2.5 宏基因组测序与分析

杨　　龙
lyang@sdau.edu.cn

宏基因组(16s)测序

我们能利用宏基因组做什么？

- 微生物是物种中最主要的构成
- 世界上有生命的地方都存在
- 影响：疾病、土壤性质、作物品质、食物卫生

了解微生物尤其是细菌的构成能带来：

- 一个我们不了解的世界都有谁
- 它们的变化和我们关心的事有关吗
- 它们能帮助我们预测或改造一些东西吗

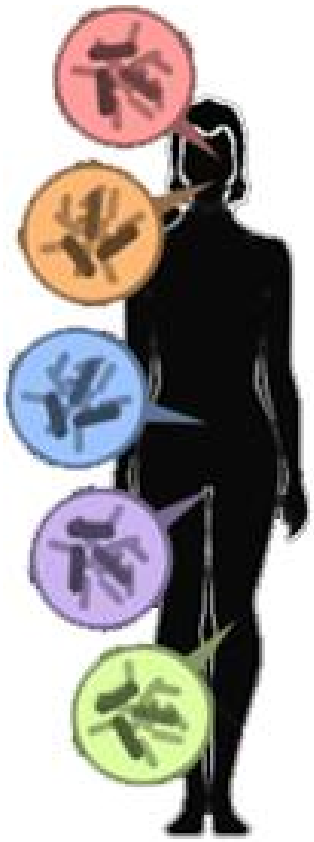
Slide by Dirk Gevers

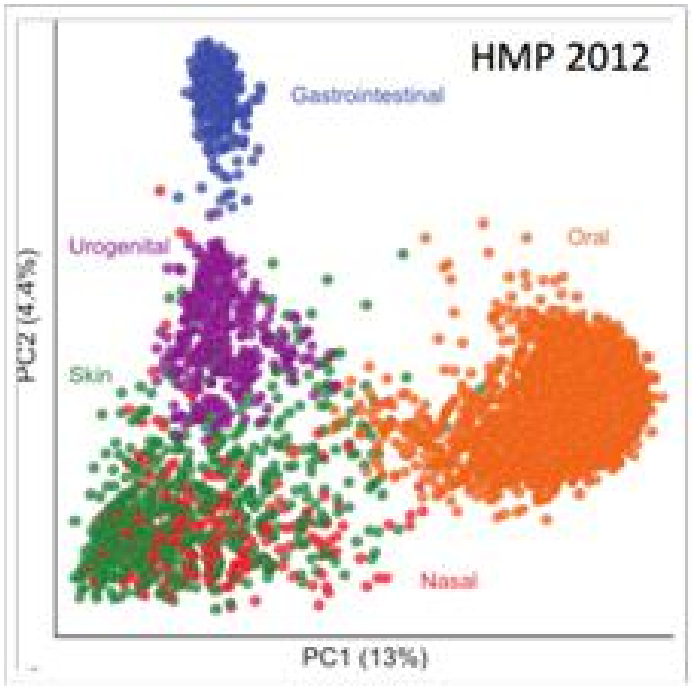
The NIH **H**uman **M**icrobiome **P**roject (**HMP**): A comprehensive microbial survey

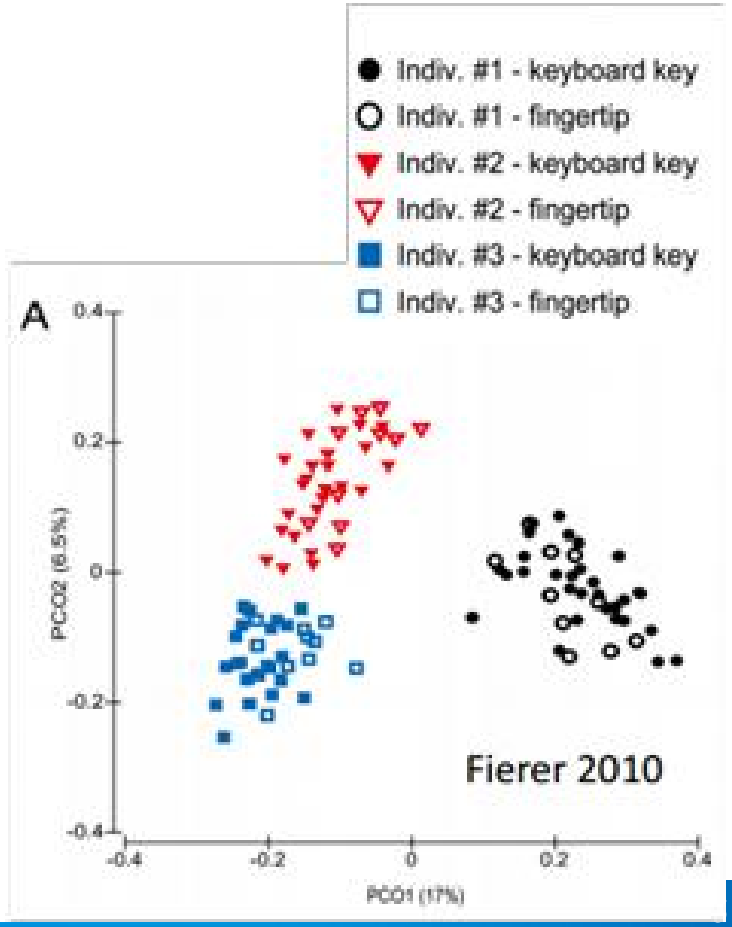
- ***What is a "normal" human microbiome?***
- 300 healthy human subjects
- Multiple body sites
  - 15 male, 18 female
- Multiple visits
- Clinical metadata

www.hmpdacc.org

微生物菌群构成，来自人体的例子







菌群多样性分析是怎么一个流程

提取DNA->扩增建库->测序（1-2周）

比对聚类->构建OUT->分析菌群构成（2天）

寻找差异和有关的标志（N天）

16S rRNA及其测序区域

区分目的，我们想要什么

- 想知道都有什么菌
- 了解菌群构成比例
- 不同分组之间存在什么差异
- 代谢途径和功能基因是否存在特征

菌群多样性分析

- 想寻找新物种和新基因
- 希望测序所有的物种序列

宏基因组测序

菌群构成有什么用

- 帮助我们知道都有什么菌
- 如果有时间或不同条件可以了解变化
- 特定菌或菌之间的比例决定某些表型
- 特定的基因或功能类决定作用
- 寻找标志物（某些菌或某些基因）
- 构建预测或干预模型，了解机理

测16s能知道基因和功能变化？

- 目前已经完成基因组测序的菌：23600多
- 借助已经测序物种推测每类菌分支的基因构成
- 标准化16s和菌的数目（1-15个16s rRNA）
- 预测样品中的每类基因和代谢途径的数目

- 准确度在84%-95%，对肠道微生物菌群和土壤菌群的功能分析接近95%

16s测序预测结果与RNA-seq比较

Supplementary Table 16. Figure 3 enzymes with predicted differences in gene abundance.

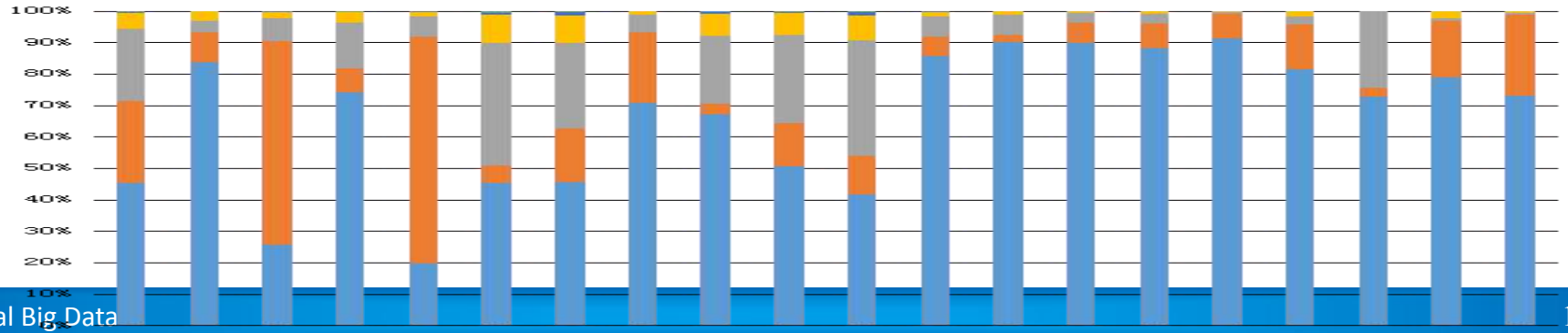
KO ID	Plant-based diet	Animal-based diet	Fold-change	p-value	Abbreviation	Figure	Consistent w/ RNA-Seq?	Full name
K00262	3361	2956	0.88	0.027	GDH	3d	Yes	Glutamate dehydrogenase
K01915	5804	5219	0.90	0.037	Gln synthetase	3d	Yes	Glutamine synthetase
K01571	2531	1952	0.77	0.022	ODx	3e	Yes	Oxaloacetate decarboxylase, alpha subunit
K01572	3624	3054	0.84	0.049	ODx	3e	Yes	Oxaloacetate decarboxylase, beta subunit
K01610	2343	2117	0.90	0.045	PEP Ck	3e	Yes	Phosphoenolpyruvate carboxykinase
K08484	1	53	53.00	0.007	PTS	3e	Yes	Phosphotransferase system, enzyme I
K08681	773	465	0.60	0.027	Gln amidotransferase	3f	No	Glutamine amidotransferase
K00599	7276	6310	0.87	0.030	Methyltransferases	3g	No	Methyltransferases

Supplementary Table 16. Figure 3 enzymes with predicted significant differences in gene abundance. Shown are genes from Figure 3 whose gene abundance was predicted to significantly differ between the animal- and plant-based diet. Gene abundances were estimated using the PICRUSt software pipeline (http://picrust.github.com.). Shown are the median predicted gene counts, fold-change, and estimated significance (Mann-Whitney U test). Genes whose fold-change direction matches observed transcript-level changes are highlighted in yellow.

David L A, Maurice C F, Carmody R N, et al. Diet rapidly and reproducibly alters the human gut microbiome[J]. Nature, 2014, 505(7484): 559-563.

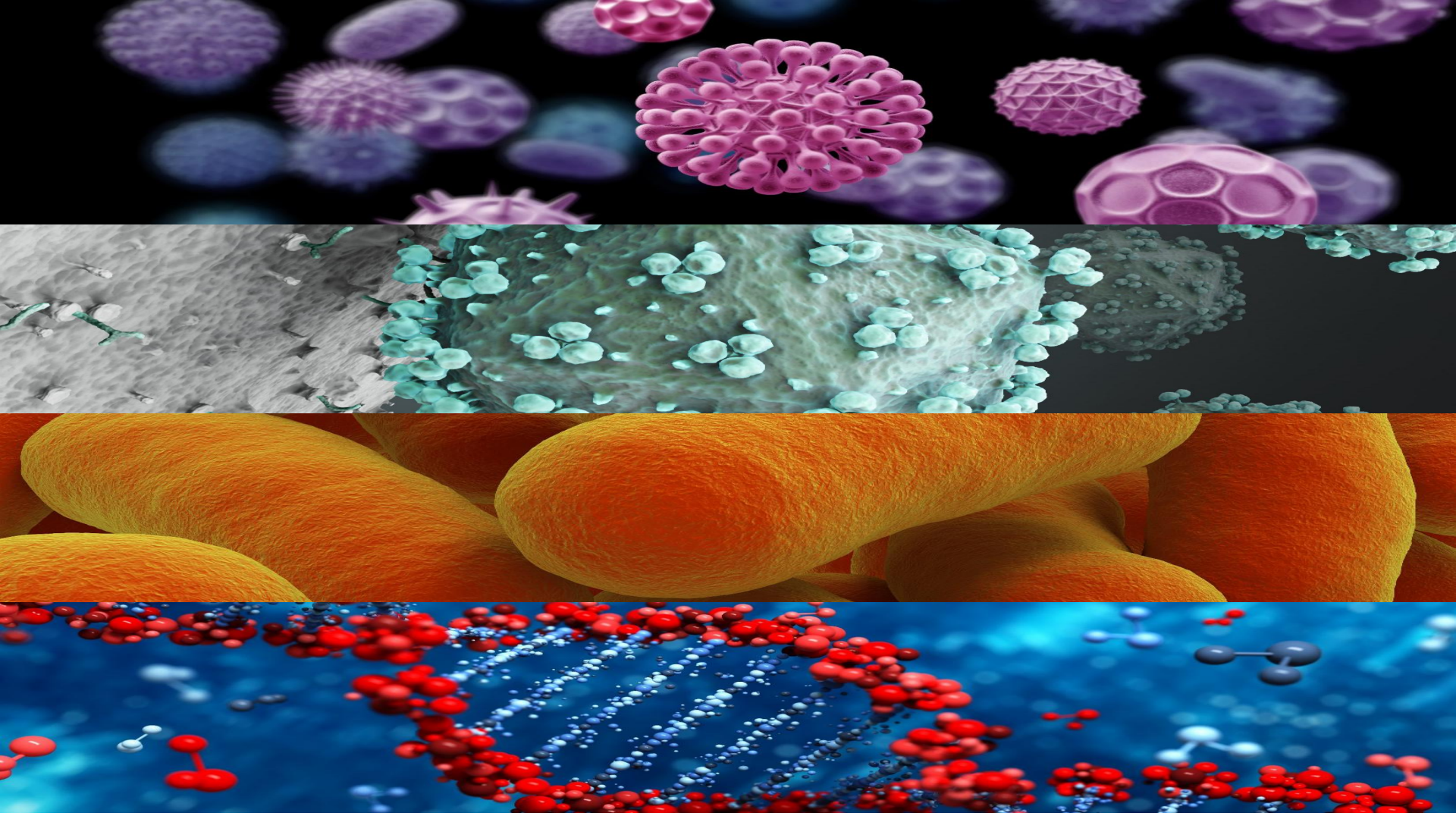
菌群多样性和宏基因组的关键

- DNA提取
- PCR扩增
- 测序深度
- 取样数目
- 对照分组的选择
- 数据的挖掘



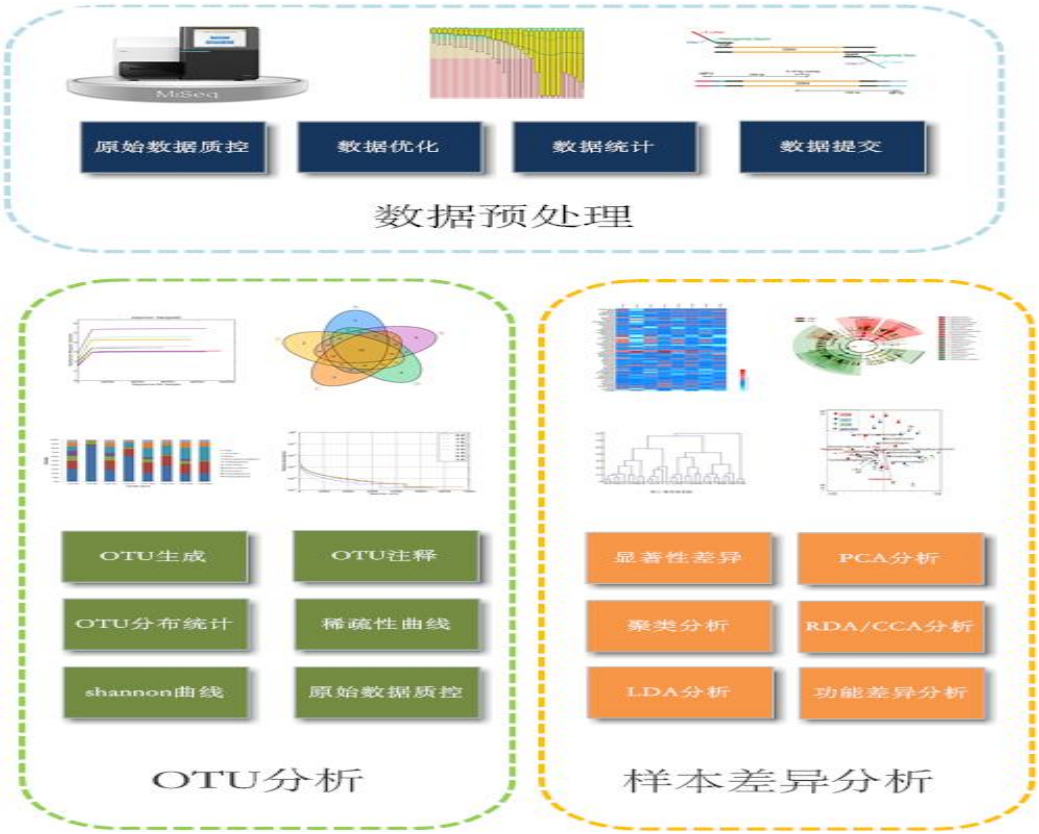
建议的策略

- 大样本的16s rRNA测序筛查
- 筛选菌群和基因
- 使用特异性引物定量PCR初步验证
- 选择代表性样本进行宏基因组或宏转录组测序



16s测序报告长什么样？

分析流程



基本概念

16S rRNA及其测序区域

16S rRNA

16S rRNA 基因是编码原核生物核糖体小亚基的基因，长度约为1542bp，其分子大小适中，突变率小，是细菌系统分类学研究中最常用和最有用的标志。16S rRNA基因序列包括9个可变区和10个保守区，保守区序列反映了物种间的亲缘关系，而可变区序列则能体现物种间的差异。16S rRNA基因测序以细菌16S rRNA基因测序为主,核心是研究样品中的物种分类、物种丰度以及系统进化。

OTU

operational taxonomic units (OTUs)在微生物的免培养分析中经常用到，通过提取样品的总基因组DNA，利用16S rRNA或ITS的通用引物进行PCR扩增，通过测序以后就可以分析样品中的微生物多样性，那怎么区分这些不同的序列呢，这个时候就需要引入operational taxonomic units，一般情况下，如果序列之间，比如不同的 16S rRNA序列的相似性高于97%就可以把它定义为一个OTU，每个OTU对应于一个不同的16S rRNA序列，也就是每个OTU对应于一个不同的细菌（微生物）种。通过OTU分析，就可以知道样品中的微生物多样性和不同微生物的丰度。

我们对v3-v4双可变区域进行扩增和测序。首先对每个样板库的序列进行OTU注释，使用QIIME（version 1.8.0）工具包进行注释，数据库使用GreenGene（version gg_13_8）。
项目的研究目的之一就是比较不同样本之间的微生物组成差异，OTU比对完成后进行注释，由于16S序列的高度相似性，根据V3-V4区的测序长度，只能将序列准确的归类到属（genus）这个级别，注释结果中虽然有部分的种（species）结果，但是并不完全。

OTU结果统计

OTU数目和测序统计表

SampleName	SampleSize	OTUsNumber	OTUsSeq	Coverage
T12h	91490	2094	88346	1
T36h	79556	2131	75006	0.99
T60h	89470	2481	86094	1
T108h	72822	1170	64286	0.99
T84h	91438	1744	85785	0.99
25212h	100956	2090	95463	1
25284h	84117	1178	77962	0.99
25260h	94979	1181	89000	0.99
252108h	54265	1066	47533	0.99

SampleName：样本名称
SampleSize：样本序列总数
OTUsNumber：注释上的OTU数目
OTUsSeq：注释上OTU的样本序列总数

Coverage：是指各样品文库的覆盖率，其数值越高，则样本中序列没有被测出的概率越低。该指数实际反映了本次测序结果是否代表样本的真实情况。
计算公式为：C=1-n1/N　其中n1 = 只含有一条序列的OTU的数目；　N = 抽样中出现的总的序列数目。

表解读：该表主要是对每个样本的测序数量和OTU数目进行统计，并且在表格中列出了测序覆盖的完整度（显示前10个样本）

OTU分类水平统计表

Sample	Phylum	Class	Order	Family	Genus	Species
25212h	2208	2202	2135	1856	963	98
25260h	1059	1059	1054	1037	478	37
25284h	995	995	994	975	460	72
252108h	638	636	624	572	299	88
T12h	1441	1435	1384	1203	812	194
T36h	1876	1870	1802	1472	711	110
T60h	1827	1820	1782	1501	563	141
T84h	1427	1421	1393	1274	650	78
T108h	1323	1323	1317	1266	612	35

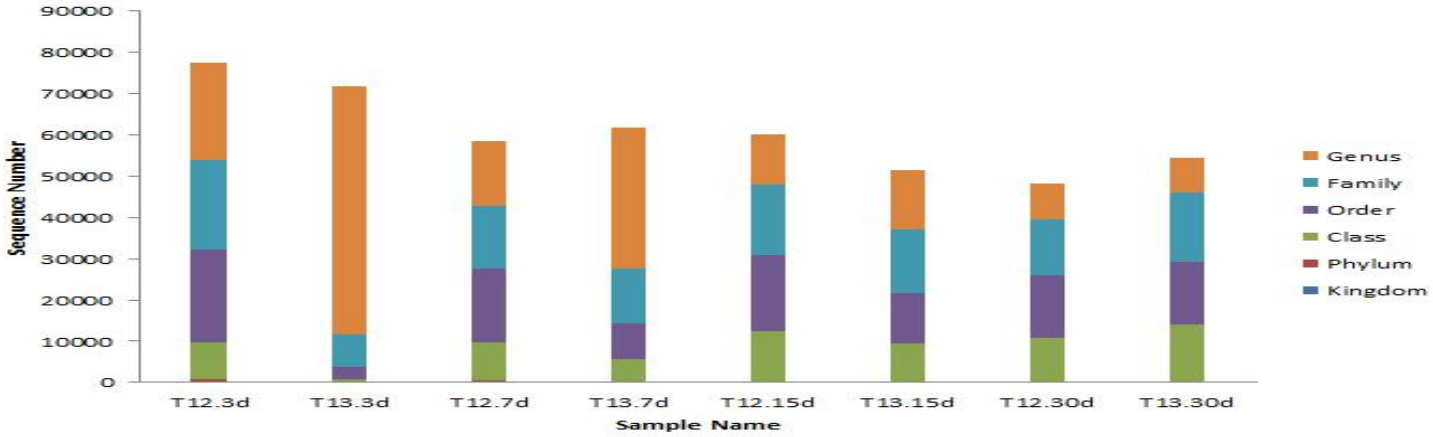
SampleName：样本名称
Phylum：分类到门的OTU数量
Class：分类到纲的OTU数量
Order：分类到目的OTU数量

Family：分类到科的OTU数量
Genus：分类到属的OTU数量
Species：分类到种的OTU数量

表解读：该表主要是对每个样本在分类学水平上的数量进行统计，并且在表格中列出了在每个分类学水平上的物种数目（显示前10个样本）

OTU构成

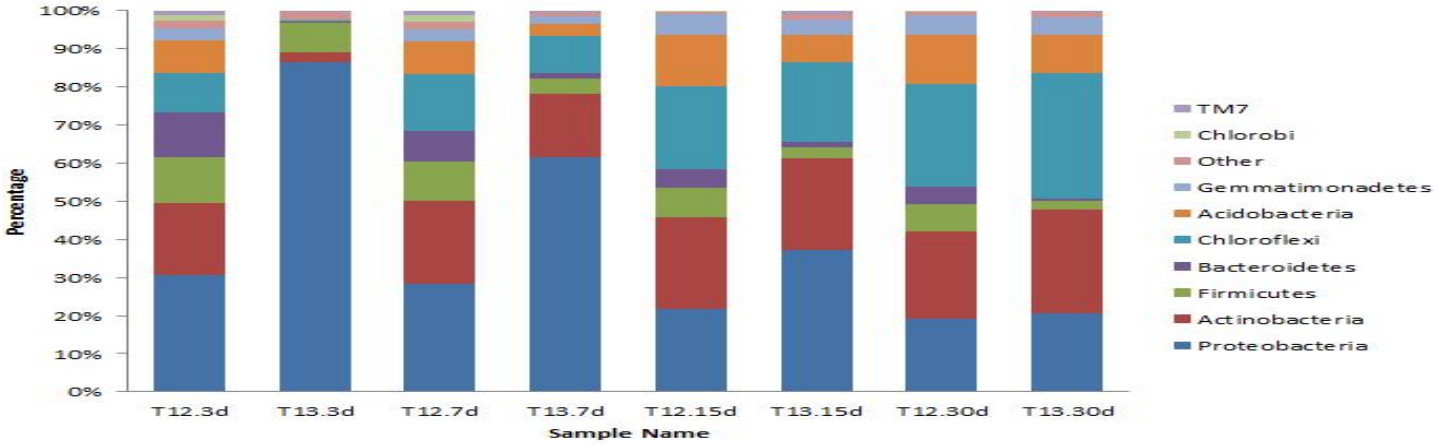
样本种属比例



图中的每根柱子中的颜色表示该样本在不同级别（门、纲、目等）的序列数目，序列数目只计算级别最低的分类，例如在属中计算过了，则在科中则不重复计算。

图解读：横坐标中每一个条形图代表一个样本，纵坐标代表该分类层级的序列数目。同一种颜色代表相同的分类级别。

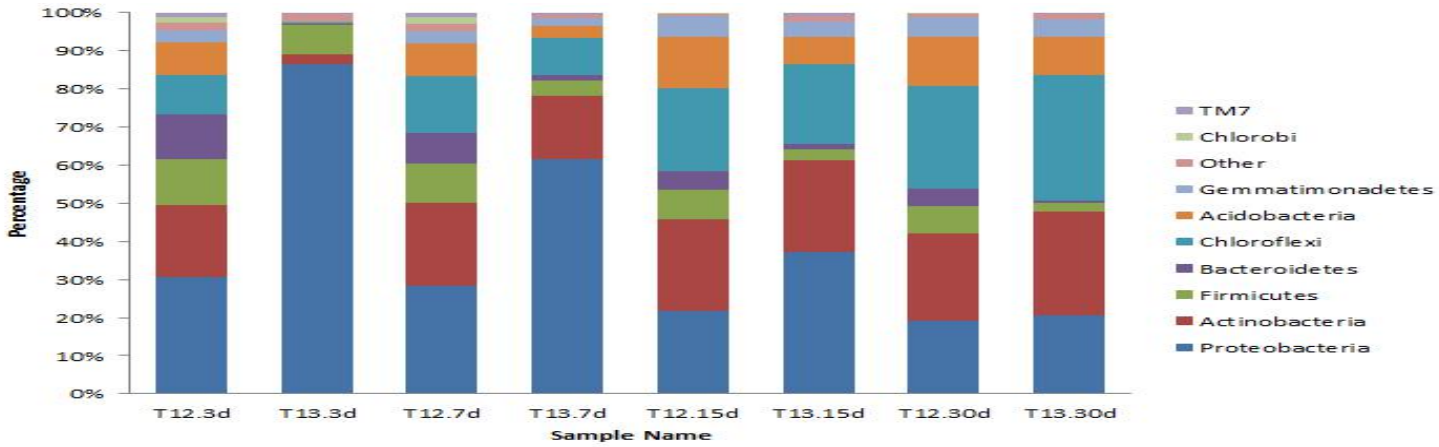
比例最高的前10个属构成柱状图



图解读：横坐标中每一个条形图代表一个样本，纵坐标代表该分类单元的所占比例。同一种颜色代表相同的分类单元。

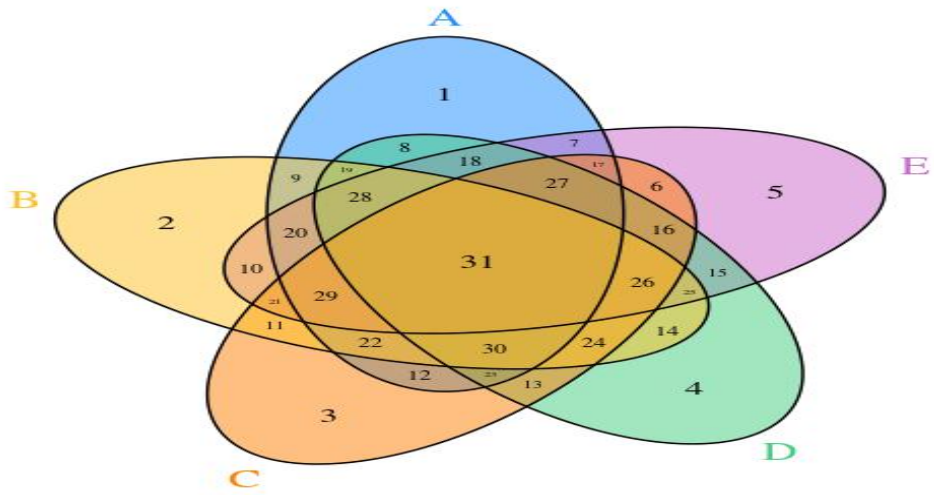
OTU构成及比较

比例最高的前10个属构成气泡图



图解读：横坐标中每一个列代表一个样本，纵坐标将比例最高的10个属从上到下以气泡形式排列，气泡大小代表该分类单元所占的比例。同一种颜色代表相同的分类单元。

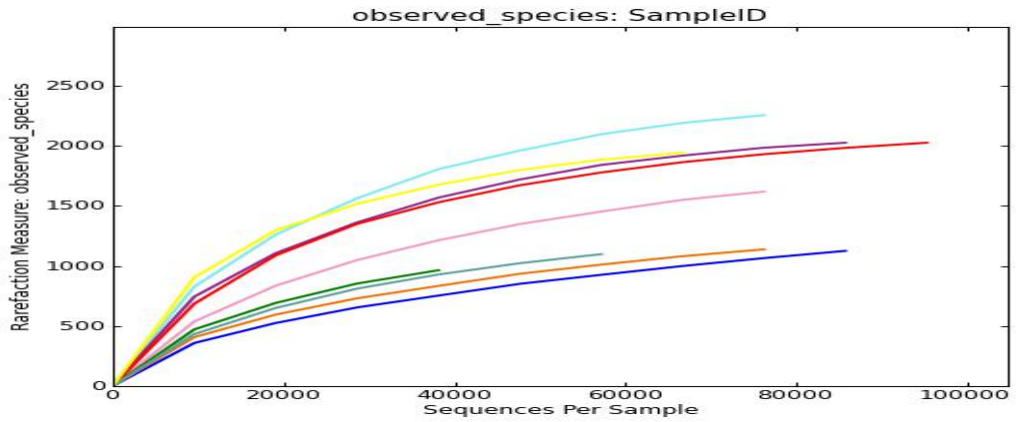
OTU比较韦恩图（ <=5个样本 ）



图解读：每个圈代表一个样本，圈之间的重叠区域表示样本间共有的OTUs，每个区域的数字大小表示该区域对应的OTUs数目。

样本构成丰度

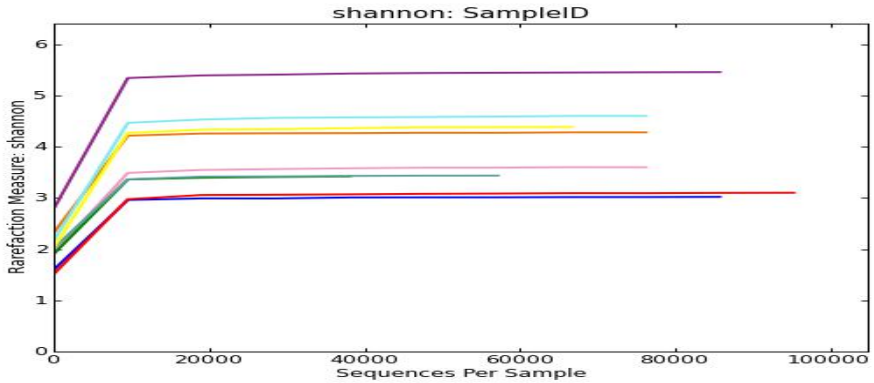
稀释曲线



稀释曲线
微生物多样性分析中需要验证测序数据量是否足以反映样品中的物种多样性，稀释曲线（丰富度曲线）可以用来检验这一指标。稀释曲线是用来评价测序量是否足以覆盖所有类群，并间接反映样品中物种的丰富程度。稀释曲线是利用已测得16S rDNA序列中已知的各种OTU的相对比例，来计算抽取n个（n小于测得reads序列总数）reads时出现OTU数量的期望值，然后根据一组n值（一般为一组小于总序列数的等差数列）与其相对应的OTU数量的期望值做出曲线来。当曲线趋于平缓或者达到平台期时也就可以认为测序深度已经基本覆盖到样品中所有的物种；反之，则表示样品中物种多样性较高，还存在较多未被测序检测到的物种。
在我们的结果报告目录中提供四种不同方法的丰富度曲线，其中包括最常用的obseved species指标来绘制稀疏曲线和shannon指数绘制的shannon-winner曲线，详细结果在结果目录中alpha_rarefaction中。

图解读：横坐标代表随机抽取的序列数量；纵坐标代表观测到的OTU数量。样本曲线的延伸终点的横坐标位置为该样本的测序数量，如果曲线趋于平坦表明测序已趋于饱和，增加测序数据无法再找到更多的OTU；反之表明不饱和，增加数据量可以发现更多OTU。

Shannon-Winner曲线

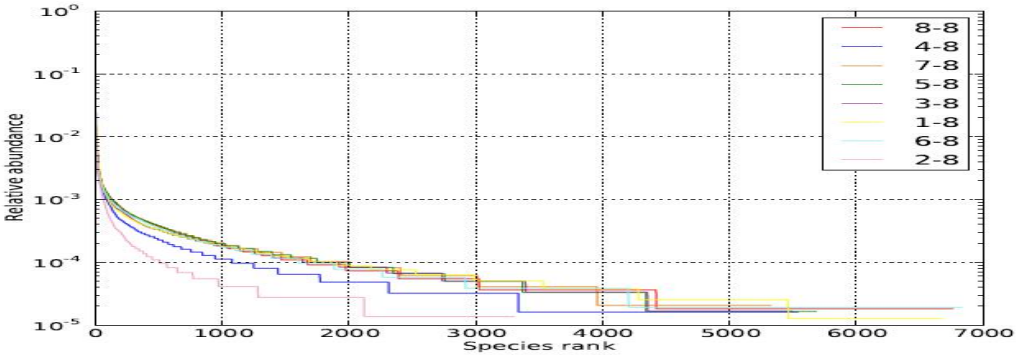


Shannon-Wiener 曲线
Shannon-Wiener 曲线，是利用shannon指数来进行绘制的，反映样品中微生物多样性的指数，利用各样品的测序量在不同测序深度时的微生物多样性指数构建曲线，以此反映各样本在不同测序数量时的微生物多样性。 当曲线趋向平坦时，说明测序数据量足够大，可以反映样品中绝大多数的微生物物种信息。

图解读：与上图一样，横坐标代表随机抽取的序列数量；纵坐标代表的是反映物种多样性的Shannon指数。样本曲线的延伸终点的横坐标位置为该样本的测序数量，如果曲线趋于平坦表明测序已趋于饱和，增加测序数据无法再找到更多的OTU；反之表明不饱和，增加数据量可以发现更多OTU。其中曲线的最高点也就是该样本的Shannon指数，指数越高表明样品的物种多样性越高。

样本多样性

Rank-Abundance曲线
丰度分布曲线



Rank-Abundance 曲线
用于同时解释样品多样性的两个方面，即样品所含物种的丰富程度和均匀程度。 物种的丰富程度由曲线在横轴上的长度来反映，曲线越宽，表示物种的组成越丰富； 物种组成的均匀程度由曲线的形状来反映，曲线越平坦，表示物种组成的均匀程度越高。一般超过20个样本图就会变得非常复杂而且不美观，所以一般20个样本以下会做该图，**图片保存为结果目录中rank.pdf。**

图解读：横坐标代表物种排序的数量；纵坐标代表观测到的相对丰度。样本曲线的延伸终点的横坐标位置为该样本的物种数量，如果曲线越平滑下降表明样本的物种多样性越高，而曲线快速陡然下降表明样本中的优势菌群所占比例很高，多样性较低。

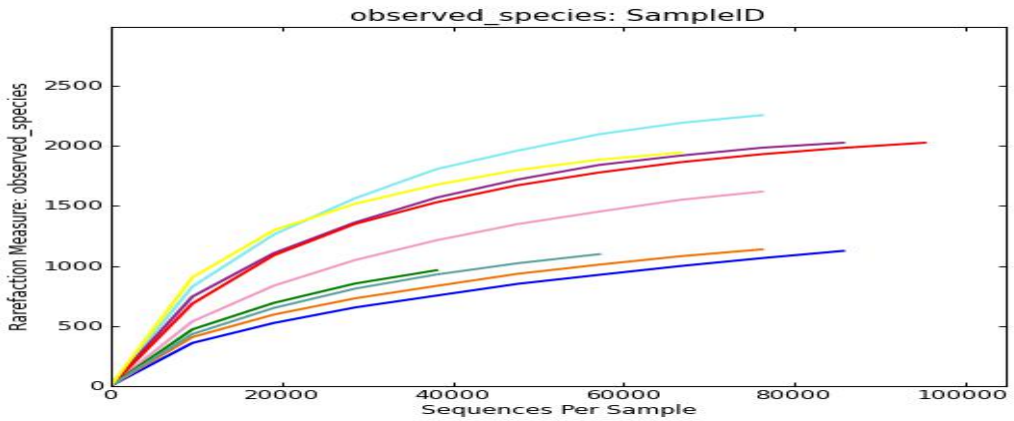
Alpha多样性（样本内多样性）

Sample	chaos	shannon	ACE	simpson
1-8	7297.36770921	10.5356331043	7400.34282551	0.99513826132
2-8	4130.08353222	4.68427385873	4541.90839788	0.655272782023
3-8	6164.50831601	10.1907194457	6220.91797855	0.99362658987
4-8	7903.66666667	7.52911010602	8228.52451228	0.861331961682

表解读：Alpha多样性是指一个特定区域或者生态系统内的多样性，常用的度量指标有Chao、ACE、Shannon、Simpson等，Simpson指数值越大，说明群落多样性越高；Shannon指数越大，说明群落多样性越高。文件保存在结果目录中的alpha_div.txt。

Beta多样性分析（样品间差异分析）

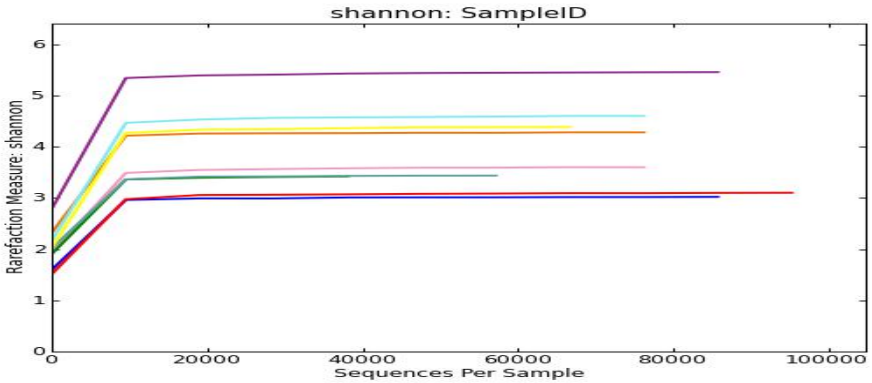
PCoA分析



PCoA（principal co-ordinates analysis）是一种研究数据相似性或差异性的可视化方法，通过一系列的特征值和特征向量进行排序后，选择主要排在前几位的特征值，PCoA 可以找到 距离矩阵中最主要的坐标，结果是数据矩阵的一个旋转，它没有改变样品点之间的相互位置关系，只是改变了坐标系统。通过PCoA可以观察个体或群体间的差异。

图解读：图中每一个点代表一个样本，相同颜色的点来自同一个分组，两点之间距离越近表明两者的群落构成差异越小。

NMDS分析（非度量多维尺度分析）

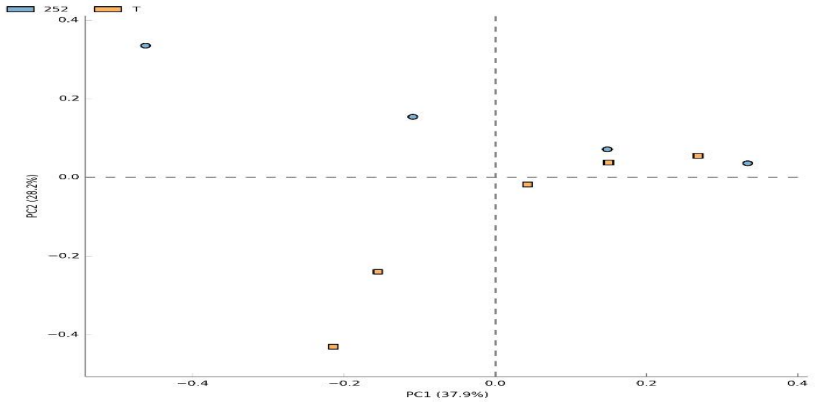


NMDS（Nonmetric Multidimensional Scaling）常用于比对样本组之间的差异，可以基于进化关系或数量距离矩阵。
横轴和纵轴：表示基于进化或者数量距离矩阵的数值在二维表中成图。
与PCA分析的主要差异在于考量了进化上的信息。

图解读：图中每一个点代表一个样本，相同颜色的点来自同一个分组，两点之间距离越近表明两者的群落构成差异越小。

Beta多样性分析及差异分析

PCA分析

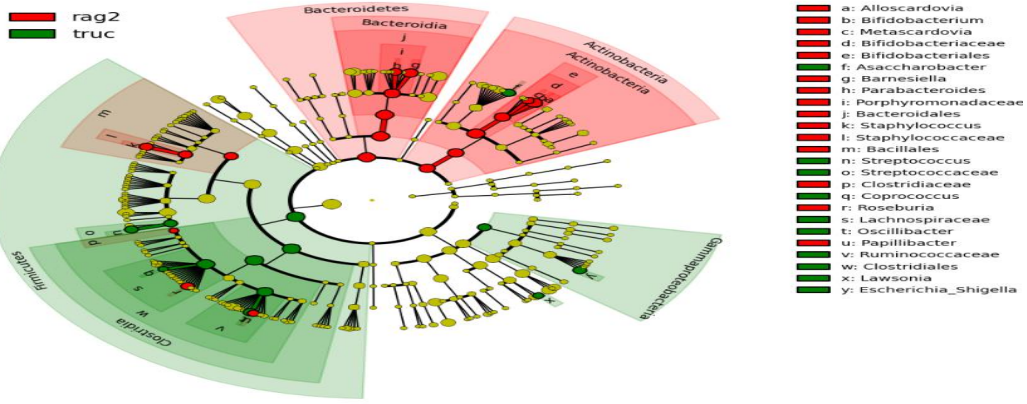


主成分分析PCA

主成分分析PCA（Principal component analysis）是一种研究数据相似性或差异性的可视化方法，通过一系列的特征值和特征向量进行排序后，选择主要的前几位特征值，采取降维的思想，PCA 可以找到距离矩阵中最主要的坐标，结果是数据矩阵的一个旋转，它没有改变样品点之间的相互位置关系，只是改变了坐标系统。详细关于主元分析的解释推荐大家看一篇文章，http://blog.csdn.net/ayw_hehe/article/details/5736659。通过PCA 可以观察个体或群体间的差异。

图解读：图中每一个点代表一个样本，相同颜色的点来自同一个分组，两点之间距离越近表明两者的群落构成差异越小。

LDA差异贡献分析



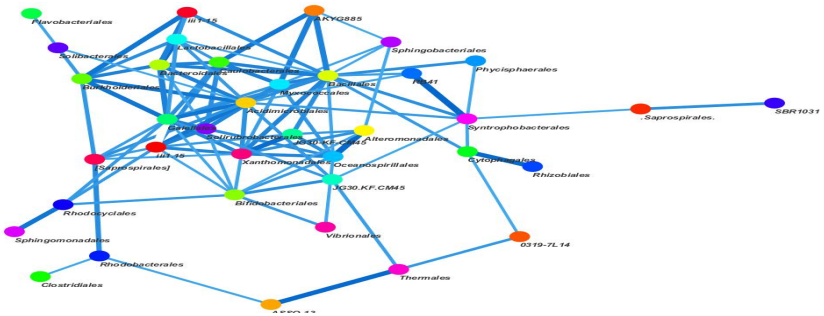
LDA差异贡献分析

PCA和LDA的差别在于，PCA，它所作的只是将整组数据整体映射到最方便表示这组数据的坐标轴上，映射时没有利用任何数据内部的分类信息，是无监督的，而LDA是由监督的，增加了种属之间的信息关系后，结合显著性差异标准测试(克鲁斯卡尔-沃利斯检验和两两Wilcoxon测试)和线性判别分析的方法进行特征选择。除了可以检测重要特征，他还可以根据效应值进行功能特性排序，这些功能特性可以解释顶部的大部分生物学差异。详细说明可以参考这篇文章 http://blog.csdn.net/sunmenggmail/article/details/80715022。

图解读：图中不同颜色代表不同样本或组之间的显著差异物种。使用LefSe软件分析获得，其中显著差异的logarithmic LDA score设为2。

物种相关性分析

物种相关性网络图A



物种相关性分析

根据各个物种在各个样品中的丰度以及变化情况，计算物种之间的相关性，包括正相关和负相关。然后对相关性最高的前100组关系进行绘制网络图。节点以不同颜色表示，连线的粗细和颜色深浅均表示相关性的高低。

相关性分析使用CCREPE进行Spearman秩相关分析并进行统计检验，详细的相关性结果见表文件：simulation.txt

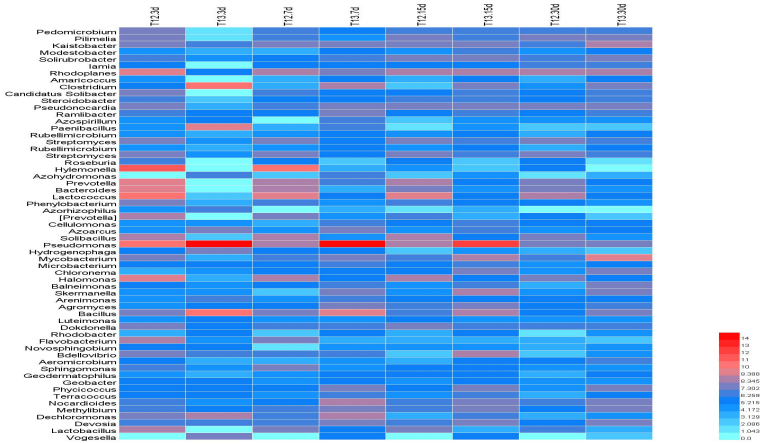
图解读：图中每一个点代表一个物种，存在相互关联的物种用连线相连，线的粗细和颜色的深浅都表示其相关性的高低。

物种相关性网络图B

图解读：物种相关性的另一种表示图，图中每一个点代表一个物种，大小表示与其他物种的关联关系的多少，关联越多物种以及相关性越高则节点半径和字体越大，线的粗细表示物种之间的相关性大小，越大线越粗。

聚类分析

无聚类热图



图解读：热图中的每一个色块代表一个样品的一个属的丰度，样品横向排列，属纵向排列，此图没有进行聚类，样品排序按照分组直接显示。

聚类热图

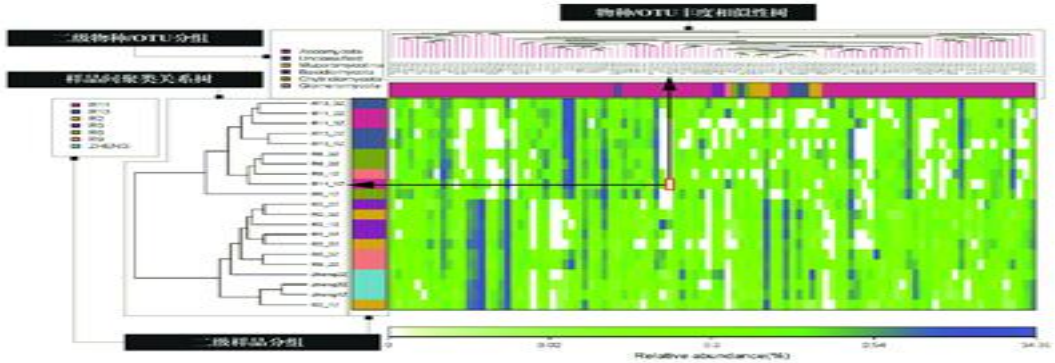
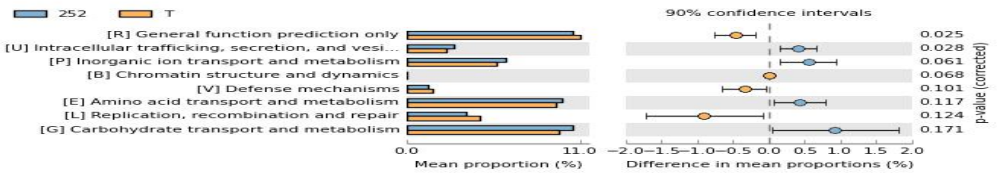


图 8

图解读：热图中的每一个色块代表一个样品的一个属的丰度，样品横向排列，属纵向排列，此图分别对样品和属进行聚类，从中可以了解样品之间的相似性以及属水平上的群落构成相似性。

群落功能差异分析

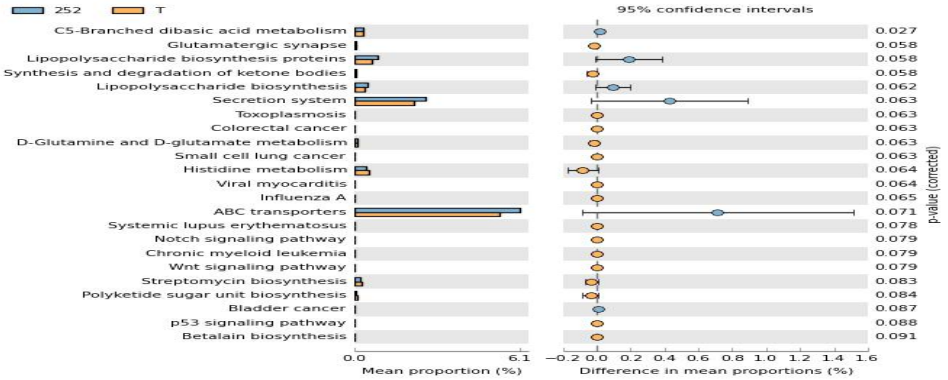
COG构成差异分析



通过对已有测序微生物基因组的基因功能的构成进行分析后，我们可以通过16s测序获得的物种构成推测样本中的功能基因的构成，从而分析不同样本和分组之间在功能上的差异（PICRUSt *Nature Biotechnology*, 1-10. 8 2013）。
通过对宏基因组测序数据功能分析和对应16s预测功能分析结果的比较发现，此方法的准确性在84%-95%，对肠道微生物菌群和土壤菌群的功能分析接近95%，能非常好的反映样品中的功能基因构成。
为了能够通过16s测序数据来准确的预测出功能构成，首先需要对原始16s测序数据的种属数量进行标准化，因为不同的种属菌包含的16s拷贝数不相同。然后将16s的种属构成信息通过构建好的已测序基因组的种属功能基因构成表映射获得预测的功能结果。
此处提供COG，KO基因预测以及KEGG代谢途径预测。用户也可自行使用我们提供的文件和软件（STAMP）对不同层级以及不同分组之间进行统计分析和制图，以及选择不同的统计方法和显著性水平。

图解读：图中不同颜色代表不同的分组，列出了COG构成在组间存在显著差异的功能分类以及在各组的比例，此外右侧还给出了差异的比例和置信区间以及P-value。

KEGG代谢途径第三层分类差异分析

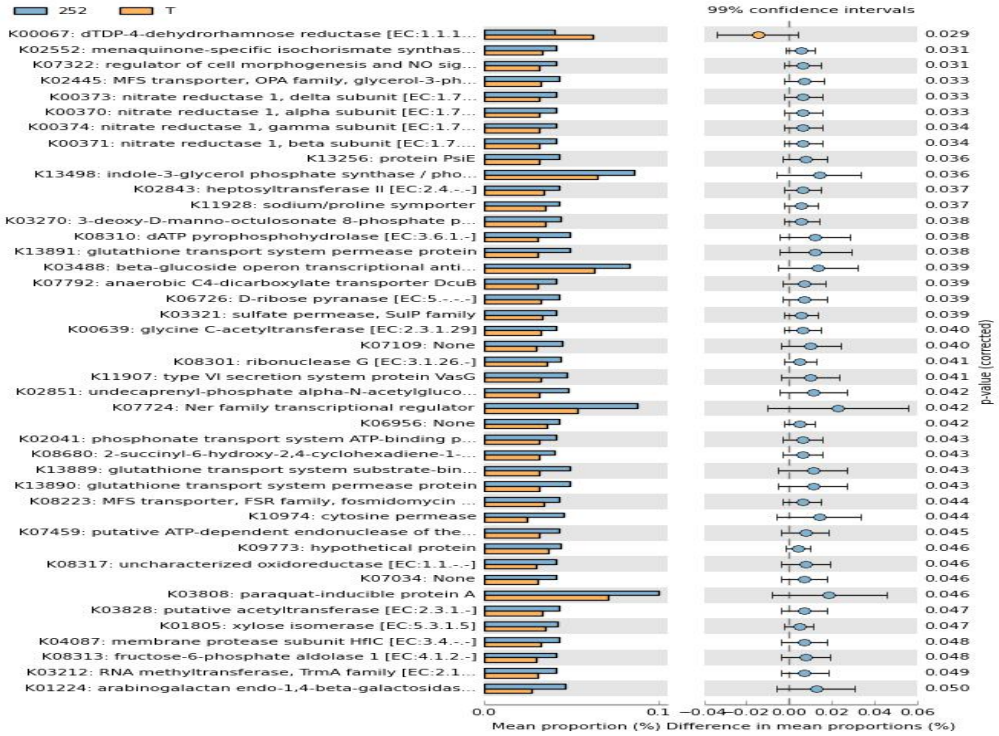


通过KEGG代谢途径的预测差异分析，我们可以了解到不同分组的样品之间在微生物群落的功能基因在代谢途径上的差异，以及变化的高低。为我们了解群落样本的环境适应变化的代谢过程提供一种简便快捷的方法。

图解读：图中不同颜色代表不同的分组，列出了在第三层级的构成在组间存在显著差异的KEGG代谢途径第三层分类以及在各组的比例，此外右侧还给出了差异的比例和置信区间以及P-value。

群落功能差异分析

KEGG功能基因KO构成差异分析

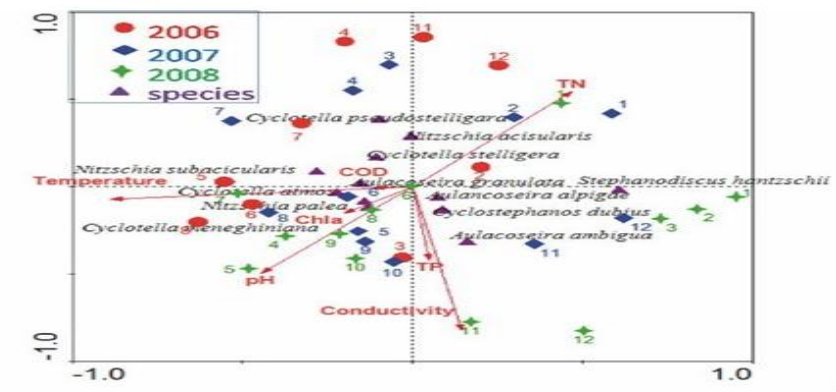


除了能对大的基因功能分类和代谢途径进行预测外，我们还能提供精细的功能基因的数量和构成的预测，以及进行样本间以及组间的差异分析，并给出具有统计意义和置信区间的分析结果。这一分析将我们对于样本群落的差异进一步深入到了每一类基因的层面。

图解读：图中不同颜色代表不同的分组，列出了在组间/样本间存在显著差异的每一个功能基因（酶）以及在各组的比例，此外右侧还给出了差异的比例和置信区间以及P-value。

环境因子分析

RDA分析



CCA/RDA分析

基于对应分析发展的一种排序方法，将对应分析与多元回归分析相结合，每一步计算均与环境因子进行回归，又称多元直接梯度分析。主要用来反映菌群与环境因子之间的关系。

RDA 是基于线性模型，CCA是基于单峰模型。分析可以检测环境因子、样品、菌群三者之间的关系或者两两之间的关系。

横轴和纵轴：RDA 和CCA 分析，模型不同，横纵坐标上的刻度为每个样品或者物种在与环境因子进行回归分析计算时产生的值，可以绘制于二维图形中。

图解读：冗余分析可以基于所有样品的OTU作图，也可以基于样品中优势物种作图；
箭头射线：箭头分别代表不同的环境因子；
夹角：环境因子之间的夹角为锐角时表示两个环境因子之间呈正相关关系，钝角时呈负相关关系。环境因子的射线越长，说明该影响因子的影响程度越大； 不同颜色的点表示不同组别的样品或者同一组别不同时期的样品，图中的拉丁文代表物种名称，可以将关注的优势物种也纳入图中； 环境因子数量要少于样本数量，同时在分析时，需要提供环境因子的数据，比如 pH值，

参考文献

Blaxter, M.; Mann, J.; Chapman, T.; Thomas, F.; Whitton, C.; Floyd, R.; Abebe, E. (Oct 2005). "Defining operational taxonomic units using DNA barcode data.". Philos Trans R Soc Lond B Biol Sci 360 (1462): 1935–43.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet allocation" . Journal of Machine Learning Research 3 (4–5): pp. 993–1022.

Gotelli, Nicholas J.; Colwell, Robert K, Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness (2001)Ecology Letters 4 (4): 379–391.

H. Tuomisto, Dept of Biology, A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity, Ecography 33: 2_22, 2010

Herfindahl, O. C. (1950) Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University.

Huang J, Wang Z, et al. (2014) Identification of Microbial Communities in Open and Closed Circuit Bioelectrochemical MBRs by High-Throughput 454 Pyrosequencing. PLoS ONE 9(4): e93842.

J Gregory Caporaso1,12, Justin Kuczynski2,12,QIIME allows analysis of highthroughput community sequencing data, nature methods(2010), DOI:10.1038/NMETH.F.303

John Venn (1880) "On the employment of geometrical diagrams for the sensible representations of logical propositions," Proceedings of the Cambridge Philosophical Society, 4 : 47-59.

Jonathon Shlens，A Tutorial on Principal Component Analysis，Systems Neurobiology（2005），CA 92093-0402.

Kellert, Stephen H. (1993). In the Wake of Chaos: Unpredictable Order in Dynamical Systems. University of Chicago Press. p. 32. ISBN 0-226-42976-8.

Legendre, P. and Legendre, L. 1998. Numerical Ecology. Second English Edition. Developments in Environmental Modelling 20. Elsevier, Amsterdam.

Lindsay I Smith，A tutorial on Principal Components Analysis ,(2002)

McGarigal, K., S. Cushman, and S. Stafford (2000). Multivariate Statistics for Wildlife and Ecology Research. New York, New York, USA: Springer.

M.Vi˜nasd,*，S. Lladóa,b ,etc.Pyrosequencing reveals the effect of mobilizing agents and lignocellulosic substrate amendment on microbial community composition in a real industrial PAH-polluted soil. Journal of Hazardous Materials 283 (2015) 35–43

Shannon, C. E. and Weaver W. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423 and 623–656.

Simpson, E. H. (1949). "Measurement of diversity". Nature 163:688.

Venn, J. (1880) "On the diagrammatic and mechanical representation of propositions and reasonings". Philosophical Magazine and Journal of Science. 5 10 (59): 1–18.

Weixiang Wu*, Xiaohui Guo,.A comparison of microbial characteristics between the thermophilic and mesophilic anaerobic digesters exposed to elevated food waste loadings. Bioresource Technology 152 (2014) 420–428

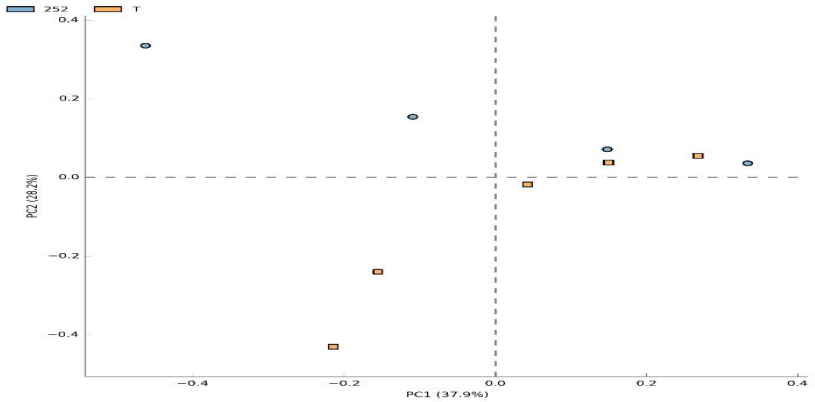
Whittaker, R. H.. Evolution and Measurement of Species Diversity. (1972) Taxon, 21, 213-251.

Zhi-Pei Liua,* strategyYang Xua,b, etc. Successful bioremediation of an aged and heavily contaminated soilusing a microbial/plant combination strategy.Journal of Hazardous Materials 264 (2014) 430– 438.

Zongxin Ling1*, Xia Liu1,2*,etc. Impacts of infection with different toxigenic Clostridium difficile strains on faecal microbiota in children.SCIENTIFIC REPORTS 4 : 7485

图例说明与信息解读

PCA分析

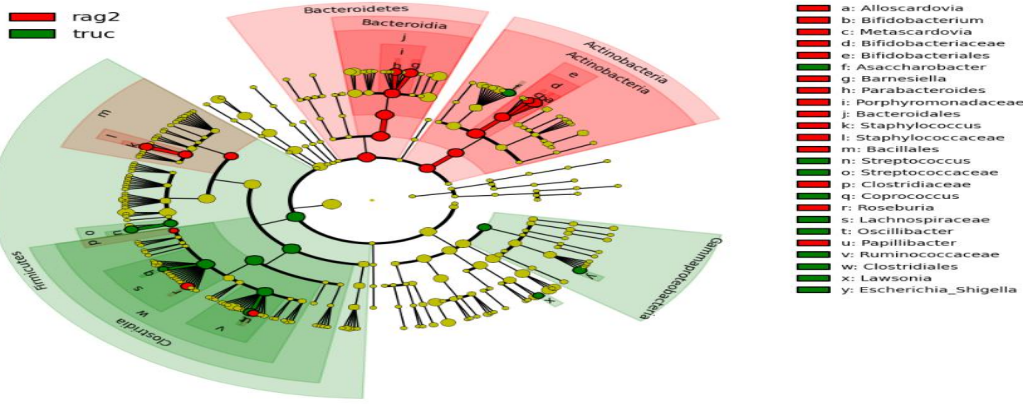


主成分分析PCA

主成分分析PCA（Principal component analysis）是一种研究数据相似性或差异性的可视化方法，通过一系列的特征值和特征向量进行排序后，选择主要的前几位特征值，采取降维的思想，PCA 可以找到距离矩阵中最主要的坐标，结果是数据矩阵的一个旋转，它没有改变样品点之间的相互位置关系，只是改变了坐标系统。详细关于主元分析的解释推荐大家看一篇文章，http://blog.csdn.net/ayw_hehe/article/details/5736659。通过PCA 可以观察个体或群体间的差异。

图解读：图中每一个点代表一个样本，相同颜色的点来自同一个分组，两点之间距离越近表明两者的群落构成差异越小。

LDA差异贡献分析

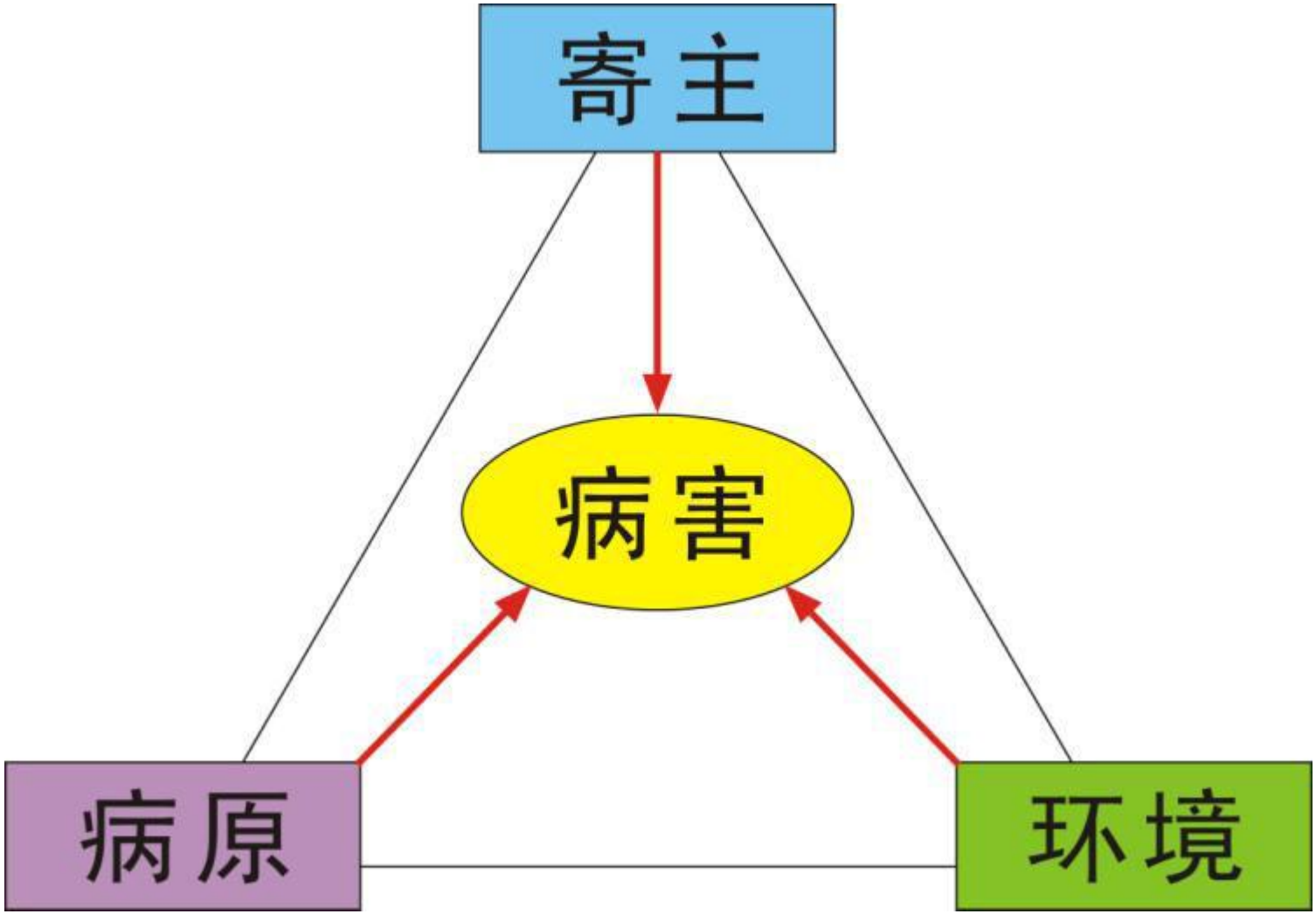


LDA差异贡献分析

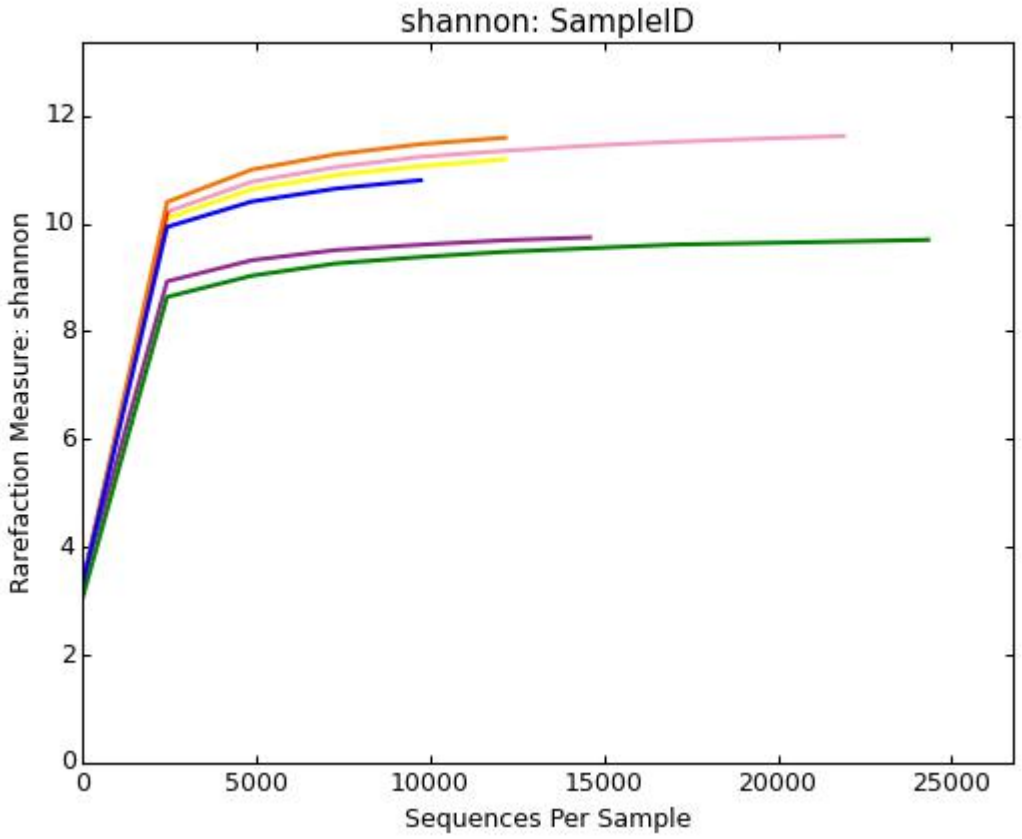
PCA和LDA的差别在于，PCA，它所作的只是将整组数据整体映射到最方便表示这组数据的坐标轴上，映射时没有利用任何数据内部的分类信息，是无监督的，而LDA是由监督的，增加了种属之间的信息关系后，结合显著性差异标准测试(克鲁斯卡尔-沃利斯检验和两两Wilcoxon测试)和线性判别分析的方法进行特征选择。除了可以检测重要特征，他还可以根据效应值进行功能特性排序，这些功能特性可以解释顶部的大部分生物学差异。详细说明可以参考这篇文章 http://blog.csdn.net/sunmenggmail/article/details/8071502。

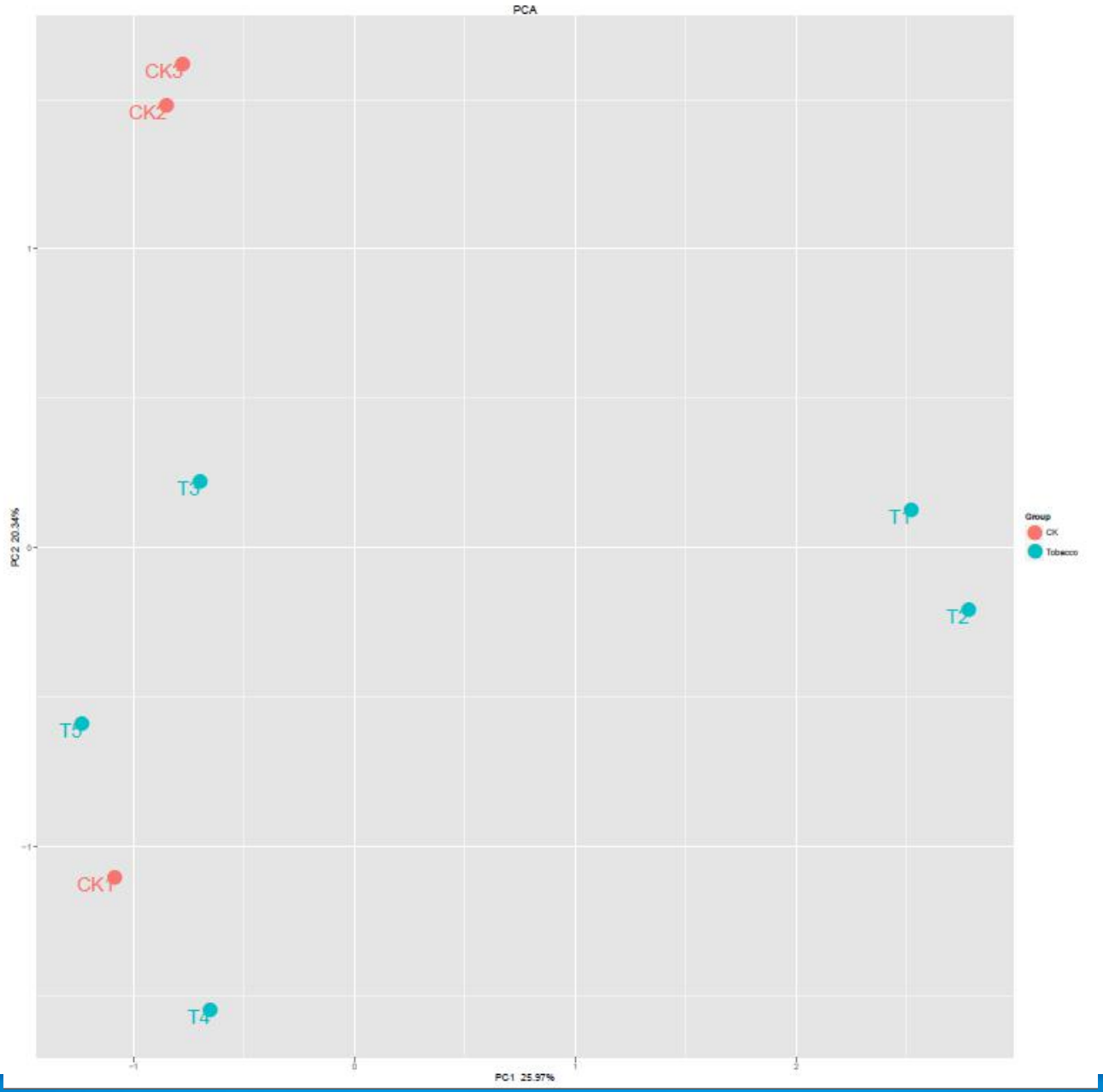
图解读：图中不同颜色代表不同样本或组之间的显著差异物种。使用LefSe软件分析获得，其中显著差异的logarithmic LDA score设为2。

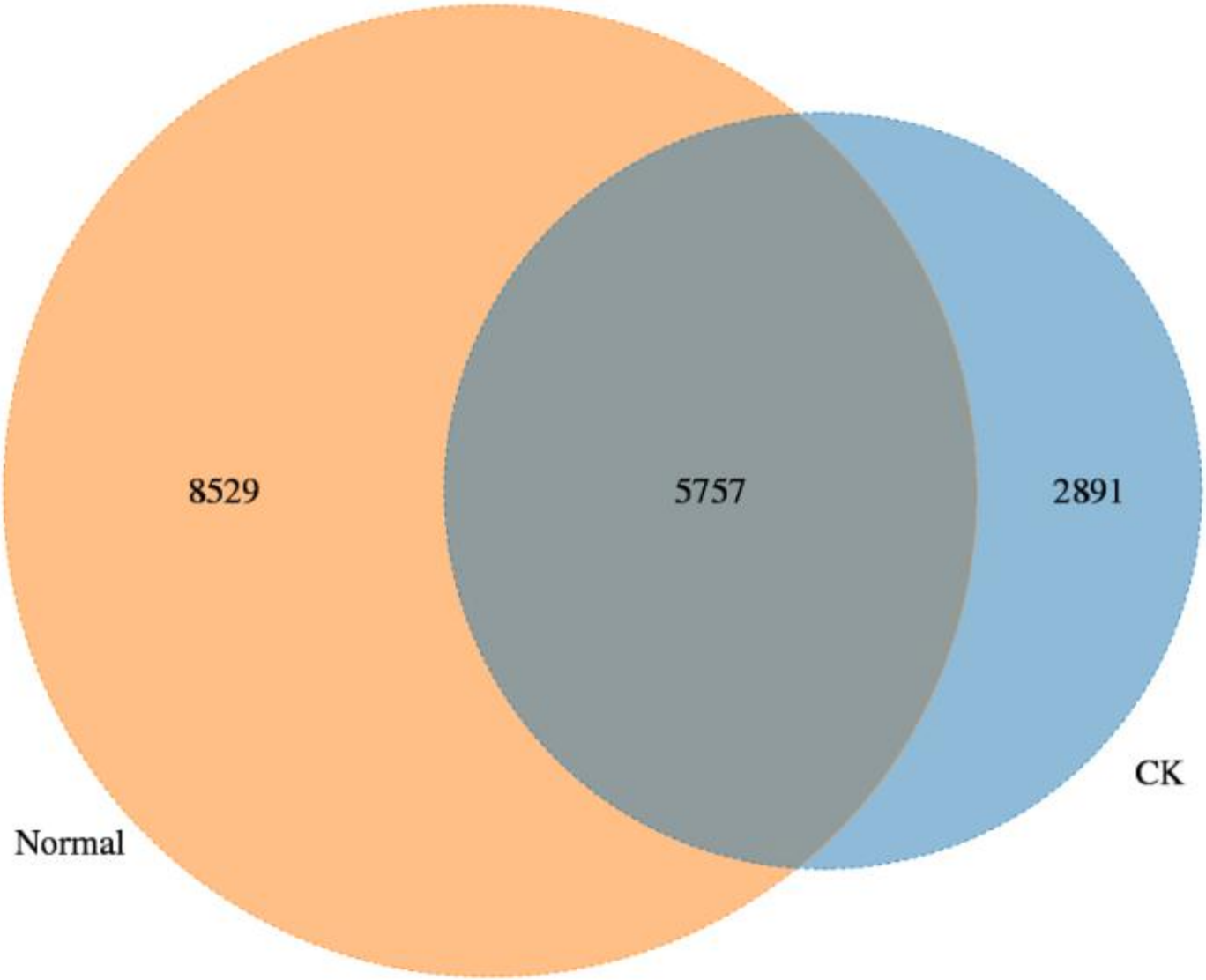
宏基因组(16s)测序实例

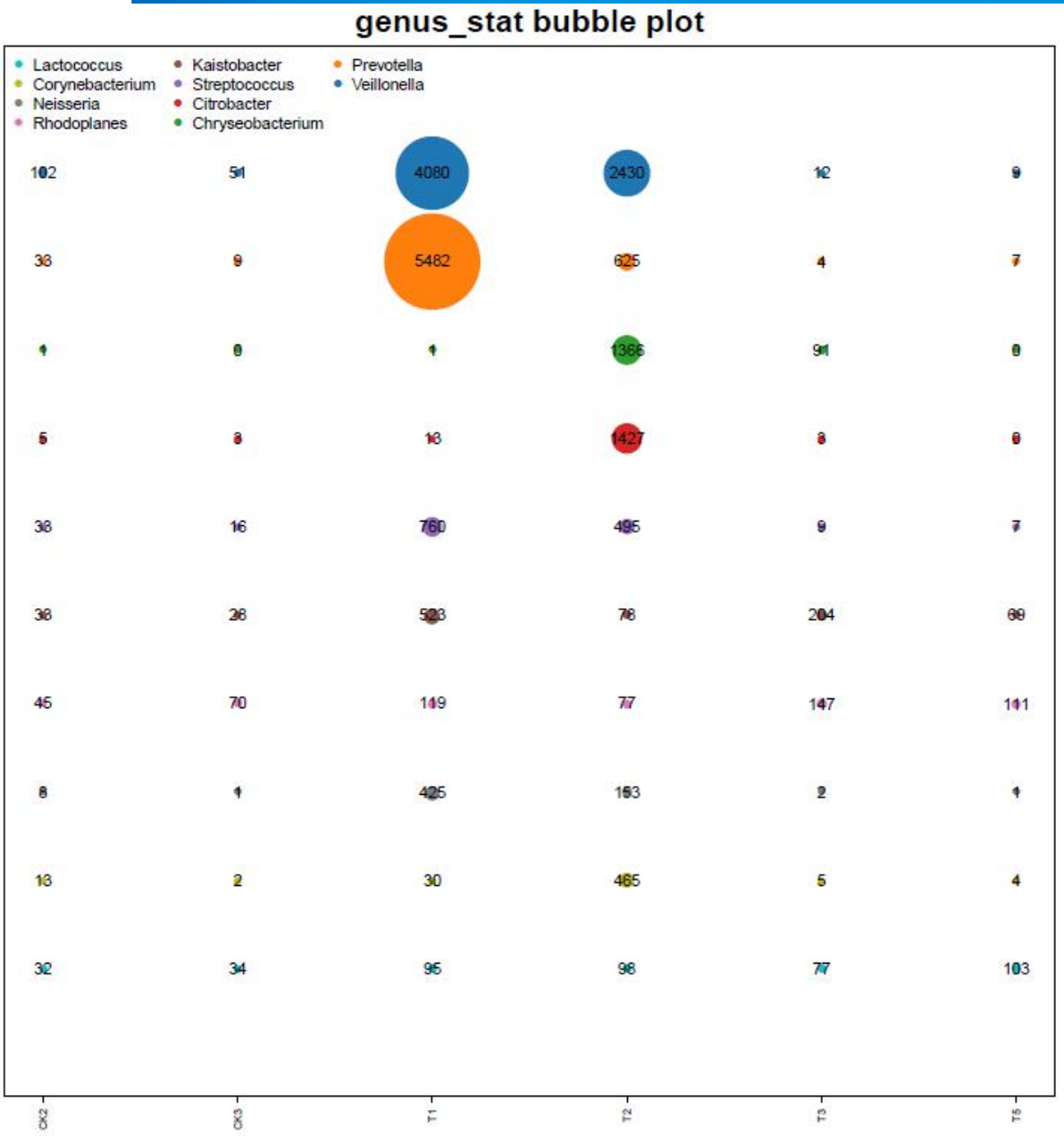


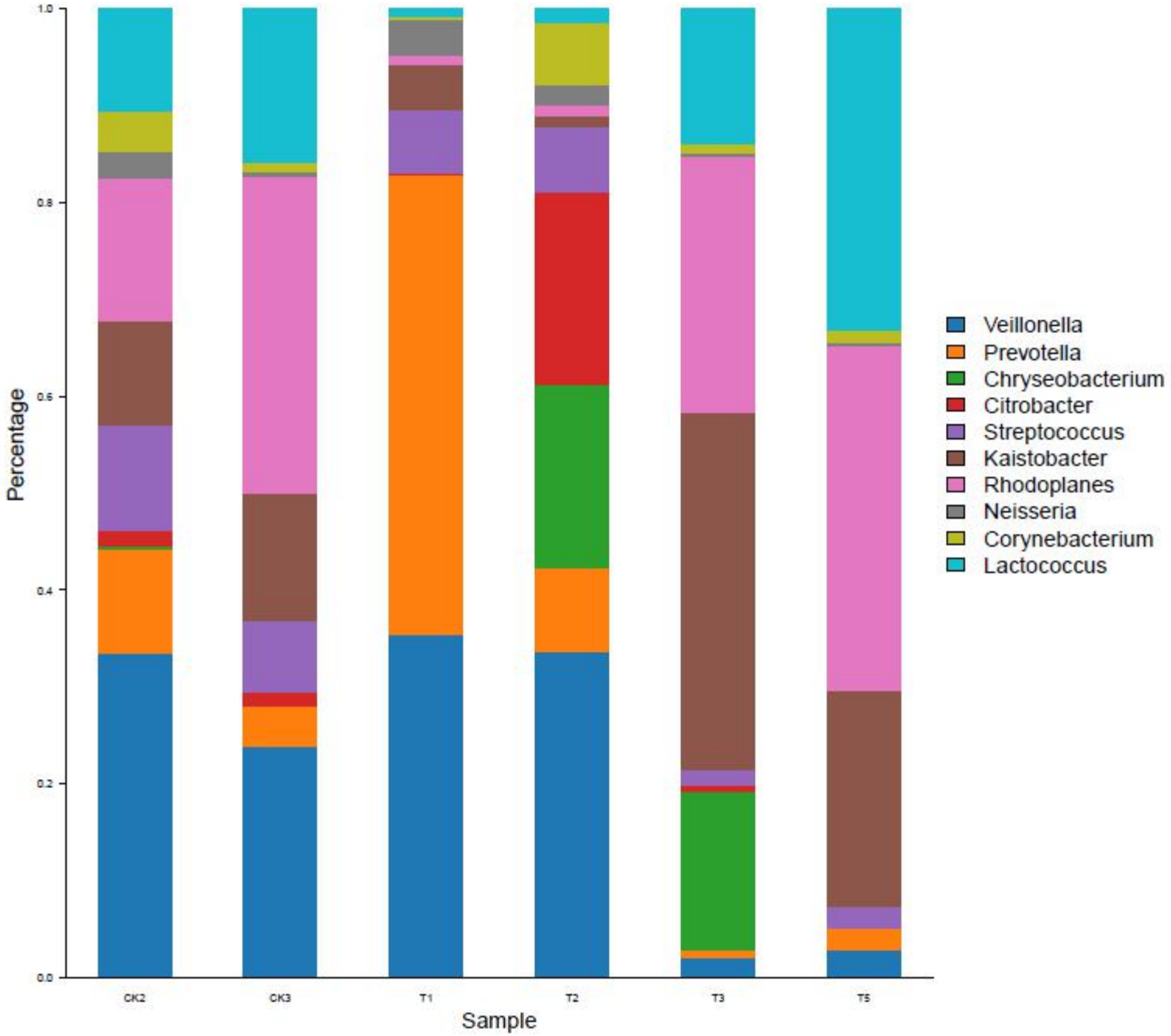
病害发生的三要素


shannon: SampleID
Rarefaction Measure: shannon
Sequences Per Sample
0
2
4
6
8
10
12
0
5000
10000
15000
20000
25000

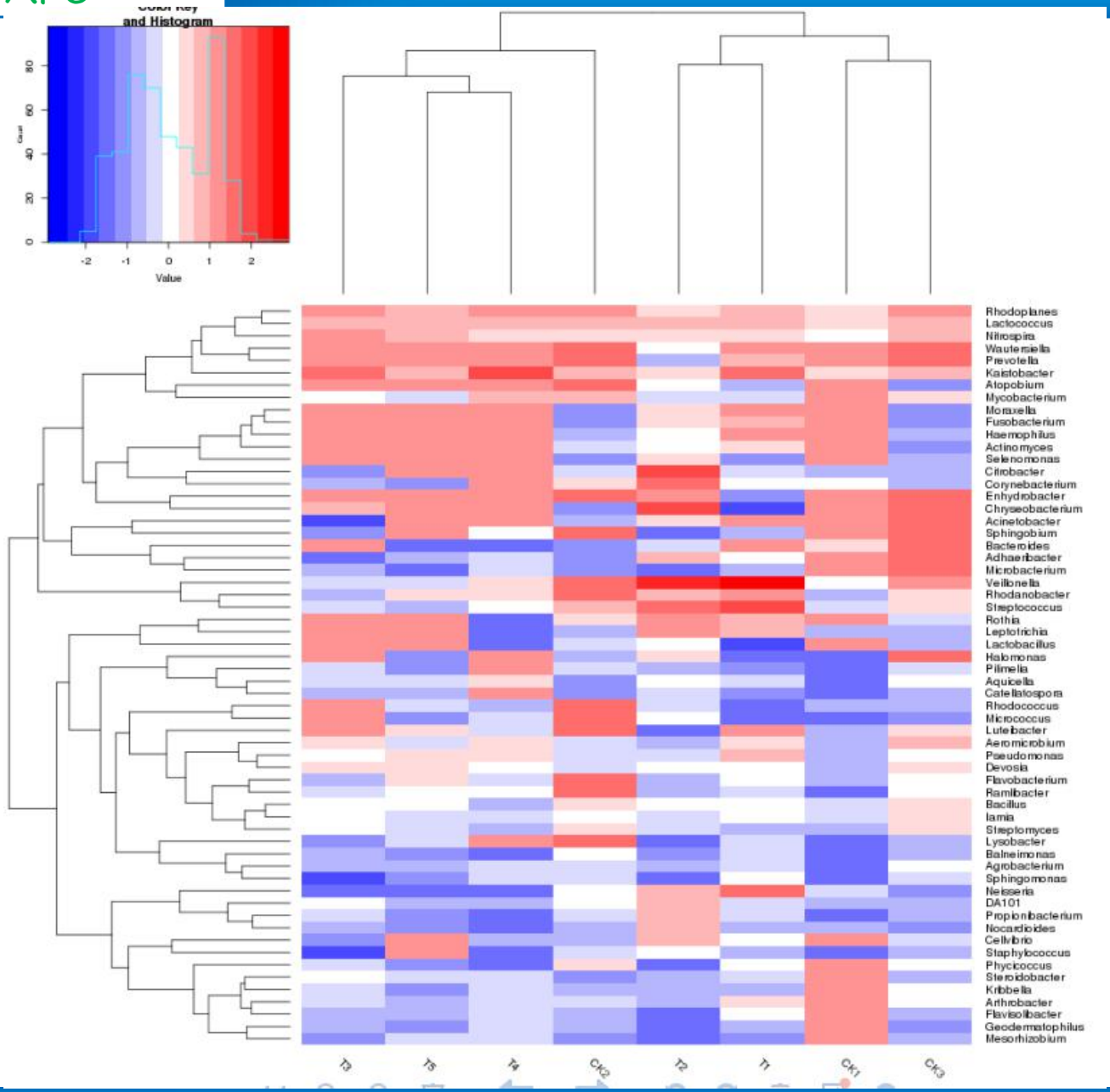

PCA
CK3
CK2
T3
T1
T2
T5
CK1
T4
PC2 20.34%
PC1 25.97%
Group
CK
Tobacco

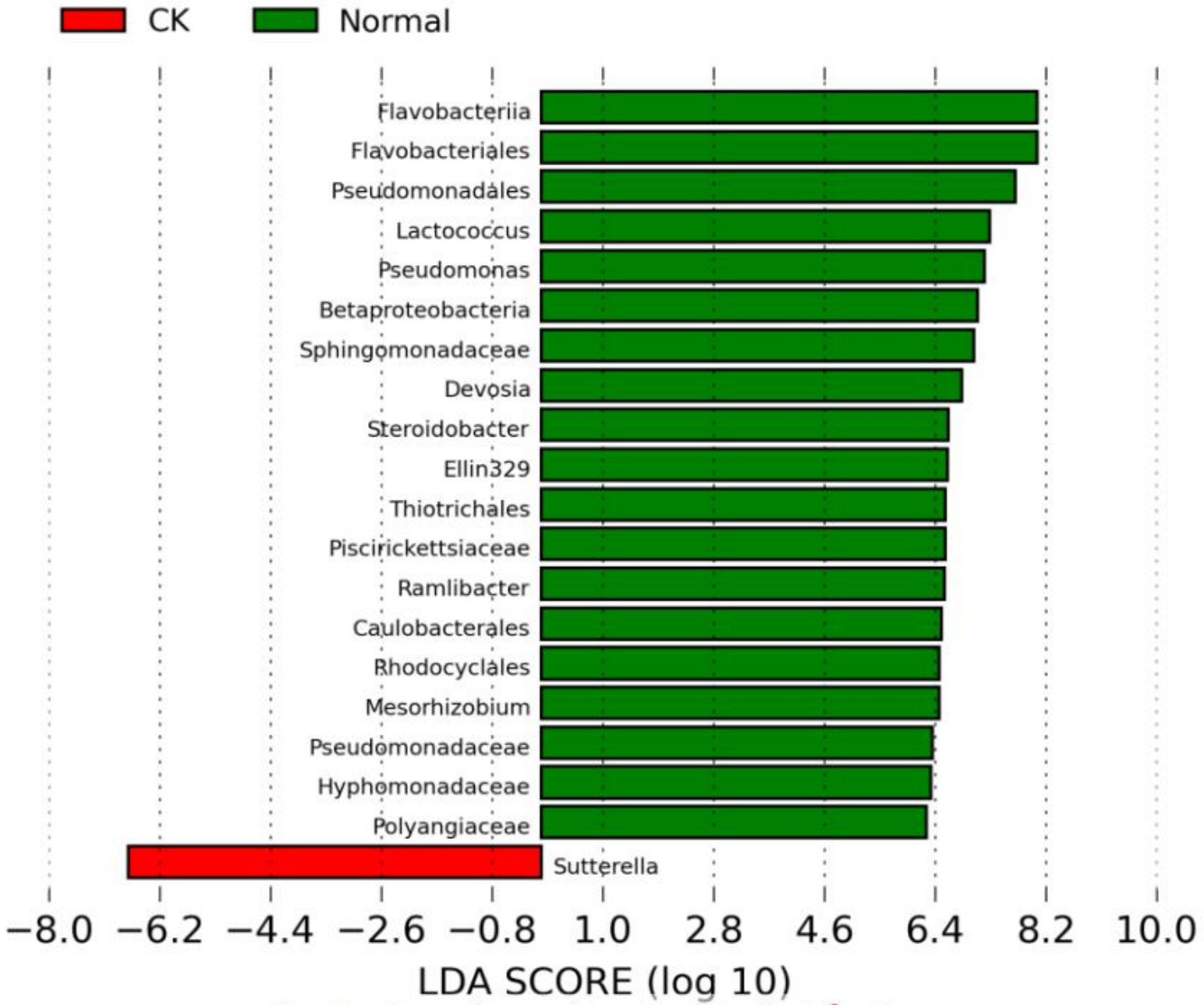

8529
5757
2891
Normal
CK


genus_stat bubble plot
Lactococcus
Corynebacterium
Neisseria
Rhodoplanes
Kaistobacter
Streptococcus
Citrobacter
Chryseobacterium
Prevotella
Veillonella
4080
2430
5482
CK2
CK3
T1
T2
T3
T5

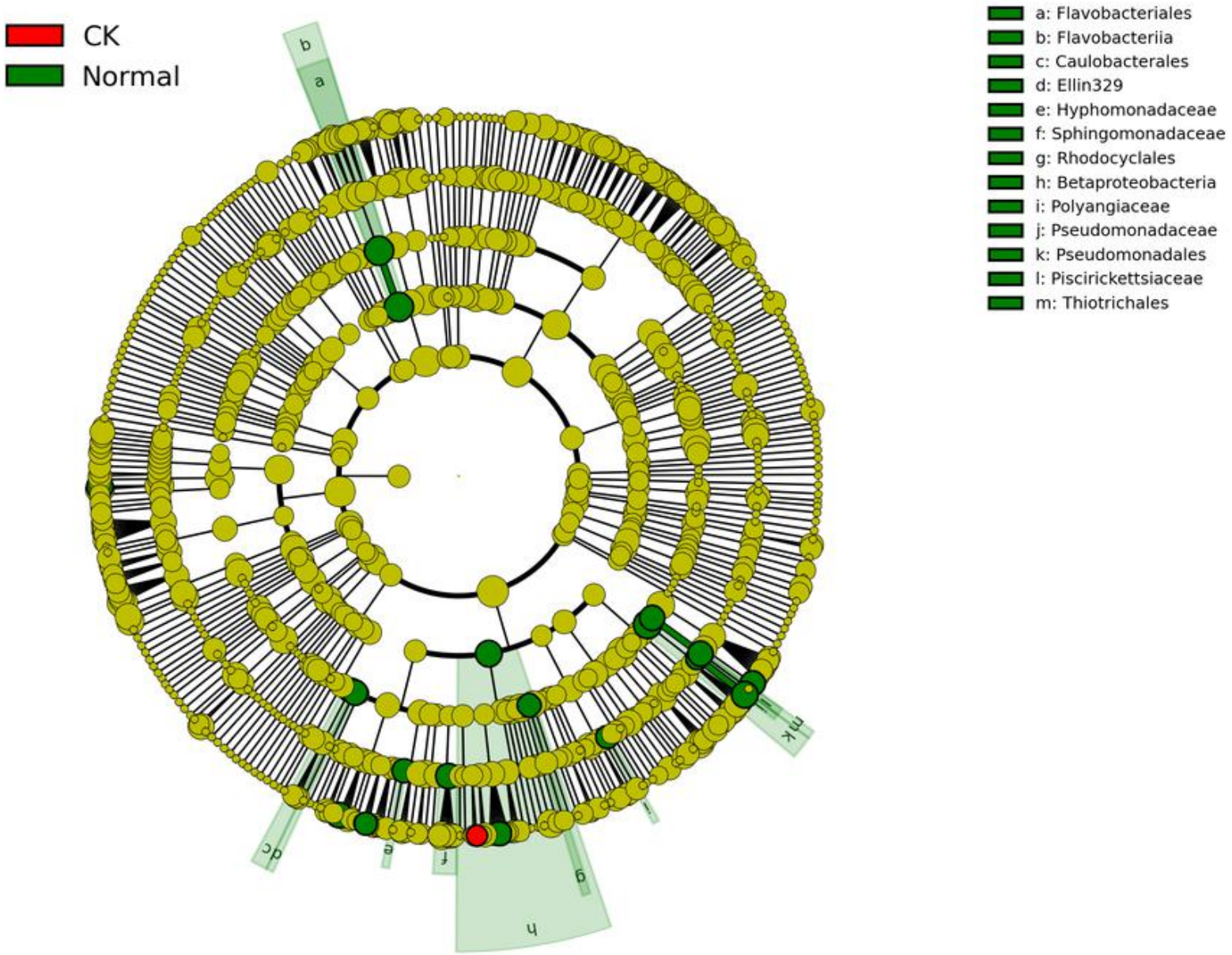

Percentage
1.0
0.8
0.6
0.4
0.2
0.0
CK2
CK3
T1
T2
T3
T5
Sample
Veillonella
Prevotella
Chryseobacterium
Citrobacter
Streptococcus
Kaistobacter
Rhodoplanes
Neisseria
Corynebacterium
Lactococcus

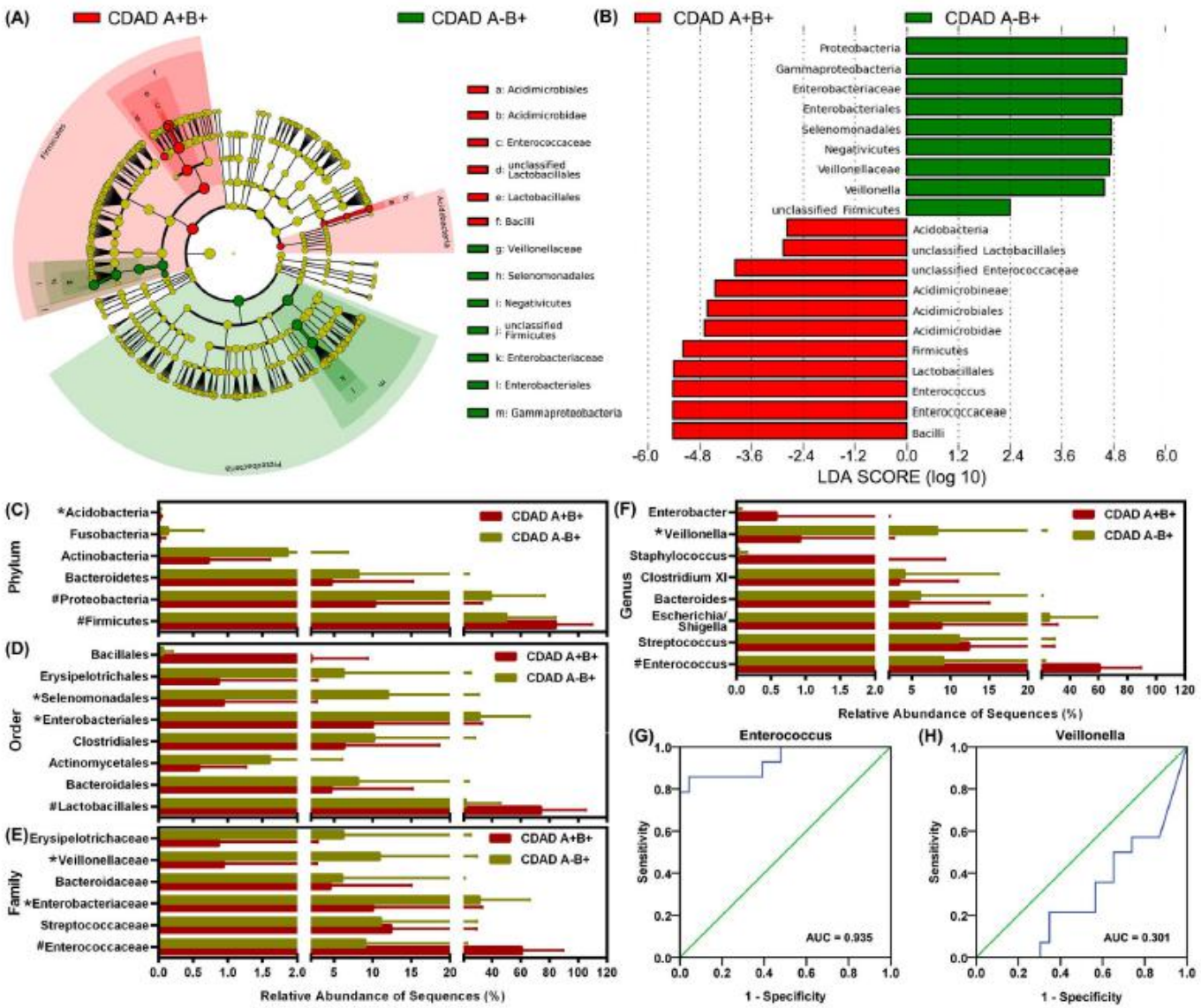

A: Veillonellaceae
B: Prevotellaceae
C: Weeksellaceae
D: Enterobacteriaceae
E: Sphingomonadaceae
F: Streptococcaceae
G: Hyphomicrobiaceae
H: Corynebacteriaceae
I: Nitrospiraceae
J: Neisseriaceae
K: Leptotrichiaceae
L: Moraxellaceae
M: Micrococcaceae
N: Pseudomonadaceae
O: Mycobacteriaceae
Species tree
Acidobacteria
Actinobacteria
Armatimonadetes
Bacteroidetes
Chloroflexi
Cyanobacteria
Firmicutes
Fusobacteria
Nitrospirae
Planctomycetes
Proteobacteria
TM7
Thermi
Verrucomicrobia
WS3
CK2
CK3
T1
T2
T3
T5
Abundance when present(log2)


Color Key and Histogram
Count
Value
Rhodoplanes
Lactococcus
Nitrospira
Wautersiella
Prevotella
Kaistobacter
Atopobium
Mycobacterium
Moraxella
Fusobacterium
Haemophilus
Actinomyces
Selenomonas
Citrobacter
Corynebacterium
Enhydrobacter
Chryseobacterium
Acinetobacter
Sphingobium
Bacteroides
Adhaeribacter
Microbacterium
Veillonella
Rhodanobacter
Streptococcus
Rothia
Leptotrichia
Lactobacillus
Halomonas
Pilimelia
Aquicella
Catellatospora
Rhodococcus
Micrococcus
Luteibacter
Aeromicrobium
Pseudomonas
Devosia
Flavobacterium
Ramlibacter
Bacillus
Iamia
Streptomyces
Lysobacter
Balneimonas
Agrobacterium
Sphingomonas
Neisseria
DA101
Propionibacterium
Nocardioides
Cellvibrio
Staphylococcus
Phycicoccus
Steroidobacter
Kribbella
Arthrobacter
Flavisolibacter
Geodermatophilus
Mesorhizobium
T3
T5
T4
CK2
T2
T1
CK1
CK3


CK
Normal
Flavobacteriia
Flavobacteriales
Pseudomonadales
Lactococcus
Pseudomonas
Betaproteobacteria
Sphingomonadaceae
Devosia
Steroidobacter
Ellin329
Thiotrichales
Piscirickettsiaceae
Ramlibacter
Caulobacterales
Rhodocyclales
Mesorhizobium
Pseudomonadaceae
Hyphomonadaceae
Polyangiaceae
Sutterella
−8.0 −6.2 −4.4 −2.6 −0.8 1.0 2.8 4.6 6.4 8.2 10.0
LDA SCORE (log 10)

Cladogram


CK
Normal
a: Flavobacteriales
b: Flavobacteriia
c: Caulobacterales
d: Ellin329
e: Hyphomonadaceae
f: Sphingomonadaceae
g: Rhodocyclales
h: Betaproteobacteria
i: Polyangiaceae
j: Pseudomonadaceae
k: Pseudomonadales
l: Piscirickettsiaceae
m: Thiotrichales


(A)
CDAD A+B+
CDAD A-B+
a: Acidimicrobiales
b: Acidimicrobidae
c: Enterococcaceae
d: unclassified Lactobacillales
e: Lactobacillales
f: Bacilli
g: Veillonellaceae
h: Selenomonadales
i: Negativicutes
j: unclassified Firmicutes
k: Enterobacteriaceae
l: Enterobacteriales
m: Gammaproteobacteria
(B)
CDAD A+B+
CDAD A-B+
Proteobacteria
Gammaproteobacteria
Enterobacteriaceae
Enterobacteriales
Selenomonadales
Negativicutes
Veillonellaceae
Veillonella
unclassified Firmicutes
Acidobacteria
unclassified Lactobacillales
unclassified Enterococcaceae
Acidimicrobineae
Acidimicrobiales
Acidimicrobidae
Firmicutes
Lactobacillales
Enterococcus
Enterococcaceae
Bacilli
-6.0 -4.8 -3.6 -2.4 -1.2 0.0 1.2 2.4 3.6 4.8 6.0
LDA SCORE (log 10)
(C)
Phylum
*Acidobacteria
Fusobacteria
Actinobacteria
Bacteroidetes
#Proteobacteria
#Firmicutes
(D)
Order
Bacillales
Erysipelotrichales
*Selenomonadales
*Enterobacteriales
Clostridiales
Actinomycetales
Bacteroidales
#Lactobacillales
(E)
Family
Erysipelotrichaceae
*Veillonellaceae
Bacteroidaceae
*Enterobacteriaceae
Streptococcaceae
#Enterococcaceae
CDAD A+B+
CDAD A-B+
0.0 0.5 1.0 1.5 2.0 5 10 15 20 40 60 80 100 120
Relative Abundance of Sequences (%)
(F)
Genus
Enterobacter
*Veillonella
Staphylococcus
Clostridium XI
Bacteroides
Escherichia/ Shigella
Streptococcus
#Enterococcus
CDAD A+B+
CDAD A-B+
0.0 0.5 1.0 1.5 2.0 5 10 15 20 40 60 80 100 120
Relative Abundance of Sequences (%)
(G)
Enterococcus
Sensitivity
1 - Specificity
AUC = 0.935
(H)
Veillonella
Sensitivity
1 - Specificity
AUC = 0.301

基因组重（re-seq）测序

基因组重测序概念

- 基因组重测序是对基因组序列已知的个体进行基因组测序，并在个体或群体水平上进行差异性分析的方法。
- 优点：数据量大，可以获取大量的有用信息，构建超精细图谱，可以进行不同的性状关联分析，为基因定位提供大量高效依据，可以对物种的起源，进化以及变异情况进行详细了解。

测序选择

- 少量个体测序：主要针对植物特异性状的改变，也就是SNP，INDEL或者基因加倍等对植物性状的影响。（资金投入少）
- 群体重测序：

  1）遗传群体重测序：构建遗传图谱，进行QTL定位，同时进行相关性状的基因定位和GWAS分析。

  2）自然群体重测序：了解物种的起源，进化，人工选择对基因的影响等。

试验设计原则

- 尽量选择与已经全基因组测序的参考基因组亲缘关系近的品种。
- 在资金允许情况下尽量增加测序深度。（人类80X，植物10X）
- 群体选择要合理，性状研究尽量规避非目的性状差异，进化分析尽量包含不同来源不同背景的群体。

影响因素

- 测序丰度：测序丰度不够会导致有些区域的SNP统计缺失和不精确，导致结果误差。也可能造成假阳性。
- 参考基因组质量：参考基因组的拼接质量会严重影响重测序质量。烟草的测序质量太低导致重测序难度非常大。
- 亲缘关系远近：过远的亲缘关系可能导致某些差异较大的大片段无法比对到基因组上，导致片段缺失。

几个术语

- **Reads：** 高通量测序平台产生的序列标签就称为reads
- **Map：** 将测序得到reads比对到参考基因组上的过程。
- **QC：** quality control，对于下机数据进行质量控制，去除低质量reads以及接头序列等的处理方式。
- **SNP：** 个体间基因组DNA序列同一位置单个核苷酸变异(替代、插入或缺失)所引起的多态性。不同物种、个体基因组DNA序列同一位置上的单个核苷酸存在差别的现象。有这种差别的基因座、DNA序列等可作为基因组作图的标志。 癌症中特异的单核苷酸变异是一种体细胞突变（somatic mutation），称做SNV 。
- **Indel：** 基因组上小片段（<50bp）的插入或缺失 。

几个术语

- **CNV（copy number variation ）：** 基因组拷贝数变异是基因组变异的一种形式，通常使基因组中大片段的DNA形成非正常的拷贝数量。
- **SV（structure variation ）：** 染色体结构变异是指在染色体上发生了大片段的变异。主要包括染色体大片段的插入和缺失（引起CNV的变化），染色体内部的某块区域发生翻转颠换，两条染色体之间发生重组（inter-chromosome trans-location）等。

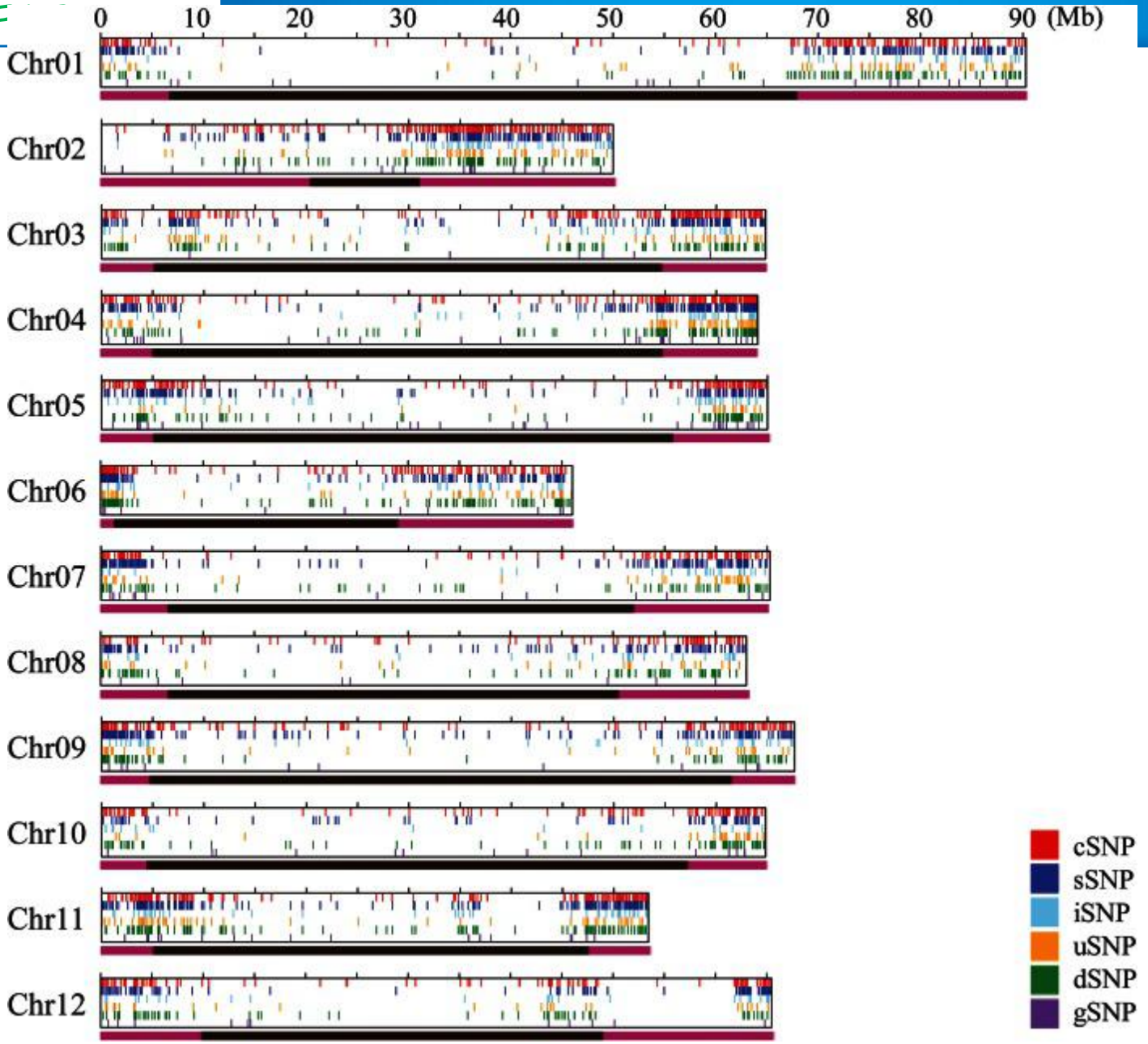
重测序数据处理软件

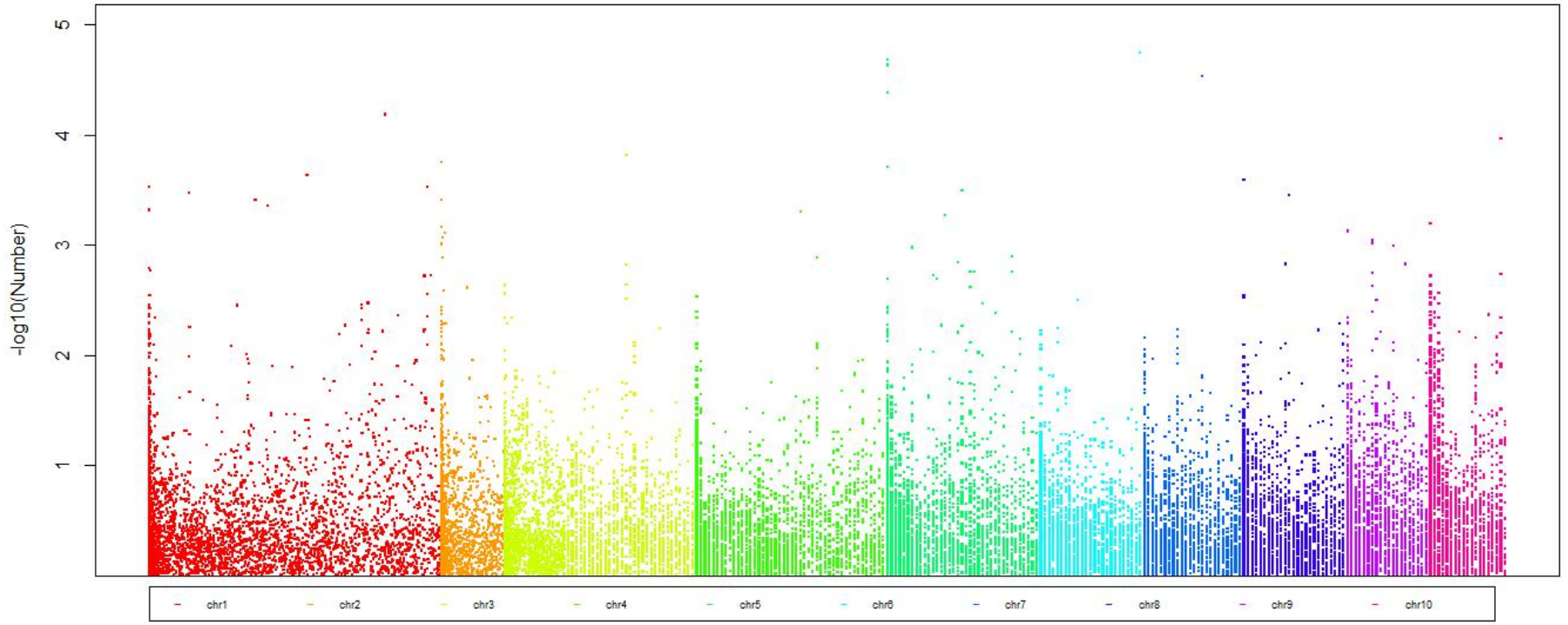
1. BWA + Samtools

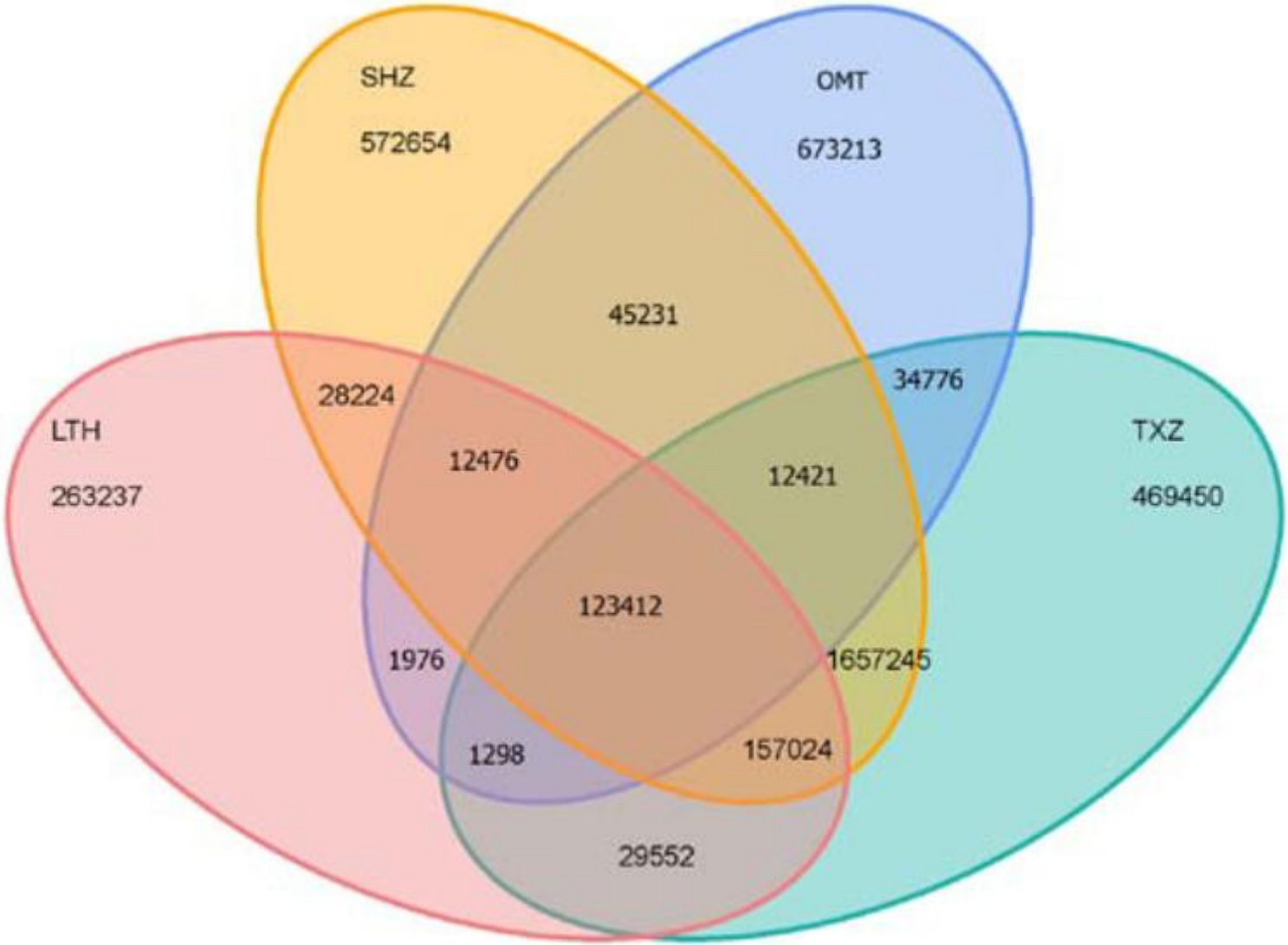
2. Soapsnp

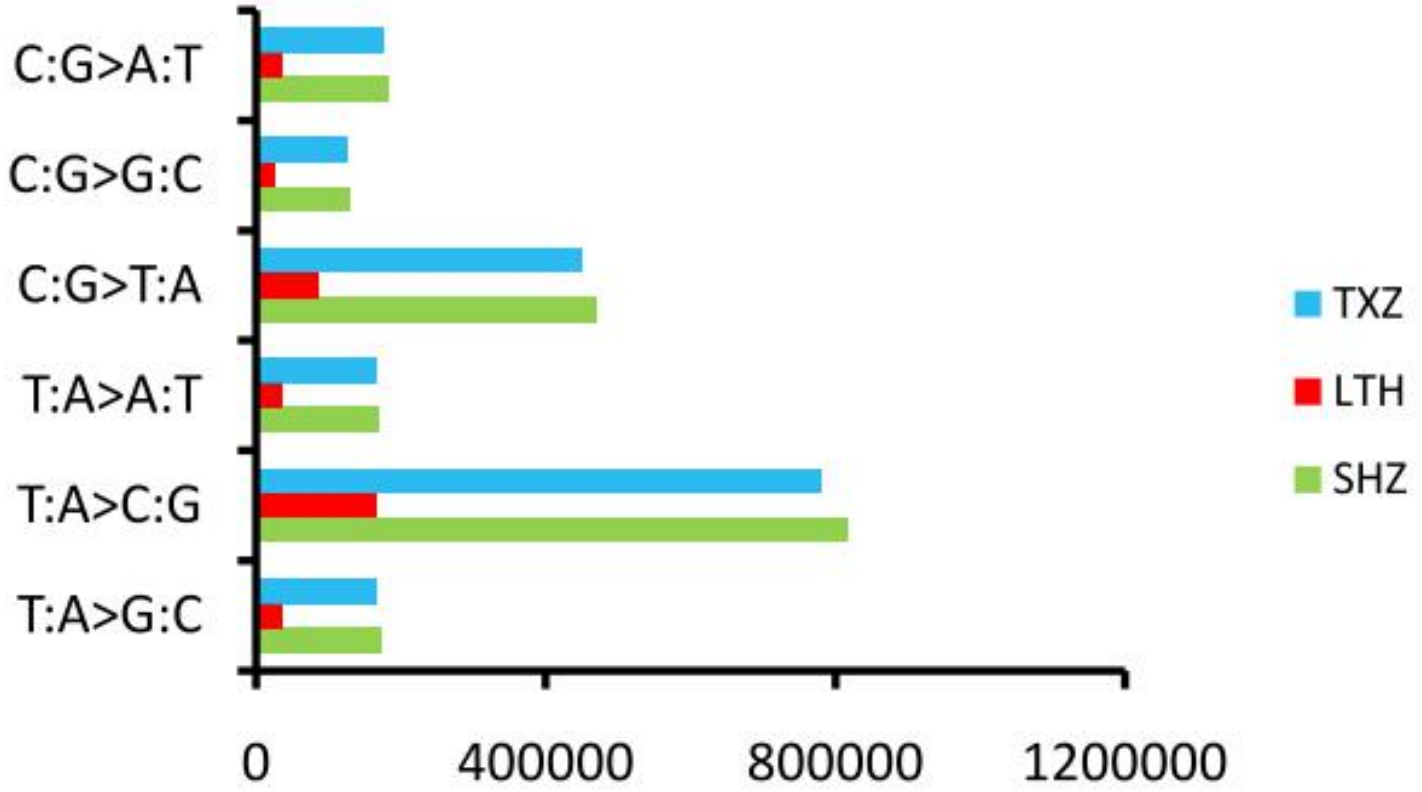
3. GATK

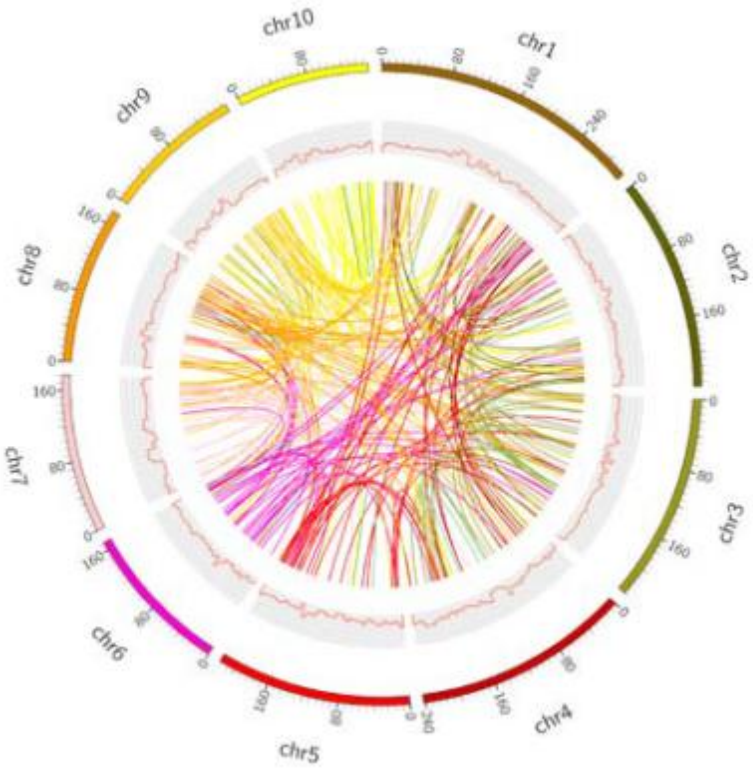
重测序变异分析

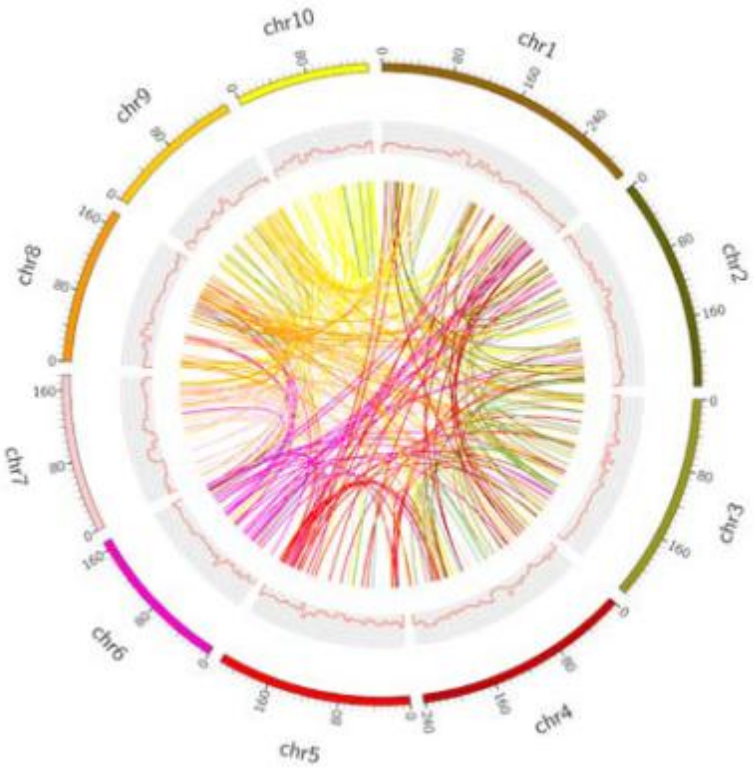

0
10
20
30
40
50
60
70
80
90 (Mb)
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
cSNP
sSNP
iSNP
uSNP
dSNP
gSNP

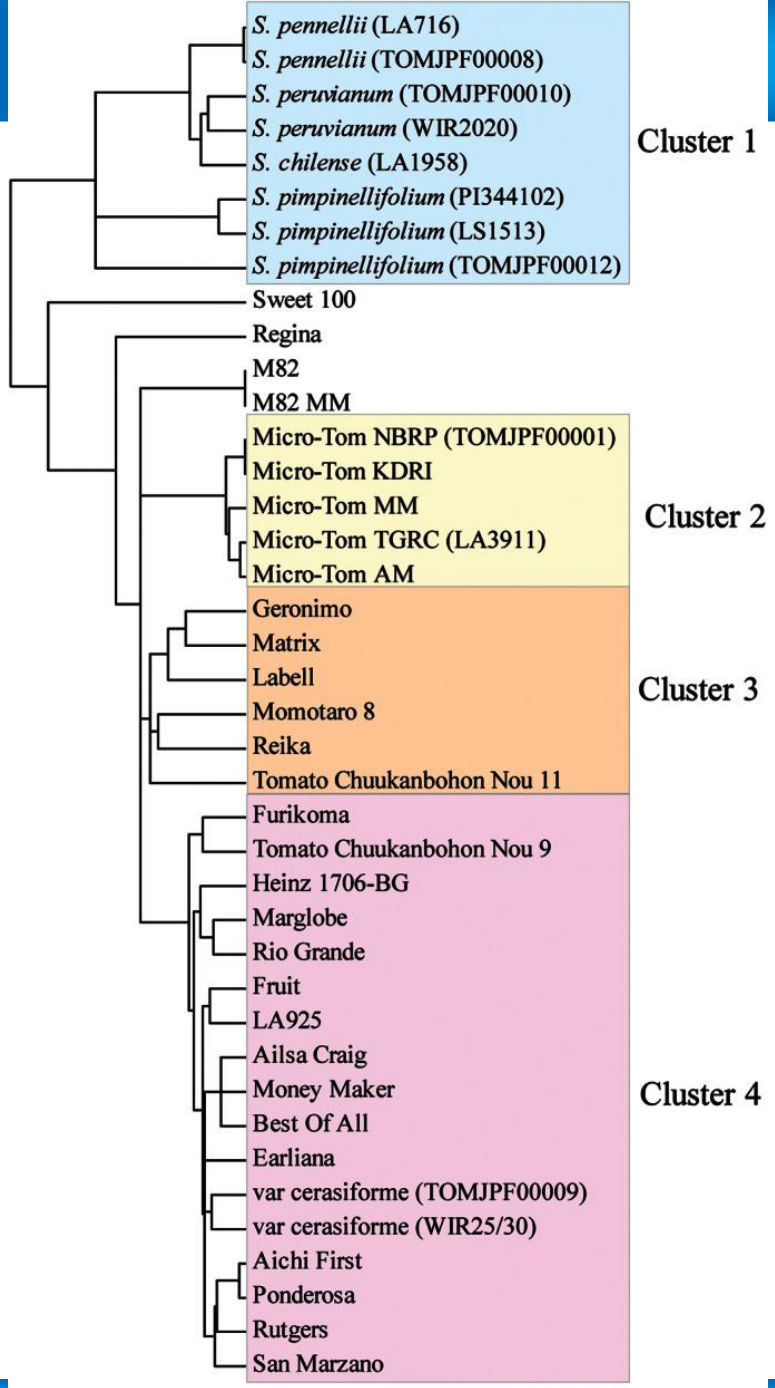

-log10(Number)
1
2
3
4
5
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10


SHZ
572654
OMT
673213
LTH
263237
TXZ
469450
45231
28224
34776
12476
12421
123412
1976
1657245
1298
157024
29552

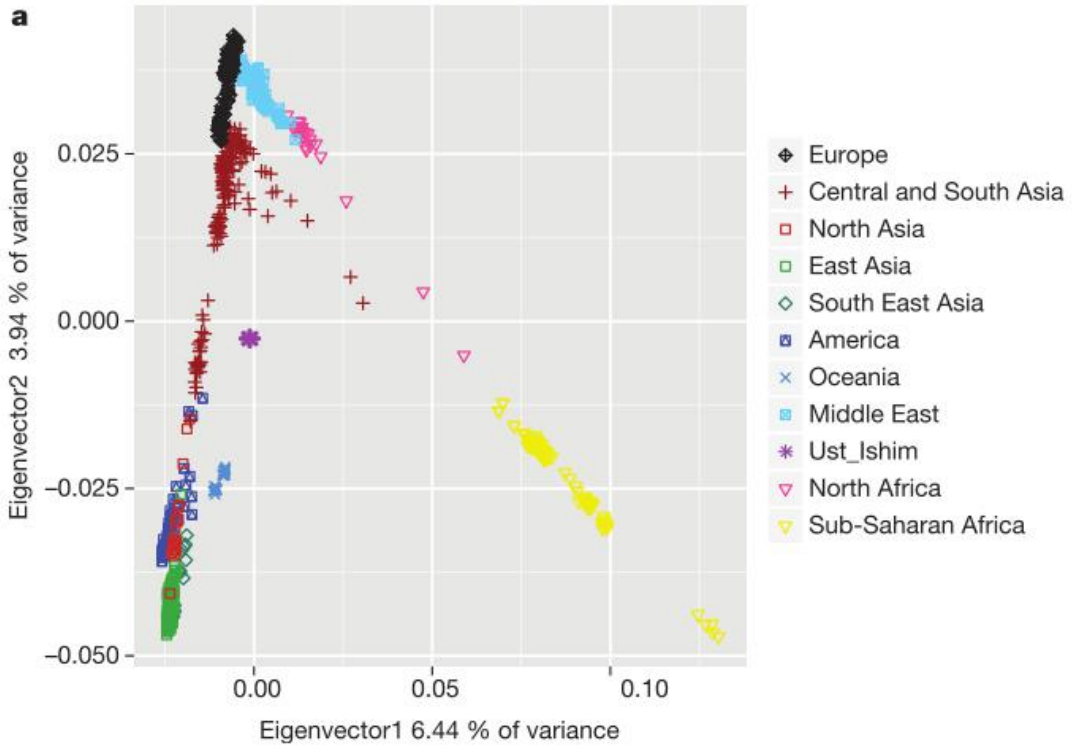

C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
0
400000
800000
1200000
TXZ
LTH
SHZ

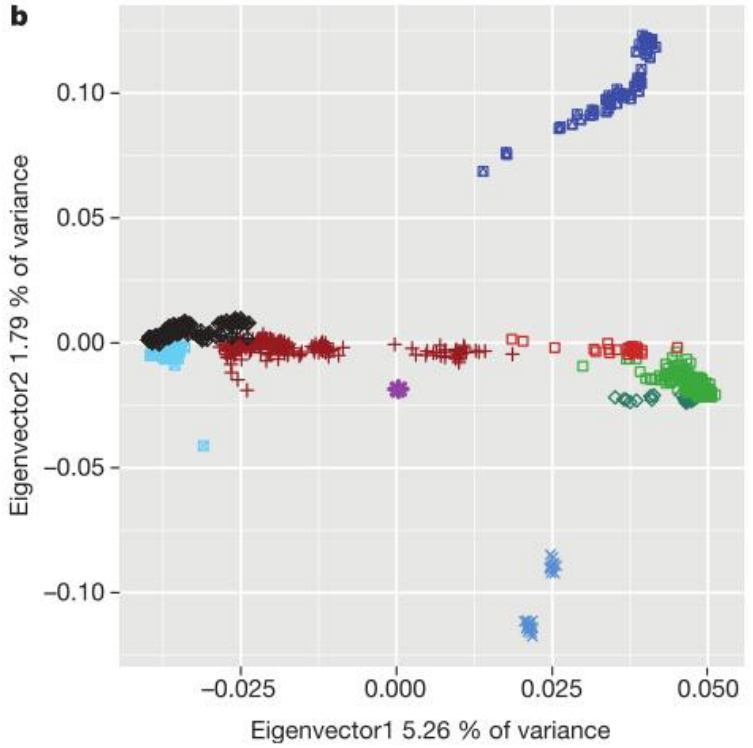

chr10
chr1
chr9
chr2
chr8
chr3
chr7
chr6
chr4
chr5


chr10
chr1
chr9
chr2
chr8
chr3
chr7
chr6
chr4
chr5

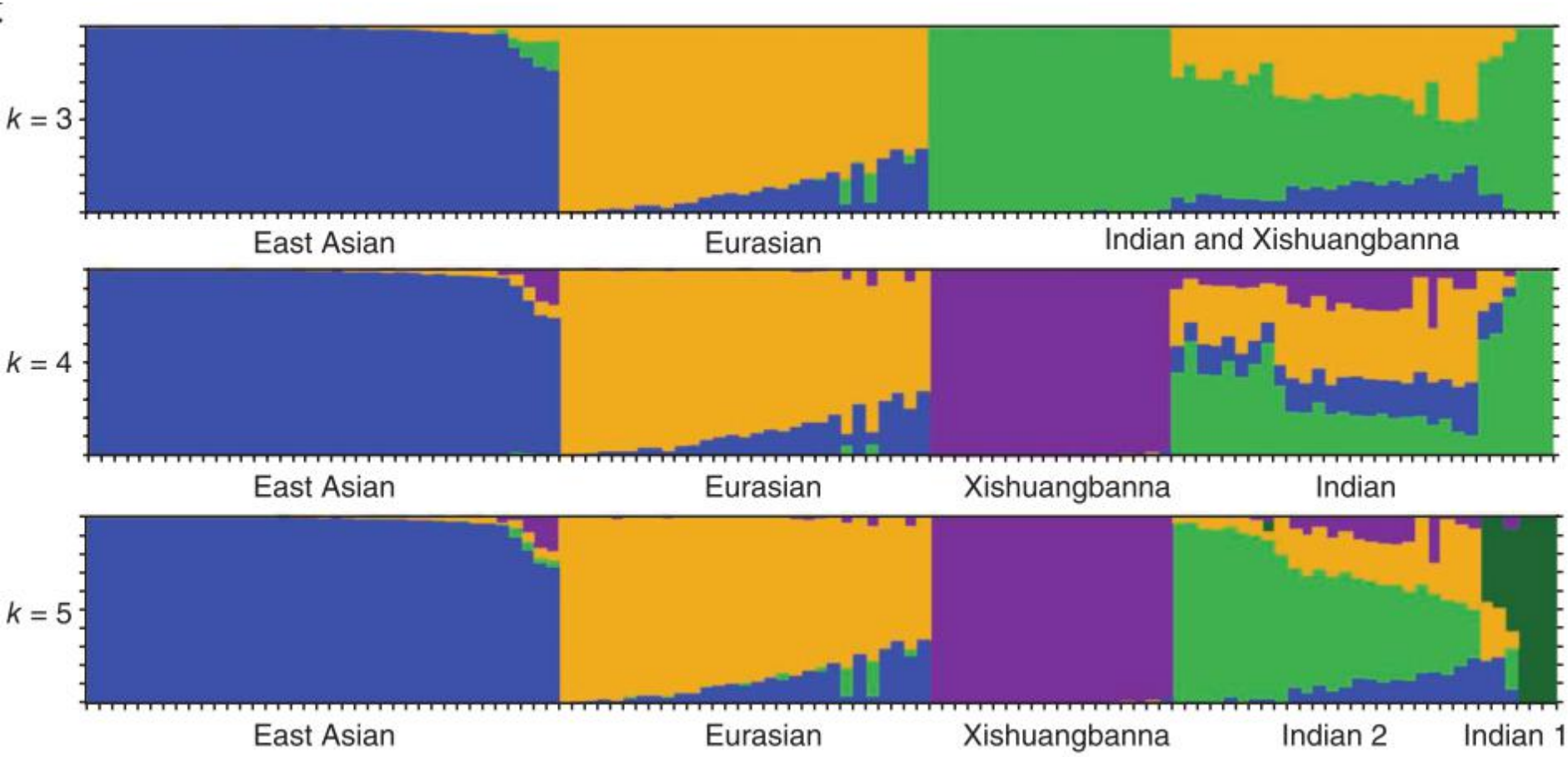

S. pennellii (LA716)
S. pennellii (TOMJPF00008)
S. peruvianum (TOMJPF00010)
S. peruvianum (WIR2020)
S. chilense (LA1958)
S. pimpinellifolium (PI344102)
S. pimpinellifolium (LS1513)
S. pimpinellifolium (TOMJPF00012)
Cluster 1
Sweet 100
Regina
M82
M82 MM
Micro-Tom NBRP (TOMJPF00001)
Micro-Tom KDRI
Micro-Tom MM
Micro-Tom TGRC (LA3911)
Micro-Tom AM
Cluster 2
Geronimo
Matrix
Labell
Momotaro 8
Reika
Tomato Chuukanbohon Nou 11
Cluster 3
Furikoma
Tomato Chuukanbohon Nou 9
Heinz 1706-BG
Marglobe
Rio Grande
Fruit
LA925
Ailsa Craig
Money Maker
Best Of All
Earliana
var cerasiforme (TOMJPF00009)
var cerasiforme (WIR25/30)
Aichi First
Ponderosa
Rutgers
San Marzano
Cluster 4
1600 1200 800 400 0

重测序主成分分析


a
Eigenvector2 3.94 % of variance
0.025
0.000
−0.025
−0.050
0.00
0.05
0.10
Eigenvector1 6.44 % of variance
Europe
Central and South Asia
North Asia
East Asia
South East Asia
America
Oceania
Middle East
Ust_Ishim
North Africa
Sub-Saharan Africa

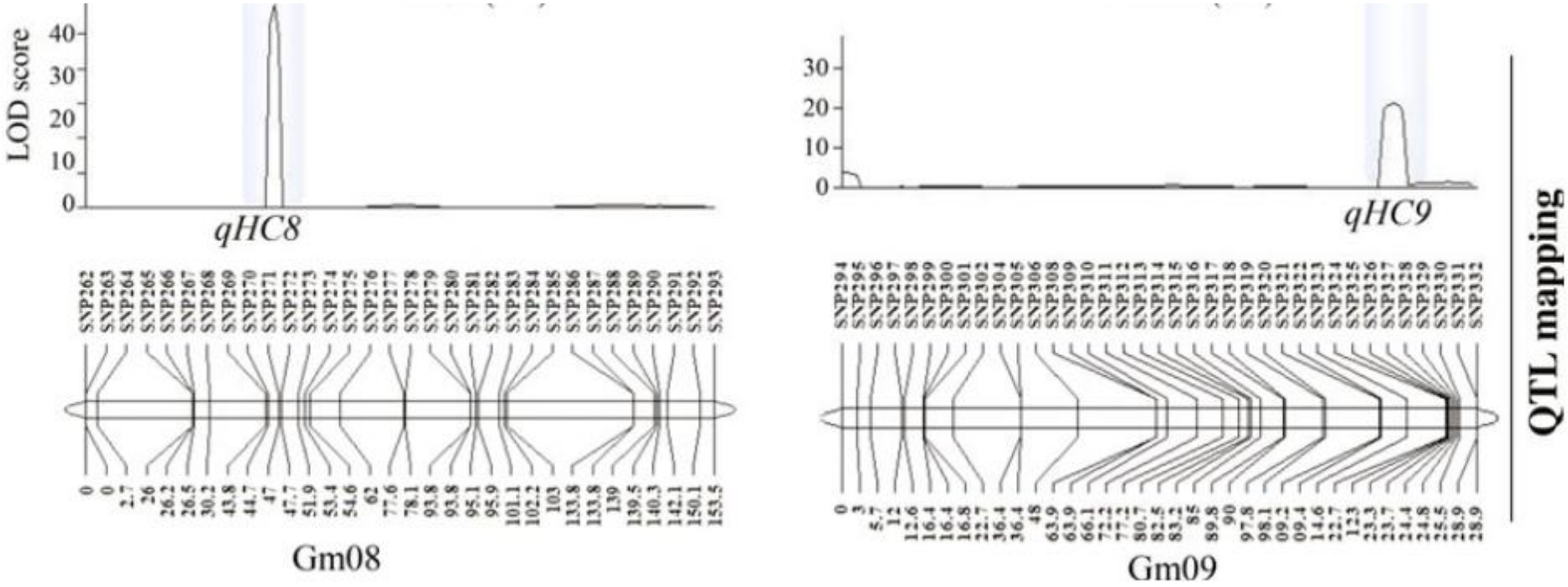

b
Eigenvector2 1.79 % of variance
0.10
0.05
0.00
−0.05
−0.10
−0.025
0.000
0.025
0.050
Eigenvector1 5.26 % of variance

重测序群体结构分析

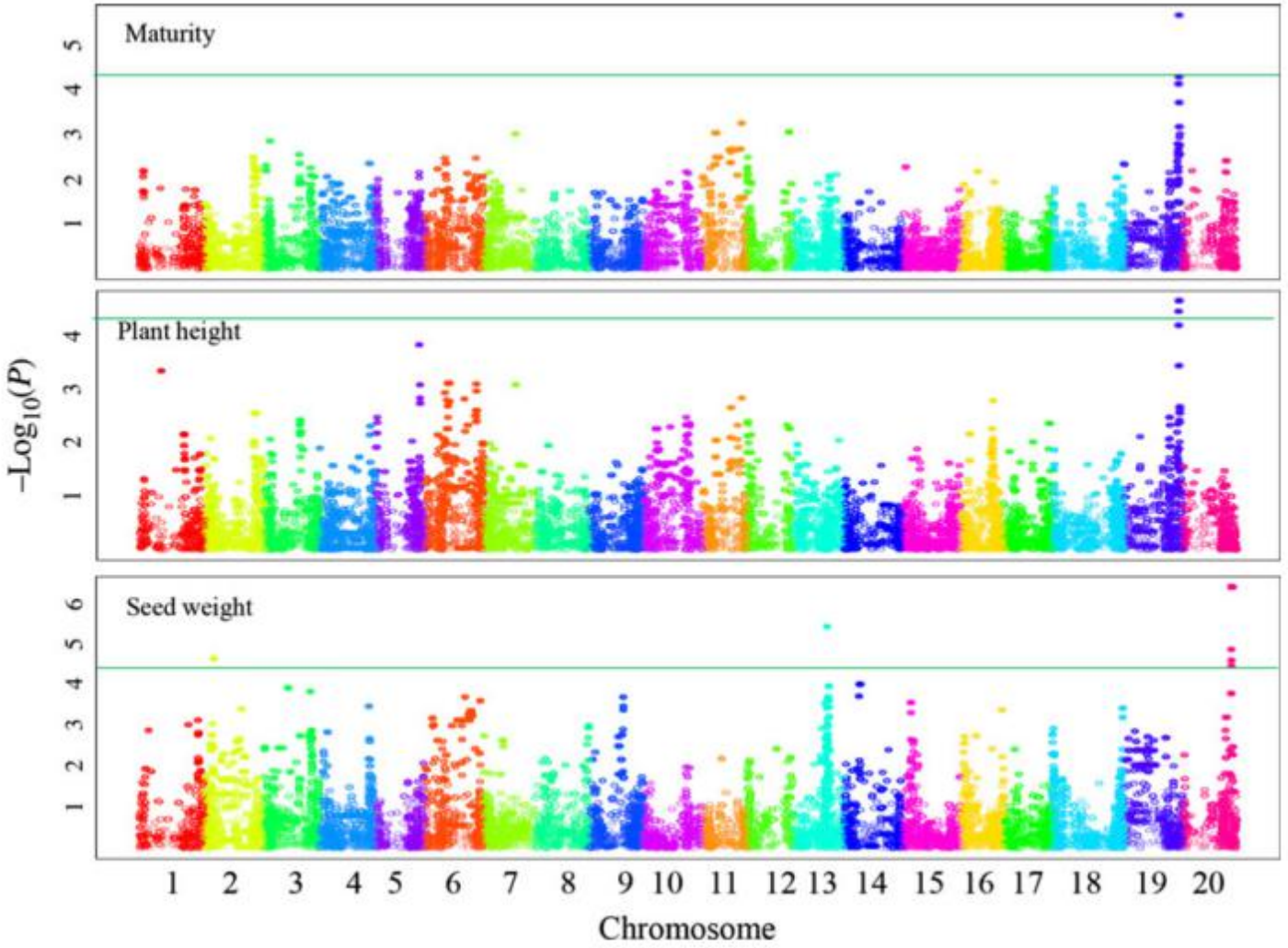

k = 3
East Asian
Eurasian
Indian and Xishuangbanna
k = 4
East Asian
Eurasian
Xishuangbanna
Indian
k = 5
East Asian
Eurasian
Xishuangbanna
Indian 2
Indian 1

构建遗传图谱和QTL分析

- Joinmap + winQTLcart



GWAS关联分析


Maturity
Plant height
Seed weight
−Log10(P)
1 2 3 4 5
1 2 3 4
1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Chromosome

谢谢大家！