Genomics 108 (2016) 18–24

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Exploring the evolutionary rate differences between human disease and non-disease genes

Sandip Chakraborty, Arup Panda, Tapash Chandra Ghosh *

Bioinformatics Centre, Bose Institute, P-1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form 29 October 2015
Accepted 3 November 2015
Available online 11 November 2015

Keywords:
Evolutionary rates
Human
Disease genes
Protein complex
Expression breadth
Multifunctionality

ABSTRACT

Comparisons of evolutionary features between human disease and non-disease genes have a wide implication to understand the genetic basis of human disease genes. However, it has not yet been resolved whether disease genes evolve at slower or faster rate than the non-disease genes. To resolve this controversy, here we integrated human disease genes from several databases and compared their protein evolutionary rates with non-disease genes in both housekeeping and tissue-specific group. We noticed that in tissue specific group, disease genes evolve significantly at a slower rate than non-disease genes. However, we found no significant difference in evolutionary rates between disease and non-disease genes in housekeeping group. Tissue specific disease genes have a higher protein complex number, elevated gene expression level and are also associated with conserve biological processes. Finally, our regression analysis suggested that protein complex number followed by protein multifunctionality independently modulates the evolutionary rate of human disease genes.

1. Introduction

The preliminary aim of medical research is to explore the genetic basis of human diseases to improve the remedies of disease prevention and their treatment [1]. In the last decade, researchers analyzed the genetic basis of the human diseases by exploring different aspects of disease genes such as their evolutionary pattern, functional property, gene essentiality, gene duplication, protein disorder content, and protein–protein interaction network [2–12]. Evolutionary studies on human disease genes could provide important clues on their phenotypic connectivity and therefore, a large number of studies have compared evolutionary rates of human disease genes with that of non-disease genes [3–7,13]. Initially, Smith and Eyre-Walker showed that human disease genes evolve at a faster rate as compared to non-disease genes [3]. However, in a subsequent study Huang and Winter found no significant difference in protein evolutionary rates between human disease and non-disease genes [4]. Even an opposite trend, i.e., disease genes are evolutionarily conserved relative to non-disease genes was observed by López-Bigas and Ouzounis in their independent study [5]. These conflicting results draw further attention to characterize the evolutionary forces operating on human disease genes. Parallel studies delineating the effects of gene expression pattern on protein evolutionary rate emphasized that evolutionary constraints on human proteins vary widely depending upon their gene expression breadth (number of tissues in which a gene is expressed). Housekeeping genes were shown to be evolutionary conserved, whereas tissue-specific genes were found to be evolutionarily faster [14,15]. In recent studies, researchers analyzed the evolutionary rates of different disease gene classes like monogenic, polygenic and neurodegenerative disease genes to unveil the signatures of molecular evolution in human disease genes [7,8]. Despite, all these studies nature of the evolutionary forces acting on human disease genes remains a controversial issue till date. Maximum Genetic Diversity (MGD) hypothesis developed by Huang and his collogues may provide important insights in this regard [16–18]. Interestingly, majority of earlier studies concluded that the disease genes are tissue-specific by nature [3,19], i.e., these genes are expressed in a narrower ranges of tissues. Thus, this suggests that disease genes would evolve at a faster rate, a trend observed by Smith et al. [3]. However, the majority of earlier studies reported that human disease genes evolve at a slower rate in spite of their tissue-specific nature [5,7,19,20]. Thus, it raised the question, how disease genes remained conserved in spite of being expressed in the fewer number of tissues?

Most of the previous analyses of gene expression of human disease and non-disease genes were based on microarray experiments. However, statistical methods for analyzing microarray data are less capable of differentiating low gene expression pattern from experimental noises [21]. Thus, there is a high possibility of error included in the previous studies when lowly expressed genes were considered. Therefore, misclassification may occur in distinguishing the tissue-specific and housekeeping genes. Interestingly, using EST and microarray dataset Zhu et al. concluded that the information of the total number housekeeping genes was less documented than it is actually present [22]. Using the next-generation RNA-sequencing data Emig et al. also estimated higher

* Corresponding author.
E-mail address: tapash@jcbose.ac.in (T.C. Ghosh).

http://dx.doi.org/10.1016/j.ygeno.2015.11.001
0888-7543/

proportion of housekeeping genes in their study compared to the microarray-based experiments [21]. Therefore, in this study we used RNA-sequencing data to classify human genes in housekeeping and tissue-specific categories. Moreover, studies that dealt with the evolutionary rate of human disease genes were mainly concentrated on specific disease types and analyzed approximately 500 to 2000 human disease genes [1,3,7]. Therefore, the apparent conflict in their results may be due to the inconsistency in the datasets. Notably, majority of those studies used OMIM database to collect human disease genes. However, Wang et al. [23] reported that the disease genes such as cancer and type 2 diabetes are underrepresented in the OMIM database. Thus, majority of the complex disease genes are not included in this database. Recently, to get a global view of human disease gene network Goh et al. [24] collected more than 12,000 human disease genes from two different databases, viz., OMIM, and GAD. To establish a global trend, in this study, we took this approach and collected human disease genes from three different databases viz., OMIM, GAD, and HGMD databases.

A number of parameters were shown to dictate the evolutionary rate of human proteins [25–27]. Among these possible determinants, gene expression levels were considered to have the strongest influence in constraining the rate of sequence evolution [27,28]. In our previous study, we established that protein complex forming ability has substantial contribution in determining their evolutionary rates [29,30]. In particular, we noticed that proteins participating in the large number of protein complexes are evolutionarily conserved. Although, it was noted that mutations in these proteins are susceptible to diseases [31]; however, how protein complex forming nature dictates the evolutionary rate of human disease proteins still remains elusive. Recent progress in the field of evolutionary biology also emphasized that the rate of protein evolution can be constrained by their functional requirements [25]. Along with protein multifunctionality, the involvement in core or regulatory processes of the proteins also shown to have a profound influence on their evolutionary rates [7]. Therefore, in this study we considered the interplay of all these factors to analyze their relative contribution to human disease protein evolution in housekeeping and tissue-specific groups.

Finally, our results revealed that the human disease and non-disease genes are shaped by different evolutionary constraints in tissue-specific and housekeeping groups. Here, we observed that disease genes evolve slower than non-disease genes in the tissue-specific groups. However, in housekeeping gene both disease and non-disease genes were found to evolve at a similar pace. Interestingly, our study revealed that disease genes are evolutionarily constrained from their higher gene expression level, higher protein complex association, and their higher protein multifunctionality in tissue-specific group. Independent influence of all these factors on the evolutionary rate of human disease proteins was confirmed from regression analysis. Based on results presented in this communication, here we proposed to consider the expression profile of the disease and non-disease genes for better understanding of evolutionary constraints operating on these genes.

2. Results

2.1. Tissue-specific disease genes are evolutionarily conserved

In this study, we compared the average evolutionary rates (dN/dS) of 10,291 human disease genes with that of 4957 non-disease genes. Here, we noticed that the average dN/dS value of human disease genes is significantly lower than the non-disease genes (average $dN/dS_{disease} = 0.2918$ (± 0.0024), average $dN/dS_{non\text{-}disease} = 0.3502$ (± 0.0040); Mann–Whitney U test, $P < 1.0 \times 10^{-6}$) [Fig. 1]. Previously, López-Bigas et al. [5] observed a similar trend with a much smaller dataset (they studied only 1567 disease genes). However, Smith et al. [3] found that disease genes evolve at a faster rate, and they are tissue specific (TS) by nature. In accordance with the observation of Smith et al. [3], we also found that most of the disease genes are expressed in tissue-specific fashion as compared to non-disease genes (61.59% disease and 52.09% non-disease genes are TS, Fisher's exact test, $P < 1.0 \times 10^{-6}$). However, it has already been reported that TS genes evolve at a faster rate than the housekeeping (HK) genes [14]. Indeed, here we also noticed that TS genes evolve at a higher rate than the HK genes (average $dN/dS_{TS} = 0.3526$ (± 0.0046), average $dN/dS_{HK} = 0.2525$ (± 0.0037); Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). The negative relationship between evolutionary rates and expression breadth (EB) (Spearman's $\rho_{EB\ vs.\ dN/dS} = -0.1133$, $P < 1.0 \times 10^{-6}$) further confirmed that ubiquitously expressed genes evolve at a slower pace than the tissue-specific genes [3,15]. Considering the tissue-specific nature of human disease genes it could be expected that these genes would evolve at a higher rate than non-disease genes. However, their slower evolutionary rate raises an important question regarding the fundamental relationship between protein evolutionary rates and gene expression breadth. In order to resolve this controversy, we analyzed the evolutionary rates of disease and non-disease genes in TS and HK group separately. We observed that disease genes evolve significantly slower than non-disease genes in TS group (average $dN/dS_{disease} = 0.3271$ (± 0.0050), average $dN/dS_{non\text{-}disease} = 0.4279$ (± 0.0103); Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). However, no significant differences of protein evolutionary rate between disease and non-disease genes was observed in HK group (average $dN/dS_{disease} = 0.2525$ (± 0.0046), average $dN/dS_{non\text{-}disease} = 0.2525$ (± 0.0064); Mann–Whitney U test, $P = 7.0 \times 10^{-1}$) [Fig. 1].

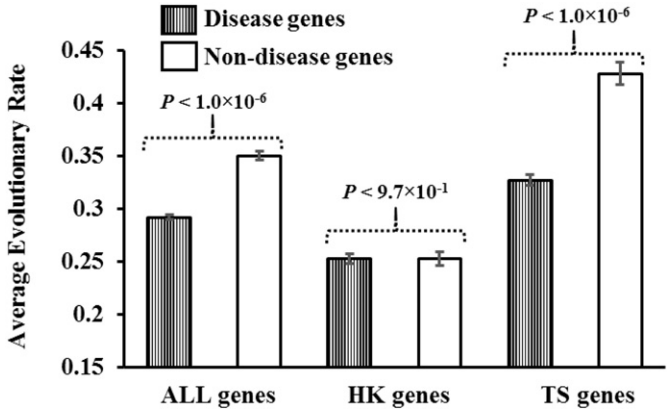

Fig. 1. Average evolutionary rates (dN/dS) of human disease and non-disease genes. The statistical comparison was performed by two-tailed Mann–Whitney U test.

High gene expression level imposes a strong selective constraint on protein evolutionary rate [27,30]. Thus, highly expressed genes are often found to have a lower evolutionary rate than the genes that expressed at lower level [28,29]. Accordingly, we also found a significant negative correlation between expression level (EL) and evolutionary rates (dN/dS) (Spearman's $\rho_{EL\ vs.\ dN/dS} = -0.0780$, $P < 1.0 \times 10^{-6}$) in our study. Therefore, it is likely that expression abundance may modulate the evolutionary rate of disease and non-disease genes in TS and HK gene pool. Interestingly, when we calculated their average expression level, disease genes showed 6 to 7 fold increased expression level than non-disease genes (average $EL_{disease} = 44.9373$ (± 6.2505), average $EL_{non\text{-}disease} = 7.0786$ (± 0.6774); Mann–Whitney U test, $P < 1.0 \times 10^{-6}$) in TS group. However, in HK group we found negligible difference in the expression level between disease and non-disease genes (average $EL_{disease} = 6.0260$ (± 0.2245), average $EL_{non\text{-}disease} = 5.8015$ (± 0.4096); Mann–Whitney U test, $P < 9.0 \times 10^{-6}$). Thus, these results emphasize that disease genes evolve at slower rate than the non-disease genes might be due to the constraints imposed by their higher gene expression level in TS group.

The aforementioned results indicate that both EB and EL imposes strong selection pressure on the rates of protein evolution. However,