

PriVar documentation

PriVar is a cross-platform Java application toolkit to prioritize variants (SNVs and InDels) from exome or whole genome sequencing data by using different filtering strategies and information of external databases.

PriVar contains four modules : Annotation, Quality control, Candidate Gene Identification and Prediction of functional impact of variants.

● **Annotation:**

Step1: Download the resource.zip from website and extract it to a local directory.

Step2: Annotate variants calls in VCF format.

```
java -jar PriVar.jar  
    -resourcedir /path/resource  
    -inputvcf    /path/data.vcf  
    -outputdir  /path/  
    -module Anno
```

In the /path directory you will find two output files: "variant_annotation_exome.csv", "variant_annotation_genome.csv" which stand for exonic and splicing variants and all the genomic variants, respectively. This two files can be used to manually inspect and prioritize the variants. The prioritization score will be also provided for each variant. The vcf file can include more than one sample, PriVar will provided the information of genotype and quality separately.

● **Quality control**

1. Summary

```
java -jar PriVar.jar  
-inputanno /path/ variant_annotation_genome.csv  
[-varqual 10]  
[-mapqual 10]  
[-depth 10]  
-outputdir /path/  
-module QC
```

The input file of quality control is the annotated file "variant_annotation_genome.csv", the output file is "variant_QC.txt" in the given output directory. The known variants are defined by dbSNP.

2. SNP chip concordance

```
java -jar PriVar.jar  
-inputanno /path/ variant_annotation_genome.csv  
-outputdir /path/  
[-varqual 10]  
[-mapqual 10]  
[-depth 10]  
-module Chipconcor  
-chip test.ped  
-pileup test.pileup
```

The input file of SNP chip concordance is the annotated file "variant_annotation_genome.csv" and the output file is "variant_chip.txt" in the given output directory.

3. **Mendelian error**

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        -module MenErr
        -triororder 2:0:1
```

The input file of quality control is the annotated file "variant_annotation_genome.csv", the output file is "variant_Mendel.txt" in the given output directory. "-triororder" is used to assign the order of father, mother and child in the input file, respectively. In the example, the father is the third sample, mother is the first and child is second in the input file.

● **Candidate Gene Identification**

1. **Linkage-based strategy**

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-population CEU/ASI/YRI]
        [-varqual 10]
        [-mapqual 10]
```

```
[-depth 10]
[-allelefreq 0.01]
[-deletercut 0.5]
-module Pri
-pricand IBD
-resourcedir /path/resource
```

Because PriVar used Markov model to calculate haplotype allele frequency, so control data is needed, the default is the Caucasian from hapmap phaseII (CEU), the user can also select Asian (ASI) or YRI (African). The allele frequency is defined in all the subjects of 1KG project. The input annotated file should contain two individuals.

2. Run of homozygosity (ROH)-based strategy

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-population CEU/ASI/YRI]
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-allelefreq 0.01]
        [-deletercut 0.5]
        -module Pri
        -pricand Homo
        -resourcedir /path/resource
```

This module can only process one individual each time in this version.

3. **"Double-hit"-based strategy**

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-allelefreq 0.01]
        [-deletercut 0.5]
        -module Pri
        -pricand Double
```

This module can only process one individual each time in this version.

4. **Mutation burden**

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-allelefreq 0.01]
        [-deletercut 0.5]
        -module Pri
        -pricand Burden
```

5. **De novo mutations**

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-deletercut 0.5]
        -module Pri
        -pricand Denovo
        -triorder 0:1:2
```

6. Literature-based strategy

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-deletercut 0.5]
        -module Pri
        -pricand Candidate
        -disease C0024138
```

Or users can specify their customized gene list as the format below one gene name per line.

For example:

Customized gene list named "candidate.txt":

GSTP1

HLA-DRB1

ETS1

STAT4

```
java    -jar PriVar.jar
        -inputanno    /path/ variant_annotation_genome.csv
        -outputdir    /path/
        [-varqual 10]
        [-mapqual 10]
        [-depth 10]
        [-deletercut 0.5]
        -module Pri
        -pricand Candidate
        -customlist /path/candidate.txt
```

Parameter description

-Help

Help information

-buildver

The version of reference genome (default:hg19).

-resourcedir

The directory of resource folder.

-outputdir

Output directory.

-inputvcf

Input file for annotation (vcf format).

-inputanno

Input file for gene prioritization (csv format), this should be the output of annotation module.

-module

Functional modules: Ann: annotation, QC: quality control, Pri: gene prioritization, MenErr: mendelian error, Chipconcor: SNP chip concordance.

-pricand

Sub modules in gene prioritization including: IBD (linked based strategy),Homo (ROH-based strategy),Double (Double-hit -based strategy), Burden (Mutation burden), Denovo (*De novo* mutations) Candidate (Literature based Candidate Genes).

-triorder

The order of trio labeled (father: first position, mother: seconde position, child: third position, format as: 1:2:0 for the corresponding position in csv file (second, third, first).

-disease

Provide the Unified Medical Language of candidate disease which is used in identifying literature based candidate genes.

-population

Specify population (ASI: Asian, CEU: European, YRI: African) for calculating haplotype allele frequency .

-proporind

The minimum proportion of individuals shared the same mutated gene (default:0.8) for mutation burden module.

-allelefreq

The requirement of population allele frequency for defining common sites.(default:1%)

-customlist

The customized candidate gene list, should provide full directory.

-iBDcut

The minimum length requirement for evaluating shared IBD region (default:1Mb).

-homocut

The minimum length requirement for evaluating homozygous region (default:1Mb).

-pileup

pileup file for obtaining the covered sites.

-chip

SNP chip file (plink format) for SNP concordance evaluation (ped file).

-varqual

quality cutoff for variant quality (Phred Score) (default:10).

-mapqual

quality cutoff for mapping quality (Phred Score) (default:10).

-depth

cutoff for sequencing depth (default:10).

-deletercut

cutoff for deleterious effect prediction score (default:0.5).