**Tempus Bioinformatics Technical Challenge**

Jun Wei, 4-21-2019

For this challenge, you are asked to prototype a variant annotation tool. We will provide you with a VCF ﬁle, and you will create a small software program to output a table annotating each variant in the ﬁle. Each variant must be annotated with the following pieces of information:

1. Type of variation (Substitution, Insertion, Silent, Intergenic, etc.) If there are multiple possibilities, annotate with the most deleterious possibility.

2. Depth of sequence coverage at the site of variation.

3. Number of reads supporting the variant.

4. Percentage of reads supporting the variant versus those supporting reference reads.

5. Allele frequency of variant from Broad Institute ExAC Project API (API documentation is available here: <http://exac.hms.harvard.edu/>)

6. Additional optional information from ExAC that you feel might be relevant.

**Illustration of program:**

The code is written with Python (it is tested in Python3.5.4 version).

The following libraries were used:

Datetime: to record date that file is annotated.

Pandas: to convert vcf file to dataframe, which is easier to manipulate, and file can easily convert to different form, including csv, excel, html etc.

Requests: to access ExAC API.

Collections: to create ordered dictionary.

The annotation has basically 4 steps:

1. Access to file. The file must be stored under ”\ OriginalFile\” subdirectory. Users will be ask input file name without path and vcf extension. Program will search for file for annotation. Users can stop program by input ‘n’.
2. Read vcf file and convert file to metainformation string and all data will convert to dataframe.
3. Add required information to dataframe and make ExAC queries. There are some variants, which have more than one change at the sites. Here I only choose one variant, which have largest length difference comped to reference, to query (assume insertion or deletion will cause amino acid loss or frame shift, which have more deleterious impact than substitution. There some potential pitfalls: SNP causing non-sense mutation may be more deleterious; frameshift may be more deleterious than amino acid loss. Possible solution is that query all variants and find most deleterious form). One of other pitfall of this program is that only ExAC database is queried. A lot of variants can not be found in this data base. There are some other tools can be used for query, such as VEP, ANNOVAR (not sure if I should use these tools, since they can do better annotation than this program). All data will be combined to dataframe and the end of the process.
4. Save dataframe as csv (or any possible file).

Information provided by this annotation:

1. compTwoSampl: Comparison of two samples. Compare GT:GQ:DP:DPR:RO:QR:AO:QA data of these two samples. If there is any difference, the value will be set to 0, otherwise the value is 1. (Since the possible use of NGS to find meaningful variant in tumor. Here we can find if there is any difference between tumor and normal tissue. In some case, the quality values show difference, it might rise red flag in sample process or analysis.)
2. #CHROM: as shown in original vcf file.
3. POS: as shown in original vcf file.
4. rsID: rsID in ExAC database.
5. REF: as shown in original vcf file, sequence of reference allele.
6. ALT: as shown in original vcf file, sequence of alteration allele.
7. QUAL: as shown in original vcf file.
8. FILTER: as shown in original vcf file.
9. VariantType: variant types were extract from vcf INFO.
10. Gene Symbol: the query result, here only showed gene with canonical transcript.
11. CCDS ID: CCDS ID of the gene.
12. Motif Name: the query result.
13. Mutation type: mutation type in gene with canonical transcript.
14. Amino acid change: if there is any amino acid change in gene with canonical transcript.
15. alleleFrequency：variant frequency in whole population in ExAC database.
16. Clincal Significance: possible clinical significance in gene with canonical transcript.
17. PubMed: PubMed id of queried variant.
18. ENSP ID:ENSP id of gene.
19. CDC Position: cdc position if variant is on cdc
20. Intron: in which intron the variant is located (canonical transcript).
21. Exon: in which exon the variant is located (canonical transcript).
22. STRAND: the DNA strand that gene is located.
23. SIFT: SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.
24. PolyPhen: Predicting possible impact of an amino acid substitution on the structure and function of a human protein.
25. GTof1st: genotype of 1st sample.
26. DPof1st: read depth of 1st sample.
27. ReadSupport1st: total reads support reference and variant alleles in 1st sample.
28. ratioAO1st (%): percentage of queried variant among reads supports in 1st sample.
29. ratioRO1st (%): percentage of reference allele among reads supports in 1st sample.
30. GTof2nd: genotype of 2nd sample.
31. DPof2nd: read depth of 2nd sample.
32. ReadSupport2nd: total reads support reference and variant alleles in 2nd sample
33. ratioAO2nd (%): percentage of queried variant among reads supports in 2nd sample.
34. ratioRO2nd (%): percentage of reference allele among reads supports in 2nd sample.
35. vepAnnotationMostImpactTranscripts: the first vep information from ExAc query of variant. According to ExAC guide, this showed most deleterious impact of variant on any possible transcripts of the gene. It could be canonical one, but not necessary.
36. vepAnnotationCanonicalTranscript: the vep information of canonical transcript from ExAc query of variant. Canonical transcript is more often used..
37. INFO: as shown in original vcf file.
38. FORMAT: as shown in original vcf file.
39. Normal: as shown in original vcf file.
40. vaf5: as shown in original vcf file.

rsID, Gene Symbol, Motif Name, Mutation type, Amino acid change are most information that patients and physician are interested in. Here I shown only information from canonical transcript, because canonical transcripts are more often used in clinic and research. I have some experience that non-canonical annotation caused confusion. CCDS id and ENSP id will help to identify which transcript or isoform is shown here. CDC position, intron, exon, may help experts to have idea where variants are, compared to well-known variants, such as some EGFR, KRAS mutations. But all information of transcript on which variant is predicted to have most deleterious impact and canonical transcript will be annotated at the end.

Usually I prefer to provide SIFT and PolyPhen score (or classification). In some case, these information can help research have more clear idea the impact of variant on gene function.