SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Power of Our Genome-wide CMS Test Compared to Previously Published Methods, Related to Figure 1

We compared the power of genome-wide CMS to previously published long-range haplotype tests (iHS and XP-EHH) using simulations developed by (Schaffner et al., 2005) for three different populations (CEU, CHB+JPT, YRI). In carrying out the analysis, we identified and corrected an error in the code that simulated gene conversion during a selective sweep. Since power analysis on the long-range haplotype tests were previously published using the code with this error, we are presenting our power comparisons for both the (A) original and (B) corrected simulations across a wide range of causal allele ages (5,000 – 35,000 years). For both the original and the corrected version of simulations, CMS (red) provides power comparable to, and often better than the long-range haplotype methods (iHS: green and XP-EHH: blue). The corrected simulations demonstrate the greater power of CMS and the long-range haplotype tests for nearly all derived allele ages, and particularly for older derived alleles (20,000 – 35,000 years)

SUPPLEMENTAL TABLE LEGENDS

Table S1. Regions identified by genome-wide CMS, **Related to Figure 1**
A list of regions identified in 1000G data by a genome-wide extension of the CMS test (CMS_{GW}), at an FPR of 0.1%.

Table S2. Localized CMS regions, **Related to Figure 1**
A list of novel and previously discovered regions under positive selection that have been localized using the CMS method. The localized regions have an average size of 62 kb, on the order of single genes.

Table S3. Enrichment of functional variants in selected regions, **Related to Figure 1**
Enrichment analysis of different classes of variants. Since our 412 selected regions each contain at most one selected variant and a median of 46 other variants, enrichment analysis for different classes of functional elements will not be well-powered to detect an over-representation of functional variants. Fold-enrichment indicates enrichment over the genomic average.

Table S4. Gene pathway enrichment in selected regions, **Related to Figure 1**
Enrichment analysis of specific gene pathways using INRICH. Enrichment analysis of regions under selection may be less powered than typical analyses of regions associated with a particular phenotype (e.g. from GWAS), since selection acts on a number of different phenotypes and pathways.

Table S5. lincRNAs in candidate selective regions, Related to Figure 3
A list of lincRNAs that overlap with regions under positive selection.

Table S6. Characterization of high-scoring non-synonymous SNPs, **Related to Figure 2**
A list of non-synonymous SNPs that were high-scoring in regions under positive selection.

Table S7. Candidate selective regions that overlap eQTLs, **Related to Figure 3**
A list of eQTLs that overlap with regions under positive selection, including variants that are high-scoring by both CMS and eQTL analysis.

Table S8. CMS high-scoring SNPs in active enhancers or promoters, **Related to Figure 3**
A list of CMS high-scoring SNPs that disrupt enhancers and promoters.

Table S9. Trait-associated CMS high-scoring SNPs , **Related to Figure 3**
SNPs that are high-scoring by CMS and have also been associated with resistance to leprosy or tuberculosis or with a phenotype in the NHGRI GWAS catalogue.

EXTENDED EXPERIMENTAL PROCEDURES

Coalescent Simulations

We used the simulations described earlier in (Grossman et al., 2010), with one change: a coding error was fixed in the code that simulated gene conversion during a selective sweep (neutral simulations were unaffected). The corrected simulations have improved the performance of long-range haplotype tests, such as iHS test, compared to earlier reports (Voight et al., 2006). We have thus included power calculations based on simulations with the bug and without for comprehensive comparison. We include a complete description of our simulations below for completeness.

We simulated population genetics datasets designed to mimic the 1000 genomes project data. Simulations were done using the *cosi* coalescent simulator(Schaffner et al., 2005), extended to simulate full and partial selective sweeps(Sabeti et al., 2007), and corrected as described above. *cosi* performs coalescent simulations similarly to the widely used *ms* tool(Hudson, 2002), but allows recombination rate variation along the region. Like *ms* and unlike some other coalescent simulators(Marjoram and Wall, 2006), *cosi* does not put any restrictions on which chromosomes within a population may coalesce (except during selective sweeps, when coalescence between chromosomes carrying selected and unselected alleles is forbidden).

There were two aspects to the simulation: the base neutral model, and the model of selective sweep.

Base neutral model

For the neutral simulations, we used a demographic model previously shown to replicate HapMap data on several metrics (allele frequency spectrum, relationship between allele frequency and ancestral state, Fst, r^2, fraction of marker pairs with D'=1, and heterozygosity)[49]. The model has been used in many previous studies(He et al., 2012; Sabeti et al., 2007). The model includes three present-day populations -- West African, East Asian and European -- with present-day effective sizes of 24000, 7700 and 7700 respectively. The populations were formed via the following history: an ancestral population (effective size 12,500, expanding to 24,000 at time 17,000 generations ago) split into an African and a Eurasian population 3500 generations ago; the Eurasian population then split into European and Asian populations 2000 generations ago.

Several population bottlenecks were modeled: one in the Eurasian population shortly after its creation (inbreeding coefficient = 0.085); and one in each present-day population shortly after the European/Asian split (inbreeding coefficients = 0.008, 0.67 and 0.02 for the African, Asian and European populations, respectively).

Two-way symmetric migration was modeled for the (African, European) and (African, Asian) population pairs, with probability of 32e-6 and 8e-6 per chromosome per generation respectively, in the 500 generations following the Asian/European split. (In the published model, migration continues until the present time; migration rates were found to be the least important parameters for matching HapMap data(Schaffner et al., 2005).

For each simulation, a recombination map was generated using the 'recosim' program from the cosi simulator distribution. The model included regional variation in recombination rates (estimated from deCode data) and local hotspots of recombination(Frazer et al., 2007). The local recombination hotspots were modeled using a gamma distribution, with the shape parameter 0.35, mean hotspot spacing of 8500 bp, size of region of local variation of 100000 bp, and the fraction 0.12 of the mean recombination rate kept constant across the region. The recosim parameter file appears below:

```
model 1
shape 0.35
space 8500
bkgd 0.12
local_size 100000
```

The distribution giving regional variation in recombination rates, estimated from decode data, was as follows:

```
0.0   0.2   0.05052
0.2   0.4   0.14890
0.4   0.6   0.27919
0.6   0.8   0.40295
0.8   1.0   0.50370
1.0   1.2   0.59713
1.2   1.4   0.66145
1.4   1.6   0.71358
1.6   1.8   0.77198
1.8   2.0   0.80874
2.0   2.2   0.84086
2.2   2.4   0.86372
2.4   2.6   0.88945
2.6   2.8   0.91396
2.8   3.0   0.92996
3.0   3.2   0.93987
3.2   3.4   0.95006
3.4   3.6   0.95487
3.6   3.8   0.96286
3.8   4.0   0.96884
4.0   4.2   0.97386
4.2   4.4   0.97876
4.4   4.6   0.98233
4.6   4.8   0.98540
4.8   5.0   0.98787
5.0   5.2   0.98971
5.2   5.4   0.99101
5.4   5.6   0.99351
5.6   5.8   0.99531
5.8   6.0   0.99575
6.0   6.2   0.99633
6.2   6.4   0.99706
6.4   6.6   0.99740
```

```
6.6   6.8   0.99786
6.8   7.0   0.99842
7.0   7.2   0.99898
7.2   7.4   0.99920
7.4   7.6   0.99962
7.6   7.8   0.99981
7.8   8.0   1.00000
```

Gene conversion was modeled at a fixed uniform rate of 4.5e-9 per chromosome per generation. The mutation rate was fixed at 1.5e-8.

The specification of the neutral model in cosi simulator input format appears below:

```
gene_conversion_rate 0.0000000045
mutation_rate 0.000000015
length 1000000

pop_define 1 european
pop_define 4 asian
pop_define 5 african

#european
pop_size 1 7700
sample_size 1 120

#asian
pop_size 4 7700
sample_size 4 120

#african
pop_size 5 24000
sample_size 5 120

pop_event migration_rate "afr->eur migration" 5 1 1505 .000032
pop_event migration_rate "eur->afr migration" 1 5 1505 .000032
pop_event migration_rate "afr->as migration" 5 4 1505 .000008
pop_event migration_rate "as->afr migration" 4 5 1505 .000008

pop_event bottleneck "african bottleneck" 5 1997 .008
pop_event bottleneck "asian bottleneck" 4 1998 .067
pop_event bottleneck "european bottleneck" 1 1999 .02

pop_event split "asian and european split" 1 4 2000
pop_event migration_rate "afr->eur migration" 5 1 1996 0
pop_event migration_rate "eur->afr migration" 1 5 1995 0
pop_event migration_rate "afr->as migration" 5 4 1994 0
pop_event migration_rate "as->afr migration" 4 5 1993 0

pop_event bottleneck "OoA bottleneck" 1 3499 .085
pop_event split "out of Africa" 5 1 3500

pop_event change_size "african pop size" 5 17000 12500
```

Modeling the Selective Sweep
Selective sweeps in a single population are modeled in cosi using the structured coalescent approach(Braverman et al., 1995; Kim and Stephan, 2002; Przeworski, 2002), in which the population undergoing the sweep is partitioned into two pools (chromosomes with and without the selected allele) and coalescence is restricted to occur only within the same pool. Recombination can move a sequence segment from a chromosome in one pool to a chromosome in the other, with the probability of such transitions determined by the frequency of the selected allele. The frequency trajectory of the causal allele is modeled by the deterministic approximation of (Stephan et al., 1992). Our implementation closely follows that of (Kim and Stephan, 2002), with two differences: 1) we choose the initial and final frequencies of the beneficial allele to be 1-1/2Ne and 1/2Ne, respectively; and 2) partial (soft) sweeps are supported, letting the final frequency of the causal allele be an arbitrary specified value. For partial sweeps, the present-day chromosomes are randomly assigned to be in the causal allele pool or the non-causal allele pool based on the target final frequency of the causal allele.

Implementation of the sweep algorithm was validated by direct inspection of the beneficial allele frequency trajectory and the associated coalescence and recombination rates. We tested the code's large-N_e behavior by comparing the predicted heterozygosity within a selective sweep with the approximate model in (Durrett and Schweinsberg, 2004) (Proposition 1, in that paper) and found excellent agreement.

Our simulations included sweeps in each population; the age of the selected allele ranged over 5ky, 10ky, 20ky; and the final frequency of the selected allele ranged over 0.2, 0.4, 0.6, 0.8 and 1.0. These two parameters determine the selective coefficient. For each combination of these parameters, we created 300 simulation replicas, with a single selected SNP in the middle of the 1MB region. We also simulated 1300 neutral replicas (300 for constructing likelihood tables used in CMS computation, and another 1000 for evaluating the false positive rate of selection detection by CMS).

While the simulations capture many aspects of the 1000G data, they still represent an idealized version of the real data. In particular, they do not include:

- Sequencing, phasing or imputation errors
- Uncertainty in the genetic map: the genetic map used in the analysis of each simulation is the true map, while the real data is analyzed using an estimated map constructed from haplotype data
- Missing information about ancestral state of any SNPs
- Regions with more complex selection scenarios, such as multiple positively selected SNP or other modes of selection

Composite of Multiple Signals (CMS)
Two versions of the CMS test were used: the original (within-region) CMS test(Grossman et al., 2010) for localizing the selected variant within a candidate region, and a modified test (denoted CMS_{GW}) for identifying candidate regions within the genome.

iHS, XP-EHH, and iHH were calculated as described in (Voight et al., 2006) and (Sabeti et al., 2007) for all bi-allelic SNPs in the 1000G dataset, with the following modification for full sequence data. The extended haplotype homozygosity, which forms the basis for these three tests, measures the probability that two randomly chosen chromosomes in a given population or with a specific core allele are identical from the position of the core allele to the position of a given marker. This statistic aims to capture the length of the haplotype as it decays due to recombination. When calculating this statistic from full sequence data, we found that haplotypes would often break due to a single low-frequency mutation, although it was clear from inspecting the data that the sequence continued to be nearly identical for a much longer stretch. These premature breaks reduced the signal-to-noise ratio. To mitigate this effect, we excluded all rare SNPs below 5% from our EHH calculations.

Mean pairwise F_{ST} and difference in derived allele frequency (ΔDAF) between the putative selected population and the two other populations was calculated for each SNP using allele frequencies from 1000G data.

Five tests were included in the composite score: iHS, XP-EHH, iHH difference, F_{ST}, and ΔDAF. The scores for each test except ΔDAF were normalized as follows: for within-region CMS, scores were normalized within each simulated or real region; for CMS_{GW}, scores were normalized to scores in simulated neutral regions (for simulations) or to the genome-wide distribution (for real data). For each of the tests, we generated three empirical distributions from the simulations: (1) selected SNPs, (2) neutral SNPs within 500kb of a selected SNP, and (3) SNPs in neutral regions. Each distribution was represented using 60 bins. For within-region localization, the value ranges used to define the bins were as follows: iHS, [-3,3]; XP-EHH, [-3,3]; iHH difference, [-3,3]; F_{ST}, [-2,2]; and ΔDAF, [-1,1]; for CMS_{GW}, the value ranges were as follows: iHS, [-6,6]; XP-EHH, [-3,8]; iHH difference, [-3,5]; F_{ST}, [-1,6]; and ΔDAF, [-1,1]. Values outside the range were binned into the nearest bin at the end of the range.

For each test, we used these empirical distributions to calculate the posterior probability a given SNP was selected conditional on its score for that test. Let $t \in \{iHS, XPEHH, iHH, F_{ST}, \Delta DAF\}$ denote the test, v_t denote the score of test t at the SNP, $bin_{t,k}$ denote bin k of the distribution of values of test t, N_{SNP} denote the number of SNPs in the region, and $selected / unselected$ denote the event that the SNP is/is not selected. Then, assuming exactly one SNP in the region is selected,

$$P(selected \mid v_t \in bin_{t,k}) = \frac{P(v_t \in bin_{t,\kappa} \mid selected) * \frac{1}{N_{SNP}}}{P(v_t \in bin_{t,\kappa} \mid selected) * \frac{1}{N_{SNP}} + P(v_t \in bin_{t,\kappa} \mid unselected) * \frac{N_{SNP} - 1}{N_{SNP}}}$$

The values of $P(v_t \in bin_{t,k} \mid selected)$ and $P(v_t \in bin_{t,k} \mid unselected)$ are estimated from the fraction of simulated selected and unselected SNPs which fall into that bin. (Note,

computing the probability that a test score falls into a particular bin, rather than that it equals a particular value, solves the problem that the probability of seeing a particular value is zero for continuous-valued test scores. Approximate Bayesian Computation (ABC) methods can also be used to address this problem (Csillery et al., 2010)).

The composite likelihood is the product of the posterior probabilities that a given SNP is selected from each of the five tests:

$$CMS = \prod_{t \in tests} P(selected \mid v_t \in bin_{t,k})$$

Because this is not a true likelihood (it assumes independence between the tests when in fact some are correlated), we calculated significance empirically from the distribution of scores in simulated neutral regions.

CMS_{local} vs. *CMS_{GW}*

When using CMS to localize regions, we used the distribution of neutral SNPs within 500kb of selected SNPs as the "unselected" distribution and assumed exactly one selected SNP per region. To use CMS as a genome-wide method to detect selected regions, we made the following modifications:

(1) SNPs in neutral regions were used as the "unselected" distribution
(2) We did not assume any prior hypothesis about how many SNPs are under selection. Therefore instead of calculating the posterior probability, we calculated the Bayes factor for each test

$$BF_t = \frac{P(v_t \in bin_{t,k} \mid selected)}{P(v_t \in bin_{t,k} \mid unselected)}$$

and defined the composite score as the product of the Bayes factor of each test:

$$CMS_{GW} = \prod_{t \in tests} BF_t$$

(3) Scores were normalized to neutral simulations (for simulated data) or to the whole genome (for real data), rather than within each region.
(4) Bin boundaries were adjusted as described earlier.

We identified 100kb regions in which 30% of SNPs had a normalized score above 3, a threshold which corresponded to a 0.1% FPR in simulations (i.e. in 1000 neutral simulations of a 1MB region no more than 1 contained a 100KB region meeting this criterion; the upper bound of the 95% binomial confidence interval for the FPR is 0.6%), and used this threshold to detect selected regions in the 1000 Genomes data. We note that since the 1000G data includes 2.42Gb, we expect 24 false positive regions at this threshold.

1000 Genomes (1000G) Project

Quality-controlled phased SNP and indel calls for the CEU, YRI and CHB+JPT populations released by the low-coverage portion of the 1000G project were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1/ , representing the March 2010 data release. The phasing and imputation had been done by the 1000G project using IMPUTE2 software. All genetic variants with more than two alleles were converted to biallelic variants, by mapping all alleles to two alleles while preserving alleles' ancestral state where known. Ancestral state was taken from the ancestral state data released by the 1000 genomes project at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments, constructed from a four-way alignment of human, chimp, orangutan and rhesus macaque. Monomorphic SNPs omitted from 1000G data but present in HapMap Phase II data were added back into the data. Only SNPs genotyped in all three HapMap populations were used for CMS analyses. Copy number variant calls were also taken from the 1000genomes project, available at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/companion_papers/mapping_structural_variation/. All CNVs with genotype information were used: specifically deletions, mobile element insertions, and tandem duplications. Rare CNVs with a frequency of less than 5% were excluded.

CNV Overlap Analysis

An instance of overlap is defined as a unique combination of a localized region and CNV that overlap one another. We found 60 such overlaps in the selected regions. The probability of an overlap between a localized region of length and a CNV of length b is (a+b)/g, where g is the size of the genome. Using the size and number of selection regions and summing over all possible pair-wise probabilities, the expected number of overlaps is 49.9885. Using a Poisson distribution, the probability that 60 or more overlaps would be seen at random is 0.226.

We also calculated an empirical p-value for CNV enrichment by simulating 1000 sets of randomly-located regions of the same size and number as the true regions, and recording the number of overlaps with the CNVs. We found that real regions fell in the 85th percentile of these simulations ($P_{empirical}$=0.15).

While our analysis of candidate causal variants focuses on SNPs, we catalogued CNVs as they could themselves be targets of selection. They can be found on our website at: http://www.broadinstitute.org/mpg/cms.

Coding Variant Analysis

We used 1000G project annotation of coding variants. Since only 93% of the coding regions of the genome were covered in the dataset, we calculated the 95% confidence interval for the number of high-scoring non-synonymous variants present in the selected regions. First we note that there are 317 observed non-synonymous SNPs in the selected regions and a total of 862,752 coding bases (using RefSeq exons). If 93% of these are covered by sequencing, then 802,359 coding bases are observed in the dataset. We can estimate the frequency of non-synonymous SNPs in coding regions to be 317/802359 = 3.95×10^{-4}. Using a binomial distribution with N=802,359 and x=317, the 95%

confidence interval for the frequency is ($3.53x10^{-4}$, $4.41x10^{-4}$). Hence in the unobserved coding regions we expect between 21.3 and 26.6 non-synonymous SNPs. Assuming the fraction of high-scoring non-synonymous SNPs in the unobserved regions is the same as in the observed regions, the confidence interval for the number of unobserved high-scoring non-synonymous SNPs is (2.3,2.9). Therefore we expect at most 38 high-scoring non-synonymous SNPs in the regions.

Enrichment Analysis of Functional Variants

We picked random sets of 412 non-overlapping regions in the genome that matched our selected regions in size. For each set, we calculated the fraction of variants that were within exons or lincRNAs, or had been previously associated with gene expression (eQTLs) or a phenotype (GWAS). We repeated these simulations 10000 times. P-values were calculated as the proportion of simulations where the fraction of variants with a particular annotation was higher than what we observed in our selected regions.

Enrichment Analysis of Gene Pathways

We manually defined functional categories that previous literature suggests may have played an important role in recent human adaptation. These include functions such as skin pigmentation, immune system processes, response to infectious disease, sensory perception, and metabolism. We searched the EntrezGene, Uniprot, and MGI mouse knockout descriptions of every gene in the human genome for certain keywords related to the function of interest. The keywords included the name of the pathway and its derivatives (e.g. immune, immuno-, immuni-, etc.), the cell-types associated with the pathway (e.g. hair cell for Hearing), and common molecules associated with the pathway (eg. melanin for Pigmentation, cytokine for Immune System). The keywords were as permissive as possible to generate a comprehensive initial list.

We then manually reviewed the results of this search to remove any genes that were clearly not related to the category and assemble a set of genes for the function or phenotype of interest. There was some overlap between related pathways in our final lists. Between the Immune System and Response to Infectious Disease categories, there were 127 overlapping genes. The Sensory Perception category included all of the genes in the individual senses (Vision, Olfactory, Hearing) as well as a short list of genes related to taste perception. All of the gene sets have been made available online at http://www.broadinstitute.org/mpg/cms/.

INRICHv1.0 was used to test for enrichment of these gene sets and functional categories (Lee et al., 2012). INRICH uses a two-step permutation algorithm to test for enrichment of pathways defined by the user or derived from published databases. On the first pass, for every region in the input list, the program randomly selects another region of the genome that matches the input region in terms of size, SNP number, and number of genes. For our analysis, this process was repeated ten million times, generating ten million random lists of regions that matched the input list. From these permuted lists, the program then calculates empirical p-values for enrichment in all of the input categories. On the second pass, INRICH randomly selects 10,000 of the permuted lists and calculates which categories and how many of them are enriched. This second step was also repeated

ten million times. The program then uses the results from the second-pass permutations to correct p-values from the first step, and reports the corrected and uncorrected p-values as its output. See http://atgu.mgh.harvard.edu/inrich/ for more information.

Protein Structure Modeling

We performed a blast search for solved protein structures that closely matched the amino acid sequence of our non-synonymous variants. Sequences that had a similarity of at least 25%, and preferably over 40%, to the target sequence were selected as templates for homology modeling. Modeller9v8 was used to align the target sequence to the template sequences, and the alignments were further optimized through manual manipulation (Eswar et al., 2006). The final alignment was used to generate homology models using Modeller9v8. At least 10 models were generated for every protein. All homology models were assessed by their DOPE and GA341 scores, and the model with the lowest DOPE score was selected as the most reliable homology model. Modeller9v8's loop refinement algorithm was then used to reduce the energy of unfavorable loops and generate a more stable and reliable final model. For toll-like receptor 5 (TLR5), Modeller9v8 could not accurately predict the characteristic crescent-shaped structure of the ectodomain. This was mainly due to a poor alignment between the leucine-rich repeat domains (LRRs) of TLR5 and the other TLR proteins used as templates. We therefore used a published computationally derived model of human TLR5, provided by Wei and colleagues (Wei et al., 2011).

Multiple Sequence Alignments

Multiple-species sequence alignments were generated by pulling the UCSC 44-way vertebrate nucleotide alignment for the region of interest and translating the alignment in the appropriate reading frame.

LincRNA Expression

RNA reads aligned to the hg18 genome were counted across the 4,421 previously detected regions of interest. Reads were RPKM normalized against both the length of the region and the total read count in the lane (Mortazavi et al., 2008) to provide a baseline expression level for each region. A region was considered to be an expressed lincRNA if it contained non-zero expression in at least half of the individuals in a population. If the median expression was zero, the region was no longer considered, consistent with the methods of (Pickrell et al., 2010).

A reference list of human lincRNAs was obtained by integrating publically available lncRNA annotations with transcriptome assemblies of RNA sequencing data from 24 tissues and cell lines and processing those through a lincRNA calling pipeline (Cabili et al., 2011).

RNA reads for YRI (Pickrell et al., 2010) were obtained from http://eqtl.uchicago.edu/RNA_Seq_data/. RNA reads for CEU (Montgomery et al., 2010) were obtained from http://jungle.unige.ch/rnaseq_CEU60/. Reads for both populations were obtained in fastq format and aligned to hg19 using the BWA aligner (Li and Durbin,

2009), version 0.5.7. Read counts for each region were obtained using SAMtools (Li et al., 2009), version 0.1.16.

All non-zero expression levels were quantile-normalized within each population in order to produce a normal distribution of expression. All individuals with zero aligned reads for a given ncRNA were excluded from the quantile-normalization for that ncRNA. For individuals where multiple lanes of reads were available, their quantile-normalized expression level was averaged.

eQTL Analysis

To date, there have been several studies that have measured gene expression levels across the genome in the 1000G individuals. We obtained expression intensities of 47293 probes representing the majority of the human gene complement from (Stranger et al., 2007). We also downloaded the normalized gene expression levels for 22032 genes in the YRI individuals measured by RNA seq by (Pickrell et al., 2010), and the significant p-values for the CEU individuals measured by RNA seq by (Montgomery et al., 2010) (expression levels not released). We used intensities or normalized read counts from each gene within 1 MB of each SNP within the selected regions as quantitive traits in a standard association test (Purcell et al., 2007), and recorded all regions that contained a significant eQTL SNP (using the significance thresholds defined in each of the studies). We also tested for significant associations with lncRNA expression, using the expression levels defined above.

Chromatin State Analysis

We identified all variants in the CMS regions that fall in enhancer and promoters along with their motif disruptions from HaploReg (Ward and Kellis, 2012). Within these motif disruptions, we identified those affect motifs that are candidate regulators in the combination of cell types where the enhancer or promoter is active.

GWAS Datasets (TB, Leprosy, and NHGRI GWAS Catalogue)

We compiled a database of polymorphisms associated with susceptibility to diseases and various other traits by combining hits reported in the NHGRI catalogue of genome-wide association studies (http://www.genome.gov/gwastudies/) with genome-wide significant SNPs from several published GWAS of a variety of infectious diseases (Davila et al., 2010; Fellay et al., 2007; Ge et al., 2009; Jallow et al., 2009; Kamatani et al., 2009; Mbarek et al., 2011; Png et al., 2011; Zhang et al., 2009). We intersected these hits with the selected regions, and identified all SNPs significantly associated with phenotypes that lie within the selected region.

We examined in more depth a recent Wellcome Trust Case Control Consortium study in the Gambia (Thye et al., 2010) with 1,498 confirmed TB cases and 1,496 controls, genotyped on the Affymetrix GeneChip 500K Array comprising 500,568 SNPs using the CHIAMO algorithm. Multi-dimensional scaling of identity-by-state (IBS) metrics, calculated between each pair of individuals using a subset of 100,715 uncorrelated SNPs passing QC filters, identified three axes of genetic variation that distinguish most common ethnic groups in this study. The primary analysis focused on single-locus tests

of association using 1,320 TB cases compared to 1,384 Gambian controls for all 405,226 SNPs passing QC filters with a study-wide MAF > 1%. The trend test was performed in a logistic regression modeling framework, which was adjusted for the three axes of multi-dimensional scaling, by inclusion as covariates in the logistic regression model, reducing the over-dispersion of trend tests from $\lambda = 1.13$ (no adjustment) to $\lambda = 1.05$.

The host genetics study of leprosy in Indians (Wong et al., 2010) consisted of 258 confirmed cases of leprosy and 300 controls from New Delhi. All individuals in this study were genotyped with the Illumina IBC gene-centric 50K array. The microarray genotypes more than 48,000 SNPs distributed in approximately 2,100 genes throughout the genome, including 3,470 non-synonymous markers. The data quality control and analysis were performed using PLINK. Multi-dimensional scaling (MDS) and principal component analysis (PCA) were carried out with PLINK and EIGENSTRAT to remove population outliers. A total of 209 leprosy cases and 239 controls were carried forward for analysis after quality control filters. The primary test of association in the New Delhi and Kolkata cohorts was carried out with the Pearson's χ^2 allelic test, Cochran-Armitage trend test and logistic regression.

For both the TB and leprosy studies we then identified SNPs with evidence of association with susceptibility to these pathogens ($P < 0.01$) within the regions CMS_{GW} detected to be under selection.

Functional Testing of rs5744174

Cell Lines

293FT cells were grown in D-MEM GlutaMAX supplemented with 10% fetal bovine serum (FBS), 0.1 mM MEM Non-Essential Amino Acids Solution, 1 mM MEM Sodium Pyruvate Solution and 500 μg/mL Gentamicin. Jurkat E6.1 cells were grown in IMDM GlutaMAX supplemented with 10% FBS, 50 μM 2-mercaptoethanol and 50 μg/mL gentamicin.

Reporter Construction

Transgenes carrying either the ancestral (*tlr5a;* leucine) or derived (*tlr5d;* phenylalanine) form of TLR5 were synthesized and cloned into the retroviral vector m6pg carrying GFP as a transgene to create m6pg[*tlr5a*] and m6pg[*tlr5d*](Andersen et al., 2008). For the measurement of NF-κB activity in Jurkat cells the retroviral reporter m3pkb[*luc*] carrying an NF-κB inducible luciferase reporter was used as previously described (Loizou et al., 2011). In 293FT cells NF-κB activity was measured using pGL4.32 and pGL4.74 (Promega).

Stable Cell Construction

Stable 293FT and Jurkat cell lines were created by transducing the cells with either m6pg[*tlr5a*] or m6pg[*tlr5d*]. The transduction efficiency was around 30% in all cases. Based on the expression of the GFP transgene, the transduced cells were then subjected to multiple rounds of cell sorting using a MoFlow (Beckman Coulter), until a purity of transduced cells were >95%. Using primers TGCATCCAGATGCTTTTCAG and CCAGCCATTTCCAGAAACAT the expression of TLR5 was measured by qPCR using

the PerfeCta SYBR Green SuperMix, ROX (Quanta) according to the manufacturer's instructions.

293FT Luciferase Assays

293FT cells stably expressing either the ancestral or derived forms of TLR5 were plated at 5 X 10^4 cells per well in 100 μL complete media one day prior to transfection in 96-well tissue-culture treated plates (Nunclon). For each well of co-transfection, 100 ng each of pGL4.32 and pGL4.74 DNA was combined with 0.12 μL Plus reagent in 20 μL of Opti-MEM I Reduced Serum Medium for 5 minutes. 0.4 μL of Lipofectamine LTX was then added, mixed and incubated at room temperature for an additional 30 minutes before being added drop-wise to cells. Cells were incubated at 37°C, 5% CO_2 and after four hours 75 μL of media was aspirated from each well and replaced with complete media and allowed to incubate for an additional 22 hours. Following incubation cells were stimulated for an additional 24 hours with 800 ng/mL PMA or increasing levels of flagellin at 1, 5, 10 or 100 ng/mL. Cells were then lysed using 75 μL of Dual-Glo® Luciferase Reagent and after 10 minutes firefly luminescence was measured in a 96-well plate reader (Top Count). 75 μL of Stop & Glo® Substrate was added to each well and after 10 minutes *Renilla* luminescence was measured in a Top Count machine.

Jurkat Luciferase Assays

One day prior to transfection 293FT cells were plated in a 10 cm^2 dish so that they were 90% confluent at the time of transfection. 6.6 μg each of m3p[Luc], retroviral packaging pCL-Eco and viral envelope VsV-g DNA were combined with 12 μL Plus reagent in 2.5 mL of Opti-MEM I Reduced Serum Medium for 5 minutes. 40 μL of Lipofectamine LTX was then added, mixed and incubated at room temperature for an additional 30 minutes before being added drop-wise to cells. Cells were incubated at 37°C, 5% CO_2 for four hours and the media was aspirated from the plate and replaced with complete media. The cells were incubated for an additional 20 hours at which point the supernatant containing viral particles was removed and filtered using a 0.45 μM filter. Filtered virus was stored in the -80°C freezer.

Jurkat cells stably expressing either the ancestral or derived forms of TLR5 were seeded at 2 X 10^5 cells per well in 100 μL complete media on the day of transduction. 50 μL of m3pkb[*luc*] viral supernatant and 6 μg/mL protamine sulfate were combined and added to each well. The 96-well plates were spun at 400 x *g* for 2 hours at 32°C. Plates were incubated for 20 hours at 37°C, 5% CO_2. Following incubation, plates were centrifuged for 5 minutes at 400 x *g* and 100 μL of media was aspirated and replaced with 50 μL complete media. Following an additional 6 hour incubation, cells were stimulated with 400 ng/mL PMA and 1.5 μg/mL ionomycin or 10 ng/mL flagellin. After a 24 hour incubation, cells were lysed with the Bright-Glo™ Luciferase Assay Reagent (Promega) and after 10 minutes firefly luminescence was measured in a Top Count machine.

SUPPLEMENTAL REFERENCES

A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Andersen, K.G., Butcher, T., and Betz, A.G. (2008). Specific immunosuppression with inducible Foxp3-transduced polyclonal T cells. PLoS biology *6*, e276.

Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H., and Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics *140*, 783-796.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.

Csillery, K., Blum, M.G., Gaggiotti, O.E., and Francois, O. (2010). Approximate Bayesian Computation (ABC) in practice. Trends in ecology & evolution *25*, 410-418.

Davila, S., Wright, V.J., Khor, C.C., Sim, K.S., Binder, A., Breunis, W.B., Inwald, D., Nadel, S., Betts, H., Carrol, E.D.*, et al.* (2010). Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet *42*, 772-776.

Durrett, R., and Schweinsberg, J. (2004). Approximating selective sweeps. Theor Popul Biol *66*, 129-138.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics *Chapter 5*, Unit 5 6.

Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A.*, et al.* (2007). A whole-genome association study of major determinants for host control of HIV-1. Science *317*, 944-947.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Ge, D., Fellay, J., Thompson, A.J., Simon, J.S., Shianna, K.V., Urban, T.J., Heinzen, E.L., Qiu, P., Bertelsen, A.H., Muir, A.J.*, et al.* (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature *461*, 399-401.

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

He, Y., Wang, W.R., Xu, S., Jin, L., and Snp Consortium, P.A. (2012). Paleolithic Contingent in Modern Japanese: Estimation and Inference using Genome-wide Data. Sci Rep *2*, 355.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337-338.

Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M.*, et al.* (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet *41*, 657-665.
Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H.*, et al.* (2009). A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. Nat Genet *41*, 591-595.
Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics *160*, 765-777.
Lee, P.H., O'Dushlaine, C., Thomas, B., and Purcell, S.M. (2012). INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics *28*, 1797-1799.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
Loizou, L., Andersen, K.G., and Betz, A.G. (2011). Foxp3 interacts with c-Rel to mediate NF-kappaB repression. PloS one *6*, e18670.
Marjoram, P., and Wall, J.D. (2006). Fast "coalescent" simulation. BMC genetics *7*, 16.
Mbarek, H., Ochi, H., Urabe, Y., Kumar, V., Kubo, M., Hosono, N., Takahashi, A., Kamatani, Y., Miki, D., Abe, H.*, et al.* (2011). A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. Hum Mol Genet.
Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773-777.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.
Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768-772.
Png, E., Thalamuthu, A., Ong, R.T., Snippe, H., Boland, G.J., and Seielstad, M. (2011). A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region. Hum Mol Genet.
Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. Genetics *160*, 1179-1189.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.
Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R.*, et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913-918.

Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome research *15*, 1576-1583.
Stephan, W., Wiehe, T.H.E., and Lenz, M.W. (1992). The Effect of Strongly Selected Substitutions on Neutral Polymorphism - Analytical Results Based on Diffusion-Theory. Theor Popul Biol *41*, 237-254.
Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D.*, et al.* (2007). Population genomics of human gene expression. Nat Genet *39*, 1217-1224.
Thye, T., Vannberg, F.O., Wong, S.H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M.A.*, et al.* (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat Genet *42*, 739-741.
Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS biology *4*, e72.
Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research *40*, D930-934.
Wei, T., Gong, J., Rossle, S.C., Jamitzky, F., Heckl, W.M., and Stark, R.W. (2011). A leucine-rich repeat assembly approach for homology modeling of the human TLR5-10 and mouse TLR11-13 ectodomains. J Mol Model *17*, 27-36.
Wong, S.H., Gochhait, S., Malhotra, D., Pettersson, F.H., Teo, Y.Y., Khor, C.C., Rautanen, A., Chapman, S.J., Mills, T.C., Srivastava, A.*, et al.* (2010). Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog *6*, e1000979.
Zhang, F.R., Huang, W., Chen, S.M., Sun, L.D., Liu, H., Li, Y., Cui, Y., Yan, X.X., Yang, H.T., Yang, R.D.*, et al.* (2009). Genomewide association study of leprosy. N Engl J Med *361*, 2609-2618.

Figure S1.pdf