Bioinformatics Advance Access published August 24, 2016

Bioinformatics, 2015, 1–4
doi: 10.1093/bioinformatics/btv602
Original Paper

Genome analysis

LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals

João Fadista[1,2,*], Nikolay Oskolkov[2], Ola Hansson[2] and Leif Groop[2,3]

[1]Department of Epidemiology Research, Statens Serum Institut, 2300 Copenhagen S, Denmark, [2]Department of Clinical Sciences, Lund University Diabetes Centre, Lund University, Sweden and [3]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 8, 2015; revised on October 12, 2015; accepted on October 13, 2015

Abstract

Motivation: Depletion of loss-of-function (LoF) mutations may provide a rank of genic functional intolerance and consequently susceptibility to disease.

Results: Here we have studied LoF mutations in 60 706 unrelated individuals and show that the most intolerant quartile of ranked genes is enriched in rare and early onset diseases and explains 87% of *de novo* haploinsufficient OMIM mutations, 17% more than any other gene scoring tool. We detected particular enrichment in expression of the depleted LoF genes in brain (odds ratio = 1.5; *P*-value = 4.2e−07). By searching for *de novo* haploinsufficient mutations putatively associated with neurodevelopmental disorders in four recent studies, we were able to explain 81% of them. Taken together, this study provides a novel gene intolerance ranking system, called LoFtool, which may help in ranking genes of interest based on their LoF intolerance and tissue expression.

Availability and implementation: The LoFtool gene scores are available in the Supplementary data.

Contact: joaofadista@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next generation sequencing technology has generated a vast amount of genomic data, but for clinical applications, interpretability of the results remains a major challenge. Whole exome sequencing has recently started to detect variants related to rare Mendelian and complex diseases, usually in a trio setting where both the patient and the parents are exome sequenced in order to identify *de novo* disease-associated mutations (Allen *et al.*, 2013; Fromer *et al.*, 2014; Neale *et al.*, 2012; Zaidi *et al.*, 2013; Zhang, 2014). Presently several tools exist to predict a single variant's pathogenicity, like PolyPhen (Adzhubei *et al.*, 2013) and SIFT (Ng and Henikoff, 2003), but they do not extrapolate their predictions to the gene level. There are also tools that prioritize candidate disease-causing genes such as ENDEAVOUR (Tranchevent *et al.*, 2008) or Prioritizer (Franke *et al.*, 2006), but these need a priori genetic knowledge of the disease. Recently, a new set of tools that prioritize candidate disease-causing genes with no a priori disease knowledge has emerged. By using the whole exome sequences of 6503 individuals, from the NHLBI GO Exome Sequencing Project (ESP) (Tennessen *et al.*, 2012), Petrovski *et al.* (2013) devised a 'residual variance intolerance score' (RVIS) from common missense and loss-of-function (LoF) variants versus total number of protein coding variants regardless of their frequency in the ESP population. A complementary approach for gene-level prioritization of disease genes calibrates *de novo* mutation rates per gene based on exome-sequenced trios in order to prioritize real excess of *de novo* mutations (Samocha *et al.*, 2014). EvoTol (Rackham *et al.*, 2015), another gene score, measures the gene's intolerance to mutations using evolutionary conservation of protein sequences. Here we developed an alternative gene score method called LoFtool, based on the mutation patterns in the exomes of 60 706 unrelated individuals from the Exome Aggregation Consortium (ExAC) dataset (http://exac.broadinstitute.org/). Although LoF intolerance is not a synonym of general intolerance, LoF mutations have the greatest pathogenic potential and hence are a good proxy for general genic intolerance to functional variation. LoFtool is based on the ratio of LoF to synonymous mutations for each gene, adjusting for the gene *de novo* mutation rate and evolutionary protein conservation (Section 2). We evaluate LoFtool against RVIS (Petrovski. *et al.*, 2013), *Z*-score (Samocha *et al.*, 2014) and EvoTol (Rackham *et al.*, 2015), and show that it clearly outperforms them in identifying *de novo* haploinsufficient disease-causing genes from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005), and from studies of neurodevelopmental disorders (Lee *et al.*, 2014; Soden *et al.*, 2014; Wright *et al.*, 2015; Yang *et al.*, 2014).

2 Methods

The LoFtool gene score. Our method to prioritize candidate disease-causing genes is based on the Exome Aggregation Consortium (ExAC) dataset (http://exac.broadinstitute.org/). This dataset comprises exome sequencing data for 60 706 unrelated individuals from several studies, reprocessed and jointly variant-called with the same analysis pipeline. We retrieved from ExAC all the synonymous and high confident LoF mutations called as such with the LOFTEE tool (https://github.com/konradjk/loftee). High confident LoF mutations were here defined as stop-gained, splice site disrupting and frame-shift variants that are not: (i) the ancestral allele (across primates), (ii) in the last 5% of the transcript, (iii) in exons and introns with non-canonical splice sites around it, (iv) in introns less than 15 bp, (v) in genes with only a single exon and (vi) in acceptor sites rescued by in-frame acceptor site. Then for each RefSeq known protein coding gene (except ribosomal proteins) with a HUGO Gene Nomenclature Committee (HGNC) name and belonging to the consensus coding sequence (CCDS) transcripts database (Pruitt *et al.*, 2009) we obtained the ratio of the number of LoF versus synonymous mutations. Afterwards, we corrected for each gene *de novo* mutation rate by dividing the LoF/synonymous ratio by the -log10 LoF *de novo* probability of each gene (Samocha *et al.*, 2014). Then we multiplied this LoF intolerance percentile by the EvoTol percentile and obtained a value that was transformed into our final LoFtool intolerance percentile. The genes with the lower LoFtool percentiles represent the genes that are most intolerant to functional variation (Supplementary Table S1). We only probed genes that had at least a 10-fold mean coverage in 70% of their CCDS in the ESP database, and had available EvoTol percentiles to allow comparisons with RVIS and EvoTol. In total we obtained 14515 genes with LoFtool scores. Because of the stringent LoF filters applied, we chose not to further analyze 906 genes with no LoF in 60 706 individuals from the ExAC database (non-ribosomal RefSeq known protein coding genes with at least a 10-fold mean exonic coverage in ExAC). Nevertheless, we report them in Supplementary Table S2 as these might be enriched in essential genes.

3 Results

3.1 LoFtool discrimination of different kinds of Mendelian diseases

To assess if LoFtool could differentiate between genes that do or do not cause Mendelian diseases, we compared the LoFtool percentiles in the different OMIM gene categories: 'haploinsufficiency,' 'dominant-negative,' 'de novo' disease causing, and 'recessive' (as adopted from Dataset S1 (Petrovski. *et al.*, 2013)). Genes that do not belong to these categories were defined as non-disease related ('non OMIM'). We observed that Mendelian diseases genes have significantly lower LoFtool percentiles than non-disease related genes (Fig. 1), with the strongest associations seen in haploinsufficient and *de novo* diseases (Wilcoxon rank sum test *P*-value = 1.7e − 49 and 1.5e − 91, respectively).

We also tested the better curated ClinGen haploinsufficient (HI) gene list available at http://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/, although the LoFtool scores from the OMIM HI versus ClinGen HI genes were not significantly different (*P*-value = 0.3078, Wilcoxon rank sum test). In agreement and as a positive control for this haploinsufficient feature, LoFtool percentiles were also significantly lower for genes in chromosome X (*P*-value = 1.17e − 27, Wilcoxon rank sum test). This suggests that genes intolerant to functional variation in the human population are more prone to cause Mendelian diseases than genes that endure functional variation. We then benchmarked our method against RVIS (Petrovski. *et al.*, 2013), *Z*-score (Samocha *et al.*, 2014) and EvoTol (Rackham *et al.*, 2015) for these different OMIM gene categories, as previously done to compare RVIS and EvoTol (Rackham *et al.*, 2015). In Figure 2 the percentage of OMIM genes detected by each tool's first gene

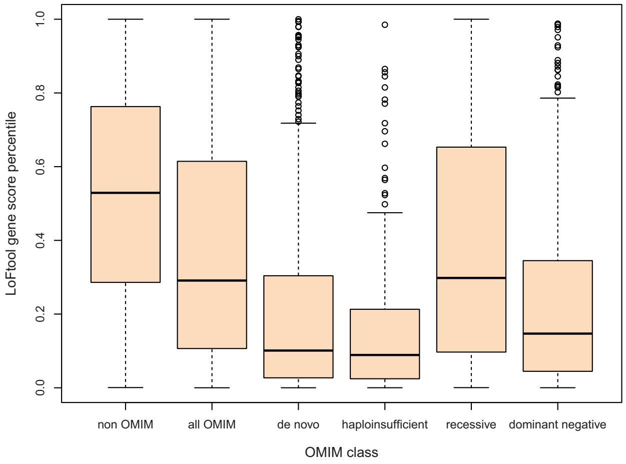

Fig. 1. Distribution of the different OMIM categories of Mendelian diseases in the LoFtool gene score percentile. The lower the LoFtool percentiles, the most intolerant are genes to functional variation

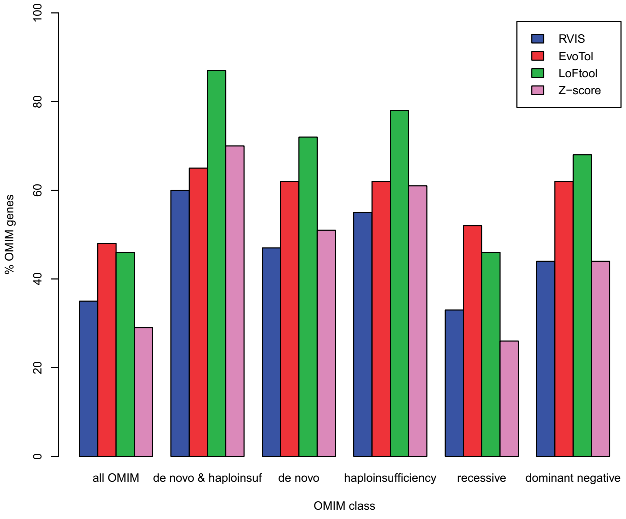

Fig. 2. Comparison between LoFtool, RVIS, EvoTol and *Z*-score intolerance scores in the different OMIM categories of Mendelian diseases (percentage of OMIM genes in each tool most intolerant quartile)

intolerance quartile is depicted. Overall, LoFtool has similar power in identifying all Mendelian disease genes as EvoTol (46% LoFtool, 48% EvoTol, 35% RVIS). However, LoFtool is better in identifying *de novo*, haploinsufficient and dominant negative OMIM mutations, with a clear superior performance for the *de novo* haploinsufficient genes, 87% of which were detected by LoFtool compared with 70% by *Z*-score, 65% by EvoTol and 60% by RVIS (Fig. 2). Despite the fact that online versions of RVIS and *Z*-score based on ExAC database already exist (http://chgv.org/GenicIntolerance/, ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/), their ExAC scores are very much similar to their original scores here tested.

To test whether LoFtool is better at predicting early or late onset diseases with autosomal dominant mode of inheritance we analyzed a hand-curated version of OMIM that has information about disease age of onset (Blekhman *et al.*, 2008). As can be seen for the most intolerant genes quartile (25%) in Supplementary Figure S1, the earlier the age at onset of disease the better is the predictive value of LoFtool, particularly before the age of reproduction (Fisher's Exact Test *P*-value = 4.72e − 12, odds ratio = 3.1). Since LoFtool excels in identifying *de novo* haploinsufficient mutations with relatively early age of onset, we hypothesized that the LoFtool gene score would be enriched in predicting rare diseases. To test this, we extracted genes from the ClinVar database where the gene mutation is classified by the Office of Rare Diseases Research at NIH (Landrum *et al.*, 2014). LoFtool's most intolerant gene quartile was enriched in rare disease genes (Fisher's Exact Test *P*-value = 9.9e − 34, odds ratio = 2.1), suggesting that it could be helpful in ranking new rare Mendelian disease genes. Moreover, LoFtool percentiles were also significantly lower for genes that produce lethal knockout mouse lines from the International Mouse Phenotype Consortium (Brown and Moore, 2012) (*P*-value = 4.2e − 18, Wilcoxon rank sum test). Taken together, this is in line with LoFtool's intolerance to functional variation, with consequent deleterious effect on fitness.

3.2 Enrichment of expression in brain and in genes causing neurodevelopmental disorders

We then asked if the most intolerant LoFtool genes showed enriched expression in any human tissue by exploring the Genotype-Tissue Expression project database (GTEx Consortium, 2015), which provides a comprehensive atlas of gene expression in different human tissues. For each of the 14 515 genes with LoFtool scores we computed if any of those were particularly expressed in a specific tissue (non-sex related tissue with minimum of 30 samples), defined as being at least 3 times more expressed than in any other tissue (Supplementary Table S1). As expression of the most intolerant LoFtool quartile was enriched in brain (Fisher's Exact Test *P*-value = 4.2e − 07, odds ratio = 1.5), we tested whether LoFtool would be a good ranking classifier of genes involved in neurodevelopmental disease. To do this we analyzed all *de novo* haploinsufficient genes from four recent large studies with exome sequencing data in cases (mostly trios) with neurodevelopmental disorders (Lee H *et al.*, 2014; Soden *et al.*, 2014; Wright *et al.*, 2015; Yang *et al.*, 2014) (Supplementary Table S3). The most intolerant LoFtool quartile of ranked genes explained 81% of these genes (ROC AUC = 0.86), while *Z*-score explained 84% (ROC AUC = 0.87), RVIS 64% (ROC AUC = 0.80) and EvoTol 54% (ROC AUC = 0.64) (Fig. 3). *Z*-score (AUC = 0.87) seems to perform slightly better that LoFtool (AUC = 0.86) in the capability to predict *de novo* haploinsufficient genes associated with neurodevelopmental disorders, although this difference is not statistical significance (*P*-value = 0.5808, Wilcoxon rank sum test).

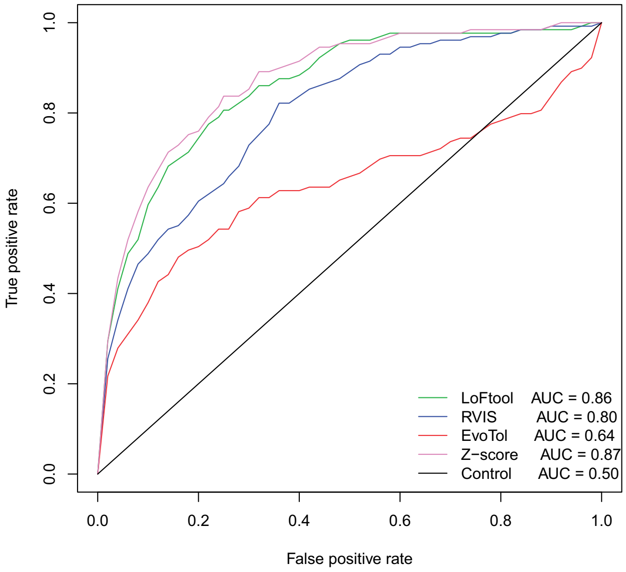

Fig. 3. ROC curves of LoFtool, RVIS, EvoTol and *Z*-score intolerance percentiles capability to predict *de novo* haploinsufficient genes associated with neurodevelopmental disorders (Yang *et al.*, 2014; Lee H *et al.*, 2014; Wright *et al.*, 2015; Soden *et al.*, 2014)

Moreover, the most intolerant LoFtool quartile was also enriched in previous independently reported extreme *de novo* mutations in neurodevelopmental disorders, as stored in NPdenovo database (Li *et al.*, 2015) (Fisher's Exact Test *P*-value = 2.3e − 3, odds ratio = 1.9). Most importantly, LoFtool have significantly lower scores (more intolerant) for genes associated with diseases reported in the NPdenovo database versus the respective controls, giving validity to LoFtool's performance. LoFtool performs the best when comparing its gene score percentile of controls versus autism (*P*-value = 0.004, Wilcoxon rank sum test) and controls versus schizophrenia (*P*-value = 6.45e − 05, Wilcoxon rank sum test), while it is second when comparing its gene score percentile of controls versus epileptic encephalopathy and intellectual disability, after Samocha's *Z*-score (Supplementary Fig. S2).

4 Discussion

We have shown that LoFtool discriminates between genes that do and do not cause disease (mostly early onset), outperforming other state-of-the-art gene scoring methods. LoFtool predicts genome-wide *de novo* haploinsufficient mutations accurately and could be of help in search for genetic causes of rare Mendelian diseases. Moreover, its brain expression enrichment coupled to a ROC AUC of 0.86 in detecting neurodevelopmental disorder genes makes LoFtool also an attractive method for investigating complex brain diseases with strong genetic effects. In addition, our LoFtool score has already been adopted by popular variant annotation tools like ANNOVAR, SnpEff, dbNSFP, UCSC Genome Browser Variant Integrator and Ensembl Variant Effect Predictor (Cingolani *et al.*, 2012; Karolchik *et al*, 2014; Liu *et al*, 2013; McLaren *et al.*, 2010; Wang *et al.*, 2010), which highlights the wide applicability of our gene intolerance score. Future work could focus on functional protein domains, rather than the overall gene mutation burden, in order to more easily prioritize the avalanche of rare variants putatively associated with a phenotype of interest (Lee *et al.*, 2014). Taken together, this study provides a novel gene intolerance ranking system that ranks genes based on their loss-of-function intolerance and tissue expression specificity.

Acknowledgements

The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about.

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115006 (IMI-SUMMIT), resources of which are composed of financial contribution from the European Union's 7th Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. This research has also received support from the European Research Council (EU 7th Framework Programme) under grant agreement n° 269045.

Conflict of Interest: none declared.

References

Adzhubei,I. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **7**, Unit7.20.

Allen,A.S. *et al.* (2013) *De novo* mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.

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

Brown,S.D. and Moore,M.W. (2012) The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. *Mamm. Genome*, **23**, 632–640.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Fromer,M. *et al.* (2014) *De novo* mutations in schizophrenia implicate synaptic networks. *Nature*, **506**, 179–184.

GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Karolchik,D. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.

Lee,H. *et al.* (2014) Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*, **312**, 1880–1887.

Lee,S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.

Li,J. *et al.* (2015) Genes with *de novo* mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry.*, doi: 10.1038/mp.2015.40.

Liu,X. *et al.* (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E23402.

McLaren,W. *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.

Neale,B.M. *et al.* (2012) Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature*, **485**, 242–245.

Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Petrovski,S. *et al.* (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**, e1003709.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Rackham,O.J. *et al.* (2015) EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res.*, **43**, e33.

Samocha,K.E. *et al.* (2014) A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.*, **46**, 944–950.

Soden,S.E. *et al.* (2014) Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci. Transl. Med.*, **6**, 265ra168.

Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

Tranchevent,L.C. *et al.* (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**, W377–W384.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wright,C.F. *et al.* (2015) Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, **385**, 1305–1314.

Yang,Y. *et al.* (2014) Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, **312**, 1870–1879.

Zaidi,S. *et al.* (2013) *De novo* mutations in histone-modifying genes in congenital heart disease. *Nature*, **498**, 220–223.

Zhang,X. (2014) Exome sequencing greatly expedites the progressive research of Mendelian diseases. *Front. Med.*, **8**, 42–57.