诺禾致源微生物 Meta 项目售后 FAQ

目录

一、名词解释

1. Metagenomic（宏基因组）

宏基因组是基因组学一个新兴的科学研究方向。宏基因组学（又称元基因组学，环境基因组学，生态基因组学等），是研究直接从环境样本中提取的基因组遗传物质的学科。

传统的微生物研究依赖于实验室培养，宏基因组的兴起填补了无法在传统实验室中培养的微生物研究的空白。过去几年中，DNA 测序技术的进步以及测序通量和分析方法的改进使得人们得以一窥这一未知的基因组科学领域。

Magenomics 研究的对象是整个微生物群落。相对于传统单个细菌研究来说，它具有众多优势，其中很重要的两点：

(1) 微生物通常是以群落方式共生于某一小生境中，它们的很多特性是基于整个群落环境及个体间的相互影响的，因此做 Metagenomics 研究比做单个个体的研究更能发现其特性；

(2) Metagenomics 研究无需分离单个细菌，可以研究那些不能被实验室分离培养的微生物。

广义宏基因组是指特定环境下所有生物遗传物质的总和 ，它决定了生物群体的生命现象 。它是以 生态环境中全部 DNA 作为研究对象 ，通过克隆 、异源表达来筛选有用基因及其产物 ，研究其功能和彼此之间的关系 和相互作用 ，并揭示其规律的一门科学 。

狭义宏基因组学则以生态环境中全部细菌和真菌基因组 DNA 作为研究对象 ，它不是采用传统的培养微生物的基因组 ，包含了可培养和还不能培养的微生物的基因 ，通过克隆 、异源表达来筛选有用基因及其产物 ，研究其功能和彼此之间的关系和相互作用 ，并揭示其规律 。

2. Raw Data

测序得到的原始数据，但是由于数据量太大，我们不会长久保存，一般在短期内就会删除。

3. Clean Data

对原始数据进行过滤处理，得到用于分析的有效数据。

4. Contig

拼接软件基于 reads 之间的 overlap 区，拼接获得的序列称为 Contig（重叠群）。

5. Scaffold

基因组 de novo 测序，通过 reads 拼接获得 Contigs 后，通过使用具有 paired-end 关系的 reads 对 Contig 序列集进行连接后得到的序列集。

6. Scaftig

将组装得到的 scaffolds 从 N 连接处打断，得到不含 N 的序列片段，称为 scaftig。

7. Contig N50

Reads 拼接后会获得一些不同长度的 Contigs。将所有的 Contig 长度相加，能获得一个 Contig 总长度。然后将所有的 Contigs 按照从长到短进行排序，如获得 Contig 1，Contig 2，Contig 3……Contig 25。将 Contig 按照这个顺序依次相加，当相加的长度达到 Contig 总长度的一半时，最后一个加上的 Contig 长度即为 Contig N50。举例：Contig 1+Contig 2+ Contig 3+Contig 4=Contig 总长度*1/2 时，Contig 4 的长度即为 Contig N50。Contig N50 可以作为基因组拼接的结果好坏的一个判断标准。

8. Scaffold N50

Scaffold N50 与 Contig N50 的定义类似。Contigs 拼接组装获得一些不同长度的 Scaffolds。将所有的 Scaffold 长度相加，能获得一个 Scaffold 总长度。然后将所有的 Scaffolds 按照从长到短进行排序，如获得 Scaffold 1，Scaffold 2，Scaffold 3……Scaffold 25。将 Scaffold 按照这个顺序依次相加，当相加的长度达到 Scaffold 总长度的一半时，最后一个加上的 Scaffold 长度即为 Scaffold N50。举例：Scaffold 1+Scaffold 2+ Scaffold 3 +Scaffold 4 +Scaffold 5=Scaffold 总长度*1/2 时，Scaffold 5 的长度即为 Scaffold N50。Scaffold N50 可以作为基因组拼接的结果好坏的一个判断标准。

9. Scaffold N90

N50 和 N90 是基因组组装中的常用组装指标。Scaffold N90 含义为，将所有的 Scaffold 长度相加，能获得一个 Scaffold 总长度。然后将所有的 Scaffolds

按照从长到短进行排序，如获得 Scaffold 1，Scaffold 2，Scaffold 3……Scaffold 25。将 Scaffold 按照这个顺序依次相加，当相加的长度达到 Scaffold 总长度的 90% 时，最后一个加上的 Scaffold 长度即为 Scaffold N90。该数值反映了基因组 90% 以上的区域，都能被该数值以上长度的序列覆盖，体现了组装对于后续分析的质量贡献。

10. Q 值及 Q20、Q30

Q 值是 Illumina 在做碱基测序过程中，从测序原始数据转换为碱基的过程中评估出的质量分数取整的结果，E 为错误率，则质量分数为 Q =-10 log10(E)，以 Q 值 20 为例，折合错误率约为 0.01，Q 值 30 时错误率则为 0.001。而 Q20 和 Q30 则是 Reads 中 Q 值高于 20 或 30 的碱基所占的比例，反映了整体测序的质量情况。

11. GC 分布图

GC 分布图是展示 GC 随 reads 读长不同位置的分布比例变化，不同颜色曲线分别表示了 A，T，G，C 及 N 的比例。由于采用了随机 PCR 扩增和双端测序，因此，AT 比例之间和 GC 比例之间大体上应该是一致的，不过测序前端序列由于受到引物连接的一些偏好性影响，可能有一定的波动。

12. K-mer 及其作用

K-mer 就是一个长度为 K 的 DNA 序列，K 为正整数。如 K=17，则称为 17-mer。K-mer 有多种用途，用于纠正测序错误；构建 contig；估计基因组大小、杂合率和重复序列含量等。假设测序 read 读长为 L，则一条 read 上能取出 (L-K+1) 个 K-mer。用 K-mer 估算基因组大小时，基因组的大小= K-mer 总数/K-mer 期望深度。

13. K-mer 值的选择

通过对大量文献的研究，我们发现一般组装设置的 k-mer 值为 49、55、65，但是由于宏基因组组装为杂菌组装，并非某一单菌，因此组装过程更为复杂，消耗集群资源也会更大，时间也会更长。在结合大量文献以及项目经验综合统计分析，我们目前选择的 k-mer 值一般为 55，组装效果还是比较好的。

14. 测序深度、基因深度和覆盖度

测序深度是指测序得到的总碱基数与待测基因组大小的比值。假设一个基因大小为 2M，测序深度为 10X，那么获得的总数据量为 20M。再比如要测的细菌基因组大小 5M，实际测序得到 raw data 1.2Gb，质量剪切后 clean data 1Gb，那么测序深度就是 200X。

基因深度是根据单碱基位点的深度计算得来的，从基因来源的 Scaftigs 在各样品中的单碱基位点深度数据出发，根据基因在 Scaftigs 上的起始位点和终止位点，可以得到各基因在各样品中所对应的深度。既是基因各个碱基平均测序深度之和。

覆盖度是指测序获得的序列占整个基因组的比例。由于基因组中的高 GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖有所的区域，这部分没有获得的区域就称为 Gap。例如一个细菌基因组测序，覆盖度是 98%，那么还有 2%的序列区域是没有通过测序获得的。

15. 各个数据库的介绍：

15.1 KEGG

KEGG 数据库（Version: 2018.01，http://www.kegg.jp/）是一个综合性数据库，其中最核心的为 KEGG PATHWAY 和 KEGG ORTHOLOGY 数据库。在 KEGG PATHWAY 数据库中，将生物代谢通路划分为 7 类，分别为：细胞过程（Cellular Processes）、环境信息处理（Environmental Information Processing）、遗传信息处理（Genetic Information Processing）、人类疾病（Human Diseases）、新陈代谢（Metabolism）、生物体系统（Organismal Systems）、Drug Development（药物发展），其中每类又被系统分类为二、三、四层。第二层目前包括 66 个种子 pathway；第三层即为其代谢通路图；第四层为每个代谢通路图的具体注释信息。

在 KEGG ORTHOLOGY 数据库中，将行使相同功能的基因聚在一起，称为 Ortholog Groups (KO entries)，每个 KO 包含多个基因信息，并在一至多个 pathway 中发挥作用。不是所有的 KO 都有 EC 编号；Module 比 pathway 更精细。

15.2 eggNOG

eggNOG 数据库（Version：4.5，http://eggnogdb.embl.de/）是利用 Smith-Waterman 比对算法对构建的基因直系同源簇（Orthologous Groups）及其功能进行注释， eggNOG V4.5 涵盖了 2,031 个物种的基因，构建了约 19 万个 Orthologous Groups。

综合了 COG 和 KOG 数据库，直系同源簇 (Orthologous Groups)功能数据库；1）通过已知蛋白对未知序列进行功能注释；2）种系发生图谱，以此确定特定 OG 中某给定物种是否存在这些功能蛋白，以用来确定在一个物种中是否一个特定的代谢途径。

15.3 CAZy

CAZy 数据库（Version: 2018.01，http://www.cazy.org/）是研究碳水化合物酶的专业级数据库，主要涵盖 6 大功能类：糖苷水解酶(Glycoside Hydrolases ，GHs)，糖基转移酶(Glycosyl Transferases，GTs)，多糖裂合酶(Polysaccharide Lyases，PLs)，碳水化合物酯酶（Carbohydrate Esterases，CEs），辅助氧化还原酶(Auxiliary Activities ， AAs)和碳水化合物结合模块（Carbohydrate-Binding Modules，CBMs）。

15.4 CARD

CARD 数据库（Version:2.0.1，https://card.mcmaster.ca/）目前包括 4008 个本体术语（Ontology Terms），2498 条参考序列，1211 个 SNP，2437 个出版物（Publications）和 2545 个 AMR 检测模型，能够对 76 种病原体，4194 条染色体，4666 个质粒，66095 个 WGS 组装，132747 个等位基因进行抗体预测。

二、质控、组装过程

1. 基因丰度如何进行均一化

过滤掉在各个样品中支持 reads 数目<=2 的基因，获得各样本中基因的 reads 数目表，即 Unigenes.readsNum.xls，从 reads 数目及基因长度出发，采用类 TPM 算法（公式在结题报告中有），计算获得各样本中基因的相对丰度，生成 Unigenes.readsNum.relative.xls 表格。具体均一化过程为：流程默认采用 Unigenes.readsNum.xls 表的所有样本中基因 reads 数目之和的最大值作为阈值，

最大值除以样本的所有基因的相对丰度之和即为该样本的放大倍数，对 Unigenes.readsNum.relative.xls 表中各样本的基因丰度乘以放大倍数，生成均一化后的基因丰度表，即 Unigenes.readsNum.even.xls 表格。

2. 增加数据量是否可以组装出低丰度物种

在 Metagenome 中，高丰度物种和低丰度物种分布的不均匀性，物种的多样性会对组装过程造成困难，组装软件在进行组装时，若高丰度和低丰度物种差异太大，组装时，会把低丰度物种所属的 kmer 作为分支也剪切掉，这时会造成低丰度的物种难以组装出来，这也是现在宏基因组研究中的一个难点。测序深度的增加，对于低丰度物种的组装能否有利，需要看该物种的具体的丰度情况，若丰度太低，即使增加数据量，也不一定能够组装出来。

3. 测序数据量（测序深度）限制宏基因组组装效果，那为什么不把所有样本数据合并在一起组装？

1）资源限制，所有样本数据合并组装将消耗极大的内存，并且时间相当长；

2）不同样本的高丰度物种差异很大，所有样本都混合在一起进行组装，那么无疑会大大增加数据的复杂度，组装效果可能会更差[1]；

[1] Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing[J]. nature, 2010, 464(7285): 59-65.

4. 组装效果以及 reads 利用率

1）目前土壤水体等环境复杂度高的样本组装 N50 在 600bp-800bp 左右，组装效果好的可以达到 1.5k 左右，reads 利用率与组装结果直接相关，一般在 20%左右，也会有达到或超过 50%的项目；

2）肠道样本：N50 目前基本能达到或者超过 1.5k，实际 reads 利用率超过 80%；

不同的样本类型采用的组装方法也不同，生物样本采用预处理后得到 Clean Data，使用 SOAP denovo 组装软件进行组装分析（Assembly Analysis）；生态样本的微生物组成更复杂，采用 Megahit 进行组装。

5. 为什么有污染的情况下组装结果较差

由于组装软件在组装过程中是将测序数据看作来自同一个基因组的前提下进行的，如果有外源 DNA 混杂，其中不同来源的 DNA 中会有不同程度的相似性序列和非相似性序列，这些复杂的关系会对组装软件产生干扰，而软件为保证组装的准确性，只能将可疑的部分切断成不同的碎片序列，而这也导致最终的组装只能拿到碎片化的序列，而失去了组装本身想要达到的效果。

6. 预测基因的完整性问题

（1）为什么结果文件 Total.gene.CDS 中几乎 99%以上都不能同时找到起始密码子（ATG、ATT、ATA）及终止密码子（TAA TAG TGA），或者其反向序列，一个完整的 CDS 框是一个完整的编码序列。如果要编码一个蛋白必须同时含有起始密码子和终止密码子？

答：对于基因预测，我们采取的是以下分析策略：

将序列长度为 500bp 以下的 scaftigs 进行过滤，对过滤后的序列，我们采用的是 MetaGeneMark 软件进行基因预测，该软件在宏基因组训练模型，在宏基因组研究中较为广泛使用。

对于该软件进行基因预测的结果，会存在只有起始密码子或终止密码子，或起始密码子和终止密码子都没有的情况，这些都是预测出来的片段基因，还有起始密码子和终止密码子都存在的情况，为预测出来的完整基因。

为了对比不同软件的基因预测结果，我们采用该项目的 scaftigs 序列作为输入，采用了 GeneMark, MetaGeneMark 这两个软件进行对比（这两个软件的训练模型和算法是不一样的，MetaGeneMark 拥有宏基因组训练模型）。对比结果如下所示：

	GeneMark	MetaGeneMark
start	0	9055
end	53	10618
none	0	15430
all	151	2710

表格中，start 指的是只含有起始密码子的序列，end 指的是只含有终止密码子的序列，none 指的是起始密码子和终止密码子都没有，all 指的是起始密码子和终止密码子都含有的。

（2）组装拼接基因为什么不完整？

答：目前大型的项目一般采用的都是 soapdenovo 软件进行的序列组装，如上面提到文献，我们采用的也是 soapdenovo 软件进行分析。从上面的基因预测结果统计表中可以看出，all 比例（即起始密码子和终止密码子都存在）在 7.2%左右，其他的片段基因在 93%左右。

在样品组装时，我们选取的是 soapdenovo 软件进行的组装，组装的效果主要跟以下几个因素有关：样品的测序数据量，物种的多样性，物种的丰度分布的不均匀性等，这些因素造成了宏基因组组装比细菌等单物种的组装更加困难，这些问题也是目前宏基因组科研中研究的重点。

（3）在组装拼接过程中是否是共有的高丰度基因组装出来了，而个体特有的低丰度基因没有拼接出来？（这种结果一般会导致个体间无任何显著差异）

答：受到测序深度及测序成本的影响，在现在的宏基因组文章中，一般选择 6G 的测序数据量，能够测量出样品中的绝大多数的微生物，但是，一些低丰度的物种，因为测序深度的原因，确实很有可能会组装不出来。在宏基因组分析中，一般较多的关注的是较高丰度的物种的构成情况，如果要对低丰度物种进行特殊分析，一般需要加大测序数据量，或者在前期提取过程中经过一些特殊的处理，尽可能的富集出多的低丰度物种，再进行测序分析。

7. 为什么选择 500bp 作为过滤条件

设置 500bp 的过滤条件是基于大量项目经验以及文献支撑进行的，是目前分析过程中应用比较普遍的方法，认可度也比较高，具体参考文献如下。

[1] Li J, Zhao F, Wang Y, et al. Gut microbiota dysbiosis contributes to the development of hypertension[J]. Microbiome, 2017, 5(1):14.
[2] Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis[J]. Nature, 2014.
[3] Karlsson F H, Fåk F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome[J]. Nature communications, 2012, 3: 1245

8. 宏病毒组装问题

1）目前环境样本中病毒丰度都非常低（海水相对高一点），如果直接提取环境样本进行测序组装，以目前的测序深度，很难组出来病毒序列；

2）为避免丰度低问题而进行的病毒随机引物扩增，富集的样本进行测序组装效果也不会太好，首先引物扩增出来的片段本身就有可能是断断续续的片段，使得组装效果不好；其次，随机引物很大可能扩增出细菌序列，拼接后较长的片段很大可能是细菌基因组序列；

3）病毒核酸序列本身变异度很大，组装难度较大；

9. 相对丰度和注释方法

（1）结果中相对丰度的数值都非常小（一般都是 10^{-5}～10^{-9}），是否会影响差异比较？

答：在我们给出的相对丰度表中，有丰度较高的（30%左右），也有丰度较低的（比如 10^{-5}），属于正常现象。从 metastats 的计算方法可以看出，数值的大小（比如 0.01vs0.03，100VS300）不同，计算结果中 p value 会有差异（http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000352，http://metastats.cbcb.umd.edu/detect_DA_features.r），在其计算过程中，首先会对数据进行均一化处理，处理后，若原始数据为绝对丰度矩阵，则会转化为相对丰度矩阵，随后，软件会对原始输入矩阵进行判断，若输入的数值非常小，会执行 fisher test，反之，则会执行 nonparametric t - test，因此，对于相对丰度非常低的物种，若将其放大 100000 倍再进行计算，其 p vaule 的计算结果会不一样。

（2）04.TaxAnnotation 文件夹下，MAT 文件夹中列出了界门纲目科属种的相对丰度，请问里面的 others 项是没有列完全，还是没有注释上？如若未注释上，比例是否太高？（比如在 species 表中，列出了 2797 种微生物，但 others 几乎占到了 90%）。

答：在 MAT 文件夹中，显示的是不同分类层级的相对丰度表，others 表示在某个分类层级（比如 species 层级）未被注释上的比例。我们注释时，采用的是基于 NT 库，采用 lca 算法进行的物种注释，在物种注释过程中，如果该基因同时

注释到了两个不同的 species，这两个 species 分别属于相同的 class，但是属于不同的 order，那么，在注释的时候该基因会被归属于 order 这个分类层级，而在 order 以下的分类层级，则会归为 others 当中了。

由上述过程可以看出，我们采用的 lca 算法，计算的是最保守的物种分类层级，保证了结果的准确性，但是损失了注释的精确性（精确到 species 层级等）。如果采用另外其他的算法，比如说以 best hits 作为注释结果，那么很有可能是保证了注释的精确性，但是损失了结果的准确性。综合以上的考虑，因此，我们采用了 lca 算法来进行物种注释。

10. 比对的基因数目与统计的数据不符

在“report”中的“功能注释”提到“功能注释结果概述：原始去冗余后的预测基因共有 917,971 个，有 397,818(43.34%)个基因能够比对上 KEGG 数据库，其中，有 213,050(23.21%)个基因能够比对上数据库中 2,909 个 KEGG orthologgroup（KO）；有 425,815(46.39%)个基因能够比对上 eggNOG 数据库；有 39,298(4.28%)个基因能够比对上 CAZy 数据库”。但是在表格 unigenes.absolute.level1.xls 中提到 GH 基因 24854 个，8147 个 GT 基因，CBM 基因 5545 个，CE 基因 2940 个，PL 基因 512 个，AA 基因 6 个，合计是 42004 个。与上面提到的 39298 个基因不相符。

“有 39,298(4.28%) 个基因能够比对上 CAZy 数据库”，这只是对所有样品中 unigenes 比对到数据库的整体情况进行描述，但同一个 unigenes 可能同时比对到数据库中同一层级下的不同功能，因此，同一层级下不同功能注释到基因数目的加和要大于等于整体注释情况。

11. 如何根据结果中生成的 CDS 或者蛋白质的序列找到其相所属的物种的序列

宏基因组测序实际是对复杂环境样本的微生物构成进行分析，由于测序深度和环境复杂度的影响，无法对样本的某种微生物基因组进行组装；所以 CDS 无法与具体的某个来源物种的 DNA 序列进行对应。

三、物种和功能注释

1. 基因功能注释过程

将 result\03.ReadsMapping\total.scaf.screening.fa 文件中 scaftigs 进行过滤，去除<500bp 的 scaftigs，用 MetaGeneMark 软件进行基因预测，并从预测结果出发，过滤掉长度小于 100nt 的信息，将预测的结果统计在 result\05.GeneComponet\GenePredict\Total.gene.rename.mgm.gff 文件中，并再次进行了序列的重命名。然后对各样品及混合组装的基因预测结果，用 CD-HIT 软件进行去冗余，将 result\05.GeneComponet\GenePredict\Total.gene.rename.mgm.gff 中的基因进行聚类，过滤掉在各个样品中支持 reads 数目<=2 的基因，根据这些选择的代表性基因，再进行各功能数据库的功能注释。

2. 真菌注释信息较少，如何解释，是否有其他解决策略？

真菌注释信息较少的可能原因有以下几个方面：

1）样本中真菌丰度本来就很低，测序深度有限造成真菌序列组装不起来而影响后续物种注释；

2）真菌基因组杂合度高，较原核基因组复杂很多，单个真菌组装难度就很大，而在宏基因组如此复杂环境中组装真菌序列难度更大，因此可能在组装时就只有很少的真菌序列被组装出来从而影响后续的注释；

3）由于基因预测软件对原核和真核基因的预测模型不同，而宏基因组基因预测软件（MetaGeneMark）偏向于原核生物的基因预测，所以在基因预测部分造成预测出来的真菌基因偏少从而影响后续注释；

候补解决策略：

可以尝试基于 reads 进行物种和功能分析，跳过组装和基因预测的限制；例如利用 MetaPhlAn 进行物种注释，MetaPhlAn 的 marker 基因集来源于~17,000 基因组序列（~13,500 细菌和古菌, ~3,500 病毒, 和~110 真核），可在一定程度上解决上述问题。

3．功能丰度分析柱形图纵坐标不到“1”

流程默认去除未被注释到的信息，使注释到的信息更加直观，因此纵坐标不到1。

4．16S 注释结果与 Meta 注释结果的比较

关于“16S 分析和 Meta 分析的注释结果相差较大”，我们进行了排查以及查阅了相关文献，回复如下：

16S 分析和 meta 分析结果存在差别，主要有两个方面的原因：

1）分析方法存在较大差异；Shah N 等在 2011 年发表的一篇文献 [1] 指出，尽管先前有研究表明 16S 和 Meta 的分析得到相似的物种组成，但是，他们的研究结果表明，大多数不同来源的环境样品经过 16S rRNA 与 meta 分析得到的物种注释结果都存在着明显的差别，主要原因是，16S 是经过扩增的，而且不同物种的 DNA 扩增的倍数不一致（biased because of unequal amplication of species' 16S rRNA genes ...）；而宏基因组的 DNA 测序深度可能不是十分足够（may not be deep enough to detect...），并且，宏基因组分析得到的相对物种丰度因 DNA 的提取方法以及测序方法有很大差异（vary signicantly depending on the DNA extraction and...）；

2）物种注释方法以及数据库存在着一定的差别；16S 采用的是将 16S rRNA 与 Silva 数据库进行比对注释，而 meta 是将预测得到的基因与 NR 数据库进行比对从而进行注释；

另外，16S 分析和 Meta 分析的注释结果也存在一定的相似点，比如说门水平相对丰度排名靠前的物种的类别是相似的等；

综上所述，分析方法本身存在一定差异，是导致 16S 分析和 Meta 分析的注释结果存在一些差别的主要原因，但同时两者也有一定的相似之处。

参考文献：

Shah N, Tang H, Doak T G, et al. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics.[J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2011, 16(6):165.

因此 16S 与 meta 由于本身测序手段、数据库和分析方法存在很大差异，实际上不具可比性！

针对一个问题项目（3 个样本，不具普适性，仅供参考）做得如下测试：

1）原始的扩增子标准流程分析；

2）原始的 META 标准流程分析；

3）从原始的扩增子序列进行 16S 拷贝数矫正后的分析；

4）基于宏基因组 reads 的 MetaPhlAn 分析

5）基于宏基因组分离 16S 序列的分析

测试结果：

1）基于 metaphlan 软件的结果比较接近于扩增子的分析结果；

2）拷贝数矫正后的扩增子分析结果与标准扩增子分析结果有差异；

3）基于宏基因组分离 16S 的分析结果与宏基因组标准分析结果一致性高

5．宏病毒的分类

病毒在纲和门上没有分类，目层级也只有个别有分类。

6．病毒注释结果与 NCBI 网站比对结果不同

宏病毒分析采用 3 个数据库（Virus Refseq、VirusNT、Acalme）进行比对，最终的注释结果是选取的最优注释。在使用 NCBI 网站比对时，是与 NCBI 中的 NT/NR 全库进行比对，其数据库中不仅包含病毒序列信息，也包含了真核、原核生物的序列信息，由于测序 reads 较短，用 reads 序列比对时很可能比对到其他物种上，且在线比对的参数与流程中的参数不一致。ncbi 上的数据库实时更新，我们的数据无法实时更新，数据库的差异也会导致比对结果的差异。

7．物种和功能对应关系的查找

以 KEGG 的注释结果为例，可查找结果文件 05.FunctionAnnotation/KEGG/KEGG_Anno/Unigenes.KEGG.tax.xls

该文件为通过 KEGG 注释到不同层级的基因对应到其相关的物种注释信息，可以通过该文件筛选关注功能的物种对应信息。

8. Metastats 如何进行差异比较分析

Metastats 在计算过程中，首先会对数据进行均一化处理，处理后，若原始数据为绝对丰度矩阵，则会转化为相对丰度矩阵，随后，软件会对原始输入矩阵进行判断，若输入的数值非常小，会执行 fisher test，反之，则会执行 nonparametric t－test。（参考网址为：

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000352 ，http://metastats.cbcb.umd.edu/detect_DA_features.r）。

9. 如何查找每一个样本中的基因数目

关于每个样本中实际 unigene 数目的问题，因为宏基因组组装包括个体组装和混合组装，再把所有样本的 clean reads 比对到组装后的 scaftigs 上，去掉所有 readnum 小于 2 的即 cutoff<=2 的基因，再进行信息分析。因此可能会出现部分基因在 mapping 时 mapping 不到的情况，也就是说会有部分基因并未参与后续分析，而由于无法确定这些基因来自于哪一个样本，因此不能够精确的统计每个样本中的基因数目。

但是我们可以统计实际参与信息分析的基因数目，具体统计表格可见结果文件\03.GenePredict\GeneTable\Total\Unigenes.readsNum.even，该表格为均一化之后的基因丰度，如下所示：

Reference_ID	A1	A2	A3	A4	A5	A6	A7
A1_1	232.3588	391.9214	338.6748	180.5658	236.069	121.8372	220.47
A1_100	569.2508	96.8435	43.48409	204.3427	163.9483	223.8845	508.48
A1_1000	863.04	38.4788	30.25439	1006.277	568.8864	119.1789	695.05
A1_100003	27.17588	29.23883	0	4.83787	12.24343	0	
A1_10001	80.651	4.311716	2.300449	24.96965	20.01076	95.49196	70.301
A1_100011	30.47762	15.61486	9.997277	0	0	3.952271	
A1_100013	25.88179	0	0	0	0	0	
A1_100022	22.33634	36.62006	146.5354	2.650888	13.41745	2.896527	4.8940
A1_100032	31.23664	2.880673	0	0	0	0	
A1_100035	93.3313	192.394	275.531	29.32042	49.46839	272.3174	154.27

A2 样本中丰度不为 0 的基因 ID 即为该样本存在的基因类型，老师可以利用 Excel 的数据统计功能，统计该样本中所有丰度不为 0 的数值，看下具体的数目。

由于我们均一化是基于 TPM 计算公式进行的，因此会出现不取整的情况，具体公式信息及参考文献可见宏基因组结题报告。

$$G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

（说明：r 为比对上基因的 reads 数目，L 为基因的长度）

另外，由于我们所选取展示的基因均是某一基因的代表序列，例如 A1_1，该基因可能会存在于多个样本中。

10. 如何根据结果中生成的 的 CDS 或者蛋白质的序列找到其相所属的物种的序列

宏基因组测序实际是对复杂环境样本的微生物构成进行分析，由于测序深度和环境复杂度的影响，无法对样本的某种微生物基因组进行组装，所以 CDS 无法与具体的某个来源物种的 DNA 序列进行对应。

四、结果文件解读

1. 03.GenePredict 文件夹下 GeneTable 中 coverage.depth.table.xls 的各列数据代表什么意思

该表格是各基因覆盖度总体情况统计，各列数据分别为：基因名称、基因长度、read 比对到基因的碱基个数、覆盖率、测序深度（单碱基覆盖的加和除以基因长度）、单碱基覆盖的加和。

2. 03.GenePredict 文件夹下 GeneTable 文件，如何查看基因是样本间共有还是样本特有

我们在进行组装的过程中有一步混合组装，为提高组装效果，我们将每个样品组装后未利用上的数据再进行混合组装，混合组装数据作为一个独立样品再进行基因预测，其预测出的基因 ID 命名为 NOVO_MIX。预测基因后，进行基因去冗余，每一类序列相似的基因会选择出一个代表序列作为该类基因的代表基因，其代表基因 ID 的命名一般选择在哪个样品中丰度较高，命名为哪个样品。每个样品预测出来的基因可能在两个样品中都会存在。根据基因的 ID 无法看出该基因是某

个样品独有或者是两个样本共有，可以根据结果文件中 03.GenePredict/GeneTable/Total/Unigenes.readsNum.relative.xls，查看基因在每个样品中的丰度，来推断该基因是共有还是特有。

3．CAZy_MAT 文件夹下 Absolute 文件中各列数据代表什么意思

Absolute 文件夹内容是在 CAZY 数据库中对不同层级的各个功能类在各样品中的绝对丰度进行的统计。第一行为样品信息，第一列为各层级功能的名称，中间的数字则代表某个功能类在某个样品中的绝对丰度，最后一列为功能的详细说明。

4．CAZy_MAT 文件夹下 GeneNums.BetweenSamples 基因数目是否是拷贝数

这里的基因数目是我们在进行组装、基因预测后，将每个样品预测出的基因与 CAZY 数据库进行比对，注释到数据库中各个层级上的基因数目。

5. eggNOG_Anno 文件夹下 Unigenes.blast.m8.filter.anno.xls 的 Subject ID 是什么

Subject ID 是比对上的序列号， 是数据库中该基因的 ID 号，可在该数据库的官网中查找该基因的具体信息。

6．如何查找显著差异基因的具体序列

先根据 Metastat 表格中挑选出感兴趣的 KO 号，根据 KO 号在结果文件 /result/05.FunctionAnnotation/KEGG/KEGG_Anno/Unigenes.blast.m8.filter.anno.xls 找到其对应的基因 ID 号，根据其基因 ID 号在 /result/03.GenePredict/UniqGenes/Unigenes.CDS.cdhit.fa 文件中寻找其具体的序列信息。

7．E-Value 和 Score 是什么意思

E-Value 和 Score 是进行比对后产生的一个分值，用于评估比对结果， E-Value 值越低越好，Score 值越高越好。

8. Pathway overview 图是否有分辨率高的图

代谢通路的比较分析，由于结果文件过大，在结题报告中展示了部分结果，全部分析结果在最终释放的结果文件中，以矢量图形式展示，保证放大缩小不会失真。

9. KO 号输入 KEGG 网址分析，部分基因找不到代谢图

KEGG 数据库中注释到的基因有一部分是参加代谢网络的，有代谢通路图，可以在 KEGG 的 Pathway 数据库中找到，但是有一部分基因是不参加代谢通路网络的，或者是 KEGG 的 pathway 数据库现有的代谢通路图中没有该基因参与的代谢通路图，这部分基因只能在 KEGG 的 gene 库中找到，不能在 pathway 数据库中找到。

10. 实际 map 图中的具体酶类功能描述与 overview 整体描述不符

在 KEGG 的 overview 展示中，某一 map 描述为具体的某一功能，实际上在该 map 中的具体的酶类物质，其在 KEGG 数据库中的功能描述与 map 功能描述不符，例如某一 map 描述可能与抗生素生物合成相关，但是具体的酶类物质可能只是某一脂质或者参与脂肪酸代谢。

原因可能为：

一、这些具体的酶类物质可能参与某一大的生物合成通路中的某些前体反应，因此从目前字面上的结果来看，两者属于不同的功能。

二、我们在进行 overview 数据抓取时已经抓取了涵盖的所有结果，只是根据某一 EC 号或者 KO 号去 KEGG 数据库进行查找的时候，查找的内容会涉及多个 map，而且这种查找方式相当于查找某一种化合物参与的所有通路信息，涵盖面非常广，没有针对于所关注的该 map，因此结果会有差异。

11. Metastat 中 P value 与 Q value 的计算方式

P-value：用 Metastats 软件，对组间的物种或功能丰度数据进行假设检验得到 p 值；

Q-value：Meta 项目中的 Metastat 分析所用的 Q 值计算参考文献《Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples》，计算步骤：

1）计算π 0（λ ）值：$\hat{\pi}_0(\lambda)=\frac{\#\{p_j>\lambda\}}{m(1-\lambda)}$.

λ 为 0 到 0.95 间隔为 0.01 的数组，m 为 P 值个数，#{Pj>λ }为所有 p 值大于λ 的个数；

2）用λ 、π 0（λ ），拟合三次样条函数，得到 $\hat{\pi}_0=\hat{f}(1)$.；

3）对 P 值进行从小到大排序，并计算 $\hat{q}(p_{(m)})=\min(p_{(m)}\times\hat{\pi}_0,1)$.；P 值中最大值乘以π 0 与 1 比较，取最小值作为 q 的最大值；

4）对 P 值从大到小依次进行计算，$\hat{q}(p_{(i)})=\min\left(\frac{\hat{\pi}_0\times m\times p_{(i)}}{i},\hat{q}(p_{(i+1)})\right)$，P 值乘以π 0 乘以 m 除以秩序的数值与前一个 P 值计算出的 q 值，取最小数作为当前的 Q 值；

参考文献：

【1】Audic, S. and J. M. Claverie (1997).The significance of digital gene expression profiles. Genome Res 7(10): 986-95.

【2】Benjamini, Y. and D. Yekutieli (2001).The control of the false discovery rate in multiple testing under dependency.The Annals of Statistics.29: 1165-1188.

12．如何打开结果文件

我们在对应文件夹内有较为详细的文件说明。结果文件分两类，一类是普通文本文件（或压缩的文本文件），可以用文本编辑器打开（或解压后用文本编辑

器打开），推荐文本编辑器 Notepad2，下载网址为：

http://www.flos-freeware.ch/notepad2.html；另外一类是图片文件，可以用图片浏览器打开，如 windows 中的图片预览工具，其中一类图片文件是 svg 文件，是用于网络的矢量图格式，可以用 8.0 以上的 IE，或者 firefox，chrome 之类的第三方浏览器打开。更老版本的 IE 需要安装插件才能显示，推荐使用 firefox。

此外超大文件的打开，如 cleandata 的 fastq 文件，这些文件也是纯文本文件，可用文本编辑器打开，不过由于 windows 下读取文本文件的内存机制，会将整个文件读入内存，瞬间将内存占满，因此通常读取会产生障碍。实际上这些文件是测序的原始数据，结果文件中已包括具体的统计结果，建议老师不要在 windows 下尝试打开类似文件。如果老师有需求，我们可以提供截取部分文件内容作为示例。

13．物种及功能注释结果中 others、Unclassified、Candidatus

1、others 表示分类时，程序无法根据规则判断应该属于哪一类，可能是注释到该水平，注释信息却是 Unclassified；也可能是没有注释到该水平。

2、 在种水平上能够有具体的注释信息，但是在上层水平上属于 Unclassified，这种情况一般表示在比对到的数据库的某一参考序列有具体的种水平注释信息，但是在上一层级的分类水平上却无法区分或所属上一层级没有定义好的注释名称（Un--s-），这种情况在微生物中较为常见。

3、Candidatus 也是微生物分类学中的一个分类层级，一般是不可培养的微生物，是一种临时的分类层级；在二代测序中，即便是接近完整的 16S 基因组，也有可能注释到 Candidatus。

14．物种注释 others 比例高的问题

对于物种或者功能的注释结果，我们除了更改一些标点符号或者特殊字符外，基本上是完全复制的数据库中的信息，而这些信息也都是先前研究的结果，因此能够注释到的结果，均是选择了阈值范围之内得分最高的，结果比较可靠。

目前我们物种注释比对使用 2018.1.18 的 NR 数据库，涵盖信息比较全，因此不会因为数据库版本或者注释流程的问题导致 others 比例高，可能是属于样本本身的特性，这些没有注释结果的序列有可能是未知生物，或者是已知生物，但数

据库中信息太少。而我们由于受限于数据库中的相关信息，因此具体属于哪一类还不能明确。

五、常见高级分析或个性化分析

1. contig-binning

contig binning 适用产品类型：

1. 少样本量，深度测序样品：大数据量，少样品量，比如红树林项目，2 个样品，数据量 120G

2. 大样本量样品：一般数据量，大样品量，参考 concoct 的文章，样品量达到 50 以上，平均数据量 5G 左右

Contig binning 具体步骤：

第一步：binning 前期评估，一般选取物种注释结果 top100 的物种，使用 Checkm 软件评估混杂度和完整性，也可以针对于老师感兴趣的某一物种进行评估；

第二步：contig binning 分析，使用 Concoct/MetaBAT 软件对组装结果进行 binning，通过 CheckM/SCGs 的方法对 binning clusters 进行质量评估；

第三步：单菌草图分析，对质量评估合格的每个 cluster 进行单菌组装，并进行基因组组分分析和基因组功能注释分析。

补充说明：

完整度-completeness: estimated completeness of genome as determined from the presence/absence of marker genes and the expected collocalization of these genes

混杂度-contamination: estimated contamination of genome as determined by the presence of multi-copy marker genes and the expected collocalization of these genes

一般情况下，完整度≥75%，混杂度≤10%，定义为 LowRisk，很可能获得高质量 Binning 结果；完整度≥75%，混杂度>10%且≤200%，定义为 Neutral，通过一定数量的样品有机会获得高质量 Binning 结果；完整度<75%，或混杂度>200%，很难通过少量样品获得高质量 binning 结果。可根据以上内容进行判断是否可以进行 contig-binning。

2. MetaPhlAn 物种注释流程

MetaPhalAn2 能够直接通过微生物宏基因组质控数据与 marker 基因集比对，对复杂样品（细菌、古菌、真核和病毒）进行精确到种水平的物种注释，与基于组装和基因注释进行宏基因组物种注释的方法（如 LCA）相比，省却了组装和基因预测过程的资源消耗，注释的准确度也大幅提高。

MetaPhlAn2 的 marker 基因集来源于~17,000 基因组序列（~13,500 细菌和古菌, ~3,500 病毒, 和~110 真核），具有以下特点：1. 可以对样品进行明确的物种分类；2. 准确评估物种相对丰度；3. 对细菌、古菌、真核生物和病毒进行种水平物种鉴定；4. 对所有样品进行特定菌株的鉴定和追踪；5. 和已有方法相比能注释到更多的低丰度物种。

Metaphlan 流程共分第一步质控、第二步 metaphlan 物种注释、第三步 diversity 分析三个部分。