Further applications of Paralog Annotation

Contents

0.0.1 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 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

0.0.2 Results and Discussion

0.0.2.1 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.
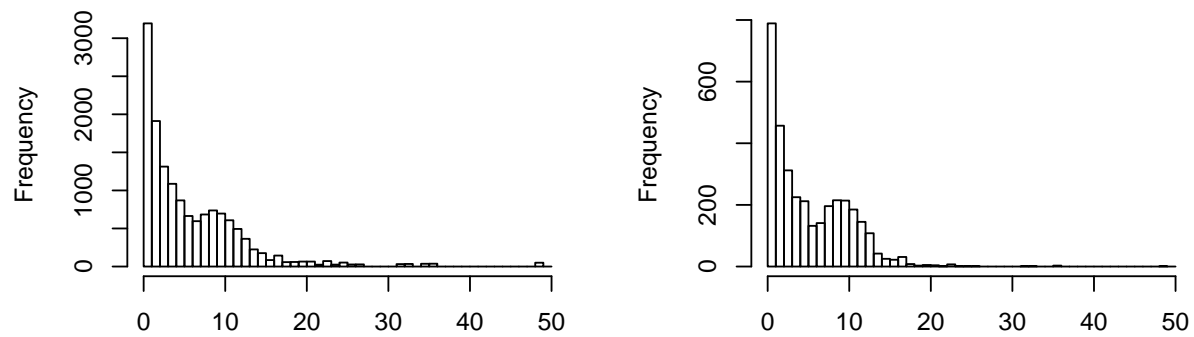
Performing a simple kolmogorov smirnov test between the distribution of pathogenic variants in genes with paralogues and benign variants shows a p-value of 0.9329528 suggesting that the null hypothesis of the distributions being identical cannot be rejected. From this, there appears to be no statistical 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.

#### 0.0.2.2 Annotation of Clinvar

Table 1: Table continues below

Variant	Total	Paralogue_Annotation_no_QC
Pathogenic variants	22583	17477
Benign variants	13070	605
PPV	NA	0.966541311801792
Sensitivity	NA	0.773900721781871
P value	NA	0

f protein_genes_w_paralogues_wide$numbe **linvar_P_LP_genes_w_paralogs_wide$numl**

protein_genes_w_paralogues_wide$number_of_paralog   linvar_P_LP_genes_w_paralogs_wide$number_of_paral

linvar_B_LB_genes_w_paralogs_wide$numl
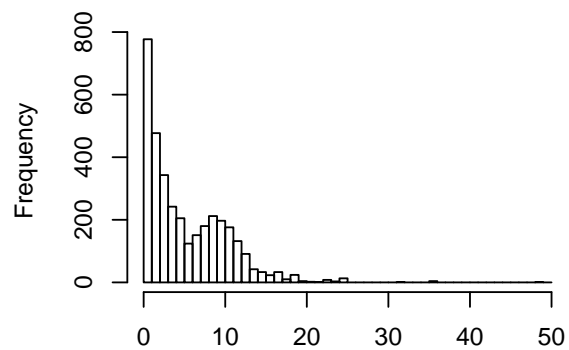
linvar_B_LB_genes_w_paralogs_wide$number_of_paral

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

Table 2: Table continues below

Variants_remaining_after_PA_QC1	Variants_removed_after_PA_QC1
16356	1121
183	422
0.988935243968801	0.72650680492547
0.724261612717531	NA
0	3.80673775046922e-14

Table 3: Table continues below

Variants_remaining_after_PA_QC2	Variants_removed_after_PA_QC2
7220	9136
40	143
0.994490358126722	0.984588856557819
0.319709516007616	NA
0	0

Variants_remaining_after_PA_QC3	Variants_removed_after_PA_QC3
3170	4050
3	37
0.99905452253388	0.990946904820161
0.140371075587832	NA
0	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.

0.0.2.3   Para-Z scores
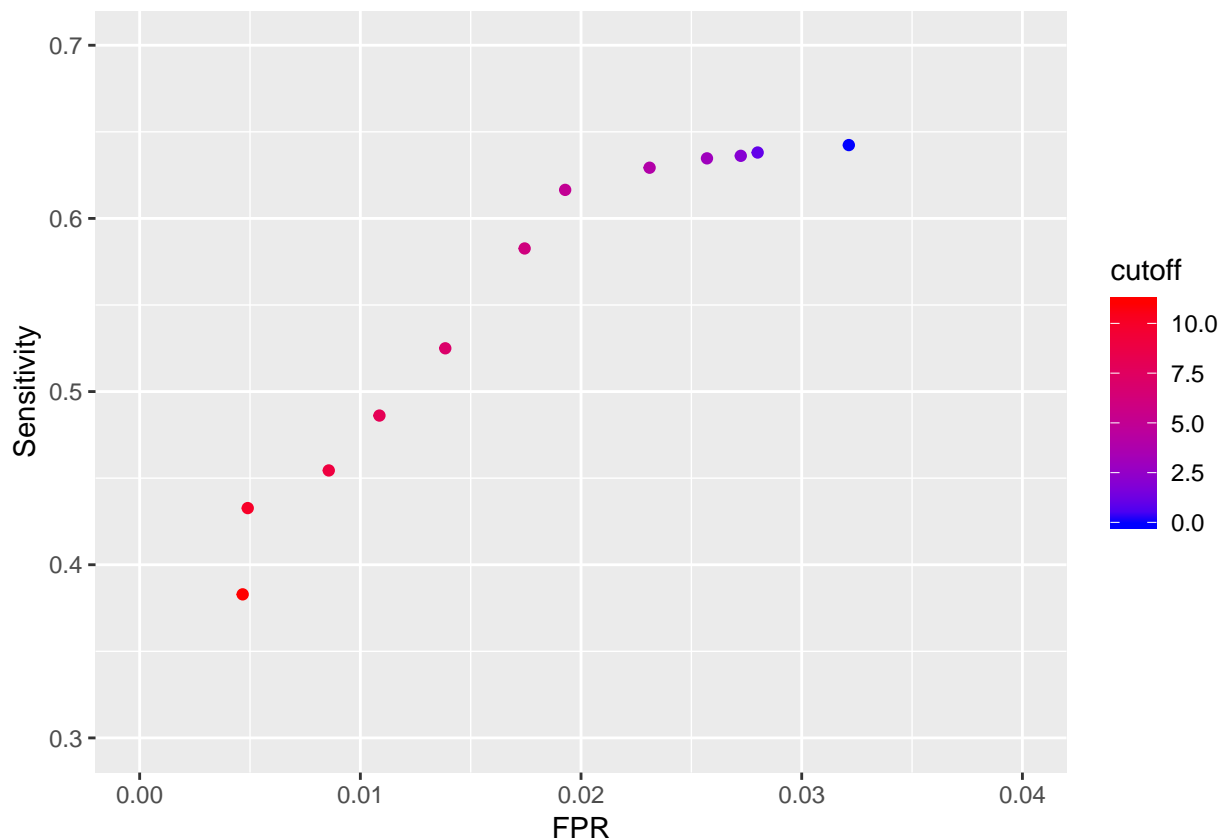
The filtering steps outlined above take a more binaray path into taking account the conservativeness of amino acid positions in the alignments. They only consider if amino acids in question share the the same amino acid or not. The Para-Z scores on the other hand take a more quantitative approach to 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.

0.0.2.4 Subset of 8 sarcomeric genes and calculation of EF

Taking only the 8 sarcomeric genes:

Variant	Total	Paralogue_Annotation_no_QC	Variants_remaining_after_PA_QC1	Variants_removed
Pathogenic variants	454	16	16	0
Benign variants	34	1	1	0
PPV	NA	0.941176470588235	0.941176470588235	NaN
Sensitivity	NA	0.0352422907488987	0.0352422907488987	NA
P value	NA	1	1	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, A
MYBPC3	MYBPC2, MYBPHL, MYBPH, MYBPC1, IGSF22, IGFN1, MYOM2, MYOM3, MYOM1
MYH7	MYH6, MYH4, MYH3, MYH13, MYH8, MYH1, MYH2, MYH15, MYH7B, MYH14, MYH11, MYH
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. For the 8 sarcomeric genes

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.

0.0.2.5   References

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.

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.