Current Biology *18*, 883–889, June 24, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.04.074

Report

Natural Selection on Genes that Underlie Human Disease Susceptibility

**Ran Blekhman,[1] Orna Man,[2] Leslie Herrmann,[3]
Adam R. Boyko,[4] Amit Indap,[4] Carolin Kosiol,[4]
Carlos D. Bustamante,[4] Kosuke M. Teshima,[1,5]
and Molly Przeworski[1,]***
[1]Department of Human Genetics
University of Chicago
Chicago, IL 60637
[2]Department of Structural Biology and Department
of Molecular Genetics
Weizmann Institute of Science
Rehovot 76100
Israel
[3]Warren Alpert Medical School
Brown University
Providence, RI 02912
[4]Department of Biological Statistics and
Computational Biology
Cornell University
Ithaca, NY 14853

Summary

What evolutionary forces shape genes that contribute to the risk of human disease? Do similar selective pressures act on alleles that underlie simple versus complex disorders [1–3]? Answers to these questions will shed light onto the origin of human disorders (e.g., [4]) and help to predict the population frequencies of alleles that contribute to disease risk, with important implications for the efficient design of mapping studies [5–7]. As a first step toward addressing these questions, we created a hand-curated version of the Mendelian Inheritance in Man database (OMIM). We then examined selective pressures on Mendelian-disease genes, genes that contribute to complex-disease risk, and genes known to be essential in mouse by analyzing patterns of human polymorphism and of divergence between human and rhesus macaque. We found that Mendelian-disease genes appear to be under widespread purifying selection, especially when the disease mutations are dominant (rather than recessive). In contrast, the class of genes that influence complex-disease risk shows little signs of evolutionary conservation, possibly because this category includes targets of both purifying and positive selection.

Results and Discussion

Diseases are thought to persist in human populations primarily because of a balance between mutation, genetic drift, and natural selection, with alleles that contribute to disease introduced by mutation, governed in part by random genetic drift, but eventually eliminated from the population by purifying selection [5, 7, 8]. For simple, highly penetrant disorders, purifying selection might be quite strong. For complex diseases, however, individual alleles might contribute little to overall risk and be only weakly deleterious [9]. Similarly, alleles that cause exclusively late-onset Mendelian disorders might not impose an evolutionary-fitness cost and thus could be under little or no selection. Disease susceptibility could also arise, not from a balance between mutation and purifying selection but as a consequence of adaptation. For example, there is evidence of heterozygote advantage (e.g., at β-globin) and of the fixation of compensatory alleles [10] in genes that cause Mendelian disorders, as well as indications that environmental shifts have led to changes in selection pressures over time. In particular, in a subset of genes associated with complex-disease risk, the susceptibility allele is ancestral, and population-genetic analyses suggest that the derived, protective allele is selectively advantageous ([3] and references therein). Finally, alleles could be subject to balancing selection if they increase the risk of one disease but decrease the risk of another or if there are important interactions between genotype and environment. These considerations raise the possibility that a fraction of loci that underlie contemporary human diseases have been the target of positive, as well as of purifying, selection.

The main approach to evaluating these hypotheses has been contrasting evolutionary rates in genes associated with Mendelian-disease phenotypes with those in all other genes by use of D_n/D_s, the ratio of nonsynonymous to synonymous substitutions. Assuming that synonymous substitutions are mostly neutral, D_n/D_s reflects the proportion of amino acid changes in a gene that reach fixation and therefore are not strongly deleterious. Thus, $(1 - D_n/D_s)$ is often thought of as an estimate of the evolutionary constraint acting on a gene (an underestimate if adaptations are frequent), which reflects the extent of purifying selection and, to a lesser extent, its strength. To date, results of comparisons between disease and "nondisease" genes have been conflicting: Two studies found significantly lower D_n/D_s values in genes that cause Mendelian disease than in other genes [8, 11], two found significantly higher values [12, 13], and one found no significant difference [14]. These divergent answers might be due to the reliance of most studies on the OMIM database. Although OMIM is the most exhaustive publicly available resource, its phenotypic information is sometimes outdated and is not entered in a standard format, rendering automated searches unreliable (see Supplemental Data, available online). A second limitation, for a subset of papers, might be the use of comparisons between human and rodent, because it is hard to estimate D_n/D_s reliably for such distantly related species. In addition, many genes classified as nondisease genes might nevertheless be under strong and widespread purifying selection, reducing the power to detect a difference between categories [8].

To overcome these limitations, we created a hand-curated version of OMIM (hereafter "hOMIM"), including only highly penetrant diseases caused by a mutation in an autosomal or X-linked gene (see Experimental Procedures). Because the vast majority of mutations currently known to underlie simple diseases are in exons, we focused on the coding regions, assessing levels of constraint by estimating D_n/D_s between

*Correspondence: mfp@uchicago.edu
[5]Present address: Graduate University for Advanced Studies, Kanagawa-ku, Yokohama 221-8686, Japan

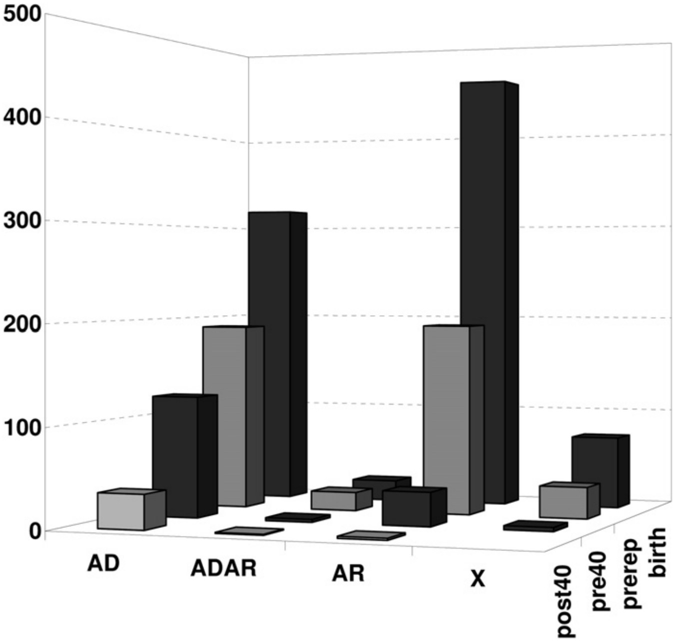


Figure 1. Mode of Inheritance and Age at Onset of Disease Phenotypes in Our Hand-Curated Version of the OMIM Database

Categories are autosomal dominant (AD), autosomal recessive (AR), both (ADAR), and X-linked (X). We also classified genes by the age at onset of the disorder (see Supplemental Experimental Procedures for details).

Data are in Table S4.

human and rhesus macaque. This Old World monkey last had a common ancestor with humans over 25 Mya [15], which is long enough for the comparison to be informative but short enough for the estimates of D_n/D_s to be reliable and for the two species to be more likely to share similar pathophysiologies. Finally, we used a classification of essential genes in mice to identify a subset of genes that are not currently associated with human disease but that are nonetheless likely to be conserved in mammals [16].

Analysis of hOMIM Genes

We first compared rates of protein evolution among genes in hOMIM (see Experimental Procedures), with the prediction that, all else being equal, genes in which mutations cause solely late-onset disorders should be less conserved than those in which mutations cause earlier-onset disorders. We further expected that if weak purifying selection is common (i.e., if the selection coefficients acting on homozygotes are often in the range $-8 < N_e s < 0$, where N_e is the effective population size) [7, 11], genes in which mutations cause recessive disorders should have higher D_n/D_s values than do those in which mutations lead to dominant disorders (e.g., Figure 8 in [17]). We therefore tabulated from OMIM entries information about the age at onset of the disorder and the mode of inheritance (see Experimental Procedures), then we assessed whether this information predicts the evolution of genes underlying simple disorders. Because the entire coding region is used to test these predictions, a key assumption is that the mode of inheritance and age at onset are predictive of these attributes for other mutations in the same gene.

As expected, most Mendelian disorders with a known genetic basis are early-onset disorders, with only a small set manifesting themselves after age 40 (Figure 1). Overall, 45.3% of the disease phenotypes are recessive; the data further suggest that early-onset disorders are more likely to be recessive and that late-onset disorders are more likely to be dominant, but these findings could also reflect ascertainment bias (e.g., the greater difficulty in mapping loci underlying early-onset, dominant disorders).

Considering divergence between human and rhesus macaque (hereafter, "human-rhesus macaque divergence"), we found no evidence that genes in which mutations cause earlier-onset disorders have lower D_n/D_s values than those in which mutations cause later-onset disorders (Table S1). This could simply reflect a lack of power, given that we have data on very few genes (14) that cause *exclusively* late-onset disorders; alternatively, mutations in the genes might have pleiotropic effects, or the age at onset might have been earlier in the past [18].

In contrast, we found a highly significant effect of the mode of inheritance on conservation levels of the protein ($p << 10^{-3}$; see Supplemental Data): D_n/D_s values tend to be higher in genes with recessive disease mutations (median = 0.184, n = 452) than in those with dominant disease mutations (median = 0.084, n = 294), and they tend to be intermediate in X-linked genes (median = 0.138, n = 64) (Figure 2; see also [19]). This association could reflect a confounding factor. In particular, the mode of inheritance is known to vary markedly among GO functional categories (e.g., Table S2; see Experimental Procedures for details); however, it remains a highly significant predictor of D_n/D_s values after these and other possible covariates are controlled for (Table S3).

We then combined human polymorphism data and data on human-rhesus macaque divergence to estimate the fraction of amino acid sites that are not strongly deleterious, ω. We also estimated the selection coefficient acting on homozygote mutations in disease genes, γ (assuming a fixed selective effect); this value can be thought of as a summary of the pooled polymorphism and divergence data for genes in a given category (see Experimental Procedures). As shown in Figure 3, there appears to be more-widespread and stronger purifying selection on genes associated with dominant rather than with recessive disease phenotypes.

Comparison of Genes Associated with Simple versus Complex Diseases

Next, we compared conservation levels of genes in hOMIM to those of genes in which mutations are associated with cancer or contribute to other complex-disease susceptibility, genes for which knockouts are inviable or cause sterility in mice [16] (hereafter, "essential genes"), and "other" genes not known to influence disease risk (see Experimental Procedures). Comparisons of D_n/D_s values suggest that, as a class, proteins that are essential in mouse and those in which mutations are associated with cancer evolve the most slowly (Figure 4; median Genome D_n/D_s = 0.077 and 0.061, respectively). In turn, the coding regions of hOMIM genes tend to be slightly, but significantly, more slowly evolving than are genes not associated with disease (median Genome D_n/D_s = 0.133 versus 0.139, respectively; see Table S1 for p values).

The polymorphism data further suggest widespread purifying selection on amino acid sites in these gene categories (Figure 4). Notably, in all three sets of genes, nonsynonymous

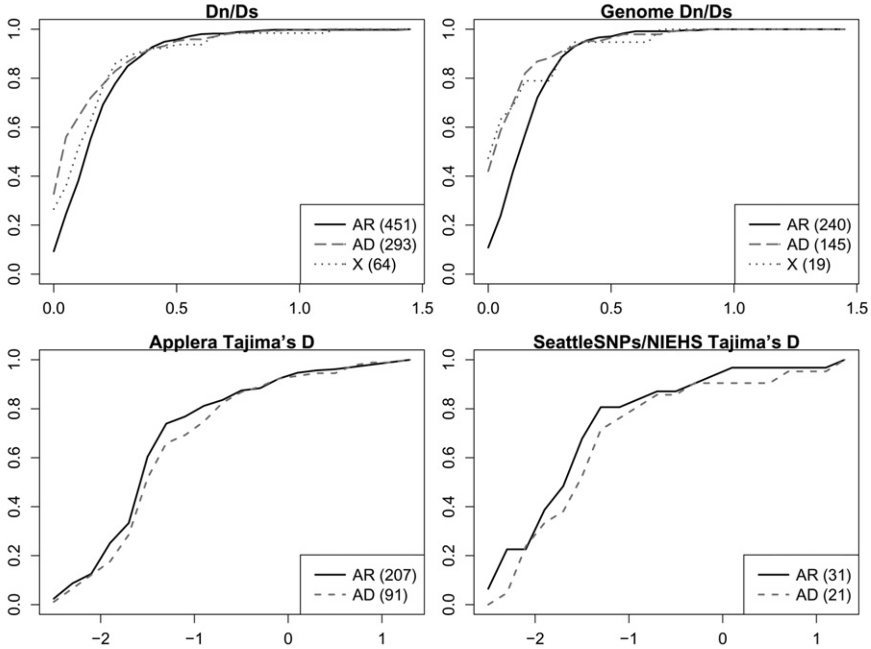


Figure 2. Cumulative Distributions of D_n/D_s and Tajima's *D* as a Function of the Mode of Inheritance

The value of the statistic is given on the *x* axis. AR refers to autosomal recessive and AD to autosomal dominant. In parentheses are the numbers of genes in each category. D_n/D_s plots are shown for two sets of human-rhesus macaque alignments. The distributions of D_n/D_s for AD and AR categories are significantly different from one another, but the distributions of Tajima's *D* values are not (see Table S1). Tajima's *D* was calculated for amino acid variants with the use of a European population sample; when an African-American sample is used instead, the order of AR and AD is reversed, but again, the distributions are not significantly different (not shown).

variants occur at significantly lower frequencies than do variants at synonymous sites (see Figure S2). A similar conclusion emerges when we combined polymorphism and divergence data to estimate selection parameters ω and γ (Figure 3). Thus, our findings lend further support to the hypothesis that proteins that underlie Mendelian-disease or are associated with human cancers evolve primarily under purifying selection.

Although a model of mutation-selection balance was also proposed for genes that influence complex-disease risk [7], this group does not show evidence of conservation greater than that of non-disease-associated genes. Instead, it tends to have a higher D_n/D_s ratio (median Genome $D_n/D_s = 0.203$) than do hOMIM genes or "other" genes, consistent with one of the earlier reports regarding comparisons between human and mouse [14]. This difference between genes associated with complex versus Mendelian diseases is still significant after correction for GO categories and after exclusion of genes associated with immune response (median Genome D_n/D_s after exclusion = 0.172; see Experimental procedures and Figure S1).

In addition, in genes associated with complex-disease susceptibility, the frequencies of amino acid alleles tend to be higher than in other categories of genes, including genes not associated with disease (Figure 4). Additionally, the amino acid allele frequencies do not differ significantly from those of silent variants (Figure 2). These findings do not appear to be explained solely by the ascertainment bias of complex-disease-gene discovery or by the smaller number of genes in this category (see Supplemental Data). Together, they suggest that genes associated with complex-disease susceptibility tend to be under less-pervasive purifying selection than are other classes of essential or disease genes. In further support of this conclusion, the estimate of ω is higher for genes associated with complex-disease risk than for Mendelian or even for non-disease-associated genes, as is the estimate of the selection coefficient, γ (Figure 3).

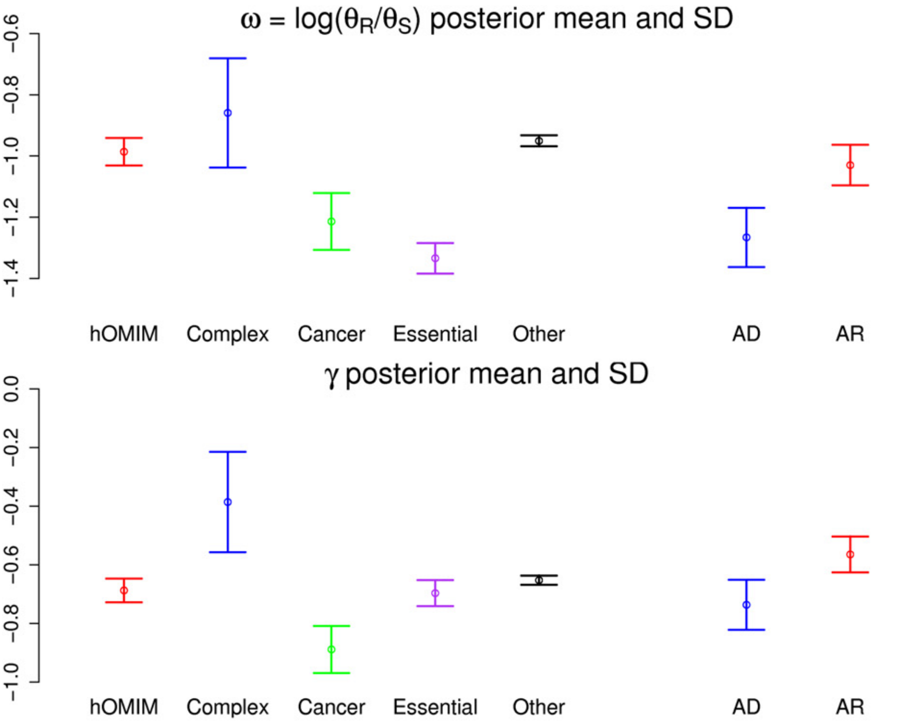
Why would this be the case? Two (non-mutually exclusive) explanations are: (1) A substantial fraction of "other genes," although not known to be essential in mouse or to be associated with human disease, are in fact under widespread and strong purifying selection. In contrast, alleles that contribute exclusively to complex diseases tend to explain only a small proportion of disease risk [9] and to have late-onset effects, so they might have few fitness consequences. If so, changes in genes associated with complex-disease risk could be under very weak, if any, purifying selection. (2) Genes that influence complex-disease susceptibility include loci under widespread purifying selection but are also enriched for targets of positive selection, thus appearing to be less conserved when considered as a class. For example, if we consider all candidate loci evaluated for evidence of selection by Sabeti et al. [20], there appears to be an enrichment for targets of selection among genes associated with complex-disease risk relative to Mendelian-disease genes: 8.3% (six out of 72) of genes fall in the empirical 5% tail of the distribution of at least one statistic in at least one population, whereas only 1.1% (11 out of 1004) of genes in hOMIM do so ($p = 5 \times 10^{-4}$, by a one-tailed Fisher's exact test). Complex-disease mapping is in its infancy, so it is too early to reliably distinguish between hypotheses–especially given that the genes that have been found to date are probably an unrepresentative subset (see Experimental Procedures). Nonetheless, existing data raise the possibility that, whereas simple disorders are generally well-described by models of purifying selection, complex-disease susceptibility is tied, at least in part, to evolutionary adaptations.

Experimental Procedures

Hand-Curating OMIM

Our goal was to create a list of all genes that contribute to human diseases with a simple genetic basis. To do so, we used the Online Mendelian Inheritance in Man database (OMIM; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM/). OMIM is the most exhaustive, publicly available repository of information about human-disease phenotypes. However, it suffers from a number of limitations: for example, entries do not have a standard format, and outdated information is supplemented with new data rather than replaced. Moreover, although most phenotypic entries are Mendelian or at least have a simple genetic basis, a nonnegligible fraction are clearly complex in etiology (e.g., autism). These features make automated queries highly unreliable.

We therefore decided to create a hand-curated summary of the OMIM database (hereafter referred to as "hOMIM"), consisting of a list of pairs (gene, phenotype) together with phenotypic information about the mode of inheritance and age at onset. A description of how the list was constructed is



Complex	Fixed	Segregating
Silent	545.8	51
Replacement	395.6	49

Odds Ratio: 0.7545 (95% CI: 1.1663 – 0.4885)

Other	Fixed	Segregating
Silent	114515.7	9879
Replacement	57440.4	8885

Odds Ratio: 0.5576 (95% CI: 0.5749 – 0.5409)

Essential	Fixed	Segregating
Silent	11492.5	1023
Replacement	3943.6	636

Odds Ratio: 0.5519 (95% CI: 0.6142 – 0.4961)

hOMIM	Fixed	Segregating
Silent	10912.8	961
Replacement	4979.7	798

Odds Ratio: 0.5494 (95% CI: 0.6079 – 0.4967)

Cancer	Fixed	Segregating
Silent	2665.3	213
Replacement	872.8	173

Odds Ratio: 0.4029 (95% CI: 0.5027 – 0.3233)

Figure 3. Selection Parameters Estimated from Polymorphism and Divergence Data

In the left panel are estimates of two parameters, ω and γ, obtained from pooled polymorphism and divergence data in different categories of genes, including those in hOMIM, those associated with complex-disease susceptibility ("complex"), those associated with cancer ("cancer"), those for which knockouts are inviable or cause sterility in mice ("essential"), and those in none of the above categories ("other"). Genes in hOMIM are further broken down into two categories, depending on whether mutations cause dominant ("AD") or recessive ("AR") disease phenotypes. Shown are the mean and the standard deviation of the posterior-distribution estimate for each parameter. The parameter ω = log(θ_R/θ_S) can be thought of as the fraction of amino acid mutations that contribute to polymorphism, i.e., that are neutral or nearly neutral (θ_R is the effective mutation rate at replacement sites, and θ_S is that at synonymous sites), and γ is the selection coefficient acting on amino acid mutations in a category of genes. The estimates are obtained by the assumption of one selection coefficient γ for all mutations within a category; given this unrealistic assumption, the value of the γ estimate is less informative than is the ordering for the different categories (see Supplemental Data for details). In the right panel are summaries of the pooled polymorphism and divergence data (i.e., McDonald-Kreitman tables) for genes in each category (see Experimental Procedures for details). We note that γ can also be thought of not as a parameter estimate but as a summary of the table for each category, thereby capturing information similar to that captured by the odds ratio.

provided in the Supplemental Experimental Procedures, and the list is available in Table S4. This process yielded a list of 1685 unique pairs (gene, phenotype), corresponding to 1039 distinct genes, for examination. To run our analyses, we excluded phenotypes that were clearly complex or caused by triplet-repeat expansions; 1613 pairs remained.

In our analysis of Mendelian-disease genes, we also tried using a smaller list of OMIM pairs (gene, phenotype), compiled independently by Jimenez-Sanchez et al. (2001) [21] with the use of slightly different criteria; the qualitative conclusions were the same (results not shown).

List of Genes that Contribute to Complex-Disease Susceptibility

To create a list of genes that influence complex-disease susceptibility, we relied on two sources. First, we used compilations in three surveys of association studies [2, 22, 23]. To create a more stringent set of genes, we used only genes for which the associations had been replicated at least once or for which a meta-analysis supported the original association (i.e., bolded entries in Table 2 of [22], as well as entries in Table 2 of [23] and Table 1 of [2]). Second, we tabulated results from genome-wide association studies of complex-disease susceptibility that were published by June 7, 2007 (see Table S5 for references). Of the associations reported in these studies, we retained only cases in which the association had been replicated and a specific candidate gene had been identified by the investigators. From these sources, we found 72 genes that are associated with complex diseases but are not known to cause Mendelian diseases (i.e., not included on our hand-curated version of OMIM), of which 46 met our more stringent criteria. In our analysis, we considered genes that contribute both to complex-disease risk and to Mendelian diseases as Mendelian-disease genes.

In addition, we analyzed a set of 363 genes in which somatic or germline mutations are associated with cancer susceptibility (the complete working list is available from http://www.sanger.ac.uk/genetics/CGP/Census/) as of Feb. 13 2007, as well as a set of genes for which knockouts were inviable or caused sterility in mice [16] (downloaded from http://www.umich.edu/~zhanglab/download/Liao_MBE2006_update/essential.txt). When comparing classes of genes, we classified genes that belong to multiple categories in the following order of priority: hOMIM, complex, cancer, essential, other; such that genes are only in the "other" category if they are not associated with any type of disease and not known to be essential in mouse. We also ran the D_n/D_s analyses, excluding the genes that belonged to multiple categories, and the results were unchanged (not shown).

GO Categories and Patho-Physiologies

In order to examine the functional annotation of genes, we used the Gene Ontology (GO) database (http://www.geneontology.org/). Specifically, we retrieved the (level 2) GO assignment of each gene by examining the specific GO terms with which each gene is associated, as determined by the European Bioinformatics Institute (http://www.ebi.ac.uk/). We then located each of these terms on the overall directed acyclic graph (DAG) structure of GO and traced them back to their ancestral terms at this level of annotation. Both the EBI annotations of the genes and the entire DAG structure were downloaded from the database site on September 21, 2006. In one analysis, we excluded genes associated with immune response by removing all genes that are associated with the immune-system-process ontology (GO:0002376) or with any of its subontologies. We also used the pathophysiology classifications of Huang et al. [13]. The GO categories for each gene are available in Table S6.

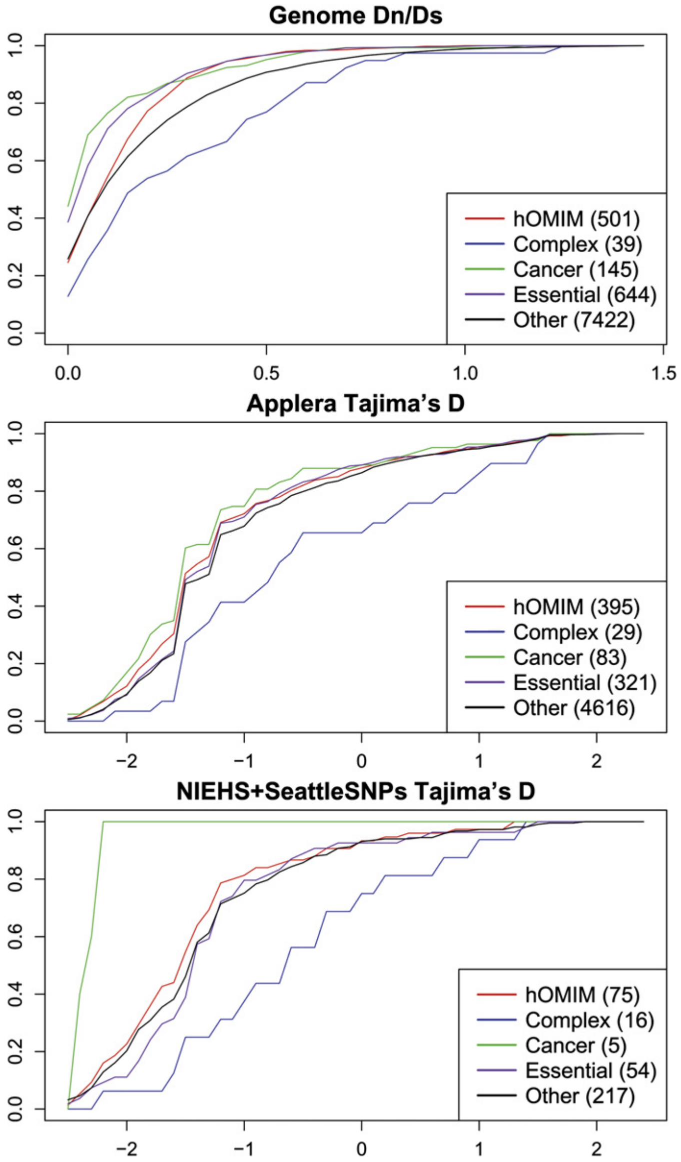


Figure 4. Conservation of Genes in Different Disease Categories

Cumulative distributions of D_n/D_s and Tajima's *D* for genes in hOMIM, those associated with complex-disease susceptibility ("complex"), those in which mutations are associated with cancer ("cancer"), those for which knockouts are inviable or cause sterility in mice ("essential"), and those in none of the above categories ("other"). For other details, see the legend of Figure 2. The distributions of D_n/D_s are significantly different for all pairwise comparisons (at the 5% level) except those of "essential" genes versus "cancer" genes and those of "other" genes versus "complex" disease, in which significance is marginal (see Table S1). The distributions of Tajima's *D* values in the larger Applera dataset (shown here for the European samples) are significantly different (at the 5% level) in genes associated with complex diseases versus either hOMIM genes or generic genes (see Table S1); all other pairwise comparisons are also significant, other than those of "cancer" genes versus "hOMIM" genes, "hOMIM" genes versus "essential" genes, and "other" genes versus "essential" genes.

Human Coding Sequences

The Refseq collection of human transcripts was downloaded from ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot on March 18, 2006. For each gene on our list, we examined all records corresponding to it and selected the longest coding sequence for the gene. In the case of *IGKC*, Ig kappa chain C region, which does not have a record in refseq, we used the coding sequence in GenBank record BC073791.1.

Estimates of Human-Rhesus Macaque D_n/D_s

For divergence data, we used human-rhesus macaque alignments taken from 10,376 1:1:1 orthologous alignments between human, chimp, and rhesus macaque [24], kindly provided by Adam Seipel at Cornell University. We estimated D_n/D_s for each gene by using the PAML package [25] with the default parameters for nuclear DNA. We excluded cases in which synonymous divergence was 0 and set D_n/D_s to 0 when nonsynonymous divergence was 0. This set of estimates is referred to throughout as the "Genome D_n/D_s" values. To map the genes to those on our compilations of disease associations, we used all known gene symbols and aliases from the kgAlias table at the UCSC Genome Database. Genes from the Cornell dataset for which we could not find a symbol were not included in the analysis.

These data only provided alignments for 50% of genes in hOMIM. To increase the number of Mendelian- and complex-disease genes for which we could estimate D_n/D_s, we also built our own alignments. For this purpose, unassembled sequence of the rhesus macaque genome was downloaded on Feb. 17, 2006 from http://www.hgsc.bcm.tmc.edu/projects/rmacaque/; for details on how orthology was determined, see the Supplemental Data. Each human gene sequence was aligned to its rhesus macaque ortholog with the use of the GAP program [26]. Using the translation of the coding sequence of the human gene, we retained only positions corresponding to whole codons. If an insertion in the rhesus macaque sequence occurred within codons, the codons affected by the insertion were removed, as were codons in which the rhesus macaque sequence contained a stop codon. We used the PAML package [25] to estimate the D_n/D_s ratio for the resultant pairs of aligned orthologous sequences. Only genes for which the rhesus macaque sequence covered at least 50% of the human sequence were included in the analyses. The D_n/D_s estimates for all genes analyzed are available in Tables S7–S10.

Human Polymorphism Data

We analyzed polymorphism data from two resequencing efforts, the NIEHS SNPs (http://egp.gs.washington.edu/) and the SeattleSNPs (http://pga.gs.washington.edu/) databases, on August 21, 2006. We analyzed European samples and African (or African-American) samples separately. Sub-Saharan African populations do not appear to have experienced a recent bottleneck, in contrast to European populations (e.g., [27]), so their allele frequencies might be closer to mutation-selection balance. On the other hand, much of the anecdotal evidence for selection on genes associated with complex-disease risk is in regard to Europeans (e.g., [28]).

In addition, we analyzed the resequencing polymorphism data in the Applera dataset [11], a genome-wide resequencing effort, considering European-American or African-American samples separately. We also ran the same analyses pooling all population samples, and the qualitative conclusions were unchanged (results not shown). The Applera project sequenced a chimpanzee to infer the ancestral state, and we used their inference to construct a derived frequency spectrum (see below). We mapped the Applera dataset genes to genes in our lists of Mendelian- and complex-disease genes as described for the Rhesus Macaque Genome Sequencing and Analysis Consortium alignments.

We used the set of nonsynonymous polymorphisms to calculate Tajima's *D* [29], a summary of the (folded) allele-frequency spectrum known to be sensitive to the effects of purifying selection [29]. To do so, we excluded SNPs with small sample sizes (< 10 individuals) and more than 10% missing data, as well as genes with 0 nonsynonymous polymorphisms; see the Supplemental Data for the formula used. The Tajima's *D* values are available in Tables S9 and S10. We also calculated the frequency spectrum for each gene by creating 20 bins of allele frequencies (< 5%, 5%–10%, etc.) and tabulating the number of alleles in each bin. We then created an "average frequency spectrum" for each category (e.g., "autosomal dominant") by summing the number in each bin over all genes in that category (effectively concatenating all genes in a given category).

Statistical Analyses

To assess whether the distributions of a statistic (D_n/D_s or Tajima's *D*) differed between two groups of genes (e.g., those in which mutations cause autosomal-dominant versus autosomal-recessive disorders), we used a Kolmogorov-Smirnov test. Details are provided in the Supplemental Data. To test whether D_n/D_s or Tajima's *D* predicted the odds of belonging to a given category, we performed logistic regressions by using the R function glm with the binomial parameter (http://www.r-project.org). A p value was calculated with the use of the anova function.

To examine the selective pressures acting on amino acid variants, we calculated the mean derived allele frequency for synonymous and for nonsynonymous SNPs for each gene. To assess whether they differed, a Wilcoxon matched-pairs signed-rank test was performed on the two paired-value lists with the use of the wilcox.test function in R, with only genes that had both synonymous and nonsynonymous SNPs in the sample considered.

Estimates of γ and ω

We estimated two selection parameters, γ and ω, by using a Bayesian method (mkprf) that relies on the entries of a McDonald-Kreitman Table [11]. The parameter $\gamma = 2N_es$ (in which N_e is the effective population size) is the scaled selection coefficient acting on homozygous carriers of amino acid mutations. In turn, $\omega = \log(\theta_R/\theta_S)$ is a measure of constraint on amino acid mutations (cf [30]): θ_R and θ_S are estimates of the effective rate of replacement and silent mutations, so that their ratio indicates what fraction of amino acid mutations can contribute to polymorphism (i.e., is not strongly deleterious).

The mkprf approach uses the number of synonymous and nonsynonymous polymorphisms with humans and the number of synonymous and nonsynonymous fixed differences between species (here, human and rhesus macaque). Attractive features of the method are that it uses information from polymorphism and divergence jointly and that it depends on only the number of polymorphisms, not their frequency, and so should be insensitive to possible ascertainment-bias effects on the frequency spectrum of genes associated with complex disease. We relied on the polymorphism data from the Applera project, pooling population samples; more details are provided in the Supplemental Data. Specifically, we summed the entries of the MK tables for all genes within a category (e.g., all genes associated with complex-disease susceptibility), excluding X-linked genes (see [11] for details). This approach assumes a fixed selection coefficient across mutations and all genes, effectively averaging over the distribution of selective effects of mutations that contribute to polymorphism or divergence. This highly restrictive assumption makes the absolute value of γ difficult to interpret; however, its ordering across categories is meaningful for a wide variety of distributions of selection coefficients (see Supplemental Data). Moreover, γ can also be thought of not as a parameter estimate but as a summary of the pooled MK tables for each category, thereby capturing information similar to the odds ratio (see Figure 3). For all genes, we assumed a dominance coefficient $h = ½$, but we note that, other than in the case of overdominance (i.e., $h > 1$), this assumption does not affect estimates of the selection coefficients acting on homozygotes [17].

The Allele-Frequency Spectrum of Genes Associated with Complex Disease

In the analysis of the allele frequency in genes associated with complex disorders, it is important to note a number of ascertainment biases. Indeed, genes known to influence complex-disease risk have been identified mainly by association studies, so they are likely to harbor at least one common allele [5]. We ran resampling analyses to assess the possible effect of this ascertainment bias on Tajima's *D* for amino acid sites and found it to be relatively minor (see Supplemental Data), whereas the effects on D_n/D_s and estimates of γ from the mkprf method are expected to be negligible (see above).

A second consideration is that genes first discovered to influence complex-disease risk probably have unusually large effects on the disease phenotype, which implies that common alleles yet to be discovered are likely to explain a smaller proportion of the variance. If so, one might predict that the genes yet to be discovered will be under weaker selection. This said, there might also remain unknown genes associated with complex-disease risk that harbor rare alleles of large effect and are relatively more conserved than genes identified to date.

Evidence for Positive Selection in Genes Associated with Disease

Sabeti et al. (2006) [20] considered all genes previously reported to be under positive selection and assessed whether patterns of polymorphism and divergence were unusual relative to background patterns of variation in the genome. For each gene, they reported the percentiles of the distribution of various test statistics designed to detect signatures of selection (their Table S4). We used their criterion, considering a gene as showing evidence for selection if it fell in the 5% tail of at least one statistic in at least one of the three populations. This included six genes on our list of complex-disease genes (out of 72) but only 11 genes in hOMIM (out of 1004). We note that the study by Sabeti et al. predates the publication of one of the best-characterized cases of positive selection in a gene associated with complex disease, *TCF7L2* [28]. Moreover, the few genes in hOMIM that showed evidence of selection might be unusual, given that they include *HFE* and *BRCA1* (which others have considered as associated with complex, rather than Mendelian, disorders [14]), as well as genes such as *G6PD* and *HBB*, which are known to be involved in the resistance to malaria.

Supplemental Data

Additional experimental procedures, three figures, and ten tables are available at http://www.current-biology.com/cgi/content/full/18/12/883/DC1/.

Acknowledgments

Thanks to G. Coop, N. Cox, A. Di Rienzo, Y. Gilad, C. Ober, G. Sella, M. Stephens, J. D. Wall, W. Wen, and especially J. Pritchard for helpful discussions and/or comments about the manuscript. We are also grateful to the Rhesus Macaque Genome Sequencing and Analysis Consortium in general and to A. Siepel in particular for providing a set of human-rhesus alignments, to K. Bullaughey and A. Fledel-Alon for help with bioinformatics, and to M. Groat, L. Stephens, and B. Rashidi for their aid in the curation of the OMIM database. This work was supported by an Alfred P. Sloan fellowship in Computational Molecular Biology and by a National Institutes of Health (NIH) grant GM79558 to M.P.; R.B. is supported by a NIH grant GM077959 to Y. Gilad.

Received: October 9, 2007
Revised: April 27, 2008
Accepted: April 30, 2008
Published online: June 19, 2008

References

1. Zwick, M.E., Cutler, D.J., and Chakravarti, A. (2000). Patterns of genetic variation in Mendelian and complex traits. Annu. Rev. Genomics Hum. Genet. *1*, 387–407.
2. Lohmueller, K.E., Mauney, M.M., Reich, D., and Braverman, J.M. (2006). Variants associated with common disease are not unusually differentiated in frequency across populations. Am. J. Hum. Genet. *78*, 130–136.
3. Di Rienzo, A. (2006). Population genetics models of common diseases. Curr. Opin. Genet. Dev. *16*, 630–636.
4. Keller, M.C., and Miller, G. (2006). Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best? Behav. Brain Sci. *29*, 385–404.
5. Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: Common disease-common variant…or not? Hum. Mol. Genet. *11*, 2417–2423.
6. Cohen, J.C. (2006). Genetic approaches to coronary heart disease. J. Am. Coll. Cardiol. *48*, A10–A14.
7. Kryukov, G.V., Pennachio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.
8. Kondrashov, F.A., Ogurtsov, A.Y., and Kondrashov, A.S. (2004). Bioinformatical assay of human gene morbidity. Nucleic Acids Res. *32*, 1731–1737.
9. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., et al.Wellcome Trust Case Control ConsortiumAustralo-Anglo-American Spondylitis Consortium (TASC) (2007). Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat. Genet. *39*, 1329–1337.
10. Kondrashov, A.S., Sunyaev, S., and Kondrashov, F.A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA *99*, 14878–14883.
11. Bustamante, C.D., Fledel-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.

12. Smith, N.G., and Eyre-Walker, A. (2003). Human disease genes: Patterns and predictions. Gene *318*, 169–175.
13. Huang, H., Winter, E.E., Wang, H., Weinstock, K.G., Xing, H., Goodstadt, L., Stenson, P.D., Cooper, D.N., Smith, D., Alba, M.M., et al. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol. *5*, R47.
14. Thomas, P.D., and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proc. Natl. Acad. Sci. USA *101*, 15398–15403.
15. Goodman, M., Porter, C., A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. (1998). Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. *9*, 585–598.
16. Liao, B.Y., Scott, N.M., and Zhang, J. (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol. Biol. Evol. *23*, 2072–2080.
17. Williamson, S., Fledel-Alon, A., and Bustamante, C.D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics *168*, 463–475.
18. Fogel, R.W. (2005). Changes in the disparities in chronic diseases during the course of the 20th century. Perspect. Biol. Med. *48*, S150–S165.
19. Furney, S.J., Alba, M.M., and Lopez-Bigas, N. (2006). Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics *7*, 165.
20. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
21. Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human disease genes. Nature *409*, 853–855.
22. Hirschhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genet. Med. *4*, 45–61.
23. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. *33*, 177–182.
24. 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.
25. Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.
26. Huang, X. (1994). On global sequence alignment. Comput. Appl. Biosci. *10*, 227–235.
27. Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA *102*, 18508–18513.
28. Helgason, A., Palsson, S., Thorleifsson, G., Grant, S.F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., et al. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat. Genet. *39*, 218–225.
29. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
30. Gilad, Y., Bustamante, C.D., Lancet, D., and Paabo, S. (2003). Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am. J. Hum. Genet. *73*, 489–501.