1. 目的

规范壹基因动植物研究部进行群体重测序群体结构分析的操作流程，降低分析风险，以达到高质量、高标准的完成项目的目的。

2. 职责

项目执行人：严格按照本标准操作规程进行动植物群体重测序的群体结构分析。

信息负责人：参照本标准对项目执行人的操作进行复核，确保产品质量。

3. 适用范围

适用于壹基因动植物研究部进行群体重测序群体结构分析操作。

4. 术语和定义

4.1 主成分分析（PCA）

数学中的一种多元统计分析方法，研究如何通过少数几个主成分来代表多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量信息，实际应用非常广泛。在群体重测序中，主要利用群体 SNP 数据进行 PCA 分析，编程计算出能代表所有 SNP 变异结构的一些主要成分（特征向量），然后画图将分析结果直观展示出来，示意图如下：

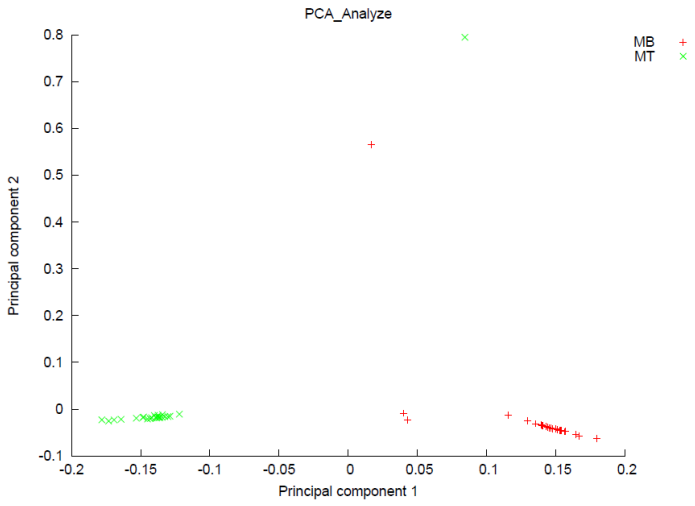

图 1 PCA 图

上图中一个点代表一个样品，点的横纵坐标是该样品对应的 1、2 特征向量中同一顺序元素的值，相应的特征值大小代表该主成分在整个关系中所占的比例，通过该图可以跟样品的实际分组进行对比，看出样品分组好坏，同一个亚群内的个体在图上应该是能聚在一起的，该结果对于后续需要分群进行研究的点起到一定的指导作用。

4.2 进化树(Tree)

用于衡量群体中样品之间的进化关系远近，示意图如下：

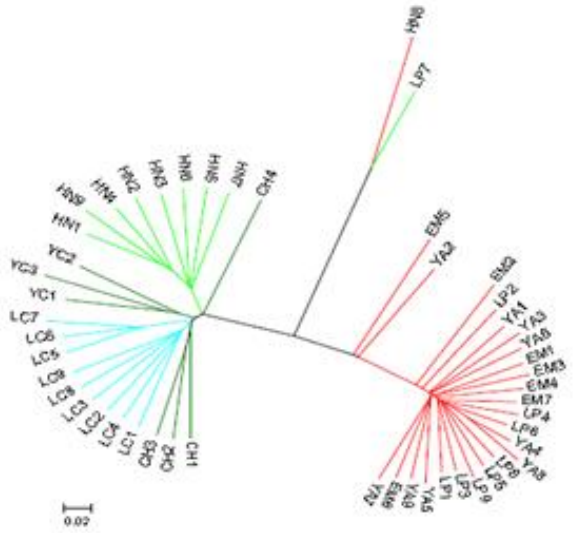

图 2 Tree 图

对于同一亚群内的样品，图上显示应该能很好的分在一起或离得不远，如图上红色分枝显示，通过该图可以说明品种之间的进化关系远近，由此衍生出一些专门针对物种进化过程的研究点。

4.3 群体结果推断(Structure)

由 K（2...7）个亚群组成，每个样品中属于同一亚群的部分用同一颜色标识，如果一个样品对应两种不同的颜色，则表示该样品可能是两个亚群之间的中间品种。当 K 值取得越大时，样品之间的差异性越被放大，分得越细，但并一定要画 2 到 7 的所有 K 值，可根据实际结果图来决定 K 值取到哪就可以完全体现出所有样品的结构关系。K=2 时的示意图如下：

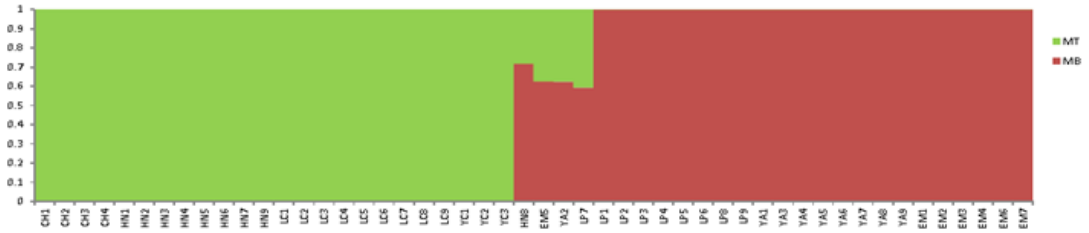

图 3 structure 图

该图与上述的 PCA 图及 Tree 图可以相互对应，K 值取得越大，群体不断分化，慢慢产生了该物种的其它品种。为了使同一亚群内的样品能聚在一起，可以调整横坐标对应的样品顺序来使得图上颜色成块显示，只是显示方式的改变对结果并没有影响。

4.4 连锁不平衡(LD)

不同座位上某两个等位基因同时遗传的频率明显高于预期的随机频率的现象，称为连锁不平衡，用 r^2 来衡量连锁不平衡程度，r^2 为 1 时表示完全连锁，r^2 为 0 表示完全不连锁，等位基因随机组合，示意图如下：

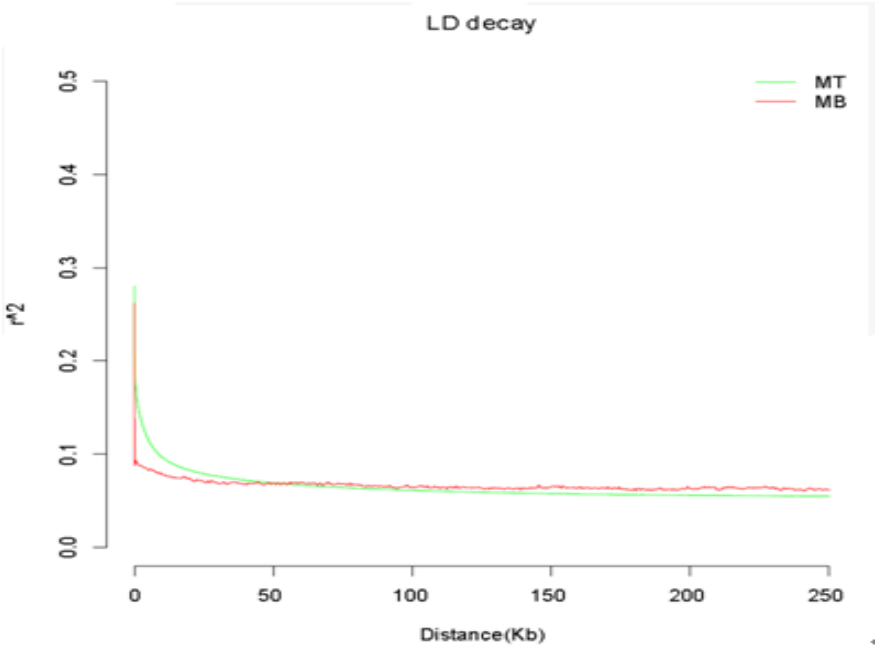

图 4 LD 衰减图

图中横坐标表示两个 SNP 位点之间的距离，纵坐标表示 r^2，由图上也可看出，距离越短，连锁不平衡越强，即 r^2 取值越大，一般通过 r^2 下降到最大值一半时对应横坐标大小来衡量不同物种或同一物种不同品种之间的连锁不平衡程度差异。r^2 的取值水平直接影响到后续 tag SNP 的挑选、单体型以及 Hapmap 的分析。

5 操作流程

5.1 总流程图

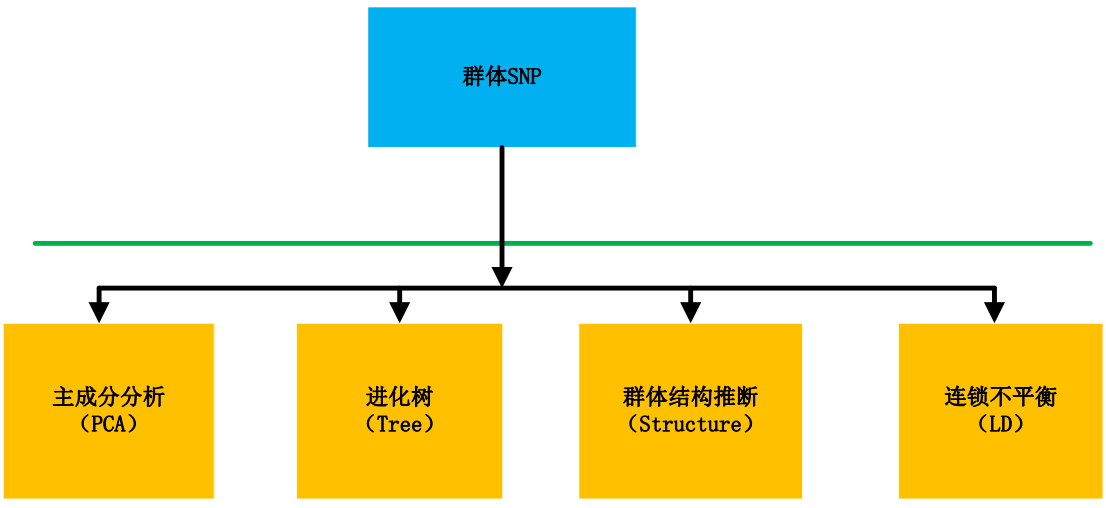

图 5 群体结构分析总流程图

5.2 流程路径

/nfs/pipe/Resequencing/2.population_structure/

5.3 具体操作

5.3.1 基本群体结构分析(PCA/Tree/Structure)

5.3.1.1 程序

PopuStructV1.1.pl

群体结构中的三个基本点：PCA、Tree、Structure 可用该程序完成分析，得到相应的结果图。

5.3.1.2 用法

perl PopuStructV1.1.pl －Struct * -OutDir * -indi * -list * -qsubPara *

两种方式：将公共程序总目录链接到项目目录下或运行时设置程序全路径。配置好参数写入 sh 脚本后 nohup *.sh &命令后台运行，程序将自动生成所要分析内容的脚本以及 qsub 自动投递部分任务，有些分析点的第二、三步或者最后一步任务还需要手动操作。

5.3.1.3 输入文件

共两个：包含所有染色体群体基因型文件路径的列表文件和群体对应样品的信息文件。

群体基因型文件格式：

染色体 位点 参考序列碱基 样品 1 基因型 样品 2 基因型…

前三列以 tab 分割，从第三列开始后面用空格隔开，将各条染色体的群体基因型文件路径列到一个文件里就是所需要的群体基因型列表文件。

样品信息文件格式：

样品编号 样品编号 样品编码 样品编码 对应群体 对应亚群

各列以 tab 键分开，第一列样式 1 2 3 4 5 10 11...，第二列样式 01 02 03...09 10 11...，第三列为样品全称，第四列为样品简称，这两列根据客户所提供样品信息写入，若一开始客户提

供的样品名就是简写代码，则两列相同，第五列为样品所属大群，第六列为更细划分后样品所属亚群，若大群下没有亚群划分，则两列相同。

5.3.1.4 参数说明

-Struct　五个可选参数设置值：PCA/Tree/Frappe/Structure/All，设置为 All 表示分析全部

-OutDir　输出目录，全路径

-indi　群体中所有样品信息文件，全路径

-list　列表文件，该列表文件里每行对应的群体基因型文件都需要提供全路径

-qsubPara　自动投递任务时-q 和-P 参数的设置，格式为：q:bc.q_P:aaptest

其中-Struct 参数分别设置为 Frappe 和 Struct 时，进行的都是群体结构推断分析(Structure)，只是用的软件不同，但 Struct 耗时太长，可能跑几月才出结果，因此一般不推荐此分析方法，而推荐用 Frappe。

5.3.1.5 结果说明

在设置的输出目录下，有对应分析内容的文件夹，包含中间结果文件、脚本和最终结果等内容，具体信息可参见下面的分步解析。

5.3.1.6 分步解析

5.3.1.6.1　PCA

5.3.1.6.1.1 脚本

perl PopuStructV1.1.pl －Struct PCA -OutDir * -indi * -list * -qsubPara *，写入 sh 脚本 nohup 后台运行。

5.3.1.6.1.2 主程序

调用主程序“flow_pcaV1.1.pl”生成 PCA 分析所有脚本及对脚本的自动投递，待任务运行完成后，nohup 运行脚本“Sum.sh”，得到最终结果图。画图所用结果文件为“*eigenvector.xls”，这些文件都给出了前 4 个特征向量的取值，如“1_2.eigenvector.xls”，画 1-2 主成分分析图时，每个点横纵坐标取值为 1、2 特征向量同一顺序元素的值，因此两个向量长度一致并且跟样品

个数相同，可自己编程根据该方法画图。

5.3.1.6.1.3 分析程序

共四个：pcaclean.pl、pcaMtM.pl、pcasumX.pl、pcaPlot.pl，最后还有个对 PCA 结果画 3D 图的程序“new_3DPCA.pl”，同一结果的不同展示效果供选择。

5.3.1.6.1.4 软件说明

用 EIGENSOFT 软件提供的 twstats 程序进行 Tracey-Widom 检验得到特征向量的显著性分析。

5.3.1.6.1.5 结果说明

在设置的输出目录下，程序自动新建文件夹 PCA，再在这个文件夹下新建三个子文件夹：“data”，“plot”，“shell”，其中 data 用于放置流程中间结果，shell 目录放置所有脚本，plot 目录下的 pdf 文件是我们所需的最终结果图，包括二维和三维图，“*eigenvector.xls”文件是对应的特征向量结果，画图程序的输入文件。

5.3.1.6.2　Tree

5.3.1.6.2.1 脚本

perl PopuStructV1.1.pl －Struct Tree -OutDir * -indi * -list * -qsubPara *，写入 sh 脚本后 nohup 后台运行。

5.3.1.6.2.2 主程序

总脚本“Run_Tree_shell.sh”串联起所有分析流程，该脚本通过程序“PopuStructV1.1.pl”自动投递到大型机。

5.3.1.6.2.3 分析程序

共用到四个：cat_genotype.pl、Com_for_tree_V3.pl、whole_pdistance.pl、fneighbor。

5.3.1.6.2.4 软件说明

fneighbor 通过邻接法（neighbor-joining method）构建有根树，详细的软件说明见网址：

http://genome.csdb.cn/cgi-bin/emboss/help/fneighbor。

5.3.1.6.2.5 结果说明

“*.tre”文件为最终结果文件，将其从大型机下到 windows 后，导入 MEGA 软件查看。通过 MEGA 软件的一些菜单调试，可设置不同的图形显示方式和分支颜色。

5.3.1.6.3 Frappe

5.3.1.6.3.1 脚本

perl PopuStructV1.1.pl －Struct Frappe -OutDir * -indi * -list * -qsubPara *，写入 sh 脚本后 nohup 后台运行。

5.3.1.6.3.2 主程序

调用主程序“Frappe_FlowV1.1.pl”，生成群体结构分析的所有脚本及脚本的自动投递。待第一步分析脚本自动运行完成后，需要手动 nohup 运行第二步的脚本投递程序“F_step2K_qsub.sh”，不一定要跑 K 从 2 到 7 的所有脚本，可自行根据物种而定，一般最大就跑到 K=7，然后 nohup 运行第三步的分析脚本“F_step3.sh”，完成后将得到最终的结果文件“result.final”。

5.3.1.6.3.3 分析程序

changeToGenotype.pl， newFormat.pl， ReadLog2txt.pl，plink，frappe_linu-x64， osh.pl，nsh.pl，structure_arrange.pl 等。

5.3.1.6.3.4 软件说明

两个核心软件：plink， frappe_linux64.软件 plink 的参数设置为：--pedped_file －recode12 --geno 0.5，frappe 软件构建群体结构和群体世系的信息。

5.3.1.6.3.5 结果说明

在设置的输出目录下新建文件夹“Frappe”，在其下面又新建了四个子文件夹：“ data”，“frappe”，“ plink”，“ shell”，分别放置中间文件，frappe 软件输出，plink 软件输出以所有脚本文件。文件“result.final”即为我们所要的最后结果，将其从大型机下到 windows 桌面，

导入 excel 后分别画出 K=2..7 时的柱状比例图即可。如画 K=2 时的结构图，用第四列和第五列来画柱状比例图，第三列为每一柱子对应的样品，以此类推画出 K=3..7 时的群体结构图。

5.3.1.6.4　Structure

5.3.1.6.4.1 脚本

perl PopuStructV1.1.pl －Struct Structure -OutDir * -indi * -list * -qsubPara *，写入 sh 脚本后 nohup 后台运行。

5.3.1.6.4.2 主程序

shell 总脚本为“Run_Structure_shell.sh”，串联起 Structure 分析的所有步骤，脚本中最后的“Start.sh”将对任务进行自动投递，待所有脚本运行完成后，要再一次运行该总脚本能得到最终结果文件。

5.3.1.6.4.3 分析程序

result_report.pl、all_cov.pl、all.pl、rand_small_position.pl、get_structure_input.pl、new_mainparams.pl、osh.pl 以及 shell 脚本 Start.sh。

5.3.1.6.4.4 软件说明

核心软件：STRUCTURE

5.3.1.6.4.5 结果说明

第一次运行总脚本后，在设置的输出目录下程序自动新建文件夹 Struct，在这个目录另外新建了 6 个子文件夹：structure_K02、structure_K03...，分别对应 K=2...7 时的分析脚本，跑完这些脚本后，需将基因型列表文件和样品信息文件链接到脚本“Run_Structure_shell.sh”所在的同级目录下，再重新 nohup sh 总脚本，在文件夹“Struct”下有个 final.out 的文件，将其从大型机上下载下来后导入 excel 按 5.3.1.6.3.5 所述方法画图即可。

5.3.2　LD 连锁不平衡

为了分析群体的连锁不平衡水平，用 java 软件 Haploview 计算等位基因的相关系数（r^2），

再用画图工具（如 R）画出对应的 LD 衰减图。由于 Haploview 这步分析比较耗时，需对基因组进行切分以染色体为单位进行，因此分别可画出各条染色体数据对应的 LD 衰减图和整个群所有数据对应的 LD 衰减图。

5.3.2.1 程序用法

perl Run_LD_decay_flow.pl

-input　群体基因型列表文件

-outDir 输出目录

-maxdis SNP 位点之间的最大距离，对应 LD 衰减图中横坐标大小，单位 kb，该值需要根据具体物种而设定，在分析前可找该物种文章中相应的 LD 衰减图看一下，默认 500kb，此时输出文件较大，当设为 1Mb 的最大距离时文件大小是 500kb 时的 10 倍，需安排好存储，但一般不会设到那么长。

-BinDir LD 分析程序路径

-qsub 设置即为自动投递任务，不设为手动 qsub 投

-mem 内存，默认 1.68G.

配置好参数写入 sh 脚本 nohup 运行后，程序会自动投递所有染色体对应 LD 分析脚本，得到每条染色体对应的画图结果文件，“plot”目录下的“*.info”文件，用第 1 列和第 5 列画图可得到 r^2 结果值，用第 1 列和第 6 列画图可得到 D' 结果图。

如果要画的是所有群体 SNP 数据计算 r^2 后的 LD 衰减图，需要先 nohup 运行脚本“Sum.sh 得到最终统计结果文件“Sum.info”，对文件第 1 列和第 5 列画 r^2 图。

流程中默认使用的画图程序是“Rplot2Decay.pl”，只对一条染色体或一个亚群按上述所说的各列来画图，得到的图形往往不理想，需要调整，往往还想将不同亚群的 LD 曲线放在一张图中比较，综合这些情况，需要根据项目情况自己 R 编程画图。

5.3.2.2 输入文件

只有一个群体基因型列表文件，如上 5.3.1.3 所述。

5.3.2.3 结果说明

在设置的输出目录下，新建四个文件夹：“LDOUT”，“plot”，“shell”和“stat”。

“LDOUT”：流程中各步产生的中间结果文件，“*.info”，“*.ped”，“*.log”。

“plot”：最后的结果图和图形对应的输入统计文件。

“shell”：所有脚本，按染色体画的脚本和所有数据合起来画图的程序脚本(“Sum.sh”)。

“stat”：统计结果的各种中间文件和列表。

5.3.2.4 流程解析

为了更加明白分析流程，对 LD 分析脚本中的内容按顺序进行分步解析，说明如下：

5.3.2.4.1 genotype2pedigree.pl

5.3.2.4.1.1 用法

perl genotype2pedigree.p <genotype data><ped output><info output>

5.3.2.4.1.2 输入文件

群体基因型文件，格式与 5.3.1.3 所述格式基本一致，但不包括参考序列碱基这一列，前两列以 tab 键分割，从第三列开始以空格隔开；对于大群下面的亚群进行 LD 分析，需要总的基因型文件里面提取出亚群样品所对应的各列基因型。

5.3.2.4.1.3 输出文件

*.ped：群体基因型转化格式后的 ped 文件

*.info：群体各位点处等位基因频率统计信息

5.3.2.4.2 Haploview.4.2.jar

5.3.2.4.2.1 用法

jar －jar Haploview.4.2.jar －n －pedfile *.ped －info *.info －log *.log －maxdistance 250 －minMAF －hwcutoff 0.001 －dprime －memory 2096

参数说明：

-log：Haploview 软件运行的日志文件

-maxdistacne：最大距离，如上所述。

-minMAF：最小等位基因频率，一般设为 0.1。

-hwcutoff：检验哈代温伯格（hw）平衡的 p-value 过滤值，一般设为 0.001，根据物种满足哈代温伯格平衡情况而定，也可增大该值。

-dprime：输出 LD 文本文件到*.LD

-memory：脚本所用内存，单位 M，默认 512M.

5.3.2.4.2.2 输入文件

上步即 5.3.2.4.1 步产生的*.ped 和*.info 结果文件。

5.3.2.4.2.3 输出文件

*.LD：等位基因 r^2 结果文件，第五列为 r^2 的值，详细介绍可参见官网：

http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview。

*.log：软件运行的日志文件，包括设置参数和输出文件信息。

5.3.2.4.3 LD_decay_RR_D_step1.pl

5.3.2.4.3.1 用法

perl LD_decay_RR_D_step1.pl <*.LD 文件> <*.info 文件> <bin 大小> <符合要求的位点 r^2 等信息的统计文件> <不符合要求位点的统计结果文件> <长度>。

参数说明：

bin 大小：默认值为 1

长度：跟跑 Haploview 时设置的-maxdistacne 参数值一致，单位为 bp

这步将对上步产生的*.LD 文件进行过滤并求 r^2 和 D' 的平均值，过滤条件为第 8 列的距离小于指定长度，根据用于计算 r^2 值的两个位点 MAF 是否相等来分别输出统计结果。

5.3.2.4.3.2 输入文件

5.3.2.4.1 步产生的*.info 文件和 5.3.2.4.2 步产生的*.LD 结果文件。

5.3.2.4.3.3 输出文件

*.match：符合第 8 列距离小于指定长度且用于计算 r^2 的两个位点 MAF 相等的距离处 r^2 和 D' 的平均值。

*.unmatch：符合长度要求但 MAF 不等的距离处 r2 和 D' 的平均值。

上述两个文件的输出格式可参见程序本身。

5.3.2.4.4　LD_decay_RR_D_step2.pl

5.3.2.4.4.1 用法

perl LD_decay_RR_D_step2.pl <列表文件> <输出文件:*.info>

5.3.2.4.4.2 输入文件

列表文件：上步产出的*.match 和*.unmatch 文件全路径列表。

5.3.2.4.4.3 输出文件

*.info：最终用于画图的 r^2 和 D' 统计信息文件，输出格式为：距离、个数、r^2 总和、D' 总和、r^2 平均值、D' 平均值。

5.3.2.4.5　Rplot2Decay.pl

流程默认画图程序，一般需要另外重新编程画，为了阐述完整性，在这里也对该程序进行说明。

5.3.2.4.5.1 用法

perl Rplot2Decay.pl –input *.info –output *.pdf

5.3.2.4.5.2 输入文件

*.info：即为 5.3.2.4.4 步产生的结果文件

5.3.2.4.5.3 输出文件

plot 目录下的*.pdf 文件。

5.4 任务监控

5.4.1 qstat 查看任务运行状态，对于长时间没有运行的任务，需要 qstat －j 任务 ID 检查-q 和-P 参数设置的正确性；“qalter －l vf=*G”修改内存；后台运行程序监控：“ps －ef|grep 个人 id”，“kill 任务代码”可杀掉后台运行程序。

5.4.2 当需要根据具体物种情况修改上述流程分步解析说明中的参数时，对于默认自动投任务的分析内容，可待产生完所有脚本后杀掉总进程，再用 qdel 取消所有任务，修改好参数后按上述说明一步步手动重投。

5.4.3 查看 nohup.out 文件，程序运行过程中的一些报错信息会输出到该文件，如格式不正确或者样品个数不一致等。根据报错信息对应到相应的程序和文件进行修改。

5.4.4 查看脚本运行后的*.sh.*文件，一般会有报错信息，可对比不同染色体脚本的两个输出文件，没有明显异常时表示该分析完成，否则按报错信息提示去尝试解决。

5.5 错误分类

5.5.1 Unable to run job: job rejected: no access to project "paptest" for user…

进行 Struct 和 LD 分析时，程序设置了固定的-q(bc.q)和-P(paptest)参数，需要根据个人项目和组别进行修改。

Structure：修改脚本“Start.sh”中 qsub 语句的-P 参数。

LD：修改总脚本“Run_LD_decay_flow.pl”中 qsub 语句，添加-q 和-P 参数。

5.5.2 跑完某步分析没有结果图

先确定输入文件的正确性，再检查画图程序是否适合于该项目数据，找到问题修改相应程序。

5.6 注意事项

5.6.1 该文档只提供了群体 snp 以后的群体结构分析内容流程说明，其它分析如群体 snp 可参见相应的 SOP 文档。

5.6.2 群体基因型文件由群体 snp 结果转化而来，而群体 snp 的分析方法有很多种，如 SOAP +Soapsnp+GLFmulti，bwa+GATK，SOAP+caSFS，bwa+samtools 等，可根据各自项目进行方

法选择，将 snp 结果整理成群体基因型文件格式，即可开始后面的分析内容。

5.6.3　文档中用到的所有软件在公共程序路径下均可以找到。

5.6.4　最后仔细检查各步的分析结果，当得到我们所要图形或者画图所需文件后，方可完成分析。

6　相关软件

表 5　核心软件及版本

编号	软件	版本
1	PopuStructV1.1.pl	1.1
2	EIGENSOFT	4.2
3	fneighbor	3.6b
4	plink	1.07
5	frappe_linux64	1.1
6	STRUCTURE	2.2
7	Haploview.4.2.jar	4.2

7　相关文件

无。

8　相关记录

无。

9　附录

无。

联系我们

项目执行人：韩雪莲

项目执行人联系方式：hanxuelian@1gene.com.cn

扫一扫，更多精彩等着您

电话：0571-87885727

网址：www.1gene.com.cn

地址：杭州市滨江区江陵路 88 号万轮科技园 7 幢（310051）