**PERs**

Pathogenic variant enriched regions (PERs) tutorial, examples and additional files

**Introduction**

Missense variant interpretation is challenging. Essential regions for protein function are conserved among gene family members, and genetic variants within these regions are potentially more likely to confer risk to disease. Here, we generated 2,871 gene family protein sequence alignments involving 9,990 genes and performed missense variant burden analyses to identify novel essential protein regions. We mapped 2,219,811 variants from the general population into these alignments and compared their distribution with 76,153 missense variants from patients. With this gene family approach, we identified 465 regions enriched for patient variants spanning 41,463 amino acids in 1,252 genes. These regions were identified as Pathogenic variant Enriched Regions (PERs). Our entire set of results are available at the PER viewer (<http://per.broadinstitute.org/>) where the user can explore PERs across genes and gene families. Here we provide additional files associated to the study and the source code used to call PERs from a single gene family alignment.

**Contents**

* **db/**: Database folder containing the consolidated set of missense variants from ClinVar-HGMD databases (input.clinvar-hgmd) and the missense variants from gnomAD (input.gnomad.1-7). Paralog score database and amino acid code are contained in the folder for annotation purposes. Since HGMD private professional version was used, the genomic information and disease associations coming from HGMD missense variants are marked as “NA” (Not available). The db/ folder also contains all the burden analysis results for all family-wise aligments and all gene-wise protein sequences (family-wise-results-ALL,gene-wise-results-A-L,gene-wise-results-M-Z).
* **Glutamate-receptors.clw**: Example .clw file containing the protein sequence alignment of four glutamate receptors (*GRIN2A*, *GRIN2B*, *GRIN2C* and *GRIN2D*). The format of the file is the output of the MUSCLE aligner (<https://www.ebi.ac.uk/Tools/msa/muscle/>).
* **Voltage-gated-sodium-channels.clw**: Example .clw file containing the protein sequence alignment of ten Voltage-gated-sodium-channels (*SCN11A*, *SCN7A*, *SCN10A*, *SCN5A*, *SCN4A*, *SCN8A*, *SCN9A*, *SCN3A*, *SCN2A*, *SCN1A*). The format of the file is the output of the MUSCLE aligner (<https://www.ebi.ac.uk/Tools/msa/muscle/>).
* **SOX9.fasta**: Example .fasta file containing the protein sequence of the SOX9 gene.
* **Example-output/**: Expected output files from the examples provided in this repository.
* **Part-1-missense-aligner.pl**: Perl script able to format gene-family alignments “.clw” or gene sequences “.fasta” and map missense variants over them.
* **Part-2-burden-analysis.R**: R script able to perform missense burden analysis, PER identification and missense burden plots.

**Example**

The pipeline to detect PERs starts from any alignment of protein sequences performed with the software MUSCLE (.clw files) or any single gene protein sequence (.fasta files). This repository contains two example alignments: Glutamate-receptors with four members and the Voltage-gated-sodium-channels with ten members (see file contents). Additionally, we included the protein sequence of the SOX9 gene as a gene-wise example. The pipeline is structured in two parts: missense alignment and burden analysis which are performed by a Perl and R scripts, respectively.

**Part 1:**

* **Download**: Download the repository and uncompress the db folder contents. Make sure you have installed the Perl modules “*Data::Dumper*” and “*List::MoreUtils*” and the R packages "*ggplot2*", "*readr*" and "*ggrepel*" before running the code.
* **Command**: From terminal and inside the repository directory run:

$ perl Part-1-missense-aligner.pl

* **Output**: In the db/ folder, two files will be produced per aligment, one of them accounting for clinvar-hgmd missense variant mapping over the alignment (family-name.clinvar-hgmd.binary) and the other accounting for gnomad missense variants mapping over the same alignment (family-name.gnomad.binary).
* **Output legend**: The output file's *family-name.clinvar-hgmd.binary* and the *family-name.gnomad.binary* have the following structure:
  1. *Index*: Absolute position of the aligned aminoacids.
  2. *Gene-Sequence*: Canonical protein sequence in the format: “aminoacid\_position” for each of the genes belonging to the family (one per column).
  3. *Parazscore*: Paralog score observed for that Index position.
  4. *Gene-Binary Annotation*: At least one missense variant observed at corresponding aminoacid (YES=1, NO=0). One column per gene-family member.
  5. *Gene:Disease*: Gene ID coupled with the corresponding disease association observed. This is a collapsed field and more than one gene can have disease associations at the same index positions. Since no disease associations are present in the *family-name.gnomad.binary* file, the tag "GeneID:gnomad" is collapsed on this column.

**Part 2:**

* **Command**: From terminal and inside the repository directory run:

$ Rscript Part-2-burden-analysis.R

* **Output**: The R script will calculate the missense burden analysis over the whole alignment and identify PERs when the difference between the burden of the general population’s missense variants and patient’s pathogenic missense variants becomes significant. Burdens plots are produced in the same format as the one shown in the PER viewer (<http://per.broadinstitute.org/>). The complete burden analysis are written in a single file with “bin9.stats” extension denoting the 9 amino acid bin size used for the burden analysis.
* **Output legend**: The file *family-name.bin9.stats* has the following structure:
  1. *Index*: Absolute position of the aligned aminoacids.
  2. *Sequence*: Canonical protein sequence in the format: “aminoacid\_position” for each of the genes belonging to the family (one per column).
  3. *Parazscore*: Paralog score observed for that Index position.
  4. *Adj\_bin\_count*: Adjusted burden of missense variants observed in the general population.
  5. *DM\_adj\_bin\_count*: Adjusted burden of missense variants observed in patients from CLinvar-HGMD (Pathogenic, Likely Pathogenic and/or Disease Mutations).
  6. *Gene:Disease*: Gene ID coupled with the corresponding disease association observed. This is a collapsed field and more than one gene can have disease associations at the same index positions.
  7. *fisher.p*: Nominal p-value from fisher exact test.
  8. *or*: odd ratio.
  9. *ci1*: 95% lower confidence interval.
  10. *ci2*: 95% upper confidence interval.
  11. *adj.p*: Bonferroni adjusted p-value considering the amount of bins tested in the alignment (the longer the alignment, lower the alpha).
  12. *log.adj.p*: Logarithm of the adjusted p-value.
  13. *aa.per*: Index position belongs to a PER.
  14. *proxy*: Index position is the anchor of the bin tested. The bin size is 9 aminoacids, with the structure -4aa anchor-aa +4aa. Anchor aminoacid determine the stats of the whole bin.
  15. *per.tag*: Number of PER. If overlapping bins are significant, the bins are fused together keeping the strongest proxy.
  16. *per.start*: Start index Position of the PER.
  17. *per.end*: End index position of the PER.
  18. *size*: Aminoacid size of corresponding PER.

**Final remarks**

After running the provided example the user should be able to produce the files contained in "*Example-output*" folder. The complete study and detailed method description is currently available as a preprint entitled: “Identification of pathogenic variant enriched regions across genes and gene families”.