ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Method development for cross-study microbiome data mining: Challenges and opportunities

Xiaoquan Su [a,b,*], Gongchao Jing [b], Yufeng Zhang [a,b], Shunyao Wu [a]

[a] College of Computer Science and Technology, Qingdao University, Qingdao, Shandong 266071 China
[b] Single-Cell Center, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong 266101 China

交叉研究微生物组数据挖掘的方法开发：挑战与机遇

Method development for cross-study microbiome data mining: Challenges and opportunities

Computational and Structural Biotechnology Journal [6.018]

原文链接： https://doi.org/10.1016/j.csbj.2020.07.020

第一作者： Xiaoquan Su

通讯作者： Xiaoquan Su

主要单位： 青岛大学计算机科学与技术学院，中国科学院青岛生物能源与过程研究所单细胞中心

摘要

在过去的十年中，已经产生了大量的微生物组测序数据用来研究微生物的组成与环境之间的动态关联。如何准确、高效地解读大规模微生物组数据，并进一步发挥其优势，已成为当前微生物组研究的一个重要瓶颈。这篇综述重点介绍了整合分析多个研究的微生物组数据集的三个关键步骤，包括微生物组成分分析、数据整合和数据挖掘。通过介绍现有的生物信息学方法并讨论其局限性，本文展望了这三个步骤计算方法的发展机遇，并从多组学数据分析层面提出了可能的解决方案，以便从不同角度全面理解和快速研究微生物组，从而通过提供“微生物组数据空间”的更广阔视野来促进数据驱动的研究。

关键词

微生物组（Microbiome），鸟枪宏基因组（Shotgun metagenome），扩增子测序（Amplicon sequencing），数据挖掘（Data mining），微生物组搜索（Microbiome search），多组学数据（Multi-omics data）

前言

近年来，为研究微生物组与自然环境 [2, 3]，人体健康[4-7]，农业[8, 9]等的动态联系，已对大量微生物群落样本进行了测序。如何有效、全面地发现隐藏在大规模数据下的生物学故事已成为目前微生物组研究最本质的瓶颈之一[10, 11]。从序列比对和机器学习等通用算法，到微生物组的特定分析方法如 OTU（Operational Taxonomy Unit）聚类[12]和基于系统发育的菌群距离[13, 14]等，生物信息学工具的发展进步为解密微生物组数据带来了机遇。而从另一个角度来讲，大量的微生物组数据也带来了新的挑战，特别是在整合多个研究和平台产生的数据集[15]，样本之间的比较[16]以及通过训练大规模数据集进行状态或疾病的分类和预测[17, 18]这些问题上。

对多个交叉研究数据集进行 meta-analysis，能够产生稳定且可重复的结果是进一步研究和应用的基础[19-21]。其中，三个分析步骤（**图 1**）在处理微生物组大数据中起着至关重要的作用：***成分分析***，从序列中解码微生物组物种成分和功能组成（**图 1a**）；***数据整合***，整理、规范和统一现有数据集（**图 1b**）；***数据挖掘***，通过从整合数据中学习到的微生物组特征，对给定样本的状态进行识别和分类（**图 1c**）。通过分别回顾用于微生物组成分分析、数据整合和数据挖掘的计算方法和工具（**表 1** 和**表 2**），本文总结了这三个方面的挑战和机遇，并且通过多组学数据分析，对从不同角度全面理解和快速研究微生物群落，提出更具前瞻性的解决方案。

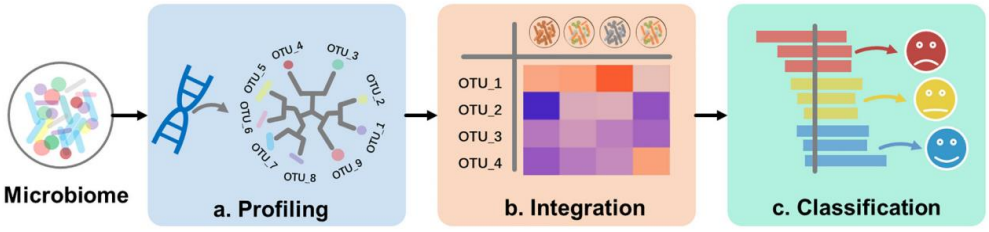

图 1. 微生物组大数据 meta-analysis 的关键步骤。（a）成分分析，从序列中解码微生物组物种成分和功能组成。（b）数据整合，整理、规范和统一现有数据集。（c）数据挖掘，通过从整合数据中学习到的微生物组特征，对给定样本的状态进行识别和分类。

微生物组组成分析

DNA 测序是目前用于解析微生物群落组成特征的主要方法。目前广泛使用

的有两种测序策略：扩增子测序（Amplicon Sequencing），用标记基因（如 16S rRNA, 18S rRNA, ITS 等）来实现物种的分类和鉴别；以及鸟枪法宏基因组全基因测序（whole-genome sequencing; WGS），可以获取样本中所有生物全部的遗传信息序列。

对于扩增子测序序列的物种组成分析，基于序列相似性的 OTU 聚类算法已广泛用于对微生物群落物种组成的分析，如 UPARSE[12]和 Usearch[23]等算法。而诸如 DADA2[24]、Deblur[25]和 UNOISE3[26]等扩增子序列变异（Amplicon sequence variants; ASVs）工具，进一步提高了扩增子序列在单核苷酸水平上的分析精度，比常规 OTU 的有更高的可靠性、可重复性和全面性。对于微生物组的功能解析，PICRUSt[28, 29]、Tax4Fun[30]等类似的软件可以利用扩增子标记基因和其已知的参考全基因组之间的联系，推测整个群落的功能组成。上述方法大多已被整合到集成的分析流程中，如 QIIME[31, 32]、Mothur[33]或者 Parallel-META3[34]，并增加了对微生物群落 α 和 β 多样性的统计分析。作为一个经济高效的方法，扩增子测序分析已经用于大规模微生物组的研究，但由于 PCR 扩增偏好性、标记物短序列的分辨率不足以及全基因组信息的缺失，所以准确性也受到限制。例如，对 16S rRNA 基因某些可变区进行测序所获得的短序列，其生物分类注释往往只能到“属”水平[36, 37]，而且对于缺乏参考全基因组信息的环境微生物物种，其功能推测也不尽人意。

由于全基因测序（WGS）的信息量更大，因此可以利用 WGS 短序列来进行“种”甚至“株”水平的物种分类注释[38, 39]（如 Karken [40]、mOTUs [41]、MetaPhlAn2 [42]）和功能解析（如 HUMANn2 [43]）。同时，基于分箱（binning）或拼装（assembling）的工具（如 metaSPAdes[44]、meta-IDBA[45]和 MetaWRAP[46]）能够进行物种基因组重建，从而实现新基因预测和单核苷酸多态性（Single Nucleotide Polymorphism; SNP）分析。然而由于 WGS 的成本较为高昂，包括测序、数据存储和共享、序列质量控制[47, 48]、分类和功能分析[38, 43] 等生物信息学处理，总成本比扩增子[28, 34, 49, 50]要高出 3-10 倍，目前也难以大规模地进行使用。近日，一种新的浅鸟枪测序（shallow shotgun sequencing）策略，通过较少的测序序列，获得近似于常规 WGS 测序的“种”水平的微生物组结构和功能组成解析，从而能够以更经济的方式来获得宏基因组测

序序列[51]。

不同于常规扩增子测序仅针对 16S rRNA 基因的某些可变区，PacBio 或 Oxford Nanopore 测序平台可以对 16S rRNA 基因进行全长测序，可以将微生物群落的结构解析分辨率提高到“种”甚至于“株”的水平 [52]。与此同时，随着越来越多的全长 16S rRNA 基因序列和其完整参考基因组的发布[53]，将扩增子测序所获得的标记基因比对到统一的参考数据库上，可以在更加广泛范围内进行高分辨率的微生物组分析。为了结合长序列测序平台数据的优势，新的序列降噪、聚类、注释算法及策略也同样应该更新以适应新数据的特点。综上所述，微生物组分析方法的快速发展为人们更广泛地了解“微生物数据世界”提供了基础。

数据仓储和数据整合

目前，在微生物组相关的项目和研究中产生了数量巨大的数据集，这些样本大多被存储于在线的数据仓储中，如 NCBI-SRA[57]、MG-RAST[58]、EBI Metagenomics[59]、JGI-IMG/M[60]以及 MPD[61]等。这些庞大的数据为全球微生物多样性和分布研究提供了素材，但也给数据整合和再利用等方面带来了新的问题。在这些数据仓储中，大多样本是按其来源的研究或项目进行组织管理，并存储其原始或质控过的 DNA 序列，而且其元数据（meta data）中采样信息的命名和记录也并不完整和统一，从而很难寻找或获取来自特定条件下的或具有某些结构功能特征的微生物组样本。

为了重新利用这些宝贵的微生物组大数据以进行进一步的分析和比较，许多工作重新整理了具有统一格式的元数据[62, 63]，并利用标准操作流程（Stand Operating Procedures; SOP）对微生物组样本的测序数据进行重新分析处理。GMrepo[65]是一个组织良好且经过精心整理的人类肠道宏基因组的数据库，具有统一注释的元数据。GcMeta[66]拥有一个数据管理系统，该系统与数据分析工具和工作流集成在一起，能够以标准化的方式存储和发布数据。Qiita[67, 68]允许用户跨研究进行 meta-analysis，并利用类似 SQL 的检索，查找包含特定特征（例如元数据、物种分类信息和序列片段等）的微生物组样本。

然而，当产生了新的微生物组样本的测序数据时，仍然很难回答在目前的数据仓储或数据库中，是否已经存在与该样本在群落结构整体上很相似的样本，进

而根据这些已有样本的采样信息，对新样本的环境条件或健康状况等特点进行预测。为了解决这个问题，科研人员研发了微生物组搜索引擎（Microbiome Search Engine; MSE）[69]，用于快速的“群落对群落”的比较和匹配。通过动态索引策略和一系列微生物组整体水平的相似性计算方法[70, 71]，MSE 实现了在海量数据中对具有特定结构的目标微生物组的实时级搜索访问。

对交叉研究的微生物组数据集进行整合的另一个障碍是不同来源、不同批次的扩增子测序数据之间的技术差异。技术因素，如 DNA 提取方法、标记基因 PCR 引物的选取、标记基因扩增的区域、测序平台和测序类型等，会显著影响数据集之间的比较[72]。对于具有较大效应量的生物问题（例如对来自于多个栖息地的环境微生物组，或者来自于不同身体部位以及不同年龄、地域和具有不同饮食习惯宿主的人类的微生物组进行比较），技术差异可以将测序序列与参考 16S rRNA 基因进行比对来抵消[73, 74]（例如，将短序列段映射到全长 16S rRNA 来实现 referenced-based OTU picking），使交叉研究整合变得有意义。然而，对更细微影响的生物问题进行研究，仍然要求按照统一的标准和实验方案来产生所有的扩增子数据集。相比之下，鸟枪法 WGS 数据在研究微生物组的疾病关联和时间序列动态变化方面对技术差异的敏感度要低[19, 75]，对于交叉研究数据的集成和比较也是一种值得考虑的替代选择。

数据挖掘以进行状态识别和分类

正是因为微生物群落与生态系统紧密相连，微生物组具有很强的潜力能够将菌群成分的变化与其表型和生理状态联系起来，从而促进疾病诊断、生态失衡的检测、治疗效果评估等领域中新技术的发展。先前的研究已经证明了机器学习方法如 XGBoost，随机森林（Random Forest; RF），支持向量机（Support Vector Machine; SVM），K-最近邻（K-nearest Neighbor; KNN）等利用人体微生物组数据实现疾病检测和分类的可行性，并应用于炎症性肠病（Inflammatory Bowel Disease）[77]、结直肠癌（Colorectal Cancer）[19]和龋齿（Caries）[78]等。作为一种定量方法，基于机器学习计算出的微生物组相关指数还能够用于评估潜在疾病的风险，并评估不同治疗方法之间的效果[79, 80]。

通常来讲，基于微生物组的检测必须对给定样本的特定状态（如疾病）作出

预先假设，并寻找出在疾病样本和对照样本之间分布不同的结构或功能特征（如物种或基因）作为生物标记，然后用这些标记物训练和构建机器学习模型以进行疾病识别。由于在这种模型中检测范围仅限于给定的状态类型，因此很难广泛地确定样本是否健康。此外，由于人群中微生物组数据的异质性，将疾病的特定模型扩展到其他人群，在可行程度上非常具有挑战[81]。此外，相同的生物标记可以与多种不同的疾病有关，这也可能导致多种疾病分类中的错误[82]。

近日，一种基于搜索的疾病检测和分类策略，将待检测样本在大规模的健康菌群库中进行搜索，实现其健康状态的检测。如果待检测样本与大量的健康菌群结构都不相似，便可能是由于其异常的健康状态造成的。接下来，该方法也通过在多种疾病样本中进行搜索比对，并根据菌群整体程度上的最佳匹配来对异常样本具体的疾病类型进行识别[83]。这种微生物组整体水平的搜索和匹配策略，在多人群、多测序平台以及数据存在污染的情况下也能鉴定微生物组状态。不足的是，目前该方法仅适用于扩增子序列，并且要求扩增子序列与 16S rRNA 参考数据库进行比对来获得 OTU。

卷积神经网络(Convolutional Neural Network ; CNN)、深度神经网络（Deep Neural Network; DNN）等深度学习方法的应用已从计算机视觉拓展到了微生物学领域[17]。通过支持并行计算的多核 CPU 和众核 GPU 的硬件提升，深度学习在大数据整合和对异构数据的鲁棒性方面表现出优势，但模型构建中的特定参数仍需要针对解决不同的问题而进行优化。TensorFlow（https://www.tensorflow.org/）和 PyTorch（https://pytorch.org/）包可通过 Python 轻松实现人工智能（AI）技术，从而推动了深度学习在微生物分析中分类识别[85]、生物标志物选择[86]、多疾病检测与分类[87]等方面的应用。深度学习在微生物组研究中的另一个潜力是多标签分类，该功能已广泛应用于图像处理[88]。目前对微生物组的疾病研究主要集中在单标签分类上，即单个样本只有一种特定的状态。然而在现实中单个微生物组样本，其来源的宿主可能同时患有多种不同的疾病[56, 89]，这种情况可以通过在微生物组领域进一步推广 AI 技术来解决。

多组学数据分析的前景

对“微生物群落中存在什么生物”和“微生物群落有什么功能”的研究不再足以

充分理解微生物组与环境之间的相互作用。尽管对 DNA 测序序列的分析能够获取微生物群落中的功能基因，但反映生物合成特征的细胞功能活性和基因表达以及代谢产物尚不清楚。微生物组的多组学数据分析利用化学和生物学方法提供了“微生物群落正在做什么”的全面视图，它从宏转录组学[91]、宏蛋白质组学[92]、宏代谢组学[93]和病毒学[94]进一步研究微生物群落。以前的一些工作已经证明了多组学数据在理解人类微生物组方面具有深入而独特的见解[95, 96]。然而，其产生的数据类型和计算工具大多是特定于组学的，例如用于宏基因组测序的软件与宏转录组学的 RNA-seq 数据以及代谢组学的质谱数据之间并不兼容，这使得多种工具的组合具有针对特定情况、不可扩展和不可复制的特点。IMP（Integrated Meta-omic Pipeline）工作流程可以来执行自动化、标准化和灵活性的分析，以整合宏基因组学和宏转录组学等多组学数据[97]。这种开放式开发框架策略增强了不同类型数据分析的集成以及从多个方面对结果的解释，并促进了微生物组多组学研究模式的发展。

然而，基于序列的分析目前在临床或产业应用中仍然未普及，主要原因之一是因为测序仪生成数据通常需要至少 2 天的时间[98]。目前，荧光激活细胞分选（Fluorescence-Activated Cell Sorting; FACS）方法，基于细胞中靶蛋白、代谢物或核酸的标记，能够实现细胞的快速功能性分选[99]。同时，基于拉曼激光的细胞分选（Raman-Activated Cell Sorting; RACS）方法，基于细胞成像，无需对细胞进行标签处理，不依赖于特定生物标记，就能实现微生物群落中单细胞精度的物种分类或状态鉴定[100, 101]。更重要的是，由于 FACS 或 RACS 只需花费几秒钟即可对每个细胞进行分析，因此该类技术可被视为以高通量和低时间成本监测微生物组的单细胞分辨率方法。

表 1. 微生物组数据分析的挑战与机遇

方法	主要挑战和局限性	机会和前景
微生物成分分析	**基于扩增子测序的组成分析** •生物分类注释往往只能到“属”水平 •功能分析的适用范围有限	**16S rRNA 基因全长序列** •将微生物群落结构解析分辨率提高到“种”甚至于“株”的水平 •增加扩增子标记基因和参考全基因组之间的联系 •将标记基因比对到统一的参考数据库和明确的系统发育树上，在更广泛范围内进行微生物组分析
	基于全基因组测序的组成分析 •高昂的测序成本 •组成分析在数据和计算上都很复杂	**浅鸟枪测序（shallow WGS）** •以近似扩增子测序的成本，获得“种”水平的微生物组结构和功能组成解析
数据整合	**通用数据仓储** •大多数数据仓储中只存储原始 DNA 序列 •缺少统一的元数据和注释 •很难寻找来自特定条件下的或具有某些结构功能特征的微生物组样本	**精心整理的数据库** •标准化测序质量控制 •统一的微生物结构分析和元数据注释 •查找包含特定特征的微生物组样本 **微生物组搜索引擎（Microbiome Search Engine）** •在整个微生物水平上进行“群落对群落”的比较和匹配 •实时级搜索访问
状态分类和预测	**机器学习** •很难广泛地确定微生物组样本是否健康 •在多标签分类问题上表现欠佳 •很难将疾病的特定模型扩展到其他人群	**基于搜索的策略** •不需要状态假设和生物标记 •对微生物组数据的异质性和被污染数据的鲁棒性 **深度学习** •为大数据训练模型提供更好的硬件和系统环境的支持 •多标签分类问题的优化 •开发完善的程序扩展包

表 2. 当前用于微生物组数据分析的工具

Tool name	Type	URL	Parallel computing	Installation	Reference
UParse	OTU clustering tool	https://drive5.com/uparse/	Multi-threads parallel computing	Binary package	[12]
Usearch	Integrated sequence analysis tool for amplicons (e.g. OTU clustering, denoising)	https://www.drive5.com/usearch/	Multi-threads parallel computing	Binary package	[23]
Vsearch	Alternative implementation of Usearch	https://github.com/torognes/vsearch	Multi-threads parallel computing	Source code / Binary package	[49]
DADA2	Amplicon sequence variants (ASVs) tools	https://benjjneb.github.io/dada2/	Multi-threads parallel computing	Bioconda / Source code / Binary package	[24]
Deblur	Amplicon sequence variants (ASVs) tools	https://github.com/biocore/deblur	Multi-threads parallel computing	Conda / Source code	[25]
UNOISE3	Amplicon sequence variants (ASVs) tools	http://www.drive5.com/usearch/manual/unoise_algo.html	Multi-threads parallel computing	Binary package	[26]
PICRUSt/PICRUSt2	Functional profiles prediction from amplified marker genes	http://picrust.github.io/picrust/	Multi-threads parallel computing	Bioconda / Miniconda / Source code / Online service (galaxy)	[28,29]
Tax4Fun	Functional profiles prediction from amplified marker genes	http://tax4fun.gobics.de/	Not appliable	R package	[30]
QIIME/QIIME2	Integrated microbiome bioinformatics workflow	http://qiime.org/https://qiime2.org/	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Conda / Miniconda / VirtualBox / Docker	[31,32]
Mothur	Integrated microbiome bioinformatics workflow	https://mothur.org/	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Binary package / Source code	[33]
Parallel-META3	Integrated microbiome bioinformatics workflow	http://bioinfo.single-cell.cn/parallel-meta.html	Multi-threads parallel computing	Source code	[34]
Karken	Taxonomical annotation of WGS short reads	http://ccb.jhu.edu/software/kraken/	Multi-threads parallel computing	Source code	[40]
mOTUs	Taxonomical annotation of WGS short reads	https://motu-tool.org/	Multi-threads parallel computing	Conda / Source code	[41]
MetaphlAn2	Taxonomical annotation of WGS short reads	https://huttenhower.sph.harvard.edu/metaphlan	Multi-threads parallel computing	Bioconda / Source code	[42]
HUMANn2	Functional annotation of WGS short reads	https://huttenhower.sph.harvard.edu/humann	Multi-threads parallel computing	Source code / Python-pip / Conda	[43]
metaSPAdes	Assembling of WGS short reads	https://github.com/ablab/spades	Multi-threads parallel computing	Source code / Binary package	[44]
Meta-IDBA	Assembling of WGS short reads	https://github.com/loneknightpy/idba	Multi-threads parallel computing	Source code	[45]
MetaWARP	Extraction and interpretation of high-quality metagenomic bins	https://github.com/bxlab/metaWRAP	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Conda / Bioconda / Docker / Source code	[46]
NCBI-SRA	Online general-purpose bio-data repository	https://www.ncbi.nlm.nih.gov/sra	Not appliable	Online service	[57]
MG-RAST	Online microbiome data repository	https://www.mg-rast.org/	Not appliable	Online service	[58]
EBI-Metagenomics	Online microbiome data repository	https://www.ebi.ac.uk/metagenomics/	Not appliable	Online service	[59]
JGI-IMG/M	Online microbiome data repository	https://img.jgi.doe.gov/	Not appliable	Online service	[60]
MPD	Pathogen genome and metagenome database	http://data.mypathogen.org	Not appliable	Online service	[61]
GMrepo	Curated database of human gut metagenomes	https://gmrepo.humangut.info/home	Not appliable	Online service	[65]
GcMeta	Integrated microbiome research platform	https://gcmeta.wdcm.org/	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Online service	[66]
Qiita	Online microbiome study management platform	https://qiita.ucsd.edu/	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Online service	[67,68]
MSE	Microbiome search engine	http://mse.ac.cn/	Multi-threads parallel computing	Online Service / Source code	[69]
TensorFlow	Open source platform for machine learning	https://www.tensorflow.org/	GPU parallel computing	Python-Pip / Docker / Source code	
PyTorch	Library for deep learning	https://pytorch.org/	GPU parallel computing	Conda / Python-pip / Source code	
IMP	Integrated *meta*-omic pipeline framework	https://r3lab.uni.lu/web/imp/	Partially with multi-thread parallel computing, depends on the specific tool in the pipeline	Conda / Docker / Source code	[97]

参考文献

1. Blaser, M.J., et al., *Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges.* mBio, 2016. **7**(3).
2. Bork, P., et al., *Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction.* Science, 2015. **348**(6237): p. 873.
3. Wu, L., et al., *Global diversity and biogeography of bacterial communities in wastewater treatment plants.* Nat Microbiol, 2019. **4**(7): p. 1183-1195.
4. Forslund, K., et al., *Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota.* Nature, 2015. **528**(7581): p. 262-266.
5. Halfvarson, J., et al., *Dynamics of the human gut microbiome in inflammatory bowel disease.* Nat Microbiol, 2017. **2**: p. 17004.
6. Poore, G.D., et al., *Microbiome analyses of blood and tissues suggest cancer diagnostic approach.* Nature, 2020. **579**(7800): p. 567-574.
7. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
8. Gao, P., et al., *Feed-additive probiotics accelerate yet antibiotics delay intestinal microbiota maturation in broiler chicken.* Microbiome, 2017. **5**(1): p. 91.
9. Zhang, J.Y., et al., *NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice.* Nature Biotechnology, 2019. **37**(6): p. 676-+.
10. Kyrpides, N.C., E.A. Eloe-Fadrosh, and N.N. Ivanova, *Microbiome Data Science: Understanding Our Microbial Planet.* Trends in Microbiology, 2016. **24**(6): p. 425-427.
11. Wood-Charlson, E.M., et al., *The National Microbiome Data Collaborative: enabling microbiome science.* Nat Rev Microbiol, 2020.
12. Edgar, R.C., *UPARSE: highly accurate OTU sequences from microbial amplicon reads.* Nat Methods, 2013. **10**(10): p. 996-8.
13. Lozupone, C. and R. Knight, *UniFrac: a new phylogenetic method for comparing microbial communities.* Appl Environ Microbiol, 2005. **71**(12): p. 8228-35.
14. Su, X., J. Xu, and K. Ning, *Meta-Storms: Efficient Search for Similar Microbial Communities Based on a Novel Indexing Scheme and Similarity Score for Metagenomic Data.* Bioinformatics, 2012.
15. Sinha, R., et al., *Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium.* Nat Biotechnol, 2017. **35**(11): p. 1077-1086.
16. Comin, M., et al., *Comparison of microbiome samples: methods and computational challenges.* Brief Bioinform, 2020.
17. Cammarota, G., et al., *Gut microbiome, big data and machine learning to promote precision medicine for cancer.* Nat Rev Gastroenterol Hepatol, 2020.
18. Goecks, J., et al., *How Machine Learning Will Transform Biomedicine.* Cell, 2020. **181**(1): p. 92-101.
19. Wirbel, J., et al., *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer.* Nature Medicine, 2019. **25**(4): p. 679-+.
20. Bisanz, J.E., et al., *Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet.* Cell Host Microbe, 2019. **26**(2): p. 265-272 e4.
21. Armour, C.R., et al., *A Metagenomic Meta-analysis Reveals Functional Signatures of Health*

and Disease in the Human Gut Microbiome. mSystems, 2019. **4**(4).

22. Knight, R., et al., *Best practices for analysing microbiomes.* Nat Rev Microbiol, 2018. **16**(7): p. 410-422.
23. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.
24. Callahan, B.J., et al., *DADA2: High-resolution sample inference from Illumina amplicon data.* Nat Methods, 2016. **13**(7): p. 581-3.
25. Amir, A., et al., *Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.* mSystems, 2017. **2**(2).
26. Edgar, R.C., *UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing*. bioRxiv, 2016: p. 081257.
27. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, *Exact sequence variants should replace operational taxonomic units in marker-gene data analysis.* Isme Journal, 2017. **11**(12): p. 2639-2643.
28. Langille, M.G., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nat Biotechnol, 2013. **31**(9): p. 814-21.
29. Douglas, G.M., et al., *PICRUSt2 for prediction of metagenome functions.* Nat Biotechnol, 2020. **38**(6): p. 685-688.
30. Asshauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.* Bioinformatics, 2015. **31**(17): p. 2882-4.
31. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nat Methods, 2010. **7**(5): p. 335-6.
32. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 (vol 37, pg 852, 2019).* Nature Biotechnology, 2019. **37**(9): p. 1091-1091.
33. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl Environ Microbiol, 2009. **75**(23): p. 7537-41.
34. Jing, G., et al., *Parallel-META 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities.* Scientific Reports, 2017. **7**: p. 40371.
35. Jones, M.B., et al., *Library preparation methodology can influence genomic and functional predictions in human microbiome research.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(45): p. 14024-14029.
36. Edgar, R.C., *Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.* PeerJ, 2018. **6**: p. e4652.
37. Yarza, P., et al., *Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences.* Nat Rev Microbiol, 2014. **12**(9): p. 635-45.
38. Ye, S.H., et al., *Benchmarking Metagenomics Tools for Taxonomic Classification.* Cell, 2019. **178**(4): p. 779-794.
39. Scholz, M., et al., *Strain-level microbial epidemiology and population genomics from shotgun metagenomics.* Nat Methods, 2016. **13**(5): p. 435-8.
40. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biol, 2014. **15**(3): p. R46.
41. Sunagawa, S., et al., *Metagenomic species profiling using universal phylogenetic marker*

genes. Nature Methods, 2013. **10**(12): p. 1196-+.

42. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nat Methods, 2012. **9**(8): p. 811-4.
43. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nat Methods, 2018. **15**(11): p. 962-968.
44. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
45. Peng, Y., et al., *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 2012. **28**(11): p. 1420-8.
46. Uritskiy, G.V., J. DiRuggiero, and J. Taylor, *MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis.* Microbiome, 2018. **6**(1): p. 158.
47. Zhou, Q., X. Su, and K. Ning, *Assessment of quality control approaches for metagenomic data analysis.* Sci Rep, 2014. **4**: p. 6957.
48. Zhou, Q., et al., *RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.* BMC Genomics, 2018. **19**(1): p. 144.
49. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4**: p. e2584.
50. Lu, J. and S.L. Salzberg, *Ultrafast and accurate 16S microbial community analysis using Kraken 2.* bioRxiv, 2020: p. 2020.03.27.012047.
51. Hillmann, B., et al., *Evaluating the Information Content of Shallow Shotgun Metagenomics.* Msystems, 2018. **3**(6).
52. Johnson, J.S., et al., *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.* Nat Commun, 2019. **10**(1): p. 5029.
53. Haft, D.H., et al., *RefSeq: an update on prokaryotic genome annotation and curation.* Nucleic Acids Res, 2018. **46**(D1): p. D851-D860.
54. Integrative, H.M.P.R.N.C., *The Integrative Human Microbiome Project.* Nature, 2019. **569**(7758): p. 641-648.
55. Thompson, L.R., et al., *A communal catalogue reveals Earth's multiscale microbial diversity.* Nature, 2017. **551**(7681): p. 457-463.
56. McDonald, D., et al., *American Gut: an Open Platform for Citizen Science Microbiome Research.* mSystems, 2018. **3**(3).
57. Kodama, Y., et al., *The Sequence Read Archive: explosive growth of sequencing data.* Nucleic Acids Res, 2012. **40**(Database issue): p. D54-6.
58. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**: p. 386.
59. Harrison, P.W., et al., *The European Nucleotide Archive in 2018.* Nucleic Acids Res, 2019. **47**(D1): p. D84-D88.
60. Chen, I.A., et al., *IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes.* Nucleic Acids Res, 2019. **47**(D1): p. D666-D677.
61. Zhang, T., et al., *MPD: a pathogen genome and metagenome database.* Database (Oxford), 2018. **2018**.
62. Yilmaz, P., et al., *Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.* Nature Biotechnology,

2011. **29**(5): p. 415-420.
63. Buttigieg, P.L., et al., *The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation.* J Biomed Semantics, 2016. **7**(1): p. 57.
64. Ten Hoopen, P., et al., *The metagenomic data life-cycle: standards and best practices.* Gigascience, 2017. **6**(8): p. 1-11.
65. Wu, S., et al., *GMrepo: a database of curated and consistently annotated human gut metagenomes.* Nucleic Acids Res, 2020. **48**(D1): p. D545-D553.
66. Shi, W., et al., *gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data.* Nucleic Acids Res, 2019. **47**(D1): p. D637-D648.
67. Gonzalez, A., et al., *Qiita: rapid, web-enabled microbiome meta-analysis.* Nat Methods, 2018. **15**(10): p. 796-798.
68. McDonald, D., et al., *redbiom: a Rapid Sample Discovery and Feature Characterization System.* mSystems, 2019. **4**(4).
69. Su, X., et al., *Identifying and Predicting Novelty in Microbiome Studies.* MBio, 2018. **9**(6).
70. Jing, G., et al., *Dynamic Meta-Storms enables comprehensive taxonomic and phylogenetic comparison of shotgun metagenomes at the species level.* Bioinformatics, 2019.
71. Su, X., et al., *GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU.* Bioinformatics, 2014. **30**(7): p. 1031-3.
72. Costea, P.I., et al., *Towards standards for human fecal sample processing in metagenomic studies.* Nat Biotechnol, 2017. **35**(11): p. 1069-1076.
73. Hacquard, S., et al., *Microbiota and Host Nutrition across Plant and Animal Kingdoms.* Cell Host Microbe, 2015. **17**(5): p. 603-16.
74. Lozupone, C.A., et al., *Meta-analyses of studies of the human microbiota.* Genome Res, 2013. **23**(10): p. 1704-14.
75. Voigt, A.Y., et al., *Temporal and technical variability of human gut metagenomes.* Genome Biol, 2015. **16**: p. 73.
76. Statnikov, A., et al., *A comprehensive evaluation of multicategory classification methods for microbiomic data.* Microbiome, 2013. **1**.
77. Gevers, D., et al., *The treatment-naive microbiome in new-onset Crohn's disease.* Cell Host Microbe, 2014. **15**(3): p. 382-392.
78. Teng, F., et al., *Prediction of Early Childhood Caries via Spatial-Temporal Variations of Oral Microbiota.* Cell Host & Microbe, 2015. **18**(3): p. 296-306.
79. Sun, Z., et al., *A Microbiome-Based Index for Assessing Skin Health and Treatment Effects for Atopic Dermatitis in Children.* mSystems, 2019. **4**(4).
80. Huang, S., et al., *Predictive modeling of gingivitis severity and susceptibility via oral microbiota.* ISME J, 2014. **8**(9): p. 1768-80.
81. Duvallet, C., et al., *Meta-analysis of gut microbiome studies identifies disease-specific and shared responses.* Nat Commun, 2017. **8**(1): p. 1784.
82. Jackson, M.A., et al., *Gut microbiota associations with common diseases and prescription medications in a population-based cohort.* Nat Commun, 2018. **9**(1): p. 2655.
83. Su, X., et al., *Multiple-Disease Detection and Classification across Cohorts via Microbiome Search.* mSystems, 2020. **5**(2): p. e00150-20.
84. Zitnik, M., et al., *Machine Learning for Integrating Data in Biology and Medicine: Principles,*

Practice, and Opportunities. Inf Fusion, 2019. **50**: p. 71-91.

85. Fiannaca, A., et al., *Deep learning models for bacteria taxonomic classification of metagenomic data.* BMC Bioinformatics, 2018. **19**(Suppl 7): p. 198.
86. Kather, J.N. and J. Calderaro, *Development of AI-based pathology biomarkers in gastrointestinal and liver cancer.* Nat Rev Gastroenterol Hepatol, 2020.
87. LaPierre, N., et al., *MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction.* Methods, 2019. **166**: p. 74-82.
88. Wei, Y., et al., *HCP: A Flexible CNN Framework for Multi-Label Image Classification.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016. **38**(9): p. 1901-1907.
89. He, Y., et al., *Regional variation limits applications of healthy gut microbiome reference ranges and disease models.* Nat Med, 2018. **24**(10): p. 1532-1535.
90. Bikel, S., et al., *Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome.* Comput Struct Biotechnol J, 2015. **13**: p. 390-401.
91. Bashiardes, S., G. Zilberman-Schapira, and E. Elinav, *Use of Metatranscriptomics in Microbiome Research.* Bioinformatics and Biology Insights, 2016. **10**: p. 19-25.
92. Kleiner, M., *Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities.* Msystems, 2019. **4**(3).
93. Abubucker, S., et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome.* PLoS Comput Biol, 2012. **8**(6): p. e1002358.
94. Garretto, A., T. Hatzopoulos, and C. Putonti, *virMine: automated detection of viral sequences from complex metagenomic samples.* PeerJ, 2019. **7**: p. e6695.
95. McHardy, I.H., et al., *Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships.* Microbiome, 2013. **1**(1): p. 17.
96. Franzosa, E.A., et al., *Relating the metatranscriptome and metagenome of the human gut.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. E2329-38.
97. Narayanasamy, S., et al., *IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses.* Genome Biol, 2016. **17**(1): p. 260.
98. Quinn, R.A., et al., *From Sample to Multi-Omics Conclusions in under 48 Hours.* mSystems, 2016. **1**(2).
99. Rinke, C., et al., *Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics.* Nat Protoc, 2014. **9**(5): p. 1038-48.
100. Ho, C.S., et al., *Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning.* Nature Communications, 2019. **10**.
101. Teng, L., et al., *Label-free, rapid and quantitative phenotyping of stress response in E. coli via ramanome.* Sci Rep, 2016. **6**: p. 34359.