Part 1.1 - Comparative Analysis of Personal Genomes

James Diao (*Writing*), Hussein Mohsen (*Coding*), and Nir Neumark (*Pipeline*)

May 8, 2017

1 Instructions

Compare the variants in Carl's genome with those in the gnomAD and 1000 Genomes databases. Compare and contrast the results from the two databases.

Writing: Describe the 1000 Genomes and gnomAD databases and what can be learned from collections of variants observed in large populations. How have these databases evolved over the past decade?

Coding: Propose a tool that finds information on a subset of Carl's variants in the gnomAD and 1000 genomes databases.

Pipeline: Identify and run a tool that finds the population frequencies for each of the variants observed in Carl's genome and also look to see how many are private variants. How do these number change based on the two databases used?

2 Introduction

With the completion of the Human Genome Project, scientists have made great progress in cataloging the breadth of human genetic variation. Large-scale reference data for genetic variation plays a crucial role for the clinical interpretation of sequenced DNA variants. In recent years, such data has been used to inform and validate pathogenicity metrics for various classes of mutation. For example, comparing the

2.1 1000 Genomes Project

The 1000 Genomes Project was an international effort to sequence, for the first time, a large and diverse collection of human genomes (at least 1000). By the end of the 3 project phases, the consortium had sequenced 2,504 genomes from 26 global populations, including African, Admixed American, European, East Asian, and South Asian. Although

newer datasets contain many times more individuals, the 1000 Genomes Project remains a useful and high-quality dataset for population-level genome analysis.

2.2 Genome Aggregation Database

The Genome Aggregation Database (gnomAD) is an expansion from the Exome Aggregation Consortium (ExAC), developed by the MacArthur Lab at the Broad Institute. gnomAD seeks to aggregate germ-line genome sequences from a large number of sources (including the 1000 Genomes Project). It does this by uniformly processing and filtering genomes from these different studies. The result is 138,632 genomes from 5 major continental populations. gnomAD provides a high-resolution picture of human genetic variation, allowing the analysis of very rare variants that had previously been undetected.

2.3 Other Databases

- DGV: Database of Genomic Variants
- DECIPHER: DatabasE of genomiC VarIation and Phenotype in Humans using Ensembl Resources
- EVA: EBI Variation Archive
- ExAC: Exome Aggregation Consortium

2.4 Applications to Personal Genome Analysis

Analysis of a personal genome involves evaluating its differences from the normal range of genetic variation. This normal range is inferred by comparison to many other genomes in some database. A variant identified as disease-causing or protein-knockout can be searched up in such a database to find how common it is. Very common variants may be dismissed as uninformative. Rare variants, on the other hand, are often considered to be highly penetrant.

3 Coding

Online browser and API access are not straightforward for users interested in using their own variant call files (VCFs) to retrieve a slice from a very large database. We propose a tool, vcfR, to help users to use VCFs to query certain variants within a reference dataset. The reference can be a standalone file or a link. The reference can be customized and the searching process is fast given proper reference and sorted order of query by genome position. The usage of vcfR is as follows:

```
python vcfR.py -i <input.vcf> -f <ref.vcf.gz> -o <output.vcf>
```

The matched terms in input file will be output to a file named `input_matched.file` in case of the comparison between matches and output. As mentioned earlier, vcfR can also deal with online reference, e.g., to query chr1 variants in 1000 Genomes and gnomAD:

Query gnomAD chr1

```
$ python vcfR.py -i sample_input.vcf -f
  https://storage.googleapis.com/gnomad-public/release-170228/vcf/
  genomes/gnomad.genomes.r2.0.1.sites.1.vcf.gz -o sample_out.vcf
```

Query genome1000 chr1 phase.1

```
$ python vcfR.py -i sample_input.vcf -f
  ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/
  integrated_call_sets/ALL.chr1.integrated_phase1_v3.20101123.
  snps_indels_svs.genotypes.vcf.gz -o sample_out.vcf
```

Hope this information could help. Please feel free to ask me if you have any questions about this part.

4 Pipeline

We have built a tool to find population frequencies for the variants observed in Carl's genome, and see how many are private variants ...

There are ... differences between the 1000 Genomes and gnomAD datasets ...

5 Conclusion

...

References

[1] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi:10.1038/nature15393.

[2] Lek M, Karczewski KJ, Samocha KE, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2016;536(7616):30338. doi:10.1101/030338.

[3] Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17(9):507-522. doi:10.1038/nrg.2016.86.