Database, 2020, 1–6
doi:10.1093/database/baaa105
Original article

Original article

ncVarDB: a manually curated database for pathogenic non-coding variants and benign controls

Harry Biggs[1,#], Padmini Parthasarathy[1,#], Alexandra Gavryushkina[1,2] and Paul P. Gardner[1,2,*]

[1]Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand and
[2]Bio-Protection Research Centre, University of Otago, PO Box 56, Dunedin 9054, New Zealand

[*]Corresponding author: Tel: +64 3 479 extn 7264; Fax: +64 3 479 7866; Email: paul.gardner@otago.ac.nz

Citation details: Biggs, H., Parthasarathy, P., Gavryushkina, A. *et al.* ncVarDB: a manually curated database for pathogenic non-coding variants and benign controls. *Database* (2020) Vol. XXXX: article ID baaa105; doi:10.1093/database/baaa105

[#]These authors contributed equally to this work.

Received 28 April 2020; Revised 13 October 2020; Accepted 12 November 2020

Abstract

Variants within the non-coding genome are frequently associated with phenotypes in genome-wide association studies. These non-coding regions may be involved in the regulation of gene expression, encode functional non-coding RNAs, or influence splicing and other cellular functions. We have curated a list of characterized non-coding human genome variants based on the published evidence that indicates phenotypic consequences of the variation. In order to minimize annotation errors, two curators have independently verified the supporting evidence for pathogenicity of each non-coding variant in the published literature. The database consists of 721 non-coding variants linked to the published literature describing the evidence of functional consequences. We have also sampled 7228 covariate-matched benign controls, that have a population frequency of over 5%, from the single nucleotide polymorphism database (dbSNP151) database. These were sampled controlling for potential confounding factors such as linkage with pathogenic variants, annotation type (untranslated region, intron, intergenic, etc.) and variant type (substitution or indel). The dataset presented here represents a curated repository, with a potential use for the training or evaluation of algorithms used in the prediction of non-coding variant functionality.

Database URL: https://github.com/Gardner-BinfLab/ncVarDB.

Context

The advent of high-throughput sequencing has allowed the capture of millions of genome variants (1–3). The accessibility of genome variation data has spawned an industry of genome-wide association studies (GWAS), where genetic variation and phenotypic variation, such as disease susceptibility, are linked by statistical association tests (4). The combined results of GWAS have revealed that many variants that are linked to phenotypic consequences reside outside the protein-coding regions (5–7). These non-coding

genetic variants can contribute to the phenotypic variation in a multitude of ways, including influencing alternative splicing and altering gene expression (8, 9). The study of non-coding variation has been hampered by a lack of molecular and computation tools for analysing the consequences of these variants (6).

Non-coding variants may influence gene expression and splicing, be functional non-coding RNAs (ncRNAs) or, as is frequently the case, be of unknown importance (6, 10). The experimental validation of every non-coding variant discovered *in silico* is currently not feasible; therefore, computational methods that can prioritize variants that are likely to have functional impacts are a research priority (11–13). These computational methods require reliable training and evaluation data to learn features that are indicative of a functional impact. Many of the existing non-coding variant annotation tools have been built using training and evaluation datasets constructed from public repositories such as ClinVar (14).

Recent benchmarks show that while these tools perform well with ClinVar variants (e.g. area under the curve [AUC] values >0.95), the tools do not perform as well against other databases such as the Catalogue Of Somatic Mutations In Cancer (15) (AUC values <0.78) (16, 17). The functional probing of saturation mutation of disease-associated gene promoters and enhancers has provided further independent data for evaluating the accuracy of non-coding pathogenicity prediction tools. This approach has also highlighted the relatively poor predictive performance of these methods (AUC values between 0.53 and 0.75) (18). The benchmarking of pathogenicity prediction tools for protein coding variants has highlighted the issue of overtraining of methods on the evaluation data, which is likely to also be a problem for non-coding methods (19).

Errors in biological databases are an ever-present concern for database curators, which impact the training, development and evaluation of different methods (20). The accuracy of functional annotations of genes (21, 22), taxonomic origins of sequences (23), and variant classifications (24) can unduly influence the conclusions of research that relies on accurate database information. This issue has led to calls to allow researchers to directly edit entries in leading sequence databases in order to correct errors (25, 26).

In order to partially address the issues of overtraining upon existing databases and minimize the number of errors in human non-coding variant classification databases, we have produced a manually curated variant classification database (ncVarDB). Two separate data curators have curated non-coding pathogenic variants directly from the published literature and from public data repositories. The current database release contains 721 pathogenic variants and 7228 the single nucleotide polymorphism database (dbSNP)-derived benign variants.

Data description

Two datasets were generated, one containing pathogenic variants supported by the published literature and one containing presumed benign variants. Multiple publicly available data repositories were used in the curation of these datasets. The pathogenic dataset was generated using the October 2019 release of ClinVar (14) and OMIM (27), also accessed between April and October 2019.

We selected non-coding variants from the ClinVar database (see Supplementary Data for details) and assessed the cited (by ClinVar) literature for confirmation of each variant. In cases where there was no citation available for the entry, the entry was excluded. In cases where the citation either did not contain genomic position information or contained information for a different mutation, the variant was also excluded.

To identify the well-characterized disease-associated variants that lie within non-coding genes, Online Mendelian Inheritance in Man (OMIM) was mined for ncRNA variants. We manually identified ncRNA variants and again confirmed that the pathogenicity of each variant was correctly mentioned in the citation. With further literature searches, additional three variants were included, such as variants in RMRP, the variation that may cause cartilage hair hypoplasia. After these methods, 721 pathogenic variants were kept for use.

We generated a set of benign non-coding variants from variants in the dbSNP151 database (1) using the University of California, Santa Cruz (UCSC) table browser tool (28, 29). Variants with a minor allele frequency (MAF) between 5 and 20% are likely to be benign, as stated in the 2015 American College of Medical Genetics guidelines (30). Any variant with a MAF between 5 and 20% in the entire dbSNP151 database was selected, with no alternate chromosomes included. This set was then randomly sampled with 10 benign variants being sampled for each pathogenic variant, the proportions of variant positions (e.g. intergenic, untranslated region [UTR] or intronic) and variant types (e.g. substitution, insertion or deletion) were kept the same. In order to control for linkage, no variant within 30 kb of a ncVarDB pathogenic variant was selected. These are estimated to have a <1% chance of being in linkage with the non-coding pathogenic variants (31). A comparison of the datasets is provided in Figure 1.

A potential confounding factor in this database is the lack of pathogenic variants that lie in intergenic regions. There is a low number of intergenic variants in the ncVarDB pathogenic dataset in comparison to other variant positions. Some potential reasons for this are discovery bias or verification bias. Because the original variant discovery was performed using database searches in ClinVar and OMIM,

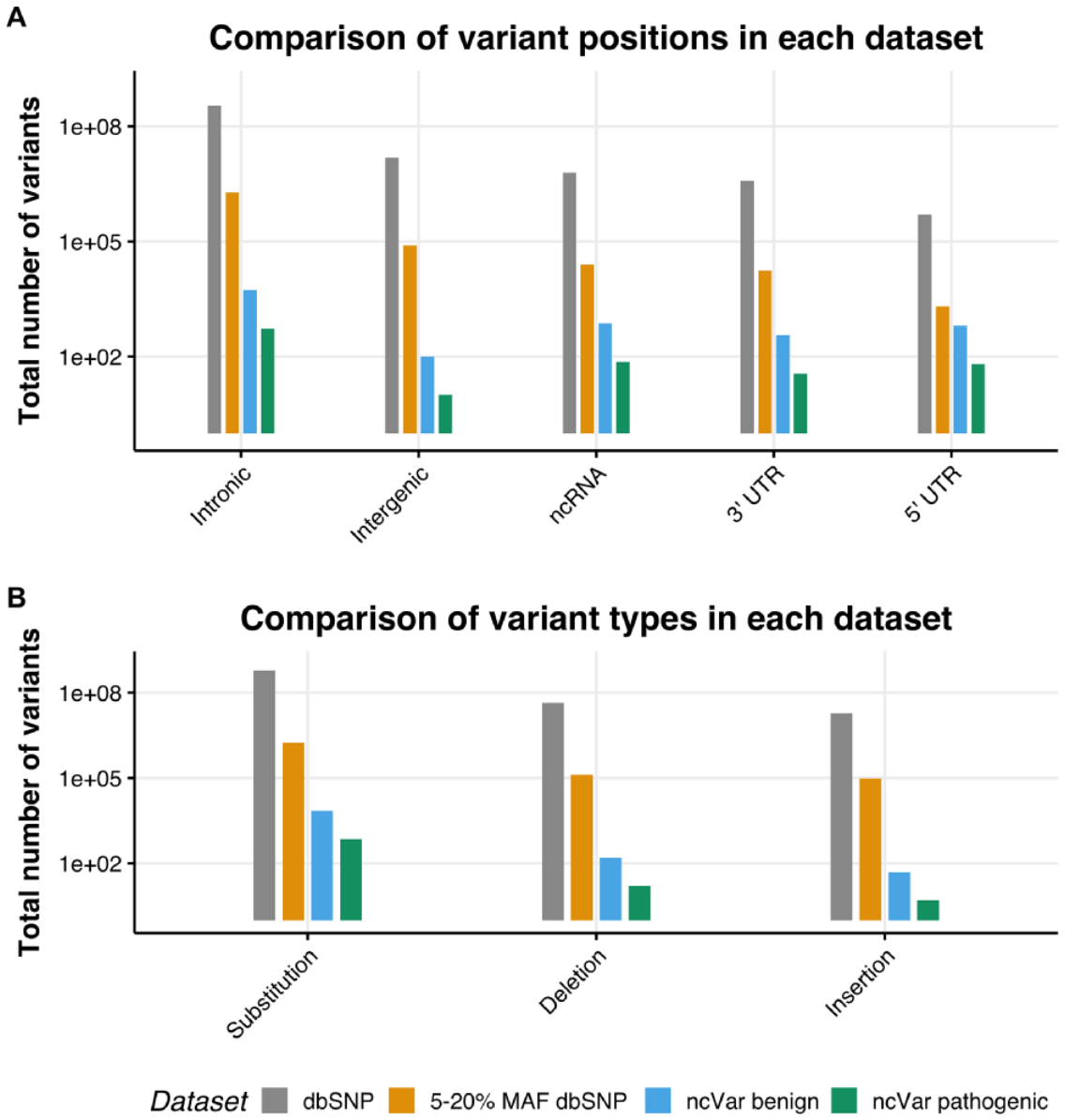

Figure 1. The location and single-nucleotide polymorphism (SNP) types of ncVarDB variants in comparison to variants from the dbSNP database. A comparison of the variant positions and the type of variants in every SNP in dbSNP dataset excluding variants from alternate contigs (dbSNP), every non-coding SNP with a MAF between 5 and 20% (5–20% MAF dbSNP), the ncVar benign dataset and the ncVar pathogenic dataset. (A) A comparison of the frequency of genomic positions of variants present in each dataset. Positions are based on the genomic notation submitted with the variant in either dbSNP or ClinVar. (B) A comparison of the frequency of variant types for each dataset. Variant types have been simplified to three types to avoid type expansion.

without searching specifically for intergenic regions, the variants lying in those regions may have not been captured by the original variant screening process. Another possibility is that as the variants in the database have been biologically validated for phenotypic changes, these variants are more likely to be in genic regions as these regions are traditionally of more interest to researchers studying genetic diseases.

The control benign dataset has been assembled automatically and not manually curated. Rare errors can occur in the benign dataset due to errors or ambiguities in the dbSNP database (see Supplementary data). There are very few variants on the mitochondrial chromosome in the dbSNP database that have a known molecular function and MAF value; as a result, ncVar benign dataset by chance does not contain variants on the mitochondrial chromosome.

The two datasets contain the following:

ID: An ID for this database

Genome: The genome that the variant was found in

Chr: The chromosome the variant is in

Pos: The starting position of the variant (referring to the first affected nucleotide). In case of a substitution (or a deletion), the starting position is the first nucleotide in the substituted (or deleted) sequence. In case of an insertion the starting position is the position of the nucleotide after which a new sequence is inserted.

Ref: The reference genome sequence

Alt: The variant sequence

Mutation_type: the type of mutation of the variant (substitution, insertion, deletion)

Mutation_position: The genomic position of the variant (intronic, 5utr, 3utr, ncRNA, intergenic)

MAF: The frequency of the minor allele (Alt)

X_ref: Any ID's from other databases e.g. dbSNP [REF] ClinVar [REF], OMIM [REF], Literature

The pathogenic dataset has extra two columns:

Pubmed_ID: A pubmed identifier that relates to literature that confirms the pathogenicity of the variant

Phenotype: The phenotype associated with the variant (sourced from XXXX)

The database can be found in ncVarDB (32).

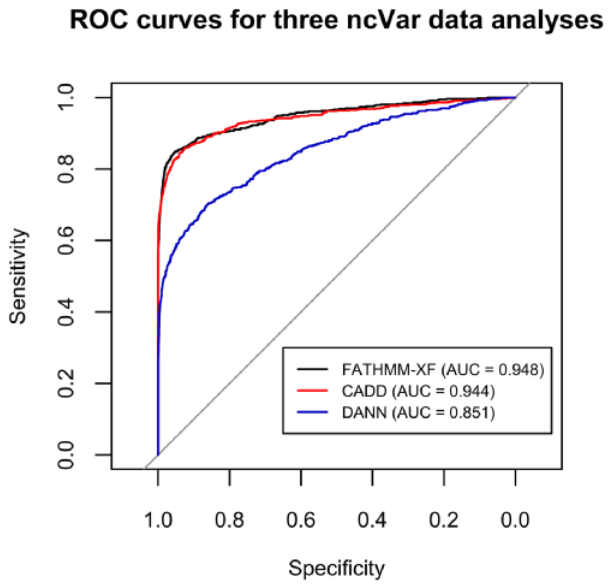

Figure 2. ROC curves for the classification analyses of the ncVar dataset by three different software tools: FATHMM-XF, CADD v1.4 and DANN. FATHMM-XF and CADD predict the pathogenicity of the ncVarDB variants with noticeably higher specificity and sensitivity than DANN. Overall good performance of all three tools additionally validates the ncVar dataset.

Data analysis

We classified ncVarDB variants using popular software tools: Functional Analysis Through Hidden Markov Models with an eXtended Feature set (FATHMM-XF) (33), Combined Annotation Dependent Depletion (CADD) v1.4 (11, 34) and Deleterious Annotation of genetic variants using Neural Networks (DANN) (12). FATHMM-XF and CADD v1.4 use statistical learning techniques (a support vector machine and a logistic regression model, respectively) to assign scores to variants based on conservation scores and other genomic features. Although there are differences in classification methods and the sets of features in the two software tools, the main difference is in the training sets. FATHMM-XF used previously identified pathogenic and benign variants from public databases. The training set for CADD consists of high frequency derived alleles in the human genome (compared to the inferred genome of the human-ape ancestor) as a 'proxy-benign' (neutral) group and simulated, free of selective pressure, variants as a 'proxy-pathogenic' group. DANN uses the same training set as CADD but uses a deep neural network algorithm for the classification.

The online FATHMM-XF tool was used for scoring our pathogenic and benign variants. This programme does not score insertions, deletions and more than one nucleotide long substitutions. It does not score variants on chromosomes X, Y and M. Excluding these variants and several additional variants that caused an error (see Supplementary data), we performed a receiver operating characteristic (ROC) curve analysis on 569 (79% of all ncVarDB pathogenic variants) pathogenic and 6823 (94% of all ncVarDB benign variants) benign ncVarDB variants that received FATHMM-XF score.

We used online CADD scoring implementation with raw scores for ROC curve analysis. CADD does not score variants on chromosome M. In total, 656 (91%) pathogenic and 7228 benign variants (all variants) were scored by CADD.

For DANN analysis, we downloaded precomputed scores provided by the authors for single nucleotide variants. DANN does not score variants on chromosomes Y and M. The positions of the DANN scored variants are provided relative to GRCH37/hg19 assembly. We converted positions of ncVarDB variants to positions relative to GRCH37/hg19 assembly using the UCSC hgLiftOver tool (29). Due to conversion errors, several variants were further excluded (see Supplementary data), which resulted in 633 (88%) pathogenic and 6989 (97%) benign DANN scored variants.

FATHMM-XF and CADD performed well on the ncVar dataset with AUCs of 0.948 and 0.944, respectively (Figure 2). The AUC for DANN analysis was 0.851 (Figure 2). Several previous comparisons of scoring methods showed closer AUC values (up to 0.06 difference) for CADD and DANN analyses (12, 34, 35).

The accurate classification of the ncVar dataset by the three popular scoring tools additionally validates the

dataset. These analyses are also an example of the potential use of the ncVar dataset for evaluation of the scoring method performance.

Data validation and quality control

To ensure a high level of fidelity, each variant was inspected by two different data curators. Each variant in this database contains a link to a PubMed article that was used to verify that variant.

Re-use potential

This database contains a test set and a control set containing pathogenic variants and benign variants. This has a wide range of potential uses, such as training algorithms for predicting a non-coding variant functionality.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

None declared.

Funding

This work was supported by a Dean's Bequest Fund, Otago School of Medical Sciences and a New Zealand Tertiary Education Commission Centre of Research Excellence grant to the Bio-Protection Research Centre.

Availability of supporting data

All entries in these datasets link back to several publicly available data repositories.

Conflict of interest The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author contributions

Harry Biggs: Data curation, visualization, writing, validation, investigation.
Padmini Parthasarathy: Data curation, writing, methodology, validation, investigation.
Alexandra Gavryushkina: Supervision, methodology, data analysis, data compilation, writing.
Paul P. Gardner: Conceptualization, funding acquisition, supervision, writing.

Data description

Two TSV files and two VCF files containing variant information. Pathogenic variant information has been manually curated.

References

1. Kitts, A., Phan, L., Ward, M. *et al.* (2014) *The Database of Short Genetic Variation (dbSNP)*. NCBI Bookshelf.
2. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
3. Sudmant, P.H., Rausch, T., Gardner, E.J. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
4. Visscher, P.M., Wray, N.R., Zhang, Q. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Human Genet.*, **101**, 5–22.
5. Hindorff, L.A., Sethupathy, P., Junkins, H.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.
6. 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.
7. Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Meth.*, **12**, 931–934.
8. Suzuki, H., Kumar, S.A., Shuai, S. *et al.* (2019) Recurrent non-coding U1-snRNA mutations drive cryptic splicing in Shh medulloblastoma. *Nature*, 574, 707–711. https://doi.org/10.1038/s41586-019-1650-0.
9. Muniz, L., Deb, M.K., Aguirrebengoa, M. *et al.* (2017) Control of gene expression in senescence through transcriptional read-through of convergent protein-coding genes. *Cell Rep.*, **21**, 2433–2446.
10. MacArthur, D.G., Manolio, T.A., Dimmock, D.P. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
11. Kircher, M., Witten, D.M., Jain, P. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
12. Quang, D., Chen, Y. and Xie, X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
13. Shihab, H.A., Rogers, M.F., Gough, J. *et al.* (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
14. Landrum, M.J., Lee, J.M., Riley, G.R. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
15. Bamford, S., Dawson, E., Forbes, S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.
16. Li, J., Drubay, D., Michiels, S. *et al.* (2015) Mining the coding and non-coding genome for cancer drivers. *Cancer Lett.*, **369**, 307–315.
17. Drubay, D., Gautheret, D. and Michiels, S. (2017) Abstract 388: a benchmark study for identifying cancer drivers in the non-coding part of the genome. *Cancer Res.*, **77**, 388.
18. Kircher, M., Xiong, C., Martin, B. *et al.* (2019) Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.*, **10**, 3583.

19. Grimm, D.G., Azencott, C.-A., Aicheler, F. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.
20. Weber, L.M., Saelens, W., Cannoodt, R. *et al.* (2019) Essential guidelines for computational method benchmarking. *Genome Biol.*, **20**, 125.
21. Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.
22. Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
23. Nilsson, R.H., Ryberg, M., Kristiansson, E. *et al.* (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One*, **1**, e59.
24. Shah, N., Hou, Y.-C.C., Yu, H.-C. *et al.* (2018) Identification of misclassified ClinVar variants via disease population prevalence. *Am. J. Human Genet.*, **102**, 609–619.
25. Pennisi, E. (2008) DNA DATA P: proposal to 'wikify' GenBank meets stiff resistance. *Science*, **319**, 1598–1599.
26. Finn, R.D., Gardner, P.P. and Bateman, A. (2012) Making your database available through Wikipedia: the pros and cons. *Nucleic Acids Res.*, **40**, D9–D12.
27. OMIM - Online Mendelian Inheritance in Man. *OMIM - Online Mendelian Inheritance in Man*. https://www.omim.org/ (5 November 2019, date last accessed).
28. Karolchik, D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
29. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
30. Nykamp, K., Anderson, M., Powers, M. *et al.* (2017) Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet. Med.*, **19**, 1105–1117.
31. Lynch, M., Xu, S., Maruki, T. *et al.* (2014) Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*, **198**, 269–281.
32. ncVarDB. *ncVarDB*; Github. https://github.com/Gardner-BinfLab/ncVarDB.
33. Rogers, M.F., Shihab, H.A., Mort, M. *et al.* (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.
34. Rentzsch, P., Witten, D., Cooper, G.M. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
35. Drubay, D., Gautheret, D. and Michiels, S. (2018) A benchmark study of scoring methods for non-coding mutations. *Bioinformatics*, **34**, 1635–1641.