BMS 320: Introduction to Bioinformatics

SNPs Calling: R Workflow

Mayar Mansour Mohamed Ahmed

Computational Biology and Genomics, Biomedical Sciences, University of Science and Technology, Zewail City, Giza, Egypt, s-mayarmansour@zewailcity.edu.eg

Abstract

Motivation: Single Nucleotide Polymorphisms (SNPs) are one of the common genetic variations in which there is a single nucleotide base different among individuals. They occur randomly throughout the genome, and different individuals have different SNPs. Most SNPs are neutral and do not affect human's health. However, these mutations can affect the susceptibility to disease or cancer or response to a certain drug. That is why, they can be dealt with as biological markers that help scientists in locating genes associated with diseases. In this project, I aim to detect the SNPs in genomic sequences (SNP calling analysis) which could be used in determining whether these variations cause a predisposition to certain diseases and their effect on the protein in case of non-synonymous variants. Using a variant call format file (VCF) extracted from the 1000 genome project database, I will investigate SNPs in Transient Receptor Potential Vanilloid (TRPV1, TRPV2, TRPV3) gene family on chromosome 17, Breast cancer 1 gene (BRCA1) on chromosome 17 and Breast cancer 2 gene (BRCA2) on chromosome 13. This project is implemented using R.

Results: There were a total of 132 variants in the TRPV gene family where 32 were found in the TRPV2, 49 in the TRPV3 and 51 in the TRPV1. Moreover, there were 97 variants in BRCA1 and 173 in BRCA2. The variants located in all of them was mostly located in the coding region (nonsynonymous SNPs). This means that they affect the structure of the resulting protein and will lead to predispositions for disease which in case of BRCA1 and BRCA2 will be breast and ovarian cancer.

Availability: The quick brown fox jumps over the lazy dog.

Contact: s-mayarmansour@zewailcity.edu.eg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Detecting genetic variation in the human genome is very important for understanding the variation in phenotype, especially in coding regions or regulatory regions. These mutations can affect the susceptibility to disease or cancer. Single Nucleotide Polymorphisms (SNPs) are the most common genetic variation where a single nucleotide either adenine, thymine, cytosine or guanine changes from one individual to the other in the same species. Scientists found that they occur almost once every 1000 Nucleotide [1].

Breast malignant growth has risen as the second significant disease type, including practically 25% of all tumors among ladies [2]. Breast cancer is the most frequently occurring disease among ladies with frequency rates differing broadly over the world, having rates going from 27 for each 100,000 in Middle Africa and Eastern Asia to 92 in Northern America [2]. BRCA1 gene located on chromosome 17q21 and BRCA2 gene located on chromosome 13q12-13 act as a tumor suppressor as they play a role in preventing abnormal cell growth and proliferation [3,4]. Moreover, they are involved in the pathway that repairs double-strand breaks in DNA via homologous recombination, indicating that they play a role in maintaining genomic stability. Studies showed that inheriting specific mutations in BRCA1 and BRCA2 play a significant hazard factor in breast malignancy. It has been evaluated that 5–10% of all breast tumors are because of inherited inclination [5]. BRCA1 and BRCA2 have been appeared to incline to familial breast malignancy. Breasts cancer has been connected to BRCA1 in an expected 52% of high-hazard breast malignant growth families gathered by the Breast Cancer Linkage Consortium and to BRCA2 in 32% of families [5]. Mutations in BRCA1 and BRCA2 are uncommon in sporadic breast tumors. However, allelic irregularity in BRCA1 and BRCA2 loci, have been observed often in sporadic breast tumors [6]. As a result of the decline in BRCA1 expression in sporadic breast malignancy, different instruments, like CpG island methylation, were examined, but they did not explain the loss of heterozygosity [6]. This confirmed that the tumor suppressors (BRCA1 & BRCA2) play a significant role in breast malignancy.

Transient Receptor Potential Vanilloid (TRPV) gene family is one of the seven subfamilies of the transient receptor potential (TRP) which includes six members TRPV1, TRPV2, TRPV3, TRPV4, TRPV5 and TRPV6. TRP channels are cationic channels acting as signal transducers by altering membrane potential or intracellular calcium (Ca2+) concentration [7]. However, the TRPV subfamily members are expressed in sensory neurons, activated by capsaicin, the active ingredient in chillies, by temperature higher than 42 °C and by protons [7]. In this paper, I will focus on three members: TRPV1, TRPV2 and TRPV3. They

are located on chromosome 17. TRPV1 is expressed in regions of the brain, spinal cord, skin, and tongue; TRPV2 is in the brain, vascular smooth muscle cells, intestines, and macrophages and TRPV3 is in keratinocytes, brain, and testis [8]. As they act as thermoreceptors, the mutations in the TRPV family can lead to multiple complications depending on the tissue they are in. For example, TRPV1activation is involved in the regulation of different physiological functions, such as the release of inflammatory mediators in the body, gastrointestinal motility function, and temperature regulation, and mutation in TRPV1 can lead to digestive diseases like gastric ulcer [9].

Variant Call Format (VCF) is a format for handling all DNA polymorphism data like SNPs, indel and structural variants that is highly annotated [10]. It was first developed by the 1000 genome project but currently widely used in SNPs calling. In this paper, a VCF file downloaded from [11] will be used to investigate SNPs found in TRPV1, TRPV2, TRPV3, BRCA1 and BRCA2 exploring their location in respect to the gene structure whether they in exon, introns, untranslated regions (UTR), splice site, promotor or intergenic regions. Moreover, in case of non-synonymous variants, the amino acid change will be predicted.

2 Related work and Survey

The most used SNP identification methodologies for Next Generation Sequencing (NGS) nowadays are MAQ [12], SOAPsnp [13], SNVmix2 [14] and Bcftools [15]. MAQ and SOAPsnp are fast assortment alignment instruments that utilize the mass division concluded assortment and the alignment. MAQ utilizes the blending statistics to assess the chance of every alignment read mistake, and the Bayesian factual model to evaluate the absolute last Genotype error likelihood. The joined binomial model is utilized to discover the SNP for SNVmix2, giving the confidence score for each SNP. bcftools and GATK use Samtools as the thought for assessing SNPs utilizing Bayesian possibility gauges. Burrows-Wheeler Aligners (BWA) or Bowtie differentiate is by and by an unbelievably decent correlation of the product; however, there are various issues [16], for example, false positive, and false negative. That is why the workflow used in the paper is based on using a VCF file build based on mrsfast, not BWA, which is a hash-based mapper. Msrfast, as a hash-based indexer provides higher sensitivity and higher precision when dealing with mismatch and indel errors [17].

3 Problem Definition

As mentioned before, SNP Calling is very important as SNPs can be used as biomarkers for a predisposition for a disease or cancer and susceptibility to a specific drug. Breast and ovarian cancer predispositions are very important to detect as it affects the lives of many women worldwide as they can easily remain undetected until the last stages. Moreover, TRPV family as their expression affect the response to pain, anxiety and different surrounding temperature, it is very urgent to detect the SNPs found in their gene in order to predict the different consequences of their malfunction. That is why, the proposed workflow will take the VCF file entered by the user and investigate chromosome 17 specifically the genome ranges for BRCA1, TRPV1, TRPV2 and TRPV3 and chromosome 13 (13q12-13) for BRCA2. After extracting the genome ranges for the four genes the program print the SNPs found in them then identify the location of the variant relative to the gene function. The program also shows how many variants affect multiple gene along with the consequences of the variant whether it will lead to non-synonymous or synonymous or missense mutation or untranslated or frameshift and their effect on the protein in case of non-synonymous mutations.

4 The R Workflow

First, in order for the program to work, the R version installed should be version 4.

The coding pipeline for the workflow is as follows:

The first step is reading the VCF file entered by the user using the scanVcfHeader function and printing a part of the fixed columns in the VCF file along with the metadata.

The second step is searching using the gene symbol of the targeted genes to be able to extract their ENTREZ gene id and use it to extract genomic ranges later on

The third step is loading the txdb object that holds the genomic annotation for hg19 from UCSC. It will be be used later as a reference to extract the chromosome ranges for the targeted genes and to locate the position of the variant relative to the gene function (splice site, Five UTR, intron, exon,etc.). There is also another txdb object that can be loaded if the VCF file used was initially aligned to GRCh38 instead of GRCh37 (TxDb.Hsapiens.UCSC.hg38.knownGene)

The fourth step is to make sure that the VCF and the txdb used are of the in the same format as the VCF file was initially aligned to the NCBI genome GRCh37, but the txdb used in the program was build from UCSC, and there are some differences between both of them. That is why, in this step, the txdb was modified to match the VCF file.

The fifth step is to use the data extracted from step two, three and four to extract the ranges of the targeted genes from the VCF file and return them in a new VCF object then showing a fixed column of this VCF that correspond to these ranges.

The sixth step is using the locatevariant() function, with the txdb and the VCF created in step five as its arguments, to give the region of the variants of the genes related to their function (splice site, Five UTR, intron, exon,etc.) and then show a summary for the number of variants in each location. This step includes a summary of the number of variants per gene and the number of variants found in more than one gene.

The seventh step is the last step where predictCoding() function is used, with the modified VCF from step five, the txdb and the species to which the data belongs, to find the consequences of these variants on the amino acids in case of nonsynonymous variants and mention if they will lead to synonymous, frameshift, missense or non-translated variants. However, before applying the predictCoding() function, both the VCF file and the txdb object must be in UCSC format. Moreover, this step provides a summary of the consequences of the variants per gene and the number of variants that match more than one gene.

5 Results/Evaluation/Experiments

5.1 Data Details of Data set utilized

The VCF file used in this project was downloaded from [11], where it was produced from phase 1 of the 1000 Genomes Project. The date in the file resulted after aligning the samples from 1092 individuals from 14 different populations [18]. VCF file has three segments: a metadata region, a fixed region and a GT area. The metadata is at the start of the document. The data in the meta area characterizes the abbreviations utilized across the file. Underneath the metadata region, the information is in a tabular form. The initial eight sections of this table contain data about every variation. This information might be basic over all variations, for example, its chromosomal position, or a synopsis over all examples, for example, quality measurements. This information is fixed over all files. The fixed area is required in a VCF record, however there are other optional fields that can also be found [19].

```
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage
from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype
Likelihoods">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/ref
erence/ancestral_alignments/README">
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele
Frequency based on AC/AN">
##INFO=<ID=AMR_AF,Number=1,Type=Float,Description="Allele
Frequency for samples from AMR based on AC/AN">
##INFO=<ID=ASN_AF,Number=1,Type=Float,Description="Allele
Frequency for samples from ASN based on AC/AN">
##INFO=<ID=AFR_AF,Number=1,Type=Float,Description="Allele
Frequency for samples from AFR based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele
Frequency for samples from EUR based on AC/AN">
##INFO=<ID=VT,Number=1,Type=String,Description="indicates what
type of variant the line represents">
##INFO=<ID=SNPSOURCE,Number=.,Type=String,Description="indicates
if a snp was called when analysing the low coverage or exome
alignment data">
##reference=GRCh37
##INFO=<ID=VA,Number=.,Type=String,Description="Variant
Annotation, gencode7.coding.interval">
#CHROM     POS   ID    REF   ALT   QUAL  FILTER     INFO
1     13302 rs180734498     C     T     100   PASS  THETA=0.0048;AN=
2184;AC=249;VT=SNP;AA=.;RSQ=0.6281;LDAF=
0.1573;SNPSOURCE=LOWCOV;AVGPOST=0.8895;ERATE=0.0058;AF=
0.11;ASN_AF=0.02;AMR_AF=0.08;AFR_AF=0.21;EUR_AF=0.14;VA=
1:AL627309.2:ENSG00000249291.2:
+:synonymous:1/1:AL627309.2-201:ENST00000515242.2:384_225_75_H->H
1     13327 rs144762171     G     C     100   PASS  AVGPOST=
0.9698;AN=2184;VT=SNP;AA=.;RSQ=0.6482;AC=
59;SNPSOURCE=LOWCOV;ERATE=0.0012;LDAF=0.0359;THETA=0.0204;AF=
0.03;ASN_AF=0.02;AMR_AF=0.03;AFR_AF=0.02;EUR_AF=0.04;VA=
1:AL627309.2:ENSG00000249291.2:
+:nonsynonymous:1/1:AL627309.2-201:ENST00000515242.2:384_250_84
_G->R
```

Fig.1: Shows Sample of the data in the VCF file used

5.2 The List of questions the experiments are designed to answer

The project was investigating the TRPV gene family and the BRCA1/2 to answer the following questions

(1) How to find the SNPs in these genes?

(2) What is their location relative to the gene function? (splice site, Five UTR, intron, exon,etc.)

(3) Are they shared with more than one gene?

(4) What are the consequences of these variants? (synonymous, frameshift, missense or non-translated)

(5) What will the protein be affected in case of nonsynonymous protein?

5.3 Observations and the Interpretations

The coding pipeline of the workflow can be divided into 3 parts:1- preparing the input file to be used to answer the designed questions, 2-finding the SNPs in the targeted genes and relate their location to gene function and 3- Predicting the consequences of these variants and their effect on the amin acid

5.3.1 Preparing the input file to be used to answer the designed questions

As the file used in the experiment contains data for the whole genome not the specific regions we need to investigate, it was important to parse the file first to extract the desired regions. Figure 2 shows the code that is used to extract the gene id for the desired gene using them to locate the ranges from the txdb object which acts as the reference and forming a new VCF file containing the SNPs in the desired regions only.

```
71 - ## ---------------------------------------------------------------------
72   hdr <- scanVcfHeader(file)
73   info(hdr)
74   fixed(hdr)
75   geno(hdr)
76
77 - ## ---------------------------------------------------------------------
78   meta(hdr)
79
80 - ## ---------------------------------------------------------------------
81   ## get entrez ids from gene symbols
82   genesym <- c("TRPV1", "TRPV2", "TRPV3", "BRCA1","BRCA2")
83   geneid <- select(org.Hs.eg.db, keys=genesym, keytype="SYMBOL",
84            columns="ENTREZID")
85   geneid
86 - ## ---------------------------------------------------------------------
87   txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
88 - ## ---------------------------------------------------------------------
89
90   txdb <- renameSeqlevels(txdb, gsub("chr", "", seqlevels(txdb)))
91   txdb <- keepSeqlevels(txdb, c("17","13"))
92
93 - ## ---------------------------------------------------------------------
94   txbygene = transcriptsBy(txdb, "gene")
95
96 - ## ---------------------------------------------------------------------
97   gnrng <- unlist(range(txbygene[geneid$ENTREZID]), use.names=FALSE)
98   names(gnrng) <- geneid$SYMBOL
99
100 - ## ---------------------------------------------------------------------
101  param <- ScanVcfParam(which = gnrng, info = "DP", geno = c("GT"))
102  param
103
104  ## Extract the TRPV1 ,TRPV2 TRPV3 BRCA1 BRAC2 ranges from the VCF file
105  vcf <- readVcf(file, "hg19", param)
106
107  ## Inspect the VCF object with the 'fixed', 'info' and 'geno' accessors
108  vcf
109  info(vcf)
110  head(fixed(vcf))
111
112  geno(vcf)
```

Fig.2: Code for achieving Part 1 (Preparing the input file)

The results from this part was shown in figures 3 and 4 where it is clear that there was significant decrease in the size of the data after the filtration (402 SNPs) and shows the ENTREZID for the targeted gene.

```
>
> txdb <- renameSeqlevels(txdb, gsub("chr", "", seqlevels(txdb)))
> txdb <- keepSeqlevels(txdb, c("17","13"))
>
> ## ---------------------------------------------------------------------
> txbygene = transcriptsBy(txdb, "gene")
>
> ## ---------------------------------------------------------------------
> gnrng <- unlist(range(txbygene[geneid$ENTREZID]), use.names=FALSE)
> names(gnrng) <- geneid$SYMBOL
>
> ## ---------------------------------------------------------------------
> param <- ScanVcfParam(which = gnrng, info = "DP", geno = c("GT"))
> param
class: ScanVcfParam
vcfWhich: 2 elements
vcfFixed: character() [All]
vcfInfo: DP
vcfGeno: GT
vcfSamples:
>
> ## Extract the TRPV1 ,TRPV2 TRPV3 BRCA1 BRAC2 ranges from the VCF file
> vcf <- readVcf(file, "hg19", param)
ScanVcfParam ♦info♦ fields not found in  header: ♦DP♦
>
> ## Inspect the VCF object with the 'fixed', 'info' and 'geno' accessors
> vcf
class: CollapsedVCF
dim: 402 0
rowRanges(vcf):
  GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
info(vcf):
  DataFrame with 1 column: INFO
  Fields with no header: INFO
geno(vcf):
  List of length 1: GT
geno(header(vcf)):
```

Fig.3: Shows size of the data after filteration and extracting the SNPs in the targeted genes

	SYMBOL	ENTREZID
1	TRPV1	7442
2	TRPV2	51393
3	TRPV3	162514
4	BRCA1	672
5	BRCA2	675

Fig.4: Shows the ENTREID for the targeted genes

5.3.2 Finding the SNPs in the targeted genes and Relating their location to gene function

After extracting the genomic ranges for the targeted genes, the new VCF created was used to show a summary for number of SNPs per gene (the distribution for the 402 SNPs among the genes). Moreover, using locatevariant() function, the program also show information related to each SNP and their location relative to the gene function. Figure 5 shows the coding for this part.

```
## ---------------------------------------------------------------------
#Use the 'region' argument to define the region of interest.
cds <- locateVariants(vcf, txdb, CodingVariants())
cds
five <- locateVariants(vcf, txdb, FiveUTRVariants())
five
splice <- locateVariants(vcf, txdb, SpliceSiteVariants())
splice
intron <- locateVariants(vcf, txdb, IntronVariants())
intron
## ---------------------------------------------------------------------
#Extract all variants related to all gene function location
all <- locateVariants(vcf, txdb, AllVariants())
 all

## ---------------------------------------------------------------------
## Did any variants match more than one gene?
table(sapply(split(mcols(all)$GENEID, mcols(all)$QUERYID),
       function(x) length(unique(x)) > 1))

## Summarize the number of variants by gene:
idx <- sapply(split(mcols(all)$QUERYID, mcols(all)$GENEID), unique)
sapply(idx, length)

```

Fig.5: Shows coding for extracting the SNPs accroding to gene function

Figure 6 shows that there were 132 variants in the TRPV gene family: 32 were found in the TRPV2, 49 in the TRPV3 and 51 in the TRPV1. Moreover, there were 97 variants in BRCA1 and 173 in BRCA2. There are also 8 SNPs that are found in more than one gene. As for the location for each SNP in the gene function, in case of TRPV1, 50 SNPs where in the coding region while 1 where in introns. TRPV2, had all 32 SNPs in the coding regions. TRPV3 had 47 in the coding regions,2 in the introns while 1 SNPs was in the splice site. BRAC1 had 38 in the coding regions, 59 in introns and 1 in splice sites. BRCA2 had 165 in the coding regions, 2 in introns and 6 in the five UTR regions.

```
> ## Did any variants match more than one gene?
> table(sapply(split(mcols(all)$GENEID, mcols(all)$QUERYID),
+       function(x) length(unique(x)) > 1))

FALSE  TRUE
  394     8
>
> ## Summarize the number of variants by gene:
> idx <- sapply(split(mcols(all)$QUERYID, mcols(all)$GENEID), unique)
> sapply(idx, length)
 10230 162514  51393    672    675   7442  84690
     4     49     32     97    173     51      4
>
> ## Summarize variant location by gene:
> sapply(names(idx),
+        function(nm) {
+          d <- all[mcols(all)$GENEID %in% nm, c("QUERYID", "LOCATION")]
+          table(mcols(d)$LOCATION[duplicated(d) == FALSE])
+        })
          10230 162514 51393 672 675 7442 84690
spliceSite     0      1     0   1   0    0     0
intron         3      2     0  59   2    1     0
fiveUTR        0      0     0   0   6    0     0
threeUTR       0      0     0   0   0    0     0
coding         0     47    32  38 165   50     0
intergenic     0      0     0   0   0    0     0
promoter       1      0     0   0   0    0     4
>
```

Fig.6: Summary for the information extracted in Part 2 of the pipeline

```
> cds
GRanges object with 1897 ranges and 9 metadata columns:
                 seqnames    ranges strand | LOCATION  LOCSTART    LOCEND   QUERYID
                    <Rle> <IRanges>  <Rle> | <factor> <integer> <integer> <integer>
      rs146183848       17   3470167      - |   coding      2495      2495         1
      rs146183848       17   3470167      - |   coding      2282      2282         1
      rs146183848       17   3470167      - |   coding      2456      2456         1
      rs146183848       17   3470167      - |   coding      2462      2462         1
      rs146183848       17   3470167      - |   coding      2462      2462         1
              ...      ...       ...    ... .      ...       ...       ...       ...
 13:32972673_C/A       13  32972673      + |   coding     10023     10023       398
 13:32972760_G/A       13  32972760      + |   coding     10110     10110       399
       rs56309455      13  32972771      + |   coding     10121     10121       400
 13:32972852_C/T       13  32972852      + |   coding     10202     10202       401
       rs1801426       13  32972884      + |   coding     10234     10234       402
                      TXID      CDSID      GENEID       PRECEDEID         FOLLOWID
                <character> <IntegerList> <character> <CharacterList> <CharacterList>
      rs146183848     62409     182516        7442
      rs146183848     62410     182516        7442
      rs146183848     62411     182516        7442
      rs146183848     62412     182516        7442
      rs146183848     62413     182516        7442
              ...       ...        ...         ...              ...              ...
 13:32972673_C/A     49484     147353         675
 13:32972760_G/A     49484     147353         675
       rs56309455    49484     147353         675
 13:32972852_C/T     49484     147353         675
       rs1801426     49484     147353         675
  -------
```

Fig.7: Shows the sample of the SNPs found in coding regions along with their positions

5.3.3 Predicting the consequences of these variants and their effect on the amin acid

Since figure 6 shows that most of the SNPs are in the coding regions, it was important to investigate their effect on the amino in order to find what consequences they will lead to. In case of nonsynonymous variants i.e. the variants will lead to a different amino acid than in the reference genome, it is crucial to find what the new amino acid will be to infer what will happen to the protein resulting from this gene or the gene function in general. These all will be done using predictCoding() function. The coding part for this step is show in figure 8. Moreover, this step offers a summary for the number of variants per gene.

```
seqlevelsStyle(vcf) <- "UCSC"
seqlevelsStyle(txdb) <- "UCSC"
aa <- predictCoding(vcf, txdb, Hsapiens)
aa
## ---------------------------------------------------------------------
## Did any variants match more than one gene?
table(sapply(split(mcols(aa)$GENEID, mcols(aa)$QUERYID),
  function(x) length(unique(x)) > 1))

## Summarize the number of variants by gene:
idx <- sapply(split(mcols(aa)$QUERYID, mcols(aa)$GENEID, drop=TRUE), unique)
sapply(idx, length)

## Summarize variant consequence by gene:
sapply(names(idx),
       function(nm) {
         d <- aa[mcols(aa)$GENEID %in% nm, c("QUERYID","CONSEQUENCE")]
         table(mcols(d)$CONSEQUENCE[duplicated(d) == FALSE])
       })
```

Fig.8: Shows coding for predicting the consequences for these variants

Figure 9 shows the consequences of the variants as a Granges object with 17 columns. Form these columns, REF and ALT shows the SNP in the reference genome and the alternative SNP in the sample respectively. Moreover, columns as CONSEQUENCE, REFCODON, VARCODON, REFAA and VARAA are also important as one shows what type of variants this is (nonsynonymous, synonymous, nonsense), the triplet codon in the reference genome, the triplet codon in the caused by the SNP, the original amino acid the genome and the amino acid change caused by the SNP respectively.

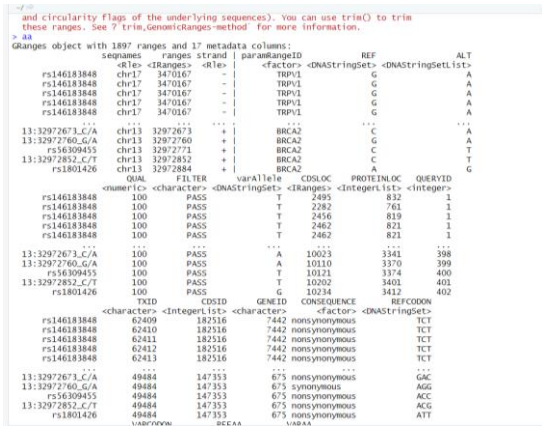

Fig.9: Result from the PredictCoding() function of the consequences for the SNPs

The workflow after using predictCoding to compare the file with the reference shows that there are 48 variants in TRPV3 (28 are nonsynonymous and 20 synonymous), 32 in TRPV2 (18 are nonsynonymous and 14 are synonymous), 51 in TRPV1 (28 are nonsynonymous and 23 are synonymous), 90 in BRCA1 (63 are nonsynonymous and 27 are synonymous) and 171 in BRCA2 (2 are nonsense, 28 are nonsynonymous and 23 are synonymous)

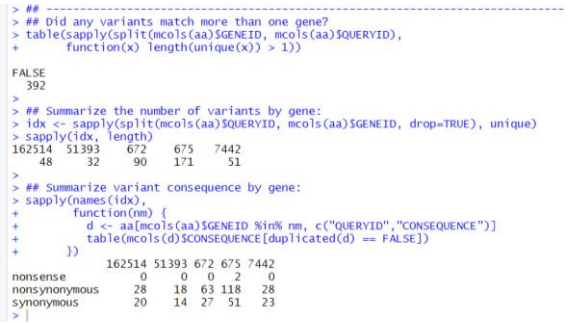

Fig.10: Summary for the reults of predictCoding() function

Figure 10 shows that most of the variants in the genes are nonsynonymous ensuring that they act as predispositons for disease and improper gene function. For example in case of BRCA1 and BRCA2 they act as predisposition for breast and ovarian cancer.

5.4 Comparison with another tools

As mentioned earlier, the most common tools used in SNP calling are samtool, bcftools and GATK. Both samtool and bcftools are command line based while other tools like Variant Effect Predictor (VEP) on ensemble are web-based tool. Web based tools in general form an issue as there is a limit for the size of files you can upload. Despite this disadvantage in web-based tool, they are still a good option when dealing with small dataset as they integrate different database to offer more information related to your query. When comparing command line-based tools with the web based one, they do not suffer from the size limit problem but some functions can only run on Linux system and do not support windows. This can cause a problem to those who work with windows not Linux. There is also another disadvantage with the command line tools which is the lack of interaction with various databases that is found in web-based tools. Now, when comparing this workflow this both sides, it takes some advantages from both, as it is independent to the operating system, it only depends on the R version used. It also does not have a size limit. However, the downfall of the workflow is that it is not connected to that many databases to further examine the variants. For example, it can't predict whether these SNPs have been clinically linked to a disease or not.

6. Conclusion

To conclude, the workflow works efficiently and very user friendly. It can parse the VCF file entered by the user to find out the SNPs found in the targeted genes. Even though the targeted genes are determined in the workflow to be TRPV1, TRPV2, TRPV3, BRCA1 and BRCA2 on chromosomes 17 and 13 respectively, the user can change them easily to his targeted ones. It can predict the consequences of these SNPs and the change in amino acid resulted from the nonsynonymous variants.

7. Suggestion and Improvement

To further investigate these SNPs, the workflow should be connected with the OMIM database or the Clinical database in NCBI for example to be able to ensure that they are clinically linked to diseases.

Acknowledgements

This was done under the supervision of Dr Eman Mostafa.I would like to thank Dr Eman for her effort and guidance through out the semester along with Ms. Nehal and Mr. Mohamed Kerdawy.

Funding

This work was not funded

Conflict of Interest: none declared.

References

[1] G. Reference, "What are single nucleotide polymorphisms (SNPs)?", Genetics Home Reference, 2019. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp?fbclid=IwAR0tJhUIP2QNuNm9bjo5Yq4zbO2ltlAVPtuK6NBkU8xf9EAeQNJK2VU5uCY.

[2] S. Mylavarapu, A. Das and M. Roy, "Role of BRCA Mutations in the Modulation of Response to Platinum Therapy", Frontiers in Oncology, vol. 8, 2018. Available: 10.3389/fonc.2018.00016

[3] G. Reference, "BRCA1 gene", Genetics Home Reference. [Online]. Available: https://ghr.nlm.nih.gov/gene/BRCA1.

[4] G. Reference, "BRCA2 gene", Genetics Home Reference. [Online]. Available: https://ghr.nlm.nih.gov/gene/BRCA2.

[5] A. Forsti, "Allelic imbalance on chromosomes 13 and 17 and mutation analysis of BRCA1 and BRCA2 genes in monozygotic twins concordant for breast cancer", Carcinogenesis, vol. 22, no. 1, pp. 27-33, 2001. Available: 10.1093/carcin/22.1.27
[6] T. Rebbeck et al., "Association of Type and Location ofBRCA1andBRCA2Mutations With Risk of Breast and Ovarian Cancer", JAMA, vol. 313, no. 13, p. 1347, 2015. Available: 10.1001/jama.2014.5985
[7] B. Nilius and G. Owsianik, "The transient receptor potential family of ion channels", Genome Biology, vol. 12, no. 3, p. 218, 2011. Available: 10.1186/gb-2011-12-3-218 .
[8] C. Colton and M. Zhu, "2-Aminoethoxydiphenyl Borate as a Common Activator of TRPV1, TRPV2, and TRPV3 Channels", Transient Receptor Potential (TRP) Channels, pp. 173-187. Available: 10.1007/978-3-540-34891-7_10
[9] Q. Du, Q. Liao, C. Chen, X. Yang, R. Xie and J. Xu, "The Role of Transient Receptor Potential Vanilloid 1 in Common Diseases of the Digestive Tract and the Cardiovascular and Respiratory System", Frontiers in Physiology, vol. 10, 2019. Available: 10.3389/fphys.2019.01064
[10] P. Danecek et al., "The variant call format and VCFtools", 2011.
[11] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/annotated_vcfs/
[12] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008; 18: 1851-1858.
[13] Li R, Li Y, Fang X, Yang H, Wang J. SNP detection for massively parallel wholegenome resequencing. Genome Res 2009; 19: 1124-1132.
[14] Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008; 24: 713-714.
[15] Goya R, Sun MG, Morin RD, Leung G, Ha G. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 2010; 26: 730-736
[16] Chen, G. & Xie, X.. (2017). A light weight SNP detection algorithm for the breast cancer targeted sequencing data. Biomedical Research (India). 28. 3574-3579.
[17] F. Hach, I. Sarrafi, F. Hormozdiari, C. Alkan, E. Eichler and S. Sahinalp, "mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications", Nucleic Acids Research, vol. 42, no. 1, pp. W494-W500, 2014. Available: 10.1093/nar/gku370
[18] "Data | 1000 Genomes", Internationalgenome.org, 2015. [Online]. Available: https://www.internationalgenome.org/data.
[19] "VCF (Variant Call Format) version 4.0 | 1000 Genomes", Internationalgenome.org, 2020. [Online]. Available: https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/.