A Genome-Wide Perspective of Human Diversity and Its Implications in Infectious Disease

Jérémy Manry and Lluis Quintana-Murci

Cold Spring Harb Perspect Med 2013; doi: 10.1101/cshperspect.a012450

Subject Collection  Bacterial Pathogenesis

General Aspects and Recent Advances on Bacterial Protein Toxins
Emmanuel Lemichez and Joseph T. Barbieri

Bacterial Quorum Sensing: Its Role in Virulence and Possibilities for Its Control
Steven T. Rutherford and Bonnie L. Bassler

Epigenetics and Bacterial Infections
Hélène Bierne, Mélanie Hamon and Pascale Cossart

Entry of Listeria monocytogenes in Mammalian Epithelial Cells: An Updated View
Javier Pizarro-Cerdá, Andreas Kühlbacher and Pascale Cossart

For additional articles in this collection, see http://perspectivesinmedicine.cshlp.org/cgi/collection/

Copyright © 2013 Cold Spring Harbor Laboratory Press; all rights reserved
A Genome-Wide Perspective of Human Diversity and Its Implications in Infectious Disease

Jérémy Manry1,2,3 and Lluis Quintana-Murci1,2

1Institut Pasteur, Unit of Human Evolutionary Genetics, Department of Genomes and Genetics, F-75015 Paris, France
2Centre National de la Recherche Scientifique, CNRS URA3012, F-75015 Paris, France
3Université Pierre et Marie Curie (UPMC), Cellule Pasteur, F-75015 Paris, France

Correspondence: lluis.quintana-murci@pasteur.fr

Progress in genomic technologies, such as DNA arrays and next-generation sequencing, is allowing systematic characterization of the degree of human genetic variation at the scale of individual genomes. Public efforts, such as the International HapMap Project and the 1000 Genomes Project, have provided a realistic picture of the levels of genetic diversity in individuals and populations. These genomic techniques are also making it possible to evaluate the contribution of host genetic diversity to differences in susceptibility to both rare and common infectious diseases. Recent studies have revealed the power of whole-exome sequencing for dissecting the immunological mechanisms underlying the pathogenesis of severe, rare infectious diseases. Likewise, genome-wide association studies on common viral, bacterial, and parasitic infections have shed light on the host genetic basis of susceptibility to infectious diseases and, in some cases, of disease progression and drug responses.

Dissection of the genetic basis of the phenotypic diversity observed in the human host is essential if we are to understand our relationship to disease, a major goal in biology and medicine. The release of the reference human genome sequence in the early 2000s (Lander et al. 2001; Venter et al. 2001) provided the first foundations for studies of the genetics of the human host, but provided little insight into the extent of naturally occurring genetic variation between the genomes of different individuals and populations that could account for their phenotypic variation. The advent of new genomic technologies, such as DNA microarrays and next-generation sequencing, has allowed comparative studies of variation at the level of the entire genome, between populations of different geographic origins and ethnic backgrounds, thereby increasing our understanding of the level of naturally occurring variation of the human genome and its relationship to disease. These genome-wide studies have yielded important insight into population histories and patterns of migration and admixture, into the genetic mechanisms underlying the adaptation of our species to changing environments, such as climate, nutritional resources, and pathogen pressures, and into the contribution of human genetics to disease susceptibility, severity, and response to treatment.
THE DIVERSITY OF THE HUMAN GENOME
Genome-Wide Data Sets for Human Genetic Variation

The Human Genome Project (Lander et al. 2001), the SNP Consortium (Sachidanandam et al. 2001), the International HapMap Project (The International HapMap Consortium 2005; Frazer et al. 2007; Altshuler et al. 2010), and, more recently, the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010) have collectively identified \( \approx 15 \) million common DNA variants, mostly single-nucleotide polymorphisms (SNPs), constituting a key tool for investigations of the extent of genetic variation in human populations. The first large-scale studies of human genetic variation at the genome-wide level were based on genotyping technologies, such as analyses of variation at hundreds of thousands of SNPs surveyed only in the genomic regions probed by the array. The International HapMap Project has been the most popular and successful project of this type. It was launched in 2002, with the aim of providing genotype and haplotype data for common SNPs in 270 individuals from four geographically diverse populations (Yoruba from Nigeria, individuals of Western European descent from Utah, Han Chinese, and Japanese). Phase I of the project involved the genotyping of \( \approx 1.3 \) million SNPs (The International HapMap Consortium 2005), a figure extended to \( >3.1 \) million SNPs during phase II (Frazer et al. 2007). The goal of phase III has been to increase the number of samples and populations investigated, rather than the number of SNPs, by genotyping 1.6 million SNPs in 1184 individuals from 11 global populations, and sequencing 10 100-kb regions in 692 of these individuals (Altshuler et al. 2010). The International HapMap Project has thus provided a high-resolution catalog of the allele frequencies of hundreds of thousands of SNPs distributed throughout the genome, across human populations. These data have provided important insight into the fine-scale structure of linkage disequilibrium (LD; i.e., the pattern of correlation between SNPs located close together on the chromosome) in the genome (Conrad et al. 2006b; Frazer et al. 2007), the distribution and causes of recombination hotspots (McVean et al. 2004; Myers et al. 2005), the prevalence of structural variation such as copy number polymorphisms (CNPs) (Conrad et al. 2006a; Redon et al. 2006), and the identity of genes that have been targeted by different forms of natural selection in the human genome (Voight et al. 2006; Frazer et al. 2007; Sabeti et al. 2007; Barreiro et al. 2008).

These genotyping resources have had a profound impact on the design and interpretation of genome-wide association studies (GWAs), by making it possible for the vast majority of common variants—defined as having a frequency \( >5\% \)—to be tested for association with various traits or diseases. However, the down side of these genotyping approaches is that they largely ignore low-frequency (from 0.5% to 5%) and rare (\( <0.5\% \) frequency) variants, which make up a large proportion of the human genome and may also contribute to disease susceptibility (Manolio et al. 2009). The 1000 Genomes Project was launched in 2008, to improve our understanding of the genetic contribution to human phenotypes through a consideration of the entire allelic spectrum of sequence-based variation (The 1000 Genomes Project Consortium 2010). Its aim was to provide a large number of complete human genome sequences and, therefore, accurate information on all forms of DNA polymorphism in major population groups from West Africa, Europe, East and South Asia, and the Americas. More specifically, the goal was to characterize \( >95\% \) of the variants with an allele frequency of 1% or higher (the accepted, conventional definition of polymorphism), or a frequency of 0.5% or lower within genes (the probability to find deleterious variants in this frequency range is the highest). The results of the pilot phase of the project have recently been released, providing information about the location, allele frequency, and local haplotype structure of \( \approx 15 \) million SNPs, 1 million short insertions and deletions, and 20,000 structural variants, most of which were newly identified in this project (Fig. 1). Overall \( \approx 8.4 \) million SNPs and \( \approx 840,000 \) insertions/deletions were described for the first time (The 1000 Genomes Project Consortium 2010). Collectively, these
sequencing data, together with the genotyping data obtained from the International HapMap Project (Altshuler et al. 2010) and the HGDP-CEPH panel (Li et al. 2008), have provided a realistic picture of the existing levels of genetic variation in our species, including the identification of rare variants, both at the interindividual and interpopulation levels.

**Human Genome Diversity and Population History**

The level of genetic variation of a population reflects its demographic history. Characterization of the patterns of genetic diversity in large numbers of human populations from diverse ethnic backgrounds should therefore provide important information about human origins and migration patterns. In this context, genome-wide data have confirmed that African populations have the highest levels of genetic diversity and that non-African diversity is essentially a subgroup of the diversity observed in Africans (Frazer et al. 2007; Li et al. 2008; Altshuler et al. 2010; The 1000 Genomes Project Consortium 2010). This provides conclusive support for a recent African origin of modern humans ~200,000 years ago, as suggested by other pieces of evidence from studies of uniparentally inherited mitochondrial DNA and Y-chromosome DNA and sets of neutrally evolving regions of the genome (Ingman et al. 2000; Thomson et al. 2000; Cavalli-Sforza and Feldman 2003; Voight et al. 2005; Fagundes et al. 2007; Campbell and Tishkoff 2008; Laval et al. 2010). The 1000 Genomes Project has shown, for example, that populations with African ancestry have the highest proportion of novel variants, reflecting the greater diversity of African populations (The 1000 Genomes Project Consortium 2010). However, as Asian and European populations have been expanding more rapidly in the recent past than the African population, it has been suggested that the sequencing of larger numbers of individuals from each population might yield a larger number of polymorphic sites in Eurasians than in Africans (Gravel et al. 2011). All the existing genome-wide data sets for human genetic variation have increased our understanding of the demographic history of human populations and provided insight into the degree of population genetic structure and levels of admixture in many populations worldwide. This information is essential for studies of association with disease.

**Human Genome Diversity and Natural Selection**

The observed patterns of population diversity or population structure can be explained by random genetic drift in most cases, reflecting the demographic history of the population under study. However, natural selection also influences our genomic diversity (Sabeti et al. 2006; Nielsen et al. 2007). Indeed, many genome-wide scans for selection have shown that the observed genetic patterns are incompatible with neutrality and have identified a number of genomic regions that have been subject to positive selection (Voight et al. 2006; Frazer et al. 2007; Sabeti et al. 2007; Barreiro et al. 2008; Akey 2009; Pickrell et al. 2009; Hancock et al. 2011). The detection of genes subject to positive selection facilitates
the identification of genes harboring functional variants that have conferred a selective advantage on the host. Studies of this type have identified genes and gene variants potentially accounting for differences in human phenotypes, which may be benign (e.g., skin color, height, hair development) or disease related. Collectively, these massive genome-wide data sets have shed light on the intensity and types of natural selection acting on the human genome, providing information complementary to that generated by clinical and epidemiological genetic approaches, through the identification of genes playing an important role in disease susceptibility and progression and, more generally, in host survival (Sabeti et al. 2006; Nielsen et al. 2007; Quintana-Murci et al. 2007; Gilad et al. 2009; Barreiro and Quintana-Murci 2010).

The Burden of Deleterious Mutations in the Human Population

Both demographic factors and natural selection shape the patterns of genetic diversity in populations, and the cumulative effects of these two forces have an impact on the burden of potentially deleterious mutations present in the human population. For example, the 1000 Genomes Project has estimated that each of us carries 190–210 in-frame insertions/deletions, 80–100 premature stop codons, 40–50 splice-site-disrupting variants, and 220–250 frameshift deletions, most of which lead to premature stop codons (Table 1) (The 1000 Genomes Project Consortium 2010). Even more surprisingly, this project has estimated that each of us carries 50–100 variants classified by the Human Gene Mutation Database as responsible for inherited disorders, albeit in the heterozygous state. Thus, 250–300 of the ~20,500 genes carried by a given individual harbor putative loss-of-function mutations (Clamp et al. 2007; The 1000 Genomes Project Consortium 2010). However, despite the enormous progress made in recent years, many questions remain unanswered. For example, most international consortia have focused on major ethnic groups, whereas ethnic minorities are largely underrepresented in these studies. The inclusion of adequate numbers of individuals from these minorities is of crucial importance for the identification of population-specific or region-specific functional variants associated with disease or drug response (Tishkoff et al. 2009). In addition, most of the signals of natural selection detected to date remain unexplained. The functional impact of the selected variants is usually postulated on the basis of predictive algorithms, such as PolyPhen (Adzhubei et al. 2010), but the actual role of most of the variants of interest remains unknown (Lohmueller et al. 2008).

THE INFLUENCE OF HOST GENETIC DIVERSITY ON HUMAN DISEASE

Studies of the degree of genetic variation at the level of the genome are making it possible to test blindly for genetic associations with disease

<table>
<thead>
<tr>
<th>Genic variants</th>
<th>Total</th>
<th>Total rangea</th>
<th>African (YRI)</th>
<th>European (CEU)</th>
<th>East-Asian (CHB + JPT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous SNPs</td>
<td>55,217</td>
<td>10,572–12,126</td>
<td>12,334–12,591</td>
<td>10,533–10,652</td>
<td>10,357–10,471</td>
</tr>
<tr>
<td>Nonsynonymous SNPs</td>
<td>61,284</td>
<td>9966–10,819</td>
<td>11,048–11,422</td>
<td>9690–9776</td>
<td>9932–10,115</td>
</tr>
<tr>
<td>Stop-introducing SNPs</td>
<td>951</td>
<td>88–101</td>
<td>93–100</td>
<td>74–78</td>
<td>102–108</td>
</tr>
<tr>
<td>Splice-site-disrupting SNPs</td>
<td>500</td>
<td>41–49</td>
<td>41–47</td>
<td>34–38</td>
<td>53–58</td>
</tr>
<tr>
<td>Genes disrupted by large deletions</td>
<td>143</td>
<td>28–36</td>
<td>32–35</td>
<td>34–37</td>
<td>33–37</td>
</tr>
</tbody>
</table>

SNP, single-nucleotide polymorphism.

* These ranges correspond to the interquartile range of the number of variants of specified type per individual.
susceptibility, pathogenesis, or response to treatment, without having to make prior assumptions as in candidate-gene studies. Such studies have been done for both rare (“Mendelian”) and common (mostly “complex”) diseases. The dividing line between Mendelian and complex diseases may be considered vague and simplistic, but Mendelian diseases are usually rare and mostly result from highly penetrant, deleterious alleles that segregate in families (Zwick et al. 2000; Pritchard and Cox 2002). However, most phenotypic differences in morphology, physiology, and disease are quantitative rather than discrete and follow a continuous, almost normal distribution (Pritchard and Cox 2002). These so-called “complex traits” are generally caused by a number of genetic and environmental factors and they therefore do not follow simple Mendelian rules of inheritance (Badano and Katsanis 2002; Botstein and Risch 2003). Genes involved in human disease may vary considerably in the number and population frequency of the corresponding disease-predisposing alleles (Fig. 2) (Reich and Lander 2001). For severe Mendelian diseases, the frequency of susceptibility mutations is usually very low (<0.1%) and is inversely proportional to the severity of the disease (Pritchard and Cox 2002). Such simple rules cannot be applied to complex susceptibility to disease. In this context, two principal nonmutually exclusive hypotheses have been put forward. According to the common disease–rare variant hypothesis, a significant proportion of the inherited susceptibility to disease results from the additive effects of a number of low-frequency, independent variants of different

**Figure 2.** Scheme representing the frequency of the variants and their effects in the framework of genome-wide association studies based on genotyping by next-generation sequencing. Most of the variants identified by GWAs are common variants with a moderate effect but, with advances in next-generation sequencing, it is now possible to identify rarer variants, as reported in recent exome studies detecting very rare variants responsible for Mendelian diseases. Rare variants with weak effects are likely to be difficult to identify genetically, as a number of such variants will probably be required to account for a disease or trait and this will require high levels of coverage of the entire genome. Common variants with a large effect are unusual as they would be expected to be eliminated by natural selection and should therefore not be present at such high frequencies. (From McCarthy et al. 2008; adapted, with permission.)
genes, with each variant making a moderate contribution to disease risk (Bodmer and Bonilla 2008). According to the common disease–common variant hypothesis, disease susceptibility is often owing to variants with relatively high population frequencies (Reich and Lander 2001).

In recent years, genome-wide studies of disease susceptibility have involved the genotyping of hundreds of thousands of SNPs, using array technologies such as those of Illumina or Affymetrix, and have focused principally on common, complex diseases or traits. By comparing the genome-wide frequencies of SNPs between affected and unaffected individuals, it is possible to identify genomic regions, and specific SNPs in particular, that are associated with the disease or trait studied, through the use of various statistical thresholds (i.e., the \( P \) values used to define a SNP as “robustly” associated with a given disease may differ in the extent to which they are conservative). By January 2012, >1000 GWAs had been published and >5000 SNPs had been found to be associated, to various extents, with numerous phenotypes (www.genome.gov). In total, 237 phenotypes have been associated with precise genomic regions, at \( P \) values \(< 5 \times 10^{-8} \). The diseases, syndromes, or traits overrepresented in these studies are complex traits or late-onset disorders, such as asthma (21 studies), Alzheimer’s disease (27), breast cancer (20), height (19), prostate cancer (19), schizophrenia (25), and type 2 diabetes (29). Most of these GWAs investigated polymorphisms with an allele frequency of >5% and largely neglected rare variants. This has led to an overassociation of common variants with disease. It has also been suggested that some GWA signals may reflect the effects of multiple, rare variants that have been credited to common variants (Dickson et al. 2010).

From a genome-wide perspective, studies of Mendelian disorders, which are expected to result from low-frequency variants, have progressed much more slowly than those focusing on complex traits or diseases. High-throughput sequencing is now making it possible to take into account these rare, sometimes population-specific variants. In this context, the whole-genome sequencing of affected individuals would be the most effective method, because it can capture causal variants located both within genes and elsewhere in the genome (e.g., regulatory variants). However, the technology required remains costly. In contrast, whole-exome sequencing, in which only exons, accounting for \(~1\%\) of the genome, are resequenced, has become technically feasible and affordable. It is now the most widely used genome-wide method in research into Mendelian disorders. In 2009, Ng and colleagues published the first exome-based study, providing proof of concept that causal genetic variants could be identified by whole-exome sequencing (Ng et al. 2009). A comparison of the exomes of individuals with Freeman-Sheldon syndrome (FSS), a rare autosomal-dominant disorder, with those of eight HapMap individuals led to the identification of \( MYH3 \), which had previously been implicated as the causal gene of FSS. Subsequent studies have identified variants or genes involved in numerous Mendelian disorders with no previously identified genetic cause, such as the \( DHODH \) gene, which Ng et al. showed to be the causal agent of Miller syndrome (Ng et al. 2010). More than 40 studies of Mendelian disorders based on whole-exome sequencing or the sequencing of whole genomes for a small number of individuals have been published. Ku and coworkers recently reviewed many of the exome studies done to date, highlighting the power and efficiency of whole-exome sequencing as an approach to identification of the genes underlying rare Mendelian disorders (Ku et al. 2011). One of the shortcomings of this method is that it cannot pick up possible regulatory variation outside genes that might also contribute to disease.

**GENOME-WIDE STUDIES OF MENDELIAN DEFECTS IN IMMUNITY AND INFECTION**

The Mendelian inheritance of numerous severe, rare infectious diseases or immunological defects has been dissected mostly by candidate-gene approaches. As this topic has been extensively reviewed elsewhere (Casanova and Abel 2005; Hill 2006; Fellay et al. 2009), we will focus here exclusively on studies based on genome-
wide approaches. The first identification of a Mendelian defect related to an infectious disease by a genome-wide approach was reported in 2010 (Byun et al. 2010). By whole-exome sequencing in a single child, the investigators identified a mutation responsible for fatal Kaposis sarcoma (KS), which although not infectious in itself, is directly related to human herpesvirus-8 (HHV-8) infection. After quality control and the exclusion of variants corresponding to simple polymorphisms, a splice-site mutation in the STIM gene, which is known to be involved in a primary T-cell immunodeficiency (Picard et al. 2009), was the only remaining potential disease-causing variant. STIM1 mRNA splicing, protein production, and Ca\(^{2+}\) influx were abolished in cell lines from the patient, but rescued by the expression of wild-type STIM1. This work showed that childhood KS could result from a single-gene defect and provided proof of concept that whole-exome sequencing in single patients could be used to decipher the genetic basis of rare inborn errors of immunity (Byun et al. 2010).

Since this pioneering study, other studies on Mendelian susceptibility to infectious diseases have been published. For example, a combination of genome-wide linkage analysis and whole-exome sequencing led to the identification of a homozygous missense mutation in the FADD gene of patients with autoimmune lymphoproliferative syndrome (ALPS) (Bolze et al. 2010). In this case, ALPS was observed in a context of severe bacterial and viral diseases, recurrent liver disease and encephalopathy, together with cardiac malformations. Supporting their genetic observations with functional analyses, the investigators determined not only the genetic basis of this complex disorder, but also the mechanisms underlying its pathogenesis. They showed that the observed bacterial infections resulted partly from functional hyposplenism, and that the observed viral infections resulted from impaired interferon (IFN) immunity. A similar approach has been used to show that germline mutations of STAT1 are responsible for chronic mucocutaneous disease (CMCD) (Liu et al. 2011). Loss-of-function STAT1 mutations had previously been implicated in predisposition to intracellular bacterial and viral diseases, impairing the STAT1-dependent response to IFNs (Zhang et al. 2008). However, the more recent exome-based study showed that some patients have a gain-of-function STAT1 mutation that leads to an impairment of IL-17 immunity, causing autosomal-dominant CMCD (Liu et al. 2011).

A recent study based on whole-exome sequencing in a patient suffering from a Crohn disease-like illness not only identified the causal mutation but also highlighted the potential use of this technology in a clinical setting (Worthey et al. 2011). The exome study identified a missense mutation in the X-linked inhibitor of apoptosis gene (XIAP). The XIAP protein plays a key role in the proinflammatory response and in bacterial sensing via the NOD signaling pathway. The results of functional analyses were consistent with aberrant XIAP function, with impaired IL-8 production in response to the stimulation of the patient's cells with muramyl dipeptide, and enhanced apoptosis after the CD3 stimulation of peripheral blood mononuclear cells (PBMC). Based on these findings, allogeneic hematopoietic progenitor cell transplantation was performed, in accordance with recommendations for the treatment of XIAP deficiency. An assessment more than 42 days after transplantation showed that the child had recovered, and no recurrence of gastrointestinal disease was observed (Worthey et al. 2011).

Collectively, these studies show that whole-exome sequencing in a small number of individuals, or even an isolated individual, is an efficient and cost-effective strategy for investigating the mechanisms underlying the pathogenesis of severe infectious diseases or immunity-related disorders, and for improving disease diagnosis, potentially leading to better treatment and personalized medicine.

**GENOME-WIDE STUDIES OF COMPLEX SUSCEPTIBILITY TO INFECTIOUS DISEASES**

GWAs of susceptibility to infectious diseases have progressed much more slowly than similar studies of other diseases, such as cancer,
psychiatric disorders, or metabolic diseases, with only seven complex infectious diseases having been studied in this way to date (for an extensive review, see Vannberg et al. 2011). However various GWAs for viral, bacterial, and parasitic infectious diseases have, in some cases, shed light on the contribution of host genetic diversity to the differences in response to both treatment and vaccine administration between individuals.

Viral Infections

The first reported GWAs of an infectious disease focused on HIV-1 infection and found that various SNPs in the human leukocyte antigen (HLA) region were associated with viral load (Fellay et al. 2007). They investigated HIV-1 viral load set point during the asymptomatic period of infection or the viral load “resting state” after the peak of viremia following acute infection. The strongest correlation was obtained for HLA-B and HLA-C, which were independently associated with a lower viral load set point (Fellay et al. 2007). A subsequent study by the same investigators confirmed these association signals, with some variants in this genomic region showing the strongest associations with both viral set point and progression (Fellay et al. 2009). The most likely causal variants appear to be the HLA-B*5701 allele and the HLA-C promoter SNP-35C, which has been shown to regulate the production of HLA-C mRNA (Fellay et al. 2007, 2009). The investigators estimated that human genetic variation and demographic variables could account for up to 22% of the variability in HIV-1 viral load, at least in the European population (Fellay et al. 2009).

Since this initial study, other GWAs of viral infections have been conducted. The association between hepatitis C virus (HCV) infection and the IL28B locus, which encodes interferon-\(\lambda_3\), merits particular attention. Several SNPs within or close to IL28B have been found to be associated with spontaneous HCV clearance or a better response to pegylated IFN\(\alpha\)-ribavirin treatment for chronic HCV infection in various human populations (Ge et al. 2009; Suppiah et al. 2009; Tanaka et al. 2009; Thomas et al. 2009; Rauch et al. 2010). The frequency of rs12979860, a SNP located 3 kb upstream of IL28B and associated with both spontaneous viral clearance and better response to drug treatment, varies significantly between populations, with the protective C allele being most prevalent in Asians, present at an intermediate frequency in Europeans and least prevalent in individuals of African descent (Fig. 3) (Ge et al. 2009; Thomas et al. 2009). Furthermore, it has recently been shown that this high frequency of the protective allele in Eurasian populations is the result of positive selection (Manry et al. 2011). These studies also pave the way for improvements in the treatment of HCV, to minimize adverse effects. Indeed, ribavirin-induced hemolytic anemia is a major side effect of pegylated IFN\(\alpha\)-ribavirin treatment, necessitating dose modification in up to 15% of patients. It has recently been shown that ITPA gene variants protect against hemolytic anemia in patients with hepatitis C treated with ribavirin (Fellay et al. 2010), again highlighting the importance of host genotype in response to treatment.

Another study identified 11 SNPs in a region including HLA-DPA1 and HLA-DPB1 as associated with hepatitis B virus (HBV) infection in a Japanese cohort (Kamatani et al. 2009). These associations were replicated in additional cohorts, and specific haplotypes associated with risk and protection were identified (Kamatani et al. 2009; Nishida et al. 2012). These findings suggest that antigen presentation on HLA-DP molecules may be critical for virus elimination, the pathogenesis of chronic hepatitis B. In the context of public health, the standard course of vaccination against HBV fails to elicit protective antibody levels in almost 10% of adult recipients. A GWAS of antibody titers in a cohort of hepatitis B vaccine recipients from Indonesia has recently been conducted, to identify genetic variants explaining such differences in vaccine responses (Png et al. 2011). The investigators identified a number of hits within the HLA region, including HLA-DR, HLA-DP, and a gene-rich HLA class III interval. The concordance between the alleles associated with protective effects against chronic HBV infection and those associated with a response to the vaccine (i.e., HLA-
suggests that the genetic bases of the two phenotypes may have major aspects in common. More recently, another GWAs has identified several SNPs associated with variation in cytokine responses in smallpox vaccine recipients, suggesting the complexity of the host response to vaccinia virus involving multiple genes and gene families (Kennedy et al. 2012a,b; Ovsyannikova et al. 2012). The most significant result they obtained was an intronic SNP in the \( \text{WDR92} \) gene, which is associated with variation in IFN-\( \alpha \) secretion. This, together with the reported associations of HLA-DP variations with HBV infection (Kamatani et al. 2009; Nishida et al. 2012) and of \( \text{IL28B} \) variants with both spontaneous viral clearance and response to treatment in HCV infection (Ge et al. 2009; Suppiah et al. 2009; Tanaka et al. 2009; Thomas et al. 2009; Rauch et al. 2010), suggests that the identification of host genetic variants associated with different susceptibilities to infection may generally provide useful information for improving both the prevention and treatment of infectious diseases.

Bacterial Infections

Three major bacterial infections have been studied from a genome-wide perspective: tuberculosis, leprosy, and meningococcal disease. The initial tuberculosis study combined two GWAs from Ghana and the Gambia and found weak to moderate associations with several genes, including those of the HLA-DQ region (Thye et al. 2010). The most significant hit concerned a non-HLA region located in a gene-poor region on chromosome 18q11.2. After imputation analysis of the data from the 1000 Genomes Project on the same Ghanian cohort, a possible resistance locus on chromosome 11p13 has been more recently identified (Thye et al. 2012), but the findings remain overall conflicting (Mahasirimongkol et al. 2012; Png et al. 2012). The apparent lack of success of these studies may conceal an interesting and more promising result. Gene-poor regions, such as the one on chromosome 18q11.2 associated with tuberculosis (Thye et al. 2010), may harbor regulatory elements (The ENCODE Project Consortium...
that may play a major role in disease susceptibility in addition to that of classical missense mutations in genes.

A first genome-wide association for leprosy detected significant associations with SNPs in the $CCDC122$, $C13orf31$ (LACC1), $NOD2$, $TNFSF15$, $HLA-DR$, $RIPK2$ and, possibly, $LRRK2$ genes in a cohort from China (Zhang et al. 2009). Subsequent studies achieved replication of these findings for $CCDC122$ and $C13orf31$ (LACC1) (Wong et al. 2010b) or obtained signals for other loci, such as $HLA-DRB1$, $HLA-DQA1$, and $TLR1$ (Wong et al. 2010a) in the Indian population. New susceptibility genes for leprosy have recently been identified and include regions encompassing the $IL23R$ and $RAB32$ genes (Zhang et al. 2011). The $RAB32$ gene has been implicated in the response to $Mycobacterium tuberculosis$ infection (Seto et al. 2011), suggesting that there may be genetic variants involved in susceptibility to both of these mycobacterial infections (Zhang et al. 2011). The conflicting results obtained in leprosy studies may reflect genuine biological differences in leprosy susceptibility between populations, or more technical issues relating to study design, such as the small sample sizes and low power of some studies.

Finally, a GWAS of meningococcal disease in various European cohorts found and replicated an association with a cluster of complement factor SNPs, including a missense mutation in the complement factor H gene ($CFH$) and an intronic mutation in the CFH-related protein 3 gene ($CFHR3$) (Davila et al. 2010). These results suggest that genetic variation in complement genes may play an important role in determining whether a subject presents invasive disease or asymptomatic $Neisseria meningitides$ colonization, opening up new avenues of research relating to possible treatment of the disease by pharmacological modulations of CFH protein levels in vivo.

**Parasitic Infections**

Malaria, particularly its severe forms, is the only parasitic disease to have been studied from a genome-wide perspective. A large GWAS of severe malaria in a cohort of children from the Gambia identified the strongest signal of association as being located close to the $HBB$ gene on chromosome 11p15 (Jallow et al. 2009), to which the well-known HbS polymorphism (rs334) maps. HbS homozygotes suffer from sickle-cell disease, but heterozygotes have a 10-fold-reduced risk of severe malaria, which explains the high frequencies (up to 30%) of this allele in Africa (Allison 1954).

By combining genome-wide studies with the targeted resequencing of the $HBB$ genomic region in a subset of individuals, the investigators identified the causal HbS polymorphism as the SNP displaying the strongest association. This study provided proof of principle that the causal polymorphism within a GWA signal can be identified by regional sequencing of the corresponding genomic region. In addition to the $HBB$ hit, the investigators identified 19 additional but weaker association signals, including regions encompassing or located close to genes such as $SPATA3$, $LOC257407$, $PSMD1$, and $GPR55$. Some of these variants may be authentic loci with a modest genetic effect on malaria susceptibility, but further validation is required.

**INSIGHT FROM THE EVOLUTIONARY APPROACH TO INFECTIOUS DISEASE RESEARCH**

The evolutionary genetics of immunity is a field of research investigating whether and how natural selection has shaped the variability of host defense or immunity-related genes in present-day human populations. This approach has proved an indispensable complement to clinical and epidemiological genetics for understanding the biological relevance of host genes in host survival (Casanova and Abel 2005; Nielsen et al. 2007; Quintana-Murci et al. 2007; Sabeti et al. 2007; Barreiro and Quintana-Murci 2010). Indeed, the burden of infectious diseases has been massive throughout history, with life expectancy not exceeding 20 to 25 years of age until the advent of Pasteur’s microbial theory of disease and the resulting control of infections by hygiene, vaccines, and antibiotics (Casanova and Abel 2005). It is therefore hardly
surprising that immunity-related genes appear to be among those with biological functions most strongly targeted by natural selection in the human genome, both with respect to the chimpanzee and within different human populations (Barreiro and Quintana-Murci 2010). For example, several genome-wide scans for positive selection have detected more than 5000 loci presenting at least one genomic signature of positive selection (see Akey 2009 and Barreiro and Quintana-Murci 2010 for reviews). Within the genomic regions targeted by positive selection, 360 genes with immune-related functions have been identified, including 186 genes identified as targets of positive selection in at least two studies (Barreiro and Quintana-Murci 2010). This group of “selected genes” may display functional variation that is differentially distributed between human populations and is therefore highly likely to be involved in the differences in susceptibility to infectious diseases or, more generally, to immunity-related disorders, such as autoimmune, inflammatory, or allergic disorders, observed today in human populations. The most obvious source of selection pressure on these genes is the presence of the pathogen (e.g., pathogen-driven selection). Proof of concept is provided by HLA class I, blood group antigen, and interleukin genes, for which genetic variation has been shown to be associated with pathogen species richness and diversity (Prugnolle et al. 2005; Fumagalli et al. 2009a,b). Balancing selection, which is considered to be rare in humans, is known to have acted on several immune-related genes; the best known example is major histocompatibility complex (MHC) genes, which display an extraordinary amount of variation in the population because heterozygous individuals are able to recognize a wider variety of microbes (Prugnolle et al. 2005; Cagliani et al. 2008; Andres et al. 2009; Fumagalli et al. 2009a,b).

Past selection, leading to an increase in resistance to the presence of the pathogen may lead, in some cases, to higher susceptibility to autoimmune or inflammatory disorders, as predicted by the hygiene hypothesis (Sironi and Clerici 2010). For example, several interleukin/interleukin receptor genes involved in the pathogenesis of inflammatory bowel disease (IBD) and celiac disease (CeD) have been subjected to pathogen-driven selective pressure. Specifically, the IBD/CeD risk alleles of IL18RAP, IL18R1, IL23, IL18R1, and the intergenic region between IL2 and IL21 are more frequent in populations exposed to high microbial/viral loads, suggesting that these variants play a protective role in the host response to these pathogens (Fumagalli et al. 2009b; Sironi and Clerici 2010). Furthermore, several alleles or haplotypes associated with greater susceptibility to autoimmune disorders, including type I diabetes (T1D), CeD, and multiple sclerosis (MS), have been shown to have been subject to recent positive selection (Barreiro and Quintana-Murci 2010). This suggests that some unknown selective pressure, possibly pathogen driven, has resulted in an increase in the frequency of risk alleles/haplotypes in the general population or even in specific populations. For example, European populations display high frequencies of variants conferring susceptibility to celiac disease (Abadie et al. 2011).

CONCLUDING REMARKS

The relatively slow pace at which genome-wide research into infectious diseases has progressed reflects, at least partially, the complex nature of infectious diseases, which have diverse genetic bases, extending from single genetic defects to the presence of complex susceptibility alleles with various degrees of penetrance and much more complex situations in which many common or rare variants are involved, each having a modest effect. In the next few years, whole-genome sequencing technology is likely to become more affordable, and this will probably lead to a huge increase in the number of studies based on these techniques. This technology will be used not only in the context of rare infectious diseases or immunological defects probably resulting from single-gene lesions, but also in the context of complex susceptibility to infectious diseases, which may be caused by numerous low-frequency variants. The genetic data obtained from these whole-genome sequence-
based association studies will then be combined with other independent sources of data, in large cohorts of individuals presenting clinically well-defined infectious phenotypes. Multidisciplinary approaches, as those used by the ENCODE Project Consortium (The ENCODE Project Consortium 2012), integrating whole-genome sequencing data, expression, and proteomic profiles obtained with arrays, RNA-Seq and mass spectrometry, epigenetic markers such as genome-wide methylation profiles, and evolutionary and population genetics data should facilitate the identification of functionally important genes and variants involved in host immunity to infection and uncover the immunological mechanisms favoring our past and present survival in the natural setting. This, together with improvements in statistical methods and modeling, will facilitate the dissection and more precise quantification of the contribution of host genetics to the considerable heterogeneity between individuals and populations in terms of susceptibility to infection, disease progression, and drug responses.

ACKNOWLEDGMENTS

This work is supported by the Institut Pasteur, the Agence Nationale de la Recherche (ANR-08-MIEN-009-01), the Fondation pour la Recherche Médicale (FRM), the Centre National de la Recherche Scientifique, and an EPFL Debiopharm Life Sciences Award to L.Q.-M. J.M. is supported by an FRM fellowship.

REFERENCES


