

Natural Selection on Genes that Underlie Human Disease Susceptibility

Ran Blehman, Orna Man, Leslie Herrmann, Adam R. Boyko, Amit Indap, Carolin Kosiol, Carlos D. Bustamante, Kosuke M. Teshima, and Molly Przeworski

Supplemental Experimental Procedures

Creating a Hand-Curated Version of the OMIM Database

We downloaded all phenotypic entries with a known genetic basis (i.e., all entries preceded by the symbols + and #) containing the phrases “autosomal dominant,” “dominant,” “autosomal recessive,” “recessive,” or “X-linked,” and retrieved the OMIM phenotype identifiers (Nov. 2004); we found a couple of additional entries by using (incorrect) variations on the spellings. We then used EnsMart [S1] to retrieve the HUGO gene symbols and Entrez Gene identifiers for all genes associated with the list of OMIM phenotype

entries. We turned this set of genes and phenotypes into a list of unique pairs (HUGO gene symbols, OMIM phenotype numbers) and added OMIM gene numbers.

Once the list was created, 2–4 people independently read each phenotype entry that we had downloaded and recorded phenotypic information (see below). We consulted additional online sources (including GeneCards, eMedicine, Gene reviews, Gene clinics, Dynomed, and WebMD) to resolve discrepancies among readers and, when possible, to fill in missing information. We excluded genes with a tenuous link to the phenotype or in which the phenotype was caused by a large rearrangement (e.g., the deletion of several

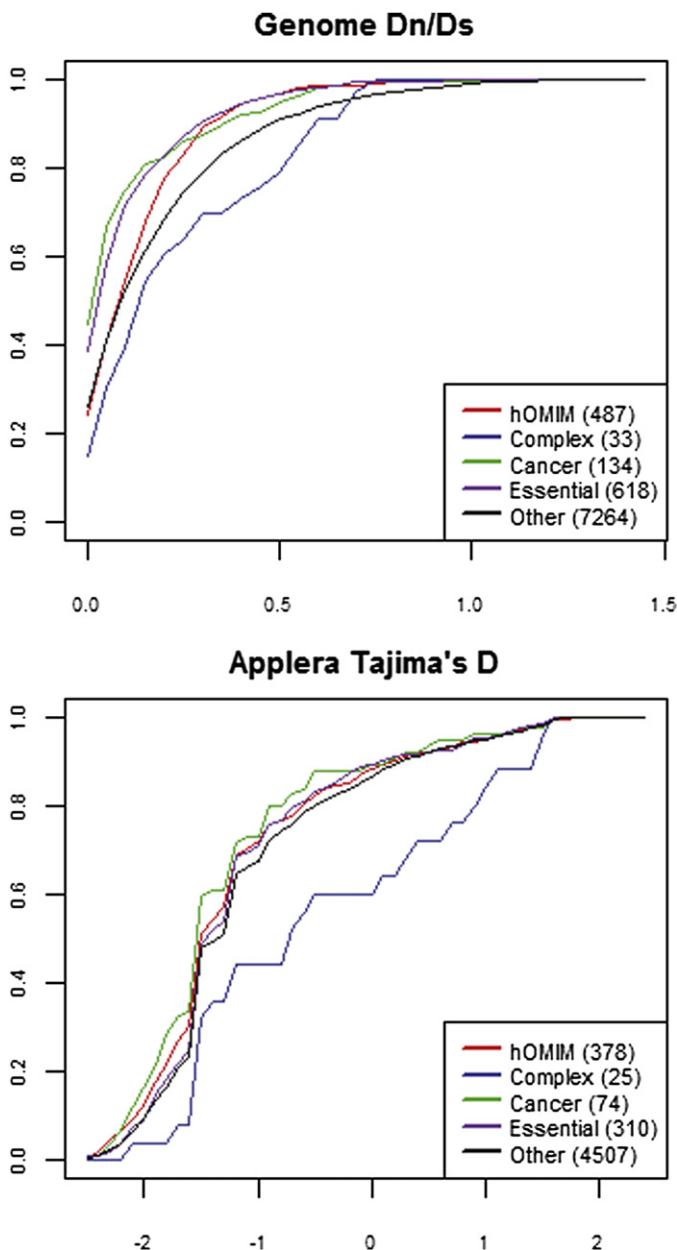


Figure S1. Cumulative Distributions of D_n/D_s and Tajima's D for the Five Groups of Genes, Excluding Genes that are Associated with Immune Response

The order of the five lines remains the same when compared to the distributions including all genes (Figure 4), and the same comparisons remain significant (results for “hOMIM” versus “complex” are shown in Table S1). The legend of Figure 4 and the main-text Experimental Procedures provide more details.

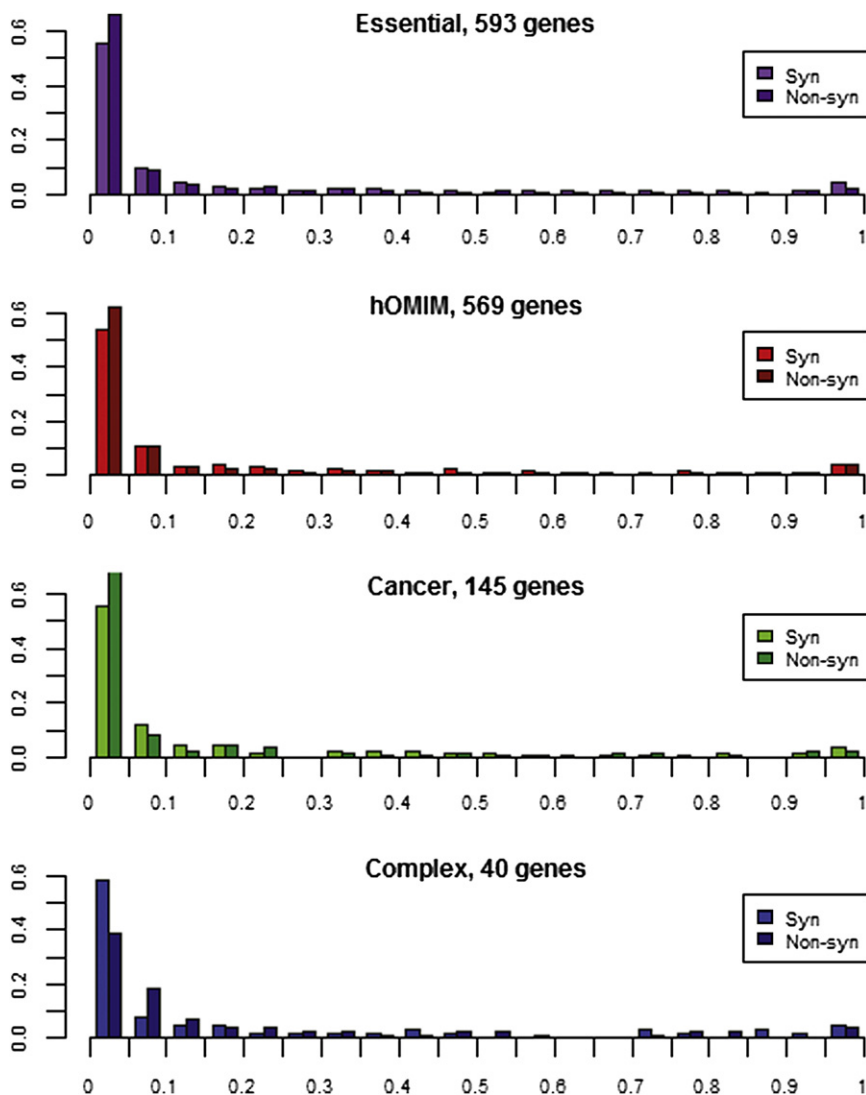


Figure S2. The Frequency Spectrum for Synonymous and Nonsynonymous Alleles in Each Category of Genes

The frequency spectrum was tabulated in the European population sample of the Applera dataset; see main-text Experimental Procedures for details. Very similar results were obtained with the use of the African and African-American population samples or the SeattleSNP and NIEHS resequencing projects instead (not shown). For assessment of the difference between synonymous and nonsynonymous variants, a one-tailed Wilcoxon matched-pairs signed-rank test was performed on the median allele frequencies for each gene (see main-text Experimental Procedures). p values are 5×10^{-6} for essential genes, 0.013 for hOMIM genes, 0.026 for cancer-associated genes, and 0.842 for complex-disease-associated genes.

genes) rather than a point mutation; we also excluded cases in which the disease was only associated with somatic mutations in the gene.

For the mode of inheritance, our categories were “autosomal dominant,” “autosomal recessive,” or “X-linked.” We did not distinguish among X-linked mutations, because such distinctions do not always have a clear meaning [S2]. For the age at onset, our categories were “in utero or shortly after birth,” “before 15 years of age,” “between 15 and 40 years of age,” and “post-40.” We considered the age at which the symptoms first become apparent in the absence of medical intervention, and for cases in which OMIM records were anecdotal, we used the earlier instances as the age at onset. We took 15 years to be the start of the reproductive age, and we took 40 years to be the end because it corresponds approximately to the end of female reproductive age [S3] and to male life expectancy in pre-industrial societies [S4].

We used this information to test for differences in the levels of constraint between genes that cause dominant and recessive disorders and genes that cause early (before 40 years of age)- and late (after 40)-onset disorders. Only 14 genes are known to cause *exclusively* late-onset disorders, and for two (*SNCB* and *ABCC9*), the link to the phenotype was weak. We did not consider comparisons between other age groups, because it is unclear whether age at onset would have a monotonic effect on fitness (e.g., in comparing the fitness loss of the parent if an offspring died at birth versus at age 10). However, we provide this information in Table S4.

We also tried to record information about the severity of the disease phenotype—again, in the absence of medical intervention. However, because the information in OMIM is often sparse and it is very difficult to establish

the fitness consequences of certain phenotypes (e.g., hereditary fructose intolerance) in pre-industrial societies, let alone in pre-agricultural ones, these severity measures are likely to be highly unreliable. We therefore based our analyses on mode of inheritance and age at onset only.

Finding Human-Rhesus Macaque Orthologs

We ran BLASTN searches [S5] of the human coding sequences against the rhesus macaque genome sequence, without any filtering of low-complexity sequences, and we set the threshold of expectation values (e-values) at 0.001. For each human sequence, we then used the alignments produced by the BLASTN search to construct the orthologous rhesus macaque sequence. For this purpose, we used only alignments with at least 85% identity between the human and rhesus macaque sequences. If an alignment was more than 500 bp long, it was broken into overlapping alignments, each one shifted by one bp from the previous one, and only those subalignments displaying at least 85% identity were retained. Because the BLASTN search can produce more than one alignment for the same query, there could be cases in which, with the consideration of these multiple alignments, there would be an ambiguity as to the identity of the nucleotide at a specific position in the orthologous sequence. In such cases, the identity of the nucleotide was determined by the alignment with the lowest e-value. If, in the alignment used to choose a specific nucleotide, this nucleotide was followed by an insertion in the rhesus macaque sequence, this insertion was introduced into the orthologous sequence.

A concern is that the inferred rhesus macaque orthologous sequence could contain nucleotides originating from different contigs in the genome

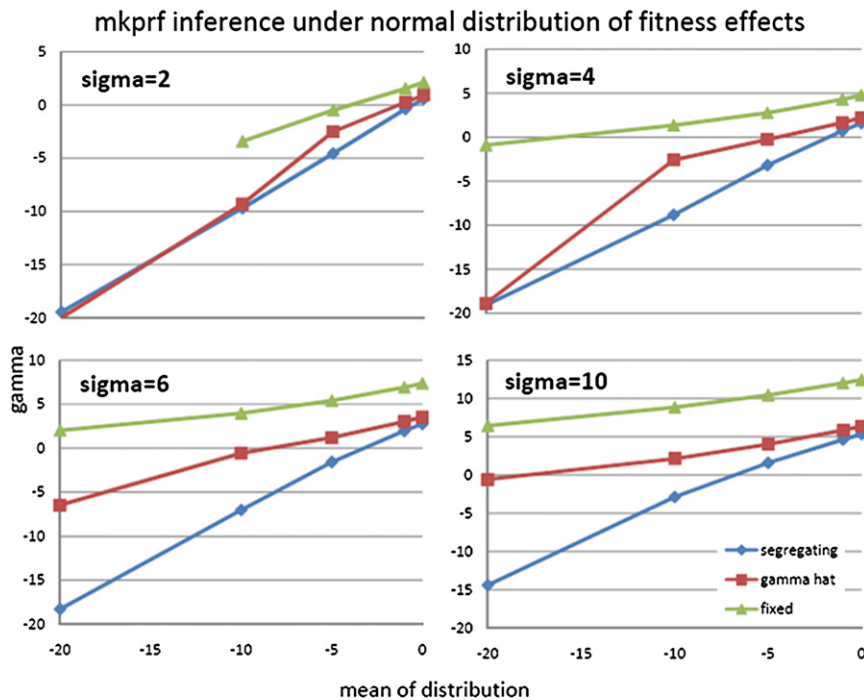
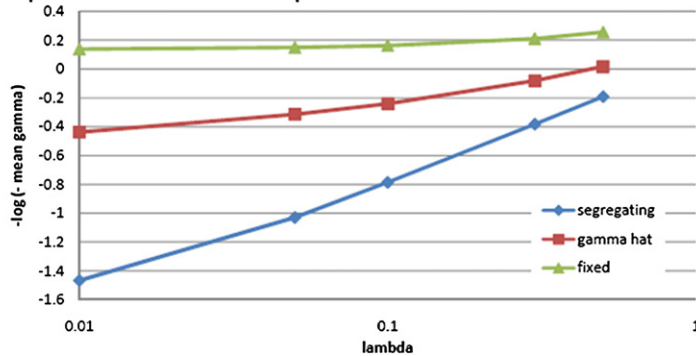


Figure S3. Mkprf Inference of γ under a Distribution of Fitness Effects

In the top panel, the fitness effect of new mutations is drawn from a normal distribution $N(\mu, \sigma)$. In the bottom panel, the fitness effect of new mutations is drawn from an exponential distribution $EXP(\lambda)$. Shown in blue is the mean γ of mutations observed to be segregating in a sample of 78 chromosomes, shown in green is the γ estimated by mkprf, and shown in red is the mean γ of mutations that become fixed in the population. Note that essentially no mutations become fixed when drawn from $N(-20, 2)$.

mkprf inference under exponential distribution of fitness effects



draft. We therefore performed consistency checks on the sequence. For this purpose, each position in the rhesus macaque orthologous sequence was associated with the following information: the contig from which it came, the strand on the contig, its position in the contig, and the e-value obtained for the alignment of the contig against the human gene sequence. A pair of positions $i < j$ on the inferred rhesus macaque orthologous sequence was then defined to be consistent if they were associated with the same strand of the same contig and if the contig positions associated with them, $p(i)$ and $p(j)$, respectively, fulfilled $p(i) < p(j)$ if they are on the forward strand or $p(i) > p(j)$ if they are on the reverse strand. The positions of the inferred rhesus macaque orthologous sequence were then arranged into the longest-possible consistent subsets (i.e., within a subset, every pair of positions was consistent), and each subset was associated with the minimal e-value associated with any of the positions it contained. If two sets of positions were associated with the opposite strands of the same contig, we removed from the sequence the nucleotides associated with the subset with the highest e-value. We proceeded similarly in cases in which two sets of positions overlapped on the inferred rhesus macaque orthologous sequence and were associated with different contigs in the rhesus macaque genome draft. If two sets overlapped on the inferred rhesus macaque orthologous sequence and were associated with the same contig, we resolved the inconsistency by finding a position in one of the sets (designated here as "set A") that is consistent with all the positions in the other set ("set B") and then removing all the positions in set A that are not consistent with at least one position in set B.

As a final quality check, we used BLASTN [S5] searches of the resultant orthologous sequences against the human refseq mRNA collection to

ensure that the ortholog pairs were mutual best hits. In twelve cases, the rhesus macaque sequence was not a mutual best hit of the human gene (because the ortholog does not seem to exist in the genome sequence) and our search picked up a close homolog instead: *PRSS1*, *CYPC19*, *SLC25A15*, *ACTC*, *ACTA1*, *GCSH*, *SFTPA1*, *USP6*, *DUX4*, *SSX1*, *SSX2* and *CIC*. *HBG1* and *HBG2* are part of a cluster of globin genes. They are the result of a duplication event that occurred 20 to 40 million years ago [S6] and have undergone gene-conversion events [S7], resulting in highly similar paralogous sequences (only two nucleotide differences between the human paralogs). Our search recovered the same rhesus macaque sequence for both *HBG1* and *HBG2*, having nine and eight nucleotide differences with *HBG1* and *HBG2*, respectively. Given that their evolution is not independent, we discarded these genes from further analysis.

Calculation of Tajima's D in the Presence of Missing Data

The following formula was used to calculate Tajima's D for each gene:

$$D = \frac{\sum_{i=1}^M (\hat{\theta}_{\pi(i)} - \hat{\theta}_{w(i)})}{W},$$

in which

$$\hat{\theta}_{\pi(i)} = \frac{n_i}{n_i - 1} 2p_i(1 - p_i),$$

$$\hat{\theta}_{w(i)} = 1/a_{n_i}, a_{n_i} = \sum_{k=1}^{n_i-1} 1/k,$$

Table S1. Tests of Differences between Categories

Contrast	Data, Statistic	No. of Genes	Medians	KS Test ^a	Logistic Regression ^a	KS Bootstrap
AR vs. AD	Genome D_n/D_s ^b	240, 145	0.169, 0.078	1.00E-10	1.29E-06	NT
post40 vs. pre	Genome D_n/D_s	6, 466	0.033, 0.136	0.022		0.009
Complex vs. OMIM	Genome D_n/D_s	39, 501	0.203, 0.133	0.003	4.78E-06	0.003
Complex (short) ^c vs. OMIM	Genome D_n/D_s	25, 501	0.209, 0.133	0.008	2.07E-04	0.01
Cancer vs. OMIM	Genome D_n/D_s	145, 501	0.061, 0.133	8.45E-09	1.44E-03	
Other vs. OMIM	Genome D_n/D_s	7482, 501	0.139, 0.133	1.05E-05	7.84E-08	
Cancer vs. Complex	Genome D_n/D_s	145, 39	0.061, 0.203	5.34E-06	3.24E-06	<1E-04
Cancer vs. Complex (short) ^c	Genome D_n/D_s	145, 25	0.061, 0.209	2.22E-04	9.03E-05	<1E-04
Other vs. Complex	Genome D_n/D_s	7482, 39	0.139, 0.203	0.079	0.156	0.061
Other vs. Complex (short) ^c	Genome D_n/D_s	7482, 25	0.139, 0.209	0.130	0.289	0.109
Other vs. Cancer	Genome D_n/D_s	7482, 145	0.139, 0.061	7.58E-11	8.16E-09	
Essential vs. Cancer	Genome D_n/D_s	645, 145	0.077, 0.061	0.085	0.398	
OMIM vs. Essential	Genome D_n/D_s	501, 645	0.133, 0.077	5.49E-09	1.93E-03	
Other vs. Essential	Genome D_n/D_s	7482, 645	0.139, 0.077	0	4.60E-23	
Complex vs. Essential	Genome D_n/D_s	39, 645	0.203, 0.077	2.64E-05	1.27E-05	<1E-04
Complex (short) ^c vs. Essential	Genome D_n/D_s	25, 645	0.209, 0.077	1.04E-03	1.10E-03	<1E-04
AR vs. AD	D_n/D_s ^b	452, 294	0.184, 0.084	1.11E-16	1.85E-04	
post40 vs. pre	D_n/D_s	14, 926	0.120, 0.149	0.575	0.994	0.502
Complex vs. OMIM	D_n/D_s	65, 952	0.224, 0.149	0.003	1.18E-05	
Complex (short) ^c vs. OMIM	D_n/D_s	40, 952	0.217, 0.149	0.010	0.010	0.006
Cancer vs. OMIM	D_n/D_s	326, 952	0.091, 0.149	1.49E-05	0.860	
Cancer vs. Complex	D_n/D_s	326, 65	0.091, 0.224	6.93E-07	3.52E-04	
Cancer vs. Complex (short) ^c	D_n/D_s	326, 40	0.091, 0.217	2.73E-05	0.032	<1E-04
AR vs. AD	Appl. Tajima's D	214, 95	-1.42, -1.31	0.314	0.372	
post40 vs. pre	Appl. Tajima's D	4, 378	0.223, -1.42	0.276		0.15
Complex vs. OMIM	Appl. Tajima's D	29, 399	-0.65, -1.42	0.008	5.12E-04	0.005
Complex (short) ^c vs. OMIM	Appl. Tajima's D	17, 399	-1.12, -1.42	0.317	0.023	0.238
Cancer vs. OMIM	Appl. Tajima's D	84, 399	-1.42, -1.42	0.581	0.347	
Other vs. OMIM	Appl. Tajima's D	4644, 399	-1.28, -1.42	0.029	0.036	
Cancer vs. Complex	Appl. Tajima's D	84, 29	-1.42, -0.65	0.007	4.0E-04	0.003
Cancer vs. Complex (short) ^c	Appl. Tajima's D	84, 17	-1.42, -1.12	0.217	0.015	0.126
Other vs. Complex	Appl. Tajima's D	4644, 29	-1.28, -0.65	0.015	1.62E-03	0.004
Other vs. Complex (short) ^c	Appl. Tajima's D	4644, 17	-1.28, -1.12	0.370	0.048	0.266
Other vs. Cancer	Appl. Tajima's D	4644, 84	-1.28, -1.42	0.064	0.037	
Essential vs. Cancer	Appl. Tajima's D	323, 84	-1.39, -1.42	0.172	0.201	
OMIM vs. Essential	Appl. Tajima's D	399, 323	-1.42, -1.39	0.356	0.623	
Other vs. Essential	Appl. Tajima's D	4644, 323	-1.28, -1.39	0.438	0.208	
Complex vs. Essential	Appl. Tajima's D	29, 323	-0.65, -1.39	0.011	6.68E-04	0.004
Complex (short) ^c vs. Essential	Appl. Tajima's D	17, 323	-1.12, -1.39	0.349	0.027	0.237
AR vs. AD	SN Tajima's D ^d	33, 22	-1.65, -1.41	0.685	0.409	0.560
post40 vs. pre	SN Tajima's D	5, 71	-1.06, -1.48	0.103		0.068
Complex vs. OMIM	SN Tajima's D	16, 78	-0.57, -1.48	7.54E-04	7.15E-04	<1E-04
Complex (short) ^c vs. OMIM	SN Tajima's D	6, 78	-0.90, -1.48	0.142		0.092
Cancer vs. OMIM	SN Tajima's D	6, 78	-2.29, -1.48	1.1E-03		<1E-04
Other vs. OMIM	SN Tajima's D	224, 78	-1.39, -1.48	0.408	0.326	
Cancer vs. Complex	SN Tajima's D	6, 16	-2.29, -0.57	1.87E-04		<1E-04
Cancer vs. Complex (short) ^c	SN Tajima's D	6, 6	-2.29, -0.90	0.026		0.002
Other vs. Complex	SN Tajima's D	224, 16	-1.39, -0.57	0.003	0.002	0.001
Other vs. Complex (short) ^c	SN Tajima's D	224, 6	-1.39, -0.90	0.251		0.196
Other vs. Cancer	SN Tajima's D	224, 6	-1.39, -2.29	3.73E-04		<1E-04
Essential vs. Cancer	SN Tajima's D	57, 6	-1.39, -2.29	6.55E-04		<1E-04
OMIM vs. Essential	SN Tajima's D	78, 57	-1.48, -1.39	0.221	0.502	
Other vs. Essential	SN Tajima's D	224, 57	-1.39, -1.39	0.739	0.935	
Complex vs. Essential	SN Tajima's D	16, 57	-0.57, -1.39	0.004	0.003	0.007
Complex (short) ^c vs. Essential	SN Tajima's D	6, 57	-0.90, -1.39	0.175		0.117
Complex (I) vs. OMIM (I) ^f	Genome D_n/D_s	33, 487	0.172, 0.133	0.054	0.002	0.035
Complex (short I) ^c vs. OMIM (I) ^f	Genome D_n/D_s	21, 487	0.203, 0.133	0.102	0.006	0.079
Complex (I) vs. OMIM (I) ^f	D_n/D_s	55, 922	0.210, 0.149	0.045	0.001	
Complex (short I) ^c vs. OMIM (I) ^f	D_n/D_s	34, 922	0.217, 0.149	0.028	0.036	0.03
Contrast	Data, Statistic	No. of Alleles ^e	Medians	KS Test ^a	Logistic Regression ^a	KS Bootstrap
OMIM vs. Complex	Nonsyn. Allele Freq.	1173, 70	0.025, 0.075	0.001	0.120	
OMIM vs. Other	Nonsyn. Allele Freq.	1173, 11527	0.025, 0.026	0.300	0.935	
Complex vs. Other	Nonsyn. Allele Freq.	70, 11527	0.075, 0.026	0.003	0.107	
OMIM vs. Complex	Syn. Allele Freq.	1392, 65	0.027, 0.025	0.982	0.975	

Table S1. *Continued*

Contrast	Data, Statistic	No. of Alleles ^e	Medians	KS Test ^a	Logistic Regression ^a	KS Bootstrap
OMIM vs. Other	Syn. Allele Freq.	1392, 11899	0.027, 0.026	0.971	0.727	
Complex vs. Other	Syn. Allele Freq.	65, 11899	0.025, 0.026	0.985	0.913	

Shown in red are cases in which the permutation test or logistic regression is significant at the 5% level; in orange at the 10% level (not corrected for multiple tests). Polymorphism-data results are presented for the European population samples; qualitatively similar results were obtained for the African and African-American population samples (not shown).

^a p values obtained with the use of a Kolmogorov-Smirnov (KS) test or logistic regression. For cases with fewer than 50 genes in each category, we estimated the p value of the KS test by a permutation test (see “Statistical Analyses” section). We only fit a logistic regression when there were at least ten genes in each category.

^b D_r/D_s and Genome D_r/D_s refer to two sets of alignments (see main-text Experimental Procedures).

^c Only genes with replicated associations are included (see main-text Experimental Procedures).

^d Refers to polymorphism data from SeattleSNPs and NIEHS projects (see main-text Experimental Procedures).

^e The frequencies of nonsynonymous alleles are compared between the groups. Numbers refer to the total numbers of alleles in each group.

^f Genes associated with immune response were excluded from both groups of genes in this analysis.

n_i is the sample size at site i , and

p_i is the allele frequency at site i .

W was defined following Tajima (1989)[S8], as:

$$W = \sqrt{e_1 S + e_2 S(S-1)},$$

in which

$$e_2 = \frac{c_2}{a_1^2 + a_2}, e_1 = \frac{c_1}{a_1},$$

$$c_2 = b_2 - \frac{n_{\max} + 2}{a_1 n_{\max}} + \frac{a_2}{a_1^2},$$

$$c_1 = b_1 - \frac{1}{a_1}, b_2 = \frac{2(n_{\max}^2 + n_{\max} + 3)}{9n_{\max}(n_{\max} - 1)},$$

$$b_1 = \frac{n_{\max} + 1}{3(n_{\max} - 1)},$$

$$a_2 = \sum_{k=1}^{n_{\max}-1} 1/k^2,$$

$$a_1 = \sum_{k=1}^{n_{\max}-1} 1/k,$$

S is the number of segregating sites, and

n_{\max} is the maximum sample size over all sites.

Statistical Analyses

To compare D_r/D_s or Tajima’s D values between categories, we excluded genes that were assigned to more than one mode-of-inheritance category (e.g., genes in which mutations lead to both recessive and dominant disease phenotypes). In turn, genes that belong to more than one age-at-onset category (i.e., genes that cause multiple diseases with different ages at onset) were assigned to the youngest age.

We used a Kolmogorov-Smirnov (KS) test to assess the significance of the difference between distributions, with the `ks.test` function in the R software environment for statistical computing (<http://www.r-project.org>). Although all the statistics that we considered are continuous, when there are few polymorphic or divergent sites per gene, they take on a limited number of values and effectively become discretized. To address this potential problem, in cases with fewer than 50 genes in each category, we estimated the p value by permutation, using the maximum difference between cdfs, Z , as our test statistic. Specifically, we randomly divided all the values into two groups of the same size as observed, then calculated Z . This permutation was repeated 10,000 times, and each time the maximal difference between the cdfs of the two randomly selected groups (Z_i) was recorded. The test p value was defined as the number of times in which $Z_i \geq Z$, divided by 10,000.

A second concern was that, in genes with few divergent sites, our estimate of D_r/D_s might be unreliable, leading to differences between categories driven by few outliers. To address this, we calculated a global D_r/D_s for each category as

$$G = \frac{\sum K_A / \sum N_A}{\sum K_S / \sum N_S}$$

in which K_A is the number of nonsynonymous mutations across all genes in the category, K_S is the number of synonymous mutations, N_A is the estimated number of nonsynonymous bases, and N_S is the estimated number of synonymous bases. Each sum was calculated over all genes in the tested category, with G calculated with the use of the four sums. We then considered the difference in G between categories as a test statistic. The significance of the difference was assessed by permutation; qualitative results were similar to those obtained with the use of a KS test (not shown).

Finally, we fit linear models to the D_r/D_s data, using the R software. To do so, values of D_r/D_s were Box-Cox transformed using

$$y^{(\lambda)} = (y^\lambda - 1)/\lambda.$$

The parameter λ was calculated with the use of the `box.cox.powers` R function (in the package “car”). The `lm` function was then used to fit linear models to the transformed values. Age at onset was represented as a binary categorical variable that stated whether a gene is in the “pre-40” or the “post-40” age category. As before, if a gene caused phenotypes with varying ages of onset, we assigned it to the youngest one. Mode of inheritance was represented by two categorical variables, the first stating the specific category (either autosomal dominant [AD], autosomal recessive [AR], or X-linked), and the second stating whether a gene is assigned to both AD and AR (in which case, the first variable would mark the gene as AD). GO categories (or pathophysiologies) were represented by N binary categorical variables, indicating the inclusion of a gene in a specific GO (or pathophysiology) category (N is the number of GO or pathophysiology categories for the included genes). We note that genes can, and often do, belong to more than one GO (pathophysiology) category and that this feature is taken into account in our analysis. In tests involving genes that contribute to complex-disease risk or are associated with cancer, we used a categorical variable that stated whether the gene is assigned to, for example, a “complex” or a “simple” phenotype. To compare the fit of two nested linear models, we used the `anova` function in R.

Estimating γ and ω from Human Polymorphism and Human-Rhesus Macaque Divergence Data

Human polymorphism data were taken from Bustamante et al. (2005) [S9], for which Celera Genomics resequenced the same 39 individuals for approximately 20,362 human genes. For divergence data, we used human-rhesus alignments taken from 10,376 1:1 orthologous alignments between human, chimpanzee, and rhesus macaque generated by the Rhesus Macaque Genome Sequencing and Analysis Consortium [S10]. Regions in these orthologous alignments that were not surveyed for polymorphism by Bustamante et al. were masked out. When a SNP is segregating, one base pair matches the rhesus macaque sequence and the other base matches the hg18 assembly, the nucleotide was altered to match the same base as in rhesus macaque.

We note that when comparing human-rhesus sequences, by chance alone some sites will have undergone multiple substitutions. Because synonymous mutations tend to saturate first, a parsimony approach tends to

Table S2. GO Categories that are Overrepresented in “Autosomal Dominant” versus “Autosomal Recessive” Categories and in “Autosomal Recessive” versus “Autosomal Dominant” Categories

Overrepresented in AR versus AD	
GO:0044444	cytoplasmic part
GO:0006091	generation of precursor metabolites and energy
GO:0019752	carboxylic acid metabolism
GO:0006082	organic acid metabolic process
GO:0006807	nitrogen-compound metabolic process
Overrepresented in AD versus AR	
GO:0050794	regulation of cellular process
GO:0050789	regulation of biological process
GO:0016070	RNA metabolic process
GO:0019219	regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process
GO:0045449	regulation of transcription

The five most-significant GO terms are shown in each case. The Gostat program [S17] (<http://gostat.wehi.edu.au/>) was used for this analysis. As noted previously (see [S18] and references therein), regulatory proteins tend to be overrepresented among the AD category, whereas enzymes tend to be overrepresented among the AR category (see Kondrashov et al. [S18] for possible explanations).

underestimate the synonymous versus the replacement mutations and, as a result, the MK test rejects neutrality more often than expected (R. Hernandez and C.D.B., unpublished work). Therefore, we obtained maximum-likelihood estimates of the expected number of silent and replacement fixed differences for each gene by using the codeml program of the PAML software package [S11], with codon frequencies estimated from nucleotide frequencies at each of the three codon positions, a single nonsynonymous-synonymous ratio $\omega = D_n/D_s$ across sites and branches, and independent transition-transversion ratios for each gene. The human-rhesus macaque comparisons are pairwise, so the tree consists of a single branch of evolutionary distance, t . The total number of silent sites and the total number of replacement sites were estimated from the codeml output with the use of the Nei-Gojobori method [S12].

Interpreting Estimates of γ When There is a Distribution of Selective Effects within a Gene

The mkprf method [S9, S13] estimates selection parameters assuming all mutations in a gene are subject to the same selection intensity, γ . This assumption is unlikely to be realistic. To examine the effect of a distribution of fitness effects on our inference of γ , we generated MK tables under various normal and exponential distributions of fitness effects by using the method of [S14]. We then compared our inference of γ with the mean selection intensity of segregating mutations (γ_S) and the mean selection intensity of mutations that had become fixed (γ_F) by using the equations of [S15] integrated over the distribution of fitness effects. Except for distributions tightly centered on a single γ , γ_S and γ_F will be less negative than the mean fitness effect of a new mutation (i.e., the mean of the distribution of fitness effects), because strongly deleterious alleles are lost from the population.

As expected, we found that the mkprf estimate closely matches γ_S when mutations are so deleterious that they rarely fix and that it is intermediate between γ_S and γ_F when the fitness distribution includes nearly neutral or positively selected sites (see Figure S3). Under an exponential distribution of fitness effects, the inferred γ is closer to γ_F than to γ_S because the distribution is not symmetric but, rather, skewed toward less-negative selective effects. On a log scale, however, the inferred γ under an exponential distribution is more intermediate. Most studies of the distribution of fitness effects conclude that it is not symmetric but, rather, exponential or gamma distributed [S16].

Resampling Analyses for Evaluation of the Effect of Ascertainment Bias

Complex-disease genes have been discovered primarily through association mapping and are therefore likely to harbor at least one allele at intermediate frequency. We wanted to assess the possible effect of this

Table S3. Comparison of Nested Linear Models

Models Compared	Dependent Var.	No. of Genes	p Value
Mode + GO2 versus GO2	D_n/D_s	917	10^{-5}
Mode + Pathophys. versus Pathophys.	D_n/D_s	719	$<10^{-5}$
Mode + %Mut versus %Mut	D_n/D_s	867	$<10^{-5}$
Age + GO2 versus GO2	D_n/D_s	936	0.470
Age + Pathophys. versus Pathophys.	D_n/D_s	730	0.558
Age+Mode+GO2 versus Age + GO2	D_n/D_s	910	$<10^{-5}$
Age+Mode+Pathophys.+GO2 versus Age+Pathophys.+GO2	D_n/D_s	713	$<10^{-5}$
Age+Mode+Pathophys.+GO2+%Mut versus Age+Pathophys.+GO2+%Mut	D_n/D_s	690	$<10^{-5}$
Disease type + GO2 versus GO2	D_n/D_s	1190	0.002
Disease type + Pathophys. versus Pathophys.	D_n/D_s	747	0.345

“Mode” refers to the mode of inheritance, “age” to the age at onset, “GO2” to the GO level-two gene category, “pathophys.” to the pathophysiology category, and “%Mut” to the number of known disease mutations in the protein divided by the total number of amino acid positions (derived from the humsavar.txt file, which was downloaded from Swissprot at <http://www.expasy.org> on Sept. 12, 2006). “Disease type” refers to whether the disorder is classified as “simple,” “complex,” or “cancer” (see “Creating a Hand-Curated Version of the OMIM Database” section for details).

The mode of inheritance remains significant after correction for functional categories (as captured by GO categories), pathophysiology, and the fraction of amino acid sites known to cause disease.

Disease type is significant after controlling for functional categories but not pathophysiology. However, we have only pathophysiology data for 19 genes associated with complex diseases, and there are 17 pathophysiology categories, so this might reflect lack of power.

ascertainment bias on analyses of the frequency spectrum in these genes. First, we examined whether this might account for the difference in the distribution of Tajima’s D values observed between hOMIM and complex-disease genes (see main-text section “*Comparison of Genes Associated with Simple versus Complex Diseases*”). To do so, we tried to mimic the discovery process by performing the following resampling analysis: Using the Applera polymorphism data, we identified hOMIM genes with at least one allele at intermediate frequency (i.e., between 0.1 and 0.9); this implicitly assumes that all genes were found through association studies, which is somewhat extreme but conservative for these purposes. From these genes, we randomly selected 29 genes (to match the number of complex-disease genes for which we had polymorphism data) and compared amino acid Tajima’s D values in the European sample for this random set with the entire hOMIM set by using a KS test, logistic regression, and a bootstrap KS test (see above). This procedure was repeated 10,000 times, and each time the p values for the three tests were recorded. In only 4.23%, 0.79%, and 2.85% of cases (for the three tests, respectively) did we observe a p value lower than or equal to the actual p value for the difference between the actual hOMIM and complex-disease genes. This analysis suggests that the ascertainment-bias effect alone is unlikely to explain the observed differences in Tajima’s D values between complex-disease genes and genes in other categories.

Next, we wanted to evaluate whether a small sample size or ascertainment bias in the discovery of genes associated with complex-disease risk could explain our finding that synonymous and nonsynonymous allele frequencies do not differ significantly (see main-text section “*Comparison of Genes Associated with Simple versus Complex Diseases*”). To do this, we selected 20 genes (the number of complex-disease genes used in this analysis) at random from hOMIM, conditional on their harboring at least one allele (synonymous or nonsynonymous) at a sample frequency between 0.1 and 0.9. Applying the same test as that applied to the actual data (i.e., a Wilcoxon matched-pairs signed-rank test), we then asked how often we obtained a p value as high as or higher than that observed for the difference between synonymous and nonsynonymous sites (i.e., $p = 0.842$ for the European sample). This occurred in 5.69% of cases. This analysis suggests that ascertainment bias and small sample size alone are also unlikely to account for the lack of difference in frequency between synonymous and nonsynonymous alleles in genes that contribute to complex-disease risk.

Supplemental References

- S1. Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Mellis, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: A generic system for fast and flexible access to biological data. *Genome Res.* 14, 160–169.
- S2. Dobyns, W.B., Filauro, A., Tomson, B.N., Chan, A.S., Ho, A.W., Ting, N.T., Oosterwijk, J.C., and Ober, C. (2004). Inheritance of most X-linked traits is not dominant or recessive, just X-linked. *Am. J. Med. Genet. A.* 129, 136–143.
- S3. Hawkes, K. (2003). Grandmothers and the evolution of human longevity. *Am. J. Hum. Biol.* 15, 380–400.
- S4. Livi-Bacci, M. (2001). A concise history of world population (Oxford, UK: Blackwell).
- S5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- S6. Barrie, P.A., Jeffreys, A.J., and Scott, A.F. (1981). Evolution of the beta-globin gene cluster in man and the primates. *J. Mol. Biol.* 149, 319–336.
- S7. Goodman, M., Koop, B.F., Czelusniak, J., and Weiss, M.L. (1984). The eta-globin gene. Its long evolutionary history in the beta-globin gene family of mammals. *J. Mol. Biol.* 180, 803–823.
- S8. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- S9. Bustamante, C.D., Fedel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157.
- S10. Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316, 222–234.
- S11. Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556.
- S12. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.
- S13. Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olson, K.M., Purganan, M.D., and Hartl, D.L. (2002). The cost of inbreeding in *Arabidopsis*. *Nature* 416, 531–534.
- S14. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. Assessing the evolutionary impact of amino-acid mutations in the human genome. *PLoS Genet.*, in press. Published online May 30, 2008. 10.1371/journal.pgen.1000083.
- S15. Sawyer, S.A., and Hartl, D.L. (1992). Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176.
- S16. Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618.
- S17. Beissbarth, T., and Speed, T.P. (2004). GOstat: Find statistically over-represented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465.
- S18. Kondrashov, F., and Koonin, E.V. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* 20, 287–290.