Manual for PopLDdecay

PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files

Version 3.40

2018-10-16

hewm2008@gmail.com / hewm2008@qq.com

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/D' statistics file will be obtained, and the file will be used as input file in step2.

The parameters are shown in Figure 1.

```
bc_rd@hk-software-install PopLDdecay]$ ./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/D' 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]
            -EHH            <str>     To Run EHH Region decay set StartSite [NA]
            -OutFilterSNP             OutPut the final SNP to calculate
            -OutType        <int>     1: R^2 result 2: R^2 & D' result 3:PairWise LD Out[1]
                                      See the Help for more OutType [1-8] details

            -help                     Show more help [hewm2008 v3.40]
```

Figure 1. Parameters for PopLDdecay

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 "*-OutType 3*"
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'*

H. With VCF files, PopLDdecay can also be used to calculate extended haplotype homozygosity (EHH) (Sabeti, et al., 2002) with the option of "*-EHH*"

```
[-OutType 1]    is the fastest for only cal (Dist ~ R^2) for MeanBin method plot
[-OutType 2]    will OutPut the Stat (Dist ~ r^2 & D') result for R^2 & D' MeanBin method plot
[-OutType 3]    will OutPut one more result of PairWise LD compaire result(with Dist~r^2)
[-OutType 4]    will OutPut the Stat (Dist ~ r^2 & D' ~ Number) result for R^2 & D' MeanBin/HW/MedianBin/PercentileBin plot
[-OutType 5]    will OutPut the Stat (Dist ~ r^2 ~ Number) result for R^2 MeanBin/HW/MedianBin/PercentileBin plot
[-OutType 6]    will OutPut one more result of PairWise LD compaire result(with Dist~r^2/D' )
[-OutType 7]    will OutPut one more result of PairWise LD compaire result(with Dist~r^2/D'/LOD)
[-OutType 8]    will OutPut one more result of PairWise LD compaire result(with Dist~r^2/D'/LOD/CIlow/CIhi )
```

Examples:

```
# 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 SNP.vcf.gz -OutStat Lddecay.stat.gz -SubPop
GroupA_sample.list
# 4)  To calculate EHH with phased VCF Files
     ./bin/PopLDdecay -InVCF  SNP.vcf.gz  -OutStat Lddecay.stat.gz –EHH chr1:235687
```

Step2

In this step, the main task is to plot the result, here we provide two Perl scripts '*plot_OnePop.pl*' and '*Plot_MultiPop.pl*' to use for different situations. Users can change the 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 results can also be used to plot the result.

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

The parameters of two '*plot_OnePop.pl*' and '*Plot_MultiPop.pl*' are similar. The parameters are shown in Figure 2.

```
perl Plot_OnePop.pl
        2016-04-22       hewm@genomics.cn

        Usage:    perl Plot_OnePop.pl  -inFile  LDdecay.stat.gz  -output OUT

        Options
        -inFile    <s> :  Input PopLDDecay OutPut Stat File
        -inList    <s> :  Input FileList if multi-File of PopLDDecay OutPut Stat
        -output    <s> :  Output Figure File Prefix

        -bin1      <n> :  the size bin for mean r^2/D' of Short Dist[10]
        -bin2      <n> :  the size bin for mean r^2/D' of Long Dist [100]
        -break     <n> :  break point to distinguish Short or Long Dist[100]
        -maxX      <n> :  max X coordinate Dist to plot LDdecay[kb] [maxDist]
        -measure   <s> :  use the [r2/D/both] to measure LD decay [r2]
        -method    <s> :  Plot methold (MeanBin/HW/MedianBin/PercentileBin)[MeanBin]
        -percent   <f> :  percent ratio(0-1) for PercentileBin method [0.5]
        -keepR         :  keep the R script for draw the LDdecay Fig
        -help          :  show this help
```

Figure 2. Parameters for the Per 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. By default, r^2 is used to generate the final plot, users can also use D' with "*-measure D*"
I. By default, the mean LD measure is used to generate the final plot, several other methods are also supported, including the mathematical function proposed by Hill and Weir (Hill and Weir, 1988) with "*-method HW*", median LD measure with "*-method MedianBin*", and percentile LD measure with "*-method PercentileBin*". Caution: With more than **100,000** observations, plotting with the **Hill and Weir** method will be memory and time consuming, and the default MeanBin method is recommended.
F. Users can keep the R script to modify the figure by their self with command "*-keepR*"

Examples

```
#  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*"

Multiple 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*' .

Multiple 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

Reference

Hill, W.G. and Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations, *Theor Popul Biol*, **33**, 54-78.

Sabeti, P.C.*, et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**, 832-837.