人类微生物组研究设计、样本采集和生物信息分析指南

钱旭波[1]，陈同[2]，徐益萍[1]，陈雷[3]，孙馥香[4]，卢美萍[1]，刘永鑫[5,6]

1. 浙江大学医学院附属儿童医院风湿、免疫和变态反应科
2. 中国中医科学院中药资源中心
3. 首都医科大学附属复兴医院
4. 易汉博基因科技（北京）有限公司
5. 中国科学院遗传与发育生物学研究所
6. 中国科学院大学，生物互作卓越创新中心

钱旭波和陈同为共同第一作者

通讯作者：卢美萍，浙江大学医学院附属儿童医院风湿、免疫和变态反应科，中国浙江杭州竹竿巷 57 号，邮编：310003，邮箱：meipinglu@zju.edu.cn

A guide to human microbiome research: study design, sample collection, and bioinformatics analysis

Chinese Medical Journal [IF: 1.585]

DOI: https://doi.org/10.1097/CM9.0000000000000871

摘要

这篇综述的目的是为医学研究人员，特别是那些没有生物信息学背景的研究者提供简单易懂的微生物组学知识，包括研究中常用的概念、技术和分析方法等。首先，我们介绍了基本概念，例如微生物群（microbiota）、微生物组（microbiome）和宏基因组（metagenome）等。然后，我们讨论了研究设计方案、样本量计算方法以及提高研究可靠性的方法。我们特别强调了阳性和阴性对照的重要性。接下来，我们讨论了微生物组研究中常用的统计分析方法，重点关注多重比较的问题以及组间β多样性分析的方法。最后，我们介绍了生物信息学分析的具体流程。总之，严谨的研究设计是获得有意义结果的关键步骤，而适当的统计方法对于准确解释微生物组数据很重要。通过阅读这篇文章，研究者能获得研究设计、样本采集和生物信息分析等全方位的微生物组学知识。

关键词：微生物组、研究设计、统计分析、样本量、生物信息分析、分析流程

1. 前言

随着测序技术和数据分析方法的发展，近几年医学微生物组研究领域出现了一些令人瞩目的成果[1-3]，比如微生物组与代谢性疾病[4-6]、消化系统疾病[7-10]和心血管系统疾病[11]之间的关系日益明确。这些发展和发现增加了医生在微生物组研究方面的兴趣，进而也涌现出了大量有价值的论文[12]。另外，随着 QIIME 2[13]和多组学方法[1, 9]等先进技术和分析流程的出现，微生物组分析方法也不断进步。然而，理解和掌握这些技术和分析流程并非易事，特别对于医生来说更是如此。

本文的目的是为研究者，特别是那些没有生物信息学背景的医生提供易懂的微生物组学知识，这些知识包括详细的微生物组学基本概念、科研设计方法、样本采集和保存方法、统计分析方法以及生物信息分析方法。我们希望医生们通过阅读此文能够快速掌握以上知识和方法，进而有效地挖掘数据背后的生物学意义。

2. 基本概念

2.1 Microbiota、Microbiome 等术语

Microbiota 是指~~寄生~~定植在人体特定部位的微生物，包括细菌、古菌、病毒、真菌和原生动物[14, 15]。在医学研究中，如果测序技术采用的是 16S rRNA 基因（又称为 rDNA），则 microbiota 是指细菌和古菌。Microbiome 是指整个微生境，包括微生物、基因组和周围环境[14, 15]。不过，microbiota 和 microbiome 有时存在混用情况。我们建议，如果你的研究仅涉及微生物本身，则应该使用 microbiota，否则应该使用 microbiome（图 1）。例如，如果研究者想探索肠道短链脂肪酸与微生物的关系，使用 microbiome 更合适。宏基因组（metagenome）是指微生物基因组的集合[14]，一般用鸟枪法宏基因组测序获得，宏基因组学则是研究宏基因组的学科[12, 14]。病毒组（virome）指人体内或表面的病毒集合，包括内源性逆转录病毒、真核生物病毒和噬菌体[16]。研究病毒组的学科就是病毒组学。*作者注：Microbiota 国内有些学者翻译为“微生物群”，microbiome 翻译为“微生物组”。不过中文文献用“微生物组”或“××菌群”即可，多数情况下不需要区分是 microbiota 或 microbiome*。

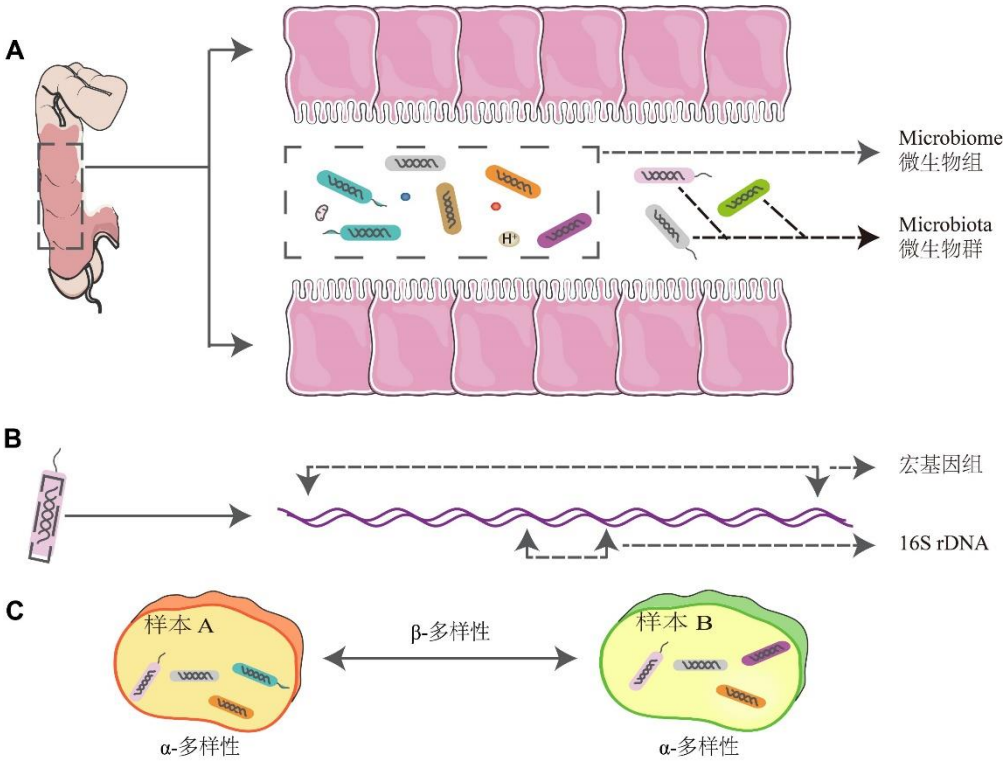

图 1：微生物组、微生物群、宏基因组和 16S rDNA 的概念。（A）微生物组（microbiome）的概念不仅涵盖微生物，而且涵盖周围的环境条件。微生物群（microbiota）仅指微生物本身。（B）宏基因组是指微生物的所有基因组，而 16S rDNA 仅涵盖基因组的一部分。（C）α 多样性衡量样本中的多样性，而 β 多样性比较样本之间的物种差异。

2.2 细菌层级分类

细菌分类最常用的层级为门、纲、目、科、属、种、株。例如，临床上十分常见的大肠埃希菌的层级分类见表 1。

表 1: 大肠埃希菌细菌层级分类

分类层级	分类名称
门	变形菌门
纲	丙型变形菌纲
目	肠杆菌目
科	肠杆菌科
属	*埃希氏杆菌属*
种	*埃希氏菌*
株	EIEC112ac 株

2.3 操作分类单元和扩增子序列变异

操作分类单元（operational taxonomic units，OTUs）的构建对于标记基因（扩增子）数据分析非常重要[17]。OTU 是指一组高度相似的序列，通常将具有 97%相似性的一组序列归为一个 OTU[18, 19]。不过，这种 OTU 的方法有显著的缺点，它人为地设置一个相似性阈值，漏掉了细微的和真正的生物学序列差异[20]。最近开发的扩增子序列变异（amplicon sequence variants, ASVs）方法可以解决这些问题，它使用序列变异信息将序列数据解析为准确的序列特征。ASV 具有单核苷酸分辨率，并且具

有比 OTU 相似或更好的敏感性和特异性[20]。注意，OTU 或 ASV 不等于物种，一个 OTU / ASV 可能包括多个物种，反之亦然[21]。

2.4 α-多样性

α-多样性是指样本内的多样性，常见的样本有粪便，唾液或支气管肺泡灌洗液等[15]。医学研究中经常使用 3 种 α 多样性指数：Chao 1 指数、香农指数和辛普森指数。Chao 1 指数主要反映物种数量（richness），它计算时考虑以下三个因素：物种数量、单条序列数量和双条序列数量[22]。这意味着它不能反映微生物组的丰度（abundance）。香农指数结合了丰度和均匀度信息[23]，它赋予稀有物种更多的权重[22]，这意味着当稀有物种的数量增加时，它的值会更大。香农指数的值通常不超过 5.0；它的值越高，α 多样性就越丰富[22]。辛普森指数也整合了丰度和均匀度，不过与香农指数比较，计算时它对常见物种有更大权重。它的值介于 0-1 之间，这个值越大，α 多样性越丰富[22]。在以上指数中，richness 是指一个样本中物种的数量[17, 24]，而 abundance（丰度）指物种的原始序列读数[24]。如果原始序列读数被转换成百分比后，它就称为相对丰度。

2.5 β-多样性

β-多样性是指样本或组间的微生物组差异，通常用于了解两组微生物组组成的差异是否显著。在这里，我们关注两个常用的 β 多样性指数：Bray-Curtis 相异性和 UniFrac 距离。Bray-Curtis 相异性是一种用于量化两个样本或组间的物种组成差异的指标，其值的范围是 0 到 1，其中 0 表示两个样本或组间具有相同物种，而 1 则表示它们不共享任何物种[25]。此外，它在计算时给予常见物种更大的权重[23]。请注意，Bray-Curtis 相异性不是真正的距离度量指标，因此用“Bray-Curtis 相异性”的叫法比“Bray-Curtis 距离”更恰当[22]。

UniFrac 距离可以不加权，也可以加权，它基于系统发育距离估算微生物组样本或组间的差异[26]。未加权的 UniFrac 距离只考虑了物种是否存在，它对于检测稀有物种的数量变化很敏感，但是在计算中忽略了丰度信息[27]。加权 UniFrac 距离计算时纳入了丰度信息[28]，并减少了稀有物种的权重[29]。

2.6 排序

排序用于探索数据结构，由降维后的正交轴图形表示。排序图是可视化 β 多样性的有效方法。排序可以分为 2 大类：非约束排序和约束排序[30-32]。如果图形上的点不受环境因素（样本元数据）的约束，这种排序叫做非约束排序，否则叫约束排序[32]。常用的非约束标准包括主成分分析（principal component analysis, PCA）、对应分析（correspondence analysis, CA）、主坐标分析（principal coordinate analysis, PCoA）和非度量多维标度（non-metric multidimensional scaling, NMDS）[30, 32]。常用的约束排序有冗余分析（redundancy analysis, RDA）和典范对应分析（canonical correspondence analysis, CCA）[31, 32]。

微生物组信息是高维数据。PCA 通过将数据以几何方式投影到较少的维度上来简化复杂性，它在计算中使用欧几里得（Euclidean）距离[30]。通常情况下它并不适用于物种丰度数据的分析，因为 PCA 分析的数据必须是线性的[30]。但是如果物种数据经过 Hellinger 转换，则 PCA 可以用于物种数据分析[30]。相反，CA 适合于物种丰度数据分析，而且无需预先转换数据。在 CA 分析中，所有样本均使用 Pearson 卡方距离进行排序[30]。但是请注意，稀有物种可能会对 CA 分析产生过大影响[33]。如果研究人员希望基于相异性指标来对样本或特征进行排序，那么 PCoA 是一个不错的选择。在微生物组研究中，PCoA 分析最常使用 Bray-Curtis 相异性和 UniFrac 距

离。NMDS 用于表示排序图中样本的相对位置。与 PCoA 相似，NMDS 分析可以使用任何距离或相异矩阵。参考文献[30]详细介绍了 PCoA 和 NMDS 之间的差异，在大多数情况下 PCoA 比较常用。

RDA 是一种结合了 PCA 和回归的约束排序，它的响应矩阵是微生物组数据，解释矩阵是临床指标（样本元数据）。RDA 对于显示微生物组数据是否受临床指标影响很有用。但是请注意，由于 PCA 计算过程要求响应矩阵的数据结构必须是线性的，因此可能需要对数据进行预转换。最后，CCA 其实就是 CA 的约束版本，它具有 CA 的基本特性和缺点[31]。

3. 研究设计

3.1 研究设计方案

严谨的研究设计对于获得准确而有意义的结果很重要。医学微生物组研究中最常使用的研究方法包括横断面研究、病例对照研究、纵向研究和随机对照试验（randomized controlled trial, RCT）。前 3 种是不应用干预因素的观察性研究，而最后一个是典型的实验性研究。

横断面研究分为描述性横断面研究和分析性横断面研究[34]。前者仅是描述性的，主要用于调查一个或多个人群中的微生物组成，而后者则用于探讨微生物组与健康结果之间的关联。但是，微生物组与健康结果之间的关联可能源于混杂因素，例如性别[35]、年龄[36]、体重指数（body mass index, BMI）[37]、饮食[5, 38]、季节[39]和药物治疗[40, 41]。此外，横断面研究时，微生物组和结果是同时测量的，因此很难确定它们之间的因果关系。通常，横断面研究仅用于探索微生物组的基本特征，并且可以作为后续研究的初步实验。

在大多数情况下，微生物组被视为暴露（exposure），疾病被视为结局（outcome）。在这些假设下，传统的病例对照研究很少用于微生物组研究，因为以前的暴露（微生物组）信息很难获得。但是，如果暴露和结局对调，则可以使用病例对照研究设计方案。

同样，在上述假设下进行前瞻性队列研究也很困难，因为很难知道哪些微生物是潜在的暴露。而且，定义可用作暴露或非暴露因素的特定微生物组并非易事，因此难以将研究对象确定为暴露或非暴露个体。在实践中，有或没有疾病的个体通常归入研究组或对照组，然后在不同时间点前瞻性地收集含有微生物组的样本[17]。也就是说，前瞻性队列研究中的研究对象通常根据临床结局而不是特定的微生物组模式进行分组。

RCT 或其他实验研究的目的是评估干预措施的有效性。干预措施可以是药物或微生物组。例如，粪菌移植研究中的干预措施是微生物群[42, 43]。

值得注意的是，对照组的选择应恰当。以上这些研究设计中应注意匹配混杂因素，这部分内容将在下面讨论。有时对照的选择很困难，尤其是在临床研究中干预措施是微生物群本身的情况下。在这种情况下，如果其他研究设计不合适，那么进行有对照组的前后自身对照试验（controlled before-after trial）或历史对照试验将是一个不错的选择[44]。

3.2 定义纳入和排除标准

定义确切的纳入标准和排除标准可以使组间更好地匹配，并且有利于控制混淆因素，比如年龄[36, 45]、性别[35]、BMI[46]、饮食[47]、季节因素[39]、药物治疗[40, 41]、种族[48]、地理区域[45]和共存疾病等[7]。年龄可显著影响微生物组，对于那些小于 16 岁

的人更是如此[36, 45]。因此，对于涉及儿童的研究，年龄必须很好地匹配。饮食是另一个对微生物组改变有影响的因素，所以也要进行匹配[47]。为了增加组间的可比性，地理区域因素在研究设计时也需要考虑在内[45]。由于药物治疗对于微生物组有显著影响，所以入组前数月内接受过药物治疗的患者应该排除在外[41, 49]，这里讲的数月通常指入组前 3~6 个月[49]。

3.3 微生物组研究的样本量和检验效能计算

在进行研究设计时估计样本量大小非常重要。适当的样本量可使微生物组研究识别出组间的差异，并节省资源和时间。但是，样本量和检验效能计算对于研究者来说仍然是一个挑战[50]。微生物组研究中最常用的样本量和效能计算方法可以用 t 检验、方差分析、χ^2 检验和 Dirichlet 多项式模型[51]。以 t 检验为例，分 3 个步骤确定样本大小和效能计算。首先，通过初步实验获得少量扩增子数据。其次，使用 R 包 vegan 计算出每个样品的香农指数[52]。最后一步是使用 R 软件包 pwr 中的 power.t.test()函数计算样本量和效能。当研究者仅关注两组之间物种多样性的差异时，可使用 t 检验计算样本量和效能。在参考文献[51]中有样本量和效能计算的详细介绍。

3.4 阴性和阳性对照的重要性

微生物组研究的结果可能会受到多种因素的影响，例如 DNA 提取试剂盒、采样方法、污染和测序方法等[53]，不过可以通过使用阴性和阳性对照来减少这些影响。不幸的是，以前的研究中只有 30%报告使用了阴性对照，只有 10%报告使用了阳性对照[53]。使用对照对于准确认识微生物组非常重要，尤其是当样本的微生物含量较低时。以前的研究发现，过去被认为是无菌的标本（例如胎盘和关节液）可能会被微生物定植[54]。但是，这些阳性的结果可能是由其他因素导致的，例如污染。有趣的是，这些低生物含量标本在采用阴性和／或阳性对照后已被证明是无菌的[55]。因此，我们建议当样本为低生物含量样本（例如血液、羊水、脑脊液、关节液和胎盘等）时，应考虑使用阴性和阳性对照。值得注意的是，阴性对照和阳性对照在病毒学研究中也很重要，因为病毒和细菌通常是同时进行检测的[16]。此外，R 包 *decontam* 可用于鉴定和去除扩增子和宏基因组学数据中的污染物序列[56]。

3.5 测序方法的选择

微生物组研究中使用的测序方法包括扩增子测序、宏基因组测序和转录组测序。扩增子测序包括适用于细菌和古菌的 16S rDNA 测序以及适用于真菌的内部转录间隔区（internal transcribed spacer, ITS）测序。每种测序方法的优缺点在这 2 篇参考文献中有详细讨论[17, 57]。简而言之，扩增子测序很便宜，可应用于受宿主 DNA 污染的低生物含量标本，但一般仅能注释到“属”层级，并且易受某些固有偏倚来源的影响，例如 PCR 循环数[58]。宏基因组测序方法对样品中存在的所有 DNA 进行测序，包括细菌、病毒、真核生物和宿主的 DNA。它不仅将其分类学分辨率扩展到“种”或“株”的水平，而且还提供了潜在功能信息[17]。但是，扩增子和宏基因组测序方法都无法区分死微生物或活微生物[17]。转录组测序仅产生群落的活跃功能信息。鉴于这些测序方法的优缺点不同，建议将多种测序方法整合在一起以优化研究设计。简而言之，测序方法的选择主要取决于实验成本和样本质量。扩增子测序通常用于获得微生物群落的概况[59]，并且通常适用于大规模研究[6, 60]。如果您有足够的项目资金，并且想要获得菌株水平的分辨率和潜在功能，甚至想要恢复整个基因组，宏基因组测序是一种首选方法[61-65]。

3.6 提高研究可靠性的方法

简单的横断面研究在微生物组研究中的意义有限。在本小节中我们讨论了提高研究可靠性的方法。首先，首选纵向研究或 RCT 研究，而不是横断面研究或病例对照研究[17,66]。其次，应计算样本量[51]。第三，混淆因素应匹配，元数据（即各种临床指标等信息）应仔细收集。第四，应详细定义纳入和排除标准。例如，幼年特发性关节炎有几种亚型，每种亚型可能代表不同的疾病[67]。研究者应确定患者组中是否包括所有亚型。第五，最好考虑使用阴性和/或阳性对照[68]。第六，整合其他组学方法，例如代谢组学、转录组学和蛋白质组学，这对于全面了解微生物群落的结构和功能至关重要[17]。因此，应考虑获取微生物群落代谢物概况和 / 或其他多组学数据。目前，仅探索微生物群落结构的研究不被视为论证效率强的研究设计[17]。最后，建议在动物模型中验证从临床试验获得的初步结果。

表 2 列出了设计临床微生物组研究需要考虑的因素，图 2 展示了典型的工作流程。实验研究需要考虑的因素见参考文献[49]。

表 2：临床微生物组研究设计需要考虑的要素核对表

需要考虑的要素	核对详情
研究设计类型	☐横断面研究　☐病例对照研究　☐队列研究　☐RCT　☐其他：
性别	☐已匹配　☐未匹配　☐其他：
年龄	☐已匹配　☐未匹配　☐其他：
BMI	☐已匹配　☐未匹配　☐其他：
种族	☐已匹配　☐未匹配　☐其他：
地理区域	☐已匹配　☐未匹配　☐其他：
饮食	☐组间已经均衡并已记录：列出详细信息 ☐未记录
季节因素	☐样本收集自相同季节 ☐样本收集自不同季节
药物治疗	入组前使用了哪些药物？使用了多久？
纳入标准	☐已定义好　☐定义不清晰
排除标准	☐已定义好　☐定义不清晰
样本量	☐已计算　☐未计算
测序方法	☐扩增子　☐宏基因组　☐其他
阴性和 / 或阳性对照	☐有阴性对照　☐无阴性对照 ☐有阳性对照　☐无阳性对照
多组学方法	☐代谢组　☐转录组　☐蛋白组
样本类型	☐粪便　☐结肠灌洗液　☐腔内刷　☐组织钳出物　☐粘膜下组织 ☐关节液　☐尿液　☐牙菌斑　☐唾液　☐皮肤　☐其他：
动物模型验证	☐结果将在动物模型中验证 ☐结果不将在动物模型中验证

RCT：随机对照试验

4. 样本类型、保存和储藏

4.1 样本类型

人类微生物组研究的样本类型包括粪便、结肠灌洗液和腔内刷等（表 2）。样本类型的选择取决于感兴趣的研究假设。例如，粪便样本易于收集，可用于大规模和

纵向研究。另一方面，活检样本对于探索微生物群与宿主之间的相互作用更有用[69]。注意，在一项研究中应该固定采样位置，因为人体的不同部位~~寄生~~定植着不同的微生物群[70, 71]。

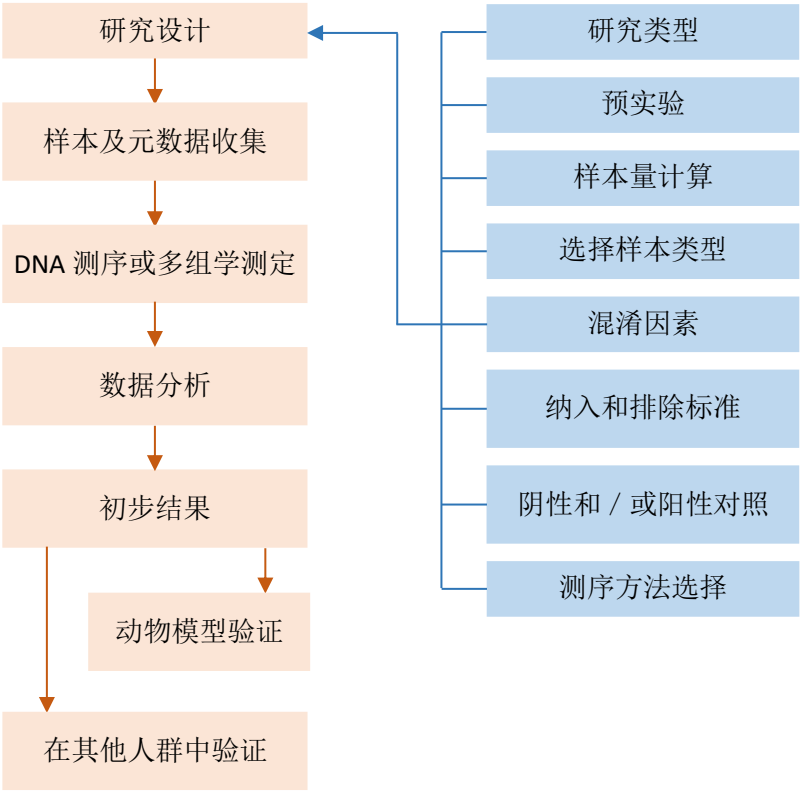

图 2：人类微生物组研究的典型流程。

4.2 保存和储藏

样品保存和储藏的方法应适合实验方法和样品类型。最通用的方法是直接冷冻样品，它可用于各种测序和实验方法，例如扩增子、宏基因组、转录组测序和代谢组学测定。建议将样品收集后 15 分钟内保存在-20℃下[72, 73]，然后在收集 24 小时内用干冰转移到-80℃冰箱中储藏。不过样本通常是在家里而不是在医院收集的，在这种情况下可以使用保存液。保存液中保存的样本可以在环境温度下保存一周以上[74]。请注意，样品的保存和储藏方法应一致，以最大程度地减少潜在的混淆因素干扰。

5. 微生物组研究中的统计分析方法

医学研究者通常熟悉单变量统计方法，例如 t 检验、方差分析、x^2 检验和秩和检验。因此，我们在这里仅讨论与多重比较和其他多元统计方法有关的问题。我们首先讨论多重比较会遇到的问题及其解决方案，包括 *P* 值调整和使用错误发现率。然后，我们讨论其他多元统计方法，例如置换多元方差分析（permutational multivariate analysis of variance, PERMANOVA）和 Mantel 检验。

5.1 多重比较的问题及解决方法

由于微生物组数据是高维的，因此多重比较经常在微生物组研究中使用。例如，特征表（feature table）具有成百上千个 OTU 或 ASV，并且每个 OTU 或 ASV 都可

以进行多次比较。医学研究者经常遇到的另一个例子可能更容易理解。假设一项研究分为 3 组，例如 A 组、B 组和 C 组，而研究者想比较这 3 组之间的差异。在这种情况下就应调整 P 值，因为每个组都进行了 2 次比较，即 A 组与 B 组，A 组与 C 组，B 组与 C 组。如果有任何组或变量需要进行多次比较则必须进行 *P* 值调整，以便减少假阳性率[75]。

调整 *P* 值的经典方法是控制 family-wise 错误率，即 I 类错误或 α 水平。Bonferroni 是校正 α 水平最常用的方法。校正 *P* 值的计算非常容易：单个检验的 α 值除以检验次数。因此，对于上述具有 3 个检验次数的例子，调整后的 P 值为 0.05 / 3 = 0.017，即只有 *P* <0.017 的检验结果才被认为是有意义的[75]。请注意，Bonferroni 校正仅适用于多重比较次数较少的假设检验，否则会导致较高的假阴性率（图 3）[75]。

解决多重比较问题的另一种方法是控制错误发现率（false discovery rate, FDR），它是 I 类错误或假阳性的数量与所有被拒绝的无效假设的预期比例。例如，如果 100 个阳性假设检验结果中有 5 个是错误发现，则 FDR 为 5%。在微生物组研究中，通常使用“Benjamini-Hochberg（BH）校正的 *P* 值”而不是原始 *P* 值。校正后的 *P* = 原始 *P** m/i，其中 m 是检验次数，i 是每个 *P* 值从小到大排序的序号[75]。如果校正后的 *P* 值小于你选择的所选 FDR，则认为该检验是有统计学意义的。与 Bonferroni 方法相比，BH 方法不那么保守（即校正强度不是很大），BH 法通常用于微生物组特征的多重比较。Bonferroni 和 BH 是最常用的 *P* 值校正方法[76]，这两种 *P* 值校正方法的校正强度见图 3 所示。

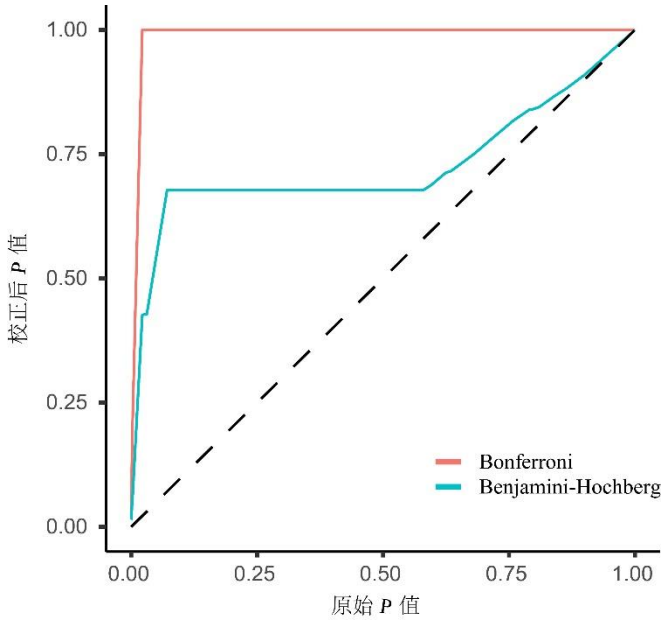

图 3：不同 *P* 值校正方法的校正强度。该图显示，Benjamini-Hochberg 校正强度小于 Bonferroni。随着原始 *P* 值的增加，Bonferroni 校正法生成的校正后 *P* 值快速接近 1.0。

5.2 PERMANOVA 检验

有几种统计方法或模型可以用于组间 β 多样性比较，比如 PERMANOVA、Mantel 检验、相似性分析（ANOSIM）和多响应置换程序（multi-response permutation procedures, MRPP）。PERMANOVA 最常用，并且被认为是以上检验方法中检验效能最大的一种[77]，它可通过 R 包 vegan 中的函数 adonis()实现[52]。vegan 包可计算 4 种

相异性或距离度量：Bray-Curtis 相异性、Jaccard 距离以及加权和未加权 UniFrac 距离[29]。如果 PERMANOVA 检验的 *P* 值小于 0.05，则表明不同组间的 β 多样性差异具有统计学意义；该检验的另一个输出结果是 R^2，它表示总方差可以用分组因素来解释的比例[29]。

5.3 Mantel 检验

Mantel 检验通常用于分析元数据矩阵和微生物组矩阵之间的关联[77]，它可使用 R 包 vegan 中的 mantel()函数实现[52, 77]。该检验的输出至少 2 个主要统计量：*P* 值和 r。与其他类型的相关系数类似，r 的值范围是-1～+1[29]。例如，假设研究人员想知道元数据种的分组因素（例如吸烟状态）是否对肠道微生物组的组成产生影响。如果 $P<0.05$ 并且 $r>0$，这表明吸烟组和不吸烟组之间肠道微生物组的组成不同，元数据矩阵和微生物组矩阵呈正相关。

6. 生物信息分析

6.1 扩增子数据分析：从原始数据到物种分类表

有几种流行的软件或分析流程（pipeline）可用于扩增子数据分析，例如 QIIME 2[13]、USEARCH[78]、VSEARCH[79]和 mothur[80]。前两者具有许多优点，并已被许多研究者使用和推荐。每种软件或分析流程的优缺点已在我们先前的论文中详细描述[81]。扩增子分析的主要步骤见图 4A。我们通常从 fastq 格式的原始双端 Illumina 数据开始，最终输出是一个特征表，也称为 OTU 表或 ASV 表。

第一步是从原始数据中恢复纯净的扩增子序列，因为原始数据包括人工产物（artifact），例如引物和标签（barcode）。它包括 3 个主要过程：合并双端序列，通过标签拆分序列和去除引物。由于原始数据没有统一的标准格式，因此我们需要设计适合上述过程的分析流程。另外，我们也可以使用基因测序公司提供的纯净扩增子数据。图 4B 显示了用于恢复纯净扩增子序列的典型分析流程。

第二步是滤除低质量序列，以便减少背景“噪音”。

第三步是识别非冗余序列并且计数。高质量序列仍然有许多人工产物，例如错误序列和嵌合体。非冗余序列的计数是找出可靠序列的关键信息。

第四步是选择代表性序列（特征）。此步骤基于唯一序列，并通过将序列聚类成 OTU 或降噪生成 ASV 来实现[18, 82]。此步骤还包括 de novo 检测和去除嵌合体。

第五步是有参嵌合体检测，这是第四步的替代选项[83]。通过将序列比对到数据库中，例如 rRNA 数据库 SILVA[84]，可以进一步过滤特征序列。应当指出的是，该步骤可以降低假阳性率但易于导致假阴性结果。

最后，通过将纯净的扩增子数据与特征序列进行比较来生成特征表（图 4A）。然后使用基于 RDP[85]或 Greengenes[86]数据库的分类器实现特征序列的物种分类。此外，基于 16S rRNA 基因谱，使用 PICRUSt[87, 88]、FAPROTAX[87, 89]和 BugBase[90]等工具可实现功能预测。

6.2 宏基因组分析：从原始数据到物种和功能分类表

扩增子测序仅产生分类学信息，而且 PCR 过程很容易产生偏倚和嵌合体[83]。鸟枪宏基因组测序比扩增子测序提供更详细的基因组信息和更高的分类学分辨率[66]。与扩增子方法相比，宏基因组学分析更为复杂，但是它提供了更准确的物种分类、多维度的功能信息和无法培养微生物的基因组草图。宏基因组分析流程如图 4C 所示。

第一步是预处理原始序列数据。原始数据包含低质量的污染序列以及与宿主相关序列。我们可以使用 FastQC 软件（http://www.bioinformatics.babraham.ac.uk/projects/fastqc/）进行数据质量检查，然后使用 KneadData 流程进行质量控制[91]并去除宿主 DNA[92]。有关更多 KneadData 的信息，请访问 http://huttenhower.sph.harvard.edu/kneaddata。

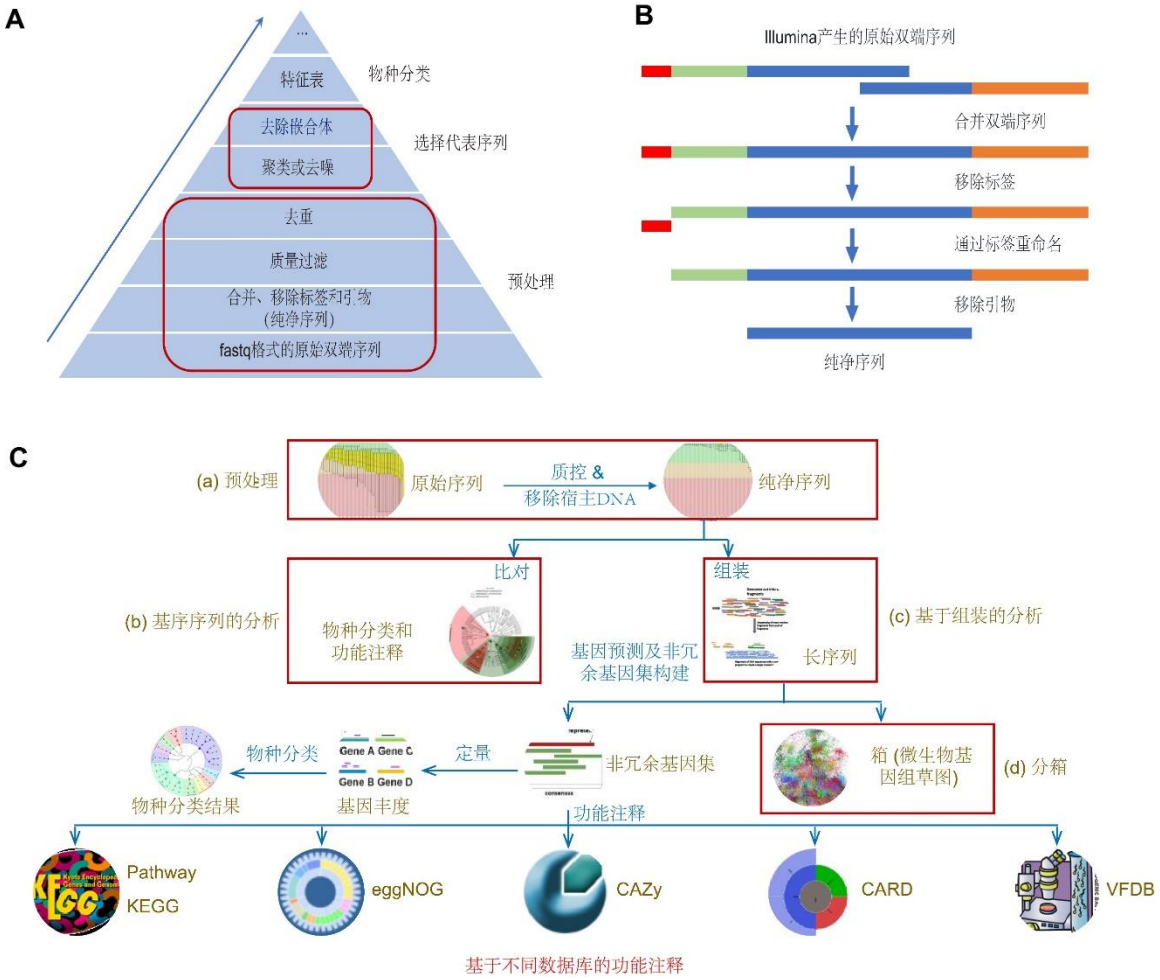

图 4：人类微生物组研究的生物信息学分析流程。（A）扩增子数据分析的主要步骤。（B）扩增子数据预处理的典型流程图：从原始的双端序列到纯净的扩增子。（C）宏基因组测序数据的分析流程。（a）预处理。它涉及删除低质量序列、接头和宿主序列。输出文件是纯净序列。（b）基于序列的分析。它将序列与数据库比对来推断物种分类和代谢特征。（c）基于组装的分析。它将短序列组装为长序列，预测基因，构建非冗余基因目录，并与数据库比对进行物种分类和功能注释。（d）分箱。它涉及恢复无法培养微生物的基因组草图，并重建系统发育和代谢通路。KEGG：京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes）；eggNOG：基因进化谱系：非监督直系同源群（Evolutionary genealogy of genes: non-supervised orthologous groups）；CAZy：碳水化合物活性酶数据库（Carbohydrate-active enzymes database）；CARD：抗性基因综合数据库（Comprehensive antibiotic resistance database）；VFDB：毒力因子数据库（Virulence factor database）。

第二步是使用基于序列的方法分析物种分类和代谢特征。人类微生物组具有高质量的基因目录（gene catalog）和基因组[64, 65]，因此我们建议使用 HUMAnN2[93]工

具并采用基于序列的方法进行物种分类和代谢通路分析，该方法高效且易于操作。但是，这种方法只使用一小部分序列信息，而且分析结果受到已知数据库的限制[66]。

如果需要发现新物种或基因功能，则需要进行第三步。有几个好的软件工具可以用于将纯净序列组装为长序列（contigs），例如 MEGAHIT[94]和 metaSPAdes[95]。然后通过 MetaProdigal[96]或 Prokka[97]从长序列中预测基因。另外，其他软件工具也可以用于从短序列中预测编码基因，例如 MetaGeneAnnotator[98]、MetaGeneMark[99]、Glimmer-MG[100]、MetaGUN[101]、FragGeneScan[102]和 Orphelia[103]。为了减少重复基因，在分析多个样品或批次时需要使用 CD-HIT 构建非冗余基因集[104]。通过采用 Bowtie 2[92]或 Salmon[105]工具进行比对的方法可以计算基因丰度。目前至少有 20 个软件工具可用于宏基因组数据物种分类[106]。我们建议使用超快速分类器 Kraken 2，它可以提供快速、准确和“种”级别的分类结果[107]。至于功能注释，许多研究人员都推荐使用 DIAMOND[108]，它是一种快速、敏感的蛋白质比对工具[108]。每个数据库都提供了独特的功能视角，例如，京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes, KEGG）[109]、EggNOG（一个提供直系同源关系、功能注释和基因进化历史的数据库）[110]、碳水化合物活性酶数据库（Carbohydrate-Active enZYmes Database, CAZy）[111]、致病菌的毒力因子（Virulence Factors of Pathogenic Bacteria, VFDF）[112]和综合抗生素抗性数据库（Comprehensive Antibiotic Resistance Database, CARD）[113]。宏基因组通常包含 100~1000 个物种[64]，很难厘清彼此关系。分箱算法可以恢复无法培养的高丰度菌的基因组草图，并重建系统发育和代谢通路。

最后一步是使用 metaWRAP[114]或 DASTool[115]执行分箱流程（图 4C）。这些软件工具有逐步操作教程，并且在其网站上提供了有关人类微生物组的一些样本数据集[81]。另外，几个集成的分析流程，例如 MOCAT 2[116]、bioBakery[98]、IMP[117]和微生物组助手（Microbiome Helper）[118]，可以执行上述部分或全部分析步骤。你可以在微信公众号“宏基因组”中找到最受欢迎软件的中文教程。

现在你已经获得了物种分类和功能信息文件。通过 STAMP 或 LEfSe 可以轻松找到你感兴趣的生物标记[119, 120]。使用 R 语言或 ImageGP（http://www.ehbio.com/ImageGP）可以将所有结果可视化。

7. 病毒组在人类疾病中的作用

近年来病毒组在人类疾病中的作用吸引了医学研究者的关注[121]。使用病毒组学的方法已发现了许多令人信服的研究成果[122]，其中一些技术已经用于临床[123]。在微生物组研究中，病毒组学与其他多组学方法整合后显示出广阔的应用前景。但是，病毒组学研究仍然面临一些挑战。例如，至少 40％的病毒序列无法注释[124]。此外，病毒的测序结果容易受到背景噪音的影响[17]。最后，很难获得用于病毒组研究的商业化阳性对照，即病毒模拟群落[16]。

8. 总结和结论

本文讨论了用于微生物组研究的研究设计、样本收集、统计方法和生物信息学分析方法。在“研究设计”部分，我们强调了研究设计的重要性，特别是设计方案、样本量计算以及用于提高研究可靠性的多种措施。研究设计非常重要，因为不好的研究设计可能会产生无意义的数据。在“统计分析”部分，我们介绍了详细的多重比较 *P* 值校正方法。选择合适的统计方法对于准确解释微生物组数据很重要。最后，“生物信息学分析”部分介绍了用于分析微生物组数据分析的方法。本文图中使用的脚本可从 https://github.com/YongxinLiu/Qian2020CMJ 获得。

综上所述，对于微生物组研究而言，严谨的研究设计在获得有意义的结果方面具有举足轻重的作用，而适当的统计方法对于准确解释微生物组数据非常重要。循序渐进的分析流程为研究者掌握最新生物信息学分析方法提供了帮助。

参考文献

1. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. Nature 2019; 569: 641-648. doi: 10.1038/s41586-019-1238-8.
2. NIH Human Microbiome Portfolio Analysis Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome 2019; 7: 31. doi: 10.1186/s40168-019-0620-y.
3. Xu Y, Zhao F. Single-cell metagenomics: challenges and applications. Protein Cell 2018; 9: 501-510. doi: 10.1007/s13238-018-0544-5.
4. Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vich Vila A, Vosa U, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat Genet 2019; 51: 600-605. doi: 10.1038/s41588-019-0350-x.
5. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science 2018; 359: 1151-1156. doi: 10.1126/science.aao5774.
6. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016; 48: 1396-1406. doi: 10.1038/ng.3695.
7. Wang Z, Xu CM, Liu YX, Wang XQ, Zhang L, Li M, et al. Characteristic dysbiosis of gut microbiota of Chinese patients with diarrhea-predominant irritable bowel syndrome by an insight into the pan-microbiome. Chin Med J (Engl) 2019; 132: 889-904. doi: 10.1097/CM9.0000000000000192.
8. Dong LN, Wang M, Guo J, Wang JP. Role of intestinal microbiota and metabolites in inflammatory bowel disease. Chin Med J (Engl) 2019; 132: 1610-1614. doi: 10.1097/CM9.0000000000000290.
9. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 2019; 569: 655-662. doi: 10.1038/s41586-019-1237-9.
10. Yang J, Yu J. The association of diet, gut microbiota and colorectal cancer: what we eat may imply what we get. Protein Cell 2018; 9: 474-487. doi: 10.1007/s13238-018-0543-6.
11. Chen X, Li HY, Hu XM, Zhang Y, Zhang SY. Current understanding of gut microbiota alterations and related therapeutic intervention strategies in heart failure. Chin Med J 2019; 132: 1843-1855. doi: 10.1097/CM9.0000000000000330.
12. Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. BMJ 2017; 356: j831. doi: 10.1136/bmj.j831.
13. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019; 37: 852-857. doi: 10.1038/s41587-019-0209-9.
14. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome 2015; 3: 31. doi: 10.1186/s40168-015-0094-5.

15. Gilbert JA, Lynch SV. Community ecology as a framework for human microbiome research. Nat Med 2019; 25: 884-889. doi: 10.1038/s41591-019-0464-9.
16. Santiago-Rodriguez TM, Hollister EB. Human Virome and Disease: High-Throughput Sequencing for Virus Discovery, Identification of Phage-Bacteria Dysbiosis and Development of Therapeutic Approaches with Emphasis on the Human Gut. Viruses 2019; 11: doi: 10.3390/v11070656.
17. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018; 16: 410-422. doi: 10.1038/s41579-018-0029-9.
18. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013; 10: 996-998. doi: 10.1038/nmeth.2604.
19. Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 2015; 3: e1487. doi: 10.7717/peerj.1487.
20. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 2017; 11: 2639-2643. doi: 10.1038/ismej.2017.119.
21. Liu YX, Qin Y, Bai Y. Reductionist synthetic community approaches in root microbiome research. Curr Opin Microbiol 2019; 49: 97-102. doi: 10.1016/j.mib.2019.10.010.
22. Xia Y, Sun J, Chen D. Community diversity measures and calculations. In: Xia Y, Sun J, Chen D, eds. Statistical Analysis of Microbiome Data with R. Singapore: Springer Singapore; 2018: 167-190.
23. Borcard D, Gillet F, Legendre P. Community diversity. In: Borcard D, Gillet F, Legendre P, eds. Numerical Ecology with R. Switzerland: Springer International Publishing; 2018: 369-412.
24. Xia Y, Sun J, Chen D. Introductory overview of statistical analysis of microbiome data. In: Xia Y, Sun J, Chen D, eds. Statistical Analysis of Microbiome Data with R. Singapore: Springer Singapore; 2018: 43-75.
25. Bray JR, Curtis JT. An ordination of the upland forest communities of southern wisconsin. Ecol Monogr 1957; 27: 326-349. doi: 10.2307/1942268.
26. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005; 71: 8228-8235. doi: 10.1128/AEM.71.12.8228-8235.2005.
27. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 2012; 28: 2106-2113. doi: 10.1093/bioinformatics/bts342.
28. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 2007; 73: 1576-1585. doi: 10.1128/AEM.01996-06.
29. Xia Y, Sun J, Chen D. Multivariate community analysis. In: Xia Y, Sun J, Chen D, eds. Statistical Analysis of Microbiome Data with R. Singapore: Springer Singapore; 2018: 285-330.
30. Borcard D, Gillet F, Legendre P. Unconstrained ordination. In: Borcard D, Gillet F, Legendre P, eds. Numerical Ecology with R. Switzerland: Springer International Publishing; 2018: 151-201.
31. Borcard D, Gillet F, Legendre P. Canonical ordination. In: Borcard D, Gillet F, Legendre P, eds. Numerical Ecology with R. Switzerland: Springer International Publishing; 2018: 203-297.
32. Xia Y, Sun J, Chen D. Exploratory analysis of microbiome data and beyond. In: Xia Y, Sun J, Chen D, eds. Statistical Analysis of Microbiome Data with R. Singapore: Springer Singapore; 2018: 191-294.
33. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. Oecologia 2001; 129: 271-280. doi: 10.1007/s004420100716.

34. Aryal S. Cross-Sectional Study. 2019. Available from: https://microbenotes.com/cross-sectional-study/. Last accessed on May 30, 2020.
35. Rizzetto L, Fava F, Tuohy KM, Selmi C. Connecting the immune system, systemic chronic inflammation and the gut microbiome: The role of sex. J Autoimmun 2018; 92: 12-34. doi: 10.1016/j.jaut.2018.05.008.
36. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao JZ, et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol 2016; 16: 90. doi: 10.1186/s12866-016-0708-5.
37. Sun L, Ma L, Ma Y, Zhang F, Zhao C, Nie Y. Insights into the role of gut microbiota in obesity: pathogenesis, mechanisms, and therapeutic perspectives. Protein Cell 2018; 9: 397-403. doi: 10.1007/s13238-018-0546-3.
38. Kolodziejczyk AA, Zheng D, Elinav E. Diet-microbiota interactions and personalized nutrition. Nat Rev Microbiol 2019; doi: 10.1038/s41579-019-0256-8.
39. Davenport ER, Mizrahi-Man O, Michelini K, Barreiro LB, Ober C, Gilad Y. Seasonal variation in human gut microbiome composition. PLoS One 2014; 9: e90731. doi: 10.1371/journal.pone.0090731.
40. Willmann M, Vehreschild M, Biehl LM, Vogel W, Dorfel D, Hamprecht A, et al. Distinct impact of antibiotics on the gut microbiome and resistome: a longitudinal multicenter cohort study. BMC Biol 2019; 17: 76. doi: 10.1186/s12915-019-0692-y.
41. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018; 555: 623-628. doi: 10.1038/nature25979.
42. Wang Y, Wiesnoski DH, Helmink BA, Gopalakrishnan V, Choi K, DuPont HL, et al. Fecal microbiota transplantation for refractory immune checkpoint inhibitor-associated colitis. Nat Med 2018; 24: 1804-1808. doi: 10.1038/s41591-018-0238-9.
43. Zhang F, Cui B, He X, Nie Y, Wu K, Fan D, et al. Microbiota transplantation: concept, methodology and strategy for its modernization. Protein Cell 2018; 9: 462-473. doi: 10.1007/s13238-018-0541-8.
44. Sedgwick P. Before and after study designs. 2014. Available from: https://www.bmj.com/content/349/bmj.g5074. Last accessed on May 30, 2020.
45. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012; 486: 222-227. doi: 10.1038/nature11053.
46. Haro C, Rangel-Zuniga OA, Alcala-Diaz JF, Gomez-Delgado F, Perez-Martinez P, Delgado-Lista J, et al. Intestinal Microbiota Is Influenced by Gender and Body Mass Index. PLoS One 2016; 11: e0154090. doi: 10.1371/journal.pone.0154090.
47. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014; 505: 559-563. doi: 10.1038/nature12820.
48. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med 2018; 24: 1526-1531. doi: 10.1038/s41591-018-0160-1.
49. Marques FZ, Jama HA, Tsyganov K, Gill PA, Rhys-Jones D, Muralitharan RR, et al. Guidelines for Transparency on Gut Microbiome Studies in Essential and Experimental Hypertension. Hypertension 2019; 74: 1279-1293. doi: 10.1161/HYPERTENSIONAHA.119.13079.
50. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol 2016; 17: 217. doi: 10.1186/s13059-016-1086-x.

51. Xia Y, Sun J, Chen D. Power and sample size calculations for microbiome data. In: Xia Y, Sun J, Chen D, eds. Statistical Analysis of Microbiome Data with R. Singapore: Springer Singapore; 2018: 129-166.
52. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2019. Available from: https://cran.r-project.org/web/packages/vegan/index.html. Last accessed on May 30, 2020.
53. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol 2019; 95: doi: 10.1093/femsec/fiz045.
54. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med 2014; 6: 237ra265. doi: 10.1126/scitranslmed.3008599.
55. de Goffau MC, Lager S, Sovio U, Gaccioli F, Cook E, Peacock SJ, et al. Human placenta has no microbiome but can contain potential pathogens. Nature 2019; 572: 329-334. doi: 10.1038/s41586-019-1451-5.
56. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018; 6: 226. doi: 10.1186/s40168-018-0605-2.
57. Rausch P, Ruhlemann M, Hermes BM, Doms S, Dagan T, Dierking K, et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 2019; 7: 133. doi: 10.1186/s40168-019-0743-1.
58. Sze MA, Schloss PD. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. mSphere 2019; 4: doi: 10.1128/mSphere.00163-19.
59. Wang J, Zheng J, Shi W, Du N, Xu X, Zhang Y, et al. Dysbiosis of maternal and neonatal microbiota associated with gestational diabetes mellitus. Gut 2018; 67: 1614-1625. doi: 10.1136/gutjnl-2018-315988.
60. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018; 24: 1532-1535. doi: 10.1038/s41591-018-0164-x.
61. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 2017; 357: 802-806. doi: 10.1126/science.aan4834.
62. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464: 59-65. doi: 10.1038/nature08821.
63. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature 2011; 473: 174-180. doi: 10.1038/nature09944.
64. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 2014; 32: 834-841. doi: 10.1038/nbt.2942.
65. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 2019; 176: 649-662.e620. doi: 10.1016/j.cell.2019.01.001.
66. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 2017; 35: 833-844. doi: 10.1038/nbt.3935.
67. Wu EY, Bryan AR, Rabinovich CE. Juvenile Idiopathic Arthritis. In: Kliegman RM, Stanton BF, St Geme III JW, Schor NF, eds. Nelson Textbook of Pediatrics. the United States: Elsevier; 2015: 1160-1170.

68. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol 2019; 27: 105-117. doi: 10.1016/j.tim.2018.11.003.
69. Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. Nat Rev Gastroenterol Hepatol 2017; 14: 585-595. doi: 10.1038/nrgastro.2017.97.
70. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. Nat Rev Microbiol 2016; 14: 20-32. doi: 10.1038/nrmicro3552.
71. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018; 24: 392-400. doi: 10.1038/nm.4517.
72. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. PLoS One 2015; 10: e0134802. doi: 10.1371/journal.pone.0134802.
73. Choo JM, Leong LE, Rogers GB. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 2015; 5: 16350. doi: 10.1038/srep16350.
74. Han M, Hao L, Lin Y, Li F, Wang J, Yang H, et al. A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. Microbiome 2018; 6: 43. doi: 10.1186/s40168-018-0429-0.
75. McDonald JH. Multiple tests. In: McDonald JH, ed. Handbook of Biological Statistics. Baltimore, Maryland, U.S.A.: Sparky House Publishing; 2014: 257-263.
76. Arnold T, Emerson J. The R Stats Package. 2019. Available from: https://www.rdocumentation.org/packages/stats/versions/3.6.1. Last accessed on May 30, 2020.
77. Xia Y, Sun J. Hypothesis Testing and Statistical Analysis of Microbiome. Genes Dis 2017; 4: 138-148. doi: 10.1016/j.gendis.2017.06.001.
78. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010; 26: 2460-2461. doi: 10.1093/bioinformatics/btq461.
79. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016; 4: e2584. doi: 10.7717/peerj.2584.
80. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009; 75: 7537-7541. doi: 10.1128/AEM.01541-09.
81. Liu Y, Qin Y, Guo XX, Bai Y. [Methods and applications for microbiome data analysis (in Chinese)]. Yi Chuan 2019; 41: 845-862. doi: 10.16288/j.yczz.19-222.
82. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016; 13: 581-583. doi: 10.1038/nmeth.3869.
83. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011; 27: 2194-2200. doi: 10.1093/bioinformatics/btr381.
84. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013; 41: D590-596. doi: 10.1093/nar/gks1219.
85. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 2014; 42: D633-642. doi: 10.1093/nar/gkt1244.

86. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012; 6: 610-618. doi: 10.1038/ismej.2011.139.
87. Zhang J, Liu YX, Zhang N, Hu B, Jin T, Xu H, et al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nat Biotechnol 2019; 37: 676-684. doi: 10.1038/s41587-019-0104-4.
88. Zheng M, Zhou N, Liu S, Dang C, Liu Y-X, He S, et al. N2O and NO emission from a biological aerated filter treating coking wastewater: Main source and microbial community. Journal of Cleaner Production 2019; 213: 365-374. doi: 10.1016/j.jclepro.2018.12.182.
89. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science 2016; 353: 1272-1277. doi: 10.1126/science.aaf4507.
90. Ward T, Larson J, Meulemans J, Hillmann B, Lynch J, Sidiropoulos D, et al. BugBase predicts organism-level microbiome phenotypes. bioRxiv 2017; 133462. doi: 10.1101/133462.
91. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114-2120. doi: 10.1093/bioinformatics/btu170.
92. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012; 9: 357-359. doi: 10.1038/nmeth.1923.
93. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods 2018; 15: 962-968. doi: 10.1038/s41592-018-0176-y.
94. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015; 31: 1674-1676. doi: 10.1093/bioinformatics/btv033.
95. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Research 2017; 27: 824-834. doi: 10.1101/gr.213959.116.
96. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 2012; 28: 2223-2230. doi: 10.1093/bioinformatics/bts429.
97. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014; 30: 2068-2069. doi: 10.1093/bioinformatics/btu153.
98. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 2008; 15: 387-396. doi: 10.1093/dnares/dsn027.
99. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 2010; 38: e132. doi: 10.1093/nar/gkq275.
100. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 2012; 40: e9. doi: 10.1093/nar/gkr1067.
101. Liu Y, Guo J, Hu G, Zhu H. Gene prediction in metagenomic fragments based on the SVM algorithm. BMC Bioinformatics 2013; 14 Suppl 5: S12. doi: 10.1186/1471-2105-14-S5-S12.
102. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 2010; 38: e191. doi: 10.1093/nar/gkq747.
103. Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 2009; 37: W101-105. doi: 10.1093/nar/gkp327.

104. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28: 3150-3152. doi: 10.1093/bioinformatics/bts565.
105. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 2017; 14: 417-149. doi: 10.1038/nmeth.4197.
106. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 2019; 178: 779-794. doi: 10.1016/j.cell.2019.07.010.
107. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. bioRxiv 2019; 762302. doi: 10.1101/762302.
108. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods 2015; 12: 59-60. doi: 10.1038/nmeth.3176.
109. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 2000; 28: 27-30. doi: 10.1093/nar/28.1.27.
110. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 2019; 47: D309-D314. doi: 10.1093/nar/gky1085.
111. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 2014; 42: D490-D495. doi: 10.1093/nar/gkt1178.
112. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Research 2019; 47: D687-D692. doi: 10.1093/nar/gky1080.
113. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research 2017; 45: D566-D573. doi: 10.1093/nar/gkw1004.
114. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 2018; 6: 158. doi: 10.1186/s40168-018-0541-1.
115. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 2018; 3: 836-843. doi: 10.1038/s41564-018-0171-1.
116. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 2016; 32: 2520-2523. doi: 10.1093/bioinformatics/btw183.
117. Narayanasamy S, Jarosz Y, Muller EE, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol 2016; 17: 260. doi: 10.1186/s13059-016-1116-8.
118. Comeau AM, Douglas GM, Langille MG. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems 2017; 2: doi: 10.1128/mSystems.00127-16.
119. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 2014; 30: 3123-3124. doi: 10.1093/bioinformatics/btu494.
120. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biology 2011; 12: R60. doi: 10.1186/gb-2011-12-6-r60.
121. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, et al. The Global Virome Project. Science 2018; 359: 872-874. doi: 10.1126/science.aap7463.

122. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579: 270-273. doi: 10.1038/s41586-020-2012-7.
123. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet 2019; 20: 341-355. doi: 10.1038/s41576-019-0113-7.
124. Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res 2017; 239: 136-142. doi: 10.1016/j.virusres.2017.02.002.