**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 disease. The format of the input files and output files are explained below.

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

**A) Installing ARVIN**

To install ARVIN, please follow these steps:

1. Go to ARVIN github ( <https://github.com/gaolong/arvin/> ) and download the entire (master) directory. From here on, we call this directory */home/programs/*

2. Extract the package into your file system and name it “arvin”, so you will get /*home/programs/arvin/ .* If you want to use a different path, please go to line 3 in  *scripts/arvin\_wrapper.sh* and point “*arvin\_software\_dir*” to the ARVIN directory.

3. Open R. Install R [igraph](http://igraph.org/r/) package: install.packages(‘igraph’, dependencies = TRUE)

4. Install ARVIN R package via this command:

> install.packages(‘/home/programs/arvin/source/ARVIN\_0.1.tar.gz’, dependencies = TRUE)

5. Download ARVIN annotation data from [our repository](https://chopri.box.com/s/8pmpoeq8kucm9ioo37cdbyc8vkejawtl): and extract it into /*home/programs/arvin/* so that you will have */home/programs/arvin/arvin\_annotation\_data/*

6. Add the path */home/programs/arvin/perl\_libraries/*  to your perl path, as follows:

export PERL5LIB=$PERL5LIB:/home/programs/arvin/perl\_libraries/

**B) Preparing ARVIN Input Files:**

**1**. **Input SNPs file:**

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:

#chr start end ref\_allele alt\_allele snp\_id

chr6 138230039 138230040 T A rs200820567

chr1 160807714 160807715 T C rs3766379

chr1 160809002 160809003 G A rs6682654

**2**. **Enhancer-promoter interaction file:**

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

In this file, “Target” column may be either target transcript ID (as shown above) or target gene ID or gene symbol.

Alternatively, if you are working on one of the diseases that is included in the [ARVIN paper](https://www.nature.com/articles/s41467-018-03133-y), you can use the pre-computed differential expression results from the study. These pre-computed enhancer-promoter interactions are in here: */home/programs/arvin/arvin\_annotation\_data/built\_in\_ep\_data/ .* The enhancer promoter interactions are matched between diseases and related tissues. You can find the detailed information in Supplementary Materials of the paper.

**3. GWAVA features file:**

ARVIN uses sequence features for the input SNPs generated by GWAVA. GWAVA is an open-source software developed by Sanger Institute. For running GWAVA online navigate to <https://www.sanger.ac.uk/sanger/StatGen_Gwava> , upload the list of input SNPs and download the features in tab separated (TSV) format (in CSV format, coordinates include thousand separator commas causing problem, that’s why avoid CSV format), which will be input for ARVIN.

**4. FunSeq features file:**

ARVIN also uses sequence features generated by FunSeq. FunSeq can 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 Funseq 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, the first two separators need to be double 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.

**5. Differential expression p-values file:**

ARVIN employs differential expression data to weight gene nodes in the network.You need to provide the differential expression p-value for the input genes in tab separated format, gene symbol followed by differential expression p-value (adjusted) as follows:

UNC119 0.001247

NHSL2 0.001475

ANGEL2 0.00165

Alternatively, if you are working on one of the diseases that is included in the [ARVIN paper](https://www.nature.com/articles/s41467-018-03133-y), you can use the pre-computed differential expression results from the study. In this case, while running the wrapper script, instead of the differential expression file name, you need to specify three letter notation for the disease as follows:

|  |  |
| --- | --- |
| **Disease** | **Notation** |
| Alzheimer’s Disease | AD |
| Autism Spectrum Disorder | ASD |
| Asthma | AST |
| Bladder Cancer | BLC |
| Breast Cancer | BRC |
| Coronary Artery Disease | CAD |
| Cystic Fibrosis | CF |
| Chronic Obstructive Pulmonary Disease | COPD |
| Coloreectal Cancer | CRC |
| Crohn’s Disease | CRH |
| Hypercholesterolemia | HC |
| Myocardial Infarction | MI |
| Melanoma | MLN |
| Multiple Sclerosis | MS |
| Neuroblastoma | NBL |
| Obesity | OBE |
| Parkinson’s Disease | PD |
| Prostate Cancer | PRC |
| Psoriasis | PSO |
| Rheumatoid Arthritis | RA |
| Systemic Lupus Erythematosus | SLE |
| Schizophrenia | SZA |
| Type 1 Diabetes | T1D |
| Type 2 Diabetes | T2D |
| Thalassemia Beta | TSB |
| Ulcerative Collitis | ULC |

**C) Running ARVIN with custom input**

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 arvin\_wrapper.sh is stored and execute it as follows:

./arvin\_wrapper.sh \

input\_snps\_file.bed \

ep\_interaction\_file.bed \

gwava\_features\_file.tsv \

funseq\_feautes\_file.bed \

de\_adj\_p.txt \

intermediate\_directory \

output\_directory

The input files should be formatted as in previous section. In addition, you need to specify the two directory paths: intermediate\_directory (for intermediate results) and output\_directory (for final output files).

If you want to use pre-computed differential expression results you can specify the disease (e.g. SLE) with its notation listed above, instead of the differential expression file name:

./arvin\_wrapper.sh \

input\_snps\_file.bed \

ep\_interaction\_file.bed \

gwava\_features\_file.tsv \

funseq\_feautes\_file.bed \

SLE \

intermediate\_directory \

output\_directory

The wrapper script will generate three output files in the output directory:

**1. arvin\_scores.txt:** These are the risk prediction scores computed by ARVIN. Higher score for a SNP means higher association with disease risk:

rs77508451 0.524

rs229554 0.421

rs71459332 0.362

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

chrom start end ref alt snp\_id gene\_symbol entrez\_id int\_score

chr22 37595485 37595486 G A rs1001810 CYTH4 27128 0.733

chr12 56428243 56428244 G A rs77508451 ANKRD52 283373 0.731

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

snp\_id ref alt TF motif log\_disruption\_q position

rs3788013 C A RFX5 M4575\_1.02 1.01412464269161 1

rs1001810 G A ZNF263 M4604\_1.02 0.581826236058033 8

**D) Running ARVIN for example input data**

We have prepared example input to run ARVIN. It may help you to understand the workflow. To run the example input, navigate to */home/programs/arvin/example\_2/* and run it as *./run\_example\_2* . If ARVIN is installed correctly, it should make two directories named “*intermediate*” and “*output*” and you should be able to see the SNP scores in the “*output*” directory. This example computes ARVIN scores for 300 SNPs using built in enhancer-promoter interaction set for autoimmune diseases and built in differential expression analysis results for type 1 diabetes (T1D).