

LDBlocksShow

中英文使用手册

a fast and effective tool for Show Region Linkage disequilibrium
heatmap Figure analysis based on variant call format files

一种基于 vcf 格式文件快速计算某一区间的
连锁程度热度图的工具

Version 1.20

2020-04-23

hewm2008@gmail.com / hewm2008@qq.com

Contents

| | |
|--|----|
| LDBlocksShow 中英文使用手册..... | 1 |
| 背景介绍..... | 1 |
| 下载安装..... | 2 |
| 下载: | 2 |
| 安装: | 2 |
| 参数说明..... | 2 |
| LDBlockShow: | 3 |
| ShowLDSVG: | 4 |
| 输出文件..... | 6 |
| 经典实例..... | 7 |
| 实例 1: 倒三角+默认 LD Blocks..... | 7 |
| 实例 2: 倒三角+默认 LD Blocks + GWAS | 8 |
| 实例 3: 倒三角+plinks LD Blocks..... | 9 |
| 实例 4: 倒三角+plinks LD Blocks + GWAS..... | 9 |
| 实例 5: 倒三角+默认 LD Blocks (+ GWAS)+GeneStuct..... | 10 |
| 常见问题..... | 10 |
| 问题 1 结果准确性..... | 10 |
| 问题 2 显示 region 的其它 Pi 可否? | 11 |

注: 倒三角 就是该区域的两两 SNP 的 LDHeatMap

背景介绍

群体全基因组关联分析 GWAS 中，常关联到一些点，要看看这些点带动周边其它点连锁程度，一般须要计算这一区间里面的 SNP 两两之间的关联系数(r^2)和 LD blocks 等。现存在两个比较常用的软件 Haploview4.2.jar 和 R 包 Ldheatmap，但这两个软件都存在很多不够之处。如区域较大，里面的 SNP 多的话，则这两个软件的用到的计算资源极大。还有格式转化等不方便。现以 1000 个位点 1000 个样品来评价。

- 1 Haploview 要用到的内存是 95G，其中计算时间长达 36 小时。
- 2 Ldheatmap 虽然内存只用到 1G，但计算时长也达 48 小时并且没有 LD Blocks 的输出信息
- 3 两个软件的结果和 GWAS 的图要整合在一起时，须要手动 AI 操作浪费人力时间。
- 4 输出的图须要小修时，须又要重新计算，浪费时间。

基于上面等原因，特此开发本软件，在同等 1k 位点和 1k 的样品中，只须用到 1G 内存同时只须要 1 小时的计算时间就可以出结果同时可以存出中间重要文件，软件还同时提供 用户自己作图修改的程序，可以传递参数控制最终图的效果。即主要达到如下几个效果

- 1 计算资源少，内存和时间 综合起来都比原来的好
- 2 计算后自动画图之后，还要支持自己画图，即可以通过参数来控制最终图
- 3 通用易用，可以直接读 vcf 格式，同时画图可以支持别的软件的 blocks 格式
- 4 更多功能，如将 GWAS 结果图示，即显示该区域的一些其它统计量等特征

即有 LDBlocksShow 相对其它两个软件在结果是一致的前提下，用到的计算资源更少，更容易使用。

| Software | Mem (G) | Cpu Time (h) | Result | Statistics |
|-------------------|---------|--------------|----------------------|------------|
| Haploview4.2.jar | 95 | 36 | heatMap+Block | R^2 |
| R: Ldheatmap 0.99 | 1 | 48 | heatMap | R^2 |
| LDBlocksShow 1.02 | 1 | 1 | heatMap+Block+(GAWS) | R^2/D' |

其中 LDBlocksShow 全新的新算法画图可以画到上万个位点，生成的文件极小。
算法介绍具体见后面软件解析。

下载安装

下载:

后期将会在 github 网址部署, 可以 `git clone` 等下载
<https://github.com/BGI-shenzhen/LDBlockShow/>

安装:

方法 1 linux/Unix 和 macOS 下的

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

****Note:**** 最后 link 失败的话,可以试重装安装 zlib 库(<https://zlib.net/>)

方法 2 linux/Unix 和 macOS 下的

```
tar -zxvf LDBlockShowXXX.tar.gz
cd LDBlockShowXXX;
cd src;
make ; make clean # or [sh make.sh]
../bin/LDBlockShow
```

****Note:**** 最后 link 失败的话,可以试重装安装 zlib 库(<https://zlib.net/>)

方法 3

我们提供了 linux/Unix 下 64 位的静态编译, 可以解压直接运行, 若如果存在安装麻烦的话, 可以直接联系我~

参数说明

除了主程序 LDBlockShow 之外, 我同时也提供多了一个 ShowLDSVG 程序, 用户若对 LDBlockShow 输出的结果做一些细节调整, 或者添加 GWAS 等其它信息, 可以再运行 ShowLDSVG 美化结果, 丰富图片结果

LDBlockShow:

简要参数

程序 LDBlockShow，其中程序只须要一个输入文件 **In.vcf** 和指定区域就可以计算并画出对应的 LDheatMap 图，很中画图的参数除了默认之外，还会根据文件的 SNP 数据自动优化。如下是程序 LDBlockShow 的简要参数，

```
[heweiming@cngb-ologin-25 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: R^2 2: D' [1] |
| -BlockType | <int> | Method to detect Block [beta] 1: Gabriel Method 2 Solid Spine of LD [1] 3: Block by Created plink |
| -help | | Show more Parameters and help [hewm2008 v1.20] |

| | |
|------------|---|
| -InVCF | 输入群体 VCF 格式，即入文件 |
| -Region | 输入一个指定的区间，即显示该区间的 LDHeatMap 图， 格式为 【chr:start:end】 中间用冒号隔开： |
| -OutPut | 输出文件路径 |
| -BlockType | 检测 LD blocsk 的方法，初步有三个选项， 第三种是调动外部 plinks 软件 |
| -SeleVar | 默认是计算 R^2 ， 用户可以选择为 D' |
| -help | 查看更多的参数说明，如过滤 SNP 的，如画图的 |

详细参数

```
./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 [-RegionOne],

| | | |
|------------|---------|--|
| -SubPop | <str> | SubGroup Sample File List[ALLsample] |
| -MAF | <float> | Min minor allele frequency filter [0.05] |
| -Het | <float> | Max ratio of het allele filter [0.88] |
| -Miss | <float> | Max ratio of miss allele filter [0.25] |
| -BlockCut | <float> | 'Strong LD' high confidence interval RR/D' cutoff for Block[0.90] |
| -TagSNPCut | <float> | 'Strong LD' high confidence interval RR/D' cutoff for TagSNP[0.98] |
| -Ratio | <float> | Ration of Gabriel Blocks for 'Strong LD' pairwise SNP [0.90] |
| -SBMinLen | <float> | Min Length for samall Blocks [10] |

| | |
|------------|--|
| -SubPop | 如果只用到一些样品(子群)来算, 则样品放在一文件, 以此参数传递 |
| -MAF | 过滤位点, 把低频的位点过滤掉, 默认是 0.05 |
| -Het | 过滤位点, 把高杂合的位点过滤掉, 默认是 0.88 |
| -Miss | 过滤位点, 把碱基缺失多的位点过滤掉, 默认是 0.25 |
| -BlockCut | 定义超连锁,强关联的两个 SNP 的 R^2/D' 大于某值, 默认是 0.90 |
| -TagSNPCut | 定义 Blocks 里面 TagSNP 的选择条件, 两个超强关联 SNP, 只取一个 |
| -Ratio | 检测 blocks 的一个标准, 当这个 blocks 里面的有 ratio 的比例是强关联 |
| -SBMinLen | 过滤长度太短的 LD blocks, 默认为 10, 即几乎不过滤。 |

ShowLDSVG:

简要参数

ShowLDSVG 主要对程序自动出来的图如果不满意, 可以手动传递参数以美化结果。如下是期简要的参数

```
./ShowLDSVG
```

Options

| | | |
|-----------|-------|------------------------------------|
| -InPreFix | <s> : | InPut Region LD Result Prefix |
| -OutPut | <s> : | OutPut svg file result |
| -help | : | Show more help with more parameter |

- InPreFix 输入文件，即就是 LDBlockShow 的输出文件
- OutPut 输出文件，SVG 结果，同时也提供 png 格式文件
- help 查看更多参数

详细参数

```
./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 Pvalue

-InGFF       <s>      : InPut GFF3 file to show Gene CDS and name
-NoGeneName  : No show Gene name,only show stuct

-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 stroke [white]
-WidthGrid   <s>      : the stroke-width of gird [1]
-NoGrid      : No Show the gird col
-ShowRR      : Show the R^2 in the heatmap
```

GWAS:

- InGWAS 输入文件，即要结果 gwas 的文件，联合上下作图，
格式为三列文件 (chr site Pvalue)
- NoLogP 对 Pvalue 值不进行 取 -log() 进行转换
- Cutline gwas 中的 cut off 线
- InGFF 输入 GFF 文件，即可以标出基因的结构和名字
(blue 为 CDS; Orange 为 intron yellow 为 UTR,上面有基因名)
- NoGeneName 不显示基因名，好显示结构

Cor:

- crBegin 开始颜色,对应弱关联($R^2=0$),默认为 白色
- crMiddle 中间颜色, 对应弱关联($R^2=0.5$),默认为 黄色
- crEnd 终止颜色,对应弱关联的($R^2=1$),默认为 红色

-NumGradien 开始颜色到终止颜色的渐变份数

Grid:

-CrGrid 网格边缘的颜色，默认为白色

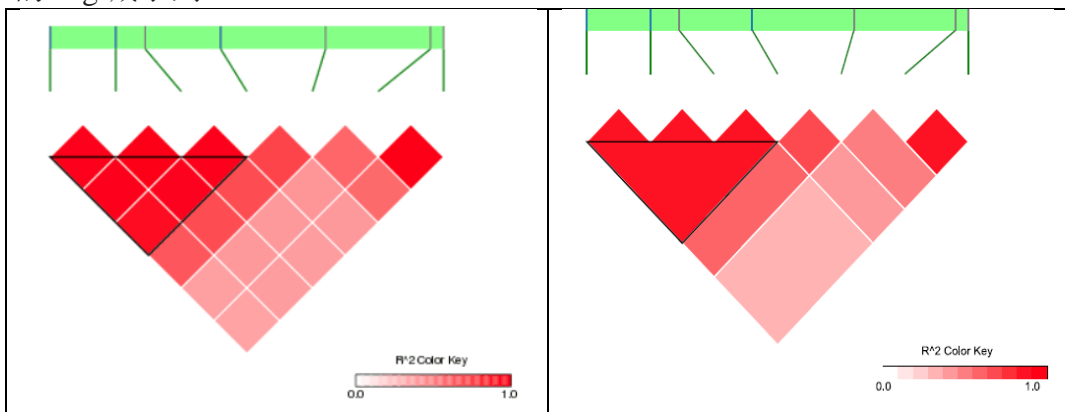
-WidthGrid 网格边缘线条的宽度，默认为1

-NoGrid 不显示网格边缘信息

-ShowRR : 在 SNP 数小于 50 时 加上此参数显示 R^2 的数字上图。

在 ShowLDSVG 大于 1.08 以上的版本，在位点在于 100 个 SNP 以上，ShowLDSVG

采用了全新的算法。具体之前每一个比对组 R^2 就是点。若 $1w$ 个位点，则倒三角的个数为 $1w*1w/2 = 5kw$ 。输出的 svg 极大。现新的算法为把周边同等颜色合并成多边形，合并成一个点后再输出的，此时 svg 会更小。如下的可以从 26k 的 svg 减小到 8k。



新的算法可能会花点时间在找周边共颜色的点 并成一个多边形（累递归+归并），但出来的点少，即生成的 svg 大大减小。其中 颜色区分度（0-1 默认按 100 分）参数调少点（如 0-1 按 10 等分），周边颜色块会更多相同，即可以合并更多位点到一个多边形里面，出来的图更小。

输出文件

，输入文件为群体的 VCF 变异文件，自然不须要介绍，对输出文件，程序为了让用户自己可以根据需求调整作图，美化图片，连同中间输出结果都输出来，如有如下几个文件。

| 输出文件 | 说明和格式 |
|------------------|--|
| out.site.gz | 为过滤后参与计算的 SNP 位点，格式为两列，【chr Site】 |
| out.blocks.gz | 为程序判断得到的 Blocks 结果文件，主要前面三列【Chr Start End】 |
| out.TriangleV.gz | 为两两之间的连锁程序 R^2/D' ，为矩阵的一半，倒三角 |
| out.svg | 为最终要看的图片，svg 为矢量图 |
| out.png | 为最终要看的图片，png 为位点图,以防 svg 过大打不开 |

| | |
|-----------|---|
| 输入文件 | 说明和格式 |
| In.vcf | VCF 格式，群体检测变异的结果文件 |
| gwas.pvlu | GWAS 的结果，格式主要有三列【chr Site Pvalue】，可选项，非必须 |

新功能介绍

- 1 可以 -lnGFF 把基因结构和名字画上
- 2 可以 -lnGwas 把这一区域的 gwas 信息画上
- 3 可以 -ShowRR 在 SNP 数据把 R^2 表上
- 4 添加了 -crMiddle，实现三种颜色的渐变
- 5 全新的画图算法，似递归找同相邻颜色块 + 找最大回路 算法。
可在大数据可以大大减少输出 svg 的大小，现可画数几万个位点。
- 6 可以直接调用 plinks 的生成 blocks 信息并画上图
- 7 可以选择 计算 R^2 还是 D' 。

经典实例

下面均都是以默认 -SeleVar 为 R^2 来计算，如果要用 D' 只须加上参数 -SeleVar 2 即可以替换为 D'

实例 1：倒三角+默认 LD Blocks

程序考虑为了大家易用方便使用，仅输入群体 VCF 格式和指定区域即可。
具体可以进入程序的 LDBlocksShow/example/Example1 里面查看，如下是代码

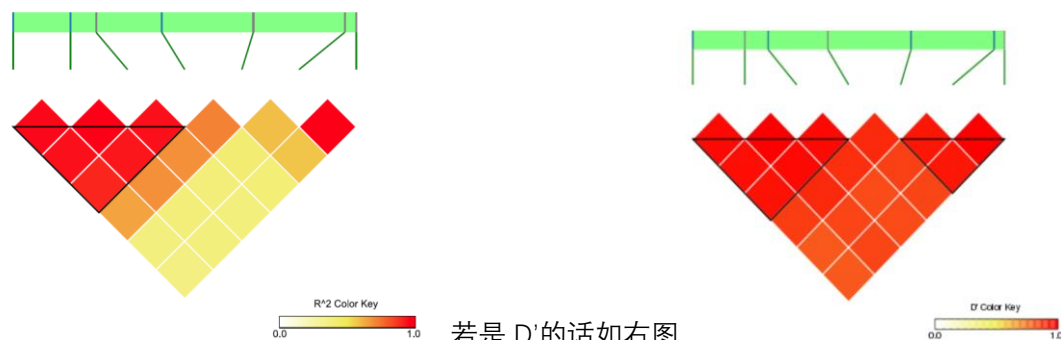
```
../bin/LDBlockShow -lnVCF Test.vcf.gz -OutPut out -Region Ghir_D11:24100000:24200000
```

如下截图：

```
[heweiming@cngb-ologin-25 Example1]$ ll
total 384K
-rwxr-xr-x 1 heweiming bc_pap 206 Feb 7 16:58 run.sh
-rw-r--r-- 1 heweiming bc_pap 377K Jan 20 16:02 Test.vcf.gz
[heweiming@cngb-ologin-25 Example1]$ cat run.sh
#!/bin/sh
#$ -S /bin/sh
#Version1.0 hewm@genomics.cn 2020-01-10
echo Start Time :
date
../bin/LDBlockShow -lnVCF Test.vcf.gz -OutPut out -Region Ghir_D11:24100000:24200000
echo End Time :
date
[heweiming@cngb-ologin-25 Example1]$ sh run.sh
Start Time :
Tue Feb 18 13:40:21 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 :
Tue Feb 18 13:40:22 CST 2020
[heweiming@cngb-ologin-25 Example1]$ ls
out.blocks.gz out.png out.site.gz out.svg out.TriangleRR.gz run.sh Test.vcf.gz
[heweiming@cngb-ologin-25 Example1]$
```

说明:

只要运行就可以得到上面列的 5 个文件，其中 out.svg/out.png 为最终要的图

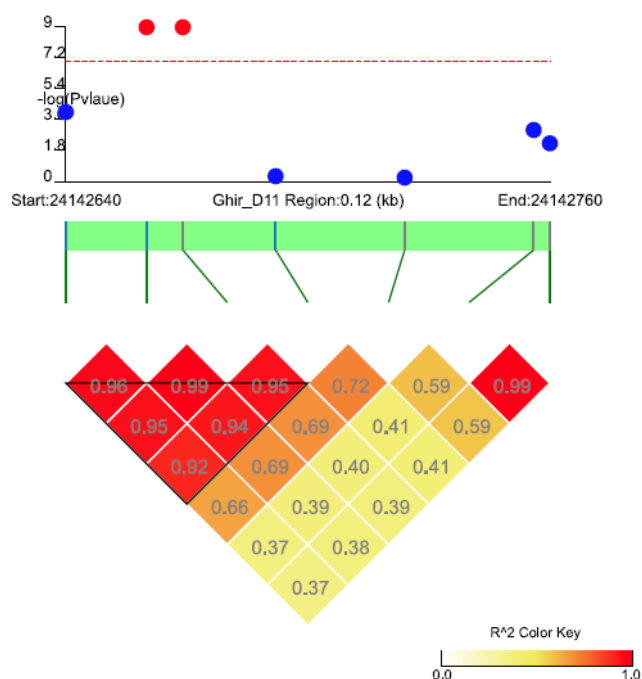


实例 2: 倒三角+默认 LD Blocks + GWAS

如上面实例 1 虽然得到倒三角解图和 LD blocks 的结果，但常还要加上一些 GWAS 的信息上去，那么只要把 gwas 格式整理成三列【chr site Pvalue】，即可以用-InGWAS 参数传。具体可以进入程序的 LDBlocksShow/example/Example2 里面查看，如下是代码

```
#../bin/ShowLDSVG -InPreFix ../Example1/out -OutPut out.svg -InGWAS gwas.pvalue  
../bin/ShowLDSVG -InPreFix ../Example1/out -OutPut out.svg -InGWAS gwas.pvalue -Cutline 7
```

如上只用到 ShowLDSVG 重新画图即可，得到的最图同共是是 out.svg 和 out.png



实例 3: 倒三角+plinks LD Blocks

本程序找 LD block ,相对 Haploview 和 plinks 在大的 blocks 差别很大, 但碎的 LD blocks 均存在一点差别, 目前本程序检测 LD blocks 的算法和细节取取后期调整改进, 期望达到 plinks 的结果一模一样的结果, 此外由于 plinks 在计算 LD blocks 十分快, 本程序已经兼容其 LD blocks 的结果, 只须须要把结果替换进去, 同时重新画画一下图。具体可以进入程序的 LDBlocksShow/example/Example3 里面查看。

在 1.06 之后 可以直接添加 **-BlockType 3** 即可以直接调用 plinks 1.9 生成 block 信息。

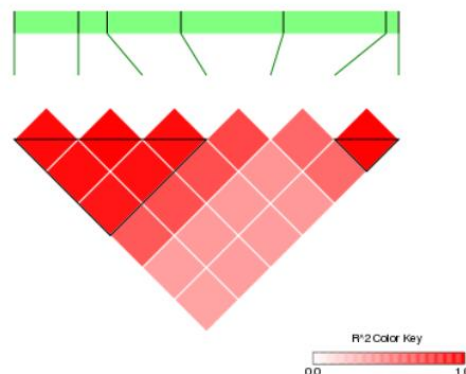
具体可以进入程序的 LDBlocksShow/example/Example1 里面查看, 如下是代码

```
../../bin/LDBlockShow -InVCF Test.vcf.gz -OutPut out -Region Ghir_D11:24100000:24200000 -BlockType 3
```

```
## 这是以前建议的做法,  
#./plink-1.9/plink --vcf ../Example1/Test.vcf.gz --geno 0.1 -maf  
0.01 --blocks no-pheno-req --out plink --allow-extra-chr  
#cp ../Example1/out.* ./;  
#rm out.blocks.gz ; mv plink.blocks.det out.blocks ; gzip out.blocks ; rm plink.*  
#../../bin/ShowLDSVG -InPreFix ./out -OutPut out.svg
```

如上将 block 结果文件件替换换一下, 然后后 ShowLDSVG 重新画图得到的最图

同共是 out.svg 和 out.png。 仅 blocks 有不一致而已。



实例 4: 倒三角+plinks LD Blocks + GWAS

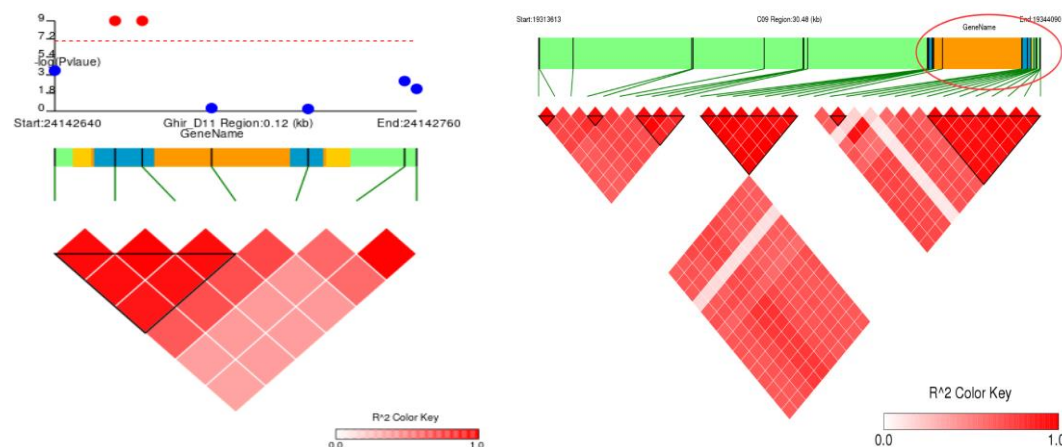
把实例 2 的输入文件 (即实例 1 的)切换为 (即实例 3 的)的输入文件即可, 输出的结果结果和实例 2 类同, 在这不过过多显示, 具体进入 LDBlocksShow/example/Example4 查看

实例 5: 倒三角+默认 LD Blocks (+ GWAS)+GeneStuct

如上面实例 1 虽然得到倒三角解图和 LD blocks 的结果，但常还要加上一些基因的名字和结构上去，即可以用-InGFF 参数传 gff 格式文件。具体可以进入程序的 LDBlocksShow/example/Example5 里面查看，如下是代码

```
../bin/ShowLDSVG -InPreFix ../Example1/out -OutPut out.svg -InGWAS gwas.pvalue -Cutline 7 -InGFF In.gff
#../bin/ShowLDSVG -InPreFix ../Example1/out - OutPut out.svg -InGFF In.gff
```

如上只用到 ShowLDSVG 重新画图即可，得到的图最终同是 out.svg 和 out.png



如上 blue 为 CDS; Orange 为 intron , yellow 为 UTR,上面有基因名 GeneName.

常见问题

本软件使用起来十分方便，对输入文件没有过多，只须要提供一个群体的 SNP 文件（VCF 格式）和指定区域的就行，总体来说使用十分方便，占用计算资源也十分少。

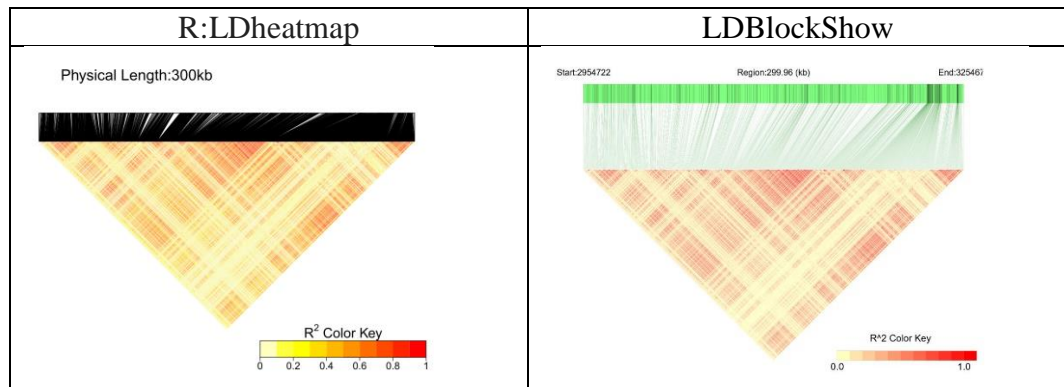
问题 1 结果准确性

程序计算如此快，是否结果有问题，是否倒三角和 LDheatmap 的是一致的。

答：逻辑存在问题，并不是计算越久就越好，plink 算 LD blocks 时速度相当快，结果也和 Haploview 一致。

- 1 本程序经过和 Haploview 和 R:LDheatmap 比较，其倒三角的数据均是一致的，即结果完全无误，如下是用户用了 LDheatmap 的图，和本程序画的倒

三图比较，均无两异，如下所示



- 2 至于 LD block，目前软件找到的 ld blocks 均相对 plink 和 Haploview 在默认下大致是一致的，特别是大的 block 是对的，可能是碎的 blocks 存在小差别

问题 2 显示 region 的其它 Pi 可否？

这一个可以 -InGwas 的第三列 Pvalue 值用为你自己的数据即可，加上-NoLogP 即可，对不对这一值进行-log 置换，和实例 2 运行即可。

若还要同时显示更多其它特性的话，可能要对代码进行小修改，在这若有更多需求的话，后面可以考虑添加这一功能
步骤中，主要的功能就是画图，根据各种组合各种需求各种情景画的

Reference^{1,2}

1. Zhang, C. *et al.* PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* (2018).
2. Evans, L.M. *et al.* Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet* **46**, 1089-96 (2014).