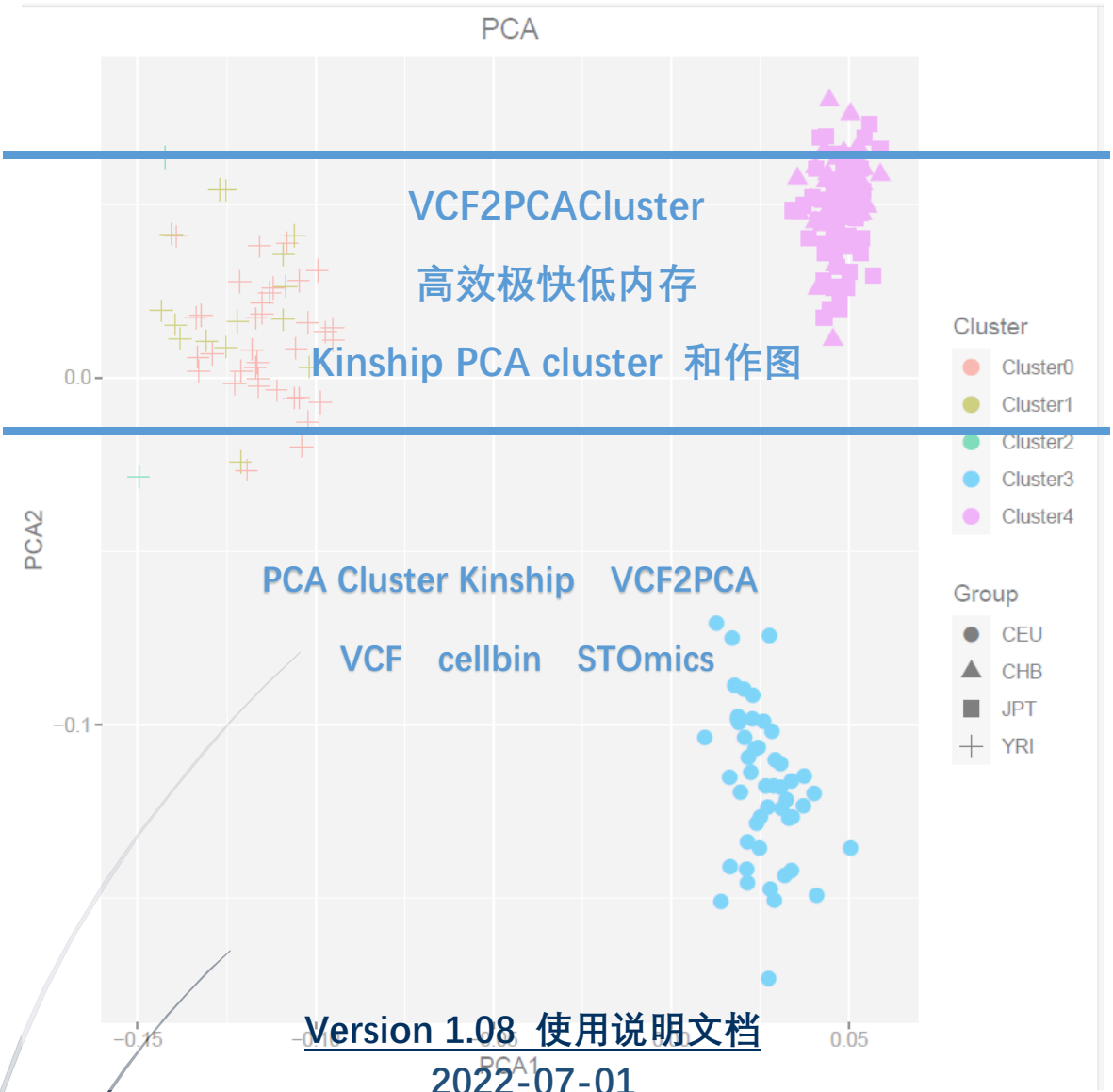
-

使用手册

# MingPCACluster




Version 1.08 使用说明文档

2022-07-01

hewm2008@gmail.com / hewm2008@qq.com

微信 打赏　　QQ 入群: ***125293663***　　***微信公众号***

# 目录

# 1.简介

MingPCACluster 是于基于 VCF 开发的 **PCA 分析**和**聚类**软件，同时兼并了 Genotype 等格式软件，同时开发针对时空单细胞表达量的格式（xx.gem.gz）文件（beta 功能）。
即只要对应的一个输入文件进来，这 PCA 和作图分组等一位到位。

主要亮点：
A:
一步高效生成 PCA 和聚类图。其中为了强调核心是高效**低内存**，一步操作，一个输入 到位 PCA 结果和图，对用户友好。

简易强调说明：
1 结果和 tassel gapit 和 gcta 这三个软件比对，**结果是一样，仅精度差别**。
2 功能有 1 多种 kinship 矩阵 2 PCA 结果 3 聚类结果 和 4 以 cluster 染色并作图。
3 一个 VCF 输入，一步到位，方便用户使用.
4 边读边算，内存剥离受位点多少的影响（时空组是剥离受基因数量多少的影响），内存只受样品量影响，故上 100k 的样品当也行，在这个基础上开发时空细胞 PCA 和聚类,虽然时空组学上主要是样品多。（80K 60G 内存）
5 Kmean 聚类分析，并找出最佳 K 值，和 Structure 和 K 值一样. 作图以此染色。
6 提作作图小脚本，可以用这个脚本优化作图等。

程序提供了 两个应用场景实例之后，重测序千人 VCF 和时空 cellbin 的细胞聚类 pca

**知乎专栏**也收集了一些用户的配置和说明。

## 各版本新功能和计划

A. 初放版本，大家试用 有 bug 联系和更的需求有空再丰富
B. 作图初步依赖于 R 和 ggplot 包，**后面看须要可以开发独立的 svg 画图功能和软件**
C. .

# 2.应用场景示例

初步开发主要针对群体重测序和时空单细胞表达量 pca 分析和聚类分析。
其中其它只要造相应的数据作为输入，即可以分析
-

## 2.1 千人 VCF 重测序 SNP 基因型

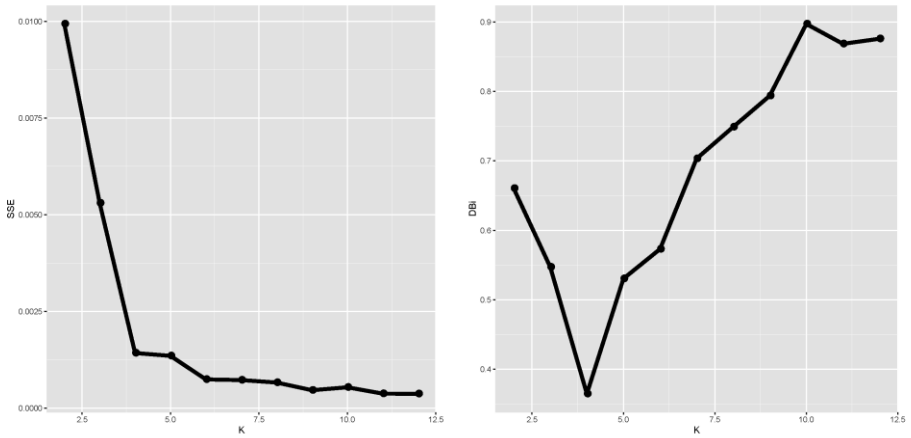
共从 K 人数据 chr22 dbsnp 里面随机挑出了 1194 个位点，挑 CEU（49），CHB（46），JPT（56）和 YRI （52）共 **203** 个样品来分析。

用法如下： 结果 pca 和用 gatc 的结果是一致的，其中 CHB 和 JPT 分不开的原因极可能是位点太多，这儿不深研究，已经确保同一数据，pca 的结果是相当的了。 那么我们这的

*MingPCACluster -InVCF Khuman.vcf -OutPut OUT*

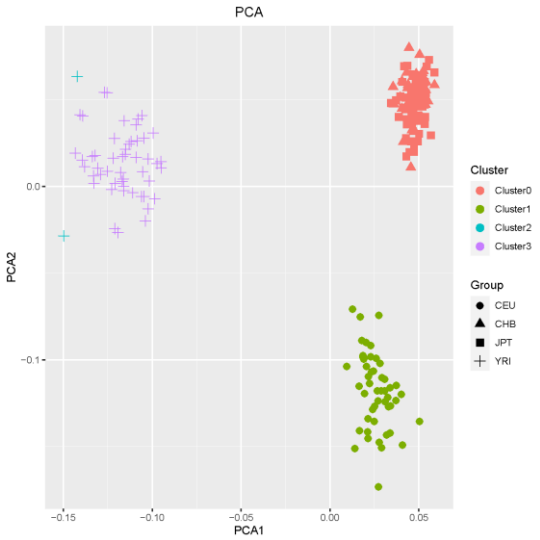
其中可以用 **-InSampleGroup** 输入样品和分群信息看和聚类结果是否一致。
MingPCACluster *-InVCF Khuman.vcf -OutPut OUT* ***-InSampleGroup*** *pop.info*

文件 pop.info 的格式为两列，第一样为样品名，第二列为分群名。
结果： 软件根据 K 和 SSE 的关系，认为最佳的聚类为 K=4, 4 后面之后**就平缓了** or 高了，软件可以取 best K=4. 软件默认输出了 K3-12 的结果，用户若认为聚类不好，要用 K=3 可以从文件 OUT.cluster 里面挑 K=3 的结果出来即可。




其中 best K 一般认为 是 SSE **第一次开始平滑 变缓**那时对应的 K. 并不是最小值 ，在 1.07 的版本后更新多计算 Dbi 的功能，软件根据 K 和 Davies–Bouldin index (DBI)的关系**,DBi 最小值**就是 best K. 其中 SSE 和 DBi 的相关说明和介绍可以点击打开查看。

那么 K=4 的结果和 pca 画图如下：




如上，主要多出来的两个聚类主要是在 YRI,相当 wild 种 聚类看起来较乱（其实当在 pca3 时 YRI 能分出来）。若要手动把 K=3 拿出来重新画图（软件提供了 ploteig 画图软件，有很多参数，对默认的图不满意，可以单独重新用这个软件快速作图），脚本如下 ：

*perl aa.pl OUT.cluster OUT.eigenvec NewOUT.eigenvec*

*perl ../../bin/ploteig -InPCA NewOUT.eigenvec -OutPrefix NewFig*

（aa.pl 见 example1）

另外 **ploteig** 有 **-ColShap** 的参数，可以把上面的 颜色和 形状 **对换**一下。

有时边缘点在聚类分析中有时因为初始散点质心而有所影响，即找 best K 时会有所变化，在 1.08 后的版本提供了可以 repeat 多次聚类的参数（***-RandomCenter***）查看聚类结果。用户可以多次使用 看看各种的效果。

./MingPCACluster -InKinship in.BaldingNicols.Kinship -RandomCenter -OutPut NewOUT

## 2.2 cellbin 时空细胞表达量 pca 和聚类（beta）

我的软件内存主要受样品量 n 的影响，n*n . 见常问题 1,

时空分析我初了解主要是：seurat ，我很浅淡的了解，这个包用到的 n*m (n 是样品，m 是位点)的稀疏矩阵，好像周边的做时空的人总说内存很大，我这没有对时空数据敏感，对表达量进行了取 log10. 也用了稀疏矩阵 和 n*n，由于时空 n 是样品量很大，怕难下降。

初以 我这用了文件大于(File.gem.gz：177M )， 范围： XXmin: 4975 XXmax: 23374 YYmin: 2525 YYmax: 20724 )。取 bin 50， n 达到的 88507，即主要 88507*88507 的矩阵 double 上，占用 60.742G (稀疏矩阵 5G 矩阵:55G) 。

另外 2w*2w 的 内存约为 2G， 时间后面再优化 等

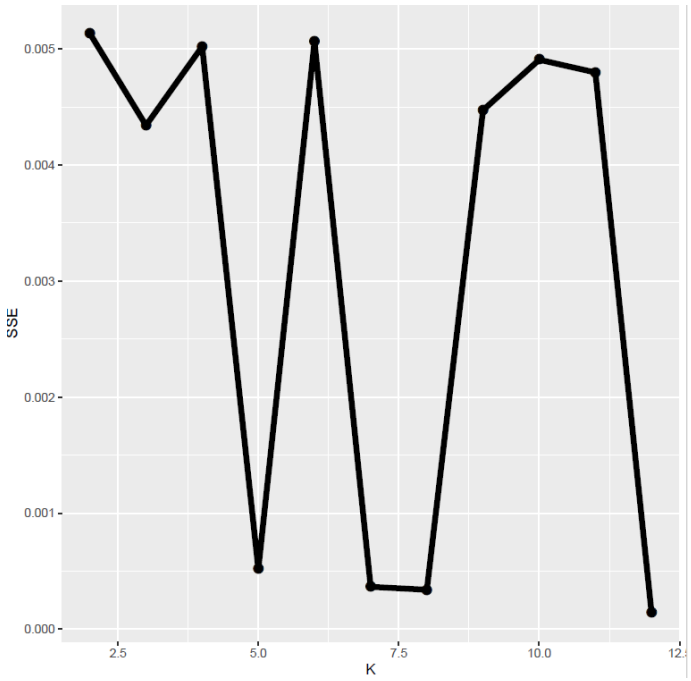
额外话：
但我这聚类结果和 pca 还没有和 seurat 等其它比较差异和准确和时间存等，建议有人来测和处理，速度时间暂时估计 beta 版本 ，可能不会快多少，内存可能会降低一点，但不会低太多，最近回家开发这软件( 家人对我意见很大，怕是没有多的时间弄这，即使弄会很慢，后面有空再优化更新这儿的功能)。 这儿自己还不是很不知道详细比较信息，而且估计还没有绝对的质的提高，故记为 beta 功能

直接上脚本和结果：用法如下：

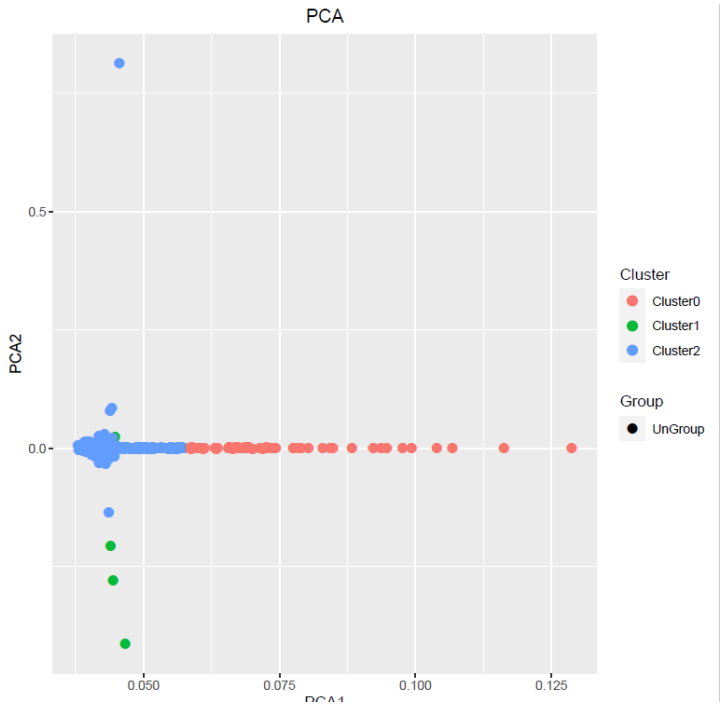
MingPCACluster -InSTOgem Test.gem.gz -OutPut Test -CellBin 500

下面是测序可以跑通，取 bin=500，数据和结果 准确性还盼其它高手测试之

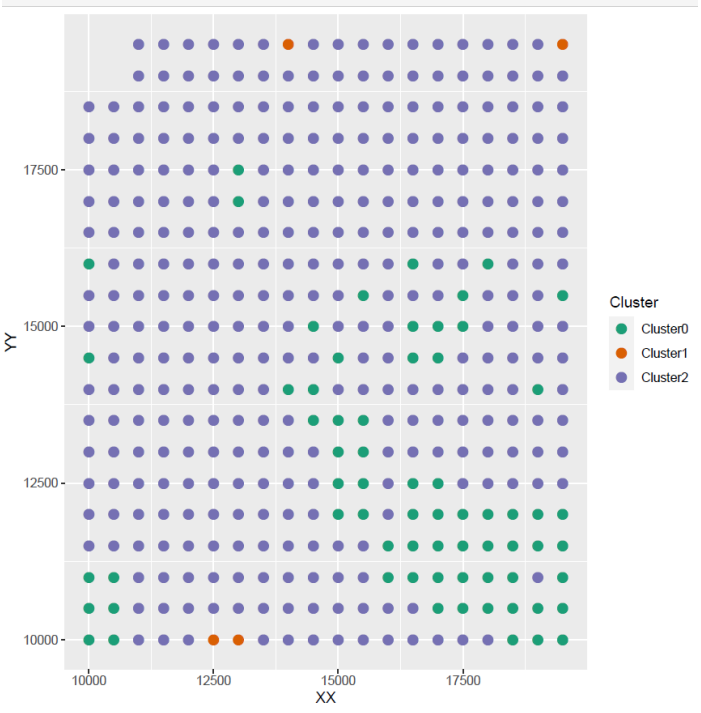
软件以 K=3 认为最优




PCA 1-2 维图：




对应坐标用 cluster 染色如下:




绿的偏右下　　红的两边几个

# 3. 下载与安装

## 3.1 下载网址

这主要是防止大家还在用 beta 版本，还在用可能存在 bug 的程序，强迫***大家定期更新***。
/hwfssz4/BC_PUB/Software/08.Centos7/MingPCACluster-*

里面有 example1-4

**持续更新**在这：https://github.com/hewm2008/MingPCACluster 下载时记得***加个星星***哦

## 3.2 预先安装

MingPCACluster 适用于 Linux/Unix/macOS 系统。在安装之前，请先安装以下使用条件：
1) R 和 ggplot2 包，（后期有空将看要不要独立来开）
2) convert: 系和 ps2pdf 系统命令其中一个，可以将 svg 转 png。莫有也无关系，有则更佳
3 ) g++ : g++ with --std=c++11 > 4.8+ is recommended

## 3.3 安装

使用者可采用以下直接 chmod 755 运行：
1)

```
git clone https://github.com/hewm2008/ MingPCACluster
cd MingPCACluster; chmod 755 bin/*
./bin/MingPCACluster  -h  #  直接运行
```

# 4. 用法和参数说明

## 4.1 MingPCACluster 参数

程序 MingPCACluster 很简单，一个输入和一个输出. 具体如下。

```
[heweiming@cngb-ologin-25 bin]$ ./MingPCACluster

        Usage: MingPCACluster   -InVCF   <in.vcf.gz>   -OutPut <outPrefix>

        -InVCF          <str>       Input SNP VCF Format
        -InGenotype     <str>       InPut Genotype File
        -InKinship      <str>       Input SNP K Kinship File Format
        -OutPut         <str>       OutPut File Prefix(Kinship PCA etc)

        -Method         <int>       Method of Kinship [1-4],defaut [1]
                                    1:BaldingNicolsKinship(VanRaden/Normalized_IBS)
                                    2:IBSKinshipImpute 3:IBSKinship 4:p_dis

        -help       v1.03           Show more Parameters and help [hewm2008]
```

只要一个 VCF　or　STOgem 文件作为输入即要可
　　一个输出文件前缀。

其中对 VCF 可以传多一个文件，只计算里面部分样品的 pca 的结果。
-Method　　是生成样品之间 kinship 的算法，也是 pca 依赖的计算矩阵。

### 4.1.1 更多详细参数

上面是简要输入，更多详情可以 `-h` 查看。

```
[heweiming@cngb-ologin-25 bin]$ ./ MingPCACluster    -h

More Help document please see the doc/pdf
              Para [-i] is show for [-InVCF], Para [-o] is show for [-OutPut]

      Usage: MingPCACluster    -InVCF    <in.vcf.gz>    -OutPut <outPrefix>

            -InVCF            <str>          Input SNP VCF Format
            -InGenotype       <str>          InPut Genotype File
            -InKinship        <str>          Input SNP K Kinship File Format
            -OutPut           <str>          OutPut File Prefix(Kinship PCA etc)

            -Method           <int>           Method of Kinship [1-4],defaut [1]
                                              1:BaldingNicolsKinship(VanRaden/Normalized_IBS)
                                              2:IBSKinshipImpute 3:IBSKinship 4:p_dis

            -help          v1.03              Show more Parameters and help [hewm2008]


            -InSampleGroup <string>     In File of sampleGroup info,format(sample groupA)

            -MAF                <float>      Min minor allele frequency filter [0.001]
            -Fchr               <str>        Filter the chrX chr[chrX,chrY,X,Y]
            -Miss               <float>      Max ratio of miss allele filter [0.25]
            -Het                <float>      Max ratio of het allele filter [1.00]
            -HWE                <float>      Exact test of Hardy-Weinberg Equilibrium for SNP Pvalue[0]
            -SubPop             <str>        Sub Sample File List to PCA[ALLsample]
            -KeepRemainVCF                    keep the VCF after filter

            -InSTOgem           <str>        InPut STOmics gem File of MIDCounts(beta)
            -STOName            <string>     STOmics Sample Name STOName
            -CellBin            <int>        STOmics cell bin[100]


            -PCANum             <int>         Num of PCA eig [10]
            -MaxCluNum          <int>         Max Cluster Num to find Best K [12]
            -BestKRatio         <float>      Get the best K Cluster by deta-SSE Ratio[0.15]
```

参数当一看即明，不过多说明。  中间主要是过滤 SNP 的参数 和 STO 的 bin 50 参数

下面一般参数  如设 K 的最大值，cluster，可以理解 structer 的 K 值

### 4.1.2 其它脚本参数

程序也提供了作图软件 perl 作图脚本（这个脚本后面将会优化更动较大，主要是最近时间较忙），作图脚本的简要参数说明如下：

```
[heweiming@cngb-ologin-25 bin]$ ./ploteig

        Version:1.03              hewm2008@gmail.com

        Usage: ploteig    -InPCA    pca.eigenvec -OutPrefix Fig

            Options

            -InPCA            <s>     : InPut File of PCA
            -OutPrefix        <s>     : OutPut file prefix

            -help                     : Show more help [hewm2008]
```

即输入 MingPCACluster 的 pca 结果即可。更详细 help 可以-h 查看，主要作图参数，如

```
[heweiming@cngb-ologin-25 bin]$ ploteig    -h

            Version:1.03              hewm2008@gmail.com

        Usage: ploteig    -InPCA    pca.eigenvec -OutPrefix Fig

            Options

            -InPCA            <s>     : InPut File of PCA
            -OutPrefix        <s>     : OutPut file prefix

            -help                     : Show more help [hewm2008]

            -columns          <s>     : the columns to plot a:b [3:4]
            -ColShap                  : colour <=> shape for cluster or subpop
            -pops             <s>     : Populations to plot, eg    -p GA:GB:GC [ALL]
            -border           <i>     : how to plot the border (1,2,4,8,3,31 ) [3]
            -title            <s>     : title (legend) [PCA]
            -keystyle         <s>     : put key at top right    default(in) [outside]box [outside]
            -pointsize        <i>     : point size for plot [3]

            -BinDir           <s>     : The Bin Dir of gnuplot/R/ps2pdf/convert [$PATH]
```

画图的点的大小等

## 4.2 输入文件

1 常规群体 VCF 格式，多样品的
2 群体 genotype 格式
3 时空表达量 xx.gem.gz 格式

### 4.2.1 数据文件

VCF 文件格式

具体 VCF 格式见这儿，[点击打开查看]
文件截图：

**1.1 An example**

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

**genotype 文件格式**

A 样品和位点 的基因型格式，直接见下面截图即知

| #CHROM | POS | REF | 1 | 10 | 100 | 100Y | 101 | 101Y | 102 | 103 | 103Y | 104 | 105 | 105Y | 106 | 107 | 108 | 109 | 109Y | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr1 | 2551 | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | M | C | C | C | C | C | - | C | C | - | C | - | C | C | C | C | C | C | C | C | C |
| Chr1 | 2554 | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | - | T | T | - | T | - | T | T | T | T | T | T | T | T | T |
| Chr1 | 2557 | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | - | G | G | - | G | - | G | G | G | G | G | G | G | G | G |
| Chr1 | 2560 | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | - | C | C | - | C | - | C | C | C | C | C | C | C | C | C |
| Chr1 | 2565 | A | A | A | A | G | A | R | A | A | R | A | A | A | A | A | A | R | A | A | A | A | R | - | A | R | - | A | - | A | A | A | R | A | A | A | A |
| Chr1 | 2566 | A | A | A | A | A | A | A | A | A | W | A | A | A | A | A | A | W | A | A | A | A | A | - | A | W | - | A | - | A | A | A | W | A | A | A | A |
| Chr1 | 2572 | C | C | C | C | A | C | M | C | C | M | C | C | C | C | C | C | M | C | C | C | C | M | C | C | M | - | C | C | C | C | C | M | C | C | C | C |
| Chr1 | 2574 | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | M | C | C | - | C | C | C | C | C | C | C | C | C | C |
| Chr1 | 2581 | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | - | C | C | C | C | C | T | C | C | C | C |
| Chr1 | 2584 | T | T | T | T | - | T | Y | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T |
| Chr1 | 2585 | T | T | T | T | - | T | T | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T |
| Chr1 | 2589 | A | A | A | A | A | A | A | A | A | A | A | A | - | A | A | A | A | R | A | A | A | A | A | A | A | - | A | A | A | A | A | A | A | A | A | A |
| Chr1 | 2590 | T | T | T | T | C | T | T | T | T | T | T | T | - | T | T | T | T | Y | T | T | T | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T |
| Chr1 | 2591 | T | T | T | T | T | T | T | T | T | T | T | T | - | T | T | T | T | W | T | T | T | T | T | T | T | - | T | T | T | T | T | T | T | T | T | T |
| Chr1 | 2594 | C | C | C | C | C | C | C | C | C | T | C | C | - | C | C | C | C | Y | T | C | C | C | C | C | T | - | C | C | C | C | C | T | C | C | Y | C |
| Chr1 | 2611 | T | T | T | T | T | T | A | T | T | T | T | T | T | T | T | T | W | T | T | T | T | W | T | T | T | - | T | T | T | T | T | T | T | T | T | T |

**时空 xx.gem.gz 文件格式**

如下是，共四列，基因 x ，y 和表达量

| geneID | x | y | MIDCounts |
|---|---|---|---|
| A1BG | 20870 | 10517 | 3 |
| A1BG | 16276 | 13190 | 2 |
| A1BG | 17095 | 10127 | 3 |
| A1BG | 10024 | 10804 | 1 |
| A1BG | 15667 | 9103 | 1 |
| A1BG | 18740 | 12537 | 3 |
| A1BG | 9007 | 15766 | 1 |
| A1BG | 7762 | 5636 | 1 |
| A1BG | 10275 | 10030 | 4 |

### 4.2.2 样品分群信息(可选)

有时已经知道样品的分群信息了，要看已知分群信息和聚类的差异。传入和各聚类结果比较。
格式很简单，就两列。 第一列 样品名 第二列 分群名。 如下截图

| | |
|---|---|
| NA06984 | CEU |
| NA06986 | CEU |
| NA06994 | CEU |
| NA07048 | CEU |
| NA07051 | CEU |
| NA07347 | CEU |
| NA07357 | CEU |

用 -InSampleGroup 输入上面的文件
另顺说 -SubPop 只要样品名一列就行，这儿是计算某些样品之间的 pca

### 4.3 输出文件

主要结果有如下文件:

| 输出文件 | 说明 |
| --- | --- |
| **out.kinship** | 输出的亲缘矩阵，各样品的两两关系 |
| **out.eigenvec** | 输出最优聚类和 pca 结果 |
| **out.eigenval** | 输出 pca 结果的特征向量 |
| **out.PCA1_PCA2.pdf** | 输出按 cluster 染色后的 pca 1 2 图 |
| **out.K.pdf** | 输出 cluster K 图 |
| **out.cluster** | 输出的各种 K 的 cluster 聚类结果 |
| **Out.cellbin.gz** | 输出 bin50 cell 的结果，若是 -InSTOgem |
| **Out.cluster pdf/ png** | 输出坐标 cluester 图，若是-InSTOgem |

示例图见上面应用场景给的图。示例图和格式当一看即明，相关图可以见 example 1 和 2

# 5.实例

**Example1 是抽了小样品的，可以当实例**

**Example2 的实便数据较大， 在这只提取缩小区域取了**

更多实例 随时更新，见 website 网页，具体见这:
https://github.com/Hewm2008/MingPCACluster 里面的

# 6.优势

1 快速少内存 可以用于多种场景
2 用法简易，对小白用户十分友好。
3 应用场景广泛，应用场景多种多样，自主性较大，用户可以结合自己的数据画图。
4 免安装，使用方便

# 7.常见问题

## 7.1 为什么要重复开发这个 PCA 分析软件

A . PCA 分析很多软件，但很多须要原始数据转格式，对小白不友好。
B tassle gapit 的软件是：n*m（n 是样品，m 是位点） n*n 的矩阵去算 kinship 和 pca 的 由于用到 m 位点一多 内存极大。 我这是连读连算 只是两个 n*n 的内存。 速度更快 结果比较会 除了精度问题 kinship 是一致的。 和 gcta 是一样的。 我这添加了新的 聚类分析 k-means 算法和染色
C 时空用到的工具 seurat 也是 R 包，我猜也是用到 n*m 的**稀疏矩阵**，但还是耗大量的内存，同时在时空组学中，样品多(n)，聚类耗的内存也大，我这主要使之不受 m 的影，n 大，n*n 的矩阵也是很耗内存，在这还没有和 seurat 进行过系统比较和测评（本人还对这数据不敏感），希望能够用，同时降低内存。

## 7.2 MingPCACluster 和准确性如何?

用同一数据做 pca 分析,同 gcta 和 tassle gapit 的 kinship 是一致的，pca 的结果也是一样的。再次 Example 也是很好把各人群分开，说明软件的准确性不用质凝，效率从开发时就是考虑要高效低内存的，具体后面有空再给出测评比较图表。
By the way 其中 tassle gapit 主要做 gwas,是过滤 maf 0.05 后去做的 pca，对特异的样品（离群的）影响大，mingPCA 默认 maf 为 0.001，几不过滤 maf， PCA 能把离群样品找到。

## 7.3 联系与打赏

随意 是缘是福，一切随风

- ✉ hewm2008@gmail.com / hewm2008@qq.com
- join the ***QQ Group : 125293663***

微信 打赏

QQ 入群：***125293663***

群名称:Reseqtools (itools)
群 号:125293663

*微信公众号*