# Mutational Landscape and Identification of Deleterious Variants from Personal Whole-exome Sequencing (WXS)

Zepeng Mu

## Introduction

At the start of Human Genome Project (HGP), scientists believed that much of the information in human genome and disease-related loci can be revealed with the whole-genome sequence of a single person at hand. However, when HGP was eventually accomplished, it was realized that a single person's genome can merely serve as reference of human genome without much information with regards to any particular diseases. Therefore, it soon became a consensus that in order to unravel the genetic architecture that leads to diseases, we need to, in a case-by-case manner, sequence a sufficiently large patient cohort and compare it to the reference genome. Only by having a sufficiently large sample can we observe the association between variant frequencies and health states. What is more, with accumulation of knowledge in diseases and genetic variants, it is now becoming a reality to predict disease risks of health subjects based on personal genomes. Thus, DNA sequences from one individual is also becoming more and more informative. This is important because it may push evident-based medicine forward to Precision Medicine in the near future.

In fact, our understanding of human genome has always been a direct result of technology progress. In the past 40 years [1], DNA sequencing technologies has witnessed a great leap forward, during which cheaper, faster and more precise methods were introduced. Without these technology innovations, it would be impossible to sequence the genome of an individual painlessly. For example, despite the on-going technology innovations, HGP was not finished until after more than 10 years of work by thousands of scientists and technicians [2]. Needless to say, back then it is impossible to sequence a personal genome for clinical purposes with an affordable price and within a tolerable length of time for any patient. With next-generation sequencing, on the contrary, sequencing can be done in a couple of weeks, where gDNA is sheared into small pieces randomly, sequenced, and then mapped to reference genome using computer algorithms. At the same time, the development of high performance computing (HPC) and data storage and sharing strategies has made it feasible to manage and access the whole genome of individuals regardless of time and locations.

To further bring down the cost of sequencing, whole-exome sequencing (WXS) may be used. Exons account for merely 1 % of human genome, but is regarded as a major player in human diseases. Compared to whole-genome sequencing (WGS), WXS only requires an additional step during sample preparation, when a set of probes were used to enrich exon sequences. Therefore, compared to WGS, personal genome sequencing can be done at a much lower price, or at higher coverage with the same amount of cost by WXS.

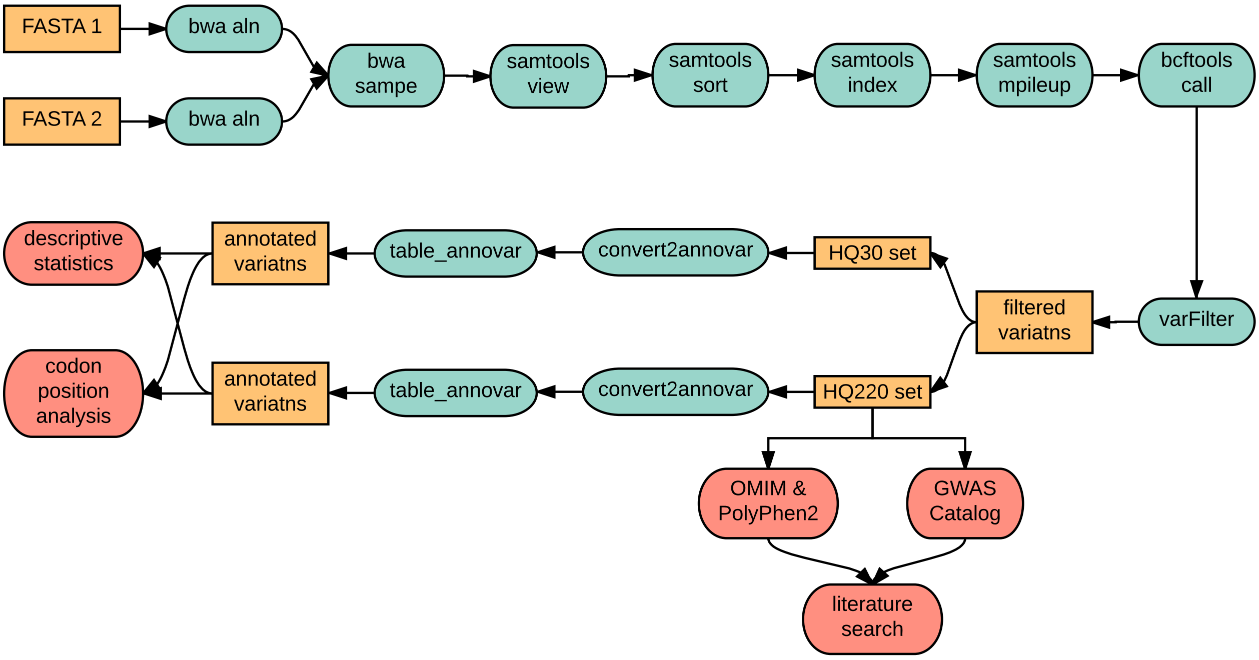

Figure 1. Summary of analysis pipeline. Rectangles colored orange show data used or generated, while rounded rectangles colored in blue show scripts and commands provided by TA, and rounded rectangles colored in red show self-performed analysis.

Here we describe the analysis of an individual's WXS. The data comes from 1000 Genomes Project [3], and the sequences we used here (Accession Number: SRR765980; Seq04 in class) come from a Spanish as part of the Spanish Hapmap Project. We explored the general variant landscape in this exome, identified a large number of SNVs and indels, and performed a functional analysis on some of the most intriguing and significant ones.

# Methods

All the scripts used in this project except for those provided by TA can be found in the online project repository (https://github.com/Zepeng-Mu/BIOS26120-Fall2017-ZepengMu). A brief summary of analysis pipeline is shown in Figure 1.

## Sequence alignment, genotyping, variant annotation and selection of high quality variants

Sequence alignment, genotyping and variation annotation was accomplished by scripts provided by TA. Briefly, *bwa aln* [4] was used to align the paired-end sequencing raw reads separately. Reference genome used was hg38. Then, two mapped reads were used to generate a SAM file with *bwa sampe* and subsequently sorted by *SAMtools sort*. Then, genotype and filtering was done with *SAMtools mpileup* and *varFilter*, respectively, which was passed to ANNOVAR [5] for variant annotation. Detailed commands and parameters used can be found in Box 1.

High quality variants were selected using shell commands using 30 and 220 as two cutoffs.

## Filtering variants associated with Mendelian diseases

Risk alleles for Mendelian diseases were selected based on two criteria. First, it must be recorded by OMIM database. Second, it must has a predicted score from HumVar-trained PolyPhen-2 [6]. SNPs satisfying these two criteria were identified and sorted by ranked PolyPhen-2 score. All analysis was performed with R version 3.3.2.

## Functional analysis of significant eQTLs from GTEx database

In order to find highly confident eQTLs from GTEx portal, we used the ANNOVAR output to get the p-values for all the genes in all the tissues reported by GTEx v6. To do so, ANNOVAR output entries with a GTEx annotation was selected. Then, the SNPs, ENSIDs and tissues were formated and written to a comma separated values (CSV) file. The file was uploaded to GTEx portal, tested for eQTLs and the results were downloaded. SNPs surviving Bonferroni Correction were remained.

To find the SNPs that has possible biological functions, we queried the GWAS Catalog [7]. All entries in GWAS Catalog were downloaded as a file and loaded into R. We then looked at the overlap between SNPs reported in GWAS Catalog and our candidate eQTLs.

The analysis was accomplished with R version 3.3.2. For scripts used, refer to final report repository online.

# Results

## Summary of genotyping

In genotyping process, 212334 variants were called in total, with 187274 of them being high-quality (QUAL larger than 30) and allocated to a known chromosome. The distribution of high-quality variants on chromosomes are shown in Figure 2a. The number of variation on each chromosome is roughly correlated to the size of the chromosomes. We then studied the distribution of quality score of all the genotyped variants. At low quality range, the distribution peaked at 50-100 and then decreased (Figure 2b). The overall shape of distribution was nearly normal bur skewed to the left. Surprisingly, there was an enrichment of variants with extremely high quality scores (Figure 2b), with 31128 of them larger than 220, accounting for 16.4 % of the variants with a score larger than 30. We defined these variants as *HQ220* set, and they shall be of our primary interests in subsequent analysis. The remaining variants with quality score higher than 30 were defined as the *HQ30* set (158426 variants). The sub-classification of these variants is summarized in Table 1.

Table 1. Summary of SNP and indel numbered called.

| | **HQ30** | **HQ220** | **Total** |
|---|---|---|---|
| SNPs | 147,110 | 29,837 | 176,947 |
| Indels | 11,594 | 1291 | 12,885 |
| Total | 158,426 | 31,128 | 189,554 |

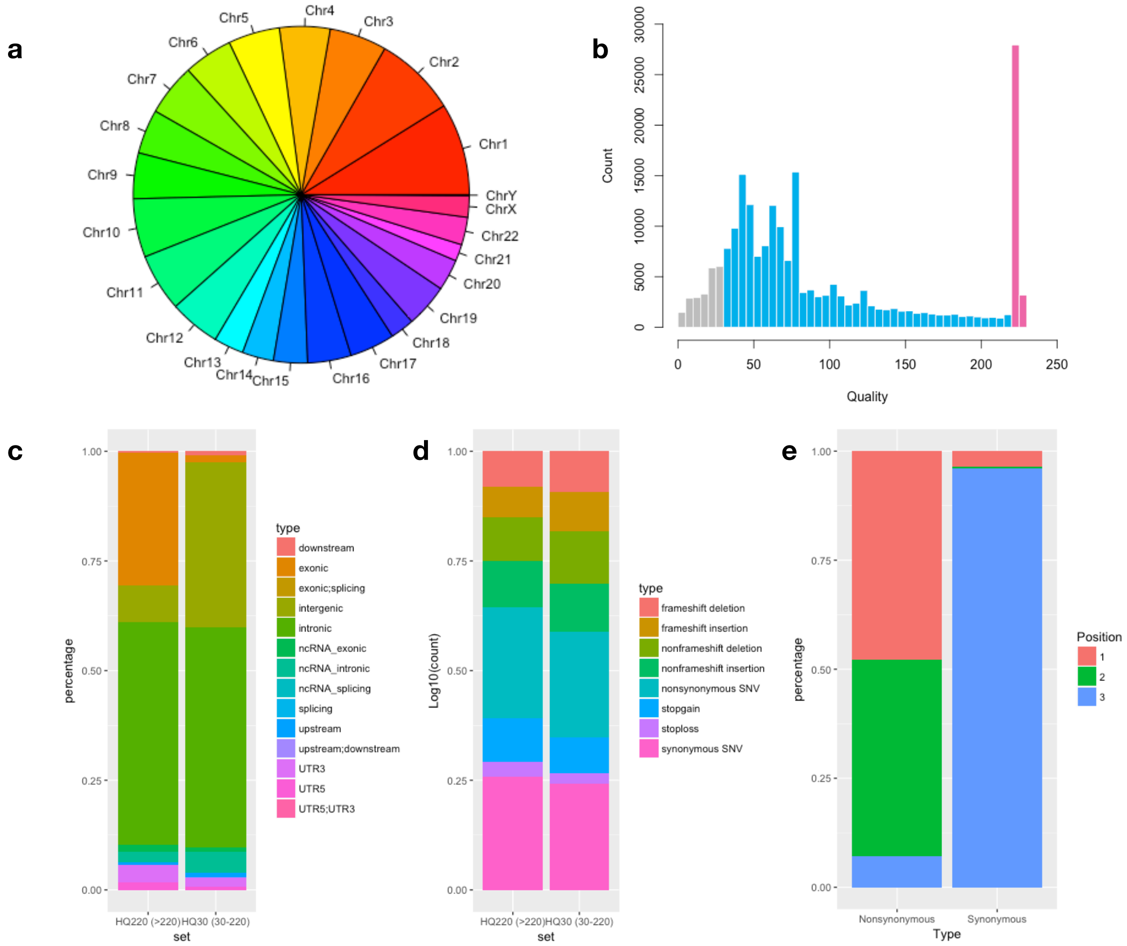

Figure 2. Graphic summary of variant calling and annotation. (a) The number of variants located on each chromosome is roughly correlated with chromosome sizes. (b)Histogram of quality score of variants in genotyping. Grey: score less than 30; Red: score larger than 220; Blue: in between. (c) The localization of variants in HQ220 and HQ30 sets. (d) The function distribution of variants in HQ220 and HQ30 sets. (e) The distribution of nucleotide substitutions in the position of three-letter codons between synonymous and nonsynonymous mutations.

## High quality variants are enriched with exonic SNPs

We performed variant annotation using ANNOVAR for HQ220 set and HQ30 set separately, and then compared the distribution of variant localization (Figure 2c) between the two sets. Although HQ220 set contained much less variants than HQ30 set, it is enriched with exonic variants both in number (7115 vs 2604) and in proportion (22.8 % vs 1.39 %). This is in-line with the theory that exonic mutations are more likely to have a profound biological function. In summary, we totally identified 9719 exonic variants in total, some of the sub-classification of exonic mutation was summarized in Table 2. From Figure 2d, we also see that HQ200 set contained more stoploss and stopgain variants than HQ30 set.

Table 2. Summary of variant functions.

| | **HQ30** | **HQ220** | **Total** |
|---|---:|---:|---:|
| Synonymous SNV | 1,229 | 3,749 | 4,978 |
| Nonsynonymous SNV | 1,233 | 3,227 | 4,460 |
| Frameshift | 29 | 22 | 51 |
| Stopgain | 11 | 23 | 34 |
| Stoploss | 2 | 3 | 5 |
| Non-frameshift Indel | 59 | 53 | 112 |
| Total | 2,563 | 7,077 | 9,640 |

## Codon position for synonymous and nonsynonymous mutations

It has become a widely acknowledged fact that the degenerate nucleotide at third position of codons tend to be less prone to nonsynonymous mutations. Therefore, assuming that each nucleotide has the same tendency to mutate in a three-letter codon, more synonymous mutations would be observed in the third nucleotide than the first two. The vast number of synonymous and nonsynonymous SNVs in our dataset enabled us to validate this theory. Indeed, we discovered that almost all of the synonymous mutations arose from nucleotide substitutions at the third position in codons. In fact, only six nonsynonymous mutations were due to mutations at the second position. On the other hand, mutations in the first and second position of codons accounted for nearly 80 % of all the nonsynonymous mutations, and they occurred in approximately equal frequencies (Figure 2e).

## Functional inferences in HQ220 set

In order to narrow down the candidate variants for functional inferences, we only focused on the HQ220 set. We first studied the variants that might lead to Mendelian diseases. A mutation is selected only if it was both reported in OMIM according "CLNDSDB" and had a score calculated from HumVar-trained PolyPhen-2 [6]. Interestingly, all the 177 variants selected are nonsynonymous SNVs in exonic region, with 16 of them being predicted to be deleterious. There are two examples in these high quality nonsynonymous SNPs. A nonsynonymous SNV (rs1800566, C→T) on chromosome 16 at position 69711242 causes a P115S mutation in Quinone Oxidoreductase-1 (NQO1). Another exonic mutation, on chromosome 1 at position 203225058 (rs2297950), causes a G→A substitution that leads to G102S mutation in Chitinase 1 (CHIT1). We delve deeper into these two SNPs in the Discussion section.

We next set out to analyze the intronic and intergenic variants in HQ220 set. There has been a consensus that a large proportion of non-coding DNA is not "junk DNA". On the contrary, it contains many regulatory elements, with enhancers being most of them [8]. We reasoned that if a SNP significantly changes the activity of an enhancer, then the SNP should have a profound function, since it might influence the expression of multiple genes. Therefore, we focused on eQTLs reported in the GTEx v6 database. However, ANNOVAR does not report p-values from GTEx experiments in its output, so we tested p-values online (https://www.gtexportal.org/home/testyourown, see Methods for details). The results were sorted by p-values and closely scrutinized. This led to a set of 65 eQTLs targeting 151 genes in 41 different tissues with Bonferroni corrected significance. In order to further investigate the functions of these eQTLs, we looked them up in GWAS Catalog (https://www.ebi.ac.uk/gwas/) to see if published GWAS has identified any of them. We found two eQTLs that has been reported by previous studies [9, 10]. The first one was rs1023252 at position 11899033 on chromosome 1. The second one was rs6502557 at position 16960978 on chromosome 17. We talk about these SNPs in detail in Discussion.

# Discussion

## SNPs with strong functional implications

Through our analysis, we identified several common variants that are likely to have profound functions. Nonsynonymous mutation rs1800566 causes a P115S substitution in protein NQO1. NQO1 is a human quinone oxidoreductase. The mutation is thought to compromise the activity of NQO1, which could lead to a higher level of sensitivity to cisplatin-related chemotherapy drugs. This is because one of the ways in which cisplatin function is to cause oxidative stress, and quinone acts as a radical group scavenger, thus is protective against cisplatin [11, 12, 13]. NQO1 could reduce quinone in cell, possibly plays a role in the regeneration of quinone.

Another intronic variant rs1023252 was reported before [9] in a genome-wide association study. In that research, del Greco and colleagues studied variants associated with N-terminal cleavage product of the B-type natriuretic peptide (NT-proBNP) concentration in serum, which is indicative of cardiac diseases. This SNP, rs1023252, reached genome-wide significance [9]. It locates in the intron of *CLCN6* gene, which encodes for chloride channel. The *CLCN6* gene also forms a gene cluster together with *MTHFR*, *NPPA*, *NPPB* [9], in a region with high level of linkage disequilibrium (LD). In our analysis, *MTHFR* and *NPPA-AS1* also appeared to be significant eQTL target of rs1023252, alone with *CLCN6*, indicating genes in this cluster may be co-regulated by one enhancer covering rs1023252. However, no experiment has proved that rs1023252 resides in an enhancer, as suggested by GTEx. To test this possibility, we used data from ENCODE and RoadmapEpigenetics to see if it is predicted to be an enhancer. Indeed, the region flanking rs1023525 is likely to have enhancer activity in various tissues including liver, fetal heart, and right ventricle (Figure 3), as predicted by the ChromHMM [14] algorithm.

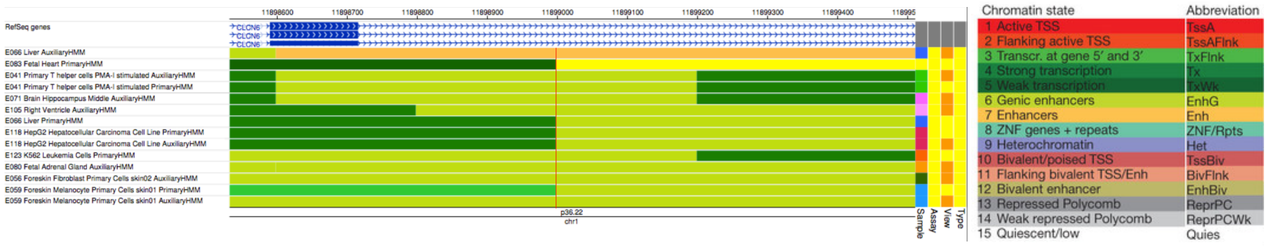
11898600
11898700
11898800
11898900
11899000
11899100
11899200
11899300
11899400
118995
RefSeq genes
CLCN6
CLCN6
CLCN6
E066 Liver AuxiliaryHMM
E083 Fetal Heart PrimaryHMM
E041 Primary T helper cells PMA-I stimulated AuxiliaryHMM
E041 Primary T helper cells PMA-I stimulated PrimaryHMM
E071 Brain Hippocampus Middle AuxiliaryHMM
E105 Right Ventricle AuxiliaryHMM
E066 Liver PrimaryHMM
E118 HepG2 Hepatocellular Carcinoma Cell Line PrimaryHMM
E118 HepG2 Hepatocellular Carcinoma Cell Line AuxiliaryHMM
E123 K562 Leukemia Cells PrimaryHMM
E080 Fetal Adrenal Gland AuxiliaryHMM
E056 Foreskin Fibroblast Primary Cells skin02 AuxiliaryHMM
E059 Foreskin Melanocyte Primary Cells skin01 PrimaryHMM
E059 Foreskin Melanocyte Primary Cells skin01 AuxiliaryHMM
p36.22
chr1
Sample
Assay
View
Type
Chromatin state
Abbreviation
1 Active TSS
TssA
2 Flanking active TSS
TssAFlnk
3 Transcr. at gene 5′ and 3′
TxFlnk
4 Strong transcription
Tx
5 Weak transcription
TxWk
6 Genic enhancers
EnhG
7 Enhancers
Enh
8 ZNF genes + repeats
ZNF/Rpts
9 Heterochromatin
Het
10 Bivalent/poised TSS
TssBiv
11 Flanking bivalent TSS/Enh
BivFlnk
12 Bivalent enhancer
EnhBiv
13 Repressed Polycomb
ReprPC
14 Weak repressed Polycomb
ReprPCWk
15 Quiescent/low
Quies

Figure 3. Prediction of chromatin states by ChromHMM. The red line indicates the position of rs1023252. Figure was generated by WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/). Legend credited to http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html.

### Limitations to this work

Although the advances in theories and methods of personal genomics might promise a future of precision medicine, there are some problems that are inherent to the field that need to be concerned. To begin with, NGS and genotyping are able to identify hundreds of thousands of common variants in an individual genome, but the genetic effects or the excessive risk to a given disease attributable to each of these variants are subtle. In fact, even the most significant SNPs can explain only a small proportion of the variance in the phenotype for most of the published GWAS. Therefore, based solely on exome-sequencing data, the statistic power we have to predict disease risks is really limited. This makes us question whether the SNPs identified in personal genome are actually clinically actionable. It might be plausible to interrogate *multi-omics* data at the same time [15]. Secondly, without extensive phenotype data, it is difficult to make correct and meaningful inferences to the variants identified, which further limits the utility of these variants in a clinical setting.

What is more, SNPs in non-coding regions can hardly be tested in WXS, since most of them are not enriched in the library preparation. This may significantly reduces our ability to report risk alleles, since many regulatory elements also plays an important role in disease onset.

## Conclusions

Through our analysis of a whole-exome sequencing results of a health Spanish individual from 1000 Genomes Project, we identified a number of risk alleles reported by various databases like OMIM, and predicted to be deleterious by methods like SIFT and PolyPhen-2. These discoveries, as a proof-of-principle, showed the power to identify diseases risks on a personal level and to invoke preventive measures before disease onset.

However, we acknowledge that using whole-exome data the power to detect truly deleterious and medically actionable risk alleles is quite restricted. Therefore, it is necessary to develop new sequencing technologies that are cheaper, faster, and more precise, as well as developing computer algorithms, statistical methods and data management tools to enable the effective integration of various types of personal data. Also, it should be noted that as whole-exome sequencing was used to generate the data we used for analysis, the proportion of exonic variants to non-coding variants may be distorted. If we assume that nucleotides in exons and non-coding regions have the same propensity to mutate and they are under neutral selection, then the number of mutations in non-coding regions should be 99 times higher than that in exons. If we take into account the higher selective load on exon regions, then the number mutations in exons would be further dwindled. This is certainly not the case in Figure 2c, thus we must be wary when interpreting the results.

## Acknowledgment

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund (http://commonfund.nih.gov/GTEx) of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 12/1/2017.

**Box 1.** Detailed commands and methods used for alignment, genotyping and annotation was listed here.

**Alignment**

```
bwn aln                                        # align raw reads
    -q 5                                       # used for read trimming
    -t 28                                      # specify the number of nodes

bwa sampe                                      # create bam file
    -P                                         # load index into memory

samtools view          # print alignment following a specified format
    -b                                         # output in BAM format
    -S      # input in SAM format, but ignored in latest version of samtools

samtools sort            # sort alignments by leftmost coordinates
    -@ 28                                      # number of threads
    -m 1500M             # maximum required memory per thread
    -T                                         # write temporary files
    -o                                         # output filename
```

**Genotyping**

```
samtools mpileup    # Generate VCF, BCF or pileup for one or multiple BAM
    files
    -t SP     # set output to SP (Phred-scaled strand bias P-value, FORMAT)
    -uv                          # generate uncompressed VCF output
    -f                      # specify reference file in the FASTA format.

bcftools call                                  # call SNP and indels
    -mv      # another model for multiallelic and rare-variant calling and output
    variant sites only

varFilter               # filtering short variants and output .flt.vcf file
    -D100
```

**Annotation**

```
convert2annovar.pl               # convert vcf format to annovar format
    -format                                # designate input format
    -outfile                               # designate output file

table_annovar.pl          # tab-delimited output file with many columns
    --buildver hg38                            # reference genome
    --protocol refGene,avsnp150,clinvar_20170905,dbnsfp33a              #
    annotation database
    --operation g,f,f,f        # gene-based for refGene, filter-based for other 3
```

# References

[1] Jay Shendure et al. "DNA sequencing at 40: past, present and future History of DNA sequencing technologies". *Nature Publishing Group* (2017).

[2] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome". *Nature* 409.6822 (2001), pp. 860–921.

[3] Adam Auton et al. *A global reference for human genetic variation.* 2015.

[4] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[5] Xiao Chang and Kai Wang. "wANNOVAR: annotating genetic variants for personal genomes via the web". *Journal of Medical Genetics* 49.7 (2012), pp. 433–436.

[6] Ivan A. Adzhubei et al. "A method and server for predicting damaging missense mutations". *Nature Methods* 7.4 (2010), pp. 248–249.

[7] Jacqueline MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". *Nucleic Acids Research* 45.D1 (2017), pp. D896–D901.

[8] Len A Pennacchio et al. "Enhancers: five essential questions." *Nature reviews. Genetics* 14.4 (2013), pp. 288–95.

[9] M. Fabiola del Greco et al. "Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster". *Human Molecular Genetics* 20.8 (2011), pp. 1660–1671.

[10] Gordan Lauc et al. "Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers". *PLoS Genetics* 9.1 (2013).

[11] H. J. Kim et al. "Augmentation of NAD+by NQO1 attenuates cisplatin-mediated hearing impairment". *Cell Death and Disease* 5.6 (2014).

[12] Gi Su Oh et al. "Pharmacological activation of NQO1 increases NAD + levels and attenuates cisplatin-mediated acute kidney injury in mice". *Kidney International* 85.3 (2014), pp. 547–560.

[13] Tae-Won Kim et al. "NQO1 Deficiency Leads Enhanced Autophagy in Cisplatin-Induced Acute Kidney Injury Through the AMPK/TSC2/mTOR Signaling Pathway." *Antioxidants & redox signaling* 24.15 (2016), pp. 867–883.

[14] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization." *Nature methods* 9.3 (2012), pp. 215–6.

[15] Rui Chen et al. *Personal omics profiling reveals dynamic molecular and medical phenotypes.* 2012.