About PERCH

[Back to top]

How to cite PERCH?

PERCH: a unified framework for disease gene prioritization. Feng BJ. *Hum Mutat*. 2017;38(3):243-251.

Input Files

[Back to top]

Do you have a script to create a Genotype File from an existing annotated VCF?

I am afraid it is not possible for me to write a one-size-fits-all script to convert all kinds of output files from different annotation programs (AnnoVar, VEP, SnpEff, CAVA, VAT, Ingenuity, SVS). Moreover, each annotation program may have a variety of options, parameters, and databases. Hence, the output file format may be different from lab to lab even when they use the same program. And I don't know what kinds of format people have. So, if you could send me an example of your file, I may be able to help. If your file format is general enough I will change my programs, otherwise I could write a script to convert the file formats. So please don't hesitate to contact me.

Alternatively, please consider using my VICTOR pipeline, which performs analyses starting from a raw VCF so that you don't have any file format problems. Please see VICTOR's Tutorial page for details.

Parameters

[Back to top]

How should I filter variants by BayesDel and MaxAF?

Although allele frequency and deleteriousness are quantitatively considered in variant weights, it is still better to filter variants based on these two factors in the association and segregation analysis. The default values are --filt-MaxAF=0.01 and --filt-del=-0.059, which are suitable for low frequency to rare variants conferring a intermediate to high risk. The BayesDel cutoff -0.059 is the value that maximizes specificity and sensitivity at the same time in the test data. If you want to include variants conferring a modest risk, you may want to increase the MaxAF cutoff to --filt-MaxAF=0.05. Accordingly, you may also lower the --filt-del threshold.

Allele Frequency

[Back to top]

Is the allele frequency quantitatively or qualitatively (with transformation) integrated with other components in BayesDel?

It's quantitatively integrated.

Deleteriousness

[Back to top]

Why you don't include more deleteriousness predictors?

Currently, we only include predictors that were readily available for large-scale annotations; so the predictors that require multiple sequence alignments, such as AlignGVGD and MAPP, are not considered. In addition, we doubt that the inclusion of even more deleteriousness predictors could dramatically increase the accuracy, since they more or less measure the same characteristics of a variant. Nevertheless, we will re-train our model with new deleteriousness scores when they are available. Please support us because we need fundings to do so.

Segregation

[Back to top]

I want to use SEQlinkage instead of the vSEG, how do I do that?

You can run SEQlinkage, extract LOD scores, then populate these scores to the BayesSeg column by matching with the Func_Gene column.

Association

[Back to top]

I want to use other association methods, how do I do that?

We plan to add those functions soon. Please support us because we need fundings to do so.

Variant call quality

[Back to top]

Does the integration of variant call quality make a big difference?

If you do variant prioritization in a genomic scale, the performance improvement by integrating variant call quality is marginal, because the proportion of false variant calls is very small. However, if you do a gene-based rare-variant association test, this could make a bigger difference. For example, if you found only 2 variants in a gene, one of which is a false variant, then the proportion of false variant is 50%. Integrating call quality may change the rare-variant association test result substantially.

Biological relevance

[Back to top]

Does biological knowledge make a big difference in gene prioritization?

That depends on your sample size. The power of segregation and association analysis is a function of sample size, while biological relevance is not. So the more samples, the less relative importance of any biological relevance assessment. Actually, you don't even want it to play too much role in the analysis of a complex disease. If you observe that biological relevance scores swamp the other components, i.e. the highest association+segregation scores are relatively small comparing to the highest biological relevance scores, that means your sample size is small. That said, it doesn't mean that the biological relevance assessment is useless. It still may move genes in or out of your candidate list when they are at the boundary.

Error messages

[Back to top]

I got one of the following warning messages:

```
## Warning: <50% of the variants have a FATHMM score <=0.5. FATHMM might not be converted.
## Warning: >50% of the variants have a SIFT   score <=0.5. SIFT might not be converted.
## Warning: >50% of the variants have a MT     score <=0.5. MT might not be converted.
## Warning: >50% of the variants have a LRT    score <=0.5. LRT might not be converted.
```

This happens only when you use "vDEL --ensemble=nsfp23" (not by default). vDEL tries to keep you out of trouble by checking whether the deleteriousness scores are converted. However, this checking is not perfect and may give false alarms, which normally happens when you are working on a deleteriousness variant database rather than an unselected variant database. You can ignore these messages if you are sure the deleteriousness scores are converted.

The program doesn't run correctly and I got the following error message: Illegal Instruction

It's likely that you are running the program on Mac OS 10.6 or earlier.

Results

[Back to top]

I see a lot of OR../HLA../MUC.. genes at the top, and the scores are very high.

It is very likely that the cases and controls are not comparable. Did they come from different exome enrichments, sequencing platforms, bioinformatics pipelines, or populations? You may want to restrict the analysis to the intersection of captured regions, and exclude difficult regions for next-generation sequencing. But these methods can only reduce the bias, not eliminate them. Besides olfactory receptor genes, you can also spot a false positive if the score is too high, or the gene is a polymorphic gene, paralogous gene, or pseudogene. Also, be careful about the gene that has a pseudogene.

In the final results, can I remove the top genes that have a negative BayesHLR or negative BayesSeg score?

When you have a mixture of pedigrees and independent case and controls, both BayesHLR and BayesSeg are computable. Normally, you would expect that both scores are positive for the causal genes. So it may be reasonable to filter out the genes that either one of the scores is negative. However, in our experiences we found that the top genes with one negative one positive score may be due to bad data quality. Careful QC procedures would make them go away. Therefore, I would suggest thinking more about quality control rather than filtering results. Another potential explanation for negative result at the top is heterogeneity. If your samples include both high-risk pedigrees and sporadic cases without a family history nor a young age of onset, then the results of BayesHLR and BayesSeg will not match. And I am not sure this study design would work.