Human Molecular Genetics, 2017, Vol. 26, No. 3 489–500

doi: 10.1093/hmg/ddw405
Advance Access Publication Date: 4 January 2017
Original Article

ORIGINAL ARTICLE

Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology

Nino Spataro[1], Juan Antonio Rodríguez[1], Arcadi Navarro[1,2,3,4,†] and Elena Bosch[1,†,*]

[1]Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain, [2]National Institute for Bioinformatics (INB), Barcelona Biomedical Research Park (PRBB), Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona Biomedical Research Park (PRBB), Barcelona, Spain and [4]Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona Biomedical Research Park (PRBB), Barcelona, Spain

[*]To whom correspondence should be addressed at: Elena Bosch, Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain. Tel: +34 93 3160841; Fax: +34 93 3960901; Email: elena.bosch@upf.edu

Abstract

Do genes presenting variation that has been linked to human disease have different biological properties than genes that have never been related to disease? What is the relationship between disease and fitness? Are the evolutionary pressures that affect genes linked to Mendelian diseases the same to those acting on genes whose variation contributes to complex disorders? The answers to these questions could shed light on the architecture of human genetic disorders and may have relevant implications when designing mapping strategies in future genetic studies. Here we show that, relative to non-disease genes, human disease (HD) genes have specific evolutionary profiles and protein network properties. Additionally, our results indicate that the mutation-selection balance renders an insufficient account of the evolutionary history of some HD genes and that adaptive selection could also contribute to shape their genetic architecture. Notably, several biological features of HD genes depend on the type of pathology (complex or Mendelian) with which they are related. For example, genes harbouring both causal variants for Mendelian disorders and risk factors for complex disease traits (Complex-Mendelian genes), tend to present higher functional relevance in the protein network and higher expression levels than genes associated only with complex disorders. Moreover, risk variants in Complex-Mendelian genes tend to present higher odds ratios than those on genes associated with the same complex disorders but with no link to Mendelian diseases. Taken together, our results suggest that genetic variation at genes linked to Mendelian disorders plays an important role in driving susceptibility to complex disease.

[†]These authors contributed equally.

Received: April 12, 2016. **Revised:** November 10, 2016. **Accepted:** November 23, 2016

Introduction

Improving our understanding of genetic mutations and potential genetic risk factors directly causing, or contributing to, human diseases is at the core of modern medical genetics. In that context, evolutionary theory can be used to characterize the selective forces that have acted on the causal and susceptibility alleles underlying human genetic disorders. Complex disorders are common in the general population and result from the interaction of several susceptibility loci and environmental factors. In contrast, Mendelian disorders are typically rare and have predictable inheritance patterns as they usually result from a single causative mutation in a gene. However, heterogeneity and incomplete penetrance suggest that the classical distinction between Mendelian and complex diseases is not always absolute and that a continuum exists between the purely Mendelian and the most complex diseases. Although several unclear situations are found, since many genes harbour mutations that are unequivocally linked to particular Mendelian diseases, one can still define a set of genes linked to Mendelian diseases. Recent advances in medical research have resulted in large catalogues of mutations causing Mendelian hereditary disorders and of susceptibility loci contributing to complex diseases, such as those of the Online Mendelian Inheritance in Man (OMIM) database (http://www.omim.org/) (1) and the Genome-Wide Association Studies (GWAS) Catalogue (https://www.ebi.ac.uk/gwas/) (2,3). This biomedical information, when coupled with the increasing public availability of whole genome sequences, can be used to elucidate the genetic architecture and natural history of human genetic diseases.

Alleles causing disease are generally thought to be introduced by random mutations and eventually eliminated by purifying selection, as this is the evolutionary force that keeps deleterious alleles at low frequencies (4). This simple mutation-selection balance model will not apply for diseases appearing after the reproductive age but is well suited for strong and highly penetrant mutations causing monogenic disorders early in life. Indeed, as predicted by the model, Mendelian diseases have high allelic heterogeneity and low prevalence (5). However, highly deleterious traits are not only the target of negative selection. Other models of selection do apply to the allele frequency dynamics of highly penetrant Mendelian alleles in particular populations, such as balancing selection in those causing sickle cell anaemia (6), *G6PD* deficiency (7) and thalassaemias in populations inhabiting geographical regions where malaria is endemic (8,9).

For several reasons, it is even more difficult to conceive a general evolutionary model for complex polygenic disease traits. Each single susceptibility allele contributes only a small fraction of the overall disease risk and thus it would be subject only to very weak purifying selection, which prevents its extinction and may allow its increase in frequency (10–12). Also, many complex diseases affect individuals only after their reproductive period and, thus, the underlying causal alleles could be invisible to the action of purifying selection (13,14). Furthermore, most complex disorders have low penetrance and a relevant fraction of their total heritability may be explained by gene-environment interactions (15,16) and epistatic mechanisms (17), making the understanding of their genetic architecture even more problematic (18). Moreover, some well-known disease susceptibilities result from adaptive scenarios, including heterozygote advantage (19), antagonistic pleiotropy (20), or from environmental shifts that revert an ancestral protective allele to a deleterious variant in modern conditions (21), as suggested by the 'thrifty genotypes' (22) and the 'sodium retention' hypotheses (23). Finally, demographic processes could also contribute to shape the genetic architecture of disease genes across ethnic groups (24,25), even if the main genetic variants underlying complex diseases are shared across continents (26). In particular, the several bottlenecks and founder events occurred in non-African populations could have increased the frequencies of weakly deleterious alleles (27,28), and the recent explosive population growth in humans could have also facilitated the introduction of multiple low frequency variants with possible deleterious effects (29–31). Overall, these observations suggest that, even if the mutation-selection balance model does fit the evolutionary history of most highly penetrant deleterious mutations in genes linked to Mendelian diseases, it is insufficient to explain the selective pressures acting on the full set of alleles related to diseases.

In this study, we set out to characterize several biological properties and the evolutionary forces acting on human disease (HD) genes. To that end, we first tested whether HD genes have different evolutionary and functional properties when compared to non-disease genes and to putatively essential human genes. Then, we proceed to explore whether different biological patterns emerge when comparing the three following non-overlapping subsets of HD genes: (i) genes linked only to Mendelian disease traits, (ii) genes uniquely associated with complex diseases and, (iii) genes linked to Mendelian disorders and also associated with complex diseases. Previous seminal works focused on the properties of susceptibility or causal SNPs (32–35) and on particular subsets of human diseases (32,33, 36–42). However, recent datasets (1–3,43–47) allow a more refined analysis of this subject. Here, we provide novel insights on the evolutionary pressures acting on both coding and regulatory regions of HD genes by taking advantage of the full genome resequencing data from the 1000 Genomes Project (48), of the recent advances in genome annotation of regulatory elements (49) and of the exhaustive information available after the GWAS era regarding susceptibility loci associated with complex diseases (2,3). In addition, dN/dS values, network properties, gene expression patterns across different tissues and other biological features were explored for each individual gene, compared across all gene sets and interpreted in the same integrated evolutionary framework. Whereas HD genes show a general pattern of functional constraint relative to non-disease genes, several biological features of HD genes mainly depend on the type of disease, Mendelian or complex, with which they are associated. Interestingly, even if only a fraction of genes linked to Mendelian disorders is also involved in complex diseases, our results suggest that variation within these genes may have a more important role in driving complex disease susceptibility than that of the remaining genes involved in the same diseases, which have never been linked to Mendelian disease. Overall, this study provides evidence for the existence of functional links between Mendelian and complex diseases.

Results

Properties of human disease (HD) genes

A total list of 3,275 unique protein coding genes related to human disease was obtained by merging (i) the whole list of human genes known to harbour causal variants for Mendelian disorders, available from a hand-curated version of the OMIM database (referred to as hOMIM) (50); with (ii) the list of genes with variants contributing to risk for complex diseases, available from the GWAS catalogue (Supplementary Material, Table S1). To obtain a first insight to the biological properties of HD genes as a subset, we compared their evolutionary and protein network features with those of two different sets of genes (for details, see Materials and Methods): (i) putatively essential genes, defined as orthologues of mouse essential genes detected by knock-out experiments (51) and not involved in any human disease (Essential Non-Disease, END, 1,572 genes), and (ii) the rest of genes in our genome neither associated with any human disorder nor found to be essential (Non-Disease and Non-Essential, NDNE, 13,135 genes). Details on p-values for all comparisons between gene groups regarding protein network properties, dN/dS values and neutrality tests can be found in Fig. 1 and in Supplementary Material, Supplementary Note 1. Protein network parameters, dN/dS values and neutrality statistics for each single gene can be found in Supplementary Material, Table S2. Henceforth, all reported observations will be statistically significant, unless specifically highlighted.

In agreement with the functional relevance of essential genes, END genes show the highest level of degree, are the most central in the network (*i.e.* display the higher closeness) and present the highest betweenness (Fig. 1A). Conversely, NDNE genes are clearly the less connected, less central and present lower importance for the information flow in the network (*i.e.* they present the lowest levels of betweenness). Thus, HD genes present intermediate values comprised between the extremes represented by END and NDNE genes for the three network properties (Fig. 1A). These results suggest that HD genes are a special subset of genes in our genome, since they need to be functionally relevant to be associated with a disease phenotype but not as much as the END genes, which, on average, occupy more important positions in the protein network.

Similar observations can be made from the analysis of rates of protein evolution (as measured by dN/dS, see Materials and Methods), which allows assessing the ancestral selective pressures acting on each group of genes. After considering the corresponding orthologous pairs between human-chimpanzee, human-macaque and human-mouse, END genes show the lowest rate of protein evolution over three different evolutionary scales. In contrast, NDNE genes show the highest dN/dS values, while the HD group displays intermediate rates of evolution comprised between the extremes of END and NDNE genes (Fig. 1B). Hence, even if HD genes are less constrained than the END subset, they are under stronger purifying selection than the remaining genes in our genome.

When exploring the genetic variation at intra-species level, END genes display the lowest Tajima's D values at both their coding and regulatory sequences and also marginally higher Fay and Wu's H values (Fig. 1C). The higher proportion of rare variants in the allele frequency spectra of END genes when compared to the remaining genes, as reflected in the Tajima's D distributions, probably indicates that END genes are still under stronger purifying selection. Conversely, HD genes show the highest Tajima's D values, indicating higher proportions of intermediate frequency variants at both sequence types. Moreover, while no differences on the Fay and Wu's H distributions were detected between HD and END genes, NDNE genes show the lowest values for that summary statistic, suggesting a higher amount of derived high frequency variants in both their coding and regulatory elements.

A potential source of bias in the analysis above can be due to the allele frequencies and the linkage disequilibrium (LD) properties of variants detected through association studies to be related to human diseases. Indeed, the susceptibility variants identified through GWAS tend to have relatively high frequencies because association tests have more power to identify high frequency variants and because genotyping arrays contain pre-ascertained SNPs with intermediate frequency alleles (52). Similarly, increased linkage disequilibrium is known to provide more power to detect associations over a larger region of the genome (53). Since these discovery biases could affect the site frequency spectrum of HD genes, we repeated the above comparisons taking into account the MAF (Minor Allele Frequency) of each associated variant as well as the LD between each associated variant and the reported genes (see details in Supplementary Material, Supplementary Note 2). HD genes in direct LD ($r^2 > 0.8$) with the associated variants described in the GWAS catalogue were clearly affected by this ascertainment bias in both their regulatory and exonic regions when compared to NDNE genes matched by MAF and LD. However, the regulatory regions of those genes reported in the GWAS catalogue that presented lower LD ($r^2 \leq 0.8$) with the corresponding associated variants, which represent 60% of the GWAS reported genes, still displayed higher Tajima's D values when compared to the remaining NDNE genes (Supplementary Material, Fig. SN2.2.1). Overall, these results suggest that HD genes are the target, not only of purifying selection, but also of other adaptive forces,

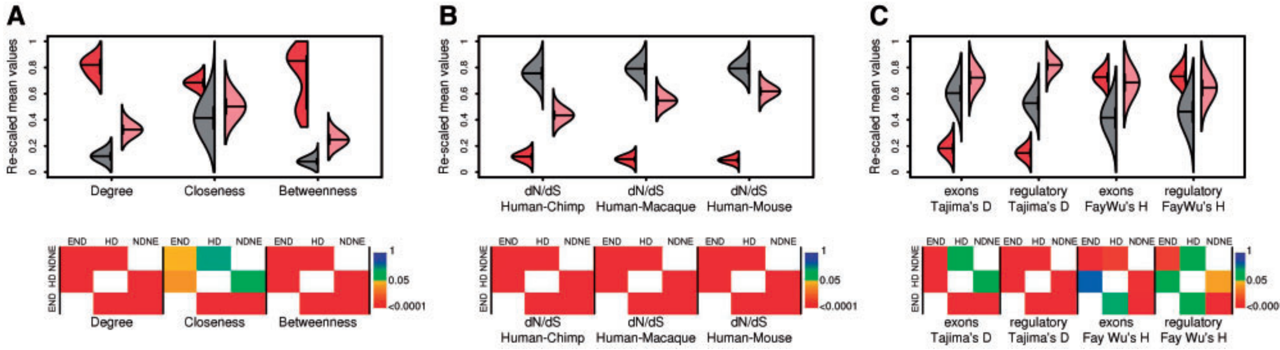

Figure 1. Protein network and evolutionary properties of human genes. Scaled resampled mean values and resampling p-values for three different protein network parameters (**A**), dN/dS (**B**), and Tajima's D and Fay and Wu's H (**C**). END genes are shown in dark red, HD genes in light red, NDNE genes in grey. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are represented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

particularly in their regulatory regions, where a higher amount of intermediate frequency variants is detected. Indeed, the Tajima's D distribution of the regulatory sequences of HD genes is clearly shifted toward positive values (Supplementary Material, Fig. SN2.2.3). Another potential confounding factor such as gene length does not have any relevant effect on the analyses performed and all the differences reported between HD and NDNE genes remain significant when considering only genes with a similar length (Supplementary Material, Supplementary Note 2).

We also compared the levels of gene expression among the three groups of genes by jointly considering the expression patterns previously described on 16 different human tissues (44,45). Expression data reveal that while END and NDNE genes are, respectively, the most and least expressed genes, HD genes show intermediate levels of expression (Fig. 2A). Additionally, END, HD, and NDNE genes tend to be expressed in decreasing numbers of tissues (p-value = 5.47×10^{-14} for END-HD, p-value = 5.05×10^{-56} for END-NDNE and p-value = 2.83×10^{-18} for HD-NDNE in a two-sided T-test, data not shown). When analysing separately the single tissues included in the Expression Atlas database, END genes are generally the most expressed in each single tissue, while NDNE genes tend to be the less expressed. However, HD genes are the most expressed in the liver, probably reflecting their important functional role in metabolism (Supplementary Material, Fig. S1A).

Next, we used the PhyloPat database (46) to explore the corresponding phylogenetic-based age of each gene and classify them in three predefined categories (see details in Materials and Methods). We found that most of the END genes originated before the appearance of vertebrates. In particular, END genes were found enriched in the "Old" but depleted in the "Young" categories. Conversely, HD genes were found over-represented in the "Intermediate" and "Young" categories, suggesting that most HD genes emerged more recently than the rest of the genes in the genome (Fig. 2B).

Finally, we also investigated whether these three groups of genes encode different biological functions (Fig. 2C). Among the protein categories listed in the PANTHER database (47), END genes show enrichment in nucleic acid binding and transcription factor proteins; so they could be involved in ancestral functions related to the interaction with the genetic material. While both END and HD genes are enriched in signalling molecules, enzymes, and transcription factors when compared to NDNE genes, six additional biological categories (*i.e.* receptors, transporters, transfers/carriers, cell junction, cell adhesion and extracellular matrix components) are also over-represented among HD genes. Notably, HD genes were also found specifically enriched in proteins involved in the immune and defence system relative to the rest of the genome (Fig. 2C). This suggests that HD genes could not only have a major structural and a cell communication role, but that they are especially important in our immune response.

Properties of disease gene subgroups

A detailed list of the genes in the HD set and the Mendelian and/or complex disorders with which they are related is provided in Supplementary Material, Table S1. Up to ~23% of the

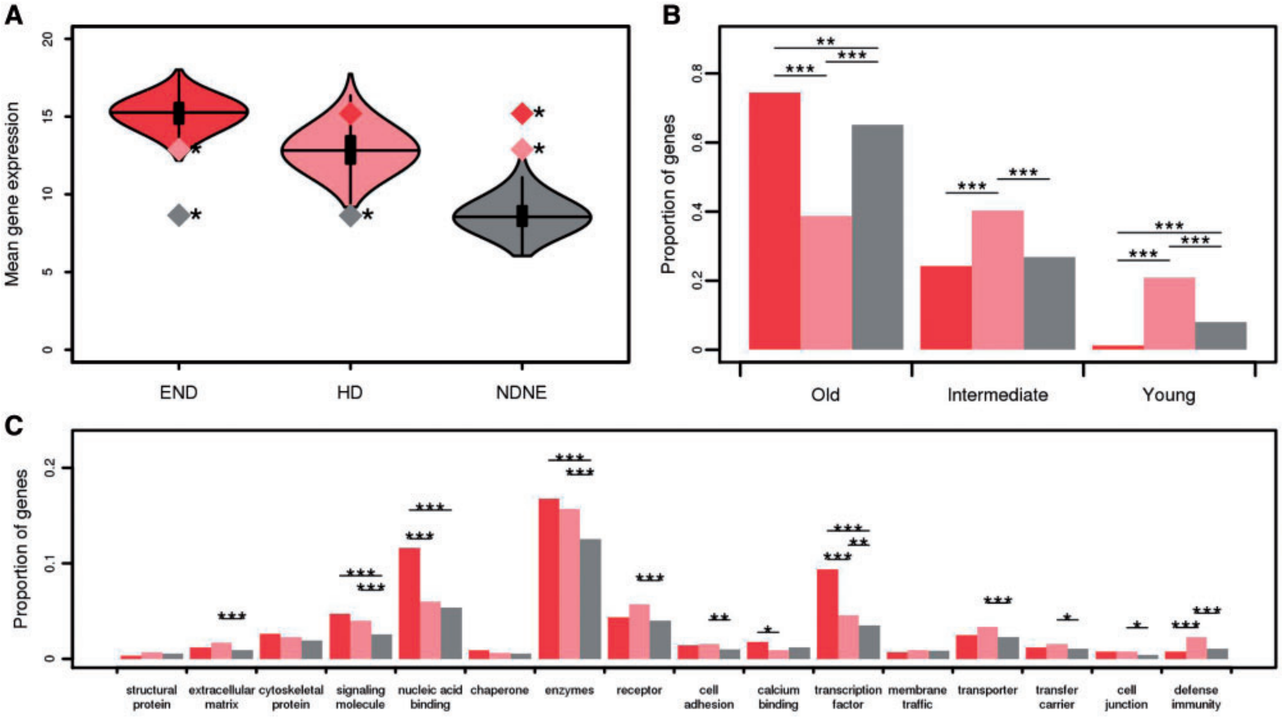

Figure 2. Biological features of human genes. (**A**) Resampling expression levels over 16 different human tissues reported in the Expression Atlas. END genes are shown in red, HD genes in light red, NDNE genes in grey. At each of the 10,000 resamplings, 1,000 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 1,000 genes was thus calculated and the 10,000 mean values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database. (C) Proportions of genes in different protein functions considered in PANTHER database. END genes are shown in red, HD genes in light red, NDNE genes in grey. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels respectively for a chi-square test (B and C).

genes that are linked to highly penetrant Mendelian diseases have been associated with at least a complex disorder. HD genes were thus divided in three different mutually excluding groups: (i) Complex-Mendelian (CM) genes, 203 genes found to be associated with both complex and Mendelian disorders; (ii) Mendelian Non-Complex (MNC) genes, 684 genes uniquely causing Mendelian disease traits; and (iii) Complex Non-Mendelian (CNM) genes, 2,388 genes uniquely associated with complex diseases. P-values for comparisons of different protein network properties, dN/dS values and neutrality tests among the three considered subgroups of HD genes can be found in Fig. 3 and in Supplementary Material, Supplementary Note 1. As above, all reported observations will be statistically significant, unless specifically highlighted.

When considering separately the three subgroups of HD genes, we found that CM genes show the highest levels of degree and are the most relevant for the information flow in the network (Fig. 3A). However, CM genes present no significant evolutionary differences from other HD genes, neither in their long-term protein evolution rates nor in their site frequency spectra (Fig. 3B and C). Even if not statistically significant, MNC genes show a trend towards lower dN/dS and Tajima's D when compared with genes found to be associated with at least one complex disease (CM and CNM genes), a pattern that could suggest stronger purifying selection acting on MNC genes.

Next, we compared the expression levels among the three HD gene subgroups using the Expression Atlas database (44,45). Whereas MNC genes are significantly more expressed than CNM genes, CM genes show intermediate expression levels compared to both MNC and CNM, even if no significance is reached with resampling tests (Fig. 4A). However, when performing a T-test to assess differences in gene expression among groups, CM genes were significantly more expressed than CNM genes (p-value = 8.19×10^{-3}), but less expressed than MNC genes (p-value = 3.82×10^{-4}). When considering the number of tissues where each single gene is expressed, we found that on average CNM genes are expressed in a higher number of tissues in comparison to genes linked only to Mendelian disorders (p-value = 0.0182 in a two-sided T-test). In turn, CM genes also seem to be expressed in a higher number of tissues compared to MNC but statistical significance is not reached (data not shown). Moreover, when comparing the three subgroups of HD genes separately for each single tissue in the Expression Atlas database, we found that MNC genes tend to be more expressed than CNM in at least seven different tissues (Supplementary Material, Fig. S1B). Interestingly, MNC genes are very highly expressed in liver; and so they can explain the high liver expression of HD genes as a whole (Supplementary Material, Fig. S1). This finding may reflect the involvement of MNC genes in many essential metabolic processes and it agrees with the enrichment of their encoded protein products in basic biological functions, including structural and cytoskeletal proteins, extracellular matrix components, enzymes, proteins involved in cellular communication such as receptors, transporters and transfers/carriers, and proteins involved in the immune and defence system. As MNC, CM genes also show enrichment in enzymes, extracellular matrix components, transporters and immune system proteins when compared to CNM. Thus, among those genes associated with at least one complex disease, CM genes seem to have more relevant functional roles (Fig. 4C).

Studying the phylogenetic-based ages obtained from the PhyloPat database (46) among the different subgroups of HD genes, we found that Mendelian genes tend to be older than those genes associated only with complex disorders. Specifically, MNC genes are enriched in the "Old" category, probably indicating that these genes encode most of the basal metabolic processes that appeared before the emergence of vertebrates. While CM genes are over-represented in the "Intermediate" category compared to the MNC group, genes associated only with complex diseases (the CNM subgroup) were found enriched in the "Intermediate" and "Young" categories compared to MNC. Thus, the genes involved in susceptibility to complex disorders seem to encode for functions that emerged more recently in the evolutionary scale (Fig. 4B).

Specific biological properties of CM genes

All results above indicate that CM genes are a very specific subgroup of HD genes: they are the most relevant HD genes in the protein network, they tend to present higher expression levels and they are enriched in specific relevant protein function categories when compared to CNM genes. To assess further differences between the two subgroups of genes associated with complex disease (CM and CNM), we compared the allelic Odds-Ratios (ORs) of all associations so far reported in these genes. For each gene associated with a given complex disease, the mean and the maximum OR of all the SNPs mapping close to or on that gene were obtained from all the GWAS studies that reported an association between such a gene and a given disease

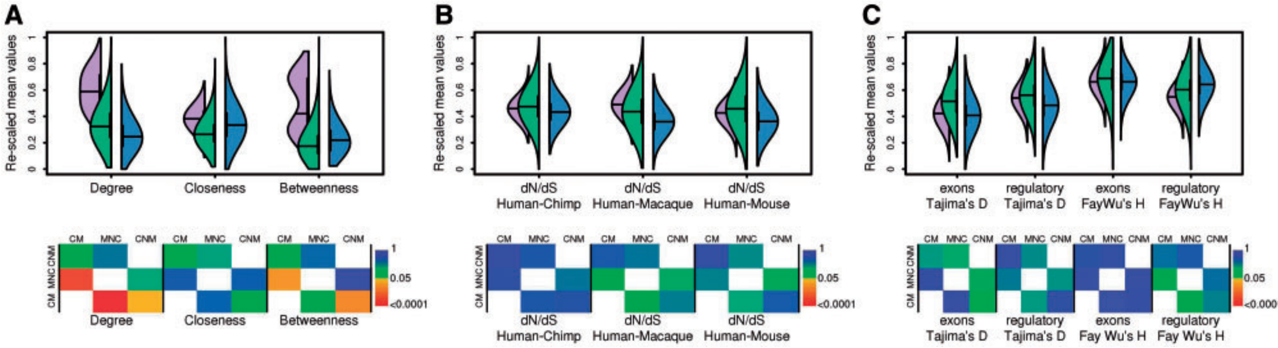

Figure 3. Protein network and evolutionary properties of human disease genes. Scaled resampled mean values and resampling p-values for three different protein network parameters (**A**), dN/dS (**B**), and Tajima's D and Fay and Wu's H (**C**). CM genes are shown in violet, MNC genes in blue, CNM genes in green. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are presented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

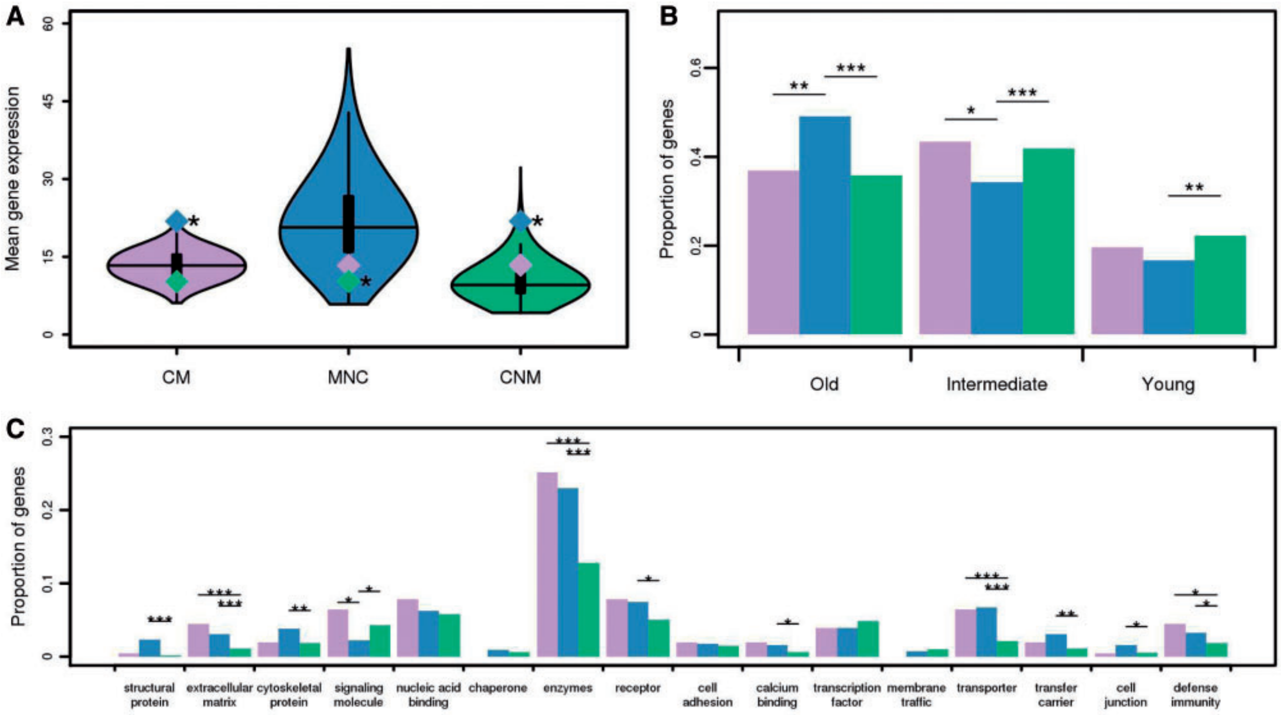

Figure 4. Biological features of different subgroups of human disease genes. (**A**) Resampling expression levels over 16 different human tissues reported in the Expression Atlas. CM genes are shown in violet, MNC genes in blue, CNM genes in green. At each of the 10,000 resampling, 100 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 100 genes was thus calculated and the 10,000 mean values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is found either on the top or bottom 2.5% of the resampling distribution. (**B**) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database. (**C**) Proportions of genes in different protein functions considered in PANTHER database. CM genes are shown in violet, MNC genes in blue, CNM genes in green. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels respectively for a chi-square test (B and C).

(Supplementary Material, Table S3). Interestingly, we found that CM genes tend to have higher ORs than those genes associated only with complex disorders, suggesting that susceptibility variants around CM genes have stronger effects on the complex phenotypes they affect (Fig. 5A, see details in Supplementary Material, Supplementary Note 3).

Moreover, out of the 101 different complex pathologies considered in our study, 71 diseases displayed at least one CM gene among those increasing risk for the complex trait. To further assess the relative role of CM and CNM genes in each single complex phenotype, we compared, for those 71 complex diseases, the mean percentile of the two gene groups within each disease trait for a wide range of statistics and different biological features (ORs, network parameters, expression levels, dN/dS, Tajima's D, Fay and Wu's H and age of genes). Interestingly, for the majority of complex disease traits, CM genes not only displayed higher ORs (Fig. 5C) but also higher relevance in the network than the CNM genes associated with the same trait (Supplementary Material, Fig. S2).

Taking advantage of a previous study where clinical records of over 110 million patients were mined (54), we can assess how many of the possible combinations of complex and Mendelian phenotypes involving CM genes are found to co-occur. Blair and collaborators considered a total of 65 complex and 96 Mendelian traits and found up to a total of ~3,000 combinations of Mendelian and complex diseases to effectively co-occur in patients. Unfortunately, only 36 out of the 71 complex traits that present at least one associated gene in the CM subgroup were considered in the analysis by Blair *et al.* (2013) (54). Similarly, only 139 out of the 329 Mendelian diseases that display at least one CM gene overlap with the Mendelian traits considered by Blair *et al.* (2013) (54). A table with all the correspondences among the traits used in our study and those used by Blair and collaborators can be found in Supplementary Material, Table S4. Out of the initial subgroup of 203 CM genes, only 54 genes were showing at least one Mendelian and one complex trait included in the list of traits of Blair *et al.* (2013) (54) and could be used to test whether their associated phenotypes were co-occurring. We observed that 38 out of the 54 CM genes that could be analyzed have at least one of the possible complex and Mendelian phenotypes combinations significantly co-occurring in patients. Moreover, for 19 out of the 36 complex phenotypes that could be investigated, there is at least one CM gene associated with a Mendelian disorder that was co-occurring with the complex phenotype. Finally, a total of 90 combinations of phenotype pairs involving CM genes were already observed by Blair *et al.* (2013) (54) to co-occur beyond random expectations (Supplementary Material, Table S5). On the whole, the co-occurrence of Mendelian and complex diseases in clinical records suggest that CM genes relevant for such traits are correctly identified by both linkage and association studies.

To gain additional insight into the properties of CM genes, we also investigated the corresponding age of onset for the complex and Mendelian phenotypes with which CM genes are associated. In general, the ages of onset of Mendelian and complex phenotypes are positively, even if not significantly, correlated. Indeed, for the Birth and Late categories represented in Fig. 5B we observed a mean age of onset for complex traits of 35.7 and 44.1 years, respectively. Thus, CM genes causal for late onset Mendelian disorders tend to be associated with late onset

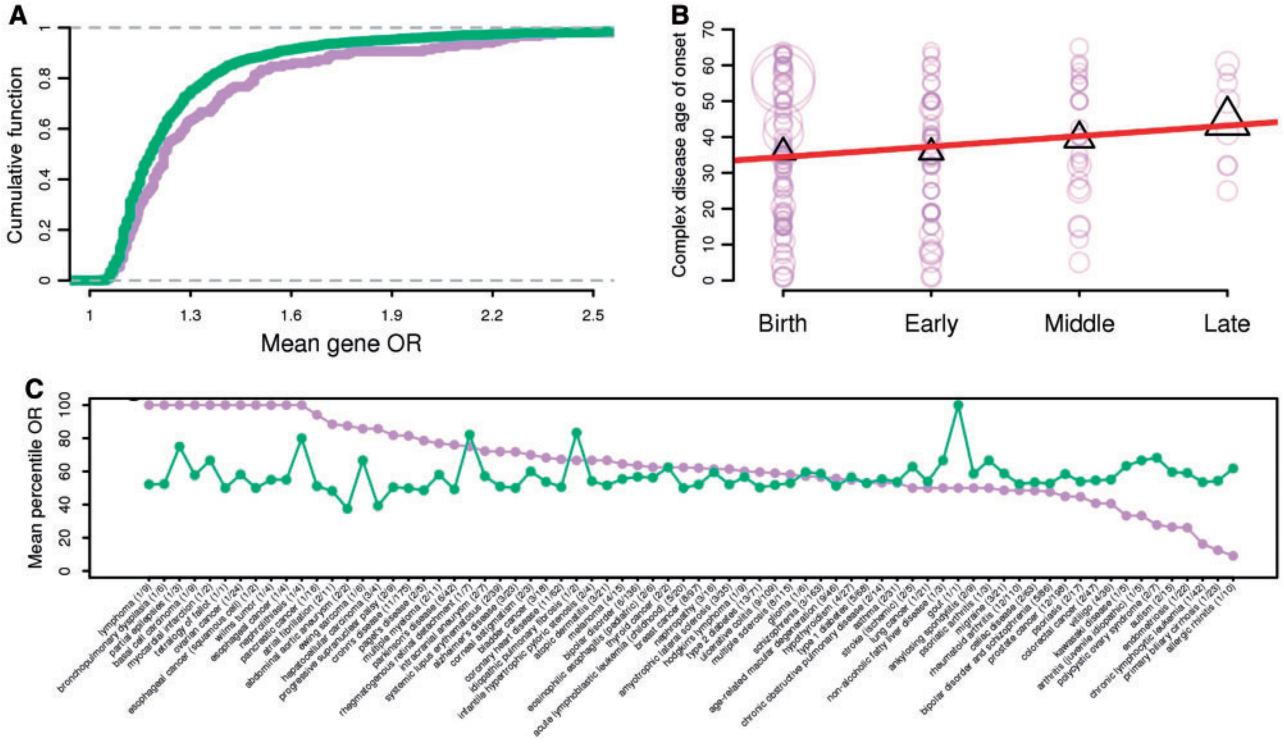

Figure 5. Odd-ratios of genes associated with complex diseases. (**A**) Comparison of mean OR for all the complex diseases CM (violet) and CNM (green) genes have been found to be associated with. (**B**) Correlation between the age of onset for the Mendelian disorders (x-axis) and the complex diseases (y-axis) CM genes have been found to be associated with. Circle sizes are proportional to the maximum odd-ratio observed for a given CM gene and the various complex diseases found associated with it. Black triangles represent the mean odd-ratio calculated over the full set of points of a given bin in the x-axis. The positions of the triangles in the y-axis represent the mean age of onset for all complex traits the CM genes of the bin have been found to be associated with. (**C**) Comparison of the mean odd-ratio percentiles of CM (violet) and CNM (green) genes associated with a given trait. For each single complex disease the number of CM and CNM genes associated with the trait is reported.

complex phenotypes. Additionally, CM genes associated with late onset Mendelian diseases showed slightly higher ORs than the rest of CM genes (*i.e.* Birth, Early and Middle bins in Fig. 5B considered jointly; p-value = 0.03 for two-sided Mann-Whitney U-test). This finding implies that the susceptibility variants on CM genes associated with late onset diseases show a trend towards higher effect sizes than those associated with early onset diseases, probably because they better evade the action of natural selection. Moreover, CM genes associated with late onset Mendelian diseases also display a trend towards higher rates of protein evolution (p-value = 0.05 for two-sided Mann-Whitney U-test; Supplementary Material, Fig. S3), indicating more relaxed long term purifying selection acting on the complete spectrum of their coding genetic variation.

Discussion

The surge in information technology that recently impacted genetics and biology provides an extraordinary opportunity for large-scale analysis and comparisons through computational methods.

Earlier studies investigated the properties of variants related to Mendelian disorders and suggested that causal mutations are generally not observed along large phylogenetic trees and occur in highly conserved regions. This pattern is reflected in the higher conservation rate and evolutionary age of Mendelian genes and it is consistent with the enrichment of Mendelian genes among those under strong purifying selection (35–37,55–57). Conflicting results were previously reported when comparing genes related to Mendelian disorders to genes increasing risk for complex diseases; some authors suggested higher evolutionary conservation in the former (50,55), which was not found by other authors (32,38,58). Similarly, contradictory observations were found when comparing the properties of genes linked to Mendelian disorders and genes related to complex diseases in the protein-protein interaction network, even if both groups of genes were significantly different from genes not involved in human pathologies (42,59).

The current availability of whole genome sequencing data, the enormous contribution of association studies to identify increasing risk factors for complex diseases and the several recently generated databases for human gene annotation provide a unique opportunity to collectively investigate the biological and evolutionary properties of genes related to human genetic diseases. We show that HD genes represent a specific subset of genes with special evolutionary and protein network properties that differ from those of genes that have never been related to human disease. In that sense, we report that HD genes are more conserved, more relevant in the protein-protein interaction network, quantitatively more expressed and expressed in a more diverse set of tissues than non-disease and non-essential genes (NDNE). However, none of these patterns is as extreme in the HD gene subset as it is for essential non-disease genes (END). In fact, a gradient of biological relevance seems to exist within the genome: while END genes represent an extremely relevant subset of genes within the original set of essential genes reported by Georgi *et al.* (51), putative essential genes that are also associated with human diseases show an intermediate functional role comprised between END and the rest of HD genes (Supplementary Material, Supplementary Note 4). Conversely, NDNE genes represent the human genes with the lowest functional relevance. It is tempting to speculate that END genes have never been associated with human disease because functionally relevant mutations on their sequences could have lethal consequences for the organism and result in miscarriage or early death; whereas, in contrast, HD genes need to be functionally relevant to be associated with a disease phenotype, but not as much as END genes.

Given the functional importance of HD genes in the protein network, we could expect that they should be the target of strong purifying selection. The observed dN/dS values suggest that HD genes are effectively under stronger long-lasting purifying selection than NDNE genes. These results agree with what was observed in previous studies (50,55,57), while contradict some others analyses (32,39,58) that were mainly conducted before the GWAS era and thus were only focused in a subset of the currently known human disease associated genes. Interestingly, Mendelian genes (MNC and CM) show similar closeness to END genes whereas for the remaining network properties, dN/dS values and neutrality statistics present the same pattern observed for the whole set of HD genes between END and NDNE genes (Supplementary Material, Supplementary Note 5). Notably, we also report unexpected higher Tajima's D values for HD genes, which reflect higher proportions of intermediate frequency variants relative to other gene sets. Even if we show that the discovery ascertainment bias of GWAS could have shifted site frequency spectra of a subset of HD genes, our results suggest that HD genes have also been the target of recent selective forces other than purifying selection especially in their regulatory elements. Moreover, the regulatory elements of Mendelian genes also show higher Tajima's D values when compared to END and NDNE genes, indicating that the enrichment of intermediate frequency variants is a shared pattern within the different subsets of HD genes (see Supplementary Note 5 for details). A previous survey detected a total of 336 different human genes showing signatures of balancing selection (60). We found an enrichment of HD genes among those genes under balancing selection ($p\text{-value} = 0.01$ for chi-square test for HD-NDNE), providing further support to the hypothesis that purifying selection is not the only evolutionary force acting on HD genes.

Both inter-species and intra-species variation suggest that all subsets of HD genes are under similar evolutionary pressures but several relevant biological features were found to depend on the type of phenotypes they are associated with. Previously, Blekhman *et al.* (50) analyzed genes within hOMIM according to their mode of inheritance and found that those genes following a dominant transmission pattern were more conserved and enriched of rare frequency variants, compared to recessive genes. These results are consistent with the hypothesis that dominant genes will be under stronger purifying selection. Accordingly, when extending this analysis here to the protein-protein interaction network data and to the expression profiles we also found higher functionality for dominant genes (see Supplementary Material, Figs. S4 and S5). The joint analysis of the CM and MNC genes reveals not only that the full set of Mendelian genes has similar properties to that of CNM genes in the protein network but that the two HD subgroups (CNM and Mendelian genes) evolve under similar evolutionary pressures at both intra-specific and inter-specific levels (Supplementary Material, Fig. S4). In contrast, as already highlighted by the separate analysis of CM and MNC, the full set of Mendelian genes is more expressed and enriched in the "Old" but depleted in the "Young" gene age categories when compared to CNM genes (Fig. 4 and Supplementary Material, Fig. S5).

The number and effect size of disease susceptibility alleles vary widely from disorder to disorder. Most rare, monogenic diseases with Mendelian inheritance are well explained by the mutation-selection balance model, under which deleterious variants are continuously introduced by random mutation in the population and subsequently removed by purifying selection. A wide range of human diseases are orders of magnitude more frequent than Mendelian diseases and the precise mechanisms maintaining variation on their susceptibility loci are still poorly understood. However, emerging observations suggest a more elusive boundary between complex and Mendelian human disorders and indicate that they are extremes in a continuum of genetic architectures (61). An increasing amount of evidence, including the present study, proves a role for genes linked to Mendelian diseases in the aetiology of complex disorders. In a previous work, we demonstrated that genes associated with Mendelian forms of Parkinson's disease are also functionally relevant for the complex form of the disease (62), suggesting that the two forms are genetically related. Here, after compiling 887 Mendelian genes, we observed that more than 23% of genes linked to a Mendelian disorder are also associated with at least one complex disease ($p\text{-value} = 2.39 \times 10^{-13}$ for chi-square test). Furthermore, the intersection between Mendelian and complex disease genes shows specific biological features, indicating a prominent role of CM genes in the aetiology of complex disease. Indeed, CM genes present higher functional importance in the protein network, higher ORs, a tendency towards higher expression levels and are enriched in relevant biological categories when compared to CNM genes.

Our study is limited by the accuracy of the genetic information currently available for human diseases as well as by our incomplete knowledge regarding the true susceptibility/causal variants and their corresponding genes. For instance, about half of all known Mendelian phenotypes still remain unsolved (63) and most of the complex disease associated variants are not fully characterized; in these cases the functional element harbouring the true causal variants has not been identified. Inevitably, the GWAS catalogue contains both false positives and false negatives. For each significant GWAS hit, a list of potential genes harbouring the causal variants is often reported based on the expertise of the authors about the biology of the disease. Thus, a fraction of the HD genes may be mis-assigned to the corresponding disease. We have assessed the robustness of our results to miss-assignment and have found that our results persist with up to a miss-assignment rate of 30–40%. By assuming that only one causal gene exists for any GWAS locus, we also checked that the unavoidable noise produced by the joint consideration of all the reported genes for any GWAS hit could not account for the observed differences. Finally, we also showed that all observed differences among gene groups remained when considering only SNPs associated at genome wide significance level ($p\text{-value} < 5 \times 10^{-8}$) (see details in Supplementary Material, Supplementary Note 2). Similarly, given that human essential genes have been ascertained from mouse essential genes detected by knock-out experiments, the putative human essential gene list could also contain both false positive and false negatives. Moreover, essential genes in mice could have been tested in a biased manner. Indeed, it is probable that genes of medical interest in humans are more likely to be investigated than the rest of the genome and thus tested for essentiality in animal models. This bias could produce an overrepresentation of well-known human disease genes, such as Mendelian genes, compared to the rest of genes related to diseases. Finally, publication bias could also affect the analysis of the protein-protein interaction network and the classification of biological functions since, indeed, genes related to human diseases are generally more investigated, resulting in the artificial inflation of the number of known protein-protein interactions and gene functional annotations for these genes. In spite of the limitations and the inevitable erroneous assignations present in the databases used, we believe that the available data provide a reliable global description and a good approximation to the human genes properties, contributing to a better understanding of the genetic architecture of human diseases.

Materials and Methods

Annotation datasets and gene groups considered

All protein-coding human genes from Ensembl were categorized in different non-overlapping groups according to the information of previously described public datasets to perform two levels of analyses. Initially, genes were categorized as Mendelian according to the information in the hOMIM database (a hand-curated version of the OMIM database that contains all genes that contribute to diseases with a simple genetic basis (50)), or as associated with complex diseases if they were described in the GWAS catalogue (2,3), which contains only variants showing a significance level $\leq 9\times10^{-6}$. We excluded infectious diseases and considered only association studies performed in populations of European ancestry, and binary disease traits, independently of the heritability levels of the complex diseases and the explained variance of the GWAS hits. For each trait, we analysed only the genes reported by the authors of each single association study in the GWAS catalog, but if a reported gene was not available, all the genes mapped by the database curators through an automated pipeline were gathered from the corresponding entry. Genes annotated with any of the 46 prenatal, perinatal or postnatal lethal phenotypes detected by knock-out experiments in mice were considered as mouse essential genes. Those human genes that were orthologues of mouse essential genes (51) were then annotated as putative essentials in humans.

In a first level of analysis, we distinguished among Human Disease (HD) genes, Essential Non-Disease (END) genes and the remaining Non-Disease Non-Essential (NDNE) genes. The HD gene set is the result of the direct merging of those genes in the hOMIM dataset (50) and the GWAS catalogue (2,3) (last accessed 18/10/2013); END genes result from removing any disease gene from the original group of genes defined as putative essentials in humans (since indeed disease genes have been directly proved to be linked or associated with disease); and, finally, NDNE genes were obtained after removing from the whole list of protein coding human genes those genes found in the disease or putative essential sets. In a second level of analysis, three subgroups of HD genes were considered: i) the Complex Non-Mendelian (CNM) subgroup is a subset of genes that comprises all genes present in the GWAS catalogue but not present in the hOMIM dataset; ii) the Mendelian Non-Complex (MNC) subgroup contains all those genes present in the hOMIM dataset but not in the GWAS catalog; and finally iii) the Complex-Mendelian (CM) subgroup comprises those genes found in both the GWAS catalogue and the hOMIM dataset and thus, it contains genes potentially associated with both complex and Mendelian disorders. For each complex disease trait, an age of onset was obtained either by considering the mean age of onset reported in the association study with the highest number of individuals and/or the information available in the Medscape database (http://www.medscape.com); for Mendelian genes, the age of onset was retrieved directly from the information provided in the hOMIM dataset (Supplementary Material, Table S3).

Evolutionary and protein network analysis

At intra-species level, evolutionary analysis was conducted differentiating coding and regulatory sequences. For each single gene, putative regulatory sequences were obtained extending 5 Kb the gene start and end coordinates and considering all the regulatory sequences retrieved from the Ensembl regulation release 78 (49) falling within the extended gene coordinates. Genomic coordinates were then converted to the hg19 assembly system using the Lift-Over Galaxy tool (64) and all the corresponding regulatory and coding regions of each gene in the human genome were downloaded in a VCF file from the 1000 Genomes Project (release 20110521) (48). All variants detected in the VCF files, including structural variants and indels, were considered in order to compute different summary statistics of neutrality in the CEU population of the 1000 Genomes project, which comprises individuals of northern and western European ancestry.

Tajima's D (65) and Fay and Wu's H (66) were calculated using a customized version of the sample_stats program (67) over both the exonic and the putative regulatory sequences of each gene. Given the correlation found between the obtained Fay and Wu's H value and the length of the fragments under analysis (Supplementary Material, Fig. S6A), we corrected the computed Fay and Wu's H statistic by dividing its value by the corresponding length of exonic or regulatory fragments. For all human-chimp, human-macaque and human-mouse orthologous genes available, values of dN/dS were extracted from the Ensembl Biomart Genes 78 (49) by considering only pairs of genes with one to one orthology relationships. Only genes with a dN/dS lower than 2 were considered for further analysis (as in (58)). Additionally, dN/dS values were corrected by the gene GC content, calculating the residuals from the correlation line obtained for dN/dS and GC content values (Supplementary Material, Fig. S6B), previously calculated using the UCSC galaxy and the EMBOSS tool GEECEE (68).

The human protein–protein interaction network (PIN) was reconstructed from the interactions available in the BioGRID database version 3.1.81 (43). Only non-redundant physical interactions were considered to calculate centrality measures. For each protein, degree was computed as the total number of interactions in which it is involved, while betweenness and closeness centralities were computed using the NetworkX Python library (69).

Expression data, ages of genes and protein functions

Gene expression data were downloaded from the Expression Atlas (version 0.1.4, E-MATB-513), which contains highly-curated and quality-checked RNA-seq experiments obtained over 16 different human tissues (44,45). Levels of expression were compared among the different groups and subgroups of genes considered across all tissues and within specific tissues. We also compared the mean number of tissues on which the genes are expressed and compared those mean numbers across groups of genes.

For each gene, phylogenetic-based ages were extracted from the PhyloPat database (version 51) (46), which contains phylogenetic trees for all human genes and thus it allows to infer when a given gene appeared in our past evolutionary history. Each corresponding gene age was thus deduced using the most distant species were the orthologous of the human gene was observed. Subsequently, genes were classified in three different age bins: i) "Old", from *Saccharomyces cerevisiae* to *Ciona savignyi*; ii) "Intermediate", from *Tetraodon nigroviridis* to *Gallus gallus*; and iii) "Young", from *Ornithorhynchus anatinus* to *Homo sapiens*. Given the non-normal distribution of gene ages in our dataset, differences between pairs of groups were assessed using a two-sided Mann-Whitney U-test. Enrichment in each bin of age was assessed using a chi-square test.

Protein functional class enrichment was performed using the categories defined in PANTHER database (47). From the whole list of categories, we excluded the "transmembrane receptor regulator", the "viral protein", the "surfactant" and the "storage proteins" classes, since less than 50 different human genes were present in them. Moreover, we obtained and used the "enzyme" class considering jointly the genes included in the "hydrolase", "isomerase", "kinase", "ligase", "lyase", "oxidoreductase", "phosphatase", "protease" and "transferase" classes. Enrichment in each functional category was assessed using a chi-square test.

Statistical analysis

Protein network parameters, summary statistics of neutrality and expression profiles were compared between gene groups using two-sided T-tests and Mann-Whitney U-tests (see details in Supplementary Material, Supplementary Note 1). Additionally, for each comparison, we also ran a resampling test, on which for each pair of groups compared we resampled 10,000 times a fixed number of genes in one of the groups (Figs 1 and 3 and Supplementary Material, Supplementary Note 1). At each resampling, the mean for a given statistic was calculated and the distribution of the 10,000 resampled means of a given group was compared to the observed mean value obtained in the non-resampled gene group of the pair under comparison. For comparisons involving END, HD and NDNE, 1,000 genes were sampled from each group at each resampling; for comparisons among CNM, MNC and CM, only 100 genes were sampled from each group at each resampling.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

The authors are grateful to Francesc Calafell for its statistical advice and David Hughes for its suggestions in data visualization and both David Hughes and Natàlia Borja Gutés for helping in manuscript editing. EB is the recipient of an ICREA Academia Award.

Conflict of Interest statement. None declared.

Funding

This work was supported by Ministerio de Ciencia e Innovación, Spain (SAF2011-29239 to EB and BFU2012-38236 to AN), by Direcció General de Recerca, Generalitat de Catalunya (2014SGR1311 and 2014SGR866), by the Spanish National Institute of Bioinfomatics of the Instituto de Salud Carlos III (PT13/0001/0026) and by FEDER (Fondo Europeo de Desarrollo Regional)/FSE (Fondo Social Europeo). Funding to pay the Open Access publication charges for this article was provided by the ICREA Award granted to EB by the Institució Catalana de Recerca i Estudis Avançats (Generalitat de Catalunya).

References

1. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, 514–517.
2. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A*, **33**, 9362–9367.
3. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, 1001–1006.
4. Haldane, J.B.S. (1937) The Effect of Variation of Fitness. *Am. Nat*., **71**, 337–349.
5. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet*., **17**, 502–510.
6. Allison, A.C. (1954) Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *Br. Med. J*., **1**, 290–294.
7. Verrelli, B.C., McDonald, J.H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A.N., Loiselet, J. and Tishkoff, S. a. (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J. Hum. Genet*., **71**, 1112–1128.
8. Oner, C., Dimovski, A.J., Olivieri, N.F., Schiliro, G., Codrington, J.F., Fattoum, S., Adekile, A.D., Oner, R., Yüregir, G.T. and Altay, C. (1992) Beta S haplotypes in various world populations. *Hum. Genet*., **89**, 99–104.
9. Hedrick, P.W. (2011) Selection and Mutation for β-Thalassemia in Nonmalarial and Malarial Environments. *Ann. Hum. Genet*., **75**, 468–474.
10. Eyre-walker, A. and Keightley, P.D. (2001) Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. *Mol. Biol. Evol*., **19**, 2142–2149.
11. Eyre-Walker, A. and Keightley, P.D. (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol*., **26**, 2097–2108.
12. Akashi, H., Osada, N. and Ohta, T. (2012) Weak selection and protein evolution. *Genetics*, **192**, 15–31.
13. Medawar, P.B. (1952) An unsolved problem of biology. *Evol. Health Dis*., **24**, 47–70.
14. Wright, A., Charlesworth, B., Rudan, I., Carothers, A. and Campbell, H. (2003) A polygenic basis for late-onset disease. *Trends Genet*., **19**, 97–106.
15. Khoury, M.J., Adams, M.J. and Flanders, W.D. (1988) An epidemiologic approach to ecogenetics. *Am. J. Hum. Genet*., **42**, 89–95.
16. Hunter, D.J. (2005) Gene-environment interactions in human diseases. *Nat. Rev. Genet*., **6**, 287–298.
17. Moore, J.H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, **56**, 73–82.
18. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M.,

Cardon, L.R., Chakravarti, A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

19. Fay, J.C. (2013) Disease consequences of human adaptation. *Appl. Transl. Genomics*, **2**, 42–47.
20. Carter, A.J. and Nguyen, A.Q. (2011) Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med. Genet.*, **12**, 160.
21. Dudley, J.T., Kim, Y., Liu, L., Markov, G.J., Gerold, K., Chen, R., Butte, A.J. and Kumar, S. (2012) Human genomic disease variants: A neutral evolutionary explanation. *Genome Res.*, **22**, 1383–1394.
22. Neel, J.V. (1962) Diabetes Mellitus: A 'Thrifty' Genotype Rendered Detrimental by 'Progress'?. *Am. J. Hum. Genet.*, **14**, 353–362.
23. Wilson, T. (1986) History of salt supplies in West Africa and blood pressures today. *Lancet*, **327**, 784–786.
24. Marigorta, U.M., Lao, O., Casals, F., Calafell, F., Morcillo-Suárez, C., Faria, R., Bosch, E., Serra, F., Bertranpetit, J., Dopazo, H., *et al.* (2011) Recent human evolution has shaped geographical differences in susceptibility to disease. *BMC Genomics*, **12**, 55.
25. Hodgkinson, A., Casals, F., Idaghdour, Y., Grenier, J., Hernandez, R.D. and Awadalla, P. (2013) Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC Genomics*, **14**, 1.
26. Marigorta, U.M. and Navarro, A. (2013) High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLoS Genet.*, **9**, e1003566.
27. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., *et al.* (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature*, **451**, 994–997.
28. Lohmueller, K. (2014) The distribution of deleterious genetic variation in human populations. *bioRxiv*, **29**, 0–20.
29. Casals, F. and Bertranpetit, J. (2012) Human Genetic Variation, Shared and Private. *Science*, **337**, 39–40.
30. Keinan, A. and Clark, A.G. (2012) Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, **336**, 740–743.
31. Gao, F. and Keinan, A. (2014) High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics*, **15 Suppl 4**, S3.
32. Thomas, P.D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U. S. A*, **101**, 15398–15403.
33. Dudley, J.T., Chen, R., Sanderford, M., Butte, A.J. and Kumar, S. (2012) Evolutionary meta-analysis of association studies reveals ancient constraints affecting disease marker discovery. *Mol. Biol. Evol.*, **29**, 2087–2094.
34. Jin, W., Qin, P., Lou, H., Jin, L. and Xu, S. (2012) A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum. Mol. Genet.*, **21**, 1611–1624.
35. Kumar, S., Dudley, J.T., Filipski, A. and Liu, L. (2011) Phylomedicine: An evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet.*, **27**, 377–386.
36. Miller, M.P. and Kumar, S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum.Mol.Genet.*, **10**, 2319–2328.
37. Subramanian, S. and Kumar, S. (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics*, **7**, 306.
38. Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T. and Sun, F. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 31.
39. Huang, H., Winter, E.E., Wang, H., Weinstock, K.G., Xing, H., Goodstadt, L., Stenson, P.D., Cooper, D.N., Smith, D., Albà, M.M., *et al.* (2004) Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol.*, **5**, R47.
40. Kryukov, G.V., Pennacchio, L.A. and Sunyaev, S.R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.
41. Podder, S. and Ghosh, T.C. (2010) Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Mol. Biol. Evol.*, **27**, 934–941.
42. Cai, J.J., Borenstein, E., Chen, R. and Petrov, D.A. (2010) Similarly Strong Purifying Selection Acts on Human Disease Genes of All Evolutionary Ages. *Genome Biol. Evol.*, **1**, 131–144.
43. Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., *et al.* (2010) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
44. Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N.A., Gonzalez-Porta, M., Hastings, E., Huber, W., Jupp, S., Keays, M., Kryvych, N., *et al.* (2014) Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, 926–932.
45. Fonseca, N.A., Marioni, J. and Brazma, A. (2014) RNA-Seq Gene Profiling - A Systematic Empirical Comparison. *PLoS One*, **9**, e107026.
46. Hulsen, T., Groenen, P.M.A., de Vlieg, J. and Alkema, W. (2009) PhyloPat: An updated version of the phylogenetic pattern database contains gene neighborhood. *Nucleic Acids Res.*, **37**, 731–737.
47. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2009) PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, 204–210.
48. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **135**, 0–9.
49. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* (2014) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
50. Blekhman, B., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M. and Przeworski, M. (2009) Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.*, **18**, 883–889.
51. Georgi, B., Voight, B.F. and Buc, M. (2013) From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.*, **9**, e1003484.
52. Lachance, J. and Tishkoff, S.A. (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, **35**, 780–786.
53. Hong, E.P. and Park, J.W. (2012) Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inf.*, **10**, 117–122.
54. Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam,

E.V., Brunak, S., *et al*. (2013) A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell*, **155**, 70–80.

55. Kondrashov, F.A., Ogurtsov, A.Y. and Kondrashov, A.S. (2004) Bioinformatical assay of human gene morbidity. *Nucleic Acids Res*., **32**, 1731–1737.
56. Lopez-Bigas, N. and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*., **32**, 3108–3114.
57. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., *et al*. (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
58. Smith, N.G.C. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions. *Gene*, **318**, 169–175.
59. Barrenas, F., Chavali, S., Holme, P., Mobini, R. and Benson, M. (2009) Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS One*, **4**, 2–7.
60. Leffler, E.M., Pfeifer, S., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J.D., Sella, G., Donnelly, P. and Mcvean, G. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, **339**, 1578–1582.
61. Rodriguez, J.A., Marigorta, U. and Navarro, A. (2014) Integrating genomics into evolutionary medicine. *Curr. Opin. Genet. Dev*., **29**, 97–102.
62. Spataro, N., Calafell, F., Cervera-Carles, L., Casals, F., Pagonabarraga, J., Pascual-Sedano, B., Campolongo, A., Kulisevsky, J., Lleo, A., Navarro, A., *et al*. (2014) Mendelian genes for Parkinson's disease contribute to the sporadic forms of the disease. *Hum. Mol. Genet*., **24**, 2023–2034.
63. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., *et al*. (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet*., **97**, 199–215.
64. Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*., **11**, R86.
65. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
66. Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
67. Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
68. Rice, P. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Curr. Opin. Colloid Interface Sci*., **14**, 126–134.
69. Hagberg, A.A., Schult, D.A. and Swart, P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. *Network*, **836**, 11–15.