Similarly Strong Purifying Selection Acts on Human Disease Genes of All Evolutionary Ages

James J. Cai, Elhanan Borenstein,*† Rong Chen,‡ and Dmitri A. Petrov**

*Department of Biology, Stanford University; †Santa Fe Institute; and ‡Department of Medicine, Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine

A number of studies have showed that recently created genes differ from the genes created in deep evolutionary past in many aspects. Here, we determined the age of emergence and propensity for gene loss (PGL) of all human protein–coding genes and compared disease genes with non-disease genes in terms of their evolutionary rate, strength of purifying selection, mRNA expression, and genetic redundancy. The older and the less prone to loss, non-disease genes have been evolving 1.5- to 3-fold slower between humans and chimps than young non-disease genes, whereas Mendelian disease genes have been evolving very slowly regardless of their ages and PGL. Complex disease genes showed an intermediate pattern. Disease genes also have higher mRNA expression heterogeneity across multiple tissues than non-disease genes regardless of age and PGL. Young and middle-aged disease genes have fewer similar paralogs as non-disease genes of the same age. We reasoned that genes were more likely to be involved in human disease if they were under a strong functional constraint, expressed heterogeneously across tissues, and lacked genetic redundancy. Young human genes that have been evolving under strong constraint between humans and chimps might also be enriched for genes that encode important primate or even human-specific functions.

Introduction

Mapping and identification of disease-causing genes in humans has a long history, predating even the discovery of DNA as the genetic molecule and the determination of the number of human chromosomes in 1950s (Haines and Pericak-Vance 1998). Today, classical map-based gene discovery has been augmented by the sequence-based gene discovery, given that the human genome project has produced high-precision tools for disease gene mapping and identification (Haines and Pericak-Vance 1998; Botstein and Risch 2003; Dean 2003; International Human Genome Sequencing Consortium 2004; Giallourakis et al. 2005). So far, the characterization of genetic defects has been successfully accomplished in more than 1,600 human Mendelian (i.e., monogenic) diseases, where one major gene has a high impact and environment or lifestyle has very little effect on the clinical outcome of patients. Mapping common and genetically complex human disease traits has proved more difficult but even in these more complex cases, a number of mutations associated with human complex diseases have been identified.

Studying the evolution of the hereditary basis of human disease can shed light onto the origins of human disorders and the factors that cause disease-causing mutations to be retained in human populations. Understanding what kind of genes are most likely to harbor disease-causing mutations, when the disease-causing alleles originated, why these disease-causing mutants segregate in human population, and how natural selection shaped the distribution of disease-causing mutations in the human genome is of great interest. Understanding the evolution of genes implicated in human inherited disorders has become one of the primary goals of evolutionary genetics.

One way to investigate the genes that harbor disease-causing mutations (which we term "disease genes") is to evaluate the way natural selection shapes their protein-coding portions. A number of studies have measured the strength of purifying selection acting on disease genes relative to non-disease genes. However, the results have been contradictory. An early study found that human disease genes have 24% higher level of *K*a/*K*s (the ratio of nonsynonymous substitution rate to synonymous substitution rate) than non-disease genes (Smith and Eyre-Walker 2003), suggesting that disease genes are subject to weaker purifying selection. However, later studies reported either that there was no difference in *K*a/*K*s between disease genes and non-disease genes (Huang et al. 2004; Thomas and Kejariwal 2004; Winter et al. 2004) or that disease genes exhibited lower *K*a/*K*s values (Kondrashov et al. 2004; Bustamante et al. 2005; Blekhman et al. 2008; Hsiao and Vitkup 2008). The discrepancy has been attributed to the small number of genes sampled in the early study (i.e., Smith and Eyre-Walker 2003) and possibly to the variation in the types of genes investigated in different studies (such as variable proportions of Mendelian and complex disease genes or genes involved in metabolic and immune diseases; Huang et al. 2004). This lack of consistency in the estimates of the rate of protein evolution in disease genes is not fully understood.

Recently, it has been shown that disease genes tend to be "old" (Domazet-Loso and Tautz 2008). Here, gene age was measured using the phylogenetic breadth of the distribution of homologous genes among different lineages. For humans, old genes are those that are present in more distantly related species like yeast and *Ciona*, whereas young genes are those that are present only in the closely related species like chimpanzee and macaque. It is known that younger genes tend to show accelerated evolutionary rates with respect to older genes (Alba and Castresana 2005; Toll-Riera et al. 2009). If most disease genes are old, then they should evolve more slowly due to their age.

Here, we readdress the question of whether disease genes are under stronger purifying selection than non-disease genes by analyzing rates of protein evolution and the strength of purifying selection of disease genes in the context of gene age. We confirmed that Mendelian disease genes tend to be older than non-disease genes and showed

Key words: human disease genes, evolutionary age of genes, strength of selection, propensity for gene loss.

E-mail: dpetrov@stanford.edu.

Genome. Biol. Evol. 1(1):131–144. 2009
doi:10.1093/gbe/evp013
Advance Access publication May 27, 2009

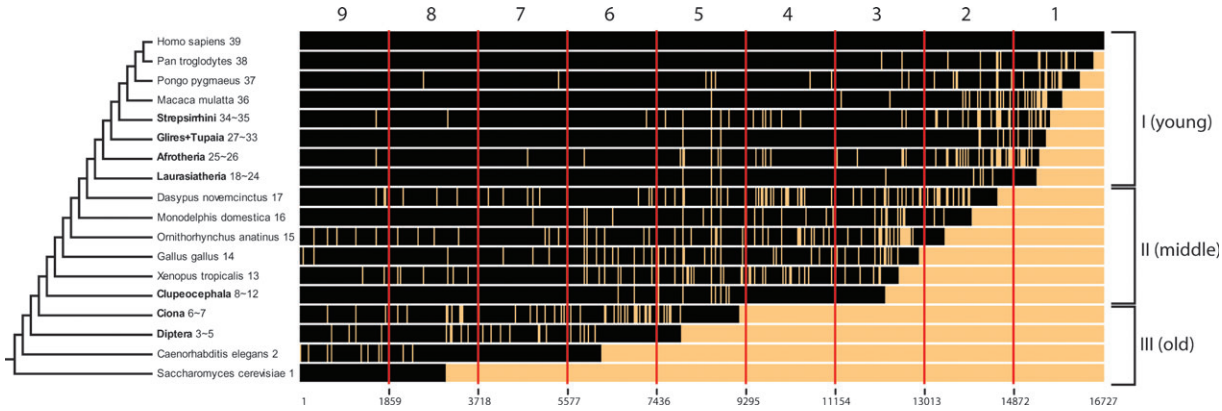


FIG. 1.—Using phylogenetic profile to define the age of genes. The left part illustrated the phylogeny of 18 eukaryotic species (including human) or lineages. The numbers following the species names are the order of 39 species given by PhyloPat (the higher the order, the closer is this species to human). Multiple species, which appeared after their common ancestor separated from the human lineage, were collapsed into one lineage (bolded). The expanded phylogeny of all 39 species is given in supplementary fig. S1 (Supplementary Material online). The right panel illustrates the phylogenetic profiles for 16,727 human genes used in this study. The panel contains 16727 × 18 cells. Each cell indicates the presence (in black) or absence (in yellow) of ortholog of the gene in the species/lineage. Here, for illustrative purpose, genes are sorted by the alphabetic order of their string representations of phylogenetic profile. Vertical red lines split genes into nine equally populated bins.

that complex disease genes tended to be middle aged. The rate of protein evolution (measured as *K*a or *K*a/*K*s) of young disease genes is substantially (1.5- to 3-fold) lower than that of young non-disease genes, whereas the rates of protein evolution of older disease and non-disease gene are indistinguishable. We also investigated gene expression patterns and genetic redundancy (as measured by the sequence identity between a gene and its closest human homolog) between disease genes and non-disease genes. We found that disease genes are expressed more heterogeneously across tissues, but the overall expression level of disease genes is not higher than that of non-disease genes. Disease genes are also less likely to have highly similar paralogs than nondiseases genes. Putting these observations together, we argue that disease genes are under strong purifying selection independently of their age because they need to be sufficiently functionally important for disruptive mutations to show sufficiently severe phenotype diagnosed as disease. At the same time, such genes cannot be ubiquitously expressed because in such cases disruptive mutations would tend to cause embryonic lethality.

Materials and Methods

Gene Sets

Two sets of human Mendelian disease genes were used in this study. First, we obtained a list of genes reported to have disease-causing mutations from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al. 2000). We filtered out genes annotated as "disease" but not as "susceptibility" or as "nondisease" in the OMIM Morbid Map. This set includes 2,011 genes (4006 MIM entries). Second, we considered the data set consisting of 952 manually curated Mendelian disease genes (namely hOMIM) from the study by Blekhman et al. (2008).

To investigate complex disease genes, we obtained 1,656 genes associated with complex diseases from the Genetic Association Database (GAD), an archive of human genetic association studies of complex diseases and disorders (Becker et al. 2004). Each of these genes has been reported at least in one genetic association study. We used the comprehensive collection of 21,528 human protein–coding genes from the Ensembl build 50 (Flicek et al. 2008) as a representative set of all well-characterized human genes. We removed from this set 4,801 pseudogenes documented in pseudogenes.org (Karro et al. 2007), leaving a total of 16,727 genes. The intersection between the all-gene set and the OMIM Morbid set contained 1,637 genes, the intersection between the all-gene set and the hOMIM set contained 803 genes, and the intersection between the all-gene set and the GAD set contained 1,347 genes. Non-disease genes are those in the all-gene sets that are not included in any of the Mendelian and complex disease gene sets.

Age of Genes

We estimated the age of each human gene based on its phylogenetic profile obtained from Phylopat database (Hulsen et al. 2006, 2009). Phylopat algorithm used the ortholog of each protein predicted by Ensembl compara database to construct a phylogenetic profile for each protein based on the presence/absence pattern of its ortholog across other proteomes. Here, we considered human lineage and 17 other lineages containing 38 species (ranging from Chimpanzee to yeast) available in Phylopat database v.50 (fig. 1 and supplementary fig. S1, Supplementary Material online). Note that, given the diverse nature of the fungal kingdom, yeast may not be an ideal representative of fungi, but it is the only representative in Ensembl build 50. Among the 17 lineages, some contained one species (e.g., *Gallus gallus*), whereas others were formed by multiple species (e.g., *Clupeocephala* were formed by *Tetraodon nigroviridis*, *Takifugu rubripes*, *Oryzias latipes*, *Gasterosteus aculeatus*, and *Danio rerio*; see supplementary fig. S1, Supplementary Material online, for detail). A phylogenetic profile can be simply conceived as an array

with 18 characters (one for each lineage in the data set), in which only "0" and "1" characters are allowed. A 0 means no ortholog of the protein is found in the corresponding proteome and a 1 indicates that an ortholog was found in the corresponding proteome. The phylogenetic profiles for all 16,727 genes were represented by a 16727 × 18 matrix of 0 and 1. Figure 1 shows the matrix panel, where 1 is in black and 0 is in yellow.

We adopted the Dollo parsimony (Le Quesne 1974; Farris 1977) to determine the age of a gene. The origin of a gene was determined by retrieving its ortholog back to the species that is most distantly related to human. To do so, we sorted the order of species/lineages by their evolutionary distance to human. Human was at the most left, and yeast, the most distant species, was at the right. Then, the age of a gene can be simply determined by the position of the last 1 in the 18 characters of the phylogenetic profile. For example, the age of the gene with phylogenetic profile '111111110101110000' is ranked as 14, and '111110000000000000' is ranked as 5.

To facilitate data analysis, we added a random variable, $\epsilon \sim \text{norm}(0, 0.001)$, to the age of all genes, making the gene age a continuous variable. The value of ϵ was small such as to not change the original rank of the age of the gene substantially, but by adding an ϵ to its age, each gene obtained a distinct rank. Using different sets of ϵ values (by applying different seeds to initialize the random number generator) did not seem to affect the results.

Next, to analyze the relationships between the age of genes and other parameters, we grouped all genes into nine bins according to their age. We used two different binning methods: To generate "equally populated bins," we adjusted the widths of nine bins so that the same number of genes would fall into each one. To generate "equally spaced bins," we defined nine bins of equal age span. The two binning methods produced qualitatively similar results.

To increase the statistical power of our analysis, we also grouped all genes into three groups, namely, (I) young-, (II) middle-, and (III) old-aged genes and repeated each analysis using these groups. Young genes included those that originated after the lineage of *Laurasiatheria* (including dog and cow), middle-aged genes included those that originated between *Clupeocephala* (including bony fishes) and *Dasypus novemcinctus*, and old genes included those that originated between *S. cerevisiae* and *Ciona* or before (fig. 1).

We also extracted the age of genes defined in the study of Domazet-Loso and Tautz (2008). Their approach for determining the age of genes was based on the same parsimony principle as ours. The difference is that they used BlastP algorithm (*E* value cutoff 0.001) to search human proteins against the National Center for Biotechnology Information non-redundant (NR) database to determine the presence/absence of homologs, whereas we directly adopted the orthologous relationship predicted in Ensembl compara database. Ensembl homologs (orthologs and paralogs) are deduced from the protein trees using the longest transcript of each gene. The detailed description of the prediction method can be found in the reference Vilella et al. (2009). Despite the technical difference, the two age estimations produced qualitatively similar results in all analyses.

In addition to estimating the age of each human gene as described above, we also estimated the tendency of a gene to be lost in evolution. This augments our age estimation, considering not only the deepest node in which the gene was present but also the information captured in the patchiness of the presence/absence patterns. Specifically, we calculated the propensity for gene loss (PGL) measure, introduced by Krylov et al. (2003). PGL is computed based on the pattern of presence/absence of genes across multiple genomes, the phylogenetic tree relating the different species, and the branch lengths. Dollo parsimony is used again to construct ancestral presence/absence states in each internal node of the tree. The PGL value of each gene is then defined as the ratio between the total length of branches in which the gene was lost and the total length of branches in which the gene could have been lost. We also calculated an alternative maximum likelihood-based measure of gene loss, the gene loss rate (GLR), introduced by Borenstein et al. (2007). The results obtained with GLR were qualitatively similar to those obtained with PGL and are not presented here.

Rates of Gene Divergence

*K*a and *K*s for human–chimpanzee orthologous pairs were obtained from BioMart database (Smedley et al. 2009). The values of *K*a and *K*s in BioMart were calculated for coding sequence alignments by using the maximum likelihood method implemented in PAML (Yang 1997). We also obtained the values of *K*a and *K*s for human–macaque orthologous pairs from the study of Blekhman et al. (2008). Major results remained qualitatively unchanged when either human–chimpanzee data or human–macaque data were used. We only reported the results derived from the human–chimpanzee comparison.

Mode of Inheritance and Gene Function

To study the influence of the mode of inheritance on selection, we divided autosomal Mendelian disease genes into genes in which mutations cause recessive disorders and genes in which mutations cause dominant disorders. This division was based on the annotation of the hOMIM data set. Forty genes were found to be both recessive and dominant and therefore excluded from our analysis.

To identify significantly over- or underrepresented gene ontology (GO) terms in a set of disease genes with respect to the set of non-isease genes, we extracted the GO terms for all the genes in our data sets using FatiGO (Al-Shahrour et al. 2004). Adjusted *P* values were calculated using the false-discovery rate (FDR) method of Benjamini and Yekutieli (2001) implemented in FatiGO. We used the adjusted $P < 0.001$ to determine significance.

mRNA Expression Data

mRNA expression data were obtained from Gene Expression Atlas (http://wombat.gnf.org; Su et al. 2004). We included normal adult samples in 54 NR tissue types in the analysis. The expression level of each probe set in a given

tissue was calculated as the mean of log (base 2) signal intensities of all samples after GC-RMA normalization (Wu et al. 2004). When multiple probes were mapped onto the same gene, the probe with the highest expression level was used as the report probe for this gene. The mean expression level of a gene (aveExp) was defined as the mean across all tissues, whereas the peak expression level (maxExp) was defined as the maximum among all tissues. The heterogeneity of expression level across all tissues (hetExp) was calculated according to (Yanai et al. 2005; Liu et al. 2008) as

$$\frac{\sum_{j=1}^{n}\left(1-\frac{\log S_j}{\log S_{\max}}\right)}{n-1},$$

where $n = 54$ is the number of human tissues included in our analysis, S_j is the expression level in each tissue, and $S_{\max}$ is the highest expression level of the probe set across all tissues.

Duplicate Sequence Homology

To understand the role of gene duplicates in robustness against deleterious human mutations, we searched for homologs of all human genes using all-against-all BlastP comparisons, following the study of Hsiao and Vitkup (2008). Sequence homologs were identified as nonself hits with E value ≤ 0.001 that could be aligned over more than 80% of both the query length and the length of identified sequence. For each query sequence, its closest human paralog was identified as the nonself hit which can be aligned over more than 80% of the length of both sequences. Sequence hits with an E value >0.001 were excluded. For human genes with identified paralogs, the distributions of amino acid sequence identities of the closest homologs were recorded.

Results

We investigated two sets of human Mendelian disease genes. First, we used the collection of 1,637 human genes involved in diseases from the OMIM Morbid Map (http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi). Second, we investigated 803 genes from the hOMIM data set—a manually curated collection of Mendelian disease genes, obtained from Blekhman et al. (2008). The hOMIM gene set is less redundant and free of complex phenotypic entries. The two disease gene sets significantly overlap: 781 genes are present in both sets. Because the two data sets generated qualitatively similar results, we only reported here results derived from the hOMIM data set.

For complex disease genes, we investigated 1,347 genes extracted from GAD database (Becker et al. 2004). The majority of genes collected in GAD are associated with complex diseases. In the study of Blekhman et al. (2008), the list of manually curated complex disease genes (supplementary table S5 of Blekhman et al. [2008]) contains 53 genes; only three of them (namely LTA4H, PALB2, and BLMH) were missing from the GAD gene set.

Non-disease genes were defined as genes that do not appear in any of the disease gene sets (including OMIM Morbid, hOMIM, and the complex disease gene sets). The non-disease gene set contained 13,864 genes (82.9% of all genes), indicating that 17.1% of human genes are known to be associated with either Mendelian or complex diseases.

Distribution of Disease Genes in Age Groups

We estimated the age for all 16,727 genes included in our analysis and split them into nine bins according to their ages, where the age group 1 contained the youngest genes and the age group 9 contained the oldest genes. The age was estimated using Dollo parsimony (Le Quesne 1974; Farris 1977) by finding the most highly divergent lineage in which an ortholog (using the Phylopat pipeline; Hulsen et al. 2006, 2009) or a homolog (using BlastP) of a particular human gene could be found (see Materials and Methods for details).

Two binning approaches, *equally populated bins* and *equally spaced bins*, were used (Materials and Methods). Figure 2 illustrates the results obtained for equally populated bins (i.e., having the same number of genes in each of the nine age bins). The bin for age groups 1 contained only 10 Mendelian disease genes (0.54%); this frequency is significantly lower than that of any other age group, which all contained at least 58 disease genes ($P < 0.001$, χ^2 test). Older groups (e.g., group ≥ 3) contained more Mendelian disease genes—3.12–7.10% of them were Mendelian disease genes. This pattern was also observed when the genes were grouped using equally spaced bins—the two binning approaches produced qualitatively similar results.

To simplify the patterns, we pooled all genes into three (including young-, middle-, and old-aged groups) instead of nine groups. The probability to contain DNA variants associated with Mendelian diseases is significantly lower in the young gene group than in the middle-aged and the old gene groups (both $P < 0.001$, χ^2 test) (fig. 2*A*). This pattern is consistent with the finding of Domazet-Loso and Tautz (2008). We further computed the fractions of complex disease genes in different age groups (fig. 2*B*). The frequency of complex disease genes in younger groups (groups 1–3) is also significantly smaller than that in middle- and old-aged groups ($P < 0.001$, χ^2 test); however, unlike Mendelian disease genes, complex disease genes are more likely to be in the middle-aged than in the old-aged groups ($P < 0.001$, χ^2 test) (fig. 2*B*).

We also obtained the age of genes from the study of Domazet-Loso and Tautz (2008). They estimated the age of genes using genes' phylostratum (Domazet-Loso et al. 2007), which focuses on homologs and determines the age of the gene family by strict parsimony assuming that a gene family can be lost but cannot reevolve independently in different lineages or be horizontally transferred. The phylostratum estimate for the age of genes match our estimates of age well (Spearman's $\rho = 0.40$, $P \ll 0.001$). All patterns obtained with phylostratum age estimate are indeed similar to those obtained with our age estimate (data not shown).

In addition to these two age estimates using strict parsimony, the PGL measure is calculated for all genes (see Materials and Method for detail). PGL captures the patchiness of phylogenetic distributions for genes that have the same age. The steady state model of gene gain and loss, assuming that genes lost have the same rate distribution

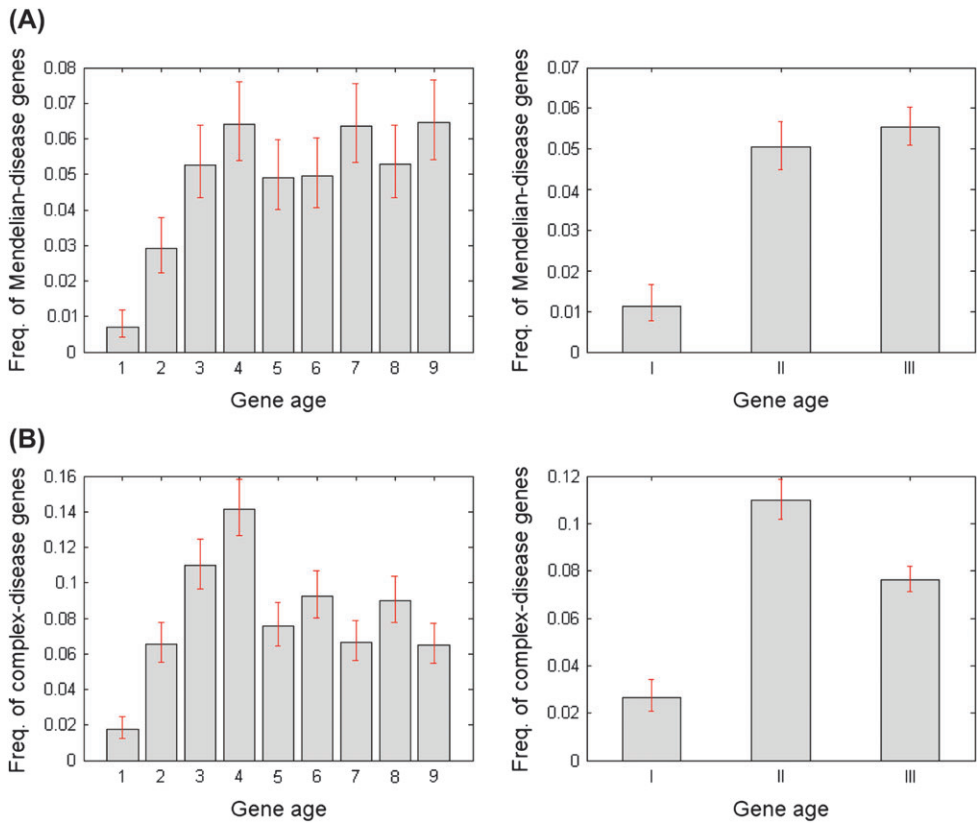


FIG. 2.—Frequencies of Mendelian disease genes (*A*) and complex disease genes (*B*) as functions of their age. Genes are partitioned into nine equally populated bins as well as (I) young-, (II) middle-, and (III) old-aged groups (Materials and Methods). The error bars represent the 95% binomial proportion confidence intervals.

as genes gained, predicts that different gene age classes have specific PGLs (Wolf et al. 2009). Indeed, we found that both our gene age and the phylostratum gene age are significantly correlated with PGL (Spearman's $\rho = -0.55$ and -0.26, respectively, both $P \ll 0.001$). We also examined the relation between the propensity of a gene to be lost (Materials and Methods for details) and the likelihood of the gene to be involved in Mendelian or complex diseases. We split genes into small, medium, and large PGL bins. In these bins (with equal number of genes), 5.7, 5.2, and 3.1% of genes are Mendelian disease genes and 5.5, 8.0, and 3.7 percent of genes are complex disease genes. The pattern resembles the one obtained with young-, middle-, and old-age groups. Next, we test whether disease genes have higher or lower PGL values than non-disease genes. Mendelian disease genes are likely to have a lower PGL values than non-disease genes (median 0.1671 vs. 0.1690 and mean 0.1756 vs. 0.2142, $P = 2.3 \times 10^{-15}$, Mann–Whitney–Wilcoxon [MWW] test). Complex disease genes are also likely to have a lower PGL values than non-disease genes but the difference is less significant (median 0.1671 vs. 0.1690 and mean 0.1910 vs. 0.2142, $P = 1.1 \times 10^{-4}$, MWW test).

Selective Pressure on Disease Genes

Mendelian disease genes have significantly lower median *K*a and *K*a/*K*s than non-disease genes (table 1). The values of *K*a and *K*a/*K*s of non-disease genes decreases when the gene age increases (fig. 3*A* and *B*). Such a negative correlation between the evolutionary rate (*K*a or *K*a/*K*s) and gene age has been well established in previous studies (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004; Alba and Castresana 2005; Wang et al. 2005; Cai et al. 2006; Kuo and Kissinger 2008; Cai and Petrov, unpublished data).

However, such an association was not observed in Mendelian disease genes. *K*a and *K*a/*K*s values for Mendelian disease genes do not decrease with gene age (for *K*a/*K*s, Spearman's $\rho = -0.0104$, $P = 0.783$; table 2). In fact, there was no difference in *K*a or *K*a/*K*s values among age groups for Mendelian disease genes ($P = 0.045$, Kruskal–Wallis [KW] test)(fig. 3*A*). These results suggest that Mendelian disease genes appear to be under strong purifying selection irrespectively of the gene age.

Given that the number of Mendelian disease genes in young age bins is very small, it is possible that the lack of correlation between *K*a or *K*a/*K*s and gene age is due to the small sample size of disease genes. To confirm that this was not the case, we randomly sampled subsets of non-disease genes in each of the nine age bins such that the number of the genes in the subset was equal to the number of Mendelian disease genes in that age bin. We repeated this subsampling process to create 10,000 replicates of non-disease gene sets and computed the Spearman's correlation coefficients between *K*a, *K*s, or *K*a/*K*s and the age of the gene for these subsets. The distribution of the correlation

Table 1
Comparison of Variables between Mendelian, Complex, and Non-disease Genes

	*K*a	*K*s	*K*a/*K*s	aveExp	maxExp	hetExp
Mendelian	0.0034	0.0163	0.237	7.443	10.983	0.256
Complex	0.0036	0.0156	0.260	7.397	11.077	0.250
Non-disease	0.0042	0.0151	0.295	7.643	10.496	0.210
Mendelian versus Non-disease	**0.000**	0.053	**0.000**	0.007	**0.000**	**0.000**
Complex versus Non-disease	0.013	0.172	$\mathbf{5.29 \times 10^{-4}}$	0.011	$\mathbf{1.89 \times 10^{-9}}$	$\mathbf{1.49 \times 10^{-20}}$
Mendelian versus Complex	0.006	0.171	0.048	0.248	0.092	0.005

NOTE.—The median values of variables: *K*a, *K*s, *K*a/*K*s, aveExp, maxExp, and hetExp are given. *P* values of Kolmogorov–Smirnov pairwise tests are given in the three bottom rows. Significant values ($P < 0.001$) appear in bold.

coefficients obtained for these subsets and the observed correlation coefficients for disease genes were plotted in fig. S2. The observed correlation coefficients between Ks values and the age of the gene fall well within the distribution of replicate correlation coefficients (fig. S2*B*). In contrast, the observed correlation coefficients between *K*a (or *K*a/*K*s) and gene age for disease genes fall far from the end of the upper tail of the resampled distributions (fig. S2*A*,*C*) ($P < 10^{-5}$), confirming that the difference reported above between disease and non-disease genes is not merely due to the small sample size.

This difference seems to be mainly driven by the significantly different *K*a (or *K*a/*K*s) values between Mendelian disease genes and non-disease genes in the young genes. In groups 1 to 3, the *K*a and *K*a/*K*s values of Mendelian disease genes are significantly lower than those in non-disease genes (both $P < 0.001$, Kolmogorov–Smirnov [KS] test) (upper panel of fig. 3*A*). Similarly, in group I, the *K*a and *K*a/*K*s values of Mendelian disease genes are almost 3-fold lower than those in non-disease genes (both $P < 0.001$, KS test) (lower panel of fig. 3*A*). In group 4–9 (or groups II and III), we did not observe significant difference in *K*a (or *K*a/*K*s) values between disease and non-disease genes ($P > 0.05$, KS test) (fig. 3*A*).

Unlike Mendelian disease genes, both the *K*a and *K*a/*K*s values of complex disease genes are negatively correlated with the age of genes (Spearman's $\rho = -0.120$ and -0.123, $P < 0.001$) in a pattern similar to that of non-disease genes (Spearman's $\rho = -0.249$ and -0.263, $P < 0.001$) (fig. 3*B*). Repeating the subsampling analysis describe above, we confirmed that the scarcity of complex disease genes in each age bin was not the reason that complex disease genes resembled non-disease genes in these patterns (fig. S3). Finally, we found significant differences in both *K*a and *K*a/*K*s values between different age groups for complex disease genes (both $P < 0.001$, KW test).

Although, as a function of gene age, the changes of *K*a and *K*s/*K*s for complex disease genes are similar to those for non-disease genes, values of *K*a and *K*a/*K*s of young complex disease genes are still significantly lower than those of young non-disease genes. For genes in groups 1–3, the *K*a and *K*a/*K*s values of complex disease genes are 1.4- and 1.5-fold lower than those of non-disease genes, respectively (both $P < 0.001$, KS test) (upper panel of fig. 3*B*). In group I, the *K*a and *K*a/*K*s values of complex disease genes are 1.5- and 1.2-fold lower than those of non-disease genes, respectively; however, the differences are less significant ($P = 0.0046$ and 0.0485, respectively, KS test) (lower panel of fig. 3*B*), underscoring the relatively weaker purifying selection acting on complex disease genes compared with Mendelian disease genes.

We obtained highly consistent results with the PGL as a complementary measure of gene evolutionary age. For non-disease genes, values of PGL are positively correlated with values of *K*a and *K*a/*K*s (Spearman's $\rho = 0.155$ and 0.167, respectively, $P \ll 0.001$ in both cases) but not correlated with values of *K*s (Spearman's $\rho = 0.021$, $P = 0.026$). This result is consistent with those from previous studies (Krylov et al. 2003; Wolf et al. 2006; Borenstein et al. 2007). In contrast, for Mendelian disease genes, PGL does not correlate with any of divergence rate measures ($P > 0.001$, Spearman correlation between PGL and *K*a, *K*s, or *K*a/*K*s). For complex disease genes, PGL are marginally significantly positively correlated with *K*a and *K*a/*K*s (Spearman's $\rho = 0.114$ and 0.139, $P = 0.002$ and 1.28×10^{-4}, for *K*a and *K*a/*K*s, respectively) but not correlated with *K*s ($P > 0.001$). These results suggest that rapidly evolved genes have a higher propensity to be lost, but the pattern is only upheld for non-disease genes. The trend is less significant in complex disease genes and completely disappears in Mendelian disease genes.

We used an additional measure of selective pressure based on polymorphism data to confirm the results derived from *K*a/*K*s. The measure is the ratio of nonsynonymous-to-synonymous polymorphisms (*P*n/*P*s). Recent accumulation of human genome–wide single nucleotide polymorphism (SNP) data enables the derivation of *P*n/*P*s (International HapMap Consortium 2003, 2007; Bustamante et al. 2005). We found that both Mendelian and complex disease genes have lower values of *P*n/*P*s computed from two SNP data sets—HapMap SNPs (International HapMap Consortium 2003, 2007) and Applera SNPs (Bustamante et al. 2005; data not shown). This is an additional line of evidence of strong purifying selection in disease genes (see also Liu et al. 2008). With either divergence or polymorphism information, we find that disease genes tend to be under stronger purifying selection than non-disease genes but only in the young gene categories.

Effects of Inheritance Mode and Gene Function

We divided Mendelian disease genes into dominant disease genes (238 hOMIM genes that are known to have dominant diseases-causing mutations) and recessive disease genes (389 genes that are known to have recessive diseases-causing mutations) as annotated by Blekhman et al.

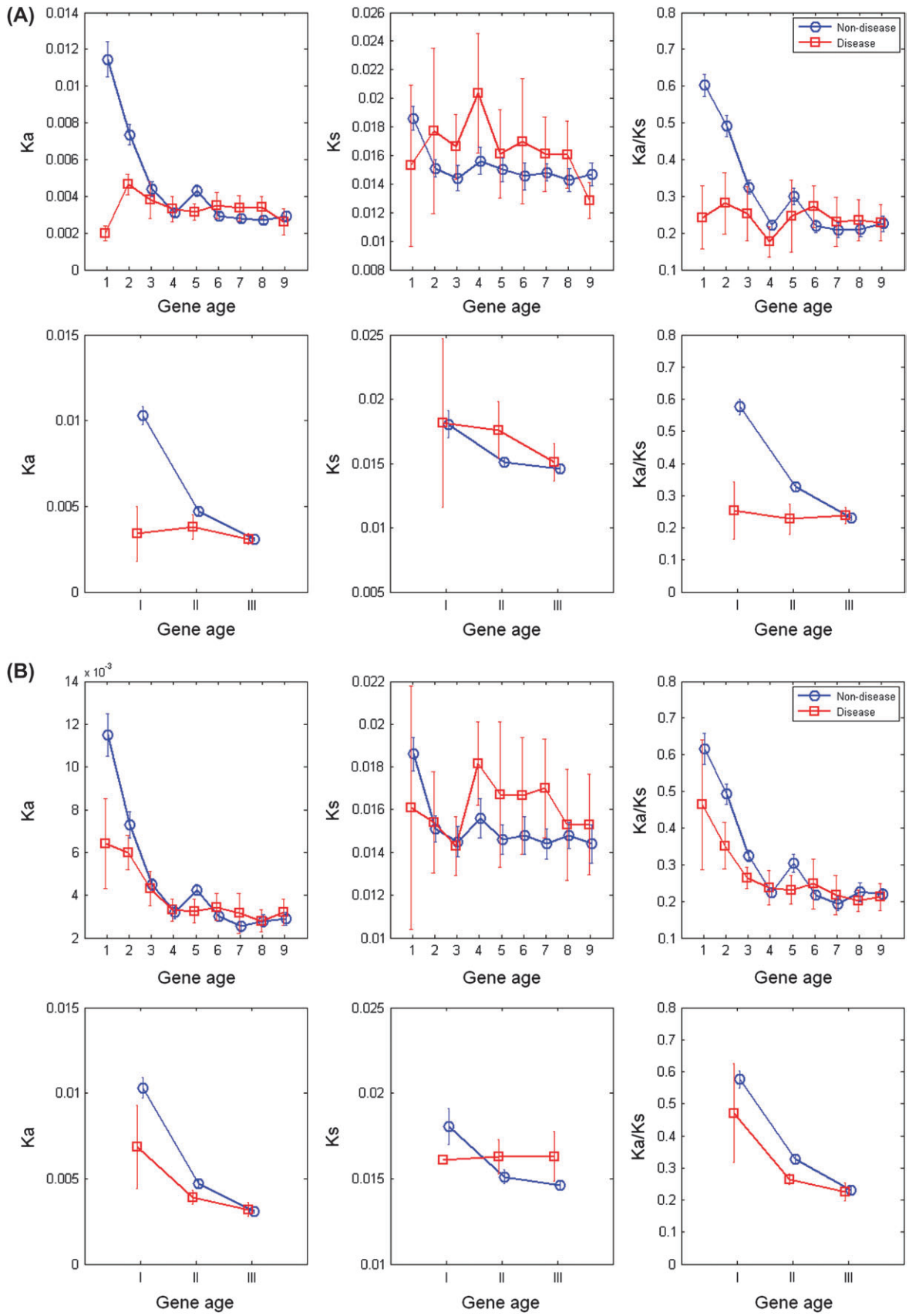


FIG. 3.—*K*a, *K*s, and *K*a/*K*s as functions of the age of genes. Mendelian disease genes (*A*) and complex disease genes (*B*) are partitioned into one–nine equally populated bins as well as (I) young-, (II) middle-, and (III) old-aged groups. Median values and 95% confidence intervals are given for disease genes (red square) and non-disease genes (blue circle).

Table 2
Correlations between Various Variables and the Gene Age

	*K*a	*K*s	*K*a/*K*s	aveExp	maxExp	hetExp
Mendelian	−0.0329 (0.376)	−0.0296 (0.433)	−0.0104 (0.783)	**0.157 (3.81×10^{-5})**	0.0300 (0.432)	−0.0495 (0.195)
Complex	**−0.120 (0.000)**	−0.010 (0.585)	**−0.123 (0.000)**	**0.127 (0.000)**	0.018 (0.346)	−0.057 (0.003)
Non-disease	**−0.249 (0.000)**	**−0.043 (4.00×10^{-6})**	**−0.263 (0.000)**	**0.179 (0.000)**	**0.114 (0.000)**	0.0152 (0.221)

NOTE.—Spearman's ρ and *P* value (in parentheses) are given. Significant values ($P < 0.001$) appear in bold.

(2008). Dominant genes have significantly lower values of *K*a/*K*s than those of recessive genes (median *K*a/*K*s 0.216 vs. 0.242; $P = 3.673 \times 10^{-4}$, KS test). This result is consistent with the results reported in two previous studies (Furney et al. 2006; Blekhman et al. 2008). Neither dominant nor recessive genes show any correlation between *K*a/*K*s and gene age (fig. S4). Collectively, dominant disease genes are younger than recessive disease genes (fig. S5).

We also examined whether the strong purifying selection acting on young Mendelian disease genes was due to the enrichment of particular biological functions in these genes (Materials and Methods). Compared with the non-disease genes in the same age group, young Mendelian disease genes were significantly enriched with anatomical structure development (GO:0048856, adjusted $P = 19 \times 10^{-5}$ and 8.31×10^{-26} for equally spaced bins and equally populated bins, respectively) and multicellular organismal development (GO:0007275, adjusted $P = 8.18 \times 10^{-5}$ and 1.57×10^{-20} for equally spaced bins and equally populated bins, respectively) genes. In addition to these two terms, some GO terms were identified to be significant only when we used equally populated bins. These terms include circulation (GO:0008015), response to stress (GO:0006950), cellular component organization and biogenesis (GO:0016043), response to external stimulus (GO:0009605), coagulation (GO:0050817), cellular developmental process (GO:0048869), as well as other terms. The complete list of enriched terms can be found in supplementary table S1. Among all these GO terms, only one term, nucleic acid binding (GO:0003676), was enriched in non-disease genes.

Effects of Gene Expression

Next, we studied the expression patterns of disease and non-disease genes in relation to gene age. We calculated the average (aveExp), maximum (maxExp), and heterogeneity (hetExp) of gene expression across 54 normal tissues for each human genes (fig. 4*A*). Mendelian disease genes show significantly higher hetExp ($P = 0.007$) and maxExp ($P < 0.001$) values than non-disease genes, whereas their aveExp ($P = 0.699$) values are similar (KS test) (table 1). This result is consistent with the hypothesis that tissue-specific genes are more likely to be involved in human disease than widely expressed genes (Winter et al. 2004; Adie et al. 2005).

Furthermore, Mendelian diseases genes show similar maxExp values across different age groups ($P = 0.699$, KW test), whereas maxExp for non-disease genes is positively correlated with the age of genes (Spearman's $\rho = 0.114$, $P < 0.001$; KW test, $P < 0.001$)(table 2). Non-disease genes in different age groups have different hetExp values ($P = 0.000443$, KW test), but hetExp values for Mendelian disease genes of different age groups show no variation ($P = 0.191$, KW test). There is no correlation between hetExp and gene age for both Mendelian and non-disease genes ($P = 0.195$ and 0.221, respectively, Spearman test) (table 2 and fig. 4*A*).

Similar to Mendelian disease genes, complex disease genes show significantly higher maxExp ($P = 1.89 \times 10^{-9}$, KS test) and hetExp ($P = 1.49 \times 10^{-20}$, KS test) values and similar aveExp values to non-disease genes (table 1). Moreover, complex disease genes show the same patterns of expression variables versus gene age as Mendelian disease genes, that is, there is a positive correlation between aveExp and gene age and there is no significant correlation between either maxExp or hetExp and gene age (table 2 and fig. 4*B*).

We conducted a survey of the tissue-specific expression patterns of disease versus non-disease genes. Distribution of genes showing peak expression in 54 tissues and portions of Mendelian and complex disease genes in all genes showing peak expression in the corresponding tissues are given in supplementary figure S6 (Supplementary Material online). We found that Mendelian disease genes are more likely to be most highly expressed in liver and kidney ($P \ll 0.001$ in both cases, Fisher's exact tests with Bonferroni correction) but less likely in testis ($P = 6 \times 10^{-6}$). Complex disease genes are more likely to be most highly expressed in liver ($P = 0.0004$).

In addition, disease genes and non-disease genes show no substantial difference in the correlation between *K*a/*K*s and gene expression, even after these genes were assigned into young-, middle-aged, and old groups (fig. S7–8).

Effects of Presence of Close Duplicates

It has been hypothesized (Lopez-Bigas and Ouzounis 2004) that proteins with similar paralogs should be less often involved in diseases because the compromised function of such proteins when mutated could be compensated for by their functional paralogs (Frenette et al. 1996; Wagner 2000; Gu 2003; Kamath et al. 2003; Dean et al. 2008; Wagner 2008). Here, we test this hypothesis using our gene sets. We used two definitions for "singleton human genes." The first considers the genes that do not have any sequence homologs, which can be identified by BlastP searches (see Materials and Methods for criteria used to define homologs). The second considers those that are not included in any Ensembl protein family (Enright et al. 2002). Using either of these definitions, Mendelian disease genes were not found more likely to be singleton human genes than non-disease genes. This result is consistent with that of Yue and Moult (2006).

Table 3
Candidate Disease Genes (*Ka*/*Ks* ≤0.30, maxExp ≥11.75, hetExp ≥0.32, and duplicate sequence similarity ≤50%)

Ensembl ID	HGNC Name	Description
ENSG00000179776	CDH5	Cadherin 5, type 2 (vascular endothelium)
ENSG00000154734	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif
ENSG00000099308	MAST3	Microtubule-associated serine/threonine kinase 3
ENSG00000172232	AZU1	Azurocidin 1
ENSG00000124006	OBSL1	Obscurin-like 1
ENSG00000039560	RAI14	Retinoic acid-induced 14
ENSG00000145555	MYO10	Myosin X
ENSG00000169347	GP2	Glycoprotein 2 (zymogen granule membrane)
ENSG00000176956	LY6H	Lymphocyte antigen 6 complex, locus H
ENSG00000169509	CRCT1	Cysteine-rich C-terminal 1

NOTE.—Those genes are not included in the list of hOMIM disease genes (Blekhman et al. 2008).

We next resorted to a different approach for testing the role and magnitude of duplicate gene contribution to robustness against deleterious human mutations. We used sequence similarity between paralogs or homologs to quantify the likelihood and magnitude of functional compensation, following Hsiao and Vitkup (2008). For nonsingleton human genes (i.e., those with identified paralogs), the distributions of amino acid sequence identities of the closest homologs are significantly different between disease and non-disease genes. The average identity of the closest homolog is 47.9% for Mendelian disease genes, 48.2% for complex disease genes, and 52.3% for non-disease genes (Mendelian vs. nondisease, $P < 0.001$; complex vs. nondisease, $P < 0.001$; Mendelian vs. complex, $P = 0.00132$, KS test). This difference between disease genes and nondisease genes seems more substantial and statistically significant for middle-age genes (fig. 5). The lack of statistical significance for young genes may be attributed to the small number of genes.

Discussion

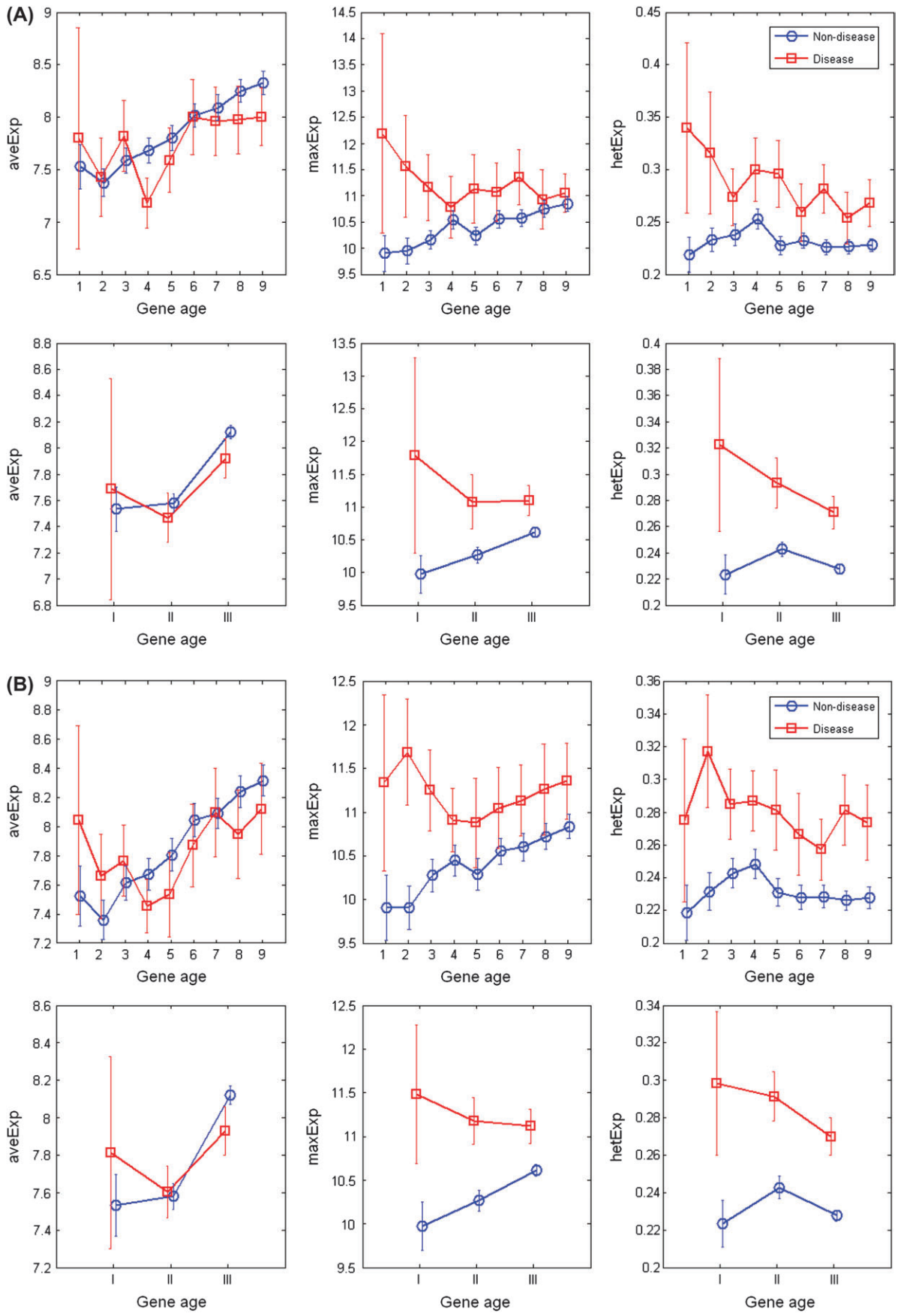
New genes can be created by many mechanisms, including exon shuffling, gene duplication, retroposition, integration of mobile elements, lateral gene transfer, gene fusion/fission, as well as de novo origination (for review, see Long et al. 2003). It is believed that we can detect only a small fraction of all the events of the formation of novel genes. What we can identify are those recent enough to be recognizable, yet old enough to be fixed or present at a high enough frequency in the population to be found in sequenced genomes (Babushok et al. 2007). However, we can use sequence similarity searches to estimate the time (in the course of evolution) when an extant gene or a gene family has appeared in the genomic sequence. Sequence similarity searches appear to be able to detect gene homologs in distantly related lineages even in cases of fast evolving genes because almost all protein-coding genes contain at least pockets of high amino acid conservation (Alba and Castresana 2007) (but see Elhaik et al. 2006).

A large portion (22%) of human genes can be detected in the yeast genome, implying that they originated before the common ancestor of human and yeast, which have diverged for more than 1.5 billion years. Other human genes can be detected only within mammals or even only within primates (Toll-Riera et al. 2009). The time of origination (the age) appears to be an important parameter in the study of molecular evolution. For instance, gene age is negatively correlated with the rate of gene evolution at the protein level. Alba and Castresana (2005) found such a negative correlation for human genes. Cai et al. (2006) found that the lineage-specific (younger) genes evolve at faster rates than widely distributed (older) genes in fungi. A similar pattern was observed in rodents (Wang et al. 2005), *Drosophila* (Domazet-Loso and Tautz 2003), parasitic protozoa (Kuo and Kissinger 2008), and bacteria (Daubin and Ochman 2004). In another study, we demonstrated that younger genes evolve rapidly primarily because they are subject to relaxed purifying selection (Cai and Petrov, unpublished).

The analysis of gene ages, however, has not been applied in the study of the evolution of disease genes, even though evolutionary age of disease genes has been investigated. For example, almost all human disease genes can be found in zebra fish genome (Hariharan and Haber 2003). Similarly, 60–80% of human disease genes can be found in the *Drosophila* genome (Fortini et al. 2000; Rubin et al. 2000; Reiter et al. 2001). Human disease genes are highly represented among human–rodent ortholog sets (Huang et al. 2004). Domazet-Loso and Tautz (2008) found that disease genes are notably absent from the younger phylostrata (i.e., age groups)—only about 0.6% of the disease genes mapped to the age since the origin of Eutheria or later—and that there was a significant negative correlation between the number and frequency of disease genes and a ranked evolutionary age. The results obtained in our study for Mendelian disease genes confirm their findings.

We confirm that complex disease genes are also underrepresented in young-aged groups. More interestingly, complex disease genes are overrepresented in middle-aged groups—a new finding that may have an important biological implication because the middle-aged groups contain more vertebrate-specific genes than other age groups. Most complex disease genes are those that originated during the emergence of vertebrates. The complicated interactions between functionally associated genes responsible for human complex diseases can, therefore, be traced back to their not-so-deep evolutionary past.

We also consider the PGL for genes in different age groups. For Mendelian disease genes, the portions of disease genes increase with the gene age and decrease with the


(A)
aveExp
maxExp
hetExp
Gene age
Non-disease
Disease
(B)
aveExp
maxExp
hetExp
Gene age
Non-disease
Disease

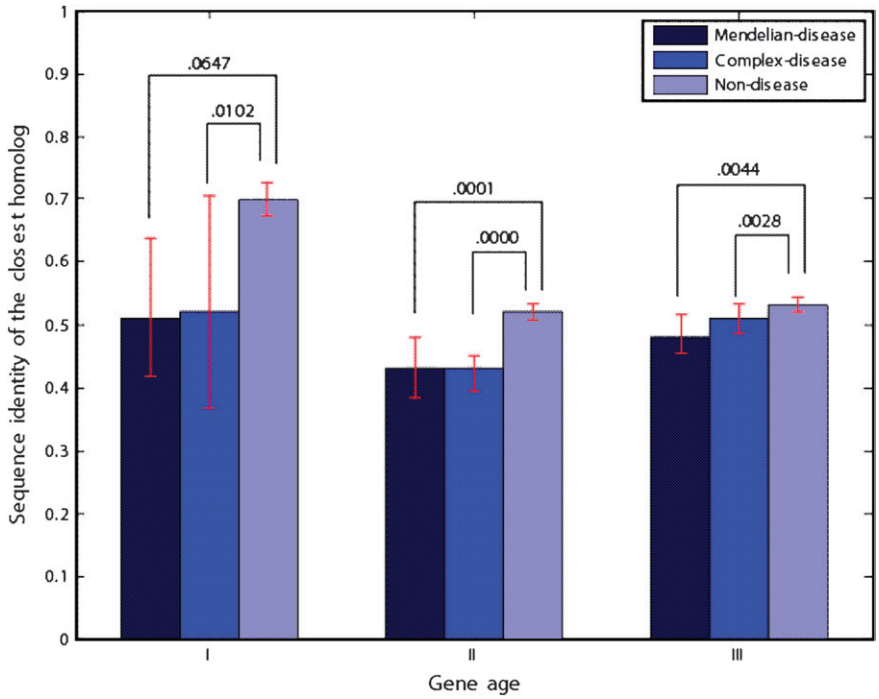


FIG. 5.—Sequence identify of the closest homolog of genes. Mendelian, complex, and non-disease genes are partitioned into (I) young-, (II) middle-, and (III) old-aged groups. Median values and 95% confidence intervals are plotted. *P* values of KS tests between groups are given.

values of PGL; for complex disease genes, the portions in the middle-PGL classes and the middle-age group are the highest. The results fit the prediction of the steady state model of gene gain and loss during genome evolution (Wolf et al. 2009), claiming that genes of different age classes (genes gained at different time during the evolution of a lineage) substantially differ in the characteristics that are correlated with the PGL.

Rate of Evolution and Gene Expression of Disease Genes as a Function of Gene Age

In this study, we analyzed human disease genes in the context of gene age and discovered several new patterns. The key pattern is that Mendelian disease genes evolve slowly regardless of their age. In contrast, non-disease genes show a strong relationship between rate of evolution and gene age, with younger genes evolving much faster than older genes. As a result, young Mendelian disease proteins evolve almost 3-fold slower than young non-disease genes. The difference is less dramatic for the young complex disease genes, but it is still highly significant. On the other hand, the older Mendelian and complex disease genes evolve at indistinguishably similar and low rates as the older non-disease genes.

The level of gene expression is an essential factor in determining the selective pressure on genes (Pal et al. 2006). It is well known that highly expressed genes tend to be under stronger purifying selection (Pal et al. 2001; Subramanian and Kumar 2004; Drummond et al. 2005; Wall et al. 2005). Slow rate of evolution of young disease genes might be due to their high levels of expression across a large array of tissues. Our observations show that this is not the case—young disease genes do not have higher median levels of aveExp than young non-disease genes ($P = 0.2981$ for Mendelian disease genes and $P = 0.1559$ for complex disease genes, KS test). However, both Mendelian and complex disease genes have significantly higher levels of tissue specificity (measured as the heterogeneity of gene expression across tissues) and significantly higher levels of peak expression across tissues than those of non-disease genes (table 1). Consistent with these findings, a recent study showed that the more experiments in which a gene was differentially expressed, the more likely it is to contain disease-associated variants (Chen et al. 2008).

Evolutionary Properties of Disease Genes

These results suggest that disease genes are a subset of genes that perform critical NR functions in some but not all tissues. Because the function of such genes is important, any disruptive mutation in these genes can lead to severe and detectable disease phenotype and would not be tolerated by purifying selection. At the same time, disruptive mutations in functionally important but widely expressed

←

FIG. 4.—Mean expression level (aveExp), expression heterogeneity (hetExp), and peak expression level (maxExp) as functions of the age of genes. Mendelian disease genes (*A*) and complex disease genes (*b*) are partitioned into one to nine equally populated bins as well as (I) young-, (II) middle-, and (III) old-aged groups. Median values and 95% confidence intervals are given for disease genes (red square) and non-disease genes (blue circle).

genes would tend to lead to embryonic lethality instead of disease. One example that supports our prediction is the susceptibility loci for Leigh syndrome. The expression levels of candidate genes for this syndrome tend to be elevated in the primary tissues or cells involved in disease (Mootha et al. 2003).

Gene duplication is known to provide genetic robustness (Frenette et al. 1996; Wagner 2000, 2008; Gu 2003; Kamath et al. 2003; Dean et al. 2008). The above reasoning suggests that disease genes should not have very close duplicates or at least no close duplicates expressed in the same tissues. However, this relationship might be more complicated because such genetic robustness might be limited to some but not other tissues or times of development and thus could allow for the expression of disruptive mutations as disease states. Both Mendelian disease genes and complex disease genes indeed tend to have more divergent homologs than non-disease genes.

The fact that very few young genes are disease genes might suggest that very few young genes perform sufficiently important functions to be disease genes. The fact that young non-disease genes evolve at a much higher rate and are subject to much weaker purifying selection supports this possibility. The sharp differences between young disease and non-disease genes allow us to make predictions about which young genes can harbor disease mutations even if they had not been identified as disease genes yet. Specifically, we chose genes (table 3) that satisfied the following criteria: $Ka/Ks \leq 0.30$, maxExp ≥ 11.75, hetExp ≥ 0.32, and duplicate sequence similarity $\leq 50\%$. This information might prove helpful in mapping of disease-causing mutations.

The identification of candidate genes within loci associated with human genetic diseases is a difficult task because the identified genomic region typically contains hundreds of genes, making experimental methods employed to identify the specific disease gene arduous and expensive. Gene prioritization is therefore critical for modern genetic medicine, and many approaches have been developed to predict disease genes, based on in-depth knowledge of phenotypic similarity (Freudenberg and Propping 2002), coexpression, genomic data fusion and protein interaction (George et al. 2006; Kohler et al. 2008), and literature-based discovery (Hristovski et al. 2005). Integrating information concerning the time of origin of genes can serve as an important tool to further improve the accuracy of gene prioritization.

Young humans genes or human genes that have a high propensity for loss in other lineages but that have been evolving under strong constrain between humans and chimps might be of particular interest in general. Such genes need to have acquired an important function that makes them evolve slower than would be predicted given their age or propensity for loss. Thus, they might be enriched for those genes that encode primate- or even human-specific functions.

Funding

This was also supported by the National Institutes of Health (NIH) [grant GM077368 to D.A.P.]; the Morrison Institute for Population and Resource Studies [to E.B.]; a grant to the Santa Fe Institute from the James S. McDonnell Foundation 21st Century Collaborative Award Studying Complex Systems; and NIH [grant GM28016].

Supplementary Material

Supplementary figures S1–S9 and table S1 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank anonymous reviewers for valuable comments. We thank Wei Yu (CDC/CCHP/NOPHG) for valuable comments on data sets of complex disease gene, Abdellali Kelil for helping in clustering gene families, and Giulietta Spudich (Ensembl) for technical support. We also thank all members of the Petrov Lab and especially Philip Bulterys for helpful comments. J.J.C. thanks the Cheung Kong group for the Endeavour Cheung Kong Research Fellowship. The URLs for data presented herein are as follows:OMIM Morbid Map, http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi.

Literature Cited

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. 2005. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics. 6:55.

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 22:598–606.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 7:53.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 20:578–580.

Babushok DV, Ostertag EM, Kazazian HH Jr. 2007. Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci. 64:542–554.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet. 36:431–432.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 29:1165–1188.

Blekhman R, et al. 2008. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 18:883–889.

Borenstein E, Shlomi T, Ruppin E, Sharan R. 2007. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. Nucleic Acids Res. 35:e7.

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet. 33(Suppl):228–237.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153–1157.

Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. J Mol Evol. 63:1–11.

Chen R, et al. 2008. FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. Genome Biol. 9:R170.

Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res. 14:1036–1042.
Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet. 4:e1000113.
Dean M. 2003. Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. Hum Mutat. 22:261–274.
Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.
Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.
Domazet-Loso T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 25:2699–2707.
Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. 102:14338–14343.
Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23:1–3.
Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.
Farris JS. 1977. Phylogenetic analysis under Dollo's Law. Syst Zool. 26:77–88.
Flicek P, et al. 2008. Ensembl 2008. Nucleic Acids Res. 36:D707–D714.
Fortini ME, Skupski MP, Boguski MS, Hariharan IK. 2000. A survey of human disease gene counterparts in the *Drosophila* genome. J Cell Biol. 150:F23–F30.
Frenette PS, Mayadas TN, Rayburn H, Hynes RO, Wagner DD. 1996. Susceptibility to infection and altered hematopoiesis in mice deficient in both P- and E-selectins. Cell. 84: 563–574.
Freudenberg J, Propping P. 2002. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics. 18(Suppl 2):S110–S115.
Furney SJ, Alba MM, Lopez-Bigas N. 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics. 7:165.
George RA, et al. 2006. Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res. 34:e130.
Giallourakis C, Henson C, Reich M, Xie X, Mootha VK. 2005. Disease gene discovery through integrative genomics. Annu Rev Genomics Hum Genet. 6:381–406.
Gu X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. 19:354–356.
Haines JL, Pericak-Vance MA. 1998. Approaches to gene mapping in complex human diseases. New York: Wiley-Liss. p. xxii, 434.
Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian Inheritance in Man (OMIM). Hum Mutat. 15:57–61.
Hariharan IK, Haber DA. 2003. Yeast, flies, worms, and fish in the study of human disease. N Engl J Med. 348:2457–2463.
Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. 2005. Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 74:289–298.
Hsiao TL, Vitkup D. 2008. Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet. 4:e1000014.
Huang H, et al. 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol. 5:R47.
Hulsen T, de Vlieg J, Groenen PM. 2006. PhyloPat: phylogenetic pattern analysis of eukaryotic genes. BMC Bioinformatics. 7:398.
Hulsen T, Groenen PM, de Vlieg J, Alkema W. 2009. PhyloPat: an updated version of the phylogenetic pattern database contains gene neighborhood. Nucleic Acids Res. 37:D731–D737.
International HapMap Consortium. 2003. The International HapMap Project. Nature. 426:789–796.
International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449:851–861.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431:931–945.
Kamath RS, et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature. 421:231–237.
Karro JE, et al. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35:D55–D60.
Kohler S, Bauer S, Horn D, Robinson PN. 2008. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 82:949–958.
Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2004. Bioinformatical assay of human gene morbidity. Nucleic Acids Res. 32:1731–1737.
Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.
Kuo CH, Kissinger JC. 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. BMC Evol Biol. 8:108.
Le Quesne WJ. 1974. The uniquely evolved character concept and its cladistic application. Syst Zool. 23:513–517.
Liu J, Zhang Y, Lei X, Zhang Z. 2008. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 9:R69.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.
Lopez-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32:3108–3114.
Mootha VK, et al. 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci USA. 100:605–610.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Reiter LT, Potocki L, Chien S, Gribskov M, Bier E. 2001. A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. Genome Res. 11:1114–1125.
Rubin GM, et al. 2000. Comparative genomics of the eukaryotes. Science. 287:2204–2215.
Smedley D, et al. 2009. BioMart–biological queries made easy. BMC Genomics. 10:22.
Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. Gene. 318:169–175.
Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 101:6062–6067.
Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373–381.

Thomas PD, 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.
Toll-Riera M, et al. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 26:603–612.
Vilella AJ, et al. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.
Wagner A. 2000. Robustness against mutations in genetic networks of yeast. Nat Genet. 24:355–361.
Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. Bioessays. 30:367–373.
Wall DP, et al. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA. 102:5483–5488.
Wang W, et al. 2005. Origin and evolution of new exons in rodents. Genome Res. 15:1258–1264.
Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res. 14:54–61.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci. 273:1507–1515.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci USA. 106:7273–7280.
Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. 2004. A model based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc. 99:909–917.
Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 21:650–659.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yue P, Moult J. 2006. Identification and analysis of deleterious human SNPs. J Mol Biol. 356:1263–1274.

Eugene Koonin, Associate Editor

Accepted May 22, 2009