An Ancient Evolutionary Origin of Genes Associated with Human Genetic Diseases

Tomislav Domazet-Lošo† and Diethard Tautz**

*Max-Planck Institut für Evolutionsbiologie, August-Thienemannstrasse 2, Plön, Germany; and †Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, Zagreb, Croatia

Several thousand genes in the human genome have been linked to a heritable genetic disease. The majority of these appear to be nonessential genes (i.e., are not embryonically lethal when inactivated), and one could therefore speculate that they are late additions in the evolutionary lineage toward humans. Contrary to this expectation, we find that they are in fact significantly overrepresented among the genes that have emerged during the early evolution of the metazoa. Using a phylostratigraphic approach, we have studied the evolutionary emergence of such genes at 19 phylogenetic levels. The majority of disease genes was already present in the eukaryotic ancestor, and the second largest number has arisen around the time of evolution of multicellularity. Conversely, genes specific to the mammalian lineage are highly underrepresented. Hence, genes involved in genetic diseases are not simply a random subset of all genes in the genome but are biased toward ancient genes.

Introduction

Most genes involved in basic cellular processes have already evolved in the unicellular ancestor of eukaryotes. Other large groups of genes, most notably those involved in signaling processes, can be traced back to multicellular metazoan evolution (King et al. 2008). At the other end of the spectrum are the orphan genes, which are restricted to particular taxonomic groups. Although it seems evident that such genes should be associated with lineage-specific adaptations, they are still not much studied and a general role could not yet be ascribed to them. We have previously shown that many orphan genes show high substitution rates, although some of them may also evolve very slowly, suggesting major lineage-specific functions (Domazet-Lošo and Tautz 2003).

To quantify and statistically analyze gene emergence at different levels of the taxonomic hierarchy, we have developed a phylostratigraphic approach (Domazet-Lošo et al. 2007). In this method, one uses a set of more or less completely sequenced genomes to represent the full tree of eukaryotic life, accentuated by key innovations reflected at phylogenetic levels, called phylostrata. This framework is used to place the genes of a given genome into those phylostrata where homologues can be found by Blast analysis. We have used this approach here to study the evolutionary origin of genes that were implicated in causing human genetic diseases.

The Online Mendelian Inheritance in Man, OMIM (TM) (2008) database lists currently more than 4,000 chromosome regions that have been associated with a genetic disease and for about half of them, the disease-causing mutation has been identified. Hence, a significant part of the human genome is susceptible to mutations that cause diseases. These genes have been further classified into essential and nonessential genes based on the comparison with data from the mouse (Goh et al. 2007; Feldman et al. 2008). A knockout of an essential gene leads to embryonic lethality, whereas nonessential genes can show a range of phenotypes when inactivated. Although there is an ongoing discussion on how predictive the comparisons with the mouse are for the function of the genes in humans (Liao and Zhang 2008), it is nonetheless clear that this general categorization can be applied to the majority of genes. Based on the analysis of Goh et al. (2007) and Liao and Zhang (2008), one can estimate that between 60% and 75% of disease genes are nonessential genes.

It was shown that nonessential disease genes are less likely to represent hubs in gene interaction networks, that they have fewer interacting partners, and that their expression patterns are less likely to be correlated with other genes (Goh et al. 2007; Feldman et al. 2008). Furthermore, it has been shown that genes at the periphery of network hubs are more often subject to positive selection (Kim et al. 2007), suggesting that they are particularly important for creating evolutionary novelties. In evolutionary terms, one could therefore place them in the category of late additions in the lineage toward humans. On the other hand, already the first genome comparisons showed that many disease genes can be found in species distantly related to humans (Rubin et al. 2000). Thus, one can ask whether disease genes are simply a random subset of all genes or whether they fall into a particular evolutionary age class that might be correlated to evolutionary innovations in the human lineage.

Methods

Phylostratigraphic Analysis

Phylostratigraphic analysis was done according to the procedure described in Domazet-Lošo et al. (2007). Human protein sequences were retrieved from Ensembl (version 45), and only the longest splicing variants were kept (22,937 unique proteins). The BlastP algorithm (0.001 *E* value cutoff) was used to compare human proteins against the National Center for Biotechnology Information (NCBI) non-redundant (NR) database. Before the sequence similarity search was done, the NR database was cleaned up with respect to sequences of uncertain taxonomic status or where the taxonomy ID is not included in the cellular organisms section of the NCBI taxonomy database. Additionally, we removed from the database sequences of metazoan taxa with unreliable phylogenetic position (Mesozoa, Myxozoa, Chaetognatha, and Placozoa). After this clean up

Key words: phylostratigraphy, orphan genes.

E-mail: tautz@evolbio.mpg.de.

Mol. Biol. Evol. 25(12):2699–2707. 2008
doi:10.1093/molbev/msn214
Advance Access publication September 26, 2008

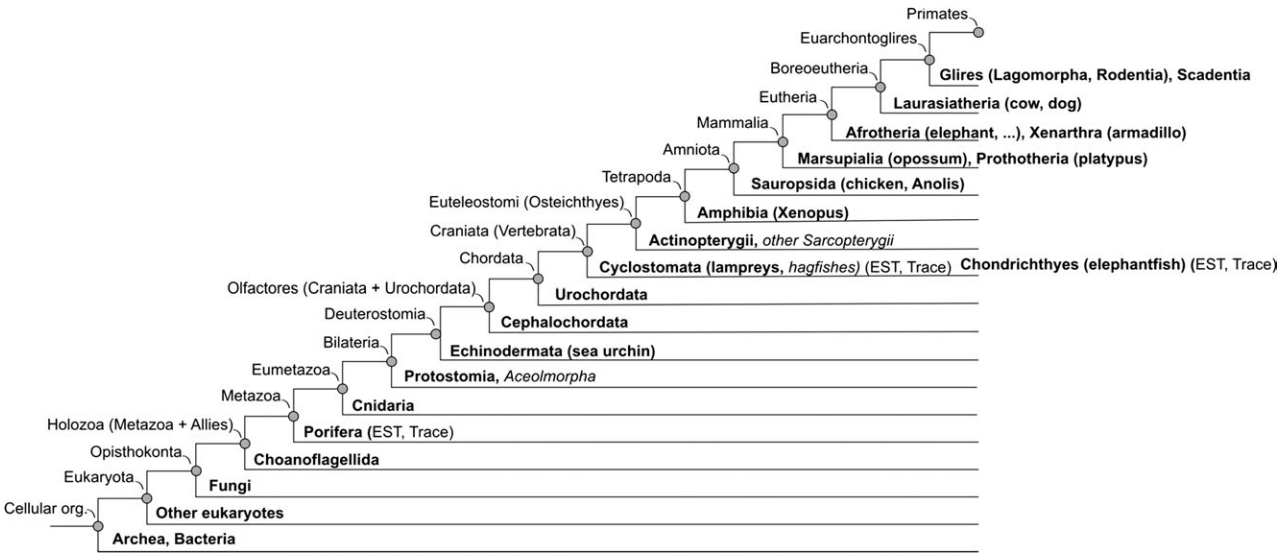

FIG. 1.—Phylogenetic framework used in the search for the human gene origins. Taxa represented in the databases with complete genomes or a substantial amount of TRACE and EST data are in bold. Taxa in italics are represented in the databases only with small numbers of highly conserved genes, and their exclusion from the analysis does not influence the results.

procedure, we filled up the NR database with complete genomes that were absent in the database but otherwise were publicly available (fig. 1). Additionally, TBlastN search (10^{-15} *E* value cutoff) was done against substantial trace and EST archives of Porifera, Cyclostomata, and Chondrichtyes (phylostrata 6 and 11) as for these internodes complete annotated genomes were lacking. The higher threshold for the trace and EST archives was necessary because of the different data structure. Using the Blast output and MS SQL database management system in a series of queries, we mapped human genes according to the evolutionary origin of their founder genes on the currently best supported phylogeny (fig. 1). Note that in a recent study, ctenophores where recovered as the earliest diverging multicellular animals (Dunn et al. 2008). However, shifting of their position among major metazoan groups does not influence our results as their genome sequence is not yet available and as almost all individual ctenophore genes which are included in the NCBI database are highly conserved and are mapping before the metazoans. Taken together, our choice of internodes is based on the availability of complete annotated genomes, the reliability of phylogenetic relationships, and the importance of evolutionary transitions. The accession numbers of all genes sorted into the phylostrata are provided in supplementary table S3 (Supplementary Material online).

Statistical Analysis of Phylostratigraphic Data

The frequency of the disease genes in every phylostratum was compared with the frequency of disease genes in the complete genome (expected frequency), and deviations are shown by calculating log-odds ratios (fig. 2). Significance of the obtained deviations from the expected frequency was tested by two-tailed hypergeometric tests (Rivals et al. 2007), and obtained *P* values were corrected for multiple comparisons via false discovery rate (FDR) at 0.05 level (Benjamini and Hochberg 1995). Similarly, we tested for significant frequency deviations of each Gene Ontology (GO) term in every phylostratum compared with the complete population of GO annotations for the whole gene set (supplementary table S1, Supplementary Material online) and the disease gene set (supplementary table S2, Supplementary Material online). As before, significance of the obtained frequency deviations was tested by two-tailed hypergeometric tests and correction for multiple comparison was done by FDR at the 0.05 level.

Precalculated protein evolutionary rates for Ensemble human–mouse orthologs were retrieved from http://www.biomart.org/. Significance of the difference in evolutionary rates between the human disease ($N = 1{,}641$) and the reference ($N = 14{,}462$) set of genes was tested by Student's *t*.

Nonessential and Polygenic Disease Genes

The total number of recovered disease genes was 1,760 based on Morbidmap (OMIM, Tag 3). This set of disease genes (Tag 3) includes only cases where the mutation was both positioned by mapping the wild-type gene and the disease phenotype itself, combined with the demonstration that a mutation in that gene is associated with the disorder (see http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html). We obtained the subsample of nonessential disease genes in two ways, one more stringent and the other more permissive. The stringent subsample (1,020 genes) was obtained by removing genes that essentially reduce the fitness to zero. These genes were annotated in the mouse knockout database (http://www.informatics.jax.org/) as "lethality-prenatal/perinatal"(MP:0005374), "lethality postnatal" (MP:0005373), or "reproductive system phenotype" (MP:0005389) and by removing genes that were annotated as essential based on human phenotype data

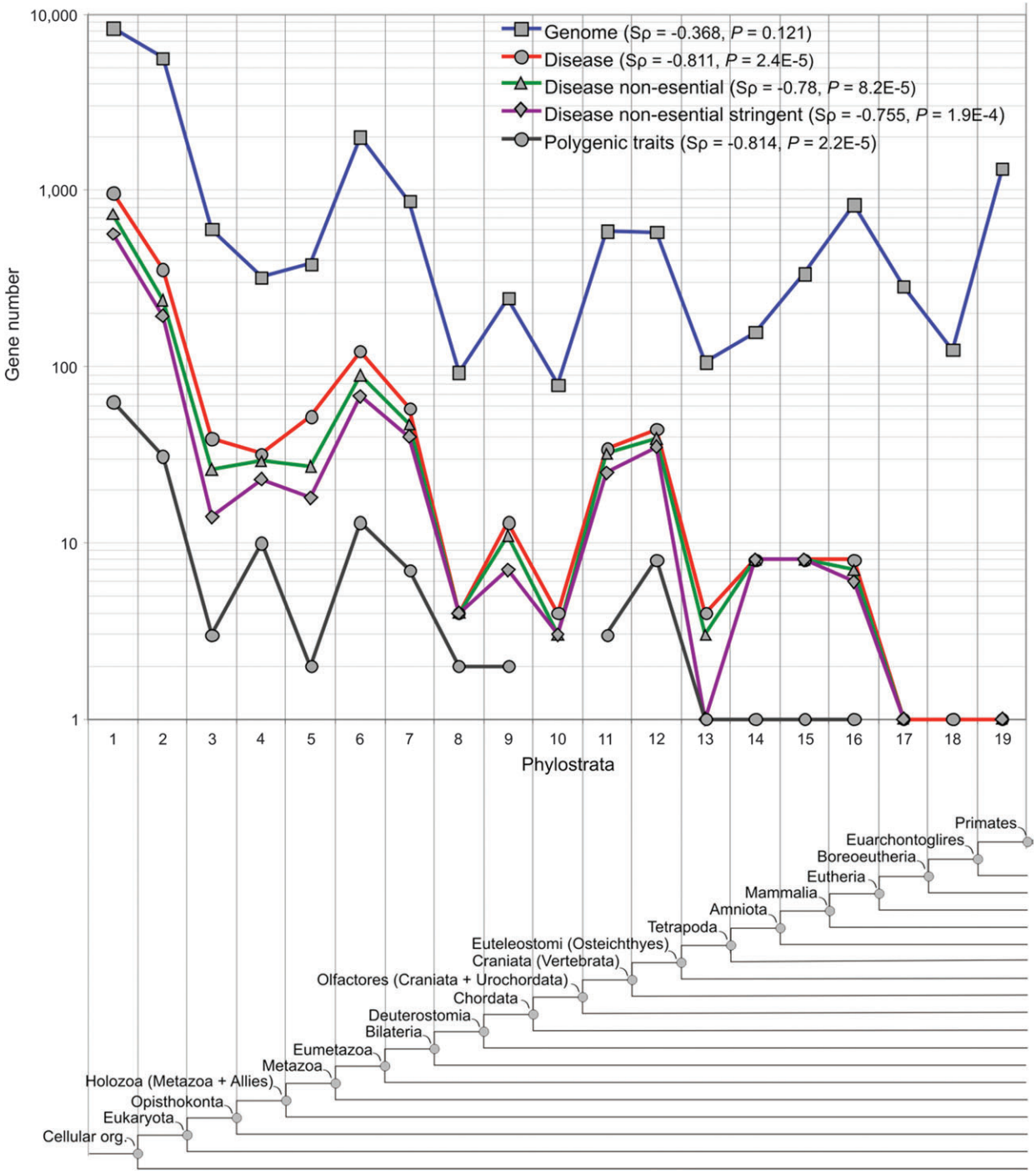

FIG. 2.—Phylostratigraphy of all human genes and different classes of disease-causing genes. The total number of human genes (N = 22,845) found in the different phylostrata is plotted (blue line—squares, note logarithmic scale on the y axis). Distribution of the total number of evaluated disease genes (N = 1,760, red line—circles). The subsample of nonessential disease genes (N = 1,305), the stringent nonessential subsample (N = 1,020), and the genes involved in polygenic traits (N = 149) are also shown. The correlation coefficient between gene count and ranked evolutionary time is listed on top (estimated by Spearman's rank correlation coefficient).

(Liao and Zhang 2008). The more permissive list of nonessential genes (1,305) was obtained by removing genes that were annotated in the mouse knockout database as lethality-prenatal/perinatal(MP:0005374). The list of 149 genes implicated in the human polygenic traits were compiled from several studies (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al. 2007; Zeggini et al. 2007; Wellcome Trust Case Control Consortium 2007a; Wellcome Trust Case Control Consortium and Australo-Anglo-American Spondylitis Consortium (TASC) 2007b; Lettre et al. 2008; Weedon et al. 2008).

Results

Based on the currently available genome data, we have created a phylostratigraphic scheme of metazoan evolution that includes 19 phylostrata (fig. 1). The taxonomic arrangement is based on a consensus phylogeny that is supported by a range of molecular and morphological data (Bourlat et al. 2006; Nikolaev et al. 2007; Dunn et al. 2008). To place all human genes into these phylostrata, we use Blast analysis with an *E* value cutoff of 0.001. This can be considered to be rather nonstringent, but we have previously found that this cutoff provides an optimal specificity with respect to the detection of orphan genes (Domazet-Lošo and Tautz 2003). It should be noted that this permissive cutoff will place all gene families that share a particular protein domain into the age class where this domain emerged first, even though a particular gene may have evolved later, for example, due to gene duplication or exon shuffling. On the other hand, because a protein domain is usually linked to a certain function, we trace the origin of this function in our analysis, irrespective of the further origin of paralogues.

Figure 2 shows the origin of 22,845 human genes plotted onto the 19 phylostrata (ps). Approximately 60% of them trace back to the origin of life and the emergence of eukaryotic cells (ps1 and ps2). These genes usually represent basic cellular functions, such as metabolic processes and transcriptional regulation (supplementary table S1, Supplementary Material online). Another peak of gene emergence is associated with the evolution of multicellularity, shortly before the Cambrian explosion (ps6). Signaling processes such as G-protein–based signaling date back to this time. Smaller peaks are evident around the emergence of bony fish/tetrapoda (ps11 and ps12) and eutherian mammals (ps16). These peaks are mainly associated with the emergence of immunity-related genes. Finally, there is an additional peak around the emergence of primates (ps19), where genes involved in spermatogenesis and keratinization are enriched. About 13% of all genes in humans have emerged in the mammalian lineage (ps15 and later) and can therefore be considered to be orphans for this taxonomic group.

To assess the distribution of disease-associated genes across the phylostrata, we have focused on the subset of genes for which a disease-causing mutation was identified (see Methods). Plotting these genes onto the same topology as for the full gene set shows a different pattern. Although the peaks of disease gene emergence mimic those of the emergence of all genes for the older phylostrata, they become notably absent from the younger phylostrata (fig. 2). Only about 0.6% of the disease genes map to ps16 or later. The genes in this class tend to be accessory proteins, such as APOC2 (apolipoprotein C-II, a component of very low density lipoprotein), MRAP (melanocortin 2 receptor accessory protein), or MUC7 (a salivary mucin thought to function in a protective capacity by promoting the clearance of bacteria in the oral cavity).

The pattern is not very different when one plots nonessential genes only. The same peaks of emergence are seen, irrespective of whether one uses a stringent or nonstringent subset (see Methods) of nonessential genes (fig. 2). In statistical terms, the comparison of the number of disease genes against the ranked evolutionary time, represented by the phylostrata, shows a significant negative correlation (Spearman's rho, fig. 2), indicating that continuously fewer disease-related genes originated over evolutionary time. Such a statistically significant correlation is not seen for the whole gene set (fig. 2).

To assess whether the functions of disease genes are a more or less random subset of the functions of the whole gene set at the different phylostrata, we compared the overrepresented GO term annotations for both groups (supplementary table S2, Supplementary Material online). Table 1 lists all overrepresented GO terms for the disease genes and compares the relative rank of the same term in the whole gene set. It is evident that there is a remarkable overlap. The most frequent terms are often the same in both lists. With few exceptions, an overrepresented term in one list is also overrepresented in the other list. The notable exception concerns the terms associated with multicellular developmental patterns in ps1 and ps2, which are not overrepresented in the full gene list. These are evidently genes that are only indirectly related to multicellularity because ps1 and ps2 cover phylostrata where only single-celled organisms existed (note that a given gene can have multiple GO terms, i.e., its function is not fully described by a single term). Further discrepancies are observed in ps15 and ps17, but these concern only very small numbers of genes.

The disease genes considered up to this point are mostly related to monogenic diseases because the systematic mapping of genes involved in common polygenic diseases has only recently become possible (Wellcome Trust Case Control Consortium 2007). We have compiled a list of 149 genes that have emerged from these studies so far, although many more are expected to come. Because each of these genes contributes only a small fraction to the full phenotype, one could consider them as modifier genes in a complex pathway, that is, orphan genes could easily be among them. However, we find again that their distribution in the phylostratigraphy mimics that of the other genes (fig. 2).

Phylostratigraphy does not only allow to plot gene emergence to particular phylostrata but allows also a statistical analysis in terms of relative over- or underrepresentation of particular gene classes in each phylostratum. This is achieved by calculating a hypergeometric statistics, the results of which are shown as odds ratios in figure 3. The results confirm that disease genes are statistically highly significantly underrepresented from ps15 onward, that is, since the evolution of mammals. Also, a highly significant statistical overrepresentation is seen for ps1 and ps5 for the whole disease gene set. Interestingly, overrepresentation in ps5 predates the peak in ps6 that is seen in the overall dataset. On the other hand, disease genes are highly significantly underrepresented among the genes that have emerged in ps2. This is somewhat surprising because this phylostratum represents the origin of eukaryotes, that is, one of the major evolutionary transitions where many novel cellular structures and processes were generated (de Duve 2007). It is also the phylostratum with the second largest absolute number of gene emergence (fig. 2), but disease-causing genes are evidently less likely to be among them.

Table 1
Comparison of Overrepresented GO Terms among the Disease Gene Set with the Rank Order of the Same Term among the Overrepresented GO Terms in the Whole Gene Set

PS	GO Term Overrepresented among Disease Genes	Rank among All Genes
1	Metabolic process	1
	Electron transport	4
	Protein amino acid phosphorylation	3
	Carbohydrate metabolic process	5
	Phosphate transport	28
	Proteolysis	3
	Ion transport	8
	Transport	15
	Homophilic cell adhesion	7
	Epidermis development	n.s.
2	Regulation of transcription, DNA dependent	1
	Multicellular organismal development	n.s.
	Protein amino acid dephosphorylation	30
	Ubiquitin cycle	3
	Regulation of Rho protein signal transduction	7
	Transcription	2
	Dephosphorylation	49
	Organ morphogenesis	n.s.
	Forebrain development	n.s.
	Regulation of transcription, DNA dependent	1
3	Transcription	2
4	Transcription	3
	Regulation of transcription, DNA dependent	2
	Positive regulation of transcription from RNA polymerase II promoter	4
	Multicellular organismal development	5
	Wnt receptor signaling pathway, calcium modulating pathway	1
5	G-protein–coupled receptor protein signaling pathway	1
	Signal transduction	3
	Activation of adenylate cyclase activity	15
	Protein–chromophore linkage	16
	Regulation of transcription	9
	Brown fat cell differentiation	41
	G-protein signaling, coupled to cyclic nucleotide second messenger	5
	G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	7
	Sensory perception of taste	n.s.
	Diet-induced thermogenesis	34
	G-protein signaling, coupled to cAMP nucleotide second messenger	14
	Heat generation	37
	Heat generation	37
	Vasodilation by norepinephrine–epinephrine involved in regulation of systemic arterial blood pressure	35
	Phototransduction	20
6	Cell communication	1
	Cell–cell signaling	3
7	Methylation	1
8	Calcium-independent cell–cell adhesion	1
9	Immune response	2
	Antigen processing and presentation	1
	Antigen processing and presentation of peptide antigen via MHC class I	4
	Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	3
	Peripheral nervous system development	23
10	Cell surface receptor–linked signal transduction	4
	Response to virus	1
	Immune response	2
	Feeding behavior	12
	Cell–cell signaling	3
	Cellular calcium ion homeostasis	8
	Neutrophil apoptosis	16
	Inflammatory response	5
	JAK–STAT cascade	20
11	None	
12	Immune response	1

Table 1
Continued

PS	GO Term Overrepresented among Disease Genes	Rank among All Genes
13	Lipoprotein metabolic process	1
	Lipid transport	4
	Negative regulation of lipid catabolic process	n.s.
	Neutrophil activation	n.s.
	Regulation of cytokine production	n.s.
	Negative regulation of lipoprotein metabolic process	n.s.
	Cell surface receptor–linked signal transduction	n.s.
	Regulation of cholesterol absorption	n.s.
	Positive regulation of interleukin-8 biosynthetic process	n.s.
14	None	
15	Keratinization	4
	Peptide cross-linking	n.s.
	Keratinocyte differentiation	n.s.
	Protein homooligomerization	n.s.
16	None	
17	None	

NOTE.—"n.s." means that the term was not significantly overrepresented in the whole gene set.

Discussion

Our data show that disease-associated genes are not simply a random subset of all genes in the human genome. There is a clear bias toward old genes among them, and the more recently evolved genes are seldomly disease associated. On the other hand, disease-associated genes appear to be a typical subset of all types of older genes, that is, their overall functional spectrum is not different from the whole set, as judged from the GO term analysis. This makes it even more surprising that newly evolved genes are rarely found among them.

It has been suggested that the ability to detect the origin of genes via Blast analysis could depend on their evolutionary rate (Elhaik et al. 2005). This effect could potentially affect our analysis, but our error would be on the conservative side. Genes with higher rates would be placed at younger phylostratigraphic levels than their true origin, that is, a fraction of the genes studied here could even be older than currently classified. On the other hand, the simulation study of Elhaik et al. (2005) assumed homogeneous substitution rates along the protein sequence, which is an unrealistic scenario. Alba and Castresana (2007) have argued that proteins usually show an among-site heterogeneity in substitution rates, owing to the presence of functional modules. Simulating such conditions, they show that the origin of a given gene can be accurately traced by Blast (Alba and Castresana 2007). Given that our nonstringent cutoff *E* value of 0.001 is likely to catch such functional modules, we conclude that our phylostratigraphic analysis does indeed trace the true origin of most genes.

Our Blast search estimates the greatest age of all genes that share a conserved domain with the focal gene, that is, a specific disease gene could have arisen later through gene duplication, whereas its paralogue might not be involved in genetic diseases. However, given the large evolutionary times that we span in our analysis, it is difficult to say which history of duplications has lead to a particular gene, that is, to distinguish the original gene from its paralogues. We have previously argued that extant genes may retain the functional properties of their founders (see detailed theoretical explanation in the main text and in the supplement of Domazet-Lošo et al. 2007). Thus, our method estimates the propensity of descendants of a particular founder to be involved in diseases. Also, we use the same analysis standards for the disease gene set and the reference set and believe therefore that our results would not be very different, if one could indeed trace the evolutionary history for each gene analyzed.

Another possible source of error could be an ascertainment bias in gene discovery. There might be a bias in disease gene discovery toward candidate genes that are conserved among species. However, because the majority of monogenic disease genes were discovered by map-based approaches with subsequent confirmation of the disease-causing mutation, this bias should not be strong. Also, it could not explain the relative absence of disease genes from phylostratum 2 or the overabundance in ps5. The situation might be different for the genes involved in polygenic phenotypes. The current studies are focused on genes with relatively strong or at least moderate phenotypic effects because the recovery of weak effect genes will require to study even larger cohorts of patients (Visscher 2008). It remains therefore open whether the pattern of gene emergence seen for the genes involved in polygenic phenotypes will change over time.

A previous analysis of evolutionary parameters in disease gene sets has suggested that they evolve faster than the nondisease gene set (Smith and Eyre-Walker 2003). When we compare mean *d*n and *d*n/*d*s calculated on the human–mouse orthologs in our data set (16,103 pairs, 1,641 disease genes), we were unable to see significant differences (*d*n, disease genes = 0.104, nondisease genes = 0.108; $P = 0.4$ Student's *t*; *d*n/*d*s, disease genes = 0.158, nondisease genes = 0.169; $P = 0.6$ Student's *t*). Kondrashov et al. (2004) found lower divergence rates for the disease genes than for the reference set, which is also different from our

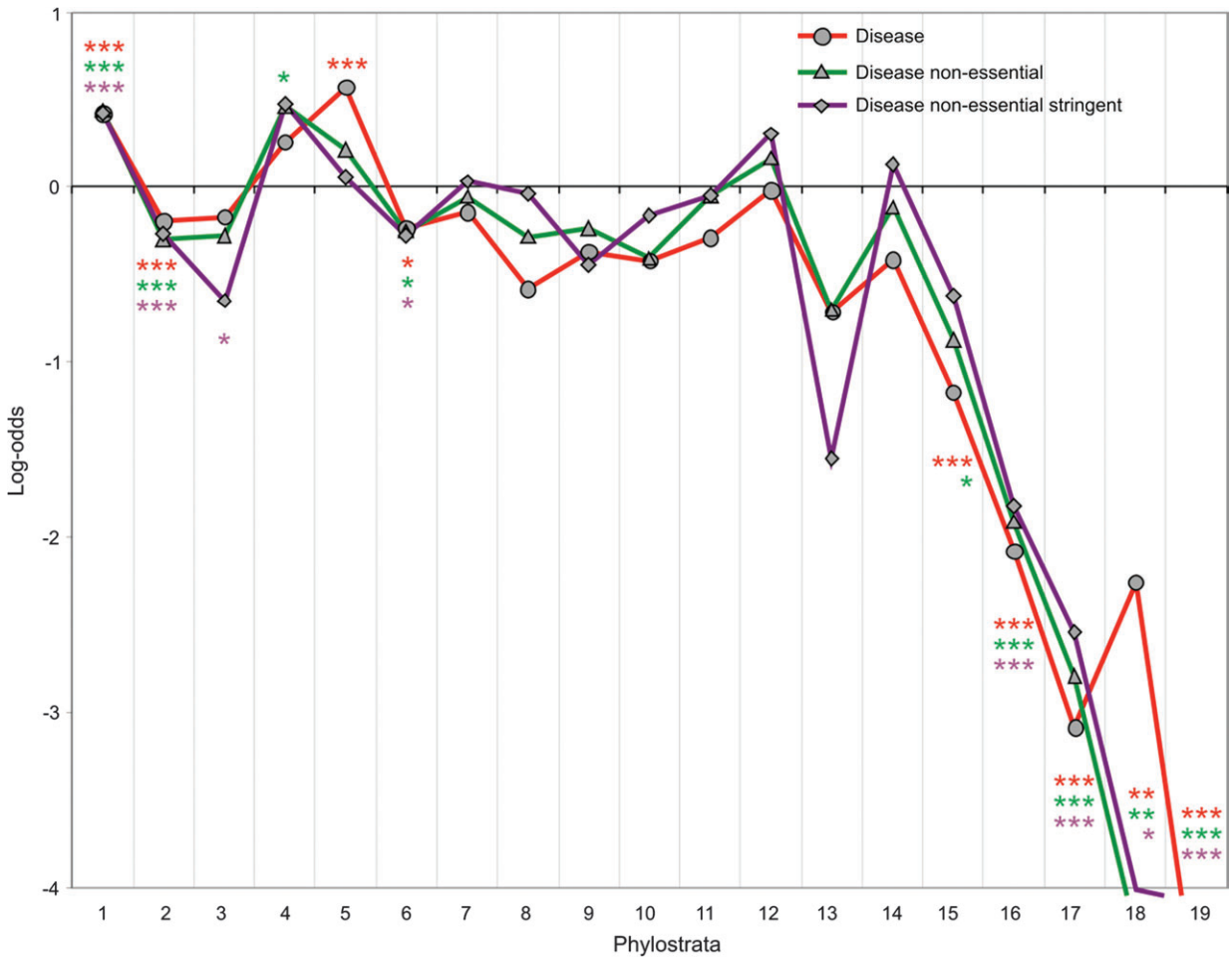

FIG. 3.—Probabilities of over- or underrepresentation of disease-causing genes in the respective phylostrata. Log-odds ratios show how the frequency of disease genes in each phylostratum deviates from the expected one estimated from the whole set of genes. Numbering of the phylostrata corresponds to those in figures 1 and 2 (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$, two-tailed hypergeometric test corrected for multiple comparison by FDR at 0.05 level).

result. However, the gene sets used in these studies are not directly comparable. The issue of faster or slower evolution of disease genes must therefore be considered as open at present. On the other hand, our data agree with the two other studies (Smith and Eyre-Walker 2003; Kondrashov et al. 2004) with respect to the finding that disease genes tend to be old and are on average longer than the genes in the reference set (for our data set: average length of disease genes = 747 amino acids, nondisease genes = 478 amino acids). To exclude the possibility that this could bias the efficiency of the Blast analysis, we restricted our search to disease genes shorter than 500 amino acids. This yielded qualitatively the same results as with the full set (supplementary fig. S1, Supplementary Material online). Interestingly, if one plots the average length of genes for all phylostrata, one gets an almost continuous decrease of length from old phylostrata to younger ones (supplementary fig. S2, Supplementary Material online). Thus, the length effect seen in the disease genes is a correlated effect of them being generally older than the reference set.

Our findings have some general evolutionary implications. Although interaction network and transcriptional analyses suggest that disease genes are not concentrated in hubs (Goh et al. 2007; Feldman et al. 2008), their ancient origin suggests that they are nonetheless involved in old biological processes. In fact, it has been shown that disease-causing mutations tend to occur at conserved positions in the proteins (Miller and Kumar 2001; Mooney and Klein 2002). Although there is—to our knowledge—currently no general theory on the emergence of genetic diseases, one would nonetheless not have suspected that biological processes that have emerged early in evolution should be most vulnerable to them. Disease-causing mutations should affect fitness and should therefore be lost over time. Thus, over extended evolutionary times, one could expect that genes that are subject to such mutations could become optimized to reduce these detrimental effects. However, because this is apparently not the case, one can conclude that genetic diseases are an inescapable component of life.

Another aspect of our finding is that there is still no new clue toward a general function of lineage-specific orphan genes. Such genes are found in practically all genome comparisons, both in eukaryotes and prokaryotes

(Domazet-Lošo and Tautz 2003; Daubin and Ochman 2004). Hence, they must be expected to play a role in the respective lineages and that they are likely to contribute to lineage-specific adaptations. But the fact that they are underrepresented in genetic screens (Domazet-Lošo and Tautz 2003) and among disease genes, as shown here, remains unexplained.

A more practical implication of our finding concerns biomedical research strategies. Given that over 90% of the disease genes have emerged before the bilaterian radiation, it seems highly justified to use organisms that are evolutionarily very remote from humans, such as nematodes or insects, as models for studying the function of disease genes. Conversely, the prevalent use of mouse as a model system would not seem as pressing as it currently is because there are less than 2% of disease genes which would not also be present in zebrafish, for example (although the functional roles of some of these genes may change over time—Liao and Zhang 2008). Furthermore, to understand the context of the biological processes in which a gene is involved, it may be advisable to use model organisms that represent the evolutionary level at which these genes emerged.

Supplementary Material

Supplementary tables S1–S3 and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

T.D.-L. was supported by a Research Fund of Republic of Croatia (grant 098–0982913–2832 to D. Ugarkovic) and Zaklada Adris as well as a postdoctoral fellowship of the Max-Planck Society. Computational resources were provided by the Isabella cluster (University computing center—SRCE), CroGrid project (RBI), and Koncar—Electronics and Informatics Inc. We thank Michael Nachmann for discussion and comments on the manuscript.

Literature Cited

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

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 57:289–300.

Bourlat SJ, Juliusdottir T, Lowe CJ, et al. (11 co-authors). 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature. 444:85–88.

Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. 14(6):1036–1042.

de Duve C. 2007. The origin of eukaryotes: a reappraisal. Nat Rev Genet. 8:395–403.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Saxuea R, Voight BF, Lyssenko V, et al. (67 co-authors). 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 316:1331–1336.

Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Dunn CW, Hejnol A, Matus DQ, et al. (11 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Elhaik E, Sabath N, Graur D. 2005. 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.

Feldman I, Rzhetsky A, Vitkup D. 2008. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci USA. 105:4323–4328.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease network. Proc Natl Acad Sci USA. 104:8685–8690.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc Natl Acad Sci USA. 104:20274–20279.

King N, Westbrook MJ, Young SL, et al. (11 co-authors). 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature. 451:783–788.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2004. Bioinformatical assay of human gene morbidity. Nucleic Acids Res. 32:1731–1737.

Lettre G, Jackson AU, Gieger C, et al. (11 co-authors). 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet. 40:584–591.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci USA. 105:6987–6992.

Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet. 10:2319–2328.

Mooney SD, Klein TE. 2002. The functional importance of disease-associated mutation. BMC Bioinformatics. 3:24.

Nikolaev S, Montoya-Burgos JI, Margulies EH. Rougemont J, Nyffeler B, Antonarakis SE. NISC Comparative Sequencing Program. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet. 5:e2.

Online Mendelian Inheritance in Man, OMIM (TM). 2008. McKusick-Nathans Institute of Genetic Medicine. Johns Hopkins University (Baltimore, MD), National Center for Biotechnology Information, and National Library of Medicine (Bethesda, MD) [Internet]. [May 2008] Available from http://www.ncbi.nlm.nih.gov/omim/

Rivals I, Personnaz L, Taing L, Potier MC. 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 15:401–407.

Rubin GM, Yandell MD, Wortman JR, et al. (11 co-authors). 2000. Comparative genomics of the eukaryotes. Science. 287:2204–2215.

Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. Gene. 318:169–175.

Visscher PM. 2008. Sizing up human height variation. Nat Genet. 40:489–490.

Weedon MN, Lango H, Lindgren C, et al. (11 co-authors). 2008. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 40:575–583.

Wellcome Trust Case Control Consortium. 2007a. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447:661–678.

Wellcome Trust Case Control Consortium and Australo-Anglo-American Spondylitis Consortium (TASC). 2007b. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet. 39: 1329–1337.

Zeggini E, Weedon MN, Lindgren CM, et al. (27 co-authors). 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 316:1336–1341.

Sudhir Kumar, Associate Editor

Accepted September 19, 2008