**ARVIN Manual**

Annotation of Regulatory Variants using Integrated Networks (ARVIN) is a general computational framework for predicting causal noncoding variants by combining sequence-based and regulatory network-based features. This manual explains how to run ARVIN framework.

ARVIN is composed of two software modules. First module processes sequence-based features and maps SNPs to genes. Second module prioritizes risk SNP based on network and sequenced based features of the candidate SNPs.

The inputs of ARVIN are list of the candidate SNPs to be analyzed, enhancer regions, FunSeq and GWAVA features and differential gene expression results comparing healthy control and patient groups for related diseases. The format of the input files and how they can be generated are explained below.

The output of ARVIN is the disease risk SNPs that are inferred based on network and sequence based features.

**A) Inputs:**

1. **Candiate SNPs:**

This is the list of all candidate SNPs to be analyzed in bed format, containg the coordinates, reference and alternate alleles and snp identifiers as follows:

#chromosome start end refrence alternate snp\_id

chr6 138230039 138230040 T A rs200820567

chr1 160807714 160807715 T C rs3766379

chr1 160809002 160809003 G A rs6682654

2. **Enhancer coordinates:**

This is the list of predicted enhancers for the cell(s) or tissue(s) being studied. It should be contain the coordinate of the enhancer center followed by the enhancer prediction probability, as follows:

#chromosome enhancer\_center enhancer\_probability

chr6 138231000 0.86

chr1 160807000 0.79

chr1 160808600 0.88

You can use CSI-ANN software for predicting the enhancers. You can download CSI-ANN from our lab web page <http://tanlab4generegulation.org/CSIANNWebpage.html> . CSI-ANN uses histone modification data as input to predict enhancer regions in genome-wide. You can use the output of CSI-ANN as input for ARVIN.

3. **Enhancer-promoter interaction data:**

This is the list of enhancer-promoter interactions. It is used to identify the genes that may be affected by the candidate SNPs. This file should contain enhancer coordinates followed by the target transcript and enhancer-promoter interaction score in tab separated format, as follows:

#Chr Start End Target Score

chr9 22124001 22126001 ENST00000452276 0.93

chr2 242792001 242794001 ENST00000485966 0.792

chr1 109816001 109818001 ENST00000534661 0.607

chr6 138230001 138232001 ENST00000509752 0.825

**4. GWAVA features**

ARVIN uses sequence features for the input SNPs generated by GWAVA. GWAVA is an open-source software developed by Sanger Institute. You can either upload the SNPs to GWAVA web page and get the output or download the source and run locally.

For running GWAVA online navigate to <https://www.sanger.ac.uk/sanger/StatGen_Gwava> , upload the list of input SNPs and get the features in csv format, which will be input for ARVIN.

If you prefer to run it locally, you need to dowload the source code from <ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/src/> and annotation data from <ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/source_data/> . Then you can run it local by running gwava\_annotate.py and generate the features, which will be input for ARVIN.

**5. FunSeq features:**

ARVIN also uses sequence features generated by FunSeq. FunSeq can also be run online or binaries can be downloaded to run locally.

For running FunSeq online, navigate to <http://funseq.gersteinlab.org/analysis> and upload the list of SNPs that you want to analyze. In the web page, it is noted that the input SNPs can be uploaded in bed format, SNP coordinates followed by reference and alternate alleles; but we discovered that it fails to process bed input. In order to have it run, you the first two seperators need to be two spaces and last two separators need to be tabs, as follows:

chr16··4526757··4526758 G A

chr14··52733136··52733137 C A

where each dot (·) represents a space.

Then, FunSeq will generate the features by selecting “bed” as the output format., which will be used as input by ARVIN.

If you prefer to run FunSeq locally, you can download FunSeq binaries from <http://funseq.gersteinlab.org/static/funseq-0.1.tar.gz> and extract it into your local. You will also need to download FunSeq annotation data from <http://funseq.gersteinlab.org/static/data/data.tar.gz> , extract it into directory that you saved the binaries. Then you can run FunSeq binary file by setting the output format to bed.

**B) Running ARVIN**

You can download ARVIN from Tan Lab web site <http://tanlab4generegulation.org/software/> . The package contains the scripts and annotation data necessary to run ARVIN to prioritize risk SNPs. To run the first module of ARVIN to process the features for the SNPs, navigate into the ARVIN directory where the executable shell script process\_features.sh is stored and execute it as follows:

./process\_features.sh \

input\_snps\_file.bed \

csi\_ann\_output\_file.txt \

im\_pet\_output\_file.bed \

gwava\_features\_file.csv \

funseq\_feautes\_file.bed \

output\_directory

This script will generate three output files, to be used as input in the second step.

**1. disruption\_p.txt** : This file contains strongest transcription factor binding disruption caused by the SNPs being analyzed in tab separated format.

snp ref alt TF disruption\_p disruption\_q log\_disruption\_q

rs4784227 C T CUX1 0.0518 0.0518 1.28567024025477

rs11568821 C T RUNX3 0.0215 0.0215 1.66756154008439

**2. snp\_target\_gene.txt**: This file lists the SNP-gene interactions in tab separated format.

snp\_id gene\_symbol entrez\_gene\_id interaction\_score

rs200820567 ADAM7 8756 0.00450758418

rs339331 MIR624 693209 0.058555728

rs1542725 C1RL 51279 0.258880096

**3. features\_gwava\_funseq.csv**: This file contains the GWAVA and FunSeq features for the input SNPs in comma separated format.

snp\_id,chr,end,start,ATF3,BATF,BCL11A,BCL3,BCLAF1,...

rs10757278,chr9,22124477,22124476,0.0,0.0,0.0,0.0,...

rs10811656,chr9,22124472,22124471,0.0,0.0,0.0,0.0,...