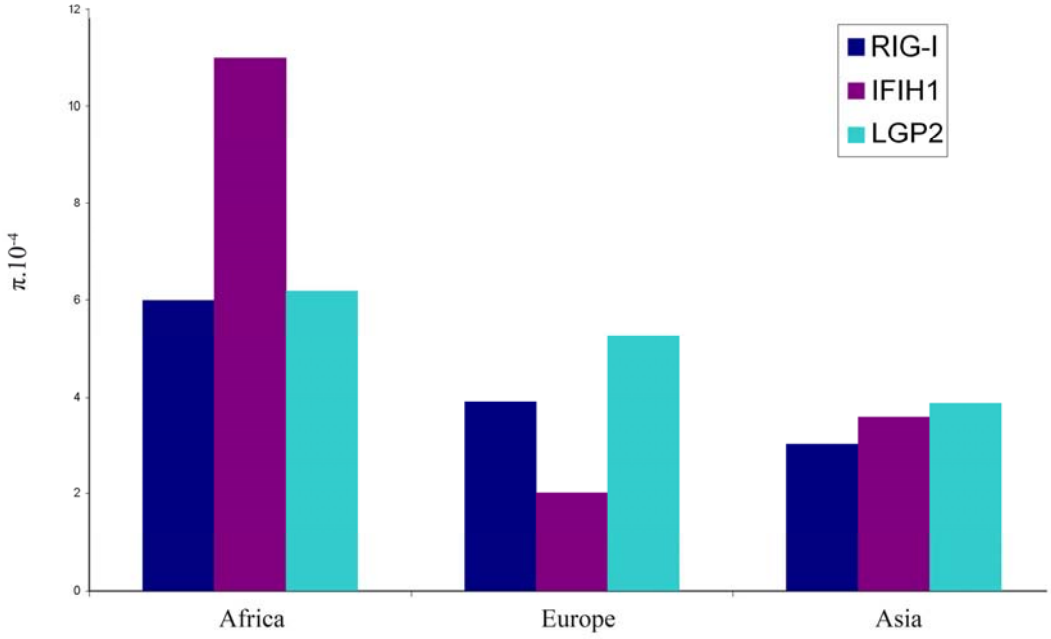


Fig. S2. Nucleotide diversity assessed by the mean number of pairwise differences between human sequences (π) in Africa, Europe and Asia for *RIG-I*, *IFIH1* and *LGP2*.

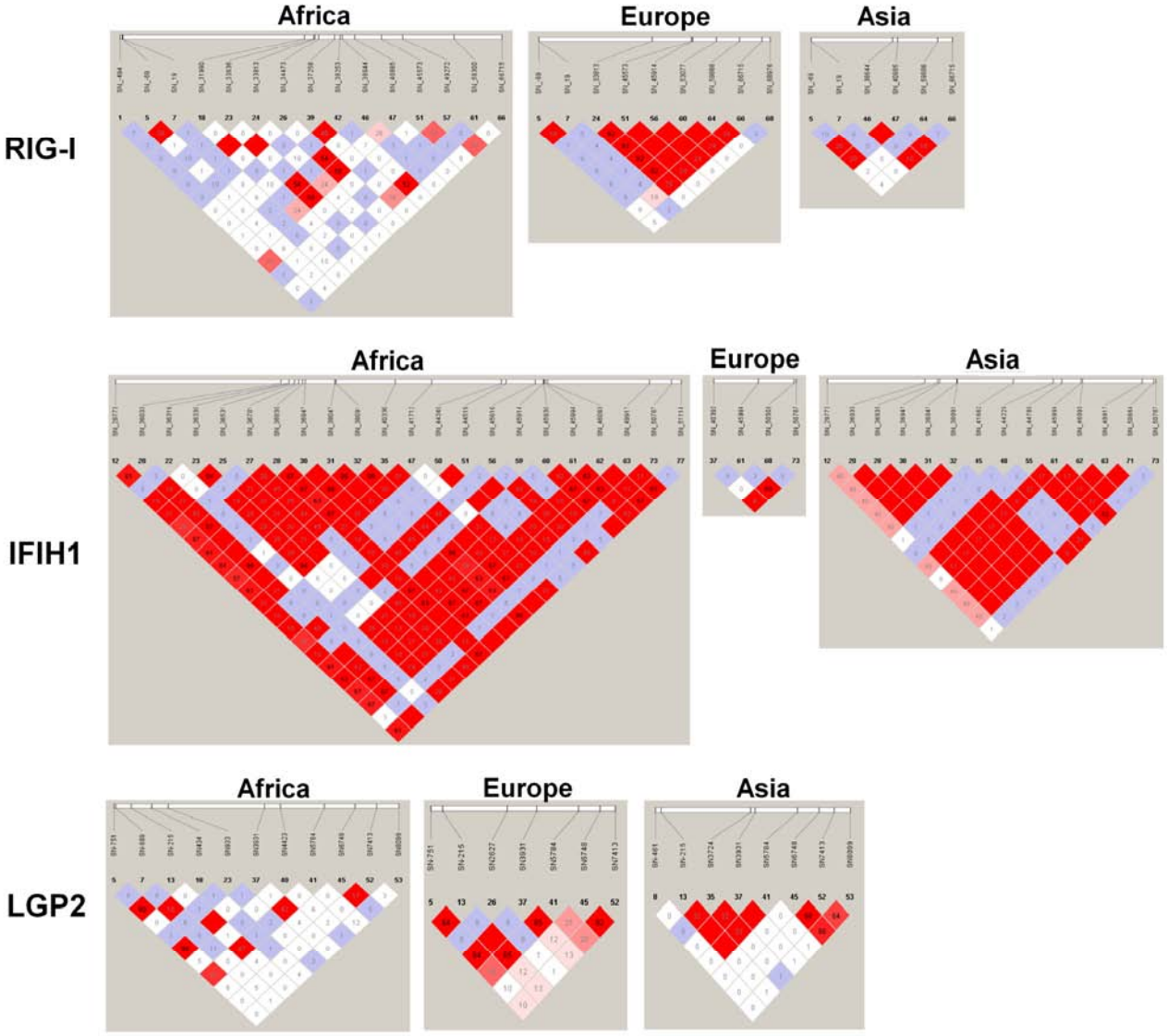


Fig. S3. Levels of linkage disequilibrium between SNPs at *RIG-I*, *IFIH1* and *LGP2* in Africa, Europe and Asia. We used a minor allele frequency cut-off of 0.05, and other Haploview settings were set at default values. Numbers represent r^2 values, whereas colours indicate *D'* levels (white, blue and red squares from non-existent/low to high LD).

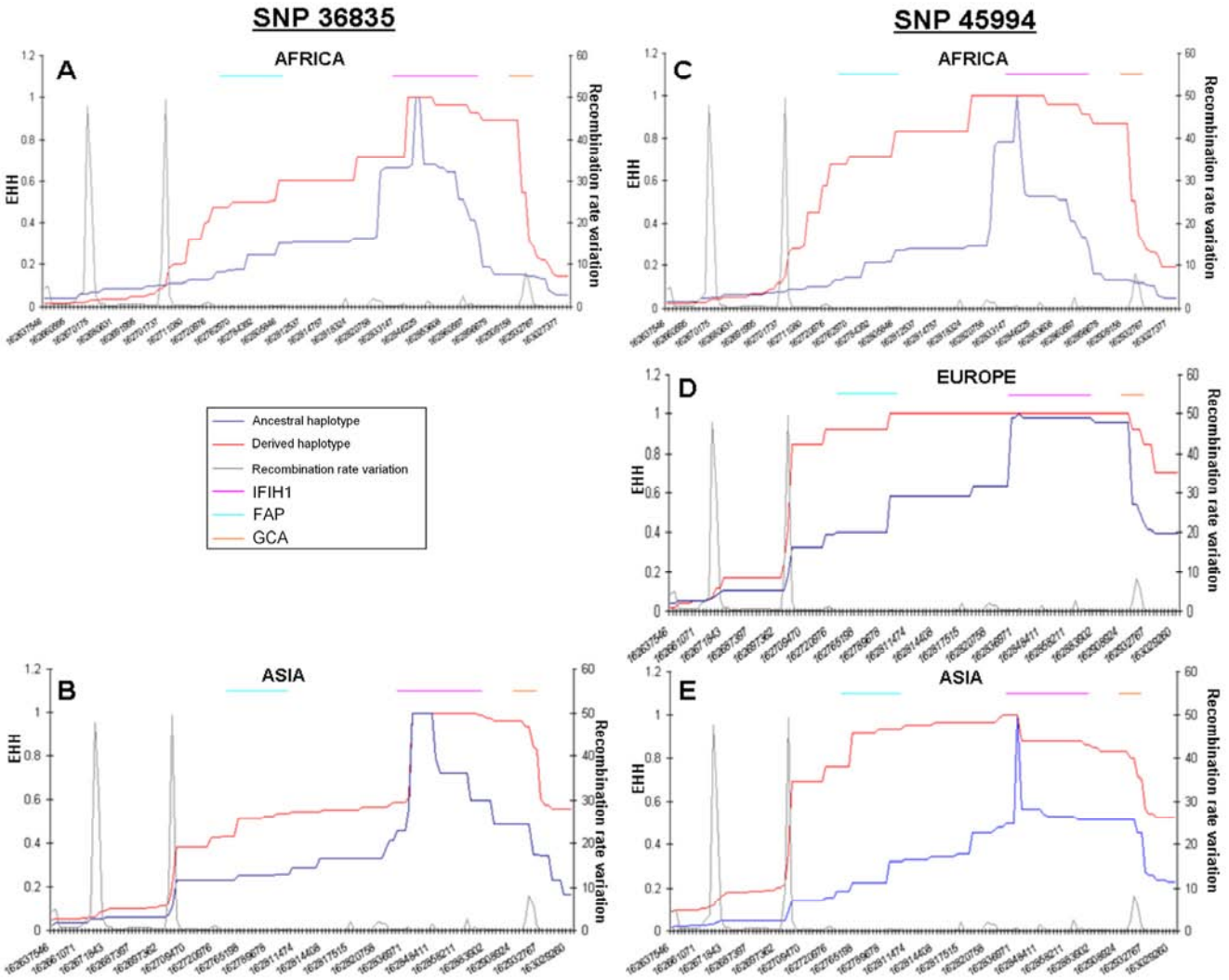


Fig. S4. EHH and recombination rate variation around *IFIH1* SNPs 36835 and 45994 in Africa, Europe and Asia. EHH is given around SNPs 36835 (**A**, **B**) and 45994 (**C**, **D**, **E)** in Africa (**A**, **C**), Europe (**D**) and Asia (**B**, **E**). EHH in Europe was not drawn for SNP 36835 as the derived allele was found to be fixed. EHH values were obtained using EHH web tool http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl. Recombination rate variation was retrieved from the HapMap recombination map. Coloured segments above the curves indicate the location of genes surrounding *IFIH1*. Ancestral and derived haplotypes are represented by dark blue and red curves, respectively.

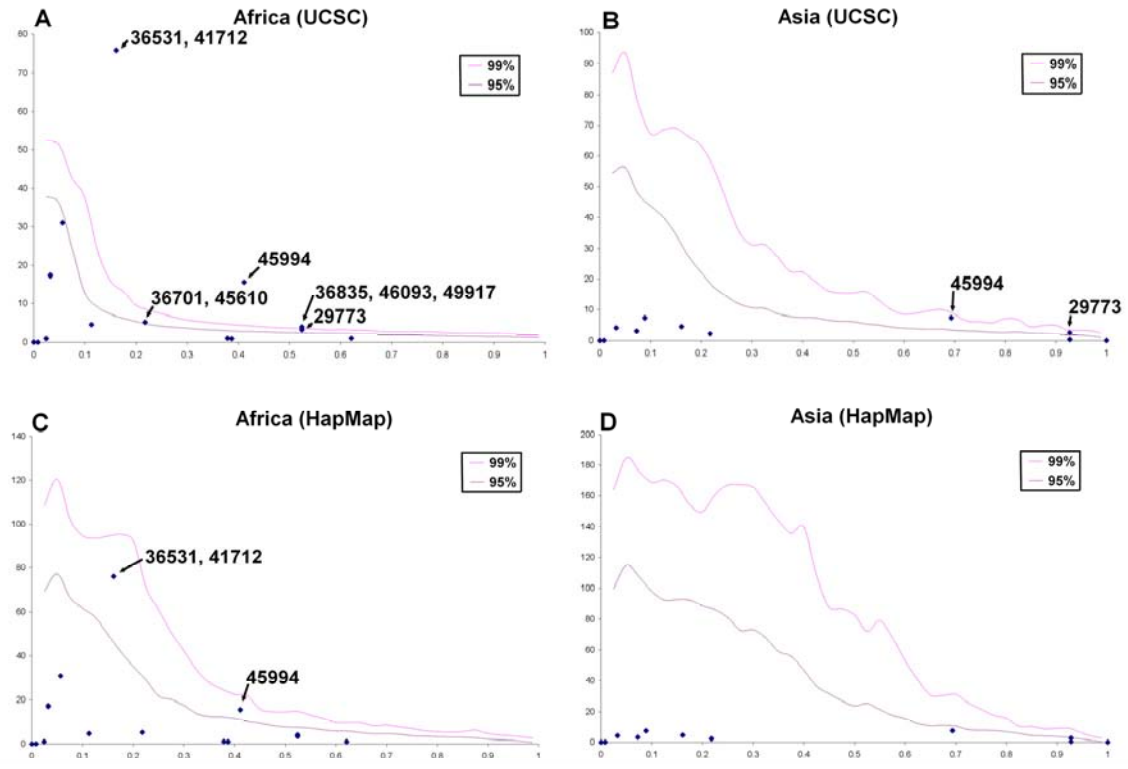


Fig. S5. Positive selection signals on *IFIH1* in Africa (A, C) and Asia (B, D) by means of the DIND test. 95 and 99% distributions were calculated using UCSC (A, B) and HapMap (C, D) recombination rates. The results represented here were obtained using the Laval's demographic model, but the same outliers were found using the Voight's model.

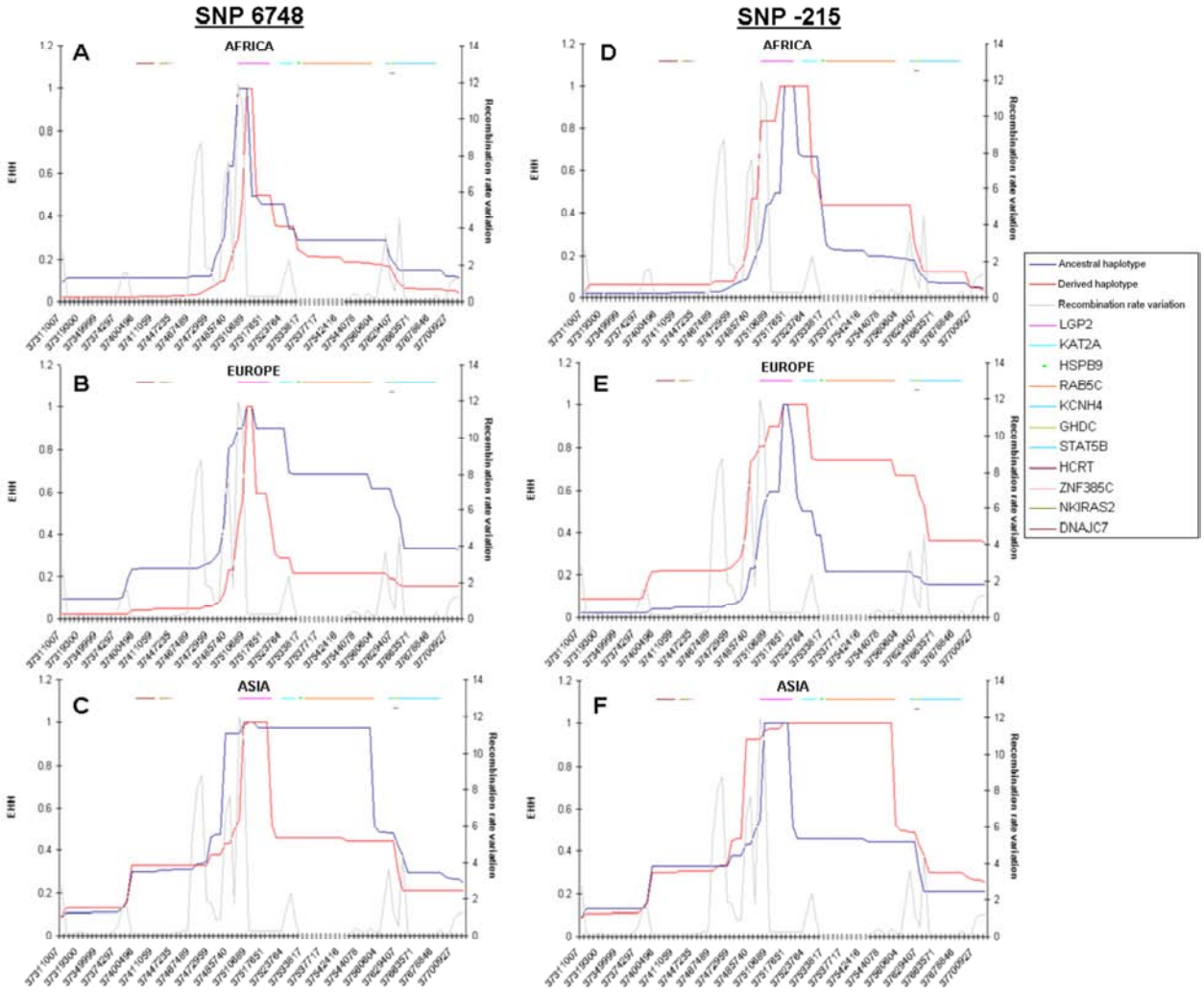


Fig. S6. EHH and recombination rate variation around *LGP2* SNPs 6748 and -215 in Africa, Europe and Asia. EHH is given around SNPs 6748 (**A**, **B, C**) and -215 (**D, E, F)** in Africa (**A**, **D**), Europe (**B, E**) and Asia (**C**, **F**). EHH for SNPs -751 and 3931 (in strong/total LD with SNP-215) gave similar results. EHH values were obtained using EHH web tool http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl. Recombination rate variation was retrieved from the HapMap recombination map. Coloured segments above the curves indicate the location of genes surrounding *LGP2*. Ancestral and derived haplotypes are represented by dark blue and red curves, respectively.

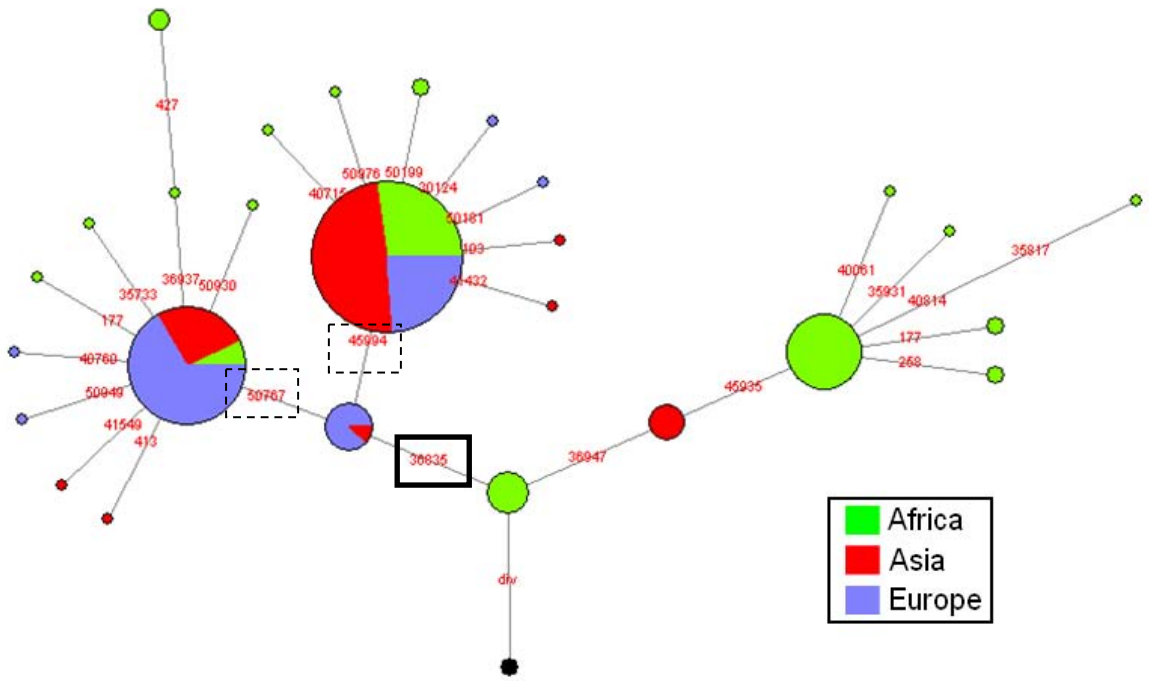


Fig. S7. Median-joining network of *IFIH1* haplotypes with Network 4.6, using coding SNPs only. Each circle represents a haplotype and circle areas are proportional to haplotype frequency. Red numbers indicate mutation positions. SNP 36835, which is found as being under positive selection in non-African populations, has been highlighted with a thick black box, whereas SNP 45994 and 50767, which have been previously found to be associated with type-I diabetes, are highlighted with a dashed-line box. It should be noticed that the derived alleles of SNP 45994 and SNP 50667 define two distinct highly frequent haplotypes and are actually never found on the same chromosome in our panel.

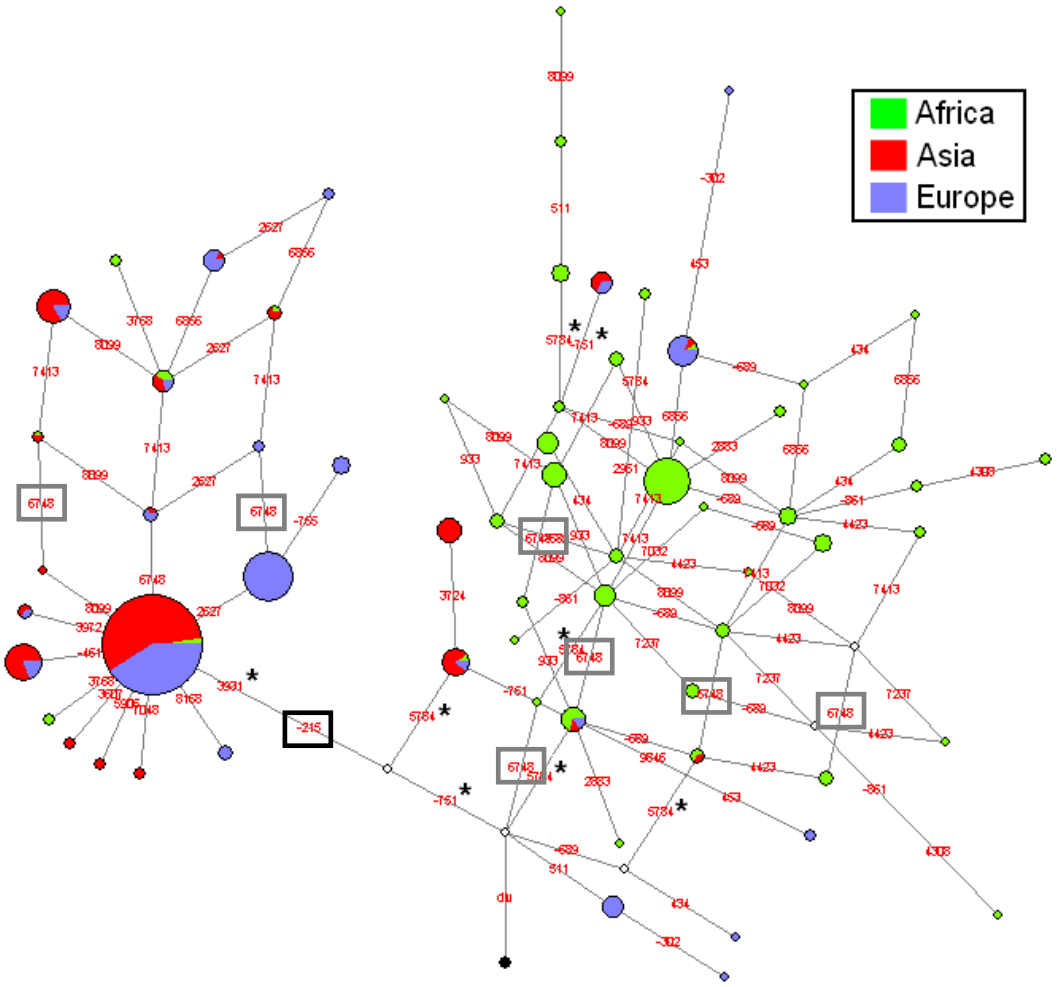


Fig. S8. Median-joining network of *LGP2* haplotypes. Each circle represents a haplotype and circle areas are proportional to haplotype frequency. Red numbers indicate positions of mutations. Mutation -215 is represented in a black frame; mutations with a star are found to be in linkage disequilibrium (r^2>0.4) with SNP -215. SNP -215 delineates two groups, strongly differentiating African (right) from European and Asian populations (left). The non-synonymous mutation 6748, which is represented with a grey frame, is observed in both groups (left and right) but at highly different heights in the tree: the derived state of SNP 6748 can be found in a minority of non-African haplotypes (left part of the tree) whereas it appears in the low part of the tree for African populations (right part of the tree).

ANNEXE 3 : *Supplementary material for article 2*

Supplemental Information

SI Materials and Methods

DNA Samples. Genetic variation was assessed at the 21 *NLR* genes in a total of 370 chromosomes from the Human Genome Diversity Panel (HGDP)–CEPH panel (1). Specifically, this subpanel includes 62 sub-Saharan Africans, 62 Europeans and 61 East Asians. Sub-Saharan African populations were composed of 19 Bantu from Kenya, 21 Mandenka from Senegal and 22 Yoruba from Nigeria; European populations include 20 French, 14 Italians, 6 Orcadians and 22 Russians; and East-Asian populations were composed of 10 Japanese, 4 Cambodians, 15 Han Chinese and 32 individuals from Chinese minorities. For a complete description of this HGDP-CEPH subpanel, see Table S1. This study was approved by the Institut Pasteur Institutional Review Board (n° RBM 2008.06).

DNA resequencing. We resequenced all the exonic regions of the 21 NLRs (except from the last exon, *i.e.* 3% of the coding region, of *NLRP8* that could not be amplified) and at least an equivalent amount of non-exonic portions including intronic, 5' and 3' regions (Table S2). The genes sequenced include the 14 members of the NALP subfamily (NALP1-14), all sharing a pyrin domain in N-terminal, and the 7 NOD/IPAF subfamily members, comprising NOD1 (CARD4), NOD2 (CARD15), NLRC3 (NOD3), NLRC5 (NOD4) and NLRX1 (NOD9) together with CIITA and NLRC4 (IPAF), which share the CARD domain in N-terminal specific to most NODs. The NLR *NAIP* was not resequenced owing to its particularly repeated genomic organization (2). The reference sequences used are indicated in Table S2. It has to be noticed that the exons we resequenced do generally correspond to the longest isoform (considered for all analyses). We inspected sequence files and chromatograms using the GENALYS software (3). To avoid SNP discovery errors, sequences were analyzed

by two different operators, and ambiguous polymorphisms were systematically reamplified and resequenced. Concerning nucleotide numbering, SNP positions correspond to the genomic DNA, taking the "A" of the ATG translation initiation codon in the reference sequence as the +1 position. To determine ancestral states at each SNP by parsimony, we used the UCSC database to retrieve the orthologous sequences of chimpanzee, gorilla, orangutan, rhesus, baboon, marmoset, tarsier, mouse-lemur and bushbaby.

Inter-species neutrality tests. To estimate the direction and strength of selection within the human species as a whole, we measured d_S and d_N — i.e. the number of silent and nonsynonymous fixed differences between humans and chimpanzees — together with p_S and p_N, — i.e. the number of silent and nonsynonymous polymorphic sites observed within humans, using DnaSP package v. 5.1 (4). Then, we used the McDonald-Kreitman Poisson Random Field (MKPRF) method (5, 6) to estimate ω (with $\omega \propto \theta_N/\theta_S$, where θ_N and θ_S are estimates of the rate of nonsynonymous and silent mutations) and γ (with $\gamma \propto \ln(d_N/p_N)$). Under neutrality, ω is not significantly different from 1. Values below 1 indicate a deficit of nonsynonymous variants, whereas values greater than 1 reflect an excess of amino-acid changes. Concerning γ, values below/above 0 indicate a deficit/an excess of non-synonymous divergence compared to polymorphism. This test was performed on the here sequenced NLRs and their results were compared with those from the RLRs (7) and TLRs (8). To assess the contributions of divergence and polymorphism to the patterns of purifying selection observed at some genes, we compared the p_N/p_S and d_N/d_S ratios between *PRRs* with significant ω values to a genomewide distribution of 1596 genes exhibiting ω values lower than 1 (5). To evaluate the power of this method, we simulated a gradient of the various possibilities for the pN/pS values, considering "neutral" dN/dS ratios (values around 1). These simulations were generated by SIMCOAL 2.0 (9) using a model of constant size. We showed that genes with ω

values significantly <1 could indeed result from “neutral” dN/dS values but very low pN/pS values. This suggests that we have enough power to assess the contributions of divergence and polymorphism to the patterns of purifying selection observed (Fig.S11).

Diversity indices and intra-species tests. Haplotype reconstruction was performed by means of the Bayesian statistical method implemented in Phase (v.2.1.1) (10). The algorithm was run five times, using different randomly-generated seeds. Consistent results were obtained across runs. The most likely run was retained for all subsequent analyses. We used Haploview software (11) to obtain and visualize levels of linkage disequilibrium (LD) in the various genomic regions of NLRs. For each population, the different summary statistics, such as the number of segregating sites (S), the number of haplotypes (H), the haplotype diversity (Hd), the average number of pairwise differences (π), and the sequence-based neutrality tests, such as Tajima’s *D*, Fu and Li’s *D**, *F** and normalized Fay and Wu’s *H* (Hn) tests were performed using DnaSP package v. 5.1 (4). Furthermore, to detect more recent signatures of positive selection, we used various haplotype-based tests and levels of population differentiation. Specifically, we used the Derived Intra-allelic Nucleotide Diversity (DIND) test based on the ratio $i\pi_A/i\pi_D$, where $i\pi_A$ and $i\pi_D$ are the levels of nucleotide diversity associated with the haplotypes carrying the ancestral and the derived allele for a given SNP, respectively (8). The rationale of this test is that a derived allele under positive selection that is at high population frequencies should present lower levels of nucleotide diversity at linked sites than expected. We also used tests based on the levels of haplotype homozygosity, such as the Extended Haplotype Homozygosity (12) using the EHH web calculator (http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl) (13) and the Cross Population Extended Haplotype Homozygosity (XP-EHH) test (14). When available, we also used the integrated haplotype Scores (iHS) (15), obtained from the HapMap Phase II dataset (16). To

determine the levels of population differentiation, we assessed the F_{ST} statistics derived from the analysis of variance (ANOVA) (17) for each SNP population pairwise.

Correction for the mimicking effects of demography. To consider the impact of demography on the patterns of diversity, we used simulation-based or empirical procedures. For the allele frequency spectrum and DIND tests, we incorporated into our neutral expectations two demographic models based on multiple, noncoding genomic regions sequenced in a set of populations similar to those used here (18, 19). The main difference between these two demographic models is that the model of (18) considers inter-continental population migration. *P*-values for the various neutrality tests were estimated from 10^4 coalescent simulations, performed using SIMCOAL 2.0 (9), under a finite-site neutral model and considering the recombination rate of the concerned region as reported in UCSC (20). Each of the 10^4 coalescent simulations was conditional on the sample size and the number of segregating sites observed in each gene. For the *DH* test, a *P-value* combining both Tajima'*D* and Fay & Wu's *H* was calculated. For the DIND analyses of *NLRP1* and *NLRP14*, the distribution of expected values were obtained using UCSC and HapMap (16) recombination rates in proportions determined by the fitting on values of low frequency alleles (<20%) that are known not to be strongly affected by natural selection. For the population differentiation tests, we compared the observed F_{ST} values at each SNP at NLRs against a background F_{ST} distribution of 650,000 SNPs genotyped in the same panel of individuals we sequenced in this study (21). Because the genome-wide F_{ST} distribution of the HGDP-CEPH dataset includes loci targeted by positive selection (22), the comparison of NLR F_{ST} against the HGDP-CEPH distribution represents a highly conservative approach to detect selection (i.e., the ''neutral'' distribution also includes selected loci). Because F_{ST} values depend on allele frequencies, F_{ST} comparisons were conditioned to SNPs presenting similar expected heterozygosity. We used

NETWORK 4.6 to construct median-joining networks in order to infer haplotype genealogy (23).

Functional diversity within protein domains. We assessed the π_N/π_S ratios using DnaSP package v. 5.1 (4). and compared them within CARD, PYD, NBD and LRR domains between all *NLRs*. As NALPs (except NALP10) and NOD/IPAF share an LRR domain in C-terminal and a central NACHT domain, we compared the π_N/π_S ratio within these domains between all *NLRs*. As NLRs differentiate in their N-terminal, we compared the π_N/π_S ratio within the PYD domain between *NALPs* only and within the CARD domain between *NLRP1, NOD1, NOD2* and *NLRC4*. The domain division was retrieved from Uniprot database (24). Note that the LRR domains include the whole region carrying the known motifs of LRRs (including inter-LRR short regions), given the constantly evolving data concerning the bounds of these motifs.

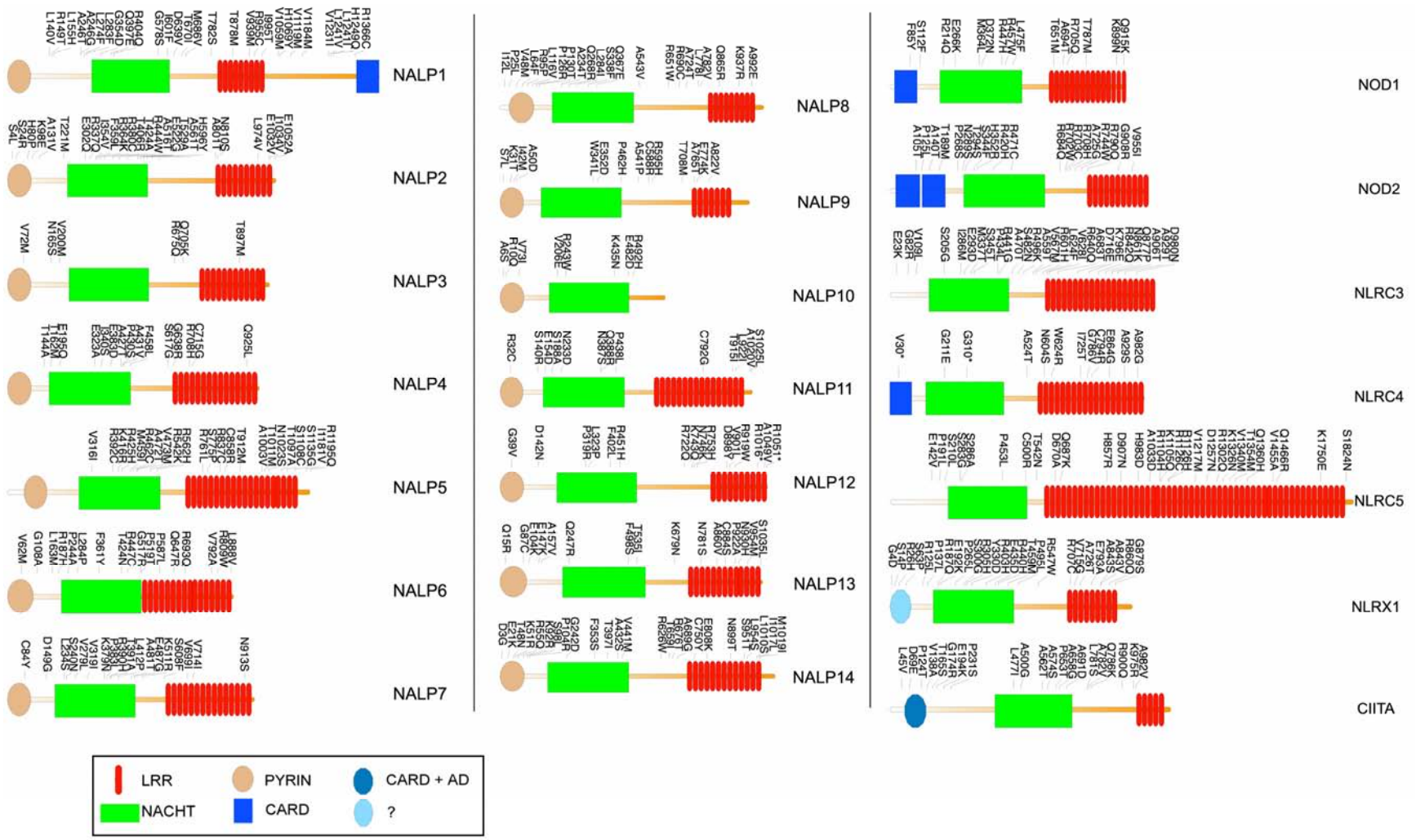


Figure S1. Distribution of the nonsynonymous and nonsense SNPs identified in this study across the *NLRs*. The location of each nonsynonymous variant within the different protein domains is shown. AD stands for activation domain. Concerning NLRX1, the N-terminal domain is neither a PYRIN nor a CARD.

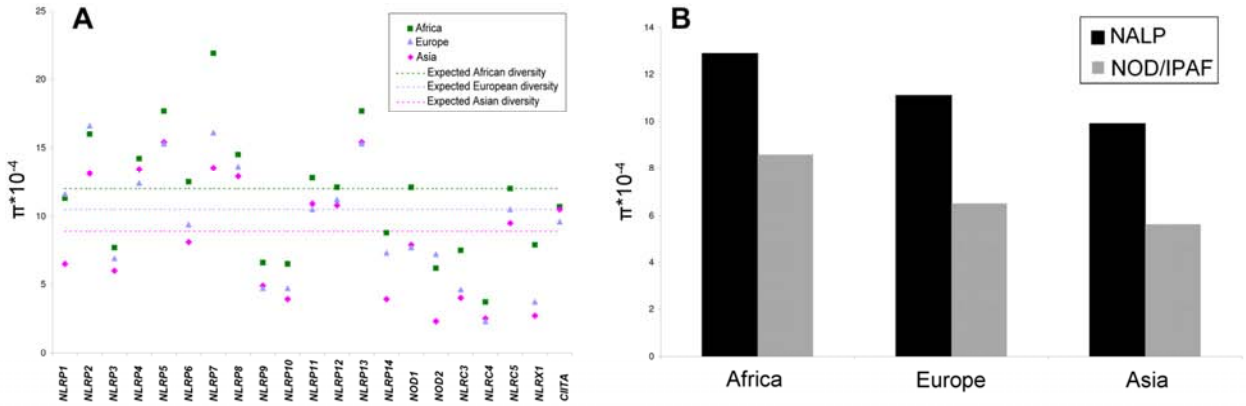


Figure S2. Patterns of nucleotide diversity (π*10^{-4}) for the *NLRs*. (*A*) Nucleotide diversity levels for the individual genes, assessed by the mean number of pairwise differences between sequences, in Sub-Saharan Africa, Europe and East Asia. Different colours represent different continents (dark green for Africa, blue for Europe and pink for Asia). The expected diversity (dashed lines) corresponds to the mean diversity levels observed for 20 autosomal noncoding regions in each geographic area (18). (*B*) Comparison of mean nucleotide diversity between the NALP (in black) and the NOD/IPAF (in grey) subfamilies.

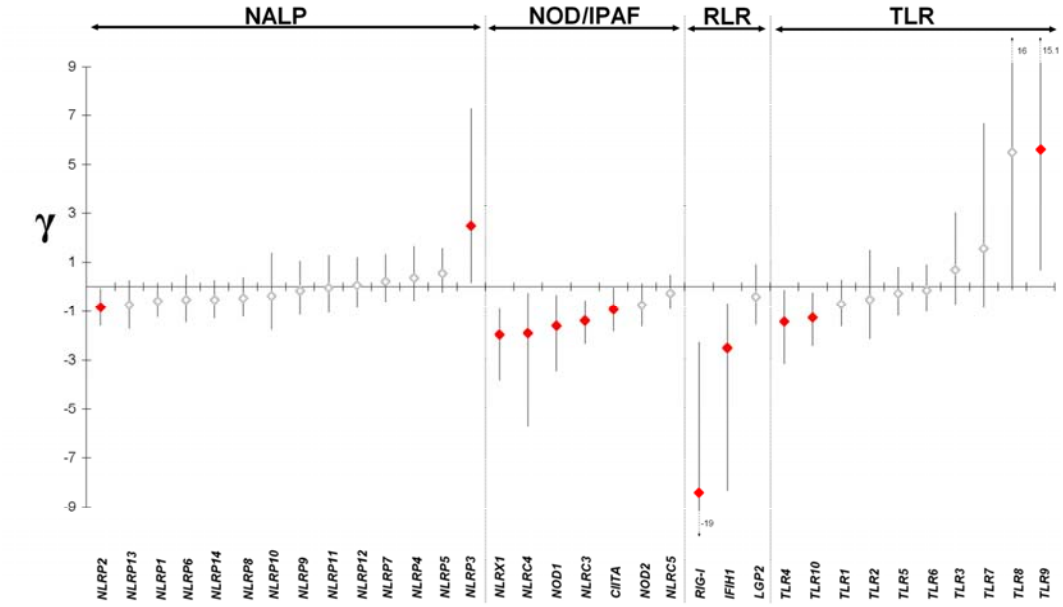


Figure S3. MKPRF results for the population selection coefficient γ. Bars indicate 95% CIs, and red diamonds indicate genes with γ estimates significantly higher/lower than 0.

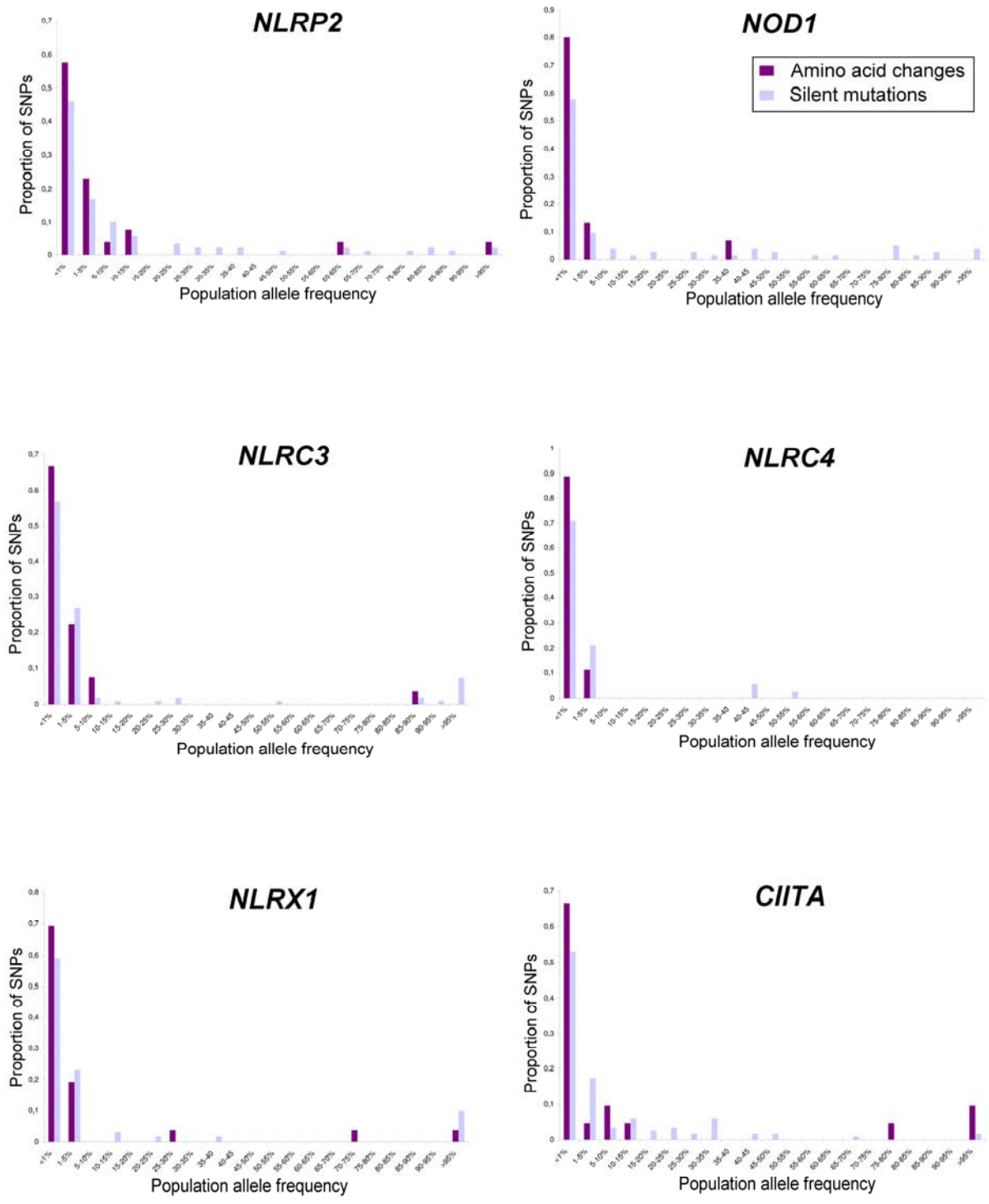


Figure S4. Allele frequency spectrum for non-synonymous and silent mutations. We present here the allele frequency spectrum for the genes that gave γ estimates significantly negative, which include *NLRP2*, *NOD1*, *NLRC3*, *NLRC4*, *NLRX1* and *CIITA*.

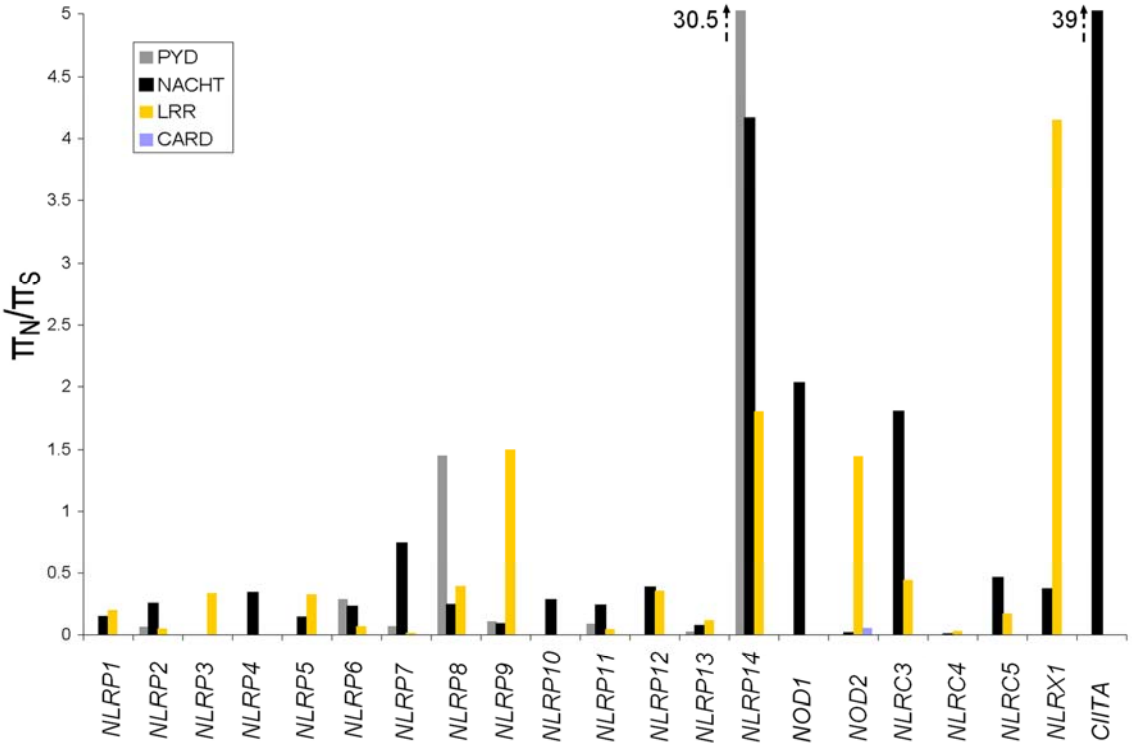
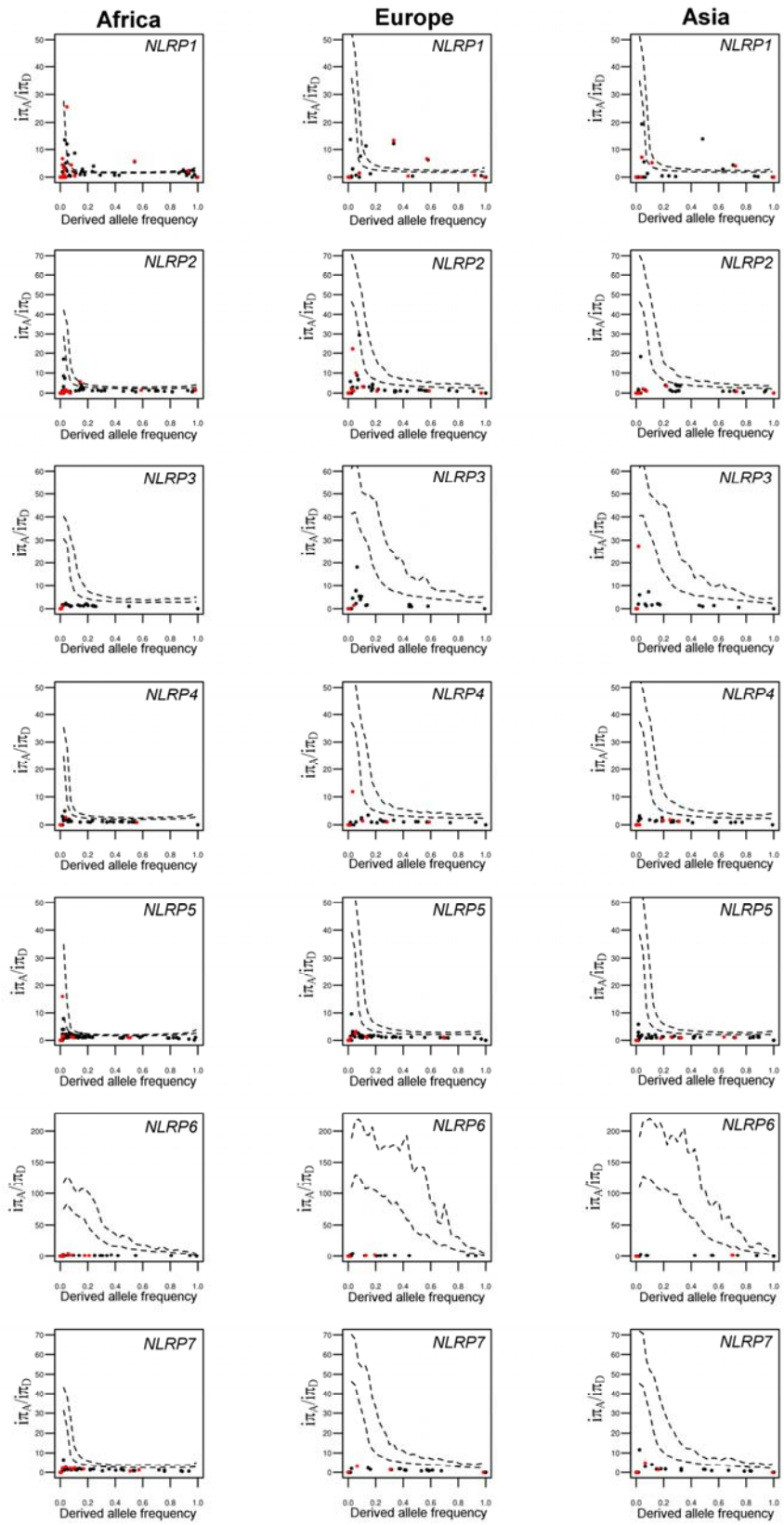
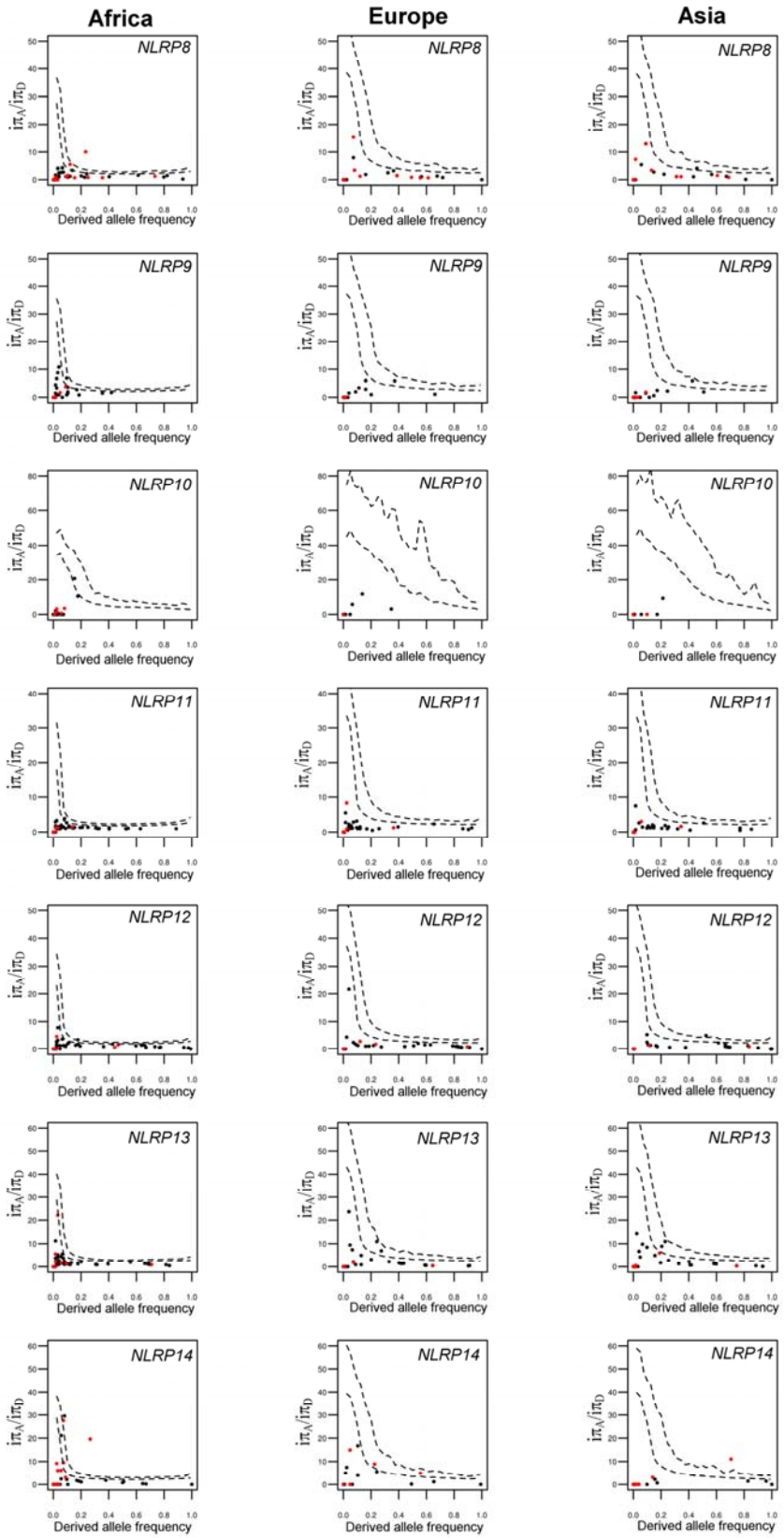


Figure S5. Nonsynonymous/synonymous diversity (π_N/π_S) across the different NLR protein domains. This ratio is based on the mean number of pairwise differences between chromosomes (π) at nonsynonymous and synonymous sites, for the 21 *NLRs* in the PYrin Domain (PYD), the NACHT domain, the LRR domain and the CARD domain. As LRR and NACHT domains are shared by all NLRs, the $\boldsymbol{\pi_N/\pi_S}$ ratios for these domains were assessed for each member of the NLR family. With respect to the PYD domain, $\boldsymbol{\pi_N/\pi_S}$ ratios were estimated for NALPs only. As to the CARD domain, we considered only *NLRP1*, *NOD1*, *NOD2* and *NLRC4*, which encode those NLRs for which this protein domain has been clearly localized in Uniprot (24).


Africa
Europe
Asia
NLRP1
NLRP2
NLRP3
NLRP4
NLRP5
NLRP6
NLRP7
iπA/iπD
Derived allele frequency


Africa
Europe
Asia
NLRP8
NLRP9
NLRP10
NLRP11
NLRP12
NLRP13
NLRP14
Derived allele frequency

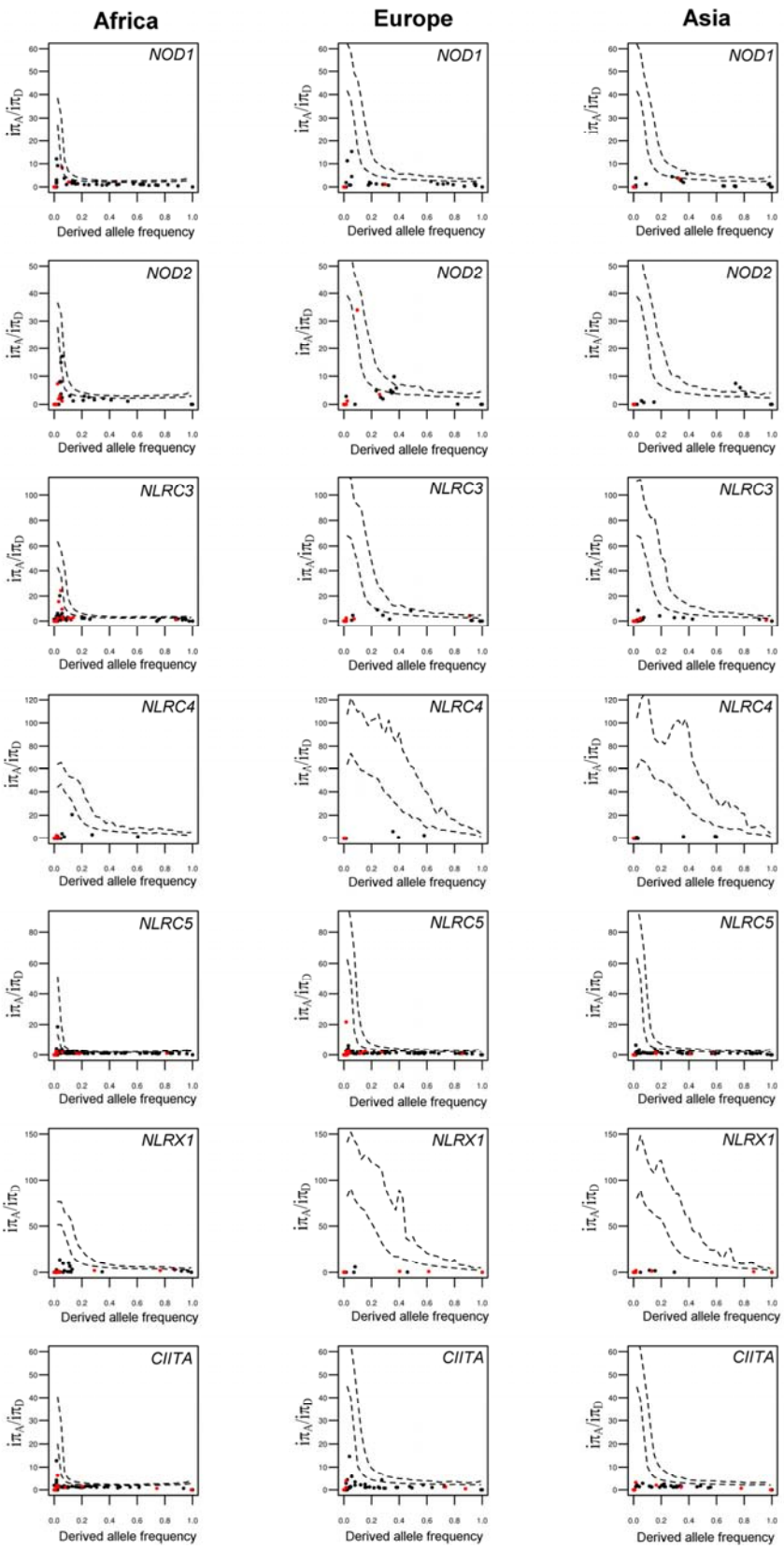

Africa
Europe
Asia
NOD1
NOD2
NLRC3
NLRC4
NLRC5
NLRX1
CIITA
Derived allele frequency

Figure S6. Detection of positive selection by the DIND test on the 21 genes encoding the NLRs in Africa, Europe and Asia. The upper dashed line on the graph corresponds to the 99th percentile, and the lower, to the 95th percentile. Black and red points represent silent and nonsynonymous SNPs, respectively. The results presented here were obtained with the Laval demographic model (18), which is more conservative, but the same outliers were obtained using the Voight model (19). Details on these analyses are given in the SI Material and Methods section.

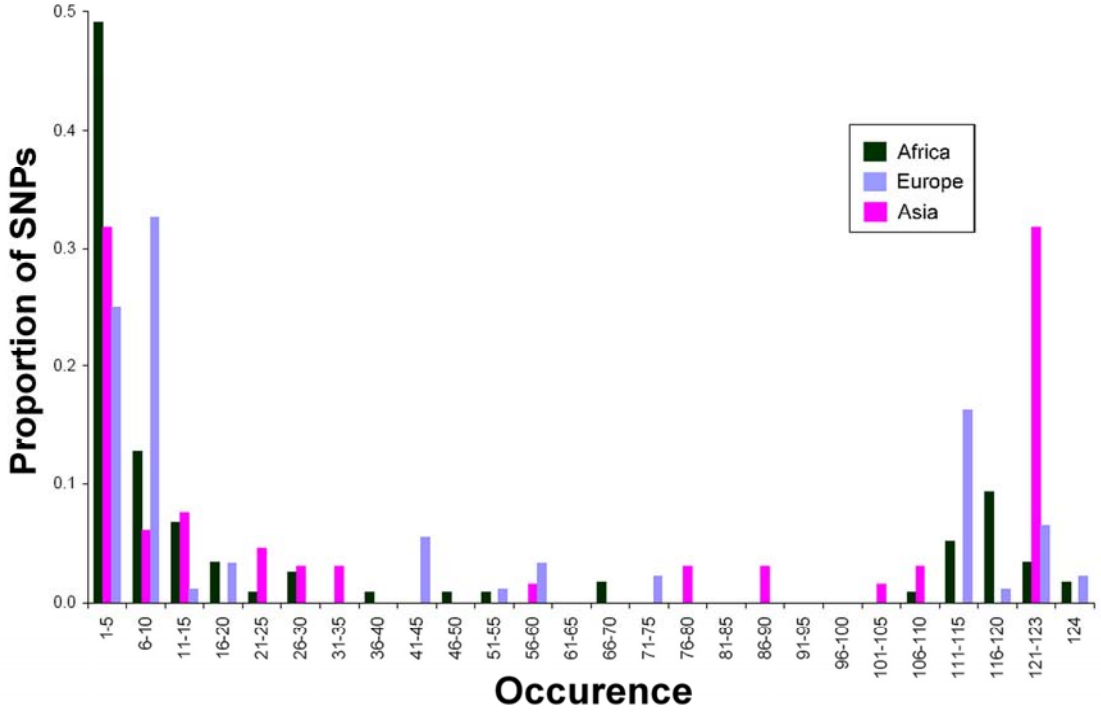


Figure S7. Allele frequency spectrum for *NLRP1* in sub-Saharan Africa, Europe and East-Asia. Note that in the case of Asia, because 122 chromosomes instead of 124 as for Africa and Europe, the bin 121-123 includes the fixed alleles in Asia. Overall, this graph highlights an excess of high frequency (or fixed) derived alleles in the three continents.

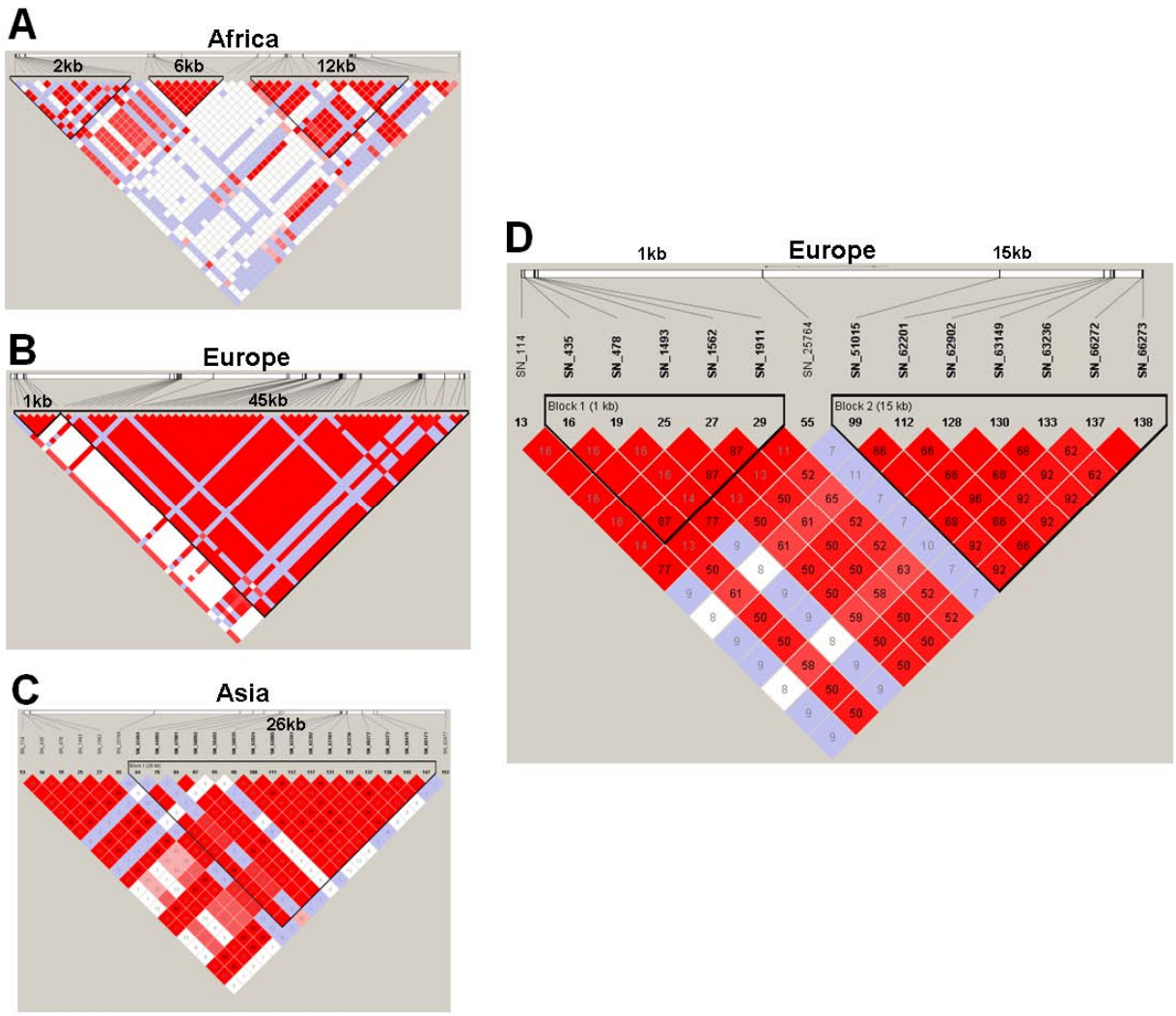


Figure S8. Levels of linkage disequilibrium (LD) assessed for sub-Saharan Africa, Europe and East-Asian in the *NLRP1* genomic region. Levels of LD, assessed using Haploview 4.0 (11), for mutations with a MAF > 0.01 in (*A*) Africa, (*B*) Europe and (*C*) Asia. In the case of Europe, LD is also represented for mutations with a MAF > 0.08 (*D*). A notable linkage disequilibrium was observed between SNP 51015 and SNP 1911 in Europe (r^2=0.52).

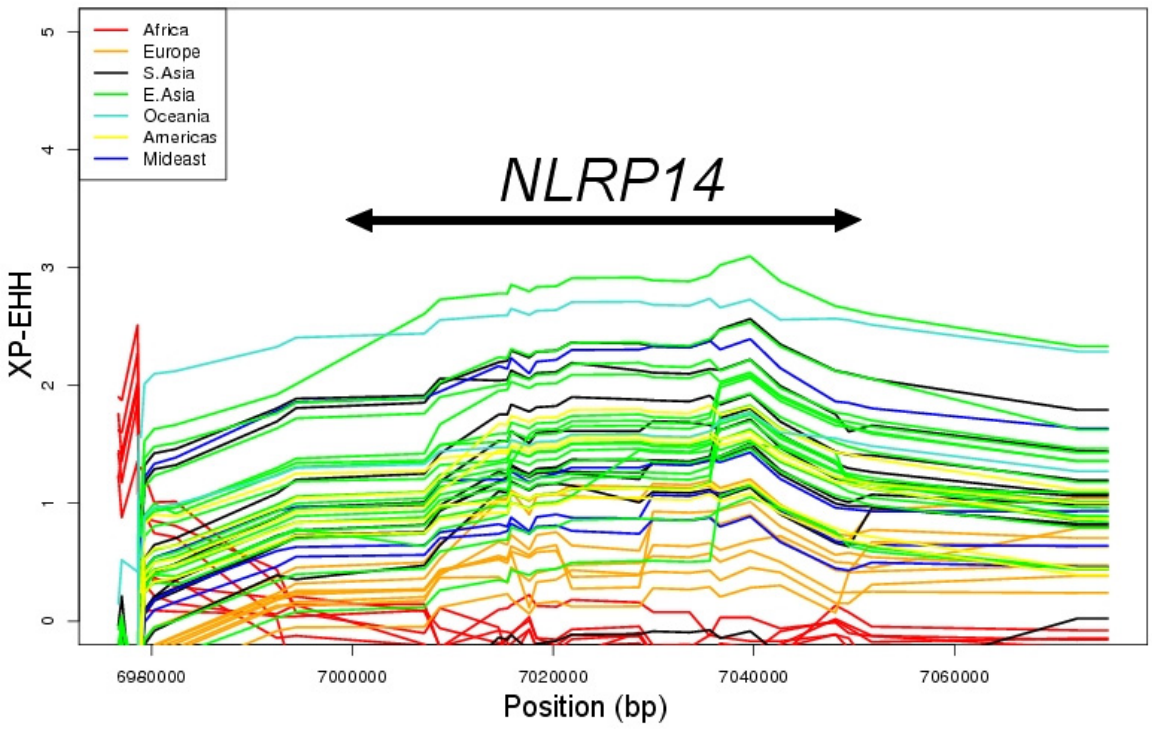


Figure S9. XP-EHH values across the *NLRP14* genomic region. This graph was obtained using the HGDP selection browser (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/). Genomic positions are given according to the hg18 (GRCh36) human assembly coordinates.

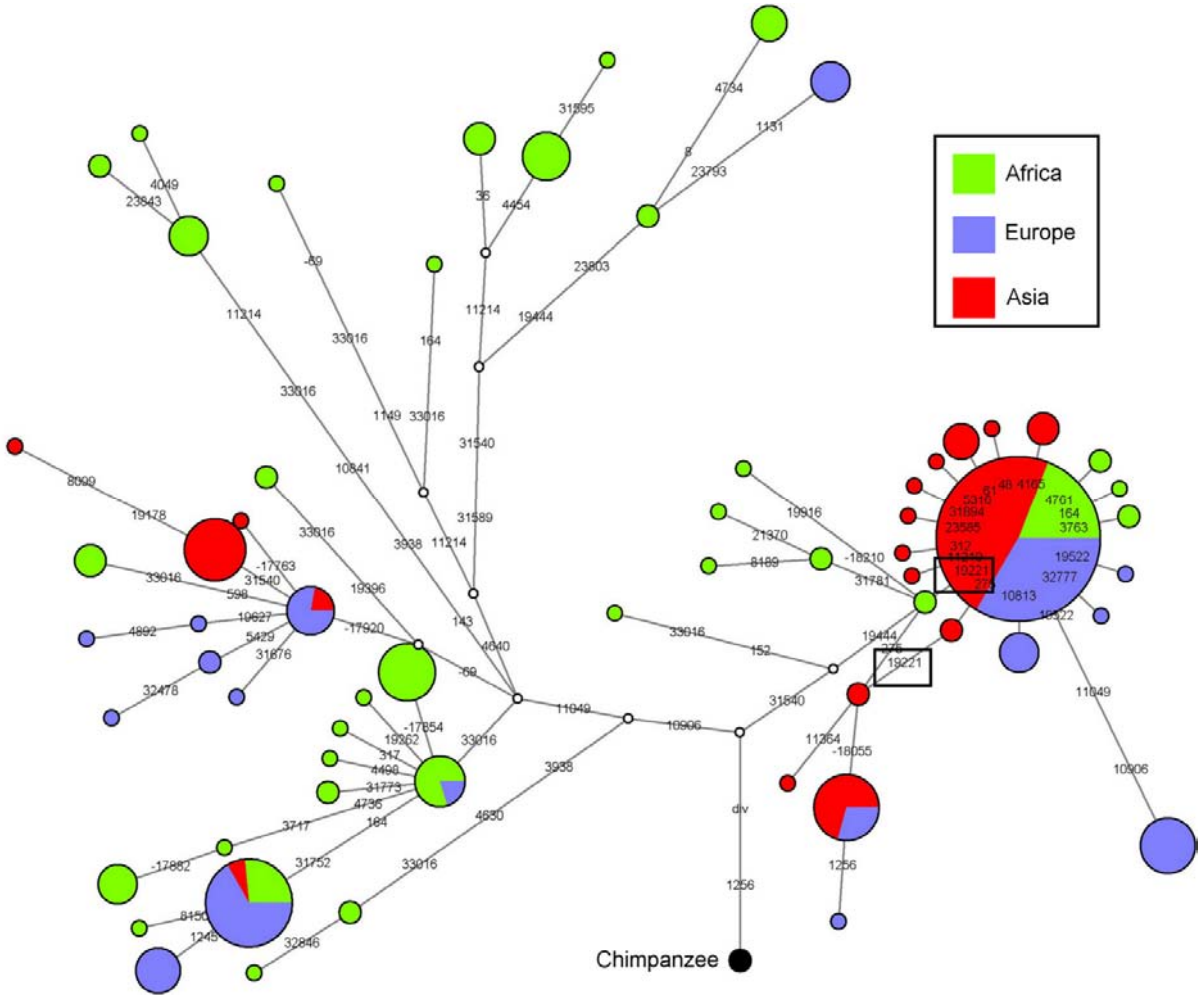


Figure S10. Median-joining network reconstructed from *NLRP14* haplotypes using coding SNPs only. Circles are proportional to haplotype absolute frequency; green areas represent the absolute frequency of the haplotype in Africans, blue in Europeans, and red in Asians. Mutation positions relative to *NLRP14* transcription start are shown. A star-like pattern starting from non-synonymous SNP 19221 (black framed) is observed, particularly in Asia.

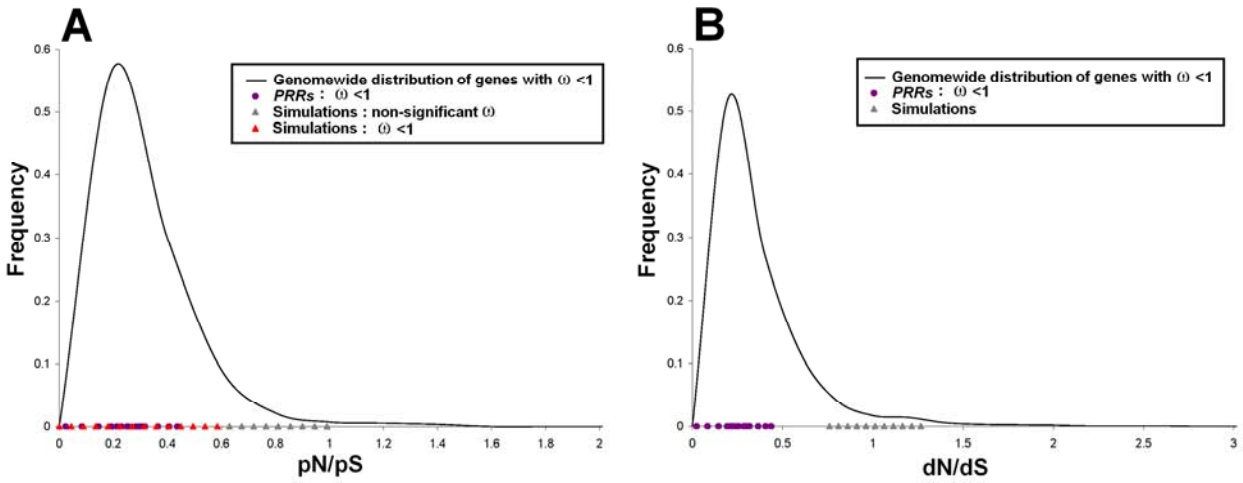


Figure S11. Location of (*A*) pN/pS and (*B*) dN/dS values of constrained *PRRs* in a genomewide distribution. To assess the contributions of divergence and polymorphism to the patterns of purifying selection observed at some genes, we compared the p_N/p_S and d_N/d_S ratios between *PRRs* with ω values significantly < 1 to a genomewide distribution of 1596 genes exhibiting ω values significantly < 1 (5). To evaluate the power of this method, we simulated a gradient of the various possibilities for pN/pS values, considering "neutral" dN/dS ratios (values around 1). We showed that genes with ω values significantly < 1 could result from "neutral" dN/dS values but very low pN/pS values. This suggests that we have enough power to assess the contributions of divergence and polymorphism to the patterns of purifying selection observed.

Table S1. Populations belonging to the HGDP-CEPH resequencing sub-panel.

Population	Geographical origin	Region	Number of individuals
Bantu	South Africa	sub-Saharan Africa	8
Bantu	Kenya	sub-Saharan Africa	11
Yoruba	Nigeria	sub-Saharan Africa	22
Mandenka	Senegal	sub-Saharan Africa	21
sub-Saharan African		***sub-Saharan Africa***	***62***
Adygei	Russia Caucasus	Europe	7
Russian	Russia	Europe	15
French	France	Europe	8
French Basque	France	Europe	12
Orcadian	Orkney Islands	Europe	6
North Italian	Italy (Bergamo)	Europe	3
Sardinian	Italy	Europe	11
European		***Europe***	***62***
Han	China	Asia	15
Dai	China	Asia	2
Lahu	China	Asia	2
Naxi	China	Asia	3
She	China	Asia	3
Yizu	China	Asia	2
Miaozu	China	Asia	4
Tujia	China	Asia	1
Tu	China	Asia	2
Xibo	China	Asia	4
Hezhen	China	Asia	3
Mongola	China	Asia	4
Daur	China	Asia	1
Oroqen	China	Asia	1
Cambodian	Cambodia	Asia	4
Japanese	Japan	Asia	10
Asian		***Asia***	***61***

Table S2. Resequenced regions in *NLR* genes. Chromosome location is given according to the hg19 (GRCh37) human assembly coordinates. The reference sequence that we used is also indicated here. Positions are given as relative to ATG position. Lengths are given in base pairs.

Gene	Chromosome location	Fragments sequenced	Total Length	Coding Length	Non-coding length
NLRP1	chr17:5,404,719-5,487,832 (isoform 5) NM_001033053.2	-615 : 504	11717	4419	7298
		1071 : 2299			
		23813 : 25764			
		30326 : 30685			
		41865 : 42512			
		44346 : 45080			
		46766 : 47214			
		49970 : 50744			
		50831 : 51289			
		53049 : 53527			
		61967 : 63258			
		66032 : 66592			
		68277 : 69386			
		82040 : 82588			
NLRP2	chr19:55,476,652-55,512,510 (isoform 2) NM_001174081.1	-58 : 650	7054	3186	3868
		4455 : 4735			
		7491 : 7926			
		11584 : 13736			
		14949 : 16440			
		19969 : 20709			
		24250 : 24614			
		27254 : 27753			
		30745 : 31122			
NLRP3	chr1:247,579,458-247,612,406 (isoform 3) NM_001079821.2	-3044 : -2599	6789	3108	3681
		-485 : 365			
		4278 : 6929			
		10632 : 11006			
		15255 : 15573			
		17112 : 17467			
		25029 : 25466			
		25753 : 26222			
		29568 : 30450			
NLRP4	chr19:56,347,944-56,393,220 NM_134444.4	-15649 : -15135	6474	2982	3492
		-121 : 352			
		5528 : 7235			
		9105 : 10101			
		15556 : 16032			
		18680 : 19203			
		24895 : 25498			
		26430 : 26956			
		29338 : 29986			

NLRP5	chr19:56,511,092-56,573,174 NM_153447.4	-510 : 111 3933 : 4453 8894 : 9204 16010 : 16275 19399 : 20128 20452 : 21060 27086 : 28831 32828 : 33375 33662 : 34066 38174 : 38553 41043 : 41404 50626 : 51052 53818 : 54171 58482 : 58918 61601 : 62245	8362	3600	4762
NLRP6	chr11:278,570-285,304 NM_138329.1	-214 : 4301 5660 : 6339 6425 : 6739	5510	2676	2834
NLRP7	chr19:55,434,877-55,458,873 NM_001127255.1 (isoform 3)	-6091 : -5752 -139 : 315 670 : 883 942 : 3724 5207 : 5519 6987 : 7223 7870 : 8243 10982 : 11252 13845 : 14135 17810 : 18412	5880	3111	2769
NLRP8	chr19:56,459,198-56,499,995 NM_176811.2	-347 : 497 4329 : 4811 6505 : 8305 13958 : 14401 18057 : 18551 22582 : 22948 25607 : 26093 28126 : 28506 31161 : 31662	5804	3045	2757
NLRP9	chr19:56,219,798-56,249,768 NM_176820.2	-523 : 399 4768 : 6440 8279 : 8626 14210 : 14559 21477 : 22117 22196 : 22627 22800 : 23381 25687 : 26630 29302 : 30064	6655	2973	3682
NLRP10	chr11:7,981,156-7,985,059 NM_176821.3	-679 : 635 1918 : 3912	3309	1965	1344

NLRP11	chr19:56,296,763-56,348,128 NM_145007.3	-18602 : -18111 -14816 : -13667 -133 : 352 7718 : 9700 9992 : 10421 16375 : 16866 21849 : 22099 25601 : 26086 28700 : 29425 32216 : 32657	6937	3099	3838
NLRP12	chr19:54,296,855-54,327,648 NM_144687.2 (isoform 2)	-348 : 304 9081 : 9344 12832 : 14650 15316 : 16106 16212 : 16714 18281 : 19062 19991 : 20637 22444 : 22993 25678 : 26042 28041 : 28392 29955 : 30808	7579	3183	4396
NLRP13	chr19:56,407,311-56,443,702 NM_176810.2	-29 : 499 7208 : 8423 18983 : 20961 21167 : 22007 24182 : 24594 27200 : 27524 30002 : 30372 33212 : 33554 36192 : 36484	6309	3129	3180
NLRP14	chr11:7,041,700-7,092,757 NM_176822.3	-18270 : -17667 -128 : 352 557 : 727 1083 : 1359 3690 : 5452 7984 : 8333 10809 : 11390 19091 : 19964 21228 : 21603 23560 : 24015 31525 : 31952 32368 : 33017	7011	3279	3732

NOD1	chr7:30,464,144-30,518,393 NM_006092	-22797 : -21828 -3320 :-2141 -239 : 278 1523 : 2016 3786 : 5807 8444 : 9097 9742 : 10409 10524 : 10840 19138 : 19750 20367 : 21019 23717 : 24449 27051 : 27653 31128 : 31676	10067	2859	7208
NOD2	chr16:50,731,050-50,766,987 NM_022162	-874 : 279 2174 : 2879 10489 : 10880 12979 : 15446 19269 : 19780 22360 : 22873 25273 : 25779 25902 : 26394 28200 : 28389 32506 : 32951 34126 : 24829	8010	3120	4890
NLRC3	chr16:3,589,038-3,627,392 NM_178844.2	-13370 : -11993 -201 : 2239 3030 : 3400 7221 : 7826 7966 : 8509 10247 : 10900 12679 : 12958 14352 : 14643 15231 : 16961 20631 : 20825 21314 : 22479 22585 : 23685	10758	3195	7563
NLRC4	chr2:32,449,518-32,490,801 NM_021209 (isoform1)	-9952 : -8717 -262 : 51 3873 : 4531 4948 : 7308 15529 : 16175 18189 : 18833 20351 : 20788 21100 : 21486 31974 : 32443	7147	3072	4075

NLRC5	chr16:57,023,410-57,117,436 NM_032206.3	-32250 : -31241	20740	5598	15142
		-190 : 389			
		1460 : 1652			
		2869 : 3515			
		4533 : 6710			
		7237 : 7725			
		8910 : 9433			
		10409 : 10861			
		12782 : 13631			
		15380 : 15709			
		16142 : 16604			
		18921 : 20264			
		20492 : 20897			
		21131 : 21710			
		22466 : 23140			
		24692 : 24835			
		25643 : 26086			
		26403 : 27092			
		30676 : 31094			
		33963 : 34526			
		34598 : 34950			
		37305 : 37580			
		38157 : 39108			
		40442 : 41051			
		44466 : 44880			
		45816 : 46027			
		46600 : 47186			
		48993 : 49408			
		49661 : 49980			
		53813 : 54065			
		55997 : 56350			
		56487 : 57376			
		58186 : 59071			
		60321 : 60942			
		61611 : 62233			
NLRX1	chr11:119039440-119054723 NM_024618.2 (isoform1)	-3512 : -2474	7057	2925	4132
		-151 : 400			
		886 : 1155			
		1243 : 1601			
		1939 : 4093			
		8209 : 8960			
		9538 : 10118			
		10594 : 11197			
		11621 : 12388			

CIITA	chr16:10971055-11018839 NM_000246.3	-1174 : 272	12070	3390	8680
		17624 : 18950			
		21225 : 21830			
		24075 : 24559			
		24605 : 25790			
		26159 : 26810			
		27129 : 27621			
		29048 : 31102			
		31447 : 32149			
		32746 : 33030			
		38102 : 39170			
		40893 : 41430			
		44576 : 45250			
		45645 : 46207			

Table S4. Mean diversity indices across the 21 genes encoding the NLRs.

	Africa (N=124)				Europe (N=124)				Asia (N=122)			
	S	**π**	**H**	**Hd**	**S**	**π**	**H**	**Hd**	**S**	**π**	**H**	**Hd**
NLRP1	116	13.257	53	0.975	90	13.525	31	0.879*****	46	7.658	33	0.863
NLRP2	73	11.305	66	0.979	63	11.685	51	0.958	50	9.26	47	0.941
NLRP3	34	5.198	64	0.981	30	4.697	39	0.952	21	4.074	30	0.926
NLRP4	61	9.163	98	0.995	40	8.049	61	0.975	36	8.648	61	0.982
NLRP5	110	15.934	100	0.995	58	12.82	86	0.991	60	11.694	89	0.993
NLRP6	49	6.905	85	0.992	34	5.204	62	0.969	22	4.459	38	0.929
NLRP7	65	12.874	85	0.993	35	9.474	36	0.943	37	7.959	40	0.96
NLRP8	54	8.41	48	0.956	34	7.889	30	0.912	33	7.488	23	0.874
NLRP9	45	4.424	41	0.953	25	3.133	26	0.868	26	3.254	25	0.872
NLRP10	29	2.157*****	25	0.89	10	1.54	10	0.748	7	1.284	7	0.718
NLRP11	57	8.87	94	0.993	39	5.531	66	0.978	38	7.542	60	0.974
NLRP12	68	9.152	66	0.981	40	8.605	48	0.948	36	8.177	28	0.828^{+}
NLRP13	77	11.195	88	0.99	40	9.636	40	0.932	37	9.738	37	0.92
NLRP14	46	6.13	34	0.92	27	5.086	17	0.8	24	2.716	18	0.656^{+}
NOD1	64	12.171	71	0.985	40	7.725	42	0.86*****	42	7.978	35	0.865
NOD2	49	4.962	43	0.959	35	5.75	29	0.87	23	1.883*****$^{/+}$	21	0.591*****$^{/++}$
NLRC3	81	8.096	70	0.987	58	4.956*****$^{/+}$	33	0.859	51	4.301*****$^{/+}$	35	0.908
NLRC4	31	2.58	30	0.847	12	1.595	15	0.621	16	1.762	20	0.68
NLRC5	175	24.969	112	0.998	112	21.762	100	0.995	113	19.79	110	0.998
NLRX1	62******	5.561	31	0.891*****	21******	2.626^{+}	20	0.709	19******	1.902^{+}	18	0.578*****$^{/+}$
CIITA	94	12.855	104	0.997	59	11.615	63	0.973	61	12.731	75	0.986

N, number of chromosomes sequenced in the corresponding population; S, number of segregating sites; H, number of haplotypes; π, nucleotide diversity per site from average pairwise differences; Hd, haplotype diversity. ******/*****/*P*-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Voight *et al*. (19); ++/+ *P*-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Laval *et al*. (18).

Table S5. Neutrality tests across *NLR* genomic regions

Gene	Africa					Europe					Asia				
	TD	**D***	**F***	**Hn**	**DH**	**TD**	**D***	**F***	**Hn**	**DH**	**TD**	**D***	**F***	**Hn**	**DH**
NLRP1	-1.24	0.08	-0.6	-2.193**	*	-0.608	-0.907	-0.933	-2.98*/+	NS	-0.324	-1.361	-1.119	-0.74	NS
NLRP2	-0.52	-0.59	-0.67	-0.75	NS	0.001	-0.836	-0.577	-0.36	NS	-0.013	-1.334	-0.941	-0.57	NS
NLRP3	-0.53	-1.81	-1.55	0.885	NS	-0.462	-1.191	-1.08	0.056	NS	0.123	-0.478	-0.296	0.259	NS
NLRP4	-0.6	-2.36	-1.94	0.538	NS	0.261	-1.282	-0.78	0.233	NS	0.884	-0.688	-0.061	-0.03	NS
NLRP5	-0.71	-0.92	-0.99	-0.3	NS	0.603	0.602	0.723	-0.29	NS	0.151	-0.359	-0.172	-0.74	NS
NLRP6	-0.75	-1.66	-1.53	-0.12	NS	-0.525	-2.769*/+	-2.236*/+	0.07	NS	0.258	-1.689	-1.124	-1.33	NS
NLRP7	0.216	-0.55	-0.27	-0.987*	NS	1.388	-0.457	0.347	-1.56	NS	0.476	-1.516	-0.848	-1.06	NS
NLRP8	-0.5	-0.43	-0.55	0.078	NS	0.757	-1.168	-0.469	0.224	NS	0.662	-0.917	-0.339	0.196	NS
NLRP9	-1.45	-1.45	-1.74	0.61	NS	-0.942	-2.131+	-1.997+	0.513	NS	-0.957	-1.211	-1.333	0.621	NS
NLRP10	-1.772*	-1.29	-1.78	0.532	NS	-0.42	-0.823	-0.81	0.661	NS	-0.03	-0.617	-0.494	0.872	NS
NLRP11	-0.5	-0.93	-0.9	0.498	NS	-0.716	-0.783	-0.907	-0.28	NS	0.205	-1.137	-0.709	0.275	NS
NLRP12	-0.87	-1.73	-1.63	-1.739**	NS	0.489	-0.432	-0.065	-2.072*	NS	0.671	-1.604	-0.82	-1.994*	NS
NLRP13	-0.69	-1.28	-1.23	-0.15	NS	0.914	-0.148	0.344	-0.59	NS	1.261	0.88	1.242	-0.87	NS
NLRP14	-0.87	-1.12	-1.22	0.47	NS	0.047	0.018	0.035	-0.19	NS	-1.134	-2.256+	-2.18*/+	-0.59	NS
NOD1	0.081	-0.98	-0.64	-0.2	NS	0.127	-0.715	-0.444	-1.39	NS	0.066	-3.024*/+	-2.107*/+	-1.44	NS
NOD2	-1.41	-3.6	-3.22	0.209	NS	-0.344	-3.278*/+	-2.509*/+	0.032	NS	-1.614*/+	-4.516**/++	-4.063**/++	-1.28	*
NLRC3	-1.47	-1.1	-1.52	-1.097*	**	-1.692*/+	-2.151	-2.35*/+	-2.344*	**	-1.704*/+	-1.737	-2.069+	-1.58	**
NLRC4	-1.64	-2.13	-2.32	0.383	NS	-0.73	-3.746**/+	-3.166**/+	0.199	NS	-1.112	-3.239*/+	-2.924*/+	0.032	NS
NLRC5	-0.76	-1.98	-1.72	0.05	NS	0.155	0.357	0.322	-0.52	NS	-0.189	-2.374*/+	-1.697*/+	-0.81	NS
NLRX1	-1.63	-1.69	-2	-1.225*	**	-0.925	-4.097**/+	-3.447**/+	0.273	NS	-1.295	-3.1*/+	-2.889*/+	0.021	NS
CIITA	-0.84	-1.76	-1.63	0.072	NS	0.194	-0.412	-0.187	-0.22	NS	0.385	-2.136+	-1.284	0.129	NS

TD, Tajima's *D*; D*, Fu & Li's *D**; F*, Fu & Li's *F**; Hn, Normalized Fay & Wu's *H*; DH, Tajima's D and Fay & Wu's H *p*-values combined. **/**P*-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Voight *et al*. (19); ++/+ *P*-values ≤ 0.01 and ≤ 0.05, respectively, according to the model of Laval *et al*. (18). Most of the members of the NOD/IPAF subfamily gave strongly negative values of Fu & Li's *D** and *F**, indicative of an excess of singletons, which was significant in Europe and Asia, in most cases. However, most gave no other significant signature of positive selection, suggesting that these patterns are the consequence of weak negative selection, as attested by significant results obtained by the MKPRF test (Fig. S3).

Table S6. SNPs included in the putatively selected *NLRP1* haplotype. The derived allele frequencies (DAF) and iHS values of *NLRP1* SNPs included in the haplotype are presented. In Europe, it consists in a 45kb haplotype block. Most of those SNPs were found in the corresponding 12kb or 6kb African haplotype blocks. Overall, 15 of those SNPs showed high frequency derived states (whose the majority reached fixation in Asia) (white rows) and the remaining 29 SNPs exhibited low frequency derived states (or were mostly absent from the Asian population studied) (grey rows). Non-synonymous mutations are in bold.

ATG position	Mutation type	AA POSITION	Mutation	Afr DAF	Eur DAF	Asi DAF	dbSNP	Afr iHS	Eur iHS	Asi iHS
23999	**NON-SYNONYMOUS**	**A246G**	**C>G**	**2,42**	**8,06**	**0**	**rs11651595**	**-**	**2,460**	**-**
24624	SYNONYMOUS	-	C>T	10,48	8,06	0	rs2001363	2,091	2,460	-
25143	SYNONYMOUS	-	G>A	89,52	91,94	100	rs12950235	-1,187	-1,750	-
25383	SYNONYMOUS	-	G>T	10,48	8,06	0	rs61753137	NA	NA	NA
25431	SYNONYMOUS	-	C>T	10,48	8,06	0	rs61753136	NA	NA	NA
25607	**NON-SYNONYMOUS**	**T782S**	**C>G**	**10,48**	**8,06**	**0**	**rs52795654**	**NA**	**NA**	**NA**
25688	NON-CODING	-	G>A	89,52	91,94	100	rs71368572	NA	NA	NA
30676	NON-CODING	-	C>T	10,48	8,06	0	rs11656080	NA	NA	NA
41964	SYNONYMOUS	-	C>T	2,42	8,06	0	rs12942931	-	2,535	-
42035	**NON-SYNONYMOUS**	**T878M**	**C>T**	**2,42**	**8,06**	**0**	**rs11657747**	**-**	**2,535**	**-**
42284	NON-CODING	-	C>T	2,42	8,06	0	rs11078570	-	2,535	-
44354	NON-CODING	-	G>A	2,42	8,06	0	rs7223362	-	2,346	-
44790	NON-CODING	-	C>T	2,42	8,06	0	rs12150292	NA	NA	NA
44791	NON-CODING	-	A>G	2,42	8,06	0	rs12150032	NA	NA	NA
44840	NON-CODING	-	C>T	2,42	8,06	0	rs12150289	NA	NA	NA
44921	NON-CODING	-	T>C	97,58	91,94	100	rs12150286	NA	NA	NA
44965	NON-CODING	-	G>A	97,58	91,94	99,18	rs12150026	NA	NA	NA
44992	NON-CODING	-	G>A	86,29	91,94	86,07	rs12150023	NA	NA	NA
46963	NON-CODING	-	A>C	2,42	8,06	0	rs7216010	NA	NA	NA
47030	SYNONYMOUS	-	G>A	99,19	91,94	100	rs11657249	-	-1,421	-
47057	SYNONYMOUS	-	G>A	99,19	91,94	100	rs7215868	NA	NA	NA
47081	SYNONYMOUS	-	G>A	91,13	91,94	88,52	rs7215856	NA	NA	NA
49993	**NON-SYNONYMOUS**	**I995T**	**T>C**	**93,55**	**91,94**	**100**	**rs34733791**	**NA**	**NA**	**NA**
50092	NON-CODING	-	A>G	89,52	91,94	88,52	rs35568355	NA	NA	NA
50161	NON-CODING	-	G>A	2,42	8,06	4,92	rs35548793	NA	NA	NA
50185	NON-CODING	-	G>T	0,00	8,06	1,64	-	NA	NA	NA
50210	NON-CODING	-	A>G	6,45	8,06	0	rs35913493	NA	NA	NA
50258	NON-CODING	-	C>T	2,42	8,06	0	rs71368558	NA	NA	NA
50371	NON-CODING	-	C>T	6,45	8,06	0	rs12943068	NA	NA	NA
50413	NON-CODING	-	C>T	2,42	8,06	0	rs71368557	NA	NA	NA
50469	NON-CODING	-	C>A	93,55	91,94	100	rs71368556	NA	NA	NA
50712	NON-CODING	-	G>A	2,42	8,06	0	rs35291181	NA	NA	NA
53186	NON-CODING	-	G>A	2,42	8,06	0	rs12940290	NA	NA	NA
53312	**NON-SYNONYMOUS**	**V1119M**	**G>A**	**93,55**	**91,94**	**100**	**rs35596958**	**NA**	**NA**	**NA**
53437	SYNONYMOUS	-	C>T	93,55	91,94	100	rs56872041	NA	NA	NA
62287	SYNONYMOUS	-	C>T	7,26	8,06	0	rs11653580	-0,310	2,051	-
62372	**NON-SYNONYMOUS**	**L1241V**	**C>G**	**92,74**	**91,94**	**100**	**rs11653832**	**0,829**	**-1,375**	**-**
62506	NON-CODING	-	C>A	1,61	8,06	0	rs11653496	-	1,630	-
62521	NON-CODING	-	A>G	7,26	8,06	0	rs11650176	0	2,381	-
62574	NON-CODING	-	A>G	7,26	8,06	0	rs11650171	0	2,381	-
62671	NON-CODING	-	C>T	7,26	8,06	0	rs11653450	-0,344	2,381	-
68479	**NON-SYNONYMOUS**	**R1366C**	**C>T**	**3,23**	**8,06**	**11,48**	**rs2137722**	**-**	**2,914**	**0,149**
69210	NON-CODING	-	G>C	96,77	91,94	100	rs34156685	NA	NA	NA
69256	NON-CODING	-	T>C	3,23	8,06	0	rs34211727	NA	NA	NA

Values of iHS < -2 represent signatures of positive selection on the derived allele, while of iHS > 2 indicate positive selection on the ancestral allele. Note however that in both cases, the real target of selection could be another variant (ancestral or derived) in high linkage disequilibrium, but for which the iHS value is not available in the HapMap Phase II database.

Other Supporting Information File

Table S3 (xls)

References for SI

1. Cann HM, *et al.* (2002) A human genome diversity cell line panel. *Science* 296(5566):261-262.
2. Boniotto M, *et al.* (2011) Population variation in NAIP functional copy number confers increased cell death upon /Legionella pneumophila/ infection. *Hum Immunol (in press).*
3. Takahashi M, Matsuda F, Margetic N, & Lathrop M (2003) Automated identification of single nucleotide polymorphisms from sequencing data. *J Bioinform Comput Biol* 1(2):253-265.
4. Rozas J, Sanchez-DelBarrio JC, Messeguer X, & Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19(18):2496-2497.
5. Bustamante CD, *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153-1157.
6. Sawyer SA & Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161-1176.
7. Vasseur E, *et al.* (2011) The selective footprints of viral pressures at the human RIG-I-like receptor family. *Hum Mol Genet* 20(22):4462-4474.
8. Barreiro LB, *et al.* (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet* 5(7):e1000562.
9. Laval G & Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20(15):2485-2487.
10. Stephens M & Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73(5):1162-1169.
11. Barrett JC, Fry B, Maller J, & Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263-265.
12. Sabeti PC, *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832-837.
13. Mueller JC & Andreoli C (2004) Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. *Bioinformatics* 20(5):786-787.
14. Sabeti PC, *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913-918.
15. Voight BF, Kudaravalli S, Wen X, & Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3):e72.
16. Frazer KA, *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851-861.
17. Excoffier L, Smouse PE, & Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131(2):479-491.
18. Laval G, Patin E, Barreiro LB, & Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5(4):e10284.
19. Voight BF, *et al.* (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* 102(51):18508-18513.
20. Fujita PA, *et al.* (2010) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 39:D876-882.

21. Li JZ, *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100-1104.
22. Pickrell JK, *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19(5):826-837.
23. Bandelt HJ, Forster P, & Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16(1):37-48.
24. Consortium TU (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37:D169-174.

ANNEXE 4 : Autre publication

Evolution of the TIR Domain-Containing Adaptors in Humans: Swinging between Constraint and Adaptation

Simona Fornarino,[1,2] Guillaume Laval,[1,2] Luis B. Barreiro,[3] Jeremy Manry,[1,2] Estelle Vasseur,[1,2] and Lluis Quintana-Murci*[,1,2]

[1]Human Evolutionary Genetics, Department of Genomes and Genetics, Institut Pasteur, Paris, France

[2]Centre National de la Recherche Scientifique, URA3012, Paris, France

[3]Department of Pediatrics, Sainte-Justine Hospital Research Center, Université de Montréal, Montréal, Canada

***Corresponding author:** E-mail: quintana@pasteur.fr.

Associate editor: Sarah Tishkoff

Abstract

Natural selection is expected to act strongly on immune system genes as hosts adapt to novel, diverse, and coevolving pathogens. Population genetic studies of host defense genes with parallel functions in model organisms have revealed distinct evolutionary histories among the different components—receptors, adaptors, and effectors—of the innate immune system. In humans, however, detailed evolutionary studies have been mainly confined to the receptors and in particular to Toll-like receptors (TLRs). By virtue of a toll/interleukin-1 receptor (TIR) domain, TLRs activate distinct signaling pathways, which are mediated by the five TIR-containing adaptors: myeloid differentiation factor-88 (MyD88), myeloid differentiation factor-88 adaptor-like protein (MAL), toll/interleukin-1 receptor domain-containing adaptor protein inducing interferon (IFN)β (TRIF), toll/interleukin-1 receptor domain-containing adaptor protein inducing IFNβ-related adaptor molecule (TRAM), and sterile α- and armadillo motif-containing protein (SARM). Here, we have examined the extent to which natural selection has affected immune adaptors in humans, using as a paradigm the TIR-containing adaptors. To do so, we characterized their levels of naturally occurring genetic variation in various human populations. We found that *MyD88* and *TRIF* have mainly evolved under purifying selection, suggesting that their role in the early stages of signal transduction is essential and nonredundant for host survival. In addition, the adaptors have been targeted by multiple episodes of positive selection, differing in timing and spatial location. *MyD88* and *SARM* display signatures of a selective sweep that has occurred in all humans, whereas for the other three adaptors, we detected signatures of adaptive evolution that are restricted to specific populations. Our study provides evidence that the contemporary diversity of the five TIR-containing adaptors results from the intermingling of different selective events, swinging between constraint and adaptation.

Key words: natural selection, immunity, adaptors, host-pathogen interactions, TIR domain.

Introduction

Throughout evolution, animals and plants have developed complex immune defense mechanisms to combat pathogens. The evolutionary dynamics of host–pathogen interactions lead to constant selection for adaptation and counteradaptation in the two competing organisms. Indeed, population genetics studies have shown that genes involved in immunity-related functions are privileged targets of natural selection, in its different forms and at different intensities (Allison 1954; Bustamante et al. 2005; Sabeti et al. 2006; Voight et al. 2006; Frazer et al. 2007; Nielsen et al. 2007; Barreiro et al. 2008; Barreiro and Quintana-Murci 2010). Any effective host defense system must have the capacity to recognize potentially pathogenic infection, signal the activation of the immune response, kill the infectious agent, and eventually, develop an immunological memory of the pathogen. Although the last two tasks are mainly fulfilled by the adaptive immune system, the first two are under the control of the innate immune system—the front-line defense system in plants, invertebrate animals, and vertebrate animals that allows immediate response to microbial assaults (Medzhitov and Janeway 2000; Janeway and Medzhitov 2002). Innate immunity relies on a diverse set of receptors referred to as pathogen- or pattern-recognition receptors, which detect molecular motifs shared by specific groups of microorganisms (i.e., pathogen-associated molecular patterns, PAMPs) (Akira et al. 2006). Upon microbial detection by the receptors, another component of the innate immune system, the adaptors, relays this information within the cell mediating signaling cascades that ultimately culminate in the induction of effector proteins. The effectors, in turn, coordinate the multifaceted innate responses to the threat that are also critical for the subsequent maturation of the adaptive immunity.

Among the different classes of innate immunity receptors, Toll-like receptors (TLRs) are the best studied from various angles (Akira and Takeda 2004; Kawai and Akira 2007; Casanova et al. 2011). They are characterized by an intracellular toll/interleukin-1 receptor (TIR) domain, an evolutionarily conserved amino acid sequence associated with defense against danger in both animals and plants (Gay and Keith 1991; Whitham et al. 1994; Lemaitre

et al. 1996; Tauszig et al. 2000; Rutschmann et al. 2002). In mammals, the TIR domain-containing proteins are organized into a superfamily that, along with the TLRs, comprises the interleukin 1 receptor, type I (IL1-RI)-like receptors and five cytoplasmic adaptor proteins. Although TLRs and IL1-RI-like receptors are involved in the direct detection of PAMPs and IL1-type ligands, respectively, the adaptors mediate the downstream signaling, culminating in the activation of transcription factors, such as nuclear factor-κB, activator protein-1 (AP-1), and members of the interferon regulatory factor family (Akira et al. 2006; O'Neill and Bowie 2007). Five TIR-containing adaptors have been identified so far: myeloid differentiation factor-88 (MyD88), myeloid differentiation factor-88 adaptor-like protein (MAL; also known as TIRAP), toll/interleukin-1 receptor domain-containing adaptor protein inducing interferon (IFN)β (TRIF; also known as TICAM1), toll/interleukin-1 receptor domain-containing adaptor protein inducing IFNβ-related adaptor molecule (TRAM; also known as TICAM2), and sterile α- and armadillo motif-containing protein (SARM; also known as SARM1) (Beutler et al. 2004; O'Neill 2008). The adaptors are involved in two main signaling pathways: the MyD88- and the TRIF-dependent pathways, according to the name of the two master adaptors (Kenny and O'Neill 2008). MyD88 is used by IL1-RI/IL-18R/ST2 and all TLRs, except TLR3, principally to induce the release of inflammatory cytokines, whereas TRIF is used by TLR3 and TLR4, principally to induce type I interferon production. MAL and TRAM function as sorting adaptors and are required to bridge the recruitment of MyD88 by TLR2 and TLR4 and of TRIF by TLR4, respectively (Kenny et al. 2009). The fifth adaptor, SARM, is functionally conserved from arthropods to humans as a negative regulator of the TRIF-dependent pathway (Carty et al. 2006; Belinda et al. 2008). Moreover, in humans, SARM inhibits the MyD88 mitogen-activated protein kinase (MAPK)-mediated AP-1 activation (Peng et al. 2010).

Studies of the extent to which natural selection operates in humans can provide important insight into the mechanisms of host defense and delineate the biological relevance/redundancy of immunity-related genes for infection in natural settings (Quintana-Murci et al. 2007; Casanova et al. 2011). With respect to the different components of the innate immune system in humans, most population genetic studies so far have been focused on the TLRs (e.g., Barreiro et al. 2009; Wlasiuk and Nachman 2010). Based on the predominant form of selection acting on the various members of the human TLR family, it has been shown that human TLRs have evolved into two distinct evolutionary groups that differ in their biological relevance (Barreiro et al. 2009). By contrast, evolutionary and population genetics data for the TIR-containing adaptors are still fragmented, despite their having been extensively studied at the immunological level (for an extensive review, see O'Neill 2008 and references therein). Here, we examined the extent to which natural selection has affected the five TIR-containing adaptors in humans. Our aims were 1) to provide an overview of the levels of naturally occurring genetic variation of the five adaptors, 2) to understand from an evolutionary perspective the biological relevance of the functions fulfilled by the adaptor proteins, and 3) to compare the evolutionary history of the different components of the innate immune system (i.e., receptors vs. adaptors), using as a paradigm the TLRs and the TIR-containing adaptors. To this end, we resequenced the five adaptor genes in a set of human populations from diverse geographic origins. Our data indicate that the contemporary diversity of the five adaptors results from multiple episodes of selective constraint and adaptive evolution, which vary in time and space, and that these evolutionary forces have had a stronger impact on the adaptors than on TLRs.

Materials and Methods

Population Samples

Sequence variation for the five TIR-containing adaptors was assessed in 183 healthy individuals from sub-Saharan Africa ($n = 61$), Europe ($n = 61$), and East Asia ($n = 61$). The DNA samples represent a subset of the Human Genome Diversity Panel—Centre d'études du polymorphisme humain (HGDP-CEPH) panel (http://www.cephb.fr/en/cephdb/). The chimpanzee sequence used as an outgroup was downloaded from the University of California-Santa Cruz (UCSC: http:// genome.ucsc.edu/). When the corresponding chimpanzee sequences were not publicly available, orthologous regions were sequenced in one specimen of *Pan troglodytes verus*. This study was approved by the Institut Pasteur Institutional Review Board (no. RBM 2008.06).

Molecular Analyses

The entire coding region was resequenced for all the genes and untranslated regions (UTRs) (except for *SARM*, for which ~16% of the gene consists of UTRs, and we therefore considered only an amount of UTRs of similar length to the coding region). A variable percentage of nonexonic portions was also added, corresponding to about 1 kb upstream from the first transcribed exon (i.e., the putative promoter), some exon-flanking intronic sites, and some noncoding sites downstream from the last 3′-UTR exon (for *TRAM*, the putative promoter region was not included for technical reasons). Chromosomal locations, genomic borders, and reference sequences for each gene were obtained from the UCSC database. DNA was amplified using polymerase chain reaction (PCR) conditions specifically optimized for each system: The protocols are available on request. All the PCR products (maximum size of ~1.5 kb) were purified with an ExoSAP procedure and sequenced with the BigDye Terminator v.1.1 Cycle Sequencing Kit and an Applied Biosystems 3730*xl* DNA Analyzer. For sequencing, we designed primers binding within the amplified sequences. Genotypes were determined by visual inspection of sequence chromatograms with GENALYS software v. 3.3b (Takahashi et al. 2003). For quality control and to avoid allele-specific amplification, both PCR and sequencing reactions were

repeated with other primers when new mutations were identified in primer-binding regions. All singletons or ambiguous polymorphisms were systematically reamplified and resequenced. The numbering system adopted identifies the first nucleotide of the gene translation start site as +1 and the nucleotide immediately upstream as −1. The same primers were used to amplify chimpanzee sequences not available from databases.

Statistical Analyses

Estimation of Diversity Indices

Haplotypes were reconstructed from unphased genotypic data by means of the Bayesian method implemented in PHASE v. 2.1.1 (Stephens and Donnelly 2003). We applied the algorithm five times, using different randomly generated seeds, and consistent results were obtained across runs. Indels were not included in statistical analyses. To assess the ancestral state of each single nucleotide polymorphism (SNP), we compared the allelic form at each nucleotide position in humans with other primates (*P. troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*; UCSC and the National Center for Biotechnology Information [NCBI] databases—http://www.ncbi.nlm.nih.gov/). As an outgroup, we used a hypothetical ancestral sequence, namely, the genuine chimpanzee sequence or a modified form corresponding to the ancestral state if the chimpanzee carried an allelic form not present in humans. Evolutionary relationships between haplotypes were inferred from a median-joining network constructed with Network v. 4.5 (Bandelt et al. 1999). Summary statistics, such as the number of segregating sites, Watterson's θ, nucleotide diversity, and haplotype diversity were calculated with the DnaSP package v. 5.0 (Rozas et al. 2003).

Prediction of the Functional Impact of Nonsynonymous Mutations

The functional impact of nonsynonymous mutations was predicted using the Polyphen algorithm v2 HumDiv (Adzhubei et al. 2010). This method, which considers protein structure and/or sequence conservation information for each gene, has been shown to be the best predictor of the fitness effects of amino acid substitutions (Williamson et al. 2005). PolyPhen predicts a mutation to be "benign" (i.e., the protein is not markedly altered), "possibly damaging," or "probably damaging" (i.e., the replacement is potentially deleterious, decreasing fitness). In the case of the adaptors, protein structures were not provided. The protein domains of the various adaptors were identified by means of the SMART database (Letunic et al. 2006) (http://smart.embl-heidelberg.de/).

Detection of Selection

We first investigated the effects of natural selection considering both intraspecies polymorphism and interspecies divergence with the McDonald–Kreitman Poisson Random Field (MKPRF) test (Sawyer and Hartl 1992; Bustamante et al. 2005), by means of the resources of the Computational Biology Service Unit of Cornell University (http://cbsuapps.tc.cornell.edu/mkprf.aspx). This extended version of the MacDonald–Kreitman (MK) test (McDonald and Kreitman 1991) uses a Markov chain Monte Carlo algorithm for the Bayesian analysis of polymorphism and divergence data under a Poisson random field setting (Bustamante et al. 2002, 2005). Specifically, we used the hierarchical model of the MKPRF test (Bustamante et al. 2002), considering information for all the genes for which data are provided to infer selection for an individual locus. Nonsynonymous and silent sites in the adaptors were analyzed together with TLR data to provide a more general biological comparative framework (Barreiro et al. 2009). Two measurements of the selective pressure acting on a given gene were supplied as output: ω and γ. The statistic ω compares nonsynonymous to silent mutation rates, taking into account the ratio of polymorphism and divergence at nonsynonymous and silent sites. The statistic γ relies on the ratio of divergence and polymorphism at nonsynonymous sites.

Next, to gain insights into the selective events having occurred specifically within the human lineage, neutrality tests for deviations from neutral expectations in the allele frequency spectrum, such as Tajima's *D* statistic (Tajima 1989), Fu and Li's *D** and *F** (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000), were performed with the DnaSP package v. 5.0 (Rozas et al. 2003). In addition, we performed the DH test, a compound of Tajima's *D* statistic and Fay and Wu's *H* (Zeng et al. 2006).

To detect more recent positive selection events, we made use of the Derived Intra-Allelic Nucleotide Diversity (DIND) test (Barreiro et al. 2009). The DIND test compares the frequency of each derived allele with the ratio of the nucleotide diversity estimated for all haplotypes carrying the ancestral ($i\pi_A$) alleles to that for all the haplotypes carrying the derived ($i\pi_D$) alleles. We also used the integrated Haplotype Score (iHS) test, which measures the haplotype homozygosity associated with a given allele (Voight et al. 2006). The values and probabilities of the iHS statistic for each candidate SNP in the different populations were obtained from HapMap Phase-II data (Frazer et al. 2007) (http://www.hapmap.org/), for a 3 Mb region, with Haplotter (http://hg-wen.uchicago.edu/selection/haplotter.htm). Linkage disequilibrium was measured with Haploview (Barrett et al. 2005), by calculating the r^2 values for all pairwise comparison between SNPs, with a minimum allele frequency $\geq$ 0.05. In parallel, we performed a long-range haplotype analysis (Sabeti et al. 2002) using the HapMap Phase-II data (Frazer et al. 2007), by computing the extended haplotype homozygosity (EHH) by means of a web calculator (http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl). EHH is defined as the probability that two randomly chosen chromosomes carrying a tested core haplotype are homozygous at all SNPs for the entire interval from the core region to the distance *x*.

Population differentiation was estimated by calculating the F_{ST} statistic from the analysis of variance (Excoffier et al. 1992). We identified SNPs presenting extreme levels of population differentiation, by comparing the observed F_{ST} values for individual SNPs in the adaptors with the calculated F_{ST} distribution for ~659,000 SNPs genotyped for

the same set of individuals of the HGDP-CEPH panel (Li et al. 2008), considering SNPs presenting similar expected heterozygosity. Empirical *P* values were estimated as previously described (Barreiro et al. 2008).

Coalescent Simulations and Demographic Models

We corrected for mimicking effects of demography and selection on patterns of genetic diversity, by considering two previously validated demographic models (Voight et al. 2005; Laval et al. 2010). The rationale of this choice was that these two models estimated population genetics parameters using resequencing data (minimizing ascertainment biases due to SNP genotyping) from noncoding regions (minimizing the effect of natural selection when resequencing genes) in a panel of populations of Sub-Saharan African, European, and East Asian descent. In contrast with the model of Voight et al. (2005), which considered each continental population separately, the model of Laval et al. (2010) considered intercontinental population migration. We therefore compared the *P* values obtained in each model because migration has been shown to increase the false positive rates of some neutrality statistics (Li 2011). For each model, *P* values for the various neutrality tests and the DIND test were estimated from 10^4 coalescent simulations with SIMCOAL 2.0 (Laval and Excoffier 2004), under a finite-site neutral model and considering the recombination rate reported by UCSC (deCODE, NCBI 36/hg38). Each of the 10^4 coalescent simulations was conditional, depending on observed sample size and the number of segregating sites for each gene (Barreiro et al. 2009). We thus estimated the significance of these tests, incorporating these two demographic models into our neutral expectations, as previously reported (Quach et al. 2009).

Results

We resequenced the five TIR-containing adaptor genes in a panel of 183 healthy individuals originating from sub-Saharan Africa, Europe, and East Asia. Each DNA sample was sequenced over 25.5 kb, 60% of which corresponded to exonic regions and the rest comprising intronic and promoter regions (supplementary fig. S1 and table S1, Supplementary Material online). We identified 219 SNPs, 14 of which were nonsynonymous mutations, and 15 indels (supplementary table S2, Supplementary Material online). Indels were excluded from the analysis; they were 1–78 bp long and were found principally in noncoding regions. Following haplotype reconstruction, we identified 209 different haplotypes (supplementary table S3, Supplementary Material online). This data set was used to estimate a number of population genetic parameters and summary statistics reflecting various aspects of the data, including levels of divergence and polymorphism, allele frequency spectra, and haplotype structure.

Levels of Polymorphism across the Five Adaptors

We first investigated the genetic diversity of the adaptors, by estimating levels of nucleotide diversity (π) over the entire sequenced region for each gene, considering the

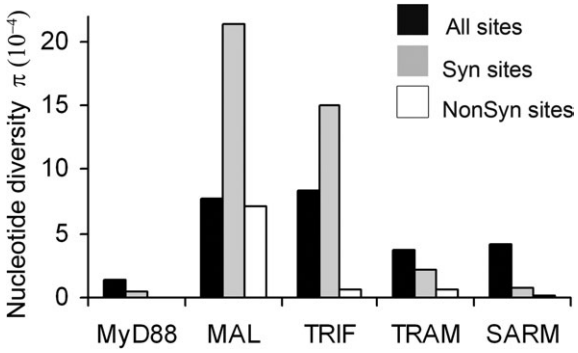


FIG. 1. Nucleotide diversity (π) values for the five human adaptors. π values for each gene were estimated considering the size of the entire sequenced region and the number of synonymous/nonsynonymous sites.

human population as a whole (fig. 1). *MyD88* was the least polymorphic adaptor, with *MAL* and *TRIF* displaying the greatest nucleotide diversity. We then compared the π values obtained for each adaptor with those of 326 genes sequenced by the Seattle SNP consortium (http://pga.gs.washington.edu/), excluding the Asian population from our results in order to make both data sets comparable in terms of population samples (supplementary table S4, Supplementary Material online). For *MyD88*, *TRAM*, and *SARM*, π values were less than half the mean value for the Seattle SNP database ($\pi = 8.59 \times 10^{-4}$). In particular, none of the Seattle genes had a π value lower than that for *MyD88* ($\pi = 1.52 \times 10^{-4}$), definitively highlighting the outlier position of this gene. We next compared patterns of nucleotide and haplotype diversity between the three continents. Both diversity indices tended to be higher in Africans than in non-Africans (table 1), as expected under the out of Africa hypothesis (Lewin 1987; Cavalli-Sforza and Feldman 2003). However, π values for *MyD88*, *TRIF*, and *SARM* were extremely uniform across populations, and π for *MAL* was much lower (2-fold lower) in Asia than in Africa and Europe.

We finally focused on the diversity of the coding region of each adaptor. Globally, synonymous were twice as frequent as nonsynonymous mutations, and no nonsynonymous mutations were observed in *MyD88* (fig. 1 and table 1). It is worth noting that, except for *MAL* that presented nonsynonymous polymorphisms at common frequencies, most nonsynonymous SNPs were found to segregate at very low frequency across the adaptors (supplementary table S2, Supplementary Material online). This suggests that, although their occurrence is tolerated, nonsynonymous polymorphisms are generally kept at low frequencies because of their likely deleterious effects. To gain further insight into the possible functional impact of these low-frequency nonsynonymous variants, we assessed their phenotypic consequences by means of the Polyphen algorithm v2 HumDiv (Adzhubei et al. 2010) and evaluated their position in the different protein domains. A considerable fraction of variants predicted to be possibly and/or probably damaging was identified (supplementary fig. S2, Supplementary Material online). Interestingly, however,

Table 1. Population Diversity Indices for the five Human Adaptors.

	Africa (chrs: 122)								Europe (chrs: 122)								Asia (chrs: 122)							
Gene	S^a	NC^b	Sy^c	NSy^d	Sg^e	θ^f	π^g	Hd^h	S^a	NC^b	Sy^c	NSy^d	Sg^e	θ^f	π^g	Hd^h	S^a	NC^b	Sy^c	NSy^d	Sg^e	θ^f	π^g	Hd^h
MyD88	16	15	1	0	10	5.43	1.82	0.50	7	7	0	0	3	2.37	1.17	0.42	7	6	1	0	4	2.37	1.13	0.53
MAL	43	37	1	5	15	15.81	8.76	0.90	21	15	3	3	5	7.72	8.20	0.81	29	23	3	3	12	10.66	4.41	0.66
TRIF	33	24	5	4	10	14.82	7.66	0.90	15	10	2	3	6	6.74	7.68	0.74	20	14	4	2	7	8.98	7.82	0.80
TRAM	23	20	1	2	4	9.72	5.12	0.85	11	11	0	0	5	4.65	3.00	0.70	9	9	0	0	4	3.80	2.48	0.61
SARM	33	27	3	3	20	9.52	4.60	0.83	15	13	1	1	6	4.33	4.03	0.75	20	16	1	3	10	5.77	3.56	0.53

NOTE.—chrs, chromosomes.
[a] Number of segregating sites (excluding indels),
[b] Number of SNPs in the noncoding region,
[c] Number of synonymous SNPs,
[d] Number of nonsynonymous SNPs,
[e] Number of singletons,
[f] Watterson θ per site ($\times 10^{-4}$),
[g] Nucleotide diversity per site ($\times 10^{-4}$),
[h] Haplotype diversity.

these variants tend to avoid the most critical domains (e.g., TIR domain) of the adaptor proteins (fig. 2 and supplementary table S2, Supplementary Material online). Altogether, these observations support the notion that nonsynonymous polymorphisms have mildly deleterious effects on individual fitness, which would account for their tolerance within the population.

Low Tolerance of the Accumulation of Amino Acid Variation Over Time

To determine the significance of the previous empirical observations and to formalize them in terms of selective constraints, we performed MK-type analyses, comparing polymorphism within humans with the divergence between human and chimpanzee at nonsynonymous and silent sites. We first used the MKPRF test (Sawyer and Hartl 1992; Bustamante et al. 2002, 2005) (see Materials and Methods). For all adaptors other than *MAL*, ω tended to be less than 1 (fig. 3*a*), with significant probabilities for *MyD88* and *TRIF*. Most nonsynonymous mutations at these two genes are thus eliminated from the population due to the likelihood of their being lethal or strongly deleterious and contribute to neither polymorphism nor divergence. This suggests a strong impact of purifying selection in driving the evolution of *MyD88* and *TRIF*. In turn, for all adaptors, except *MyD88*, γ tended to be lower than

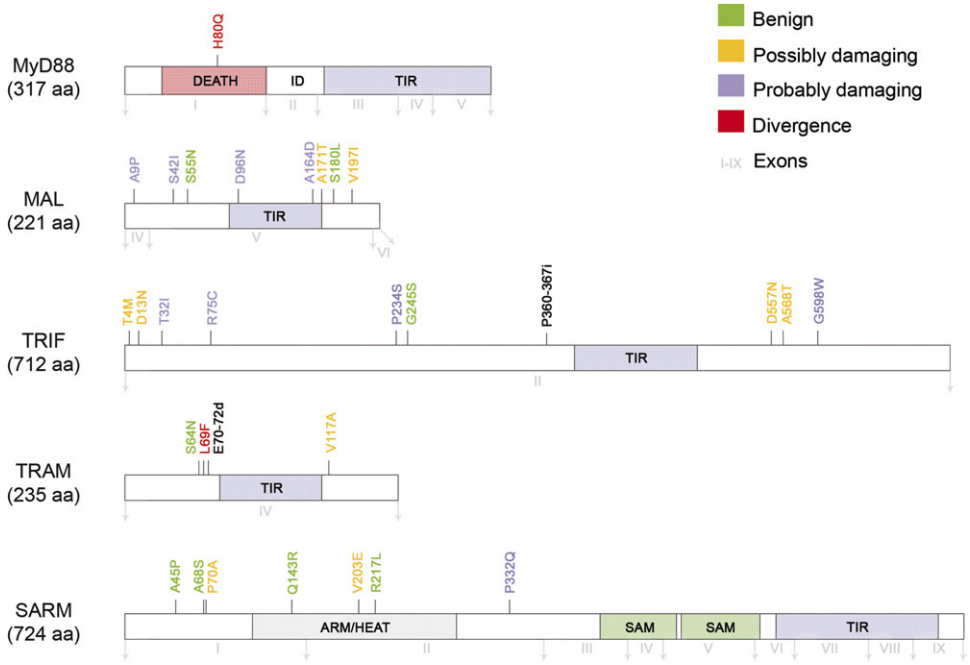


FIG. 2. Distribution of nonsynonymous variation across the five human adaptors. The functional impact of nonsynonymous SNPs was predicted using Polyphen v2 HumDiv (Adzhubei et al. 2010). Indel events occurring in the coding exons of *TRIF* (P360–367i) and *TRAM* (E70–72d) are also represented: They modify the number of proline and glutamic acid residues in the Pro- and Glu-rich portions, respectively, upstream from the TIR domain. The adaptors, like most proteins that participate in cellular signaling networks, are constructed in a cassette-like fashion and contain several domains for protein interaction, separated by hypothetical junk regions. The exact location of each nonsynonymous variant within the different protein domains is shown.

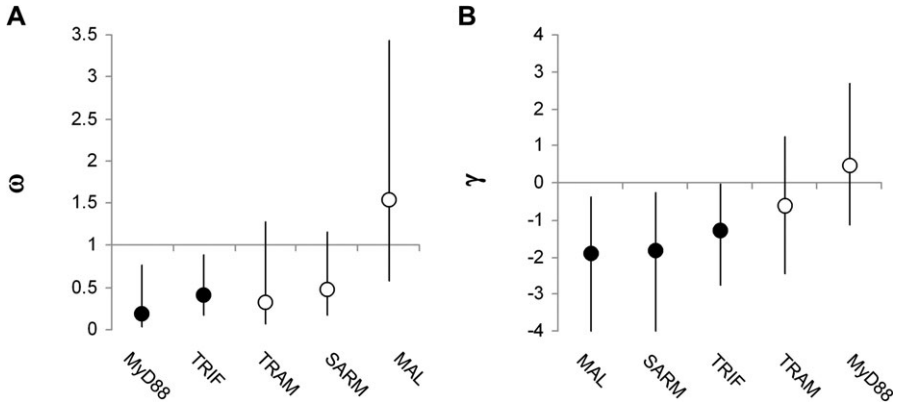


Fig. 3. Estimation of selection acting on the coding region of the five human adaptors. (*A*) Purifying selection acting on individual adaptors genes, as measured by estimated ω values (Bustamante et al. 2002). (*B*) Excess of nonsynonymous polymorphism with respect to divergence for individual adaptors genes, as measured by the population selection coefficient γ (Bustamante et al. 2002). Bars indicate 95% confidence intervals and filled circles indicate genes with ω and γ estimates significantly lower than 1 and 0 (neutral expectations), respectively. The number of nonsynonymous and silent substitutions contributing to divergence were: 3 (two of which occurred in the *Pan* lineage) and 49 for *MyD88*, 0 and 46 for *MAL*, 2 (both occurred in the *Pan* lineage) and 36 for *TRIF*, 1 and 42 for *TRAM*, and 0 and 39 for *SARM*, respectively.

0 (fig. 3*b*), and significantly, negative γ values were obtained for *MAL*, *SARM*, and *TRIF*. The pattern at *MAL*, *SARM*, and *TRIF* reflects a significant excess of nonsynonymous SNPs segregating in the human population with respect to divergence. This excess was also confirmed by the classical MK test (McDonald and Kreitman 1991), which revealed significant deviations from neutrality for the three genes (supplementary table S5, Supplementary Material online). To better clarify the nature of the selective pressures imposed on *MAL*, *SARM*, and *TRIF*, we focused on their allele frequency spectra considering silent and nonsynonymous mutations separately (supplementary fig. S3, Supplementary Material online). We observed that, for *TRIF* and *SARM*, the excess of nonsynonymous polymorphisms was specifically accounted for by low-frequency variants, suggesting that negative selection prevents them from increasing to high frequencies and eventually becoming fixed.

In summary, our results show that the TIR-containing adaptors have been subject to evolutionary constraints to varying degrees. Amino acid changes are virtually absent in *MyD88*. For *TRIF*, they are rarely tolerated and, as for *SARM*, they can only segregate at very low frequencies. The patterns of *MAL* deviate from neutral expectations but this cannot be ascribed to any specific form of natural selection, as highlighted by previous controversial observations (see e.g., Khor et al. 2007; Ferwerda et al. 2009; Nagpal et al. 2009). Finally, the patterns of *TRAM* do not deviate from neutral expectations.

Deviations of Allele Frequency Spectra from Neutral Expectations

We further tested for unexpected patterns of variation over the genomic regions encompassing the five adaptors, using several summary statistics for within-population allele frequency distribution. After correction for the mimicking effects of demography (see Materials and Methods), the hypothesis of neutrality was rejected, in at least one statistical test, for four of the five adaptors: MyD88, MAL, TRAM, and SARM (table 2). The most pronounced deviations from neutrality were those for *MyD88* and *SARM*. These genes clearly had low levels of nucleotide diversity, π, with respect to the number of segregating sites, and an excess of singletons, as attested to by significantly negative values of Tajima's *D* statistic (Tajima 1989) and/or Fu and Li's *D** and *F**

Table 2. Neutrality-Tests Results for the 5 Human Adaptors.

	Africa (chrs: 122)					Europe (chrs: 122)					Asia (chrs: 122)				
Gene	TD[a]	*D*[b]	*F*[c]	*H*[d]	DH[e]	TD[a]	*D*[b]	*F*[c]	*H*[d]	DH[e]	TD[a]	*D*[b]	*F*[c]	*H*[d]	DH[e]
MyD88	−1.81*,***	−3.78*,***	−3.64*,***	−0.11		−1.15	−1.51	−1.65	−0.87	*	−1.19	−2.41***	−2.37*,***	−1.32	*
MAL	−1.37	−1.85	−1.99	−0.10		0.19	−0.48	−0.27	−0.08		−1.74*,****	−2.35***	−2.53*,***	−0.43	
TRIF	−1.45	−1.24	−1.59	0.20		0.38	−1.79	−1.18	0.75		−0.37	−1.51	−1.29	−0.21	
TRAM	−1.36	0.13	−0.54	−0.63		−0.90	−2.01***	−1.92***	0.57		−0.85	−1.79*,***	−1.73***	0.20	
SARM	−1.55	−4.50*,***	−3.98**,****	−1.43*	**	−0.19	−1.79***	−1.43	−1.96*		−1.08***	−2.91**,****	−2.65*,****	−2.07*	***

Note.—Probability values were obtained from coalescent simulations, according to the models proposed by Voight et al. (2005), which considers each continental population separately, and Laval et al. (2010), which considers intercontinental population migration.

[a] Tajima's *D*.

[b] Fu and Li's *D**.

[c] Fu and Li's *F**.

[d] Fay and Wu's *H*.

[e] DH test probability.

*,**Indicates $P \leq 0.01$ and 0.05, respectively, according to the model of (Voight et al. 2005). ***,****Indicates $P \leq 0.01$ and 0.05, respectively, according to the model of (Laval et al. 2010).

statistics (Fu and Li 1993), respectively, in virtually all populations. The most parsimonious explanation for this is the occurrence of positive selection worldwide, which would be predicted to lead to low diversity, with an excess of rare alleles, over a selected genomic region. Because *MyD88* and *SARM* appear to be evolutionarily constrained, background selection (i.e., the selective removal of haplotypes bearing deleterious and linked neutral variation) (Charlesworth et al. 1993) may account for the observed deviations from neutrality. We therefore also carried out the DH test (Zeng et al. 2006), which has a low sensitivity to background selection and a high statistical specificity to selective sweeps. This test gave significant results for both *MyD88* and *SARM* (table 2), supporting the sweep hypothesis.

In addition, the Fay and Wu's *H* (Fay and Wu 2000) provided further insights into the nature of these sweeps. We found significant deviations from neutrality for this statistic specifically in the *SARM* gene (table 2). The Fay and Wu's *H* has considerable power to detect almost complete sweeps, with the signal rapidly falling off and disappearing after fixation of the advantageous mutation (Fay and Wu 2000; Zeng et al. 2006 and supplementary fig. S4, Supplementary Material online). However, it is worth noting that *H* was found to be significantly negative in all populations when the neutral simulations (see Materials and Methods) were performed using a model without migration (Voight et al. 2005) but lost significance when migration was considered (Laval et al. 2010). This confirmed the known effects of migration on the Fay and Wu's *H* statistic (Li 2011). Migration could therefore explain, at least partially, the negative Fay and Wu' *H* observed for *SARM* in each population independently. However, when computing for this gene, the joint probability of obtaining the observed negative Fay and Wu's *H* values in the three populations simultaneously, we found that, in the absence of selection, such a scenario is unlikely even in a neutral model considering migration (i.e., the joint *P* value under the model considering migration was equal to 0.037, see supplementary material S1, Supplementary Material online).

Taken together, our results are consistent with the occurrence of a worldwide selective sweep that is already complete for *MyD88* and not yet complete for *SARM*.

Increase in Haplotype Homozygosity in Specific Human Populations

We also searched for population-specific episodes of positive selection, through tests that make use of the low levels of intraallelic nucleotide diversity (the DIND test Barreiro et al. 2009) and elevated homozygosity (the iHS and EHH values, Sabeti et al. 2002; Voight et al. 2006) characteristic of recently selected genomic regions. Moreover, we estimated the degree of differentiation of the three populations, by calculating F_{ST} (Excoffier et al. 1992). Signatures of recent positive selection were identified for *MAL*, *TRIF*, and *TRAM*, in specific geographic areas. For *MAL*, the DIND test revealed that the haplotype harboring the derived state at three noncoding SNPs (nucleotide positions: −930, 2,073, and 2,823) displayed lower levels of nucleotide diversity than the other haplotypes, given its frequency (supplementary fig. S5*a*, Supplementary Material online). Indeed, despite being the most common haplotype in Europe, with a frequency of 27% (supplementary table S3, Supplementary Material online), it presented no internal variation, as illustrated in the phylogenetic network (supplementary fig. S6, Supplementary Material online). In addition, the derived state of the SNP 2,823 (no data were provided for SNPs −930 and 2,073) is characterized by extended levels of haplotype homozygosity among Europeans, as revealed by both the iHS (−2.43) and the EHH statistics (supplementary fig. S7, Supplementary Material online). These findings are thus consistent with the action of positive selection in Europe, which is likely targeting a long-range haplotype carrying SNP 2,823 and extending over the borders of *MAL*. For *TRIF*, the DIND test identified an SNP at nucleotide position 12 as being under positive selection in Europe (supplementary fig. S5*b*, Supplementary Material online). No HapMap Phase-II data were available for this variant, but it displayed significant levels of differentiation between African and European populations ($F_{ST} = 0.39$, $P < 0.05$) (supplementary fig. S8, Supplementary Material online). Finally, the DIND test for *TRAM* provided a significant value for SNP 864 in Asia (supplementary fig. S5*c*, Supplementary Material online). Furthermore, we observed that the EHH values associated with the derived allele at nucleotide position 864 decayed more slowly than those associated with the corresponding ancestral allele (supplementary fig. S9, Supplementary Material online). These results suggest that positive selection has been targeting a long-range haplotype that harbors SNP 864 and extends over the gene *TRAM*, in Asia.

Discussion

Depicting patterns of molecular evolution at host defense genes is a useful tool to predict the biological relevance of the specific functions performed by each component of the immune system. Results presented here show that, in humans, the adaptors as a group have been particularly targeted by natural selection, with respect to their corresponding receptors. In addition, this study reveals unique evolutionary histories for each of the five TIR-containing adaptors, increasing our understanding of the biological relevance of the signaling pathways they mediate.

Adaptors versus TLRs: Varying Selection on Two Components of the Same Defense Pathway

The major innate immunity pathways are largely conserved across large phylogenetic distances, but there is growing evidence to suggest that different components of the same pathway (i.e., receptors, adaptors, effectors) may not necessarily have followed the same evolutionary trajectories (Tiffin and Moeller 2006). In *Arabidopsis thaliana*, for example, most of the R-gene receptors are subject to balancing selection (Bakker et al. 2006), whereas adaptors are subject to selective constraint

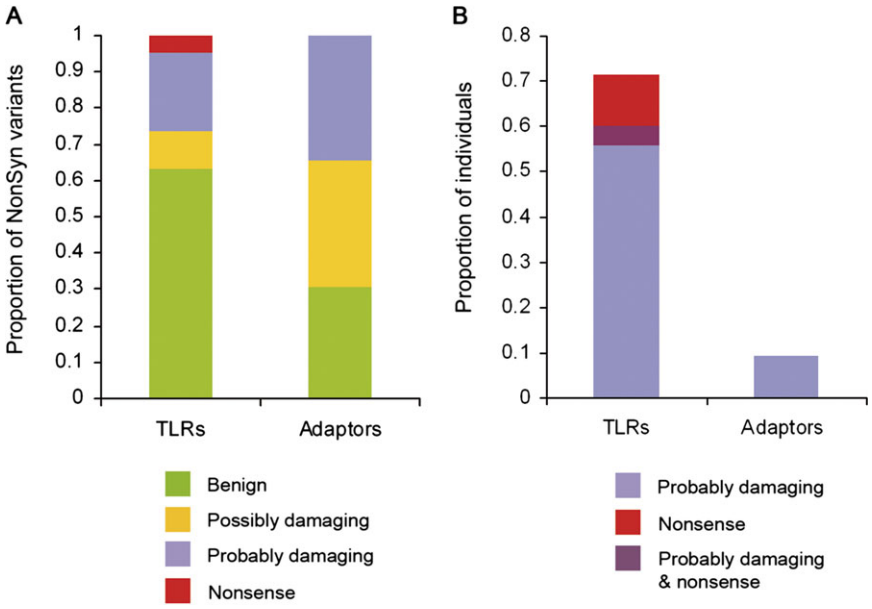


FIG. 4. Distribution of functional diversity at nonsynonymous variants for TLRs and adaptors. (*A*) Proportion of different classes of nonsynonymous SNPs. (*B*) Proportion of individuals in the general population carrying at least one probably damaging or/and nonsense mutations for the TLRs and the adaptors. The functional impact for each variant was predicted by Polyphen v2 HumDiv (Adzhubei et al. 2010).

(Bakker et al. 2008). In Drosophila, the adaptors display signals of lineage-specific positive selection at the amino acid sequence level (Lazzaro 2008). In primates, signatures of positive selection and constraint have been observed for most TLRs and the five adaptors, respectively, when evaluating the substitution rates across species (Nakajima et al. 2008; Wlasiuk and Nachman 2010), although the patterns of nucleotide variation within species are generally compatible with purifying selection (Wlasiuk and Nachman 2010). Combined episodes of constraint and positive selection have driven the evolution of human TLRs (Barreiro et al. 2009).

We show here that, in humans, the evolutionary dynamics of the adaptors reflects that of TLRs, but the strength of the various forms of selection differs between these two functional groups. The stronger selective constraint observed for the adaptors as compared with the TLRs suggests that there is a greater need to preserve the integrity of the proteins encoded by the adaptors. Indeed, although two of three nonsynonymous variants occurring in the adaptors are predicted to be "damaging" compared with one of three in the TLRs (fig. 4*a*), the frequency of possibly damaging mutations segregating in the general population is definitively higher for the TLRs (supplementary table S2, Supplementary Material online, Barreiro et al. 2009) as well as the proportion of individuals carrying probably damaging variants (fig. 4*b*). More importantly, no nonsense mutations were observed in any of the adaptors, whereas the occurrence and population frequency of stop mutations in the TLRs was nonnegligible (fig. 4). This observation clearly shows that the adaptor proteins, taken as a group, represent a more essential and nonredundant component of the human innate immune system than their corresponding receptors. In addition, the fact that events of positive selection are frequent among the adaptors but episodic for TLRs (Barreiro et al. 2009; Wlasiuk and Nachman 2010), reveal that the adaptors have been privileged targets of natural selection with respect to the TLRs in humans.

The MyD88- and TRIF-Dependent Pathways Have Been Selectively Constrained Over Time

The TIR-containing adaptors appear overall to be characterized by a substantial degree of selective constraint at the protein level. The MKPRF test revealed that *MyD88* has evolved under the strongest purifying selection, with respect to the other adaptors (fig. 3). The fact that purifying selection has been actively operating on *MyD88* suggests that the functions of the protein it encodes are essential and nonredundant. MyD88 is known as the "pan adaptor" because it orchestrates the innate and adaptive immune response by centralizing the multiple signals sensed by a great variety of receptors into a discrete number of signaling cascades (O'Neill and Bowie 2007). Specifically, MyD88 can serve 1) sensor functions, by mediating the early signal transduction following microbial detection; 2) instructive functions, by directing differentiation of T and B cells; and 3) effector functions, by triggering effective antimicrobial mechanisms (Reiling et al. 2008). These functions involve, but are not restricted to, TIR-containing proteins. MyD88 has been reported, for example, to be critical in transducing TIR-independent IFN-γ-mediated signals (Han 2006) and in controlling a TIR-independent pathway for immunoglobulin diversification (He et al. 2010). Does the signature of purifying selection observed at MyD88 reflects this broad metabolic role, regardless of the biological importance of the processes mediated by MyD88? Some MyD88-dependent processes may be individually

dispensable, but the pressures imposed by each of them on this shared adaptor, although modest, may contribute in a synergistic manner to the conservation of its protein sequence. Alternatively, the strong constraint of MyD88 may be linked to a few discernible individual processes of extreme biological relevance. So far, the specific immunological mechanism(s) in which MyD88 might play a key role remains unclear (von Bernuth et al. 2008). However, it is interesting to note that the TIR-dependent MyD88-dependent pathway is generally enriched for signals of purifying selection. Indeed, the regime observed for MyD88 perfectly mirrors the important selective constraints previously reported for four molecules operating upstream and immediately downstream from MyD88: TLRs 7-8-9 (Barreiro et al. 2009) and interleukin-1 receptor-associated kinase 4 (IRAK-4), respectively (L. Quintana-Murci, unpublished data), linking the evolutionary fate of interacting TIR-containing proteins.

From this perspective, compared with MyD88, the nature of the selective constraint depicted at the other master adaptor TRIF appears indeed much more limited, as TRIF is used only by TLR3 and TLR4 (O'Neill and Bowie 2007). TLR3, but not TLR4, has been shown to evolve under purifying selection, and clinical studies have shown that changes in TLR3 are involved in the pathogenesis of herpes simplex encephalitis (Zhang et al. 2007). Thus, here, evolutionary and clinical genetics data converge, suggesting that the entire pathway activated by TLR3 in a TRIF-dependent manner is nonredundant in host defense, at least against herpes simplex virus.

Adaptive Evolution at TIR-Containing Adaptors Attests for Advantageous Variation

The selective regime constraining the evolution of most adaptor proteins over time has not prevented these genes from undergoing episodes of positive selection for changes increasing fitness. For *MyD88*, neutrality tests support the occurrence of a worldwide selective sweep that has been completed (table 2), resulting in fixation of the target in humans. Comparisons of the coding sequences of the human MyD88 gene with those from various nonhuman primates (Nakajima et al. 2008) identified a single nonsynonymous substitution specific to the human lineage: H80Q. This amino acid substitution falls in the DEATH domain (fig. 2), which is critical for downstream protein–protein interactions and variations in this domain have been associated with defective immunity to infection (Isnardi et al. 2008; von Bernuth et al. 2008). This makes this human-specific variant an excellent candidate target for the observed selective sweep.

With respect to *SARM*, we propose that an almost complete worldwide selective sweep is the most likely explanation for the patterns observed. In this case, the selected derived allele would be expected to have a high frequency, but may not yet be fixed, in all populations. A plausible target accounting for this event is a haplotype defined by a 78-bp tandem duplication in the putative promoter of the gene (~500 bp upstream from the start codon), with a current worldwide frequency of ~80% (supplementary table S2, Supplementary Material online). Indeed, we observed a star-like shape of the haplotypes harboring this structural rearrangement in the *SARM* phylogenetic network (supplementary fig. S10, Supplementary Material online). Interestingly, the 78-bp duplication is associated with a number of expression quantitative trait loci recently identified in the 3′ region of *SARM* (Veyrieras et al. 2008) (http://eqtl.uchicago.edu/Home.html), suggesting that some component of selection targeting this haplotype could involve cis-epistatic interactions among different sites. Independent of the genuine target of selection, the selective advantage must therefore be traced back to changes in gene regulation.

Finally, haplotype-based tests all converge in pinpointing signatures of recent adaptation for *MAL* in Europe, *TRIF* in Europe, and *TRAM* in Asia (supplementary figs. S5–S7 and S9, Supplementary Material online), although statistical support was less robust than that obtained for *MyD88* and *SARM*. These signatures involve synonymous or noncoding sites, again suggesting the importance of regulatory variation in adaptors' evolution. In particular, population-specific positive selection may have recently targeted long-range haplotypes centered on *MAL* and *TRAM* but with a genomic span larger than the lengths of the two adaptor genes. This picture would then be consistent with a model of adaptation that occurs with modest advantageous changes, simultaneously across many *loci* (Pritchard and Di Rienzo 2010; Pritchard et al. 2010).

From a functional perspective, our study provides a number of potential targets of natural selection to be now experimentally tested. In vitro and ex vivo analyses will help to identify the real targets accounting for the signatures observed and delineate how they have affected the immunological mechanisms mediated by the adaptors.

Conclusion

The TIR-containing adaptors belong to a category of genes that fulfills a delicate role, driving a relay of warning signals potentially lethal for the cell. Such signals have been mostly attributed to pathogen presence because most of the biological information on the five adaptors at present concerns their role in innate immunity and translation of innate into adaptive immunity. Consequently, it seems sensible to propose that the nature of the selective pressures responsible for the evolutionary patterns depicted in this study is predominantly infectious. However, there is growing evidence, in both mice and humans, of a role of the TIR-containing adaptors in autoimmunity, inflammation, apoptosis, tumor development, and neurogenesis (Kaiser and Offermann 2005; Kim et al. 2007; Martino and Pluchino 2007; Naugler et al. 2007; Rakoff-Nahoum and Medzhitov 2007; Rolls et al. 2007; Isnardi et al. 2008; Rakoff-Nahoum and Medzhitov 2009; Ngo et al. 2010). In view of this, the selective factor(s) exerting pressures on the adaptor molecules may go well beyond nonself

immunity and involve other major biological processes related to the self.

Supplementary Material

Supplementary figures S1–S10, tables S1, S4, and S5, Tables S1 and S2 as separate Excel files, and material are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jean-Laurent Casanova, Etienne Patin, and Katherine Siddle for critical reading of the manuscript and insightful comments. This work was supported by the *Institut Pasteur*, the *Centre National de la Recherche Scientifique*, the *Fondation pour la Recherche Médicale*, Merck-Serono, the Ecole Polytechnique Fédérale de Lausanne-Debiopharm Life Sciences Award, and an *Agence Nationale de la Recherche* research grant (ANR-08-MIEN-009-01) to L.Q.-M.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. 7:248–249.

Akira S, Takeda K. 2004. Toll-like receptor signalling. *Nat Rev Immunol*. 4:499–511.

Akira S, Uematsu S, Takeuchi O. 2006. Pathogen recognition and innate immunity. *Cell* 124:783–801.

Allison AC. 1954. The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans R Soc Trop Med Hyg*. 48:312–318.

Bakker EG, Toomajian C, Kreitman M, Bergelson J. 2006. A genome-wide survey of R gene polymorphisms in Arabidopsis. *Plant Cell*. 18:1803–1818.

Bakker EG, Traw MB, Toomajian C, Kreitman M, Bergelson J. 2008. Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana. *Genetics* 178:2031–2043.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16:37–48.

Barreiro LB, Ben-Ali M, Quach H, et al. (18 co-authors) 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet*. 5:e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet*. 40:340–345.

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

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Belinda LW, Wei WX, Hanh BT, Lei LX, Bow H, Ling DJ. 2008. SARM: a novel Toll-like receptor adaptor, is functionally conserved from arthropod to human. *Mol Immunol*. 45:1732–1742.

Beutler B, Hoebe K, Georgel P, Tabeta K, Du X. 2004. Genetic analysis of innate immunity: TIR adapter proteins in innate and adaptive immune responses. *Microbes Infect*. 6:1374–1381.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors) 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. *Nature* 416:531–534.

Carty M, Goodbody R, Schroder M, Stack J, Moynagh PN, Bowie AG. 2006. The human adaptor SARM negatively regulates adaptor protein TRIF-dependent Toll-like receptor signaling. *Nat Immunol*. 7:1074–1081.

Casanova J, Abel L, Quintana-Murci L. 2011. Human TLRs and IL-1Rs in host defense: natural insights from evolutionary, epidemiological, and clinical genetics. *Annu Rev Immunol*. 29:447–491.

Cavalli-Sforza LL, Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet*. 33:266–275.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Ferwerda B, Alonso S, Banahan K, et al. (34 co-authors) 2009. Functional and genetic evidence that the Mal/TIRAP allele variant 180L has been selected by providing protection against septic shock. *Proc Natl Acad Sci U S A*. 106:10272–10277.

Frazer KA, Ballinger DG, Cox DR, et al. (250 co-authors) 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Gay NJ, Keith FJ. 1991. Drosophila Toll and IL-1 receptor. *Nature* 351:355–356.

Han J. 2006. MyD88 beyond Toll. *Nat Immunol*. 7:370–371.

He B, Santamaria R, Xu W, et al. (26 co-authors) 2010. The transmembrane activator TACI triggers immunoglobulin class switching by activating B cells through the adaptor MyD88. *Nat Immunol*. 11:836–845.

Isnardi I, Ng YS, Srdanovic I, et al. (20 co-authors) 2008. IRAK-4- and MyD88-dependent pathways are essential for the removal of developing autoreactive B cells in humans. *Immunity* 29:746–757.

Janeway CA Jr., Medzhitov R. 2002. Innate immune recognition. *Annu Rev Immunol*. 20:197–216.

Kaiser WJ, Offermann MK. 2005. Apoptosis induced by the toll-like receptor adaptor TRIF is dependent on its receptor interacting protein homotypic interaction motif. *J Immunol*. 174:4942–4952.

Kawai T, Akira S. 2007. TLR signaling. *Semin Immunol*. 19:24–32.

Kenny EF, O'Neill LA. 2008. Signalling adaptors used by Toll-like receptors: an update. *Cytokine* 43:342–349.

Kenny EF, Talbot S, Gong M, Golenbock DT, Bryant CE, O'Neill LA. 2009. MyD88 adaptor-like is not essential for TLR2 signaling and inhibits signaling by TLR3. *J Immunol*. 183:3642–3651.

Khor CC, Chapman SJ, Vannberg FO, et al. (34 co-authors) 2007. A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat Genet*. 39:523–528.

Kim Y, Zhou P, Qian L, Chuang JZ, Lee J, Li C, Iadecola C, Nathan C, Ding A. 2007. MyD88-5 links mitochondria, microtubules, and JNK3 in neurons and regulates neuronal survival. *J Exp Med*. 204:2063–2074.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20:2485–2487.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One*. 5:e10284.

Lazzaro BP. 2008. Natural selection on the Drosophila antimicrobial immune system. *Curr Opin Microbiol*. 11:284–289.

Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA. 1996. The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. *Cell* 86:973–983.

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. 2006. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*. 34:D257–D260.

Lewin R. 1987. Africa: cradle of modern humans. *Science* 237:1292–1295.

Li H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol*. 28:365–375.

Li JZ, Absher DM, Tang H, et al. (11 co-authors) 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Martino G, Pluchino S. 2007. Neural stem cells: guardians of the brain. *Nat Cell Biol*. 9:1031–1034.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.

Medzhitov R, Janeway C Jr. 2000. Innate immunity. *N Engl J Med*. 343:338–344.

Nagpal K, Plantinga TS, Wong J, Monks BG, Gay NJ, Netea MG, Fitzgerald KA, Golenbock DT. 2009. A TIR domain variant of MyD88 adapter-like (Mal)/TIRAP results in loss of MyD88 binding and reduced TLR2/TLR4 signaling. *J Biol Chem*. 284:25742–25748.

Nakajima T, Ohtani H, Satta Y, Uno Y, Akari H, Ishida T, Kimura A. 2008. Natural selection in the TLR-related genes in the course of primate evolution. *Immunogenetics* 60:727–735.

Naugler WE, Sakurai T, Kim S, Maeda S, Kim K, Elsharkawy AM, Karin M. 2007. Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. *Science* 317:121–124.

Ngo VN, Young RM, Schmitz R, et al. (31 co-authors) 2010. Oncogenically active MYD88 mutations in human lymphoma. *Nature* 470:115–119.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 8:857–868.

O'Neill LA. 2008. The interleukin-1 receptor/Toll-like receptor superfamily: 10 years of progress. *Immunol Rev*. 226:10–18.

O'Neill LA, Bowie AG. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat Rev Immunol*. 7:353–364.

Peng J, Yuan Q, Lin B, Panneerselvam P, Wang X, Luan XL, Lim SK, Leung BP, Ho B, Ding JL. 2010. SARM inhibits both TRIF- and MyD88-mediated AP-1 activation. *Eur J Immunol*. 40:1738–1747.

Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. *Nat Rev Genet*. 11:665–667.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 20:R208–R215.

Quach H, Barreiro LB, Laval G, et al . (11 co-authors) 2009. Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet*. 84:316–327.

Quintana-Murci L, Alcais A, Abel L, Casanova JL. 2007. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nat Immunol*. 8:1165–1171.

Rakoff-Nahoum S, Medzhitov R. 2007. Regulation of spontaneous intestinal tumorigenesis through the adaptor protein MyD88. *Science* 317:124–127.

Rakoff-Nahoum S, Medzhitov R. 2009. Toll-like receptors and cancer. *Nat Rev Cancer*. 9:57–63.

Reiling N, Ehlers S, Holscher C. 2008. MyDths and un-TOLLed truths: sensor, instructive and effector immunity to tuberculosis. *Immunol Lett*. 116:15–23.

Rolls A, Shechter R, London A, Ziv Y, Ronen A, Levy R, Schwartz M. 2007. Toll-like receptors modulate adult hippocampal neurogenesis. *Nat Cell Biol*. 9:1081–1088.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Rutschmann S, Kilinc A, Ferrandon D. 2002. Cutting edge: the toll pathway is required for resistance to gram-positive bacterial infections in Drosophila. *J Immunol*. 168:1542–1546.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors) 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312:1614–1620.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–1176.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. 73:1162–1169.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Takahashi M, Matsuda F, Margetic N, Lathrop M. 2003. Automated identification of single nucleotide polymorphisms from sequencing data. *J Bioinform Comput Biol*. 1:253–265.

Tauszig S, Jouanguy E, Hoffmann JA, Imler JL. 2000. Toll-related receptors and the control of antimicrobial peptide expression in Drosophila. *Proc Natl Acad Sci U S A*. 97:10520–10525.

Tiffin P, Moeller DA. 2006. Molecular evolution of plant immune system genes. *Trends Genet*. 22:662–670.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*. 4:e1000214.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A*. 102:18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4:e72.

von Bernuth H, Picard C, Jin Z, et al . (38 co-authors) 2008. Pyogenic bacterial infections in humans with MyD88 deficiency. *Science* 321:691–696.

Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, Corr C, Baker B. 1994. The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. *Cell* 78:1101–1115.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A*. 102:7882–7887.

Wlasiuk G, Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. *Mol Biol Evol*. 27:2172–2186.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.

Zhang SY, Jouanguy E, Ugolini S, et al. (33 co-authors) 2007. TLR3 deficiency in patients with herpes simplex encephalitis. *Science* 317:1522–1527.