Manual for PopLDdecay

PopLDdecay: a fast and effective tool for linkage disequilibrium

decay analysis based on variant call format files

version 3.30

2018-07-25

Contents

Usage

It is convenient for user to apply PopLDdecay to analysis the LD decay, just provide the SNP data in VCF format and perform follow two steps, users can get the decay figure.

> **Step1:** Calculate LD decay
> **Step2:** Draw the Figure

In the **Example part**, you can see the simple and clear usage to follow. Here are the instructions for the two steps.

Step1

In this step, users will use the core program named "*PopLDdecay*" , a Dist~R^2 statistics file about will be output, which will be as input file in the step2.

The parameter description are as follows:

```
[heweiming@hk-login-38-4 PopLDdecay-3.31]$ ./bin/PopLDdecay

       Usage: PopLDDecay -InVCF  <in.vcf.gz>  -OutStat <out.stat>

              -InVCF          <str>      Input SNP VCF Format
              -InGenotype     <str>      Input SNP Genotype Format
              -OutStat        <str>      OutPut Stat Dist ~ r^2 File

              -SubPop         <str>      SubGroup Sample File List[ALLsample]
              -MaxDist        <int>      Max Distance (kb) between two SNP [300]
              -MAF            <float>    Min minor allele frequency filter [0.005]
              -Het            <float>    Max ratio of het allele filter [0.88]
              -Miss           <float>    Max ratio of miss allele filter [0.25]
              -OutFilterSNP              OutPut the final SNP to calculate
              -OutPairLD      <int>      OutPut the PairWise SNP LD info [0]
                                         0/2:No_Out 1/3/4:Out_Brief 5:Out_Full
              -Method         <int>      Select the Cal agorithm [1]
                                         1:Low MEM  2 May Big MEM

              -help                      Show more help [hewm2008 v3.31]
```
<p align="center">Fig1 parameter description of PopLDdecay</p>

Note point:
A. With "*./bin/PopLDdecay –h*" command, users can get more help information.
B. Users can define the maximum distance with the command "*-MaxDist*", default 300 kb.
C. Users can also define their own filter criteria by using the command "*-MAF"*, *"-Het"*, *and "-Miss*".
D. To see detail pairwise SNP calculation information, use the command "*-OutPairLD*"
E. To calculate the **subgroup** LD decay in VCF Files, put their names into List file, and add parameters with "*–SubPop A.list*"

F. The program has two calculate algorithms, Method 1 is the optimal algorithm with low memory usage.

G. *'-i'* is short for *'–InVCF'* and *'-o'* is short for *'-OutStat'*, *'-s'* is short for *'-SubPop'*

A simple example

```
# 1)  For gatk VCF files, run PopLDdecay  directly
      ./bin/PopLDdecay    -InVCF  SNP.vcf.gz  -OutStat Lddecay.stat.gz
# 2)  For plink [.ped .map], change plink 2 genotype first  2) run PopLDdecay
      perl bin/mis/plink2genotype.pl   -inPED in.ped -inMAP in.map  -outGenotype
out.genotype ;      ./bin/PopLDdecay       -InGenotype out.genotype -OutStat LDdecay.stat.gz
# 3)  To Calculate the subgroup GroupA LDdecay in VCF Files   # put GroupA sample names into
GroupA_sample.list file
      ./bin/PopLDdecay   -InVCF    -OutStat    -SubPop   GroupA_sample.list
```

Step2

In this step, the main task is to plot the result in figure, here we provide two Perl scripts '*plot_OnePop.pl*' and '*Plot_MultiPop.pl*' to apply to different situations. And step2 only takes a few minutes to finish, user can change the drawing parameters according to their own requirements.

A. To plot one population LD decay, users can use '*plot_OnePop.pl*'. One population with multiple chromosome calculation result also can be generated to one file and plot the Result out.

B. To plot multiple populations in one figure, the scripts '*Plot_MultiPop.pl*' is recommend to plot the result.

The parameters of two '*plot_OnePop.pl*' and '*Plot_MultiPop.pl*' are similar. The parameters description are as follows:

Fig 2: Parameters description of plot PERL script

Note point:

A. User with *"-maxX"* can define theirs the max distance in the figure to plot

B. The parameter '-break' is the distance break point of *"-bin1"* and *"-bin2"*

C. The distance smaller than the break point size will use the *"-bin1"* size to smooth lines

D. The distance bigger than the break point size will use the *"-bin2"* size to smooth lines

E. Users can keep the R script to modify the figure by their self with command *"-keepR"*

A simple example

```
#    2.1  For one Population
```

```
    perl  bin/Plot_OnePop.pl  -inFile   LDdecay.stat.gz  -output  Fig
#    2.2  For one Population  multi chr          # List Format [chrResultPathWay]
    perl  bin/Plot_OnePop.pl  -inList   Chr.ReslutPath.List  -output Fig
#    2.3  For multi Populations               # List Format :[Pop.ResultPath  PopID]
    perl  bin/Plot_MultiPop.pl  -inList  Pop.ReslutPath.list  -output Fig
```

Classical case

Here, we provide four classic cases to demonstrate the application of this software, four situation will be show how to follow to get the LD decay figure out.

One population

This (one population with all chromosomes together) is most commonly used by users.
```
    ./bin/PopLDdecay -InVCF  ALLchr.vcf.gz  -OutStat  LDDecay.stat.gz
    perl bin/Plot_OnePop.pl -inFile  LDDecay.stat.gz -output  Out.Prefix
```
  Note:
     This will generate the two finale figures named "*Out.Prefix.png*" and "*Out.Prefix.pdf*"

Multi populations

For example, if there are 50 samples (wild1, wild2, wild3…wild25, cul1, cul2, cul3…cul25) in the VCF file,
To compare the LD decay of these two groups (wild vs cultivation), first of all, put their sample names into own file list for each group in one column or in one row.

```
    ./bin/PopLDdecay -InVCF  In.vcf.gz  -OutStat  wild.stat.gz  -SubPop wildName.list
    ./bin/PopLDdecay -InVCF  In.vcf.gz  -OutStat   cul.stat.gz  -SubPop culName.list
    #   created multi.list by yourself
    perl bin/Plot_MultiPop.pl -inList  multi.list  -output  OutputPrefix
```
  Note:
        A. The *<wildName.list>* can list as follows (in one row is also ok):
              wild1
              wild2
              ...
              Wild25
        B. The format of *<Multi.list>* had two columns , the file path of population result
           and the population flag, such as:
              /ifshk7/BC_PS/Lddecay/wild.stat.gz   *wild*
              /ifshk7/BC_PS/Lddecay/cul.stat.gz   *cultivation*

One population with multi-chromosomes

One population with multiple chromosome VCF files. For example, if there are 3 chromosomes VCF files (Chr1, Chr2 and Chr3) as the input.

```
        ./bin/PopLDdecay -InVCF  Chr1.vcf.gz  -OutStat  Chr1.stat.gz
        ./bin/PopLDdecay -InVCF  Chr2.vcf.gz  -OutStat  Chr2.stat.gz
        ./bin/PopLDdecay -InVCF  Chr3.vcf.gz  -OutStat  Chr3.stat.gz
        ls  `pwd`/Chr*.stat.gz   > chr.list
        perl bin/Plot_OnePop.pl -inList  chr.list  -output  OutputPrefix
```

Note:

A. Users can run in parallel when calculating the chromosomes' statistics files.

B. The files list only stores the file path, which is different from the multi-population list

C. It will generate the file 'OutputPrefix.bin' is the summary statistics file of all chromosomes, and same format with the chromosomes' statistics files.

D. the *<chr.list>* format can be generated by as above command 'ls Chr*.stat.gz > chr.list'.

Multi populations with multi-chromosomes

Multi population with multiple chromosome VCF files. For example, if there are 2 chromosomes VCF files (Chr1, Chr2) as the input.

```
        ./bin/PopLDdecay -InVCF  Chr1.vcf.gz  -OutStat  W.Chr1.stat.gz -SubPop wildName.list
        ./bin/PopLDdecay -InVCF  Chr2.vcf.gz  -OutStat  W.Chr2.stat.gz -SubPop wildName.list
        ./bin/PopLDdecay -InVCF  Chr1.vcf.gz  -OutStat  C.Chr1.stat.gz -SubPop culName.list
        ./bin/PopLDdecay -InVCF  Chr2.vcf.gz  -OutStat  C.Chr2.stat.gz -SubPop culName.list
        ls  `pwd`/W.Chr*.stat.gz   > W.chr.list
        perl bin/Plot_OnePop.pl -inList W.chr.list  -output  Wild.cat
        ls  `pwd`/C.Chr*.stat.gz   > C.chr.list
        perl bin/Plot_OnePop.pl -inList C.chr.list  -output  Cul.cat
        perl bin/Plot_MultiPop.pl -inList  multi.list  -output  OutputPrefix
```

Note:

A. The format of *<Multi.list>* had two columns , the file path of population result and the population flag, such as:

> /ifshk7/BC_PS/Lddecay/Wild.cat.bin   wild
> /ifshk7/BC_PS/Lddecay/Cul.cat.bin   cultivation