OXFORD

Databases and ontologies

RevUP: an online scoring system for regulatory variants implicated in rare diseases

Solenne Correard, Brittany Hewitson, Robin van der Lee ⓘ **and Wyeth W. Wasserman** ⓘ *****

Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Abstract

Summary: To address the difficulty in assessing the implication of regulatory variants in diseases, a scoring scheme previously published allows the calculation of the Regulatory Variant Evidence score (RVE-score). The score represents the accumulated evidence for a causative role of a regulatory variant in a disease. Regulatory Evidence for Variants Underlying Phenotypes was built to calculate the RVE-score of regulatory variants, based on the 24 criteria, with a hybrid approach combining information retrieved from public databases and user input.

Availability and implementation: RevUP is freely available at http://www.revup-classifier.ca. The source code is available at https://github.com/wassermanlab/revup.

Contact: wyeth@cmmt.ubc.ca

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

1 Introduction

To date, several DNA variant classifiers are available and most are aligned to the American College of Medical Genetics and Genomics (ACMG) consensus recommendations (Richards *et al.*, 2015). Recently, recommendations were proposed for changes to the ACMG guidelines for clinical interpretation of variants to include variants found in non-coding regions of the genome, showing the relevance of such variants for genetic conditions (Ellingford *et al.*, 2021; Turner and Eichler, 2019).

With the growing use of clinical whole-genome sequencing, there is a need for extension of the classification systems suitable for regulatory sequence variants. Diverse regulatory variant prediction methods have been developed, such as PRVCS (Li *et al.*, 2016) or GREEN-VARAN (Giacopuzzi *et al.*, 2020), have begun to aggregate predictions and prioritize variants expected to disrupt gene regulation. Such bioinformatics tools generally draw upon genomics databases for information, but often the strongest evidence is found in gene-specific publications. Furthermore, the focus on regulatory dysfunction does not necessarily relate to pathogenic impact, which is critical for clinical classification. In 2020, a review of 46 regulatory disease variants reported to disrupt the expression of 40 transcription factor genes generated a semiquantitative classification scheme that incorporates both functional and clinical evidence (van der Lee *et al.*, 2020). The scheme allows the calculation of the Regulatory Variant Evidence score (RVE-score), based on 24 criteria, which

summarizes the accumulated evidence for a regulatory variant to have a causative role in a rare disease.

Regulatory Evidence for Variants Underlying Phenotypes (RevUP, http://www.revup-classifier.ca), a web-based classifier, was built to calculate the RVE-score of regulatory variants, based on the 24 criteria presented in van der Lee *et al.* (2020). Some previously released classifiers (not regulation focused) rely on user input only (Kleinberger *et al.*, 2016), while others rely uniquely on information available in public databases (Li and Wang, 2017), RevUP relies on both. The result page is a user-friendly display, downloadable and suitable for inclusion in a scientific report, together with a FAQ on the use of RevUP.

2 Materials and methods

2.1 Architecture and hosting

RevUP is a web application composed of a front end written in React, and a backend framework written in Flask. It is hosted on Amazon Web Services using a t2.medium elastic compute (EC2) instance with Nginx being used as the web server to route traffic. All code for RevUP is available on GitHub https://github.com/wassermanlab/revup.

2.2 Scoring scheme

The scoring scheme presented in van der Lee *et al.* (2020) is composed of a clinical component 'Is there a causal link between

A General Information

Variant ID:	test variant
Patient ID:	Proband 1
hg19 position:	9-139444949-C-A
hg38 position:	9-136550497-C-A
Reference Assembly:	hg19
Patient's Genotype:	Compound Heterozygous
Patient's Phenotype:	Tetralogy of Fallot (TOF)
Identification Method:	Whole Genome Sequencing
Target Gene:	NOTCH1

External Links
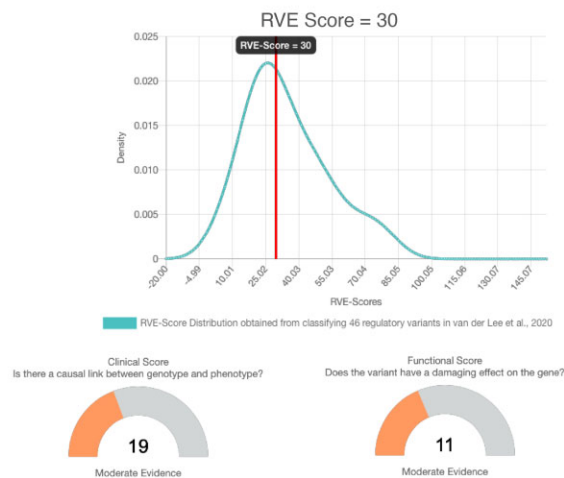
dbSNP (rs1338482227)

gnomAD

Scores

Fig. 1. RevUP report obtained for the scoring of the non-coding variant in the NOTCH1 gene reported by Wang *et al.* (2020) in proband 1. (**A**) Summary information for the variant; (**B**) clinical and functional information compiled from user input and external databases used to generate the RevUP score

genotype and phenotype?' and a functional component 'Does the variant have a damaging effect on the gene?'. Each evidence is given a score of 0 or 1 (0: The evidence does not apply to the variant or it is unknown; 1: The evidence applies to the variant, pondered based on the evidence weight). From the evidence, three scores are calculated: (i) the C-score, reflecting the clinical evidence available; (ii) the F-score, reflecting the functional evidence available and (iii) the RVE-score which is the accumulated evidence for a causative role of a regulatory variant in a disease. For some of the evidence, the information can be retrieved from public databases while other evidence requires user input (ex: C3.1, 'Variant shows familial segregation with the disease').

2.3 User input on the variant and target gene (Step 1) and databases queried by the webserver

For RevUP to query external databases, the user must input the variant details and the suspected target gene. Six external resources are then queried: PhyloP (Cooper *et al.*, 2005), PhastCons (Siepel *et al.*, 2005), gnomAD (Karczewski *et al.*, 2020), CADD (Rentzsch *et al.*, 2019), ReMap (Chèneby *et al.*, 2020) and ENCODE (Davis *et al.*, 2018). The details on the versions used and the information retrieved can be found in Supplementary Table S1.

2.4 User input for additional evidence (Step 2)

Some evidence cannot be found online. Others can be found using public databases but are too complicated to query automatically as they require interpretation by a human expert (ex: C2.1, 'Suspected target gene has been implicated in the same or a similar disease phenotype, or is otherwise relevant'). Therefore, the user will have the ability to answer 'Yes' or 'No' to these questions in order for the tool to calculate the score.

2.5 Results modification and comments (Step 3)

Users may have additional evidence that was not available in the queried public databases, therefore, the user can modify the score for each evidence level as well as add comments. Comments were created to allow citations to relevant publications, or specific figures within publications (i.e. free text). It is not advised to change the scores unless the user has strong evidence to justify it.

3 Results

Based on the user input and on the values retrieved from the external databases, RevUP calculates the C-score, the F-score and the RVE-score for the submitted variant. The result page is composed of two parts. The top portion (Fig. 1A) presents the information concerning the variant, the RVE-score and the strength of this score relative to the distribution of a curated collection of 46 regulatory variants (van der Lee *et al.*, 2020). This can indicate to the user if the score is high or low compared to published variants. Then, the C- and F-scores are displayed separately for the user to assess the strength and weakness of their analysis. The bottom portion (Fig. 1B) indicates, for each evidence level, the score, the information outputted from public databases and the comments added by the user (if any). The results are downloadable in PDF or png formats to present as-is or in modified form.

An example is shown in Figure 1, based on a recently published variant located in a regulatory region upstream of *NOTCH1* and implicated in the tetralogy of Fallot (Wang *et al.*, 2020). We input the variant characteristics and the information found in the paper in RevUP. The tool allowed for a quick scoring of the variant, and the creation of the displayed figure.

4 Conclusions and outlook

RevUP is a strong addition to the available variant classifier tools, as it will allow users to assess properties specific to regulatory variants. This scoring system does not conflict with the ACMG classification guidelines; rather it can be used as additional information when studying regulatory variants. The tool will save time for the user as it is able to query databases rapidly.

B Clinical Results

Evidence Description	Score	Comments/Additional Information
C1.1 - hg19.chr9:139444949 is evolutionarily conserved	1	phyloP Score: 1.64169 phastCons Score: 0.0
C1.2 - hg19.chr9:139444949.C>A is rare in unaffected individuals in specific sets of controls or reference population databases	1	gnomAD AF : 3.2847E-05
C1.3 - hg19.chr9:139444949.C>A or locus previously statistically associated with the same or a similar disease phenotype	0	-
C2.1 - NOTCH1 has been implicated in the same or a similar disease phenotype, or is otherwise relevant	1	NOTCH1 was reported as a transmembrane receptor and played an important role in differentiating into cardiaomyocytes (Rones et al., 2000, High and Epstein, 2008). Lack of NOTCH1 expression was associated with structural abnormalities of heart, such as bicuspid aortic valve (Garg et al., 2005) and NOTCH1 knockout mouse embryos had specific defects in heart formation (Grego-Bessa et al., 2007).
C2.2 - NOTCH1 does not contain coding variants in the same individual	1	NOTCH1 contains one coding varaint, which fits the compound heterozygous model
C2.3 - hg19.chr9:139444949.C>A is considered deleterious by computational prediction methods	0	CADD Score: 3.685
C2.4 - hg19.chr9:139444949.C>A is similar to another regulatory variant associated to NOTCH1 and implicated in the same or a similar disease phenotype	0	-
C2.5 - hg19.chr9:139444949.C>A is a striking noncoding event	0	-
C3.1 - hg19.chr9:139444949.C>A shows familial segregation with the disease	1	Compound heterozygous, variant inherited from an unaffected parent Variant segregates as expected in a trio (parents and proband)
C4.1 - hg19.chr9:139444949.C>A observed in multiple, unrelated families with the same disease phenotype	0	-
C4.2 - hg19.chr9:139444949.C>A introduction in a cell line leads to a cellular phenotype consistent with the disease phenotype	0	-
C5.1 - hg19.chr9:139444949.C>A neutralization, in a model organism or cell line, rescues or reverses the phenotype	0	-
C5.2 - hg19.chr9:139444949.C>A introduction, in a model organism or a cell line, results in a phenotype that is consistent with the human disease	0	-

Functional Results

Evidence Description	Score	Comments/Additional Information
F1.1 - hg19.chr9:139444949 is implicated in TF binding based on experimental data	0	ReMap 2020 Peaks: None
F1.2 - hg19.chr9:139444949 localizes to a regulatory region based on genome annotations	1	Regulatory region (chr9: 139444780-139446780) almost 4.5 kb upstream of NOTCH1 TSS cCREs: None
F1.3 - Regulatory region and NOTCH1 are directly linked based on annotation or experimental data	0	Supporting Experiment: -
F1.4 - hg19.chr9:139444949.C>A is statistically associated with expression levels of NOTCH1	0	Supporting Experiment: -
F1.5 - Regulatory region is shown to regulate gene expression of NOTCH1	1	Wang et al., 2020 - Figure 2
F2.1 - hg19.chr9:139444949.C>A causes a change in TF binding and/or chromatin environment	0	-
F2.2 - hg19.chr9:139444949.C>A leads to changes in expression of NOTCH1 in patient tissue	0	-
F3.1 - hg19.chr9:139444949.C>A introduction, in a cell line, leads to changes in expression of NOTCH1, a reported gene or chromatin environment	1	Wang et al., 2020 - Figure 2
F4.1 - hg19.chr9:139444949.C>A introduction, in a model organism, leads to changes in expression of NOTCH1 or a reported gene or chromatin environment	0	-

Calculations for hg19.chr9:139444949.C>A

	Value	Applicable Evidence Levels
Clinical Score	19	C1.1, C1.2, C2.1, C2.2, C3.1
Functional Score	11	F1.2, F1.5, F3.1

$$RVE\text{-}Score = \left(\begin{array}{c} \text{Weighted sum} \\ \text{of clinical evidence} \end{array} \right) + \left(\begin{array}{c} \text{Weighted sum} \\ \text{of functional evidence} \end{array} \right)$$

Clinical Score + Functional Score

Fig. 1. (continued)

References

Chèneby,J. *et al.* (2020) ReMap 2020: a database of regulatory regions from an integrative analysis of human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.*, **48**, D180–D188.

Cooper,G.M. *et al.*; NISC Comparative Sequencing Program. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

Davis,C.A. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

Ellingford,J.M. *et al.* (2021) Recommendations for clinical interpretation of variants found in non-coding regions of the genome. medRxiv 2021.12.28.21267792

Giacopuzzi,E. *et al.* (2020) GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants in whole-genome sequencing. bioRxiv 2020.09.17.301960.

Karczewski,K.J. *et al.*; Genome Aggregation Database Consortium. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

Kleinberger,J. *et al.* (2016) An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. *Genet. Med.*, **18**, 1165.

Li,Q. and Wang,K. (2017) InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.*, **100**, 267–280.

Li,M.J. *et al.* (2016) Predicting regulatory variants with composite statistic. *Bioinformatics (Oxford, Engl.)*, **32**, 2729–2736.

Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.

Richards,S. *et al.*; ACMG Laboratory Quality Assurance Committee. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Turner,T.N. and Eichler,E.E. (2019) The role of de novo noncoding regulatory mutations in neurodevelopmental disorders. *Trends Neurosci.*, **42**, 115–127.

Van der Lee,R. *et al.* (2020) Deregulated regulators: disease-causing cis variants in transcription factor genes. *Trends Genet.*, **36**, 523–539.

Wang,Y. *et al.* (2020) Family-based whole-genome sequencing identifies compound heterozygous protein-coding and noncoding mutations in tetralogy of Fallot. *Gene*, **741**, 144555.