LDBlockShow

Manual

Visualizing linkage disequilibrium and haplotype blocks based on variant call format files

Version 1.30

2020-07-28

hewm2008@gmail.com / hewm2008@qq.com

Contents

1. Introduction

LDBlockShow is a fast and effective tool to generate linkage disequilibrium (LD) heatmap from VCF files. It is more time and memory saving than other current tools. LDBlockShow can generate the plots of LD heatmap and interested statistics or annotation results simultaneously. In addition, it also supports subgroup analysis.

2. Download and Install

2.1 Download website

https://github.com/BGI-shenzhen/LDBlockShow/

2.2 Pre-install

LDblockshow is for Linux/Unix/macOS only. Before installing, please make sure the following pre-requirements are ready to use.

1) g++ : g++ with --std=c++11 > 4.8+ (https://gcc.gnu.org/) is recommended

2) zlib : zlib > 1.2.3 (https://zlib.net/) is recommended

3) Perl: The SVG.pm (https://metacpan.org/release/SVG) in Perl should be installed. LDBlockShow uses this module to plot figures. We have provided a built-in SVG module in the package.

2.3 Install

Users can install it with the 3 following options:

1) Option 1:

```
git clone https://github.com/BGI-shenzhen/LDBlockShow.git
chmod 755 configure; ./configure;
make;
mv   LDBlockShow   bin/;      #       [rm *.o]
```

2) Option 2:

```
tar -zxvf   LDBlockShowXXX.tar.gz
cd LDBlockShowXXX;
cd src;
sh make.sh          # or in Linux: make ; make clean
../bin/LDBlockShow
```

3) Option 3:

We also have the static compilation version for Linux/Unix, which can be used directly after un-compression. You can contact me (hewm2008@gmail.com or hewm2008@qq.com) to get it.
Note: If link failed, please try to reinstall the zip library (https://zlib.net/).
Note: For mac OS, if plink does no work, please re-download plink for mac OS (https://www.cog-genomics.org/plink2/) and put it in the "LDBlockShowXXX/bin" directory.

3. Parameter description

3.1 LDBlockShow

3.1.1 Main parameters

```
[heweiming@cngb-ologin-25 bin]$ ./bin/LDBlockShow
     Usage: LDBlockShow   -InVCF   <in.vcf.gz>   -OutPut <outPrefix>   -Region   chr1:10000-20000

          -InVCF          <str>       Input SNP VCF Format
          -OutPut         <str>       OutPut File of LD Blocks
          -Region         <str>       In One Region to show LD info svg Figure


          -SeleVar        <int>       Select statistic for deal. 1: D' 2: R^2 [1]
          -SubPop         <str>       SubGroup Sample File List [ALLsample]
          -BlockType      <int>       method to detect Block [beta] [1]
                                       1. Block by PLINK (Gabriel method withed D')
                                       2. Solid Spine of LD RR/D'    3. Blockcut with self-defined RR/D'
                                       4. FixBlock by input blocks files 5. No Block

          -InGWAS         <str>       InPut GWAS Pvalue File (chr site Pvalue)
          -InGFF          <str>       InPut GFF3 file to show Gene CDS and name

          -BlockCut       <float>     'Strong LD' cutoff and ratio for BlockType3 [0.85:0.90]
          -FixBlock       <str>       Input fixed block region
          -MerMinSNPNum <int>         merger color grids when SNPnumber over N[50]

          -help                       Show more Parameters and help [hewm2008 v1.30]
```

-InVCF The input file in VCF format
-OutPut The output file directory and output file name prefix (e.g., /path/pop1)
-Region The defined region to show the LD heatmap (format: chr:start:end)

-SeleVar The LD measurement (1: D' 2: R^2), the default is 1.
-SubPop A sample list for subgroup analysis

-BlockType The definition of blocks. The default 1 is called by PLINK[1] to generate the block defined by Gabriel *et al.*[2]. Solid spine of LD[3] is also supported [2]. Users can also define their own cutoff of r^2 and D' for blocks [3] combined with the option of "*-BlockCut*" or supply their own block region definition [4] combined with the option of "*-FixBlock*". 5 can be used as input if users prefer to not show the block region.

-InGWAS The statistics file (e.g., association statistics, but other values such as Tajima's D can also be accepted) for generate plot together with the LD plot. File formatted as: [chr position Pvalue]

-InGFF Input GFF3 format file for genomic region annotation

-BlockCut For block type 3, the defined cutoff for strong LD, and the ratio of strong LD SNP in one block. Default is 0.85:0.9. That's, if the user chose D' in the –SeleVar option, then in one block, the ratio of SNP pairs with D' over 0.85 is 0.9.

-FixBlock For block type 4, users can use this option to supply a self-defined block region. The file contains three columns, including chromosome, block region start position, and block region end position.

-MerMinSNPNum The minimum SNP number to merge color grids with the same color. Default is 50. Details please see Fig 1 in this manual.

-help Show more parameters

3.1.2 Other parameters

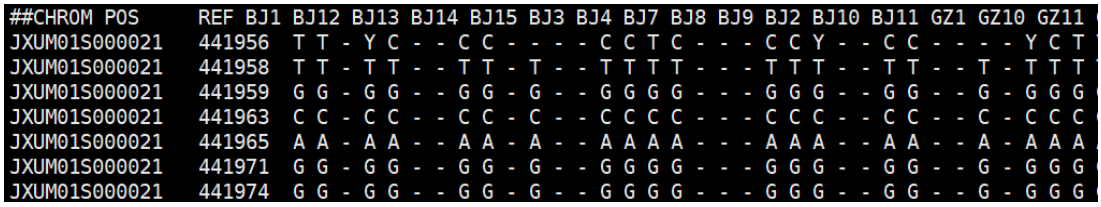
```
[heweiming@cngb-ologin-25 bin]$ ./bin/LDBlockShow  –h
        More Help document please see the Manual.pdf file
        Para [-i] is show for [-InVCF], Para [-o] is show for [-OutPut], Para [-r] is show for [-Region]

      -InGenotype     <str>       InPut SNP Genotype Format
      -InPlink        <str>       InPut Plink [bed+bim+fam] or [ped+map] file prefix

      -MAF            <float>     Min minor allele frequency filter [0.05]
      -Het            <float>     Max ratio of het allele filter [0.90]
      -Miss           <float>     Max ratio of miss allele filter [0.25]

      -TagSNPCut      <float>     'Strong LD' cutoff for TagSNP [0.80]
      -OutPng                     convert svg 2 png file
      -OutPdf                     convert svg 2 pdf file
```

-InGenotype Input file in genotype format. The format of genotype file is as follows:



```
##CHROM POS     REF BJ1 BJ12 BJ13 BJ14 BJ15 BJ3 BJ4 BJ7 BJ8 BJ9 BJ2 BJ10 BJ11 GZ1 GZ10 GZ11
JXUM01S000021   441956  T T - Y C - - C C - - - - C C T C - - - C C Y - - C C - - - - Y C T
JXUM01S000021   441958  T T - T T - - T T - T - - T T T T - - - T T T - - T T - - T - T T T
JXUM01S000021   441959  G G - G G - - G G - G - - G G G G - - - G G G - - G G - - G - G G G
JXUM01S000021   441963  C C - C C - - C C - C - - C C C C - - - C C C - - C C - - C - C C C
JXUM01S000021   441965  A A - A A - - A A - A - - A A A A - - - A A A - - A A - - A - A A A
JXUM01S000021   441971  G G - G G - - G G - G - - G G G G - - - G G G - - G G - - G - G G G
JXUM01S000021   441974  G G - G G - - G G - G - - G G G G - - - G G G - - G G - - G - G G G
```

-InPlink The prefix of input file in PLINK format.

-MAF Filter SNPs with low minor allele frequency (default ≤0.05)

-Het Filter SNPs with high heterozygosis ratio (default ≥0.9)

-Miss Filter SNPs with high missing rate (default ≥0.25)

-TagSNPCut The LD cutoff for selecting tag SNPs. Default is 0.8.

-OutPng Convert the SVG file to PNG file

-OutPdf Convert SVG file to Pdf file.

Note: If users failed to open small SVG files, please use the "-Outpdf" option to use the PDF file. For large SVG files, "-OutPng" can be used to get a relatively small figure file.

3.2 ShowLDSVG

This program is designed for users to optimize the figure (e.g., change colors) generated by LDBlockShow.

3.2.1 Brief parameters

```
./bin/ShowLDSVG
        Options
          -InPreFix     <s> :   InPut Region LD Result Frefix
          -OutPut       <s> :   OutPut svg file result
          -help             :   Show more help with more parameter
```

-InPreFix The prefix of input file (i.e., the output file of LDBlockShow)

-OutPut The out file (svg, png and pdf format plot files)

-help More parameters in detail

3.2.2 Detail parameters

-InGWAS The statistics file (e.g., association statistics, but other values such as Tajima's D can also be accepted) for generate plot together with the LD plot. File formatted as: [chr position Pvalue]

-NoLogP By default, the P value from the –InGWAS file will be –log10 transformed, with this option, the P value will not be transformed.

-Cutline The significance cutline of the –InGWAS file.

-PointSize Users can use this option (any number over 0) to set the point size.

-SpeSNPName With this option, users can input a file to indicate the names for interested SNPs, these names will be shown in the heatmap.

-ShowGWASSpeSNP Users can use this option together with the file assigned by "-SpeSNPName" to show the names of interested SNPs in GWAS plot.

```
./bin/ShowLDSVG     -h
        -InGWAS         <s>     : InPut GWAS Pvalue File(chr site Pvalue)
        -NoLogP                 : Do not get the log Pvalue
        -Cutline        <s>     : show the cut off line of Pvlaue
        -PointSize      <n>     : set the GWAS point size number
        -SpeSNPName <s>     : In File for Special SNP Name(chr site Name)
        -ShowGWASSpeSNP    : show Special SNP Name in GWAS plot with [-SpeSNPName]

        -InGFF          <s>     : InPut GFF3 file to show Gene CDS and name
        -NoGeneName             : No show Gene name,only show stuct
        -crGene         <s>     :InColor for Gene Stuct [CDS:Intron:UTR:Intergenic]
                                        default: [#e7298a:lightblue:#7570b3:#a6cee3]

        -crBegin        <s>     : In Start Color RGB [255,255,255]
        -crMiddle       <s>     : In Middle Color RGB [240,235,75]
        -crEnd          <s>     : In End Color RGB [255,0,0]
        -NumGradien     <s>     :In Number of gradien of color
        -crTagSNP       <s>     : Color for TagSNP [31,120,180]

        -CrGrid         <s>     : the color of grid edge [white]
        -WidthGrid      <s>     : the edge-width of gird [1]
        -NoGrid                 : No Show the gird edge

        -ShowNum                : Show the R^2/D' in the heatmap
        -NoShowLDist    <n>     : NoShow long physical distance pairwise[1000000]
        -MerMinSNPNum <s>    : merge color grids when SNPnumber over N[50]

        -OutPng                 : convert svg 2 png file
        -OutPdf                 : convert svg 2 pdf file
        -ResizeH                : resize image height; Width be resize in ratio[4096]
```

-InGFF The GFF file for genomic region annotation. By default, the gene name will be shown in the plot;

-NoGeneName Gene name will not be shown in the plot with this option.

-crGene Define the colors of different genomic regions. By default, CDS, intron, UTR and intergenic regions will be shown in #e7298a, light blue, #7570b3, and #a6cee3, respectively.

Parameters to optimize the color of the heatmap:

-crBegin Color for no LD (R^2/D'=0) default: white

-crMiddle Color for R^2/D'=0.5, default: yellow

-crEnd Color for complete LD (R^2/D'=1), default: red

-NumGradien The number of gradients from crBegin to crEnd

-crTagSNP Color for the tag SNP.

Parameters to optimize the grids in the heatmap:

-CrGrid Border color of the grids, default: white

-WidthGrid The width of the border, default = 1

-NoGrid No border

-ShowNum Show the LD measurement value in the grids (not recommended when SNP number is over 50).

-NoShowLDist When the distance between SNPs over this number, their pairwise LD will not be showed in the figure. Default is 10,000,000.

-MerMinSNPNum When number of SNPs over the default 50, ShowLDSVG will merge adjacent same color grids. User can change this number to any integer numbers.

-OutPng Convert the SVG file to PNG file

-OutPdf Convert SVG file to Pdf file.

-ResizeH Set the height of the image (default 4096), which can be used to adjust the resolution for PNG file. The width will be adjusted automatically.

When SNP number is large (e.g., over 100), the output SVG file might be very large. ShowLDSVG will merge adjacent same color grids. Below is an example to compress a SVG file from 26k to 8k. With smaller number of gradients (set by –NumGradien), the figure will be compressed to be smaller. –MerMinSNPNum is used to set the minimum number of SNPs, that's, if there is more SNPs than this number (default 50), the output SVG will be compressed.

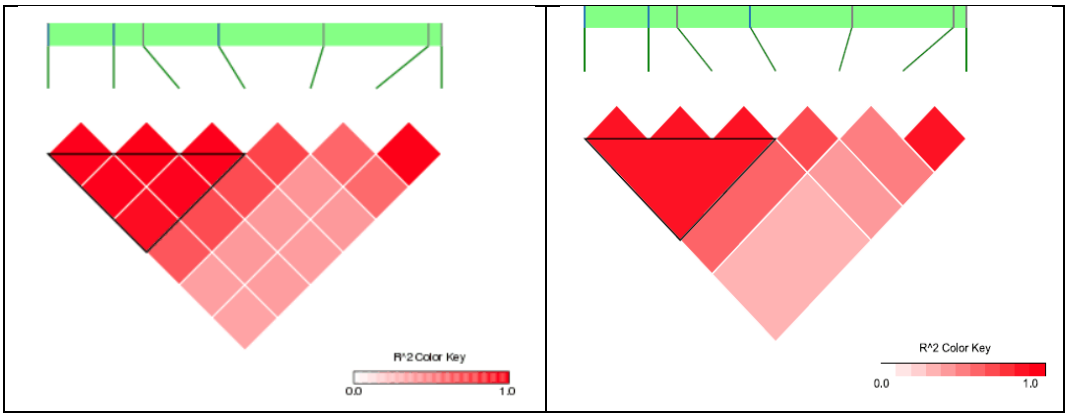


Figure 1. An example to compress LD heatmap with large number of SNPs.

3.3 Output files

Output files	Description
out.site.gz	Remained SNPs after filtering [chr site]
out.blocks.gz	Block file [chr start end block_length SNP_number SNPs]
out.TriangleV.gz	Region Pairwise R^2/D'
out.svg	Output plot in SVG format
out.png	Output plot in png format
out.pdf	Output plot in pdf format

4. Examples

All examples are using R^2 as the LD measurement, but the default measurement is D'.

4.1 Example1: Heatmap + default block generated by PLINK

In the example/Example1 directory, we have provided an example to generate the LD heatmap with the default block generated by PLINK. Example command line is shown in the run.sh file:

```
../../bin/LDBlockShow    -InVCF Test.vcf.gz    -OutPut out -Region chr11:24100000:24200000 -OutPng    -SeleVar 2
```

```
sh run.sh
Start Time :
Mon Jun  1 16:30:19 CST 2020
#Detected VCF File is phased file with '|', Read VCF in Phase mode
##Start Region Cal...    :Ghir_D11 24100000 24200000; In This Region TotalSNP Number is 7
find blocks...
Start draw... SVG info: SNPNumber :7 , SVG (width,height) = (402.5,297.5)
convert   SVG ---> PNG ...
End Time :
Mon Jun  1 16:30:19 CST 2020
```

```
ls
out.blocks.gz  out.pdf  out.png  out.site.gz  out.svg  out.TriangleV.gz
```

The final plot is shown in Figure 2.

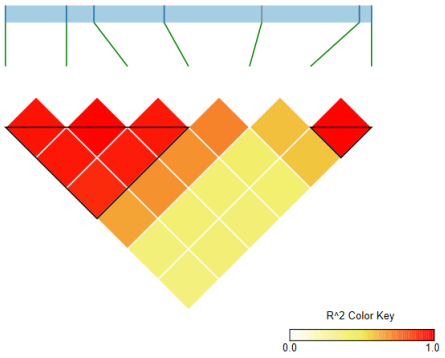


Figure 2. Plot generated in Example 1 using R^2 as the LD measurement.

If using D' as the LD measurement, the final plot is shown in Figure 3.

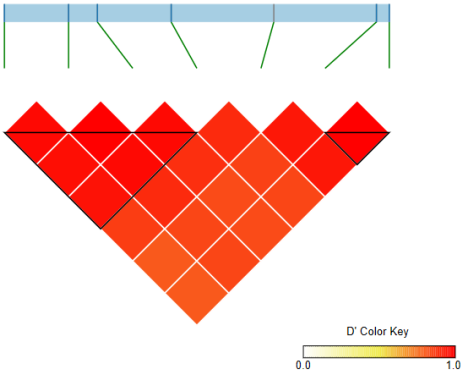


Figure 3. Plot generated in Example 1 using D' as the LD measurement.

4.2 Example 2: Heatmap + block + GWAS

In the example/Example2 directory, we have provided an example to generate the plot with the heatmap, the default block, and GWAS statistics. Example command line is shown in the run.sh file:

```
../../bin/LDBlockShow –InVCF ../Example1/Test.vcf.gz -OutPut out -Region chr11:24100000:24200000 -InGWAS gwas.pvalue -SeleVar 2
```

The generated plot is shown in Figure 4. By default, points with -log10(P value) larger than 7.3 ($P < 5 \times 10^{-8}$) are shown in red.

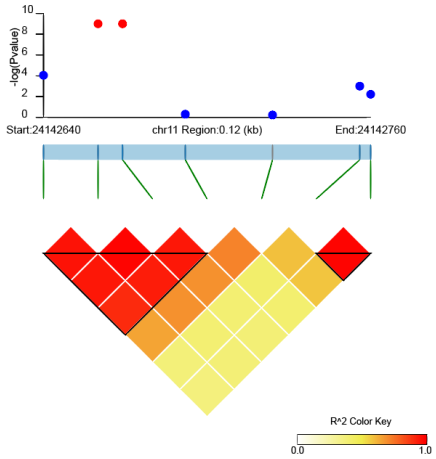


Figure 4. Heatmap + block + GWAS plot in Example 2

Users can further use ShowLDSVG to optimize the plot. Example command line is shown in the run.sh file:

```
../../bin/ShowLDSVG -InPreFix out -OutPut out -InGWAS gwas.pvalue -Cutline 7 -ShowNum -PointSize 3
```

The optimized figure is shown in Figure 5.

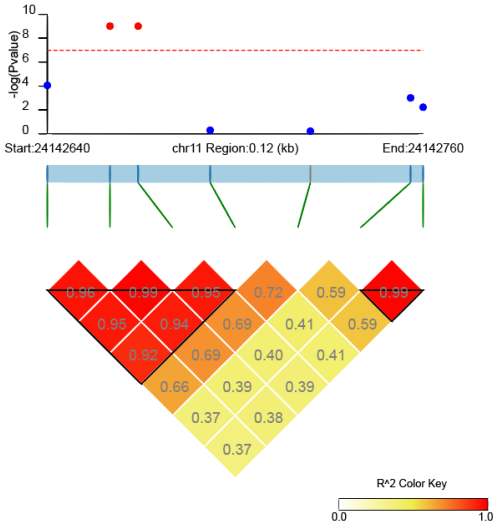


Figure 5. Optimized figure using ShowLDSVG in Example 2.

4.3 Example 3: Heatmap + block + GWAS + Annotation

In the example/Example3 directory, we have provided an example to generate the plot with the heatmap, the default block, GWAS statistics, and genomic annotation. Example command line is shown in the run.sh file:

```
../../bin/LDBlockShow -InVCF ../Example1/Test.vcf.gz -OutPut out -InGWAS gwas.pvalue -InGFF In.gff
-Region chr11:24100000:24200000 –OutPng -SeleVar 2
```

The generated plot is shown in Figure 6.

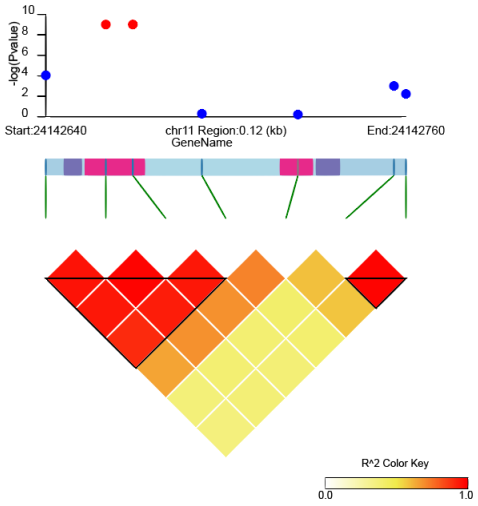


Figure 6. Heatmap + block + GWAS + Annotation plot in Example 3

Users can further use ShowLDSVG to optimize the plot. Example command line is shown in the run.sh file:

```
../../bin/ShowLDSVG -InPreFix out -OutPut out -InGWAS gwas.pvalue -Cutline 7 -InGFF In.gff -crGene
yellow:lightblue:pink:orange –showNum
```

The optimized figure is shown in Figure 7.

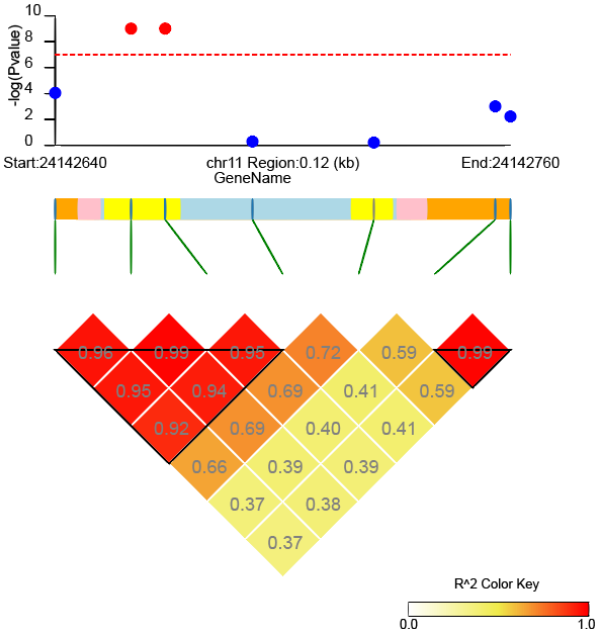


Figure 7. Optimized figure using ShowLDSVG in Example 3.

Users can also show the name of interested SNPs with the following command:

```
../../bin/ShowLDSVG -InPreFix out -OutPut out.svg -InGWAS gwas.pvalue -Cutline 7 -InGFF In.gff -crGene yellow:lightblue:pink:orange -showNum -OutPng -SpeSNPName Spe.snp -ShowGWASSpeSNP
```

The optimized figure is shown in Figure 8.

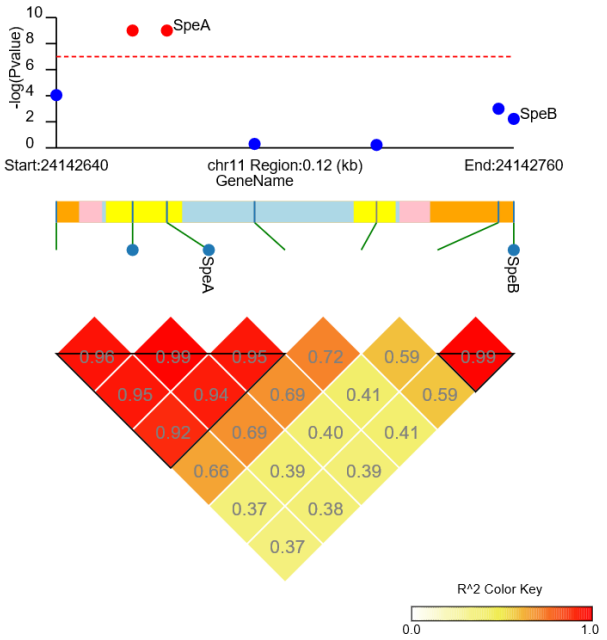
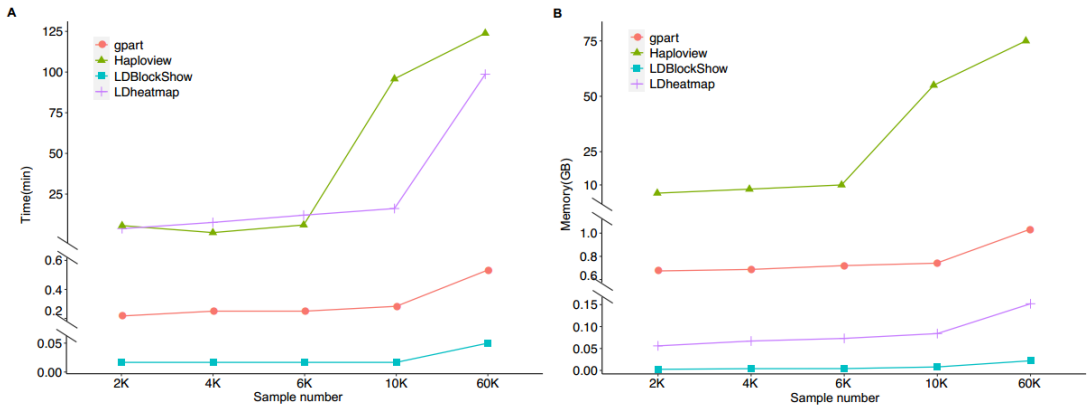


Figure 8. Optimized figure showing names for interested SNPs.

5. Advantages

To evaluate the performance of LDBlockShow, we used test VCF files to generate the LD heatmap by using LDBlockShow, Haploview[3], LDheatmap[4] and gpart[5]. The calculated r^2 and D' values of LDBlockShow is the same with other tools. As shown in Figure 9, LDBlockShow is more time and memory saving than other tools.



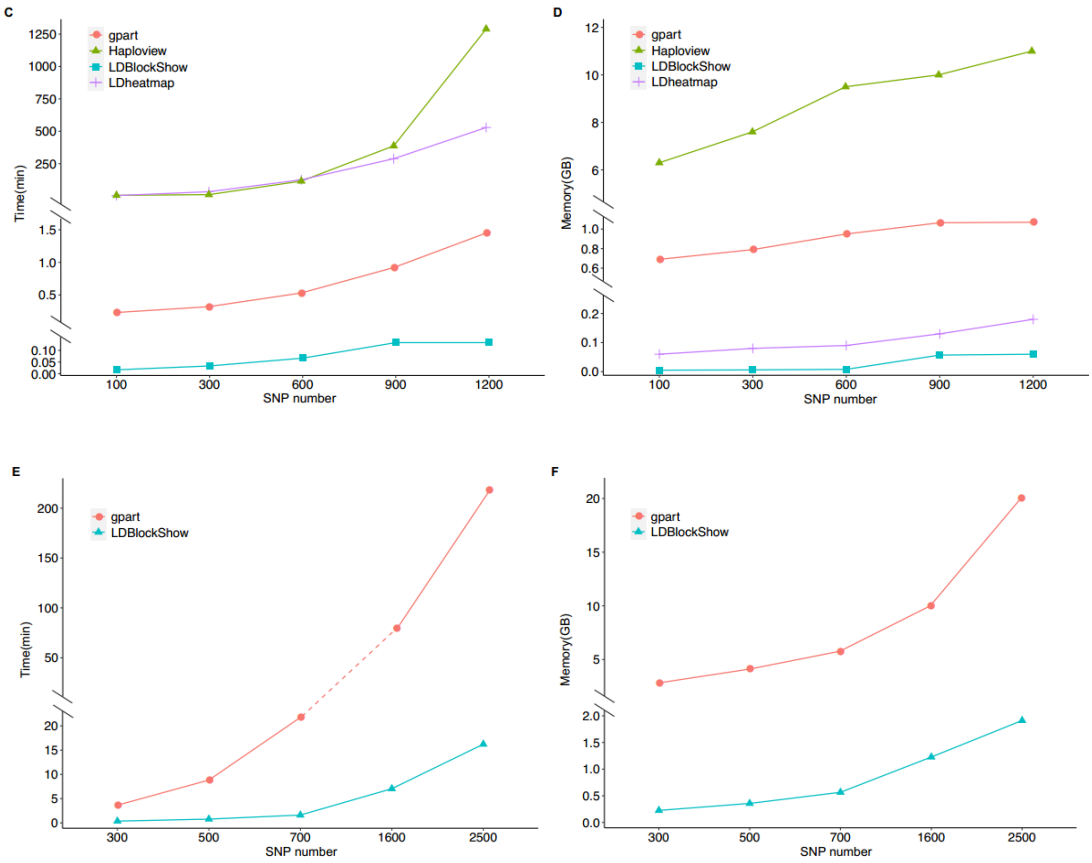


Figure 9. Comparison of computing cost for LDBlockShow, Ldheatmap, Haploview and gpart. CPU time (A) and memory cost (B) for different methods are shown with a fixed SNP number of 100 and sample size ranging from 2,000 to 60,000. CPU time (C) and memory cost (D) for different methods are shown with a fixed sample size of 2,000 and SNP number ranged from 100 to 1,200. When testing datasets in A-D, both LDBlockShow and gpart finished the analyses within reasonable time and memory. We further tested their performance when handling large dataset. CPU time (E) and memory cost (F) for these two methods are shown with a fixed sample size of 100,000 and SNP number ranged from 300 to 2,500. Computation is performed with one thread of an Intel Xeon CPU E5-2630 v4.

As shown in Table 1, LDBlockShow can generate the plots of LD heatmap and interested statistics or annotation results simultaneously. In addition, LDBlockShow also supports subgroup analysis.

Table 1. Comparison of LDBlockShow with other tools

Performance	**LDBlockShow**	**Haploview**	**LDheatmap**	**gpart**
Input				
Compressed VCF file	√	×	×	×
Uncompressed VCF file	√	×	×	√
Support subgroup analysis	√	×	×	×
Output				
Visualize additional statistics	√	×	×	×
Visualize genomic annotation	√	×	×	√

Compressed SVG	√	×	×	×
PNG file	√	√	×	√
Block region	√	√	×	√
LD measurement	D'/r^2	D'/r^2	r^2	D'/r^2

6. Frequently ask questions

6.1 How to calculate LD measurement in LDBlockShow

Similar to our previously published tool for LD decay analysis[6], pairwise LD measurements r^2 and D' were calculated according to previously reported formulas[7,8]. The calculated r^2 and D' values of LDBlockShow is the same with other tools. For example, as shown in Figure 10, the heat map we generated is the same with LDheatmap.

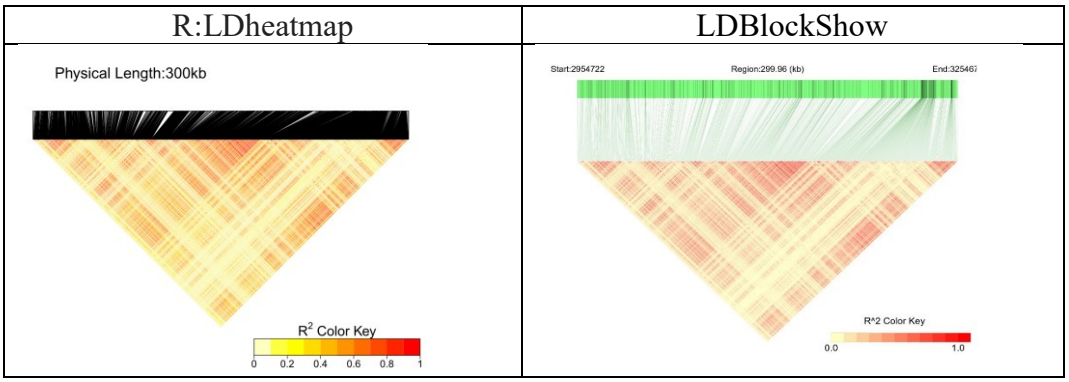


Figure 10. Comparison of the results between LDheatmap and LDBlockShow

6.2 Can another statistics rather than the GWAS results be supported?

Yes, of course. In the file supported by the option –InGWAS, the third column can be defined as any values. With the option –NoLogP, the values will not be log transformed.

Feel free to contact me for other requirements!

Reference

1. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
2. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
3. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
4. Shin, J.-H., Blay, S., McNeney, B. & Graham, J. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *2006* **16**, 9 (2006).
5. Kim, S.A. *et al.* gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics* **35**, 4419-4421 (2019).
6. Zhang, C., Dong, S.S., Xu, J.Y., He, W.M. & Yang, T.L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786-1788 (2019).
7. Lewontin, R.C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**, 49-67 (1964).
8. Hill, W.G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**, 226-31 (1968).