243.STAMP图形界面微生物组分析软件

赵丹阳，中国药科大学

2020-09-07

Table of Contents

### 243.图形界面微生物组分析软件STAMP：使用说明与实例展示

本节作者：赵丹阳，中国药科大学

版本1.0.4，更新日期：2020年9月7日

本项目永久地址： <https://github.com/YongxinLiu/MicrobiomeStatPlot> ，本节目录 243STAMP，包含R markdown(\*.Rmd)、Word(\*.docx)文档、测试数据和结果图表，欢迎广大同行帮忙审核校对、并提修改意见。提交反馈的三种方式：1. 公众号文章下方留言；2. 下载Word文档使用审阅模式修改和批注后，发送至微信(meta-genomics)或邮件([metagenome@126.com](mailto:metagenome@126.com))；3. 在Github中的Rmd文档直接修改并提交Issue。审稿人请在创作者登记表 <https://www.kdocs.cn/l/c7CGfv9Xc> 中记录个人信息、时间和贡献，以免专著发表时遗漏。

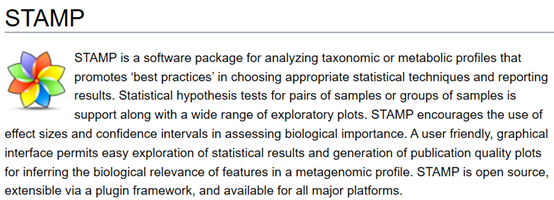
### STAMP图形界面的微生物组分析软件：使用说明与实例展示

关键字： 微生物组 统计分析 可视化 STAMP

#### 背景介绍

##### STAMP简介

STAMP是一款分析微生物物种与功能组成的可视化软件，STAMP 1.0于2010年发表在Bioinformatics杂志，后期2014年的2.0版本同样在Bioinformatics发布，目前最新版本为2.1.3。截止到2020年8月15日，两个版本STAMP的引用次数分别达到了719次和1390次。该软件除了能够绘制探索性数据分析的降维、相关图之外，还提供了假设检验的差异比较统计分析功能。此外，STAMP采用了图形化界面，对用户比较友好。



##### STAMP基本设置

* STAMP的输入文件

STAMP允许导入制表符分隔（tab-seprarated）的文件，也可以与主流生信软件如QIIME、Mothur等对接（通过Fife–Create profile from…实现）。文件包含层级注释组成表和样本信息表两部分，文件第一行为表头，含有**注释信息的列应当是从最高级到最低级排列，且必须形成严格的树型结构**。鉴于目前很多的分类分级系统（包括GreenGenes和SILVA等流行的分类法）的标签错误以及其他一些问题，STAMP网站提供了checkHierarchy.py脚本，可用于识别STAMP配置文件当中所有的非层级条目。而对于未知的条目，应记为unclassified（不区分大小写）。STAMP对于读取计数的形式没有特殊要求，可以为整数或任何实数，这使得标准化的方法可以不止一种。考虑到生物学数据低准确度、低精密度的特点，对于样本数量，STAMP的作者没有建议最小的样本数量，具体的数量应当由样品本身决定，但如需进行假设检验则必须符合相应的数据分布。

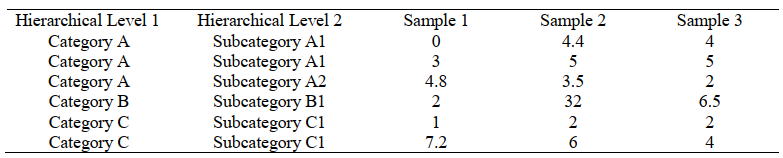


图. 输入文件1.层级物种或功能组成表

STAMP允许通过元数据（metadata）文件定义与样本相关联的其他数据。这一文件也应当是制表符分隔的文件。该文件的第一列表示每个样品的名称，并与STAMP配置文件中的样本名称一一对应，其他列可以指定为与该样本相关的任何其他数据。

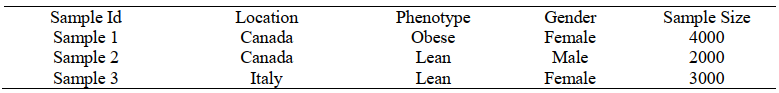
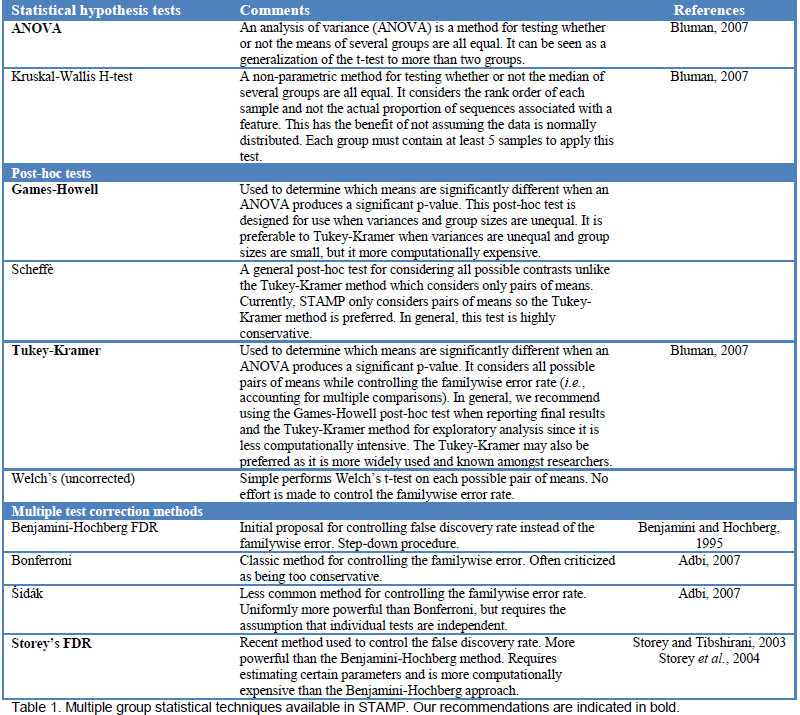


图. 输入文件2.样本元数据

* STAMP的假设检验

关于假设检验，STAMP提供了对多组、两组和两样品的统计检验方式，以及与之相应的事后检验 (Post-hoc test) 、置信区间和多重检验等。对于多组、两组以及两样品的假设检验方法分别如下面表1、表2和表3所示。对于多组样品，作者推荐使用ANOVA进行假设检验，两组样品则建议使用Welch’s t-test这一适用性更广泛的检验方式，同时建议使用Fisher精确检验应对两样品比较的情况。多重检验校正方面，可以选择传统的Benjamini-Hochberg方法，但作者更偏向使用Storey’s FDR。这一方法的计算量更大，效果较Benjamini-Hochberg也更好。



统计假设

|  |  |
| --- | --- |
| 方法 | 描述 |
| ANOVA | 方差分析(analysis of variance)的缩写，用于检验多组均值是否相等的方法。可被认为是可分析多组的t-test |
| Kruskal-Wallis H-test | 无参数的秩合检验方法，检验多组的中位数是否相等。它考虑样品排序位置而不是真实数值或比例。它不基于数据是正态分布的前提。此方法要求每组至少5个样本。 |

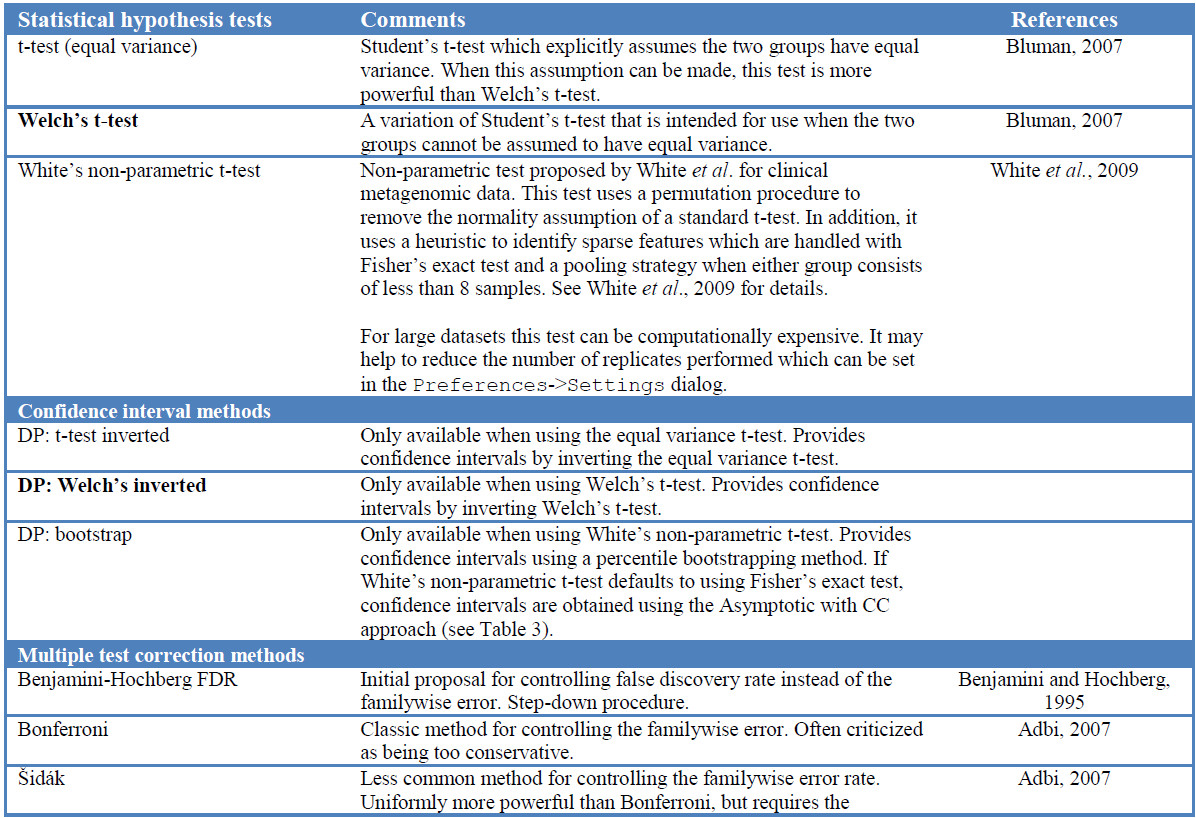
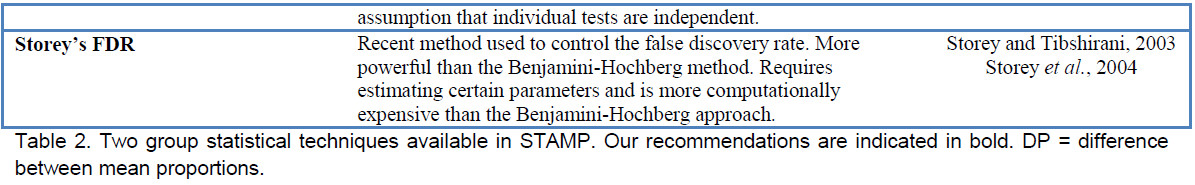
事后检验

|  |  |
| --- | --- |
| 方法 | 描述 |
| **Games-Howell** | 当ANOVA产生了显著P值后，检验具体哪两个均值显著不同。用于组样本和方差不同。当方差不同，组样本量小时推荐使用Tukey-Kramer方法 |
| Scheffe | 考虑所有可能的比较，而Tukey-Kramer只考虑成对均值。此种方法较保守 |
| **Tukey-Kramer** | 用于ANOVA显著后进一步成对比较。考虑所有可能的均值队，并考虑多次比较的错误率控制。推荐使用Games-Howell输出最终结果，而Tukey-Kramer用于探索分析。推荐此方法的另一个原因是此法使用广泛，被研究者所熟知。 |
| Welch’s(uncorrected) | 只是成队均值比较，但不进行多次比较的错误率控制 |

多重检验校正

|  |  |
| --- | --- |
| 方法 | 描述 |
| Benjamini-Hochberg FDR | 控制假阳性率FDR |
| Bonferroni | 控制整体错误率的经典方法，被批评太保守 |
| Sidak | 在整体错误率控制中使用不多，但均匀分布数据上比Bonferroni更强，但需要假设个体检验是独立的 |
| **Storey’s FDR** | 控制FDR的新方法，比BH更强。需要估计一些参数和更多的计算资源。 |

表1：STAMP提供的对于多组样本的假设检验、事后检验与多重校正方法。其中加粗为推荐方法，翻译自STAMP 2.1.3帮助文档第14页。

统计假设

|  |  |
| --- | --- |
| 方法 | 描述 |
| t检验 | T检验，亦称student t检验（Student’s t test），假设两组有相同的方差，当假设成立时，它比Welch’s检验更强，主要用于样本含量较小（例如n<30），总体标准差σ未知的正态分布。 |
| **Welch’s t-test** | t-test的一种变形，用于当两组无法满足方差相同的假设时使用。 |
| White’s无参t-test | 无参数的检验，由White为临床宏基因组数据分析提出。此方法使用排序过程移除标准t-test的正态假设。此外，它使用启法式鉴定松散的特征，可采用Fisher精确检验和pooling的策略，适合组样本一致，或小于8个样品。大数据集计算耗时。 |

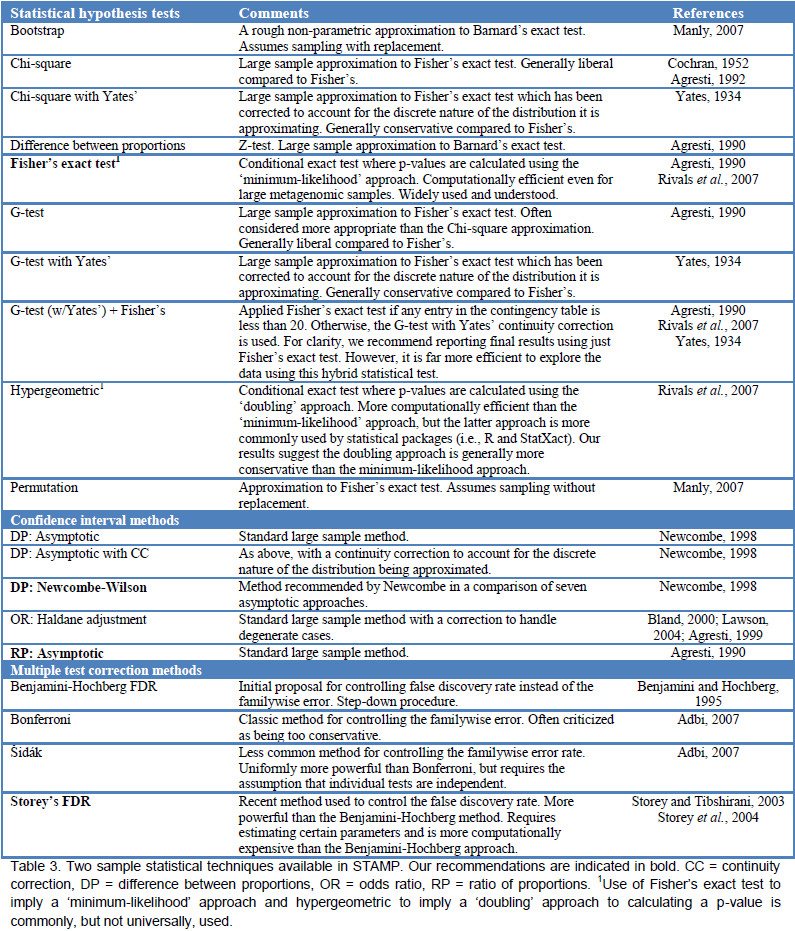
置信区间

|  |  |
| --- | --- |
| 方法 | 描述 |
| DP: t-test inverted | 只有当方差相等的t检验可用。 |
| Scheffe | 考虑所有可能的比较，而Tukey-Kramer只考虑成对均值。此种方法较保守 |
| **DP: Welch’s inverted** | 为Welch’s t检验提供置信区间。 |
| DP: bootstrap | 适合White’s 无参t-test |

多种检验校正

|  |  |
| --- | --- |
| 方法 | 描述 |
| Benjamini-Hochberg FDR | 控制假阳性率FDR |
| Bonferroni | 控制整体错误率的经典方法，被批评太保守 |
| Sidak | 在整体错误率控制中使用不多，但均匀分布数据上比Bonferroni更强，但需要假设个体检验是独立的 |
| **Storey’s FDR** | 控制FDR的新方法，比BH更强。需要估计一些参数和更多的计算资源。 |

表2：STAMP提供的对于两组样本的假设检验、置信区间与多重校正方法。其中加粗为推荐方法，翻译自STAMP 2.1.3帮助文档第17页。



统计假设

|  |  |
| --- | --- |
| 方法 | 描述 |
| Bootstrap | 一种无参方法，与Barnard精确检验相似，假设放回抽样 |
| 卡方Chi-squre | 大样本与Fisher精确检验类似，但更自由 |
| Yates卡方 | 在卡方基础上考虑了分布，比Fisher更保守 |
| **Fisher精确检验** | 条件精确检验，*P*值采用最大似然方法。宏基因组大数据样本计算速度快，应用广泛且公众认可 |
| G-test | 大样本与Fisher近似，比卡方更合适，比Fisher更灵活 |
| G-test with Yates’ | 大样本与Fisher类似，考虑自然离散校正，比Fisher更保守 |
| G-test(w/Yates’)+Fisher’s | 当列联表中小于20使用Fisher精确检验，其它使用G-test。为了结果清楚，我们推荐只使用Fisher精确检验。而在探索数据阶段，使用混合的统计方法可能更有效 |
| 超几何分布 | P值使用两种方法的条件精确检验。比最小似然法(在R和StatXact中常用)更快。但更保守。 |
| 置换 | 与Fisher类似，假定无放回抽样 |

置信区间

|  |  |
| --- | --- |
| 方法 | 描述 |
| DP: 渐近 | 标准的大样本方法 |
| Scheffe | 考虑所有可能的比较，而Tukey-Kramer只考虑成对均值。此种方法较保守 |
| DP: CC渐近 | 考虑自然离散分布和连续校正 |
| DP: Newcombe-Wilson | Newcombe推荐的7种渐近方法中最优的 |
| OR: Haldane adjustmet | 大样本方法结合校正解决退化问题 |
| RP: 渐近 | 标准的大样本方法 |

多重检验校正

|  |  |
| --- | --- |
| 方法 | 描述 |
| Benjamini-Hochberg FDR | 控制假阳性率FDR |
| Bonferroni | 控制整体错误率的经典方法，被批评太保守 |
| Sidak | 在整体错误率控制中使用不多，但均匀分布数据上比Bonferroni更强，但需要假设个体检验是独立的 |
| **Storey’s FDR** | 控制FDR的新方法，比BH更强。需要估计一些参数和更多的计算资源。 |

表3：STAMP提供的对于两样品统计检验的情况所应用的假设检验、置信区间与多重检验校正方式。推荐方法加粗。CC = 连续校正，DP = 比例差异，OR = 让步比，RP = 比例。翻译自STAMP 2.1.3帮助文档第19页。

#### 实例解读

##### 例1：扩展柱状图展示组间差异功能

本示例来自Raju等人2020年8月在Microbiome发表的文章，报导了生命的前十年使用抗菌药物会影响唾液微生物群的多样性和组成，而且男女有别。其中图4以STAMP分析和可视化以不同频次使用抗菌药物的两组儿童唾液微生物的预测功能的变化。

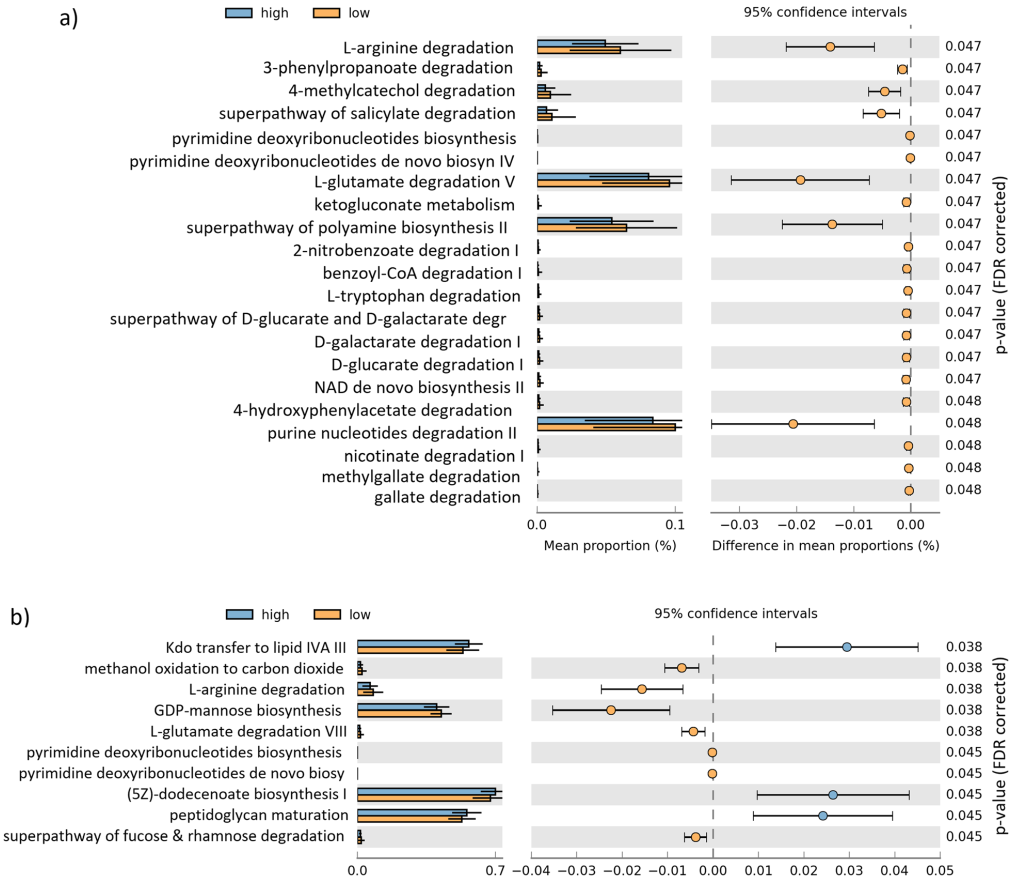


图4. 在以低频次与高频次使用a)全种类抗菌药物与b)阿奇霉素的两组儿童的唾液微生物的MetaCyc功能预测。柱状图显示以PICRUST2预测的差异性MetaCyc通路的平均占比。组间差异显示95%的置信区间，并只显示Welch’s t-test经FDR校正后q value < 0.05的部分。

Fig. 4 Functionally predicted MetaCyc pathways differing in proportions in high and low user groups of a) all AMs and in b) azithromycin. The bar plot shows mean proportions of differential MetaCyc pathways predicted using PICRUSt2. The difference in proportions between the groups is shown with 95% confidence intervals. Only p value < 0.05 (Welch’s t test, FDR adjusted), are shown and composition）

* 结果

在低频次和高频次使用全种类抗菌药物的儿童当中，功能预测鉴定出21个显著差异的metaCyc通路（图4a）。这些通路在低频组中占比更高。差异最大的通路包括了L-精氨酸降解、L-谷氨酸降解Ⅴ、多胺生物合成Ⅱ超通路以及嘌呤核苷酸降解Ⅱ。在低、高频次使用阿奇霉素的两组中，一共有十个差异通路 (图4b)。甲醇氧化至一氧化碳、L-精氨酸降解以及GDP-甘露糖生物合成通路在阿奇霉素的低频使用组中占比较高，而Kdo转移至脂质ⅣAⅢ、(5Z)-十二碳烯酸酯生物合成通路在高频使用组中占比更高。。

Functional predictions identified 21 differentially present metaCyc pathways between the low and high AM users when all AM use were combined (Fig. 4a). All of the pathways had higher proportions in the low AM use group. The largest significant differences were pathways for L-arginine degradation, L-glutamate degradation V, superpathway of polyamine biosynthesis II and purine nucleotides degradation II. Ten pathways differed between low and high azithromycin use (Fig. 4b). Methanol oxidation to carbon monoxide pathway, L-arginine degradation and GDP-mannose biosynthesis pathways showed higher proportions in the low azithromycin group, while Kdo transfer to lipid IVA III, (5Z)-dodecenoate biosynthesis and peptidoglycan maturation pathways showed higher proportions in the high azithromycin group.

##### 例2. 柱状图和PCA散点图

本文于2019年8月发表于AEM杂志，报导了不同大肠杆菌病原体引起的肠道感染的宏基因组学特征，揭示了不同致病性大肠杆菌感染造成的肠道微生物物种差异。以图5为例讲解STAMP结果中组间单菌差异柱状图、主成分分析散点图的描述。

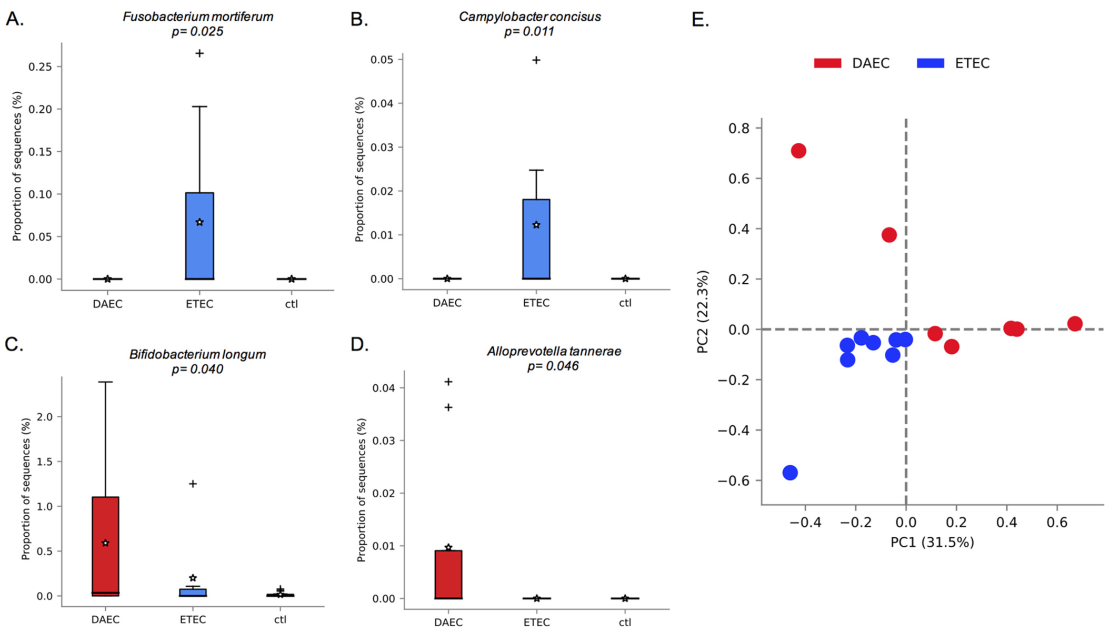


图5. 黏附性弥散型大肠杆菌（DAEC）与肠毒性大肠杆菌（ETEC）感染中的丰度差异性物种。差异性物种的筛选条件为校正后p值小于等于0.05并且效应量（即组间差异大小）为0.8。（A和B）分别表示宏基因组分析注释为死亡梭杆菌和简明弯曲菌的序列所占百分比，(C和D)则分别为长双歧杆菌和坦纳拟普雷沃菌的。（E）为去除宿主与大肠杆菌的序列之后，根据宏基因组确定的分类组成（由MetaPhlAn2根据进化分支特异性标记基因注释到物种水平）所构建的PCA图。

FIG 5 Differentially abundant (diagnostic) taxa between DAEC and ETEC infections. Differentially abundant species were reported if they had a corrected P value of ≤ 0.05 and an effect size (the magnitude of the difference between groups) of 0.8. (A and B) Proportions of metagenomic sequences assigned to Fusobacterium mortiferum and Campylobacter concisus, respectively. (C and D) Proportions of sequences assigned to Bifidobacterium longum and Alloprevotella tannerae, respectively. (E) PCA plot based on the taxonomic composition of each metagenome (annotated at the species level using clade-specific marker genes with MetaPhlAn2) after removal of human and E. coli reads from the libraries.

* 结果

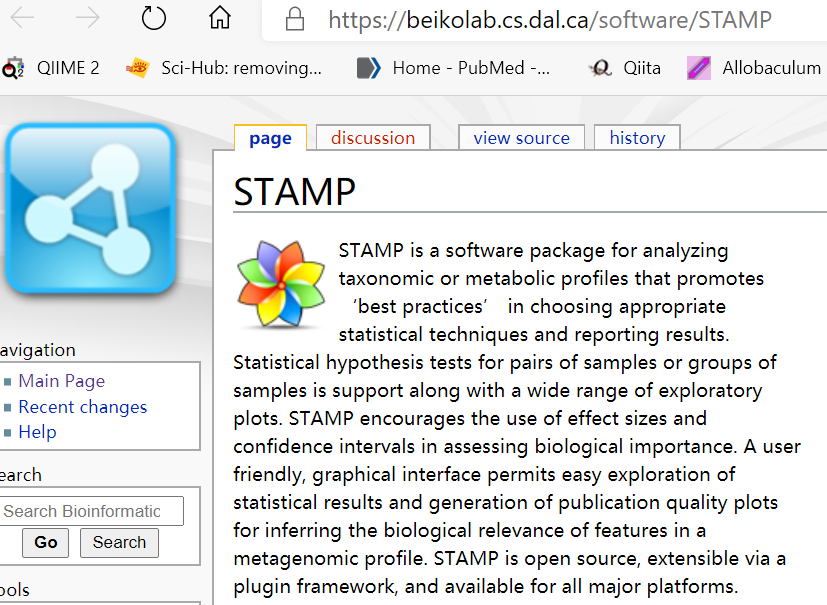
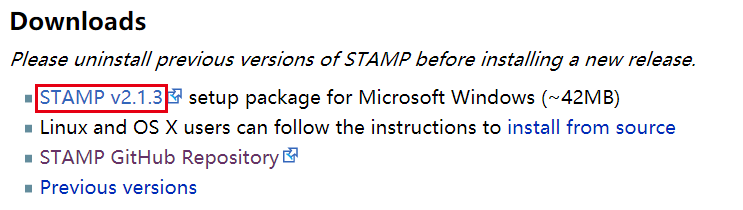
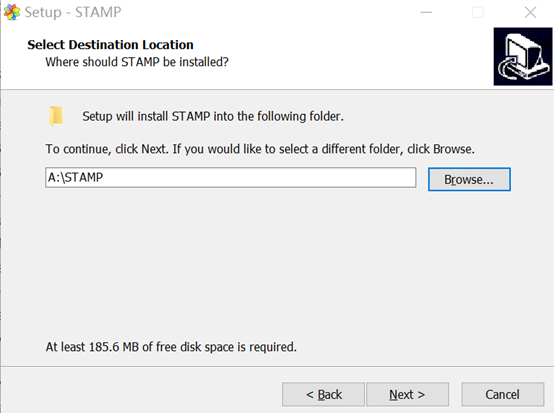
对于DAEC和ETEC感染，在最初的物种注释当中至少有四个物种出现了差异。其中，死亡梭杆菌(*P* = 0.025)和简明弯曲菌(*P* = 0.011)在ETEC感染组显著富集，而长双歧杆菌(P = 0.040)和坦纳拟普雷沃菌(P = 0.046)在DAEC感染组丰度显著上升。基于物种水平的分类组成的PCA图显示ETEC感染的样品更相似，而DAEC组的样品则显示了更强的多样性。

The initial taxonomic characterization revealed at least four species that were discriminatory of DAEC versus ETEC infections. Specifically, Fusobacterium mortiferum (P = 0.025) and Campylobacter concisus (P = 0.011) were significantly more abundant in ETEC infections (Fig. 5A and B), while Bifidobacterium longum (P = 0.040) and Alloprevotella tannerae (P = 0.046) were significantly more enriched in DAEC infections (Fig. 5C and D). A PCA based on taxonomic composition at the species level also revealed that metagenomes associated with ETEC infections tended to be taxonomically more similar among themselves, whereas DAEC samples showed more diversity.

#### 分析实战

##### 下载并安装软件

在浏览器地址栏输入https://beikolab.cs.dal.ca/software/STAMP ，在Downloads当中找到并点击STAMP v2.1.3 下载链接，保存安装程序。下载之后打开安装程序并选择路径进行安装，**注意安装路径不得含有中文字符**。

##### 数据选择

这里选取STAMP安装路径中的肠型数据(如Windows下为C:Files (x86))，Enterotypes.profile.spf为制表符分隔的特征表，由门（Phyla）和属（Genera）两个分类层级构成；Enterotypes.metadata. tsv为tsv格式的元数据，由样本编号、肠型、国籍等信息组成。部分注释信息和样品元数据分别如下图所示。

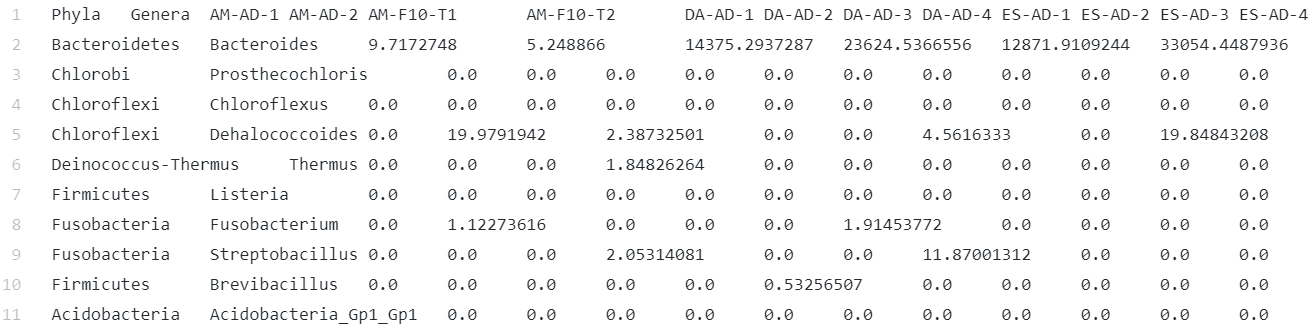


图. spf格式特征表截图。为支持层级的制表符分隔特征表，层级可以为1级，也可以为多级。

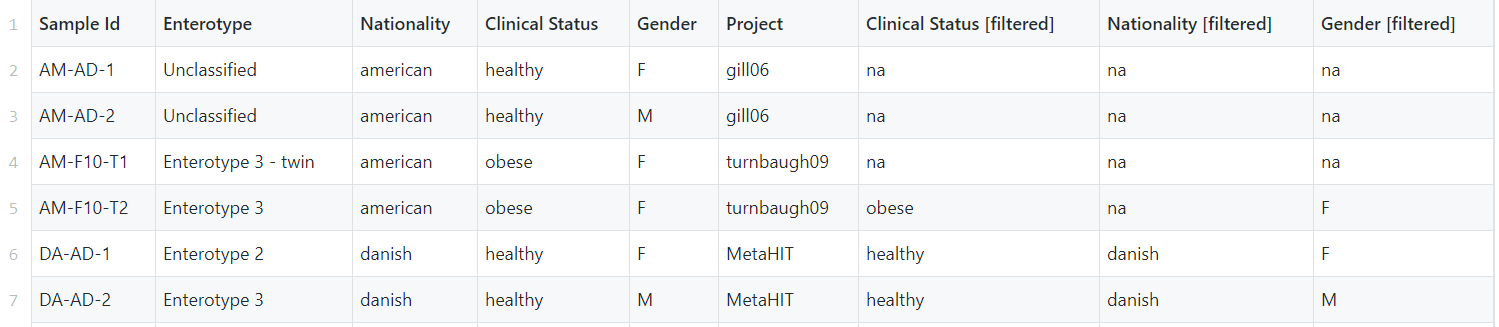
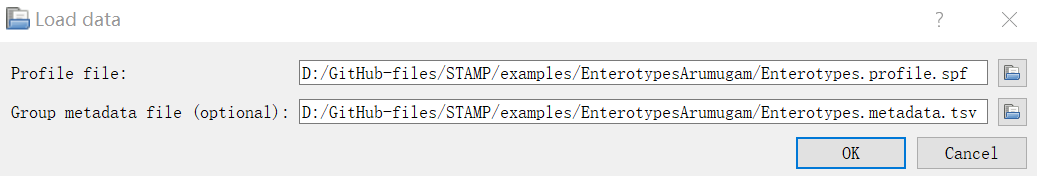


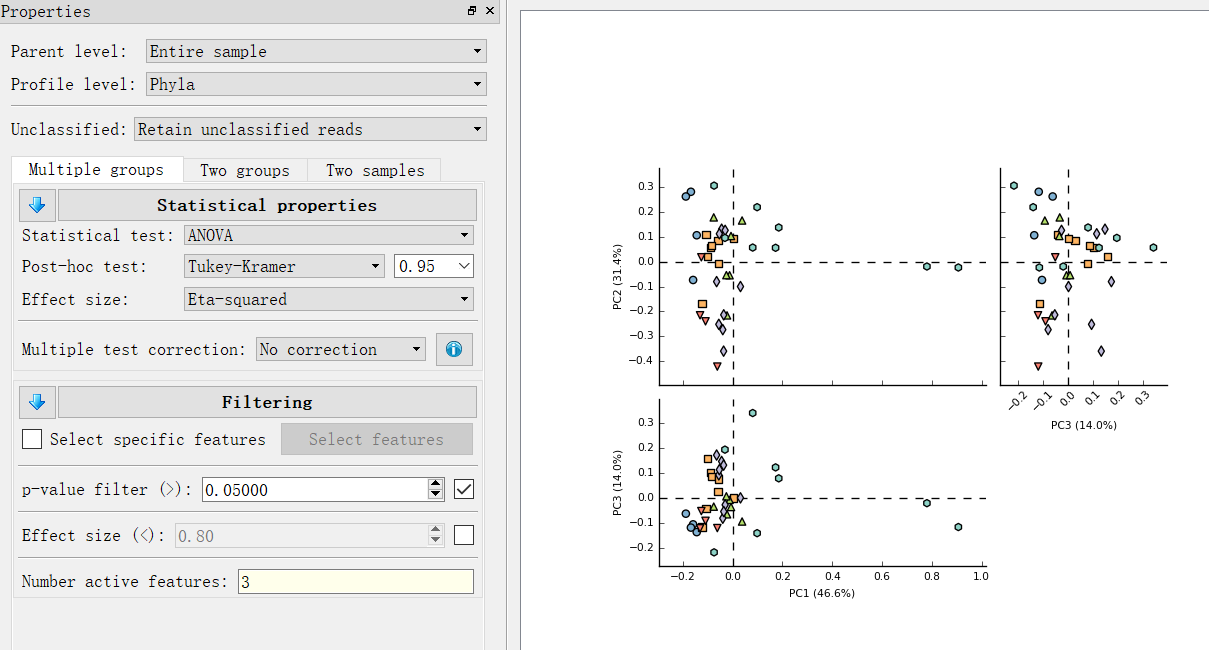
图. 元数据格式预览。即样本的分组或属性信息。

##### 多组比较

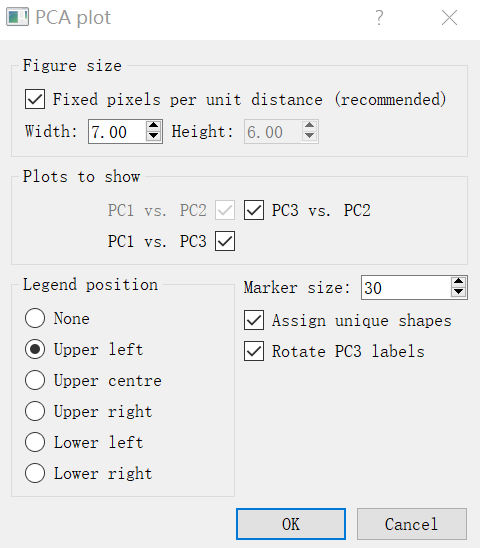
安装之后，打开STAMP，点击左上角的“file”-“load data”，分别导入Enterotypes.profile.spf和Enterotypes.metadata.tsv。

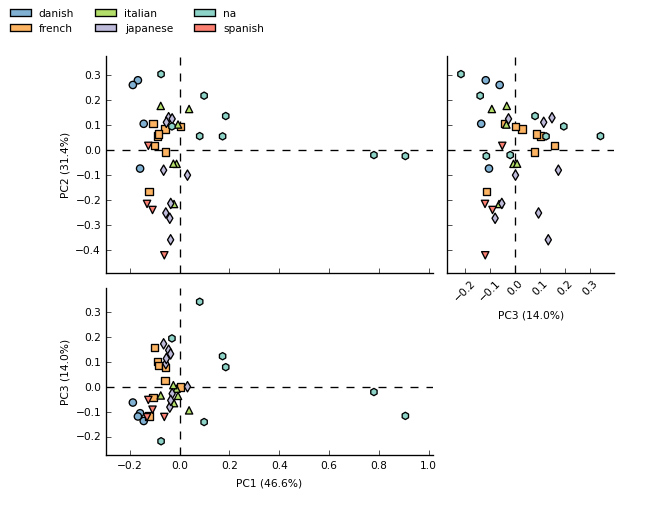


导入之后默认显示PCA结果，以散点图的形式展示门水平(注释层级的最高级)的差异：

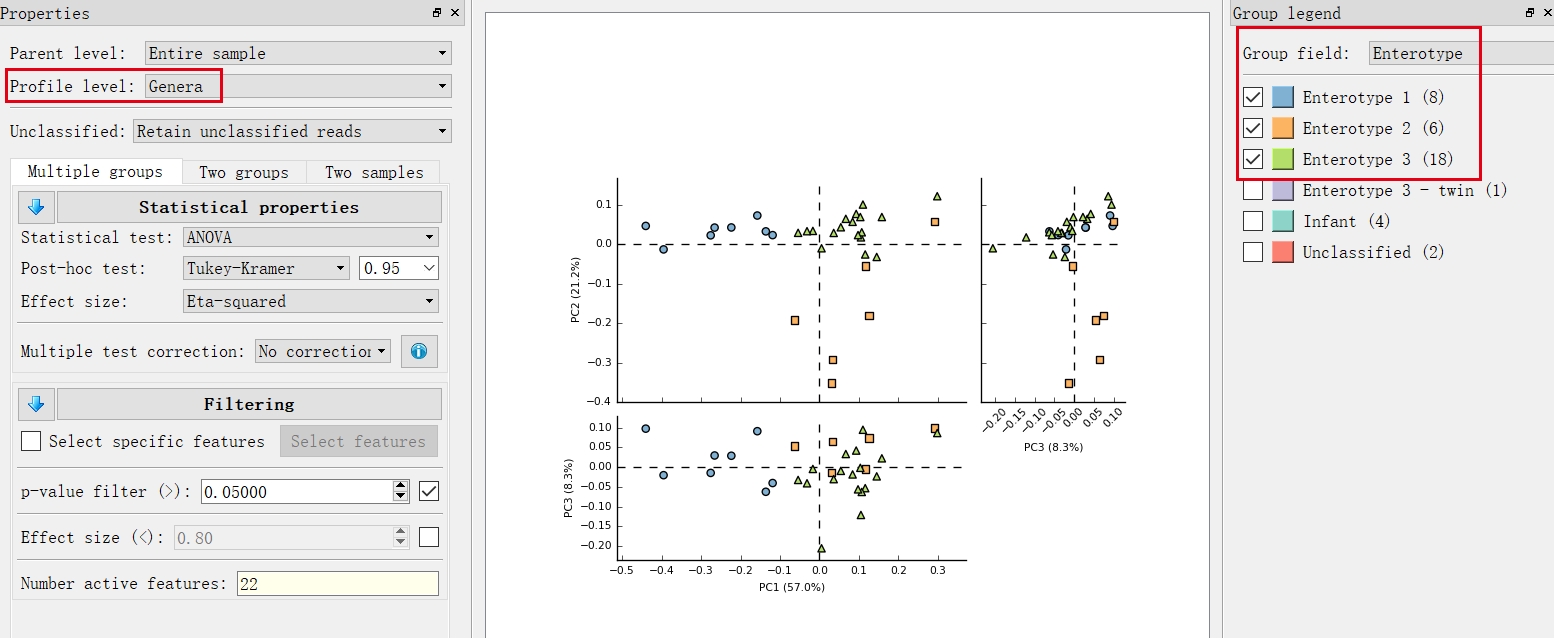


点击“Configure plot”，设置图例位置于图像左上角，也可点击“View”-“Group legend” 查看分组信息。





在右上‘Group field’选项当中重新分组，选择‘Enterotype’，并去除后三个非主要肠型，仅保留三种肠型。同时，更改左上“Profile level”为‘Genera’可以看到三种肠型在PCA图中分开较为明显。

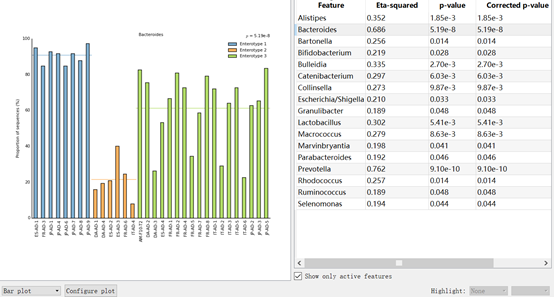


##### 切换图表类型

STAMP允许两组或多组样品以及两个样品之间的比较，支持的可视化类型除了PCA图之外还有：

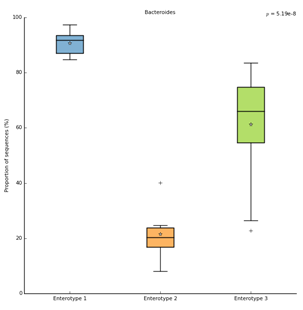
1. 柱状图

显示每个样品特征(feature)的相对比例或序列数目（通过Configure plot设置），并添加组均值，图示为三种肠型当中拟杆菌属的相对丰度图。



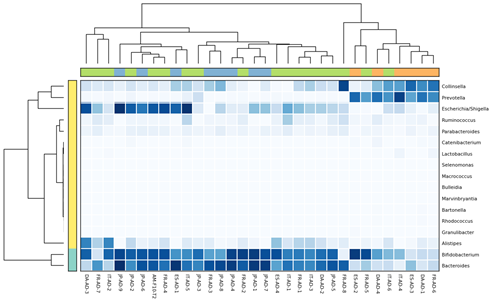
1. 箱线图

快速查看各组组内数据分布的基本情况，可通过’Show only active features’查看符合阈值的特征。



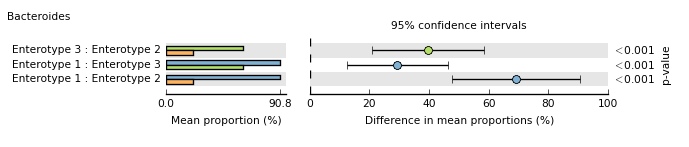
1. 热图

显示每个特征在样品中丰度的比例，不仅显示所有样本的丰度值，还可以对行与列的各单元进行聚类显示之间的关系。通过选择’Show only active features’，可以看到三种肠型的样品有部分聚到一起，和PCA的结果较为接近。



1. Post-hoc 图

在对于三组及三组以上的多组统计检验当中，零假设（即无效假设）为各组之间均值相等。在选择这种情况的时候，只能得知各组均值是否是否一致，而不能看到任意两组之间的均值是否一致。STAMP的post-hoc检验提供了四种检验方式（具体检验方式见表1，默认为Tukey-Kramer），其结果以post-hoc图的形式展出。用户可在右侧feature table当中选择感兴趣的feature进行展示，展示的结果仅包含有p值小于给定阈值的部分。图示为Bacteroides的post-hoc检验结果，可以看到Bacteroides在三组当中的均值两两之间在95%的置信区间上都不一致（p<0.001），且Enterotype 1>Enterotype 3>Enteotype 2 。



以上内容带大家熟悉了STAMP的基本使用，其他两组比较、两样本比较的结果和图也是类似的。

更多内容，可以学习宏基因组公众号之前发布的教程：

* STAMP——微生物组间差异统计分析 [简明教程](http://mp.weixin.qq.com/s/f02mPAXCobPQijkIis-c2w) [中文帮助文档](https://mp.weixin.qq.com/s/uWU-F7jUuMskM7GXOrgILA)

#### 参考文献

* STAMP. <https://beikolab.cs.dal.ca/software/STAMP>
* STAMP User’s Guide. <https://beikolab.cs.dal.ca/software/images/c/cd/STAMP_Users_Guide.zip>
* STAMP：扩增子、宏基因组统计分析神器(中文帮助文档). <https://blog.csdn.net/woodcorpse/article/details/80458077>
* 差异分析工具STAMP手册2:使用手册（汉化版）. <https://www.jianshu.com/p/331b6796f8ff>
* Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014;30(21):3123-3124. <doi:10.1093/bioinformatics/btu494>
* Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. Bioinformatics. 2010;26(6):715-721. <doi:10.1093/bioinformatics/btq041>
* Sajan C. Raju, Heli Viljakainen, Rejane A. O. Figueiredo, Pertti J. Neuvonen, Johan G. Eriksson, Elisabete Weiderpass & Trine B. Rounge. (2020). Antimicrobial drug use in the first decade of life influences saliva microbiota diversity and composition. Microbiome 8, 121, doi: <https://doi.org/10.1186/s40168-020-00893-y>
* ngela Peña-Gonzalez, Maria J. Soto-Girón, Shanon Smith, Jeticia Sistrunk, Lorena Montero, Maritza Páez, Estefanía Ortega, Janet K. Hatt, William Cevallos, Gabriel Trueba, Karen Levy & Konstantinos T. Konstantinidis. (2019). Metagenomic Signatures of Gut Infections Caused by Different Escherichia coli Pathotypes. Applied and Environmental Microbiology 85, e01820-01819, doi: <https://doi.org/10.1128/aem.01820-19>
* M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, H. I. T. Consortium Meta, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. Melo Minardi, C. M’Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich & P. Bork. (2011). Enterotypes of the human gut microbiome. Nature 473, 174-180, doi: <https://doi.org/10.1038/nature09944>

责编：刘永鑫 中科院遗传发育所

版本更新历史

1.0.0，2020/8/30，赵丹阳，中国药科大学，初稿

1.0.1，2020/9/3，刘永鑫，大修

1.0.2，2020/9/4，吴翔宇 宁波大学，全文校对

1.0.3，2020/9/4，刘永鑫，整合校对

1.0.4，2020/9/7，赵丹阳，中国药科大学，修post-hoc部分结果