Further applications of Paralog Annotation

Contents

1 Introduction

With the advancements of sequencing technology, new potential variants are being discovered constantly. However to be able to identify said variants as being pathogenic requires supporting evidence, which does not always exists especially if the variant is novel. Several commonly used bioinformatic tools for predicting the functional effect of variants, such as SIFT and Polyphen (Sim et al. 2012,Adzhubei et al. (2010)), utilize information from conservation and/or changes in biophysiocochemical properties of the substituting amino acid. 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 LQT syndrome (LQTS) genes on variants acquired from patient cohorts.

Here **Paralogue Annotation** is tested further using a Pathogenic/Likely Pathogenic (P/LP) and Benign/Likely Benign (B/LB) varaint dataset acquired from Clinvar. These variants covered a wider range of genes other than just those involved in LQTS

2 Material and Methods

2.1 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). Variants that were positioned in any of the following 8 sarcomeric: MYH7; MYBPC3; TNNT2; TPM1; MYL2; MYL3; TNNI3; and ACTC1, were also subsetted for further analysis as the case example and used for etiological fraction calculations.

Clinical significance definitions of variants - Pathogenic; Likely Pathogenic; Likely Benign; Benign - were defined according to Clinvar (Landrum et al. 2015). Variants of unknown significance were not used in the study.

For the case and control study, HCM diseased cohorts data were gathered from the OMGL and LMM datasets, which were taken from previous Walsh et al. (2017) publication. Data from ExAC were used as the control cohort Lek et al. (2016).

2.2 Annotation of variants and transfer of annotations across paralogues

2.2.1 Statistical measures

For a pathogenic paralogue alignment:

true positive (TP) = pathogenic query variant with a paralogous pathogenic hit;

false positive (FP) = benign query variant with a paralogous pathogenic hit;

false negative (FN) = pathogenic query variant with no paralogous pathogenic hit;

and true negative (TN) = benign query variant with no paralogous pathogenic hit.

Likewise for a benign paralogous alignment:

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.

Positive Predictive Values (PPV) and Sensitivties are therefore calculated by:

$$PPV = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

P values are calculated via a Fisher's exact test on a 2x2 contingency table tabulating the number of pathogenic and benign variants of interest and how many of those are predicted as pathogenic or benign.

2.3 Calculation of Etiological Fractions

Odds ratios (OR) are calculated by:

$$OR = \frac{(a/b)}{(c/d)}$$

where:

$a =$ number of variants predicted to be pathogenic by paralogue annotation in the diseased cohort

$b =$ number of variants not predicted to be pathogenic by paralogue annotation in the diseased cohort

$c =$ number of varaints predicted to be pathogenic by paralogue annotation in the control cohort

$d =$ number of variants not predicted to be pathogenic by paralogue annotation in the control cohort

and the 95% confidence intervals for OR values are calculated according to Altman (1991) via:

$$95\%\ CI = [e^{\ln(OR)-1.96\cdot SE(\ln(OR))}, e^{\ln(OR)+1.96\cdot SE(\ln(OR))}]$$

where:

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Etiological Fractions (EF) can then be calculated by:

$$EF = \frac{OR - 1}{OR}$$

and the 95% confidence intervals for EF values are calculated according to Hildebrandt et al. (2006) via:

$$95\%\ CI = [\widehat{EF} - z_{1-\frac{a}{2}}\sqrt{\widehat{VAR}(\Phi)},\ min(\widehat{EF} + z_{1-\frac{a}{2}}\sqrt{\widehat{VAR}(\widehat{\Phi})},\ 1)]$$

where the asymptomatic variance estimator of $\widehat{\Phi}$ is given by:

$$\widehat{VAR}(\widehat{\Phi}) = \widehat{\Phi}^2$$

2.4 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

2.5 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.

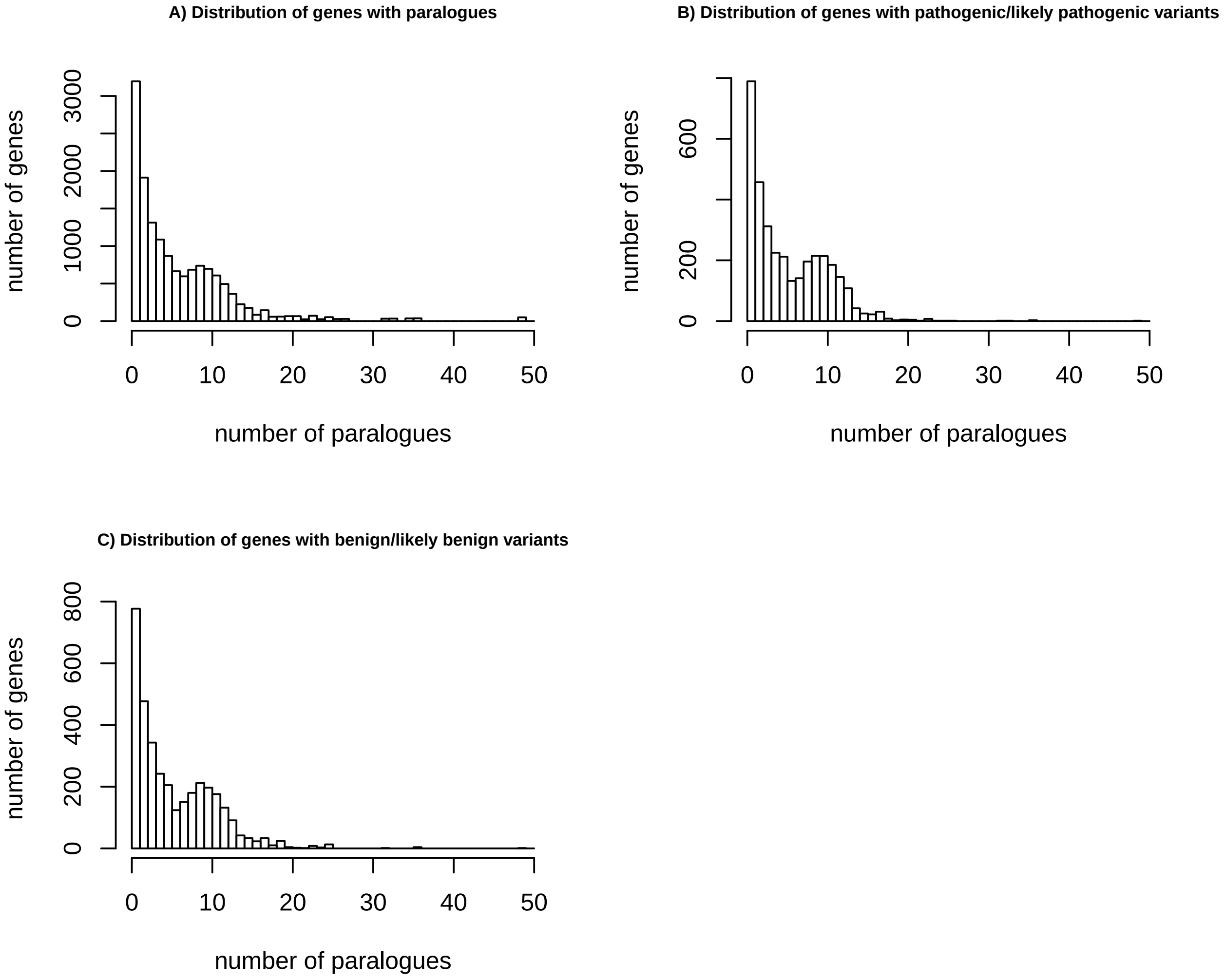

Figure 1: Distribution of genes with paralogues by the number of paralogues they're related to

3 Results and Discussion

3.1 Paralogue stats

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.

3.2 Annotation of Clinvar

	Pathogenic variants	Benign variants	PPV	Sensitivity	P value
Total variants	22583	13070	NA	NA	NA
Paralogue Annotation	17477	605	0.967	0.774	0
Variants remaining after QC1	16356	183	0.989	0.724	0
Variants removed from QC1	1121	422	0.727	NA	0
Variants remaining after QC2	7220	40	0.994	0.32	0
Variants removed after QC2	9136	143	0.985	NA	0
Variants remaining after QC3	3170	3	0.999	0.14	0
Variants removed after QC3	4050	37	0.991	NA	0

Prior to annotation, there were 22572 variants that aligned to at least one paralogous equivalent position according to ensembl (0.2203544)

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 variants was not as reliable. With 1924 known P/LP and 1926 known B/LB variants predicted to be benign. Though the proportional differene is statistically significant with a p-value of $9.6726228 \times 10^{-58}$, this lead to a PPV and sensitivity of 0.4997403 and 0.0851968 respectively.

Using the aforementioned quality control steps to increase the stringency of conservation across alignment columns in regards to reference and alternate amino acid alleles shows improvement to PPV and decrease in sensitivity over all for predicting pathogenic variants. But this does not help the case for predicting benign variants. The PPV does not improve significantly to a reliable level. Therefore, it can be concluded that at least with the dataset used in this study, paralogue annotation can be used as a variant classification method for predicting pathogenic variants, but not benign.

3.3 Para-Z scores

The filtering steps outlined above take a more binaray approach to 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 defined by Lal et al. (2017) on the other hand takes 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. As shown by **fig.** 2, one can use the Para-Z scores as cutoff thresholds for how paralogue conserved alignments must be before considering annotations. The more stringent the cutoff the lower the false positive rate is, but as expected the lower the sensitivity is also. 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.

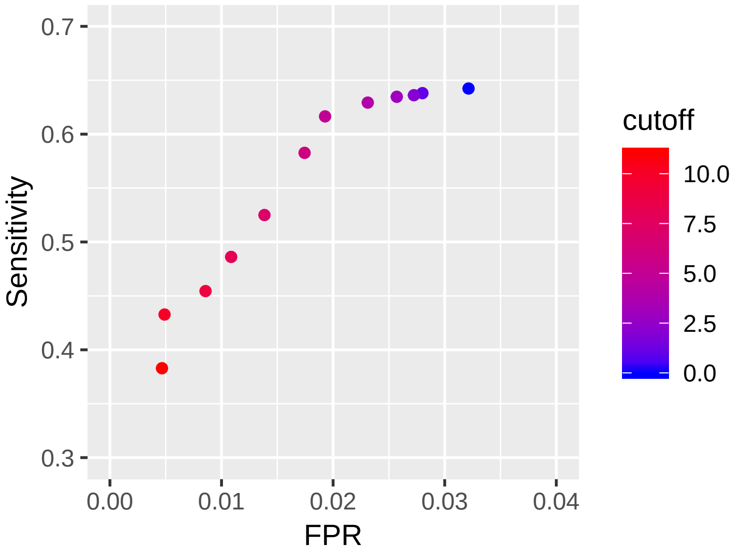

Figure 2: ROC space plot of pathogenic/likely pathogenic clinvar variants predicted as pathogenic by paralogue annotation at varying degrees of Para-Z scores

3.4 Case example: lack of paralogue annotation on 8 sarcomeric genes subset and use of etiological fractions as alternate validation

	Pathogenic variants	Benign variants	PPV	Sensitivity	P value
Total variants	454	34	NA	NA	NA
Paralogue Annotation	16	1	0.941	0.035	1
Variants remaining after QC1	16	1	0.941	0.035	1
Variants removed from QC1	0	0	NaN	NA	1
Variants remaining after QC2	9	0	1	0.02	1
Variants removed after QC2	7	1	0.875	NA	0.446
Variants remaining after QC3	3	0	1	0.007	1
Variants removed after QC3	6	0	1	NA	1

There are however limitations to this current framework. These can be listed as the following criteria: 1)the reliance on genes with paralogues; 2) for those paralogues to have pathogenic variants; and 3) for the paralogous variants to be aligned to corresponding equivalent positions. This is not always the case.

As a specific example, consider Hypertrophic Cardiomyopathy (HCM) and the 8 sarcomeric genes commonly associated with the genetic basis of HCM: MYH7; MYBPC3; TNNT2; TPM1; MYL2; MYL3; TNNI3; and ACTC1.

Annotating only these 8 sarcomeric genes with the whole clinvar P/LP dataset as before did not provide many annotations - without any quality control there were only 16 P/LP variants and 1 B/LP variants predicted to be pathogenic.

This could suggest either PA does not perform well on sarcomeric genes for the reasons stated above (paralogues to sarcomeric genes are not involed in disease) or that there is a lack of data - given more pathogenic variants to annotate with would certainly increase the likelihood of paralogous alignments.

external_gene_name	total
ACTC1	ACTA1, ACTG1, ACTB, ACTBL2, ACTL9, ACTL7B, ACTRT1, ACTRT2, ACTRT3, ACTR1A, ACTL7A, ACTR1B
MYBPC3	MYBPC2, MYBPHL, MYBPH, MYBPC1, IGSF22, IGFN1, MYOM2, MYOM3, MYOM1
MYH7	MYH6, MYH4, MYH3, MYH13, MYH8, MYH1, MYH2, MYH15, MYH7B, MYH14, MYH11, MYH10, MYH9
MYL2	MYL10, MYLPF, MYL5, MYL7, MYL12B, MYL12A
MYL3	MYL4, MYL6B, MYL6, MYL1
TNNI3	TNNI2, TNNI1
TNNT2	TNNT3, TNNT1
TPM1	TPM3, TPM2, TPM4

Looking at how many paralogues the 8 sarcomeric genes have (table `sarcomeric_genes_paralogs`), there are at least 2 for each gene with MYH7 having the most - 13. This satisfies the first criteria.

In the clinvar P/LP dataset, there are 887 known P/LP variants that lie in the 8 sarcomeric genes. But taking only the paralogues of these genes, there are only 381 variants. Assuming the clinvar dataset is complete and considering it in isolation, this would suggest that in comparison to variants in the main 8 sarcomeric genes involved in HCM, their associated paralogous variants are not as frequently involved in disease. This still does however statisfy the second criteria.

Therefore, the third criteria is where the lack of as many annotations arises. Having additional data for both known reference variants to transfer annotations from and more query variants to transfer annotations to may resolve this as there would intuitively be more alignments available to transfer annotations. Collecting additional known pathogenic variants can be difficult as that would require established interpretation and verification of such variants.

But querying every possible missense variant at all positions in the 8 sarcomeric genes can be done (30607 total variants). Doing so, predicts 1545 variants to pathogenic - a 96.6 fold increase. Though validating how many additional predictions are true positives would require additional known pathogenicity data.

However, it must be noted that even in genes with few predictions, paralogue annotation can still be functional and such predictions applicable. In the case of the 8 sarcomeric genes, one can still validate the use of paralogue annotation in a case control study.

Cases with positive (bad) outcome Number in exposed group:
39

Number in control group:
28

Cases with negative (good) outcome Number in exposed group:
41404

Number in control group:
456365

Results Odds ratio 15.3524 95 % CI: 9.4468 to 24.9499 z statistic 11.024 Significance level P < 0.0001

EFs show that even with few predictions, PA still works.

For example in the 8 sarcomeric genes involved in HCM [MYH7, MYBPC3, TNNT2, TPM1, MYL2, MYL3, TNNI3, ACTC1], taking MYH7 there were not many paralogous alignments. since most paralogues of HCM disease genes are no.

Infact performing the analysis on all possible missesnse mutations for these set of genes still shows a lack of annotation…

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.

References

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4). Nature Publishing Group: 248.

Altman, DG. 1991. "Practical Statistics for Medical Research Chapman & Hall London Google Scholar." *Haung, et Al [16] USA (Black).*

Hildebrandt, Mandy, Ralf Bender, Ulrich Gehrmann, and Maria Blettner. 2006. "Calculating Confidence Intervals for Impact Numbers." *BMC Medical Research Methodology* 6 (1). BioMed Central: 32.

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.

Landrum, Melissa J, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2015. "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants." *Nucleic Acids Research* 44 (D1). Oxford University Press: D862–D868.

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.

Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. 2012. "SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins." *Nucleic Acids Research* 40 (W1). Oxford University Press: W452–W457.

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.