On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication

Jonathan E. Dickerson and David L. Robertson*

Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

***Corresponding author:** E-mail: david.robertson@manchester.ac.uk.

Associate editor: James O. McInerney

Abstract

Over 3,000 human diseases are known to be linked to heritable genetic variation, mapping to over 1,700 unique genes. Dating of the evolutionary age of these disease-associated genes has suggested that they have a tendency to be ancient, specifically coming into existence with early metazoa. The approach taken by past studies, however, assumes that the age of a disease is the same as the age of its common ancestor, ignoring the fundamental contribution of duplication events in the evolution of new genes and function. Here, we date both the common ancestor and the duplication history of known human disease-associated genes. We find that the majority of disease genes (80%) are genes that have been duplicated in their evolutionary history. Periods for which there are more disease-associated genes, for example, at the origins of bony vertebrates, are explained by the emergence of more genes at that time, and the majority of these are duplicates inferred to have arisen by whole-genome duplication. These relationships are similar for different disease types and the disease-associated gene's cellular function. This indicates that the emergence of duplication-associated diseases has been ongoing and approximately constant (relative to the retention of duplicate genes) throughout the evolution of life. This continued until approximately 390 Ma from which time relatively fewer novel genes came into existence on the human lineage, let alone disease genes. For single-copy genes associated with disease, we find that the numbers of disease genes decreases with recency. For the majority of duplicates, the disease-associated mutation is associated with just one of the duplicate copies. A universal explanation for heritable disease is, thus, that it is merely a by-product of the evolutionary process; the evolution of new genes (de novo or by duplication) results in the potential for new diseases to emerge.

Key words: human disease genes, evolution and origins of disease, gene duplication, whole-genome duplication, gene retention.

Introduction

Sequencing of the human genome is permitting a detailed characterization of the genetics underpinning human disease (Altshuler et al. 2008) and thousands of polymorphisms associated with disease have been linked to heritable mutations (Hamosh et al. 2005). Although it is of paramount importance to understand the aetiology of specific diseases in this way, it is also interesting to consider the origins of genetic disease more generally. Results of evolutionary studies have reported that the age of disease-associated genes is heavily biased, in that they tend to be ancient, with a tendency to arise with early metazoa (López-Bigas and Ouzounis 2004; Domazet-Loso and Tautz 2008; Cai et al. 2009). As it is inherently difficult to identify the first emergence of a specific disease, it is assumed that the evolutionary age of a disease is the same as the origin of the gene the disease is associated with, that is, the inferred presence of the gene in the most recent common ancestor of extant species (López-Bigas and Ouzounis 2004; Domazet-Loso and Tautz 2008). Focusing on common ancestors without considering a gene's duplication history will, however, only permit a superficial understanding of disease origins.

This limitation is because functional innovation is rarely achieved via the de novo genesis of genes. Rather, new genes are most frequently generated by duplication either in whole or from part of existing genes (Ohno 1970; Lynch and Conery 2000; Long et al. 2003; Conant and Wolfe 2008). Duplicate genetic material is inherently redundant and so less prone to purifying selection. As a consequence, one of the duplicate copies will tend to accumulate mutations and become a nonfunctional pseudogene (termed pseudogenization). On occasion, mutations (either existing or new) can be selectively advantageous contributing to the evolution of novel function (neofunctionalization) or in the partitioning of existing functions (subfunctionalization) and become fixed. In this way, a duplicate copy of a gene can become associated with a different phenotype to its paralogous partner. This has consequences for linking disease phenotypes to genes.

Specifically, a disease-associated gene should be associated with a particular duplicate rather than the common ancestor of its gene family. As most Mendelian diseases are only characterized on evolutionary recent taxa, it is reasonable—despite being an underestimate—to use the most recent duplication node (MRD) of a human disease-associated gene as a proxy for disease origin. For example, the disease epilepsy cannot exist prior to the development of the mammalian brain and neurotransmitters, for

example, GABA1. Indeed, without considering duplication, GABA1 receptor's "age" would be misleadingly reported as its common ancestor, in this case an evolutionary ancient premetazoan origin. Such failure to include duplication history when dating disease-associated genes has the potential to misclassify the origin of a disease as erroneously ancient. To test this hypothesis, we date both the most recent common ancestor and the duplication history of known human disease genes.

We find, as previous evolutionary studies have, that the majority of disease genes have common ancestors with an origin predating animals that evolved bilaterian symmetry (Bilateria) greater than one billion years ago (López-Bigas and Ouzounis 2004; Domazet-Loso and Tautz 2008). However, 80% of these disease genes have a duplication history and the most recent of these duplications do not map back to Bilateria rather the majority came into existence with the radiation of the bony vertebrates (Chordata–Euteleostomi), some 500–600 Ma. Interestingly, the proportion of disease-associated genes with a duplicate history approximately "tracks" the overall distribution of duplicated genes until 390 Ma and is independent of host protein function or disease class. Singletons (genes present only as single copies) that are disease genes decline in numbers with recency. We also find a strikingly asymmetric distribution of disease genes in gene families, with a tendency for only one member of a gene family to be diseased associated. We discuss the implications of these findings on our understanding of Mendelian disease origins.

Methods

Assigning an Age to Human Genes

Although we cannot definitively identify the first emergence of a disease mutation, we can identify the origin, or "age," of the linked disease gene by inferring the most recent common ancestor (as indicated by taxonomic level) of all extant species from the tree of life that have homologs to this gene (supplementary fig. S1, Supplementary Material online). To identify the origin of individual human genes (disease associated or otherwise), we downloaded evolutionary trees generated from EnsemblCompara multispecies comparisons in Ensembl release 56 (ftp://ftp.ensembl.org/pub; Flicek et al. 2009), which included 51 species. In EnsemblCompara, molecular phylogenies for each gene have been inferred using orthology and paralogy gene prediction analysis. The longest translation of every identified protein-coding gene has been queried against each species protein in Ensembl using BlastP. Clusters of homologous genes, which will include gene families from each genome, have been extracted and aligned using MUSCLE (Edgar 2004), and the resulting multiple alignment used to generate a tree with TreeBeST from TreeFam (Li et al. 2006). The TreeBesT framework incorporates a number of sophisticated phylogenetic methods including DNA, codon, and protein maximum likelihood methods. A combined tree is then generated to integrate the different phylogenetic results. DNA or codon-based methods are used for parts of the phylogeny that exhibit shorter evolutionary distances, for example, intraclass comparisons, whereas protein information is used for longer evolutionary distances, for example, comparisons between classes (Vilella et al. 2009). We then query the EnsemblCompara phylogenies to identify the age of each human gene (its taxonomic level) and whether it has duplicated or is a singleton, a gene present once in the human genome with no identifiable duplication history. Note that the reliable identification of singletons assumes both detectable homology and the retention of any duplicates on multiple lineages.

For each duplicated human gene, we identified both DCA, the taxonomic level associated with the duplicate's common ancestor, and MRD, the taxonomic level associated with the "most recent duplication node." For example, we assign the chloride channel Ka gene *CLCNKA* as having an MRD of Catarrhini and a DCA of Bilateria based on its duplication history (supplementary fig. S2*A*, Supplementary Material online). For MRDs, we also identified those whose origin has been attributed to whole-genome duplication (Makino and McLysaght 2010). For each singleton, we identified SCA the taxonomic level associated with the "singleton's common ancestor." For example, we assign the histidine–proline-rich glycoprotein gene *HRG* an SCA of Coelomata (supplementary fig. S2*B*, Supplementary Material online). Putative, uncharacterized and undetermined taxonomic levels, and ambiguous mappings were excluded from our classifications. Taxonomic levels without representation amongst both duplicates and singletons were not considered for evolutionary history analyses (specifically Coelomata and Euarchontoglires), but were considered for disease type, gene function, and selection analyses if sufficient data were available.

The approximate evolutionary ages of taxonomic levels were obtained using a consensus of multiple sources (Benton and Ayala 2003; Hedges et al. 2006; Benton and Donoghue 2007; Donoghue and Benton 2007). Note that this use of divergence dates of common ancestors based on the molecular phylogeny for dating disease origin will always be a more recent estimate as they will coincide with speciation events, whereas the disease-associated mutation could have existed prior to speciation. Ensembl genes were mapped to NCBI gene identifiers wherever possible to facilitate data integration.

Mapping Diseases to Human Genes

Disease-associated genes were obtained from Online Mendelian Inheritance in Man (OMIM; Hamosh et al. 2005) and filtered to include only those genes where strong evidence that at least one mutation in the particular gene is causative of the disease. This resulted in 3,328 unique diseases across 1,755 unique genes. These data were cross-referenced with our ancestor data to yield 1,324 disease-associated genes with MRD/DCA annotations and 331 with SCA annotations. A common ancestor was not identifiable for 100 disease-associated genes.

Functional Analysis of Diseases, Genes, and Age

Medical Subject Headings (MeSH; http://www.nlm.nih.gov/mesh), which, in addition to other life science categories, hierarchically describe disease, for example, diseases to digestive system, neoplasms, etc., were downloaded and filtered for the "Diseases [C]" annotation. Nineteen top level categories were selected from the MeSH trees (http://www.nlm.nih.gov/mesh/filelist.html) that are representative of high-level classifications. These were merged with 22 disease categories and annotations from (Goh et al. 2007) to give 16 unique disease categories and provide a consistent annotation of disease. Each disease-associated gene was then labeled with the related category to yield higher level annotation (supplementary table S1, Supplementary Material online).

Using the Gene Ontology (GO)—a set of three structured controlled ontologies that describes gene products in terms of their associated biological processes, cellular compartments, or molecular functions (Harris et al. 2004)—a gene can be annotated with multiple GO terms from each of the ontologies, for example, "immune response," "plasma membrane," and "coreceptor activity," respectively. GO terms were collected for each human gene from the NCBI "gene2go" file (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA). Term ancestors were then identified for each term from "gene_ontology_edit.obo" (http://www.geneontology.org) to ensure complete coverage. A GO slim (a cut down version of GO; http://www.geneontology.org/GO.slims.shtml) term was also determined to identify higher level annotation. All GO terms were separated into the three ontologies: biological process, cellular component, and molecular function. Terms with low counts (<20 for biological process and molecular function terms and <10 for the broader cellular component terms) were excluded from this analysis, as were taxonomic levels with low counts (<15). Full GO terms were tested for overrepresentation among each taxonomic level using Fisher's exact test and adjusted for multiple testing using the Bonferroni correction method in R (http://www.r-project.org).

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000) were collected for each human gene using the mappings and descriptions available (ftp://ftp.genome.jp/pub/kegg/pathway). KEGG contains multiple related databases, including PATHWAY for exploring the higher order functions of genes in terms of the network of interactions in which they participate. KEGG pathways are manually drawn maps that can be explored from genes to functions or vice versa, providing an alternative to GO. For example, a gene could be labeled as participating in a "chemokine signalling pathway." Terms and taxonomic level with low counts (<15) were excluded and all KEGG annotations were also tested for overrepresentation among each taxonomic level as for GO terms.

Pairwise *dN*/*dS* Values

Ensembl contains pairwise *dN* and *dS* values between pairs of genes for closely related species calculated using codeml from the PAML package (model = 0, NSsites = 0) (Yang 2007; Vilella et al. 2009). Saturation of the *dS* values may occur when the species evolutionary distance is too large, potentially biasing the *dN*/*dS* ratio. Accordingly, we limit our analysis to Chimpanzee (*Pan troglodytes*) and Orangutan (*Pongo pygmaeus*) in Ensembl release 56 (Flicek et al. 2009), as downloaded from the Ensembl FTP (ftp://ftp.ensembl.org/pub). Ratios >1.5 were excluded from our analysis. For each Chimpanzee and Orangutan gene, and hence taxonomic level, a *dN*/*dS* ratio, mean, standard deviation, and standard error were calculated.

Results

Dating Disease Mutations

For all human genes for which a duplicate copy (paralog) exists, the MRD and DCA taxonomic levels (from Metazoa to *Homo sapiens*, see supplementary fig. S1, Supplementary Material online) were identified and whether or not they are associated with a disease gene recorded (table 1). For genes with no identifiable duplication history, their common ancestor, SCA, was identified and their disease association status recorded (table 1). Strikingly, of the 1,655 disease genes, 1,324 (80%) are associated with genes with a duplication history. Thus, although the majority of disease genes do have an ancient origin (58% from combined DCA and SCA) falling in the taxonomic level Bilateria or older, understanding the origin of disease has to consider the duplication history of disease genes. Focusing on MRDs, we find a high number of disease-associated genes with a duplication history at the Chordata–Euteleostomi taxonomic level (61%; table 1). Conversely, just 16% of disease-associated genes with a duplication history have an MRD pre-Bilateria (table 1). As one would expect, given the recognized importance of whole-genome duplication events in early vertebrate evolution (Blomme et al. 2006; Conrad and Antonarakis 2007; Kasahara 2007), the majority of disease genes that have been identified as whole-genome duplicates (860 from Makino and McLysaght 2010) are in the Euteleostomi taxonomic level (75%; supplementary fig. S3, Supplementary Material online).

These results apparently suggest a bias in the origin of disease genes. Specifically, not only do they indicate that disease genes tend to be ancient but also that more disease genes came into existence with the origin of vertebrates and many of these are associated with whole-genome duplication. To investigate this further, we compared the distribution of SCA, DCA, and MRD across evolutionary time to null expectations for the distribution of disease genes by sampling from all genes 10,000 times (fig. 1*A*, *C*, and *E*). Interestingly, these distributions are remarkably similar, indicating that the numbers of disease genes arising in each taxonomic level are directly correlated with the number of genes that came into existence in that taxonomic level, confirmed as statistically significant by regression analysis (fig. 1*B*, *D*, and *F*).

To visualize the relationship between disease gene and gene emergence, we plotted the ratio of the proportion of disease-associated genes in each taxonomic level, relative

Table 1. Summary of Disease-Associated, or Not, Gene Counts for Different Taxonomic Levels for MRD Nodes, DCA, and SCA.

	MRD		DCA		SCA	
Taxonomic Level	Disease	Nondisease	Disease	Nondisease	Disease	Nondisease
Homo sapiens	63	804	28	206	0	0
Homininae	5	233	1	81	0	29
Hominidae	10	179	4	62	0	46
Catarrhini	13	393	6	116	0	8
Primates	2	51	0	17	0	32
Eutheria	55	941	13	558	3	367
Theria	14	366	2	234	4	160
Mammalia	28	320	6	304	5	93
Amniota	17	293	17	439	9	223
Tetrapoda	31	233	17	250	3	73
Euteleostomi	806	6,141	367	3,696	51	760
Chordata	73	543	125	838	33	441
Bilateria	206	1,550	734	5,232	81	815
Fungi/Metazoa	1	8	4	22	142	1,004
Total	1,324	12,055	1,324	12,055	331	4,051

to the evolution of all genes at that taxonomic level (fig. 1*G*). Comparing MRD disease-associated genes to all MRD genes across the tree of life, the proportion of disease-associated genes (associated with paralogs) in each time period are relatively constant, with an average of 13% until 390 Ma (fig. 1*G*). This indicates that the emergence of disease-associated genes has been relatively stable in most of evolutionary history. Since the Eutheria taxonomic level, the proportion of disease genes drops to an average of 4% (fig. 1*G*), presumably a reflection of the low numbers of genes that have emerged since this time (7%). We also investigate the proportion of disease-associated genes in each time period for singletons (SCA) and for DCA and find a steady decline over evolutionary time (fig. 1*G*).

Our results, thus, demonstrate that the numbers of disease-associated genes at the different taxonomic levels approximately tracks the numbers of genes present at that specific taxonomic level, up to 390 Ma (fig. 1*G*). After this time, fewer novel genes came into existence either by duplication (21%) or de novo (6%) of which only 1.2% and 0.1% are associated with disease, respectively. The main MRD peak approximately 500 Ma (fig. 1*C*) is consistent with the documented contribution of gene duplication between the evolution of chordates and Euteleostomi and indicates the peak of disease-associated genes is mainly a consequence of the high levels of retained duplicates at this time. Thus, although the majority of MRD disease-associated genes emerged with Chordata–Euteleostomi, a comparably high number of genes also appeared in this period. Similarly, the additional peaks at the Catarrhini–Hominidae taxonomic level is supported by additional gene duplication events (Marques-Bonet et al. 2009).

Assessing Functional Relationships among Disease-Associated Genes

Although OMIM (Hamosh et al. 2005) provides an exhaustive list of disease and gene associations, it is difficult to compare them ad hoc. Accordingly, to explore the relationship between disease and each taxonomic level, we organized diseases into significantly broader and higher level categories (see Methods), for instance, "Tay–Sachs disease" would be classified as a "metabolic" disease. We find the distributions of the disease categories in each taxonomic level are broadly similar to the previous global trend (fig. 2). In each case, we find an overrepresentation of the pre-Metazoa and Metazoa–Bilateria taxonomic level among orthologs (i.e., SCA and DCA) (fig. 2*A*) and the Chordata–Euteleostomi taxonomic level among paralogs (i.e., MRD) (fig. 2*B*) across all disease categories. Additionally, we find no significant differences between diseases emerging at each taxonomic level for either singletons or duplicates. This highlights how different diseases emerge nonspecifically at each taxonomic level, in approximate proportion to the overall emergence of genes in the different taxonomic levels.

We also investigated the distribution of biological functions using the GO (Harris et al. 2004) for both disease-associated and nondisease–associated MRDs (see Methods). For each ontology, and each term, again, we find an overrepresentation of the Chordata–Euteleostomi taxonomic level. More significantly, we find very little difference between the distribution of GO terms present in disease-associated and nondisease–associated MRDs. Similarly, the distribution of terms across taxonomic level remains similar (supplementary fig. S4, Supplementary Material online).

To further explore the functional relationship within each taxonomic level, overrepresented terms were calculated from GO for both disease-associated and nondisease–associated MRDs (see Methods). We find overrepresented biological process terms for disease-associated MRDs to be overall higher level (hierarchically in terms of GO) for older taxonomic levels (e.g., "transport") than for nondisease–associated MRDs (supplementary tables S2 and S3, Supplementary Material online). The opposite appears to be true for younger taxonomic levels, with more low level terms being numerous (e.g., "generation of precursor metabolites and energy") among disease-associated MRDs (supplementary tables S2 and S3, Supplementary Material online). This

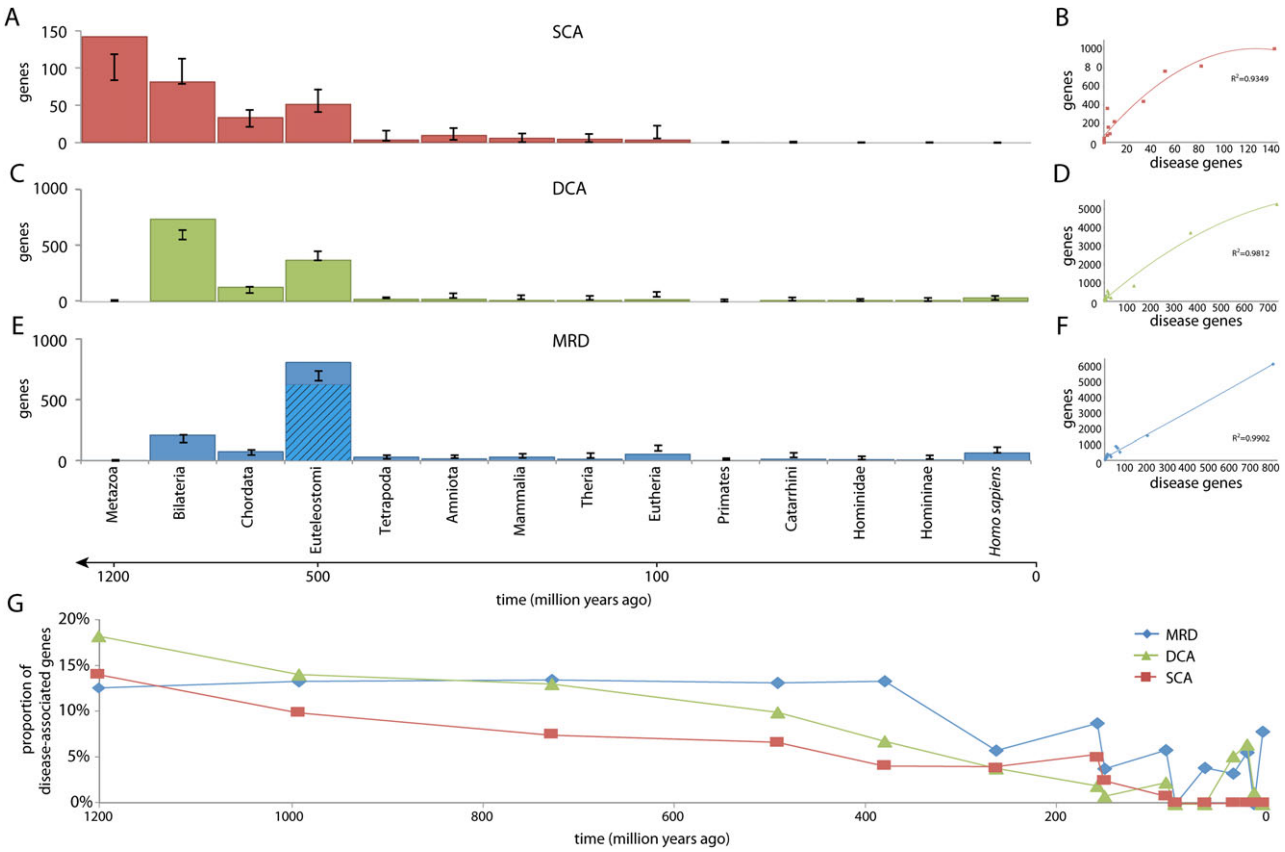

FIG. 1. The association of disease-associated genes with evolutionary history. Distribution of disease-associated genes for (*A*) SCA, (*C*) DCA, and (*E*) MRD over time. The proportion of duplicates attributed to whole-genome duplication (Makino and McLysaght 2010) are shown (hashed lines) for Euteleostomi only, as these proportions were ⩽5% for other periods (supplementary fig. S3, Supplementary Material online). Null distribution: random genes were selected for the distributions of MRD, DCA, and SCA genes, maintaining counts, from their respective nondisease–associated gene lists; this was repeated 10,000 times and the upper and lower quantiles (2.5% and 97.5%, respectively) of these distributions are shown as error bars. Taxonomic levels are indicated on the *x* axis of panel E and approximate evolutionary time below this. The proportions of disease-associated genes versus nondisease–associated genes for each taxonomic level are also shown for SCA (*B*), DCA (*D*), and MRD (*F*); polynomial regression trend lines (degree = 2) are shown in each case: SCA: $R^2 = 0.93$, *F* statistic = 78.97, P = value 3.0×10^{-7}; DCA: $R^2 = 0.98$, $F = 287.5$, P = value 3.2×10^{-10}; and MRD: $R^2 = 0.99$, $F = 558$, P = value 8.8×10^{-12}. (*G*) Ratios of the proportions of disease-associated SCA (red), DCA (green), and MRD (blue) among all SCA, DCA, and MRD, respectively, in each taxonomic level over approximate evolutionary time.

suggests that fundamental cellular processes evolving early in life are more important to the cell and so less likely be disease associated, presumably because of the probable dramatic phenotypic consequence of any disruption (Goh et al. 2007). As expected, we find multiple overrepresented terms for the Chordata–Euteleostomi taxonomic level in both disease-associated and nondisease–associated MRDs (supplementary fig. S4 and supplementary tables S2 and S3, Supplementary Material online), although noticeably more in the former, again confirming high levels of duplication at this taxonomic level dramatically contributed to disease emergence.

Disease-associated and nondisease–associated MRDs share similar molecular function terms, predominantly catalytic, binding, receptor, and transporter activity. Interestingly, overrepresented cellular compartment terms are sparingly distributed over time with notable omissions from Theria to *H. sapiens* among disease-associated versus nondisease–associated MRDs (supplementary fig. S4, Supplementary Material online).

We also investigated KEGG pathways (Kanehisa and Goto 2000), an alternative to GO, for both disease-associated and nondisease–associated MRDs (see Methods). Here, we are specifically interested in broad pathways common (or not) to each taxonomic level. Similar to the GO terms, we find comparable distributions for both disease-associated and nondisease–associated MRDs (supplementary fig. S5, Supplementary Material online). Testing for overrepresentation among complete pathways reveals the majority of KEGG terms among MRDs to be metabolic in nature across all taxonomic level. Furthermore, among disease-associated MRDs, signalling pathways, and cancer terms are overrepresented in the taxonomic level (Chordata–Tetrapoda). However, similar pathway overrepresentations were observed for nondisease–associated MRDs (supplementary fig. S5, Supplementary Material online, green asterisks). These results corroborate our finding that diseases emerge nonspecifically and have a tendency to track the overall emergence of genes at each taxonomic level.

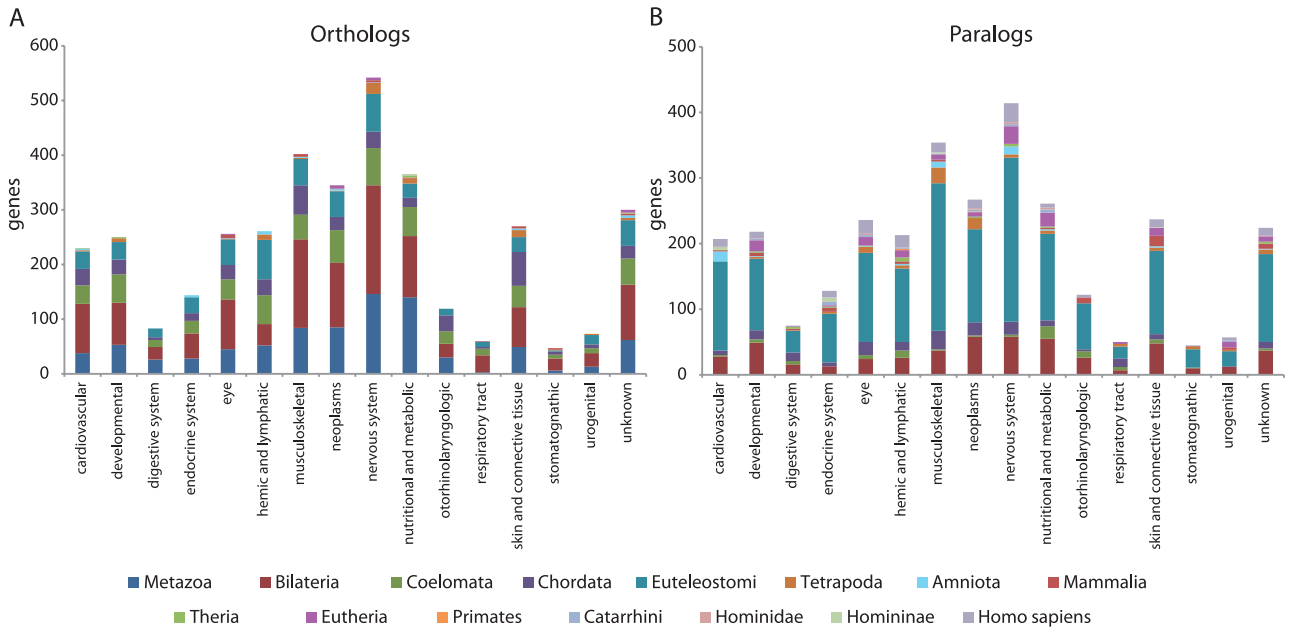

FIG. 2. The evolution of disease types. Disease class frequencies for disease-associated genes for (*A*) orthologs and (*B*) paralogs for each taxonomic level. Disease classes correspond to high-level categories.

Positive Selection Effect

As previous studies have demonstrated, we find that on average disease-associated genes have lower dN/dS (nonsynonymous substitutions/synonymous substitutions) ratio than nondisease–associated genes (Hsiao and Vitkup 2008; Cai et al. 2009). We find that this is mostly the case across evolutionary time for orthologs (i.e., SCA + DCA), with a mean dN/dS of 0.73 for nondisease–associated and 0.38 for disease-associated genes. For paralogs, that is, MRDs, we find a mean dN/dS of 0.46 for nondisease–associated and 0.28 for disease-associated duplicates. However, in the Tetrapoda–Theria (ortholog) and Primate (paralog) taxonomic levels, dN/dS are greater for disease-associated genes, highlighting the value of independently considering the many taxonomic levels, rather, than arbitrarily condensing data into fewer uniform bins (Cai et al. 2009). In addition, our results indicate that the apparent difference between "young"—defined by Cai with respect to humans and primates onward (Cai et al. 2009)—disease-associated and nondisease–associated genes is based on extremely limited data (fig. 3), 3% and 7%, respectively. As a consequence, conclusions drawn from such comparisons must be considered cautiously.

Furthermore, and confirming previous work (Domazet-Loso and Tautz 2008; Cai et al. 2009), we find an inverse relationship between evolutionary rate and gene age (fig. 3); specifically, for nondisease–associated genes, the dN/dS ratio decreases with evolutionary time ($P < 0.05$ in both cases; fig. 3). However, this is not the case for disease-associated paralogous genes ($P > 0.05$; fig. 3), again, indicating the importance of considering duplication history. These observations also need to be considered cautiously owing to limited data from Tetrapod divergence onward.

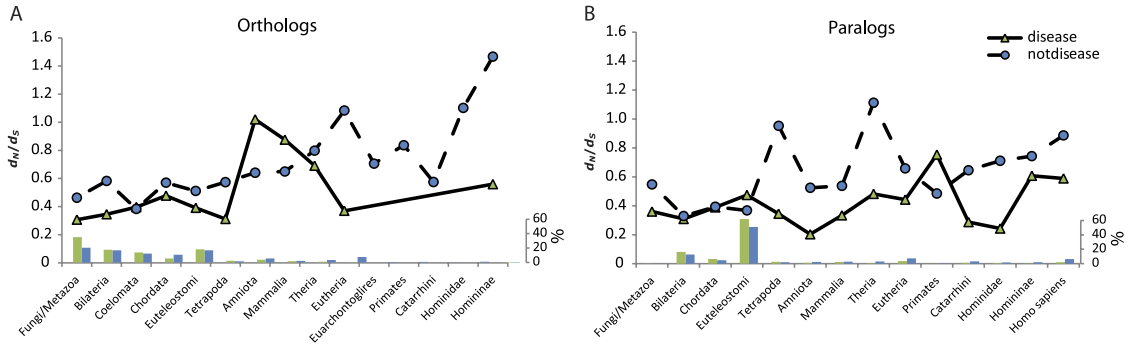

FIG. 3. Effect of positive selection on disease-associated genes. Mean dN/dS between *Homo sapiens* and *Pan troglodytes* for disease-associated (green triangles) and nondisease–associated (blue circles) orthologs (*A*) and paralogs (*B*) in each taxonomic level. Category axis labels corresponds to each taxonomic level. Inset bar chart displays percentage of both disease-associated and nondisease–associated genes in each taxonomic level.

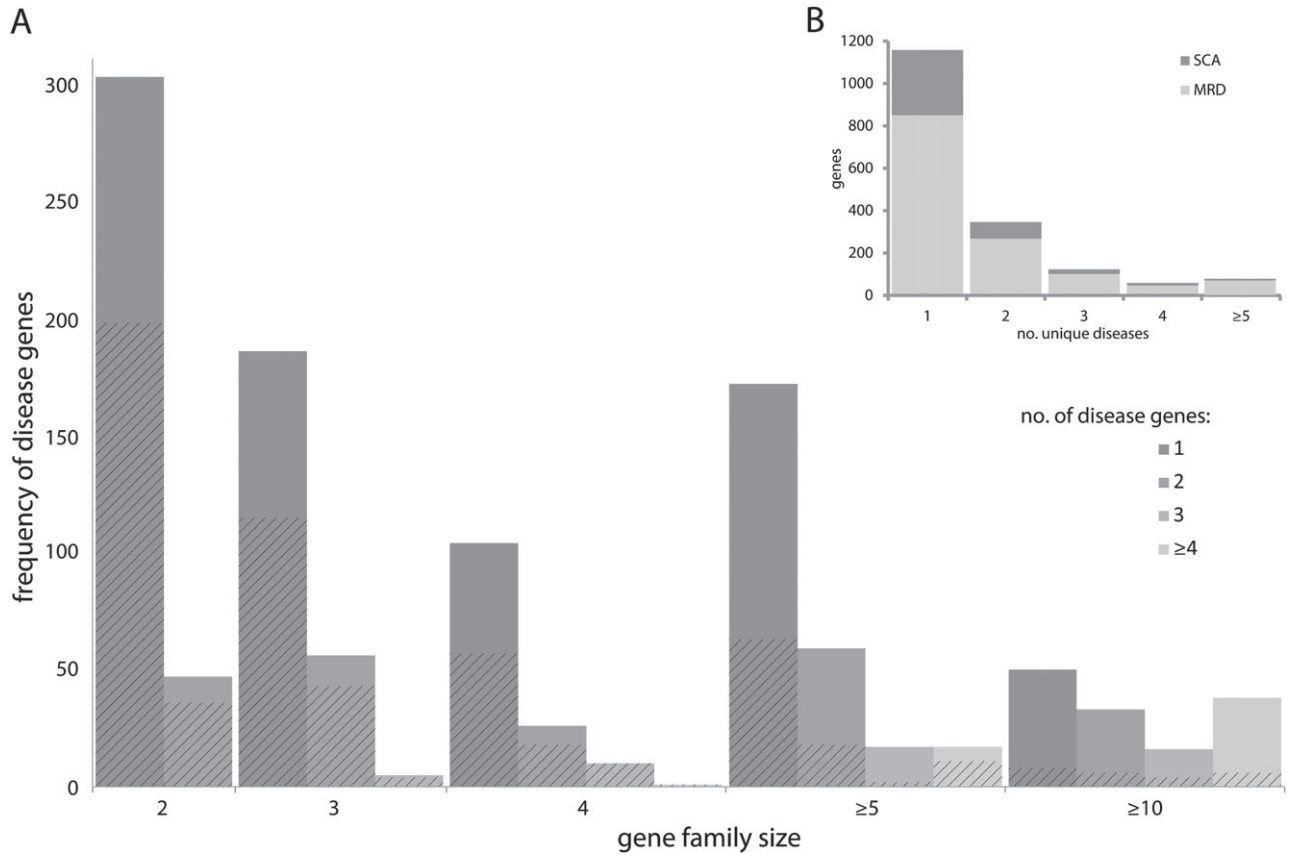

FIG. 4. (*A*) Frequencies of disease genes associated with different sizes of gene families and (*B*) frequencies of unique diseases associated with the same genes for SCA and MRD. The proportion of duplicates attributed to whole-genome duplication (Makino and McLysaght 2010) are shown for panel A (hashed lines).

Note that we obtained similar results for the Orangutan comparisons (*P. pygmaeus*; supplementary fig. S6, Supplementary Material online).

Discussion

Considering the duplication history of disease-associated genes, we find that Mendelian diseases have a propensity to be associated with genes that are paralogous, that is, associated with a specific duplicated gene in the human genome. This confirms our hypothesis on the importance of considering both common ancestors of genes and any duplication history when dating disease emergence. Although the common ancestors of disease genes do tend to be ancient, by mapping disease to duplication history, we find that the majority of duplicates cluster at the divergence of the bony vertebrates approximately 500–600 Ma (Chordata–Euteleostomi; fig. 1C) of which most of them (80%) have been identified as being generated by whole-genome duplication.

Interestingly, when we control for the differing gene emergence rates throughout evolutionary time, we observe that the emergence of disease-associated genes is in proportion to the number of nondisease-associated genes up to approximately 390 Ma. For example, the peak at Chordata–Euteleostomi corresponds to a substantial number of gene duplications, most of which are associated with whole-genome duplication. Therefore, although we observe a greater frequency of disease-associated duplicates in this taxonomic level, we similarly observe a greater number of duplicates. Correcting, thus, for gene emergence for the different taxonomic levels, we find for duplicated genes that the proportion of disease-associated genes has remained approximately constant, despite the evolution (generally) of increasing functional complexity (fig. 1).

This tendency for the emergence of disease mutations associated with genes with a duplication history to track the emergence of novel genes is presumably due to 1) the hitchhiking (Smith and Haigh 1974) of disease-causing mutations with beneficial mutations that have contributed to functional evolution and/or 2) a consequence of compensation by duplicate copies, permitting the accumulation of disease-causing mutations, either because compensation is only partial or the duplicate pairs subsequently diverge and compensation is no longer effective (Hsiao and Vitkup 2008). If this is the case, we would predict that this will lead to the tendency for an asymmetry in the relationship of disease genes to duplicates, that is a tendency for only one member of a pair to have an association with disease. Investigating this we find that this is the case for 87% of pairs and to be a general rule for different sizes of gene families (groups of paralogous genes; fig. 4*A*).

Our finding that the emergence of disease-associated duplicate genes has been relatively constant in evolutionary history is, presumably, due to an upper limit of disease mutations that populations can tolerate, that is, their mutational load (Kimura et al. 1963; Lynch and Gabriel 1990). Indeed, for the most part, diseases have been found to be associated with mildly deleterious mutations that, excluding rare beneficial disease mutations (Altshuler et al. 2008), are tolerated in the human population as they have limited impact on reproductive success (Rice 2002; Reed and Aquadro 2006). That there are relatively more disease genes associated with larger gene families (fig. 4*A*) and more cases of multiple diseases being associated with the same gene for duplicates (fig. 4*B*) supports this "disease-tolerance" hypothesis.

Our findings also suggest a hypothesis for the observed relationship between human disease and copy number variation (Conrad and Antonarakis 2007; Zhang et al. 2009; Cooper and Kehrer-Sawatzki 2011); that is, these types of changes are more likely to tolerate disease mutations because they tend to be nonlethal and so can proliferate in association with copy number variants. It is well documented that duplication of genes leads to accelerated evolution in one of the partners due to the relaxation of purifying selection, which in exactly the same way will lead to the emergence of slightly deleterious mutations, and it is these changes that have the potential to manifest as human disease (fig. 4*A*). The interesting thing in terms of understanding the origins of disease is how intimately linked disease is with functional evolution. This even influences the types of genes that proliferate in the human genome. For example, in the case of whole-genome duplicates, these have contributed significantly to evolution because dosage effects favor their retention, resulting in a greater probability that they will contribute to functional evolution (Hakes et al. 2007), which results in their subsequent duplication being constrained to a relatively greater degree due to deleterious dosage effects (Makino and McLysaght 2010).

For genes present as singletons, we find the proportion of disease-associated genes declines with recency, suggesting that the mutational load tolerated by populations with regard to singletons mostly decreases with the evolution of increasing functional complexity. Presumably, this is because singletons are less likely to be associated with functional innovation so there will be limited selection pressure acting on them, resulting in less hitchhiking of disease mutations. In addition, unlike for paralogs, there will be no opportunity for direct compensatory effects. A similar decreasing with recency trend, although with a higher proportion of disease genes, was observed for DCA (fig. 1*G*).

Exploring the functional relationships among taxonomic level reveals similar patterns for the three GOs and KEGG pathways. Disease-associated MRD tend to track in parallel to MRDs in a functionally independent and nonspecific way with a distinct peak at bony vertebrate divergence. SCA and DCA similarly track with an early pre-Bilateria peak (supplementary figs. S4 and S5, Supplementary Material online). Such similarity in the functional relationships between disease- and nondisease–associated genes has precedence, having previously been described for GO terms (Domazet-Loso and Tautz 2008). Additionally, we found that this tracking is independent of the type or class of disease (fig. 2).

Concerning the role of positive selection acting on disease-associated genes, we propose that previously observed patterns could be due to sampling effects rather than any intrinsic properties of these genes. Previous studies have found that disease-associated genes have dN/dS (nonsynonymous substitutions/synonymous substitutions) ratio higher (Smith and Eyre-Walker 2003; Huang et al. 2004), lower (Hsiao and Vitkup 2008; Cai et al. 2009), or no different to that observed for nondisease–associated genes (Thomas and Kejariwal 2004). It is has been demonstrated that disease-associated genes evolve at similar rates regardless of their evolutionary age and at a slower rate with reference to more recently evolved nondisease–associated genes (Cai et al. 2009). We also find that disease-associated genes tend to have lower mean nonsynonymous substitutions to synonymous substitutions ratio than nondisease-associated genes. However, the lack of data available after 390 Ma, let alone from Eutheria, prohibits any particularly meaningful biological conclusions on differences between age categories of disease genes. It is therefore equivocal whether or not disease genes evolve at a different rate to nondisease genes (Hsiao and Vitkup 2008; Cai et al. 2009).

In conclusion, our results demonstrate that the emergence of genes associated with disease has been relatively constant and ongoing throughout the evolution of life. Thus, the notion of fragility in evolutionary ancient genes is a misnomer. Rather, disease is inherent to biological systems as they are dependent on mutation for functional innovation. Interestingly, just as duplication is the key contributor to the evolution of function, duplicated genes contribute to more diseases than singletons and this has been relatively stable across most of evolutionary time. A unifying explanation for heritable disease, thus, is it is merely a by-product of the evolutionary process. That is, just as mutation contributes to functional innovation in concert with gene duplication, it will also lead to novel opportunities for new diseases to emerge.

Supplementary Material

Supplementary tables S1–S3 and figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgments

J.E.D. was supported by a Wellcome Trust studentship and VIP award. Thanks to Kathryn Hentges for helpful comments.

References

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* 322:881–888.

Benton MJ, Ayala FJ. 2003. Dating the tree of life. *Science* 300:1698–1700.
Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol*. 24:26–53.
Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol*. 7:R43.
Cai J, Borenstein E, Chen R, Petrov D. 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol Evol*. 1:131.
Conant G, Wolfe K. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*. 9:938–950.
Conrad B, Antonarakis SE. 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet*. 8:17–35.
Cooper D, Kehrer-Sawatzki H. 2011. Exploring the potential relevance of human-specific genes to complex disease. *Hum Genomics* 5: 99–107.
Domazet-Loso T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol*. 25:2699–2707.
Donoghue PCJ, Benton MJ. 2007. Rocks and clocks: calibrating the tree of life using fossils and molecules. *Trends Ecol Evol*. 22: 424–431.
Edgar R. 2004. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
Flicek P, Aken B, Ballester B, et al. (57 co-authors). 2009. Ensembl's 10th year. *Nucleic Acids Res* 38:D557–562.
Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease network. *Proc Natl Acad Sci U S A*. 104: 8685–8690.
Hakes L, Pinney J, Lovell S, Oliver S, Robertson D. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*. 8:R209.
Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 33: D514–D517.
Harris MA, Clark J, Ireland A, et al. (60 co-authors). 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 32:D258–261.
Hedges SB, Dudley J, Kumar S. 2006. Timetree: a public knowledgebase of divergence times among organisms. *Bioinformatics* 22: 2971–2972.
Hsiao T-L, Vitkup D. 2008. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet*. 4:e1000014.
Huang H, Winter EE, Wang H, et al. (12 co-authors). 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol*. 5:R47.
Kanehisa M, Goto S. 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28:27.
Kasahara M. 2007. The 2R hypothesis: an update. *Curr Opin Immunol*. 19:547–552.
Kimura M, Maruyama T, Crow J. 1963. The mutation load in small populations. *Genetics* 48:1303.
Li H, Coghlan A, Ruan J, et al. (15 co-authors). 2006. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 34:D572–D580.
Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 4:865–875.
López-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*. 32:3108–3114.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
Lynch M, Gabriel W. 1990. Mutation load and the survival of small populations. *Evolution* 44:1725–1737.
Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 107:9270.
Marques-Bonet T, Kidd J, Ventura M, et al. (20 co-authors). 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457:877–881.
Ohno S. 1970. Evolution by gene duplication. New York: Springer.
Reed F, Aquadro C. 2006. Mutation, selection and the future of human evolution. *Trends Genet*. 22:479–484.
Rice W. 2002. Experimental tests of the adaptive significance of sexual recombination. *Nat Rev Genet*. 3:241–251.
Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*. 23:23–35.
Smith NGC, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. *Gene* 318:169–175.
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 U S A*. 101:15398–15403.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. Ensemblcompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
Zhang F, Gu W, Hurles M, Lupski J. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 10:451–481.