An exploratory work on the annotation conflict of public available genome sequence data

The whole genome sequencing is a powerful tool for us to understand the health and diseases. Commercial products, like 23 and me, has provided the genetic test with interpretations of variants for individuals for a long time. The whole genome sequencing has been used by pharmaceutical and biotech company to identify biomarkers for disease diagnosis and treatment. And now the genomic sequencing are rapidly becoming a routine tool in the diagnostic workup of patients with diverse conditions, including cancer.

However, the main challenge for the application of genomic sequencing will be the interpreting the clinical significance of the variants observed in a given individual, as well as the consequence for the population. Every step in the variant interpretation process has difficulties and limitations, and could bring false positive and false negative results. And interpretation conflict might show among sequencing results from different conditions and among different data sources, which limited the usability of genomic sequencing in healthcare, pharmaceutical research and development as well as genetic consulting.

The exploratory study of the gene variant annotation conflict and sum up the evidence of pathogenicity from different sources, which is not well studied and understood, could provide inside for us to evaluate the quality of the variant annotations, and to assess the confidence of the relationship between variants and clinical significant.

This project will use the ClinVar data as an example to explore the variant annotation conflict in the public dataset, and dig in to understand factors that are related with annotation conflicts. The project will be able to answer questions like: Do bigger genes have more conflict? Do the number of separate submissions per gene have an impact on conflict? Do the number of alleles per gene have an impact on conflict?

Data sources

The ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) is a public archive providing access to clinically relevant variants with relationships between variants and phenotypes along with supporting evidence. Assessment of clinical significant of variants is provided by submitters using categories of benign, likely benign, pathogen, likely pathogen, uncertain, and *et al.* The conflicts in interpretation by different submitters are tagged in the aggregate records.
The latest record of conflict annotations and gene annotation summary are used in this project.

Additional, I link the variants back to the genes to explore the conflict. The human genome reference and human genome annotation in the NCBI RefSeq database are used to understand the variant annotation conflict in the gene level. The genome reference provides the standard taxonomy of genes in the whole genome, whereas the genome annotation gives the details of the genes, such as the location, starting position and *et al*.

Table 1 Data sources and related links

Data	Data source	Data URL
ClinVar Variant Conflict	ClinVar	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/summary_of_conflicting_interpretations.txt
ClinVar Gene Summary	ClinVar	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/gene_specific_summary.txt
Reference genome	NCBI RefSeq	ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/gene_RefSeqGene
Genome annotation	NCBI RefSeq	ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/GCF_000001405.28_refseqgene_alignments.gff3

(URL to data sources: https://github.com/leanawen/Data_Incubator/blob/master/README.md)

Overview of the Variant Annotation Conflict in ClinVar

Up to now, there are 22,941 conflicts in ClinVar gene variants annotation. 172 labs/groups have submitted with at least 1 conflict. 1073 genes have conflicting interpretations. BRCA2, BRCA1, TTN, MYBPC3, and MYH7 are the top 5 genes with the most conflict.

Around 70.71% conflicts are minor differences where two annotations have the same conclusion on the pathogenic of the variant. For example, the submitter 1 interpreted variant SCV000238656 as “Pathogenic” whereas the submitter 2 annotated it as “Likely Pathogenic”. For variant SCV000166793, one submitter annotated as ‘Benign’, while submitter 2 annotated as ‘Likely Benign’. 1.51% conflicts are egregious conflicts where one submitter says benign and another says pathogenic. Gene SDHD has the most egregious conflicts.

Do bigger genes have more conflict?

Now I dig into the genes to see if we can understand more about the conflict. The size of the gene in base pairs was considered first.

The length of each gene was calculated using the human genome annotation data in NCBI RefSeq database (Data source: Genome annotation). The gene length is equal to the end location in the genome extracted by the start position. If multiple splice variants occurred, the length of the longest splice was used.

Multiple data sources are needed to link the conflict data to the gene length. ClinVar has used RefSeq genome IDS, which is embedded in the free text section of the Genome annotation. A regular expression was built to extract the genome ID of each gene from the Genome annotation. Genome ID was linked to Gene Symbols using NCBI Reference genome. Then the genes in the conflict table was linked with the length using Gene Symbols.

The length of genes varies a lot, ranging from 231,163 base pairs to 7,267 base pairs with the median length of 48,769 base pairs. Three sizes of genes have extremely more conflict (Figure

1A). When taking a look at the genes with length smaller than 100,000 base pairs, smaller genes tend to have more conflicts (Figure 1B).

When normalized by length, BRAC2 still has the most conflicts, followed by MYBPC3.

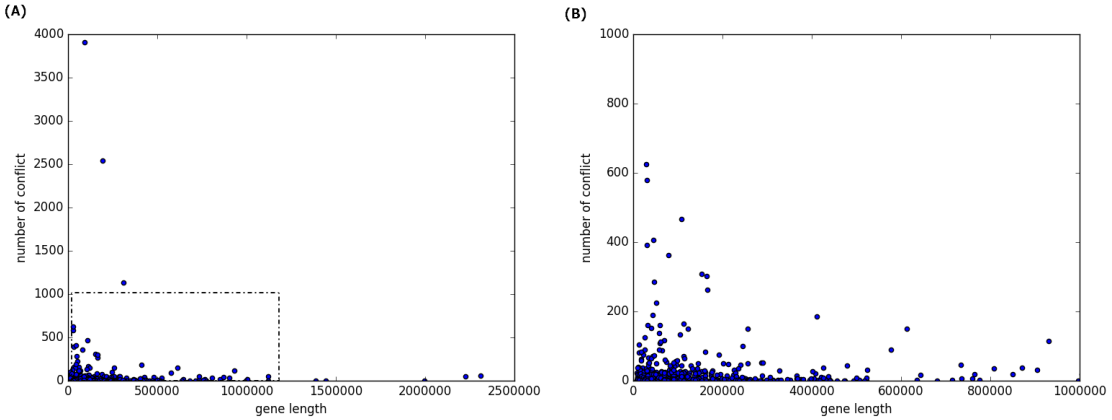

Figure 1 The number of conflict shown in the length of genes. A: total conflicts. B: conflicts in genes with size smaller than 100,000 base pairs.

(URL for Figure 1:
https://github.com/leanawen/Data_Incubator/blob/master/Figure%201_gene_length_vs_conflict.png)

Do the number of separate submissions per gene have an impact on conflict?

The number of total variant submission in ClinVar was extracted from the ClinVar Gene Summary, and then linked to genes in the conflict table.

The total submissions per genes in ClinVar is skewed right (Figure 2 A). The gene BRCA2 has the largest variants submission of 11,262, whereas CLP1 has the smallest number of submissions of 4. The median number of variant submission for each gene is 73.

When considering the number of submission, the number of conflict increases with the number of submissions (Figure 2 B). The gene BRCA2 has the most conflict with the largest number of submissions.

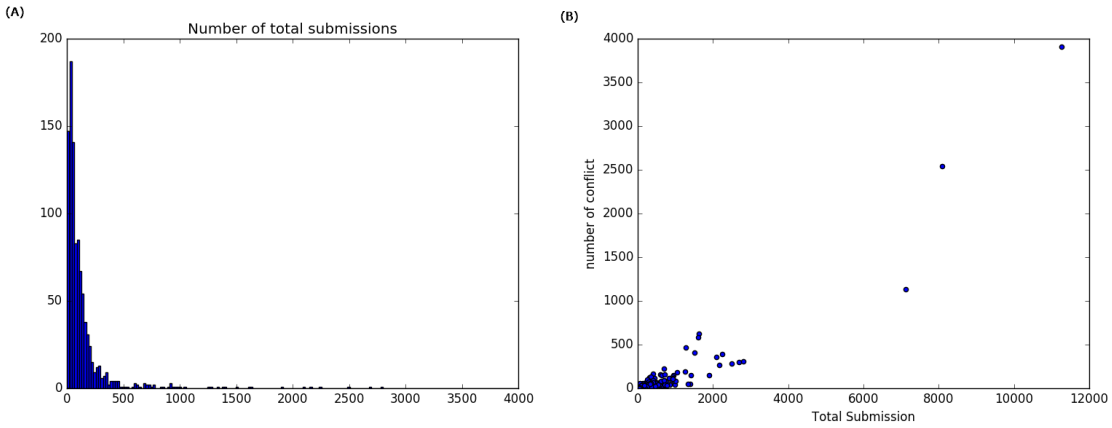

Figure 2 number of variant submission for each gene and the number of annotation conflict. A: distribution of total number of variant submission with submission smaller than 4000; B: the number of annotation conflict with the number of variant submission.

(URL for Figure 2:
https://github.com/leanawen/Data_Incubator/blob/master/Figure%202_total_submission_vs_conflict.png)

Do the number of alleles per gene have an impact on conflict?

I also take a look at other factors which could be potential features affect the number of annotation conflicts, such as the number of alleles for each gene. The number of annotation conflict increase with number of alleles (Figure 3).

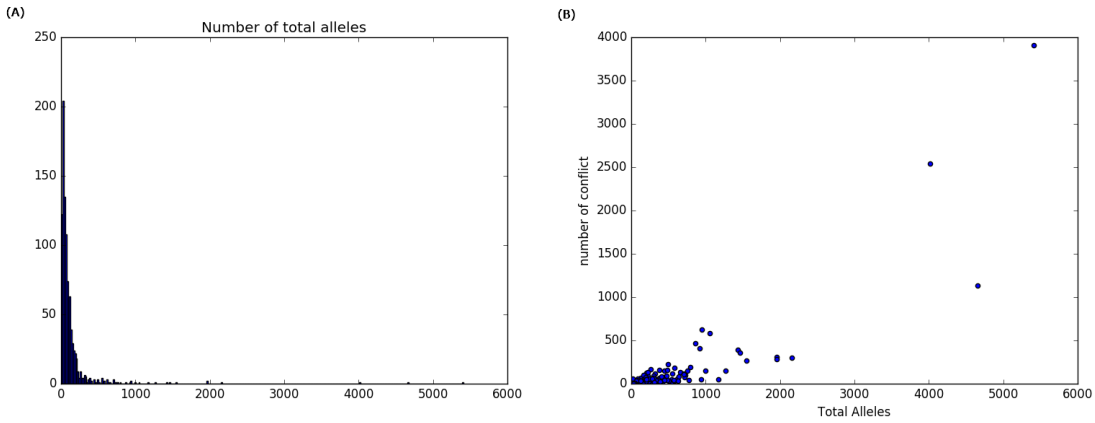

Figure 3 The number of alleles for each gene and the number of annotation conflict. A: distribution of total alleles; B: the number of annotation conflict with the number of alleles.

(URL for Figure 3:
https://github.com/leanawen/Data_Incubator/blob/master/Figure%203_total%20alleles_vs_conflict.png)

Conclusion and future work

The ClinVar data shows a great proportion of annotation conflicts. Although most of the conflicts are minor, there are genes, such as SDHD, with the most egregious conflicts. For individual gene, several factors could potentially related with the number of conflict, such as the gene size, the number of variant submissions and the number of alleles of the genes. Greater variant submission and greater number of gene alleles could bring more annotation conflicts. This could give some hints for researchers or patients when the get the genetic test results with variants.

For future work, other data sources should be evaluated, and more factors, such as gene complexity and CG content in the genes, should be studied to get deeper understanding of the annotation conflict. And a web-based searching engine to query annotation conflict across multiple data sources could be beneficial for researchers, genetic product developer, as well as patients.