BMB
reports

Human intronless disease associated genes are slowly evolving

Subhash Mohan Agarwal[1,*] *& Prashant K. Srivastava*[2]

[1]Center for Computational Biology and Bioinformatics, School of Information Technology, Jawaharlal Nehru University, New Delhi 110067, India, [2]Cellular and Molecular Pathology (G130), German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

In the present study we have examined human-mouse homologous intronless disease and non-disease genes alongside their extent of sequence conservation, tissue expression, domain and gene ontology composition to get an idea regarding evolutionary and functional attributes. We show that selection has significantly discriminated between the two groups and the disease associated genes in particular exhibit lower K_a and K_a/K_s while K_s although smaller is not significantly different. Our analyses suggest that majority of disease related intronless human genes have homology limited to eukaryotic genomes and their expression is localized. Also we observed that different classes of intronless disease related genes have experienced diverse selective pressures and are enriched for higher level functionality that is essentially needed for developmental processes in complex organisms. It is expected that these insights will enhance our understanding of the nature of these genes and also improve our ability to identify disease related intronless genes. [BMB reports 2009; 42(6): 356-360]

INTRODUCTION

For years it is well established that the gene structure of higher eukaryotic organisms including humans, constitutes of exons (protein-coding region) and introns. However, after the completion of human genome sequencing it was observed that approximately 10% of the human genes do not have introns i.e. they are coded by single exonic genes (1). This disclosure led to initiation of several studies to understand the structural and functional properties of these intronless genes. As a result, analysis revealed several interesting insights about their nature. For example, comparative analysis of these genes in humans has indicated that most of them are eukaryotic lineage specific and are evolving rapidly (2, 3). Also it has been shown that these genes are non-uniformly distributed across the various functional categories and exhibit distinct domains and molecular functions (4). In another study, comparison of human intron containing and intron lacking genes have demonstrated significant difference in the oligonucleotide composition of coding sequences suggesting the presence of novel and different exonic splicing information (5). All these studies do indicate that human intronless genes have peculiar distinct properties and should be studied as a separate class. However, the area still remains unexplored despite their sizeable amount and significance in human genome. At the same time, understanding the genetic basis of inherited disorders to improve disease prevention and treatment is still one of the primary objectives of medical researchers (6). In addition, in this genomic era the presence of various resources like OMIM (7), a catalogue of human diseases, along with information regarding homologous sequences provides us with an opportunity to better understand human biology. Therefore in continuation with our previous efforts we have addressed what evolutionary and functional attributes are associated with intronless disease genes? To this end we have examined evolutionary rate pattern of human-mouse homologous intronless disease and non-disease genes alongside their extent of sequence conservation, tissue expression, domain distribution and gene ontology composition.

*Corresponding author. Tel: 91-120-2578837; Fax: 91-120-2579473; E-mail: smagarwal@yahoo.com

Received 17 September 2008, Accepted 19 December 2008

Keywords: Gene ontology, HomoloGene, Human intronless genes, Nonsynonymous substitution rate (K_a), OMIM, Synonymous substitution rate (K_s)

RESULTS AND DISCUSSION

As it is known that the degree of selective pressure to which a gene has been subjected is reflected by the ratio of rate of non-synonymous to synonymous substitutions, in the present study the variation in the rate of nucleotide substitutions among human intronless disease and non-disease associated genes was investigated. To examine synonymous and non-synonymous substitution rate differences between 334 disease and 252 non-disease human intronless genes, the K_a/K_s ratio for each human and mouse 1 : 1 ortholog pair was calculated (Table 1). On comparing the K_a, K_s, and K_a/K_s of the two groups of genes, it was found that the disease associated human intronless genes have smaller and significantly different K_a (0.074 vs 0.103) and K_a/K_s (0.118 vs 0.160) while K_s (0.604 vs 0.622) although smaller is not significantly different, as determined using Mann-Whitney U test (Fig. 1a, Fig. 1b and Table 1). On an average, disease related intronless genes were

Table 1. Summarizes the various features investigated for disease and non-disease HIG

	Disease mean (SE)	Non-disease mean (SE)	P-value mann-whitney
K_a	0.074 (0.004)	0.103 (0.005)	9.8×10^{-6}
K_s	0.604 (0.009)	0.622 (0.011)	0.2343
K_a/K_s	0.118 (0.005)	0.160 (0.007)	5.5×10^{-6}

K_a: number of non-synonymous substitution per non-synonymous site, K_s: number of synonymous substitution per synonymous site, SE: standard error

observed to evolve 28% slower than non-disease genes. Clearly, this variation is not due to the differences in the mutation rates, as K_s of the two groups have not varied accordingly. As it is assumed that synonymous substitutions are usually neutral whereas non-synonymous substitutions are subject to selective pressures, a significant decrease in K_a of disease related genes demonstrates an increase in selective pressure on their amino acid sequence. Consequently, the finding suggests that human intronless disease genes are under strong purifying selection as a result of which these genes have mutated slower than their non-disease counterparts. On the other hand, the insignificant difference in the average K_s rates of disease genes from non-disease ones indicate that the synonymous sites of both the groups are under similar selective forces. It seems that disease associated single exonic genes have accumulated fewer mutations at the synonymous sites, but not significantly less, which could lead to major differences between the K_s average of two groups. This trend may be attributed partly to the presence of non-synonymous sites at the neighboring silent substitution sites, as K_a and K_s show a significant positive correlation between themselves, which has been noticed previously by others too (8). The results of the present study are thus in partial accordance with the previous findings of Tu *et al*. (9) who showed that disease genes are evolving at slower rate than other genes with significant differences in all the three parameters K_a, K_s, and K_a/K_s. However, it differs from that of Smith *et al*. (10) who showed that disease genes evolve faster than non-disease, at both synonymous and non-synonymous site resulting in a higher non-synonymous/synonymous rate ratio, and Huang *et al*. (11) who showed an increase in synonymous substitutions rate for disease genes and also suggested that selection has not discriminated between the two groups to a large extent. Moreover, Tu *et al*. (9) suggested that due to presence of essential genes within non-disease gene dataset, the signals regarding rate of protein evolution get blurred and are misjudged. Therefore, to ensure that observations made herein are consistent, we have compared the 252 non-disease intronless genes with the 1,789 human essential genes. The comparison revealed the presence of 8 essential genes in non-disease dataset. Again, after removing these genes when the 3 statistics were recalculated (N = 244, K_a = 0.104, K_s = 0.623 and K_a/K_s = 0.162) and compared, we observed similar results. Further, we also determined the K_a, K_s, and K_a/K_s rates of 56 morbid map disease associated intronless genes (see Materials and Methods) separately to be 0.075, 0.613 and 0.120 respectively. On comparing the non-synonymous and synonymous rates of the two groups (56 disease vs. 252 non-disease intronless genes), again it was found that disease associated human intronless genes have smaller and significantly different K_a and K_a/K_s while K_s although smaller is not significantly different. This demonstrated that observations made using both morbid map and gene map are similar and signify that human intronless disease genes are evolving slowly as compared to the other intronless genes. The results of the present study are thus in partial agreement with that of Tu *et al*. (9) but difference in terms of insignificant K_s rate has not been previously reported. Thus it indicates that human intronless genes are unique and distinct as proposed by various studies described in introduction (2-5).

To garner additional information regarding the evolution of human intronless disease genes various other sequence char-

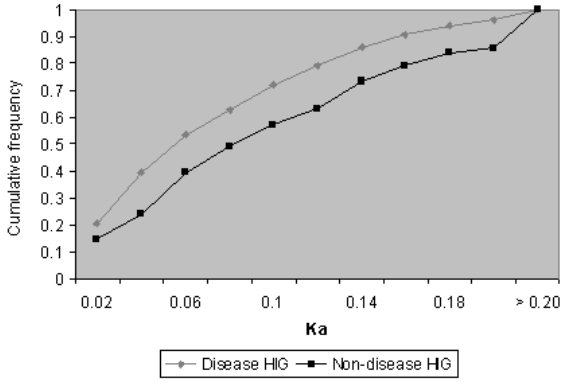

Fig. 1a. K_a distribution of ortholog pairs for disease versus non-disease human intronless genes.

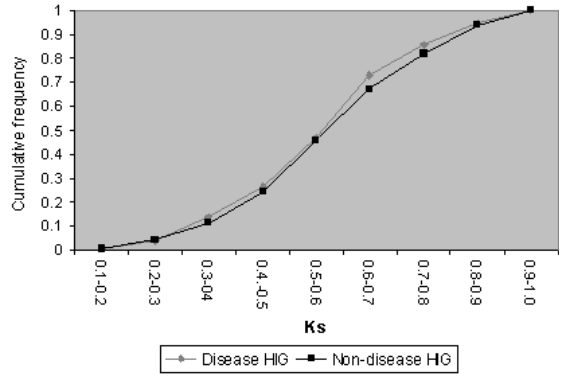

Fig. 1b. K_s distribution of ortholog pairs for disease versus non-disease human intronless genes.

acteristics were analyzed. Initially the extent of sequence conservation of the proteins across the various domains of life i.e. archaea, bacteria and eukaryotes was observed. For the purpose we utilized pre-computed sequence alignments generated from all against all BLAST comparisons available via Blink (12). It was found that majority of disease associated intronless genes (48%) exhibit homology to eukaryotic genomes only (metazoan, plants, fungi and other eukaryotes). On the other hand, the proportion of genes conserved across all the domains of life and those having homologues in bacteria and archaea only, were observed to be many more in case of non-disease genes. Thus, the above distribution of protein conservation in various taxonomic categories revealed that the proportion of extremely conserved proteins (i.e. proteins having homologs in all the domains of life) is greater in case of non-disease associated intronless genes while that having homologoues in eukaryotic genomes only are greater in disease associated intronless gene dataset. However, these observations differ from those previously reported for genome wide disease gene analysis where it has been proposed that the human disease genes are conserved across the broader phylogenetic extent than the rest of the human proteins (13). These results revalidate that the unusual nature of human intronless genes.

We further determined and compared the range of tissue expression of the disease and non-disease associated intronless genes using UniGene section in Genecards. Although, the averages of range of tissue expression did not show any significant difference between disease and non-disease dataset, however it was found that the proportion of disease associated genes expressed in either one or two tissues was higher than in non-disease genes (40% vs 34%) (Supplementary Fig. 1). This observation that the expression pattern of a substantial chunk of human intronless disease associated genes is localized is in accordance with our previous sequence conservation pattern (2) because earlier it has been suggested that domain specific proteins have increased probability of exhibiting tissue specificity (14). This signifies that majority of human intronless disease genes have homology limited to eukaryotic domain and exhibit narrow range of tissue expression.

We next asked if the trend observed was influenced by the domain composition of disease associated intronless genes. Since domain is the basic unit of protein that is self stabilizing and can exist as an independent functional feature, the Pfam database was used to assign the domains present in the disease associated proteins. The results presented in Table 2 showed that K_a/K_s is significantly affected by protein functions with histone and ion transporter associated disease genes being most slowly and cadherin domain containing disease proteins being the fastest evolving proteins. Thus the data suggests that histones and ion transporters are under negative selection which implies that high proportion of amino acids are constrained in these genes as it is expected for histones which is one of the most conserved protein family in nature (15) while transmembrane receptor and Cadherin domain containing proteins are under positive selection as shown previously (10). Moreover the 56 disease associated genes classified using morbid map were categorized into one of the 22 disorder classes described by Goh *et al*. (16). We observed that genes associated with four disorder classes namely cardiovascular, endocrine, immunological and neurological comprised nearly 50% of the total set. Further significant differences between the K_a/K_s ratio for these pathological system was observed. It was noted that genes involved in cardiovascular, endocrine and neurological disorder displayed on an average lower K_a/K_s ratio (0.083, 0.094 and 0.085 respectively) while genes associated with immunological disorder ($N = 8$, $K_a = 0.112$, $K_s = 0.732$ and $K_a/K_s = 0.153$) have on average higher K_a/K_s ratio and thus have undergone less intense purifying selection. Thus we conclude that different classes of disease related genes have experienced different selective pressures than genes that are not involved in causing diseases.

Table 2. Functional class evolutionary rate variation in disease associated HIG

S.No.	Pfam family	PFAM accession	Number of sequences	Mean K_a (SE)
1	7 transmembrane receptor	PF00001	79	0.082 (0.006)
2	Histone	PF00125	9	0.011 (0.004)
3	Helix-loop-Helix domain (HLH)	PF00010	9	0.052 (0.019)
4	Cadherin domain	PF00028	8	0.130 (0.004)
5	Connexin	PF00029	7	0.065 (0.019)
6	Ion transporter	PF00520	7	0.025 (0.008)

SE: standard error

Furthermore gene ontology enrichment analysis was applied to understand the categories that are preferentially associated with disease set. The 334 disease gene dataset was categorized into 925 gene ontology terms. The Fischer exact test was then applied on 334 genes considering the total set of genes (586 genes) as background set. This revealed that most of the disease associated intronless genes were linked to mainly 14 gene ontology terms (Bonferroni corrected P-value at 5% cutoff) that correspond to plasma membrane, signal transduction, receptors and ion channels. Additional enrichment analysis of 56 morbid map classified gene set within 334 disease set suggested that two gene ontology categories 'sensory perception of sound' and 'embryonic heart tube development' are overrepresented (Bonferroni corrected P-value at 5% cutoff). Thus we observe that intronless disease genes are enriched for higher level functionality that is essential for developmental processes in complex organisms.

CONCLUSIONS

From the study we conclude that disease genes are evolving at

lower nonsynonymous/synonymous (K_a/K_s) substitution rate and selection appears to have significantly discriminated between the two groups. It also confirms in line with some of the previous works that disease associated genes have significantly lower K_a and K_a/K_s however, for disease associated intronless genes the K_s is not significantly different validating the peculiar and distinct properties of human intronless genes. Further we demonstrate that majority of human intronless disease genes have homology limited to eukaryotic domain, exhibit narrow range of tissue expression and show varying evolutionary rate depending upon the functional category and pathological system.

MATERIALS AND METHODS

In the present study initial dataset of 1970 human intronless genes was considered (2). The presence/absence of homologous mouse protein sequence was then detected using Homologene (12). As few of intronless genes exhibit homology to intron containing genes, we have considered only those genes for analysis that have intronless gene structure in both humans and mouse. This is important as many human diseases are studied by experimenting on model organisms (such as mouse) because of the related closeness of these genomes and therefore it becomes necessary that the gene structures of homologous sequences are conserved. Further, to prevent contamination with paralogous sequences, genes that had more than one homologous sequence in mouse genome were eliminated. Thereafter, for each pair, the number of nonsynonymous substitutions per nonsynonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s) were extracted from the HomoloGene database. Pairs with high substitution rates ($K_a > 1$ and/or $K_s > 1$) were then discarded. Subsequently, the tissue expression breadth of these genes i.e. in how many out of the 12 tissues (bone marrow, spleen, thymus, brain, spinal cord, heart, skeletal muscle, liver, pancreas, prostate, kidney, lung) the gene is expressed, was investigated using UniGene section in Genecards (17). Further, genes containing expression information were selected and classified as disease or non-disease associated using OMIM (7). We have used both morbid map and gene map to identify disease association of a gene. This resulted in classification of 334 genes as disease associated (56 using morbid map and 278 using gene map) while 252 as non-disease genes. Finally, to test the statistical significance of the difference of K_a, K_s and K_a/K_s distributions, Mann-Whitney U test was performed. Also functional category to which each of the protein belongs was predicted by using protein family database (PFAM) (18) and in order to assess over representation of any of the Gene ontology terms between disease and non-disease associated intronless genes Fisher's exact test was employed (19).

Acknowledgements
We would like to thank the anonymous reviewer for his valuable inputs and suggestions.

REFERENCES

1. Sakharkar, M. K., Kangueane, P., Petrov, D. A., Kolaskar, A. S. and Subbiah, S. (2002) SEGE: a database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics* **18**, 1266-1267.
2. Agarwal, S. M. and Gupta, J. (2005) Comparative analysis of human intronless proteins. *Biochem. Biophys. Res. Commun.* **331**, 512-519. Erratum in: (2005) *Biochem. Biophys. Res. Commun.* **333**, 287.
3. Agarwal, S. M. (2005) Evolutionary rate variation in eukaryotic lineage specific human intronless proteins. *Biochem. Biophys. Res. Commun.* **337**, 1192-1197.
4. Hill, A. E. and Sorscher, E. J. (2006) The non-random distribution of intronless human genes across molecular function categories. *FEBS Lett.* **580**, 4303-4305.
5. Pozzoli, U., Riva, L., Menozzi, G., Cagliani, R., Comi, G. P., Bresolin, N., Giorda, R. and Sironi, M. (2004) Over-representation of exonic splicing enhancers in human intronless genes suggests multiple functions in mRNA processing. *Biochem. Biophys. Res. Commun.* **322**, 470-476.
6. Peltonen, L. and McKusick, V. A. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291**, 1224-1229.
7. Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V. A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic. Acids Res.* **30**, 52-55.
8. Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68-74.
9. 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.
10. Smith, N. G. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions. *Gene* **318**, 169-175.
11. Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L., Stenson, P. D., Cooper, D. N., Smith, D., Alba, M. M., Ponting, C. P. and Fechtel, K. (2004) Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol.* **5**, R47.
12. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Helmberg, W., Kapustin, Y., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2005) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **33**, D39-45.
13. Lopez-Bigas, N. and Ouzounis, C. A. (2004) Genome-wide identification of genes likely to be involved in hu-

man genetic disease. *Nucleic. Acids Res.* **32**, 3108-3114.
14. Cohen-Gihon, I., Lancet, D. and Yanai, I. (2005) Modular enes with metazoan-specific domains have increased tissue specificity. *Trends Genet.* **21**, 210-213.
15. Sullivan, S., Sink, D. W., Trout, K. L., Makalowska, I., Taylor, P. M., Baxevanis, A. D. and Landsman, D. (2002) The Histone Database. *Nucleic. Acids Res.* **30**, 341-342.
16. Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A. L. (2007) The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8685-8690.
17. Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., Ben-Dor, U., Esterman, N., Rosen, N., Peter, I., Olender, T., Chalifa-Caspi, V. and Lancet, D. (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18**, 1542-1543.
18. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic. Acids Res.* **32**, D138-141.
19. Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003) Global functional profiling of gene expression. *Genomics* **81**, 98-104.