LDkit用户使用手册

2019-03-23

目录

[1.Introduction 3](#_Toc3696595)

[2.环境搭建 3](#_Toc3696596)

[3.模拟数据 3](#_Toc3696597)

[4.软件功能操作 4](#_Toc3696598)

[4.1图形化功能操作 4](#_Toc3696599)

[4.1.1选取文件/文件夹 4](#_Toc3696600)

[4.1.2Panameten参数 5](#_Toc3696601)

[4.1.3LD type参数 5](#_Toc3696602)

[4.1.4Method参数 6](#_Toc3696603)

[4.1.5Plot程序 6](#_Toc3696604)

[4.1.6开始运算 7](#_Toc3696605)

[4.2命令行功能操作 7](#_Toc3696606)

[4.2.1 LD计算程序参数说明 7](#_Toc3696607)

[4.2.2 LD画图程序参数参数说明 8](#_Toc3696608)

[4.2.3 LD计算程序最简化命令行示例 8](#_Toc3696609)

[4.2.4 LD计算程序可选参数命令行示例 9](#_Toc3696610)

[4.2.5 LD画图程序命令行示例 10](#_Toc3696611)

[5.结果文件展示 11](#_Toc3696612)

[5.1一种染色体一种亚群体 11](#_Toc3696613)

[5.2一种染色体多种亚群体 11](#_Toc3696614)

[5.3多种染色体一种亚群体 12](#_Toc3696615)

[5.3多种染色体多种亚群体 13](#_Toc3696616)

[5.4 LDblock运算结果文件 13](#_Toc3696617)

[5.4 LDsite运算结果文件 14](#_Toc3696618)

# 1.Introduction

软件英文简单介绍

# 2.环境搭建

需要在操作系统线下配置Java环境变量，首先安装JDK1.8以上版本。根据操作系统具体下载地址：http://www.oracle.com/techneldork/java/javase/downloads/jdk8-downloads-2133151.html。

# 3.模拟数据

给定的三组测试数据test.vcf.gz ; test.vcf ; text.ped,text.map ; pop.txt

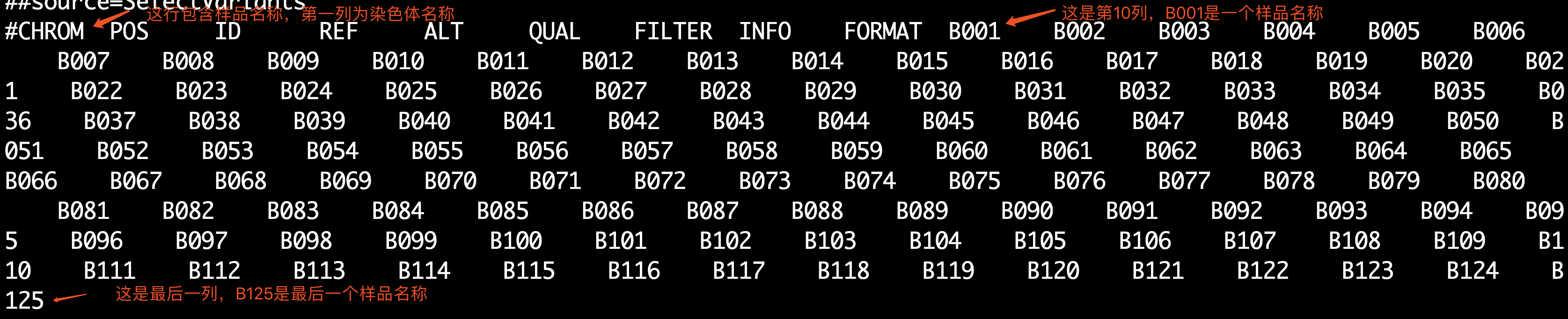
主要介绍一下gz /vcf格式的特点，还有ped map文件的特点

gz文件:其中gz是一种压缩格式，文件构成和vcf文件构成一样。

vcf文件：

##开始的都是注释文件；

#C开头的一行包含了样本信息，前9列为说明信息，第10列开始为样品名称。



#C后面的行数，每一行代表一个SNP信息。第10列为第一个SNP的类型。每个SNP有四种类型：0/0； 0/1； 1/1和 ./. 。 在计算相关性的时候，分别用0，1和2表示。./.开始的SNP为缺省值；

ped文件:每行表示一个个体，第一列到第六列分别为家庭ID,个体ID,父亲ID,母亲ID,性别，性状；从第七列开始每两列为一个位点的基因型（A A或者3 3）

map文件：每行一个标记，第一列到第四列分别为染色体，SNP ID，连锁距离，物理距离

pop.txt: 亚群体文件第一行为种群名，种群名和样本信息之间用冒号隔开，样本信息之间用逗号隔开。



# 4.软件功能操作

Windows执行图形化：双击LDkid.jar

Linux执行图形化是：Java -jar LDkid.jar

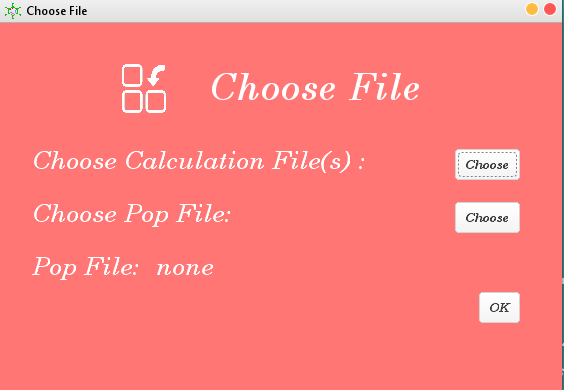
Mac执行图形化是，双击LDkid.jar

## 4.1图形化功能操作

### 4.1.1选取文件/文件夹

1.软件支持拖拽的方式选取文件，只需要将要计算的文件拖拽进软件界面，软件将自动识别可计算的文件（gz格式，vcf格式，ped格式和map格式）并在Input File界面显示出来（可同时拖拽多个文件）

2.选取文件夹的步骤和选取文件的步骤是一样的，将要计算的文件放入到文件夹下，将文件夹拖入软件，软件就会自动识别并添加要计算的文件

 3.点击软件上方红色区域（Choose File），软件将弹出文件选择框。

点击Choose Calculation File(s) ，就可以用文件选择器选择要计算的文件/文件夹了。（可同时选取多个文件。）

点击Choose Pop File，即可用文件选择器选择亚群体文件。

注意：

文件的选取一次只能选取同一种类型，即如果选取了gz格式的文件，则不能选择vcf文件和ped/map文件。否则计算将会出现错误。同理，如果选取了vcf格式的文件，则不能选择gz文件和ped/map文件。同理，ped/map文件也是一样的。

gz格式文件和vcf文件都支持同时选取多个文件，并且一起运算，但是ped/map文件只能选取一个进行计算，否则将会出错。

ped文件和map文件必须同时选择，成对出现。

Pop文件不受于上面的文件类型限制，可以同时选取。

### 4.1.2Panameten参数

点击主页面上的Panameten，即可进行Panameten参数的设定。

Windows Size：在给定的区间内算100r2。只能输入数字。默认值100

Missing Rate：缺失值的比例。输入1-0之间的小数。默认为0.2

Thread num:计算速度。默认为系统CPU核数的一半

MAF: 输入1-0之间的小数。默认为0.005

Result file name：结果文件名。默认为result

Result file Path： 结果文件输出地址。默认为为result

Chrom: 要计算的染色体。点击all计算全部染色体。点击input可以自定义要计算的染色体。用逗号隔开。默认为计算全部染色体

### 4.1.3LD type参数

点击主页面上的LD type即可进行LD type参数的设定。

默认计算方法为LD decay。

如果选择LD decay，则不用输入下方的参数。

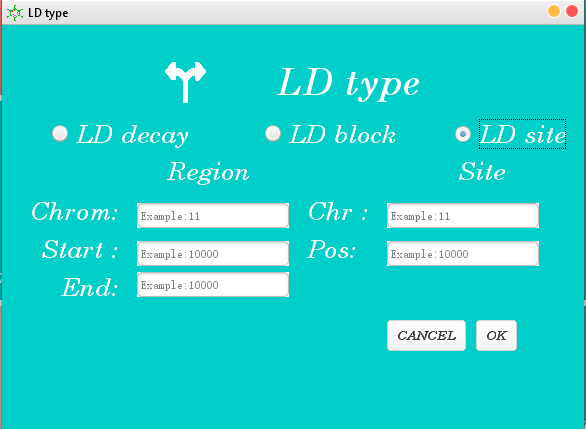
如果选择LD block则需要输入Region参数Chrom: 要计算的染色体

Start:要计算的SNP范围的开始值。

end:：要计算的SNP范围的结束值。

Intermediate:是否生成中间文件。默认是不生成

如果选择LD Site，则需要输入Region和Site参数

Chrom: 要计算的染色体

Start:要计算的SNP范围的开始值。

End:：要计算的SNP范围的结束值。

Chr：要计算的指定行的染色体。

Pos：要计算的指定行的SNP的具体值

### 4.1.4Method参数

点击主页面上的Method，即可进入Method参数的设定。

点击r-square，即可调用r-square方法进行计算。

点击D prime，即可调用D prime方法进行计算。

默认为r-square方法。

### 4.1.5Plot程序

点击主页上的Plot即可进入Plot程序

InFile:输入要进行画图的文件（由本软件输出）

Marge：是否合并不同的亚群体

Marge：是否合并不同的染色体

Bin：作图区间。默认是1kb

Result name：输出的结果文件名

Start: 点击Start开始运算

### 4.1.6开始运算

点击主页面上的Start即可开始运算。

运算只需要选取文件就可以开始。主页面下面的三个参数属于自定义参数，可以不进行选取。

运算进行时，主页面上的右侧示例图标将会进行转动，同时软件界面不允许鼠标进入。

## 4.2命令行功能操作

执行本程序命令行示例：（要求必须在英文状态下输入）

Java[空格]--jar[空格][软件地址][空格][参数名][参数数据][空格][参数名][参数数据]

### 4.2.1 LD计算程序参数说明

--infg 要计算的文件(gz格式)。可输入多个文件，用逗号隔开

--infgs 要计算的文件夹(gz格式)。只能输入一个文件夹。程序会自动读取该文件夹下的所有可计算文件

--infv 要计算的文件(vcf格式)。可输入多个文件，用逗号隔开

--infvs 要计算的文件夹(vcf格式)。只能输入一个文件夹。程序会自动读取该文件夹下的所有可计算文件

--inped要计算的ped文件。输入时候不要带文件名的后缀名。

--out 计算完成输出的文件夹。默认为要计算的文件的同级目录下。

--ws windowSize，在给定的区间内算100r2 默认为100

--subpop 要计算的亚群体文件。

--chr 要计算的染色体。可以输入多个染色体，用逗号隔开。默认为计算全部染色体

--missing 缺失值的比例。输入1-0之间的小数。默认为0.2

--threads 运算线程数。默认为1

--maf 默认为0.005

--ld 计算方法。1为LD decay，2为LD block，3为LD site。默认为1

--type 计算方式。1为r-square，2为D prime。默认为1.

--Intermediate 当计算方法为Ldblock的时候，是否生成中间文件（可能会很大）默认为no

--chra当计算方法为LD block或者LD site时，用于表示要计算的染色体

--start 当计算方法为LD block或者LD site时，用于表示要计算的SNP起始值

--end 当计算方法为LD block或者LD site时，用于表示要计算的SNP结束值

--chrb 当计算方法为LD site时，用于表示要计算的指定SNP的染色体

--siteline 当计算方法为LD site时，用于表示要计算的指定SNP的具体值

--h 输出帮助文件

注意：红色参数为必选参数，并且只能有一个必选参数，不允许出现多个必选参数。

黑色为可选参数，用户可根据需求自主确定是否输入可选参数。可选参数可以输入多个。

### 4.2.2 LD画图程序参数参数说明

--plot 程序进入画图程序

--inp 作图的输入文件（文件由LD计算程序输出）

--merge 是否合并多种亚群体。yes表示合并，no表示不合并。默认为yes

--mergechr 是否合并多种染色体。yes表示合并，no表示不合并。默认为yes

--bin 画图区间大小。建议输入0.1- windowSize之间的数。最小不要小于0.1。默认为1

--outp文件输出名称。默认为result

### 4.2.3 LD计算程序最简化命令行示例

|  |  |
| --- | --- |
| Windows | java -jar LdkitC.jar --infg C:\data\test.vcf.gz  java -jar LdkitC.jar --infv C:\data\test.vcf  java -jar LdkitC.jar --infgs C:\data\  java -jar LdkitC.jar --infvs C:\data\  java -jar LdkitC.jar --inped C:\data\test |
| Linux/Mac | java -jar LdkitC.jar --infg /mnt/data/test.vcf.gz  java -jar LdkitC.jar --infv /mnt/data/test.vcf  java -jar LdkitC.jar --infgs /mnt/data/  java -jar LdkitC.jar --infvs /mnt/data/  java -jar LdkitC.jar --inped /mnt/data/test |

**以上命令等同于：**

|  |  |
| --- | --- |
| Windows | java -jar LdkitC.jar --infg C:\data\test.vcf.gz --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infv C:\data\test.vcf --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infgs C:\data\ --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infvs C:\data\ --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --inped C:\data\test --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005 |
| Linux/Mac | java -jar LdkitC.jar --infg /mut/data/test.vcf.gz --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infv /mnt/data/test.vcf --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infgs /mnt/data/ --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --infvs /mnt/data/ --ws 100 --missing 0.2 --threads 1 --ld 1 --type 1 --maf 0.005  java -jar LdkitC.jar --inped /mnt/data/test --ws 100 --missing 0.2 --threads 1 --ld 1 --type1 --maf 0.005 |

### 4.2.4 LD计算程序可选参数命令行示例

(1)指定计算完成的文件的输出位置

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --out C:\data\

(2)指定计算的WindowsSize

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --ws 100

(3)指定计算的亚群体文件

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --subpop C:\data\pop.txt

(4)指定计算的染色体

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --chr 5,8,22

(5)指定计算的缺失比例

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --missing 0.2

(6)指定计算运行的线程数

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --threads 5

(7)指定计算的maf

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --maf 0.008

(8)指定计算方式

Java -jar LdkitC.jar --infg C:\data\test.vcf.gz --type 2

(9)指定计算方法

java -jar LdkitC.jar --infg C:\data\test.vcf.gz --chra 11 --start 13884401 --end 13889000 --ld 3 --chrb 11 --siteline 13884401

java -jar LdkitC.jar --infg C:\data\test.vcf.gz --chra 11 --start 13884401 --end 13889000 --ld 2

(10)Ldblock计算方法生成中间文件（可能会很大）

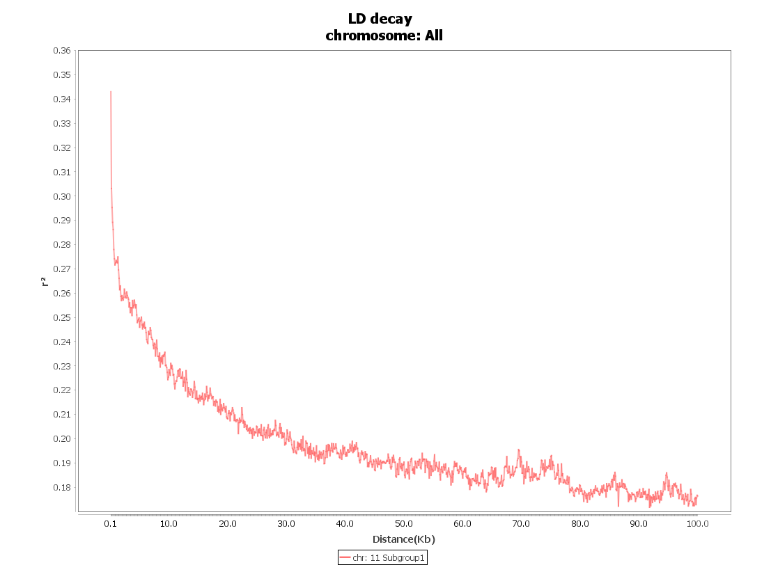
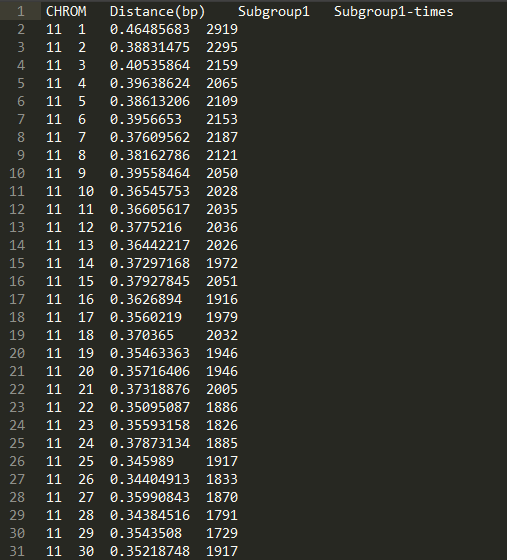
java -jar LdkitC.jar --infg C:\data\test.vcf.gz --chra 11 --start 13884401 --end 13889000 --ld 2 --Intermediate yes

### 4.2.5 LD画图程序命令行示例

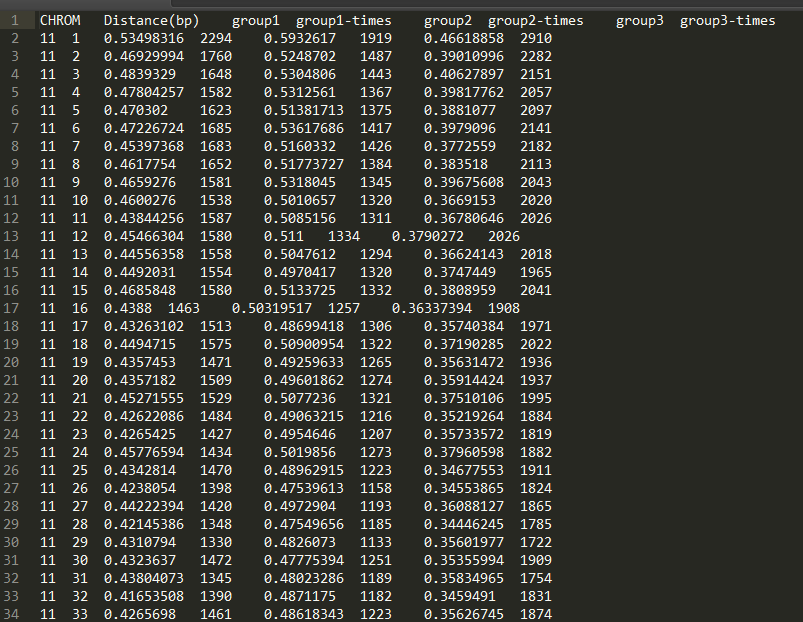
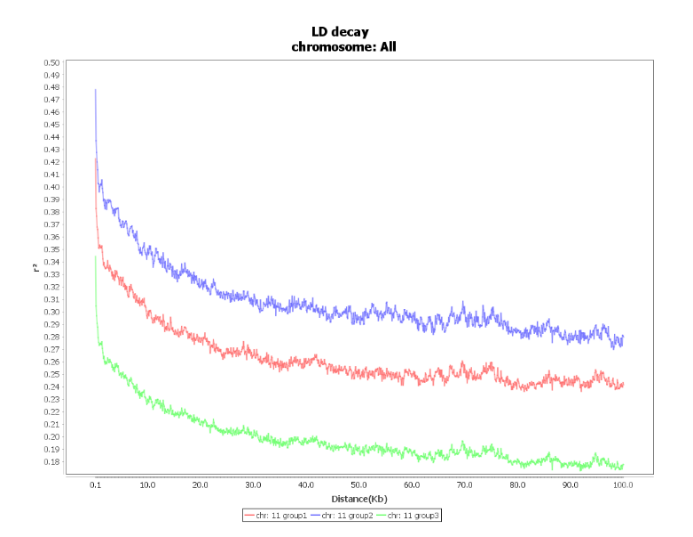
|  |  |
| --- | --- |
| Windows | java -jar LdkitC.jar --plot --inp C:\data\result.txt  java -jar LdkitC.jar --plot --inp C:\data\result.txt --merge yes --mergechr yes --bin 1 --outp result |
| Linux/Mac | java -jar LdkitC.jar --plot --inp /mut/result.txt  java -jar LdkitC.jar --plot --inp /mnt/result.txt --merge yes --mergechr yes --bin 1 --outp result |

# 5.结果文件展示

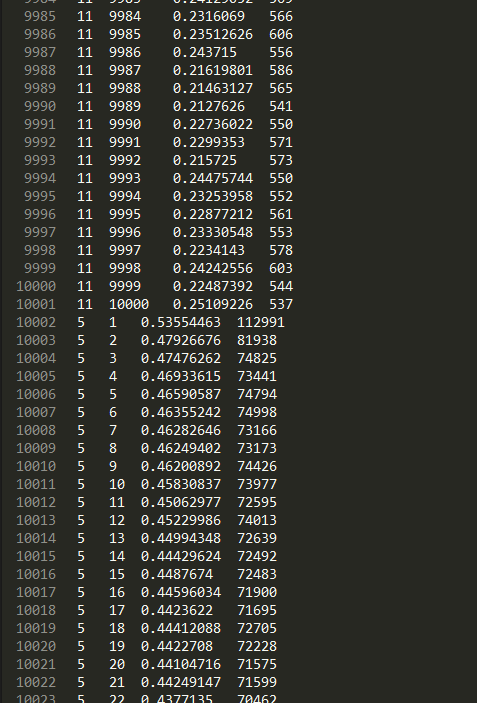
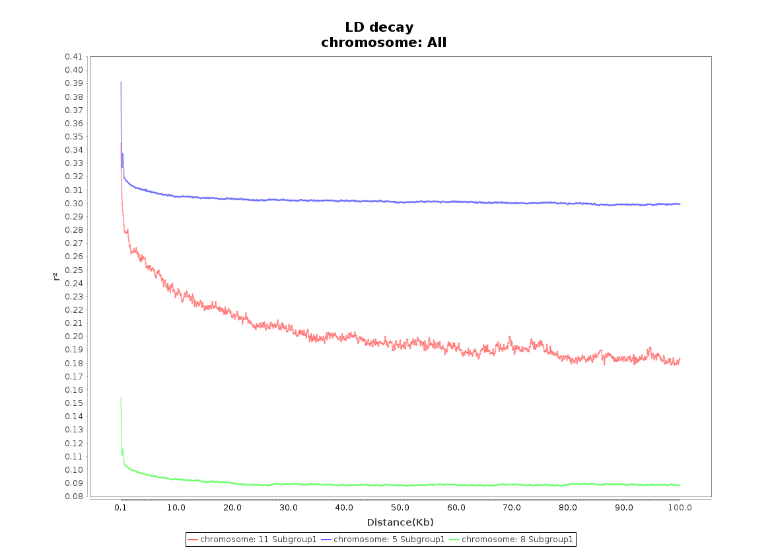
## 5.1一种染色体一种亚群体



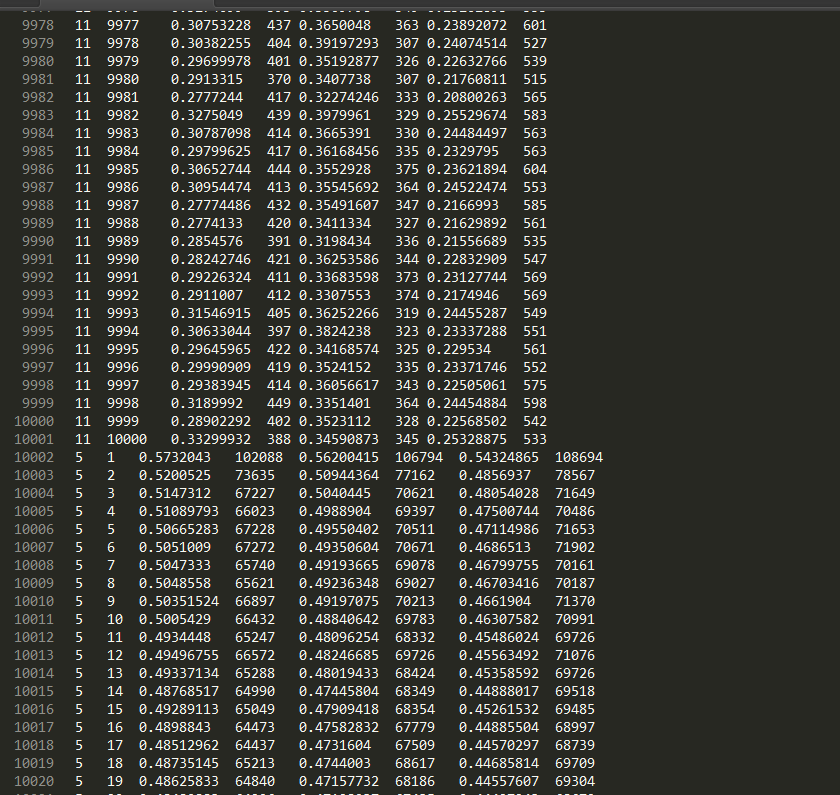
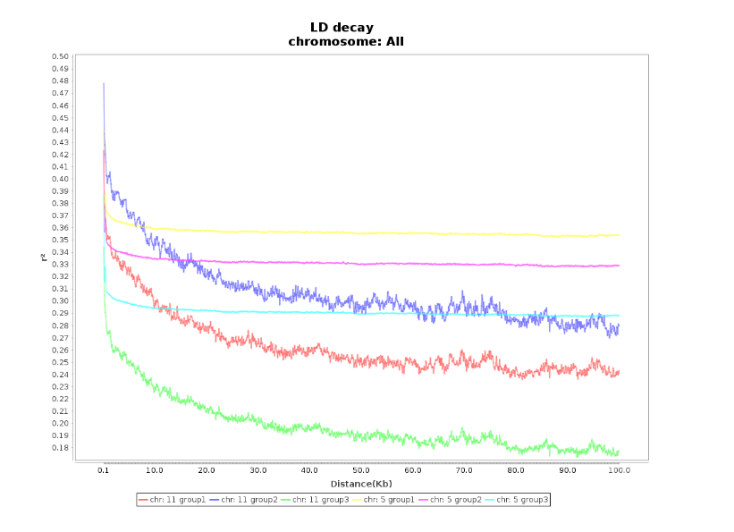
## 5.2一种染色体多种亚群体

## 5.3多种染色体一种亚群体

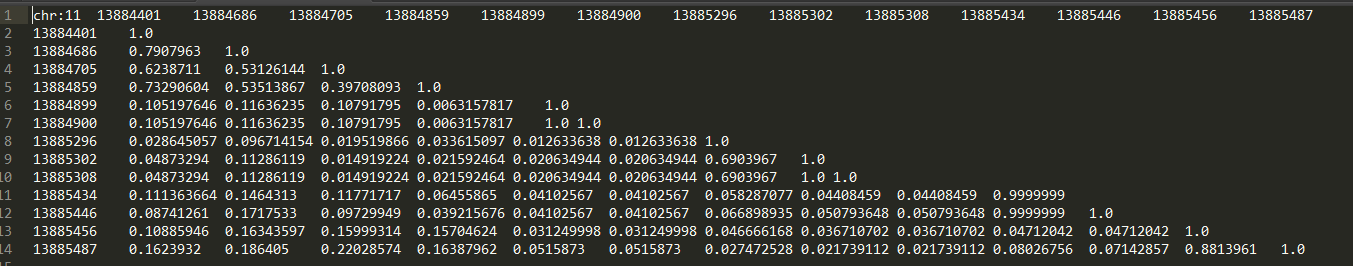
 

## 5.3多种染色体多种亚群体

## 5.4 LDblock运算结果文件





## 5.4 LDsite运算结果文件

