·微生物组测序与分析专题·

**戴磊** 博士，中国科学院深圳先进技术研究院研究员、博士生导师，深圳合成生物学创新研究院微生物组研究中心主任。国家重点研发计划青年项目首席科学家，入选《麻省理工科技评论》中国区“35 岁以下科技创新 35 人”。担任中国生物工程学会合成生物学专业委员会委员，中国医药生物技术协会合成生物技术分会委员，粤港澳肠道微生态学术联盟理事，《合成生物学》期刊编委。实验室利用合成生物学的工具，对微生物组的结构和功能进行理性设计和精准调控，致力于解决人体健康、农业生产等重大问题。

# 基于宏基因组数据的菌株分析方法及其应用

谭宇翔[1*]，胡函[2*]，李陈浩[3]，罗小舟[1]，谭验[2]，戴磊[1]

1 中国科学院深圳先进技术研究院 深圳合成生物学创新研究院 中国科学院定量工程生物学重点实验室，广东 深圳 518055
2 深圳未知君生物科技有限公司，广东 深圳 518000
3 新加坡基因组研究院，新加坡 138672

谭宇翔, 胡函, 李陈浩, 等. 基于宏基因组数据的菌株分析方法及其应用. 生物工程学报, 2020, 36(12): 2610–2621.
Tan YX, Hu H, Li CH, et al. Research progress and applications of strain analysis based on metagenomic data. Chin J Biotech, 2020, 36(12): 2610–2621.

**摘 要:** 菌株是微生物研究中最基础的生命实体，其功能多样性对宿主表型有着重要影响。随着微生物组研究的深入，对复杂微生物群落进行菌株水平的构成分析和功能分析，在基础科研、临床应用等方面都有重要的价值。文中介绍了基于宏基因组数据的菌株分析的主流算法，以及菌株分析在微生物组研究中的潜在应用和未来的发展方向。

**关键词:** 菌株，宏基因组，单核苷酸多态性，基因多样性，参考基因组

**Received:** June 24, 2020; **Accepted:** November 9, 2020
**Supported by:** National Key Research and Development Project of China (No. 2019YFA09006700), National Natural Science Foundation of China (No. 31971513).
**Corresponding author:** Lei Dai. Tel: +86-755-86392436; E-mail: lei.dai@siat.ac.cn
*These authors contributed equally to this study.
国家重点研发计划 (No. 2019YFA09006700)，国家自然科学基金 (No. 31971513) 资助。
网络出版时间：2020-12-15 网络出版地址：https://kns.cnki.net/kcms/detail/11.1998.Q.20201214.1647.002.html

# Research progress and applications of strain analysis based on metagenomic data

**Yuxiang Tan[1*], Han Hu[2*], Chenhao Li[3], Xiaozhou Luo[1], Yan Tan[2], and Lei Dai[1]**

1 *CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *Guangdong, China*

2 *Shenzhen Xbiome Biotech Co., Ltd, Shenzhen* 518000, *Guangdong, China*

*3 Genome Institute of Singapore, Singapore* 138672

**Abstract:** Strain is the fundamental unit in microbial taxonomy. The functional diversity among strains has great influence on host phenotypes. With the development of microbiome research, knowing the composition and functional capacities of complex microbial communities at the strain level has become increasingly valuable in scientific research and clinical applications. This review introduces the principles of bioinformatics algorithms for strain analysis based on metagenomic data, the applications in microbiome research and directions of future development.

**Keywords:** strain, metagenomic, single nucleotide polymorphism, gene diversity, reference genome

人体体内和体表均定植着大量微生物，其数量与人体细胞数量相近，而其功能基因的数量更是人体功能基因的百倍有余，因此人体微生物组也被视为人类基因组信息的重要补充，与人体健康息息相关[1-2]。此外，众多研究表明，人体微生物群落 (尤其是肠道微生物群落) 的紊乱，会影响宿主的免疫、代谢、神经、内分泌等多系统的功能，也是自身免疫性疾病、肥胖、糖尿病、自闭症、抑郁症、肠炎和癌症等多种慢性疾病的重要因素之一[3-5]。由此可见，微生物的构成 (Composition) 多样性和功能 (Function) 多样性，对宿主 (人体) 的表型有着重要的影响。如何能深入研究微生物群落中的构成单元 (菌株，Strain) 和功能单元 (基因，Gene) 与表型 (Phenotype) 之间的相互关系，是当前微生物组领域的一个核心问题。基于宏基因组数据 (Metagenomics data) 的有参考基因组 (以下简称“有参”) 的菌株分析算法有望成为回答该问题的最佳高通量手段。

**(1) 菌株的功能多样性**

菌株是微生物分类和研究中的最基础生命实体[6]。同一菌种 (Species) 的不同菌株之间在代谢能力、定植能力、致病能力等表型上可能存在很大的差异。例如，大肠杆菌中只有特定的菌株才会产生致癌物[7]，而其他的则是非致病性的[8]。反过来，同一菌种中也只有特定的菌株被证明对抑制肿瘤有效[9]。这些性状/功能的多样性，往往是基因多样性 (Gene diversity) 和单核苷酸多态性(Single nucleotide polymorphism, SNP) 所导致的[10-11] (图 1)。其中，SNP 可能会影响同源基因的转录和翻译等过程，从而影响蛋白的功能并导致性状表型的变化。而基因多样性，则反映了同源基因的有无，直接导致了不同菌株间特定功能的有无。而为了更好地展示这一概念，2005 年 Tettelin 等提出了微生物泛基因组 (Pan-genomoe) 的概念[12]，用以指代某一物种全部同源基因的组成及序列信息，从而清晰反映某个同源基因在不同菌株间的出现情况。根据在菌株中出现的频次可分为：1) 核心基因组 (Core genome)，指所有菌株中都存在的基因；2) 非必需基因组 (Dispensable genome)，指只在部分菌株中存在的基因。

而一个微生物群落的某项表型，往往是群落里各菌株的相关功能的集合的体现。因此，群落里的菌株构成与其功能多样性，对整个群落的表型甚至宿主的表型都有着重大的影响 (图 1)。

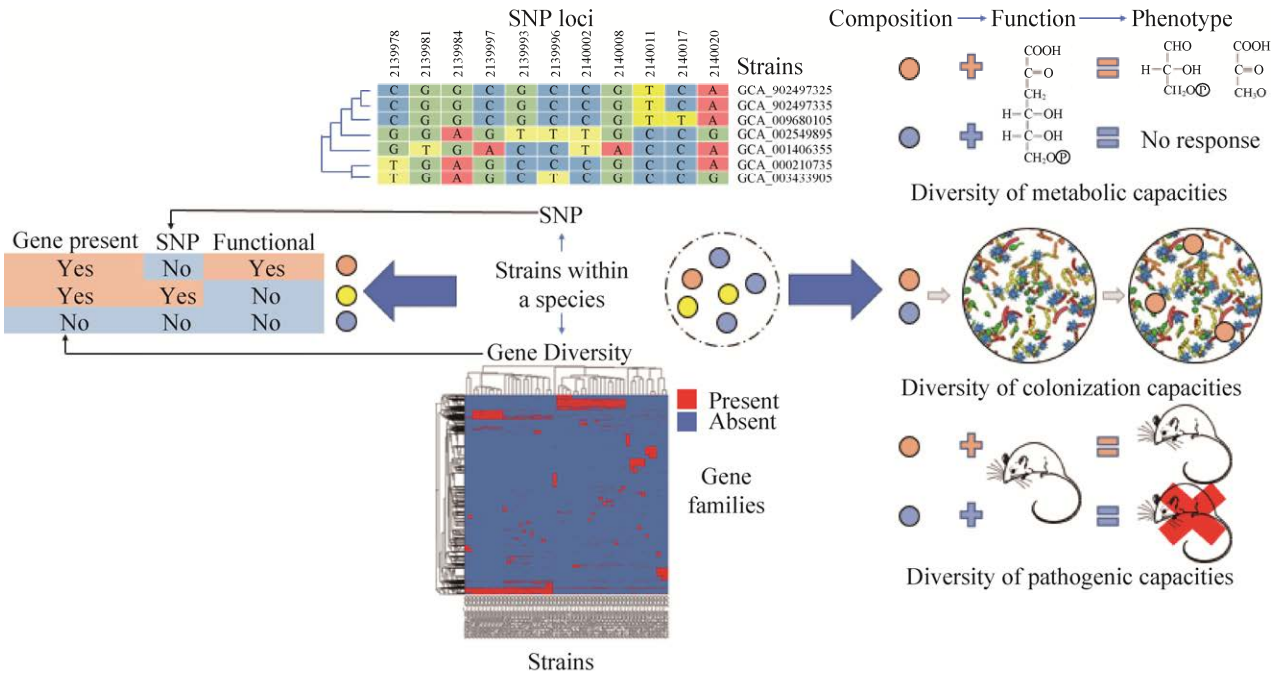


**图 1　同一物种的不同菌株之间的基因多样性和单核苷酸多态性导致功能和表型的多样性**

Fig. 1　The gene diversity and SNPs among strains within a species, and their relationship with the diversity of function and phenotype in that species.

菌株代谢能力多样性会导致不同微生物群落对相同代谢底物产生不同的代谢产物。而随着肠道微生物与药物代谢研究的深入，菌株的代谢性状对药效的影响日益明确[13]。这使得微生物组个性化有望成为继基因组个性化后的第二类个性化医疗指标。因此，从样品中高通量地判别菌株，在未来的个性化治疗领域中将有巨大的价值和需求[14]。

菌株定植能力的多样性会导致群落移植到新的生境时，可能出现只有部分菌株能定植的情况。此外，研究发现，菌群在受到扰动的情况下 (如抗生素处理) 可能出现不同菌株的演替[15]。越来越多的研究关注微生物制剂在临床使用中的定植情况[16]以及使用后原有微生物群落的变化[17]。而对于复杂群落的应用 (如粪菌移植)，更是需要对复杂菌株构成进行监控，从而确保其定植、功效、风险等[18]。

菌株致病能力的多样性会直接影响宿主的疾病表型。把某些致病菌株移植到小鼠模型上，可以观测到疾病的表型；而把同菌种的非致病菌株进行移植，则小鼠不会发病[19]。而在传染病中找到真正的致病菌株更是控制疫情的重要基础[20]。此外，研究表明，即使疾病组和对照组的菌种比例没有差异，但菌株仍可能存在差异[21]。并且菌株的选择性在疾病干预过程中有着重要作用[22]，因此菌株是一类重要的疾病判别特征[23]。而随着微生物与疾病的关系研究逐步从研究单菌株传染病向研究多菌株复杂群落的失调与非传染性疾病发生关系转移[24]，微生物构成和功能的多样性对宿主表型的影响显得更为常见和复杂，使得获取整个复杂群落的构成的需求变得日益重要。

综上所述，高通量、高分辨率 (菌株水平) 地获取微生物群落的构成和功能信息，在科学研究、工业生产和临床应用中都有着重要的价值和前景。

**(2) 基于宏基因组数据的菌株分析**

宏基因组测序是通过对样品中全部微生物的DNA进行提取后建立文库，并对文库内的随机片段进行序列测序和比对分析的一种高通量研究方法。它能很大程度上弥补培养组学的不足，也为更高通量地研究菌株的组成和功能信息提供了技术可能。宏基因组数据分析，主要可分为无参考基因组 (无参) 的组装分析和有参考基因组 (有参) 的比对分析两大类。

无参组装分析通过对测序片段进行组装 (Assembly) 并通过比较基因组学方法完成分析、解读。高质量的组装有助于获取完整微生物单倍型 (Haplotype)，不但可以获取 SNP 信息用于菌株的追踪[25]，也可用于基因水平转移、进化选择的推演[25-26]，更能挖掘出未被发现的新菌株。然而由于基因组高度相似，直接从宏基因组测序数据完成同物种的不同菌株的组装十分困难[27]。尽管研究人员近年来开发出了一些运用多样本的复杂分箱方法[28]或辅以长读长测序数据的组装算法[29]，但这些方法往往需要大量计算资源，而且对目标物种的丰度及测序深度有一定要求。还会引入嵌合体和不完整基因组，并不适用于大量样品或者低丰度样品的高通量研究[8-30]。

有参比对分析受限于参考菌株基因组的数量[8]。随着微生物培养组学和全基因组测序的不断发展，利用参考基因组进行有参比对分析已经成为现实[8,30-31]。基于宏基因组数据使用有参比对的方法，对特定的微生物菌种进行菌株构成分析和功能分析是完全可行的。本文将重点介绍有参比对的分析方法。

## 1 有参比对的菌株分析方法

对应菌株的两类多样性 (图 1)，宏基因组数据的菌株分析方法按使用及反馈的差异信息可以分为两大类 (表 1)：基于 SNP 进行识别和基于基因组成 (Genetic constitution) 差异进行识别。SNP 反映的是菌株在不同位点上的单碱基多样性；基因组成差异反映的是菌株的基因多样性。

在基于 SNP 识别的方法中，根据使用的参考基因组范围不同，又可分为两大类：使用全部基因组信息和仅使用标记基因 (Marker gene) 上的信息。使用全部的基因序列信息时，信息最全面，理论上菌株的辨识程度比用标记基因的方法高；但是因为比对的区域大，对计算资源要求很高，导致一次运行只能针对有限的菌株对象，常见情况是一次只能对一个菌种下的菌株进行区分，因此目前多用于“精细”追踪目标菌株在不同样品中的出现情况。使用标记基因时，因为只使用区分能力最好的部分基因作为标志基因，所以不容易受重组 (Recombination) 或前噬菌体 (Prophage) 等非保守区域的干扰，并且计算时可考虑更多菌株，甚至可以一次分析大部分需要考虑的多个菌

**表 1 菌株分析算法的特征汇总**

**Table 1 Summary of strain analysis algorithms**

| Algorithms | StrainEst | Strain Finder | ConStrains | StrainPhlAn | StrainSifter | PanPhlAn |
|---|---|---|---|---|---|---|
| Variation type | SNP | SNP | SNP | SNP | SNP | Genetic constitution |
| Alignment region and database | Species-specific whole genomes | Universal marker (AMPHORA) | Species-specific marker (MetaPhlAn) | Species-specific marker (MetaPhlAn) | Single whole genome | Species-specific pan-genome |
| Composition of multiple strains | Yes | Yes | Yes | No | No | No |
| Output format | Relative abundance | Relative abundance | Relative abundance | Abundance of the dominant strain | Distances among samples | Cluster of gene profiles |
| Evaluation | Yes | Yes | Yes | No | No | Yes |

种的菌株；但对于在标志基因区域高度相似的菌株，这类方法无法分辨。此外，因为不同方法使用的标志基因有所不同，所以其数据库与样品中菌株的匹配程度会对分析结果产生一定影响。

此外，虽然表 1 中的方法都可以一定程度上发现样品中存在多个菌株的情况，但并不是所有方法都能够反馈多个菌株的比例，其中 PanPhlAn 和 StrainPhlAn 都假设每个菌种中只有一个优势菌株并只反馈该菌株的信息，而 StrainSifter 只会反馈样品间的进化距离，并无具体菌株比例。且 PanPhlAn 反馈的不是组成信息，而是各样品中目标菌种的泛基因组分布情况。

### 1.1 基于 SNP 差异的菌株分析

基于 SNP 的分析方法是菌株分析算法的主流。利用的是宏基因组数据中单碱基水平的信息，通过使用宏基因组片段序列和参考序列 (全基因组或标记基因) 的比对结果进行 SNP 位点的等位基因相对比例分析，最后根据这些比例信息，推算菌株的组成。因此，这类方法只能反馈菌株构成丰度，无法直接提供功能基因相关信息。

StrainSifter[20]和 StrainEst[32]属于以全基因组为对象的方法。其中，StrainSifter 完全是用菌株的全基因组信息，一次只能追踪一个菌株及其旁系变化，并以此信息构建样品间的进化关系图，但不反馈具体的菌株丰度。它主要用于快速并准确地检测样品中是否存在特定菌株及其旁系。而 StrainEst 则是以菌种数据库为单位，对该菌种数据库内的所有已知菌株的全基因组序列进行分析，找出差异代表区域 (相当于标记基因)及 SNP 位点，然后再把宏基因组序列片段和这些区域进行比对，一定程度上结合了标记基因方法的思想，其与标记基因方法的主要差别是并非固定使用一套标记基因。此外，StrainEst 会反馈具体的菌株丰度。因此，StrainEst 适用于对特定菌种的已知菌株进行相对丰度的分析，研究菌株相对丰度和表型之间的关系。

ConStrains[33]、Strain Finder[31]、StrainPhlAn[8]都属于以标记基因为对象的方法。ConStrains 和 StrainPhlAn 使用的都是目标菌种中的特异性标记基因库 (MetaPhlAn[34])，主要差别在推算具体菌株丰度的算法上：ConStrains 使用蒙特卡罗马尔可夫模型来推算菌株的丰度并结合修正后的赤池信息量准则 (Corrected Akaike information criterion) 来选择最优模型与结果；StrainPhlAn 则侧重利用 SNP 信息计算样品间的进化关系，并不反馈各样品内的具体菌株丰度。此外 StrainPhlAn 有一个固定假设：一个样品中一个菌种只有一个代表菌株，因此该代表菌株的丰度等于该菌种丰度。Strain Finder 使用通用标记基因为参考[35]，其默认数据库只针对人肠道微生物，但用户可以重新自定义构建。它使用的是最大期望 (Expectation maximization，EM) 算法，通过拟合不同样品中的等位基因频率变化来获取不同菌株的代表 SNP 序列及各样品中对应的丰度组成。Strain Finder 被设计用于同时追踪粪菌移植前后的复杂群落样品中所有人肠道微生物各菌种内的菌株的丰度变化情况，也可用于对相近样品进行人肠道菌种的菌株普查。因为其方法允许识别数据库内不存在的菌株，所以可以一定程度上识别新的菌株。但由于该方法提取的 SNP 位置依赖于输入的样本，在不同的样本上获取的菌株信息并不能直接进行比较，因此它是整体样品分析的方法，需要相互比较的样品必须统一进行分析。此外，因为 EM 算法的特性，其结果都是局部最优，因此需要多次拟合来尽可能得到全局最优解[31]。

### 1.2 基于基因组成差异的分析方法

基于基因组成差异的菌株分析算法，利用的是宏基因组数据中基因水平的信息 (如各基因的丰度与各参考菌株中对应基因的存在或缺失情况的拟合)。PanPhlAn[30]是目前笔者所知唯一的该类方法。PanPhlAn 对目标菌种的各菌株全基因组序列中注释的基因进行聚类以合并为基因家族

(Gene family)，选取每个基因家族的代表序列构建该菌种的泛基因组图谱作为参考数据库。通过宏基因组样本数据中序列与参考数据库的比对，根据计算出的每个基因家族的覆盖度判断其在该样本中的存在或缺失情况。该方法可直接提供每个基因家族在样品各菌株中的存在与缺失情况，从而反映该菌株的功能特征信息，因此可用于菌株功能与表型关系的研究 (如菌株的毒力、抗药性等)。但该方法只提供优势菌株的基因家族组成信息，并无法提供菌株的构成丰度。而且由于复杂群落样品中物种间相似基因的干扰，样品中目标菌株的某些本应缺失的基因家族可能被误判为存在。

### 1.3 其他分析方法

除了上述两大类菌株分析方法外，还有一些基于特定假设的菌株分析算法。例如：metaMLST[36]是基于细菌鉴定中的多位点分型 (Multi-locus sequence typing，MLST) 原理，通过从宏基因组数据中获取等位子组成信息，然后与 PubMLST 数据库[37]进行比对，从而实现对菌株的鉴定。Sigma[38]是一款对病原菌菌株进行鉴定和定量的工具。在假设该菌株已确定存在的前提下，通过将宏基因组数据匹配到已知的参考基因组上，实现对最优基因组的判定、定量和变异检测。此外，MetaPhlAn2[34]和 mOTU2[39]都是较为常用的有参菌种分析方法，并提供菌株分析相关信息。其中，StrainPhlAn 是在 MetaPhlAn 的基础上进行菌种分析；mOTU2 则支持在菌种分析的基础上进行 SNP 分析，从而推断菌株组成。针对三代测序长读长数据，Dilthey 等[40]开发了 MetaMaps 进行菌种菌株的鉴定，并可以进行基因的存在/缺失判别。但是该类数据和宏基因组测序数据差异较大，不在此进行讨论。

## 2 菌株分析方法的评估体系

随着宏基因组测序应用的快速增长，宏基因组数据分析方法的评估备受重视，CAMI (Critical Assessment of Metagenome Interpretation) 组织先后发布了其模拟数据体系[41]和评估平台[42]，并对 7 个分析方法进行了评估。随后，Ye 等对 20 个宏基因组有参比对分析方法进行了详细评估[43]。但这些评估都只精细到菌种层次而非菌株层次，因此目前仍缺乏一个针对菌株分析方法的评估体系。而现有的评估体系在用于评估菌株时，会存在下面 3 个潜在问题。

首先，已有评估的菌株分析算法都只使用了模拟标准数据作为评估数据，且只考虑了以下两个因素：多菌株混合结合浓度梯度混合[31-33]。其中，ConStrains 展示了其结果与模拟标准相比，在 JS 离散量(Jensen-shannon divergence, JSD)上优于随机结果。Strian Finder 和 StrainEst 则分别显示其结果优于 Constrain，且 StrainEst 还优于 PanPhlAn 和 Sigma。但这些评估在选择菌株时，均未考虑菌株之间的序列相似度，无法深入了解和评判序列相似度对各菌株分析算法的影响，因此并不能真切反映各菌株分析算法的特点。

其次，通过实验制备的真实标准样品(如 HMP mock[44]、ILPP mock[45]和 MBARC[46])，都只是不同菌种的菌株混合，并无同菌种多菌株混合，只能反映菌种间的差异，无法用于评估菌株分析算法在同种多菌株存在场景下的精准度。

最后，目前的 6 个主流方法，数据反馈的侧重点差异较大，只有 3 个方法能反馈样品内的各菌株丰度信息，其余 3 个方法均只反馈各样品间的菌株基因组的差异信息，因此能通用的评估指标极其有限。而在反馈菌株丰度的方法中，只有 StrainEst 反映的菌株可直接与已知全基因组的参考菌株对应，而 Strain Finder 和 ConStrains 均只提供样品中代表性的 SNP 序列，导致分析结果和真实标准难以统一比较，大大提高了评估的难度。

## 3 菌株分析在微生物组研究中的应用

由于菌株在微观上具有宿主特异性，在宏观上反映样本来源的地域、饮食等环境因素的影响，

故菌株信息可用于追踪个体间传播途径，揭示环境因素的关联关系[25]。此外，由于不同菌株的基因及功能存在差异，菌株与宿主之间存在着不同的作用机制。菌株的这一特点可用于菌株作用机制研究以及关键菌株的发现筛选[47]。宏基因组菌株分析方法的出现，使得对菌株的研究脱离了实验培养的限制，为菌株研究和应用提供了新的分析思路和方法。

### 3.1 基础研究

近两年来众多大型微生物组研究项目[48]相继启动。这些项目产生的大量宏基因组数据，有助于科学家研究地域、饮食、年龄、疾病等因素与人体微生物群落的关系。通过对公共宏基因组数据集进行基于 SNP 的菌株分析，研究人员能够构建出不同菌株的进化树，并将进化树的结构与环境因素对应起来。通过对这些菌株进行泛基因组分析，研究人员能够进一步了解外界因素或者表型对应的菌株的生理功能特性。Truong 等使用 StrainPhlAn 对多个国家的人体肠道宏基因组样本进行分析，发现微生物群体结构与群体的地理结构相关联[8]。De Filippis 等使用 PanPhlAn 对意大利地区人的粪便样本进行宏基因组分析，发现素食或者杂食个体的普氏菌 *Prevotella copri* 中的菌株在基因层面存在明显差别[49]。Tett 等发现 *P. copri* 存在 4 个主要的进化支，并且它们在西方人群中出现较少[50]。Vatanen 等对 DIABIMMUNE 研究中来自芬兰、爱沙尼亚和俄罗斯的早期儿童数据集进行 SNP 以及宏基因组组装的菌株分析，发现国家之间在拟杆菌属 *Bacteroides* 与双歧杆菌属 *Bifidobacterium* 中的菌种菌株上表现出明显的地域性特点，菌株的差别会对早期婴儿肠道微生物结构带来影响[51]。Zolfo 等使用 MetaMLST、StrainPhlAn 以及 PanPhlAn 三种工具对 1 614 例城市环境样本的宏基因组数据进行了菌株分析，并从进化和功能的角度对结果进行了讨论[52]。

### 3.2 环境公共卫生

病原菌的监控与溯源是环境公共卫生领域面临的一个重要问题。许多研究开始使用宏基因组数据进行高效快捷的病原菌菌株筛查，以更好地支持政府和医疗机构的决策。Fresia 等使用 MetaMLST 对城市的海滨及污水系统的样本进行宏基因组菌株分析，发现了与临床相关的一些病原菌的存在[53]。Hamner 等使用 Nanopore MinION 测序仪以及 CosmosID 软件对河流样本进行宏基因组分析，检测出肠出血性和肠致病性大肠杆菌 *Escherichia coli* 菌株，以及其他病原性 *E. coli* 菌株[54]。一些研究尝试将菌株筛查用于食品安全领域。Walsh 等使用 MetaMLST、StrainPhlAn 以及 PanPhlAn 对刚果地区的发酵乳制品进行检测，在一些样本中发现了可能的病原性 *E. coli* 以及肺炎克雷伯氏菌 *Klebsiella pneumoniae* 菌株[55]。

此外，由于菌株在宿主个体内有高度特异性，一些研究者尝试通过宏基因组数据研究菌株的传播途径，尤其是母婴间的传递[56-59]。Tamburini 等使用菌株溯源工具 StrainSifter，并用其比较血液中分离的病原菌菌株序列与肠道宏基因组数据，从而判断这些菌株是否来源于病人肠道[20]。Olm 等使用菌株分析工具 InStrain 研究剖腹产与顺产的新生儿菌株，发现剖腹产的婴儿菌株可能来源于医院[60]。这些溯源工作对指导医院针对性地开展病原预防工作有着重要意义。

### 3.3 临床干预与治疗

随着越来越多的细菌在健康和适应症改善中的促进作用被发现，对益生菌菌株以及粪菌移植 (FMT) 的效果研究也越来越受到重视。由于益生菌菌株的基因组序列往往已经明确，菌株分析方法可用于追踪益生菌菌株在肠道的留存状况。Zmora 等通过检查宏基因组样本中是否包含益生菌菌株特异性基因来判断相关益生菌菌株在粪便样本中的有无，并由此分析益生菌菌株在不同个体肠道中的定植情况，然后从机制上阐明特定

益生菌菌株是如何通过与肠道微生物互作来发挥效果的[17]。另外，尽管许多益生菌研究已经开始使用宏基因组研究益生菌对于肠道微生物群落的影响[61-62]或者进行益生菌的成分鉴定[63]，但是宏基因组菌株构成分析方法还没有被普遍使用。

粪菌治疗 (FMT) 是近年来兴起的肠道微生物治疗方法，该方法将健康人肠道微生物移植入患者体内，从而达到系统性改变患者肠道微生态的目的。在 FMT 治疗过程中，医疗人员可以通过追踪供体粪便样本中的菌株在患者体内的留存和变化情况[20]，进行粪菌移植治疗效果的评估以及治疗方案的改进和优化[31]。此外，如果已知有益菌株的关键 SNP 或者基因，可以对供体样本进行评分和排序[64]。

### 3.4 生物标记物和药物发现

随着微生物在疾病诊断和治疗方面的研究进展，不断有微生物被发现与人体疾病存有密切关联。一些微生物的组合已经用于体外诊断。但是，即便对于同一适应症 (如肿瘤靶向药物治疗[65]和自闭症[66])，不同研究得出的疾病相关细菌可能并不一致，有的甚至相互矛盾。对于这种现象，一个可能的情况是当前主流的宏基因组分析工具主要在菌种水平进行分析，并没有考虑到同一菌种下菌株水平的差异。研究报道过许多菌种的菌株的确存在较大差异，例如 *P. copri* 作为人体常见的一类肠道微生物，与人体葡萄糖以及胰岛素耐受既表现出正相关[67]，也表现出负相关[68]关系。因此，对宏基因组数据从菌株水平进行进一步的挖掘，有助于发现与表型高度相关的特定菌株或者菌株组合。比如，Fang 等基于 MIDAS 对一名克罗恩病患者的样本进行了菌株水平的时间序列分析，发现炎症状态与特定的 *E. coli* 菌株相关联[69]。

菌株分析方法可以从新的角度 (特定 SNP、基因) 对病人表型进行分层。从机器学习和模型训练的角度看，这种处理方法相当于增加了宏基因组数据的特征数目，而且新增加的特征更加反映真实世界的情况。对于疾病/健康样本、治疗响应/不响应样本的宏基因组菌株分析，有助于进行以菌株为靶点的治疗方案的开发或者将菌株本身作为潜在药物进行开发[70-71]。

研究表明，不同肠道微生物菌株会参与药物的代谢，从而对药物的药效及毒性产生影响[72-74]。宏基因组菌株分析技术的进步，可以帮助制药公司更好地对适用人群进行评估，并采取一定的微生物调控的措施，以提高治疗的效果和减少副作用。

## 4 总结与展望

随着测序技术的发展，菌株的分析方法和结果的质量也会迎来革新式的发展。首先，长读长第三代测序技术 (PacBio、Oxford nanopore) 及高通量染色质构象捕获技术 (Hi-C) 在宏基因组数据的应用将大幅提高组装的完整性和准确性[29,75-76]，这些从菌群样本直接获得的高质量的基因组不仅可以用于无参的菌株分析，也可丰富补充现有基因组数据库，为有参的菌株分析提供更完整的参考数据。其次，这些新技术使还原质粒等可移动基因元件的菌株归属和研究其动态变化成为可能，为研究菌株的生态功能和进化发展提供了新的维度。届时，选择性地采用高质量的三代或 Hi-C 测序组装，并结合更廉价的二代测序提升研究规模，或将成为主流的微生物组研究策略，随之也将涌现出一批整合多种测序技术数据的生物信息算法及工具。另外，培养组学的快速发展，正极大地丰富着菌株基因组信息数据库。又因为通过对纯培养物进行的全基因组测序所获得的菌株基因组信息比通过宏基因组数据进行拼装所得的信息更为准确和可靠[77]，所以，培养组学将成为大量高质量菌株基因组的重要来源并将助力宏基因组数据分析迈入菌株水平时代。而随着计算能力的进一步发展，有参菌株分析方法可能会往基于

全基因组数据信息方向发展，并且能同时进行大量菌种的菌株分析，进一步提高算法的通用性和结果的信息含量。

随着应用案例的增长，基于宏基因组测序的有参菌株分析算法在科研、临床等方面都证明了其重要价值。但菌株分析算法的开发，目前仍处于起步阶段，各方法都有主要的针对性使用场景和对应的基本假设。根据应用场景，大体可分为两类： 第一类分析方法针对目的菌株进行分析。研究者需要提供菌株的有关信息，如全基因组序列，然后从样本中提取该菌株的定性和定量信息。如果已知相关菌株的背景信息，第一类方法可以很好地为临床或者政策安排提供决策依据。最常见的应用是从环境或者宿主样本中对特定菌株(如病原菌菌株) 的溯源和筛查 (如 StrainSifter)。这类方法对准确性要求高，但对结果的全面性要求较低。此外，由于肠道微生物菌株积极参与许多药物的代谢，这些预设的菌株信息可用于对相关适应症患者的分层，以优化临床治疗方案，实现精准治疗。如果菌株在适应症中的作用明确，第一类分析方法也可以在粪菌移植治疗中作为供体的指标对供体进行选择。第二类分析方法则能够从样本宏基因组数据中反映菌株的微观多样性(Microdiversity)，强调样本中有什么样的菌株构成，各自的比例是多少。这类方法可以监测菌株构成在多样本间的动态变化，应用于分析环境或者时间对菌株构成带来的影响 (如母体到新生儿环境改变，抗生素干预或者 FMT 治疗对菌株的影响等)。通过与样本相关的表型信息相关联，第二类方法可用于对表型相关的菌株的筛选，在益生菌组合工艺优化、关键菌 (如生物标记物、病原菌株、菌株药物) 发现上具有潜在的应用价值。总而言之，在选用方法时，一定要注意算法的基本假设是否符合自身课题的特点，然后根据课题的需求进行针对性选择。

然而，目前仍缺少对宏基因组数据的菌株分析方法的详细评估，导致对算法的优缺点、精准度和可靠性了解甚少，对选择合适的方法造成了阻碍。建立专门针对菌株水平的评估体系和平台，对现有的菌株分析方法进行系统评估，将有助于研究者针对数据选择合适的方法。此外，由于现有方法对 SNP 和基因组成差异的割裂使用，单一算法无法同时获取菌株水平的丰度构成信息和功能基因信息。开发新的算法实现菌株构成和基因谱的关联分析，将进一步提高研究的可靠性和应用价值。

## REFERENCES

[1] Kolde R, Franzosa EA, Rahnavard G, et al. Host genetic variation and its microbiome interactions within the human microbiome project. Genome Med, 2018, 10: 6.

[2] Sender R, Fuchs S, Milo R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell, 2016, 164(3): 337–340.

[3] Thomas S, Izard J, Walsh E, et al. The host microbiome regulates and maintains human health: a primer and perspective for non-microbiologists. Cancer Res, 2017, 77(8): 1783–1812.

[4] Zheng DP, Ratiner K, Elinav E. Circadian influences of diet on the microbiome and immunity. Trends Immunol, 2020, 41(6): 512–530.

[5] Pennisi E. Meet the psychobiome. Science, 2020, 368(6491): 570–573.

[6] Shen P, Chen XD. Microbiology. 8th ed. Beijing: Higher Education Press, 2016 (in Chinese).
沈萍, 陈向东. 微生物学. 8 版. 北京: 高等教育出版社, 2016.

[7] Dolgin E. Fighting cancer with microbes. Nature, 2020, 577(7792): S16–S18.

[8] Truong DT, Tett A, Pasolli E, et al. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res, 2017, 27(4): 626–638.

[9] Zagato E, Pozzi C, Bertocchi A, et al. Endogenous murine microbiota member Faecalibaculum rodentium and its human homologue protect from intestinal tumour growth. Nat Microbiol, 2020, 5(3): 511–524.

[10] Fehlner-Peach H, Magnabosco C, Raghavan V, et al. Distinct polysaccharide utilization profiles of human intestinal *Prevotella copri* isolates. Cell Host Microbe, 2019, 26(5): 680–690.e5.
[11] Medini D, Serruto D, Parkhill J, et al. Microbiology in the post-genomic era. Nat Rev Microbiol, 2008, 6(6): 419–430.
[12] Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA, 2005, 102(39): 13950–13955.
[13] Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, et al. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature, 2019, 570(7762): 462–467.
[14] Ferreiro A, Crook N, Gasparrini AJ, et al. Multiscale evolutionary dynamics of host-associated microbiomes. Cell, 2018, 172(6): 1216–1227.
[15] Koo H, Hakim JA, Crossman DK, et al. Individualized recovery of gut microbial strains post antibiotics. npj Biofilms Microbiomes, 2019, 5: 30.
[16] Suez J, Zmora N, Segal E, et al. The pros, cons, and many unknowns of probiotics. Nat Med, 2019, 25(5): 716–729.
[17] Zmora N, Zilberman-Schapira G, Suez J, et al. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. Cell, 2018, 174(6): 1388–1405.e21.
[18] Suez J, Zmora N, Zilberman-Schapira G, et al. Post-antibiotic gut mucosal microbiome reconstitution is impaired by probiotics and improved by autologous FMT. Cell, 2018, 174(6): 1406–1423.e16.
[19] Etienne-Mesmin L, Chassaing B, Adekunle O, et al. Toxin-positive *Clostridium difficile* latently infect mouse colonies and protect against highly pathogenic *C. difficile*. Gut, 2018, 67(5): 860–871.
[20] Tamburini FB, Andermann TM, Tkachenko E, et al. Precision identification of diverse bloodstream pathogens in the gut microbiome. Nat Med, 2018, 24(12): 1809–1814.
[21] Wang Y, Wang S, Wu CY, et al. Oral microbiome alterations associated with early childhood caries highlight the importance of carbohydrate metabolic activities. mSystems, 2019, 4(6): e00450–19.
[22] Harkins CP, Kong HH, Segre JA. Manipulating the human microbiome to manage disease. JAMA, 2019, 323(4): 303–304.
[23] Olm MR, Bhattacharya N, Crits-Christoph A, et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria. Sci Adv, 2019, 5(12): eaax5727.
[24] Finlay BB, CIFAR Humans, The Microbiome. Are noncommunicable diseases communicable? Science, 2020, 367(6475): 250–251.
[25] Chng KR, Li CH, Bertrand D, et al. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. Nat Med, 2020, 26(6): 941–951.
[26] Arnold B, Sohail M, Wadsworth C, et al. Fine-scale haplotype structure reveals strong signatures of positive selection in a recombining bacterial pathogen. Mol Biol Evol, 2020, 37(2): 417–428.
[27] Sczyrba A, Hofmann P, Belmann P, et al. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nat Methods, 2017, 14(11): 1063–1071.
[28] Cleary B, Brito IL, Huang K, et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol, 2015, 33(10): 1053–1060.
[29] Bertrand D, Shaw J, Kalathiyappan M, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol, 2019, 37(8): 937–944.
[30] Scholz M, Ward DV, Pasolli E, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods, 2016, 13(5): 435–438.
[31] Smillie CS, Sauk J, Gevers D, et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. Cell Host Microbe, 2018, 23(2): 229–240.e5.
[32] Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun, 2017, 8: 2260.

[33] Luo CW, Knight R, Siljander H, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol, 2015, 33(10): 1045–1052.
[34] Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods, 2015, 12(10): 902–903.
[35] Wu MT, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biology, 2008, 9(10): R151.
[36] Zolfo M, Tett A, Jousson O, et al. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Res, 2017, 45(2): e7.
[37] Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST. org website and their applications. Wellcome Open Res, 2018, 3: 124.
[38] Ahn TH, Chai JJ, Pan CL. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics, 2015, 31(2): 170–177.
[39] Milanese A, Mende DR, Paoli L, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun, 2019, 10: 1014.
[40] Dilthey AT, Jain C, Koren S, et al. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun, 2019, 10: 3066.
[41] Fritz A, Hofmann P, Majda S, et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome, 2019, 7: 17.
[42] Meyer F, Bremges A, Belmann P, et al. Assessing taxonomic metagenome profilers with OPAL. Genome Biol, 2019, 20: 51.
[43] Ye SH, Siddle KJ, Park DJ, et al. Benchmarking metagenomics tools for taxonomic classification. Cell, 2019, 178(4): 779–794.
[44] The Human Microbiome Project Consortium. A framework for human microbiome research. Nature, 2012, 486(7402): 215–221.
[45] Bowers RM, Clum A, Tice H, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics, 2015, 16: 856.
[46] Singer E, Andreopoulos B, Bowers RM, et al. Next generation sequencing data of a defined microbial mock community. Sci Data, 2016, 3: 160081.
[47] Yuan J, Chen C, Cui JH, et al. Fatty liver disease caused by high-alcohol-producing *Klebsiella pneumoniae*. Cell Metabo, 2019, 30(4): 675–688.e7.
[48] Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded human microbiome project. Nature, 2017, 550(7674): 61–66.
[49] De Filippis F, Pasolli E, Tett A, et al. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. Cell Host Microbe, 2019, 25(3): 444–453.e3.
[50] Tett A, Huang KD, Asnicar F, et al. The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe, 2019, 26(5): 666–679.e7.
[51] Vatanen T, Plichta DR, Somani J, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat Microbiol, 2019, 4(3): 470–479.
[52] Zolfo M, Asnicar F, Manghi P, et al. Profiling microbial strains in urban environments using metagenomic sequencing data. Biol Direct, 2018, 13: 9.
[53] Fresia P, Antelo V, Salazar C, et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. Microbiome, 2019, 7: 35.
[54] Hamner S, Brown BL, Hasan NA, et al. Metagenomic profiling of microbial pathogens in the little bighorn river, montana. Int J Environ Res Public Health, 2019, 16(7): 1097.
[55] Walsh AM, Crispie F, Daari K, et al. Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks. Appl Environ Microbiol, 2017, 83(16): e01144–17.
[56] Ferretti P, Pasolli E, Tett A, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell Host Microbe, 2018, 24(1): 133–145.e5.
[57] Korpela K, Costea P, Coelho LP, et al. Selective maternal seeding and environment shape the human gut microbiome. Genome Res, 2018, 28(4): 561–568.
[58] Yassour M, Jason E, Hogstrom LJ, et al. Strain-level

analysis of mother-to-child bacterial transmission during the first few months of life. Cell Host Microbe, 2018, 24(1): 146–154.e4.

[59] Wampach L, Heintz-Buschart A, Fritz JV, et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nat Commun, 2018, 9: 5091.

[60] Olm MR, Crits-Christoph A, Bouma-Gregson K, et al. InStrain enables population genomic analysis from metagenomic data and rigorous detection of identical microbial strains. bioRxiv, 2020, doi: 10.1101/2020.01.22.915579.

[61] MacPherson CW, Mathieu O, Tremblay J, et al. Gut bacterial microbiota and its resistome rapidly recover to basal state levels after short-term amoxicillin-clavulanic acid treatment in healthy adults. Sci Rep, 2018, 8: 11192.

[62] Hor YY, Lew LC, Jaafar MH, et al. *Lactobacillus* sp. improved microbiota and metabolite profiles of aging rats. Pharmacol Res, 2019, 146: 104312.

[63] Lugli GA, Mangifesta M, Mancabelli L, et al. Compositional assessment of bacterial communities in probiotic supplements by means of metagenomic techniques. Int J Food Microbiol, 2019, 294: 1–9.

[64] Duvallet C, Zellmer C, Panchal P, et al. Framework for rational donor selection in fecal microbiota transplant clinical trials. PLoS ONE, 2019, 14(10): e0222881.

[65] Gharaibeh RZ, Jobin C. Microbiota and cancer immunotherapy: in search of microbial signals. Gut, 2019, 68(3): 385–388.

[66] Xu MY, Xu XF, Li JJ, et al. Association between gut microbiota and autism spectrum disorder: a systematic review and meta-analysis. Front Psychiatry, 2019, 10: 473.

[67] Pedersen HK, Gudmundsdottir V, Nielsen HB, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature, 2016, 535(7612): 376–381.

[68] De Vadder F, Kovatcheva-Datchary P, Zitoun C, et al. Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. Cell Metabo, 2016, 24(1): 151–157.

[69] Fang X, Monk JM, Nurk S, et al. Metagenomics-based, strain-level analysis of *Escherichia coli* from a time-series of microbiome samples from a crohn's disease patient. Front Microbiol, 2018, 9: 2559.

[70] Tanoue T, Morita S, Plichta DR, et al. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. Nature, 2019, 565(7741): 600–605.

[71] Zhao LP. The gut microbiota and obesity: from correlation to causality. Nat Rev Microbiol, 2013, 11(9): 639–647.

[72] Lam KN, Alexander M, Turnbaugh PJ. Precision medicine goes microscopic: engineering the microbiome to improve drug outcomes. Cell Host Microbe, 2019, 26(1): 22–34.

[73] Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, et al. Separating host and microbiome contributions to drug pharmacokinetics and toxicity. Science, 2019, 363(6427): eaat9931.

[74] Javdan B, Lopez JG, Chankhamjon P, et al. Personalized mapping of drug metabolism by the human gut microbiome. Cell, 2020, 181(7): 1661–1679.e22.

[75] Kolmogorov M, Rayko MP, Yuan J, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. bioRxiv, 2019: 637637, doi: 10.1101/637637.

[76] DeMaere MZ, Darling AE. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol, 2019, 20: 46.

[77] Lagier JC, Dubourg G, Million M, et al. Culturing the human microbiota and culturomics. Nat Rev Microbiol, 2018, 16(9): 540–550.

(本文责编 郝丽芳)