

Using the plurality of codon positions to identify deleterious variants in human exomes

Author

Sankarasubramanian, Sankar

Published

2015

Journal Title

Bioinformatics

Version

Accepted Manuscript (AM)

DOI

[10.1093/bioinformatics/btu653](http://dx.doi.org/10.1093/bioinformatics/btu653)

Rights statement

© 2015 Oxford University Press. This is a pre-copy-editing, author-produced PDF of an article accepted for publication in Bioinformatics following peer review. The definitive publisher-authenticated version. Using the plurality of codon positions to identify deleterious variants in human exomes, Bioinformatics, Vol. 31 (3), 2015, pp. 301-305 is available online at: <http://dx.doi.org/10.1093/bioinformatics/btu653>.

Downloaded from

<http://hdl.handle.net/10072/69185>

Griffith Research Online

<https://research-repository.griffith.edu.au>

Using the plurality of codon positions to identify deleterious variants in human exomes

Sankar Subramanian^{1,*}

¹Environmental Futures Research Institute, Griffith University, 170 Kessels Road, Nathan Qld 4111, Australia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Motivation: A codon position could perform different or multiple roles in alternative transcripts of a gene. For instance a nonsynonymous position in one transcript could be a synonymous site in another. Alternatively, a position could remain as nonsynonymous in multiple transcripts. Here we examined the impact of codon position plurality on the frequency of deleterious single nucleotide variations (SNVs) using data from 6500 human exomes.

Results: Our results showed that the proportion of deleterious SNVs was more than twofold higher in positions that remain nonsynonymous in multiple transcripts compared to that observed in positions that are nonsynonymous in one or some transcript(s) and synonymous or intronic in other(s). Furthermore we observed a positive relationship between the fraction of deleterious nonsynonymous SNVs and the number of proteins (alternative splice variants) affected. These results demonstrate that the plurality of codon positions is an important attribute, which could be useful in identifying mutations associated with diseases.

Contact: s.subramanian@griffith.edu.au

Supplementary Information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

A major task in clinical genomics is to identify mutations associated with human diseases (Cooper and Shendure, 2011; Ward and Kellis, 2012). A number of computational prediction methods have been developed in the recent past to detect these deleterious mutations and to distinguish them from benign population polymorphisms (Adzhubei, et al., 2010; Bromberg and Rost, 2007; Cooper, et al., 2005; Kircher, et al., 2014; Kumar, et al., 2012; Ng and Henikoff, 2003; Schwarz, et al., 2010; Siepel, et al., 2005). Most of these methods examined the long-term evolutionary consequences of a mutation using the multiple-sequence alignments of human and other species. These methods predicted that mutations occurring in evolutionarily conserved positions are likely to be deleterious and hence have the potential to be associated with human diseases. Some of these methods evaluated the functional consequences of a mutation as well (Adzhubei, et al., 2010; Bromberg and Rost, 2007; Kircher, et al., 2014; Schwarz, et al., 2010). These methods proposed that mutations occurring in critical positions of a protein (such as a substrate binding site or a splice site) are likely to be deleterious as they disrupt the function and/or structure of a protein or mRNA. Furthermore these methods suggested that mutations that result in changes between dissimilar amino acids are more deleterious than those between simi-

lar amino acids. However unlike conservation-based methods, these methods are applicable only to protein-coding regions. Nevertheless these function-based methods have immense use in clinical genomics because almost 50% of the disease-associated mutations were found to be located in protein-coding regions (Stenson, et al., 2009; Subramanian and Kumar, 2006). Using conservation as well as function-based methods recent studies revealed an abundance of deleterious amino acid polymorphisms in human exomes (Coventry, et al., 2010; Fu, et al., 2013; Nelson, et al., 2012; Subramanian, 2012; Tennessen, et al., 2012).

Although these methods are useful in predicting deleterious variants, additional methods are still required to improve the accuracy of finding them. For this purpose it is important to examine the possibility of using various genomic features. In this study we focused on the roles of a position in human protein-coding genes. A codon position could perform different and/or multiple roles in alternate transcripts of a gene. For instance a nonsynonymous position in one transcript could be a synonymous or intronic position in another (Figure 1). On the other hand a site could remain nonsynonymous in multiple transcripts. It is not clear how this site plurality influences the frequency of occurrence of deleterious SNVs. Furthermore it is interesting to examine the relationship between the frequency of deleterious nonsynonymous SNVs and the number of proteins (alternative splice variants) affected. The availability of large exome datasets has enabled us to address these issues.

2 METHODS

2.1 Exome data: Exome data from 6,515 humans was obtained from the Exome Variant Server (evs.gs.washington.edu), which included exomes from 4,298 European Americans and 2,217 African Americans (Fu, et al., 2013). We extracted only single nucleotide variants (SNVs) and grouped them based on the pattern of sharing between various transcripts. We examined only the SNVs present in nonsynonymous, synonymous, intron, UTR and splice sites, which were denoted as “missense”, “coding-synonymous”, “intron”, “3-utr” (or “5-utr”) and “splice-3” (or “splice-5”) respectively. The annotations were based on SeattleSeq Annotation 137 (<http://snp.gs.washington.edu/SeattleSeqAnnotation137/HelpInputFiles.jsp>), in which splice sites were defined as two bases at 5’ and 3’ ends of introns.

We first grouped the SNVs present exclusively in one of the above genomic regions. For example the SNVs present only in nonsynonymous positions of multiple transcripts (eg. 2 in Figure 1). Then we grouped those present in any two of the above regions

*To whom correspondence should be addressed.

in different transcripts. For instance a SNV could occur in a non-synonymous position in one transcript but a synonymous position in another transcript of the same gene (eg. 1 in Figure 1). Alternatively, a SNV might be present in a splice site of one transcript but in an intron of another transcript (eg. 5 in Figure 1).

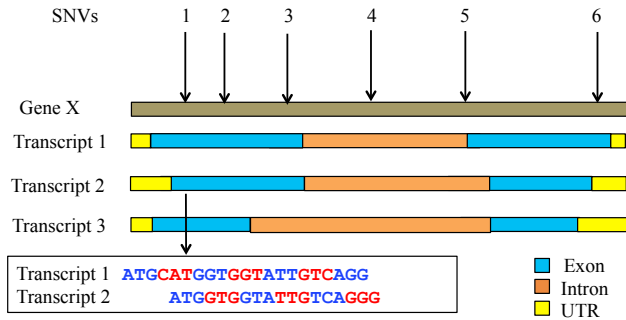


Figure 1. Illustration of various functional consequences of point mutations in a protein-coding gene.

SNVs located at:

- 1 - Synonymous+Nonsynonymous positions
- 2 - Nonsynonymous position - only
- 3 - Nonsynonymous position + Intron
- 4 - Intron -only
- 5 - Splice site + Nonsynonymous position
- 6 - Nonsynonymous position + UTR

Due to their complexity we excluded SNVs present in three or more different genomic regions. We also grouped nonsynonymous SNVs based on the number of alternatively spliced proteins affected by the mutation. For example SNV 2 in figure 1 affects nonsynonymous sites of all three transcripts whereas SNV 3 affects nonsynonymous sites of two transcripts and intron of one transcript.

2.2 Data analysis: To identify the deleterious nature of a SNV we used three methods, namely *GERP* (Cooper, et al., 2005), *Phastcons* (Siepel, et al., 2005) and *Polyphen* (Adzhubei, et al., 2010). Hence for the results shown in figures 2A&B we included only the SNVs for which both *GERP* and *Phastcons* scores were available. However for the results shown in figure 2C we included only the nonsynonymous SNVs for which a *Polyphen* prediction (such as “benign”, “possibly-deleterious” or “probably-deleterious”) was also available. To determine the deleterious nature of a SNV we used the following thresholds: *GERP* score >5.0 , *Phastcons* score >0.9 , and *Polyphen* score >0.95 or designated as “probably deleterious”. The proportion of deleterious SNVs was estimated using the ratio of deleterious SNVs to total SNVs. Standard error estimates were based on the binomial variance. A Z test was used to determine the significance of the difference between the proportions of deleterious SNVs estimated for a pair of comparisons. R studio was used for multiple regression analysis.

We used bi-allelic SNVs for all analyses reported. However restricting the analysis using only the derived alleles also produced similar results (data not shown). To identify the orientation of the SNVs we used the ancestral state of the nucleotides, which was inferred from six primate EPO alignments (Abecasis, et al., 2010). All analyses were performed using SNVs present in multiple transcripts and those present in single transcripts were only used for comparison in figure 2C (column 1).

To estimate the proportion of transcripts affected by a SNV, we assumed that synonymous sites, introns, and UTRs are largely neutral and nonsynonymous positions and splice sites are under selective constraint. If a gene has five transcripts and a SNV affects a nonsynonymous/splice site in two transcripts and a synonymous site/intron/UTR in the other three transcripts then the proportion of transcripts affected was calculated to be $2/5$ or 40%. Hence this measure was calculated as:

$$\frac{\text{No. of transcripts in which a nonsynonymous/splice site is affected}}{\text{Total number of transcripts of the gene}}$$

3 RESULTS

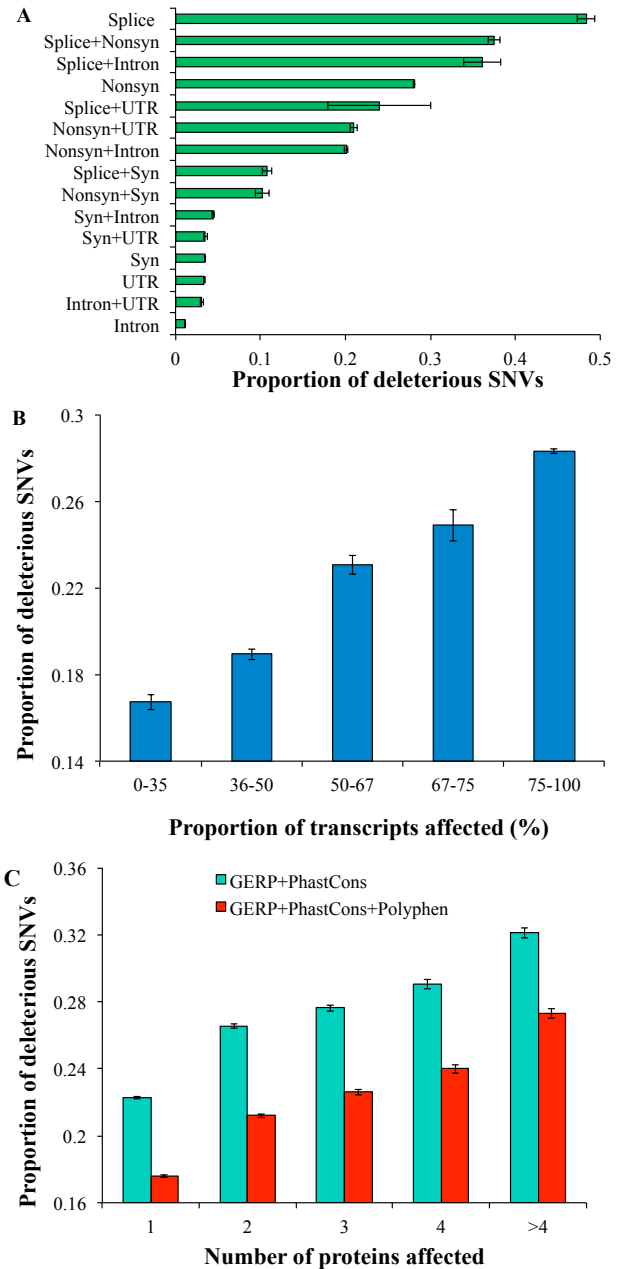


Figure 2. (A) Proportion of deleterious SNVs observed in different locations of human protein coding genes. The deleterious nature of a SNV was determined using *GERP* and *Phastcons* (see methods) (B) Relationship between the fraction of deleterious SNVs and the proportion of transcripts affected. (C) Relationship between the proportion of deleterious SNVs and the number of proteins affected. The deleterious fraction was estimated using conservation-based methods (green) and using conservation- as well as function-based methods (red). Note that the estimates shown in column 1 were computed using the SNVs present in single transcript genes. Error bars show the standard error of the mean.

3.1 Duality of codon positions

We first examined the effects of purifying selection on the codon positions that perform dual roles in different transcripts. For this purpose SNVs present in two different locations in multiple transcripts were grouped based on the type of codon positions. For example SNVs could be present in a nonsynonymous position of one transcript but in an intron of another transcript. We also grouped the SNVs present in only one type of exonic position (eg. nonsynonymous positions in multiple transcripts). We then computed the proportion of deleterious SNVs in each group of SNVs shown in Figure 2A and Supplementary Table 1.

It is evident that the SNVs present exclusively in splice sites are the most deleterious (48%). This is much higher than the proportion of deleterious SNVs affecting only nonsynonymous sites (28%), which suggests a greater purifying selection in splice sites. Interestingly the deleterious proportion (38%) of the SNVs affecting a splice site of one (or more) transcript(s) and a nonsynonymous position of the remaining transcript(s) was intermediate between the two groups mentioned above. This suggests that the magnitude of purifying selection on the sites is roughly the average of the selective constraints on the two types of positions.

The deleterious fraction was only 1-4% for the SNVs present exclusively in synonymous sites or in introns or UTRs suggesting a nearly neutral evolution at these sites. In contrast the deleterious fraction was 10-21% for the SNVs present in the sites that perform as synonymous sites/intron/UTRs in one or more transcripts and nonsynonymous sites in the remaining transcript(s). This proportion was intermediate as it was much higher than that computed using only noncoding sites (1-4%) but less than that estimated exclusively using nonsynonymous sites (28%).

3.2 Quantifying the magnitude of selection constraint in alternatively spliced genes

The above results suggest that a SNV in a genomic position is under selective constraint for one or few transcripts of a gene (if they affect a nonsynonymous or a splice site) and under nearly neutral evolution (if they affect a synonymous/intron/UTR site) for the remaining transcripts. Therefore the selection pressure on these sites is less than that on sites that are constrained (eg. nonsynonymous sites) in all transcripts but higher than that on sites that are nearly neutral or slightly deleterious in nature. Hence the sites that perform different roles in various transcripts experience intermediate or mean selection pressure. However the above results are rather qualitative and the mean selection pressure depends on the number of transcripts under selective constraint as well as those under nearly neutral evolution. Hence to systematically quantify this we estimated the proportion of transcripts affected by a SNV as shown in materials and methods. We then plotted this measure against the proportion of deleterious SNVs and observed a perfect positive correlation between them (Figure 2B and Supplementary Table 2). For example, the deleterious fraction was only 0.17 for the SNVs that affect the nonsynonymous sites (or splice sites) of 35% of the transcripts and the synonymous (or intron/UTR) sites of the remaining 65% of the transcripts of the gene. In contrast this fraction was 0.28, which is 60% higher for the SNVs that affect the nonsynonymous sites of all the transcripts of the gene. This result suggests that the proportion of transcripts affected by a SNV is an important measure to predict the frequency of deleterious mutations.

3.3 Multiplicity of codon positions

In the previous analyses we examined the selection pressure on codon positions that perform different functions in different tran-

scripts (plurality). Next we studied the pattern of purifying selection on the codon positions that perform the same function in multiple transcripts (multiplicity). For this purpose we included only the SNVs that affect nonsynonymous sites of *all* transcripts of a gene and excluded those SNVs that affect nonsynonymous sites in one (or more) transcript(s) and synonymous sites/introns/UTRs in the remaining transcripts of the same gene. The rationale for this analysis was to compare the amount of purifying selection on nonsynonymous sites of genes coding for single versus multiple proteins (splice variants). Our results showed that the deleterious fraction of nonsynonymous SNVs affecting single transcript proteins was 0.22, which was significantly ($P < 0.001$) less than that estimated for those affecting multiple splice variant proteins (0.28).

To examine this further, we divided the SNVs affecting multiple splice variant proteins into five groups based on the number of (splice variant) proteins affected and computed the proportion of deleterious SNVs for each group (Figure 2C and Supplementary Table 3). The results clearly showed a positive relationship between the fraction of deleterious SNVs and the number of proteins affected by the SNVs. The deleterious fraction of SNVs affecting more than four alternatively spliced proteins (0.32/0.27) was 44-55% higher than that of those affecting only single transcript proteins (0.22/0.18). For this analysis we also determined the fraction of deleterious nonsynonymous SNVs using the function-based method *Polyphen* in addition to the two other conservation-based methods. It is clear from figure 2C and supplementary table 3 that including this method also produced similar results. These results suggest a much higher purifying selection on the genes coding for multiple proteins compared to those coding for single proteins. Hence mutations occurring at nonsynonymous positions of multiple transcript proteins are more deleterious than those from single transcript proteins.

4 DISCUSSION AND CONCLUSIONS

In this study we examined approximately 0.75 million SNVs present in multiple transcripts (Supplementary table S1), which constitutes roughly 41% of the SNVs observed in 6,515 human exomes. This emphasizes the importance of considering the duality and multiplicity of codon positions. We showed that the duality of codon positions significantly influences the abundance of deleterious SNVs. We also quantified the magnitude of selection constraints based on the proportion of constrained transcripts and the number of alternatively spliced proteins and showed that the proportion of deleterious SNVs correlate positively with both of these measures. One of the interesting findings of this study is the much higher proportion of deleterious SNVs in splice sites compared to that in nonsynonymous positions. This is important because mutations in splice sites are known to affect more than 300 genes and are associated with over 370 diseases (Wang, et al., 2012). Furthermore over 13,000 splicing associated disease mutations have been reported in the human gene mutation database (Stenson, et al., 2014).

In this study we used a set of cutoffs to designate a deleterious SNV following previous studies and different methods use different sets of criterion to predict deleterious a SNV. However these limitations do not affect our conclusions because our results are only comparative. Nevertheless to address these issues we reanalyzed the data using a recently developed measure called Combined Annotation-Dependent Depletion (CADD) score (Kircher, et al., 2014). This method integrates diverse prediction methods (including the three methods used here) and produces a single score indicating the extent of functional consequences of a

SNV. To avoid using a cutoff value we compared the mean estimates of the CADD scores obtained for different site categories (Supplementary figure 1). Our results based this new measure also produced similar results (compare Figure 2A and Supplementary figure 1). For instance the CADD score of the SNVs affecting nonsynonymous+synonymous sites (12.2) was higher than the score estimated for the SNVs affecting exclusively synonymous sites (6.7) and lower than that obtained for the SNVs affecting nonsynonymous sites only (14.9).

In our analyses we used SNVs from human genes under strong as well as weak selection pressures. Hence it is important to examine whether the magnitude of selection pressure on genes (as opposed to that on independent sites or SNVs) influence the plurality or multiplicity of codon positions. To examine this we separated the SNVs into three groups based on the magnitude of selection constraints on their respective genes. For this purpose we used the ratio of nonsynonymous-to-synonymous divergence (dN/dS) estimated for each gene using the human-rhesus monkey comparison. This analysis produced results (Supplementary figures S2-S4 and Tables S4-S6) similar to those reported in Figure 2 (A-C). For instance the proportion of deleterious SNVs affecting exclusively nonsynonymous sites of highly constrained genes (dN/dS < 0.1) was much higher than that estimated for those affecting nonsynonymous sites in one (or more) transcripts and intron or synonymous sites in the other remaining transcripts (0.38 Vs 0.14-0.28 – Figure S2 and Table S4). A similar pattern (0.20 Vs 0.08-0.15) was also observed for weakly constrained genes (dN/dS > 0.2) and the estimates were proportionally lower than the former. We also performed logistic binomial regression analysis by using 0 or 1 as the probability of a SNV being non-deleterious or deleterious respectively. This was taken as a function of the proportion of transcripts affected (site-specific trait), dN/dS (gene-specific trait) plus the interaction between them. This analysis revealed that the proportion of affected transcripts is highly significant ($P < 10^{-16}$) in predicting the probability of a SNV to be deleterious in nature. We obtained similar highly significant result when using the number of proteins affected as a predictor. Finally we replaced the binomial variable (0 or 1) with the actual probability of a SNV to be deleterious computed by the software *Polyphen* and performed a multiple regression analysis. This result showed that the variable, the number of proteins affected is highly significant ($P < 10^{-16}$) even after controlling for the effects of dN/dS. These findings demonstrate that the gene-specific constraints do not influence our results.

To identify deleterious SNVs a number of methods have been developed in the past. These methods are either based on the evolutionary conservation of genomic positions and/or on the functional consequences of mutations. Here we showed a new genomic feature, the alternative splicing, which could potentially be used to identify deleterious SNVs. The results of this study suggest that the duality or multiplicity of codon positions could be used as an independent measure to detect deleterious variants. For instance if a mutation is located in a codon position that remains a nonsynonymous site in five or more transcripts (of a corresponding gene) then it is more likely to be deleterious than if the mutation is in a position that is nonsynonymous in only one transcript even if both positions have similar GERP (or other) scores. Hence along with other scores such as GERP, Phastcons and Polyphen the plurality and multiplicity of a codon position could be used to increase the probability of identifying disease-associated mutations.

ACKNOWLEDGEMENTS

The author is grateful to David Lambert and acknowledges the support from Environmental Futures Research Institute, Griffith University. I thank Leon Huynen for critical comments.

REFERENCES

- Abecasis, G.R., *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
- Adzhubei, I.A., *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249.
- Bromberg, Y. and Rost, B. (2007) SNAP: predict effect of nonsynonymous polymorphisms on function. *Nucleic Acids Res*, **35**, 3823-3835.
- Cooper, G.M. and Shendure, J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*, **12**, 628-640.
- Cooper, G.M., *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, **15**, 901-913.
- Coventry, A., *et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, **1**, 131.
- Fu, W., *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216-220.
- Kircher, M., *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, **46**, 310-315.
- Kumar, S., *et al.* (2012) Evolutionary diagnosis method for variants in personal exomes. *Nat Methods*, **9**, 855-856.
- Nelson, M.R., *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**, 100-104.
- Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, **31**, 3812-3814.
- Schwarz, J.M., *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*, **7**, 575-576.
- Siepel, A., *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.
- Stenson, P.D., *et al.* (2009) The Human Gene Mutation Database: 2008 update. *Genome Med*, **1**, 13.
- Stenson, P.D., *et al.* (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*, **133**, 1-9.
- Subramanian, S. (2012) The abundance of deleterious polymorphisms in humans. *Genetics*, **190**, 1579-1583.
- Subramanian, S. and Kumar, S. (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics*, **7**, 306.
- Tennessen, J.A., *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64-69.
- Wang, J., *et al.* (2012) SpliceDisease database: linking RNA splicing and disease. *Nucleic Acids Res*, **40**, D1055-1059.
- Ward, L.D. and Kellis, M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol*, **30**, 1095-1106.