### 基于高通量测序的微生物组研究技术简介

——微生物组研究，从方案设计到写作套路(一)

作者：王晓雯 凌波微课

版本1.0.2，更新日期：2020年9月22日

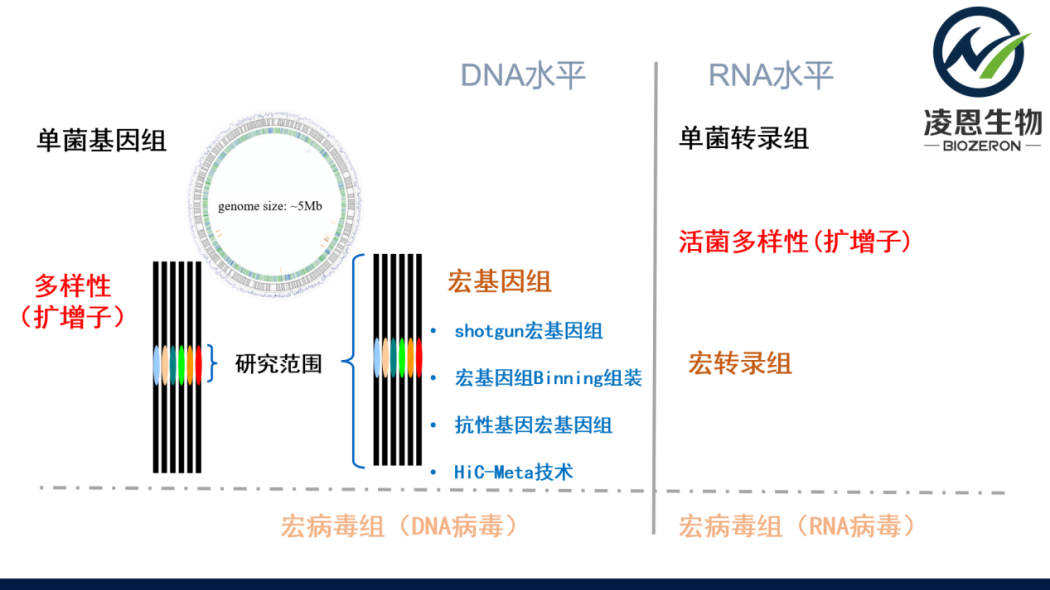
微生物组研究的热潮愈演愈烈，已经深入到我们生活的各个方面——从污染监测到人体疾病与健康，从食品检测到日常营养——微生物的研究历程，也从最开始的单菌研究，发展到近现代的群体研究，再更进一步回归到个体的功能机制研究之上。

研究的深入，首先一定是伴随着技术的完善与更新。而一项成功的研究一定少不了各种实验及分析技术的选择和优化整合。

不同的研究侧重，就涉及到不同的技术选择。包括扩增子测序，宏基因组测序、宏转录组测序等技术如何选择和开展应用——是高水平科研课题面临的第一个问题~

#### 技术的选择

基于技术原理的不同，虽然同为高通量测序的手段，但每种方法都有各自的研究侧重。实验设计上要同时考虑到研究目标、技术侧重和成本控制等多方面问题。在不同的核酸水平开展的研究，技术形式多样，首先就来讲讲这些研究复杂微生物群落的“花式”技术：



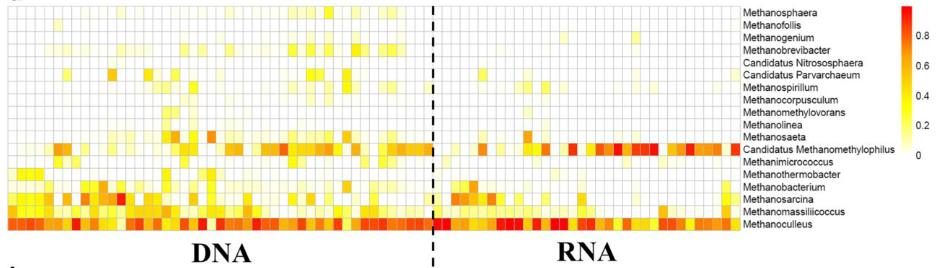
#### 基于NGS的扩增子测序技术

扩增子测序技术目前在研究中最常见，文献报道也最多的微生物组高通量测序手段。它的原理是**基于微生物基因组的特征性序列(如细菌的16S rDNA，真菌的ITS等)进行测序和分类**。简言之，就是以小见大(类似于通过指纹进行身份识别)。它的测序范围只是通过扩增检测微生物基因组上的特征性区域(如，16S rDNA上的某些高变区域)，研究目的也倾向于**物种的分类识别和相对丰度统计**。多样性项目有一个特点，靶向性明确。**根据引物来确定需要研究的微生物类群**，一对引物只针对样本中的细菌，或真菌，或某一类微生物群落。

在开展此类项目时需要注意一点，就是适宜的引物的选择：不同的引物在针对不同类型的样本研究中，也会有一些偏向性影响，进而造成分析效果的差异。例如，有报道发现，细菌V5-V7区引物可获得最低的叶绿体/线粒体比例和较高的属水平物种覆盖度1，暗示其对排除样本的叶绿体污染具有潜在价值；又如，V3-V4区引物较常见于人体及环境的微生物组研究，相较于其他引物，在二代测序的细菌扩增子研究中，属于物种分辨率最佳的引物，但其针对某些细菌及古菌物种的检测准确性不足2。

基于已有研究报道，目前细菌研究中较多见V3-V4区(341F-806R)、V4-V5区(515F-907R)引物，这两对细菌引物多见于肠道及土壤研究中。针对某些植物内生菌，比较推荐V5-V7区引物(799F-1193R)3,4，以屏蔽细胞器基因组污染。在各类细菌研究中，为了降低物种检测的偏好性，获得更精准的物种分类结果，也有研究者开始利用三代测序开展16S全长测序研究(后面有详细介绍)。真菌研究较多选择ITS引物，而针对样本中真核生物的检测较多推荐18S区的引物。当然，即便是这样有针对性的研究仍不可避免一些干扰：例如在常见的植物内生真菌等研究中，ITS和18S的引物扩增也都不可避免有很大的宿主污染的情况。这就需要我们采用多元的技术方法来降低干扰，如针对目标菌群的富集培养或通过加大测序数据量等方式获得更多的有效数据用于后续分析等。

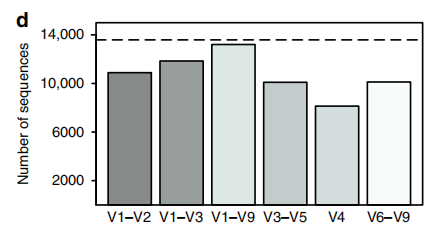
此外，基于DNA和RNA不同的理化特征，以不同类型的核酸为模板，可以获得不同目的类型的数据：**以DNA为模板开展扩增子研究，可以获得样本中所有存在的特定微生物类群；而以RNA为模板开展扩增子研究，则可获得样本中具有生物活性的特定菌群5**。



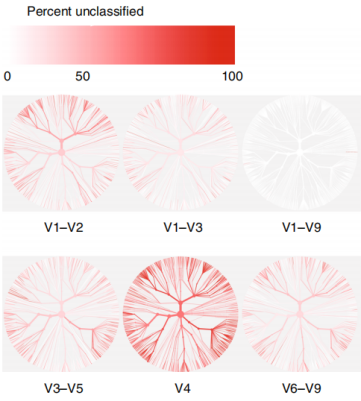
图注：同一样本中存在的细菌(DNA)及活性细菌(RNA)比较研究5

#### 基于三代测序技术的扩增子测序技术

基于三代测序技术的扩增子测序技术：伴随着三代测序技术的进步和测序成本的不断下降，利用三代测序技术平台(如PacBio)进行长片段的扩增和序列读取成为扩增子测序发展的热门方向。由于三代长读长的技术优势，使得16S全长测序成为可能。更长读长带来的是更加丰富的序列信息，显著提升了基于NGS测序所达不到的**高水平物种识别精确度、分辨率及对样本中微生物群落的还原度(下图)6**。



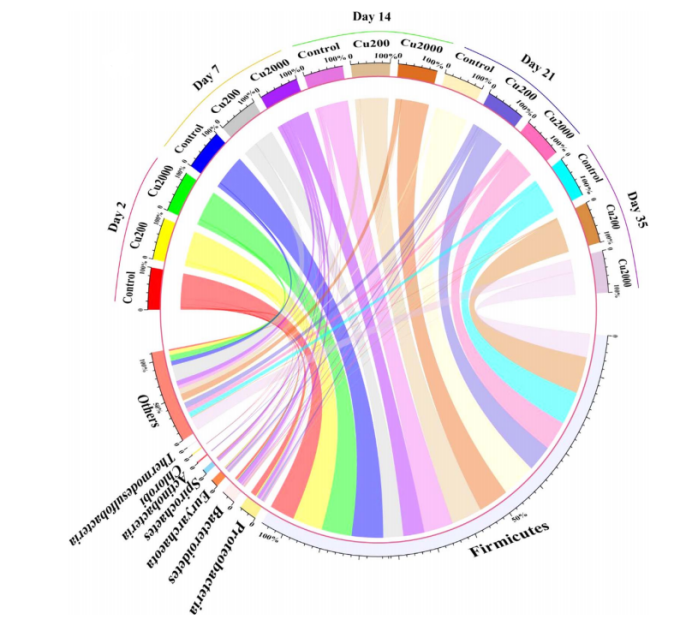
图注：以99％的序列相似性对每个可变区的序列进行聚类时创建的OTU数(虚线表示原始数据库中唯一序列的数量)6



图注：不同可变区引物在物种鉴定中分类能力存在差异，且具有分类偏好性。其中，V4区无法鉴定的物种最多，V1-V9全长区域可实现分类鉴定的物种最多6。

#### 基于扩增子技术的功能基因研究

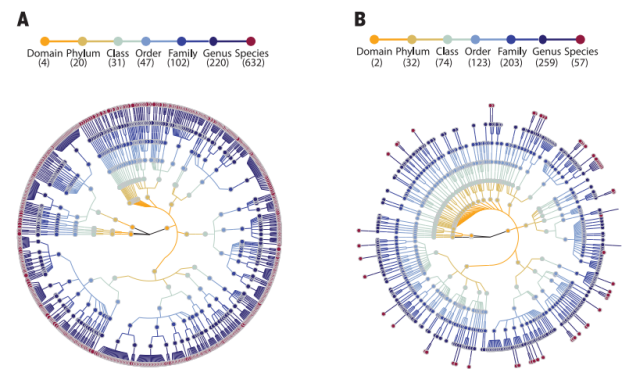
\*基于扩增子技术的功能基因研究：除了上述针对细菌真菌等微生物的扩增子研究外，也有很多研究者关注环境样本中具有特定功能类型的菌群研究，如产甲烷菌、硝化细菌、反硝化细菌等。利用特定功能基因的保守区进行引物设计，开展扩增子测序研究，也可以获得相应的功能基因类群群落结构信息(如下图)，这也是目前针对环境样本中特定功能微生物研究的常用方法。 然而，受限于功能基因的保守区进行引物设计质量不高、已知的功能基因数据库信息不足等因素，仅利用扩增子技术开展菌群功能研究还是很有限的。如何满足样本水平上真正的菌群功能研究呢？这就需要利用高通量测序技术开展宏基因组研究，一方面可以获得菌群的构成信息，另一方面可以实际测得样本中功能基因的存在分布，是深入开展微生物组功能研究的有利技术。



图注：堆肥中添加不同浓度Cu离子及不同发酵时间的样本中具有*nifH*基因功能的物种(属水平)的分布情况7

#### 宏基因组测序技术

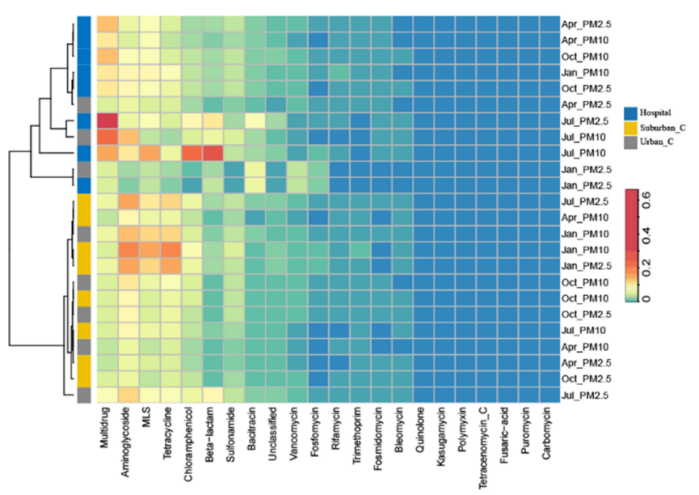
\*宏基因组测序：无需扩增，对样本中所有存在的DNA进行测序，进一步开展物种识别鉴定和基因功能研究。由于少了特异性扩增的步骤，所以研究的对象不是针对某一类型微生物，而是**样本中所有存在的广泛生物类型(可参考下图8)**。非特异性扩增研究带来的优势是，直接的检测到样本中的基因存在，可直接**获取DNA水平的基因功能的注释。因此，相对于扩增子技术而言，宏基因组的随机(shotgun)测序更有利于研究实际检测样本中微生物的功能潜力，是研究微生物组功能的不二之选。**



图注：利用宏基因组测序和16S rDNA测序实现的物种鉴定进化树。(A图为宏基因组测序鉴定的物种类型进化树，B图为16S rDNA测序鉴定的物种进化树。后者仅包含细菌群落构成，前者包含多个物种类型的鉴定)8。

#### 基于宏基因组的功能基因研究

\*基于宏基因组的功能基因研究：例如宏基因组-抗性基因研究：顾名思义，这项研究是针对复杂微生物样本中抗生素抗性基因(ARGs)的存在而开发的。目前，通过整合ARDB、BacMet及CARD三大数据库，对复杂微生物组的宏基因组数据开展针对抗生素抗性基因及抗金属基因的注释，获得样本中ARGs的种类和丰度(如下图9)，进而研究其与样本的相关性。最初，该项目主要用于环境样本中(如土壤、水体、空气等)ARGs的分布和监控，近些年也开始有医学方向的研究应用。当然，如果有其他类型功能基因的数据库，这部分工作明显可以移植，同时我们预期，随着研究的深入，各类针对不同生物、生境的功能基因数据库会越来越多，使研究内容不局限于上述几种常见的功能基因。

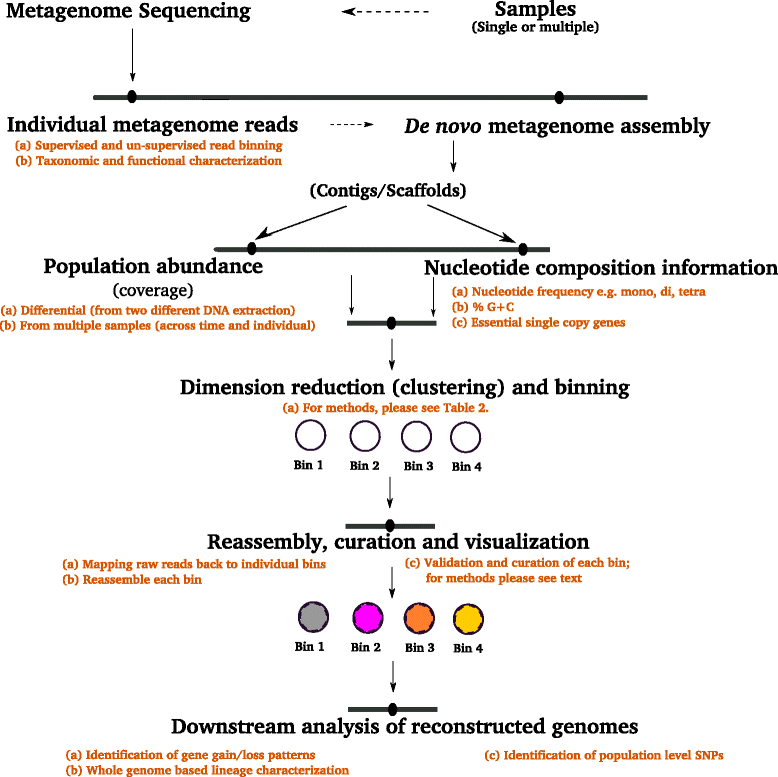


图注：所有样本中含有的ARGs种类及丰度heatmap图9

很多研究者在进行微生物群体研究的时候，都会想方设法的将其中的一些关键物种进行分离，对关键菌种开展基因组研究，来描述和评估群体中单菌的功能及贡献。但是这种样本中，并非关注的单菌都能实现理想的分离培养。在这种情况下，基于宏基因组的手段开展单菌基因组研究的技术应运而生：

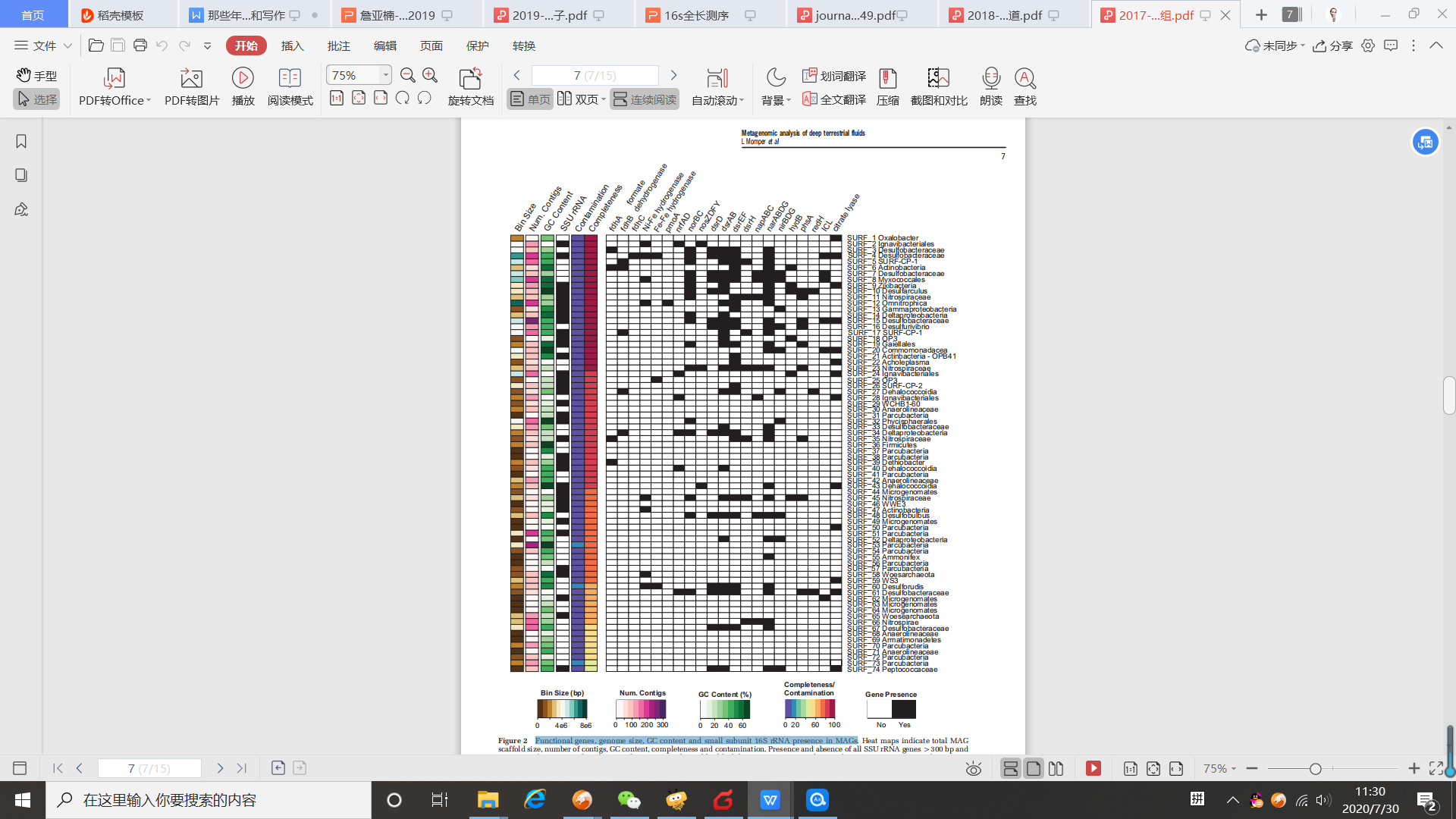
#### 宏基因组测序分箱

\*宏基因组测序binning组装：实验**技术方法及特点同常规的宏基因组测序**。在分析中引入了序列分箱 (Binning)的流程(下图10)，利用核酸的组成信息、物种的丰度模式等信息，将宏基因组获得的序列进行聚类，可获得重叠群的分箱(contig binning)。基于此，对样本中**优势的微生物种类进行基因组框架图的组装**，获得宏基因组组装的基因组(MAGs)。



图注：宏基因组分箱的基本原理和流程

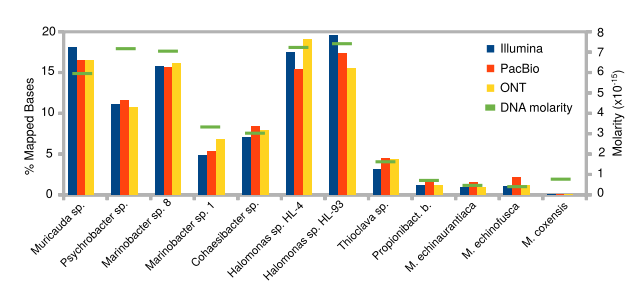
分箱可以说是纯粹基于算法开展的宏基因组单菌组装方案，在加大宏基因组测序量的前提下，分箱组装可以在获得微生物群体功能信息的同时，对其中的优势物种开展有针对性的基因组研究，可用于**将特定的功能代谢产物与特定物种、特定基因相关联(如下图11)，推动微生态研究中的因果机制的探索，在疾病监控、环境监测等领域提供菌株水平的生物靶标，特别适用于重点关注微生物中优势菌功能及下游开展单菌分离培养的培养组**等研究中。



图注：深层底下微生物群落的宏基因组研究，基于分箱组装的MAG的功能基因，基因组大小，GC含量和MAG中16S rRNA小亚基的分布11。

#### 基于三代测序的宏基因组

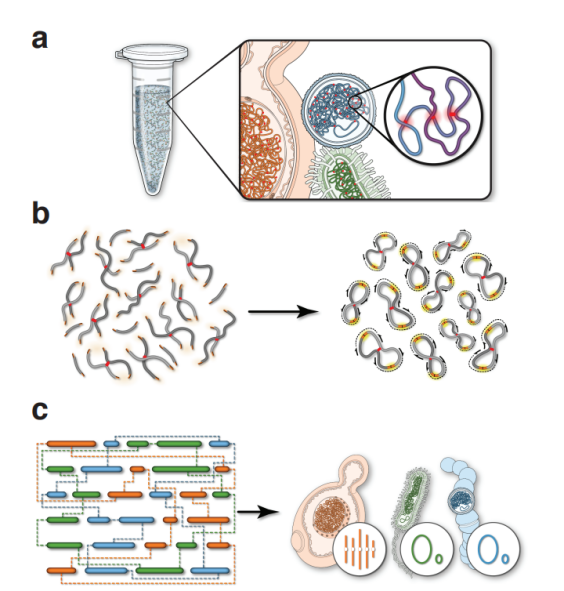
\*基于三代测序的宏基因组：伴随着三代测序技术的发展，超长读长测序技术开始走进微生物组的宏基因组研究。三代测序两大特色技术平台——Pacbio & Nanopore——可以实现3-10Kb级别的超长读取，能够轻松覆盖复杂微生物组样本中的DNA全长序列；此外，无GC偏好性的特点，可以更真实地反应菌群的复杂组成，更准确的开展微生物组的物种分布、群落结构、功能基因挖掘及代谢网络研究。



图注：模拟菌群研究中不同平台对菌种分布的还原情况12

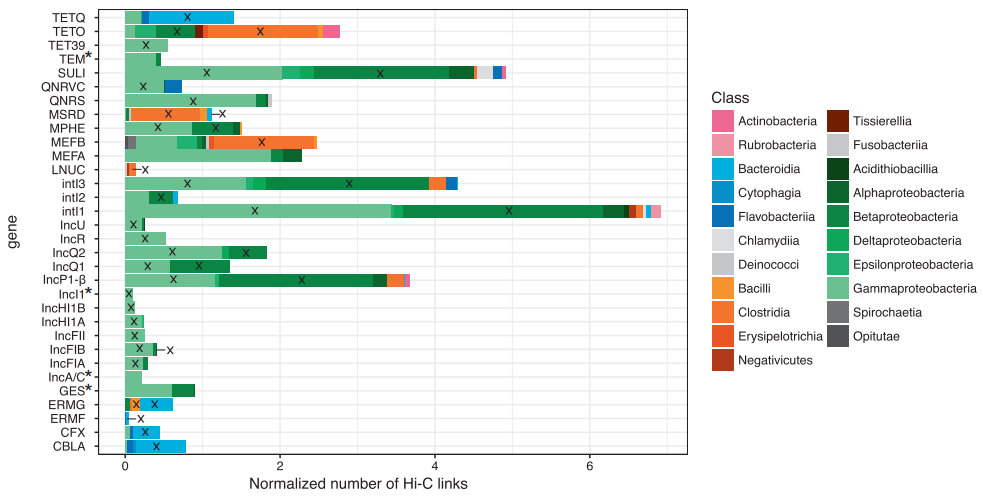
#### HiC-Meta宏基因组测序

\*HiC-Meta宏基因组测序：在宏基因组的基础上，额外利用Hi-C(染色质空间构象捕获)技术对微生物群落样本进行建库(技术原理见下图13)。



图注： 利用Hi-C技术进行宏基因组中单菌基因组组装得技术原理13

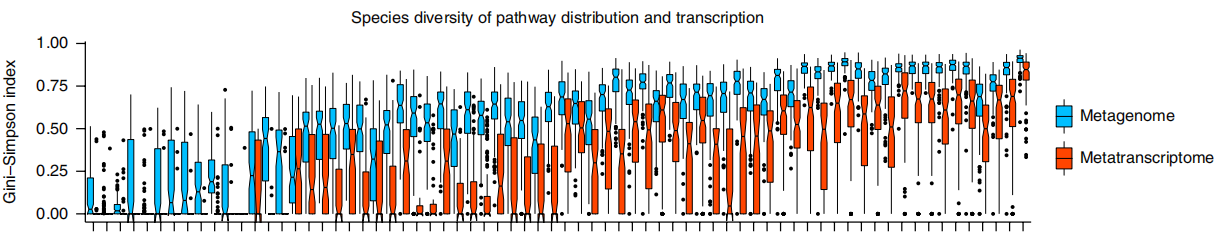
基于Hi-C的建库原理，可以更好地实现同一细胞来源的序列的划分，有助于**对宿主与质粒、噬菌体与宿主菌的对应关系开展研究(参见下图14)**，获得丰富的微生物群落元素构成之间的互作网络信息。



图注：污水研究中利用HiC-Meta技术，将分类物种与其含有的质粒、ARG及整合子进行关联对应14。

#### 宏转录组测序

\*宏转录组测序：顾名思义，有别于宏基因组以DNA为研究对象，宏转录组是对环境样本中存在的RNA开展测序。样本中存在的微生物不一定具有活性，或者的微生物也会根据生境的变化调节自身的基因表达和功能发挥(如下图示15)——这就是宏转录组研究的意义之所在。



图注：所有样本中， 通路分布的多样性等在DNA水平及RNA水平存在差异15。

针对样本中所有存在的广泛生物类型，宏转录组测序获得的是具有“活性”的信息——活性物种，表达中的活性基因和活性功能。针对复杂微生物群落的动态活性研究，宏转录组测序将是您的不二选择。

#### 宏病毒组研究

\*宏病毒组研究：又名病毒宏基因组学(Viral Metagenomics)，或宏病毒组(Metavirome)，是在宏基因组学理论的基础上，结合现有的病毒分子生物学检测技术而兴起的一个新的学科分支。宏病毒组直接以环境中所有病毒的遗传物质为研究对象，能够快速准确的鉴定出环境中所有的病毒组成。受限于目前的技术瓶颈，根据研究的目的病毒类型，分为DNA病毒的宏组学研究和RNA病毒的宏组学研究(单/双链 DNA/RNA**需要分别建库**)。研究重点关注复杂样本中潜在病毒对人类及环境的危害，主要用于样本中分布的病毒的物种分类、相对丰度及相关功能信息研究。

#### 微生物组技术概括

目前已有报道的微生物组技术的研究应用偏向及优劣势简要概括如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 技术名称 | | | 研究对象 | 研究目的(偏好) | | | 物种  分辨率 | 成本 |
| 扩增子测序 | DNA或RNA为模板 | 基于NGS平台 | 单一微生物类群 | 物种组成及丰度 | | | 较高(属水平) | 低 |
| 基于三代平台 | 单一微生物类群，多见16S全长 | 物种组成及丰度 | | | 高(种水平) | 较低 |
| 宏基因组测序 | | 随机宏基因组测序 | 样本中全部微生物类型 | 物种组成及丰度 | 基因预测、功能注释、代谢通路研究 | | 一般 | 一般 |
| 分箱组装 | 同随机宏基因组测序，倾向高丰度菌 | 同随机宏基因组测序 | 基于NGS数据可开展单菌基因组框架图组装(优势菌) | | 优势种分辨率高 | 较高 |
| 抗性基因分析 | 同随机宏基因组测序，重点关注抗性基因 | 同机宏基因组测序 | 抗生素抗性基因注释，可开展物种与抗性基因相关性研究 | | 一般 | 一般 |
| 三代测序宏基因组 | 样本中全部微生物类型和群落功能研究 | 物种组成及丰度 | 基因预测，功能注释，代谢通路研究 | | 高 | 较高 |
| HiC-Meta技术 | 关注优势菌，宿主-质粒，宿主-噬菌体关系等。 | 同宏基因组分箱组装 | 引入Hi-C建库，优化单菌基因组框架图组装 | 优势单菌、宿主与质粒和噬菌体关系研究 | 优势物种分辨较高 | 高 |
| 宏转录组测序 | | | 样本中全部活性微生物的基因功能研究 | 具有生物活性的物种组成及丰度； | 表达中的基因及通路信息等 | 不同样本之间的差异表达基因和功能 | 一般 | 较高 |
| 宏病毒组测序 | | DNA病毒/RNA病毒 | 样本中的病毒，分为DNA/RNA病毒类群研究 | 环境样本中的病毒类群 | DNA/RNA病毒需分别开展研究 | 病毒分类和功能注释方面 | 一般 | 较高 |

责编：刘永鑫 中科院遗传发育所

版本更新历史

1.0.0，2020/9/3，王晓雯，初稿

1.0.1，2020/9/8，刘永鑫，大修

1.0.2，2020/9/2，刘永鑫，小修

#### 参考文献

1 Fan Wang, Xiao Men, Ge Zhang, Kaichao Liang, Yuhua Xin, Juan Wang, Aijun Li, Haibo Zhang, Haobao Liu & Lijun Wu. (**2018**). Assessment of 16S rRNA gene primers for studying bacterial community structure and function of aging flue-cured tobaccos. ***AMB Express*** 8, 182, doi: <https://doi.org/10.1186/s13568-018-0713-1>

2 Emma K. Wear, Elizabeth G. Wilbanks, Craig E. Nelson & Craig A. Carlson. (**2018**). Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton. ***Environmental Microbiology*** 20, 2709-2726, doi: <https://doi.org/10.1111/1462-2920.14091>

3 Ancheng C. Huang, Ting Jiang, Yong-Xin Liu, Yue-Chen Bai, James Reed, Baoyuan Qu, Alain Goossens, Hans-Wilhelm Nützmann, Yang Bai & Anne Osbourn. (**2019**). A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. ***Science*** 364, eaau6389, doi: <https://doi.org/10.1126/science.aau6389>

4 Jingying Zhang, Yong-Xin Liu, Na Zhang, Bin Hu, Tao Jin, Haoran Xu, Yuan Qin, Pengxu Yan, Xiaoning Zhang, Xiaoxuan Guo, Jing Hui, Shouyun Cao, Xin Wang, Chao Wang, Hui Wang, Baoyuan Qu, Guangyi Fan, Lixing Yuan, Ruben Garrido-Oter, Chengcai Chu & Yang Bai. (**2019**). *NRT1.1B* is associated with root microbiota composition and nitrogen use in field-grown rice. ***Nature Biotechnology*** 37, 676-684, doi: <https://doi.org/10.1038/s41587-019-0104-4>

5 Jo De Vrieze, Ameet J. Pinto, William T. Sloan & Umer Zeeshan Ijaz. (**2018**). The active microbial community more accurately reflects the anaerobic digestion process: 16S rRNA (gene) sequencing as a predictive tool. ***Microbiome*** 6, 63, doi: <https://doi.org/10.1186/s40168-018-0449-9>

6 Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren & George M. Weinstock. (**2019**). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. ***Nature Communications*** 10, 5029, doi: <https://doi.org/10.1038/s41467-019-13036-1>

7 Yanan Yin, Jie Gu, Xiaojuan Wang, Kaiyu Zhang, Ting Hu, Jiyue Ma & Qianzhi Wang. (**2018**). Impact of copper on the diazotroph abundance and community composition during swine manure composting. ***Bioresource Technology*** 255, 257-265, doi: <https://doi.org/https://doi.org/10.1016/j.biortech.2018.01.120>

8 Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F. Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, Jun Wang, Floris Imhann, Eelke Brandsma, Soesma A. Jankipersadsing, Marie Joossens, Maria Carmen Cenit, Patrick Deelen, Morris A. Swertz, Rinse K. Weersma, Edith J. M. Feskens, Mihai G. Netea, Dirk Gevers, Daisy Jonkers, Lude Franke, Yurii S. Aulchenko, Curtis Huttenhower, Jeroen Raes, Marten H. Hofker, Ramnik J. Xavier, Cisca Wijmenga & Jingyuan Fu. (**2016**). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. ***Science*** 352, 565-569, doi: <https://doi.org/10.1126/science.aad3369>

9 Peng He, Yan Wu, Wenzhong Huang, Xinwei Wu, Jiayun Lv, Pengda Liu, Li Bu, Zhijun Bai, Shouyi Chen, Wenru Feng & Zhicong Yang. (**2020**). Characteristics of and variation in airborne ARGs among urban hospitals and adjacent urban and suburban communities: A metagenomic approach. ***Environment International*** 139, 105625, doi: <https://doi.org/https://doi.org/10.1016/j.envint.2020.105625>

10 Naseer Sangwan, Fangfang Xia & Jack A. Gilbert. (**2016**). Recovering complete and draft population genomes from metagenome datasets. ***Microbiome*** 4, 8, doi: <https://doi.org/10.1186/s40168-016-0154-5>

11 Lily Momper, Sean P. Jungbluth, Michael D. Lee & Jan P. Amend. (**2017**). Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. ***The ISME Journal*** 11, 2319-2333, doi: <https://doi.org/10.1038/ismej.2017.94>

12 Volkan Sevim, Juna Lee, Robert Egan, Alicia Clum, Hope Hundley, Janey Lee, R. Craig Everroad, Angela M. Detweiler, Brad M. Bebout, Jennifer Pett-Ridge, Markus Göker, Alison E. Murray, Stephen R. Lindemann, Hans-Peter Klenk, Ronan O’Malley, Matthew Zane, Jan-Fang Cheng, Alex Copeland, Christopher Daum, Esther Singer & Tanja Woyke. (**2019**). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. ***Scientific Data*** 6, 285, doi: <https://doi.org/10.1038/s41597-019-0287-z>

13 Maximilian O. Press, Andrew H. Wiser, Zev N. Kronenberg, Kyle W. Langford, Migun Shakya, Chien-Chi Lo, Kathryn A. Mueller, Shawn T. Sullivan, Patrick S. G. Chain & Ivan Liachko. (**2017**). Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. ***bioRxiv***, 198713, doi: <https://doi.org/10.1101/198713>

14 Thibault Stalder, Maximilian O. Press, Shawn Sullivan, Ivan Liachko & Eva M. Top. (**2019**). Linking the resistome and plasmidome to the microbiome. ***The ISME Journal*** 13, 2437-2446, doi: <https://doi.org/10.1038/s41396-019-0446-4>

15 Galeb S. Abu-Ali, Raaj S. Mehta, Jason Lloyd-Price, Himel Mallick, Tobyn Branck, Kerry L. Ivey, David A. Drew, Casey DuLong, Eric Rimm, Jacques Izard, Andrew T. Chan & Curtis Huttenhower. (**2018**). Metatranscriptome of human faecal microbial communities in a cohort of adult men. ***Nature Microbiology*** 3, 356-366, doi: <https://doi.org/10.1038/s41564-017-0084-4>