Elucidating the Genotype–Phenotype Relationships and Network Perturbations of Human Shared and Specific Disease Genes from an Evolutionary Perspective

Tina Begum and Tapash Chandra Ghosh*

Bioinformatics Centre, Bose Institute, Kolkata, West Bengal, India

*Corresponding author: E-mail: tapash@jcbose.ac.in, tapash_2000@yahoo.com.

Accepted: September 25, 2014

Abstract

To date, numerous studies have been attempted to determine the extent of variation in evolutionary rates between human disease and nondisease (ND) genes. In our present study, we have considered human autosomal monogenic (Mendelian) disease genes, which were classified into two groups according to the number of phenotypic defects, that is, specific disease (SPD) gene (one gene: one defect) and shared disease (SHD) gene (one gene: multiple defects). Here, we have compared the evolutionary rates of these two groups of genes, that is, SPD genes and SHD genes with respect to ND genes. We observed that the average evolutionary rates are slow in SHD group, intermediate in SPD group, and fast in ND group. Group-to-group evolutionary rate differences remain statistically significant regardless of their gene expression levels and number of defects. We demonstrated that disease genes are under strong selective constraint if they emerge through edgetic perturbation or drug-induced perturbation of the interactome network, show tissue-restricted expression, and are involved in transmembrane transport. Among all the factors, our regression analyses interestingly suggest the independent effects of 1) drug-induced perturbation and 2) the interaction term of expression breadth and transmembrane transport on protein evolutionary rates. We reasoned that the drug-induced network disruption is a combination of several edgetic perturbations and, thus, has more severe effect on gene phenotypes.

Key words: shared disease gene, specific disease gene, edgetic perturbation, drug-induced perturbation, protein evolutionary rates.

Introduction

The comparative genomics era has provided great opportunities for deciphering and comparing the genes, mutated to cause human genetic diseases, with respect to nondisease (ND) genes. Understanding the mutational basis of such human diseases helps to identify the candidate disease genes, which are still unexplored and thus benefit therapeutic drug designing (Goh et al. 2007; Barabasi et al. 2011). In this context, estimation of protein sequence evolutionary rate (the ratio of nonsynonymous nucleotide substitutions rate d*N* to synonymous substitutions rate d*S*) is of immense importance (Yang and Nielsen 2000). However, the signatures of protein evolution in human disease genes remain a controversial issue till date. In 2003, Smith and Eyre-Walker (2003) first demonstrated that disease genes evolve higher than ND genes. In 2004, Lopez-Bigas and Ouzounis (2004) noticed that disease genes have larger conservation scores compared with ND genes. Interestingly, in the same year, Huang et al. (2004) established that evolutionary rates do not vary between disease and ND genes. After dividing human disease genes into two groups according to 1) Mendelian disease phenotypes and 2) complex disease phenotypes, Blekhman et al. (2008) claimed that evolutionary rates are slow in Mendelian disease genes, intermediate in ND genes, and fast in complex disease genes. In 2009, Cai et al. (2009) confirmed that stronger purifying selection acts on human disease genes, and therefore, both (Mendelian and complex) disease genes are conserved than ND genes. Later on, considering monogenic (Mendelian) and polygenic (complex) disease genes, Podder and Ghosh (2010) found that ND genes are the most conserved group, whereas polygenic disease genes are the least conserved among the three gene sets. Although all the previous studies have emphasized on human disease genes classified based on the number of genes involved in a particular phenotype (Mendelian/complex), it will be interesting to analyze disease genes classified based on the number of phenotypes they are involved in. Hence, for the first time, we aim to investigate the evolutionary rates of human monogenic (Mendelian) disease genes after classifying them according to their number of phenotypic defects. Our study in Mendelian disease genes will also provide insights into the similar events in complex disease genes as approximately 54% of the Mendelian disease genes are also involved in complex diseases (Jin et al. 2012).

Pleiotropy, a phenomenon in the perspective of human diseasome, refers to diverse pathological effects of different mutations in a single gene causing distinct disorders in an individual (Chavali et al. 2010). Such type of gene can be termed as “shared disease (SHD)” gene (Chavali et al. 2010). Conversely, a gene can be termed as “specific disease (SPD)” gene, if it is associated with a single phenotypic defect originated from a single genetic mutation (Chavali et al. 2010). Here, the focus of this study is to know whether selective constraints differ among SHD, SPD, and ND genes, specifically within similar (monogenic) disease class. Describing the pace of such evolutionary changes is essential to understand how phenotypically heterogeneous diseases are generated and maintained in nature. If any difference in the rate of evolution is observed, it is likely that the number of phenotypic defects is playing the lead role. A number of additional evolutionary rate determinants including gene expression level, tissue expression breadth, and gene functionality may also influence our analyses (Bloom et al. 2006; Drummond et al. 2006; Pal et al. 2006; Begum and Ghosh 2010; Park and Choi 2010; Chakraborty and Ghosh 2013). Hence, it is essential to include all these features in our comparative study.

Genes and their products function as components of complex networks of macromolecules, which are linked through biochemical or physical interactions. They are often represented in “interactome” network models as “nodes" (vertices) and “edges" (links) (Zhong et al. 2009). Such network-centered approach is progressively used to interpret several pathogenic mechanisms of disease genes (Zhong et al. 2009; Wang et al. 2012). In their comprehensive study, Zhong et al. (2009) have identified distinct mutations those result in different defects/diseases due to specific loss or gain of edge(s) (edgetic perturbation model) or complete loss of gene product (node removal model) in the interactome network. Recently, using drug-targeted network, Wang, Thijssen, et al. (2013) have demonstrated that single-interface targets are more likely to generate side effects due to disruption of their only interaction interface by a drug (drug-perturbed model). However, to date, there is no systematic assessment of the evolutionary history of human diseases emerged from distinct network perturbations. We thus intend to study the evolutionary patterns of SHD and SPD genes through the lens of network perturbation models.

Our in silico study here demonstrates that the rates of protein evolution significantly differ among SHD, SPD, and ND genes irrespective of their gene expression levels. Although the disease classification depends on the number of phenotypic defects, we observed that the number of phenotypic defects has no effect on the mutational rate heterogeneity among our gene sets. We interestingly obtained that network perturbations, tissue expression breadth, and gene functionality have substantial contributions to the evolutionary rate variations of SHD, SPD, and ND genes. Further in-depth investigations on network perturbation processes revealed that drug-induced perturbation has major impact on protein evolutionary rates than edgetic perturbations.

Materials and Methods

Compiling Evolutionary Rate Information of Human Autosomal Disease/ND Genes

The study began with a list of total 4,419 hereditary disease genes (3,911 autosomal) from “morbid map” cataloged in the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al. 2009). Following Chavali et al. (2010), we distinguished disease genes according to the number of associated phenotypic errors and identified them as SPD gene and pleiotropic/SHD gene. Genes do not have any annotation in OMIM or Human Gene Mutation Database (HGMD Professional v.2011.4) (Stenson et al. 2009) or Genetic Association Database (Becker et al. 2004) were utilized as ND genes for our investigation. Evolutionary rate (dN/dS) data were retrieved from Ensembl v.74 (Flicek et al. 2013) through the BioMart interface using human–mouse orthologous pairs (>60% sequence identity) with $dS < 3$ (Tang and Epstein 2007) to avoid problems due to mutation saturation and higher estimation error. Human–mouse pair was chosen because both of them 1) are placental mammals, 2) share many common anatomical features and physiological processes, 3) demonstrate conservation of organ-specific expression, 4) have similar gene expression profiles, and 5) are functionally more similar (Liao and Zhang 2006, 2008; Gharib and Robinson-Rechavi 2011). Moreover, the regular use of mouse as a model organism to understand human biology and disease practically supports such orthology selection (Liao and Zhang 2008; Cai et al. 2009). Finally, the filtered data set of SHD ($n = 528$), SPD ($n = 1,257$), and ND ($n = 8,783$) (supplementary table S1, Supplementary Material online) genes were used for subsequent comparative analyses.

Identification of “Edgetically Perturbed” Proteins

Binary protein–protein interaction data were collected from the Human Protein Reference Database (HPRD Release 9) (Prasad et al. 2009). Identification of damaging in-frame mutations (missense mutations and in-frame insertions/deletions/indels) and deleterious single-nucleotide polymorphisms (SNPs) were achieved by using HGMD (Stenson et al. 2009) and sorting intolerant from tolerant (SIFT) prediction tool (Sim et al. 2012), respectively. Combination of such physical interaction data (at least two interaction partners) and deleterious in-frame mutations (at least one) for proteins helped us to identify proteins those are subject to edgetic perturbations (Zhong et al. 2009). By this way, we identified 267 SHD, 371 SPD, and 908 ND proteins associated with edgetic perturbations ("edgetically perturbed" proteins). For reassessing our result, we considered Protein Data Bank (http://www.rcsb.org/pdb, last accessed October 10, 2014) for structural data. We obtained 169 SHD, 209 SPD, and 340 ND proteins associated with edgetic perturbations for which structural data are available. Instead of relying on a single algorithm for deleterious mutation prediction (SIFT), we additionally checked the consistency of our result using an integrative database dbNSFP v.2.5 (Liu et al. 2013). We used the logistic regression-based deleterious mutation prediction data of dbNSFP, which includes the scores of ten different tools (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy, and PhyloP). In addition to the above criteria, we considered CADD score > 25 to identify deleterious mutations in a protein (Kircher et al. 2014; Oliver et al. 2014). Thus, we collected 298 SHD, 487 SPD, and 1,427 ND edgetic perturbation-associated proteins for our study.

Records of Protein Domains

Pfam domains employed in this study were obtained using Ensembl v.74 (Panda et al. 2012; Wang et al. 2012; Flicek et al. 2013). For interacting domain pairs, we used DIMA (Domain Interaction MAp) database v.3.0, which integrates 5,807 structurally known interactions imported from the reliable iPfam and 3did databases (Luo et al. 2011; Finn et al. 2014; Mosca et al. 2014). Proteins supplementary data of Kim et al. (2006) was used to obtain structurally resolved singlish interaction interface hubs (≥ 5 interaction partners + at most 2 interaction interfaces).

Analyses of Gene Expression Data

mRNA-seq data were retrieved from http://genes.mit.edu/burgelab/mrna-seq/ (last accessed October 10, 2014), which contains transcriptional data of 22 human tissue or cell-line samples and applied reads per kilobase of transcript per million algorithm to evaluate gene expression levels (Wang et al. 2008; Huang et al. 2013). For our study, a gene is defined as expressed in a tissue if its expression value is larger than $M + 2 \times \text{MAD}$, where M and MAD are determined by $M = \text{median}(x)$; $\text{MAD} = \text{median}(|x - M|)$ and x indicates the average expression values for the corresponding gene among all tissues (Huang et al. 2013). For each gene, we then summed up the number of over expressed tissues to compute tissue expression breadth.

Functional Categorization

For functional labeling, we used the GO biological processes of Ensembl database (He and Zhang 2006; Razeto-Barry et al. 2011); (Flicek et al. 2013). Following Lopez-Bigas et al. (2008), we subdivided gene biological functions into two categories: 1) conserved core processes and 2) less conserved regulatory processes. In addition, we considered cell adhesion, cell division, cell communication, phosphorylation, and developmental processes under regulatory processes (Beck et al. 2011). However, DNA replication, transcription, and translation-related processes are considered under core biological processes (Beck et al. 2011). For functional enrichment test, we considered nonredundant GO annotation-based GOrilla tool (http://cbl-gorilla.cs.technion.ac.il, last accessed October 10, 2014), which calculates exact *P* values by implementing a hypergeometric model and is widely used due to its fast running time and rigorous statistical analysis (Eden et al. 2009).

Collection of Drug-Related Data

DrugBank v.4.0 (Law et al. 2014) was used to retrieve drugs those have at least one known human protein target. To evaluate the side effects of the drugs, we considered SIDER2 (http://sideeffects.embl.de/, last accessed October 10, 2014) database (Kuhn et al. 2013). Thereby, we compiled a list of 996 drugs associated with 4,192 side effects. However, a target protein may have association with drugs with different number of side effects. Hence, we considered a protein as side-effect-associated protein if it is targeted by drug(s) known to have at least one side effect.

Gene Length, Protein Length, and Gene Recombination Rate Estimation

Gene length (l) and protein length were calculated using Ensembl v.75 (Flicek et al. 2013). Chromosome-wise high-resolution recombination rates were downloaded from the International HAPMAP Consortium website (International HapMap Consortium 2005). For a gene, we collected recombination rates at base position i (ρi) and computed the recombination rates across the genic regions using the formula: $(\Sigma \rho i)/l$ (Kato et al. 2008). In our gene set, we obtained recombination rate data of 435 SHD, 1,061 SPD, and 7,000 ND genes.

Analyses of Human Gene Paralogs and Protein Complex Data

We retrieved human paralogous gene pairs from Ensembl v.73 database (Flicek et al. 2013) and estimated the number of paralogs for each gene of our data set. We found paralog data of 426 SHD, 933 SPD, and 5,920 ND genes. Human protein complex data were collected from CORUM database (Ruepp et al. 2010). We define protein complex number as the count of complexes in which a particular protein participates (Das et al. 2013). Thus, we acquired complex association number for 1,090 individual genes of our data set.

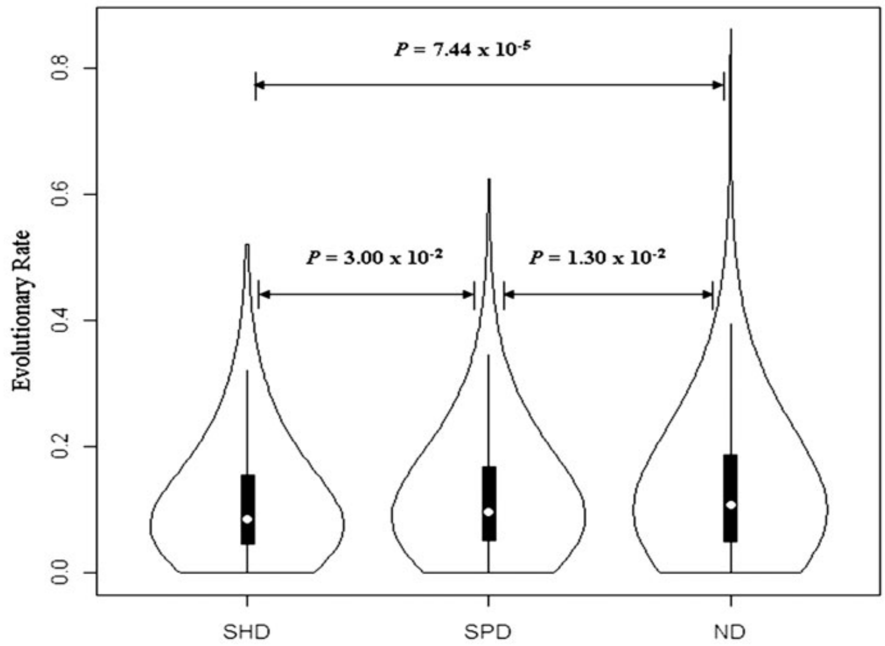

FIG. 1.—Evolutionary rate analyses of SHD, SPD, and ND genes. The violin plot depicts the relationships between protein evolutionary rates and number of phenotypic defects.

Statistical Analyses

All statistical analyses except analysis of covariance (ANCOVA) were achieved by using SPSS v.13. Throughout the study, we used nonparametric Spearman rank correlation coefficient. To calculate difference between two data sets, Mann–Whitney *U* test (MWT) was used unless mentioned otherwise. For ANCOVA analysis, considering the pair-wise interaction terms, XLSTAT 2009 was used. To generate violin plot, we used R package v. 2.13.1 (http://www.r-project.org, last accessed October 10, 2014).

Results

Evolutionary Rates of Shared, Specific, and ND Genes

The extent to which mutational changes have impacted human monogenic diseases with single (SPD) or multiple (SHD) phenotypic defect(s) can be traced by estimating the evolutionary rates of human SHD (defect number ≥ 2) and SPD (defect number = 1) genes considering ND (defect number = 0) genes as the control set. We thus observed that evolutionary rates are lower in SHD (average $dN/dS = 0.112$, $n = 528$), intermediate in SPD (average $dN/dS = 0.121$, $n = 1{,}257$), and higher in ND (average $dN/dS = 0.132$, $n = 8{,}783$) genes (Kruskal–Wallis test: $P = 3.57 \times 10^{-5}$) (fig. 1). However, the variation in evolutionary rates may be due to sample size biasness. To simplify, we pooled all the genes into three bins of equal sample size according to their evolutionary rates (bin 1 [range 0.152–0.861]: fast evolving genes; bin 2 [range 0.065–0.152]: medium evolving genes; and bin 3 [range 0.000–0.065]: slow evolving genes). We thus noticed that the proportions of ND, SPD, and SHD genes are higher in bin 1 ($n = 3{,}523$), bin 2 ($n = 3{,}523$), and bin 3 ($n = 3{,}522$), respectively, compared with the other groups of genes (fig. 2). The above observation suggests that the variations in evolutionary rates among monogenic disease and ND groups are not because of sample size differences.

In our study, classification of human diseases is based on number of observable phenotypic defects. Hence, it is reasonable to assume that number of defects contribute significantly to the group-to-group evolutionary rate differences. To examine the same, we performed correlation analysis between the number of phenotypic defects and protein evolutionary rates. Correlation study ($_{\text{Number of defects}}\rho^{dN/dS} = -0.041$, $P = 2.27 \times 10^{-5}$, $n = 10{,}568$) thus revealed that phenotypic defects may have little influence on protein evolutionary rates. However, such a weak but significant correlation could also be due to sampling bias because a large set of ND genes (phenotypic defect = 0) were included in the correlation analysis. After excluding the ND genes, the correlation becomes statistically insignificant in all three previous classified bins (bin 1: $_{\text{Number of defects}}\rho^{dN/dS} = -0.002$, $P = 9.60 \times 10^{-1}$, $n = 508$; bin 2: $_{\text{Number of defects}}\rho^{dN/dS} = 0.006$, $P = 8.79 \times 10^{-1}$, $n = 643$; bin 3: $_{\text{Number of defects}}\rho^{dN/dS} = -0.032$, $P = 4.23 \times 10^{-1}$, $n = 634$). We, therefore, conclude that some other constraints than number of defects may better explain the variation in rates of protein evolution in SHD, SPD, and ND genes.

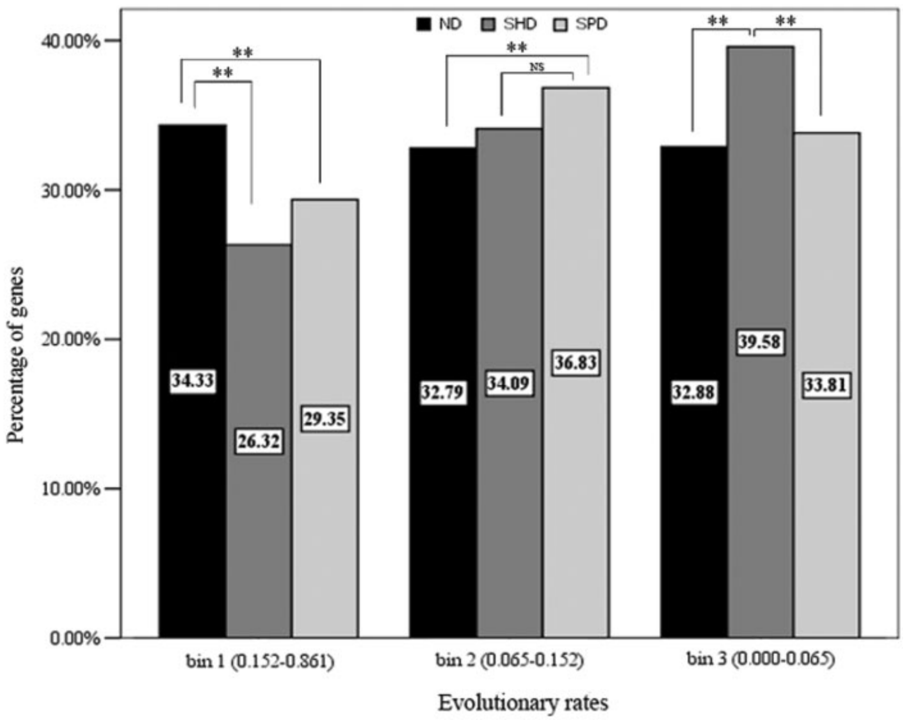

Fig. 2.—Distributions of ND, SHD, and SPD genes in different evolutionary rate bin. Genes are partitioned into three equal size bins as (I) fast evolving, (II) medium evolving, and (III) slow evolving groups. *P* values of MWTs between groups are mentioned. **Significant difference ($P < 0.01$) between groups. NS, nonsignificant difference ($P > 0.05$).

Edgetic Interactome Network Perturbation and Rates of Evolution of SHD, SPD, and ND Genes

Impact of a particular genetic abnormality is not limited to the activity of the gene product that carries it but can spread along the links of the interactome and alter the activity of gene products that do not carry any defects (Barabasi et al. 2011). Hence, it is quite reasonable that genes with more partners in the interactome tend to be involved in more diseases and phenotypes (Fraser et al. 2002; Jin et al. 2012). Our result also corroborates the same (considering ND genes: ${}_{\text{Number of interaction partners}}\rho^{\text{Number of defects}} = 0.214$, $P = 1.00 \times 10^{-6}$, $n = 4{,}739$; excluding ND genes: ${}_{\text{Number of interaction partners}}\rho^{\text{Number of defects}} = 0.164$, $P = 1.00 \times 10^{-6}$, $n = 1{,}310$). In the interactome, progression of diseases takes place by causing several molecular defects in proteins through node removal or via edge gain/loss (edgetic perturbation) (Care et al. 2009; et al. 2009). Because mutations associated with node removal are likely to generate incomplete fragments or nonfunctional gene products (Zhong et al. 2009; Wang et al. 2012), we emphasized only on edgetic perturbation (due to deleterious mutations) model for further investigations. By using SIFT algorithm (Sim et al. 2012) for predicting deleterious mutations, we found that the proportions of edgetically perturbed proteins are significantly higher in SHD (50.57%; 267/528), intermediate in SPD (29.51%; 371/1,257), and lower in ND group (10.34%; 908/8,783) (SHD vs. SPD: *Z* score = 8.471; SHD vs. ND: *Z* score = 27.037; SPD vs. ND: *Z* score = 19.073, respectively; $P < 1.00 \times 10^{-4}$ in all three cases). We reconfirmed our result by considering structural data for edgetically perturbed proteins of our data set (table 1). Although, SIFT algorithm is frequently used for its reported accuracy, it has some clear drawbacks (Lee et al. 2009; Sim et al. 2012). Hence, we considered dbNSFP database (Liu et al. 2013) to make more confident deleterious mutation predictions. When we identified edgetically perturbed proteins, we found that change in prediction algorithms did not alter the trend of our previous result (table 1). Hence, we used our previous data set (SIFT predicted) for further analyses.

Zhong et al. (2009) have claimed that edgetic perturbation is frequently associated with autosomal dominant disease genes. Moreover, it is also known that dominant disease genes are under stronger purifying selections than recessive disease genes because mutation of a single allele of the gene can adequately cause disease in the former set of genes by altering the synthesized gene product (Furney et al. 2006). Hence, to shed lights on the evolutionary conservation of SHD and SPD proteins over ND proteins, we determined the evolutionary rates of edgetically perturbed proteins compared

Table 1
Proportions of Edgetically Perturbed Proteins in Our Data Set

	Gene Pair	Percentages	*Z* Score	*P*
SIFT-predicted data[a]	SHD vs. SPD	32.01% vs. 16.62%	7.259	$<1.00 \times 10^{-4**}$
	SHD vs. ND	32.01% vs. 3.87%	27.622	$<1.00 \times 10^{-4**}$
	SPD vs. ND	16.62% vs. 3.87%	18.605	$<1.00 \times 10^{-4**}$
dbNSFP-predicted data	SHD vs. SPD	56.44% vs. 38.74%	6.875	$<1.00 \times 10^{-4**}$
	SHD vs. ND	56.44% vs. 16.25%	23.087	$<1.00 \times 10^{-4**}$
	SPD vs. ND	38.74% vs. 16.25%	18.991	$<1.00 \times 10^{-4**}$

[a] Proteins with available PDB structures.
**Significant difference ($P < 0.01$) in proportions.

with the rest of the proteins of our data set. As expected, we observed that edgetically perturbed proteins are more evolutionarily constrained than rest of the human proteins (data set using SIFT: $dN/dS_{\text{edgetically perturbed proteins}} = 0.100$ [$n = 1,546$]; $dN/dS_{\text{other proteins}} = 0.135$ [$n = 9,022$]; $P_{\text{MWT}} = 2.89 \times 10^{-39}$; data set using dbNSFP: $dN/dS_{\text{edgetically perturbed proteins}} = 0.090$ [$n = 2,212$]; $dN/dS_{\text{other proteins}} = 0.140$ [$n = 8,356$]; $P_{\text{MWT}} = 2.23 \times 10^{-108}$). However, it is also plausible that the underlying difference in evolutionary rates of edgetically perturbed and other proteins is potentially due to variations in sample sizes. To rule out the possibility that our analysis is artifactual, we again considered equal size bins of protein evolutionary rates: bin1 (fast evolving proteins, $n = 3,523$), bin 2 (medium evolving proteins, $n = 3,523$), and bin 3 (slow evolving proteins, $n = 3,522$). In support of our result, we noticed that the proportions of edgetically perturbed proteins and other human proteins are comparatively higher in bin 3 and bin 1, respectively (fig. 3*A*). Therefore, it is evident that regardless of sample size variations, an inverse correlation exists between edgetic perturbation and protein evolutionary rates. These observations provide a clue that edgetic perturbation may act as a function of protein evolutionary rates of SHD, SPD, and ND genes.

Drug-Induced Network Perturbation and Evolutionary Rates of SHD, SPD, and ND genes

Several reports (Kuhn et al. 2013; Wang, Thijssen, et al. 2013) have proposed that pharmacological treatment is necessary to restore the function of the perturbed network as drugs often bind to the interface of the disease-associated proteins. Indeed, we obtained a positive correlation between number of drugs and number of phenotypic defects associated with a protein (considering ND genes: ${}_{\text{Number of drugs/protein}}\rho^{\text{Number of defects}} = 0.251$, $P = 1.00 \times 10^{-6}$, $n = 933$; excluding ND genes: ${}_{\text{Number of drugs/protein}}\rho^{\text{Number of defects}} = 0.103$, $P = 2.20 \times 10^{-2}$, $n = 496$). In this context, a recent concept is that essentiality and centrality of target proteins may also increase the likelihood of adverse drug side effects, which in turn may help in progression of diseases among individuals (Wang, Thijssen, et al. 2013). If it is the case, then we may expect that side-effect-associated drug-targeted proteins would majorly participate in human disease progressions, because all side effects literally do not imply diseases. Accordingly, we observed that proteins with drug side effects are more frequently involved in Mendelian diseases than random expectation (odds ratio = 4.063, Z score = 12.440, $P < 1.00 \times 10^{-3}$). In general, drugs may cause side effects through network perturbation in single-interface targets by occupying the only shared interaction interface and disintegrating the interactome network (Wang, Thijssen, et al. 2013). Interestingly, previous studies have established that singlish (two at most) interface proteins are more likely to disintegrate interactome network by interrupting the links between proteins (Kim et al. 2006; Gursoy et al. 2008; Zhang and Ouellette 2011). Accordingly, we observed that proteins with singlish interaction interfaces (at least two interaction partners for each interface) are highly associated with drug side effects than random expectation (odds ratio = 1.850, Z score = 3.837, $P = 1.00 \times 10^{-3}$). Moreover, the proportions of singlish interface proteins are found to be in the order of SHD (30.30%; 160/528) > SPD (23.79%; 299/1,257) > ND proteins (10.72%; 942/8,783) (SHD vs. SPD: Z score = 2.875; SHD vs. ND: Z score = 13.526; SPD vs. ND: Z score = 13.160; $P < 5.00 \times 10^{-3}$ in all three cases). Considering the supplementary data set of Kim et al. for singlish interface hub proteins (Kim et al. 2006), we obtained similar result (table 2).

It is now widely accepted that druggable proteins are highly conserved because substitutions in the interface may obstruct target–drug interactions (Wang, Wang, et al. 2013). Moreover, singlish interface hubs evolve faster than multi-interface hubs but are generally slow evolving than rest of the proteome due to their hub nature (Kim et al. 2006; Clarke et al. 2012). Using drug-targeted proteins, our result also demonstrates that singlish interface proteins are more evolutionarily constrained than the rest ($dN/dS_{\text{drug-targeted proteins with singlish interface}} = 0.105$ ($n = 216$); $dN/dS_{\text{drug-targeted other proteins}} = 0.113$ ($n = 716$) and $P_{\text{MWT}} = 2.00 \times 10^{-3}$). Validating the above result, drug side-effect-associated proteins are found to be more conserved than proteins which do not have any known side effects ($dN/dS_{\text{proteins with drug side effects}} = 0.086$ [$n = 204$]; $dN/dS_{\text{proteins with no known side effect}} = 0.118$ [$n = 729$] and $P_{\text{MWT}} = 6.17 \times 10^{-6}$). Discrepancy may arise

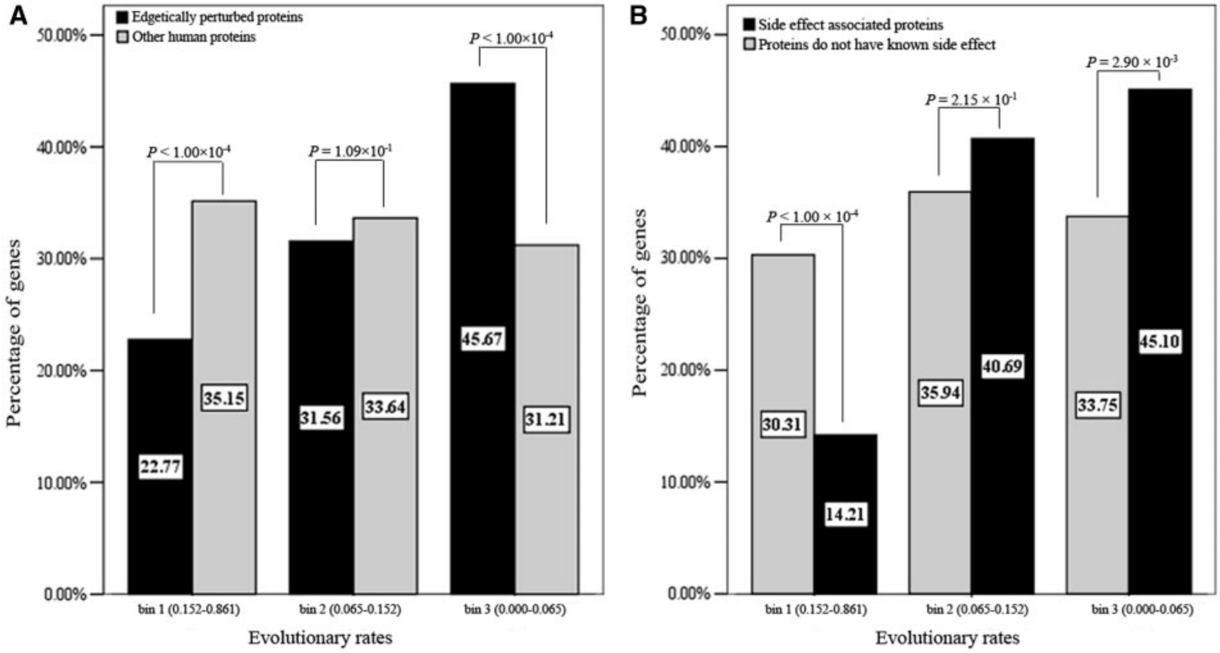

FIG. 3.—Evolutionary rate analyses of interactome network perturbed genes. The bar graphs demonstrate the distributions of (*A*) edgetically perturbed genes and (*B*) side-effect-associated genes in equal size bins classified as (I) fast evolving, (II) medium evolving, and (III) slow evolving groups. MWT was used to find difference between groups. $P < 0.05$ denotes a significant difference.

Table 2

Proportions of Singlish Interface Hubs in Structurally Resolved Network (SIN) for Our Data Set

Gene Pair	Percentages	*Z* Score	*P*
SHD vs. SPD	8.90	4.101	$<1.00 \times 10^{-4**}$
	4.05		
SHD vs. ND	8.90	22.321	$<1.00 \times 10^{-4**}$
	0.26		
SPD vs. ND	4.05	14.714	$<1.00 \times 10^{-4**}$
	0.26		

**Significant difference ($P < 0.01$) in proportions.

due to differences in sample sizes for drug side effect data in our data set. To avoid such circumstances, we again considered equal sample bins of protein evolutionary rates (bin 1: fast evolving genes; bin 2: medium evolving genes; and bin 3: slow evolving genes). Thus, a higher proportions of side-effect-associated proteins in bin 3 and proteins with no known side effects in bin 1 (fig. 3*B*) suggest that our analysis is free from sampling bias. These above observations invoke that drug-induced perturbation/drug side effect may act as a determinant of protein evolutionary rates of SHD, SPD, and ND proteins.

Impact of Gene Expression Level, Tissue Expression Breadth, and Functionality on Evolutionary Rates of SHD, SPD, and ND Genes

Among all the identified factors, gene expression level is claimed to be the most important correlate of protein evolutionary rates to date (Bloom et al. 2006; Drummond et al. 2006; Panda et al. 2012). Considering expression level as the mean expression value of a gene in the tissues (using mRNA-seq data of 22 tissue/cell-line samples) where it is found to be expressed (Kiran and Nagarajaram 2013), we also observed a negative correlation between gene expression level and protein evolutionary rates ($\rho_{\mathrm{Expression\ level}}^{dN/dS} = -0.177$, $P = 1.00 \times 10^{-6}$, $n = 7{,}055$). Hence, we expected higher expression levels of SHD genes than SPD and ND genes. Estimation of gene expression level revealed the same trend (average gene expression level: SHD = 132.899 [$n = 345$]; SPD = 92.085 [$n = 897$]; ND = 32.300 [$n = 5{,}813$]). However, SHD genes share no difference in gene expression levels with SPD genes ($P_{\mathrm{MWT}} = 5.31 \times 10^{-1}$). It may be due to the confounding effect of sample size difference. Hence, we grouped all genes equally into three bins according to their expression level (bin 1: lowly expressed [range: 5.510–10.888, $n = 2{,}351$], bin 2: medium expressed [range: 10.890–21.960, $n = 2{,}351$], and bin 3: highly expressed [range: 21.961–8,041.754, $n = 2{,}353$]). Subsequently, we noticed that in all three bins, the differences in proportions of SHD and SPD genes are statistically insignificant (bin 1: SHD vs. SPD = 17.80% vs. 20.04%; bin 2: SHD vs. SPD = 18.56% vs. 21.80%; and bin 3: SHD vs. SPD = 28.98% vs. 29.51%; $P > 5.00 \times 10^{-2}$ in all three bins). This observation confirms that gene expression level is insufficient to explain the evolutionary rate variations among genes of our interest.

In their article, Park and Choi (2010) have demonstrated that gene expression breadth has a greater influence on protein evolutionary rates than gene expression level because all broadly expressed genes (like genes those are evenly expressed at low levels in all the tissue types) are not necessarily the highly expressed ones (like genes those are expressed at high levels in specific tissue types). Moreover, from an evolutionary perspective, we expected that SHD genes are broadly expressed among all the groups, because broadly expressed genes evolve slower than tissue-specific genes (Zhang and Li 2004; Park and Choi 2010). Surprisingly, we noticed a significantly higher tissue-restricted expression of SHD (average expression breadth = 9.316, $n = 345$) genes compared with SPD (average expression breadth = 10.821, $n = 897$) and ND genes (average expression breadth = 11.606, $n = 5{,}813$) (Kruskal–Wallis test: $P = 2.99 \times 10^{-7}$), in agreement on a concept that diseases genes tend to be expressed in limited number of tissues (Goh et al. 2007; Lage et al. 2008). However, the lower evolutionary rates of tissue-restricted genes imply that tissue-restricted expression alone is inadequate to explain protein evolutionary rates.

Previously, it has been established that the protein products of genes expressed in only one or few tissues are more often involved in regulatory functions, whereas ubiquitously/broadly expressed genes mainly participate in intracellular core functions (Ramskold et al. 2009). Using biological processes of gene ontology (GO-BP) (Lopez-Bigas et al. 2008; Beck et al. 2011), we interestingly noticed that number of core functions does not differ between SHD and SPD genes and that between ND and SPD genes, whereas all the group-to-group differences are statistically significant in case of number of regulatory functions (fig. 4). Subsequently, the higher number of regulatory functions in SHD and SPD genes compared with ND genes (fig. 4) reconfirmed the statement that disease-associated mutations are more likely to affect regulatory processes (Goh et al. 2007). For a better understanding, we performed functional enrichment analysis ($P \leq 10^{-3}$) of our gene set using GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion tool [Eden et al. 2009]). By this means, the only regulatory process found to be enriched in all the three groups (out of false discovery rate corrected 40 SHD, 88 SPD, and 304 ND significant GO terms) of our data set was transmembrane transport (GO: 0055085, *P* values for GO: 0055085 in SHD, SPD, and ND genes are 1.86×10^{-8}, 6.67×10^{-4}, and 1.19×10^{-9}, respectively). At this point, it should be mentioned that genes involved in transmembrane transport are often tissue restricted, disease-causing deleterious mutation prone, and evolve slowly at protein level due to their higher buried residue content (Oberai et al. 2009; Ramskold et al. 2009; Movahedi et al. 2011). Therefore, considering transmembrane proteins (proteins containing transmembrane helices), our analyses also confirmed the same ($_{\text{Transmembrane helices/protein}}\rho^{\text{Expression breadth}} = -0.168$, $P = 1.00 \times 10^{-6}$, $n = 1{,}490$; $_{\text{Transmembrane helices/protein}}\rho^{\text{Deleterious SNPs/gene}} = 0.175$, $P = 1.00 \times 10^{-6}$, $n = 1{,}165$; $_{\text{Transmembrane helices/protein}}\rho^{dN/dS} = -0.066$, $P = 1.00 \times 10^{-3}$, $n = 2{,}577$). Moreover, in support of our studies, the proportions of transmembrane proteins are found to be in the order of SHD (38.26%; 202/528) > SPD (29.83%; 375/1,257) > ND (22.77%; 2,000/8,783) ($P < 1.00 \times 10^{-4}$ in all three cases).

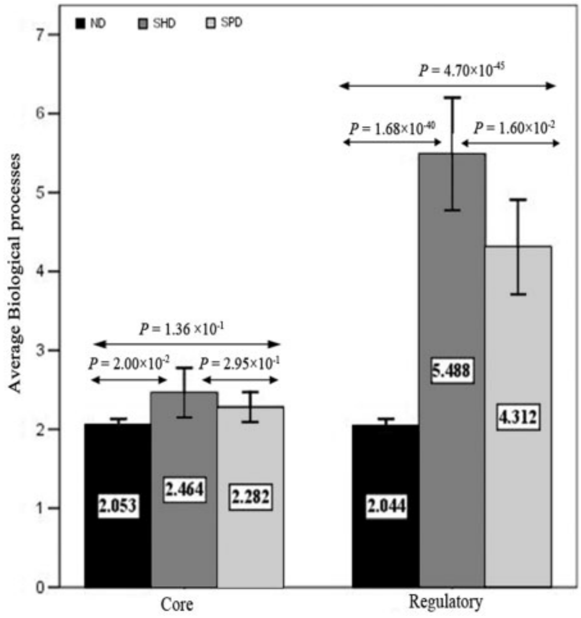

Fig. 4.—Assessment of core and regulatory processes of our gene sets. In the left cluster, the bar graph exhibits that SPD genes share no difference ($P > 0.05$) in number of core functions to SHD or ND genes. All the three groups of genes show considerable differences in their number of regulatory functions (right cluster). Error bars represent the standard error of the mean in all three groups of genes.

To understand the relationship between transmembrane transport and tissue expression breadth on protein evolutionary rate, we split all the genes into three equally spaced bins according to their tissue expression breadth (bin 1: lower breadth [range: 1–7, $n = 2{,}828$], bin 2: medium breadth [range: 8–14, $n = 1{,}444$], and bin 3: higher breadth [range: 15–22, $n = 2{,}783$]). Consequently, for tissue-restricted genes (bin 2), the correlation between expression breadth and evolutionary rates disappears ($_{\text{Tissue expression breadth}}\rho^{dN/dS} = 0.041$, $P = 1.23 \times 10^{-1}$, $n = 1{,}444$), whereas the correlation between number of transmembrane helices and protein evolutionary rates appears better than previous ($_{\text{Transmembrane helices/protein}}\rho^{dN/dS} = -0.189$, $P = 1.00 \times 10^{-3}$, $n = 334$) (fig. 5).

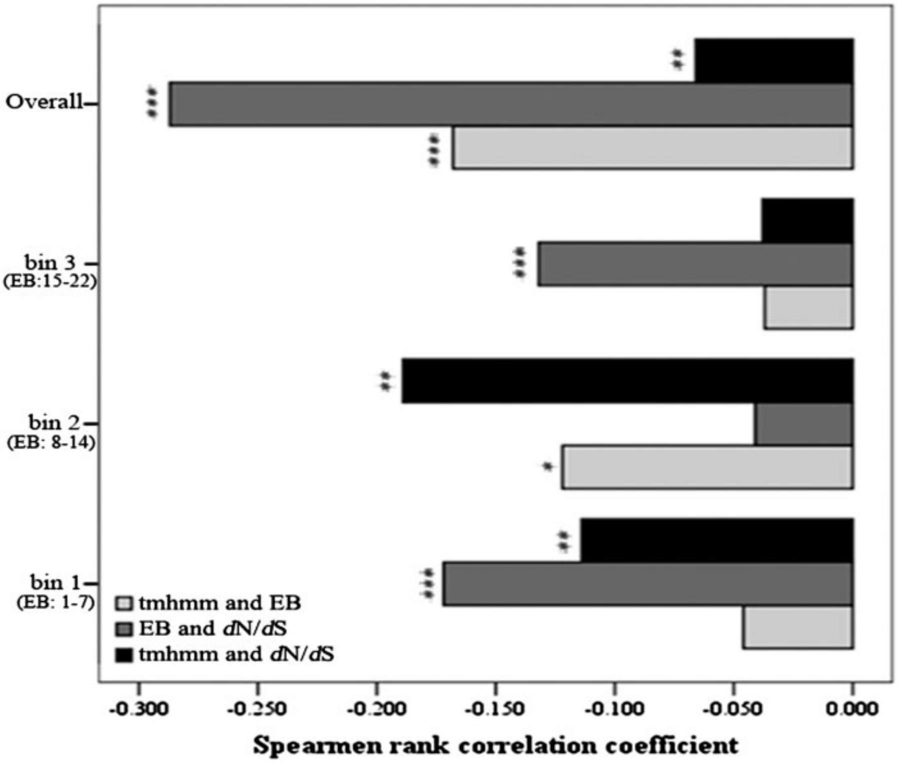

Fig. 5.—Relationships among transmembrane transport, tissue expression breadth, and protein evolutionary rates. In each cluster, tmhmm, EB, and d*N*/d*S* represent transmembrane helix count per protein, tissue expression breadth, and protein evolutionary rate, respectively. For each bin, range of tissue expression breadth has been mentioned in the parentheses. Significance: **P* value < 0.05, ***P* value < 0.01, and ****P* value < 0.001.

Moreover, the existence of a significant correlation between tissue expression breadth and the number of transmembrane helices $({}_{\text{Transmembrane helices/protein}}\rho^{\text{Expression}}_{\text{breadth}} = -0.122,\ P = 2.60 \times 10^{-2},\ n = 334)$ suggests that tissue expression breadth and transmembrane transport may simultaneously affect protein evolutionary rates (fig. 5). Using equally populated bins (according to tissue expression breadth), a similar result was obtained (data not shown). Further considering genes having both (number of transmembrane helices and tissue expression breadth) data available, we noticed considerably ($P_{\text{MWT}} = 1.10 \times 10^{-6}$) lower evolutionary rates for such genes (average d*N*/d*S* = 0.113, *n* = 1,490) compared with all human genes in our data set (average d*N*/d*S* = 0.130, *n* = 10,568). Therefore, it is evident from our analysis that the interaction term of tissue expression breadth and transmembrane transport may influence the evolutionary rates of SHD, SPD, and ND genes.

Relative Contribution of Parameters on Protein Evolutionary Rates

In this study, we identified the importance of two distinct network perturbation models and the interaction term of tissue expression breadth and transmembrane transport (three out of the six predictors) on protein evolutionary rates except number of phenotypic defects and gene expression levels. Because edgetic perturbation and drug-induced perturbation may certainly have different molecular consequences on human disease progressions and on protein evolutionary rates, it is necessary to understand the relative contributions of all the factors in dictating the variation of protein evolutionary rates. In this regard, ANCOVA helps to understand the functional relationships between the measurement variables, while at least one of the predictor variables is categorical in nature (Chen et al. 2012). However, a number of known evolutionary rate determinants (such as gene recombination rate, paralog number, protein length, and complex association number) were shown to be associated with human genetic diseases (Lopez-Bigas and Ouzounis 2004; Bloom et al. 2006; Lage et al. 2008; Cai et al. 2009; Zhou et al. 2013). Such biological parameters are likely to be confounding factors in our analysis. We, therefore, included the aforementioned four genomic covariates in our ANCOVA model. Thus, by using the backward elimination approach, our ANCOVA model ($F = 10.101$, $P < 1.00 \times 10^{-4}$, $R^2 = 13.4\%$) demonstrated that drug-induced (side effect/no side effect) perturbation model ($\beta = -0.278$, $P = 2.98 \times 10^{-4}$), gene paralog number ($\beta = -0.154$, $P = 4.20 \times 10^{-2}$), and the interaction term of expression breadth-transmembrane transport ($\beta = -0.140$, $P = 4.90 \times 10^{-2}$) independently affects protein evolutionary rates.

However, the result of an ordinary regression analysis can be misleading as it does not take into account predictor collinearity (Drummond et al. 2006). To rule out such possibility, principal component regression method can be applied (Drummond et al. 2006). Each independent principal component is linear combinations of the original predictor variables (Drummond et al. 2006). We, therefore, carried out categorical principal component analysis due to the presence of categorical predictors in our data set. Subsequent regression of the response variable (protein evolutionary rate) on the principal component scores revealed that two components (PC1 and PC2) made significant contributions to the regression model ($F = 16.138$, $P = 3.70 \times 10^{-7}$, $R^2 = 14.6\%$). Among them, 11.2% variance of protein evolutionary rates is explained by PC1 ($\beta = -0.343$, $P = 1.87 \times 10^{-6}$) and PC2 ($\beta = -0.196$, $P = 5.00 \times 10^{3}$) explains 3.4% evolutionary rate variance. PC1 primarily measures paralog numbers, protein length, and drug-induced perturbation, whereas PC2 measures the interaction term of expression breadth-transmembrane transport and protein complex association number (table 3). Thus, from both of our ANCOVA and principal component regression analyses, we infer that except edgetic perturbation (perturbed/nonperturbed) model, drug-induced (side effect/no side effect) perturbation model, and the interaction term of expression breadth-transmembrane transport have independent effects on protein evolutionary rates. It is probably due to the fact that total network disruption (all edge removals in the interactome network) by drug-induced perturbation is a combination of several edgetic perturbations. Hence, drug-induced perturbation has reasonably severe effect on gene phenotypes than edgetic perturbation.

Table 3

Result of Categorical Principal Component Analysis on Seven Predictor Variables of Our Data Set

Predictors	Percent Contributions	
	PC1	PC2
Gene paralog number	**23.78**	7.17
Gene recombination rate	19.86	14.03
EB*tmhmm	7.06	**36.81**
Protein length	**24.53**	8.29
Complex association number	4.35	**32.59**
Drug induced perturbation	**20.42**	1.11
Edgetic perturbation	0.00	0.00

Note.—PC1 and PC2 designate principal components 1 and 2, respectively. EB*tmhmm represents interaction term of expression breadth and transmembrane transport. Bold indicates that the corresponding variable contributes at least 20% to the component.

Discussion

In the field of molecular evolution, there are commonalities behind disease mutations which have been detected, but there are more complexities to disease mutations, which are yet to be discerned. In this work, we analyzed the evolutionary rates of human autosomal phenotypic disease (SHD and SPD) genes originated by mutation(s) in a single gene (monogenic) in comparison to ND genes. For evolutionary rate study, we considered widely used human–mouse orthologous pair (Liao and Zhang 2006, 2008; Cai et al. 2009; Gharib and Robinson-Rechavi 2011). The most common reason is the functional conservation between human and mouse in normal and pathological conditions (Gharib and Robinson-Rechavi 2011). However, it has been emphasized that the murid rodents have higher divergence time (~95 Ma) than primates (Clement and Arndt 2011). Hence, an alternative way is to consider human–chimpanzee or human–macaque ortholog to compute protein evolutionary rate. Again, for closely related taxa, it has been criticized that estimation of evolutionary rate can be misleading as d*S* might suffer from the uncertainty due to potentially smaller branch lengths (Wolf et al. 2009). We, therefore, continued our evolutionary rate analysis using human–mouse orthologous pair. Consistent to the previous studies (Blekhman et al. 2008; Cai et al. 2009) that disease genes are conserved than ND genes, we established that genes with more defects are strongly constrained compared with the genes with single/no defect. However, we noticed that number of defects is unable to explain the evolutionary rate differences of SHD, SPD, and ND genes. Hence, to clarify the underlying reasons of evolutionary rate variations among SHD, SPD, and ND genes, we incorporated functional (core/regulatory), gene expression related (mRNA expression level and tissue expression breadth), and structural (network perturbation due to edge removal and network perturbation due to drug binding of singlish interaction interface) features of protein evolutionary rates in this communication.

Malfunctioning/disruption of a biological function often lead to the progression of human diseases (Lopez-Bigas et al. 2008; Janjic and Przulj 2012). Distinct disease genes can be traced through interactome networks (Goh et al. 2007). To this end, edge removal/edgetic perturbation model offers an explanatory power that why different mutations from the same gene produce different phenotypes (Zhong et al. 2009). We as expected obtained higher disease association for edgetically perturbed proteins. Prior studies have already established that human disease genes evolve slower than ND genes (Blekhman et al. 2008; Cai et al. 2009). Hence, frequent disease involvements practically support the lower evolutionary rates of edgetically perturbed proteins compared with other proteins those are not found to be edgetically perturbed. In this context, it should be mentioned that drugs used for treating a disease may in turn cause diseases through increased side effects. Such side effects are promoted by drugs especially in single/singlish interface targets by occupying its interaction interface(s) (Gursoy et al. 2008; Zhang and Ouellette 2011; Wang, Thijssen, et al. 2013). In our study, for the first time, we have noticed that proteins with singlish interaction interfaces are highly associated with drug side effects and such side-effect-associated proteins are evolutionarily more conserved than rest of the proteins. However, it has been recently found that side-effect-associated proteins are essential in nature and occupy central position in the interactome networks (Wang, Thijssen, et al. 2013). Because, centrality and essentiality are negative correlates of protein evolutionary rates (Hahn and Kern 2005), our findings are reasonable in this ground. Further comparison of both the two network perturbation models revealed a higher dominance of drug-induced perturbation over the edgetic perturbation model. We reasoned that drug-induced perturbation can totally disrupt a network and, thus, is a combination of several edgetic perturbations.

One basic feature that is widely related to protein evolution is expression level of a gene, which shares a negative correlation to protein evolutionary rates (Bloom et al. 2006; Drummond et al. 2006). Interestingly, we did not notice any difference in gene expression levels between ND and SHD genes. It may be due to the fact that broadly expressed genes, which are evenly expressed at low levels in all tissues are not necessarily the highly expressed ones that are expressed at high levels in specific tissues (Park and Choi 2010). However, in contrast to the common view (Winter et al. 2004), our tissue expression study revealed a higher tissue-restricted expression of highly conserved disease genes. Although, we obtained a significant difference in tissue expression breadth within SHD, SPD, and ND genes, the above unexpected observation confirms that tissue expression breadth alone cannot explain the variation in evolutionary rates among genes of our interest. To find a reason, we concentrated on protein functions as tissue-specific components often need to perform certain regulatory functions (Ramskold et al. 2009). Further elaborative study revealed that only regulatory functions can effectively clarify the evolutionary rate differences among genes of our interest, whereas core functions are found to be inadequate. One possible scenario could be that core functions are required for basic processes within the cell and always need to be switched on. Therefore, the number of such functions required by any cell should be relatively similar (Peterson and Fraser 2001). Because disease-related mutations compatible with survival are more likely to be maintained a population, cells prefer to endure deleterious mutations in regulatory pathways (Goh et al. 2007). Interestingly, a comprehensive analysis on regulatory functions identified "transmembrane transport" as a common function present on all three groups. Proteins involve in such function is often tissue-restricted and evolve slower at the protein level (Oberai et al. 2009; Movahedi et al. 2011). Hence, presence of higher proportion of transmembrane protein in human disease category makes it evident that tissue expression breadth and transmembrane transport simultaneously influence the evolutionary rates of SHD, SPD, and ND genes.

To summarize, our analysis established a link between human phenotypic disease network and protein evolution, which will definitely help in understanding human monogenic disease etiology and the underlying mechanism of disease progressions from a single gene. The localization of disease genes in regulatory pathways is especially important for identifying new disease candidates. Till date, majority of the target-based drug identification depends on observable phenotypic defects. However, for the first time, our communication has provided useful information on the relative risk associated with drug perturbed network over edge perturbed network. Our work emphasized that further thorough investigations are needed to improve drug efficacy (i.e., drug with minimal side effects), especially for SHD and SPD genes. With this work, we can make a substantial progress in future medicine.

Supplementary Material

Supplementary table S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Council of Scientific and Industrial Research (9/15(354)/2008-EMR-I) to T.B. The authors are thankful to Dr Dan Graur and two anonymous reviewers for critical reading of the manuscript and advice. They especially thank Mr. Jyotirmoy Das and Mrs. Kamalika Sen for their help and suggestions.

Literature Cited

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's online mendelian inheritance in man (OMIM (R)). Nucleic Acids Res. 37: D793–D796.

Barabasi A-L, Gulbahce N, Loscalzo J. 2011. Network medicine: a network-based approach to human disease. Nat Rev Genet. 12:56–68.

Beck M, et al. 2011. The quantitative proteome of a human cell line. Mol Syst Biol. 7:549.

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

Begum T, Ghosh TC. 2010. Understanding the effect of secondary structures and aggregation on human protein folding class evolution. J Mol Evol. 71:60–69.

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

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 23: 1751–1761.

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. 1:131–144.

Care MA, Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. 2009. Combining the interactome and deleterious SNP predictions to improve disease gene identification. Hum Mutat. 30:485–492.

Chakraborty S, Ghosh T. 2013. Evolutionary rate heterogeneity of core and attachment proteins in yeast protein complexes. Genome Biol Evol. 5:1366–1375.

Chavali S, Barrenas F, Kanduri K, Benson M. 2010. Network properties of human disease genes with pleiotropic effects. BMC Syst Biol. 4:78.

Chen F-C, Pan C-L, Lin H-Y. 2012. Independent effects of alternative splicing and structural constraint on the evolution of mammalian coding exons. Mol Biol Evol. 29:187–193.

Clarke D, Bhardwaj N, Gerstein M. 2012. Novel insights through the integration of structural and functional genomics data with protein networks. J Struct Biol. 179:320–326.

Clement Y, Arndt P. 2011. Substitution patterns are under different influences in primates and rodents. Genome Biol Evol. 3:236–245.

Das J, Chakraborty S, Podder S, Ghosh T. 2013. Complex-forming proteins escape the robust regulations of miRNA in human. FEBS Lett. 587: 2284–2287.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:48.

Finn RD, Miller BL, Clements J, Bateman A. 2014. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res. 42:D364–D373.

Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48–D55.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296: 750–752.

Furney SJ, Alba MM, Lopez-Bigas N. 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics 7:165.

Gharib W, Robinson-Rechavi M. 2011. When orthologs diverge between human and mouse. Brief Bioinform. 12:436–441.

Goh KI, et al. 2007. The human disease network. Proc Natl Acad Sci U S A. 104:8685–8690.

Gursoy A, Keskin O, Nussinov R. 2008. Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans. 36:1398–1403.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803–806.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885–1891.

Huang H, et al. 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol. 5:R47.

Huang Y, et al. 2013. Recent adaptive events in human brain revealed by meta-analysis of positively selected genes. PLoS One 8:e61280.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Janjic V, Przulj N. 2012. Biological function through network topology: a survey of the human diseasome. Brief Funct Genomics. 11:522–532.

Jin W, Qin P, Lou H, Jin L, Xu S. 2012. A systematic characterization of genes underlying both complex and Mendelian diseases. Hum Mol Genet. 21:1611–1624.

Kato M, et al. 2008. Recombination rates of genes expressed in human tissues. Hum Mol Genet. 17:577–586.

Kim P, Lu L, Xia Y, Gerstein M. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1938–1941.

Kiran M, Nagarajaram H. 2013. Global versus local hubs in human protein-protein interaction network. J Proteome Res. 12:5436–5446.

Kircher M, et al. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 46:310–315.

Kuhn M, et al. 2013. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 9:1611–1624.

Lage K, et al. 2008. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A. 105:20870–20875.

Law V, et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42:D1091–D1097.

Lee W, Zhang Y, Mukhyala K, Lazarus R, Zhang Z. 2009. Bi-Directional SIFT predicts a subset of activating mutations. PLoS One 4:e8311.

Liao B, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol. 23: 530–540.

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

Liu X, Jian X, Boerwinkle E. 2013. dbNSFP v2.0: a database of human non-synonymous snvs and their functional predictions and annotations. Hum Mutat. 34:E2393–E2402.

Lopez-Bigas N, De S, Teichmann S. 2008. Functional protein divergence in the evolution of *Homo sapiens*. Genome Biol. 9:R33.

Lopez-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32: 3108–3114.

Luo Q, Pagel P, Vilne B, Frishman D. 2011. DIMA 3.0: domain interaction map. Nucleic Acids Res. 39:D724–D729.

Mosca R, Ceol A, Stein A, Olivella R, Aloy P. 2014. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 42:D374–D379.

Movahedi S, Van de Peer Y, Vandepoele K. 2011. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in *Arabidopsis* and rice. Plant Physiol. 156:1316–1330.

Oberai A, Joh NH, Pettit FK, Bowie JU. 2009. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. Proc Natl Acad Sci U S A. 106:17747–17750.

Oliver KL, et al. 2014. Harnessing gene expression networks to prioritize candidate epileptic encephalopathy genes. PLoS One 9: e102079.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7.337–348.

Panda A, Begum T, Ghosh T. 2012. Insights into the evolutionary features of human neurodegenerative diseases. PLoS One 7:e48336.

Park S, Choi S. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol Biol. 10:241.

Peterson S, Fraser C. 2001. The complexity of simplicity. Genome Biol. 2. 2002.1-2002.7.

Podder S, Ghosh TC. 2010. Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. Mol Biol Evol. 27:934–941.

Prasad TSK, et al. 2009. Human protein reference database-2009 update. Nucleic Acids Res. 37:D767–D772.

Ramskold D, Wang E, Burge C, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 5:e1000598.

Razeto-Barry P, Diaz J, Cotoras D, Vasquez R. 2011. Molecular evolution, mutation size and gene pleiotropy: a geometric reexamination. Genetics 187:877–885.

Ruepp A, et al. 2010. CORUM: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res. 38:D497–D501.

Sim N, et al. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40:W452–W457.

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

Stenson PD, et al. 2009. The human gene mutation database: 2008 update. Genome Med. 1:13.

Tang CSM, Epstein RJ. 2007. A structural split in the human genome. PLoS One 2:e603.

Wang ET, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476.

Wang X, Thijssen B, Yu H. 2013. Target essentiality and centrality characterize drug side effects. PLoS Comput Biol. 9:e1003119.

Wang X, Wang R, Zhang Y, Zhang H. 2013. Evolutionary survey of druggable protein targets with respect to their subcellular localizations. Genome Biol Evol. 5:1291–1297.

Wang X, et al. 2012. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 30:159–164.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res. 14:54–61.

Wolf JBW, Kuenstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (d(N)) and synonymous (d(S)) substitution rates affects inference of selection. Genome Biol Evol. 1: 308–319.

Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.

Zhang K, Ouellette B. 2011. CAERUS: Predicting CAncER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. PLoS Comput Biol. 7: e1001114.

Zhang LQ, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236–239.

Zhong Q, et al. 2009. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 5:321.

Zhou T, Hu Z, Zhou Z, Guo X, Sha J. 2013. Genome-wide analysis of human hotspot intersected genes highlights the roles of meiotic recombination in evolution and disease. BMC Genomics 14:67.

Associate editor: Dan Graur