TASSEL5.0 用户手册

Cornell 大学 Buckler 实验室

(June 12, 2014)

翻译：陈建国

湖北大学生命科学学院

www.maizegenetics.net/tassel

声明： 虽然 Cornell 大学 Buckler 实验室已经进行了广泛的测试并且一般来说结果是可靠的、正确的或合适的，但是对于任何一套特定的数据不能保证一定能够得到你想要的结果。强烈地建议用户利用其它软件来验证 TASSEL 的结果。

更多的帮助： 除了这个文档以外还可以得到额外的帮助。欢迎用户报告软件的缺陷，通过 TASSEL 网址申请新的性能。也欢迎对我们现在的团队成员提出问题。要想得到更快速和更准确的答案，请把你的问题提交给最相关的人：

Tassel 用户群（推荐）	**http://groups.google.com/group/tassel**
	tassel@googlegroups.com
一般的信息	**Ed Buckler（项目领导人）**
	esb33@cornell.edu
数据输入，Pipeline	**Terry Casstevens**
	tmc46@cornell.edu
统计分析	**Peter Bradbury**
	pjb39@cornell.edu

Contributors: Ed Buckler, Terry Casstevens, Peter Bradbury, Zhiwu Zhang, Dallas Kroon, Jeff Glaubitz, Kelly Swarts, Jason Wallace, Fei Lu, Alberto Romero, Cinta Romay, Eli Rodgers-Melnick, Alexander Lipka, Sara Miller, James Harriman, Yogesh Ramdoss, Michael Oak, Karin Holmberg, Natalie Stevens, and Yang Zhang.

Citations:

Overall Package:

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Genotyping by Sequencing:

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9(2): e90346

Mixed Model GWAS:

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. (2010) Mixed linear model approach adapted for genome-wide association studies. Nature Genetics 42:355-360.

TASSEL 项目由国家科学基金和 USDA-ARS 资助

相关的链接：

主网站：http://www.maizegenetics.net/tassel

开源代码：https://bitbucket.org/tasseladmin/tassel-5-source

Wiki：https://bitbucket.org/tasseladmin/tassel-5-source/wiki

目录

引言

虽然自从 2001 年开始公开发行以来 TASSEL 已经发生了相当大的变化，但是它的主要功能仍然是为研究表现型和基因型之间的关系提供工具[1]。TASSEL 的功能有：关联研究，评价进化关系，分析连锁不平衡，主成分分析，聚类分析，估算缺失数据，数据可视化。TASSEL 的开发由玉米遗传学和基因组学的一个课题组领导，因此这个软件的设计和计算上的优化都是为了解释很多植物和育种情况中存在的生物学现象。与人类遗传学相比，很多作物在核苷酸水平和结构变异上都是非常多样的（多样性比人类大 10 50 倍），近交和大的家系也是常见的，并且全基因组预测正在日益应用于现实世界的问题。这些生物学的差异导致一些不同的优化，这些优化对作物之外的很多生物学系统也有用处。

驱动 TASSEL 开发的设计要点之一是对更大的数据集进行分析的需要。TASSEL5 的核心对大数据进行了很多设计优化，包括：

- 核苷酸的位级编码（bit level encoding），这样可以非常迅速地获得遗传距离和连锁不平衡估计值（速度增加 20 50 倍）。
- 广泛应用 HDF5 文件格式，它已经作为很多气候模型的一个稳健的元件对矩阵形式数据开发。
- 从大规模的测序基因分型（Genotyping-by-Sequencing）数据中提取和调用 SNPs 的工具（通过超过 2.5 百万个 SNPs 和 96 百万个序列等位基因对 60,000 个样本进行了测试）。
- 对作物中的大家系进行了优化的投影（projection）和估算方法。这些优化中的一些可以使内存和计算的改良>100,000 倍。
- 以 DNA 亲缘关系为基础的混合模型已经开始主宰 GWP（Meuwissen et al，2001）和 GWAS（Yu et al，2006），然而这些模型的求解可能是缓慢的。TASSEL 已经成为一个试验台，实现了最优化方法中的一些，比如 EMMA（Kang at al，2008），加上最优化方差分量的方法，一旦使用 P3D（Zhang et al，2010）和 EMMAX（Kang et al，2010）。压缩算法也是可用的（Zhang et al，2010）。当正确地使用时，这些优化使得强大的 GWAS 在计算上成为可能。
- 代码正在不断地对更多的处理器（core）和集中站（cluster）而优化。例如，我们通常在 64 核计算机上运行估算。虽然 Java 的优点是系统之间的互操作性，但是它的

代码大约比最优化的 C 语言库慢 2 倍，对于一些问题的处理比 GPU 慢 10 倍。TASSEL5 正在构建直接到本机码的连接层，当需要这些效率时。

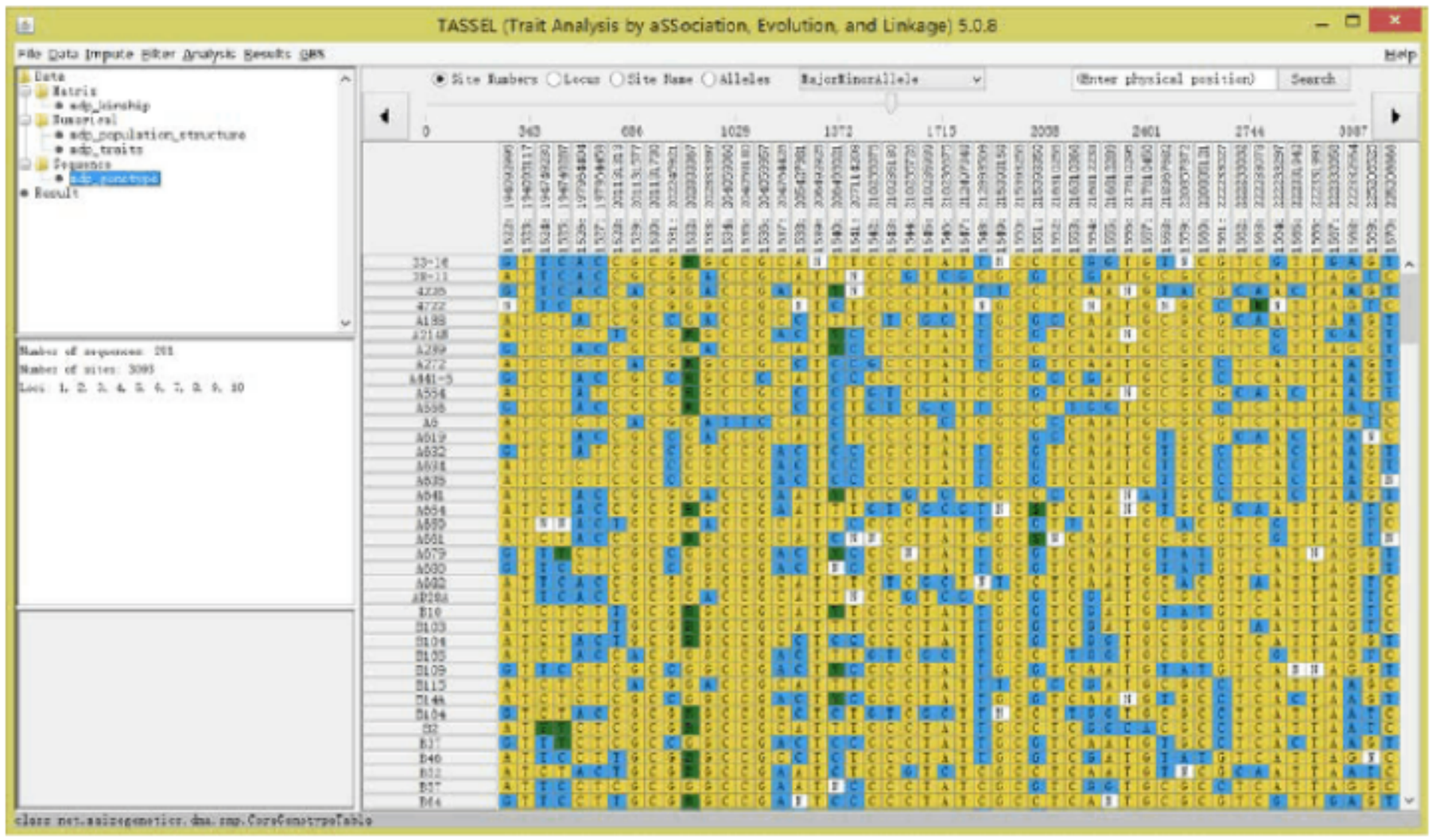
TASSEL 是为各种各样的用户设计的，包括那些对统计遗传学或计算机科学不熟练的用户。利用图形界面，通过点击适当的选项，就可以在少数几个步骤中完成 GWAS，这种 GWAS 利用混合线性模型方法来结合有关群体结构[6-8]和隐藏的亲缘关系[9]的信息。分析所需的所有处理过程都是自动进行的，包括导入表型数据和基因型数据，估算缺失数据（表现型或基因型），按次要等位基因频率过滤标记，产生主成分和亲缘关系矩阵来代表群体结构和隐藏的亲缘关系，优化压缩水平和进行 GWAS。

TASSEL 的命令行版本称为 Pipeline（管道），为用户提供对任务编制程序的能力，利用脚本而不是图形用户界面（GUI）。这个特征允许科研工作者利用少许代码行来定义任务，并提供把 TASSEL 作为一个分析管道的组成部分来使用或进行模拟计算的能力。我们也建立了一个大的开发者社区，为这个平台增加功能，并合作来改进该系统。因此在整个用户手册中你将看到如何通过三种不同的方式来完成大多数事情：利用 GUI，利用管道，利用 API（应用编程接口）。

TASSEL 是用 Java 编写的，因此几乎可以在任何操作系统中使用。通过单击 www.maizegenetics.net/tassel 上的一个链节，可以利用 Java 网页启动（Java Web Start）技术来安装它。也可以下载 TASSEL 的单机版本，以管道模式使用，或者在用户想要启动该软件的任何情形中从命令行使用。

1 入门指南

开始使用 TASSEL 的一个快速的方法是加载教学数据并尝试进行分析。然而，因为一些必要的步骤可能不是直观的，我们建议新用户按照这个手册后面的指南去做。本节提供安装和启动 TASSEL 软件的信息，并对界面进行简短的概述。



1.1 执行 TASSEL

http://www.maizegenetics.net/tassel/docs/ExecutingTassel.pdf

1.2 开源代码

在下面的网站上可以得到 TASSEL 的开源代码：https://bitbucket.org/tasseladmin/tassel-5-source。该套装软件使用了内含在 TASSEL 发行版的很多其它的库。这些包括 PAL 库的一个修改版（http://www.cebl.auckland.ac.nz/pal-project/），COLT 库（http://dsd.lbl.gov/~hoschek/colt/），jFreeChart（http://www.jfree.org/jfreechart/），Guava（Google Core Libraries）（https://code.google.com/p/guava-libraries），JUnit（http://junit.org），Archaeopteryx（https://sites.google.com/site/cmzmasek/home/software/archaeopteryx），以及 BioJava（http://www.biojava.org）。

1.3 软件开发工具

jProfiler (http://www.ej-technologies.com/products/jprofiler/overview.html)

install4j (http://www.ej-technologies.com/products/install4j/overview.html)

NetBeans IDE (https://netbeans.org)

Eclipse (http://www.eclipse.org)

IntelliJ (http://www.jetbrains.com/idea)

Structure101 (http://structure101.com)

TeamViewer (http://www.teamviewer.com)

Bitbucket (https://bitbucket.org)

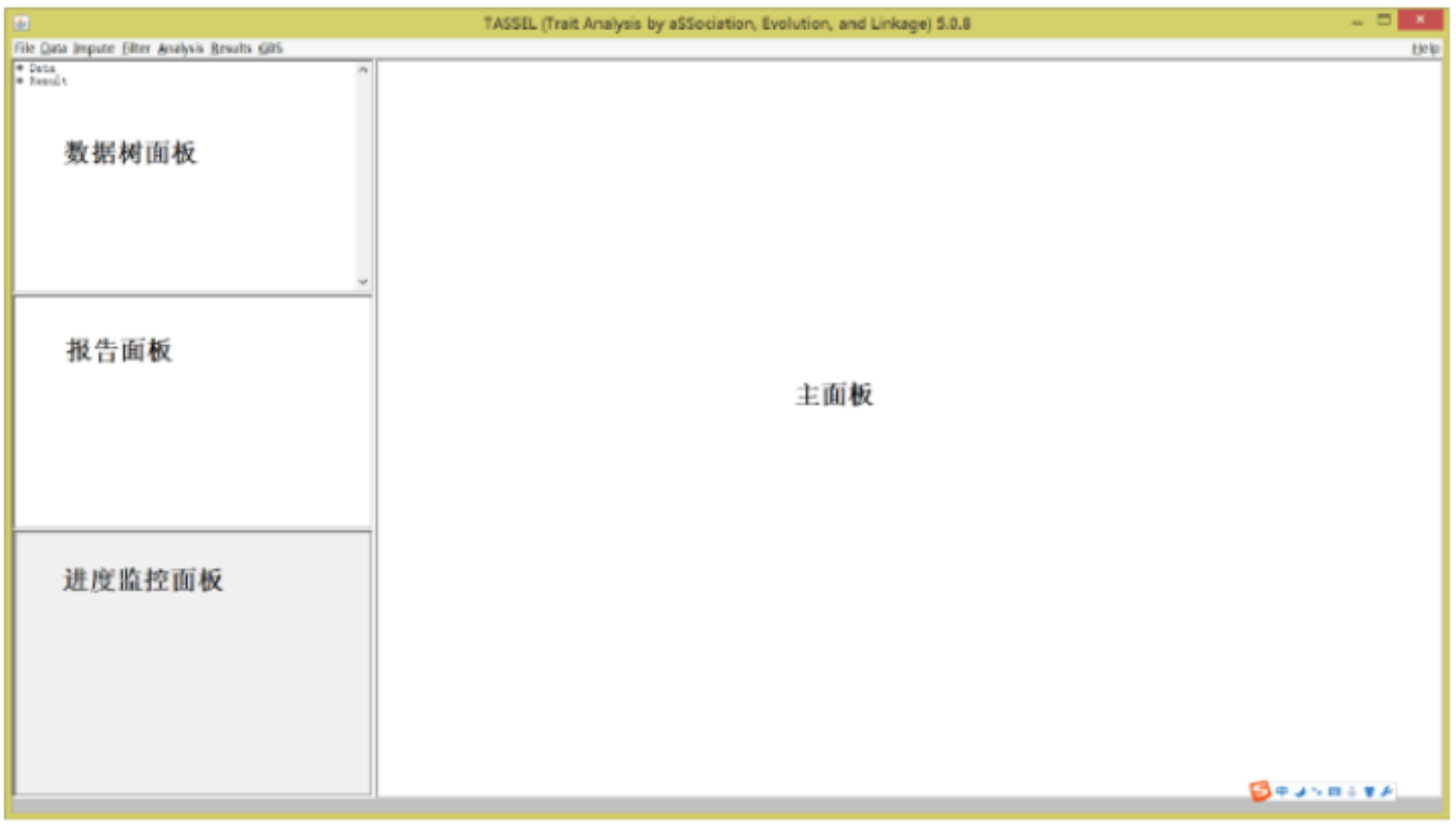
sourceforge (http://sourceforge.net)

JIRA (https://www.atlassian.com/software/jira)

Tower (http://www.git-tower.com)

1.4 图形界面

TASSEL 被组织成五个主要面板。1）顶部的菜单控制功能。2）左边顶部的数据树，组织数据集和结果。在执行一个想要的功能或分析之前必须首先选择数据树中显示的数据集。要选择多个数据集，按下 CTRL 键然后选择数据集。3）报告面板，位于数据树面板下面。它显示从数据树中选择的数据集的有关信息，比如数据的类型以及它是如何创建的。4）进度监控面板，在报告面板下面，显示运行任务的进度，具有能够取消任务的按钮。5）主面板，占据视图区域的右侧，显示从数据树中选择的数据集的内容。



1.5 管道（命令行界面）

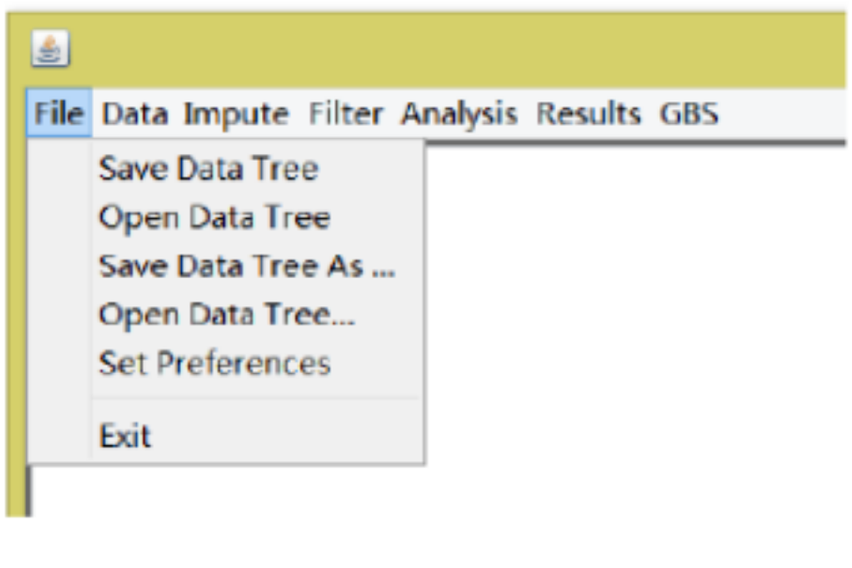
http://www.maizegenetics.net/tassel/docs/TasselPipelineCLI.pdf

1.6 GBS 管道

http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf

2 File（文件）菜单

数据树可以按照二进制格式保存。



2.1.1 Save Data Tree（保存数据树）

这个命令允许你将数据树面板的全部内容保存到一个默认位置。当用户下一次启动该程序时如果他们不想再创建一个已经用信息填充过的数据树，那么这个命令是有帮助的。为了保存一个数据树，选择 **File**（文件）> **Save Data Tree**（保存数据树）。

2.1.2 Open Data Tree（打开数据树）

为了恢复一个先前保存的数据树，选择 **File**（文件）> **Open Data Tree**（打开数据树）。

2.1.3 Save Data Tree As…（数据树另存为…）

为了把数据树保存到一个特定的位置或者要给它一个特定的名称，选择 **File**（文件）> **Save Data Tree As…**（数据树另存为…）。

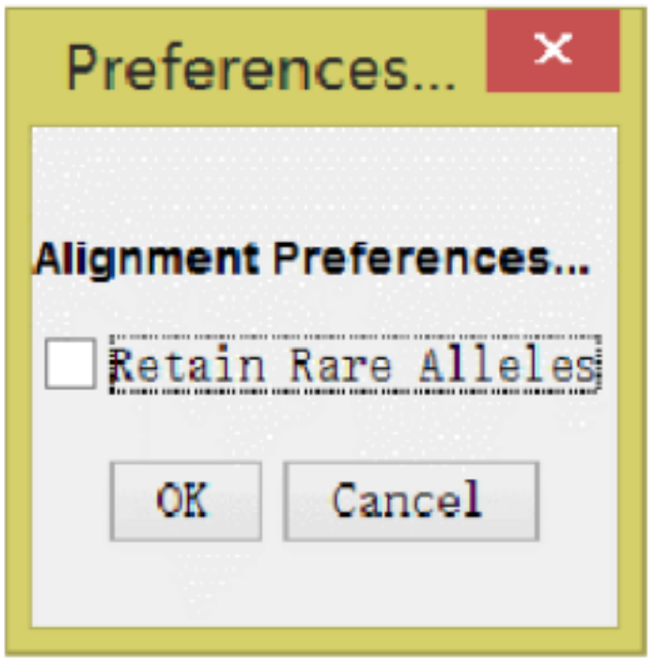
2.1.4 Open Data Tree…（打开数据树…）

为了从一个特定的位置恢复一个数据树，选择 **File**（文件）> **Open Data Tree…**（打开数据树…）

注意：上面提到的保存数据树的信息通常适用于特定的版本。当 TASSEL 的一个新版本发布时，用一个以前的版本保存的数据树可能不能加载到该版本。为了长期贮存，最好的作法是保存单独的数据集而不是整个数据树。

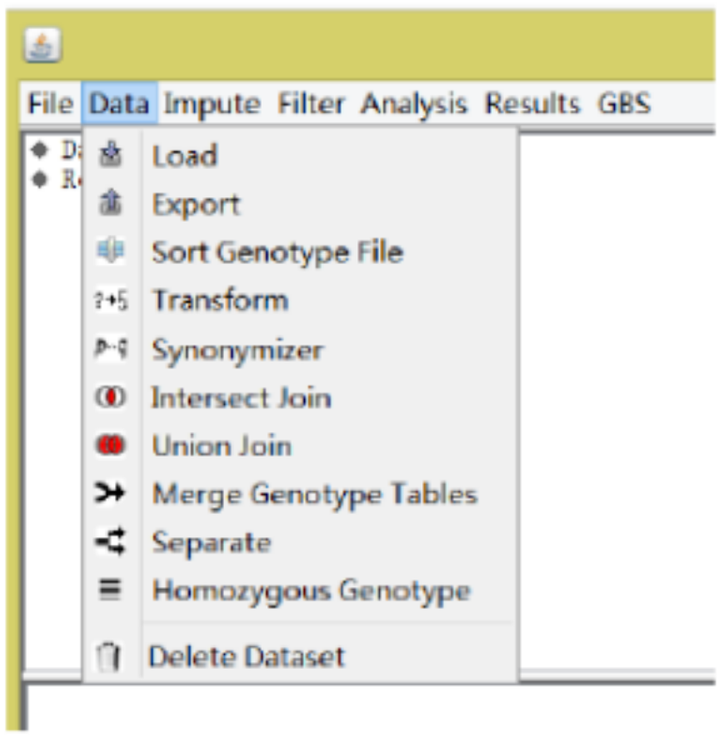
2.1.5 Set Preferences（设置首选项）

目前只有一个首选项，即是否要保留“稀有”等位基因。这是对于核苷酸数据是不相干的，因为在那个状态的数目上没有数据丢失。对于其它类型的数据，等位基因状态的最大数目（每个位点 0 可能超过 14。如果你“保留稀有等位基因”（Retain Rare Alleles），较低频率的等位基因的值将被固定到一个稀有的状态（Z）。否则，那些较低频率的等位基因被转变为未知的（N）。



3 Data（数据）菜单

Data（数据）菜单具有用来导入和导出数据集的选项，还有其它的数据处理功能。



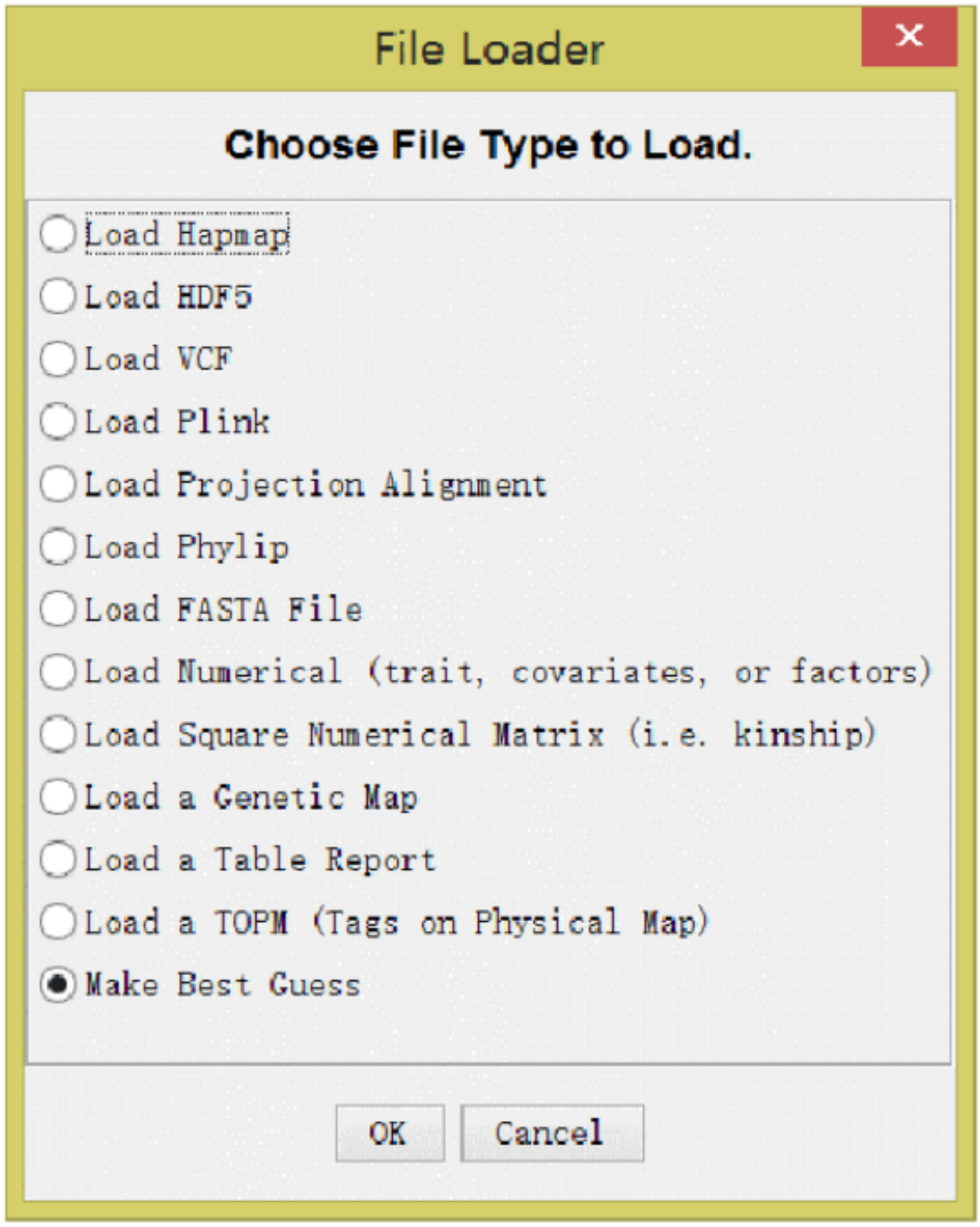
3.1 Load（加载）

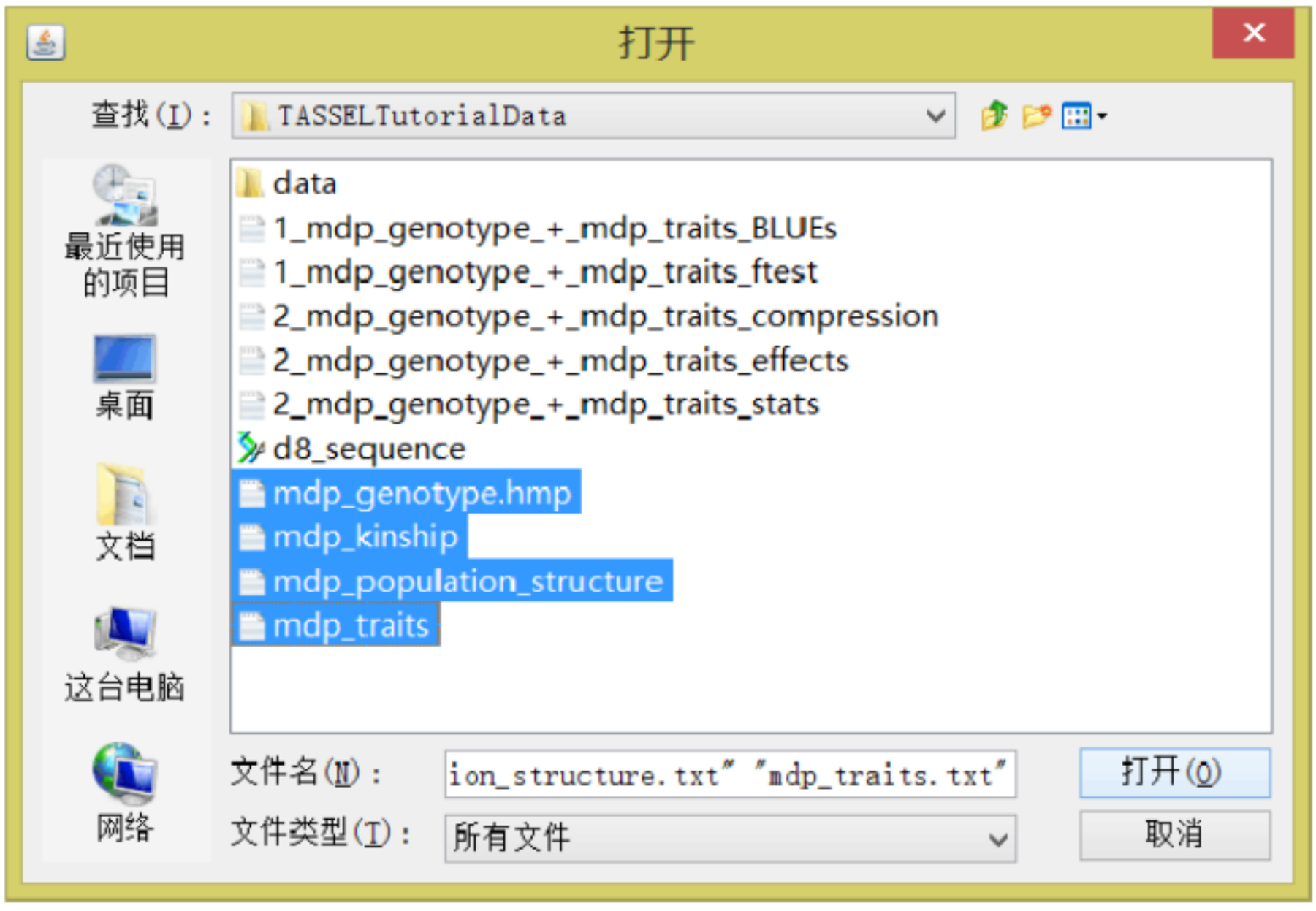
Load（加载）提供选项来导入基因型、表现型、群体结构、以及亲缘关系矩阵、等等的文件。

可以从 TASSEL 网址下载教学数据，链节为：

http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip.

要使用这些数据，压缩文件必须被解压，并保存在你的本地计算机上。利用“Make Best Guess”选项将这些教学文件正确地加载。可以同时导入多个文件，通过首先加亮它们（单击时按住 Shift 或 Ctrl 键）然后单击 Open 按钮。


File Loader
Choose File Type to Load.
Load Hapmap
Load HDF5
Load VCF
Load Plink
Load Projection Alignment
Load Phylip
Load FASTA File
Load Numerical (trait, covariates, or factors)
Load Square Numerical Matrix (i.e. kinship)
Load a Genetic Map
Load a Table Report
Load a TOPM (Tags on Physical Map)
Make Best Guess
OK
Cancel


打开
查找(I):
TASSELTutorialData
最近使用的项目
桌面
文档
这台电脑
网络
data
1_mdp_genotype_+_mdp_traits_BLUEs
1_mdp_genotype_+_mdp_traits_ftest
2_mdp_genotype_+_mdp_traits_compression
2_mdp_genotype_+_mdp_traits_effects
2_mdp_genotype_+_mdp_traits_stats
d8_sequence
mdp_genotype.hmp
mdp_kinship
mdp_population_structure
mdp_traits
文件名(N):
ion_structure.txt" "mdp_traits.txt"
打开(O)
文件类型(T):
所有文件
取消

3.1.1 Hapmap

Hapmap 是一个基于文本的文件格式，用于存储序列数据。一系列 SNP 以及种质品系的全部信息被保存在一个文件里。第一行包含标题标签，每个额外的行包含与单个 SNP 有关的全部信息。最前面 11 列描述 SNP 的属性，接下来的列描述单个种质品系的 SNP 值。第一行的最前面 12 列看起来应该像这样的，其中“Line 1”是种质品系名称的开始。

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panelLSID	QCcode	Line 1

虽然所有 11 个标题列是必需的，但对于 TASSEL 要正确地解释数据并不需要填写所有 11 列。仅仅需要的字段是“chrom”（染色体名称）和“pos”（位置）。在下面的例子中，基因型值是用 2 个字符（即 AA）代表的。注意你可以把那些基因型值作为单个字符值记录（见附录中的“Nucleotide Codes”（核苷酸代码））。

为了让 TASSEL 正确地读取 Hapmap 数据，数据必须按照每个染色体内部的位置次序排列，文件应该用制表符分隔（下面的例子为 Excel 格式，仅仅为了容易查看）。如果一些数据缺失，还是必须给出制表符的正确数目，以便 TASSEL 可以正确地将数据分派到列。

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	33-16	38-11	4226	4722	A188
PZB00859.1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	AA
PZA01271.1	C/G	1	1947984	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	GG	CC	GG	CC
PZA03613.2	G/T	1	2914066	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG	GG	GG
PZA03613.1	A/T	1	2914171	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA03614.2	A/G	1	2915078	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG	GG	GG
PZA03614.1	A/T	1	2915242	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA00258.3	C/G	1	2973508	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	CC	CC	CG	CC
PZA02962.13	A/T	1	3205252	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA02962.14	C/G	1	3205262	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	CC
PZA00599.25	C/T	1	3206090	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	TT	CC	TT	TT
PZA02129.1	C/T	1	3706018	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	CC	CC	CC	CC
PZA00393.1	C/T	1	4175293	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	CC	TT
PZA02869.8	C/T	1	4429897	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	TT	CC	NN	CC
PZA02869.4	C/G	1	4429927	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	NN	GG
PZA02869.2	C/T	1	4430055	+	AGPv1	Panzea	NA	NA	maize282	NA	NN	TT	TT	CC	TT
PZA02032.1	A/T	1	4490461	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	TT	AA	AA	AA
zag1.5	A/T	1	4835434	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	NN	AA	AA	AA
zag1.2	A/C	1	4835558	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	CC
zag1.6	C/T	1	4835658	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZD00081.2	C/T	1	4836542	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	CC
zag1.1	A/C	1	4912526	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	AA	AA	AA	AA
PZB00919.1	A/C	1	5353319	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	AA
PZB00919.2	G/T	1	5353655	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG	GG	GG

3.1.2 HDF5（层次数据格式，版本 5）

http://www.hdfgroup.org/HDF5

3.1.3 VCF（Variant Call Format 变异体召唤格式）

http://www.1000genomes.org/wiki/analysis/variant-call-format/vcf-variant-call-format-version-42

3.1.4 Plink

Plink 是一个全基因组关联分析工具箱，它带有它自己的基于文本的数据格式。数据被保存在两个文件中，一个.map 文件和一个.ped 文件。

.ped 文件包含所有的 SNP 值，具有六个强制性的标题列，家系标识符、个体标识符、父本的标识符、母本的标识符、性别以及表现型。TASSEL 仅仅要求个体标识符字段被填写。.ped 文件的每一行描述单个种质品系。注意在 Plink 中，一个未知的字符是用“0”代表的。然而在 TASSEL 中一个未知的字符是用“N”代表的，“0”代表杂合的 indel。TASSEL 将自动地在“0”和“N”之间转换。任何导出的 Plink 文件将用“+”（插入）和“-”（缺失）来代表杂合的 indel。

.map 文件描述关联的.ped 文件中的全部 SNP，其中每一行提供一个 SNP 的有关信息。.map 文件必须准确地包含四个列：Chromosome（染色体），rs#，Genetic distance（遗传距离）和 Position（位置）。TASSEL 不要求 Genetic distance（遗传距离）字段被填写。

两个文件都应该用制表符分隔。

对于数据格式的更详细说明，请访问 Plink 基本用法和数据格式网页：

（http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml）。

3.1.5 投影校准（Projection Alignment）

3.1.6 Phylip

有关 Phylip 格式的详情在以下网址上：

http://evolution.genetics.washington.edu/phylip/doc/sequence.html

3.1.7 FASTA

3.1.8 Numerical Data（数值数据）

这类格式被用于性状和协变量数据（比如群体结构）。与序列比对基因型数据相似，数值数据也由两个部分组成：一个标题定义数据结构，一个包含主要数据的主体。应该用制表符作为分隔符。然而，任何空白字符（比如空格）也将被当作一个分隔符。因此，名称中嵌入的空格将导致数据被错误地导入。我们建议用“NA”或“NaN”来代表缺失值。然而，任何文本值（例如“?”）将被作为缺失数据解释。有若干数值数据的格式来满足建模的要求。性状数据（因变量）可以被导入，通过利用“<Trait>”启动第一行然后利用性状名称。额外的分类器（classifier）也可以被包括在随后的标题行中，通过利用“<Header name = xxx>”启动行，继之以数据的每一列的名称。例如，为了定义环境，利用“<Header name = env>”启动第二个标题行。

可以在文件的开头插入注释行。注释行以字符“#”开始。

3.1.8.1 性状格式

这个格式不要求用户提供有关行数和列数的信息。文件以关键词<Trait>开始，后面是列的名称。品系的列不应该被标签。

例 1，性状值的简单列表：

<Trait>	EarHT	dpoll	EarDia
811	59.5	NA	NA
33-16	64.75	64.5	NA
38-11	92.25	68.5	37.897
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
A188	27.5	62	31.419
...			

例 2，在多个环境中收集的性状：

<Trait>	EarHT	PlantHT	EarHT	PlantHt
<Header name=env>		Loc1	Loc1	Loc2	Loc2

811	59.5	NA	NA	NA
33-16	64.75	121.5	NA	NA
38-11	92.25	153.8	37.897	83.4
4226	65.5	130.1	32.21933	82.1
4722	81.13	165.7	32.421	90.1
A188	27.5	110.2	31.419	79.6
...				

3.1.8.2 协变量格式

除了第一行必须是“<Covariate>”之外，协变量数据的格式和性状数据一样。这一行告诉 TASSEL 这个文件中的变量将被作为协变量使用而不是作为因变量使用。这是用于群体结构协变量的格式。

<Covariate>			
<Trait>	Q1	Q2	Q3
33-16	0.014	0.972	0.014
38-11	0.003	0.993	0.004
4226	0.071	0.917	0.012
4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
...			

3.1.8.3 作为数值协变量的标记值

有时候，用户可能想要把标记值当做数值协变量。如果文件的第一行是“<Numeric>”，那么数据将将被作为数值数据导入，但是在 GLM 和 MLM 中用作标记数据。

<Numeric>					
<Marker>	m1	m2	m3	m4	m5
33-16	0	1	1	0	0
38-11	0	0	1	0.3	0
4226	0	1	1	0.5	0

3.1.9 Square Numerical Matrix（数值方阵）

亲缘关系可以用外部软件计算，比如利用 SAS Proc Inbreeding 从系谱计算[18]，或者利

用其他软件从标记计算。用下面的格式来导入得到的亲缘关系估计值：

如果 n 代表分类单元的数目，则亲缘关系文件的格式如下：

```
n
Taxa1Name   r11  r12  ...  r1n
Taxa2Name   r21  r22  ...  r2n
...
TaxanName   rn1  rn2  ...  rnn
```

这里 rij（i, j= 1, 2, ..., n）是亲缘关系矩阵中位于第 i 行和第 j 列的元素。亲缘关系矩阵不允许有缺失值。

重要提示：当前的格式与 TASSEL 2.0 或更低的版本中使用的格式不同。

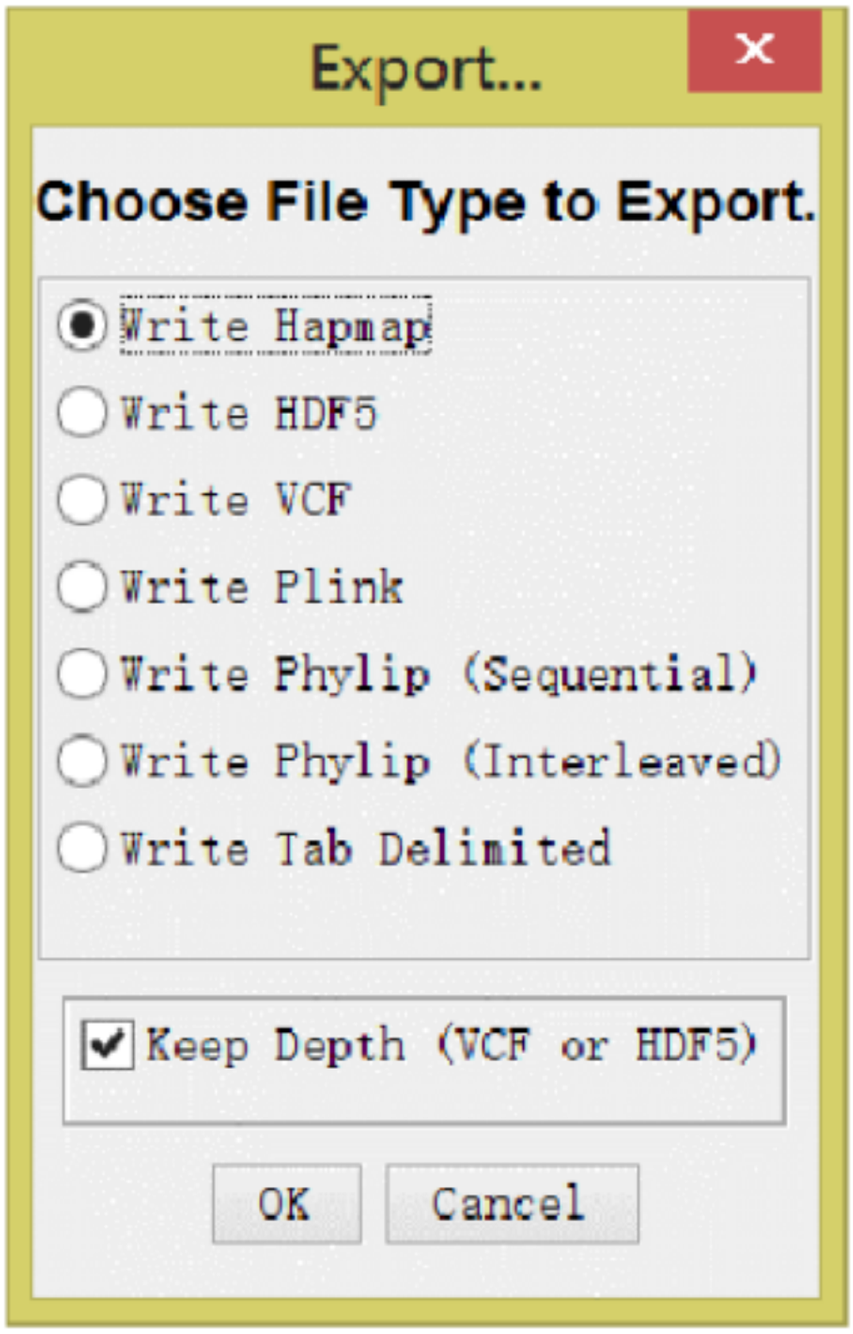
3.1.10 Table Report（表格报告）

数据可以作为制表符分隔的文本文件导入。文件的第一行将被解释为列标签，剩余的行为表格中的行。

3.1.11 TOPM（Tags on Physical Map，物理图谱上的标签）

3.2 Export 导出

提供了选项来导出序列数据：Hapmap、Plink、Phylip（顺序的或间隔的）。表现型和协变量数据被作为数值的性状数据导出。表格报告（Table Reports）被作为一个制表符分隔的表格导出。对于数值数据，导出（Export）的功能与结果（Results）模式中的表格（Table）功能相似。

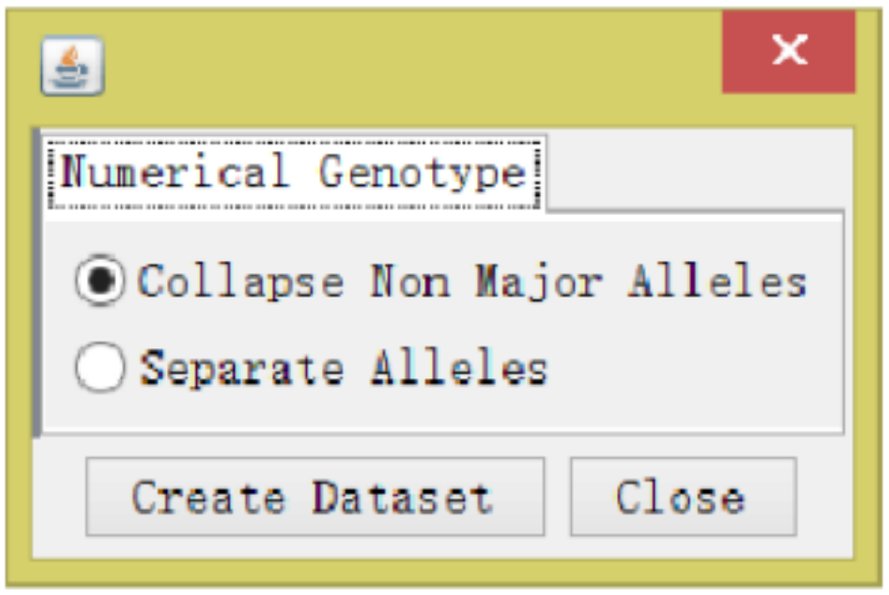


3.3 转换（Transform）

这一组功能允许对基因型和（数值的）表现型数据进行多个数据操作。当一个基因型数据集被选择时，数据被转换为数字。当一个数值数据集被选择时，可以进行数学的转换、数据估算以及主成分分析（PCA）。在一个数据（Data）对话框中将显示转换列（Transform columns）标签页，具有三个标签：Trans、Impute 和 PCA。

3.3.1 Genotype Numericalization（基因型数字化）

提供了两个选项来将基因型从字符转换为数值，如下面的对话框所示。



3.3.1.1 Collapse Non Major Alleles（折叠非主要等位基因）

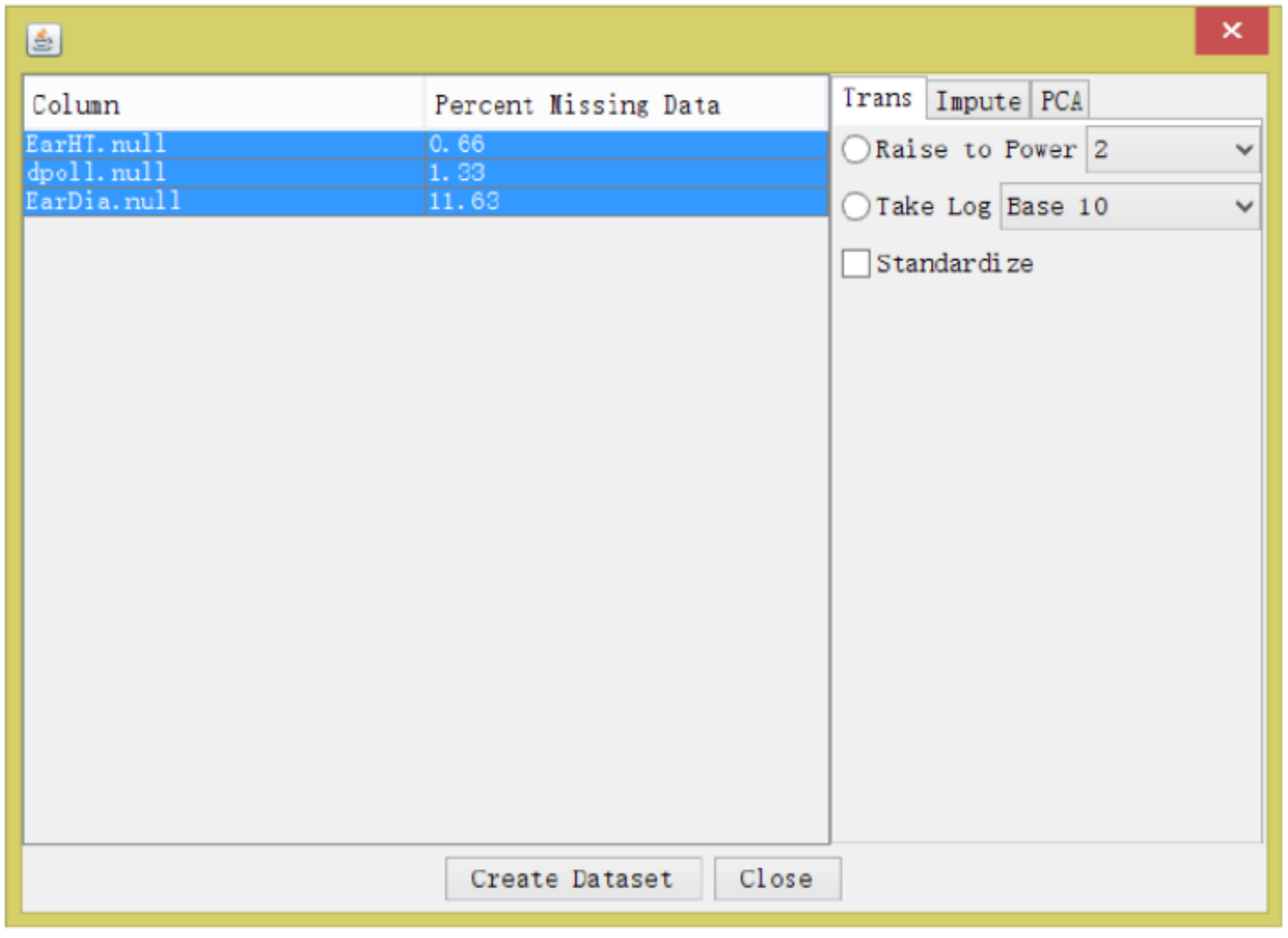
这个功能把 1 分派给主要等位基因，把 0 分派给任何其它的等位基因。转换的基因型被保存在一个新的数值数据集中。

3.3.1.2 Separate Alleles 分离等位基因

这个功能对每个等位基因指定一个指示符（存在为 1，不存在为 0）。转换的基因型被保存在一个新的数值数据集中。

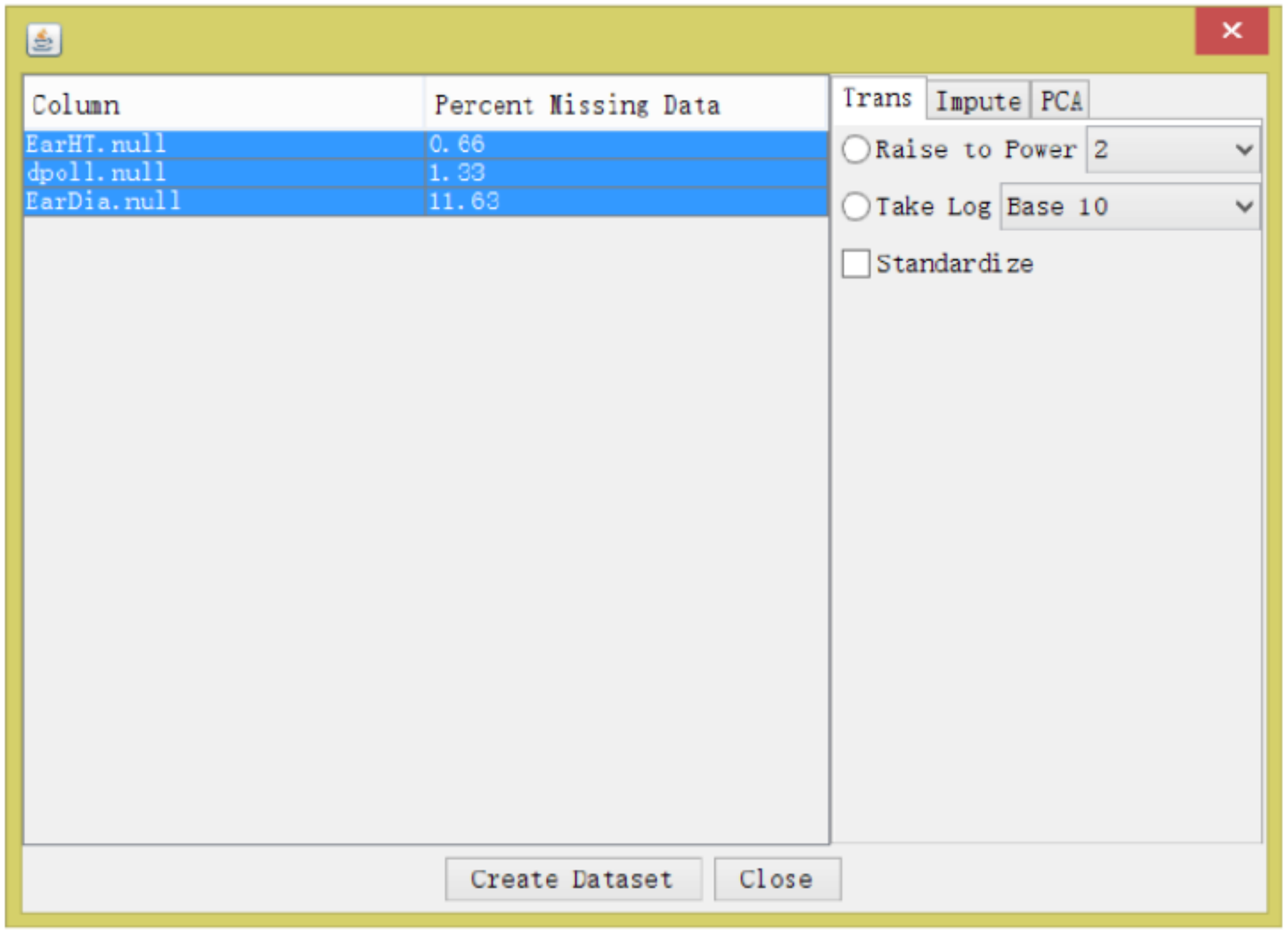
3.3.2 Transform and/or Standardize Data 转换和/或标准化数据

Trans 对话框是默认的选择，如同下面所示。在列（Column）列表中，选择你想要转换的一列或几列。然后选择你想要执行的转换类型。选择标准化（Standardize）复选框将通过从性状值中减去列平均数然后除以该列的标准差来转换数据。单击产生数据集（Create Data set）按钮将在数据树中产生一个只包含选定列的数据集的位置。



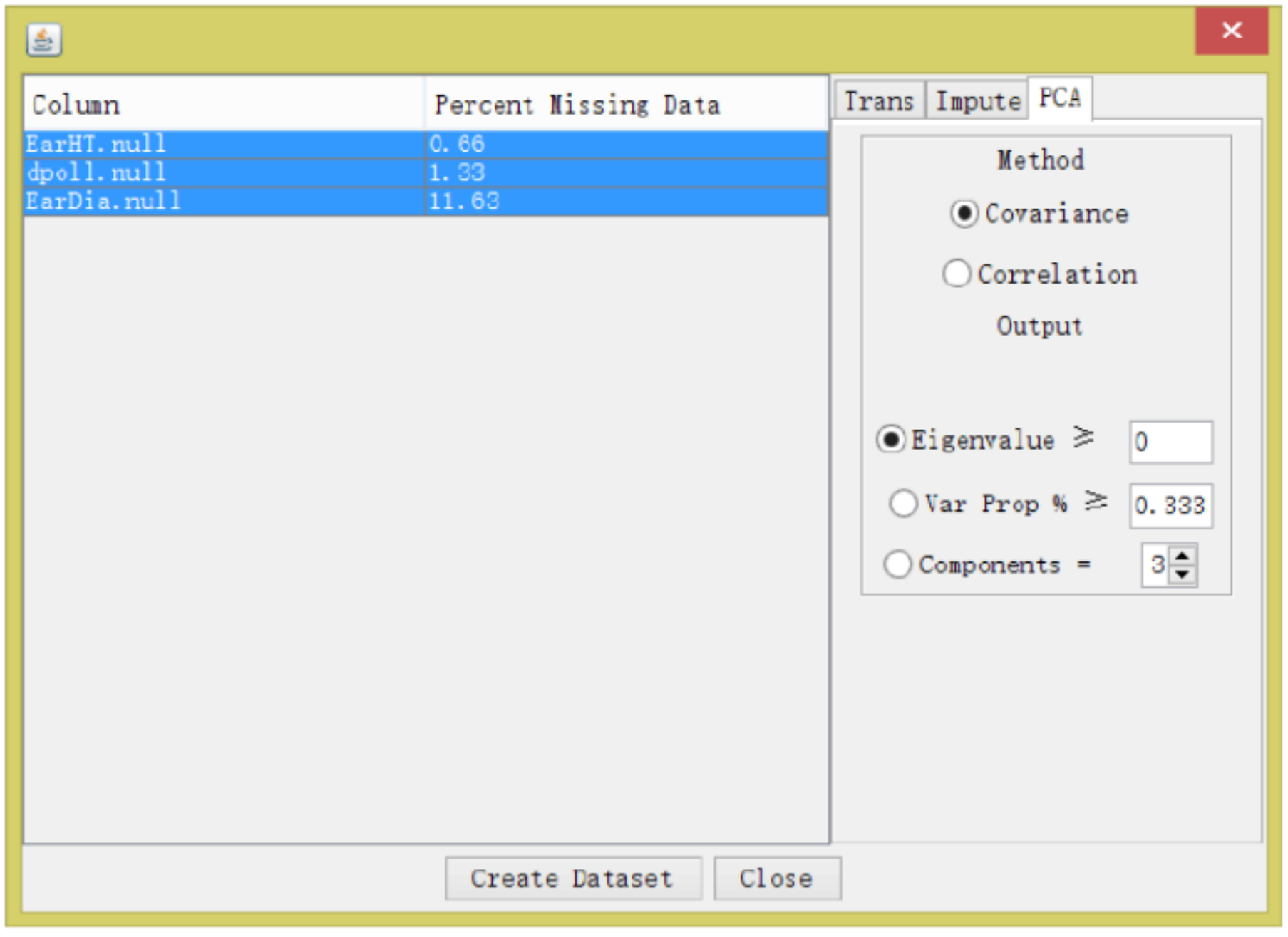
3.3.3 Impute Phenotype 估算表现型

k-最近邻算法（k-nearest-neighbor algorithm）[20]被用来估算缺失的表现型数据。如果数据是对一个分类单元的性状之一缺失，该算法寻找与它最像的其他分类单元（邻近者）来估计缺失的性状。它运用邻近者的平均数来估算缺失数据。单击估算（Impute）标签来显示以下对话框 ：



3.3.4 PCA（主成分分析）

只能对没有缺失值的数值数据集进行主成分分析（PCA）。有两种方法：相关系数（correlation）或协方差。这确定是用相关系数矩阵还是用协方差矩阵将作为该分析的基础。默认为相关系数，这是遗传数据的一个合理的选择。通过选择以下任何一个，可以控制输出数据集中 PCA 轴的数目：与每个轴有关的最小的特征值，由一个轴捕获的方差的最小百分率，或者轴的数目。产生的轴将按照每个捕获的方差的大小排序。

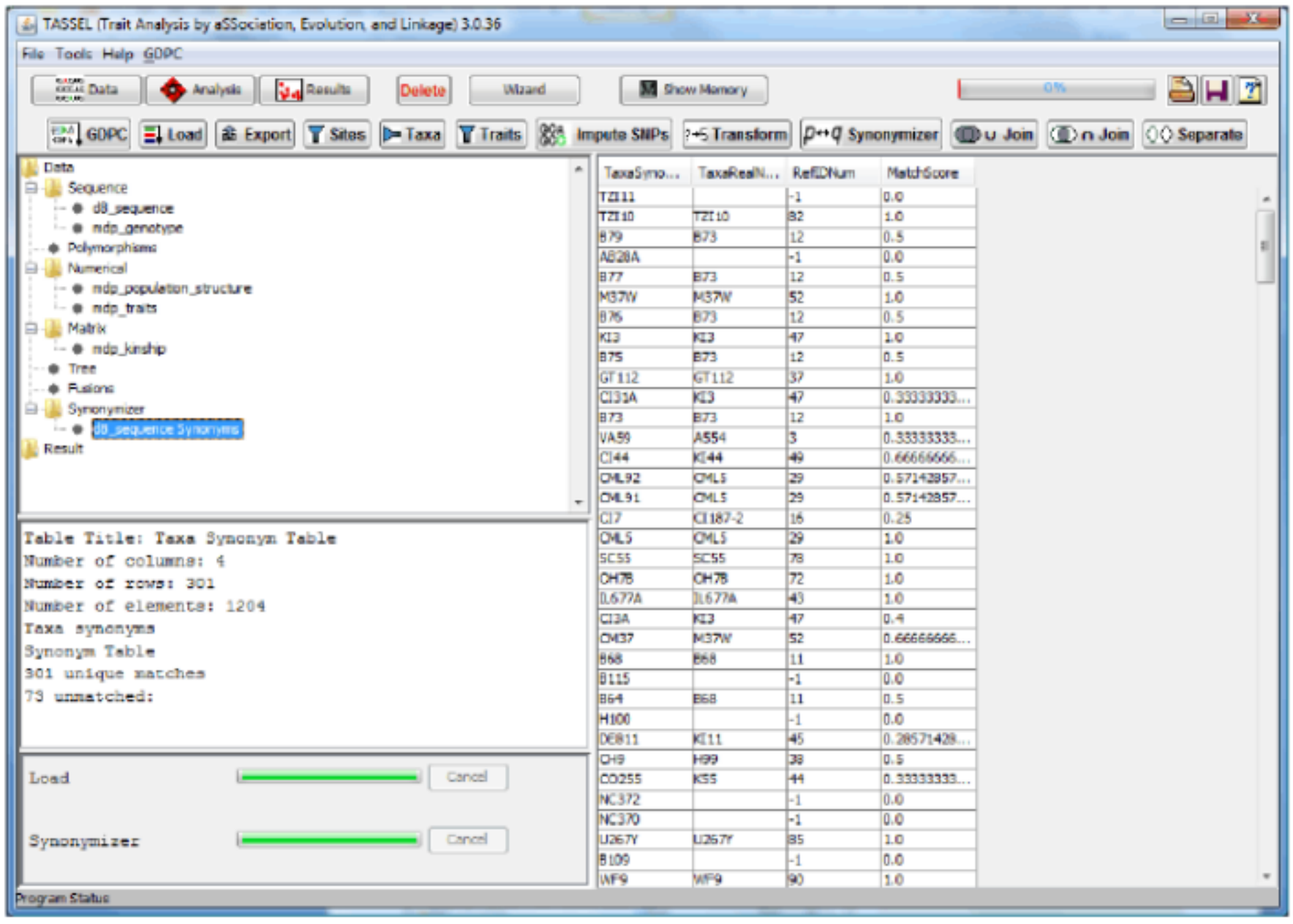


3.4 Synonymizer（举出分类单元名称的同义词）

这个按钮使分类单元名称统一，以便可以进行数据集的合并。

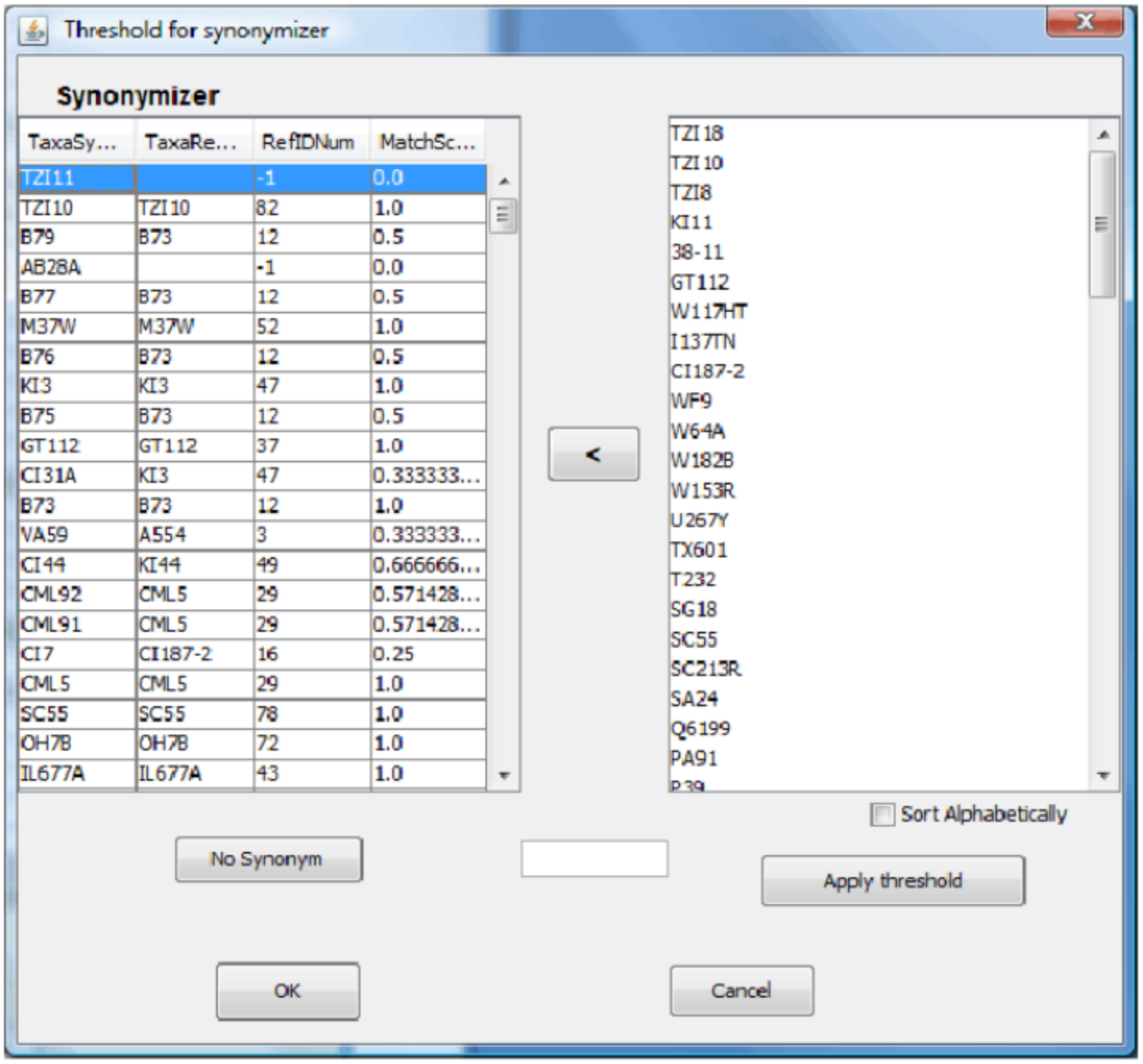
合并功能产生合并的数据集，它是通过匹配分类单元名称来进行的。因而，如果一个给定的分类单元存在多个名称（一个添加的后缀、另一种拼写、不同的命名规则、等等）那么两个数据集就不能被正确地合并。为了帮助补救这个问题，Synonymizer 功能允许用一个数据集的分类单元名称来替换第二个数据集中相似的分类单元名称。它依赖于一个算法来计算名称之间的相似度，利用来自第一个数据集、而与第二个数据集最相似的名称。

使用 Synonymizer 时要记住，选择的次序对结果有影响。总是首先选择具有你想要使用的名称（“real”名称）的数据集，然后，在按住 CTRL 键的同时，单击你想要改变其分类单元名称（“synonym”）的第二个数据集。然后单击 Synonymizer 按钮。一个同义词数据集将被放在数据树面板上，在 Synonyms（同义词）下。数据集中第二次选择的每个名称现在被列在 TaxaSynonym 列中。紧挨着这一列的是一个 TaxaRealName 列，列出由“real”名称数据集获得的最高得分匹配。MatchScore 列表示两个名称之间相似性的大小（其中 0 是没有相似性，1.0 是完全相同）。



注意! 在使用同义词之前，我们强烈地建议用户核对匹配得分，尤其对于那些匹配得分低的分类单元。要那么做，用户选择同义词文件，然后单击“Synonymizer”按钮。此时，不正确的匹配（通常是匹配得分低的那些）可以被拒绝。首先对匹配得分列排序可以使这个过程相当容易。

如果一些分类单元没有被正确地解释，可以手工地修改匹配。在左边选择你想要修改的分类单元，然后从右侧选择一个替换分类单元。单击箭头按钮 < 来替换该分类单元。没有同义词的分类单元可以通过选择然后单击“No Synonym”被识别。单击 OK 来保存改动。



一旦确定了分类单元名称是正确匹配的，就可以使用同义词了。选择了同义词以后，按下 CTRL 键同时单击第二个/同义词数据集（你想要改变其名称的数据集）。然后再一次单击 Synonymizer 按钮来把新的名称应用到数据集。

3.5 Intersect Join（交集合并）

命令

./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt -combine3 -input1 -input2 -intersect -export group1_group2_intersect.hmp.txt -runfork1 -runfork2 -runfork3

这个命令按照分类单元的交集合并多个数据集。分类单元必须在两个要被包含的数据集中都存在。利用 CTRL 键结合鼠标单击选择多个数据集，然后单击交集（intersection）按钮

来合并数据集。因为这个功能运用分类单元名称来合并数据集，分类单元名称中的任何变异都可能妨碍正确的合并。利用“Synonymizer”可以使分类单元名称统一。

3.6 Union Join（并集合并）

命令

./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt -combine3 -input1 -input2 -intersect -export group1_group2_union.hmp.txt -runfork1 -runfork2 -runfork3

这个命令按照分类单元的并集合并多个数据集。如果分类单元从一个数据集中缺失，则将插入缺失数据。利用 CTRL 键结合鼠标单击选择多个数据集，然后单击并集（union）按钮来合并数据集。因为这个功能运用分类单元名称来合并数据集，分类单元名称中的任何变异都可能妨碍正确的合并。利用“Synonymizer”可以使分类单元名称统一。

3.7 Merge Genotype Tables（合并基因型表格）

命令

./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt -combine3 -input1 -input2 -intersect -export group1_group2_merge.hmp.txt -runfork1 -runfork2 -runfork3

其他选项（使用这些选项（即 after -Xmx5g）进行更多的控制。）

- -retainRareAlleles true | false

这定义是否保留稀有等位基因。如果为 false，超过 14 个较低频率的等位基因将被改变为未知的（Unknown）。如果为 true，则它们将被改变为 Z（Rare，稀有的）。这个选项对核苷酸数据不起作用。

- -exportType Hapmap | HapmapDiploid

在-export <filename>标签之后使用这个选项。如果为 Hapmap，则按照 IUPAC（国际理论和应用化学联合会）代码用单个字符代表杂合的和纯合的二倍体值。如果为 HapmapDiploid，则二倍体值被作为两个字符输出。

注释

- 不明确的分类单元 / 位点等位基因值被设置为未知的（UNKNOWN）。
- 重复的分类单元 / 位点设置为最后处理的校准（Alignment）。
- 这映射到图形用户界面上的“Data -> Merge Genotype Tables”菜单。
- 如果同一文件中有重复的位点名称则会出错。
- 位点是通过基因座（Locus）、物理位置（Physical Position）和位点名称（Site Name）识别的

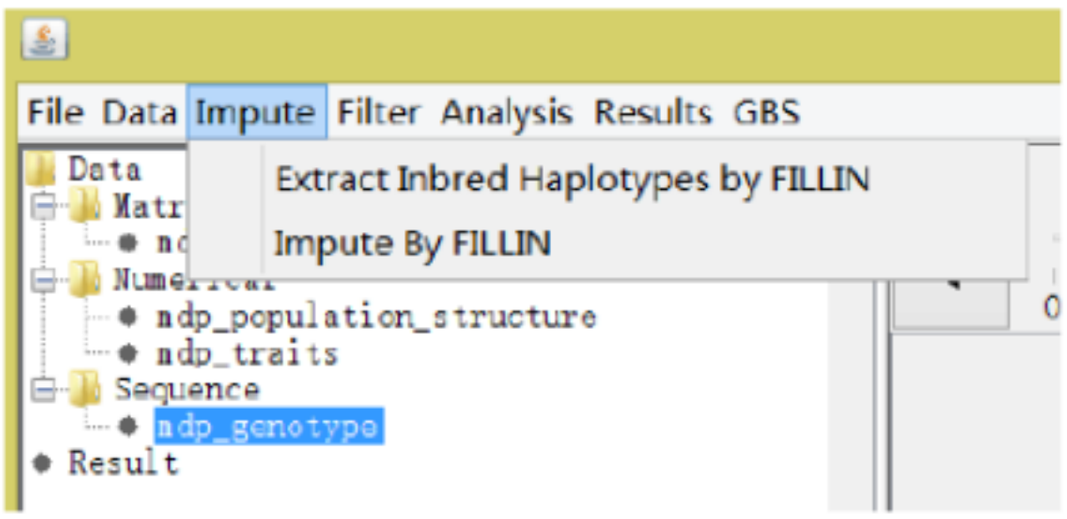
3.8 Separate（分离）

这个命令把选择的数据集分离成它的组分。例如，一个基因型表格将被分离成个别的染色体。

3.9 Homozygous Genotype（纯合的基因型）

这把所有杂合的值改变为未知的（N）。

4 Impute（估算）菜单



TASSEL5 包括两个估算缺失的基因型信息的方法，一个是一般化的方法，适合于各种类型的群体，但是为那些具有较高的近交系数的群体进行了优化（FILLIN），另一个是专门为发现全同胞家系中的重组断点优化过的（FSFHap）。在下面文献中可以找到有关这两种方法的更多信息：

Swarts et al. FSFHap (Full-Sib Family Haplotype Imputation) and FILLIN (Fast, Inbred Line Library ImputatioN) optimize genotypic imputation for low-coverage, next-generation

sequence data in crop plants, Plant Genome, in review.

FSFHap（Full-Sib Family Haplotype Imputation，全同胞家系单倍型估算）：

FSFHap 估算全同胞家系中缺失的基因型并纠正近交个体的基因型鉴定错误。它对于低覆盖度的 GBS 数据中的单倍型的调用（calling）是非常有用的。个体必须是近交的，至少部分近交，因为该方法依赖于发现近交的片段（inbred segments）来识别单倍型。它不直接使用亲本的基因型，但是将亲本包括在内对于结果的解释可能是有用的。

系谱文件格式：

对 FSFHap 专化的唯一的文件格式是系谱文件。分类单元名称必须与基因型数据中的名称准确地匹配。如果基因型数据包含系谱文件中没有包含的分类单元，则只有系谱文件中列出的个体将被分析。输入基因型可以是 TASSEL 接受的任何格式。系谱文件必须包含要被分析的个体的分类单元名称、每个个体所属的家系、亲本、亲本贡献、以及平均的近交系数。文件的第一行必须为列标题。列中的数值应该用制表符分隔，并且应该按以下顺序：家系、分类单元、亲本 1、亲本 2、亲本 1 贡献、亲本 2 贡献、F。F 值不是必需的，但是其他所有列都是必需的。

例子：

family	taxonName	parent1	parent2	contribution1	contribution2	F
fam1	t0001	par1	par2	0.5	0.5	.92
fam1	t0002	par1	par2	0.5	0.5	.92
...						
fam2	t0201	par1	par3	0.5	0.5	.92
fam2	t0202	par1	par3	0.5	0.5	.92
fam2	t0203	par1	par3	0.5	0.5	.92

贡献 1、贡献 2 和 F 的值是家系平均数。那些值是从一个家系的第一行中读取，然后应用于整个家系。

使用 FSFHap 的命令行：

FSFHap 包括三个 TASSEL 插件：CallParentAllelesPlugin，ViterbiAlgorithmPlugin，WritePopulationAlignmentPlugin，它们是按顺序调用的。运行 FSFHap 的一个典型的命令如

下（以实际参数值替换<>中的项）：

run_pipeline.pl -h <genotypeFilename> CallParentAllelesPlugin -p <pedigreeFilename> -m 0.9 -r 0.5 -logfile <logFilename> -endPlugin -ViterbiAlgorithmPlugin -g true -endPlugin WritePopulationAlignmentPlugin -f <outputFilename> -m false -o parents -endPlugin

CallParentAllelesPlugin 的选项：

选项所取的参数值用“Value = []”表示：

-p 或-pedigrees：系谱文件。Value = [文件名]

-w 或-windowSize：每个 LD 簇（LD cluster）要考察的 SNPs 的数目。Value = [整数] (默认 = 50)

-r 或-minR：用来按照 LD 过滤 SNPs 的最小的 R。Value = [0 和 1 之间的数字]。(默认 = 0.2，use 0 for no ld filter)

-m 或-maxMissing：一个 SNP 容许的缺失数据的最大比例。Value = [0 和 1 之间的数字]。(默认 = 0.9)

-f 或-minMaf：用来过滤 SNPs 的最小的次要等位基因频率。如果为负数，则按照根据亲本的贡献预期的分离比例过滤。Value = [1 和-1 之间的数字]。(默认 = -1)

-b 或-bc1：使用 BC1 专化的过滤器。Value = [true 或 false] (默认 = true)

-n 或-bcn：使用多个回交专化的过滤器。Value = [true 或 false] (默认 = false)

-logfile：日志信息文件的名称。Value = [文件名].

不取参数值的选项：

-cluster：使用聚类算法。minMaf 默认为 0.05.

-subpops：过滤亚群体中的杂合位点。

-nohets：估算前从原始数据中删除杂合位点（het calls）。

“-cluster”、“-subpops”和“-nohets”选项不取参数，但是只作为在分析中包含某些特性的标签起作用。对于那些分析，聚类（cluster）是最有用的。当使用聚类选项时，将使用一个不同的算法，它可以更好地处理亲本中剩余的杂合性。然而，对于仅仅自花受粉了一两个世代的部分近交的 RILs 来说，它的表现并不好。如果被估算的 RILs 是 F2 或 F3，则不应该使用“- cluster”选项。只有当输入由玉米多样性项目（Maize Diversity Project）培育的 NAM

群体的家系时才应该使用“-subpops”选项。包含“-nohets”选项是为了检验错误的 het calls 是否导致太多的 hets 被估算。看起来它对结果影响很小。

推荐使用“-logfile”选项。输出可用于识别和诊断可能的问题。对于具有两个或更多回交的群体应使用“- bcn true”。然而，不一定需要使用“- bc1”选项，因为默认的行为通常是最好的。

ViterbiAlgorithmPlugin 的选项：

-g 或 fillgaps：如果为 true 则由来自同一亲本的 SNPs 侧翼的缺失值将被估算到那个亲本，如果为 false 则以另外方式估算。值= [true 或 false]（默认= true）

-h 或-phet：杂合基因座的期望频率。只有当系谱文件中未标明近交系数时才使用。值= [0 和 1 之间的数字]（默认= 0.07）

WritePopulationAlignmentsPlugin 的选项：

必需的：

-f 或-file：ouput..hmp.txt 将附加到其上的基本文件名。值= [文件名]

可选的：

-m 或-merge：如果为 true 则家系被合并成单个文件，如果为 false 则每个家系被输出到一个单独的文件。值= [true 或 false]（默认= true）

-o 或-outputType：如果值= parents（亲本），则输出亲本召唤（calls），如果值= nucleotides（核苷酸），则输出核苷酸，如果值= both，则在单独的文件中输出两者（默认= both）

-d 或-diploid：如果为 true 则输出是 AA/CC/AC，如果为 false 则输出是 A/C/M。值= [true 或 false]（默认= false）

-c 或-minCoverage：一个单态的 SNP 要被包括在核苷酸输出中的最低覆盖度。值= [0 和 1 之间的数字]（默认= 0.1）

-x 或-maxMono：用来调用单态 SNP 的最大的次要等位基因频率（默认= 0.01）

对于单独的家系，只估算多态的 SNPs。当 merge = false 时，只有那些 SNPs 出现在输出中。当 merge = true 时，在任何家系中为多态的 SNPs 将被存入到输出。对于任何位点，如果一个家系中的 SNP 覆盖度足够高，有把握确定它对于那个家系是单态的，那么那个家系中的全部个体将被估算到那个位点上的单态的值。-minCoverage 和-maxMono 选项用来确

定阈值，用于确定一个家系中的一个位点是否将被称作单态的。如果这些选项中的任何一个的值被设置为 NaN，那么单态位点上的缺失值不会被估算。

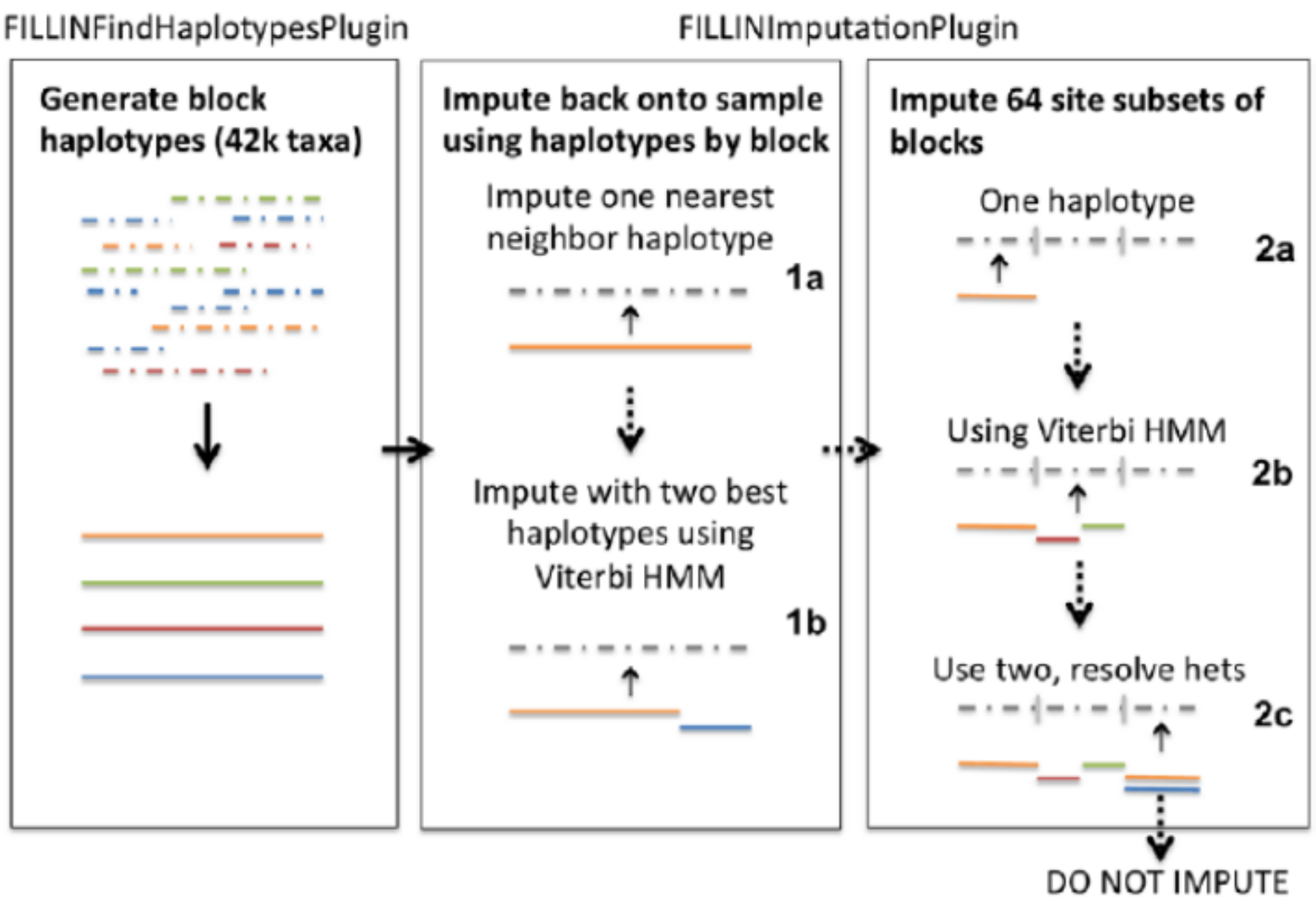
FILLIN（Fast, Inbred Line Library ImputatioN，自交系库的快速估算）：一般的方法

FILLIN 按照两个步骤估算缺失的基因型：1）单倍型世代（FILLINFindHaplotypesPlugin）；2）产生的单倍型向后估算到目标样本上（FILLINImputationPlugin）。

单倍型是通过折叠低覆盖度、但近交的片段产生的，这些片段通过位点窗口（默认：8k）按一个用户提供的、可选的阈值共享状态相同（identity by state）；这是通过第一个插件（FILLINFindHaplotypesPlugin）进行的。因为短的 IBD 片段在一个物种内（甚至在不同的个体之间）可能是广泛重复的，我们建议为这个步骤提供一个物种内的全部可用的信息。

第二个插件（FILLINImputationPlugin）使用这些单倍型来估算目标个体中缺失的基因型。它通过多个步骤来做到这一点。首先在整个位点窗口之内寻找将次要等位基因匹配到一个阈值的单倍型（在以下的图解中为 1a），然后，如果这个过程失败，则寻找两个单倍型来解释位点窗口，并且假定这代表两个近交的单倍型之间的一个重组断点，运用一种 Viterbi HMM 算法来对重组断点建模（2a）。如果不能找到两个单倍型来解释整个位点窗口，该算法接下来搜索单倍型来解释一个较小的焦点窗口，在该位点窗口之内每次以 64 个位点为中心，向右边和左边搜索直到找到足够的信息性的次要等位基因。它首先按一个阈值寻找一个单倍型（2a），然后寻找对近交的片段之间的一个重组断点建模的两个单倍型（2b），最后按一个更高的阈值寻找两个单倍型，并对 64 个焦点位点窗口作为杂合位点建模，将两个单倍型结合在一起。2a - c 的阈值也被不同地设置，根据目标分类单元的整个序列是否超过或低于一个用户提供的杂合性阈值。对于被看作异交（超过阈值）的分类单元，Viterbi 选项 2b 从来不被使用，因为在一个异交的分类单元中如果两个单倍型解释一个片段则那个两个单倍型更可能是杂合的。如果该算法找不到满足这些阈值要求的单倍型，则片段不会被估算。焦点块（focus block）估算的阈值是根据输入的 mxInbErr 和 mxHybErr 值设置的（或默认）：

	Below mxHet (inbred)	Above mxHet (outbred)
2a	3/10*mxInbErr	1/10*mxInbErr
2b	1/3*mxHybErr	0
2c	mxInbErr	mxInbErr



运行 FILLIN:

FILLIN 包括两个 TASSEL 插件，FILLINFindHaplotypesPlugin 和 FILLINImputationPlugin，它们是按顺序调用的。如果你想掩蔽你的数据并计算准确性，则使用 FILLINImputationPlugin 的 -accuracy 标签。如果估算玉米，则可以在 Panzea 网址（http://www.panzea.org/lit/data_sets.html）上找到一个来自 40k+分类单元的单倍型的供体文件。FILLIN 可以在 TASSEL 图形用户界面之内运行，也可以通过命令行运行。两者的选项是一样的。

通过命令行运行 FILLIN 的一个典型的命令序列如下（以实际参数值替换<>中的项）：

run_pipeline.pl FILLINFindHaplotypesPlugin -hmp <genotypeFilename> -o <outDonorFile.gX.hmp.txt>

run_pipeline.pl -FILLINImputationPlugin -hmp <genotypeFilename> -d <outDonorFile.gX.hmp.txt> -o <outFile.hmp.txt.gz>

要从图形用户界面运行 FILLIN，转到 Impute -> FILLINFindHaplotypesPlugin 或 FILLINImputationPlugin

FILLINFindHaplotypesPlugin 的选项：

-hmp <目标文件>：输入基因型来从中产生单倍型。通常最好使用来自一个物种的全部可用的样本。接受由 TASSEL5 支持的所有文件类型（必需的）

-o <供体文件夹/文件基本名称>：输出文件目录名称，或新建目录路径；目录将被创建，如果不存在的话。输出文件将被放在该目录中，并且给予相同的名称附加子串“.gc#s#.hmp.txt”来表示染色体和部分（section）（必需的）。

-mxDiv <来自创始者的最大的遗传趋异>：从创始者单倍型到聚类序列（cluster sequences）的最大的遗传趋异（默认：0.01）

-mxHet <Max heterozygosity of output haplotypes>：输出单倍型的最大杂合性。杂合性来自聚类的序列，这些序列要么具有剩余的杂合性，要么具有不共享全部次要等位基因的聚类的序列。（默认：0.01）

-minSites <要聚类的最少位点>：要比较遗传距离以便评价相似性用于聚类的两个分类单元中存在的位点的最小数目（默认：50）

-mxErr <估算两个供体的最大的合并误差>：容许对分类单元聚类的最大的遗传趋异（默认：0.05）

-hapSize <偏爱的单倍型大小>：位点中首选的单倍型块大小（最小为 64）；将使用等于或低于提供的值 64 的最接近的倍数（默认：8192）

-minPres <要对匹配进行检验的最少位点>：要进行搜索的输入序列内存在的位点的最小数目（默认：500）

-maxHap <每个片段的最大单倍型>：每个片段单倍型的最大数目（默认：3000）

-minTaxa <产生一个单倍型的最少分类单元>：产生一个单倍型的分类单元的最小数目（默认：2）

-maxOutMiss <缺失的每个单倍型的最大频率>：输出单倍型中缺失数据的最大频率（默认：0.4）

-nV <true | false>：抑制系统输出（默认：false）

-extOut <true | false>：包括在系统输出的每个单倍型中的分类单元的详情（默认：false）

FILLINImputationPlugin 的选项：

-hmp <目标文件>：输入要估算的目标基因型的 HapMap 文件。接受 TASSEL5 支持的所有文件类型（必需的）

-d <供体文件夹>：包含来自 FILLINFindHaplotypesPlugin 的输出的供体单倍型的目录。

文件名中带有'.gc '的所有文件将被读取，只有那些具有匹配位点的被使用（必需的）

-o <输出文件名>：输出文件；接受 hmp.txt.gz 和.hmp.h5。（必需的）

-hapSize <偏爱的单倍型大小>：位点中首选的块大小（和在 FILLINFindHaplotypesPlugin 中一样使用）（默认：8000）

-hetThresh <杂合性阈值>：每个分类单元杂合性的阈值，用于把分类单元作为杂合的处理（不是 Viterbi，het 阈值）。（默认：0.01）

-mxInbErr <估算一个供体的最大误差>：最大误差比率，用于应用一个单倍型到整个位点窗口（默认：0.01）

-mxHybErr <估算两个供体的最大合并误差>：最大误差比率，用于将 Viterbi 应用到整个位点窗口（默认：0.03）

-mnTestSite <要对匹配进行检验的最少位点>：要对单倍型和焦点块中的目标之间的 IBS 进行检验的位点的最小数目（默认：20）

-minMnCnt <要比较的次要等位基因的最小数目>：搜索窗口中信息性次要等位基因的最小数目（或 10X 的次要等位基因）（默认：20）

-mxDonH <最大供体假说>：要被探索的供体假说的最大数目（默认：20）

-hybNN <true | false>：如果为 true，则在焦点区块中运用组合模式，否则不估算（默认：true）

-ProjA <true | false>：对高密度的标记产生一个投影校准（projection alignment）（默认：false）

-impDonor <true | false>：估算供体文件本身（默认：false）

-nV <true | false>：抑制系统输出（默认：false）

计算准确性的选项：

-accuracy <true | false>：在估算之前遮蔽输入文件，并以遮蔽的基因型为基础计算准确性（默认：false）

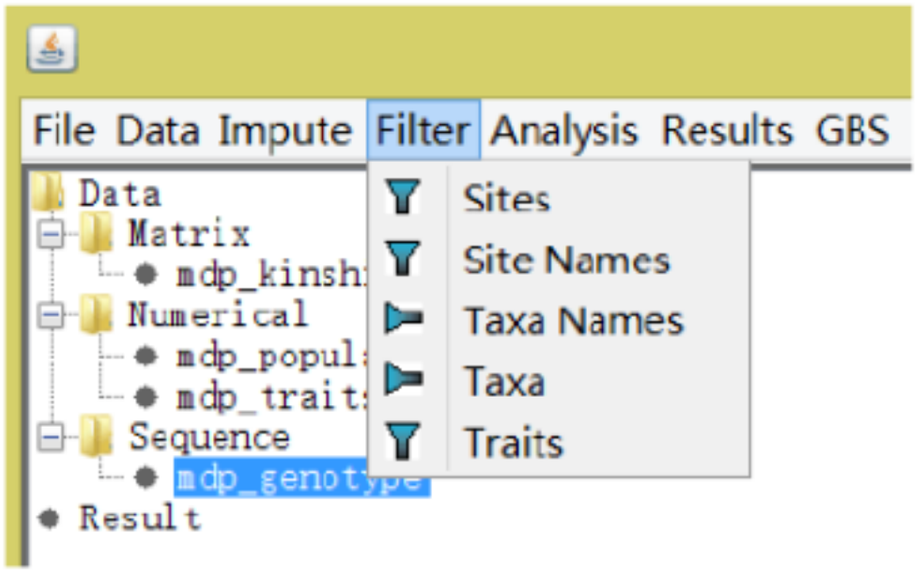
-propSitesMask <要遮蔽的基因型的比例，如果没有深度>：为计算准确性而遮蔽的基因型的比例，如果深度（depth）不可用的话（默认：0.01）

-depthMask <要遮蔽的基因型的深度>：为计算准确性而遮蔽的基因型的深度，如果深度信息可用的话（默认：9）

-propDepthSitesMask <要遮蔽的深度基因型的比例>：为计算准确性而遮蔽的给定深度

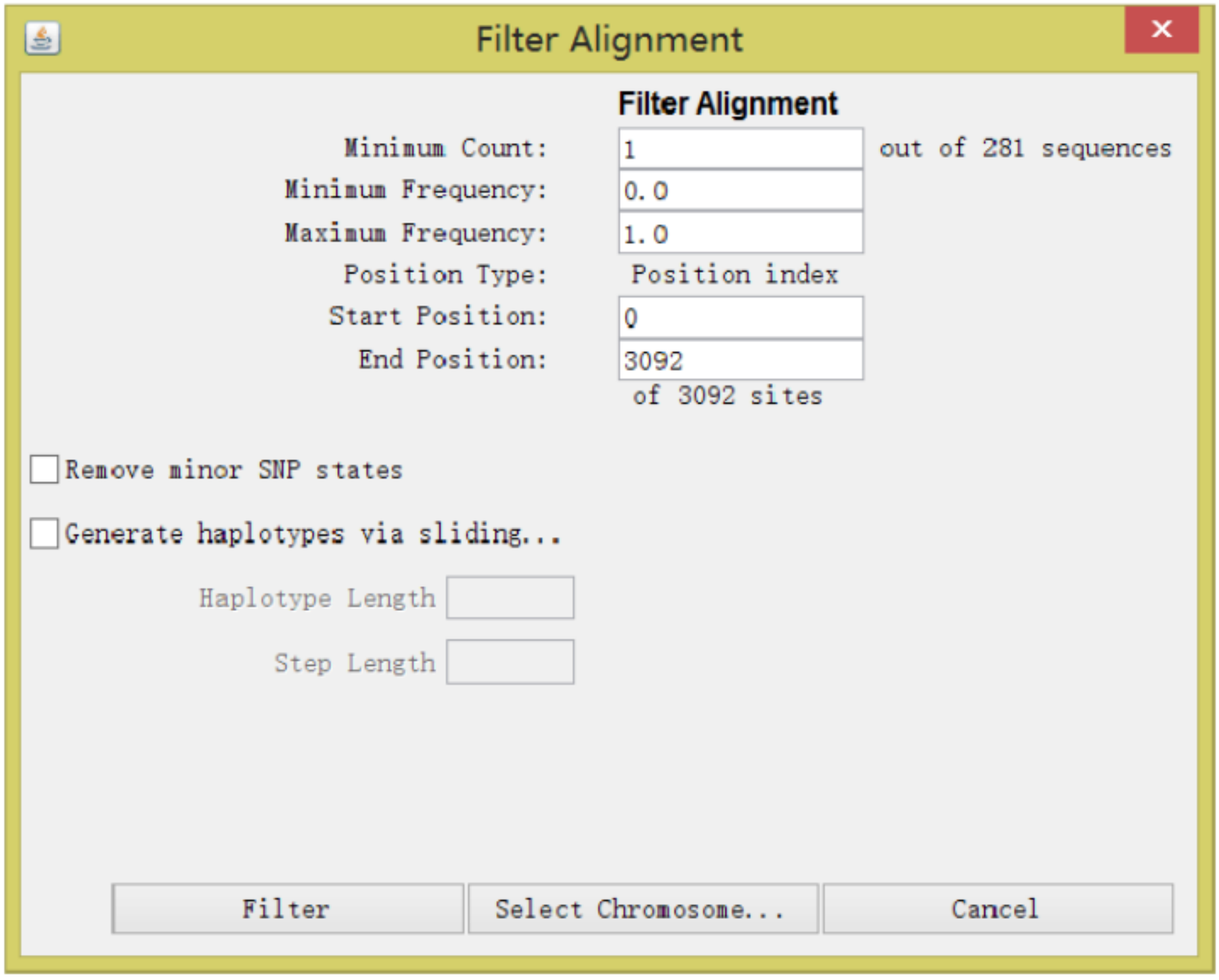
的基因型的比例，如果深度可用的话（默认：0.2）

5 Filter（过滤）菜单



5.1 Sites（位点）

基因型表格可以按照若干方法被过滤。例如，单态的位点可以被淘汰，并且可以淘汰一个序列的区域（一段序列？）。



Minimum Count（最小的计数）——分类单元的最小数目，在其中位点必须已经被评分以便被包括在过滤的数据集内（空隙（GAP）或缺失数据不计数）。

Minimum Frequency（最低的频率）——要被包括在过滤的数据集中的位点的少数多态性（minority polymorphism）的最低频率。

Start Position（开始位置）、End Position（结束位置）——确定用于过滤的位点的范围。

Extract Indels（提取 indels）——如果选择了，则从序列比对中提取 indels。如果未选择，则只提取点替换（point substitution）。

Remove minor SNP states（删除次要的 SNP 状态）——把第三的和稀有的状态转换成缺失数据（“?”），因此强制使位点在一个基因座上只有两个类型的分离的位点。这可以帮助消除测序的误差。

Generate haplotypes via sliding window（通过滑动窗口产生单倍型）——从 SNP 的一个有序集产生单倍型。

例子：删除 MAF（Minimum Allele Frequency，最低等位基因频率）小于 5%的 SNP 的

管道命令

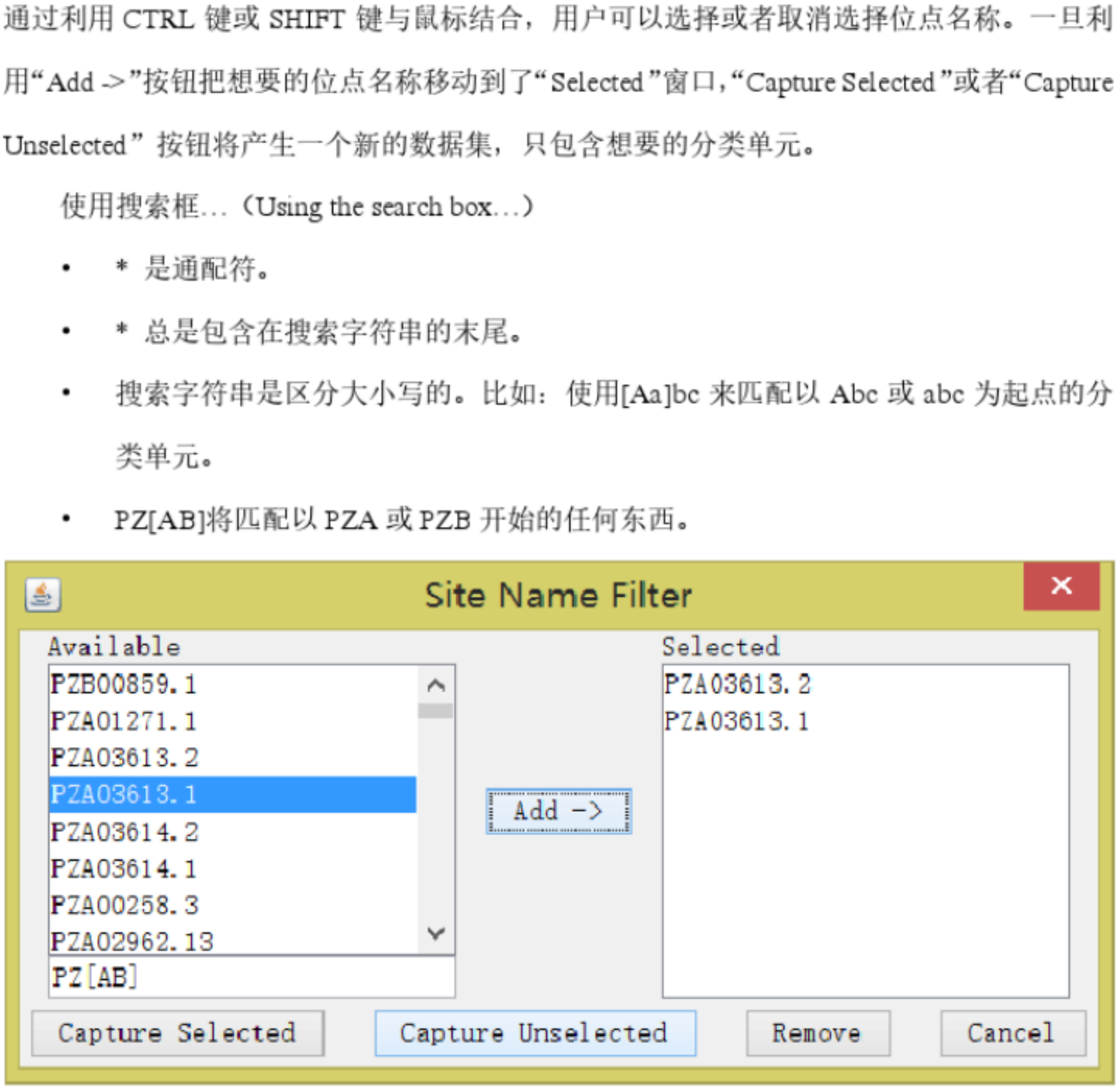
run_pipeline.pl -fork1 -h mdp_genotype.hmp.txt -filterAlign -filterAlignMinFreq 0.05 -export filtered_genotype -runfork1

5.2 Site Names（位点名称）

首先从数据树中选择基因型数据。产生的对话框显示与选择的数据有关联的位点名称。通过利用 CTRL 键或 SHIFT 键与鼠标结合，用户可以选择或者取消选择位点名称。一旦利用“Add ->”按钮把想要的位点名称移动到了“Selected”窗口，“Capture Selected”或者“Capture Unselected”按钮将产生一个新的数据集，只包含想要的分类单元。

使用搜索框…（Using the search box…）

- * 是通配符。
- * 总是包含在搜索字符串的末尾。
- 搜索字符串是区分大小写的。比如：使用[Aa]bc 来匹配以 Abc 或 abc 为起点的分类单元。
- PZ[AB]将匹配以 PZA 或 PZB 开始的任何东西。



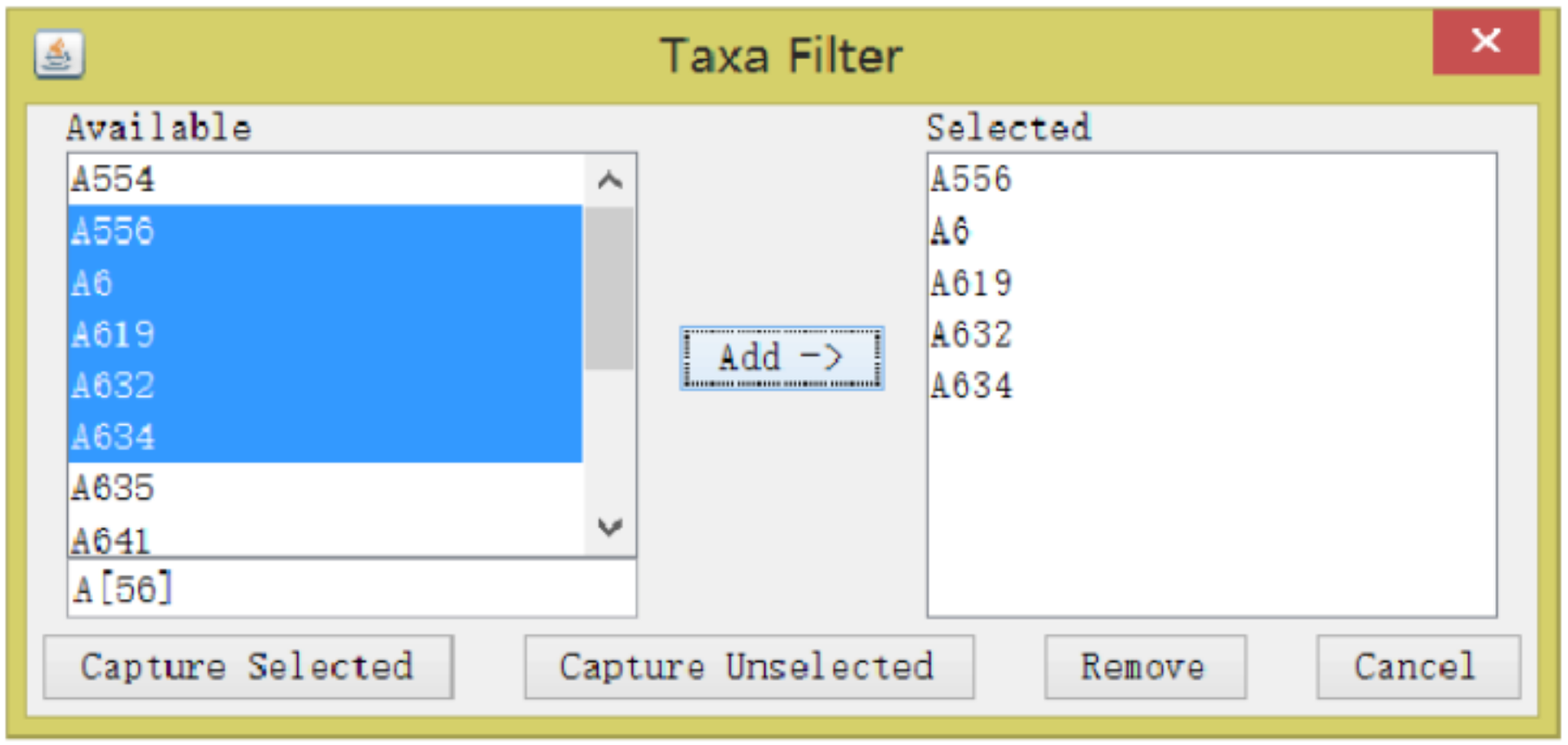
5.3 Taxa Names（分类单元名称）

首先从数据树中选择基因型、表型、或群体结构数据。产生的对话框显示与选择的数据有关联的分类单元。通过利用 CTRL 键或 SHIFT 键与鼠标结合，用户可以选择或者取消选

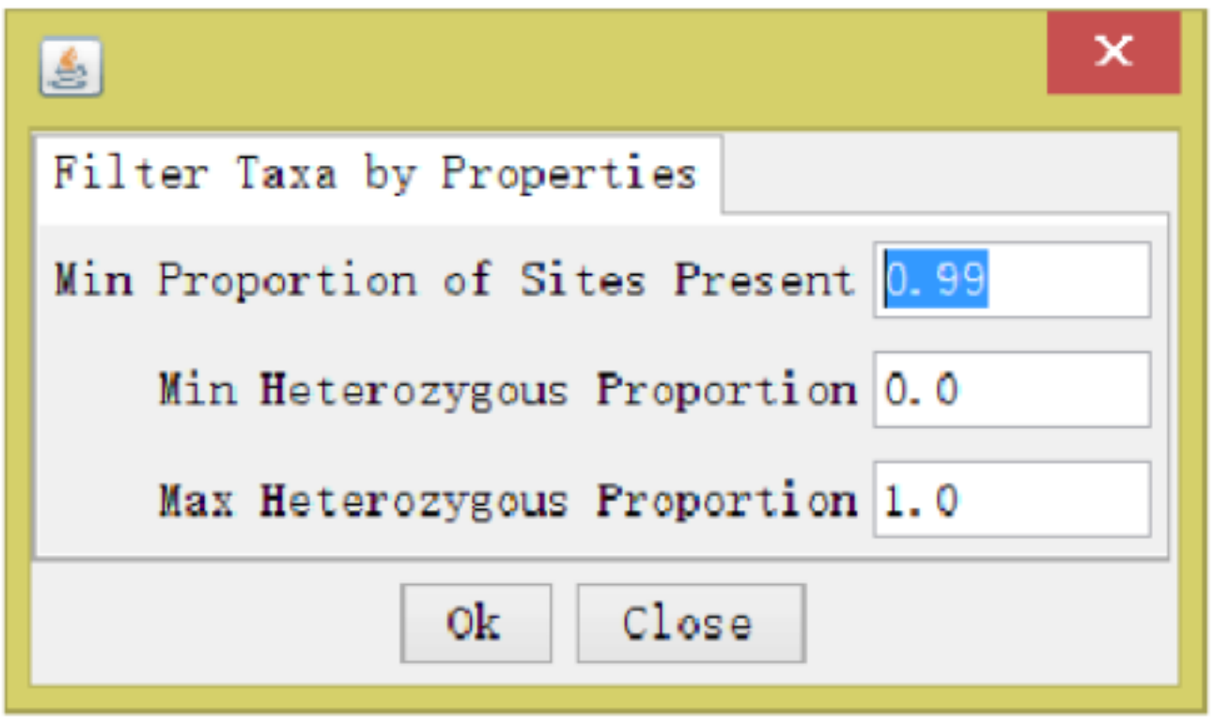
择分类单元。一旦利用“Add ->”按钮把想要的位点名称移到了“Selected”窗口，“Capture Selected”或者“Capture Unselected”按钮将产生一个新的数据集，只包含想要的分类单元。

使用搜索框…（Using the search box…）

- * 是通配符。
- * 总是包含在搜索字符串的末尾。
- 搜索字符串是区分大小写的。比如：使用[Aa]bc 来匹配以 Abc 或 abc 开始的分类单元。
- A[56]将匹配从 A5 或 A6 开始的任何东西



5.4 Taxa（分类单元）



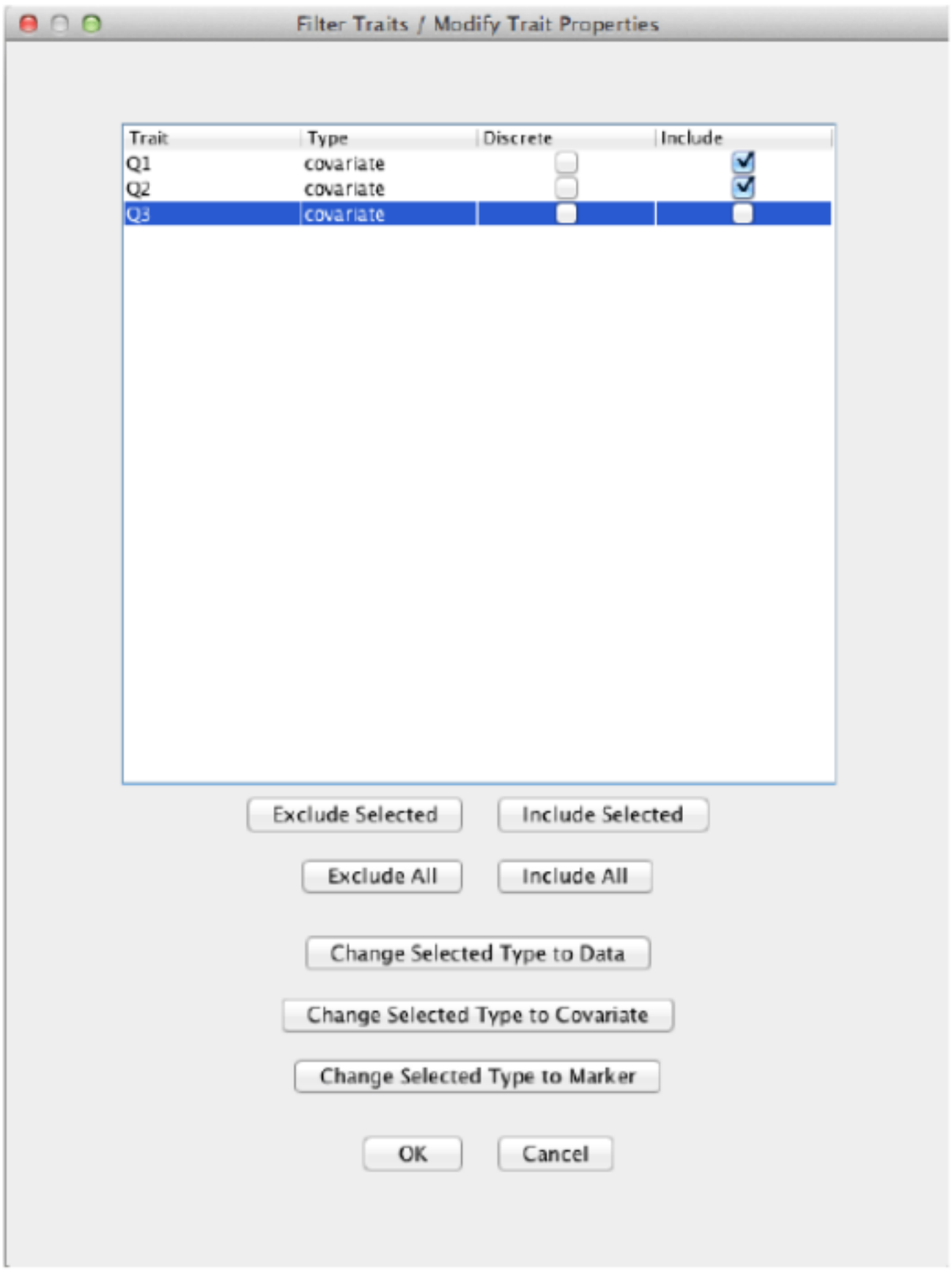
5.5 Traits（性状）

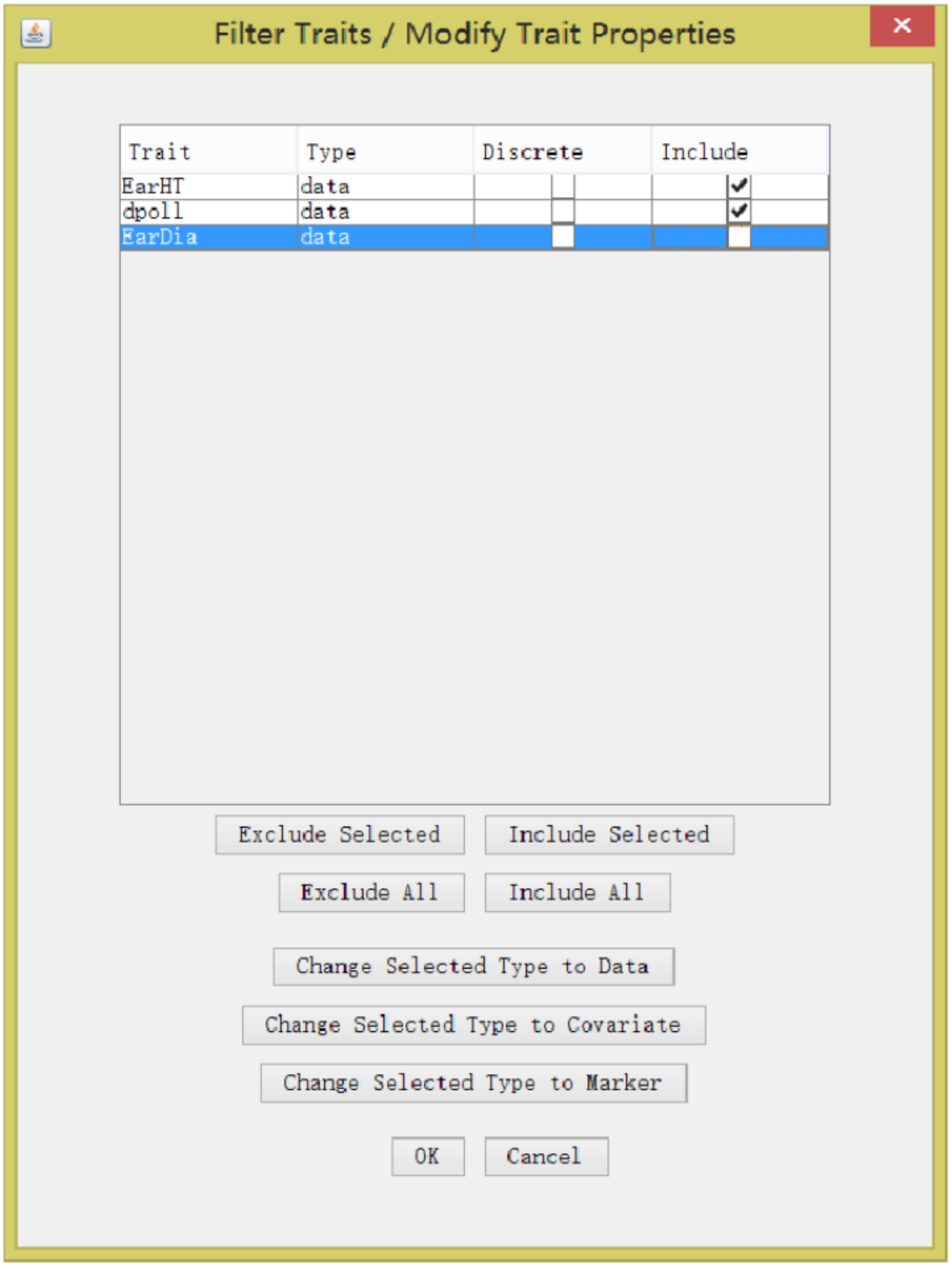
单击“Data”工具栏上的“Traits”按钮启动性状过滤（Trait Filter）对话框。这个对话框被用于数值数据集，（1）改变性状类型，（2）查看但是不改变性状是离散的还是连续的，（3）从数据集中删除一个或多个性状。此外，该对话框可用于查看性状的特性而不改变它们。如果单击了“OK”按钮，则产生一个包括了改变的新的数据集，原来的数据集没有变化，并且对话框关闭。如果单击了“Cancel”按钮，则不产生新的数据集，原来的数据集没有变化，并且对话框关闭。

容许的性状类型有数据、协变量、因子和标记。通常，数据和协变量性状是连续的，而因子是离散的。一个数值数据集中的标记将是连续的。离散值的标记最好是作为基因型被导入，并利用“Sites”过滤器过滤。

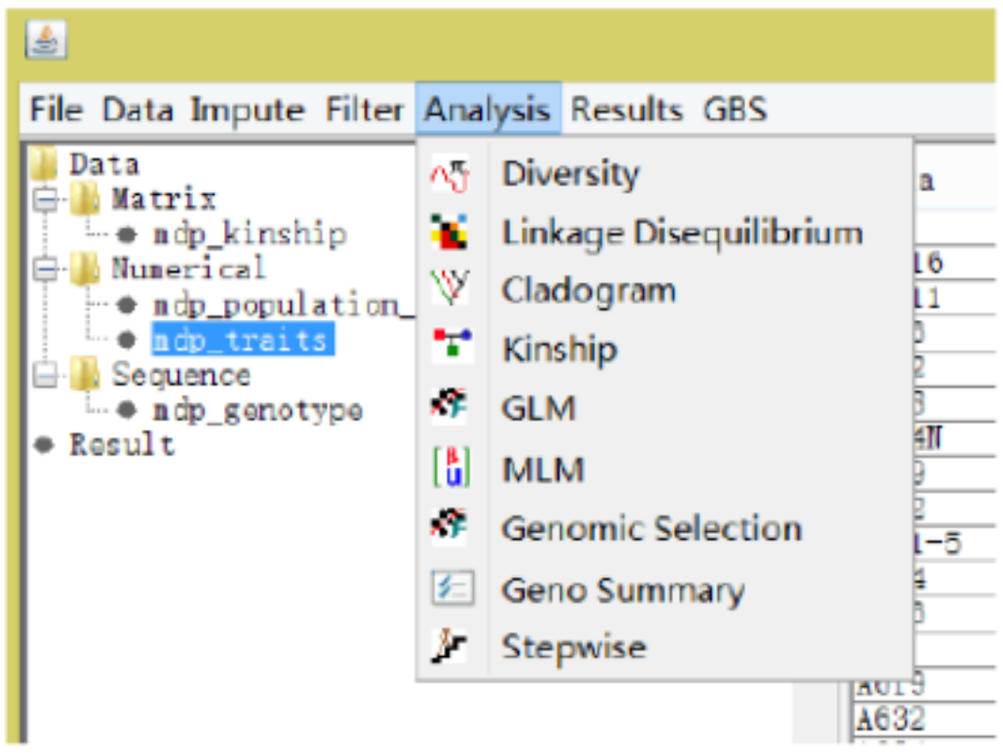
单击“Exclude All”对所有的性状清除“Include”框。单击“Include All”对所有的性状复选“Include”框。“Exclude Selected”和“Include Selected”按钮对通过鼠标选择加亮的那些性状做相同的事情。通过在那个性状的类型列的下拉框中选择一个值，可以对个别的性状改变类型。通过选择那些性状然后单击“Change Selected Type to …”按钮中的一个，也可以对多个性状改变类型。

重要提示：一旦一个数值数据集已经与基因型合并，它就不能再利用性状过滤功能加以修改。


Filter Traits / Modify Trait Properties
Trait
Type
Discrete
Include
Q1
covariate
Q2
covariate
Q3
covariate
Exclude Selected
Include Selected
Exclude All
Include All
Change Selected Type to Data
Change Selected Type to Covariate
Change Selected Type to Marker
OK
Cancel


Filter Traits / Modify Trait Properties
Trait
Type
Discrete
Include
EarHT
data
dpoll
data
EarDia
data
Exclude Selected
Include Selected
Exclude All
Include All
Change Selected Type to Data
Change Selected Type to Covariate
Change Selected Type to Marker
OK
Cancel

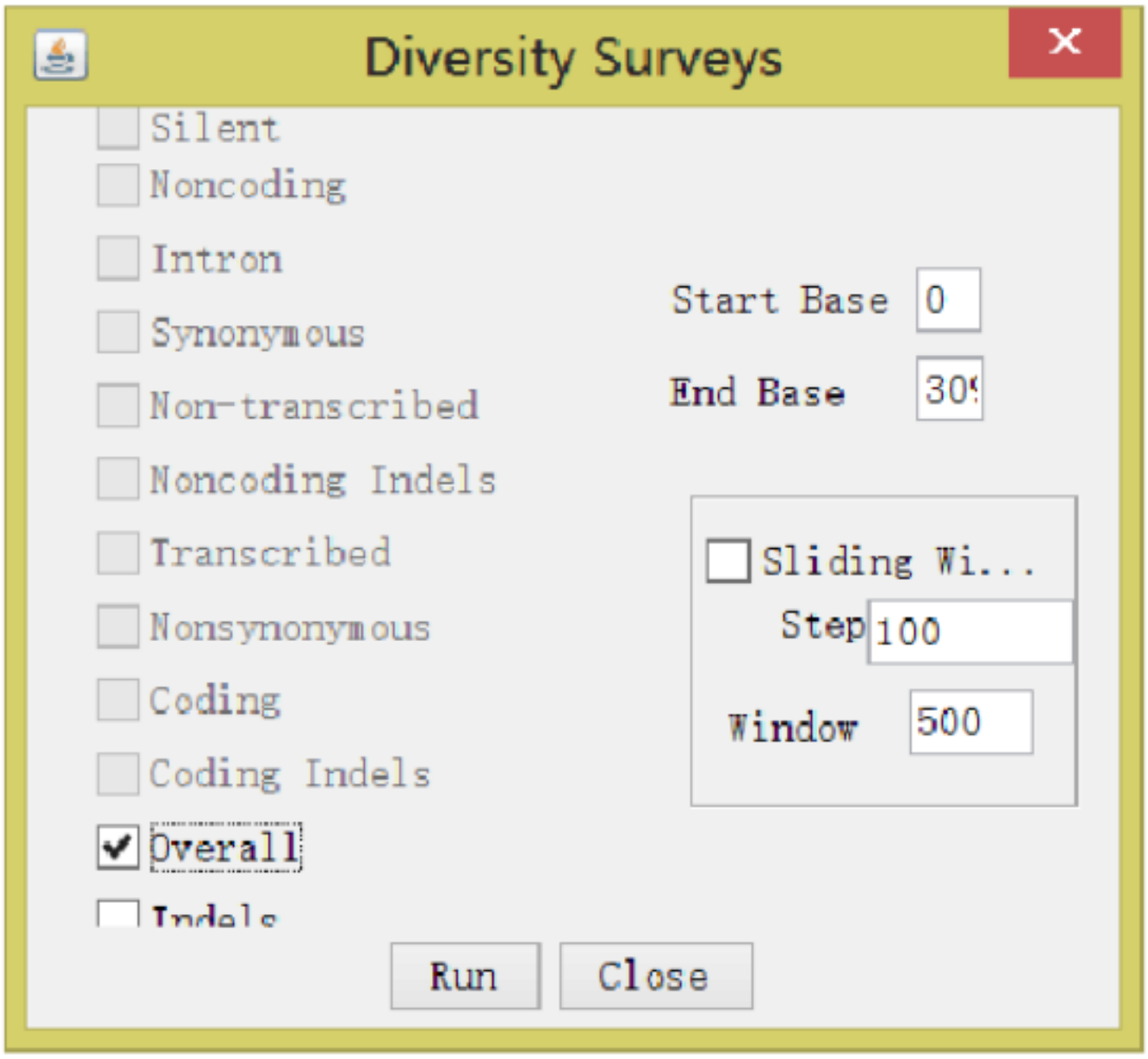
6 分析（Analysis）菜单



6.1 Diversity（多样性）

这个命令执行基本的多样性分析。可以计算平均成对多样性（Average pairwise divergence，π），分离的位点（segregating sites），以及 θ 估计值（4Nμ），还有多样性的滑动窗口（sliding window）。

为了进行多样性分析，在一个原始的序列比对上单击，然后选择 Analysis（分析）->Diversity（多样性）。



在出现的“Diversity Surveys”（多样性考察）对话框中，可用于分析的不同的位点类别被列在左边。如果序列没有注释，那么只有“Overall”和“Indels”选项将是有效的。

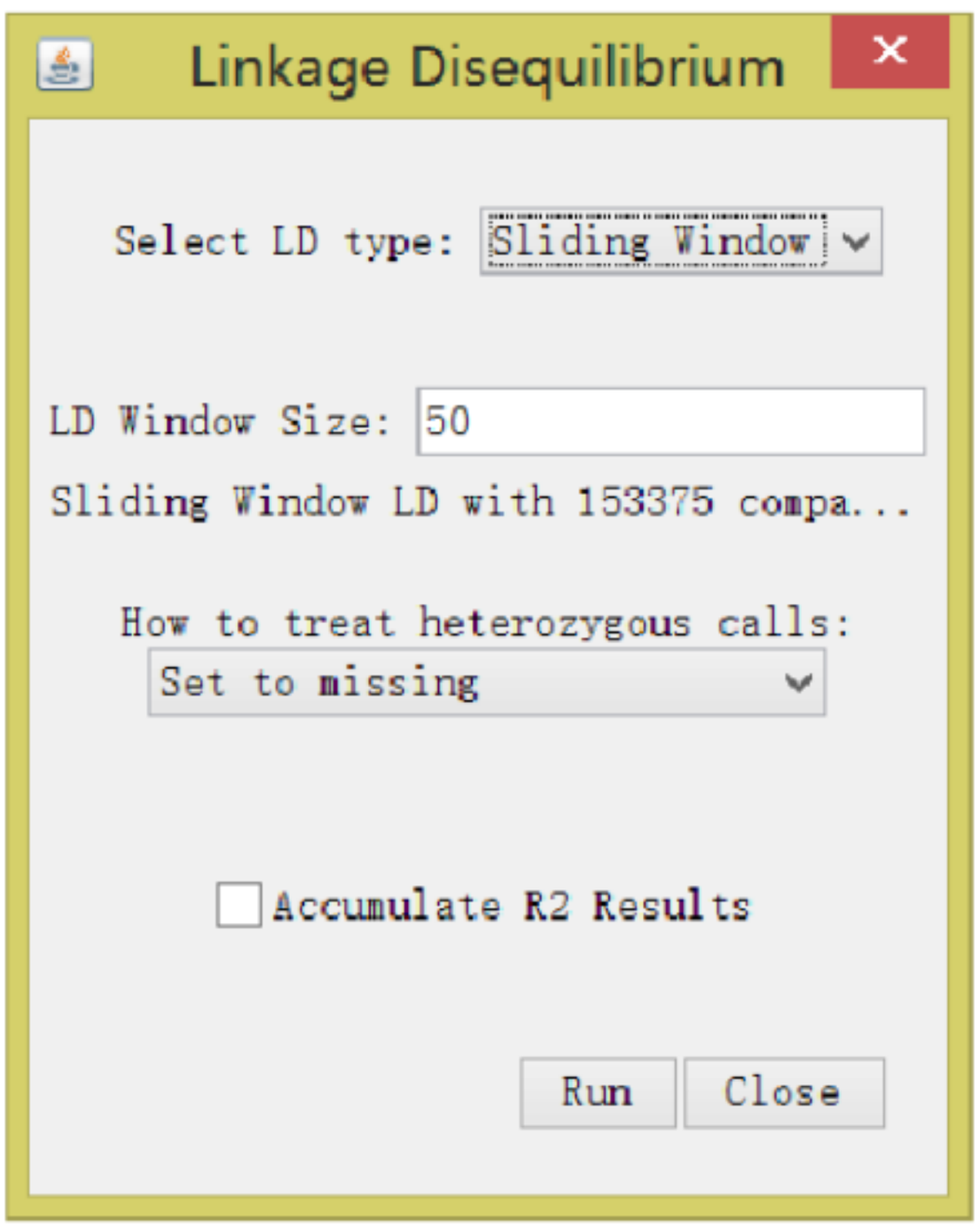
还可以跨越该区域计算多样性的一个滑动窗口。要产生一个滑动窗口，复选紧挨着“Sliding Window”的复选框，然后输入想要的步长和滑动窗口的大小。

结果可以利用 Results（结果）->Chart（图表）绘图，或通过 Results（结果）->Table（表格）查看。

6.2 Linkage Disequilibrium（连锁不平衡）

这个命令从 SNP 数据中产生一个连锁不平衡数据集。

注意：重要的是，估计连锁不平衡时只使用过滤的数据集（首先应用 Filter（过滤）->**Sites（位点）**），因为与很多不变的碱基的原始的序列比对将花费非常长的时间，并且计算要消耗大量内存。



通过单击一个过滤的多态性数据集然后使用 Analysis（分析）->Link. Diseq.（连锁不平

衡），可以估计任何多态性数据集之间的连锁不平衡。在这时候，D'、r^2 和 P 值将被估计。当前的版本只计算具有已知连锁相的单倍型之间的LD(不支持连锁相未知的二倍体基因型；对于基因型支持见 PowerMarker 或 Arlequin)。

D'是标准化的不平衡系数，对于确定一对等位基因之间是否发生过重组或平行演化（homoplasy）是一个有用的统计量。

r^2 代表两个基因座上的等位基因之间的相关性，它对于评价关联方法的分辨率是信息性的。

当只存在两个等位基因时才可以计算 D'和 r^2 [21]。如果存在复等位基因，则计算两个基因座之间 D'或 r^2 的一个加权平均数[22]。这个加权平均数是通过对等位基因的全部可能的组合计算 D'或 r^2，然后按照等位基因频率对其加权。注意：还不能完全肯定这个过程是否充分地解释等位基因数目效应。

P 值是通过两个方法确定的。如果两个基因座上都只有两个等位基因，则计算一个双侧的 Fisher 精确检验。注意：TASSEL 的以前的版本使用单侧检验，但是 TASSEL 版本 1.0.8 及以后的版本都使用双侧检验。

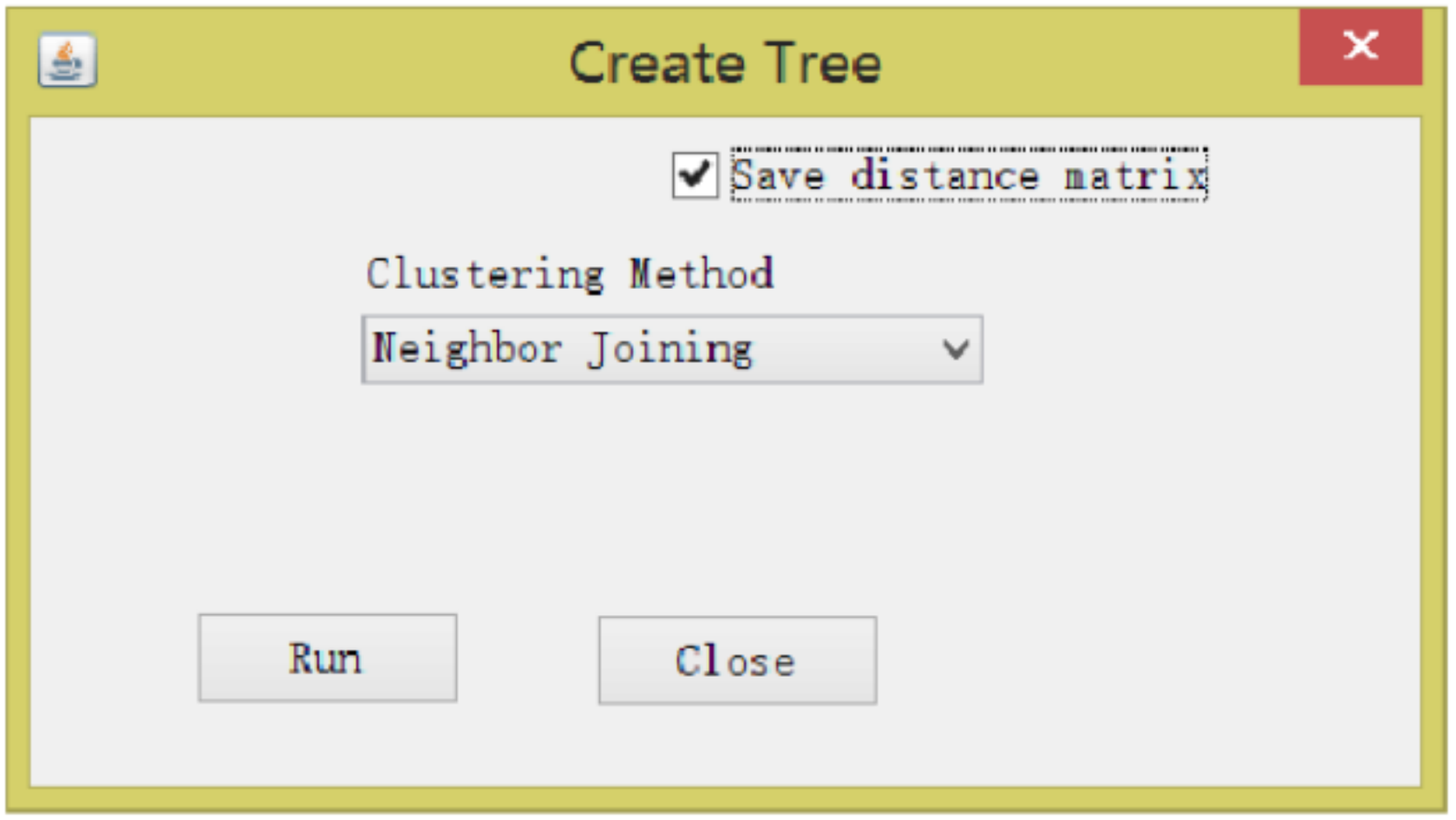
如果等位基因超过两个，则用排列（permutation）来计算比独立性的零假设下观测到的配子分布较少可能的排列的配子分布的比例[21]。

当计算连锁不平衡时，用户可以使用“Rapid Permutations”（快速排列）选项。如果选择了这个选项，则该算法就要么计算一个固定的排列次数，要么运行直到发现比观测到的 P 值更显著的 10 次排列为止。虽然这略略降低了 P 值，但是节省了大量的计算时间。如果想要一个无偏的 P 值，那么用户必须取消选择“Rapid Permutations”复选框。

“Full Matrix LD”（全矩阵 LD）对序列比对中的位点的每个组合计算 LD。“Sliding Window LD”（滑动窗口 LD）对当前位点周围位点的一个窗口内部的位点计算 LD。“LD Window Size”（LD 窗口大小）确定当前位点一侧的窗口的宽度。

连锁不平衡分析的结果可以利用 Results -> LD Plot 来绘图，或通过 Results -> Table 在一个表格中查看。

6.3 Cladogram（进化分枝图）



这个功能产生一个树或进化分枝图数据集。TASSEL 只使用简单的简约替换（parsimony substitution）模型产生邻近值-连接树。

为了重新得到进化分枝图数据，首先从数据树中选择基因型数据，然后单击 Analysis（分析）->Cladogram（进化分枝图）。产生的树数据和相应的矩阵将作为单独的数据集出现在数据树上。可以使用 Results（结果）-> Archaeopteryx Tree（始祖鸟树）对结果绘图。

6.4 Kinship（亲缘关系）

这个功能从一个基因型产生一个亲缘关系矩阵。要这样做的话，首先加亮 SNP 数据，然后单击“Analysis/Kinship”子菜单。产生的对话框将提供选项来选择“scaled IBS”或“pairwise IBS”。单击“OK”产生一个亲缘关系矩阵。

当选择了一个基因型文件并选择了“pairwise IBS”选项时，产生的亲缘关系矩阵的每个元素 i, j 等于在分类单元 i 和分类单元 j 之间不同的 SNPs 的比例。对每一对分类单元计算距离，忽略对一个分类单元具有缺失值的任何位点。通过用 2 减去所有值然后尺度化把距离矩阵转换成一个相似矩阵，矩阵中的最小值为 0、最大值为 2。亲缘关系可以由一组随机的 SNP 数据获得（建议最少几百个覆盖全基因组的 SNPs）。这个 ad-hoc 重新尺度化的方法在 TASSEL 的一个早期版本实现了，以便提供加性遗传方差的一个合理估计，但是趋向于过高估计那个值。重新尺度化不影响它对于校正群体结构的作用。它只影响加性遗传方差的估计值，因而，只影响遗传率的估计值。

为了提供加性遗传方差的更好的估计，可以使用一个替代的方法，通过选择“scaled IBS”。这个方法（根据 Endelman 和 Jannink, 2012）把基因型编码为 2 、1 或 0，等于那个基因座上的一个等位基因的计数。然后在估计关系矩阵之前用那个基因座上的平均基因型得分取代缺失的基因型值。计算亲缘关系之前估算基因型的其它方法可能提供更好的结果。例如，在计算亲缘关系之前，不使用缺失值的这个默认的处理，而是使用数值的基因型方法继之以 3.3 节中描述的估算，也是一个合理的替代方案。当使用数值基因型时，总是应用“scaled IBS”方法来计算亲缘关系。

用户也可以使用 Data -> Load 来加载自己的亲缘关系数据。亲缘关系矩阵可以使用 SPAGeDi 软件包计算（http://www.ulb.ac.be/sciences/ecoevol/spagedi.html）。在文献中可以找到计算亲缘关系方法的比较（例如 Stich et al. 2008）。

6.5 GLM（一般线性模型）

这个功能使用一个最小二乘固定效应线性模型进行关联分析。

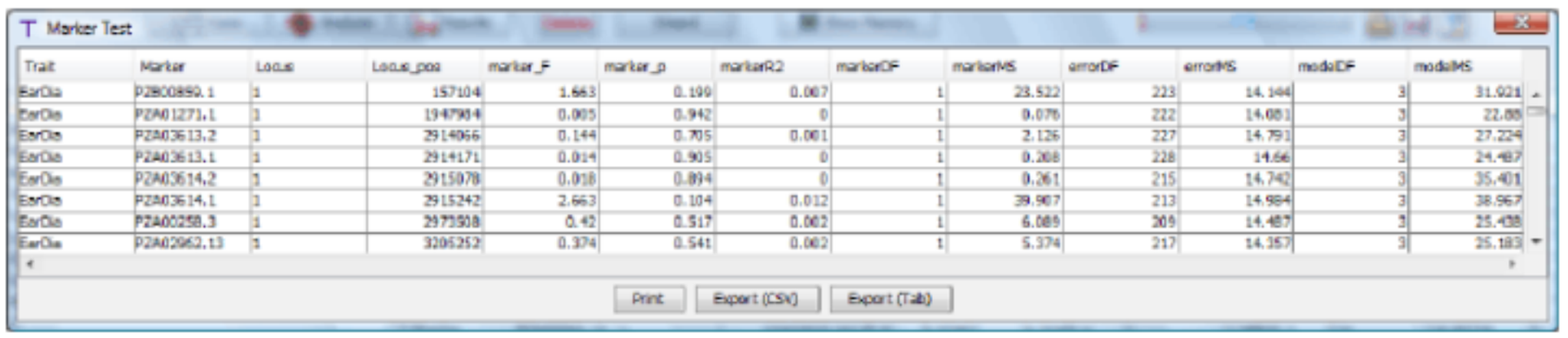
TASSEL 利用一个固定效应线性模型来对分离的位点和表现型之间的关联进行检验。该分析利用表示内在群体中的从属关系程度的协变量来选择性地解释群体结构。一个只有主效应的模型被自动地建立，利用输入数据中的全部变量。对每个性状和标记组合建立一个单独的模型并求解。任何因子、协变量、重复或地点被作为主效应包括在每个模型中。数据被怎样使用必须被定义，要么在输入数据文件中，要么利用性状过滤器（Trait Filter），在数据已经被导入之后，但是在它被与一个基因型合并之前。

一般线性模型（GLM)）可以只利用一个数值数据集来运行，或者利用数值数据与基因型数据合并后再运行。如果只选择了数值数据，则将对分类单元的每个性状产生最佳线性无偏估计（BLUEs 或最小二乘平均数）。[注意：在这个阶段只应该包含用来控制田间变异的因子和协变量。只有当标记也在分析中时才应该包含群体结构协变量（它被用来控制标记效应）。如果带有基因型的数值数据被分析，每个性状-标记组合将被检验，将产生两个报告，一个报告包含性状-标记的 F 检验，另一个报告包含等位基因估计值。

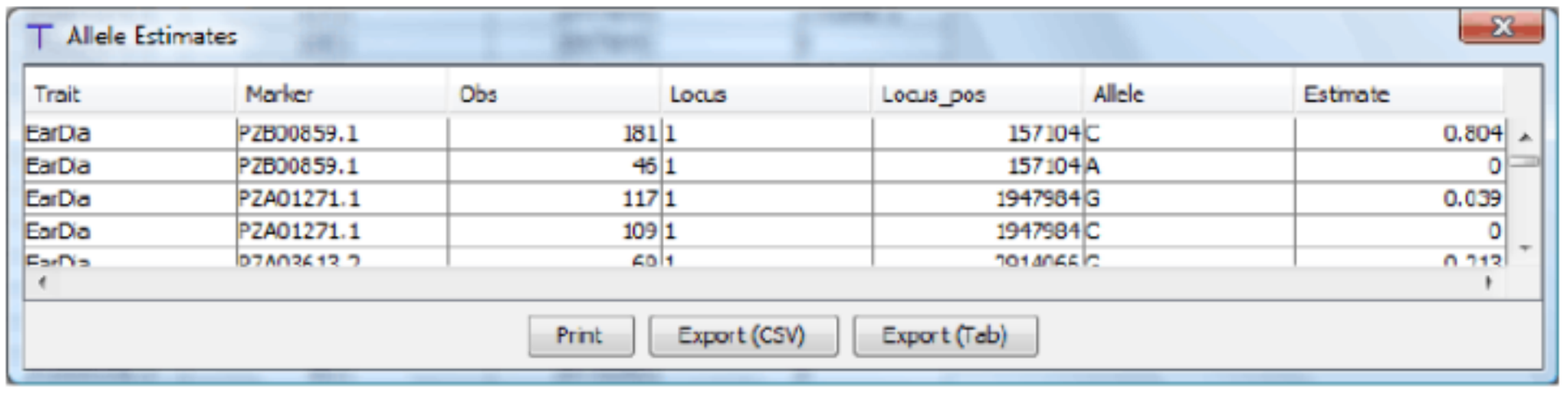
为了运行 GLM，选择一个数据集然后单击 GLM 按钮。将弹出一个对话框，允许用户指明一个排列检验应该被运行，以及允许改变排列的次数。排列检验将使用 Anderson 和 Ter Braak（2003）提出的方法运行，它计算缩减模型（包含除标记之外的所有项）的预测值和残差值，然后排列残差值并把它们添加到预测值。当 GLM 选项对话框被关闭时，用户得到

一个对话框，允许将输出保存到一个文件，而不是保存在内存中，并且由 TASSEL 显示。当预期输出非常大并且有超出可用内存的风险时，这个选项是有用的。

下面的表格显示标记检验（Marker Test）输出的一个例子，是用 Results/Table 查看的：



除了对要求的 F 检验显示 F 统计量和 p 值之外，该表格也包含标记 R^2（MarkerR2），标记效应、模型（对平均数校正）以及误差的均方（MS）和自由度（DF）。如果分类单元是重复的（跨越重复或环境），那么标记被检验，使用标记均方内部的分类单元。如果分类单元没有重复，则使用残差均方。MarkerR2 是标记的边缘 R 平方，计算公式为：SS Marker (在配合所有其他的模型项之后) / SS Total，其中 SS 代表平方和。下面的表格显示等位基因估计值输出的一个例子，是用 Results/Table 查看的：



对于每个标记和性状组合，每个标记等位基因被列表，与携带那个等位基因的分类单元的观测值的数目（Obs）、基因座（通常为染色体）和那个标记的基因座位置、等位基因、以及那个等位基因的效应估计值一起。由于 GLM 编码等位基因的方式，一个标记的最后的等位基因估计值总是零，其他的等位基因估计值则以那个为基准。

6.6 MLM（混合线性模型）

这个命令通过一个混合线性模型（MLM）来进行关联分析。

混合模型是包含固定效应和随机效应的模型。包含随机效应使 MLM 可以结合有关个体之间的亲缘关系信息。当一个基于亲缘关系矩阵（K）的遗传标记被与群体结构（Q）一起

使用时，与只用“Q”的方法相比“Q+K”方法提高了统计功效[9]。MLM 可以按照 Henderson 的矩阵符号表示法[23]描述如下：

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

其中 **y** 是观测值的向量；**β** 是包含固定效应的未知向量，包括遗传标记和群体结构（Q）；**u** 是个体/品系的随机加性遗传效应的一个未知向量，来自多个背景 QTL；**X** 和 **Z** 是未知的设计矩阵；**e** 是未观测到的随机剩余效应的向量。**u** 和 **e** 向量假定为正态分布的，平均数为零，方差为

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

其中 $\mathbf{G}=\sigma_a^2\mathbf{K}$，$\sigma_a^2$ 为加性遗传方差，**K** 为亲缘关系矩阵。对于剩余效应假定方差是同质的，这意味着 $\mathbf{R}=\mathbf{I}\sigma_e^2$，其中$\sigma_e^2$ 是剩余方差。遗传方差对总方差的比例被定义为遗传率（h^2）。

当 **K** 由系谱衍生而来时，**K** 的元素等于 2 ×概率（IBD），其中 IBD 意味着随机取出的两个等位基因是血统相同的。通常，由标记计算的 **K** 是一个 IBS 矩阵。得到的乘数（multiplier）则不是 σ_a^2，而是某个未知的常数乘上 σ_a^2。一些用于计算 **K** 的方法，比如在 SPaGEDI 中实现的那些，实际上使用标记来衍变出 IBD 关系矩阵的一个估计。对于 **K** 的那些值，得到的方差估计值可以被当做 σ_a^2 的一个估计，只要用来导出 **K** 的方法的假设对于被分析的群体没有违反。一个含意是两个不同的 **K** 矩阵可以给出 σ_a 和遗传率的非常不同的估计值，然而产生相同的模型配合和标记关联的检验。

TASSEL 实现若干方法来提高统计功效和减少计算时间。σ_a^2 和σ_e^2 的限制最大似然（REML）估计值是通过高效的混合模型关联（Efficient Mixed-Model Association，EMMA）算法[24]得到的，它比期望和最大化（EM）算法[25]快得多。

TASSEL 也实施一个称为压缩（compression）的方法，它降低亲缘关系矩阵的维数，以便减少计算时间，并改进模型拟合。当 MLM 被使用而没有压缩（compression= 1）时，每个分类单元属于它自己的组群。在另一个极端，GLM 可以被解释为最大压缩（compression = n），其中全部分类单元在单个组群中。在那种情况下，不可能独立于误差来估计随机效应，并且 σ_a^2 被吸收到σ_e^2 中去。在这两个极端之间，分类单元可以被归组，利用以亲缘关系为基础的聚类分析。当 n 个个体被压缩到 s 个类（组群）中时，个体之间的亲缘关系被组群之间的亲缘关系取代。在一些组群水平上，取决于被分析的性状和群体，与常规的 MLM 相比，这个压缩的 MLM 提高了统计功效[4]。在没有配合遗传标记的情况下，具有 MLM 的最好的模型配合的最优的归组对于标记的关联检验具有最好的统计功效[4]。TASSEL 允许用户指定

压缩水平（每个组群的平均个体数目），或者让程序确定最优的组群。

与 GLM 相似，MLM 对性状和标记的每个组合进行关联检验。TASSEL 为用户提供了若干选项：1）来对每个组合估计遗传方差和剩余方差；2）来对每个性状得到这些估计值一次，在没有配合遗传标记的情况下，然后使用那些估计值来检验标记；来使用一个由用户提供的先验的遗传率估计值。第二个选项，称为 P3D（**p**opulation **p**arameters **p**reviously **d**etermined，总体参数预先确定），与第一个选项具有相同的统计功效[4]。利用 P3D 方法或者使用先验的遗传率可以比对每个标记计算遗传率快得多。

MLM 的使用与 GLM 非常相似。区别是除了选择合并的数据集（或者数值数据集）之外，在单击 MLM 按钮来显示 MLM 选项对话框之前，亲缘关系数据也必须被加亮。“No Compression”的选项是常规的 MLM，它等价于“Custom level = 1”。对于具有大量分类单元的数据集，最佳的压缩选项可能比不压缩或者用户提供的压缩慢很多。这是因为该算法对一系列压缩水平的每一个都求解模型一次以便确定最佳的一个。

所有的 MLM 分析产生两个输出表格，模型统计量和模型效应。如果使用了压缩，该分析则产生三个表格。

MLM_statistics_for_Filtered_mdp_traits + Filtered_mdp_population_structure + mdp_genotype_chr1_157104-3706018

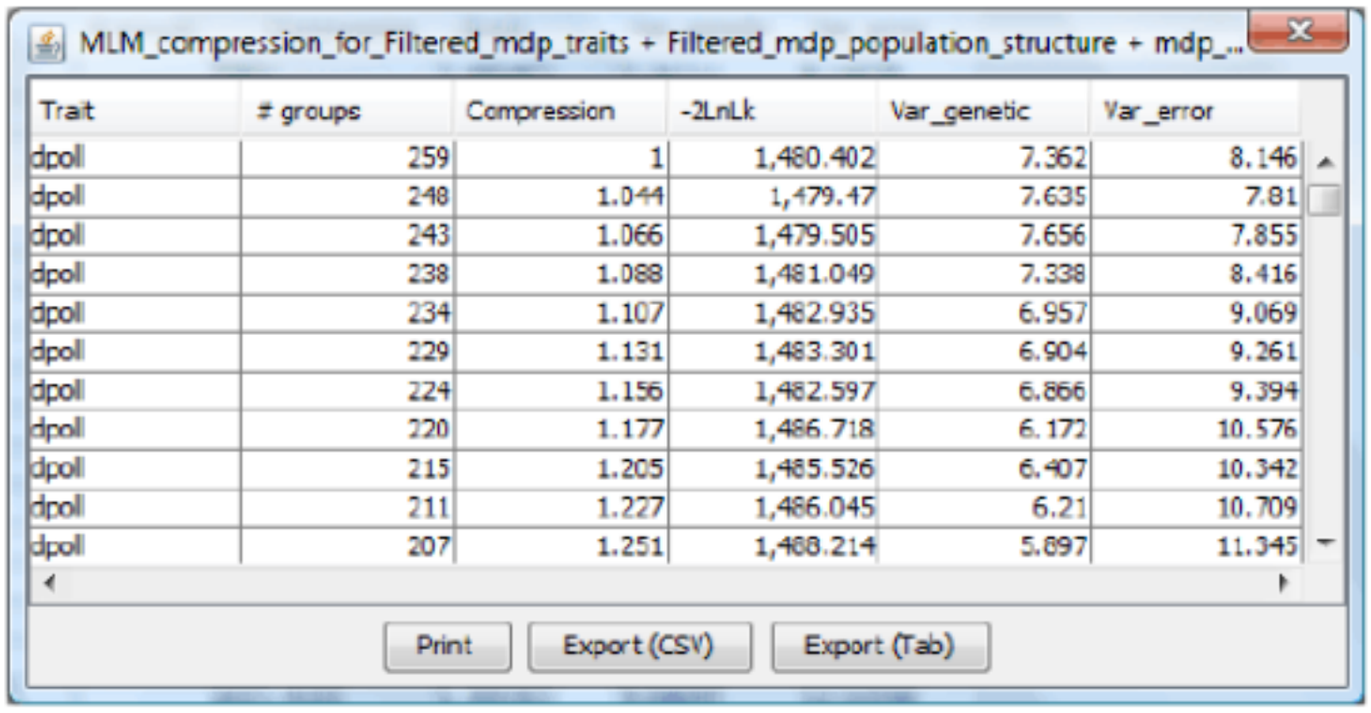
Trait	Marker	Locus	Site	df	F	p	errordf	markerR2	Genetic Var	Residual Var	-2LnLikelihood
dpoll	None			0			257		8.068	14.585	1,477.183
dpoll	PZB00859.1	1	157104	1	0.001	0.979	250	0	8.068	14.585	1,477.183
dpoll	PZA01271.1	1	1947984	1	4.339	0.038	248	0.015	8.068	14.585	1,477.183
dpoll	PZA03613.2	1	2914066	1	0.132	0.716	255	0	8.068	14.585	1,477.183
dpoll	PZA03613.1	1	2914171	1	2.829	0.094	256	0.01	8.068	14.585	1,477.183
dpoll	PZA03614.2	1	2915078	1	0.044	0.834	243	0	8.068	14.585	1,477.183
dpoll	PZA03614.1	1	2915242	1	0.788	0.375	241	0.003	8.068	14.585	1,477.183
dpoll	PZA00258.3	1	2973508	1	0.732	0.393	240	0.003	8.068	14.585	1,477.183
dpoll	PZA02962.13	1	3205252	1	0.967	0.326	244	0.004	8.068	14.585	1,477.183
dpoll	PZA02962.14	1	3205262	1	0.026	0.873	239	0	8.068	14.585	1,477.183

Print　Export (CSV)　Export (Tab)

统计量表显示对每个性状进行检验的结果。第一行用于没有标记的模型。后面是每个检验的标记占一行。列标签“df”、“F”、“p”是自由度、F 值和 p 值，来自对标记进行检验的 F 分布。列“errordf”是 F 检验的分母使用的自由度。列标签“markerR2”是标记的 R2，根据一个广义最小平方法（GLS）模型的 R2 的公式计算，如同这里所示。

列“Genetic Var”，“Residual Var”，和“-2LnLikelihood”分别列出 σ_a^2，σ_e^2，以及模型似然的负二倍。当使用 P3D 选项时，全部值对一个给定的性状是一样的，因为它们只计算一次。第二个表格对每个标记列出每个等位基因的效应估计值，与 GLM 的输出相似。压缩结果表格显示优化过程中检验的每个压缩水平的似然函数、遗传方差和误差方差，如下所示。组群和压缩的用意在上面对压缩方法的描述中讨论过了。具有最低的- 2LnLk 值的压缩水平

被用于检验标记。



6.7 基因组选择（使用岭回归）Genomic Selection (using Ridge Regression)

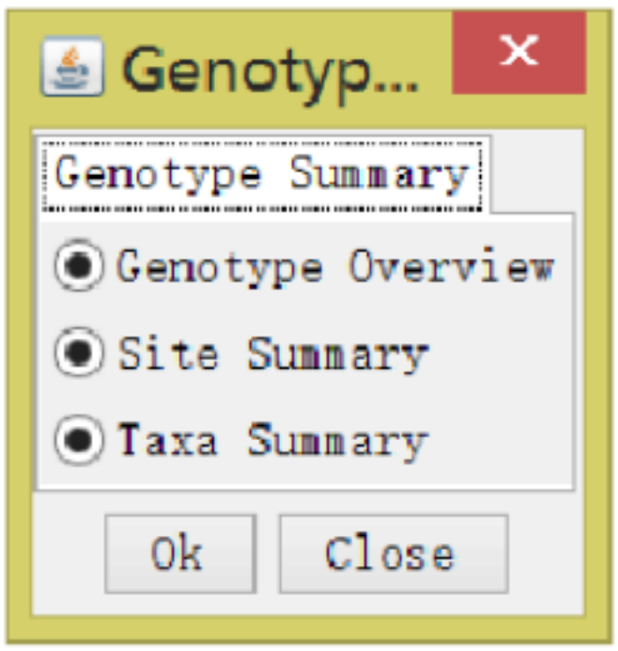
这个功能进行岭回归来根据基因型预测表现型。它是基因组选择（genomic selection，GS）使用的方法之一。

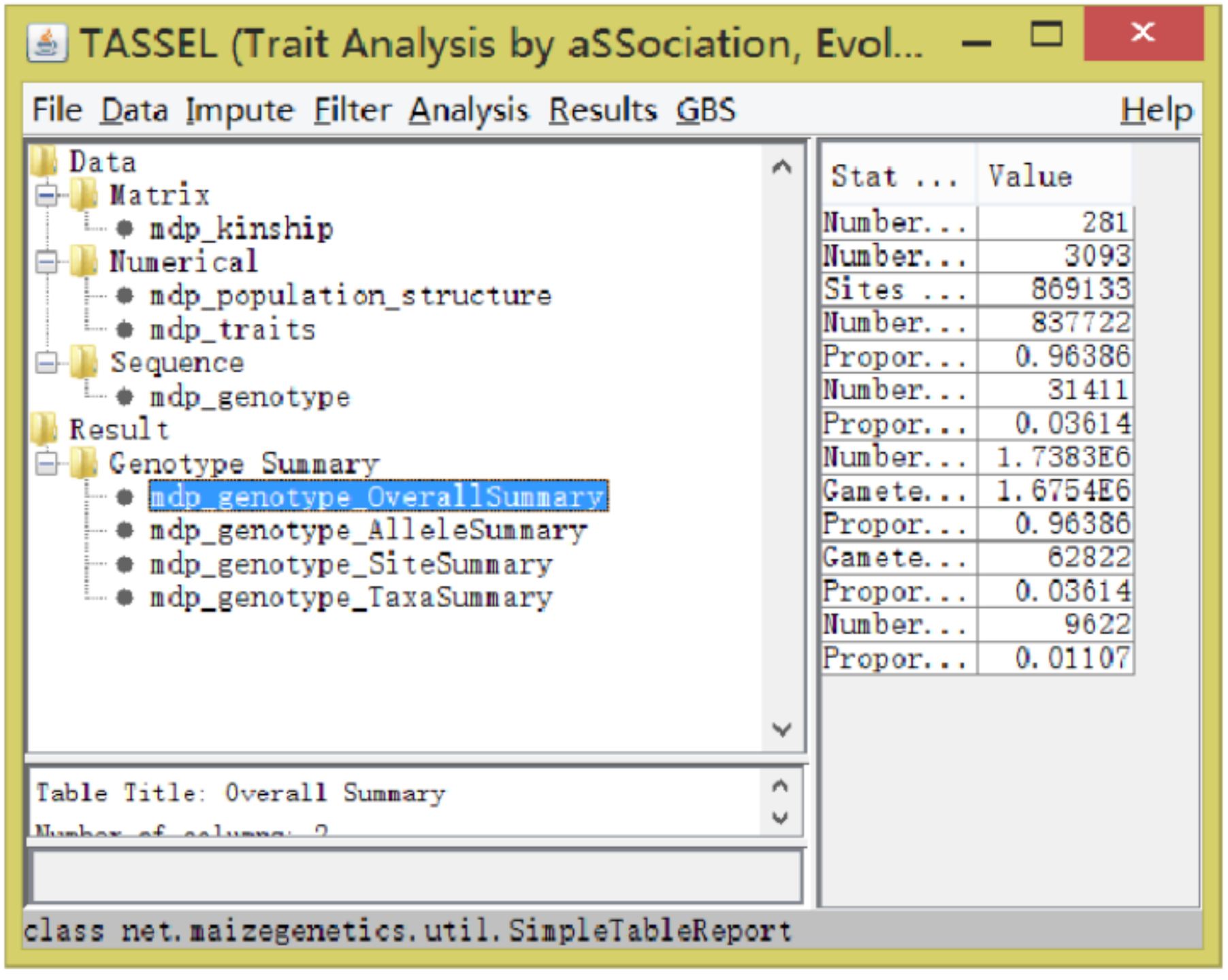
输入数据集必须包含一个或多个表现型和数值的标记数据。它还可以包含因子和协变量，不过这是可选地。通过选择输入数据集然后单击“GS”按钮来运行该分析。因为不需要额外的用户输入，在按钮被单击之后该分析就立即运行。所有的性状将被利用数据集中的全部基因型、因子和协变量单独地分析。对于每个性状，输出将由两个新的数据集组成。一个数据集包含每个分类单元的基因组育种值的估计（genomic estimated breeding values，GEBVs），另一个数据集包含基因型文件中每个标记的最佳线性无偏预测（BLUPs）。输出数据集将出现在“Numerical”文件夹中，它也存放输入数据。输出数据集可以依次被用于后续的分析。例如，它可以与输入数据合并，以便预测值能够对原始值绘图。

理解输入数据的要求对保证得到正确和有用的分析结果是重要的。基因型必须是数值的，每个标记具有一个列。期望标记是二等位基因的，纯合体编码为 1 和- 1，杂合体编码为 0。然而，任何合理的编码方案都可以起作用。例如，缺失数据可以用根据估算产生的一个概率来取代。如果任何基因型是缺失的，它将被估算，作为标记的平均得分，对那个标记跨越所有的分类单元。如果一个用户更喜欢使用一个不同的估算方法，那么在将数据导入到 TASSEL 之前缺失的基因型必须被估算。

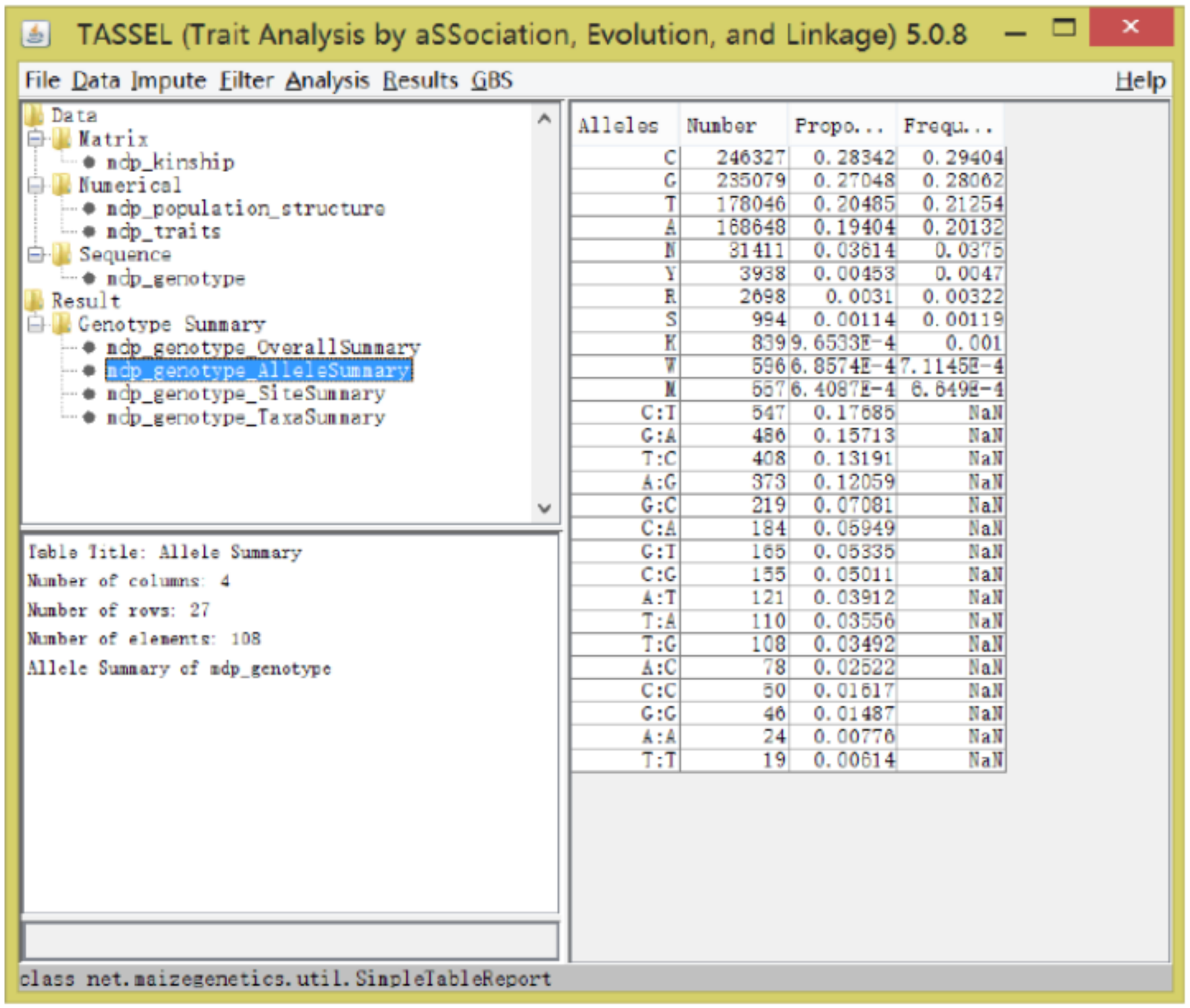
将对数据集中的全部分类单元计算 GEBV，包括具有缺失表现型数据的任何品系。基因组选择的一个典型的用途是根据一个训练集的表现来对一组未鉴定表型的品系预测 GEBV。要那么做，一个既包含要被预测的基因型又包含训练集的基因型的数据集可以与一个包含训练集的表现型的数据集合并，利用并集合并命令。表现型数据集中的全部分类单元都应该具有基因型。如果包含了一个没有基因型数据的个体，则将对那个个体估算所有的标记数据，这通常不是一件有用的事情。

6.8 Geno Summary（基因型汇总）

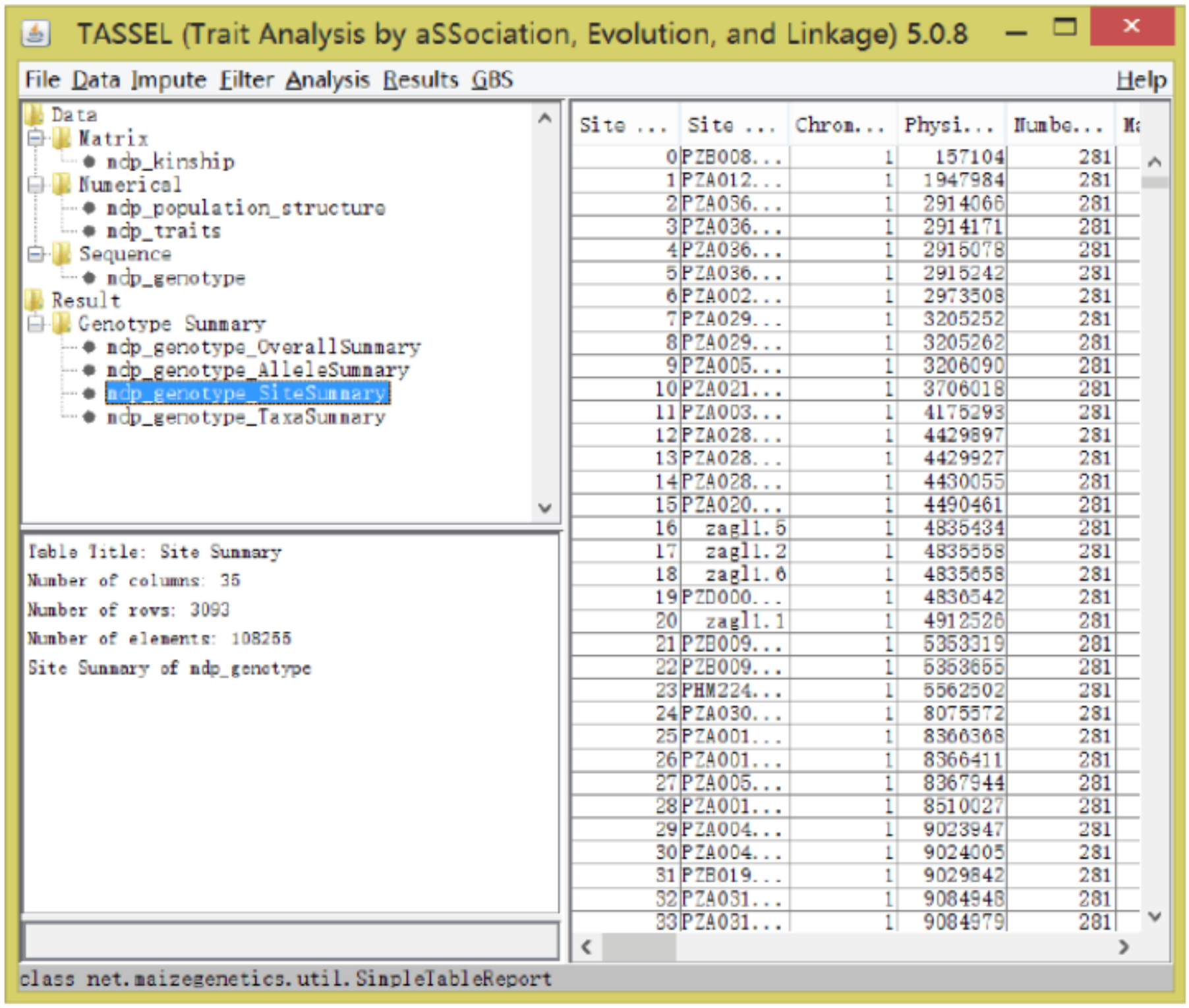




- Number of Taxa - 数据集中分类单元的数目。
- Number of Sites - 数据集中位点的数目。
- Sites x Taxa - 位点数目乘以分类单元的数目。
- Number Not Missing - 值不是未知的等位基因值数目（NN）
- Proportion Not Missing - 未缺失的数目 / 位点 x 分类单元
- Number Missing - 未知值的数目（NN）
- Proportion Missing - 缺失的数目 / 位点 x 分类单元
- Number Gametes - 位点的数目 x 分类单元数目 x 2
- Gametes Not Missing - 非未知的配子的数目
- Proportion Gametes Not Missing - 非缺失的配子 / 配子数目
- Gametes Missing - 未知的配子数目（N）
- Proportion Gametes Missing - 缺失的配子 / 配子数目
- Number Heterozygous - 杂合值的数目
- Proportion Heterozygous - 杂合的数目 / 位点 x 分类单元



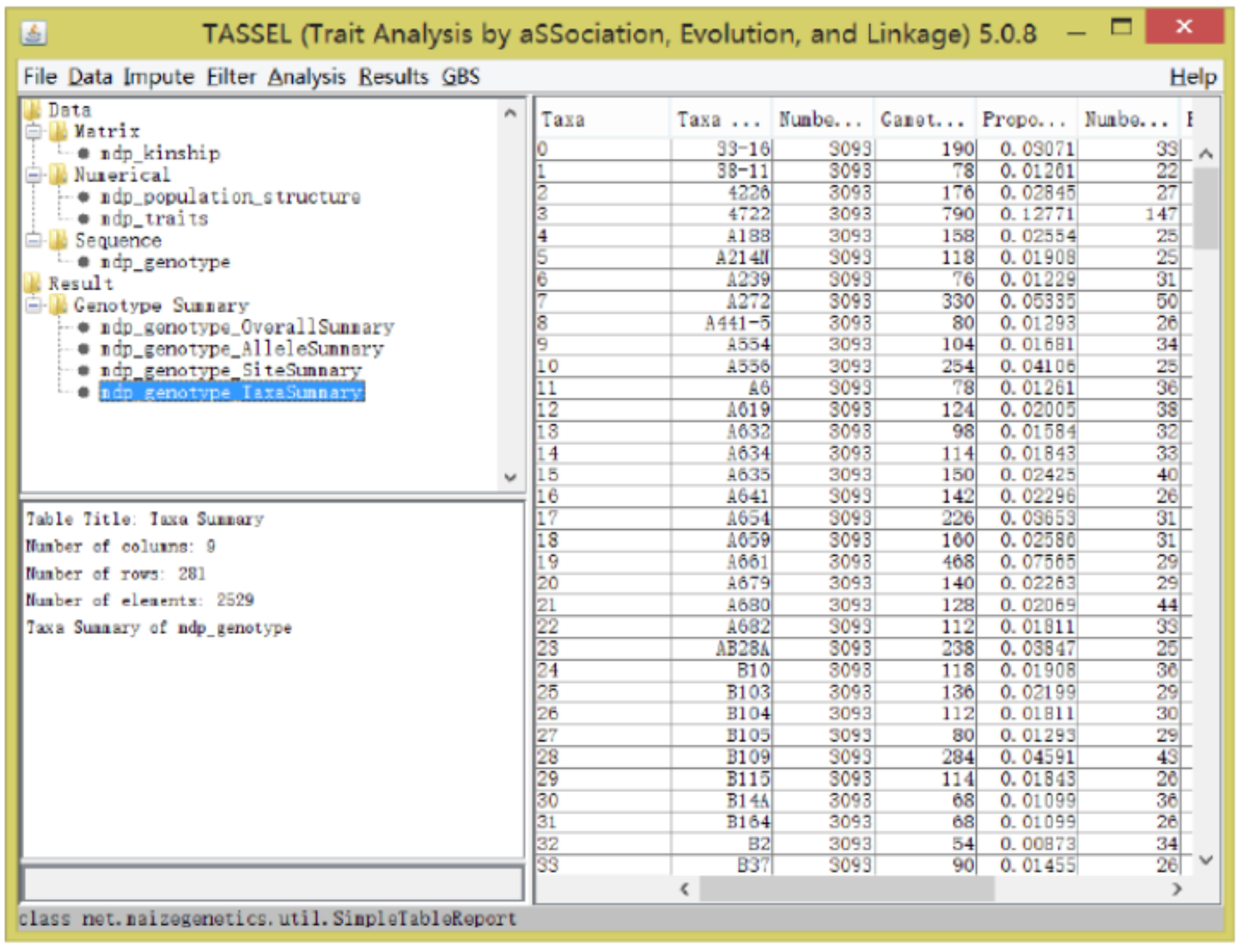
- Alleles - 数据集中给出的等位基因值。单字母值是二倍体，其中一些字母代表杂合的。二字母值是与位点的计数的主要的 / 次要的组合。
- Number 出现的次数
- Proportion - 数值出现在数据集中的百分率。
- Frequency - 出现在数据集中没有计数未知的（N）数值的百分率。



- Site Number - 位点的索引
- Site Name - 位点的名称
- Chromosome - 染色体
- Physical Position - 染色体上的物理位置
- Number of Taxa - 位点的分类单元的数目（所有的都相同）
- Major Allele - 位点的主要等位基因
- Major Allele Gametes - 位点的主要等位基因出现的次数（高达分类单元数目的两倍）
- Major Allele Proportion - 主要等位基因配子 / （分类单元的数目×2）。分类单元的数目×2是一个位点的配子的数目。
- Major Allele Frequency - 主要等位基因配子 / ((分类单元的数目×2)-缺失的配子)
- Minor Allele - 位点的次要等位基因
- Minor Allele Gametes - 位点的次要等位基因出现的次数
- Minor Allele Proportion - 次要等位基因配子 / （分类单元的数目×2）。分类单元

的数目×2是一个位点的配子的数目。

- Minor Allele Frequency - 次要等位基因配子 / ((分类单元的数目×2)-缺失的配子)
- Gametes Missing - 具有未知的（N）数值的配子的数目
- Proportion Missing - 缺失的配子 / （分类单元的数目×2）。
- Number Heterozygous - 对位点杂合的分类单元的数目。
- Proportion Heterozygous - 杂合的数目 / 分类单元的数目（没有计数未知的（NN）分类单元）
- Inbreeding Coefficient 近交系数
- Inbreeding Coefficient Scaled by Missing -



Taxa	Taxa ...	Numbe...	Gamet...	Propo...	Numbe...
0	33-16	3093	190	0.03071	33
1	38-11	3093	78	0.01261	22
2	4226	3093	176	0.02845	27
3	4722	3093	790	0.12771	147
4	A188	3093	158	0.02554	25
5	A214N	3093	118	0.01908	25
6	A239	3093	76	0.01229	31
7	A272	3093	330	0.05335	50
8	A441-5	3093	80	0.01293	26
9	A554	3093	104	0.01681	34
10	A556	3093	254	0.04106	25
11	A6	3093	78	0.01261	36
12	A619	3093	124	0.02005	38
13	A632	3093	98	0.01584	32
14	A634	3093	114	0.01843	33
15	A635	3093	150	0.02425	40
16	A641	3093	142	0.02296	26
17	A654	3093	226	0.03653	31
18	A659	3093	160	0.02586	31
19	A661	3093	468	0.07565	29
20	A679	3093	140	0.02263	29
21	A680	3093	128	0.02069	44
22	A682	3093	112	0.01811	33
23	AB28A	3093	238	0.03847	25
24	B10	3093	118	0.01908	36
25	B103	3093	136	0.02199	29
26	B104	3093	112	0.01811	30
27	B105	3093	80	0.01293	29
28	B109	3093	284	0.04591	43
29	B115	3093	114	0.01843	26
30	B14A	3093	68	0.01099	36
31	B164	3093	68	0.01099	26
32	B2	3093	54	0.00873	34
33	B37	3093	90	0.01455	26

- Taxa - 分类单元的索引。
- Taxa Name - 分类单元的名称
- Number of Sites - 分类单元位点数（所有都相同）。
- Gametes Missing - 具有未知（N）数值的配子的数目 每个分类单元 / 位点组合有两个配子。
- Proportion Missing - 缺失的配子 / （位点数目×2）。

- Number Heterozygous - 分类单元的杂合的位点数
- Proportion Heterozygous - 杂合的数目 / 位点的数目（没有计数未知的位点（NN））
- Inbreeding Coefficient -
- Inbreeding Coefficient Scaled by Missing -

6.9 Stepwise（逐步的）

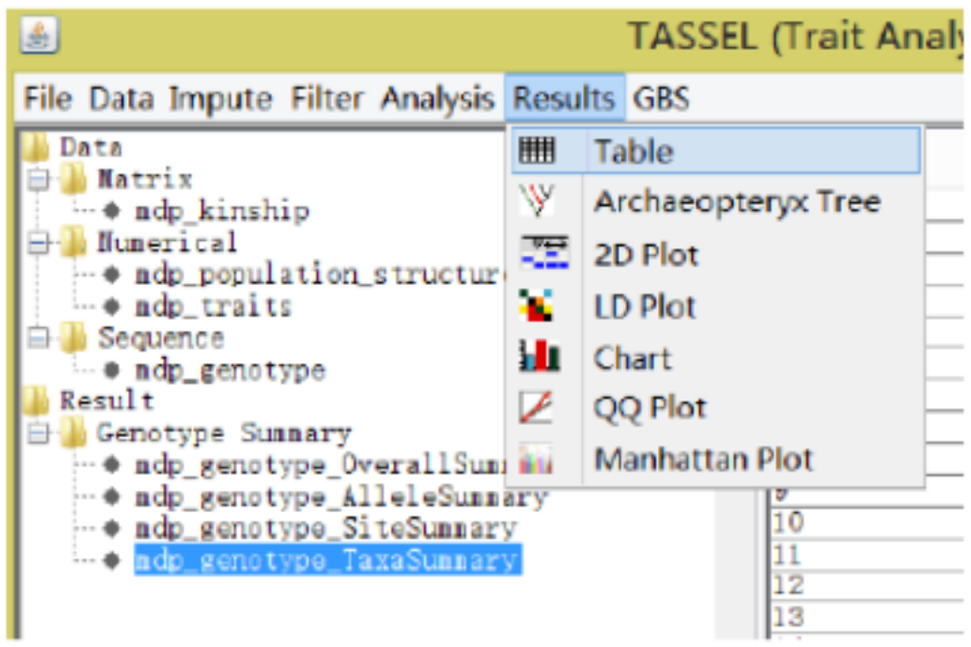
7 Results（结果）菜单

Results（结果）包括将数据作为表格或图形输出的功能。

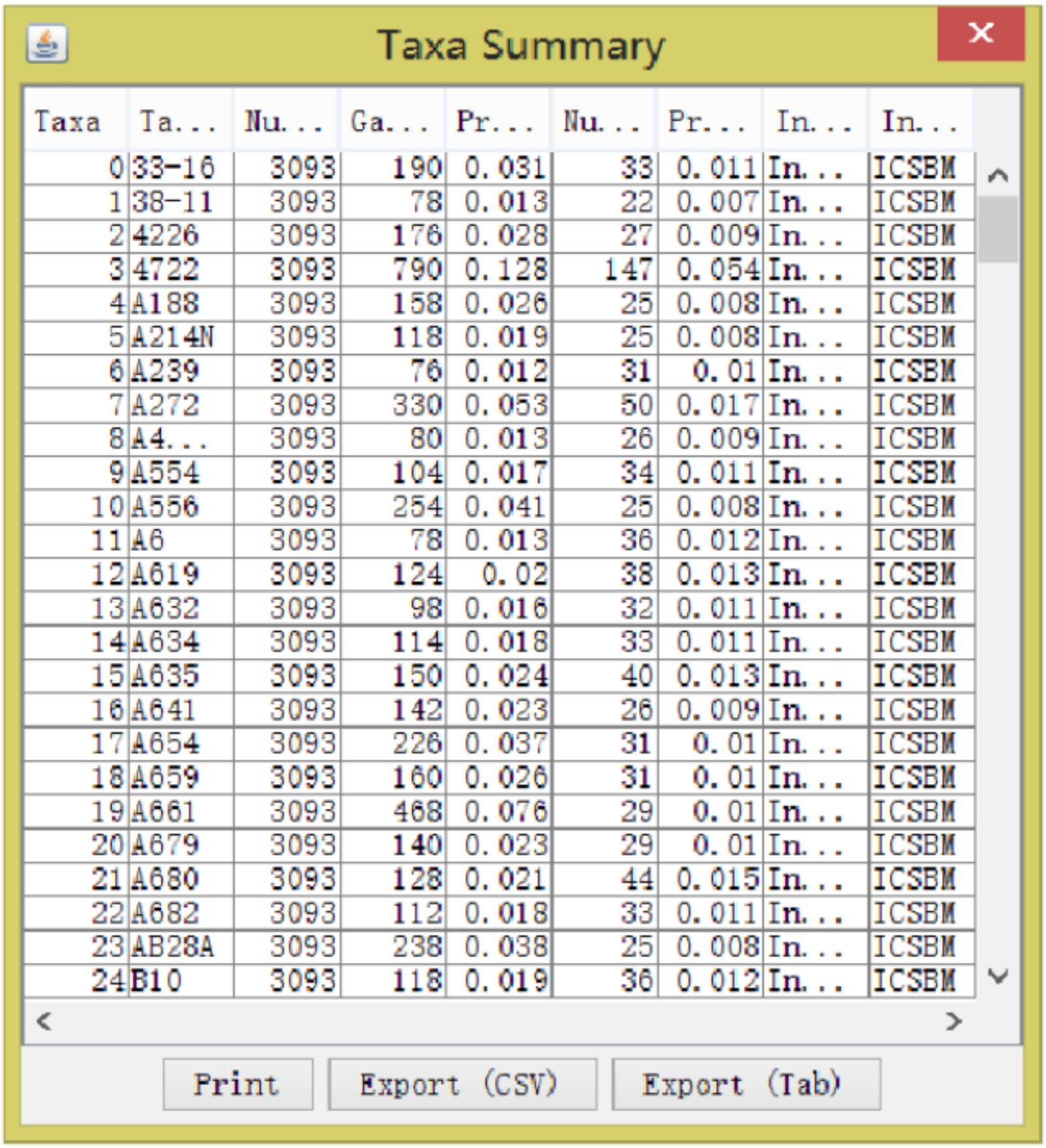
7.1 Table（表格）

允许数据被显示在一个电子表格视图中，并且导出到一个纯文本文件里。

要产生一个表格，从数据树面板中选择一个数据集，然后单击菜单“Results -> Table”。



下面显示一个例子，其中显示了分类单元汇总。



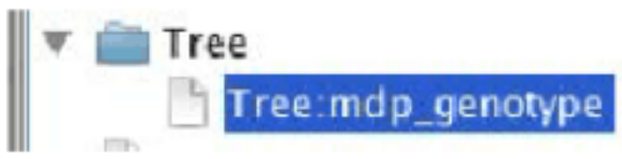
通过单击所关心的列标头可以对数据排序。通过按下 CTRL 键并单击第二列可以进行第二次排序。

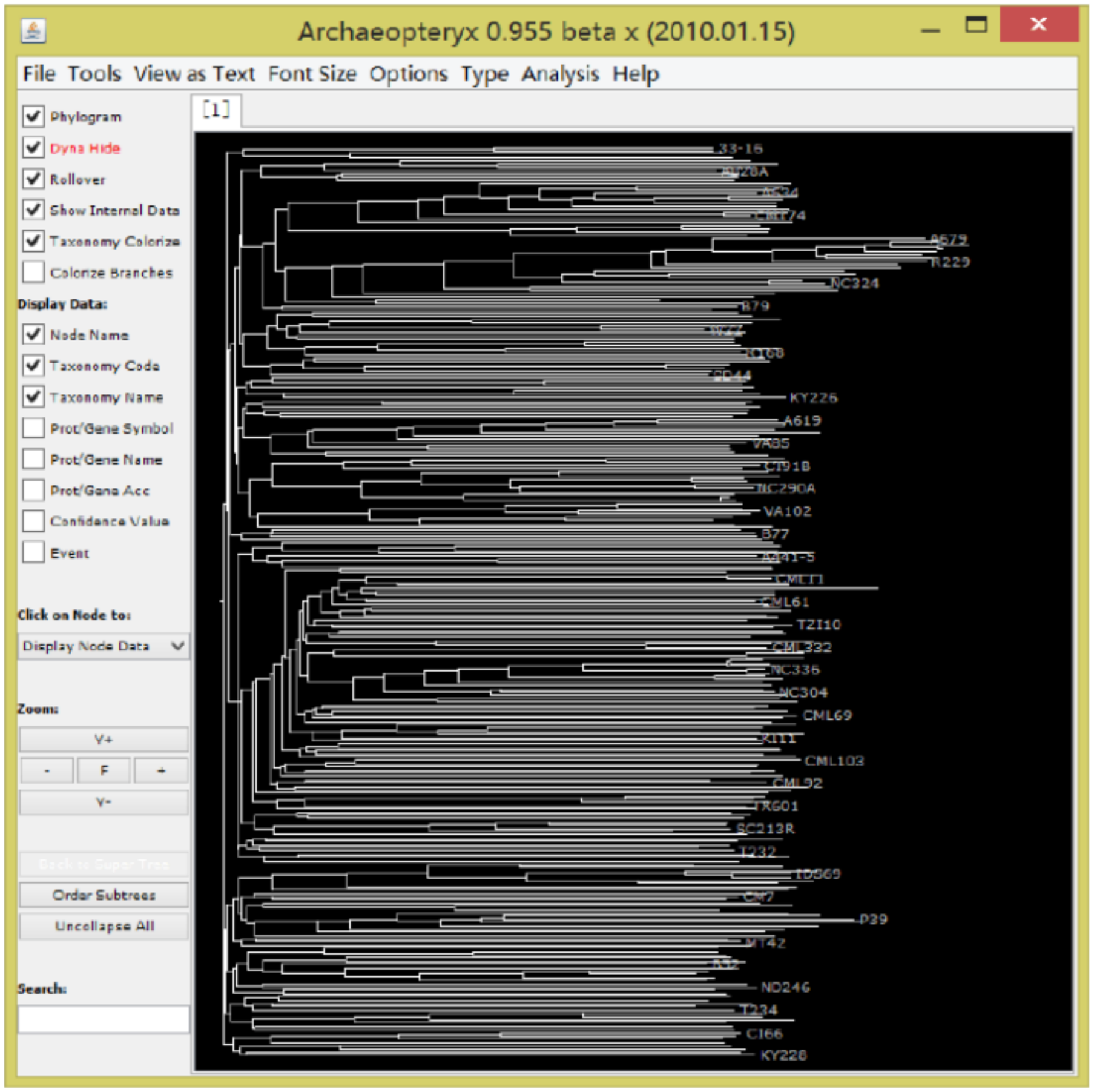
数据可以被导出到由逗号分隔的（CSV）或者用制表符分隔的纯文本文件。这些格式都可以被导入到电子表格程序里，比如 Excel。表格还可以被打印。

7.2 Archaeopteryx Tree（始祖鸟树）

选择一个要使用的“Tree:...”数据集。

https://sites.google.com/site/cmzmasek/home/software/archaeopteryx

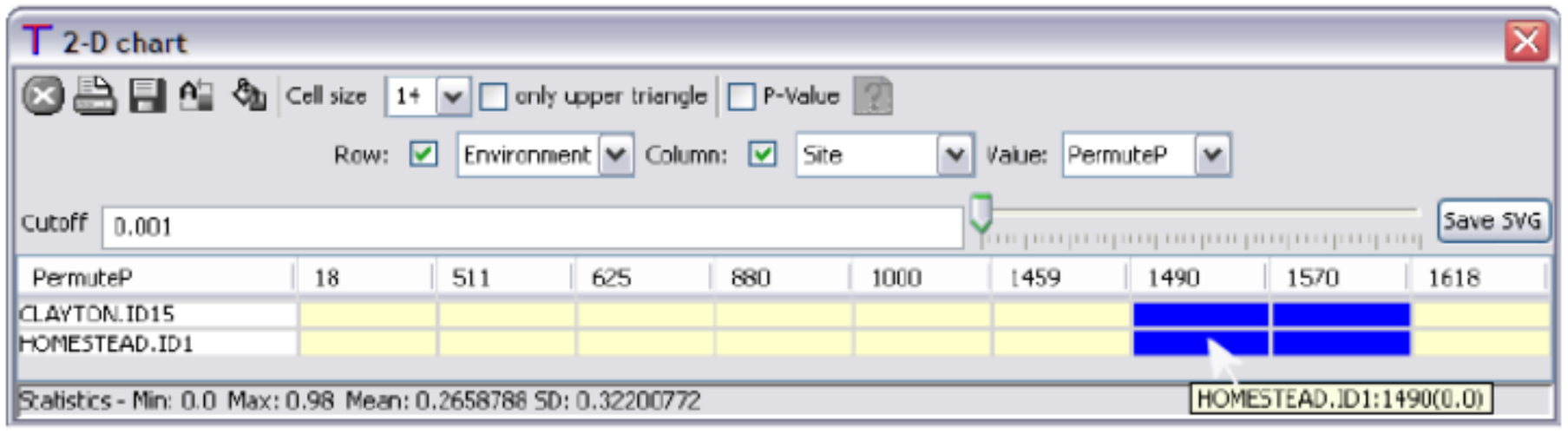




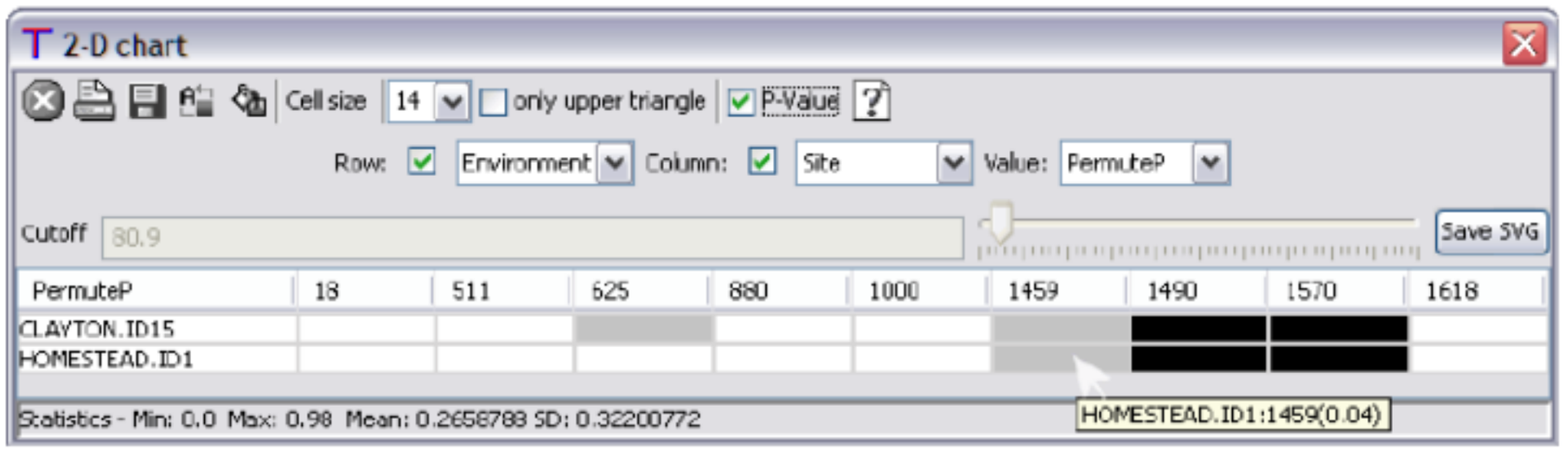
7.3 2D Plot（2D 图）

显示二维图并确定色差阈（color threshold）。这个功能对绘图多个环境中的关联是有用的。

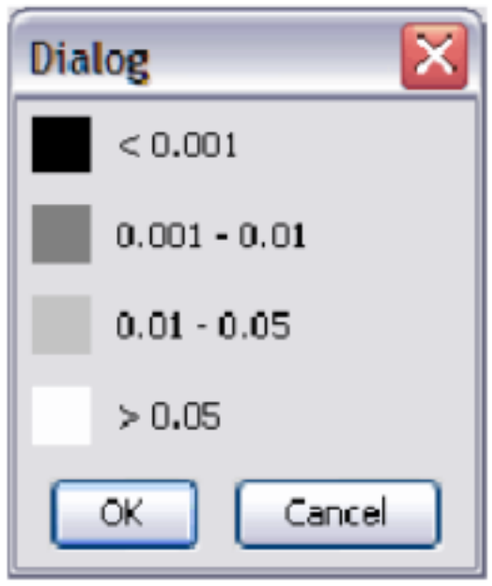
首先，选择想要的结果集合。使用提供的下拉框，用“Environment,”填充行，用“Site,”填充列，用“PermuteP”填充数值。颜色的边界值（cutoff value）可以通过在文字框中输入一个数值或者使用文字框右边的滑块工具来选择。用户可以“mouse over”任何框来查看与那个框有关的数值，如同这里所示：



如果想要对 P 值着色，就复选 P 值框，如下所示：



通过复选 P 值框，参数截止值选择工具将被禁止，区域将被按照以下灰度级着色：



通过单击紧靠着 P 值复选框的“?”图标可以显示这个键。

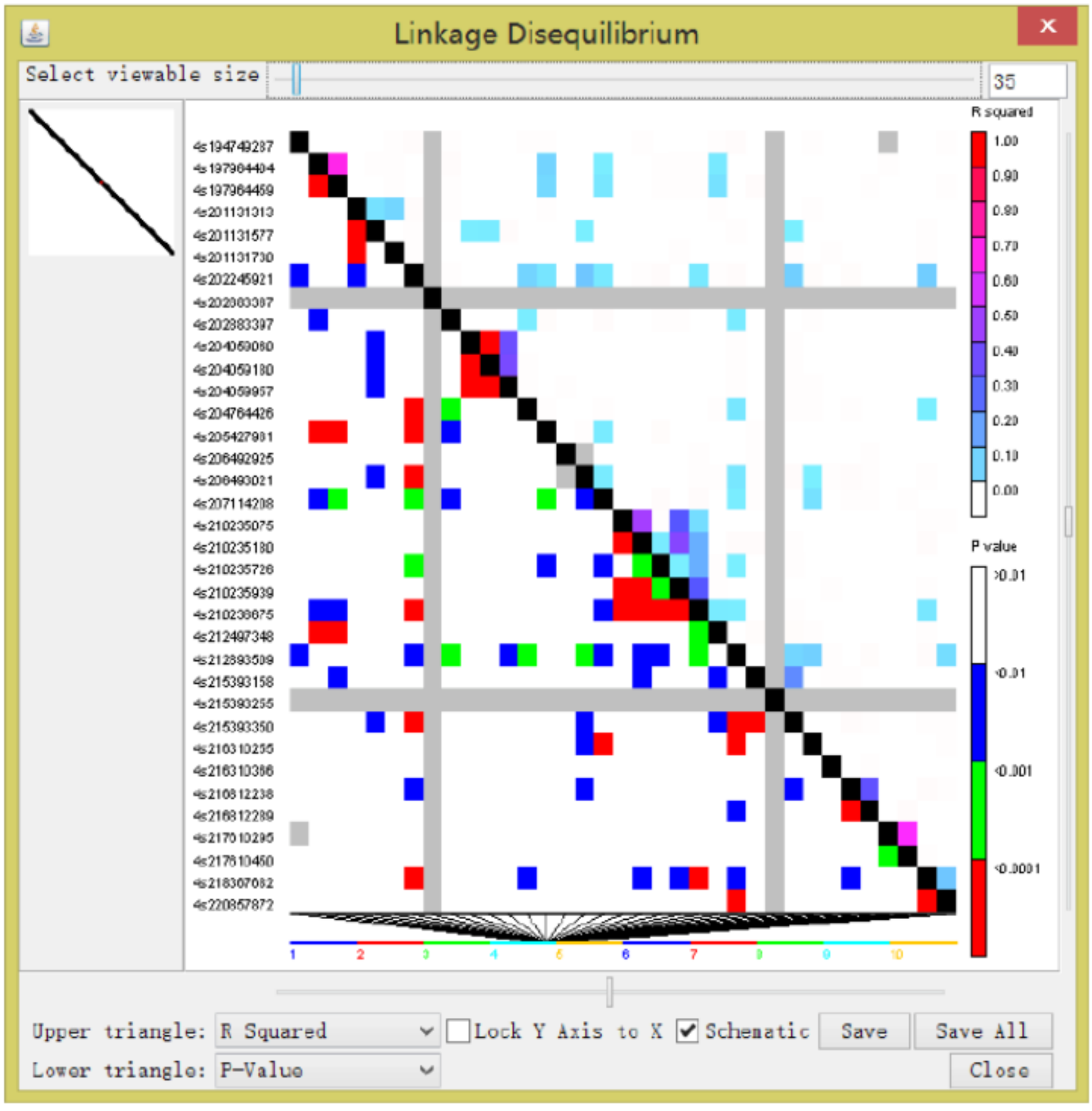
7.4 LD Plot（LD 图）

显示连锁不平衡分析的结果。

在从数据树中选择了想要的结果以后，选择 Results-> LD Plot。

产生的图显示分析步骤计算的成对位点之间的 LD。黑色对角线代表每个位点和它自己之间的 LD。默认设置绘图在右上方对 r^2 绘图，在左下方对 p 值绘图。这个默认可以通过单

击左下的按钮而修改。图的左侧包含染色体和位点名称的一个文本描述。在图的底部显示了沿着染色体的每个位点的位置。通过取消选择“Schematic”复选框可以隐藏这个显示。描述颜色方案的图例出现在图的右侧。



可以通过在右上角的白色框中输入一个数字或者通过紧挨着它的滑动杆来选择显示的位点数目。使用右边和底部的滑动条在图上移动。左上角的小的白色窗口中的红色框将显示该图显示了什么部分。为了只围绕对角线移动，选择“Lock Y Axis to X”复选框（建议通过滑动窗口直观化 LD 时使用）。

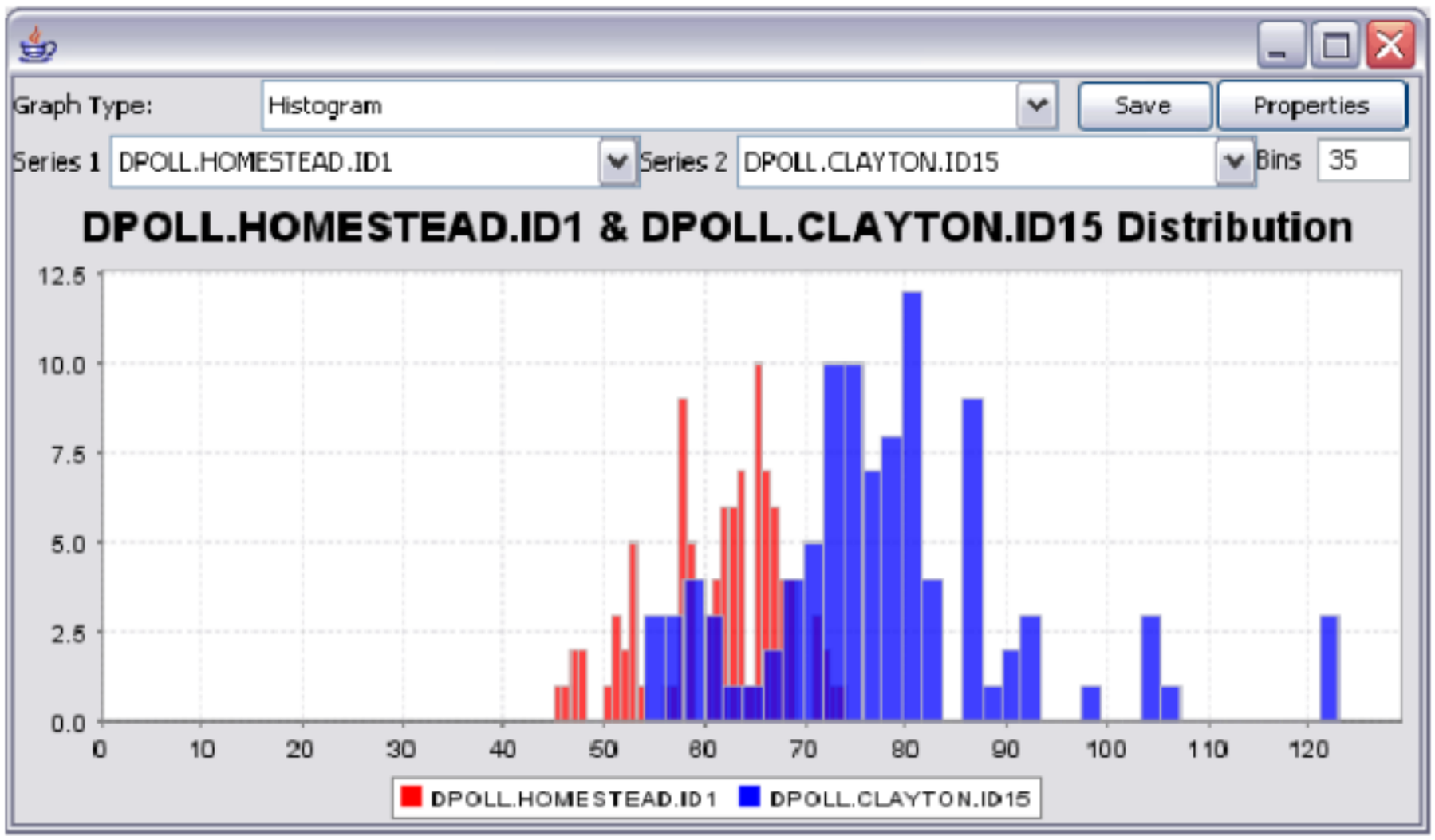
可以用若干格式保存 LD 图。Save 按钮将保存显示在屏幕中的图形区域，而 Save All 按钮将保存整个图形。

7.5 Chart（图表）

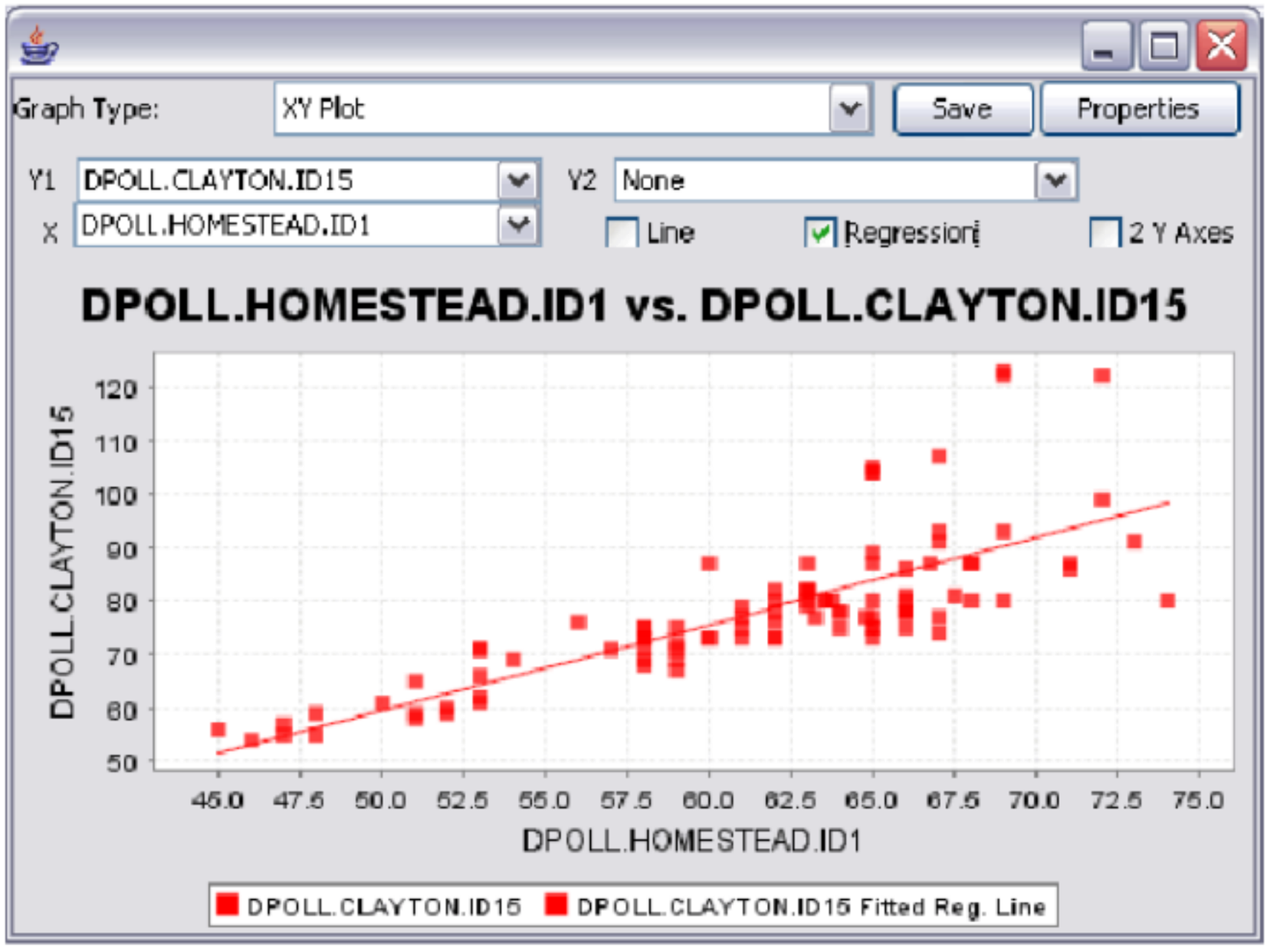
Chart（图表）提供直观化数值数据的各种图形。

这个特性可用于显示直方图、XY 图、条形图和/或饼图。任何数值列表数据都可以被制成图表，包含 LD 结果、表型数据、多样性结果以及关联结果。

Histograms（直方图）：使用图类型组合框来从选项列表中选择想要的图类型（直方图）。高达二种不同系列的数据可以被一起绘图。用户可以指定要在直方图中使用的 bins 的数目。



Scatter plots（散点图）：使用图类型组合框来从选项列表中选择想要的图类型（XY Plot）。用合适的下拉框选择要在 X 和 Y 轴绘图的数据。如果二个数据序列被同时绘图在 Y 轴上，“2 Y Axes”复选框就会为每个数据序列提供一个轴。



7.6 QQ Plot（QQ 图）

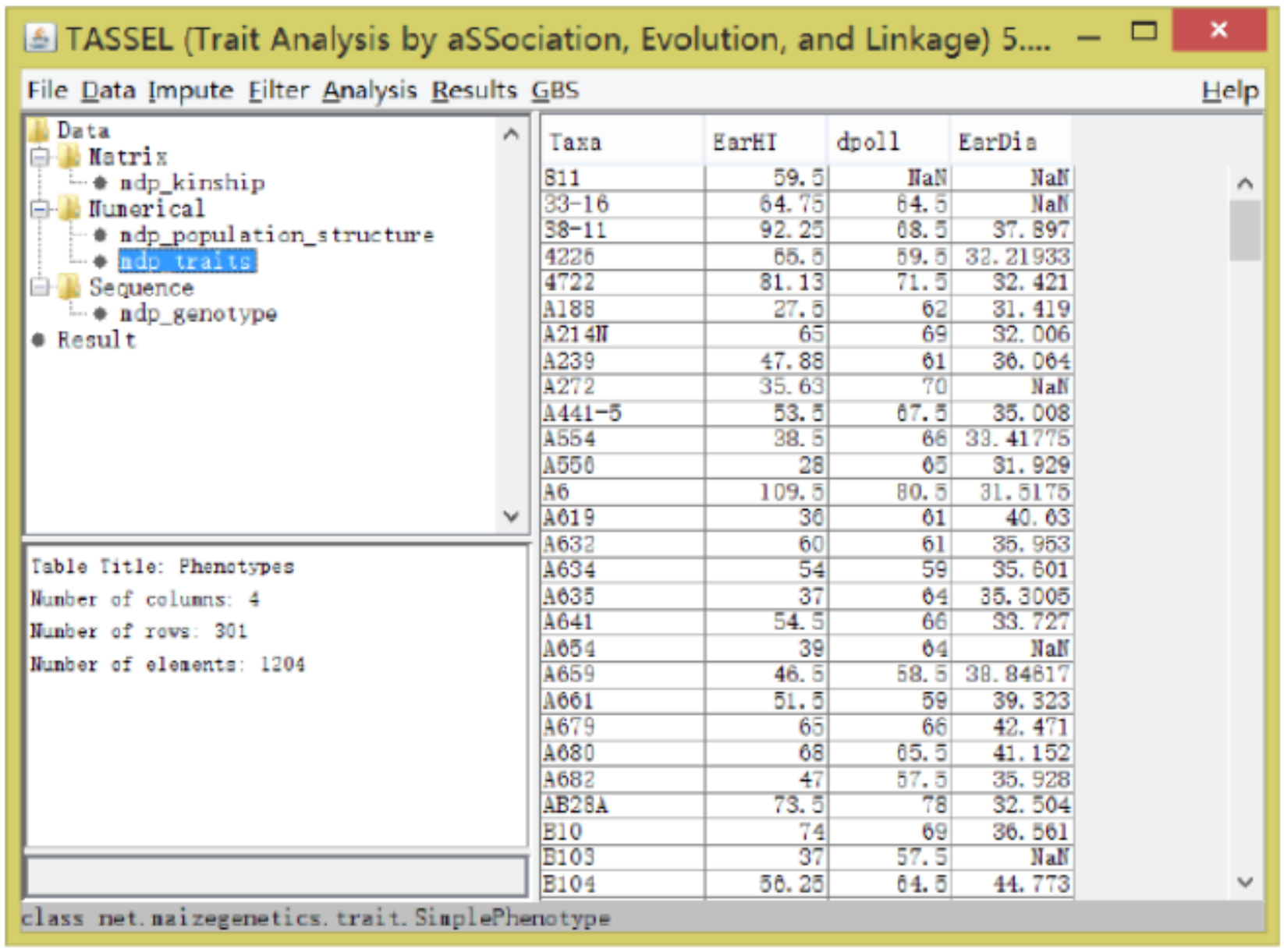
7.7 Manhattan Plot（曼哈顿图）

8 教程

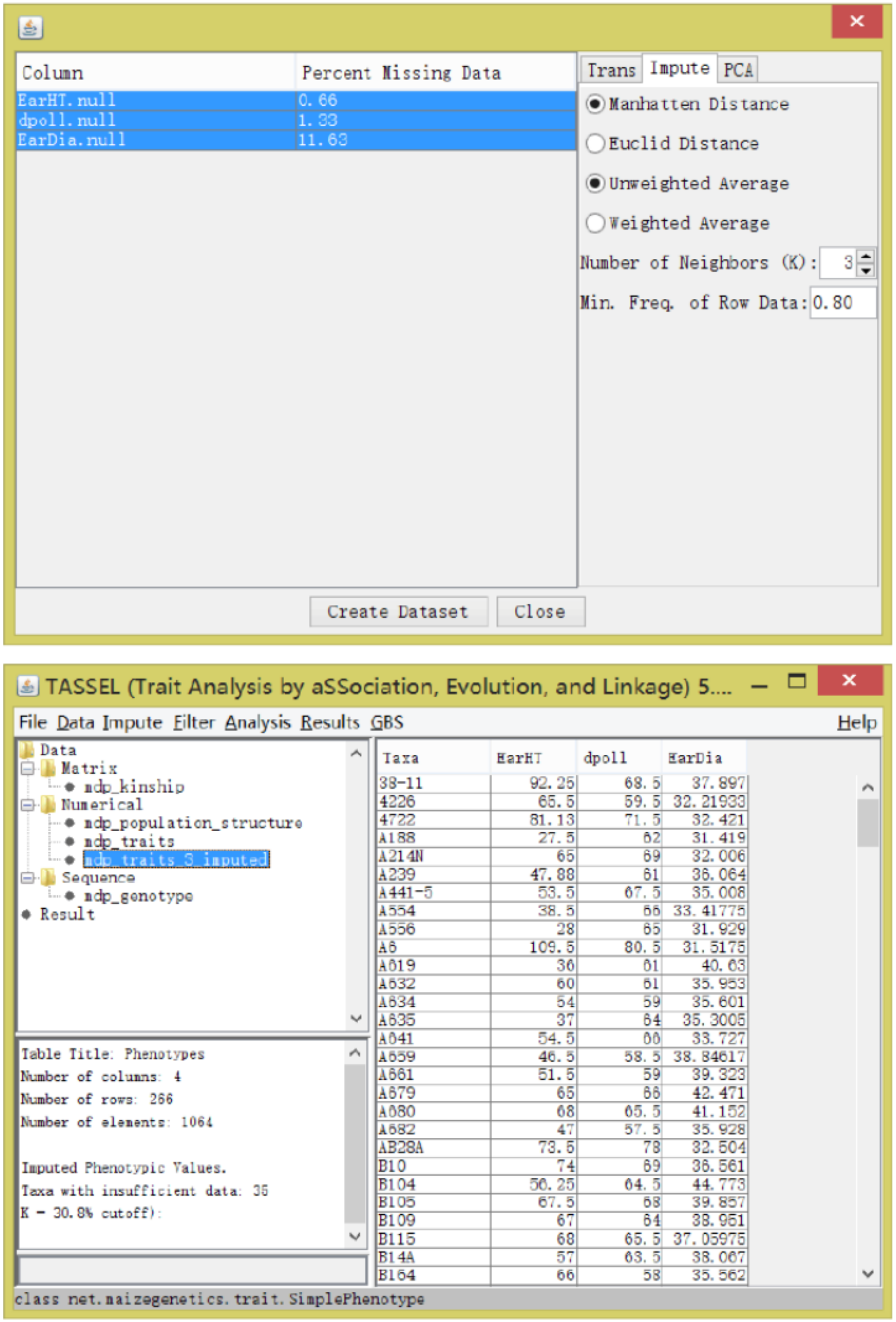
这个教程综述使用 TASSEL 的若干常见的情况，以便帮助用户更好地理解它的数据处理和关联分析的能力。TASSEL 软件包包含一个教学数据集，可以从 TASSEL 网址下载（请将所有文件解压到你选择的一个目录）。这个教学数据集包含表现型、基因型、群体结构以及亲缘关系的数据。

8.1 缺失表现型的估算

将用表现型文件 mdp_traits 来说明估算缺失数据的过程。注意下面的数据集包含缺失值（NaN）。



为了估算缺失数据，首先在数据树面板中选择 mdp_traits 数据集，然后单击 Transform（转换）按钮（Data -> Transform）。将打开“Transform Column Data”窗口。在这个窗口中单击 Impute（估算）标签。最后，单击“Create Dataset”按钮来产生具有估算的缺失值的新数据集。

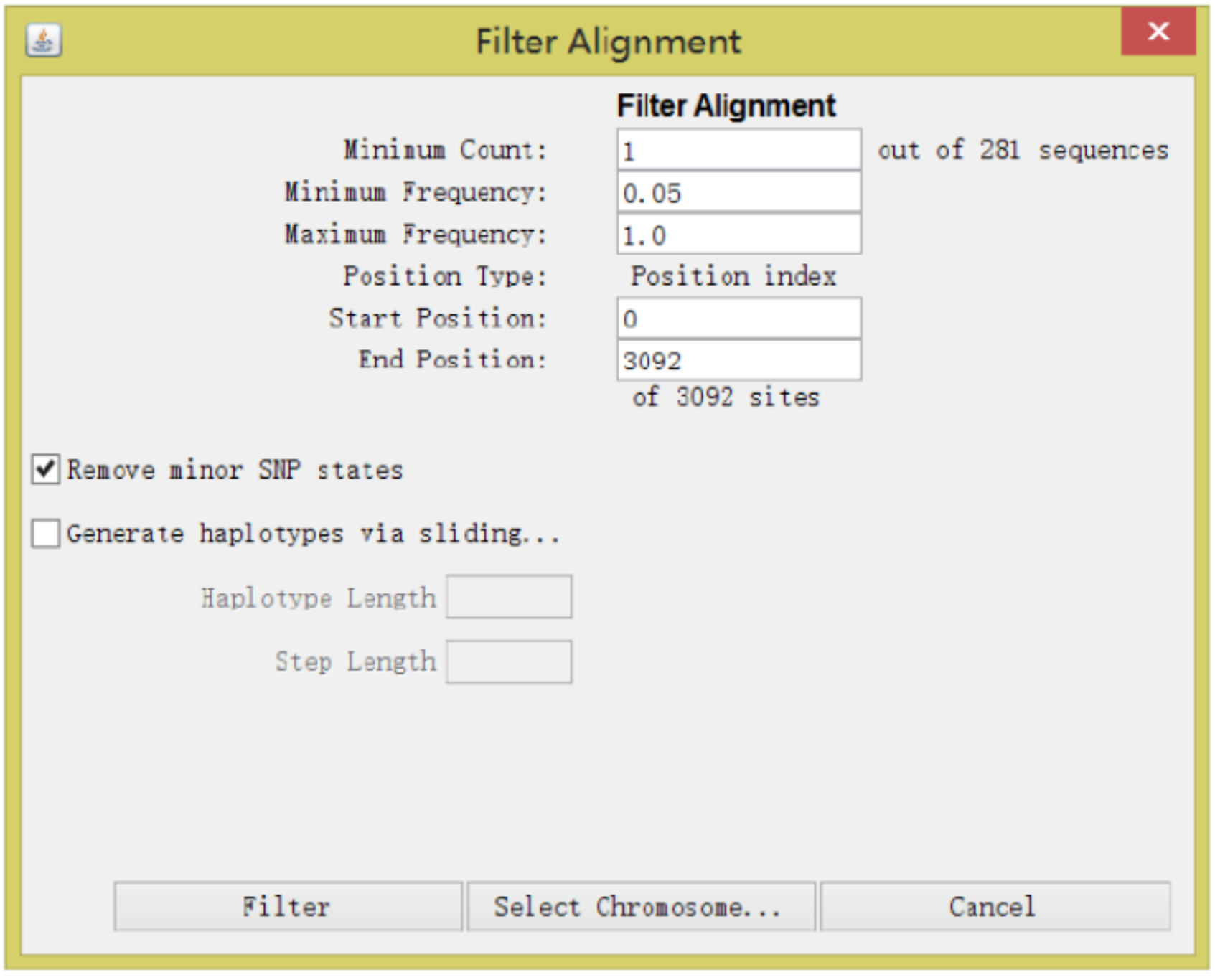


注意现在缺失值被填充了。

8.2 主成分分析

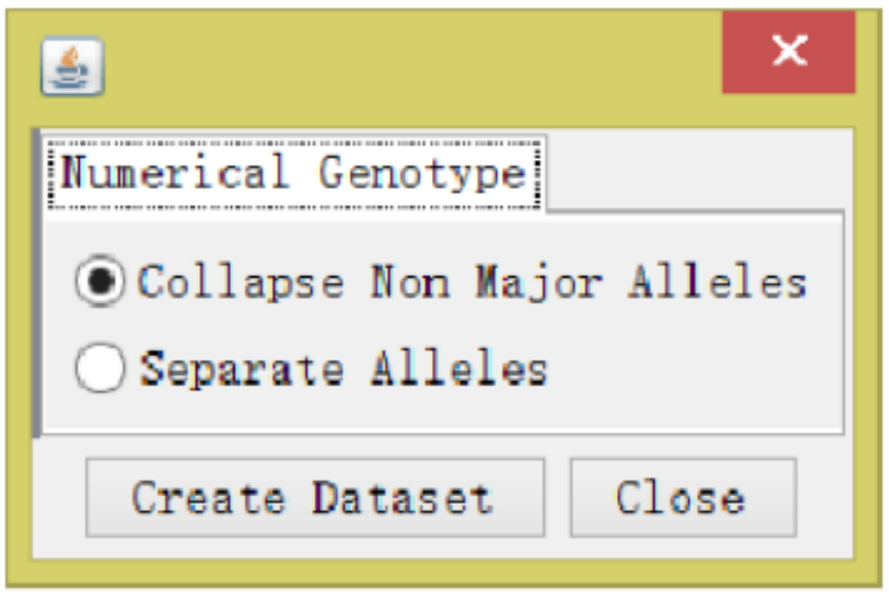
主成分分析（PCA）是一个统计工具，它将一组相关的变量转换成少量不相关的变量（称为主成分（PC））。第一 PC 捕获尽可能多的变异，后继的 PC 解释剩余方差的一个减少的百分率。PCA 的另一个应用是使用由遗传标记获得的 PC 来代表群体结构。这个方法需要的计算时间比最大似然估计少得多。由于大多数标记数据是字符，必须首先进行数字化。转换字符标记得分的一个常用的方法是把一个纯合体设置为 0，另外一个纯合体设置为 2，杂合体设置为 1。对于单倍体，可以通过把一个等位基因编码为 0、另外一个等位基因编码为 1 来进行转换。TASSEL 中的 TRANSFORM 功能把主要等位基因转换为 0，所有其它的等位基因被折叠到单个类别并编码为 1。PCA 要求全部变量应该具有变异并且不应该具有缺失值。因此，过滤基因型来剔除单态的标记并估算缺失值可能是必需的。估算缺失值可以在数字化之前或之后进行。这里我们将演示如何从教学数据的基因型文件中产生 PC。

1. **删除单态的位点：**确保 TASSEL 是处于 Data 模式之中的。加亮基因型文件然后单击 Filter -> Site（位点）。把“Minimum Frequency”（最低频率）设置为 0.05，复选“Remove minor SNP status”。单击 Filter（过滤器）按钮。

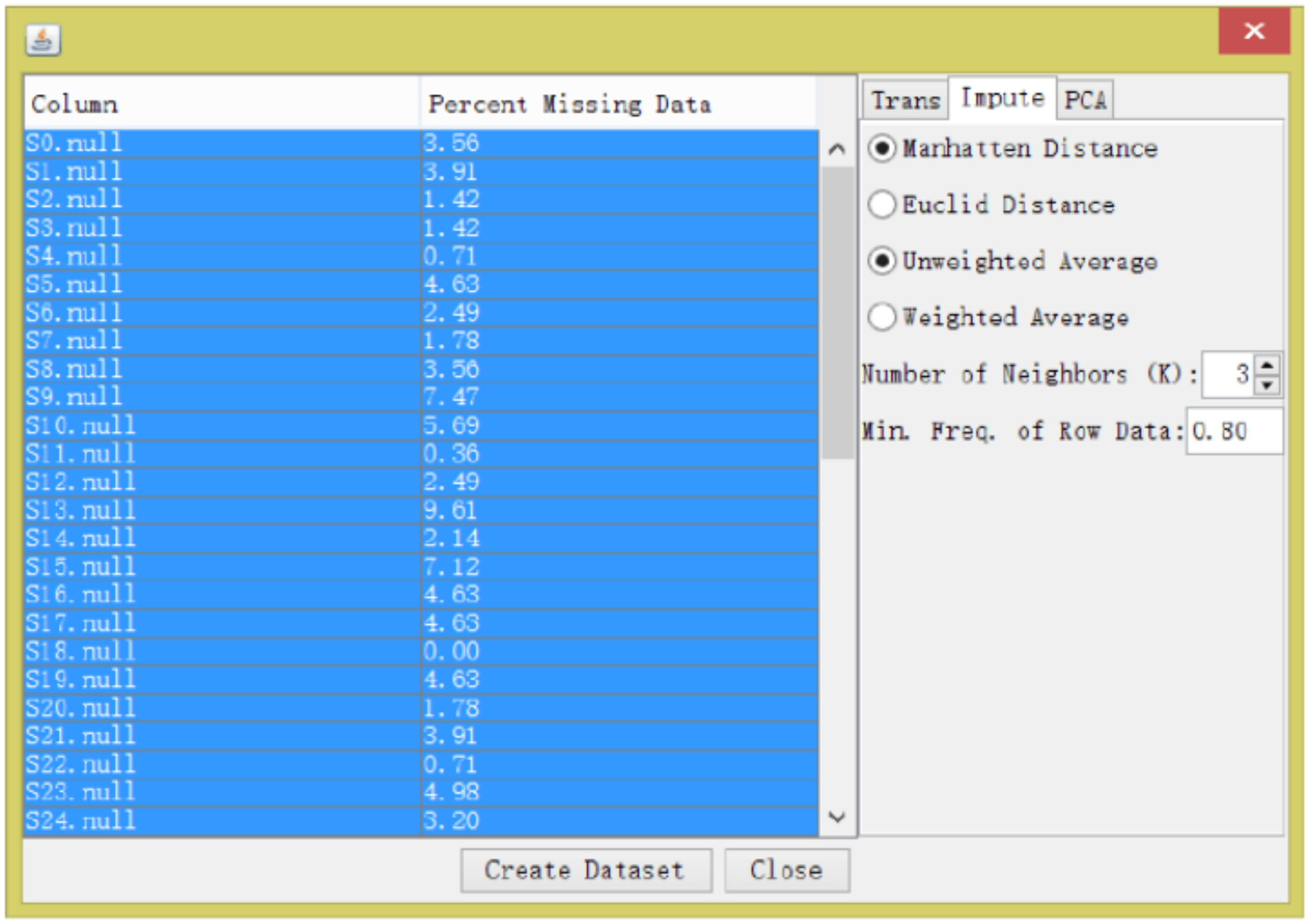


2. **数字化：**加亮过滤了的基因型文件然后单击 Data -> Transform（转换）。使用默认

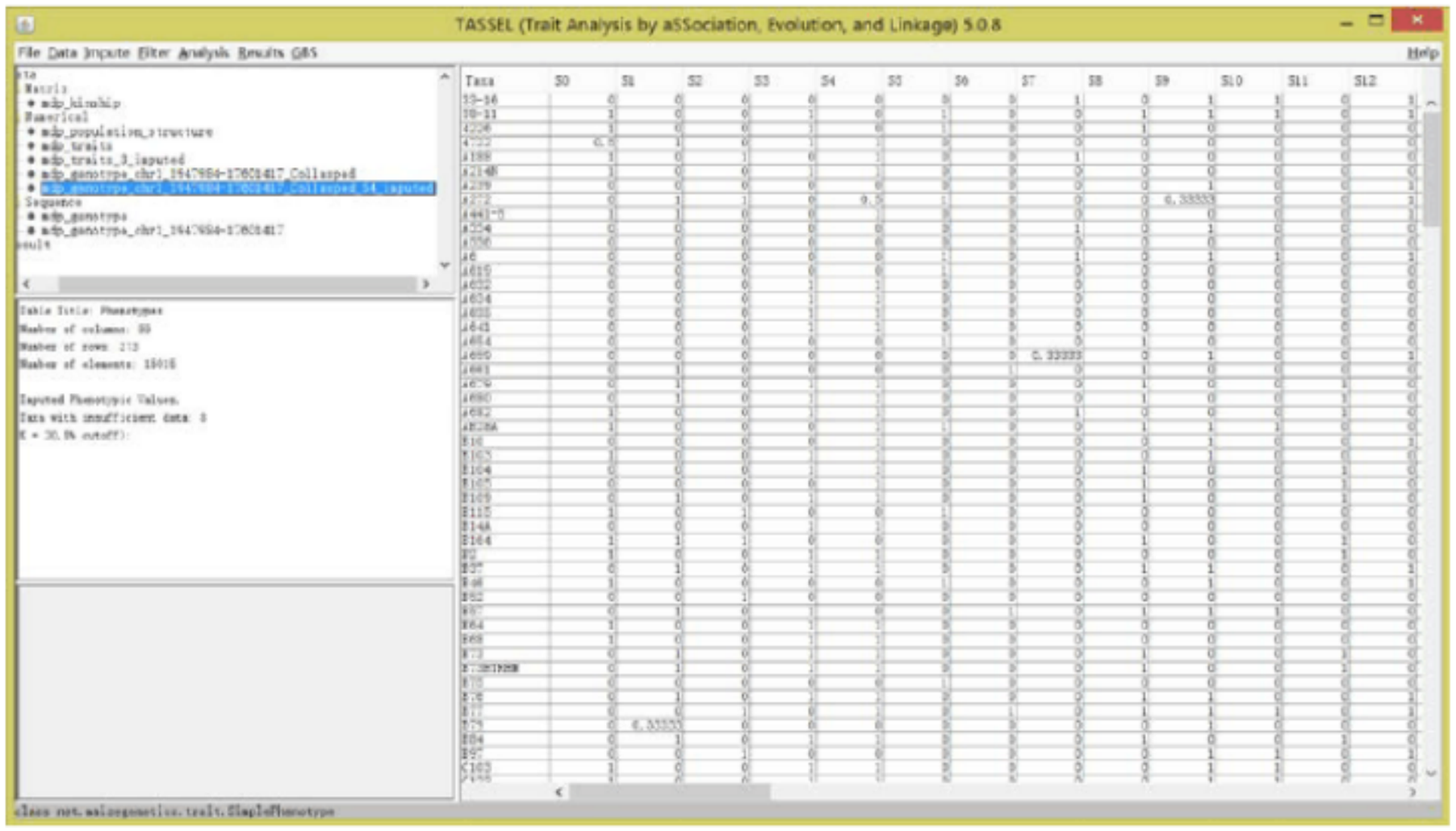
选项“Collapse non major alleles.” 单击“Create dataset”。

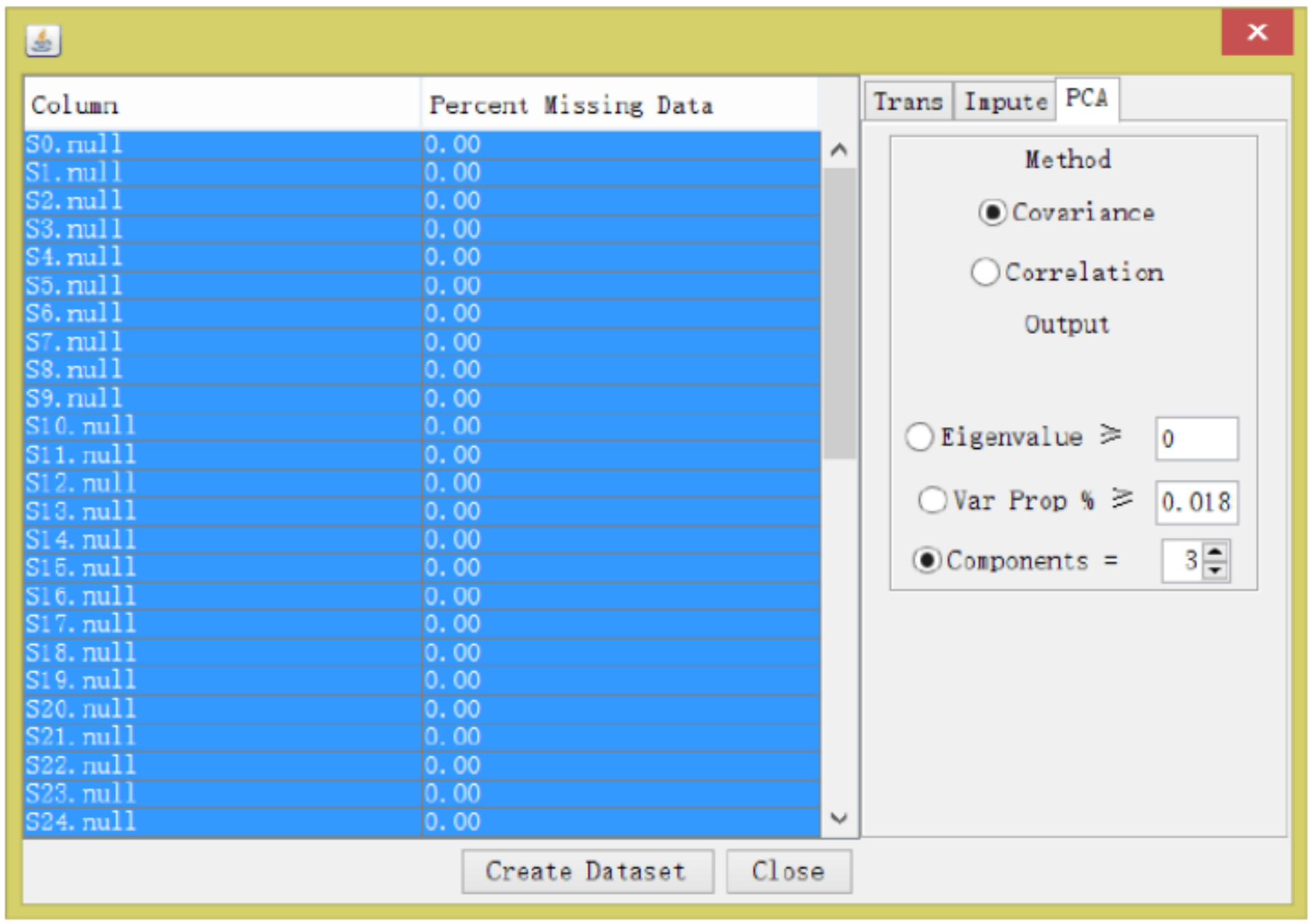


3. **估算缺失值：**加亮数值的基因型文件，单击 Data -> Transform（转换），然后单击 Impute（估算）标签。使用默认选项。单击“Create dataset”。

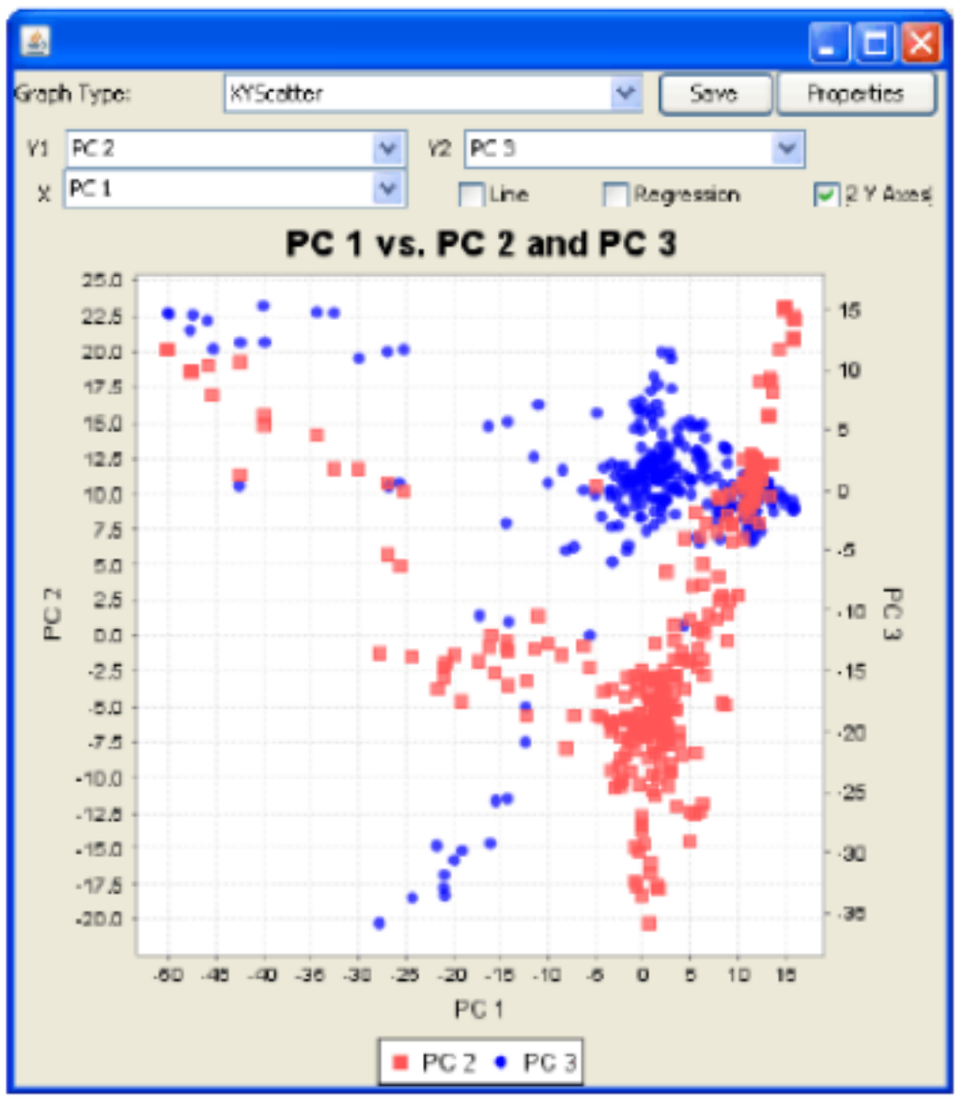


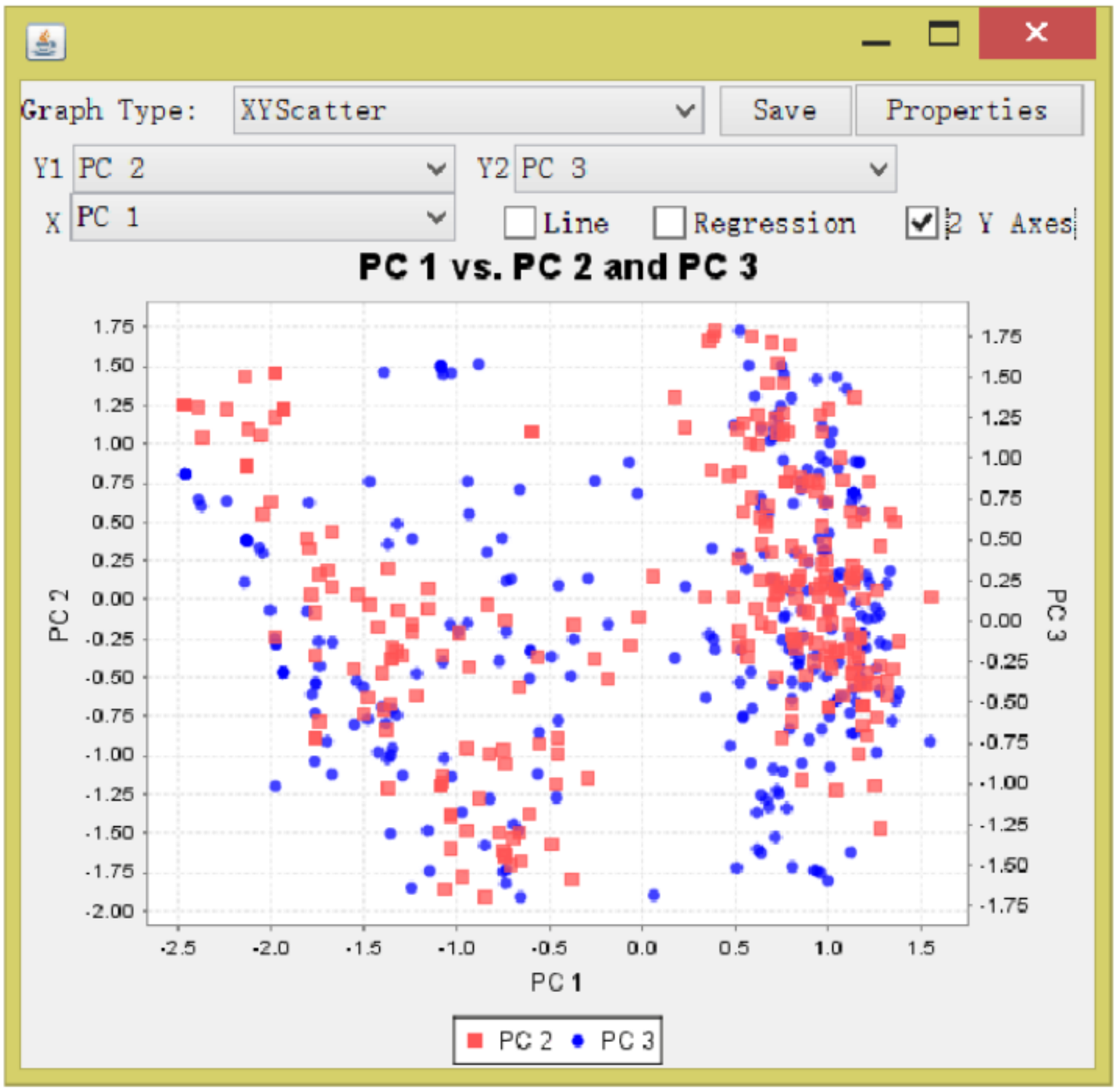
4. **PCA：**加亮估算的数值基因型，单击 Transform（转换），然后单击 PCA 标签。通过选择 Components（成分）然后在文字框中键入 3，把默认选项改为“Components＝3”。单击“Create dataset”。

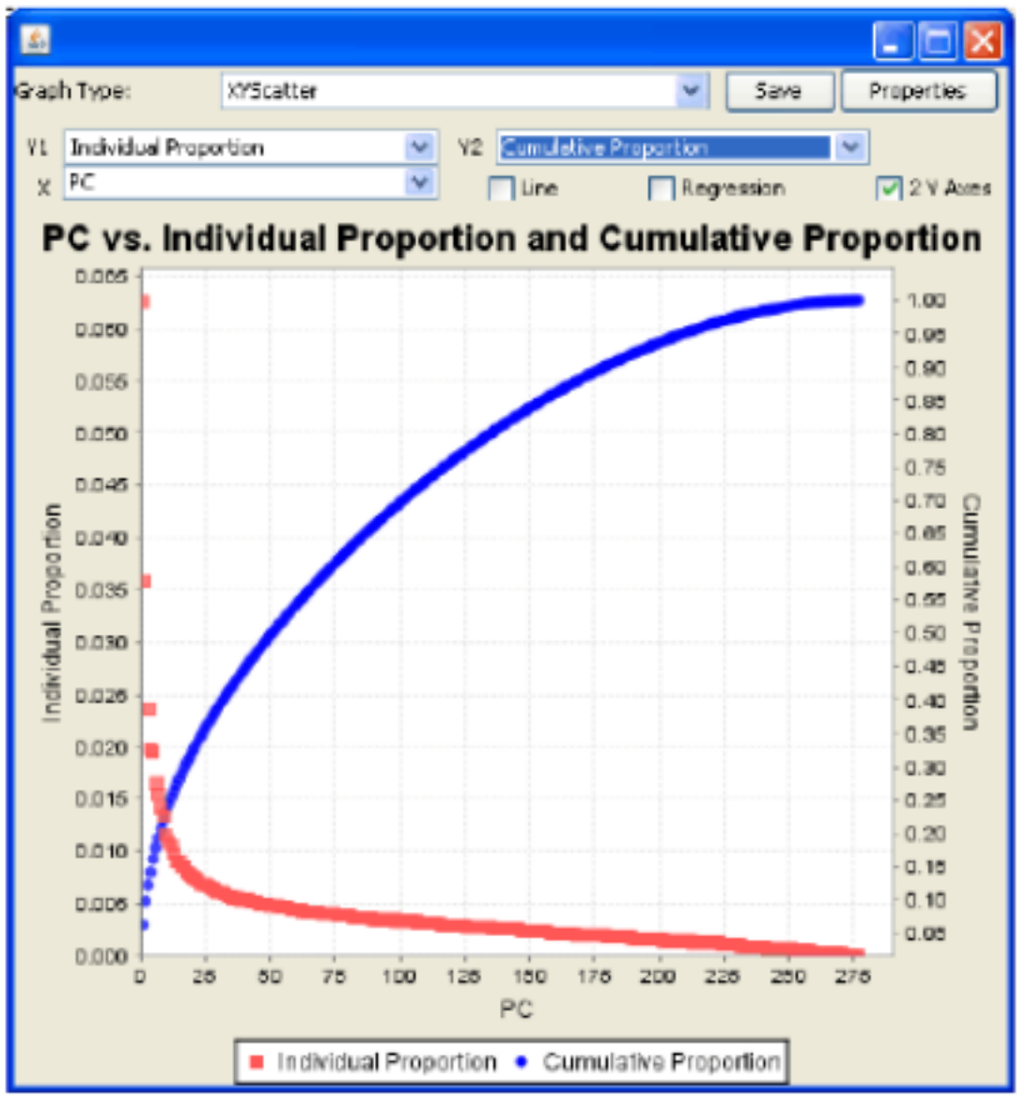


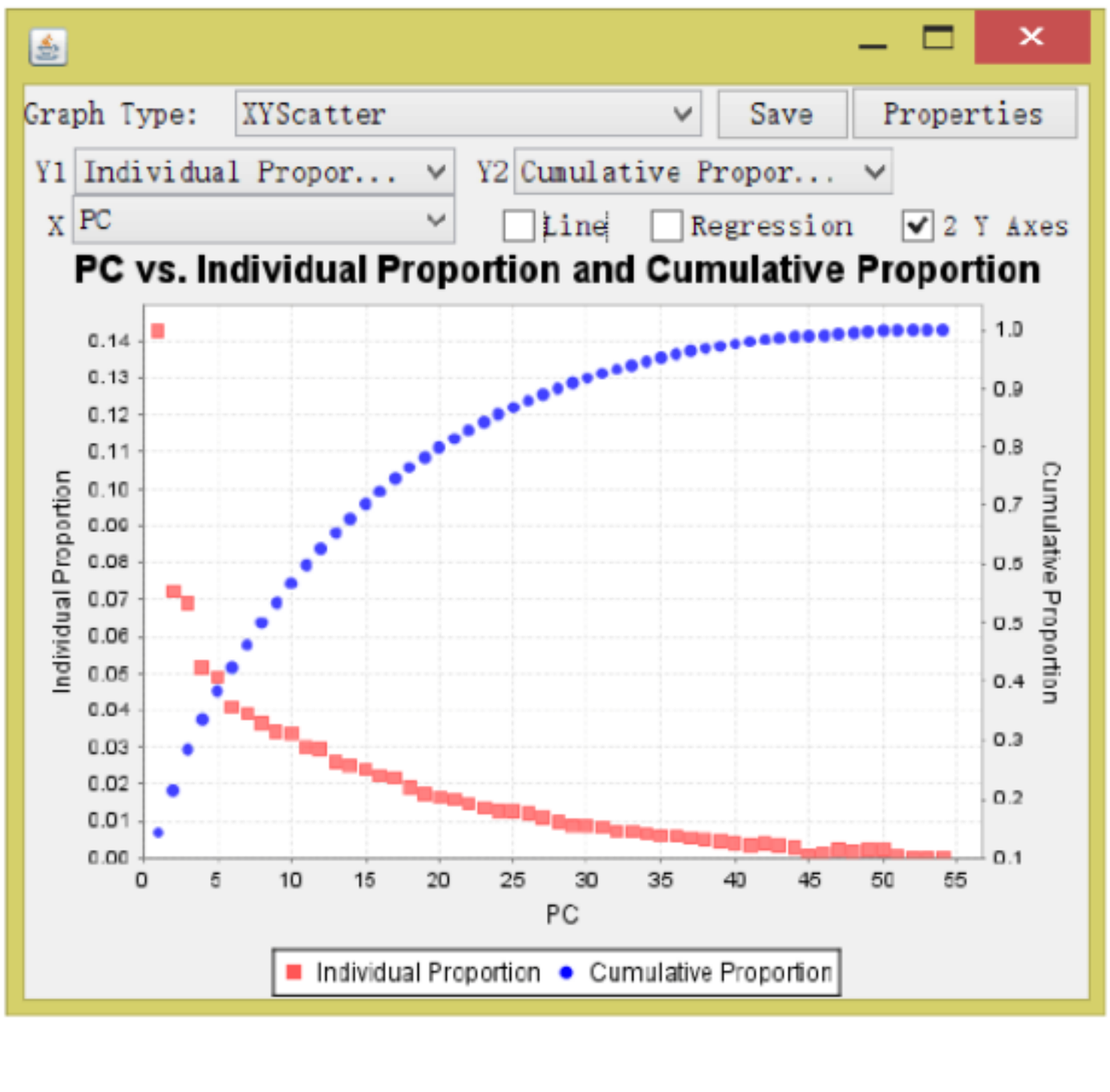


在运行 PCA 之后三个项将被添加到数据树。第一个是 PCs。第二个是特征值。最后一个是特征向量。这里我们使用 Result（结果）模式中的 Chart（图表）功能来对前面三个 PCs、各别的特征值贡献（有时称为一个 skree 图）以及累积的特征值贡献绘图。特征值是所关心的，因为它们等于由每个 PCs 解释的方差。


Graph Type:
XYScatter
Save
Properties
Y1
PC 2
Y2
PC 3
X
PC 1
Line
Regression
2 Y Axes
PC 1 vs. PC 2 and PC 3
PC 2
PC 3
PC 1
PC 2
PC 3


Graph Type:
XYScatter
Save
Properties
Y1
PC 2
Y2
PC 3
X
PC 1
Line
Regression
2 Y Axes
PC 1 vs. PC 2 and PC 3
PC 2
PC 3
PC 1
PC 2
PC 3

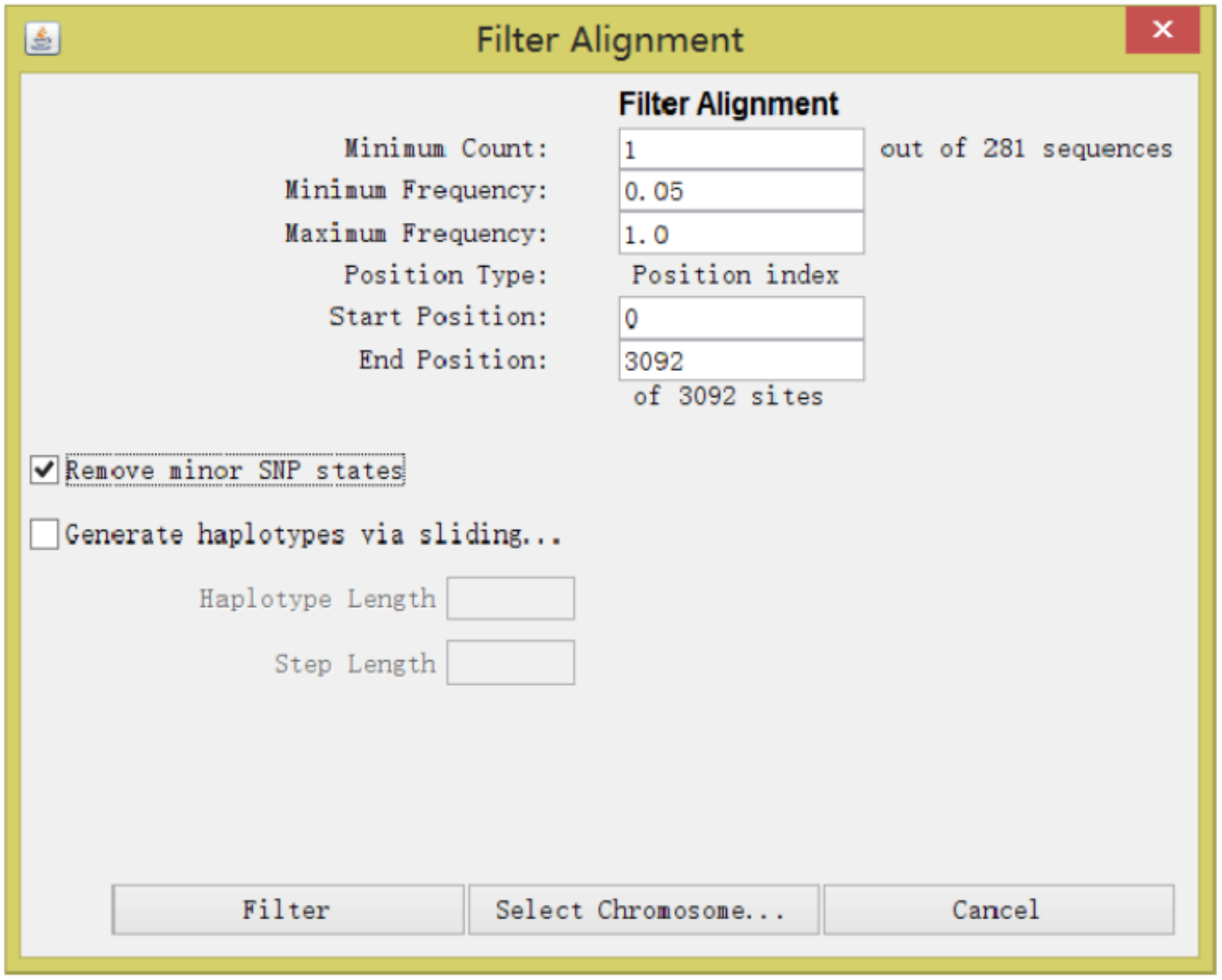




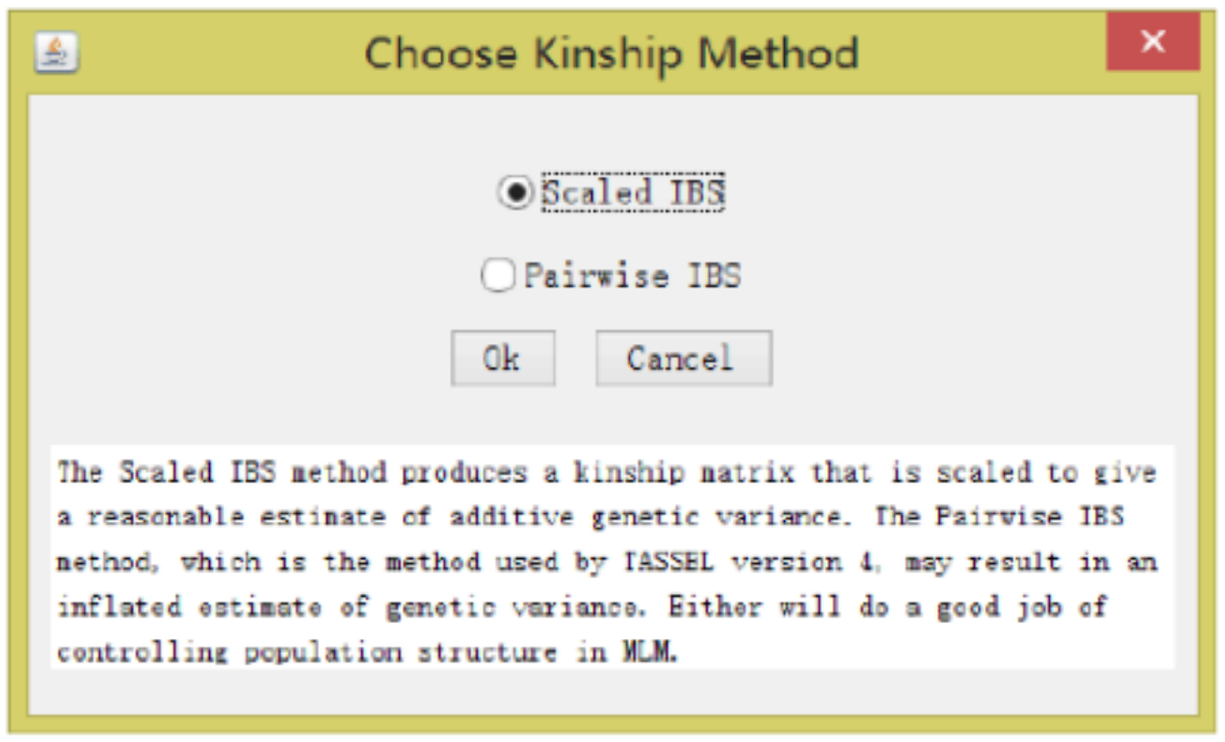
8.3 利用遗传标记估计亲缘关系

虽然 PC 可用于捕获主要的群体再分，但是亲缘关系可用于捕获更微妙的关系。本节说明如何根据计算 PC 时使用的相同的 SNP 数据来产生一个亲缘关系矩阵。

1. **删除单态的位点：**加亮基因型文件，在菜单栏上选择 Filter -> Sites。在 MAF 上把阈值设置为 0.05，复选“Remove minor SNP status,”然后单击 Filter（过滤器）。

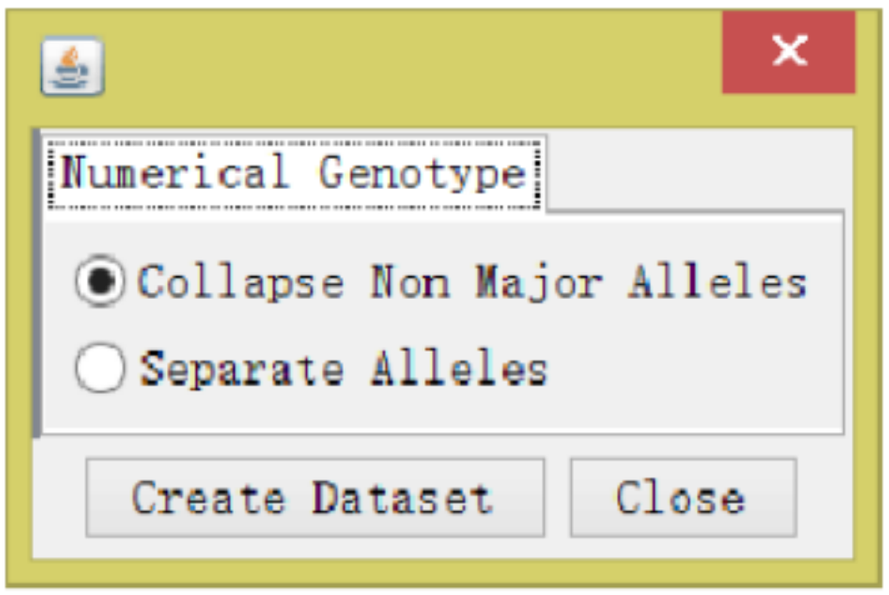


2. **估计亲缘关系：**加亮过滤了的基因型文件然后单击 Analysis -> Kinship。在“Choose Kinship Method”对话框中选择“Scaled IBS”然后单击 OK。一个亲缘关系矩阵将被添加到数据树，在 Matrix（矩阵）类别下。

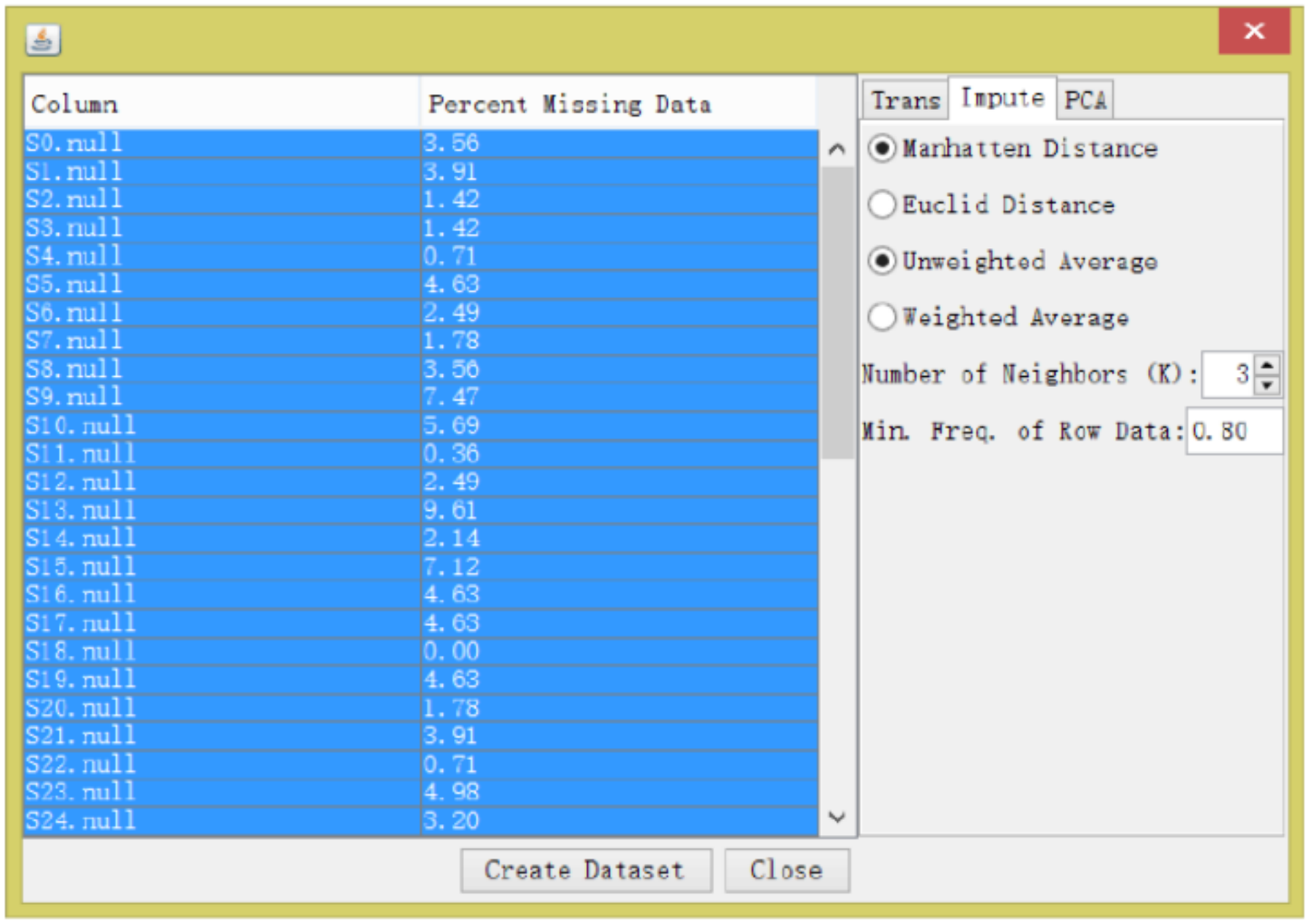


3. 也可以首先估算缺失的基因型数据然后利用估算的数据产生亲缘关系矩阵。要估算缺失数据，加亮过滤的基因型，选择 Data -> Transform，选中“Collapse Non-Major Alleles”，然后单击“Create Dataset”。一个附加了“_Collapse”的新数据集将出现在“Numerical”文

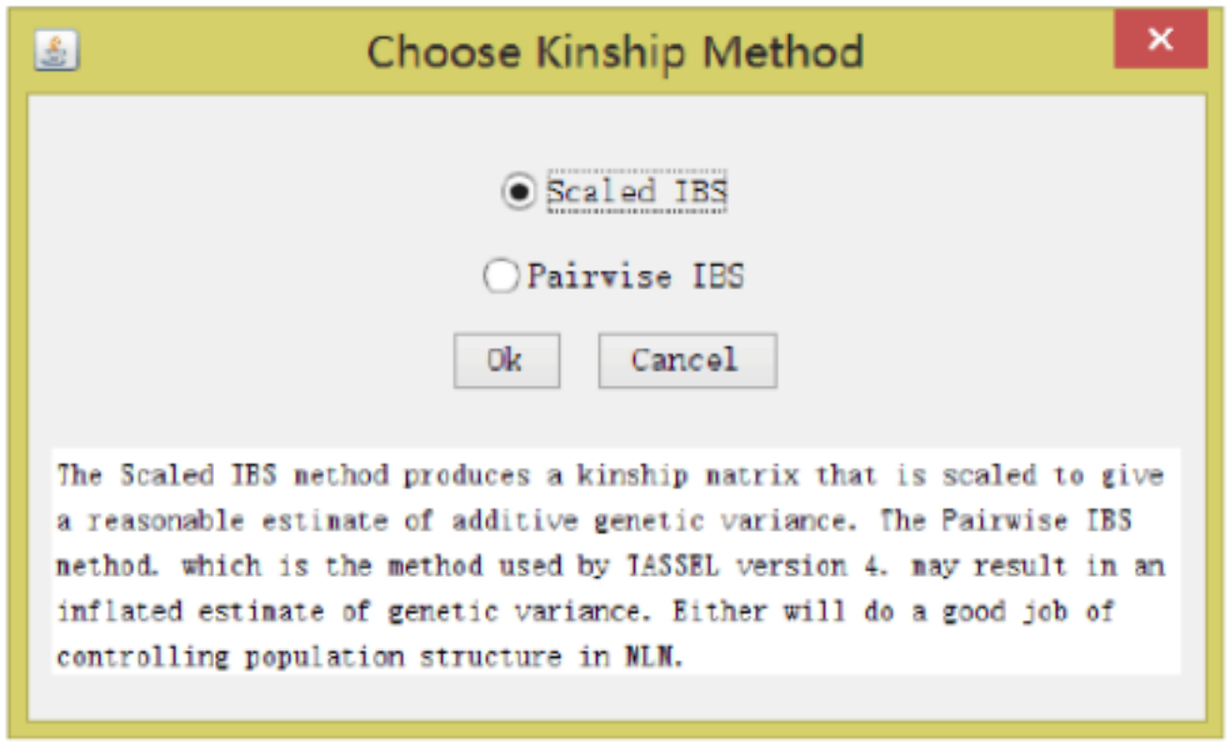
件夹中。

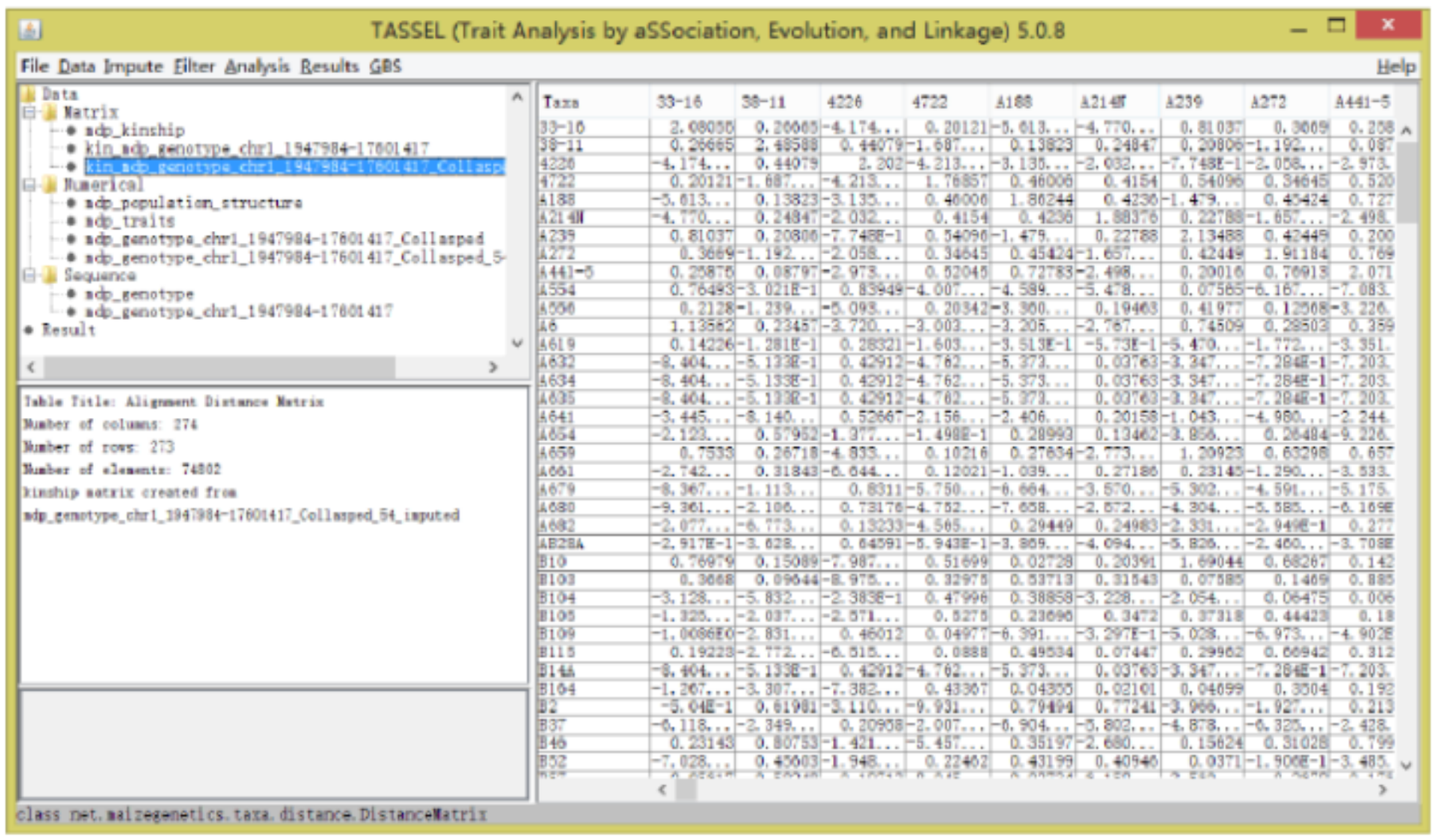


加亮折叠的数据集，选择 Data -> Transform，选择 Impute（估算）标签，然后单击“Create dataset”。



加亮产生的估算数据然后选择 Analysis -> Kinship。



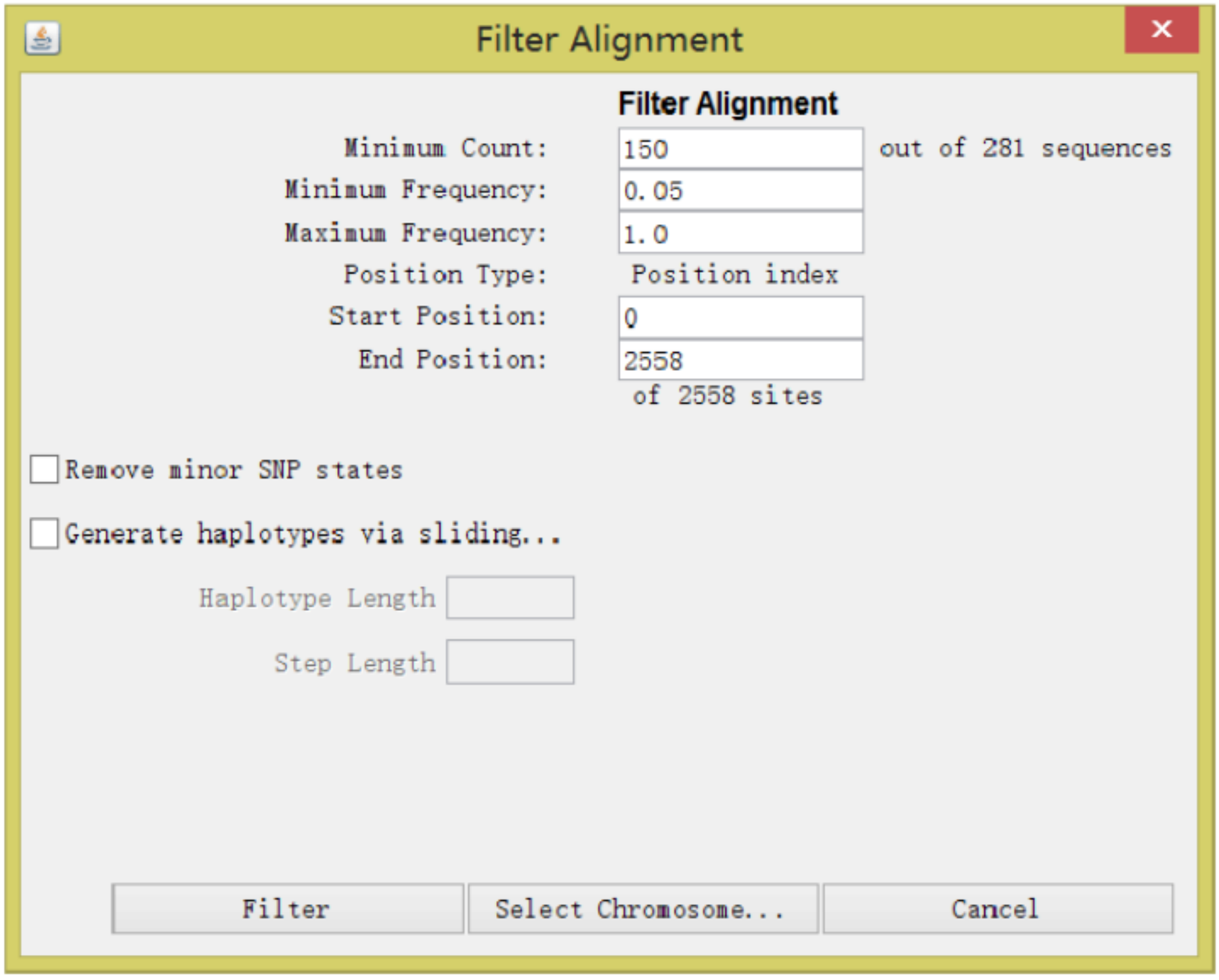


8.4 利用 GLM 进行关联分析

我们使用来自教学数据集的三个文件，利用 GLM 进行关联分析。第一个文件是 mdp_genotype.hmp.txt，在 3093 个位点、281 个玉米自交系上的一套 SNPs 得分。第二个文件是 282 个玉米自交系的群体结构（mdp_population_structure.txt）。最后一个文件是 282 个玉米自交系的三个性状的表现型（mdp_traits.txt）。统计模型是：

开花期 = 群体结构 + 标记效应 + 残差

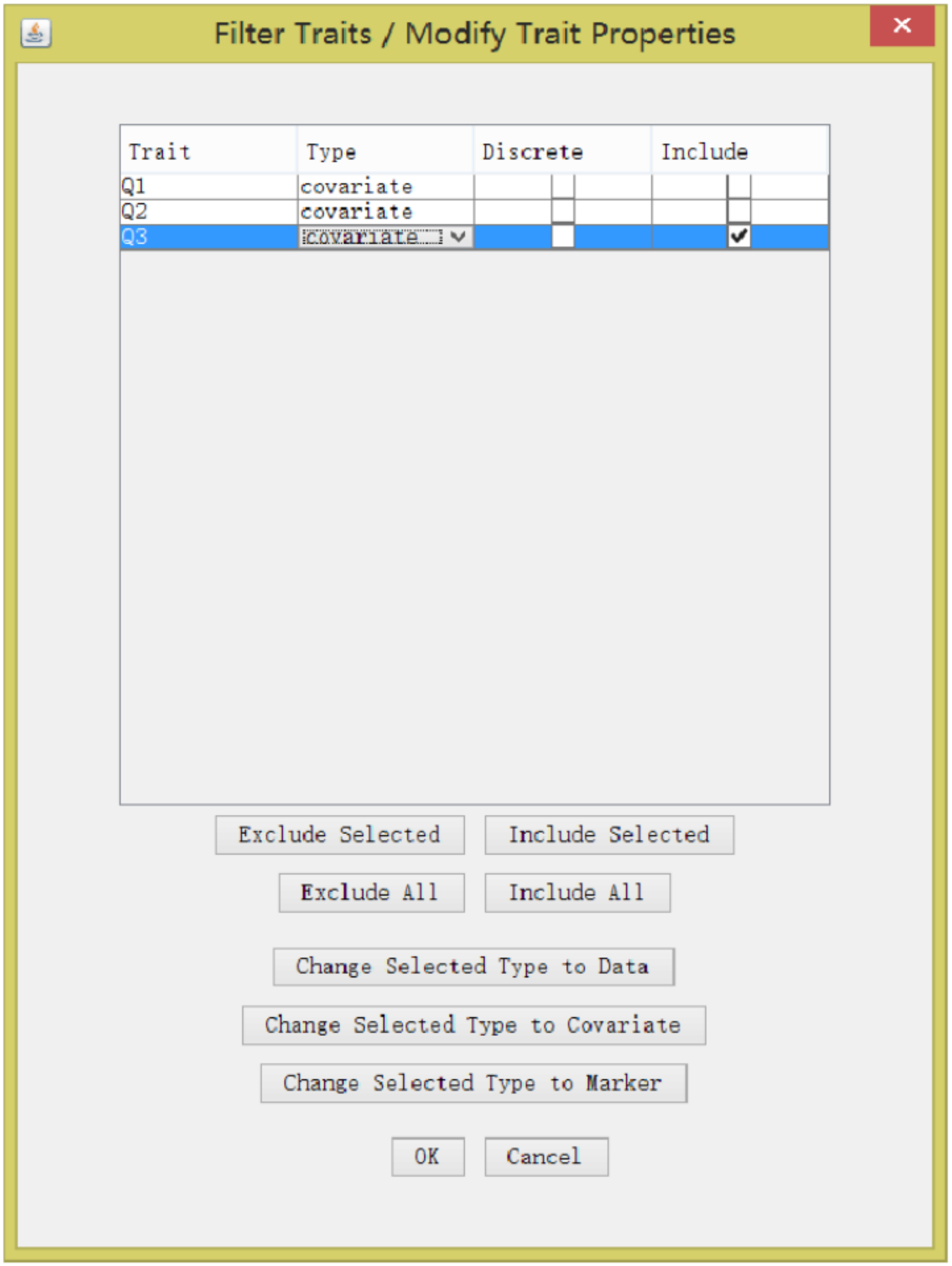
1. **删除单态的和低覆盖度的位点：**加亮 mdp_genotype，在菜单栏上单击 Filter -> Sites。把“Minimum Frequency”设置为 0.05、“Maximum Frequency”设置为 1.0，“Minimum Count”设置为 150。单击 Filter（过滤器）来产生一个过滤了的基因型数据集。



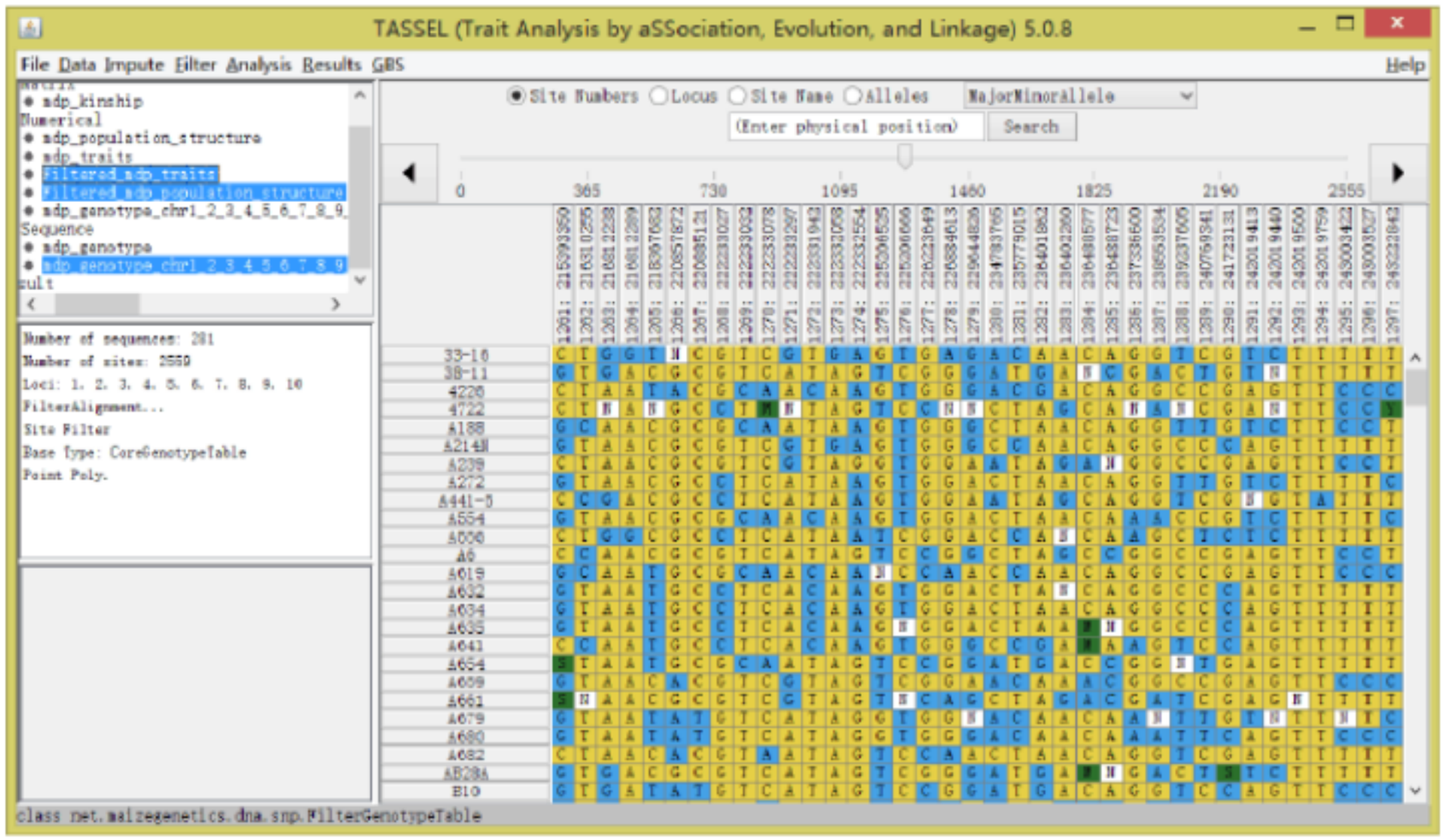
2. **性状选择：**加亮表现型，单击菜单项 Filter -> Traits。除开花期（DPOLL）之外清除所有的性状。确保 Type（类型）被设置为 Data（数据）。单击 OK 来产生一个过滤的表现型。

3. **协变量选择：**群体结构被作为每个群体的比例给出。有三个群体，表示为 Q1、Q2 和 Q3。它们的总和为 100%。如果我们把它们全部作为协变量使用，这会产生线性相依性（dependency）。虽然 GLM 可以正确地处理这种相依性，但是它将导致 MLM 抱怨并且拒绝完成你的分析。我们可以通过删除 Q 变量的一个来消除相依性。在这个演示中，我们除去最后一个。加亮 mdp_population_structure 然后单击 Filter -> Traits。清除最后一个群体（Q3）。确保 Type（类型）被设置为 Covariate（协变量）。然后单击 OK 来产生一个过滤的群体结构

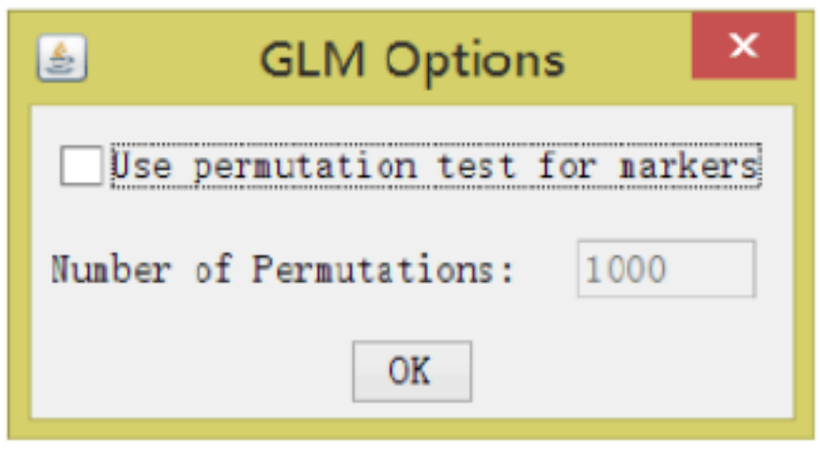
数据。

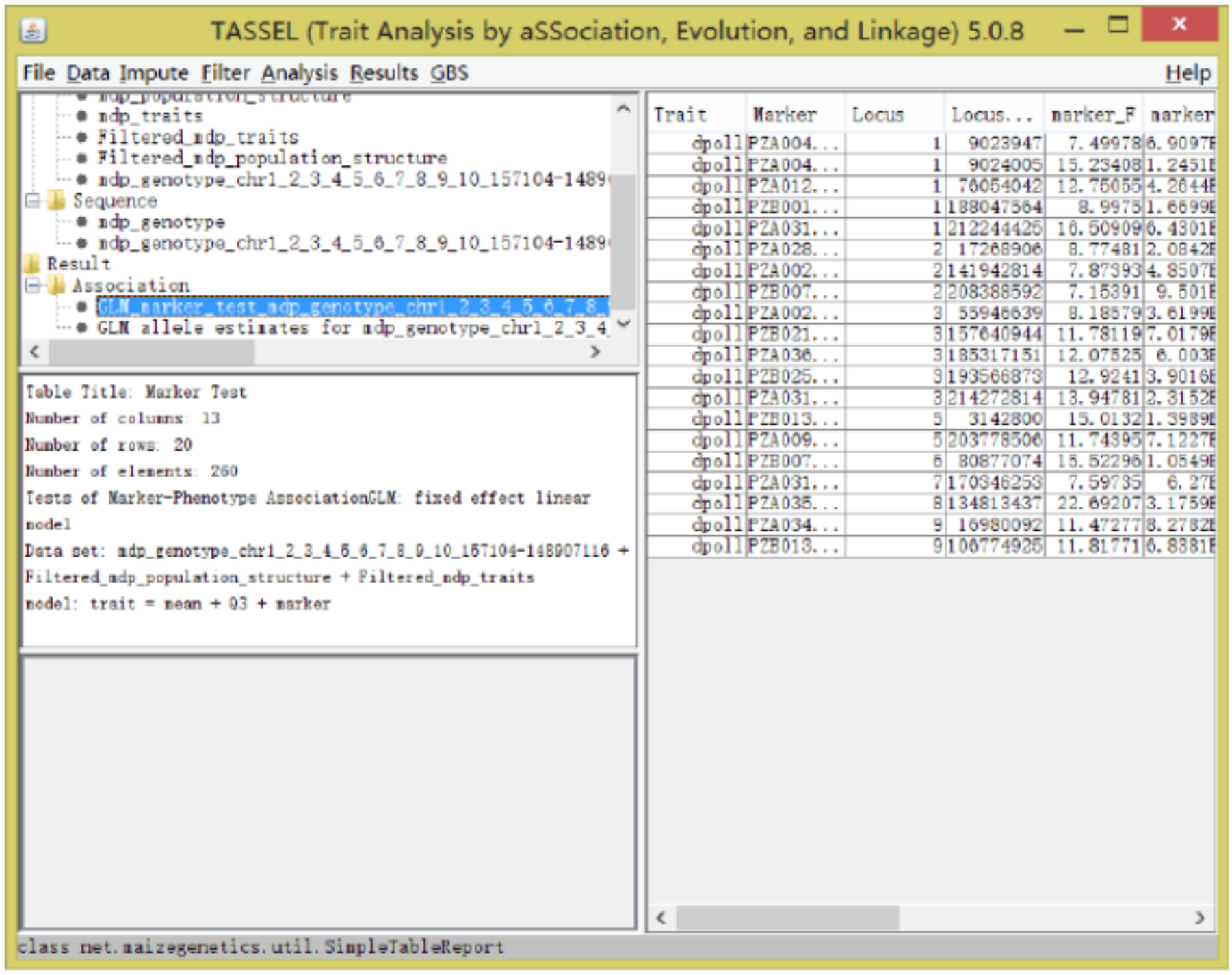


4. **合并数据：**通过按住 Ctrl 键同时选择各别的数据集来加亮三个过滤的数据集。然后单击菜单项 Data -> Intersect Join 来产生一个合并的数据集。



5. **关联分析：**加亮合并的数据集然后单击菜单项 Analysis -> GLM 来进行关联分析。两个报告将被添加到数据树。





添加到数据树的一个报告被标签为“GLM_Marker_Test_”继之以合并数据的名称。除了性状和标记的信息之外，该数据集还包含下列统计量：

marker_F：对标记进行 F 检验的 F 值；

marker_p：对标记进行 F 检验的 P 值；

markerR2：在配合其它的模型项（群体结构）之后标记的 R2；

markerDF：标记的自由度；

markerMS：标记的均方；

errorDF：剩余误差的自由度；

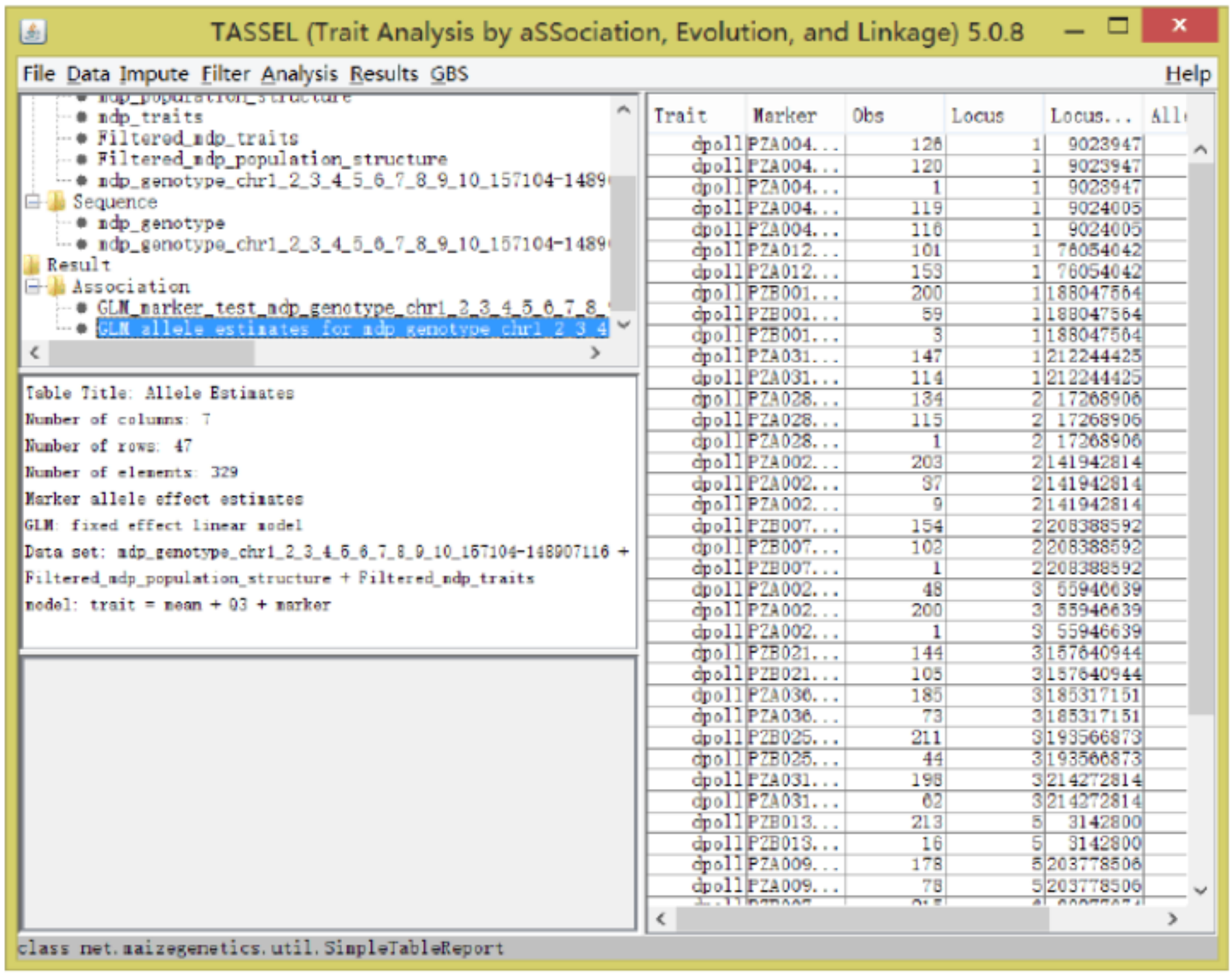
errorMS：剩余误差的均方；

modelDF：模型的自由度；

modelMS：模型的均方。

单击“marker_p”将按照 P 值对表格排序。最小的 P 值是 3.5963×10^{-6}。一个合理的显著性阈值是 1.9×10^{-5}，它是在 Bonferroni 多重检验矫正之后的 5%（0.05/2559）。Bonferroni 矫正中的分母是检验的 SNPs 的总数。该关联是显著的。

添加到数据树的另一个报告被标签为“GLM_Allele_Estimates_”继之以合并数据的名称。对于最显著的 SNP（在下面的图中加亮显示），有两个基因型（AA 和 GG）。有 220 个品系具有基因型 AA，41 个品系具有等位基因 GG。对于性状 dpoll（到传粉的天数），两个纯合体之间的差数是 3.86 天。



8.5 利用 MLM 进行关联分析

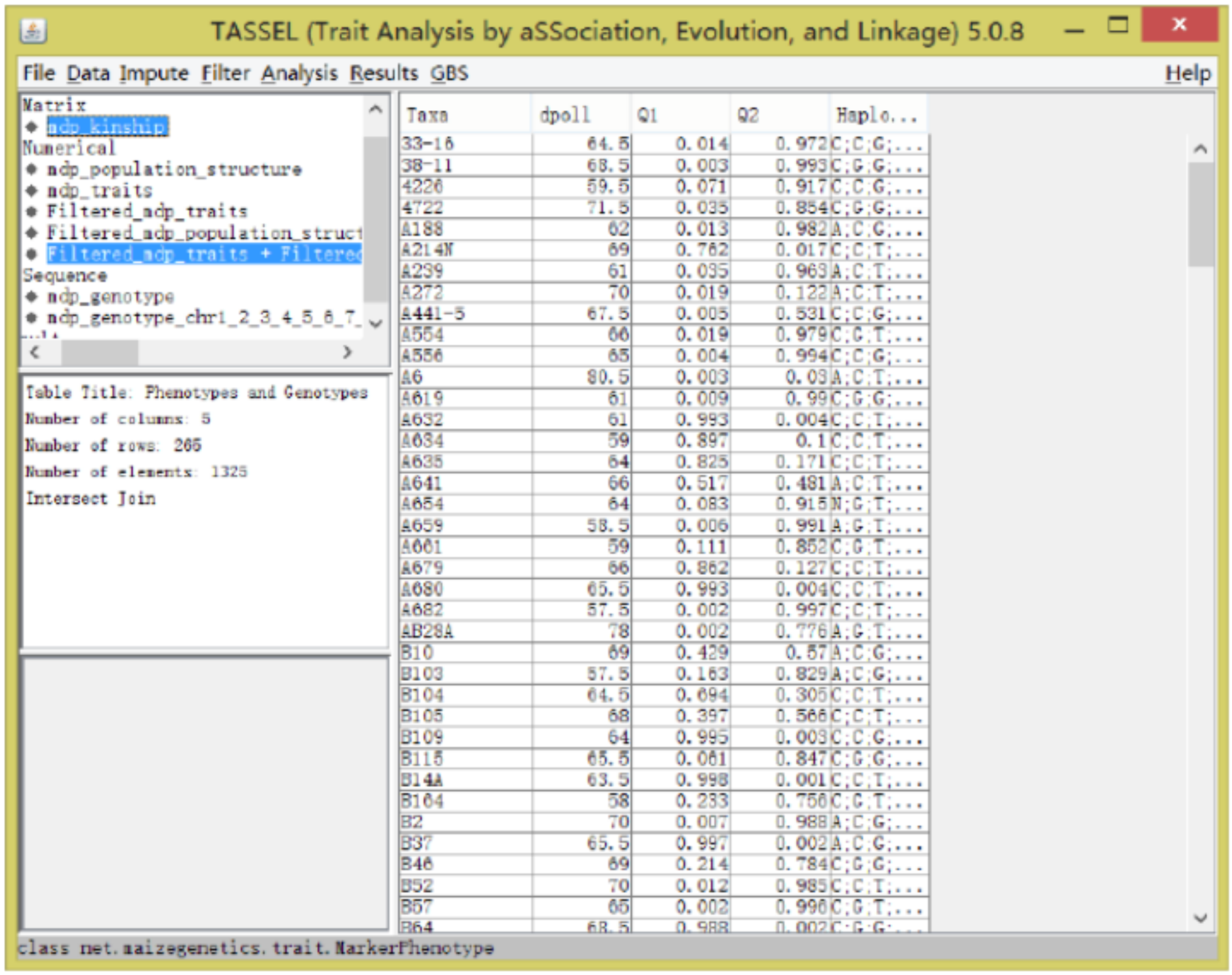
在 TASSEL 中运行 MLM 与运行 GLM 相似。区别是除了合并的数据（或者数值数据）之外，MLM 需要亲缘关系数据来定义个体之间的关系。亲缘关系矩阵乘以一个参数等于个体之间的协方差矩阵。这里我们使用来自教学数据集的亲缘关系文件来配合下列统计模型。

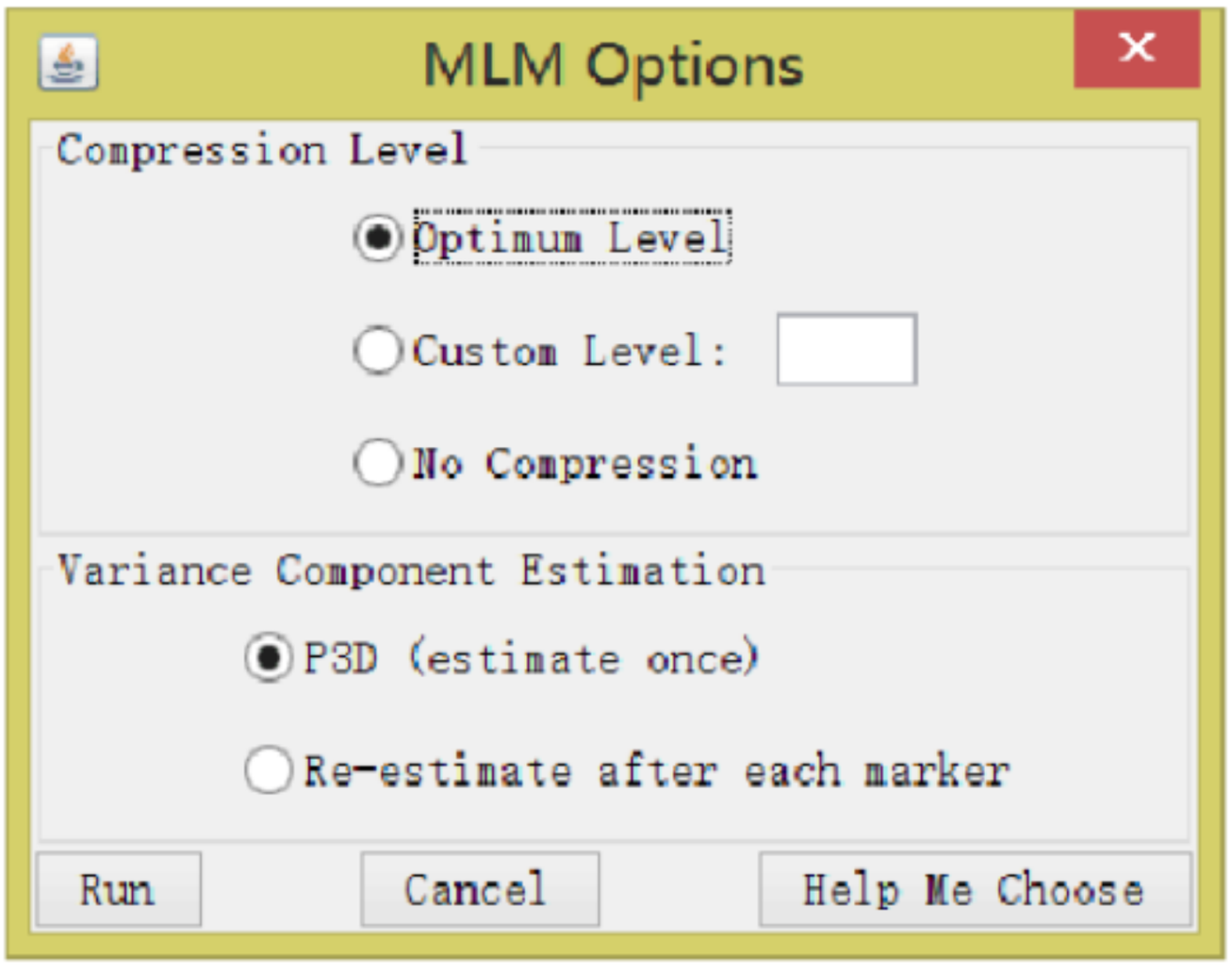
开花期 = 群体结构 + 标记效应 + 个体 + 残差

个体和残差是作为随机效应配合的。其它的项被当作固定效应。

至于标记效应，我们将利用两组标记来演示该分析。一组是在 GLM 教程中使用过的 dwarf8 基因序列。另外一组是跨越玉米基因组分布的一套 3093 个 SNPs。

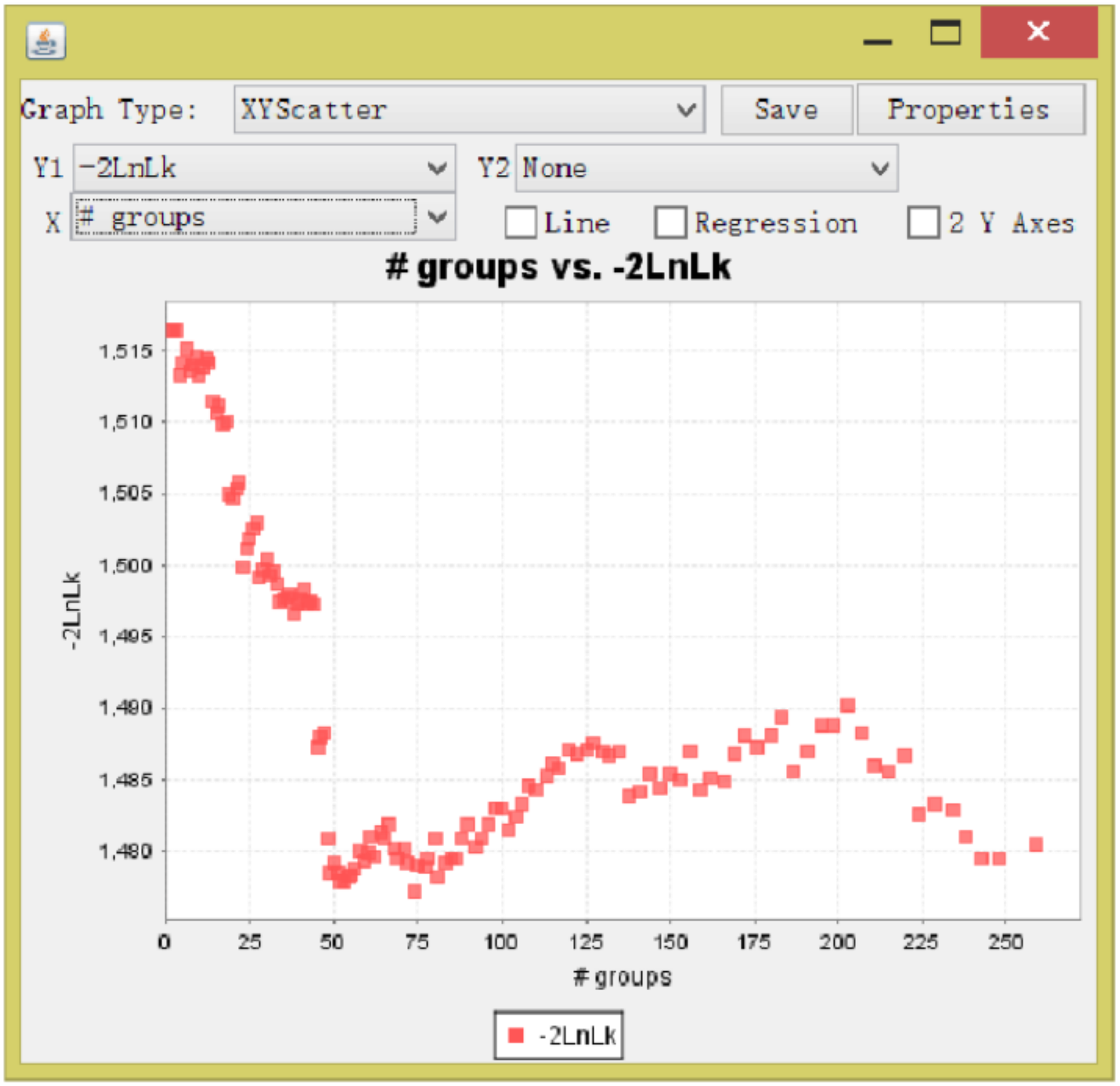
对于 dwarf8 基因序列，使用按照 GLM 的教程创建的合并数据集。通过加亮合并数据集和亲缘关系数据然后在 Analysis（分析）模式中单击 MLM 按钮来求解混合线性模型。

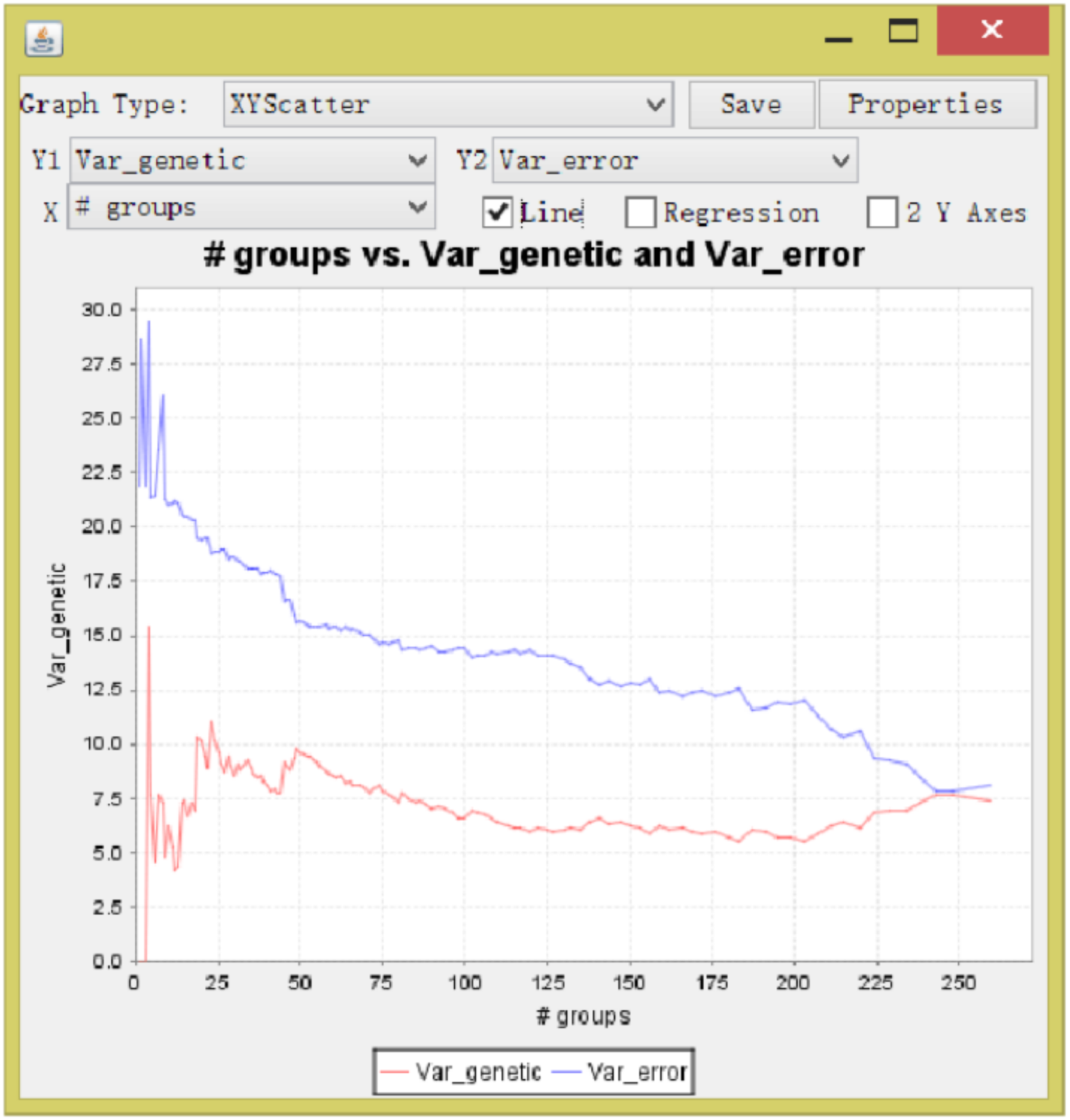




将弹出一个 MLM 选项对话框，如上所示。选择默认选项，它使用 P3D 和最优压缩水平上的压缩。在单击 Run 按钮之后，进度条将开始移动。需要的时间取决于样本容量、性状数目、标记数目，以及在 MLM 选项对话框中选择的选项。在进度条被重置为零之后，表示 MLM 完成了，三个报告将被添加到数据树。前面两个与由 GLM 产生的报告相似。最显著的 SNP 仍然是相同的，然而关联的强度要弱些，P 值为 7.199×10^{-4}（与来自 GLM 的 1.1021×10^{-4} 对照），它没有超过 Bonferroni 多重检验阈值（5×10^{-4}）。

第三个报告包含 MLM 专化的统计量，包括不同压缩水平下的 -2 对数似然函数（-2 Log Likelihood）、遗传方差和剩余方差分量。在 Result（结果）模式上通过 Chart（图表）功能对这些统计量进行作图，如同下述。

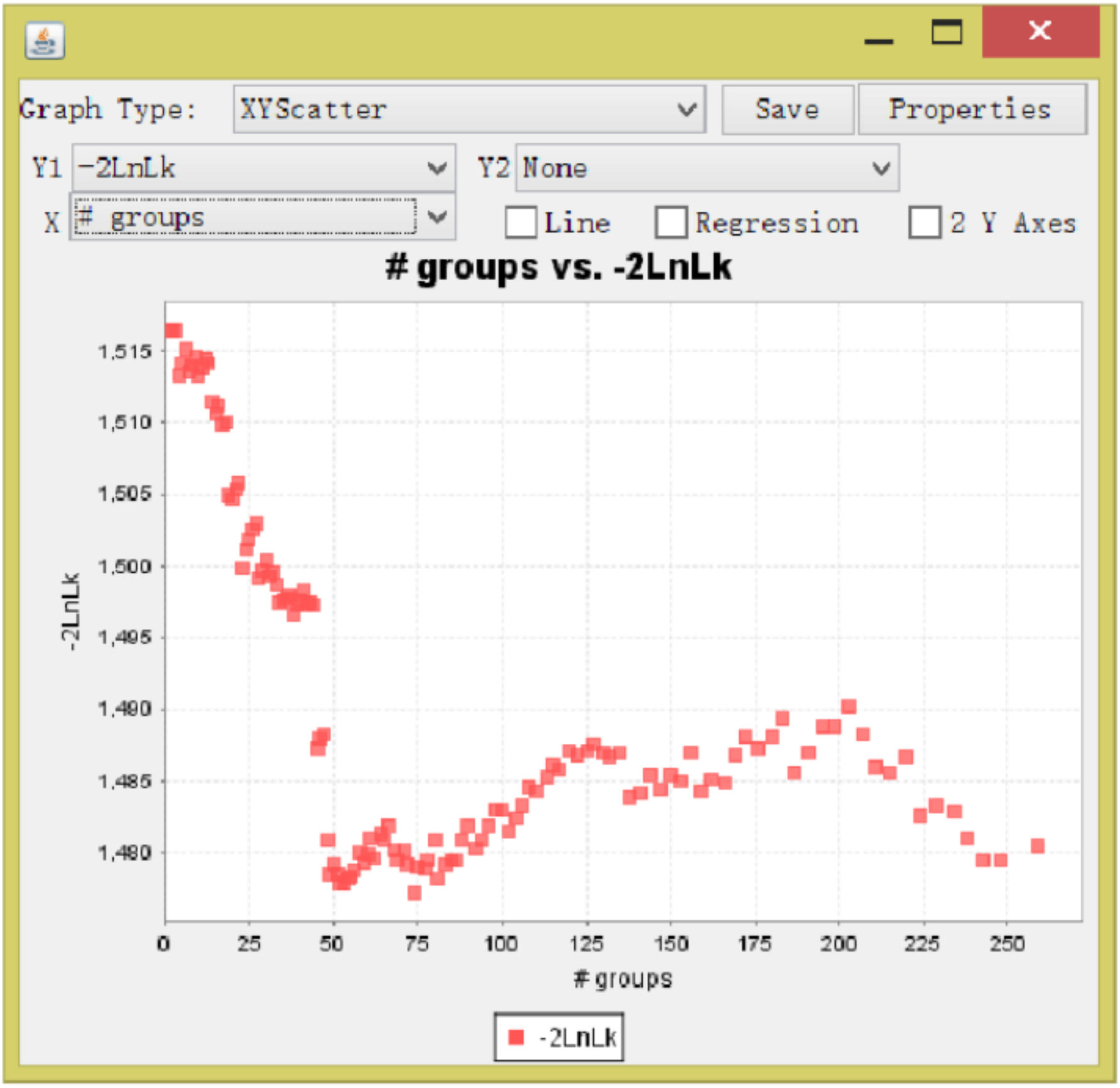


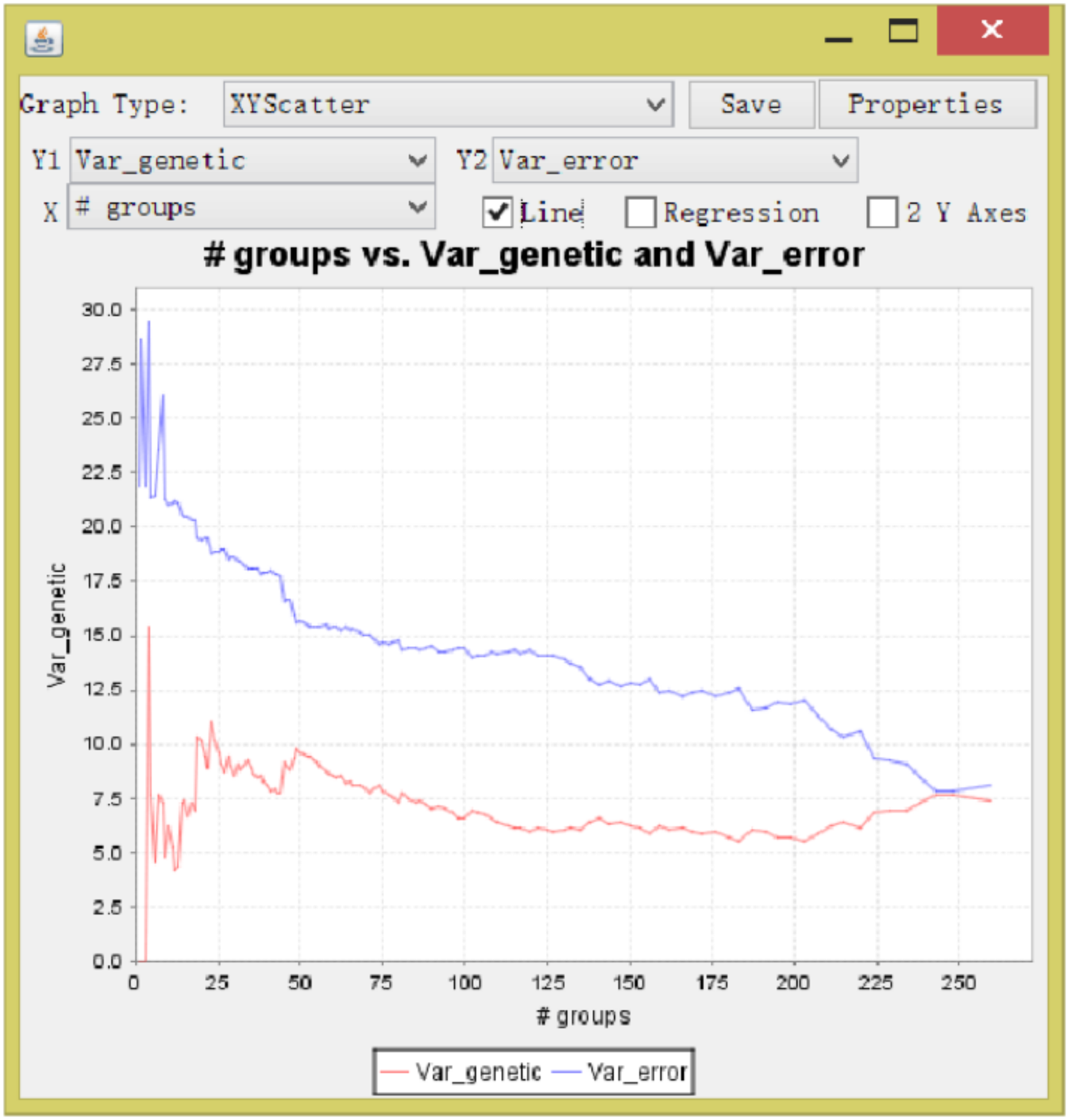


在该例子中，79 被包含在最后的分析中。当它们被聚类成 44 个组群时，-2 对数似然函数达到最小，这表明配合了最佳的模型。在这个最优的压缩水平上进行 SNPs 的筛选。

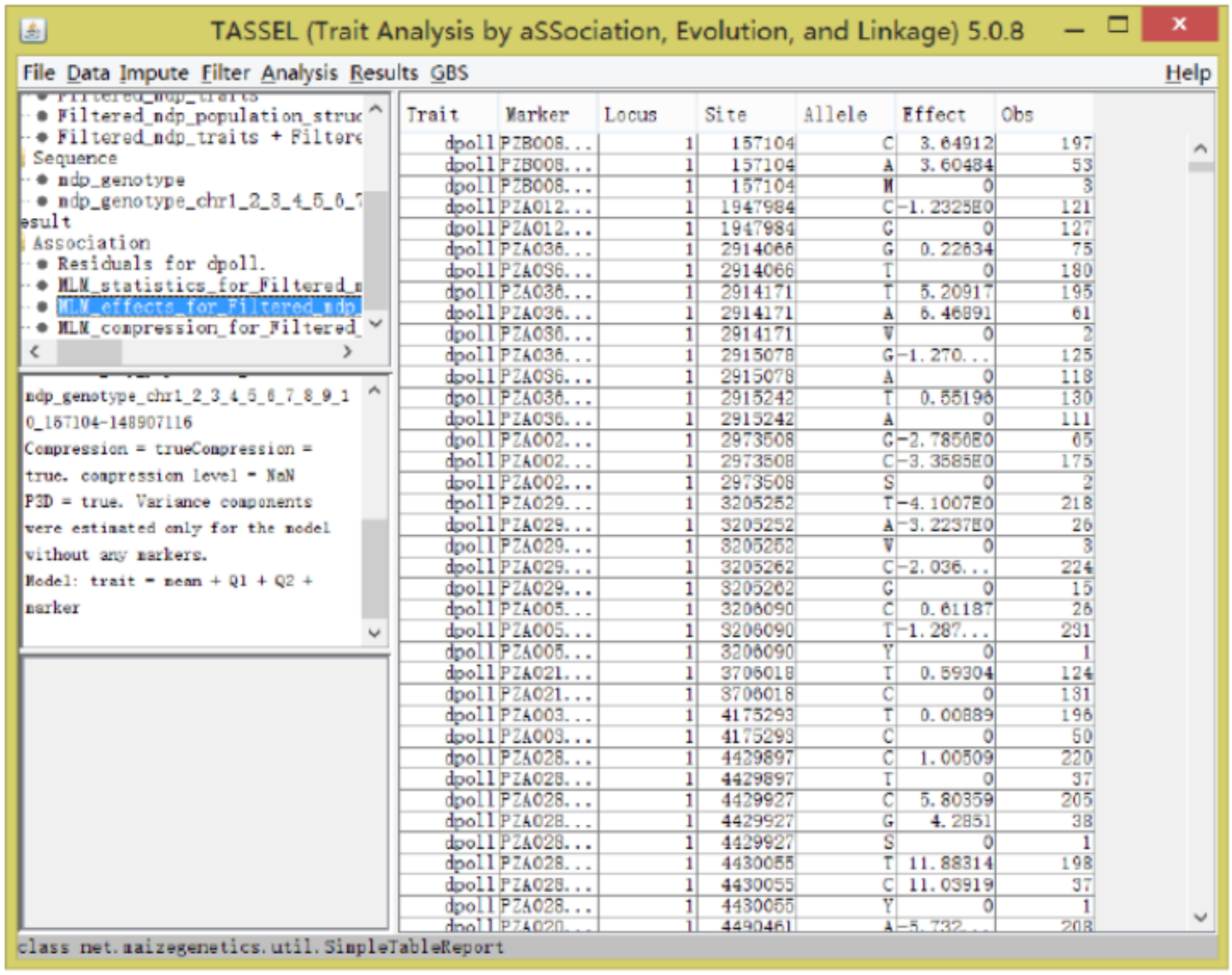
注意：当两个或更多个体被聚类到一个组群时，随机效应的方差分量不等于没有压缩时的这个方差分量。因而，得出的遗传率不应该被作为基于个体的遗传率解释。

为了在 3093 个 SNPs 上进行全基因组关联研究（GWAS），我们需要产生一个新的合并数据集，包含过滤的表现型、群体结构、以及全基因组基因型。加亮新的合并文件和亲缘关系数据，然后单击 MLM 按钮。在 MLM 选项对话框中选择默认选项。该分析将花费一两分钟。标签为“MLM_compression”的输出报告表明在该分析中使用了 259 个品系。具有 74 个组群，来自最佳模型的统计量被绘图如下。


Graph Type:
XYScatter
Save
Properties
Y1
-2LnLk
Y2
None
X
groups
Line
Regression
2 Y Axes
groups vs. -2LnLk
-2LnLk
1,515
1,510
1,505
1,500
1,495
1,490
1,485
1,480
0
25
50
75
100
125
150
175
200
225
250
groups
-2LnLk



关联最强的 SNP 在染色体 3 上 193565357 bp 处。P 值是 1.3027×10^{-4}。Bonferroni 多重检验矫正之后 1%显著性水平上的阈值是 3.2331×10^{-5}（0.01/3093）。该关联是不显著的。如同下面举例说明的，标签为“GLM_Allele_Estimates”的输出显示归因于每个 SNP 的基因型的标记效应（GLM 也是同样的）。例如，染色体 1 上 157104 bp 上的第一个 SNP 具有三个基因型（AA、CC 和 AC），根据 IUPAC 代码编码为 A、C 和 M，见附录（核苷酸代码）。



Trait	Marker	Locus	Site	Allele	Effect	Obs
dpoll	PZB008...	1	157104	C	3.64912	197
dpoll	PZB008...	1	157104	A	3.60484	53
dpoll	PZB008...	1	157104	M	0	3
dpoll	PZA012...	1	1947984	C	-1.2325E0	121
dpoll	PZA012...	1	1947984	G	0	127
dpoll	PZA036...	1	2914066	G	0.22634	75
dpoll	PZA036...	1	2914066	T	0	180
dpoll	PZA036...	1	2914171	T	5.20917	195
dpoll	PZA036...	1	2914171	A	6.46891	61
dpoll	PZA036...	1	2914171	V	0	2
dpoll	PZA036...	1	2915078	G	-1.270...	125
dpoll	PZA036...	1	2915078	A	0	118
dpoll	PZA036...	1	2915242	T	0.55196	130
dpoll	PZA036...	1	2915242	A	0	111
dpoll	PZA002...	1	2973508	G	-2.7850E0	65
dpoll	PZA002...	1	2973508	C	-3.3585E0	175
dpoll	PZA002...	1	2973508	S	0	2
dpoll	PZA029...	1	3205252	T	-4.1007E0	218
dpoll	PZA029...	1	3205252	A	-3.2237E0	26
dpoll	PZA029...	1	3205252	V	0	3
dpoll	PZA029...	1	3205262	C	-2.036...	224
dpoll	PZA029...	1	3205262	G	0	15
dpoll	PZA005...	1	3206090	C	0.61187	26
dpoll	PZA005...	1	3206090	T	-1.287...	231
dpoll	PZA005...	1	3206090	Y	0	1
dpoll	PZA021...	1	3706018	T	0.59304	124
dpoll	PZA021...	1	3706018	C	0	131
dpoll	PZA003...	1	4175293	T	0.00889	196
dpoll	PZA003...	1	4175293	C	0	50
dpoll	PZA028...	1	4429897	C	1.00509	220
dpoll	PZA028...	1	4429897	T	0	37
dpoll	PZA028...	1	4429927	C	5.80359	205
dpoll	PZA028...	1	4429927	G	4.2851	38
dpoll	PZA028...	1	4429927	S	0	1
dpoll	PZA028...	1	4430055	T	11.88314	198
dpoll	PZA028...	1	4430055	C	11.03919	37
dpoll	PZA028...	1	4430055	Y	0	1
dpoll	PZA020...	1	4490461	A	-5.732...	208

9 附录

9.1 核苷酸代码（来源于国际理论和应用化学联合会（IUPAC））

代码	含义
A	**A:A**
C	**C:C**
G	**G:G**
T	**T:T**
R	**A:G**
Y	**C:T**
S	**C:G**
W	**A:T**
K	**G:T**

M	A:C
+	+:+ (插入纯合子)
0	+:-
-	-:- (缺失纯合子)
N	未知的

9.2 TASSEL 教学数据集

http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip

文件名	类型	格式
d8_sequence.phy	基因型	Phylip Alignment
mdp_genotype.hmp.txt	基因型	Hapmap Alignment
mdp_genotype.plk.ped mdp_genotype.plk.map	基因型	Plink Alignment
mdp_kinship.txt	亲缘关系	数值方阵
mdp_population_structure.txt	群体结构	数值的性状数据
mdp_traits.txt	表现型	数值的性状数据

File #1 是 dwarf8 基因序列，具有 2466 个位点，在 91 个玉米自交系上。该数据由有关 Dwarf8 和开花期之间的关联论文描述[26]。

File #2-6 是 281 个玉米关联自交系的 3093 个 SNPs。数据以三种格式给出（Hapmap、Plink 和 Flapjack）。数据是由 PANZEA 科研项目产生的，由 NSF（美国国家科学基金会）资助。数据的详情可以在下面的网址上找到：http://www.panzea.org。

File #5 和 6 对于 Plink 的格式是成对的。

File #7 是由 Yu 等产生的亲缘关系[9]。

File #8 是 282 个玉米自交系的群体结构[27]。

File #9 是 282 个玉米自交系的三个性状的表现型，包括开花期[9]。

9.3 经常被问的问题

1. 如果 TASSEL 不正常运转我该怎么办？

TASSEL 是一个开源软件项目，寄放在 SourceForge 上，在 http://sf.net/projects/tassel 上有一个错误跟踪列表，你在那儿可以把问题报告给开发者团队。为了使一个错误被纠正，我们必须能够重复该问题。因此，重要的是要记录下产生错误时所采用的操作步骤。如果你分析的数据不是非常敏感的，请附上有错误的操作中使用过的文件。如果你不愿把你的数据文件贴在 SourceForge 上，你可以通过电子邮件将它发给一个该软件的开发者。

2. 我在什么地方可以获得更多的信息 ？

如果你对于 TASSEL 的某个方面有困难，你可以要么发电子邮件给 www.maizegenetics.net 上列出的一个该软件的开发者，要么在 SourceForge（http://sf.net/projects/tassel）上查看 TASSEL 论坛，因为别的用户可能已经考虑过一个相似的问题。在 http://groups.google.com/group/tassel 上还有一个 TASSEL 讨论群。

3. 我如何加入 SourceForge 上的 TASSEL?

TASSEL 是一个开源项目，在 GNU 一般公用许可证下发行（GNU general public license）。这意味着源代码是可用的，用户可以自由地修改代码来适合他们特定的需要。我们欢迎来自开发者的投入，也欢迎那些想要参与这个软件的改进的人。该项目是寄放在 SourceForge（http://sf.net/projects/tassel）上的，因此允许任何人访问代码的最新的变化。这个设置使得任何人便于添加特殊的功能到 TASSEL，如果它们愿意这样的话。对于想要参与一个生物信息学软件开发项目的任何人来说，它也充当一个好的平台。

4. 当我单击 TASSEL 网络启动的当前版本时，出现一个以前的版本。我该如何?

TASSEL 网络启动的以前的版本被缓存在你的计算机中。要用当前版本取代它，单击 Windows 中的 Start 按钮，继之以 Run。键入 javaws 然后单击 OK。在打开的窗口中，保留 TASSEL 的当前版本，删除其余者。

5. 在 TASSEL 中我应该用什么来代替缺失值?

对于版本 3 格式的数值数据，用 NA 或者 NaN。对于版本 2 格式的数值数据，用“-999”代表缺失值。对于 SNP 数据，使用“N”。对于 SSR 数据，则用“?”。亲缘关系不允许有缺失值。

6. 可以改变数据树中的数据名称吗？

可以。在数据树上单击想要的数据名称，等待一秒，然后重新单击它或者立即按 F2 键。重新命名数据集然后按回车键来保存该改变。

7. 我怎样才能在桌面上创建一个 TASSEL 图标呢？

在微软 Windows 上单击“Start”，选择“控制面板”，然后双击 Java 来显示“Java 控制面板”。在“Temporary Internet Files”部分，单击“View”按钮显示“Java Cache Viewer”。移动鼠标到 TASSEL 应用上，单击右键，选择“Install Shortcuts”。

8. 为什么我在 MLM 关联分析中得到空的正方形？（square）？

空的正方形意味着无信息（null information）。主要原因包括方差分量估计中的不收敛性或者没有计算成为问题的统计量。例如，标记 F、p 和 R2 不被计算，当没有标记被包括在模型中时。

9. 为什么我应该除去群体结构的一个列？

对于计算群体结构的一些方法，比如软件 STRUCTURE，群体比例总和为 1。这产生群体协变量之间的线性相依性（linear dependence）。虽然 GLM 使用的算法容忍那个相关性，但是 MLM 将会失败，因为设计矩阵将不是可逆的。除去一个列就消除了列之间的线性相依。利用 PC 轴来代表群体结构不导致线性相依，因为所有的 PC 列被确保是独立的。

10. 亲缘关系可以取代群体结构吗？

有时可以。对于一些性状和群体，只有 K 的模型可能与 Q+K 模型同样好，或者比 Q+K 模型更好。对于其它的来说，Q+K 可能要好些。只有 Q 的模型对于控制群体结构没有另一种那样有效。可惜，没有通用准则用于预测哪个模型将表现最好。因此，一个研究者可能希望配合所有三个模型然后比较结果。如果消除假阳性是非常重要的，那么可以接受最保守的模型。然而，如果目的是识别候选者用于更进一步的研究，并且按一个错误的引导继续研究的代价比较低，则最不严格的模型可能是更喜欢的。

11. 为什么 TASSEL 和 SPAGeDi 给出不同的亲缘关系估计值？

首先，计算亲缘关系的算法有很多，它们的估计值会彼此不同。其次，TASSEL 中的算

法把每个基因型作为一个单倍型处理。不推荐用 TASSEL 来从杂合的基因型产生一个亲缘关系矩阵。在不久的将来，TASSEL 亲缘关系算法将被修改来处理杂合的二倍体。

12. 我可以利用 SAS Proc Mixed 或者 TASSEL MLM 得到标记 R 平方吗？

SAS Proc Mixed 不产生一个 R2 统计量。TASSEL 中的 MLM 产生 R2 统计量。用户手册描述了它是如何计算的。

13. MLM 比 GLM 发现更多的关联吗？

有时可以。MLM 比 GLM 具有更高的统计功效，可以检测更多真实的关联。当检验的遗传标记与亲缘关系结构混杂时，GLM 不能象 MLM 一样有效地校正那个亲缘关系结构，可能产生更多的假阳性。

14. 对于来自 Tassel 的 p 值，我需要多重检验矫正吗？

是的。

15. TASSEL 可以处理二倍体基因型数据吗？

虽然 TASSEL 接受大多数常见的序列比对格式，它处理多倍体基因型数据，包括单倍体和二倍体在内，但是一些分析对于杂合的数据是不适当的。GLM 或者 MLM 配合 SNPs，一次一个，把每个不同的基因型作为一个单独的类别处理。这具有配合一个加性效应加上显性模型的作用。把两个效应分开正在考虑中。因为把杂合体作为第三个标记类别处理对于亲缘关系或者 LD 是不适当的，所以目前对于那个类型的数据不应该使用那些分析。改进软件以便处理杂合体的工作正在进行。

16. 如何引用 TASSEL？

作为一个软件包对 TASSEL 进行描述的论文[1]，和介绍 TASSEL 中实现的特定方法的论文的引用应该视情况而定，比如统一的（“Q+K”）方法、EMMA、混合线性模型的压缩和 P3D。例如：

A. Linkage disequilibrium (D', R^2 and P value) were calculated by TASSEL[1].

B. Association analyses were performed with the mixed linear model approach[9] implemented by TASSEL[1].

C. GWAS was performed with the compressed mixed linear model approach[4,9] carried by TASSEL[1] which also implemented the EMMA[3] and P3D[4] algorithms to reduce computing time.

参考文献

1. Bradbury, P.J. et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633-2635 (2007).
2. Zhang, Z., Buckler, E.S., Casstevens, T.M. & Bradbury, P.J. Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 10, 664-75 (2009).
3. Kang, H.M. et al. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics 178, 1709-1723 (2008).
4. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42, 355-60 (2010).
5. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42, 348-54 (2010).
6. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28, 286-289 (2001).
7. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. American Journal of Human Genetics 67, 170-181 (2000).
8. Zhao, K. et al. An Arabidopsis example of association mapping in structured samples. PLoS Genet 3, e4 (2007).
9. Yu, J.M. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38, 203-208 (2006).
11. Ware, D. et al. Gramene: a resource for comparative grass genomics. Nucleic Acids Research 30, 103-105 (2002).
12. Ware, D.H. et al. Gramene, a tool for grass Genomics. Plant Physiology 130, 1606-1613 (2002).
13. Jaiswal, P. et al. Gramene: development and integration of trait and gene ontologies for rice. Comparative and Functional Genomics 3, 132-136 (2002).
14. Yamazaki, Y. & Jaiswal, P. Biological ontologies in rice databases. An introduction to the activities in gramene and oryzabase. Plant and Cell Physiology 46, 63-68 (2005).
15. Zhao, W. et al. Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Research 34, D752-D757 (2006).
16. Canaran, P., Stein, L. & Ware, D. Look-Align: an interactive web-based multiple sequence alignment viewer with polymorphism analysis support. Bioinformatics 22, 885-886 (2006).
17. Du, C.G., Buckler, E. & Muse, S. Development of a maize molecular evolutionary genomic database. Comparative and Functional Genomics 4, 246-249 (2003).
18. SAS, I.I. SAS. Statistical Analysis Software for Windows, 9.0 ed. Cary, NC. USA. (2002.).
19. Hardy, O.J. & Vekemans, X. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2, 618-620 (2002).
20. Cover, T. & Hart, P. Nearest neighbor pattern classification. Proc IEEE Trans Inform Theory 13(1967).
21. Weir. Genetic Data Analysis II. Sunderland, MA. (1996).
22. Farnir, F. et al. Extensive genome-wide linkage disequilibrium in cattle. Genome Res 10, 220-

7 (2000).
23. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31, 423-447 (1975).
24. Kang, H.M. et al. Efficient control of population structure in model organism association mapping. Genetics 178, 1709-23 (2008).
25. Laird, N.M. & Ware, J.H. Random-Effects Models for Longitudinal Data. Biometrics 38, 963-974 (1982).
26. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28, 286-9 (2001).
27. Flint-Garcia, S.A. et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44, 1054-64 (2005).
28. Anderson, M.J. & Ter Braak, C.J.F. Permutations tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73, 85-113 (2003)