Paralog Annotation Notes

Contents

Aims

- Apply Paralogue annotation on other datasets
  - "Genome Wide" - Clinvar dataset and ALL possible exome variants (get from `/data/Mirror/ExAC_release/relea`
  - Cardiomyopathy genes - [MYH7, MYBPC3, TNNT2, TPM1, MYL2, MYL3, TNNI3, ACTC1] -CHECK CODE TO SEE IF ALL MISSENSE CM VARIANTS REALLY DO NOT APPEAR IN ANY OF THESE GENES AS EG. MYH6 SHOULD HAVE SOME
  - Channelopathy genes - [KCNQ1, KCNH2, SCN5A, KCNE1, KCNE2, RYR2]
- Improve precision via increasing conservation of ref/alt alleles
  - pairwise QC - ignore any individual pairwise alignments where ref alleles are not conserved
  - pairwise QC and family QC - ignore entire alignment columns if the entire family ref alleles are not conserved; NB analogous to para z scores
  - pairwise QC and family QC and alt allele QC - ignore entire alignment columns if family ref allele and alt allele isn't conserved
- investigate para z scores
- investigate pfam meta domains

Some interesting things to look at maybe

- "Ohnologs"
- Perform PA on paralogs from CPG (from Modos et al. (2016)) and non-CPG and see difference?
- Integrating ortholog data to increase confidence calling
- Gene Ontology packages - topGO?

OBSTACLES TO GET DONE:

- For the number of variants being patho or benign, how many of them actually align to anther variant? At least ones within paralogues? Difference between patho and benign variants? - Basically how many annotations there are…
- Look at if patho set have more variants that lie in genes that have more paralogs than benign
- Make sure noQC, para_con, and all_con output files have consistent total beginning number of variants
- Look at situation where only alt allele are conserved but ignore ref allele entirely
- GO - Need to map distribution of variants back to the genome, probably only take a single paralogue family as example.
- Debugging plugin
  - Have fixed "Can't call methods: start/location_from_column" by implementing `if` loops to catch errors
    - check out RBM20 and its "2" paralogues (MATR3 and MATR3…) that were example of genes which variants were causing these problems.
    - apparently there are other genes that have this same pattern where there are multiple genes with the same gene symbol but different IDs - James believes there exists a list of these genes somewhere (maybe ask Emily?)
    - see if can get a list of consistent variants that cause these specific issues for Erica.

  - “MySQL error has gone away” error seems to now have been fixed with the help from emily
    - issues seem to be due to server timeouts from running long jobs
    - sent me https://www.ensembl.org/Help/Faq?id=567 for help
  - ask emily about MART3…?
- STILL HAVE YET TO ACCOUNT FOR CONFLICTING P/B VARIANTS IN SCRIPTS
- Take a look at forking option for VEP to run faster?
- Make a list and write down overlapping genes that cause an issue like MART3, where only one of the overlapping genes has info reported back. Write it as an appendix.
  - Either report back in output file as special results or maybe TAKE OUT error catching for “Can’t call methods: start/location_from_column” errors and see if server still times out. That way warnings will be reported.
- All possible missense vcf for exome
  - be clear between all possible amino acid substitution and all possible nucleotide substitution
    - the sythetic exome from ExAC should contain all possible SNV
  - would be good to also write up stat for how many aa sub there is for every nt sub in report
- Xiaolei’s all possible missense cm vcf does not contain all possible snv
- Need to get all possible missense for genome as well, I.e. get a vcf containing all possible nt sub, not just all possible aa sub!
  - NEED TO MAKE MY OWN
- Rebenchmark with bigger test set (up to 100,000?)
  - edit the benchmarking script and rerun
- Get setup on imperial hpc and make sure plugin works
  - ~~setup perl api installation and make sure $PERL5LIB is correct; check erica chat history~~
  - ~~right installation instructions in github probably~~
  - run 30-38 failed, rerun
- ~~Data from denis~~
  - DONE - denis says that data I have is most up to date
  - DONE - Convert Para Z scores to 1 file for faster lookup and addition to tableize data.
- Data from henrike
  - waiting, will send me data when ready
- Web tool
  - Keep design simple
  - Ask Mark
- Make an experiment plan for transfering the framework over to structure space from sequence space
- Make all input data reproducible
- Look at bioconductor biomart package - easier to extract paralogue info from ensembl biomart

Manuscript Plan

- New tool to show (more likely Erica will write up)
- Contrast to previous studies, is genome wide validated
- Describe implementation and how to use
  - vep plugin lib

- Provide additional descriptive statistics of input (clinvar) data, e.g. number of genes with paralogues, number of disease genes etc.
- Paralogue annotate P/LP with P/LP; paralogue annotate B/LB with P/LP
    - generate confusion matrices for above
- can PA also predict benign variants as well as pathogenic?
- does paraZ score add additional benefit
- additional test/validation dataset
    - disease - Henrike?
    - ExAC/gnomad
    - all possible snv - synthetic vcf
- ICC genes - EFs
- Distributability
    - plugin
    - R shiny - vep web tool; integrated browser
    - integrated into gnomad
- Pfam domains - separate paper?

Paper Layout

Abstract

- *do last as usual…*

Introduction

- New variants are being rapidly discovered
- Ref previous papers and the work James/Roddy performed
- Have developed a new tool for researchers to use - ref Erica's paper
- Have expanded this to bigger data sets/genome wide
- In this paper, will show how Paralogue Annotation can be used as a way of variant classification

Material and Methods

- Ref Ensembl and Erica's paper for VEP+plugin
- Own pipeline (python/R)
- Data used
    - Ref clinvar
    - Ref Para Z scores
    - Ref Exac/Gnomad
        - all possible snv
    - Ref own clinical case/control cohorts
- statistical calculations
    - Precision/Sensitivity
    - EFs
- Webtool

Results and Discussion

- Additional descriptive statistics and background knowledge of below
- Analysis of Clinvar validation
  - pathogenic set
    - whole set; cardiomyopathy/channelopathy subset?
  - benign set (doesn't work)
  - Own filtering and Para Z scores improve precisions
- Analysis of EFs validation from OMIN data
- Gnomad/Exac
  - all possible snv
- Webtool
- Limitations
  - Quality of alignments
  - relianace on paralogues
    - reliance on variants in paralogues
- Solutions/Future Work
  - Different alignment algorithms?
  - (Don't mention Pfam meta domains specifically, but something along the lines of finding "optimal" homology)

Conclusion

- New variants being sequenced rapidly
- Concept of Paralogue Annotation works
- Novel idea of variant classification by homologous prediction
- Future work to be done

Introduction

With the advancements of sequencing technology, new potential variants are being discovered constantly. However to be able to identify said variants as pathogenic or benign requires supporting evidence, which does not always exists especially if the variant novel. Previously Ware *et al.* have developed **Paralogue Annotation** (J. S. Ware et al. 2012; R. Walsh et al. 2014), which utilizes information from paralogues (evolutionarily related genes from the same species) to help classify pathogenic variants. They verified its use in LQTS genes on variants acquired from patient cohorts.

Here **Paralogue Annotation** is tested further on a (Likely) Pathogenic/Benign varaint dataset from Clinvar.

Also have a look at Barshir et al. (2018) for more info about paralogs in diseases.

Material and Methods

The Paralog Annotation algorithm was wriiten by Erica as a perl script plugin (called **ParalogueAnno_plugin_cleanup.pr** for Ensembl's VEP version 90 (https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html).

The plugin has two arguments:

- the first parameter has 2 options:

  - `variant` (default) returns only the paralogous variants if any are present in the associated paralogs of the query gene found in the ensembl compara database
  - `paraloc` returns only paralog variant locations in the form of genomic coordinates of the corresponding codon in ALL paralogs;

- the second parameter has 2 options:
  - `all` for all variants;
  - `damaging` (default) for only damaging variant. The majority of the time `paraloc` mode is used.

The Ensembl team have touched up Erica's plugin and decrease runtime. The plugin is now called **ParalogueAnnotation.pm**

The initial output by VEP and the Plugin (VEP+Plugin) is not reader friendly for either the user nor if you want to parse informations. So a python wrapper, shwon below, for the VEP+Plugin was written to automatically parse the results, namely the paralogous variant information - **/data/Share/nick/Paralog_Anno/VEP_ParalogA**
the code below is not polished for release and is a WIP).

An intermediate python script (**File_prep_for_R.py**) was used to prep the results into R friendly data. Furthermore it could also be used to perform pairwise and family QC. Incidentally, the pairwise QC could be performed directly in R after the raw results are processed by **tableize_vcf.py** and tabulated (see below).

/data/Share/nick/Paralog_Anno/File_prep_for_R.py - formats results from **VEP_ParalogAnno.py** into tabulated format ready for R processing

As paraloc mode only returns ref alleles. The alt alleles were extracted from the VEP information. This was done by using **tableize_vcf.py**. **/data/Share/nick/Paralog_Anno/loftee/src/tableize_vcf.py** was used to format the VEP output into table format for R processing. For example:

```
python /data/Share/nick/Paralog_Anno/loftee/src/tableize_vcf.py --vcf /data/Share/nick/Paralog_Anno/data
```

If –**split_by_transcript** is used then the code above is sufficient. Otherwise a python wrapper that includes additional formatting (**/data/Share/nick/Paralog_Anno/Tableize_wrapper.py**) that tableize couldn't do, i.e. separate variants that had multiple REF and ALT alleles was used to prepare the data for R.

It is worth noting that using VEP with different versions of perl will result in slight different outputs. The difference do not seem to be detrimental to the end result as it appears that only VEP is affect but not the Plugin.

Datasets

Clinvar Likely Pathogenic/Pathogenic and Likely Benign/Benign variant vcf files were extracted and downloaded via the method developed by Zhang *et al.* (Zhang et al. 2017).

EFs calculated from OMGL and LMM datasets taken from previous Walsh et al. (2017) publication.

Exac and Gnomad data from Lek et al. (2016).

For Gnomad run:

```
library(DiagrammeR)
Gnomad_dataset_split = DiagrammeR::grViz("
digraph boxes_and_circles {
graph [overlap = true, fontsize = 10]

node [shape = plaintext, fillcolor = green, style=filled, fixedsize=false]
'RBH\ncluster: 9'; 'Imperial\nHPC: 29'; 'CX1\n(array): 19'; 'AX4\n(array): 10';
node [shape = plaintext, fillcolor = orange, style=filled, fixedsize=false]
'Total 38'; '1-9'; '10-19'; '20-29'; '30-38'

'Total 38' -> 'RBH\ncluster: 9'; 'Total 38' -> 'Imperial\nHPC: 29'; 'RBH\ncluster: 9' -> '1-9'; 'Imperia

}")
Gnomad_dataset_split
```

Benchmarking performance of the plugin

/data/Share/nick/Paralog_Anno/multi_vcf_extractor_benchmark.py is used to demonstrate speed at which VEP+Plugin takes to run

Scripts pipeline

```
library(DiagrammeR)
pipeline = DiagrammeR::grViz("
digraph boxes_and_circles {
graph [overlap = true, fontsize = 10]

node [shape = plaintext, fillcolor = green, style=filled, fixedsize=false]
'VEP_ParalogAnno.py'; 'File_prep_for_R.py'; 'Tableize_wrapper.py'; 'R markdown'

node [shape = plaintext, fillcolor = orange, style=filled, fixedsize=false]
'vcf input file'; 'paralogs file'; 'paraloc file'; 'paralogs2 file'; 'paraloc_tableized file'

'vcf input file' -> 'VEP_ParalogAnno.py'; 'VEP_ParalogAnno.py' -> 'paralogs file'; 'VEP_ParalogAnno.py'

}")
pipeline
```

Statistical terms

In context of is there a pathogenic paralogue alignment? A TP = pathogenic query variant with a paralogous pathogenic hit; FP = benign query variant with a paralogous pathogenic hit; FN = pathogenic query variant with no paralogous pathogenic hit; and TN= benign query variant with no paralogous pathogenic hit.

Likewise for a benign paralogous alignment, a TP = benign query variant with a paralogous benign hit; FP = pathogenic query variant with a paralogous benign hit; FN = benign query variant with no paralogous benign hit; and TN = pathogenic query variant with no paralogous benign hit.

Annotation of Clinvar

The Clinvar file **clinvar_20171029.vcf** was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/. Note that since the initial look at what was available there's been updated Clinvar files.

NOTE that I have noticed some descrepencies between the plugin annotations which call REFID = 1/0 and that of comparing the REF amino acid by VEP in the dataset to itself. This is due to the fact that the paralogous variant VEP is referring to is simply not in the dataset that I am annotating back to. As a result, it is best to make sure that the ref alleles are indeed the same when processing in R.

The annotataion of the entire clinvar dataset as of March 2018 release (clinvar_alleles.single.b38.vcf.gz) was taken from https://github.com/macarthur-lab/clinvar/tree/master/output/b38/single. Different clinical significance definitions were subsetted using grep e.g:

```
grep -P "CLINICAL_SIGNIFICANCE=Pathogenic" clinvar_alleles.single.b38.vcf > output1.vcf
grep -P "CLINICAL_SIGNIFICANCE=Likely_pathogenic" clinvar_alleles.single.b38.vcf > output2.vcf
cat output1.vcf output2.vcf > clinvar_alleles.single.b38.Pathogenic_and_LikelyPathogenic.vcf
```

Taking only the 8 sarcomeric genes:

Using only the 8 sarcomeric genes and joining to the whole clinvar dataset did not provide many annotations which could suggest either PA does not perform well on sarcomeric genes (paralogues to sarcomeric genes are not involed in disease) or that there is a lack of data. Therefore, it is not yet certain that PA does not work on sarcomeric genes and annotataion of additional sarcomeric data is required. See below.

Taking only the 5 channelopathy genes:

On the other hand, channelopathy genes did annotate well suggesting that their paralogues are involved in disease.

Looking at alt alleles. Taking only pairwise alignments where the alt allele is conserved leaves only 1115 individual pairwise alignments. The number of actual unique variants this equates to is less - 825.

Annotation of all possible missense variations in the 8 sarcomeric genes and calculation of EF

Limitations of this current framework are the reliance on 1) genes with paralogues and 2) for those paralogues to have pathogenic variants. This is not always the case, for example in HCM (DOUBLE CHECK THIS CLAIM), where there are not many pathogenic paralogous variants since most paralogues of HCM disease genes are not associated with diesase (according roddy….). Hence we calculated EFs in order to see for those few variants that are predicted to be pathogenic, how often do they appear to be causative of disease in a disease cohort case control study. Segway to HCM validation.

For calculating the EFs, run the all possible missense variants through VEP+plugin and return paraloc locations. Then join those locations with pathogenic clinvar variants as before. This indicates which variants from all possible missense variants are likely to be pathogenic. Then we check to see if any of these variants are present in the cases and controls. Hopefully the controls will be less but there is more control data than cases bare in mind. Calculate the EFs using that. Remember though the EFs are based on how many times an allele is seen, not the number of different alleles by themselves.

Total cases: 6140 number of affected cases: 39

Total controls: 60678 number of affected controls: 28

Odds ratio 13.7648 95 % CI: 8.4648 to 22.3833 z statistic 10.570 Significance level P < 0.0001 Attributable Risk Percent: 92.7% 95 % CI: 79.6 to 100

Paralogue stats

The additional statistics were calculated by programmatically extracting the genes of interest (using `src/check_what_clinvar_genes.py` and `src/Find_unique_genes.py`) and then retrieving relevant information manually from Ensembl's Bioimart

Alternatively, this can be reproduced using biomaRt package

Para-Z scores

For the para-z scores, will need to extract amino acid position from VEP output as well. Then look up the gene in question in para-z score folder, and using the position identify the para-z score. From my understanding, the para-z score is the same across aligned amino acids in the same gene family. Therefore, we could use a cut-off threshold to further improve our confidence in calling variants pathogenic etc. We could also then calculate ROC curves by altering the cut-off to see how that affects sensitivity/PPV.

All available para-z scores were retreived from https://git-r3lab.uni.lu/genomeanalysis/paralogs/tree/master/data (Lal et al. 2017). Para-Z score cutoff thresholds were used to remove any annotation alignments in question. Amino Acid positions that had a para-z score below the chosen cutoff threshold were not used for annotations.

Ohnologs

The "2R"" hypothesis states that some 500 million years ago, early vertebrates went through 2 rounds of whole genome duplication (WGD)(Ohno, Wolf, and Atkin 1968). Paralogues that arose from this WGD are known as ohnologs. Singh et al. (2014) showed that monogenic disease genes to be enriched in ohnologs than other paralogs that arose from small scale duplications.

Results and Discussion

Paralogue stats

```
## Warning in ks.test(clinvar_P_LP_genes_w_paralogs_wide
## $number_of_paralogues, : p-value will be approximate in the presence of
## ties
```

According to ensembl, 14514 protein coding genes are defined to have paralogues. While 6469 protein coding genes did not have paralogues. Of those genes with paralogues (**fig.** 1a) the mean had 6.297 paralogues with a standard deviation of 6.311. The maximum number of paralogues a gene had was 49.

In the clinvar pathogenic and likely pathogenic dataset, there were 102435 variants from 6665 genes. 3177 of these did not have paralogs and therefore the 28732 variants lying within these genes were not used for annotation, leaving 73703 for use in the analysis. The distribution of number of paralogues for these set of genes is shown in **fig.** 1b. The mean number of paralogues was 5.707 with a standard deviation 4.656.

For variants in the clinvar benign and likely benign dataset, there were 147115 variants from 7047 genes. 109830 variants resided in 3509 genes with paralogs. Their respective distribution is shown in **fig.** 1c, with a mean of 5.707 paralogues and a standard deviation of 4.656.

Performing a simple kolmogorov smirnov test between the distribution of pathogenic variants in genes with paralogues and benign variants shows a p-value of 0.9329528 suggesting that the null hypothesis of the distributions being identical cannot be rejected. From this, there appears to be no statistical difference between pathogenic variants being more likely to lie in genes that have more paralogs compared to benign, at least in regards to the definitions of clinical significance made by clinvar.

Annotation of Clinvar

Variant	Total	Paralogue_Annotation_no_QC	Variants_remaining_after_PA_QC1	Variants_removed
Pathogenic variants	22583	17477	16356	1121
Benign variants	13070	605	183	422
PPV	NA	0.966541311801792	0.988935243968801	0.72650680492547
Sensitivity	NA	0.773900721781871	0.724261612717531	NA
P value	NA	0	0	3.80673775046922

The full analysis of clinvar variants is shown in **table ??**In summary, 22583 Pathogenic and Likely Pathogenic (P/LP) variants and 17477 Benign and Likey Benign (B/LB) variants from clinvar had paralogue annotations. With no quality control, 17477 known P/LP and 605 known B/LB variants were predicted to be pathogenic, given a PPV and sensitivity of 0.9665413 and 0.7739007 respectively. Comparatively, predicting benign

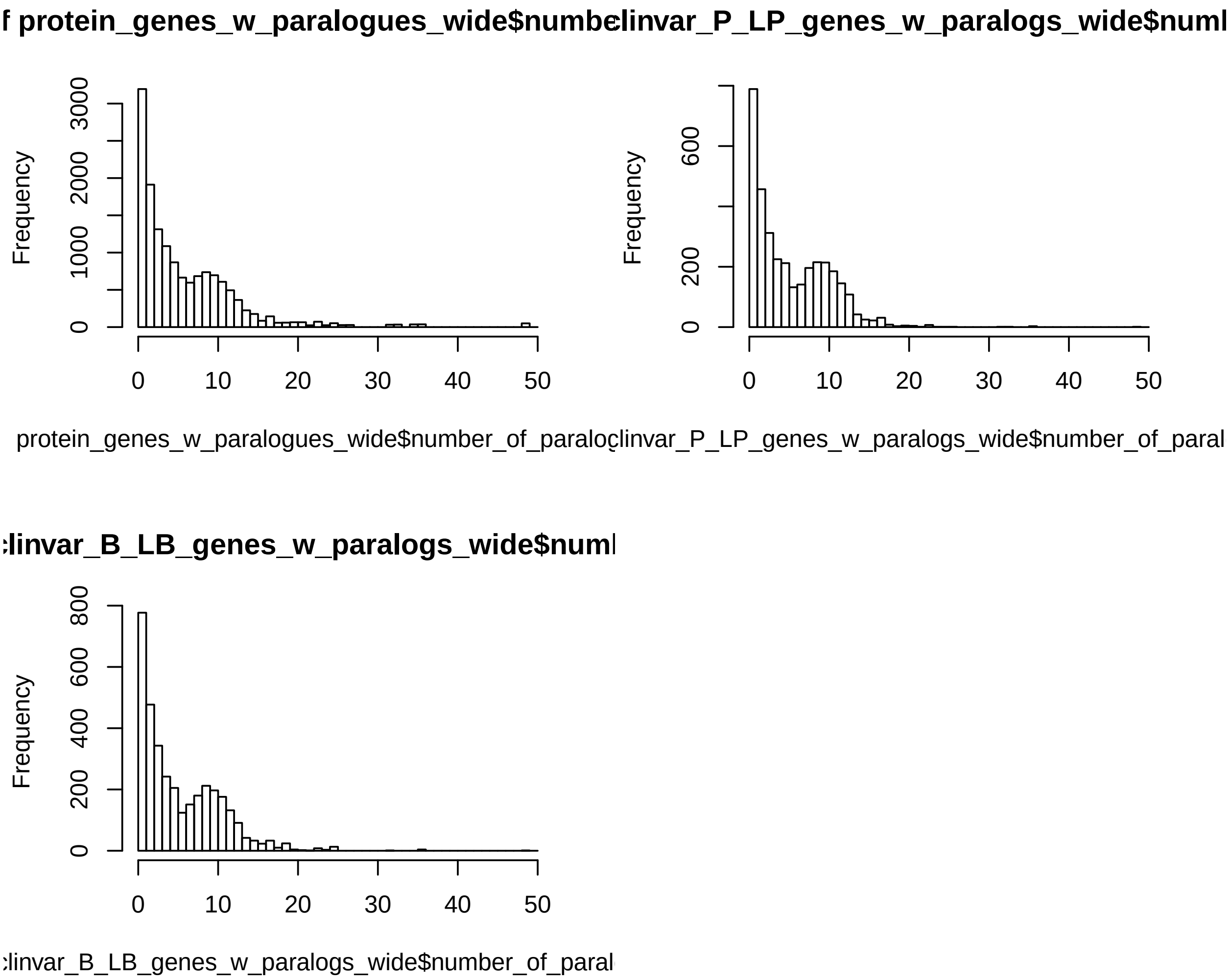
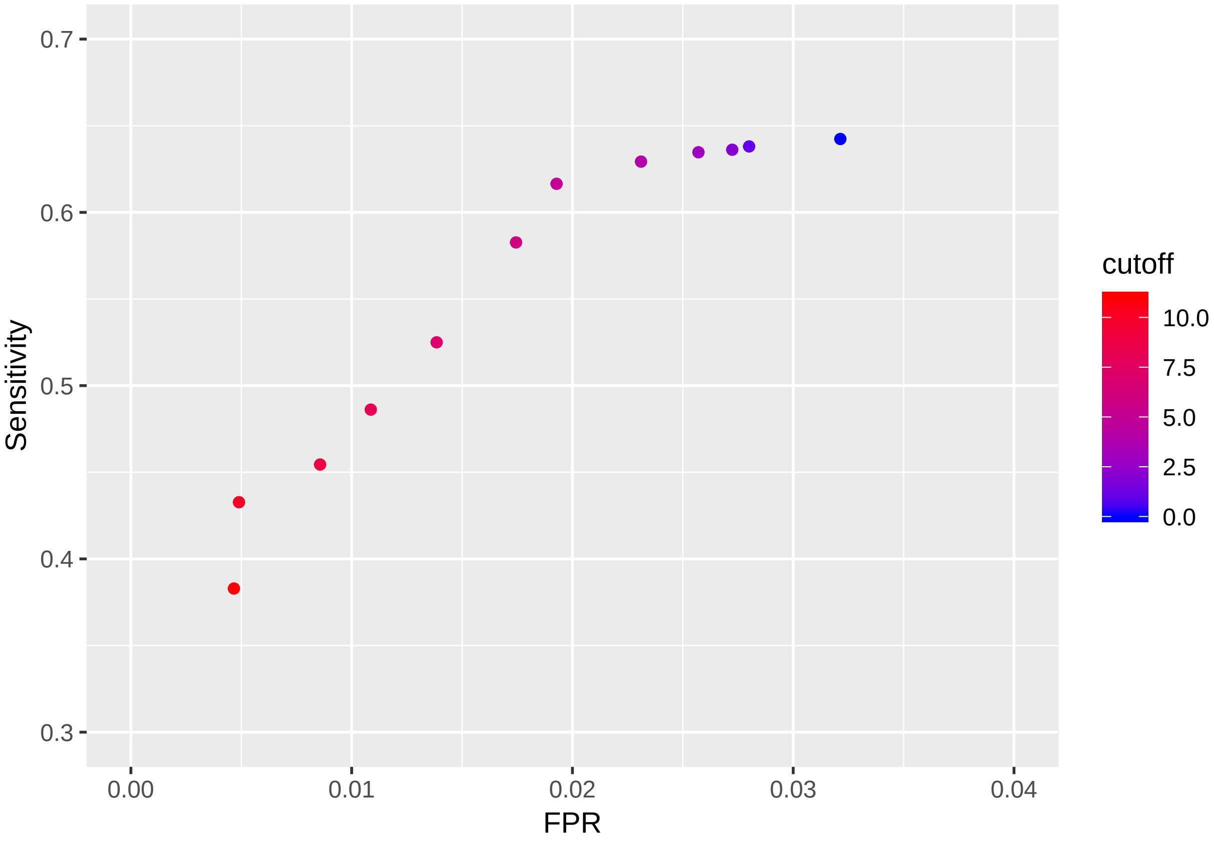


Figure 1: Distribution of genes with paralogues by the number of paralogues they’re related to

variants was not as reliable. With 1924 known P/LP and 1926 known B/LB variants

Para-Z scores



The filtering steps outlined above take a more binaray path into taking account the conservativeness of amino acid positions in the alignments. They only consider if amino acids in question share the the same amino acid or not. The Para-Z scores on the other hand take a more quantitative approach to this by representing a numeric integer value of how conserved each amino acid position is across the same paralogue family. Regardless both methods validate the concept that the more conserved amino acid positions are when transfering annotation the more likely annotations will be true positives as one would expect.

References

Barshir, Ruth, Idan Hekselman, Netta Shemesh, Moran Sharon, Lena Novack, and Esti Yeger-Lotem. 2018. "Role of Duplicate Genes in Determining the Tissue-Selectivity of Hereditary Diseases." *PLoS Genetics* 14 (5): e1007327.

Lal, Dennis, Patrick May, Kaitlin Samocha, Jack Kosmicki, Elise B. Robinson, Rikke Moller, Roland Krause, Peter Nuernberg, Sarah Weckhuysen, and Peter De Jonghe. 2017. "Gene Family Information Facilitates Variant Interpretation and Identification of Disease-Associated Genes." *bioRxiv*, 159780.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, and Beryl B. Cummings. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285.

Modos, Dezso, Johanne Brooks, David Fazekas, Eszter Ari, Tibor Vellai, Peter Csermely, Tamas Korcsmaros, and Katalin Lenti. 2016. "Identification of Critical Paralog Groups with Indispensable Roles in the

Regulation of Signaling Flow." *Scientific Reports* 6: 38588.

Ohno, Susumu, Ulrich Wolf, and Niels B. Atkin. 1968. "Evolution from Fish to Mammals by Gene Duplication." *Hereditas* 59 (1): 169–87.

Singh, Param P., Severine Affeldt, Giulia Malaguti, and Herve Isambert. 2014. "Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication." *PLoS Computational Biology* 10 (7): e1003754.

Walsh, R., N. S. Peters, S. A. Cook, and J. S. Ware. 2014. "Paralogue Annotation Identifies Novel Pathogenic Variants in Patients with Brugada Syndrome and Catecholaminergic Polymorphic Ventricular Tachycardia." *Journal of Medical Genetics* 51 (1): 35–44.

Walsh, Roddy, Rachel Buchan, Alicja Wilk, Shibu John, Leanne E. Felkin, Kate L. Thomson, Tang H. Chiaw, Calvin C. W. Loong, Chee J. Pua, and Claire Raphael. 2017. "Defining the Genetic Architecture of Hypertrophic Cardiomyopathy: Re-Evaluating the Role of Non-Sarcomeric Genes." *European Heart Journal* 38 (46): 3461–8.

Ware, James S., Roddy Walsh, Fiona Cunningham, Ewan Birney, and Stuart A. Cook. 2012. "Paralogous Annotation of Disease-causing Variants in Long Qt Syndrome Genes." *Human Mutation* 33 (8): 1188–91.

Zhang, X., E. V. Minikel, A. H. O'Donnell-Luria, D. G. MacArthur, J. S. Ware, and B. Weisburd. 2017. "ClinVar Data Parsing." *Wellcome Open Research* 2 (May 23): 33.