

May 24, 2023

Investigation and identification of somatic and germline variants for colorectal cancer exomes using the NGS pipeline: a computational analysis perspective

DOI
<https://dx.doi.org/10.17504/protocols.io.x54v9dqxqg3e/v1>
Chandrashekar K¹, Anagha S Setlur¹, Vidya Niranjana¹
¹Department of Biotechnology, RV College of Engineering, Bangalore- 560059



 **Vidya Niranjana**
R V College of Engineering

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.x54v9dqxqg3e/v1>

Protocol Citation: Chandrashekar K, Anagha S Setlur, Vidya Niranjana 2023. Investigation and identification of somatic and germline variants for colorectal cancer exomes using the NGS pipeline: a computational analysis perspective. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.x54v9dqxqg3e/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited



Protocol status: Working

We use this protocol and it's working

Created: May 23, 2023

Last Modified: May 24, 2023

Protocol Integer ID: 82294

Keywords: Colorectal cancer exomes, quality controls, somatic & germline variants, annotation & post-processing, computational analysis, germline variants for colorectal cancer exome, colorectal cancer exome, cancer exome dataset, germline variant identification, colorectal cancer, identified variant, germline mutant, human reference genome, germline variant, identified mutant, potential mutations such as single nucleotide polymorphism, potential mutation, important gene, using snpeff, variant identification, single nucleotide polymorphism, germline, haplotype caller, early disease detection, process of early disease detection

Disclaimer

This protocol was run with colorectal cancer exome datasets. However, the SRR-IDs of the datasets have not been mentioned to make this protocol universal. The steps may be followed for any cancer exome, with the intent of investigating the somatic and germline mutations.

Abstract

A thorough analysis on colorectal cancer exomes reveals potential mutations such as single nucleotide polymorphisms that can be beneficial for early detection of the disease. Thus, through a comprehensive computational protocol that identifies, investigates and analyzes the identified variants, this process of early disease detection becomes much simpler. In brief, the cancer exome datasets were retrieved from publicly available databases, followed by performing quality control checks. The datasets that qualified the quality control checks were then aligned with the human reference genome. Somatic and germline mutants were then identified and called separately, with specific tools for each case. Haplotype Caller was employed for germline variant identification, and Mutect2 for somatic. The identified mutants were then normalized, annotated and post-processed using snpEFF, WANNOVAR and VEP. This protocol helps in garnering insights on the various alterations that might possibly lead to colorectal cancer and suggests the possibility of utilizing the most important genes identified for wet-lab experimentation.

Keywords: Colorectal cancer exomes, quality controls, somatic & germline variants, annotation & post-processing, computational analysis

Guidelines

The commands mentioned in the protocol can be run in the Linux terminal.
Enough storage space is recommended while running this protocol.

Materials

Tools mentioned in the current protocol must initially be installed in the system before running the commands for the NGS pipeline. These include: SRAToolkit, Fastqc, Multiqc, Cutadapt, Bowtie2, Samtools, picard, GATK, and bcftools. wANNOVAR, VEP are available online and snpEFF is command line based.

Safety warnings

 Nil

Ethics statement

The datasets used to arrive at this protocol were all from publicly available databases such as NCBI-SRA.

Before start

All the mentioned tools in the materials section must be installed and checked for appropriate installation.

COLORECTAL CANCER EXOME RETRIEVAL

1 Retrieval of colorectal cancer exomes

This protocol was run with colorectal cancer exome datasets. However, the SRR-IDs of datasets have not been mentioned to make this protocol universal. Therefore, NCBI-SRA database (<https://www.ncbi.nlm.nih.gov/sra>) was employed to retrieve the colorectal cancer exome datasets. The SRR-IDs were noted down for further use. The exome datasets were selected based on a set of criteria. The strategy used for the datasets must be whole exome sequenced and the exomes should be of genomic source. In the present protocol, all the selected sequences were of paired layout, with Illumina used for sequencing and acquiring the reads. Other criteria as desired by the user depending on the requirement can be employed for retrieving the cancer exomes.

Additionally, the tools mentioned in the current protocol must initially be installed in the system before running the commands for the NGS pipeline. These include: SRAToolkit, Fastqc, Multiqc, Cutadapt, Bowtie2, Samtools, picard, GATK, and bcftools. wANNOVAR, VEP are available online and snpEFF is command line based.

QUALITY CONTROL

2 Quality control assessments

Prior to calling variants, the quality of raw data was assessed using the below-mentioned steps. The codes used for running the quality checks are provided in the following sections.

2.1 Quality Check using FastQC and MultiQC (for multiple datasets)

FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and MultiQC (<https://multiqc.info/>) in the present study were used to effectively assess the sequencing data quality, to identify any possible issues with the raw sequence reads and generate informative reports. Scrutiny of the mean sequence quality per reading and per base, nucleotide content per base position, distribution of GC, etc were analyzed via FastQC. A cumulative report for all FastQC outcomes was obtained from MultiQC.

For multiple datasets, FastQC and MultiQC commands used were:

Command

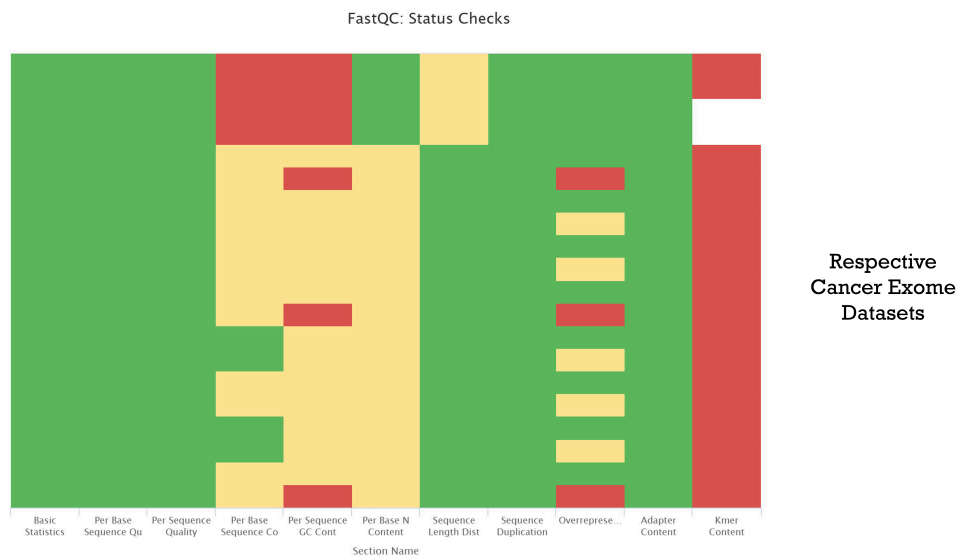
fastQC

```
$ ./fastqc
```

Command

multiQC

```
$ multiqc .
```



Expected multiQC heatmap

2.2 Adaptor sequence removal via Cutadapt

This tool identifies and eliminates the adaptor sequences from raw reads and improves the quality and accuracy of downstream analysis. All unwanted sequences were thus removed using Cutadapt (<https://cutadapt.readthedocs.io/en/stable/>) by running:

Command

Running Cutadapt

```
$ cutadapt -a adapter-sequence* -o SRRID_cut.fastq.gz SRRID.fastq.gz
```

ALIGNMENT WITH HUMAN REFERENCE GENOME

3 Retrieval and indexing of human reference genome for analysis

Indexing is a necessary step to pre-process the reference genome so the aligner can seek potential alignment locations for the reads. This is useful during alignment with Bowtie2 (<https://bowtie-bio.sourceforge.net/bowtie2/index.shtml>), where short read aligners are processed with the reference genome. The following codes were run in the terminal for obtaining Bowtie2 outputs.

Command

Unzipping the reference genome file

```
$ gunzip GRCh37_latest_genomic.fna.gz
```

Command

Indexing reference genome

```
$ bowtie2-build -f GRCh37_latest_genomic.fna.gz hg19
```

3.1 Alignment of indexed human reference genome against exome datasets

Once indexing is complete, alignment was performed using Bowtie2.

Command

Gapped alignment command with human reference genome (hg19 in this case)

```
$ bowtie2 -x hg19/38/37 -1 SRRID_1_cut.fastq.gz -2  
SRRID_2_cut.fastq.gz -S SRRID.sam $ head SRRID.sam
```

Running Bowtie2 with the specified parameters allows alignment of paired-end sequencing reads to a reference genome, enabling downstream analysis such as variant calling, differential gene expression, or identification of genomic features specific to colorectal cancer.

SAM TO BAM CONVERSION

4 SAM (Sequence Alignment/Map) to BAM (Binary Alignment/Map) conversion

To obtain the alignment outcomes in a readable format and to act as appropriate input for the next steps, SAM to BAM (<http://www.htslib.org/>) conversion was performed. The code used was:

**Command****SAM to BAM file conversion**

```
$ samtools view -bS SRRID.sam > SRRID.bam
```

4.1 **Sorting BAM**

BAM sorting and merging was performed using command:

Command**BAM sorting**

```
$ samtools sort SRRID.bam -o SRRID.sorted.bam
```

BAM sorting helps in supporting quick retrieval of alignments and has a compact size for further analysis.

4.2 **Identifying and marking duplicate reads in BAM file**

Duplicate reads may arise at times during sequencing due to several factors, inclusive of technical biases, PCR amplification artifacts and preparation of library protocols. These may skew the downstream analyses and impact the result accuracy. Therefore, marking duplicates is essential to address this issue, by using the “MarkDuplicates” function of Picard (<https://broadinstitute.github.io/picard/>) tool.

Command

Marking Duplicates

```
$ java -jar picard.jar MarkDuplicates INPUT=SRRID_sorted_reads.bam  
OUTPUT=SRRID_dedup_reads.bam METRICS_FILE=SRRID_metrics.txt
```

4.3 Sorting BAM files for deleted duplicate files

To improve the compression efficiency, another round of BAM sorting was performed for the deleted duplicate files.

Command

BAM sorting for deleted duplicate files

```
$ samtools sort SRRID_dedup_reads.bam -o SRRID_sorted_dedup_reads.bam
```

4.4 Computing and collecting alignment summary metrics from BAM file

To obtain various statistics and metrics associated with alignment of sequencing reads to a reference genome, the "CollectAlignmentSummaryMetrics" function was used. The command used to run this function was:

Command

Collecting Alignment Summary Metrics

```
$ java -jar picard.jar CollectAlignmentSummaryMetrics R=ref.fa  
I=SRRID_sorted_dedup_reads.bam O=SRRID_alignment_metrics.txt
```

ref.fa: fasta sequence file of the human reference genome

4.5 **Estimating and collecting the insert size distribution of paired-end reads from BAM file**

The primary purpose of collecting the insert sizes is to check the distribution and characteristics of the DNA fragment sizes, in a sequencing library. This enables the collection and visualization of the insert size metrics and distribution, that are useful for the quality scrutiny of the library, alignment of the reads, selection of the fragment size and quality control assessments in several genomic analyses. For this purpose, in the present work flow, the "CollectInsertSizeMetrics" function was used. The distribution of the paired end reads from BAM was analyzed using the command:

Command

Collecting Insert Size Metrics

```
$ java -jar picard.jar CollectInsertSizeMetrics  
INPUT=SRRID_sorted_dedup_reads.bam OUTPUT=SRRID_insert_metrics.txt  
HISTOGRAM_FILE=SRRID_insert_size_histogram.pdf
```

4.6 **Calculating coverage depth at each position in a sorted and duplicate-marked BAM file**

The "samtools depth" function was used to examine the depth of coverage at every position in the genome of reference depending on the aligned reads. Computation and estimation of depth of coverage important for assessing the sequence data

quality, calling of variants, structure and gene expression analysis and overall exploration of data becomes much easier with the samtools depth function. Thus, to better understand and obtain insights into the coverage depth, this function was performed via:

Command

Determining coverage depth of sorted, deleted, & duplicated BAMs

```
$ samtools depth -a SRRID_sorted_dedup_reads.bam > SRRID_depth_out.txt
```

4.7 Add or modify read group information in BAM file

To assign or update the read group information to the sequence reads in a BAM file, the add or replace read groups function was employed. This was carried out using "picard.jar AddOrReplaceReadGroups". Sample identification, tracking of the data, sequence data differentiation, integrity checks of data and overall sequence reads compatibility was obtained after this step was performed, ensuring that the reads were properly annotated and organized, facilitating meaningful data analyses.

This step was run using the following code in the terminal:

Command

Adding/modifying read group info

```
$ java -jar picard.jar AddOrReplaceReadGroups  
I=SRRID_sorted_dedup_reads.bam O=SRRID_output.bam RGID=4 RGLB=lib1  
RGPL=ILLUMINA RGPU=unit1 RGSM=20
```

4.8 Indexing BAM file- necessary before variant calling

Prior to variant calling, another indexing of the final BAM output file was carried out to allow faster retrieval of specific genomic regions, enhanced visualization, and to facilitate random access.

Command

Indexing BAM file

```
$ samtools index SRRID_output.bam
```

VARIANT CALLING - GERMLINE & SOMATIC

5 Calling of germline variants

The below sections detail preliminary steps followed for calling germline variants. Mutations from the sequenced data were identified, processed and the variants were called using PICARD (<https://broadinstitute.github.io/picard/>) and GATK (The Genome Analysis Toolkit, <https://github.com/broadinstitute/gatk>). The Haplotype caller function was used to call the germline variants.

The purpose of using the "HaplotypeCaller" function is to identify and call genetic variants, including single nucleotide variants (SNVs), insertions, deletions, and complex variants, based on the sequencing data and the provided reference genome. The following command was employed to call the variants:

Command

Germline variant calling

```
$ gatk HaplotypeCaller -R ref.fa -I SRRID_output.bam -O  
SRRID_raw_variants.vcf
```

5.1 Selection of SNPs from raw variants

The "SelectVariants" function with the "-select-type SNP" option was utilized to filter and extract only the SNPs from the original VCF file. This step enabled filtering and data refining to focus on SNPs, simplified further analyses, optimized computational resources, and ensured compatibility with SNP-specific analysis tools or pipelines. For working with larger datasets, this function works best for SNP-centric analyses. Therefore, to select the SNPs, the code run was:

Command

Selection of variants from raw mutants

```
$ gatk SelectVariants -R ref.fa -V SRRID_raw_variants.vcf --select-type-to-include SNP -o SRRID_raw_snps.vcf
```

5.2 Selection of INDELs from raw variants

The purpose of using the "SelectVariants" function with the "-select-type INDEL" option is to filter and extract only the INDELs from the original VCF file. Using the GATK "SelectVariants" with the "-select-type INDEL" option allows for the selection and extraction of INDELs from a VCF file. This step enables filtering and refining the data to focus on INDELs. This step was run prior to variant filtration.

Command

INDELs selection from raw mutants

```
$ gatk SelectVariants -R ref.fa -V SRRID_raw_variants.vcf --select-type-to-include INDEL -o SRRID_raw_indels.vcf
```

5.3 Variant filtration

To remove potential false positives and retain only the high quality variants, a series of filters were applied to the raw SNPs using the following codes:

Command

Calling SNPs

```
$ gatk VariantFiltration -R ref.fa -V SRRID_raw_snps.vcf -O  
SRRID_filtered_snps.vcf -filter-name "QD_filter" -filter "QD < 2.0" -  
filter-name "FS_filter" -filter "FS > 60.0" -filter-name "MQ_filter" -  
filter "MQ < 40.0" -filter-name "SOR_filter" -filter "SOR > 4.0" -  
filter-name "MQRankSum_filter" -filter "MQRankSum < -12.5" -filter-  
name "ReadPosRankSum_filter" -filter "ReadPosRankSum < -8.0"
```

Command

Calling INDELs

```
$ gatk VariantFiltration -R ref.fa -V SRRID_raw_indels.vcf -O  
SRRID_filtered_indels.vcf -filter-name "QD_filter" -filter "QD < 2.0"  
-filter-name "FS_filter" -filter "FS > 200.0" -filter-name  
"SOR_filter" -filter "SOR > 10.0"
```

5.4 Excluding filtered variants

To filter out the variants that do not meet the specific quality (as defined by the user), the "--exclude-filtered" option in the "SelectVariants" step was used. This ensured removal of mutants that failed specific criteria and thereby enhanced quality of the data, prioritized the high-confidence variants only, decreased the number of false positives, allowed compatibility access between datasets and also optimized the

computational resources. This is a crucial step to be followed in the data refinement process.

Command

Excluding SNPs

```
$ gatk SelectVariants --exclude-filtered -V SRRID_filtered_snps.vcf -  
0 SRRID_bqsr_snps.vcf
```

Command

Excluding INDELS

```
$ gatk SelectVariants --exclude-filtered -V SRRID_filtered_indels.vcf  
-0 SRRID_bqsr_indels.vcf
```

5.5 Recalibration of base quality

To augment the accuracy and reliability of the called variants, a base recalibration was carried out by correcting the systematic errors, addressing biases, and providing more accurate base quality scores. This step is essential in ensuring that the sequencing data quality is maintained appropriately.

**Command****Base recalibration- Step 1**

```
$ gatk BaseRecalibrator -R ref.fa -I SRRID_output.bam --known-sites  
SRRID_bqsr_snps.vcf --known-sites SRRID_bqsr_indels.vcf -O  
SRRID_recal_data.table
```

Command**Base recalibration- step 2**

```
$ gatk ApplyBQSR -R ref.fa -I SRRID_output.bam -bqsr  
SRRID_recal_data.table -O SRRID_recal_reads.bam
```

Command**Base recalibration- step 3**

```
$ gatk BaseRecalibrator -R ref.fa -I SRRID_recal_reads.bam --known-  
sites SRRID_bqsr_snps.vcf --known-sites SRRID_bqsr_indels.vcf -O  
SRRID_post_recal_data.table
```


Command

Base recalibration- step 4

```
$ gatk AnalyzeCovariates -before SRRID_recal_data.table -after  
SRRID_post_recal_data.table -plots SRRID_recalibration_plots.pdf
```

Command

Base recalibration- step 5

```
$ gatk HaplotypeCaller -R ref.fa -I SRRID_recal_reads.bam -O  
SRRID_raw_variants_recal.vcf
```

5.6 Selection of variants from recalibrated VCF files

To filter and extract only the SNPs from the recalibrated VCF files, functions "SelectVariants" along with "-selectType SNP" were used. This was useful since the data authors were working with were recalibrated variant sets.

Command

SNP selection

```
$ gatk SelectVariants -R ref.fa -V SRRID_raw_variants_recal.vcf  
-select-type-to-include SNP -O SRRID_raw_snps_recal.vcf
```

Command

INDEL selection

```
$ gatk SelectVariants -R ref.fa -V SRRID_raw_variants.vcf --select-type-to-include INDEL -O SRRID_raw_indels_recal.vcf
```

5.7 Variant filtration step

A series of filters were then applied to the raw SNPs to remove the potential false positives and retain high-quality variants. The following filters were applied:

Command

Filtration

```
$ gatk VariantFiltration -R ref.fa -V SRRID_raw_snps_recal.vcf -O SRRID_filtered_snps_final.vcf -filter-name "QD_filter" -filter "QD < 2.0" -filter-name "FS_filter" -filter "FS > 60.0" -filter-name "MQ_filter" -filter "MQ < 40.0" -filter-name "SOR_filter" -filter "SOR > 4.0" -filter-name "MQRankSum_filter" -filter "MQRankSum < -12.5" -filter-name "ReadPosRankSum_filter" -filter "ReadPosRankSum < -8.0"
```

6 Calling of somatic variants

The below sections detail preliminary steps followed for calling somatic variants.

6.1 Tumor samples with matched normal ones

Somatic variants were called using Mutect2 (<https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2>), specific for calling these variants, part of the GATK pipeline.

The purpose of this step is to identify and call somatic variants, which are genetic variations specific to the tumor sample compared to the matched normal sample. This was first carried out via:

Command

Calling somatic variants

```
$ gatk Mutect2 -R ref.fa -I tumor.bam -I normal.bam -normal
normal_sample_name --germline-resource af-only-gnomad.vcf.gz --panel-
of-normals pon.vcf.gz -O somatic.vcf.gz
```

To identify and call somatic variants for more than 2 samples (which are variations specific to the tumor samples compared to the matched normal ones).

Command

Somatic variant calling for multiple samples

```
$ gatk Mutect2 -R reference.fa -I tumor1.bam -I tumor2.bam -I
normal1.bam -I normal2.bam -normal normal1_sample_name -normal
normal2_sample_name --germline-resource af-only-gnomad.vcf.gz --panel-
of-normals pon.vcf.gz -O somatic.vcf.gz
```

6.2 Tumor only mode- without normal samples

To detect and call somatic variants particular to a tumor sample.

Command**Somatic variant calling for only tumor samples**

```
$ gatk Mutect2 -R ref.fa -I sample.bam -O single_sample.vcf.gz
```

Additionally, to call somatic variants while applying filters depending on the population frequencies and a panel of normals, the following codes were run:

Command**Somatic variant calling with filters**

```
$ gatk Mutect2 -R ref.fa -I sample.bam --germline-resource af-only-gnomad.vcf.gz --panel-of-normals pon.vcf.gz -O single_sample.vcf.gz
```

6.3 Filtering of somatic variants

Filtering variants are an essential step before the VCF files are prepared, since it ensures that the obtained variants are more likely to be true somatic mutations. In the present work, the mutations were filtered that will further help in understanding the biology of colorectal cancer.

Command**Filtration of variants**

```
$ gatk FilterMutectCalls -V raw_variants.vcf -O filtered_variants.vcf
```

The variant files obtained from both germline and somatic calling were then used for further analysis as described in the following steps.

VCF FILE PREPARATION, ANNOTATION, POST-PROCESSING

7 VCF file preparation, annotation and post-processing of variants

Once both germline and somatic mutations were identified and called, the VCF files for both these variants were prepared using bcftools for further analysis. Annotation and processing of these variants were then carried out using snpEFF (<https://pcingola.github.io/SnpEff/>), wANNOVAR (<https://wannovar.wglab.org/>) and VEF (Variant Effect Predictor) (<https://asia.ensembl.org/Tools/VEP>).

Command

VCF file prep- step 1

```
$ Vcf-validator SRRID_vcf_file.vcf
```

Command

VCF file prep- step 2

```
$ bcftools view -i 'DP>10' SRRID_vcf_file.vcf >  
SRRID_filtered_vcf_file.vcf
```



Command

VCF file prep- step 3

```
$ bcftools norm -f ref.fasta SRRID_vcf_file.vcf -o  
SRRID_normalized_vcf_file.vcf
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 20  
chr1 876499 . A G 308.06 PASS AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=28.01;SOR=3.442 GT:AD:DP:GQ:PL 1/1:0,11:11:32:322,32,0  
chr1 877831 . T C 375.06 PASS AC=2;AF=1;AN=2;DP=15;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.21;QD=25;SOR=1.112 GT:AD:DP:GQ:PL 1/1:0,15:15:44:389,44,0  
chr1 881627 . G A 570.06 PASS AC=2;AF=1;AN=2;DP=21;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.97;QD=27.15;SOR=1.508 GT:AD:DP:GQ:PL 1/1:0,21:21:63:584,63,0  
chr1 883625 . A G 428.06 PASS AC=2;AF=1;AN=2;DP=17;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=28.54;SOR=3.126 GT:AD:DP:GQ:PL 1/1:0,15:15:44:442,44,0  
chr1 888639 . T C 650.06 PASS AC=2;AF=1;AN=2;DP=23;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.4;QD=30.96;SOR=0.99 GT:AD:DP:GQ:PL 1/1:0,21:21:62:664,62,0  
chr1 888659 . T C 365.06 MQ_Filter AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=39.45;QD=33.19;SOR=2.494 GT:AD:DP:GQ:PL 1/1:0,11:11:33:379,33,0  
chr1 889158 . G C 751.06 PASS AC=2;AF=1;AN=2;DP=19;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.16;QD=28.2;SOR=1.371 GT:AD:DP:GQ:PL 1/1:0,17:17:51:765,51,0  
chr1 889159 . A C 751.06 PASS AC=2;AF=1;AN=2;DP=19;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.16;QD=25;SOR=1.371 GT:AD:DP:GQ:PL 1/1:0,17:17:51:765,51,0  
chr1 894573 . G A 930.06 PASS AC=2;AF=1;AN=2;DP=36;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.5;QD=29.06;SOR=2.765 GT:AD:DP:GQ:PL 1/1:0,32:32:95:944,95,0  
chr1 897325 . G C 2278.06 PASS AC=2;AF=1;AN=2;DP=73;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.74;QD=33.5;SOR=2.105 GT:AD:DP:GQ:PL 1/1:0,68:68:99:2292,204,0  
chr1 897564 . T C 681.06 PASS AC=2;AF=1;AN=2;DP=27;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.86;QD=28.38;SOR=1.27 GT:AD:DP:GQ:PL 1/1:0,24:24:72:695,72,0  
chr1 909238 . G C 220.64 PASS AC=1;AF=0.5;AN=2;BaseQRankSum=1.039;DP=12;ExcessHet=0;FS=0;MLEAC=1;MLEAF=0.5;MQ=41.84;MQRankSum=-0.431;QD=18.39;ReadPosRankSum=-0.165;SOR=1.609 GT:AD:DP:GQ:PL 0/1:4,8:12:89:228,0,89  
chr1 909768 . A G 1079.06 PASS AC=2;AF=1;AN=2;DP=16;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=32.7;SOR=2.833 GT:AD:DP:GQ:PL 1/1:0,33:33:99:1093,99,0  
chr1 914876 . T C 411.06 PASS AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.7;QD=29.56;SOR=2.494 GT:AD:DP:GQ:PL 1/1:0,11:11:33:425,33,0  
chr1 914940 . T C 261.64 PASS AC=1;AF=0.5;AN=2;BaseQRankSum=-0.808;DP=30;ExcessHet=0;FS=3.453;MLEAC=1;MLEAF=0.5;MQ=41.87;MQRankSum=-1.548;QD=9.02;ReadPosRankSum=-0.988;SOR=1.609 GT:AD:DP:GQ:PL 0/1:16,13:29:99:269,0,333  
chr1 915227 . A G 626.06 MQ_Filter AC=2;AF=1;AN=2;DP=22;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=39.79;QD=28.46;SOR=1.609 GT:AD:DP:GQ:PL 1/1:0,22:22:66:1640,66,0  
chr1 916549 . A G 248.64 PASS AC=1;AF=0.5;AN=2;BaseQRankSum=1.053;DP=18;ExcessHet=0;FS=0;MLEAC=1;MLEAF=0.5;MQ=41.89;MQRankSum=-0.717;QD=14.63;ReadPosRankSum=-0.44;SOR=0.33 GT:AD:DP:GQ:PL 0/1:17,10:17:99:256,0,178
```

Snapshot of results obtained from normalizing VCF files using bcftools

7.1 Annotation of variants

Command

Variant annotation

```
$ java -jar snpEff.jar GRChXX.86 SRRID_vcf_file.vcf >  
SRRID_annotated_vcf_file.vcf $ java -jar snpEff.jar -v <snpEff_db>  
SRRID_filtered_snps_final.vcf > SRRID_filtered_snps_final.ann.vcf
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 20  
chr1 876499 . A G 308.06 PASS AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=28.01;SOR=3.442 GT:AD:DP:GQ:PL 1/1:0,11:11:32:322,32,0  
chr1 877831 . T C 375.06 PASS AC=2;AF=1;AN=2;DP=15;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.21;QD=25;SOR=1.112 GT:AD:DP:GQ:PL 1/1:0,15:15:44:389,44,0  
chr1 881627 . G A 570.06 PASS AC=2;AF=1;AN=2;DP=21;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.97;QD=27.15;SOR=1.508 GT:AD:DP:GQ:PL 1/1:0,21:21:63:584,63,0  
chr1 883625 . A G 428.06 PASS AC=2;AF=1;AN=2;DP=17;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=28.54;SOR=3.126 GT:AD:DP:GQ:PL 1/1:0,15:15:44:442,44,0  
chr1 888639 . T C 650.06 PASS AC=2;AF=1;AN=2;DP=23;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.4;QD=30.96;SOR=0.99 GT:AD:DP:GQ:PL 1/1:0,21:21:62:664,62,0  
chr1 888659 . T C 365.06 MQ_Filter AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=39.45;QD=33.19;SOR=2.494 GT:AD:DP:GQ:PL 1/1:0,11:11:33:379,33,0  
chr1 889158 . G C 751.06 PASS AC=2;AF=1;AN=2;DP=19;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.16;QD=28.2;SOR=1.371 GT:AD:DP:GQ:PL 1/1:0,17:17:51:765,51,0  
chr1 889159 . A C 751.06 PASS AC=2;AF=1;AN=2;DP=19;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=40.16;QD=25;SOR=1.371 GT:AD:DP:GQ:PL 1/1:0,17:17:51:765,51,0  
chr1 894573 . G A 930.06 PASS AC=2;AF=1;AN=2;DP=36;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.5;QD=29.06;SOR=2.765 GT:AD:DP:GQ:PL 1/1:0,32:32:95:944,95,0  
chr1 897325 . G C 2278.06 PASS AC=2;AF=1;AN=2;DP=73;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.74;QD=33.5;SOR=2.105 GT:AD:DP:GQ:PL 1/1:0,68:68:99:2292,204,0  
chr1 897564 . T C 681.06 PASS AC=2;AF=1;AN=2;DP=27;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.86;QD=28.38;SOR=1.27 GT:AD:DP:GQ:PL 1/1:0,24:24:72:695,72,0  
chr1 909238 . G C 220.64 PASS AC=1;AF=0.5;AN=2;BaseQRankSum=1.039;DP=12;ExcessHet=0;FS=0;MLEAC=1;MLEAF=0.5;MQ=41.84;MQRankSum=-0.431;QD=18.39;ReadPosRankSum=-0.165;SOR=1.609 GT:AD:DP:GQ:PL 0/1:4,8:12:89:228,0,89  
chr1 909768 . A G 1079.06 PASS AC=2;AF=1;AN=2;DP=16;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=42;QD=32.7;SOR=2.833 GT:AD:DP:GQ:PL 1/1:0,33:33:99:1093,99,0  
chr1 914876 . T C 411.06 PASS AC=2;AF=1;AN=2;DP=13;ExcessHet=0;FS=0;MLEAC=2;MLEAF=1;MQ=41.7;QD=29.56;SOR=2.494 GT:AD:DP:GQ:PL 1/1:0,11:11:33:425,33,0  
chr1 914940 . T C 261.64 PASS AC=1;AF=0.5;AN=2;BaseQRankSum=-0.808;DP=30;ExcessHet=0;FS=3.453;MLEAC=1;MLEAF=0.5;MQ=41.87;MQRankSum=-1.548;QD=9.02;ReadPosRankSum=-0.988;SOR=1.609 GT:AD:DP:GQ:PL 0/1:16,13:29:99:269,0,333
```

Results obtained from snpEFF for variant annotation, using filtered SNPs as the input

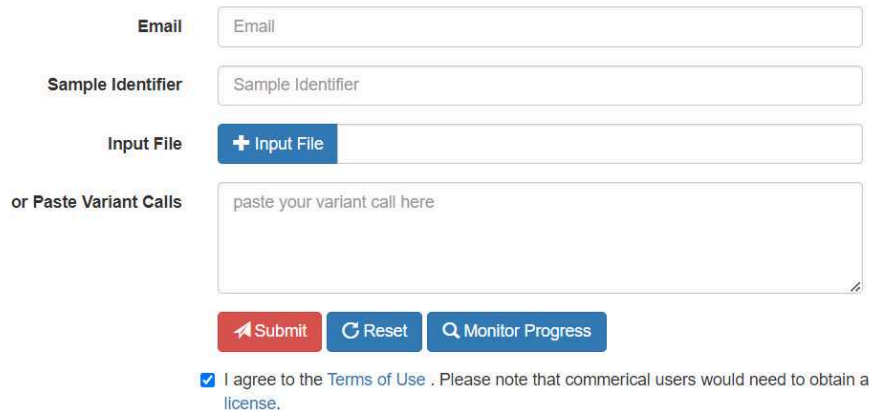
7.2 Post-processing of variants using wANNOVAR and VEP

Variant post processing was performed using two tools: wANNOVAR and VEP

Step 1- wANNOVAR:

This tool is accessible from: <https://wannovar.wglab.org/>.

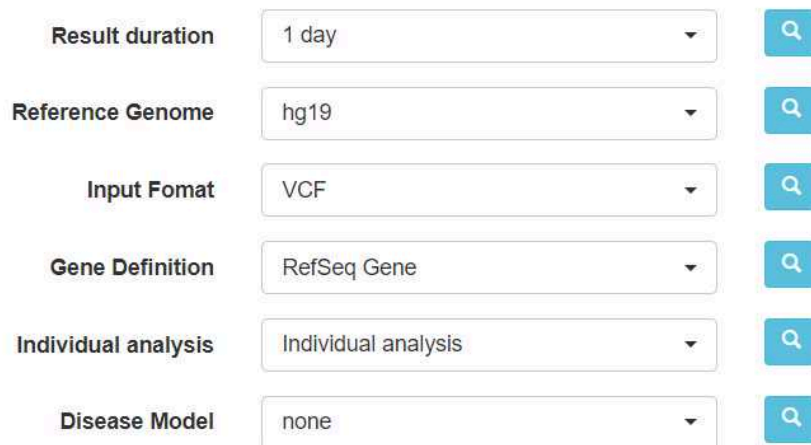
Basic Information



The screenshot shows the 'Basic Information' section of the wANNOVAR web interface. It contains four input fields: 'Email' (with placeholder text 'Email'), 'Sample Identifier' (with placeholder text 'Sample Identifier'), 'Input File' (with a blue '+ Input File' button), and 'or Paste Variant Calls' (with placeholder text 'paste your variant call here'). Below these fields are three buttons: 'Submit' (red), 'Reset' (blue with a circular arrow icon), and 'Monitor Progress' (blue with a magnifying glass icon). At the bottom, there is a checkbox that is checked, followed by the text 'I agree to the Terms of Use . Please note that commerical users would need to obtain a license.'

Input information to be filled in wANNOVAR to obtain the post-processing results using wANNOVAR

Parameter Settings



The screenshot shows the 'Parameter Settings' section of the wANNOVAR web interface. It consists of six rows, each with a label, a dropdown menu, and a search icon (magnifying glass) in a blue box. The settings are: 'Result duration' set to '1 day', 'Reference Genome' set to 'hg19', 'Input Fomat' (note the typo) set to 'VCF', 'Gene Definition' set to 'RefSeq Gene', 'Individual analysis' set to 'Individual analysis', and 'Disease Model' set to 'none'.

The default settings for running wANNOVAR

Some expected results for wANNOVAR are provided below:

chr	Start	End	Ref	Alt	Func.refGene	Gene	Exonic	Func	AAChange	1000G_A	1000G_AF	1000G_AA	1000G_SA	1000G_EU	1000G_SA	ExAc	Freq	ExAc_AFR	ExAc_AM
chr1	877831	877831 T	C	A	exonic	SAMD11	nonsynonym	SAMD11:A		1	1	1	1	1	1	0.9999	1	1	1
chr1	881627	881627 G	C	A	exonic	NOCL2L	synonymy	NOCL2L:N		0.44	0.064	0.44	0.62	0.63	0.57	0.5653	0.1397	0.478	0.484
chr1	888639	888639 C	C	A	exonic	NOCL2L	synonymy	NOCL2L:N		0.92	0.91	0.92	0.92	0.93	0.92	0.9356	0.9096	0.956	0.957
chr1	888659	888659 T	C	C	exonic	NOCL2L	nonsynonym	NOCL2L:N		0.92	0.91	0.92	0.92	0.95	0.92	0.9355	0.9096	0.9557	0.957
chr1	897325	897325 G	C	C	exonic	KLHL17	synonymy	KLHL17:W		0.86	0.7	0.88	0.92	0.94	0.92	0.9077	0.721	0.941	0.941
chr1	909238	909238 G	C	A	exonic	PLEKHN1	nonsynonym	PLEKHN1:I		0.78	0.82	0.69	0.89	0.62	0.83	0.6685	0.789	0.6912	0.6912
chr1	914876	914876 T	C	C	exonic	PERM1	nonsynonym	PERM1:N		0.97	0.94	0.95	1	0.96	0.98	0.9664	0.957	0.9554	0.954
chr1	914940	914940 T	C	C	exonic	PERM1	synonymy	PERM1:N		0.51	0.18	0.61	0.71	0.59	0.68	0.5874	0.2698	0.584	0.584
chr1	915227	915227 A	G	A	exonic	PERM1	synonymy	PERM1:N		0.92	0.76	0.94	1	0.96	0.98	0.9562	0.7839	0.9456	0.9456
chr1	916549	916549 A	G	A	exonic	PERM1	nonsynonym	PERM1:N		0.76	0.6	0.78	0.88	0.76	0.83	0.7836	0.6075	0.8048	0.8048
chr1	935222	935222 C	A	A	exonic	HS4	nonsynonym	HS4:NM		0.49	0.036	0.63	0.77	0.59	0.63	0.6611	0.2017	0.7607	0.7607
chr1	949608	949608 G	A	A	exonic	ISG15	nonsynonym	ISG15:N		0.34	0.44	0.27	0.18	0.39	0.36	0.3702	0.4111	0.2454	0.2454
chr1	949654	949654 A	G	A	exonic	ISG15	synonymy	ISG15:N		0.83	0.51	0.92	0.93	0.95	0.96	0.9122	0.5636	0.9528	0.9528
chr1	981931	981931 A	G	A	exonic	AGRN	synonymy	AGRN:NM		0.8	0.46	0.89	0.96	0.91	0.91	0.8767	0.513	0.9312	0.9312
chr1	982994	982994 A	C	A	exonic	AGRN	synonymy	AGRN:NM		0.84	0.52	0.9	1	0.92	0.97	0.8889	0.5354	0.942	0.942
chr1	984971	984971 G	A	A	exonic	AGRN	nonsynonym	AGRN:NM		0.0044	0.0023	0.0058	0.001	0.014	0.01	0.0127	0.0034	0.0034	0.016

[illegible]

gnomAD_1		gnomAD_2		gnomAD_3		gnomAD_4		gnomAD_5		gnomAD_6		gnomAD_7		gnomAD_8		gnomAD_9		gnomAD_10		gnomAD_11		gnomAD_12		gnomAD_13		gnomAD_14		gnomAD_15		gnomAD_16		gnomAD_17		gnomAD_18		gnomAD_19		gnomAD_20		gnomAD_21		gnomAD_22		gnomAD_23		gnomAD_24		gnomAD_25		gnomAD_26		gnomAD_27		gnomAD_28		gnomAD_29		gnomAD_30		gnomAD_31		gnomAD_32		gnomAD_33		gnomAD_34		gnomAD_35		gnomAD_36		gnomAD_37		gnomAD_38		gnomAD_39		gnomAD_40		gnomAD_41		gnomAD_42		gnomAD_43		gnomAD_44		gnomAD_45		gnomAD_46		gnomAD_47		gnomAD_48		gnomAD_49		gnomAD_50		gnomAD_51		gnomAD_52		gnomAD_53		gnomAD_54		gnomAD_55		gnomAD_56		gnomAD_57		gnomAD_58		gnomAD_59		gnomAD_60		gnomAD_61		gnomAD_62		gnomAD_63		gnomAD_64		gnomAD_65		gnomAD_66		gnomAD_67		gnomAD_68		gnomAD_69		gnomAD_70		gnomAD_71		gnomAD_72		gnomAD_73		gnomAD_74		gnomAD_75		gnomAD_76		gnomAD_77		gnomAD_78		gnomAD_79		gnomAD_80		gnomAD_81		gnomAD_82		gnomAD_83		gnomAD_84		gnomAD_85		gnomAD_86		gnomAD_87		gnomAD_88		gnomAD_89		gnomAD_90		gnomAD_91		gnomAD_92		gnomAD_93		gnomAD_94		gnomAD_95		gnomAD_96		gnomAD_97		gnomAD_98		gnomAD_99		gnomAD_100		gnomAD_101		gnomAD_102		gnomAD_103		gnomAD_104		gnomAD_105		gnomAD_106		gnomAD_107		gnomAD_108		gnomAD_109		gnomAD_110		gnomAD_111		gnomAD_112		gnomAD_113		gnomAD_114		gnomAD_115		gnomAD_116		gnomAD_117		gnomAD_118		gnomAD_119		gnomAD_120		gnomAD_121		gnomAD_122		gnomAD_123		gnomAD_124		gnomAD_125		gnomAD_126		gnomAD_127		gnomAD_128		gnomAD_129		gnomAD_130		gnomAD_131		gnomAD_132		gnomAD_133		gnomAD_134		gnomAD_135		gnomAD_136		gnomAD_137		gnomAD_138		gnomAD_139		gnomAD_140		gnomAD_141		gnomAD_142		gnomAD_143		gnomAD_144		gnomAD_145		gnomAD_146		gnomAD_147		gnomAD_148		gnomAD_149		gnomAD_150		gnomAD_151		gnomAD_152		gnomAD_153		gnomAD_154		gnomAD_155		gnomAD_156		gnomAD_157		gnomAD_158		gnomAD_159		gnomAD_160		gnomAD_161		gnomAD_162		gnomAD_163		gnomAD_164		gnomAD_165		gnomAD_166		gnomAD_167		gnomAD_168		gnomAD_169		gnomAD_170		gnomAD_171		gnomAD_172		gnomAD_173		gnomAD_174		gnomAD_175		gnomAD_176		gnomAD_177		gnomAD_178		gnomAD_179		gnomAD_180		gnomAD_181		gnomAD_182		gnomAD_183		gnomAD_184		gnomAD_185		gnomAD_186		gnomAD_187		gnomAD_188		gnomAD_189		gnomAD_190		gnomAD_191		gnomAD_192		gnomAD_193		gnomAD_194		gnomAD_195		gnomAD_196		gnomAD_197		gnomAD_198		gnomAD_199		gnomAD_200		gnomAD_201		gnomAD_202		gnomAD_203		gnomAD_204		gnomAD_205		gnomAD_206		gnomAD_207		gnomAD_208		gnomAD_209		gnomAD_210		gnomAD_211		gnomAD_212		gnomAD_213		gnomAD_214		gnomAD_215		gnomAD_216		gnomAD_217		gnomAD_218		gnomAD_219		gnomAD_220		gnomAD_221		gnomAD_222		gnomAD_223		gnomAD_224		gnomAD_225		gnomAD_226		gnomAD_227		gnomAD_228		gnomAD_229		gnomAD_230		gnomAD_231		gnomAD_232		gnomAD_233		gnomAD_234		gnomAD_235		gnomAD_236		gnomAD_237		gnomAD_238		gnomAD_239		gnomAD_240		gnomAD_241		gnomAD_242		gnomAD_243		gnomAD_244		gnomAD_245		gnomAD_246		gnomAD_247		gnomAD_248		gnomAD_249		gnomAD_250		gnomAD_251		gnomAD_252		gnomAD_253		gnomAD_254		gnomAD_255		gnomAD_256		gnomAD_257		gnomAD_258		gnomAD_259		gnomAD_260		gnomAD_261		gnomAD_262		gnomAD_263		gnomAD_264		gnomAD_265		gnomAD_266		gnomAD_267		gnomAD_268		gnomAD_269		gnomAD_270		gnomAD_271		gnomAD_272		gnomAD_273		gnomAD_274		gnomAD_275		gnomAD_276		gnomAD_277		gnomAD_278		gnomAD_279		gnomAD_280		gnomAD_281		gnomAD_282		gnomAD_283		gnomAD_284		gnomAD_285		gnomAD_286		gnomAD_287		gnomAD_288		gnomAD_289		gnomAD_290		gnomAD_291		gnomAD_292		gnomAD_293		gnomAD_294		gnomAD_295		gnomAD_296		gnomAD_297		gnomAD_298		gnomAD_299		gnomAD_300		gnomAD_301		gnomAD_302		gnomAD_303		gnomAD_304		gnomAD_305		gnomAD_306		gnomAD_307		gnomAD_308		gnomAD_309		gnomAD_310		gnomAD_311		gnomAD_312		gnomAD_313		gnomAD_314		gnomAD_315		gnomAD_316		gnomAD_317		gnomAD_318		gnomAD_319		gnomAD_320		gnomAD_321		gnomAD_322		gnomAD_323		gnomAD_324		gnomAD_325		gnomAD_326		gnomAD_327		gnomAD_328		gnomAD_329		gnomAD_330		gnomAD_331		gnomAD_332		gnomAD_333		gnomAD_334		gnomAD_335		gnomAD_336		gnomAD_337		gnomAD_338		gnomAD_339		gnomAD_340		gnomAD_341		gnomAD_342		gnomAD_343		gnomAD_344		gnomAD_345		gnomAD_346		gnomAD_347		gnomAD_348		gnomAD_349		gnomAD_350		gnomAD_351		gnomAD_352		gnomAD_353		gnomAD_354		gnomAD_355		gnomAD_356		gnomAD_357		gnomAD_358		gnomAD_359		gnomAD_360		gnomAD_361		gnomAD_362		gnomAD_363		gnomAD_364		gnomAD_365		gnomAD_366		gnomAD_367		gnomAD_368		gnomAD_369		gnomAD_370		gnomAD_371		gnomAD_372		gnomAD_373		gnomAD_374		gnomAD_375		gnomAD_376		gnomAD_377		gnomAD_378		gnomAD_379		gnomAD_380		gnomAD_381		gnomAD_382		gnomAD_383		gnomAD_384		gnomAD_385		gnomAD_386		gnomAD_387		gnomAD_388		gnomAD_389		gnomAD_390		gnomAD_391		gnomAD_392		gnomAD_393		gnomAD_394		gnomAD_395		gnomAD_396		gnomAD_397		gnomAD_398		gnomAD_399		gnomAD_400		gnomAD_401		gnomAD_402		gnomAD_403		gnomAD_404		gnomAD_405		gnomAD_406		gnomAD_407		gnomAD_408		gnomAD_409		gnomAD_410		gnomAD_411		gnomAD_412		gnomAD_413		gnomAD_414		gnomAD_415		gnomAD_416		gnomAD_417		gnomAD_418		gnomAD_419		gnomAD_420		gnomAD_421		gnomAD_422		gnomAD_423		gnomAD_424		gnomAD_425		gnomAD_426		gnomAD_427		gnomAD_428		gnomAD_429		gnomAD_430		gnomAD_431		gnomAD_432		gnomAD_433		gnomAD_434		gnomAD_435		gnomAD_436		gnomAD_437		gnomAD_438		gnomAD_439		gnomAD_440		gnomAD_441		gnomAD_442		gnomAD_443		gnomAD_444		gnomAD_445		gnomAD_446		gnomAD_447		gnomAD_448		gnomAD_449		gnomAD_450		gnomAD_451		gnomAD_452		gnomAD_453		gnomAD_454		gnomAD_455		gnomAD_456		gnomAD_457		gnomAD_458		gnomAD_459		gnomAD_460		gnomAD_461		gnomAD_462		gnomAD_463		gnomAD_464		gnomAD_465		gnomAD_466		gnomAD_467		gnomAD_468		gnomAD_469		gnomAD_470		gnomAD_471		gnomAD_472		gnomAD_473		gnomAD_474		gnomAD_475		gnomAD_476		gnomAD_477		gnomAD_478		gnomAD_479		gnomAD_480		gnomAD_481		gnomAD_482		gnomAD_483		gnomAD_484		gnomAD_485		gnomAD_486		gnomAD_487		gnomAD_488		gnomAD_489		gnomAD_490		gnomAD_491		gnomAD_492		gnomAD_493		gnomAD_494		gnomAD_495		gnomAD_496		gnomAD_497		gnomAD_498		gnomAD_499		gnomAD_500		gnomAD_501		gnomAD_502		gnomAD_503		gnomAD_504		gnomAD_505		gnomAD_506		gnomAD_507		gnomAD_508		gnomAD_509		gnomAD_510		gnomAD_511		gnomAD_512		gnomAD_513		gnomAD_514		gnomAD_515		gnomAD_516		gnomAD_517		gnomAD_518		gnomAD_519		gnomAD_520		gnomAD_521		gnomAD_522		gnomAD_523		gnomAD_524		gnomAD_525		gnomAD_526		gnomAD_527		gnomAD_528		gnomAD_529		gnomAD_530		gnomAD_531		gnomAD_532		gnomAD_533		gnomAD_534		gnomAD_535		gnomAD_536		gnomAD_537		gnomAD_538		gnomAD_539		gnomAD_540		gnomAD_541		gnomAD_542		gnomAD_543		gnomAD_544		gnomAD_545		gnomAD_546		gnomAD_547		gnomAD_548		gnomAD_549		gnomAD_550		gnomAD_551		gnomAD_552		gnomAD_553		gnomAD_554		gnomAD_555		gnomAD_556		gnomAD_557		gnomAD_558		gnomAD_559		gnomAD_560		gnomAD_561		gnomAD_562		gnomAD_563		gnomAD_564		gnomAD_565		gnomAD_566		gnomAD_567		gnomAD_568		gnomAD_569		gnomAD_570		gnomAD_571		gnomAD_572		gnomAD_573		gnomAD_574		gnomAD_575		gnomAD_576		gnomAD_577		gnomAD_578		gnomAD_579		gnomAD_580		gnomAD_581		gnomAD_582		gnomAD_583		gnomAD_584		gnomAD_585		gnomAD_586		gnomAD_587		gnomAD_588		gnomAD_589		gnomAD_590		gnomAD_591		gnomAD_592		gnomAD_593		gnomAD_594		gnomAD_595		gnomAD_596		gnomAD_597		gnomAD_598		gnomAD_599		gnomAD_600		gnomAD_601		gnomAD_602		gnomAD_603		gnomAD_604		gnomAD_605		gnomAD_606		gnomAD_607		gnomAD_608		gnomAD_609		gnomAD_610		gnomAD_611		gnomAD_612		gnomAD_613		gnomAD_614		gnomAD_615		gnomAD_616		gnomAD_617		gnomAD_618		gnomAD_619		gnomAD_620		gnomAD_621		gnomAD_622		gnomAD_623		gnomAD_624		gnomAD_625		gnomAD_626		gnomAD_627		gnomAD_628		gnomAD_629		gnomAD_630		gnomAD_631		gnomAD_632		gnomAD_633		gnomAD_634		gnomAD_635		gnomAD_636		gnomAD_637		gnomAD_638		gnomAD_639		gnomAD_640		gnomAD_641		gnomAD_642		gnomAD_643		gnomAD_644		gnomAD_645		gnomAD_646		gnomAD_647		gnomAD_648		gnomAD_649		gnomAD_650		gnomAD_651		gnomAD_652		gnomAD_653		gnomAD_654		gnomAD_655		gnomAD_656		gnomAD_657		gnomAD_658		gnomAD_659		gnomAD_660		gnomAD_661		gnomAD_662		gnomAD_663		gnomAD_664		gnomAD_665		gnomAD_666		gnomAD_667		gnomAD_668		gnomAD_669		gnomAD_670		gnomAD_671		gnomAD_672		gnomAD_673		gnomAD_674		gnomAD_675		gnomAD_676		gnomAD_677		gnomAD_678		gnomAD_679		gnomAD_680		gnomAD_681		gnomAD_682		gnomAD_683		gnomAD_684		gnomAD_685		gnomAD_686		gnomAD_687		gnomAD_688		gnomAD_689		gnomAD_690		gnomAD_691		gnomAD_692		gnomAD_693		gnomAD_694		gnomAD_695		gnomAD_696		gnomAD_697		gnomAD_698		gnomAD_699		gnomAD_700		gnomAD_701		gnomAD_702		gnomAD_703		gnomAD_704		gnomAD_705		gnomAD_706		gnomAD_707		gnomAD_708		gnomAD_709		gnomAD_710		gnomAD_711		gnomAD_712		gnomAD_713		gnomAD_714		gnomAD_715		gnomAD_716		gnomAD_717		gnomAD_718		gnomAD_719		gnomAD_720		gnomAD_721		gnomAD_722		gnomAD_723		gnomAD_724		gnomAD_725		gnomAD_726		gnomAD_727		gnomAD_728		gnomAD_729		gnomAD_730		gnomAD_731		gnomAD_732		gnomAD_733		gnomAD_734		gnomAD_735		gnomAD_736		gnomAD_737		gnomAD_738		gnomAD_739		gnomAD_740		gnomAD_741		gnomAD_742		gnomAD_743		gnomAD_744		gnomAD_745		gnomAD_746		gnomAD_747		gnomAD_748		gnomAD_749		gnomAD_750		gnomAD_751		gnomAD_752		gnomAD_753		gnomAD_754		gnomAD_755		gnomAD_756		gnomAD_757		gnomAD_758		gnomAD_759		gnomAD_760		gnomAD_761		gnomAD_762		gnomAD_763		gnomAD_764		gnomAD_765		gnomAD_766		gnomAD_767		gnomAD_768		gnomAD_769		gnomAD_770		gnomAD_771		gnomAD_772		gnomAD_773		gnomAD_774		gnomAD_775		gnomAD_776		gnomAD_777		gnomAD_778		gnomAD_779		gnomAD_780		gnomAD_781		gnomAD_782		gnomAD_783		gnomAD_784		gnomAD_785		gnomAD_786		gnomAD_787		gnomAD_788		gnomAD_789		gnomAD_790		gnomAD_791		gnomAD_792		gnomAD_793		gnomAD_794		gnomAD_795		gnomAD_796		gnomAD_797		gnomAD_798		gnomAD_799		gnomAD_800		gnomAD_801		gnomAD_802		gnomAD_803		gnomAD_804		gnomAD_805		gnomAD_806		gnomAD_807		gnomAD_808		gnomAD_809		gnomAD_810		gnomAD_811		gnomAD_812		gnomAD_813		gnomAD_814		gnomAD_815		gnomAD_816		gnomAD_817		gnomAD_818		gnomAD_819		gnomAD_820		gnomAD_821		gnomAD_822		gnomAD_823		gnomAD_824		gnomAD_825		gnomAD_826		gnomAD_827		gnomAD_828		gnomAD_829		gnomAD_830		gnomAD_831		gnomAD_832		gnomAD	
----------	--	----------	--	----------	--	----------	--	----------	--	----------	--	----------	--	----------	--	----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	-----------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	------------	--	--------	--

Snapshots of expected wANNOVAR results

Step 2- Variant Effect Predictor (VEP)

The VEP tool is accessible from: <https://asia.ensembl.org/Tools/VEP>

Variant Effect Predictor

New job

Species:

 Homo_sapiens

Assembly: GRCh38.p13

[Change species](#)

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Name for this job (optional):

Input data:

Either paste data:

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#), [SPDI](#)

Or upload file:

Choose file

No file chosen

Or provide file URL:

Input data settings for VEP

Transcript database to use:

☒ Ensembl/GENCODE transcripts

☐ Ensembl/GENCODE basic transcripts

☐ RefSeq transcripts

☐ Ensembl/GENCODE and RefSeq transcripts

Additional configurations:

Identifiers

Additional identifiers for genes, transcripts and variants

Variants and frequency data

Co-located variants and frequency data

Additional annotations

Additional transcript, protein and regulatory annotations

Predictions

Variant predictions, e.g. SIFT, PolyPhen

Filtering options

Pre-filter results by frequency or consequence type

Advanced options

Additional enhancements

Run

Additional configurations and the type of transcript database to be used for VEP analysis

#Uploads	Location	Allele	Consequence	IMPACT	SYMBOL	Gene	Feature_h	Feature	BIOTYPE	EXON	INTRON	HGVSc	HGVSp	cDNA_pos	CDS_pos	Protein_pos	Amino_acid	Codons	Existing_v	DISTANCE	STRAND	FLAGS	SY
-	1:876499	-G	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	protein_cc-	-	-	-	-	-	-	-	-	-	rs4372192	3085	-1	-	HC
-	1:876499	-G	intron_var	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	-	05-Nov	-	-	-	-	-	-	-	rs4372192	-	-	1 cds_start	HC
-	1:876499	-G	intron_var	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	-	Jul-13	-	-	-	-	-	-	-	rs4372192	-	-	1	HC
-	1:876499	-G	downstream	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	-	-	-	-	-	-	-	-	-	rs4372192	1828	-	1 cds_end_F	HC
-	1:876499	-G	intron_var	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	-	01-Jun	-	-	-	-	-	-	-	rs4372192	-	-	1 cds_start	HC
-	1:876499	-G	upstream	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs4372192	1047	-	-	HC
-	1:876499	-G	upstream	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs4372192	984	-	1	HC
-	1:876499	-G	non_coding	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	03-Apr	-	-	-	-	44	-	-	-	rs4372192	-	-	1	HC
-	1:876499	-G	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs4372192	3086	-1	-	HC
-	1:876499	-G	intron_var	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	processed-	-	01-Feb	-	-	-	-	-	-	-	rs4372192	-	-	1	HC
-	1:876499	-G	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs4372192	3085	-1	-	HC
-	1:876499	-G	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	processed-	-	-	-	-	-	-	-	-	-	rs4372192	4200	-1	-	HC
-	1:876499	-G	regulatory	MODIFIER	-	-	Regulatory	ENSR000001	promoter	-	-	-	-	-	-	-	-	-	rs4372192	-	-	-	-
-	1:877831	-C	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	protein_cc-	-	-	-	-	-	-	-	-	-	rs6672356	1753	-1	-	HC
-	1:877831	-C	missense	MODERAT	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	08-Dec	-	-	-	750	751	251	W/R	Tgg/Cgg	rs6672356	-	-	1 cds_start	HC
-	1:877831	-C	missense	MODERAT	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	Oct-14	-	-	-	1110	1027	343	W/R	Tgg/Cgg	rs6672356	-	-	1	HC
-	1:877831	-C	downstream	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	-	-	-	-	-	-	-	-	-	rs6672356	3160	-	1 cds_end_F	HC
-	1:877831	-C	missense	MODERAT	SAMD11	ENSG000001	Transcript	ENST000001	protein_cc-	04-Jul	-	-	-	507	508	170	W/R	Tgg/Cgg	rs6672356	-	-	1 cds_start	HC
-	1:877831	-C	non_coding	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	01-Feb	-	-	-	286	-	-	-	-	rs6672356	-	-	1	HC
-	1:877831	-C	non_coding	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	02-Feb	-	-	-	191	-	-	-	-	rs6672356	-	-	1	HC
-	1:877831	-C	non_coding	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	retained_i-	03-Apr	-	-	-	389	-	-	-	-	rs6672356	-	-	1	HC
-	1:877831	-C	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs6672356	1754	-1	-	HC
-	1:877831	-C	downstream	MODIFIER	SAMD11	ENSG000001	Transcript	ENST000001	processed-	-	-	-	-	-	-	-	-	-	rs6672356	278	-1	-	HC
-	1:877831	-C	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	retained_i-	-	-	-	-	-	-	-	-	-	rs6672356	1753	-1	-	HC
-	1:877831	-C	downstream	MODIFIER	NOC2L	ENSG000001	Transcript	ENST000001	processed-	-	-	-	-	-	-	-	-	-	rs6672356	2868	-1	-	HC

Expected result for VEP

protocols.io | <https://dx.doi.org/10.17504/protocols.io.x54v9dqxgg3e/v1>

May 24, 2023

25/27

CONCLUSIONS AND FUTURE PERSPECTIVES

- 8 Colorectal cancers have several diagnostic and treatment options to combat it, however, a delay in disease detection is life-threatening. Therefore, this protocol emphasizes on the identification and analysis of somatic and germline variants for colorectal cancer exome datasets, retrieved from publicly available databases such as NCBI-SRA. These steps describe the various tools required to run this pipeline, with the codes required to run each tool. Moreover, although this protocol is described for colorectal cancer exomes, the same set of codes and steps may be followed for the overall analysis of any cancer exome, when the desired outcome is to obtain and analyze somatic and germline variants separately. This computational pipeline requires further in-vitro and in-vivo studies, however, the outcomes will offer prospects for similar such studies that are crucial for designing a cure, prognosis or a treatment strategy for colorectal cancers, based on the scrutinized variants.

The outcomes from this protocol can be carried forward to building and designing decision support systems using artificial intelligence and machine learning (AI/ML), so that the identified somatic and germline mutants can be used for early colorectal cancer prediction through a prediction model. This will aid clinicians/researchers/diagnosticians to help take better decisions for treatment strategies.

Protocol references

1. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*. 2010 Nov 8;39(suppl_1):D19-21.
2. Andrews S. FastQC: a quality control tool for high throughput sequence data.
3. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016 Oct 1;32(19):3047-8.
4. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*. 2011 May 2;17(1):10-2.
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012 Apr;9(4):357-9.
6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *bioinformatics*. 2009 Aug 15;25(16):2078-9.
7. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in genetics*. 2012 Mar 15;3:35.
8. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols*. 2015 Oct;10(10):1556-66.
9. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. *Genome biology*. 2016 Dec;17(1):1-4.
10. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 2010 Sep 1;20(9):1297-303.