Natural selection and infectious disease in human populations

Elinor K. Karlsson[1,2], Dominic P. Kwiatkowski[3,4] and Pardis C. Sabeti[1,2,5]

Abstract | The ancient biological 'arms race' between microbial pathogens and humans has shaped genetic variation in modern populations, and this has important implications for the growing field of medical genomics. As humans migrated throughout the world, populations encountered distinct pathogens, and natural selection increased the prevalence of alleles that are advantageous in the new ecosystems in both host and pathogens. This ancient history now influences human infectious disease susceptibility and microbiome homeostasis, and contributes to common diseases that show geographical disparities, such as autoimmune and metabolic disorders. Using new high-throughput technologies, analytical methods and expanding public data resources, the investigation of natural selection is leading to new insights into the function and dysfunction of human biology.

Pathogens
Viruses, bacteria or other microorganisms that can cause disease.

[1]*Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.*
[2]*Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142, USA.*
[3]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.*
[4]*Wellcome Trust Sanger Centre for Human Genetics, Oxford OX3 7BN, UK.*
[5]*Department of Immunology and Infectious Disease, Harvard School of Public Health, Boston, Massachusetts 02115, USA.*
Correspondence to E.K.K. and P.C.S.
e-mails: elinor@broadinstitute.org; psabeti@oeb.harvard.edu
doi:10.1038/nrg3734
Published online 29 April 2014

Infectious pathogens are arguably among the strongest selective forces that act on human populations[1]. Migrations and cultural changes during recent human evolutionary history (the past 100,000 years or so) exposed populations to dangerous pathogens as they colonized new environments, increased in population density and had closer contact with animal disease vectors, including both conventionally domesticated animals (for example, dogs, cattle, sheep, pigs and fowl) and those exploiting permanent human settlement (for example, rodents and sparrows)[2,3]. Consequently, both birth and mortality rates increased markedly[4].

Host genetics strongly influences an individual's susceptibility to infectious disease[5,6]. Pathogens that diminish reproductive potential, either through death or poor health, drive selection on genetic variants that affect resistance; selection is likely to be most evident for pathogens with a long-standing relationship with *Homo sapiens*, including those that cause malaria, smallpox, cholera, tuberculosis and leprosy[7] (FIG. 1). We also contend with new threats, such as AIDS and severe acute respiratory syndrome (SARS). Some pathogens cause acute illnesses such as smallpox and cholera but, once survived, pose little additional threat. Other pathogens — for example, those causing malaria, tuberculosis and leprosy, as well as parasitic worms — can be carried as chronic infections and impair nutrition, growth, cognitive development and fertility. The timing, strength and direction (that is, positive, negative or balancing) of selection shape the patterns of variation that remain in the genome. These signatures of selection will therefore vary with the age, geographical spread and virulence of the pathogen.

For those with access, modern medicine radically diminishes exposure to various pathogens. In developed countries, vaccination, better nutrition and improved public health have eliminated diseases that were common in the past[8]. Common immune-mediated diseases may be partly caused by evolutionary adaptations for resistance and symbiosis with potentially dangerous microorganisms[9–12]. For example, decreased gut microbiome diversity in residents of developed countries[13] may alter mucosal immune responses[14]. Understanding host–pathogen interactions will inform the development of new therapies both to counter ongoing pathogen evolution and to better manage immune-mediated diseases[15].

Here, we review how the technological revolution in genomics allows us to examine human adaptation to infectious disease in new ways. Natural selection leaves distinctive signatures in the genome, as genetic variants that improve survival and reproduction increase in frequency, and detrimental variants vanish. Hundreds of candidate regions of selection were identified in early genomic data sets, but only few adaptive variants were identified[16]. High-throughput biotechnology enables large-scale surveys of genome diversity, genome-wide association studies (GWASs), next-generation sequencing, and high-throughput experimental and bioengineering approaches[17]. Together with expanding computational capacity[18], these tools offer new power to find and functionally elucidate adaptive changes. Pathogen susceptibility and immune traits are particularly amenable

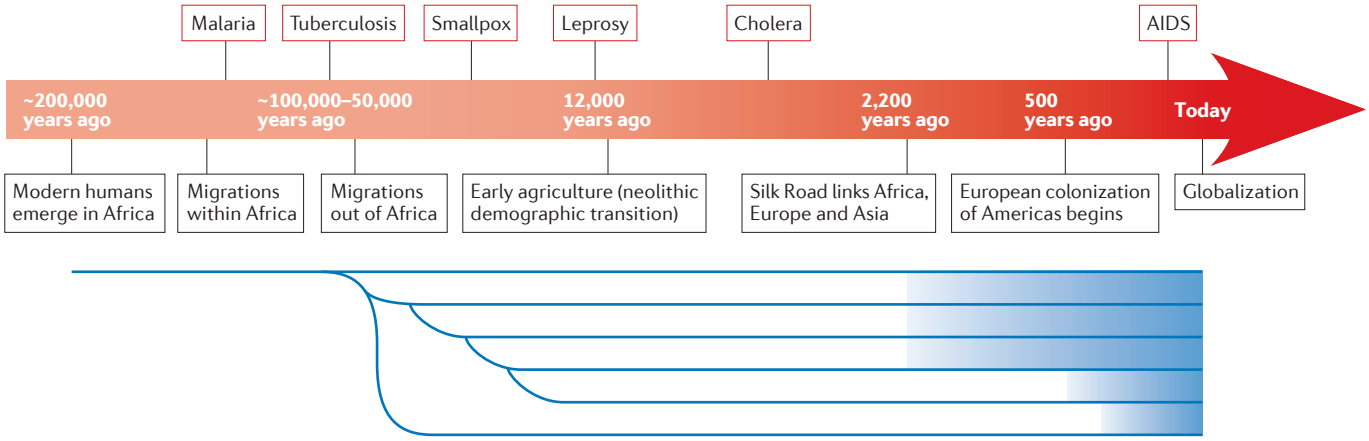


Figure 1 | **Pathogen emergence during human history.** Key events in recent human evolution (boxes outlined in black) are juxtaposed with the estimated ages of infectious disease emergence (boxes outlined in red). The fragmentation of the human lineage into genetically and geographically distinct populations (blue lines) accelerates with migration out of Africa. Later, these populations started mixing more (blue shaded regions between the populations) along trade routes (such as the Silk Road), through colonization and through high rates of global travel nowadays.

to mapping approaches that combine scans for natural selection and genetic association. We consider how these new genomic analyses provide insights into human evolution and have implications for human health. We focus primarily on examples in which selection is connected to infectious disease susceptibility through additional phenotypic associations or functional investigations.

Methods and technologies

Signatures of selection. Natural selection is the tendency for traits to increase or decrease in frequency in a population depending on the reproductive success of those exhibiting them. Positive selection increases the frequency of favoured alleles, negative selection eliminates detrimental alleles, and balancing selection favours diversity. This process leaves unusual patterns of genetic diversity that mark selected loci (that is, signatures of selection) when compared with the background distribution of genetic variation in the genome, which is assumed to evolve under neutrality to a large extent[19]. Population events that alter genetic diversity — including bottlenecks, expansions, splits and admixture — complicate accurate detection of selected loci.

Scans for natural selection have been made possible by statistical tools to detect signatures of selection and by rapidly expanding whole-genome data sets in multiple human populations. Whereas early scans relied on single-nucleotide polymorphism (SNP) genotyping arrays, next-generation sequencing technologies now enable generation of whole-genome sequence data sets for analysis. Such data sets have several advantages. They do not suffer from ascertainment bias, which is the distortion in measures of genetic diversity and neutral variation[20] created by the nonrandom sampling of SNPs on arrays. In addition, sequence data make it feasible to dissect loci with complex patterns of selection and short blocks of linkage disequilibrium (LD), such as the haemoglobin beta (*HBB*) gene that is associated with sickle cell anaemia[21]. Finally, as sequencing can detect potentially all variation throughout an individual's genome, the search for the precise causal variant driving selection is facilitated; however, as for SNP genotyping arrays, the comprehensiveness of capturing population-wide variation will depend on the number of individuals sampled.

Here, we briefly describe a few commonly used signatures that can help to elucidate human adaptations to pathogens. Various excellent resources provide more detailed background on statistical methods for detecting selection[22–26].

Signatures of positive selection. Positive selection increases the prevalence of genetic variants that improve survival and fertility. For example, a mutation that protects against malaria by disrupting expression of the Duffy antigen gene *DARC* (also known as *FY*), which encodes the receptor used by the *Plasmodium vivax* malarial parasite to enter red blood cells, has reached fixation in most of sub-Saharan Africa[27]. Positive selection can act on new variants or on standing variation that becomes favourable owing to environmental changes[28,29].

The test used to detect positive selection depends on when the selection occurred and on whether the variant is standing or new[28]. Very ancient selection may leave an excess of fixed, functional (for example, protein-coding) genetic changes that have been acquired over millions of years and through repeated selective sweeps. A selected variant that increases rapidly in frequency in the past ~250,000 years can be detected as an unusual reduction in genetic diversity. Recent positive selection (within the past 5,000–100,000 years) can be found with three different signals: unusually large allele frequency differences between populations, unusually high frequency of newly derived variants and unusually extended LD caused by the rapid increase in frequency of a single allele. The LD-based methods[30] are particularly useful for detecting incomplete sweeps (that is, variants increasing in prevalence but not to fixation), which are more common in recent human evolution than complete sweeps[31–34].

Signatures of selection
An unusual pattern of allele frequencies that marks a selected locus.

Frequency
Prevalence of an allele in a population.

Genome-wide association studies
(GWASs). Examination of variants that are distributed across the entire genome for correlation with particular traits.

Next-generation sequencing
New high-throughput, parallelized, low-cost sequencing technologies that do not use the chain termination Sanger method.

Genetic diversity
Total amount of genetic variation in a population.

Bottlenecks
Sharp decreases in the effective sizes of populations.

Admixture
Interbreeding between two genetically separated populations.

Ascertainment bias
Nonrandom selection of variants for genotyping.

Neutral variation
Genetic variation that confers no selective advantage or disadvantage and that varies in frequency by random drift.

Linkage disequilibrium
(LD). The nonrandom association of alleles at different genomic loci.

Fixation
The increase in frequency of an allele to 100% in a population.

These methods have detected hundreds of loci with signatures of selection in the human genome[16,28,30,35–41]. Their sensitivity to recent events depends on the strength of selection, as more advantageous variants increase to detectable frequencies faster. The comparison of many individuals from closely related populations using tree-based methods can help to detect smaller selection-driven changes in allele frequencies[42]. Combining several different tests for selection improves sensitivity to a wider range of selection regimes, helps to narrow down candidate regions and pinpoints a small number of top causal candidates[28,33,43].

Pathogen resistance alleles are prime candidates for discovery, as they are likely to be both recent and common, and have increased in frequency owing to the burden of disease.

Signatures of balancing selection. Balancing selection maintains multiple alleles at a locus. A variant may confer a heterozygous advantage; for example, at the *HBB* sickle cell locus, heterozygous carriers are malaria resistant to a large extent but otherwise healthy, which gives them a reproductive advantage in malaria-endemic environments over homozygous wild-type individuals (who are susceptible to malaria) and homozygous carriers (who have sickle cell disease and high childhood mortality[27]). Alternatively, the advantage conferred by a variant may depend on its prevalence. Diversifying selection favours high levels of polymorphism, as in the major histocompatibility complex (MHC). MHC diversity confers resistance to a broader range of pathogens[44] and is strongly selected for[45]. In humans, MHC diversity correlates with local pathogen diversity, which is consistent with pathogen-driven balancing selection[46]. Selection at this locus is of particular interest, as it harbours more bona fide disease associations than any other region in the human genome[47,48], including associations to infectious disease susceptibility (including AIDS[49–51], leprosy[52], leishmaniasis[53], hepatitis B[54–56], hepatitis C[57] and human papilloma virus (HPV) infection[58]), autoimmune disorders, cancers and neuropsychiatric diseases.

Recent balancing selection can resemble an incomplete selective sweep and be detected using LD-based tests for positive selection, as at the *HBB* sickle cell locus[59,60]. Long-term balancing selection reduces the number of rare alleles in a region and causes an excess of polymorphism, an excess of intermediate-frequency variants and reduced allele frequency differences between populations[61].

Cross-species comparative sequence analysis is particularly effective for detecting the subtle signals of ancient balancing selection. A comparison of humans and non-human primates found that both the MHC[61] and the blood group locus *ABO*[62,63] contain ancient multi-allelic polymorphisms maintained across species, which indicates balancing selection. A comparison of whole-genome sequences for 10 chimpanzees and 59 humans found that regions with signatures of balancing selection — MHC and 125 other loci[64] — were enriched for membrane glycoproteins. These are proteins exploited by a broad range of pathogens as receptors for cell invasion and to evade the host immune response, which suggests that the selection was pathogen driven[65–67].

Signatures of negative selection and purifying selection. Negative selection eliminates existing detrimental variation from a population[68]. For example, when human populations in the Ganges River Delta encountered pathogenic *Vibrio cholerae,* individuals of blood type O had higher risk of dying from severe cholera, which put them at a strong reproductive disadvantage. Nowadays, populations in the cholera-endemic Ganges River Delta have the lowest rates of blood type O in the world, which is consistent with negative selection[69,70]. Purifying selection is the ongoing removal of deleterious alleles as they arise. Signatures of purifying selection include decreased overall diversity, loss of functional variation and an excess of rare alleles[68]. Purifying selection also manifests as a lack of substitutions between species, and this signal is used to identify functionally important, highly conserved genomic regions in cross-species comparisons[71].

GWASs of adaptive traits. GWASs identify genomic variants that are significantly correlated with a phenotype of interest, typically in large sample cohorts. The sample size required is smaller if variants are at high prevalence, have strong effects and are in regions of extensive LD[72,73], all of which are characteristics of positively selected loci (FIG. 2). As beneficial alleles increase in prevalence, they carry nearby variants with them (known as genetic hitchhiking); these nearby variants can thus be proxies for the causal allele and enhance power to detect association.

The US National Institutes of Health (NIH) Catalog of published GWASs includes all strong SNP–trait associations ($P < 1 \times 10^{-5}$) from 1,900 curated publications[48]. However, only 68 publications are related to pathogen susceptibility, and most of those (63%) are focused on AIDS (16 studies), hepatitis B (12 studies) and hepatitis C (15 studies). The remainder includes GWASs of tuberculosis (5 studies), prion diseases (3 studies), malaria (3 studies), leprosy (2 studies), smallpox (2 studies) and 9 others (see Supplementary information S1 (table)). Before GWASs, a candidate gene approach was used to find strong-effect variants that were associated with disease susceptibility in *HBB*, *DARC* and *SLC4A1* (solute carrier family 4 (anion exchanger), member 1 (Diego blood group)) for malaria, and the *CCR5* (chemokine (C-C motif) receptor 5) gene for AIDS[74]. However, candidate gene studies were generally underpowered, and beyond these textbook examples most previous associations have so far failed to replicate in GWASs[75]. For example, of 22 tuberculosis-associated genes, none were significantly associated in a GWAS of 11,425 Africans, nor were these genes overrepresented among nominally significant results[76–78]. Candidate gene studies of infectious disease susceptibility have been reviewed in detail elsewhere[79].

Integrating selection metrics into GWASs can increase their power. A study of malaria resistance loci in Africa found that when association results are weighted on the basis of evidence of positive selection,

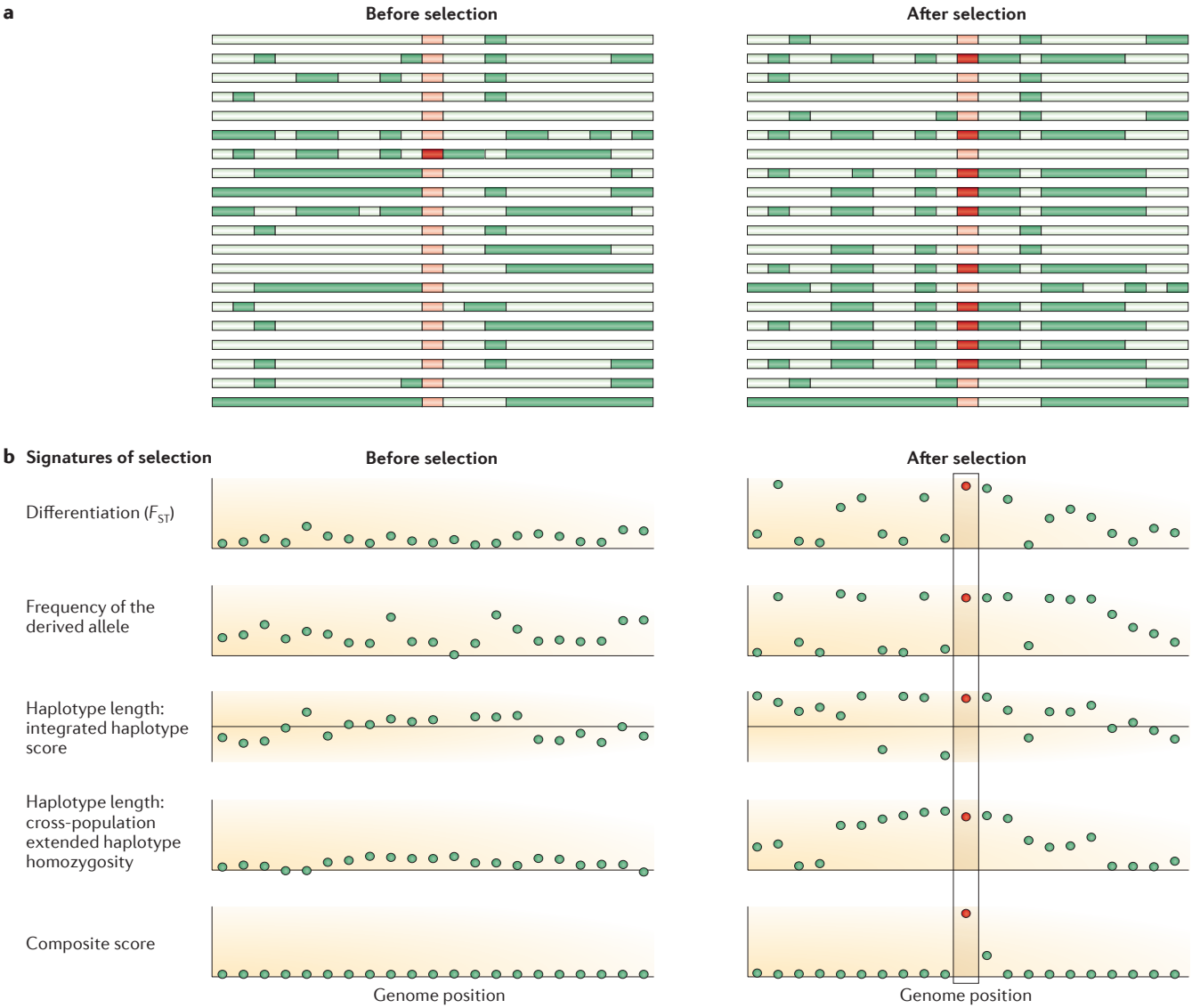
Standing variation
Existing genetic variation within a population.

Selective sweeps
Reductions in genetic variation caused by positive selection at particular loci.

Incomplete sweeps
Partial or ongoing selective sweeps of advantageous alleles to < 100% prevalence.

Complete sweeps
Selective sweeps of advantageous alleles to 100% prevalence.

Candidate gene approach
Association study that tests only variants in a pre-specified set of genes.



◀ Figure 2 | **Positive selection increases power to detect associations in GWASs.** **a** | The variants in a population-wide sample are shown for a schematic genomic locus. The red region indicates a variant that provides selective advantage (such as a host variant that confers relative resistance to an infectious agent). Positive selection of that variant rapidly increases its prevalence in a population and also the prevalence of nearby alleles that are in linkage with it. **b** | Positive selection on a variant is detectable with three types of signals: high levels of differentiation (that is, when positive selection in one geographical region causes larger frequency differences between populations than those expected for neutrally evolving alleles); high frequency of the derived allele (that is, when a new allele increases to a frequency higher than that expected under genetic drift); and long haplotypes (as determined by the integrated haplotype score and cross-population extended haplotype homozygosity) that are left when the selected allele increases in frequency sufficiently quickly that long-range associations with neighbouring variants are maintained. Combining different signals of selection into a composite score can increase resolution by up to 100-fold, which facilitates identification of the causal variants. **c** | Variants that are of high frequency, of strong effects and in regions of extensive linkage disequilibrium (LD) — all of which are characteristics of positively selected loci — are detectable with smaller sample sizes in genome-wide association studies (GWASs). Even with full sequence data, low-frequency alleles (solid blue lines) require larger sample sizes (*x* axis; equal number of cases and controls) than high-frequency alleles (solid red lines) for equivalent power over a range of effect sizes (1.2, 1.5 and 2.0). When mapping with genotyping arrays, shorter LD in unselected regions (blue dashed lines; modelled here using power simulations for the sparser Illumina 300K array) can cause a larger loss of power (relative to full sequence data) than in selected regions with longer LD (red dashed lines; modelled as denser Illumina 1M array). F_{ST}, Wright's fixation index. Part **c** is adapted from REF. 198.

power increases by 1–2 orders of magnitude[80,81]. Another approach was used to identify positively selected drug resistance loci in the malaria parasite *Plasmodium falciparum*; the sample was partitioned into two populations defined by phenotype (for example, resistant and susceptible), and regions of selection were identified in resistant individuals[82]. Furthermore, as described above, selection increases the power of GWASs themselves by driving the emergence of common alleles of strong effect. Indeed, using the Kolmogorov–Smirnov test, we found that variants in positively selected regions in the NIH Catalog of published GWASs are more significantly associated with the tested phenotype ($P_{KS} = 2 \times 10^{-7}$) than those in the rest of the genome[33].

Signatures of polygenic selection. Similar to many other human traits, adaptation to pathogens is likely to be polygenic[33,83], as mutations in many genes emerge and increase in prevalence through selection. Pathway-based approaches scrutinize candidate regions identified in GWASs[84] to elucidate functional pathways that are disrupted in complex human diseases[85]. The same approach can be applied to candidate selected regions; however, as natural selection acts on many traits, they are less powered. Only when a particular pathway is repeatedly selected for in a population, such as skin pigmentation in Europeans, do we expect significant enrichment.

Functional characterization. For an allele to be selected it must have a functional effect, but the relevant phenotype is typically unknown. Selected regions therefore present an opportunity to develop a systematic process for functionally characterizing genetic variation. In a few cases, a selected region contains well-characterized genes that offered clear functional hypotheses, for example, Toll-like receptor 5 (*TLR5*, which is involved in bacterial flagellin sensing[33]) (FIG. 3A), apolipoprotein L1 (*APOL1*, which encodes a serum factor that lyses *Trypanosoma brucei*[86]) (FIG. 3B) and *HBB* (which influences infection by *P. falciparum*[27] (see above)). For genes with unknown or diverse functions, identification of the selected trait is more complicated. Moreover, most selection signals are non-genic and are likely to alter poorly characterized regulatory regions of the genome[43,87,88].

Functional elucidation remains a difficult and mostly manual process, and benefits enormously when candidate variants are narrowed down to a small number. New methods for detecting selection are remarkably effective at identifying top functional candidates among all variants in selected regions — a list that can be further reduced by incorporating GWAS results, expression quantitative trait loci (eQTLs) and functional annotations[33,89]. Useful resources include the 1000 Genomes Project, which aims to identify all common (>1%) human genetic variation by sequencing >1,000 individuals[90], and the Encyclopedia of DNA Elements (ENCODE) Project, which aims to characterize all functional elements in the genome[91]. Breakthroughs in next-generation sequencing, high-throughput functional screens, single-cell genomics, microfluidics, chromosome conformation capture and genome engineering approaches now make it possible to test many variants in parallel, to investigate non-genic regions and to functionally screen the whole genome[17,92–99].

Genetics of infectious disease resistance

The dynamics of host–pathogen interactions (for example, length of exposure, geographical spread, morbidity and mortality, and co-occurring environmental events) influences the genetic architecture of resistance variants in modern populations (FIG. 4). Many different modes of selection shape patterns of variation in humans (reviewed in REFS 25,26), and selection scans using current methods and data sets can only detect a subset of selected loci. The most conspicuous signals are perhaps left by positive selection in recent human evolutionary history. This is the timeframe during which many major pathogens first emerged (FIG. 1; TABLE 1), which suggests that this mode of selection is particularly relevant to studies of infectious disease susceptibility. Moreover, finding beneficial variants that are favoured by recent selection could suggest new medical therapies. Transforming genetic discoveries into improved healthcare will take time, but understanding natural resistance is an important first step.

We note that some have questioned the extent of recent positive selection in humans in light of a recent paper[31]. In that paper, the authors estimated that during the past 250,000 years, ~0.5% of nonsynonymous substitutions have swept to fixation (that is, 100% prevalence), and such an observation has been interpreted to suggest that very few positively selected variants can be found in humans. In actuality, this corresponds to

Pathway-based approaches
Methods that test for joint association of genes in the same functional pathway.

Expression quantitative trait loci
(eQTLs). Genomic loci that regulate gene expression.

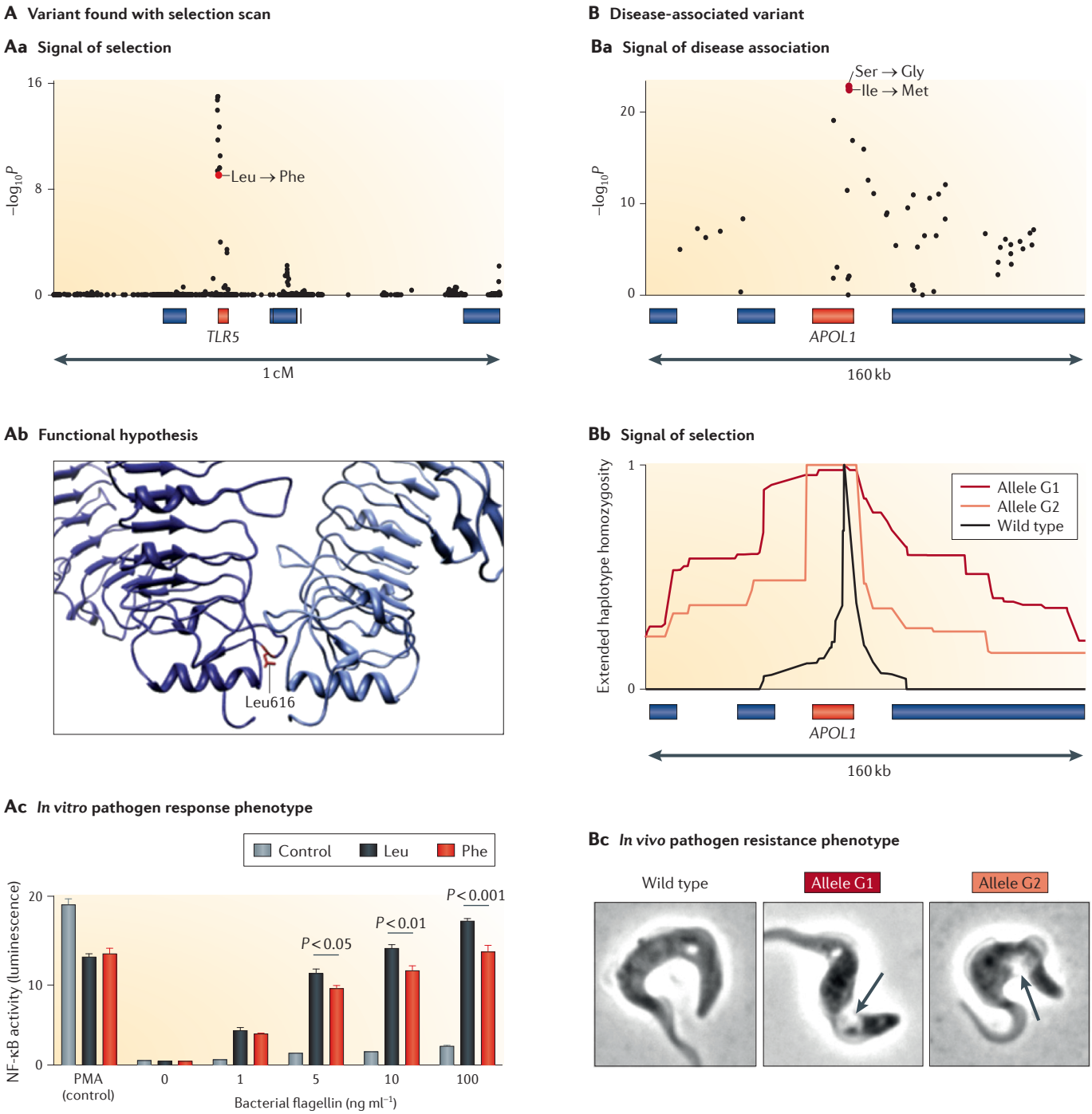


Figure 3 | **Selected variants implicated in pathogen resistance. A** | A genome-wide scan for signals of positive selection in the Yoruban population of Nigeria found a strongly selected nonsynonymous single-nucleotide polymorphism (SNP) that alters the pathogen recognition protein Toll-like receptor 5 (TLR5) (part **Aa**) and that is predicted to disrupt TLR5 activation in response to flagellated bacteria[33] (part **Ab**). Cell lines that carry the new TLR5 variant (Leu616Phe) had significantly reduced nuclear factor-κB (NF-κB) signalling in response to flagellin, which is potentially protective against some bacterial infections (part **Ac**). Error bars represent the standard error of the mean over at least three independent experiments; *P* values are indicated above the bar graphs. **B** | Two common variants of the apolipoprotein L1 (*APOL1*) gene (Allele G1 and Allele G2) that are strongly associated with kidney disease in African Americans (part **Ba**) show evidence of recent positive selection in Yorubans[86] (part **Bb**). *In vitro*, the G1 and G2 variants lyse subspecies of the *Trypanosoma* spp. pathogen that are resistant to wild-type APOL1 (part **Bc**). Arrows point to the swelling lysosome. cM, centimorgan; PMA, phorbol myristate acetate. Part **A** reprinted from *Cell*, **152**, Sharon R. Grossman *et al.*, Identifying recent adaptations in large-scale genomic data, 703–713, © (2013), with permission from Elsevier. Part **B** from Genovese, G. *et al.* Association of trypanolytic *ApoL1* variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010). Reprinted with permission from AAAS.

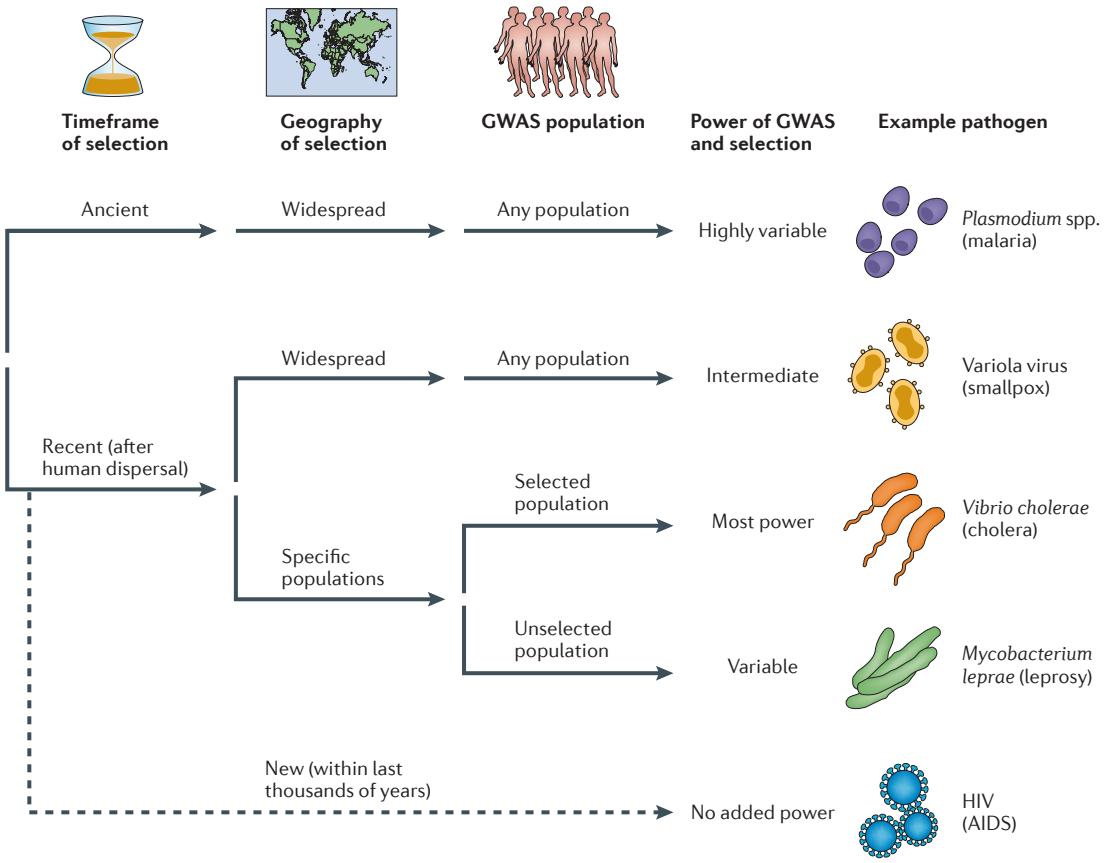


Figure 4 | **The power offered by combining natural selection with GWASs depends on the age of selection and populations chosen.** For pathogens that predate human dispersal from Africa, ancient and complex signals of selection are shared between human populations, and there are variable implications for genome-wide association studies (GWASs). For widespread pathogens that are more recent, the range of new resistance variants will be more limited, but selection is harder to detect when it is shared between populations. For recent pathogens that affect specific populations, GWASs in the selected populations will be particularly powerful, as causal variants will have been driven to high prevalence. Selection signals from one population may help to detect resistance loci in other unselected populations, but only if resistance variants arise in the same genetic loci. For GWASs of very new diseases, supplementing these studies with methods to detect selection will add no power, unless variants that confer resistance also protect against more ancient pathogens. Examples of pathogens matching each scenario are given on the right.

~340 adaptive nonsynonymous mutations in the 1000 Genomes Project data. Moreover, much (probably most) adaptive evolution occurred in regulatory, non-coding regions[100,101], and most recent selective sweeps are far from complete[102]. Thus, the actual number of positively selected loci could be much larger.

Under the framework of recent positive selection, one would anticipate that genetic variants conferring pathogen resistance that are moderately old (dating since the human migrations from Africa but at least thousands of years old), that were geographically limited in history and that exerted strong positive selective pressure will be most readily detected, provided that the studies are carried out in the population with the history of disease exposure. Many of the pathogen studies so far are imperfect fits for these criteria. Here, we review some of the prominent diseases of human history (TABLE 1), and discuss the strengths and limitations of investigating natural selection and association for these traits.

Malaria. Malaria is caused by obligate parasitic *Plasmodium* spp., which infects hundreds of millions of people and kills ~1 million children annually[103]. *P. falciparum* has afflicted humans for ~100,000 years, and a rapid upsurge of malaria ~10,000 years ago increased selective pressure on some human populations[27,104]. As a result, incidence of sickle cell disease and other inherited red blood cell disorders that are associated with malaria resistance (for example, α-thalassemia, glucose-6-phosphate dehydrogenase (G6PD) deficiency and ovalocytosis) coincides with the geographical distribution of malaria[27].

The presence of a disease in a population may indicate that the pathogen exerts selective pressure, but the inverse — absence of disease — can also be meaningful. Although *P. falciparum* is common in sub-Saharan Africa, *P. vivax* is noticeably absent. A mutation in the human *DARC* gene[105] that disrupts expression of the Duffy antigen receptor to prevent infection[106] has become 100% prevalent. In a possible example of

Table 1 | **Age and geographical origin of major human pathogens**

Pathogen	Disease	Pathogen type	Pathogen genome size (kb)	Place of origin	Approximate age of pathogen	Human mortality rate	Length of illness	GWAS?	Population used in GWAS and/or replication	GWAS refs
Plasmodium falciparum	Malaria	Protozoa	24,000	Africa before human dispersal	>100,000 years	2–30% for severe malaria	Variable	Yes	Ghanaian and Gambian	21, 111
Mycobacterium tuberculosis	Tuberculosis	Gram-positive bacteria	4,000	East Africa	40,000 years	10% develop active tuberculosis, of whom ~70% die	Years	Yes	Ghanaian, Gambian, Indonesian, Japanese, Malawian, Thai and Russian	76, 136, 199, 200
Variola virus	Smallpox	DNA virus	186	East Asia or Africa	15,000–70,000 years	1–30%	Weeks	Yes	European, African American and Hispanic	179, 180
Mycobacterium leprae	Leprosy	Gram-positive bacteria	33,000	East Africa or the Middle East	>10,000 years	Not typically lethal, but chronic infection reduces fertility	Years	Yes	Chinese	52, 126
Vibrio cholerae	Cholera	Gram-negative bacteria	4,000	Ganges River Delta	>5,000 years	5–50%	Days	No	NA	NA
HIV-1	AIDS	Lentiviral type of retrovirus	9.2	West and Central Africa	<100 years	100% (without treatment)	Years	Yes	European, African American, Hispanic, African and Malawian	49, 50, 201, 202

GWAS, genome-wide association study; NA, not available.

convergent evolution, an independent *DARC* mutation is prevalent in Southeast Asia where *P. vivax* is common[107]. As expected in the host–pathogen evolutionary 'arms race', the high prevalence of Duffy-negative hosts could be driving *P. vivax* to adapt. Strains of *P. vivax* have emerged in Africa that can infect Duffy-negative individuals[108] and that carry new variants of the gene encoding Duffy-binding protein[109].

The long history and strong selective pressure exerted by malaria may paradoxically complicate detection of resistance loci by GWASs. When selection drives multiple resistance variants to arise in a locus (for example, *HBB* and MHC), patterns of selection and association are more complex. Further complicating GWASs of malaria susceptibility, most cases occur in African populations, in which LD is short and highly variable between populations. Thus, causal variants are poorly tagged by SNP genotyping arrays[21,110]. Methods that successfully tackle these challenges are being developed, including GWASs by sequencing, improved imputation and Bayesian statistical approaches that allow heterogeneity in effect size and location[110–112]. An initial malaria GWAS in West Africa with 2,500 cases, 3,400 controls and 400,000 SNPs found no genome-wide significant associations[21]. Increasing the sample set to 3,500 cases, 4,300 controls and 800,000 SNPs (4 million after imputation) found *HBB*, *ABO* and two novel loci: one is intergenic and the other is in the *ATP2B4* gene, which encodes an erythrocyte calcium channel[111]. An international collaborative GWAS of severe malaria is likely to identify more loci[113].

Imputation
Statistical prediction of missing genetic data.

Leprosy. Leprosy is a chronic disease caused by the bacterium *Mycobacterium leprae* and causes infertility, disfiguring skin lesions, social ostracism, permanent physical disability and shortened lifespan[114–118]. *M. leprae* is an ancient and obligate human pathogen with a phylogeography that follows human migrations[119]. Although not immediately fatal, leprosy infection decreases fertility[118] by up to 85%[120], which poses a potent selective pressure. Consistent with selection that favours resistance variants, most people nowadays are genetically protected against *M. leprae*[121]. Indeed, host genetics has such a large role in susceptibility that, before the disease was proved to be caused by bacteria, leprosy was thought to be an inherited disease rather than an infectious one[122,123].

Leprosy was endemic in Europe with a prevalence of 10–40%[124] until the sixteenth century; its subsequent rapid decline is still unexplained. Currently, it remains

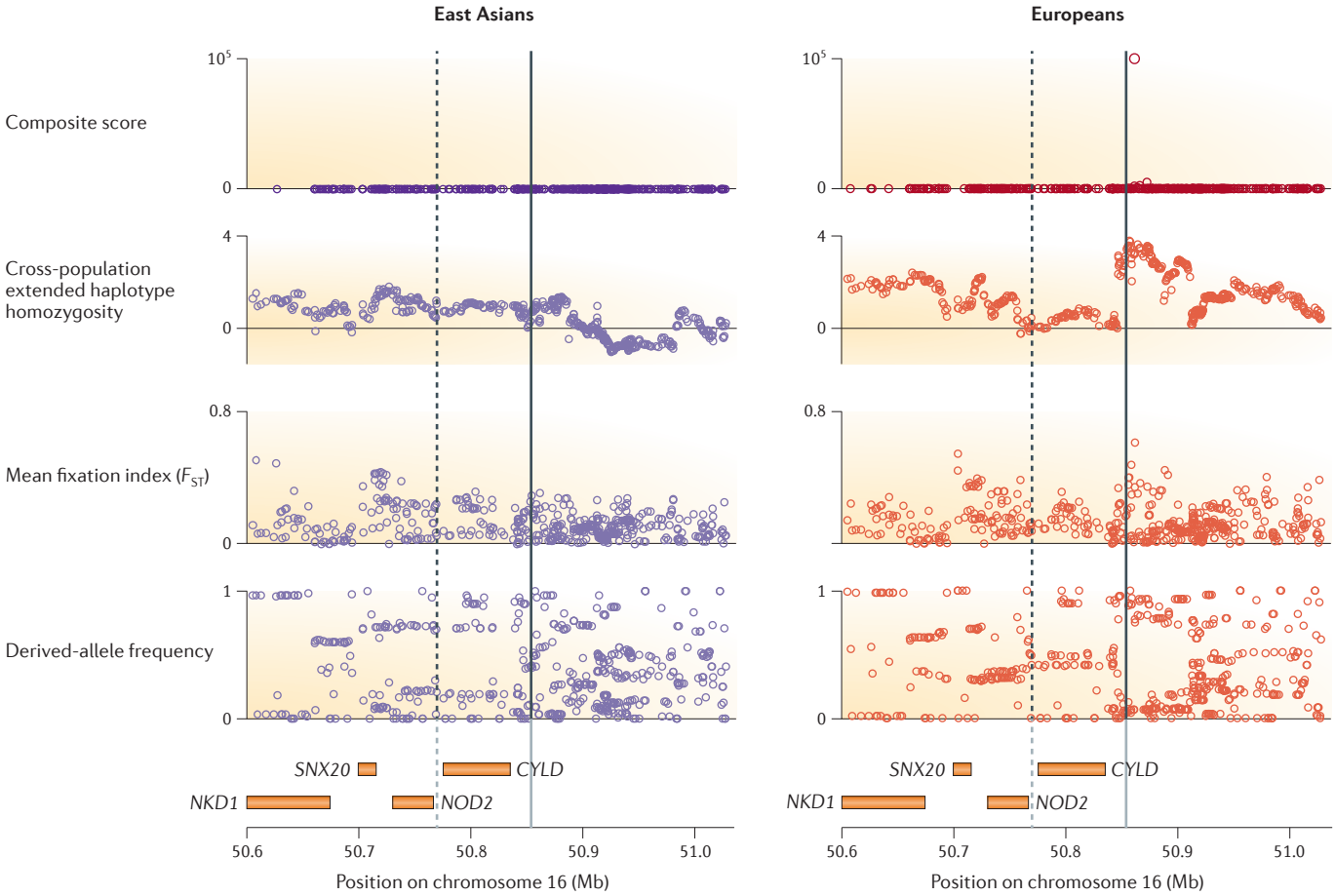


Figure 5 | **Signals of selection and association may differ between populations.** Two loci (represented by dashed and solid lines) associated with leprosy susceptibility in Han Chinese[52,126] show no evidence of positive selection in East Asians (blue circles). The association downstream of the cylindromatosis (*CYLD*) gene (solid line), but not the association near the bacterial pattern recognition receptor gene *NOD2* (dashed line), has signals of positive selection in Europeans (red circles). Selection scores for four different metrics were calculated using data from the 1000 Genomes Project and published in REF. 33. F_{ST}, Wright's fixation index; *NKD1*, naked cuticle homologue 1; *SNX20*, sorting nexin 20.

a major public health burden in India, China and South America[125]. The genetics of leprosy susceptibility differs between populations. GWASs in Han Chinese implicated the bacterial pattern recognition receptor NOD2 and three other components of NOD2-mediated innate immunity[52,126], but these associations did not replicate in an independent study of Indians and Malawians[127]. The Indian cohort instead had strong association to a functional knockout mutation in the *TLR1* gene that protects against leprosy[128]. This variant is nearly absent in the Han Chinese population (2%) and rare in Indians (9%) but extremely common in Europeans (70%)[129] owing to positive selection[130]. A locus associated with leprosy susceptibility in Chinese populations that is near the immune regulator gene cylindromatosis (*CYLD*)[131,132] may also be positively selected in Europeans (FIG. 5). Selection at leprosy-associated loci in European populations, together with the evidence from Indian and Chinese studies that risk loci differ between populations, suggests that combining European selection scans with a European GWAS might offer the most power. However, as leprosy is now rare in Europe, a GWAS would require a phenotypic proxy. Alternatively, as technology improves for analysing ancient DNA samples, thousands of medieval skeletal remains in Europe, many of which have lesions indicative of leprosy[124], may be usable in archaeological GWASs.

Tuberculosis. Tuberculosis is an often lethal disease caused by *Mycobacterium tuberculosis*, which infects one-third of the human population. Similar to the pathogens causing malaria and leprosy, *M. tuberculosis* is an ancient and obligate human pathogen that is estimated to have emerged in East Africa ~40,000 years ago[133] and to have spread around the world with ancient human migrations[134]. Consistent with long host–pathogen co-evolution, only 10% of infected individuals develop active disease; in most infected individuals the host immune system contains the pathogen. Host genetic factors are strongly implicated in susceptibility[135]. Despite this, a GWAS of tuberculosis susceptibility found only 2 significant associations in a huge data set

of 11 million SNPs and, after replication, nearly 23,000 individuals[76,136], compared with 490,000 SNPs and 9,200 individuals in the leprosy GWAS[136].

What explains the different outcomes of the tuberculosis and leprosy GWASs? One possible factor is that *M. tuberculosis* is more genetically diverse than *M. leprae*[134], and particular lineages of *M. tuberculosis* may have adapted to specific populations[137]. As individual tuberculosis status varies with how the host genetics interacts with the infecting *M. tuberculosis* strain, an accurate phenotype would include sequence data from the pathogen. In addition, tuberculosis resistance can be defined either as preventing or containing infection. Finally, co-infection with other pathogens increases the risk of developing active tuberculosis (for example, HIV infection increases risk by 21–34-fold[138]).

Phenotypic ambiguity, strain differences and strain–host interactions reduce the power of GWASs to associate host genetic factors with tuberculosis susceptibility[139], but the evidence of host–pathogen co-evolution suggests that incorporating tests of selection into GWASs may be particularly powerful. The strongest association so far is found in a GWAS carried out primarily in Africa, which implicates a region downstream of the Wilms tumour 1 (*WT1*) gene that has signatures of positive selection in East Asians[33,136].

AIDS. During the 1980s, the highly variable and fast evolving retroviral pathogen HIV-1 emerged to cause a global infectious disease pandemic that has killed >30 million people[140]. HIV-1 infects immune cells and causes a progressive and incurable failure of the immune system that allows other opportunistic infections, such as tuberculosis, to take hold. Both HIV-1 and the distantly related virus HIV-2 are predicted to be recent (<100 years) cross-species transmissions of simian immunodeficiency viruses (SIVs) into humans[141]. SIVs are mostly non-pathogenic species-specific lentiviruses carried by at least 41 African primate species[142]. Some SIVs have infected the same host for >30,000 years and probably much longer[143]. In the last century, at least ten primate-to-human SIV transmissions have been documented[142]. This suggests that human populations, particularly those in Africa, may have experienced ancient lentivirus epidemics and driven variants that confer resistance to modern HIV strains to prevalence through natural selection. The role of host genetics in HIV infection has been extensively researched, and the NIH Catalog of published GWASs lists at least 15 HIV-related publications (4 of which report significant associations of $P < 1 \times 10^{-8}$), but any connection to ancient selection remains unclear.

Among the first HIV resistance variants to be elucidated is a 32-base deletion in the cell surface receptor gene *CCR5* (known as *CCR5Δ32*) that prevents the expression of the receptor on T cells and confers complete HIV immunity on homozygous carriers[144,145]. On the bases of the high prevalence of the variant in northern Europe and the apparently high LD with nearby variants, some researchers hypothesized that *CCR5Δ32* was <1,000 years old and that positive selection for resistance to *Yersinia pestis* (the causative agent of the plague) increased its prevalence[146], which led to various follow-on studies. However, re-evaluation of the locus using newer and denser genetic maps found no evidence to support positive selection acting on this locus, which highlights the necessity of dense genome-wide data sets when identifying loci with exceptional patterns of variation[147]. Subsequent work on *CCR5* highlights the medical relevance of identifying naturally occurring resistance variants. Patients infused with autologous CD4 T cells that are engineered with a *CCR5*-disrupting mutation designed to phenotypically mimic *CCR5Δ32* showed partial resistance to HIV infection[148].

The strongest signals of HIV resistance are in the MHC loci, which are under extreme positive and balancing selection. In a GWAS comparing individuals infected with HIV who do not develop clinical disease (that is, HIV controllers) to individuals with advanced disease, MHC variants provided an explanation for 19% of the phenotypic variance[49]. They included a protective regulatory variant that is correlated with increased expression of human leukocyte antigen C (HLA-C)[149]. Although higher HLA-C expression protects against HIV progression, it also increases risk of the inflammatory disorder Crohn's disease[150], which highlights the potential for health repercussions of pathogen-driven selection.

Cholera. Cholera — a notoriously deadly disease with historic mortality rates as high as 50%[151–153] — is caused by the *V. cholerae* bacterium and endemic to the Ganges River Delta[154] of Bangladesh, where cholera is still prevalent now[155–157]. Host genetic factors strongly influence susceptibility[69,70] to this dangerous pathogen, which suggests selection favouring cholera resistance variants. Consistent with this hypothesis, the region has the world's lowest prevalence of blood type O, which is associated with an increased risk of severe cholera[69,70].

A genome-wide scan for positive selection in a Bangladeshi population identified >300 selected regions. The most strongly selected genes were also associated with cholera susceptibility in a targeted analysis[158]. A gene set enrichment analysis[159] found that two types of genes were statistically overrepresented in the selected regions compared with the rest of the genome: genes encoding potassium channels that are involved in cyclic AMP-mediated chloride secretion, and genes encoding components of the innate immune system that are involved in nuclear factor-κB (NF-κB) signalling. The success of the enrichment analysis suggests that cholera resistance in Bangladesh, similarly to pigmentation in Europe, provides an exceptionally strong evolutionary advantage and has driven selection at many different genomic loci.

The history of selection could make GWASs of cholera susceptibility in Bangladesh particularly powerful. As cholera is still common in Bangladesh, unlike leprosy in Europe, such a study is feasible. The results could help to elucidate the power for mapping positively selected resistance variants that protect against other pathogens with geographical disparity and high mortality.

Gene set enrichment
Overrepresentation of an a priori defined group of genes.

Norovirus. The single-stranded RNA viruses of the genus Norovirus are the leading cause of extremely contagious viral gastroenteritis outbreaks worldwide[160]; they are particularly dangerous to young children and cause up to 200,000 deaths each year in developing countries[161]. The origin of Norovirus is obscured by the rapid evolution of these viruses[162–164], but complex signals of selection in humans suggest that they could be very old. Individuals who are homozygous for null mutations of the fucosyltransferase 2 (*FUT2*) gene do not secrete ABO antigens and are protected against some strains[165–167]. Non-secretors are common worldwide (for example, in 20% of Caucasians). The underlying *FUT2* mutational spectrum is unexpectedly complex, as it is comprised of multiple independent mutations that vary in frequency between populations and that have diverse evolutionary signatures, from long-term balancing selection to recent positive selection[168].

Influenza. Great pandemics inflict massive mortality and are of particular interest to evolutionary geneticists. The most striking modern example is the 1918 influenza pandemic, which was caused by an unusually lethal strain of influenza A that killed 50–100 million people, including many previously healthy young adults[169,170]. Influenza is almost certainly very old: Hippocrates described a flu-like illness ~2,400 years ago[171]. Observational data suggest that host genetics influences susceptibility to severe illness[172]; for example, cases of the highly pathogenic H5N1 strain show strong familial aggregation[173]. Recently, interferon-induced transmembrane proteins (IFITMs) have been implicated in resistance to influenza A. IFITMs inhibit *in vitro* replication of some pathogenic viruses[174], and *IFITM3* expression protects against infection by multiple strains of influenza A *in vitro* and *in vivo*[175]. Hospitalized patients with severe influenza were significantly more likely to carry a splice acceptor site variant in *IFITM3* that reduces its ability to restrict influenza virus replication *in vitro*[175]. Versions of *IFITM3* that protect against influenza seem likely to confer a selective advantage, and selection scans show signals of recent positive selection in the *IFITM3* region[175].

Smallpox. Only a century ago, smallpox — caused by the Variola virus — ravaged human societies with mortality rates of up to 30%. It was an ancient and widespread scourge, and was described in historical records thousands of years old from China, India and Egypt. It is now gone and represents the only infectious disease in humans that has been eradicated by modern medicine. Variola virus has a highly conserved (>99.6% across 45 isolates) 186-kb double-stranded DNA genome[176]. Its extremely low mutation rate, simple genetic makeup and reliance on humans as its only host limited its ability to adapt and facilitated its eradication.

The age and phylogeography of smallpox is unresolved despite efforts to integrate historical records with sequence data from 45 viral isolates[176,177]. We noted that, for 32 viral isolates with documented mortality, death rates are lower in Africa (0.4–12%) than elsewhere (4–38%), even though all isolates were from a single phylogenetic clade[178]. This is consistent with selection for resistance in Africa, where the smallpox virus is predicted to have evolved from a rodent-borne ancestor tens of thousands of years ago and where outbreaks of other poxviruses continue nowadays. Human evolutionary history may help to clarify the origins of smallpox.

With smallpox eradicated, vaccine response is used as a crude phenotypic proxy for studying host resistance. Two GWASs that included European, African American and Hispanic populations identified 37 SNPs associated with cytokine response to vaccination[179,180] ($P < 1 \times 10^{-8}$). Most of the significant associations (65%) were found in African Americans, even though their sample size was half of that of the European cohorts, which is consistent with a larger effect due to selection in Africa. These results are preliminary: the studies had relatively small sample sizes (~200 African Americans), no overlap in their results and no replication. Incorporating tests for natural selection could add power for detecting true associations.

Infectious disease selection and common disease

The hygiene hypothesis proposes that autoimmune disorders are partly caused by differences between the pathogen-rich environment in which our immune system evolved and the more sterile modern world. In the absence of diverse pathogens from which to defend ourselves, our immune responses may turn on us[12]. Loci associated with common inflammatory disorders are enriched for signals of positive selection[1,181–183], and GWASs have proved particularly powerful for this class of diseases[184]. Elucidating the effect of ancient selection for pathogen resistance should help to decipher the aetiology of autoimmune diseases[89] but will require more data on immune responses to common pathogens. In cases in which selected variants have pleiotropic effects, pathogen-driven selection may even underlie diseases with no apparent immune component.

Inflammatory bowel disease. Inflammatory bowel disease (IBD) is a group of disorders, including Crohn's disease and ulcerative colitis, that are caused by autoimmune attacks on the gastrointestinal system. One hundred and sixty-three distinct loci have been significantly associated with IBD risk using meta-analyses of up to 75,000 European cases and controls, and these loci are strongly enriched for signatures of selection[11,117]. Moreover, seven of the eight leprosy susceptibility loci[52,126] are also associated with increased IBD risk[54,125]. Risk allele frequencies at some IBD loci correlate with local pathogen diversity, which is consistent with pathogen-driven selection[185].

These observations broadly support the hygiene hypothesis and connect autoimmunity to ancient evolution for pathogen resistance. However, the relationship is not straightforward. Of the four leprosy risk loci that precisely overlap IBD association peaks[11], the IBD risk variant is associated with decreased leprosy risk at only two loci; the other two associate with increased risk. Further complicating the story, one of those two — the

Pleiotropic effects
Effects on multiple unrelated phenotypes.

variant in *NOD2* — is associated with both an increased risk of Crohn's disease and protection against ulcerative colitis[186].

One potential source of the seemingly discrepant GWAS results is population differences. Whereas the 163 IBD loci were identified in cohorts of European ancestry, the leprosy GWASs were carried out in East Asian populations. The NOD2 pathway association with leprosy resistance in East Asians has not been replicated in other populations[127]. In addition, East Asians rarely carry the functional knockout mutation in *TLR1* that is common in Europe. Experimental data suggest that TLR1 and NOD2 activate distinct pathways in response to leprosy infection[187]; thus, correlating East Asian pathogen resistance variants with autoimmune disease risk in Europeans may not be informative.

A second factor is simply the lack of data on many pathogens. The remarkable overlap between GWAS loci of leprosy and IBD may reflect a bias in the available data, as leprosy is only one of the few pathogen susceptibility GWASs completed. Not a single GWAS has been carried out, for example, on susceptibility to parasitic worms, which are potentially of great relevance to gastrointestinal disorders such as IBD.

Coeliac disease. Coeliac disease is a strongly heritable[188] (~80%) inflammatory intestinal disorder triggered by gluten consumption. Despite severely affecting nutritional intake, coeliac disease occurs at 1–2% in Europe[189] and up to 6% in North African Sahrawi[190]. Loci associated with coeliac disease have signatures of positive selection[181]. A functional analysis of one selected locus in the SH2B adaptor protein 3 (*SH2B3*) gene found that individuals who are homozygous for the coeliac risk allele (~22% of the European population) have stronger activation of the NOD2 pathway and a 3–5-fold higher pro-inflammatory cytokine response to lipopolysaccharide[191]. Better protection against bacterial infection may have conferred a selective advantage that outweighed the increased risk of coeliac disease risk. Inferring selection pressure is problematic, as gluten consumption and thus the selection against coeliac disease probably changed with agriculture. A crude estimate of the age of the *SH2B3* variant based simply on haplotype length suggested that it was very recent[191] (<2,000 years old). However, simulations suggest that the long haplotype tests used to detect the selection at *SH2B3* are sensitive to selection events ~5,000–50,000 years ago[28], which implies that the *SH2B3* selection could date to either before or after the spread of agriculture through Europe ~10,000 years ago[192]. Accurately dating selected variants is challenging and requires methods that can both estimate multiple parameters and test various ancestry models, such as Approximate Bayesian Computation[193].

Non-autoimmune disease: kidney disease. African Americans suffer from kidney disease — including focal segmental glomerulosclerosis (FSGS) and hypertension-attributed end-stage kidney disease (H-ESKD) — at higher rates than European Americans. A region around the myosin heavy chain gene *MYH9* was associated with FSGS and H-ESKD in African Americans, but no causal variants were found[194,195]. One study[86] expanded the search to include an adjacent signal of African positive selection at the *APOL1* gene. Using data from the 1000 Genomes Project, they tested all polymorphisms with large frequency differences between Africans and Europeans in this expanded interval and identified two independent coding variants in *APOL1* that are strongly associated with FSGS (odds ratio = 10.5) and H-ESKD (odds ratio = 7.3). *In vitro* assays showed that the kidney disease-associated variants lyse *T. brucei rhodesiense*, which is the trypanosome parasite that causes the most acute, virulent form of sleeping sickness (FIG. 3B). The authors propose that ancient selection for resistance to sleeping sickness or a related pathogen in Africa contributes to the high rates of kidney disease in African Americans.

Conclusion

One of the oldest topics in genetics — natural selection for pathogen resistance — is being transformed by high-throughput biotechnology that offers unprecedented power to examine genome evolution. New research finds that both the most ancient signals of balancing selection (on cell surface glycoproteins) and some of the clearest signals of recent positive selection (on TLRs) implicate pathogens as the strongest selective pressure to drive the evolution of modern humans. Incorporating ancient history into disease susceptibility studies will identify functional variants and elucidate cellular mechanisms, and facilitate the development of new therapies for a surprisingly wide range of human illnesses. The range of diseases affected by ancient pathogen-driven selection extends beyond infectious diseases and immune-mediated disorders. Inflammatory and immunity genes are associated with psychiatric diseases, including schizophrenia[196] and autism[197], and the role of host genetics in maintaining a healthy microbiome is only starting to be examined[14]. At the dawn of genomic medicine, our ancient evolutionary history is one of our most powerful resources for understanding human biology towards improving human health.

1. Fumagalli, M. *et al.* Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* **7**, e1002355 (2011).
This study of genetic variation in 50 worldwide populations reveals that pathogens are primary drivers of local adaptation.
2. Polgar, S. in *Horizons of Anthropology* (ed. Tax, S.) (Aldine, 1964).
3. Armelagos, G. J., Barnes, K. C. & Lin, J. Disease in human evolution: the re-emergence of infectious disease in the third epidemiological transition. *AnthroNotes* **18**, 1–7 (1996).
4. Bocquet-Appel, J. P. When the world's population took off: the springboard of the Neolithic demographic transition. *Science* **333**, 560–561 (2011).
5. Vannberg, F. O., Chapman, S. J. & Hill, A. V. Human genetic susceptibility to intracellular pathogens. *Immunol. Rev.* **240**, 105–116 (2011).
6. Chapman, S. J. & Hill, A. V. Human genetic susceptibility to infectious disease. *Nature Rev. Genet.* **13**, 175–188 (2012).
7. Anderson, R. M. & May, R. M. Co-evolution of hosts and parasites. *Parasitology* **85**, 411–426 (1982).
8. Cohen, M. L. Changing patterns of infectious disease. *Nature* **406**, 762–767 (2000).
9. Cagliani, R. *et al.* Crohn's disease loci are common targets of protozoa-driven selection. *Mol. Biol. Evol.* **30**, 1077–1087 (2013).

10. Okada, H., Kuhn, C., Feillet, H. & Bach, J. F. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clin. Exp. Immunol.* **160**, 1–9 (2010).
11. Jostins, L. *et al.* Host–microorganism interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012). **This paper presents selection for pathogen resistance in IBD.**
12. Sironi, M. & Clerici, M. The hygiene hypothesis: an evolutionary perspective. *Microbes. Infect.* **12**, 421–427 (2010).
13. Lin, A. *et al.* Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States. *PLoS ONE* **8**, e53838 (2013).
14. Honda, K. & Littman, D. R. The microbiome in infectious disease and inflammation. *Annu. Rev. Immunol.* **30**, 759–795 (2012).
15. Segal, S. & Hill, A. V. Genetic susceptibility to infectious disease. *Trends Microbiol.* **11**, 445–448 (2003).
16. Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19**, 711–722 (2009).
17. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
18. Marx, V. Biology: the big challenges of big data. *Nature* **498**, 255–260 (2013).
19. Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
20. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **27**, 2534–2547 (2010).
21. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genet.* **41**, 657–665 (2009).
22. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
23. Ronald, J. & Akey, J. M. Genome-wide scans for loci under selection in humans. *Hum. Genom.* **2**, 113–125 (2005).
24. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111 (2003).
25. Fu, W. & Akey, J. M. Selection and adaptation in the human genome. *Annu. Rev. Genom. Hum. Genet.* **14**, 467–489 (2013).
26. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**, 97–120 (2013).
27. Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77**, 171–192 (2005). **This study shows that investigating targets of pathogen-driven selection leads to immunological discoveries and possible new therapies.**
28. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
29. Deagle, B. E. *et al.* Population genomics of parallel phenotypic evolution in stickleback across stream–lake ecological transitions. *Proc. Biol. Sci.* **279**, 1277–1286 (2012).
30. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
31. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
32. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208 –R215 (2010).
33. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
34. Granka, J. M. *et al.* Limited evidence for classic selective sweeps in African populations. *Genetics* **192**, 1049–1064 (2012).
35. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
36. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
37. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
38. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
39. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
40. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90 (2007).
41. Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nature Genet.* **40**, 340–345 (2008).
42. Bhatia, G. *et al.* Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.* **89**, 368–381 (2011).
43. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
44. McClelland, E. E., Penn, D. J. & Potts, W. K. Major histocompatibility complex heterozygote superiority during co-infection. *Infect. Immun.* **71**, 2079–2086 (2003).
45. Aguilar, A. *et al.* High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proc. Natl. Acad. Sci. USA* **101**, 3490–3494 (2004).
46. Prugnolle, F. *et al.* Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15**, 1022–1027 (2005).
47. de Bakker, P. I. & Raychaudhuri, S. Interrogating the major histocompatibility complex with high-throughput genomics. *Hum. Mol. Genet.* **21**, R29–R36 (2012).
48. Hindorff, L. A. *et al.* Potential aetiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
49. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
50. Fellay, J. *et al.* Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* **5**, e1000791 (2009).
51. Limou, S. *et al.* Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* **199**, 419–426 (2009).
52. Zhang, F. R. *et al.* Genomewide association study of leprosy. *N. Engl. J. Med.* **361**, 2609–2618 (2009).
53. LeishGEN Consortium *et al.* Common variants in the *HLA-DRB1–HLA-DQA1* HLA class II region are associated with susceptibility to visceral leishmaniasis. *Nature Genet.* **45**, 208–213 (2013).
54. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the *HLA-DP* locus associated with chronic hepatitis B in Asians. *Nature Genet.* **41**, 591–595 (2009).
55. Mbarek, H. *et al.* A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. *Hum. Mol. Genet.* **20**, 3884–3892 (2011).
56. Nishida, N. *et al.* Genome-wide association study confirming association of *HLA-DP* with protection against chronic hepatitis B and viral clearance in Japanese and Korean. *PLoS ONE* **7**, e39175 (2012).
57. Duggal, P. *et al.* Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Ann. Intern. Med.* **158**, 235–245 (2013).
58. Chen, D. *et al.* Genome-wide association study of HPV seropositivity. *Hum. Mol. Genet.* **20**, 4714–4723 (2011).
59. Hanchard, N. A. *et al.* Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* **78**, 153–159 (2006).
60. Hanchard, N. *et al.* Classical sickle β-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genet.* **8**, 52 (2007).
61. Andres, A. M. *et al.* Targets of balancing selection in the human genome. *Mol. Biol. Evol.* **26**, 2755–2764 (2009).
62. Klein, J., Satta, Y., O'hUigin, C. & Takahata, N. The molecular descent of the major histocompatibility complex. *Annu. Rev. Immunol.* **11**, 269–295 (1993).
63. Segurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl. Acad. Sci. USA* **109**, 18493–18498 (2012).
64. Leffler, E. M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).
65. Olofsson, S. & Bergstrom, T. Glycoconjugate glycans as viral receptors. *Ann. Med.* **37**, 154–172 (2005).
66. Day, C. J., Semchenko, E. A. & Korolik, V. Glycoconjugates play a key role in *Campylobacter jejuni* infection: interactions between host and pathogen. *Front. Cell. Infect. Microbiol.* **2**, 9 (2012).
67. Ko, W.-Y. *et al.* Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *Am. J. Hum. Genet.* **88**, 741–754 (2011).
68. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
69. Barua, D. & Paguio, A. S. ABO blood groups and cholera. *Ann. Hum. Biol.* **4**, 489–492 (1977).
70. Harris, J. B. *et al.* Susceptibility to *Vibrio cholerae* Infection in a cohort of household contacts of patients with cholera in Bangladesh. *PLoS Negl. Trop. Dis.* **2**, e221 (2008).
71. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
72. Hong, E. P. & Park, J. W. Sample size and statistical power calculation in genetic association studies. *Genom. Inform* **10**, 117–122 (2012).
73. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
74. Hill, A. V. Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Phil. Trans. R. Soc. B* **367**, 840–849 (2012).
75. Siontis, K. C., Patsopoulos, N. A. & Ioannidis, J. P. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur. J. Hum. Genet.* **18**, 832–837 (2010).
76. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nature Genet.* **42**, 739–741 (2010).
77. Qu, H. Q., Li, Q., McCormick, J. B. & Fisher-Hoch, S. P. What did we learn from the genome-wide association study for tuberculosis susceptibility? *J. Med. Genet.* **48**, 217–218 (2011).
78. Qu, H. Q., Fisher-Hoch, S. P. & McCormick, J. B. Knowledge gaining by human genetic studies on tuberculosis susceptibility. *J. Hum. Genet.* **56**, 177–182 (2011).
79. Kaslow, R. A., McNicholl, J. & Hill, A. V. S. *Genetic Susceptibility to Infectious Diseases* (Oxford Univ. Press, 2008).
80. Roeder, K., Bacanu, S. A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–252 (2006).
81. Ayodo, G. *et al.* Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am. J. Hum. Genet.* **81**, 234–242 (2007). **This paper shows that signals of positive selection can increase power to detect associations.**
82. Park, D. J. *et al.* Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. *Proc. Natl. Acad. Sci. USA* **109**, 13052–13057 (2012).
83. Daub, J. T. *et al.* Evidence for polygenic adaptation to pathogens in the human genome. *Mol. Biol. Evol.* **30**, 1544–1558 (2013).
84. Lee, P. H. *et al.* Multi-locus genome-wide association analysis supports the role of glutamatergic synaptic transmission in the aetiology of major depressive disorder. *Transl. Psychiatry* **2**, e184 (2012).
85. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Rev. Genet.* **11**, 843–854 (2010).
86. Genovese, G. *et al.* Association of trypanolytic *ApoL1* variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
87. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

88. Kudaravalli, S., Veyrieras, J. B., Stranger, B. E., Dermitzakis, E. T. & Pritchard, J. K. Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.* **26**, 649–658 (2009).
89. Raj, T. *et al.* Common risk alleles for inflammatory diseases are targets of recent positive selection. *Am. J. Hum. Genet.* **92**, 517–529 (2013).
This paper presents a systems-based analysis that integrates GWAS, selection, functional data and eQTL mapping.
90. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
91. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
92. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotech.* **30**, 271–277 (2012).
93. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
94. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
95. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).
96. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14**, 618–630 (2013).
97. Schneider, T., Kreutz, J. & Chiu, D. T. The potential impact of droplet microfluidics in biology. *Anal. Chem.* **85**, 3476–3482 (2013).
98. Shalem, O. *et al.* Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
99. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR–Cas9 system. *Science* **343**, 80–84 (2014).
100. Hindorff, L. A. *et al.* Potential aetiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
101. Wang, Y. *et al.* The lactase persistence/non-persistence polymorphism is controlled by a *cis*-acting element. *Hum. Mol. Genet.* **4**, 657–662 (1995).
102. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5**, e1000500 (2009).
103. World Health Organization. World Malaria Report 2011. *WHO* [online], http://www.who.int/malaria/world_malaria_report_2011/ (2011).
104. Hartl, D. L. The origin of malaria: mixed messages from genetic diversity. *Nature Rev. Microbiol.* **2**, 15–22 (2004).
105. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, *FyFy*. *N. Engl. J. Med.* **295**, 302–304 (1976).
106. Tournamille, C., Colin, Y., Cartron, J. P. & Le Van Kim, C. Disruption of a GATA motif in the *Duffy* gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nature Genet.* **10**, 224–228 (1995).
107. Shimizu, Y. *et al.* Sero- and molecular typing of *Duffy* blood group in Southeast Asians and Oceanians. *Hum. Biol.* **72**, 511–518 (2000).
108. Menard, D. *et al. Plasmodium vivax* clinical malaria is commonly observed in *Duffy*-negative Malagasy people. *Proc. Natl. Acad. Sci. USA* **107**, 5967–5971 (2010).
109. Menard, D. *et al.* Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains. *PLoS Negl. Trop. Dis.* **7**, e2489 (2013).
110. Teo, Y.-Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature Rev. Genet.* **11**, 149–160 (2010).
111. Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* **489**, 443–446 (2012).
112. Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.* **9**, e1003509 (2013).
This study presents new methods for tackling complex GWASs.
113. Achidi, E. A. *et al.* A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732–737 (2008).
114. van Brakel, W. H. Measuring leprosy stigma — a preliminary review of the leprosy literature. *Int. J. Lepr. Other Mycobact. Dis.* **71**, 190–197 (2003).
115. Rao, P. S. *et al.* Disability adjusted working life years (DAWLYs) of leprosy affected persons in India. *Indian J. Med. Res.* **137**, 907–910 (2013).
116. Guinto, R. S., Doull, J. A. & De Guia, L. Mortality of persons with leprosy before sulphone therapy, Cordova and Talisay, Cebu, Philippines. *Int. J. Lepr* **22**, 273–284 (1954).
117. Saporta, L. & Yuksel, A. Androgenic status in patients with lepromatous leprosy. *Br. J. Urol.* **74**, 221–224 (1994).
118. Leal, A. M. & Foss, N. T. Endocrine dysfunction in leprosy. *Eur. J. Clin. Microbiol. Infect. Dis.* **28**, 1–7 (2009).
119. Monot, M. *et al.* Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nature Genet.* **41**, 1282–1289 (2009).
120. Smith, D. G. & Guinto, R. S. Leprosy and fertility. *Hum. Biol.* **50**, 451–460 (1978).
121. Jacobson, R. R. & Krahenbuhl, J. L. Leprosy. *Lancet* **353**, 655–660 (1999).
122. Alter, A., Alcais, A., Abel, L. & Schurr, E. Leprosy as a genetic model for susceptibility to common infectious diseases. *Hum. Genet.* **123**, 227–235 (2008).
123. Shields, E. D., Russell, D. A. & Pericak-Vance, M. A. Genetic epidemiology of the susceptibility to leprosy. *J. Clin. Invest.* **79**, 1139–1143 (1987).
124. Boldsen, J. L. Leprosy in mediaeval Denmark — osteological and epidemiological analyses. *Anthropol. Anz.* **67**, 407–425 (2009).
125. World Health Organization. Global leprosy situation. *Weekly Epidemiol. Record* **87**, 317–328 [online], http://www.who.int/lep/resources/wer/en/ (2012).
126. Zhang, F. *et al.* Identification of two new loci at *IL23R* and *RAB32* that influence susceptibility to leprosy. *Nature Genet.* **43**, 1247–1251 (2011).
127. Wong, S. H., Hill, A. V., Vannberg, F. O. Genomewide association study of leprosy. *N. Engl. J. Med.* **362**, 1446–1447; author reply 1447–1448 (2010).
128. Johnson, C. M. *et al.* Cutting edge: a common polymorphism impairs cell surface trafficking and functional responses of TLR1 but protects against leprosy. *J. Immunol.* **178**, 7520–7524 (2007).
129. Wong, S. H. *et al.* Leprosy and the adaptation of human Toll-like receptor 1. *PLoS Pathog.* **6**, e1000979 (2010).
This paper shows selection for a leprosy protective variant in *TLR1* in Europeans, which suggests a long host–pathogen relationship.
130. Barreiro, L. B. *et al.* Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* **5**, e1000562 (2009).
131. Reiley, W. W. *et al.* Regulation of T cell development by the deubiquitylating enzyme CYLD. *Nature Immunol.* **7**, 411–417 (2006).
132. Brummelkamp, T. R., Nijman, S. M., Dirac, A. M. & Bernards, R. Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-κB. *Nature* **424**, 797–801 (2003).
133. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4**, e1000160 (2008).
134. Hershberg, R. *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* **6**, e311 (2008).
135. Moller, M. & Hoal, E. G. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis* **90**, 71–83 (2010).
136. Thye, T. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nature Genet.* **44**, 257–259 (2012).
137. Gagneux, S. *et al.* Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103**, 2869–2873 (2006).
138. World Health Organization. Global Tuberculosis Control 2011. *WHO* [online], http://www.who.int/tb/publications/global_report/2011 (2011).
139. Stein, C. M. Genetic epidemiology of tuberculosis susceptibility: impact of study design. *PLoS Pathog.* **7**, e1001189 (2011).
140. Global AIDS epidemic facts and figures. *UNAIDS* [online], http://www.unaids.org/en/resources/presscentre/factsheets/ (2013).
141. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664 (2008).
142. VandeWoude, S. & Apetrei, C. Going wild: lessons from naturally occurring T lymphotropic lentiviruses. *Clin. Microbiol. Rev.* **19**, 728–762 (2006).
143. Worobey, M. *et al.* Island biogeography reveals the deep history of SIV. *Science* **329**, 1487 (2010).
144. Liu, R. *et al.* Homozygous defect in HIV-1 co-receptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367–377 (1996).
145. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CKR5* structural gene. *Science* **273**, 1856–1862 (1996).
146. Stephens, J. C. *et al.* Dating the origin of the *CCR5–Δ32* AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62**, 1507–1515 (1998).
147. Sabeti, P. C. *et al.* The case for selection at *CCR5-Δ32*. *PLoS Biol.* **3**, e378 (2005).
148. Tebas, P. *et al.* Gene editing of *CCR5* in autologous CD4 T cells of persons infected with HIV. *N. Engl. J. Med.* **370**, 901–910 (2014).
149. Thomas, R. *et al.* HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nature Genet.* **41**, 1290–1294 (2009).
150. Apps, R. *et al.* Influence of HLA-C expression level on HIV control. *Science* **340**, 87–91 (2013).
151. Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T. & Calderwood, S. B. Cholera. *Lancet* **379**, 2466–2476 (2012).
152. Sack, D. A., Sack, R. B., Nair, G. B. & Siddique, A. K. Cholera. *Lancet* **363**, 223–233 (2004).
153. Harris, J. B. *et al.* Cholera's western front. *Lancet* **376**, 1961–1965 (2010).
154. Lee, K. The global dimensions of cholera. *Global Change Hum. Health* **2**, 6–17 (2001).
155. Chowdhury, F. *et al.* Impact of rapid urbanization on the rates of infection by *Vibrio cholerae* O1 and enterotoxigenic *Escherichia coli* in Dhaka, Bangladesh. *PLoS Negl. Trop. Dis* **5**, e999 (2011).
156. Mosley, W. H., McCormack, W. M., Ahmed, A., Chowdhury, A. K. & Barui, R. K. Report of the 1966–1967 cholera vaccine field trial in rural East Pakistan. 2. Results of the serological surveys in the study population — the relationship of case rate to antibody titre and an estimate of the inapparent infection rate with *Vibrio cholerae*. *Bull. World Health Organ.* **40**, 187–197 (1969).
157. Glass, R. I. *et al.* Seroepidemiological studies of El Tor cholera in Bangladesh: association of serum antibody levels with protection. *J. Infect. Dis.* **151**, 236–242 (1985).
158. Karlsson, E. K. *et al.* Natural selection in a Bangladeshi population from the cholera-endemic Ganges River Delta. *Sci. Transl. Med.* **5**, 192ra86 (2013).
This study uses selection for host resistance to historically localized pathogen to investigate immune response pathways.
159. Lee, P. H., O'Dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797–1799 (2012).
160. Hall, A. J. Noroviruses: the perfect human pathogens? *J. Infect. Dis.* **205**, 1622–1624 (2012).
161. Patel, M. M. *et al.* Systematic literature review of role of noroviruses in sporadic gastroenteritis. *Emerg. Infect. Dis.* **14**, 1224–1231 (2008).
162. Wertheim, J. O. & Kosakovsky Pond, S. L. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* **28**, 3355–3365 (2011).
163. Worobey, M., Bjork, A. & Wertheim, J. O. Point, counterpoint: the evolution of pathogenic viruses and their human hosts. *Annu. Rev. Ecol. Evol. Syst.* **38**, 515–540 (2007).
164. Emerman, M. & Malik, H. S. Paleovirology — modern consequences of ancient viruses. *PLoS Biol.* **8**, e1000301 (2010).
165. Lindesmith, L. *et al.* Human susceptibility and resistance to Norwalk virus infection. *Nature Med.* **9**, 548–553 (2003).
166. Carlsson, B. *et al.* The G428A nonsense mutation in *FUT2* provides strong but not absolute protection against symptomatic GII.4 Norovirus infection. *PLoS ONE* **4**, e5593 (2009).
167. Nordgren, J., Kindberg, E., Lindgren, P.-E., Matussek, A. & Svensson, L. Norovirus gastroenteritis outbreak with a secretor-independent susceptibility pattern, Sweden. *Emerg. Infecti. Diseases* **16**, 81 (2010).
168. Ferrer-Admetlla, A. *et al.* A natural history of *FUT2* polymorphism in humans. *Mol. Biol. Evol.* **26**, 1993–2003 (2009).
169. Taubenberger, J. K. & Morens, D. M. Influenza revisited. *Emerg. Infecti. Diseases* **12**, 1 (2006).

170. Johnson, N. P. & Mueller, J. Updating the accounts: global mortality of the 1918–1920 "Spanish" influenza pandemic. *Bull. Hist. Med.* **76**, 105–115 (2002).
171. Martin, P. M. & Martin-Granel, E. 2,500 year evolution of the term epidemic. *Emerg. Infect. Dis.* **12**, 976–980 (2006).
172. Albright, F. S., Orlando, P., Pavia, A. T., Jackson, G. G. & Cannon Albright, L. A. Evidence for a heritable predisposition to death due to influenza. *J. Infect. Dis.* **197**, 18–24 (2008).
173. Horby, P., Nguyen, N. Y., Dunstan, S. J. & Baillie, J. K. The role of host genetics in susceptibility to influenza: a systematic review. *PLoS ONE* **7**, e33180 (2012).
174. Brass, A. L. *et al.* The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell* **139**, 1243–1254 (2009).
175. Everitt, A. R. *et al.* IFITM3 restricts the morbidity and mortality associated with influenza. *Nature* **484**, 519–523 (2012).
176. Li, Y. *et al.* On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc. Natl. Acad. Sci. USA* **104**, 15787–15792 (2007).
177. Babkin, I. V. & Shelkunov, S. N. [Molecular evolution of poxviruses]. *Genetika* **44**, 1029–1044 (in Russian) (2008).
178. Esposito, J. J. *et al.* Genome sequence diversity and clues to the evolution of variola (smallpox) virus. *Science* **313**, 807–812 (2006).
179. Ovsyannikova, I. G. *et al.* Genome-wide association study of antibody response to smallpox vaccine. *Vaccine* **30**, 4182–4189 (2012).
180. Kennedy, R. B. *et al.* Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Hum. Genet.* **131**, 1403–1421 (2012).
181. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Rev. Genet.* **11**, 17–30 (2010).
This paper shows that pathogen-driven selection shaped the human genome.
182. Casto, A. M. & Feldman, M. W. Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? *PLoS Genet.* **7**, e1001266 (2011).
183. Hancock, A. M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* **7**, e1001375 (2011).
184. Hu, X. & Daly, M. What have we learned from six years of GWAS in autoimmune diseases, and what is next? *Curr. Opin. Immunol.* **24**, 571–575 (2012).
185. Fumagalli, M. *et al.* Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *J. Exp. Med.* **206**, 1395–1408 (2009).
186. Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H. & Butte, A. J. Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.* **5**, e1000792 (2009).
187. Schenk, M. *et al.* NOD2 triggers an interleukin-32 dependent human dendritic cell program in leprosy. *Nature Med.* **18**, 555–563 (2012).
188. Greco, L. *et al.* The first large population based twin study of coeliac disease. *Gut* **50**, 624–628 (2002).
189. Dube, C. *et al.* The prevalence of coeliac disease in average-risk and at risk western European populations: a systematic review. *Gastroenterology,* **128**, S57– S67 (2005).
190. Catassi, C. *et al.* Why is coeliac disease endemic in the people of the Sahara? *Lancet* **354**, 647–648 (1999).
191. Zhernakova, A. *et al.* Evolutionary and functional analysis of coeliac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* **86**, 970–977 (2010).
192. Pinhasi, R., Fort, J. & Ammerman, A. J. Tracing the origin and spread of agriculture in Europe. *PLoS Biol.* **3**, e410 (2005).
193. Itan, Y., Powell, A., Beaumont, M. A., Burger, J. & Thomas, M. G. The origins of lactase persistence in Europe. *PLoS Comput. Biol.* **5**, e1000491 (2009).
194. Kao, W. H. *et al. MYH9* is associated with nondiabetic end-stage renal disease in African Americans. *Nature Genet.* **40**, 1185–1192 (2008).
195. Kopp, J. B. *et al. MYH9* is a major-effect risk gene for focal segmental glomerulosclerosis. *Nature Genet.* **40**, 1175–1184 (2008).
196. Sainz, J. *et al.* Inflammatory and immune response genes have significantly altered expression in schizophrenia. *Mol. Psychiatry* **18**, 1056–1057 (2013).
197. Onore, C., Careaga, M. & Ashwood, P. The role of immune dysfunction in the pathophysiology of autism. *Brain Behav. Immun.* **26**, 383–392 (2012).
198. Spencer, C. C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
199. Png, E. *et al.* A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC Med. Genet.* **13**, 5 (2012).
200. Mahasirimongkol, S. *et al.* Genome-wide association studies of tuberculosis in Asians identify distinct at risk locus for young tuberculosis. *J. Hum. Genet.* **57**, 363–367 (2012).
201. Petrovski, S. *et al.* Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. *AIDS* **25**, 513–518 (2011).
202. Pelak, K. *et al.* Host determinants of HIV-1 control in African Americans. *J. Infect. Dis.* **201**, 1141–1149 (2010).

Acknowledgements
The authors thank S. Schaffner, E. Brown, D. Park, D. Neafsey, R. LaRocque and J. Harris for discussions.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
Catalog of published GWASs: http://www.genome.gov/gwastudies

SUPPLEMENTARY INFORMATION
See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF