Supplementary

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

version 3.30

2018-06-29

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.

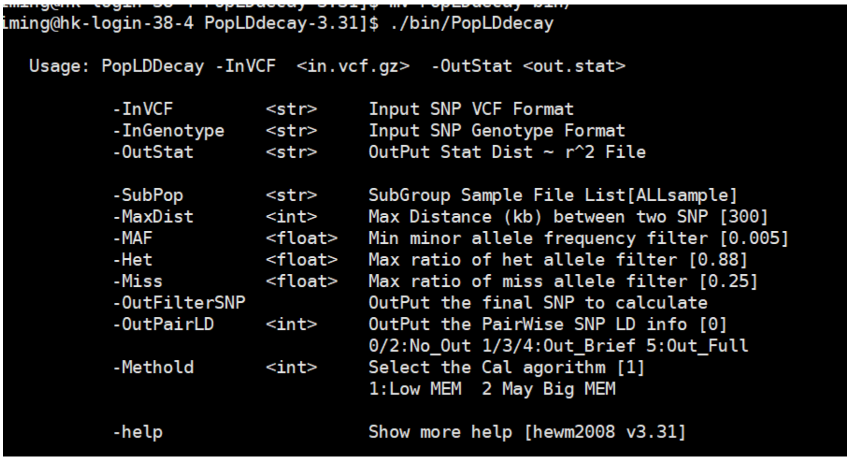
Step1: Calculate LD decay
Step2: Draw the Figure

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

Step1

In this step, User 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 as below:



```
iming@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
          -Methold       <int>     Select the Cal agorithm [1]
                                   1:Low MEM  2 May Big MEM

          -help                    Show more help [hewm2008 v3.31]
```

Fig1 parameter description of PopLDdecay

Note point:

A. User with "*./bin/PopLDdecay –h*" command will see more help, see some **Examples**.
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 -Miss*".
D. To see detail pairwise SNP calculation information, by using 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. Program had two calculate algorithm , Method 1 is the optimal algorithm at most time.
G. Parameters *'-i'* is short for *'–InVCF'* and *'-o'* is short for *'-OutStat', '-s'* is short for *'-SubPop'*

A simple example

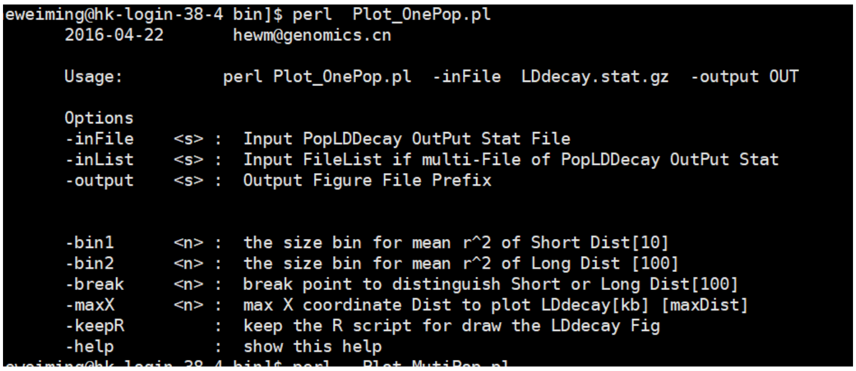
```
# 1)  For gatk VCF file deal , run PopLDdecay  direct
         ./bin/PopLDdecay    -InVCF  SNP.vcf.gz  -OutStat Lddecay.stat.gz
# 2)  For plink [.ped .map], chang plink 2 genotype first  2) run PopLDdecay
          perl bin/mis/plink2genotype.pl    -inPED in.ped -inMAP in.map  -outGenotype
out.genotype ;      ./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_MutiPop.pl*' to apply to different situations. And step2 only takes a few minutes to finish, User can change th**e drawin**g parameters until satisfied.

A. To plot one population LD decay, user can use this '*plot_**One**Pop.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_**Muti**Pop.pl*' is recommend to plot the result.

Parameters of two Perl scripts '*plot_OnePop.pl*' and '*Plot_MutiPop.pl*' are similar. Parameters description as below:



```
eweiming@hk-login-38-4 bin]$ 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 of Short Dist[10]
      -bin2      <n> :  the size bin for mean r^2 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]
      -keepR         :  keep the R script for draw the LDdecay Fig
      -help          :  show this help
```

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  muti chr         # List Format [chrResultPathWay]
     perl  bin/Plot_OnePop.pl  -inList   Chr.ReslutPath.List  -output Fig
#    2.3  For muti Population                 #  List Format :[Pop.ResultPath  PopID]
     perl  bin/Plot_MutiPop.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 situation (one population with all chromosomes together) is encountered by most users, and this situation is the simplest to carry out.

```
     ./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*"

Muti population

This is common situation in the LD decay analysis. 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, column or row is ok.

```
     ./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 manually  muti.list by yourself
     perl bin/Plot_MutiPop.pl -inList  muti.list  -output  OutputPrefix
```

Note:

A. The <*wildName.list*> can list as follow(column or row is ok):

   wild1
   wild2
   …
   Wild25

B. The format of <*muti.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-chr

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

B. The files list only store the file path, which is diff with 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 population with multi-chr

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_MutiPop.pl -inList  muti.list  -output  OutputPrefix
```

Note:

A. The format of <*muti.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
```

Evaluation

To test the accuracy and the efficiency of PopLDdecay, we used data of this web site(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502) to test follow software, only with two based site in chr22 (minimal chromosome SNP database) (**2504** sample with **1,055,401** SNP numbers)to test these software. And all the shell and Perl script can be found in attachment.

Accuracy

In order to test and evaluate the accuracy of the results, we apply PopLDdecay and Haploview to calculate LD by using the same SNP dataset with the same thresholds parameter.

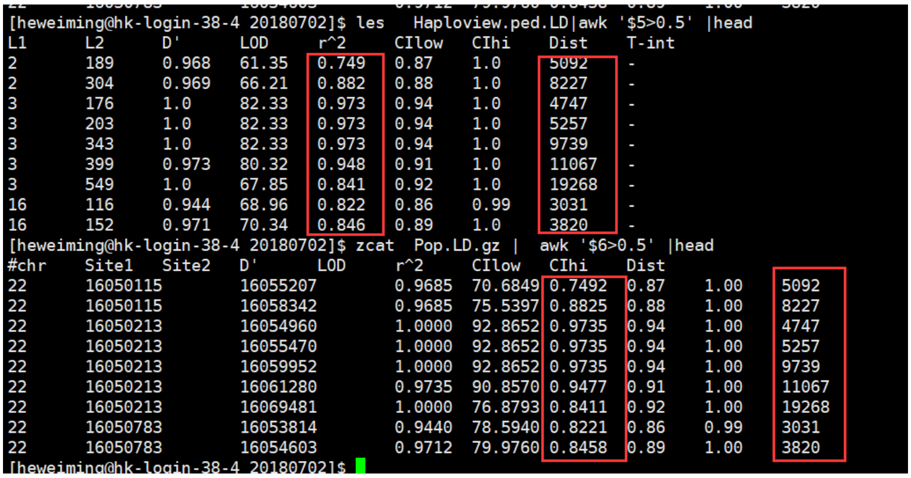
Haploview shell:

```
./iTools    Formtools       VCF2Genotype       -InPut    chr22.vcf           -OutPut Haploview.genoytpe.gz      -NoRef
perl    genotype2pedigree.pl       Haploview.genoytpe.gz      Haploview.ped      Haploview.info
java       -jar       ../Haploview.4.2.jar      -n      -pedfile    Haploview.ped -info    Haploview.info     -log    Hap.log    -
maxdistance 500    -minMAF 0.005    -hwcutoff        0.00      -dprime -memory 20960
perl    PlotLDdecay.pl    Hap.ped.LD    LDdecay.svg
```

PopLDdecay shell:

```
./PopLDdecay         -InVCF      chr22.vcf            -MaxDist           500       -OutStat            Pop        -MAF
0.005      -OutPairLD      5
```

we find the results of LD measure r^2 , D` and distance is exactly the same, as well as the number of pairwise comparisons. Therefore, we can get the conclusion that the results of PopLDdecay is accurate. Here is the top 10 line of two software results.



```
[heweiming@hk-login-38-4 20180702]$ les   Haploview.ped.LD|awk '$5>0.5' |head
L1      L2      D'      LOD     r^2     CIlow   CIhi    Dist    T-int
2       189     0.968   61.35   0.749   0.87    1.0     5092    -
2       304     0.969   66.21   0.882   0.88    1.0     8227    -
3       176     1.0     82.33   0.973   0.94    1.0     4747    -
3       203     1.0     82.33   0.973   0.94    1.0     5257    -
3       343     1.0     82.33   0.973   0.94    1.0     9739    -
3       399     0.973   80.32   0.948   0.91    1.0     11067   -
3       549     1.0     67.85   0.841   0.92    1.0     19268   -
16      116     0.944   68.96   0.822   0.86    0.99    3031    -
16      152     0.971   70.34   0.846   0.89    1.0     3820    -
[heweiming@hk-login-38-4 20180702]$ zcat  Pop.LD.gz |  awk '$6>0.5' |head
#chr    Site1   Site2   D'      LOD     r^2     CIlow   CIhi    Dist
22      16050115        16055207        0.9685  70.6849 0.7492  0.87    1.00    5092
22      16050115        16058342        0.9685  75.5397 0.8825  0.88    1.00    8227
22      16050213        16054960        1.0000  92.8652 0.9735  0.94    1.00    4747
22      16050213        16055470        1.0000  92.8652 0.9735  0.94    1.00    5257
22      16050213        16059952        1.0000  92.8652 0.9735  0.94    1.00    9739
22      16050213        16061280        0.9735  90.8570 0.9477  0.91    1.00    11067
22      16050213        16069481        1.0000  76.8793 0.8411  0.92    1.00    19268
22      16050783        16053814        0.9440  78.5940 0.8221  0.86    0.99    3031
22      16050783        16054603        0.9712  79.9760 0.8458  0.89    1.00    3820
[heweiming@hk-login-38-4 20180702]$
```

Fig 3: comparison Result of Accuracy

Here we also give out the Figure of LD Decay by *Haploview* and *PopLDdecay*, the two software lines are **overlapping**.

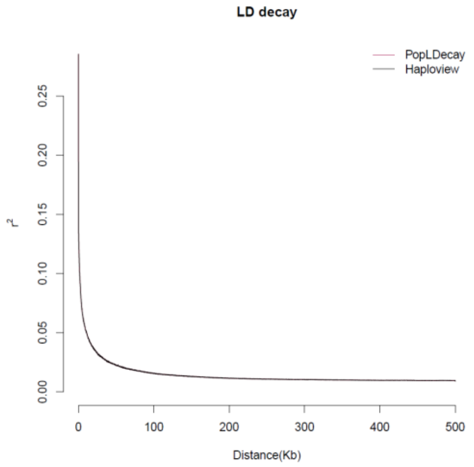


Fig 4: LD Decay of *Haploview* and *PopLDdecay result*

Efficiency

To compare the efficiency of this software, The Haploview and PLINK were taken to make a compare. Here are the shell script of running comparison.

Plinks 1.07 shell

```
../plink2.0 --vcf chr22.vcf --out chr22
../plink_v1.07 --bfile chr22 --noweb --ld-window-r20 --r2 --ld-window
        99999 --ld-window-kb 300 --maf 0.005 --out P1.7
Perl StatLD2Decay.pl P1.7.ld P1.7.ld.stat
```

Plinks 1.9 shell

```
../plink2.0 --vcf chr22.vcf --out chr22
../plink2.0 --bfile chr22 --noweb --ld-window-r20 --r2 --ld-window
        99999 --ld-window-kb 300 --maf 0.005 - -out P2.0 --threads 1
Perl StatLD2Decay.pl P2.0.ld P2.0.ld.stat
```

PopLDdecay shell:

```
./PopLDdecay -InVCF chr22.vcf -MaxDist 300 -OutStat Pop -MAF
0.005
```

Haploview shell:

```
../iTools Formtools VCF2Genotype -InPut chr22.vcf -OutPut
Haploview.genoytpe.gz -NoRef
perl ../genotype2pedigree.pl Haploview.genoytpe.gz Haploview.ped Haploview.info
time java -jar -Xmx98g -jar ../Haploview.4.2.jar -n -pedfile
Haploview.ped -info Haploview.info -log Hap.log -maxdistance 300 -minMAF 0.005 -
hwcutoff 0.00 -dprime -memory 102400 #must set 100G ,Haploview can run complete
```

The comparison result were show at the table1. Form the table.,we can see

1. The calculation time of PopLDdceay is much little, it is acceptable, although no the shortest one.
2. The average memory of PopLDdceay also takes much little, it is acceptable.
3. Since there is no intermediate file generation, the PopLDdceay output file takes up only little space.

Table 1. Computational resources statistics for chr22 LD decay

version	Average memory	Core calculation CPUs	Predicted Format conver & Statistics Time CPU	result size
Plink 1.07	1.4G	680min	5min+45min	54G
Plink 2.0	18.81G	25min	5min+45min	54G
Haploview 4.2	95.76G	3904min	5min+45min	54G
PopLDdcay 3.30	1.5G	200min	0min	4.1M

Here we also give out the Figure of LD Decay by *Haploview*, *PLINK* and *PopLDdecay*, the trend of four software lines is consistent.

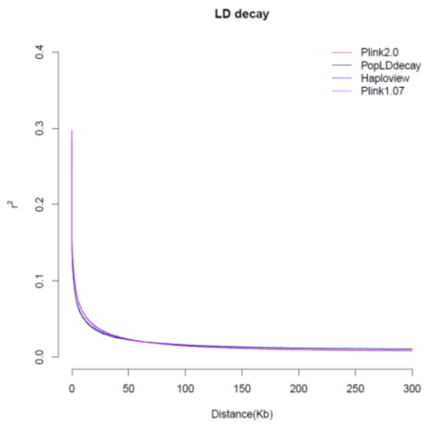


Fig 5: LD Decay of comparison software

Cited

The PopLdecay has has been cited many times. At least 6 times(update 2018-06-30) as follow(Chen, et al., 2017; Cui, et al., 2017; Li, et al., 2018; Liu, et al., 2016; Wu, et al., 2018; Zhang, et al., 2017)

Chen, W.*, et al.* Genetic Diversity, Population Structure, and Linkage Disequilibrium of a Core Collection of Ziziphus jujuba Assessed with Genome-wide SNPs Developed by Genotyping-by-sequencing and SSR Markers. *Front Plant Sci* 2017;8:575.

Cui, C.*, et al.* Genetic Diversity, Population Structure, and Linkage Disequilibrium of an Association-Mapping Panel Revealed by Genome-Wide SNP Markers in Sesame. *Front Plant Sci* 2017;8:1189.

Li, C.*, et al.* The genetic architecture of amylose biosynthesis in maize kernel. *Plant Biotechnol J* 2018;16(2):688-695.

Liu, H.*, et al.* Gene duplication confers enhanced expression of 27-kDa gamma-zein for endosperm modification in quality protein maize. *Proc Natl Acad Sci U S A* 2016;113(18):4964-4969.

Wu, Y.*, et al.* Population genomic data reveal genes related to important traits of quail. *Gigascience* 2018;7(5).

Zhang, L.*, et al.* RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nat Commun* 2017;8(1):2264.