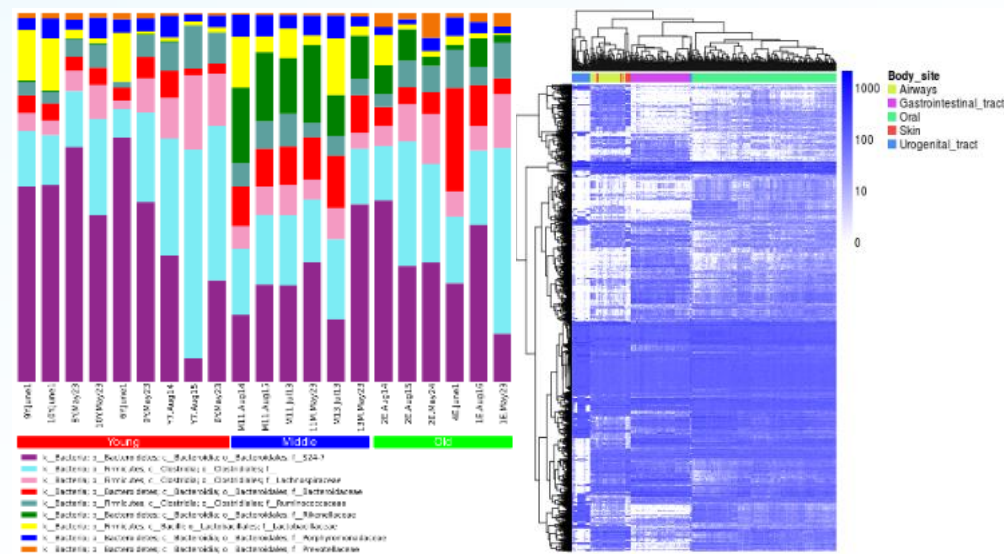




23物种和功能组成

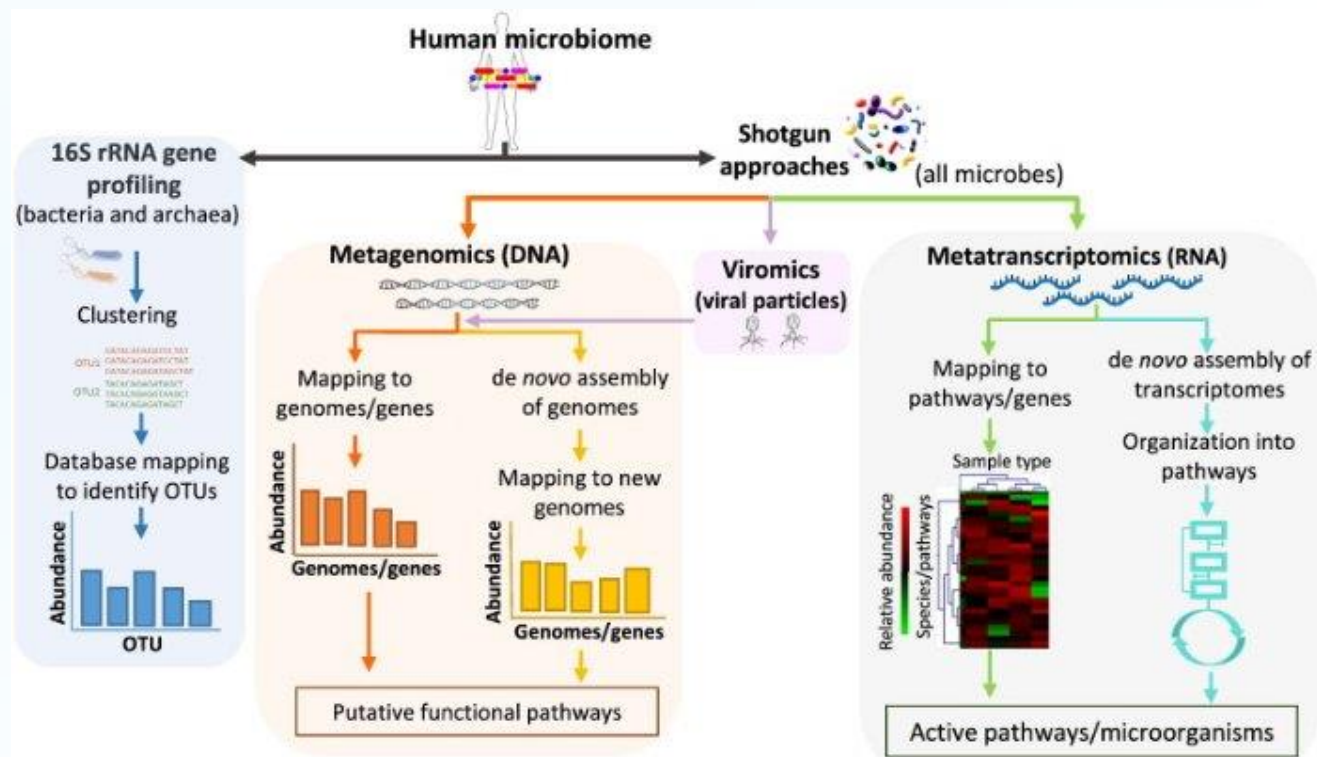
易生信
2019年6月22日



易生信，毕生缘；培训版权所有。

宏基因组有参流程——实战分析大纲

- 一. 软件安装和数据库部署
- 二. KneadData质控
- 三. **MetaPhlAn2物种组成**
- 四. HUMAnN2功能组成
- 五. GraPhlAn可视化物种
- 六. LEfSe分析物种差异
- 七. STAMP功能组成分析



易生信

意大利特伦托大学Nicola Segata组——宏基因组软件(10-12年哈佛公共卫生学院Huttenhower组博后)

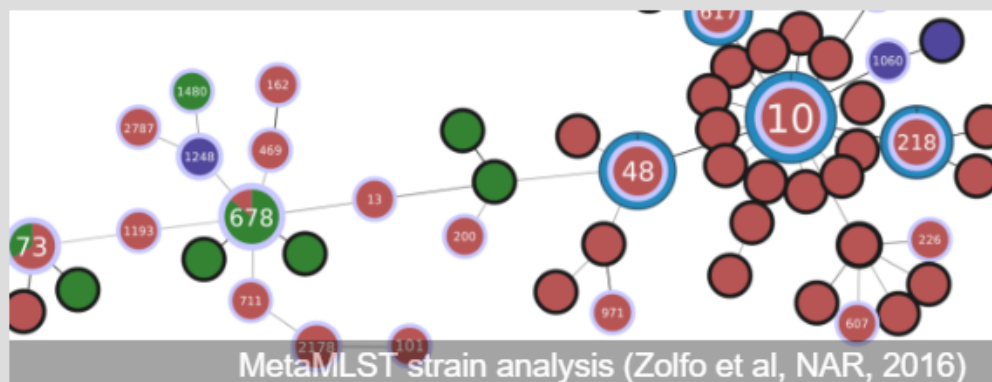
Segata lab <http://segatalab.cibio.unitn.it/>
Computational Metagenomics

Home People Publications Tools Contacts



NEW In the context of our [ERC Starting grant 2016](#) and other projects, we have **openings for four computational scientists** for human microbiome projects. Take a look at the [call for expressions of interest!](#)

Meta'omics for hacking the human microbiome



MetaMLST strain analysis (Zolfo et al, NAR, 2016)

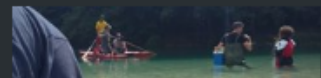


Tweets by @cibiocm

Segata Lab Retweeted

Sirio | Film Tv & Media
@siriofilm

Con i ricercatori del
@CIBIO_UniTrento per la
raccolta dei campioni
microbioma ambientale.
#Memex @RaiScuola
#luoghidellascienza con
@dcoeroborga.



开发维护和参与的软件

- [curatedMD \(2017\)](#) 微生物组物种和功能注释数据库R包
- [MetaMLST \(2016\)](#) 宏基因组多位点序列分型
- [StrainPhlAn \(2016\)](#) 菌株水平群体基因组分析
- [MetAML \(2016\)](#) 3000个宏基因组数据微生物与表型关联预测
- [PanPhlAn \(2016\)](#) 预测菌株水平基因组组成和转录活性
- [MetaPhlAn2 \(2015\)](#) 宏基因组物种组成
- [GraPhlAn \(2015\)](#) 物种或进化树圈图美化
- [ShortBRED \(2015\)](#) 蛋白归类基因家族和宏基因组功能定量
- [MicroPITA \(2014\)](#) 宏基因组样本挑选
- [MetaRef \(2014\)](#) 微生物类特异基因数据库
- [PhyloPhlAn \(2013\)](#) 新微生物基因组分类和进化关系鉴定
- [HUMAnN \(2012\)](#) 宏基因组功能组成定量
- [LEfSe \(2011\)](#) 生物标志物挖掘

• [Cell重磅发布15万人体微生物基因组！超大规模宏基因组研究揭示数千计人体微生物新物种](#)



易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

MetaPhlAn2 (2015)

- MetaPhlAn2是分析微生物群落(细菌、古菌、真核生物和病毒)组成的工具，只需一条完命令即可获得微生物的物种丰度信息。同时提供脚本可进一步统计和可视化。
- 主页：<http://segatalab.cibio.unitn.it/tools/metaphlan2/>

Metagenomic microbial community profiling using unique clade-specific marker genes

[N Segata](#), [L Waldron](#), [A Ballarini](#), [V Narasimhan](#)... - *Nature* ..., 2012 - [nature.com](#)

Metagenomic shotgun sequencing data can identify microbes populating a microbial community and their proportions, but existing taxonomic profiling methods are inefficient for increasingly large data sets. We present an approach that uses clade-specific marker genes to unambiguously assign reads to microbial clades more accurately and > 50× faster than current approaches. We validated our metagenomic phylogenetic analysis tool, MetaPhlAn2, on terabases of short reads and provide the largest metagenomic profiling to date of

★ 77 Cited by 804 Related articles All 20 versions

Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**, 811, doi:10.1038/nmeth.2066 (2012).

MetaPhlAn2 for enhanced metagenomic taxonomic profiling

[DT Truong](#), [EA Franzosa](#), [TL Tickle](#), [M Scholz](#)... - ... *methods*, 2015 - [nature.com](#)

To the Editor: MetaPhlAn (metagenomic phylogenetic analysis) 1 is a method for characterizing the taxonomic profiles of whole-metagenome shotgun (WMS) samples that has been used successfully in large-scale microbial community studies^{2, 3}. This work complements the original species-level profiling method with a system for eukaryotic and viral quantitation, strain-level identification and strain tracking. These and other extensions make the MetaPhlAn2 computational package (<http://segatalab.cibio.unitn.it> ...

☆ 77 Cited by 403 Related articles All 6 versions

Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902, doi:10.1038/nmeth.3589 (2015).

•[MetaPhlAn2一条命令获得宏基因组物种组成](#) ⁴



MetaPhlAn2的数据量和特征

MetaPhlAn2整理了超过17000个参考基因组，包括13500个细菌和古菌，3500个病毒和110种真核生物，汇编整理了100万+类群特异的标记基因，可以实现：

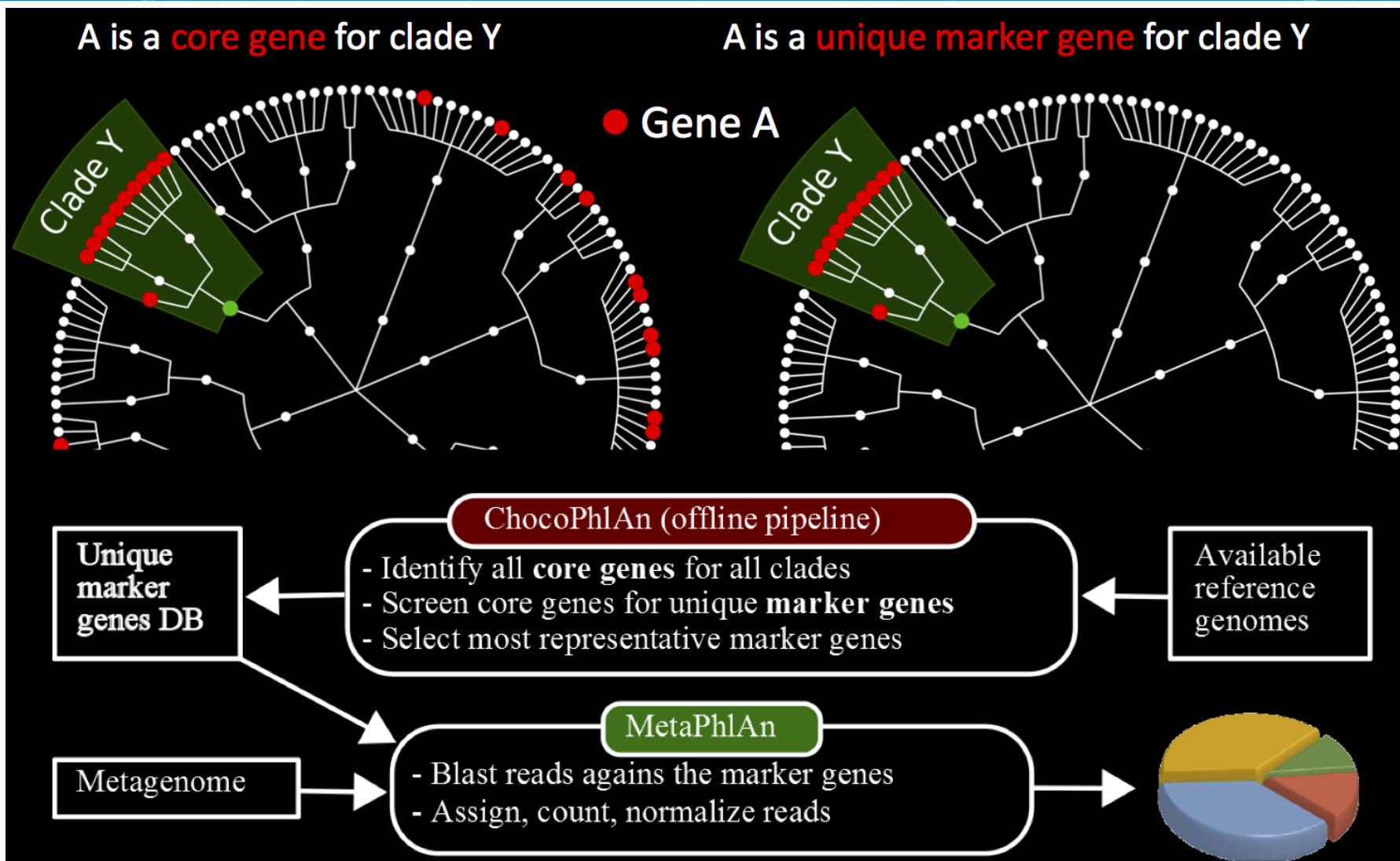
- 精确的分类群分配
- 准确估计物种的相对丰度
- 种水平精度
- 株鉴定与追踪
- 超快的分析速度

易生信 宏基因组
生信宝典



MetaPhlAn Marker的选择(核心算法)

类核心
基因

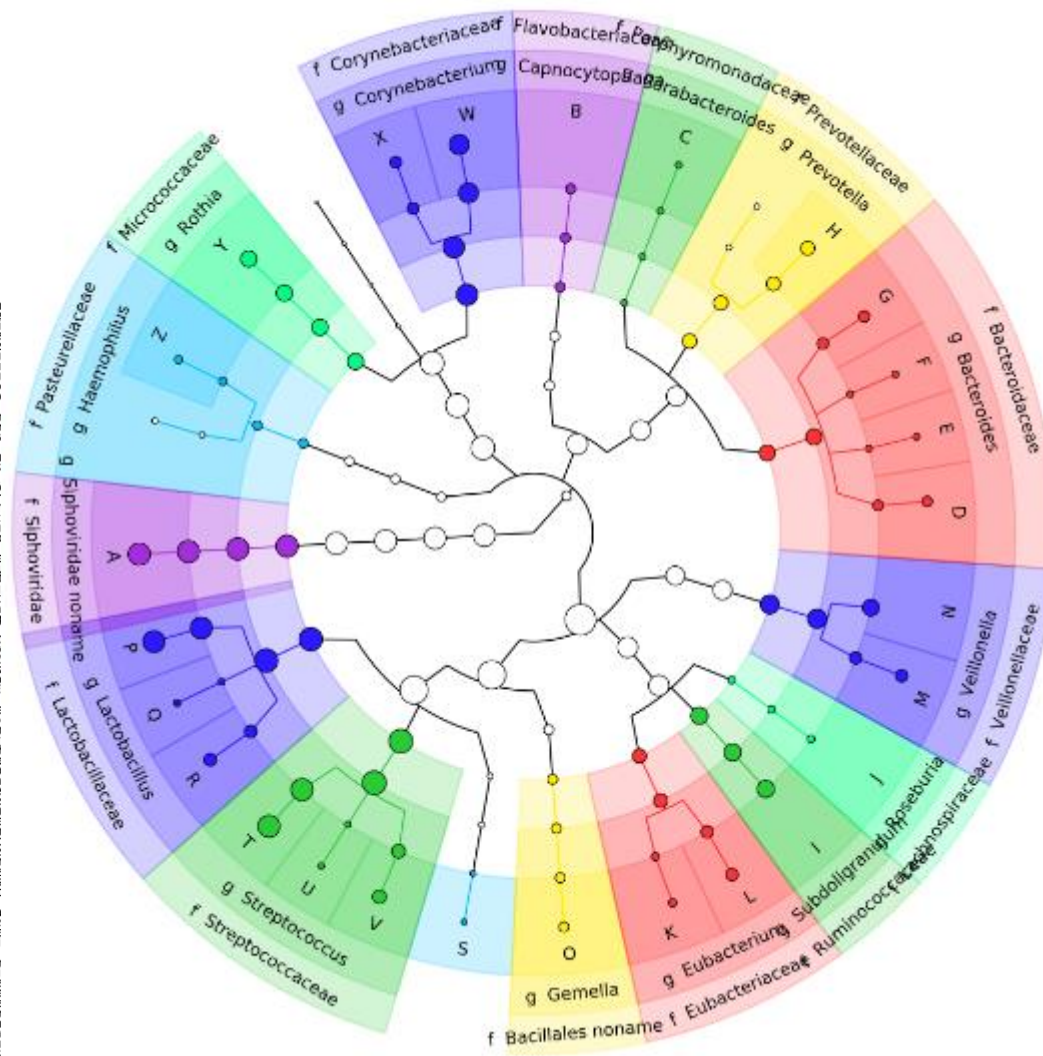
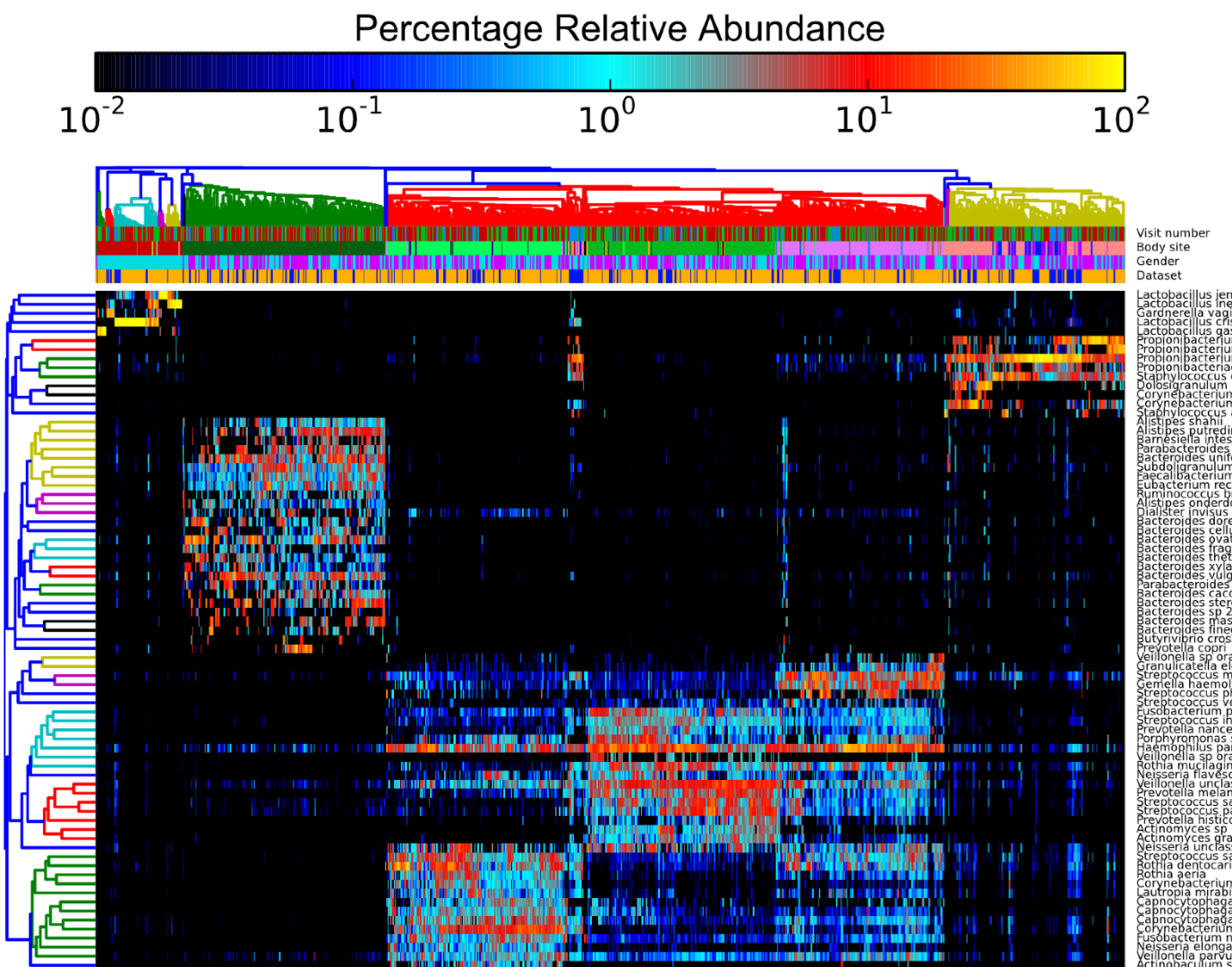


类特异
基因

泛基因组
数据库

宏基因组

MetaPhlAn2结果常用展示方式

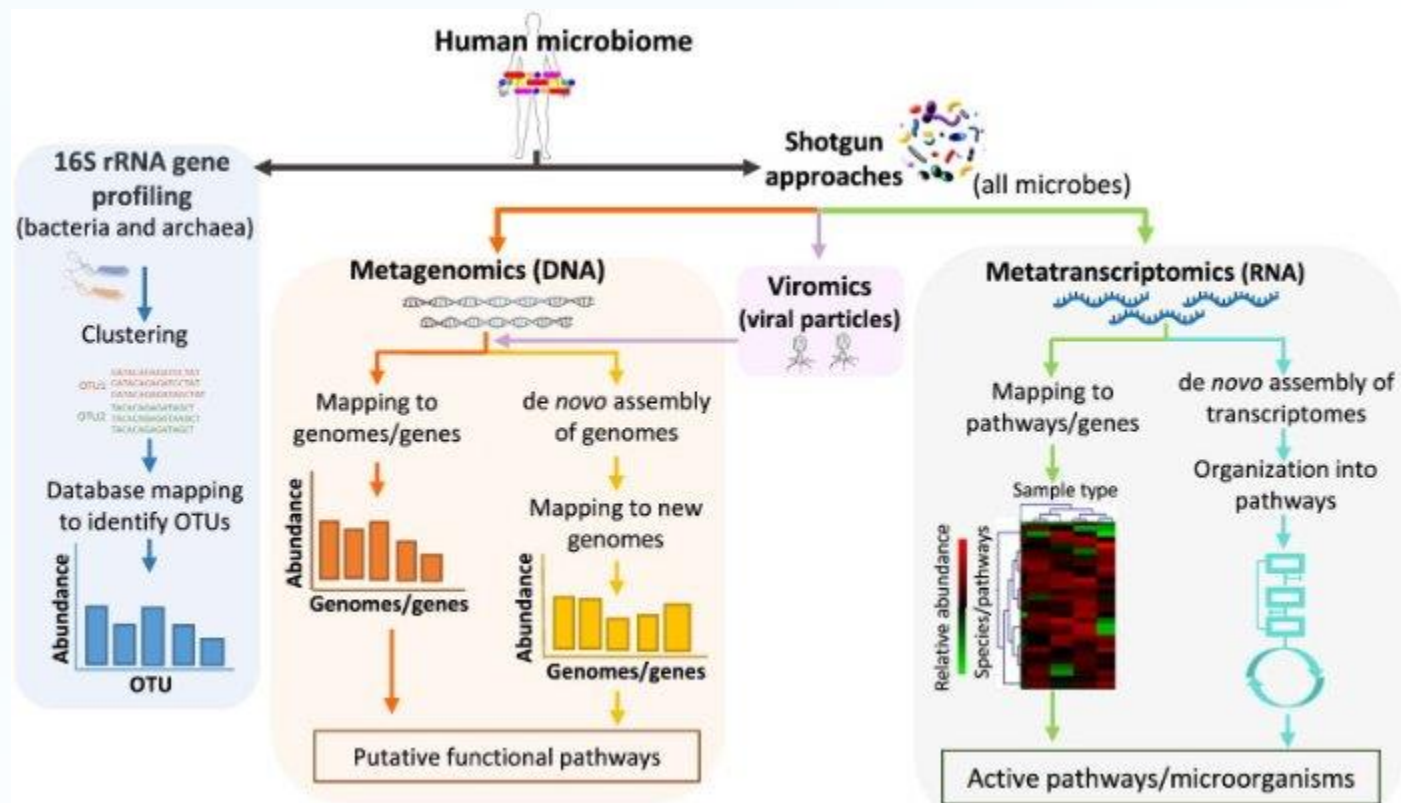


- 使用方法详见：MetaPhlAn2一条命令获得宏基因组物种组成
- 常见问题见软件安装代码soft_db.sh中human2部分或PPT附录
- 除非只关注物种组成，否则MetaPhlAn很少单独使用
- **HUMAnN2**整合了MetaPhlAn2软件，可实现一条命令完成物种、功能、以及功能对应物种组成三类文件，可多角度深入挖掘宏基因组数据。



宏基因组有参流程——实战分析大纲

- 一. 软件安装和数据库部署
- 二. KneadData质控
- 三. MetaPhlAn2物种组成
- 四. **HUMAnN2功能组成**
- 五. GraPhlAn可视化物种
- 六. LEfSe分析物种差异
- 七. 功能组成统计分析



易生信

- HUMAnN2: The HMP Unified Metabolic Analysis Network 2, HUMAnN是基于宏基因组、宏转录组数据分析微生物通路丰度的有效工具。这一过程称为功能谱，目的是描述群体成员的代谢潜能。可以回答微生物群体成员可能干什么，或在干什么的问题。

[HTML] [Metabolic reconstruction for metagenomic data and its application to the human microbiome](#)

[S Abubucker](#), [N Segata](#), [J Goll](#)... - PLoS computational ..., 2012 - journals.plos.

Microbial communities carry out the majority of the biochemical activity on the p they play integral roles in processes including metabolism and immune homeos human microbiome. Shotgun sequencing of such communities' metagenomes p information complementary to organismal abundances from taxonomic markers resulting data typically comprise short reads from hundreds of different organis best challenging to assemble comparably to single-organism genomes. Here, w

★ 被引用次数: 648 相关文章 所有 29 个版本

Abubucker, S. *et al.* Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLOS Computational Biology* **8**, e1002358, doi:10.1371/journal.pcbi.1002358 (2012).

[Species-level functional profiling of metagenomes and metatranscriptomes](#)

[EA Franzosa](#), [LJ McIver](#), [G Rahnavard](#), [LR Thompson](#)... - Nature ..., 2018 - nature.com

Functional profiles of microbial communities are typically generated using comprehensive metagenomic or metatranscriptomic sequence read searches, which are time-consuming, prone to spurious mapping, and often limited to community-level quantification. We developed HUMAnN2, a tiered search strategy that enables fast, accurate, and species-resolved functional profiling of host-associated and environmental communities. HUMAnN2 identifies a community's known species, aligns reads to their pangenomes, performs ...

★ 被引用次数: 23 相关文章 所有 5 个版本

Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* **15**, 962-968, doi:10.1038/s41592-018-0176-y (2018).

[Nature Methods: HUMAnN2实现宏基因组和宏转录组种水平功能组成分析](#)

[HUMAnN2: 人类微生物组统一代谢网络分析2](#)

[宏基因组有参流程 \(HUMAnN2\)](#)



- 可对已知和未知生物分析群体功能谱
MetaPhlAn2和ChocoPhlAn泛基因组数据库
- 可获得基因组、基因和通路层面的结果
UniRef基因家族，MetaCyc基因通路，MinPath定义最小通路集
- 简单的使用界面(单行命令实现全部工作流程)
用户只需提供质控后的宏基因组或宏转录组数据
- 加速序列比对
采用Bowtie2加速核酸水平比对，采用Diamond加速核酸翻译蛋白水平比对



- [MetaPhlAn2](#) (version \leq 2.6.0)
- [Bowtie2](#) (version \geq 2.2) (automatically installed)
- [Diamond](#) (0.9.0 > version \geq 0.8.22) (automatically installed)
- [Python](#) (version \geq 2.7)
- [MinPath](#) (automatically installed)
- [Usearch](#) (version \geq 7.0) (only required if using usearch for translated search)
- [SAMtools](#) (only required if bam input files are provided)
- [Biom-format](#) (only required if input or output files are in biom format)



HUMAnN2工作原理

- File Type 1 (a quality-controlled metagenome or metatranscriptome)

- fastq (fastq.gz)
- fasta (fasta.gz)

- File Type 2 (alignment results type 1)

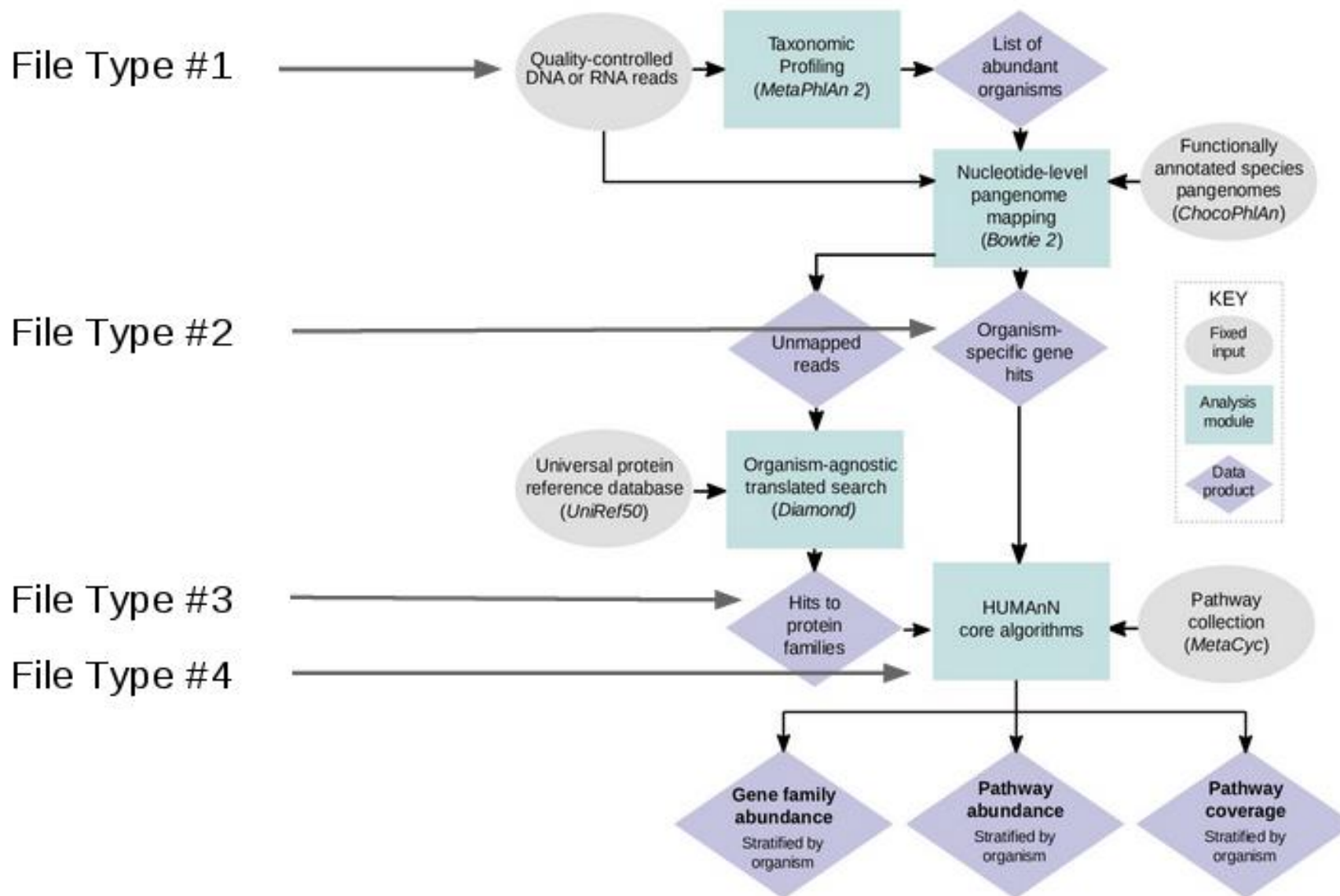
- sam
- Bam

- File Type 3 (alignment results type 2)

- blast-like tsv

- File Type 4 (gene table)

- tsv
- biom



软件安装和数据库下载

- **HUMAnN2有参宏基因组物种和功能定量流程**

conda install humann2

- **查看可用数据库并设置下载位置**

humann2_databases # 显示可用数据库

cd \${db}

mkdir -p \${db}/humann2 # 建立下载目录

- **微生物物种核心基因 5.37G**

humann2_databases --download chocophlan full \$wd

- **功能基因diamond索引 10.3G**

humann2_databases --download uniref uniref90_diamond \$wd



- 查看参数和数据库位置是否正确

显示参数

```
humann2_config --print
```

- 常用修改线程数、核酸和蛋白库位置

```
humann2_config --update run_modes threads 8
```

```
humann2_config --update database_folders nucleotide  
${db}/humann2/chocophlan
```

```
humann2_config --update database_folders protein ${db}/humann2
```

宏基因组

生信宝典

易生信



2.1 合并双端文件

- 有参宏基因组不考虑双端，将双端文件直接合并为一个文件

创建目录存放合并后的序列

```
mkdir -p temp/concat
```

for循环调用cat合并每一个样品

```
for i in `tail -n+2 result/metadata.txt | cut -f 1`;do \
```

```
    cat temp/qc/${i}*_1_kneaddata_paired_?.fastq > temp/concat/${i}.fq; done
```

查看样品数量和大小

```
ll temp/concat/*.fq
```

易生信 宏基因组



2.2 HUMAnN2计算物种和功能组成

```
mkdir -p temp/humann2
```

- 如果数据库位置正确，只需输入文件和输出目录，经parallel管理批量任务队列

```
parallel -j 3 \  
  'humann2 --input {} \  
  --output temp/humann2/ ' \  
  ::: temp/concat/*.fq > temp/log &
```

核心步骤，测序数据3X8=24线程，用时30min，真实数据可能要几小时至几天

易生信 生信宝 基因组



2.3 物种组成分析

```
mkdir -p result/metaphlan2
```

- # 样品结果合并

```
merge_metaphlan_tables.py  
temp/humann2/*_humann2_temp/*_metaphlan_bugs_list.tsv | \  
sed 's/_metaphlan_bugs_list//g' > result/metaphlan2/taxonomy.tsv
```

- # 转换为spf格式方便stamp分析

```
metaphlan_to_stamp.pl result/metaphlan2/taxonomy.tsv >  
result/metaphlan2/taxonomy.spf
```

- metaphlan_to_stamp.pl脚本来自microbiome_helper项目



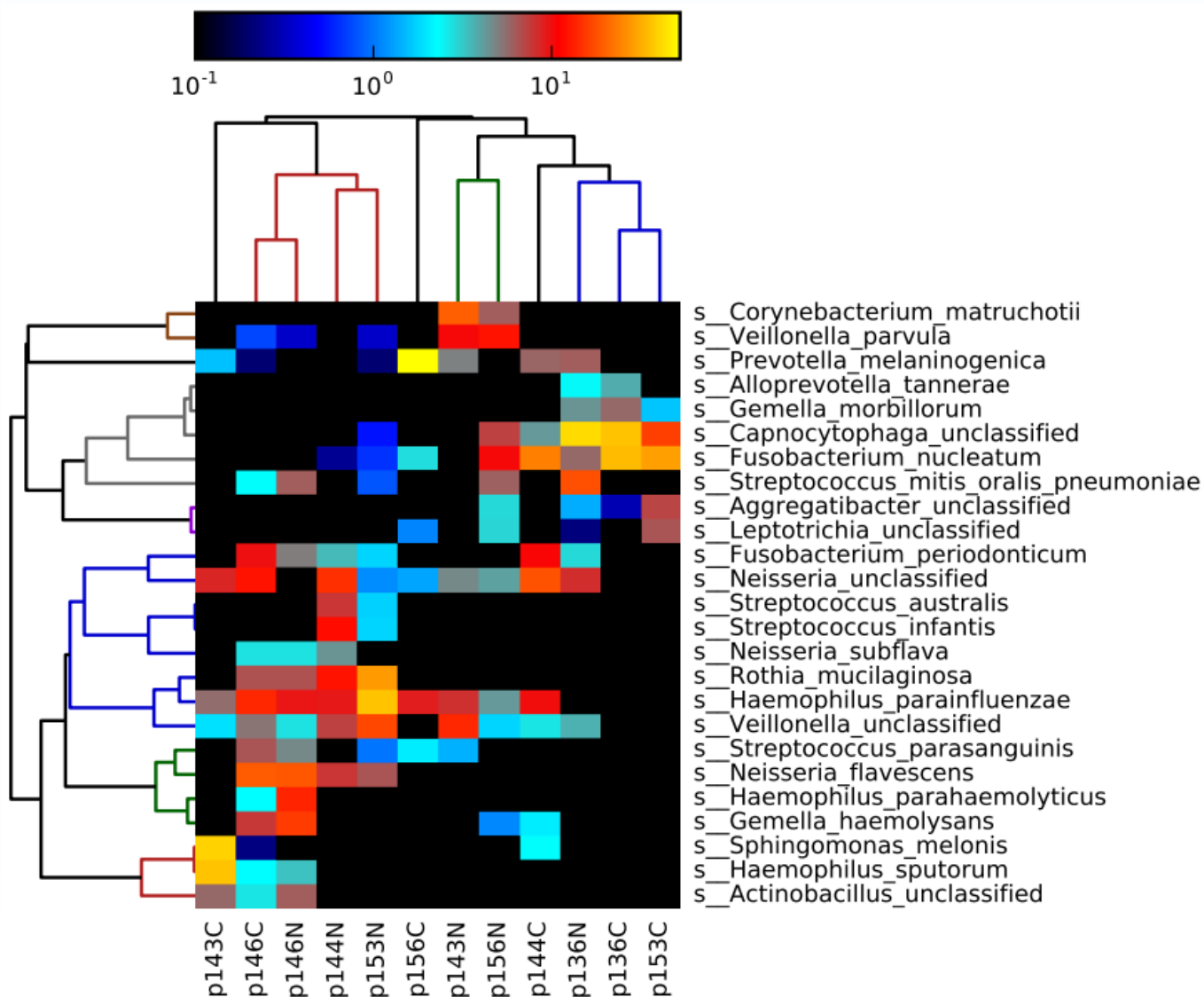
物种组成热图

- Metaphlan2提供了很多脚本，常用的有聚类热图

```
metaphlan_hclust_heatmap.py --in result/metaphlan2/taxonomy.tsv \
--out result/metaphlan2/heatmap.pdf \
-c bbcry --top 25 --minv 0.1 -s log
```
- # c设置颜色方案，top设置物种数量，minv最小相对丰度，s标准化方法，log为取10为底对数，文件名结尾可先pdf/png/svg三种图片格式。更多说明详见 `metaphlan_hclust_heatmap.py -h`
- 科研结果有时要服从人类可读性，选择合适数据的结果展示，读者容易懂你是关键



前25个物种丰度对数聚类热图



宏基因组
生信宝典

评估样本物种组成和分组聚类——更灵活绘制热图

- 方法1. metaphlan_hclust_heatmap.py 脚本服务器绘制热图，依赖关系和环境变量复杂
- 方法2. Excel筛选 metaphlan2/taxonomy.tsv 并在线绘制热图
- 方法3. R数据筛选taxonomy.spf并用pheatmap绘制热图，可指定分类级别、物种数量

```
Rscript db/script/metaphlan_hclust_heatmap.R \  
-i result/metaphlan2/taxonomy.spf \  
-t Order \  
-n 25 \  
-o result/metaphlan2/heatmap_Order
```

易生信
宏基因组
生信宝典



2.4 制作功能组成表

○ # 合并所有样品

```
humann2_join_tables --input temp/humann2 --file_name pathabundance \  
--output result/humann2/pathabundance.tsv  
sed -i 's/_Abundance//g' result/humann2/pathabundance.tsv
```

○ # 标准化为相对丰度relab或百万分数cpm

```
humann2_renorm_table --input result/humann2/pathabundance.tsv --units relab \  
--output result/humann2/pathabundance_relab.tsv
```

○ # 分层结果

```
humann2_split_stratified_table --input  
result/humann2/pathabundance_relab.tsv \  
--output result/humann2/
```

基因组
生信宝典
易生信

- HUMAnN2采用diamond比对UniRef数据库，获得完整的功能组成；注意数据库位置设置，保证其可运行；了解其依赖关系，有助于解决运行时依赖关系报错的问题；
- 结果提供功能组成，又提供了功能来源的物种，有相关join, norm, stratified脚本快速生成人类可读表格；
- HUMAnN2调用MetaPhlAn2基于bowtie2比对上万个物种的数据库获得物种组成，可快速、准确获得细菌、真菌、古菌、病毒、真核生物等的物种组成。
- 自带merge_metaphlan_tables.py、metaphlan_hclust_heatmap.py快速生成丰度矩阵和热图，metaphlan_to_stamp.pl生成STAMP格式。



- [宏基因组公众号文章目录](#)
- [生信宝典公众号文章目录](#)
- 加拿大生信网 <https://bioinformatics.ca/>
- 加拿大生信网宏基因组课程中文版——[挖掘微生物组生物标记](#)
- 美国高通量开源课程 <https://github.com/ngs-docs>
- The Huttenhower Lab <http://huttenhower.sph.harvard.edu/>
- 微生物组助手 https://github.com/LangilleLab/microbiome_helper
- Susan Holmes <http://statweb.stanford.edu/~susan/>





扫码关注生信宝典，学习更多生信知识



扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识



问题1. MetaPhlAn2数据库无法找到的问题

- 正常在首次运行时，会自动下载数据库。有时会失败，解决方法：
- 参考软件安装 # 可选：Metaphlan2数据库没成功的手动配置
- 原理：下载别人下载好的数据文件，解压并bowtie2建索引；链接到软件安装目录

```
db=~/.db
```

```
soft=~/.miniconda2
```

```
mkdir -p ${soft}/bin/db_v20
```

```
ln -s ${db}/metaphlan2/* ${soft}/bin/db_v20/
```

```
mkdir -p ${soft}/bin/databases
```

```
ln -s ${db}/metaphlan2/* ${soft}/bin/databases/
```

易生信
宏基因组

