·综述·

纳米孔测序数据比对分析方法与参考数据库研究进展

李文正[1,2]，张宁[1,2]，李卓越[2,3]，崔莉煊[2,3]，王欣博[2,3]，杜耀华[1,2*]

1 军事科学院 系统工程研究院，天津 300161
2 国家生物防护装备工程技术研究中心，天津 300161
3 天津科技大学 电子信息与自动化学院，天津 300457

李文正, 张宁, 李卓越, 崔莉煊, 王欣博, 杜耀华. 纳米孔测序数据比对分析方法与参考数据库研究进展[J]. 生物工程学报, 2026, 42(1): 77-92.
LI Wenzheng, ZHANG Ning, LI Zhuoyue, CUI Lixuan, WANG Xinbo, DU Yaohua. Research progress on nanopore sequencing data alignment analysis methods and reference databases[J]. Chinese Journal of Biotechnology, 2026, 42(1): 77-92.

摘 要：纳米孔测序作为测序技术的新兴热点，凭借其读长较长、检测快速和设备紧凑小巧等独特优势，在物种鉴定、基因组组装、变异检测、转录组分析等领域展现出巨大潜力。然而，纳米孔测序数据错误率较高，存在序列插入和缺失等问题，对传统序列比对工具运用和参考数据库构建提出了新的挑战。本文围绕纳米孔数据特征，系统梳理了适配纳米孔测序的序列比对工具，针对长读长测序、实时测序、错误率兼容、宏基因组和结构变异检测这 5 种不同应用场景，阐述了其在处理序列数据时的优势及局限性；同时，还从数据源的角度对序列参考基因组数据库进行多维度分类整理，并总结了纳米孔高质量数据库构建的关键技术。本文通过对比对工具与数据库进行协同分析，为纳米孔测序数据分析的优化与创新提供参考，推动宏基因组测序从数据生成向功能解析的深度转化。

关键词：基因测序；纳米孔测序；数据分析；序列比对；参考数据库

Research progress on nanopore sequencing data alignment analysis methods and reference databases

LI Wenzheng[1,2], ZHANG Ning[1,2], LI Zhuoyue[2,3], CUI Lixuan[2,3], WANG Xinbo[2,3], DU Yaohua[1,2*]

1 Institute of Systems Engineering, Academy of Military Science, Tianjin 300161, China
2 National Bio-protection Engineering Center, Tianjin 300161, China
3 College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300457, China

Abstract: Nanopore sequencing, as an emerging hotspot in sequencing technology, demonstrates tremendous potential in species identification, genome assembly, variant detection, and

*Corresponding author. E-mail: qsyaohua@sina.com
Received: 2025-07-18; Accepted: 2025-11-13; Published online: 2025-12-12

transcriptome analysis, owing to its distinctive advantages including extended read lengths, rapid detection capabilities, and compact instrumentation. However, nanopore sequencing data are characterized by high error rates and presence of insertions and deletions, which pose novel challenges for the application of conventional sequence alignment tools and the construction of reference databases. Focusing on the characteristics of nanopore data, this paper systematically sorts out sequence alignment tools suitable for nanopore sequencing, and elaborates on their advantages and limitations in processing sequence data for five different application scenarios: long-read sequencing, real-time sequencing, error rate compatibility, metagenomics, and structural variation detection. Meanwhile, from the perspective of data sources, this paper conducts multi-dimensional classification and organization of reference genome databases, and sorts out the key technologies for constructing high-quality nanopore databases. Through the collaborative analysis of alignment tools and databases, this paper provides references for the optimization and innovation of nanopore sequencing data analysis, and promotes the in-depth transformation of metagenomic sequencing from data generation to functional analysis.

Keywords: gene sequencing; nanopore sequencing; data analysis; sequence alignment; reference database

测序技术作为解析生物体遗传信息的重要工具，其本质在于精确测定 DNA 或 RNA 分子中碱基的排列顺序，为研究者提供生物体的遗传信息和生命活动的分子基础。自 1977 年 Sanger 测序方法问世以来，测序技术一直在不断更新迭代。第一代测序技术因通量较低、成本较高等限制了其大规模应用。第二代测序技术以 Illumina 平台为代表，读长被限制在 150−300 bp 范围内，导致基因组组装高度依赖短序列拼接[1]。第三代测序技术则凭借其超长读长(可达 10 万碱基级别)及无需 PCR 扩增等特性优势，正逐渐改变基因组学数据的解析方式[2-3]。纳米孔测序技术作为第三代测序技术的代表，与传统测序技术相比，无需荧光标记和光学检测系统，而是利用马达蛋白控制 DNA 分子通过纳米孔，基于相应的电流变化识别碱基序列，从而实现单分子实时测序，远超第二代测序技术通常几百个碱基的读长，在处理基因组高度重复区域、结构变异以及全长转录本中具有较为明显的优势[4]。

然而，纳米孔测序技术的优势也会带来数据处理的难题，高复杂度的数据使得传统序列分析算法面临计算效率与准确性的双重瓶颈[5-7]。首先原始数据错误率显著高于二代测序，且错误类型以插入缺失为主，传统“种子+扩展”的比对策略容易导致假阳性[8]。其次长读长数据对工具算力与内存效率要求更高，早期动态规划算法因时间和数据量等复杂度骤升问题，无法直接适配。目前绝大多数参考数据库是基于一代和二代测序数据构建的，存在长读长兼容性差、病原微生物(尤其是感染性真菌)数据缺失等问题，都会对比对准确性产生直接影响[9]。

从整体流程来看，纳米孔测序数据分析通常包括原始电信号采集、碱基识别、去除接头与条形码、质量控制、去除宿主序列、序列比对、从头组装、变异检测与结构变异解析等关键环节。上述步骤相互依赖，任一环节的偏差均可能影响下游环节，因此纳米孔数据的处理不应该视为若干独立模块，而需在整体架构下进行系统优化。在这一完整技术链条中，序列比对与参考数据库构建处于承上启下的枢纽位置：一方面，序列比对与参考数据库构建是物种鉴定、菌株分型、耐药与毒力基因识别以及复杂结构变异解析的基础；另一方面，它们是

纳米孔测序数据分析的核心环节，比对工具的算法设计决定了数据解析的效率与精度，参考数据库构建的完整性与标准化程度则直接影响数据分析的可靠性。即便在前端信号处理和质量控制过程中取得良好的效果，若比对策略及参考数据库体系未围绕纳米孔数据的特有误差与长读长进行优化，仍可能导致近缘物种区分模糊、低丰度病原体漏检与注释错误，进而引发下游分析结果偏差。

因此，本文在纳米孔测序数据分析整体框架的基础上，重点聚焦“纳米孔数据的比对工具适配性”与“高质量参考数据库构建”两大核心，系统分析纳米孔测序专用比对工具的技术特性，以期为临床病原检测和环境宏基因组分析等不同研究需求提供可操作的技术选型依据，推动纳米孔测序技术从“数据生成”向“临床与科研应用”的深度落地[10]。

1 纳米孔测序的序列比对工具

序列比对作为生物信息学研究的重要方向，是通过构建特定算法与数学模型解析生物序列间相似性特征的关键技术[11]。通过精准识别核酸序列间保守区域与变异位点，为揭示分子进化机制、推断功能结构域及追溯物种系统发育关系提供重要依据。随着纳米孔测序技术的快速发展，其特有的高通量、长读长和实时分析优势对传统序列比对算法提出了新的挑战。本节将系统分析现有比对工具对测序数据分析需求的适应性，深入探讨其技术瓶颈与发展方向，并为测序数据分析的优化与创新提供理论参考。针对长读长测序、实时测序、高错误率适应、宏基因组学及结构变异检测这 5 种典型应用场景，分析现有序列比对方法在算法模型、参考索引构建策略、错误处理机制及适用场景等方面的差异。

1.1 长读长序列比对方法

纳米孔测序作为第三代测序技术的代表，可产生平均 10−30 kb，最长可达 2 Mb 的连续 DNA/RNA 序列，比传统二代测序技术平均 150−300 bp 的读长具有显著优势。长读长的出现不仅推动了序列比对方法的革新，同时也为复杂的宏基因组研究带来了新的机遇。早期的序列比对算法主要基于动态规划原理，例如 1970 年提出的全局序列比对算法[4]如图 1 所示，构建了一个二维动态规划矩阵逐一计算最优匹配路径。虽然这种算法能够保证获得全局最优解，在短读长中展现出较高的精度，但在长读长场景下其时间和空间复杂度将急剧上升，因此并不适用于处理大规模宏基因组数据。

为了提升测序数据分析的准确性，减少实验中的误差，研究者们开发了多种适用于长读长的比对工具，例如 2018 年 Li 等开发了 Minimap2，革命性地解决了长读长序列比对的效率问题[12]。该算法引入了(k,w)-minimizers 概念，通过对参考序列进行 minimizer 索引来显著减少索引大小。首先利用 minimizer 作为种子链进行快速定位和动态规划延伸，在保持高精度

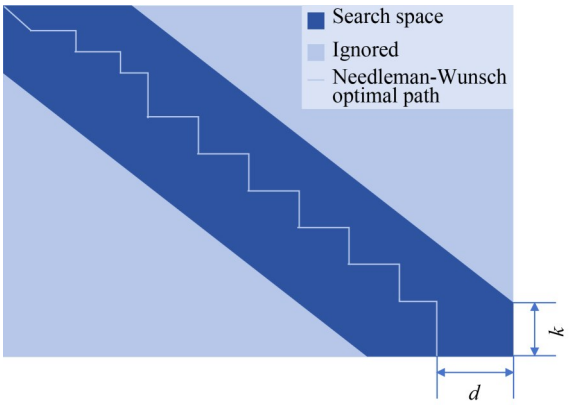

图1 Needleman-Wunsch算法最优解路径图 其中，d表示2条序列长度的差值($|m-n|$)，k表示允许的最优路径相对主对角线的最大偏移范围。

Figure 1 Needleman-Wunsch algorithm optimal solution path diagram. d denotes the difference in length between the two sequences ($|m-n|$), and k denotes the maximum allowable deviation of the optimal path from the main diagonal.

的同时显著提升比对效率。在处理超长读长(>100 kb)的纳米孔数据时，其速度比其他算法快 70 倍以上[13]。2016 年 Sović 等提出来的 GraphMap 算法采用了混合索引策略，结合 Gapped q-gram 索引和 FM (Ferragina-Manzini)索引的优势[14]，该算法能够将读长映射到参考基因组上的多个潜在位置，并根据 reads 的质量和比对位置进行综合排序，从而确定最佳比对结果。Frith 等开发了 LAST 算法通过 adaptive seeds 策略提高了比对灵敏度[15]。Sedlazeck 等提出了 NGMLR，专门针对结构变异检测进行了优化[16]。Li 和 Durbin 开发的 BWA-MEM 算法在内存效率方面表现突出[17]。而 Jain 等开发的 Winnowmap2 则在重复序列识别和复杂基因组背景下取得了显著进展[18]。这些针对长读长特性优化的比对算法，普遍采用了误差容忍性设计、结构变异感知机制以及重复序列识别策略，在高错误率环境和复杂基因组背景下，展现出更高的比对准确性与运算稳定性。

在宏基因组研究中，长读长比对方法已经成为提升组装质量和解析复杂群落结构的核心技术。首先，长读长能够有效跨越宏基因组中常见的大量重复序列和结构变异区域，从而减少拼接过程中产生的断裂或错误拼接，最终获得更完整的基因组装配结果。其次，由于宏基因组样品中常存在高度相似的物种或菌株，短读长往往难以区分，而长读长测序技术可覆盖不同物种间差异区域的大片段序列，能够显著提高菌株分辨率，避免“混拼”现象[19]。长读长可覆盖整个功能基因簇(如抗生素合成基因簇)，避免了短读长拆碎后难以归位的问题，从而能够在宏基因组水平更好地解析功能基因[20]。

1.2 实时测序比对方法

实时测序技术是 DNA 或 RNA 分子通过纳米孔通道时边测序边产生数据并同步分析，而不是等待完整数据输出后再做批量离线处理。该特性催生了面向长读长、高错误率和流式数据的实时比对算法，并为物种鉴定、基因组组装、变异检测、转录组分析等下游分析提供有力支撑，进而直接服务于临床用药决策与生物安全应急[21]，这也是本课题组致力攻克的核心科学问题。

早期工作采用动态时间规整(dynamic time warping, DTW)算法将原始电流信号与模拟参考电流进行匹配，首次实现了纳米孔自适应采样(adaptive sampling)，有效降低了无效数据量，加速了实时分析速度[22]。随后 Kovaka 等提出 Uncalled，以概率模型与 FM 索引结合，在不经过碱基识别的前提下完成了原始信号到参考的实时映射，比对速度可以达到每秒处理数千条读长的水平[23]。在此基础上，Uncalled4 进一步引入了动态条带算法结合 basecaller 对数据进行处理，又将速度提升了数倍[24]。近年来，深度学习模型被用于对短片段信号或碱基序列进行实时分类，通过人工智能(artificial intelligence, AI)辅助决策是否保留或丢弃测序片段，从而实现更高效的序列筛选。例如 Senanayake 等基于改进 1D ResNet 架构的深度学习模型开发了 DeepSelectNet，该方法能够直接对原始电流信号进行端到端分类，快速判断目标物种或片段，准确率在 91%–99% 之间(平均约 95%)，在无需参考序列的条件下即可快速筛选目标物种和片段[25]。而 Zhang 等开发的最新工具 PROFIT-seq 在测序过程中进行实时控制，将获取的数据根据测序时间、通道编号和条形码进行碱基识别、分离与比对，依赖用户提供的测序配置决定是否继续或拒绝测序过程，提升了实时比对的灵活性与智能化水平[26]。

实时比对技术与宏基因组学研究具有高度契合性。在宏基因组测序中，样本常包含复杂的微生物群落，实时比对能够在测序进行过程中即时完成物种鉴定和目标片段筛选，还可通过自适应比对策略有效去除宿主来源序列，从而显著降低背景噪声并提升检出灵敏度。尤其在低丰度病原体或稀有物种的检测中，实时比对能够增强对关键病原体的捕获能力[27]。

1.3 错误率兼容比对方法

纳米孔测序与二代测序相比，其优势在于超长读长，但在技术商业化初期，原始碱基错误率通常在 10%−15%。现如今随着技术的不断升级，碱基准确率一直在稳步提升，但错误率仍高于一代测序数据和二代测序数据。较高的错误率不仅增加了直接分析的难度，还限制了纳米孔测序在变异检测、基因组组装和宏基因组等复杂应用中的发挥。尤其在宏基因组研究中，由于样本复杂、物种多样、参考序列不完整，错误率进一步加剧了比对与组装的困难[28]。

传统短读长比对采用“seek-and-extend” (“种子+扩展”)策略对数据进行处理，在处理纳米孔测序中较多的错配和插入时，则难以展现完美的性能[29]。为此，研究者们设计了多种针对长读长错误率较高的校正方法，例如 Li 等开发的 Minimap2 针对高错误率数据做了优化设计，通过高效的锚点链式比对算法串联多个种子，避免依赖单个种子延伸，可有效减少比对过程中的错误传播[12]；同时引入启发式 Z-drop 策略，允许在遇到长缺口时中止延伸，从而减少片段插入或错配；在宏基因组分析中，Minimap2 因其速度快、内存占用低等优点而被广泛用于物种鉴定与基因组拼接；但 Li 等也指出其对超长片段插入、倒位等错误区域的敏感性仍有不足。Sović 等提出的 GraphMap 是专为早期纳米孔测序高错误率数据比对开发的比对工具，采用多模式间隔种子和图形化比对模型，提高对错配和插入区域的比对敏感性[14]。同时还设计了多套不同模式的 k-mer 模板，在每个位置提取多个重叠较少的种子序列，确保即使存在错配或小片段插入的情况，也能够在复杂基因组区域获得更优的比对结果。后续研究中，该团队推出了 GraphMap2，增加了针对长读长转录组数据的拼接可变剪切比对能力，比对过程也更加严格[30]。这种比对方法对高错配、复杂插入的外源序列具有优势，尤其适合在物种复杂度高的样本中区分微生物。Jain 等开发了 Winnowmap 及其改进版本 Winnowmap2，延续了 Minimap2 的种子-链式架构，但在种子选取方面引入了加权 minimizer 取样策略[18]。具体来说，在重复区域可避免种子爆炸，但也牺牲了种子覆盖的均一性。为了解决这一问题，他们设计了基于权重的随机种子选择方案，从而使高频 k-mer 仍有一定概率被选中。这种加权方案保留了 minimizer 分布均匀的性质，同时避免了过多的错误匹配。而后续版本 Winnowmap2 在全基因组变异检测中识别插入或缺失的数量优于 Minimap2，可提升在宏基因组中的比对准确性[31]。Sedlazeck 等开发的 NGMLR 专门针对长读长序列测序数据设计，该算法使用一种基于凸间隙成本函数的比对算法，通过优化间隙成本函数，能够更好地处理长读长序列测序数据中的错误，并提高比对精度[16]；研究表明 NGMLR 更适用于宏基因组中的高精度结构变异分析，但不太适用于常规比对[32]。

此外，Ren 和 Chaisson 于 2021 年发布 LRA (long read aligner)，结合了凹形 gap 惩罚的连接模型，用以提高长读长准确性和变异敏感性；与传统线性 gap 惩罚不同，凹形函数对 gap 打开和延伸赋予递减的边际成本，使较长的单次 gap 相对更容易接受[33]。Kiełbasa 等将 LAST 算法应用于纳米孔早期数据分析[34]；在高错误率条件下仍能产出较高质量的比对，尤其适用于跨物种或高度多态性序列的比对以及复杂微生物群落的初步分析。如果比对工具缺乏对错误率的兼容性，就容易将序列错误误判为物种间差异，从而在物种鉴定过程中产生偏差。高容错性的比对方法能够在一定程度上降低这种误判风险，从而提升物种鉴定的灵敏度与准确性，尤其针对低丰度物种的检测。通过引入加权种子选择、链式比对和图模型比对等方法，能够在保留比对均一性的同时减少错误匹配，提升基因组拼装质量[35]。

1.4 宏基因组比对方法

近年来，宏基因组学技术尝试避开分离培

养步骤，直接从环境中提取总 DNA/RNA，以便获得可培养和难以培养微生物的全部遗传信息。纳米孔测序技术的长读长优势更利于覆盖完整基因，从而增加了获取新型生物活性物质的机会。为此，发展高效的长读长比对算法可为精准物种分类提供有力的分析工具[36]。

序列比对工具经历了重要的发展历程，1990 年 Altschul 等开发的 BLAST 算法采用启发式搜索策略，显著提高了序列搜索速度[37]。随后，2002 年 Kent 开发了 BLAT 工具，针对大规模基因组比对进行了优化，在人类基因组计划中发挥了重要作用[38]。虽然这些传统比对算法具有精度和灵敏度高的特点，但是在处理宏基因组海量数据时，其基于动态规划的计算复杂度导致运行速度显著降低，难以对宏基因组的大量数据进行快速响应。为了解决速度问题，研究者们提出了基于 k-mer 索引的改进算法。这一概念最早由 Belyi 等在 de Bruijin 图理论中系统阐述，通过建立参考数据库中 k-mer 短序列的快速哈希索引，实现宏基因组序列的快速分类和比对[39]。该方法将连续序列分解为固定长度的子序列，利用哈希表实现时间复杂度的快速查找。

针对宏基因组分类需求，Li 等开发了 DeSAMBA 工具[40]；其构建参考序列的 de Bruijn 图索引使用基于稀疏近似匹配块(sparse approximate matching block, SAMB)的伪比对算法，在保持分类准确的前提下显著提升了处理速度；基准测试显示，DeSAMBA 在分类精度接近 Minimap2 的同时，处理速度比 BLAST 和 BLAT 快 5−10 倍，特别适用于病原体检测等高速处理应用场景。另一重要的宏基因组序列分类技术路线由 Wood 等在 2014 年开创，其核心载体为 Kraken 系列分类工具，Kraken2 通过构建庞大的 k-mer 数据库实现了极快的分类速度，但对高错误率数据的准确性相对较低，也容易产生假阳性结果[41]。但 KrakenUniq 等工具通过 Unique k-mer 计数方式显著改善了假阳性较高的问题。Kim 等开发的 Centrifuge 采用 BWT (Burrows-Wheeler transform)索引策略提升比对速度，满足处理大规模数据速度需求；在基准测试中，k-mer 工具报告的物种数量通常比基于种子扩展的工具多一个数量级；相比之下，基于种子扩展的工具采用更保守的匹配策略，假阳性结果明显较少[42]。

当样本中存在未知物种时，严格匹配策略有助于可减少误报，但也可能漏掉重要信息；而积极匹配策略则能提升新物种的检出率，但可能引入误报风险。实际分析中，应根据需求选用适当工具或多工具结合互补。随着算法进步，新一代工具有望在速度、准确率和资源占用间取得更佳平衡，以满足宏基因组分析需求[43]。

1.5 结构变异检测比对方法

纳米孔测序具备超长读长和直接检测原始电信号的能力，为复杂结构变异(structure variations, SV)的识别提供了前所未有的可能[44]。然而，纳米孔测序数据误差率较高，对比对方法提出了更高要求。目前，主流的 SV 检测比对流程通常包括：错误模型驱动的长读长比对、基于比对差异的 SV 候选区间识别和断点精定位与 SV 类型分类[45-46]。

针对现有的结构变异检测比对方法，Chaisson 等采用了长读长比对工具和 SV 调用器相结合的方式[47]。Minimap2-cuteSV 组合利用 Minimap2 的“种子+扩展”策略实现长读长序列的快速定位，cuteSV 基于比对缺口中的较大缺失或插入，输出 SV 候选，该组合在中高覆盖度下实现了较优的得分和断点精度[48]。Heller 等则是通过 NGMLR 构建针对纳米孔特征的分段比对模型 NGMLR-SVIM，兼顾错配和缺失的双重误差[49]。SVIM 进一步通过分簇与过滤提高 SV 类型分类的准确性，特别适合检测复杂结构变异。Winnowmap-Sniffles2 是 Winnowmap 通过去除高重复区的种子降低错误比对，再由 Sniffles2 在调用阶段支持多线程、分级过滤，在速度与

召回率上兼具优势。Graph-based Alignment 工具则是通过 GVG (genome variation graph)方法将参考序列与已知 SV 整合为图结构，再用 Giraffe 等工具将长读长映射到图上，可同时检测已知与新型 SV，显著提升对复杂重排和拷贝数变异的识别能力。此外，近期兴起的原始信号直接比对(RawAlign)技术与基于当前电流建模的对齐(HQAlign)技术，均是利用纳米孔更底层的物理特性，实现对复杂 SV 的更高召回与精度[44,50]。在纳米孔长读长数据的 SV 分析领域，HQAlign 工具展现出独特技术优势。该工具在基于 Minimap2 的框架中融入纳米孔物理模型，将电流信号偏差纳入比对评分函数，显著改善了 SV 断点识别，精度提升 10%−50%[51]。

综上，在宏基因组学中，复杂的微生物群落及参考基因组的缺乏使结构变异检测面临巨大挑战。基于图的比对方法(如 GVG-Giraffe)通过整合已知 SV 与多参考基因组，更适合解析多样性极高的群落；而信号级比对方法(如 RawAlign、HQAlign)则利用纳米孔原始电信号，有效提升了低覆盖度和高错误率条件下的检测可靠性。同时，Winnowmap-Sniffles2 等工具能够在保持速度的同时检测拷贝数变异，为群落结构和功能推断提供支持。进一步地，Uncalled 与 DeepSelectNet 实现了实时目标筛选，推动了宏基因组在环境监测和临床病原体检测中的应用。表 1 为上述比对工具的比较分析。

不同工具适用于不同场景。在实际运用中，研究者应根据分析目标、数据特点(如错误率、物种复杂性、实时需求)和资源限制(如硬件、时间)等因素合理选择或集成使用。未来比对工具的发展方向将朝着多模态融合、图结构建模与智能加速协同的综合平台迈进，其协同发展将为纳米孔测序数据的多维解析提供坚实支撑。

表1　比对工具在不同应用维度下的比较分析

Table 1　Comparative analysis of comparison tools under different application dimensions

Comparison dimensions	Representative tools	Core algorithms and features	Advantages	Limitations
Long-read sequence alignment	Minimap2, GraphMap, NGMLR, Winnowmap2	Seed-based chaining combined with dynamic programming; error-tolerant alignment strategies	Fast (Minimap2); accurate (GraphMap)	Extremely long repetitive regions can reduce alignment accuracy (Minimap2)
Real-time sequencing alignment	UNCALLED, DeepSelectNet, Read Until, PROFIT-seq	Signal-level alignment; end-to-end neural classification; differentiable modeling frameworks	Highly real-time performance; rapid target sequence identification	Requires high-performance hardware and pre-trained models
Error-rate tolerant alignment	Minimap2, GraphMap2, NGMLR, Winnowmap, LRA, LAST	Error-tolerant anchoring; heuristic z-drop; convex gap penalty models	Can accommodate error rates up to 19%	Trade-offs between speed and accuracy depending on application context
Metagenomic alignment	DeSAMBA, Minimap2, Winnowmap2, GraphMap2, Kraken2, Centrifuge	Pseudo-alignment using de Bruijn graphs; sparse k-mer indexing strategies	Efficient handling of high-throughput metagenomic datasets	Limited scalability for novel species and incomplete reference genomes
Structural variation (SV) detection	Minimap2-cuteSV, NGMLR-SVIM, Winnowmap-Sniffles2, RawAlign, HQAlign	Chained anchoring with dynamic programming and SV-specific recognition modules	High topological sensitivity for detecting complex SVs	Accuracy constrained by signal-level modeling or improper parameter tuning

2 纳米孔测序的参考数据库体系

参考序列数据库是纳米孔测序宏基因组比对分析的“基准框架”，其测序序列完整性、注释准确性、更新时效性直接决定了病原物种鉴定的精准度、分类地位界定的可靠性乃至临床诊断结果的可信度[52]。相较于传统短读长测序，纳米孔测序技术凭借长读长优势，打破了传统短读长测序中复杂群落或低丰度病原检测中需依赖基因组组装的局限，实现了宏基因组病原微生物的直接比对分析，大幅提升了检测效率与特异性。然而，这一技术优势想要得到充分发挥，需基于高质量且高度适配的参考数据库[53]。因此，本节系统梳理了当前适配纳米孔测序的参考数据库资源，明确了高质量参考数据库的特点，为纳米孔测序技术在病原检测中的规范化应用提供参考。

2.1 现有参考数据库简介

2.1.1 综合参考数据库数据源解析

2005 年，美国国家生物技术信息中心(national center for biotechnology information, NCBI)、欧洲生物信息学研究所(European bioinformation institute, EBI)与日本 DNA 数据库中心(DNA data bank of Japan, DDBJ)联合成立了国际核酸序列数据库联盟(international nucleotide sequence database collaboration, INSDC)，建立了统一的技术体系与标准规范，定期共享数据。目前国际期刊已形成共识，推荐全球研究者在发表论文时将测序数据提交至 INSDC 数据库中。但 INSDC 数据库也存在序列质量较低、物种注释不准确等问题。为此，NCBI 不仅构建了用于归档基因序列信息的公共存储库(GenBank)，还建立了包括非冗余基因和蛋白质序列的数据库(RefSeq)和包含生物名称以及系统发育谱系的数据库(taxonomy)。其中，SRA (sequence read archive)是存储原始高通量测序数据的核心国际公共资源库[54]，包含了大量宏基因组数据集，并支持纳米孔测序数据在内的多种数据格式；RefSeq 收录了多个生物类群的高质量、无冗余、完整注释基因组，通过对参考基因组进行筛选、注释与质量评估，为基因识别、功能注释和比较基因组学等研究提供了可靠参考。

EBI 专注于生命科学数据存储与分析，维护了多个核心数据库。其中 Ensembl 研究项目最初聚焦脊椎动物的基因组及相关注释，后随着测序技术的发展纳入了更多基因组数据，除此之外，还提供比较基因组学、变异与调节注释等功能。EBI 开发了 EBI Metagenomics 宏基因组分析平台[55]，近年持续更新以支持纳米孔测序数据；整合了质量控制步骤和组装策略，提供读长质量过滤、重复序列处理、分类鉴定和功能注释等标准流程。Nicholls 等曾用该平台分析了纳米孔测序的微生物群落数据，验证了其对长读长数据的处理能力[56]。

除了上述 3 家综合数据库，近年来我国国家生物信息中心下属的国家基因组科学数据中心(national genomics data center, NGDC)逐渐成为国际核心生物信息资源，作为具有自主知识产权的综合数据库体系的先行者，支持多维基因组学数据的存储、管理和共享；根据最新数据，CNCB-NGDC 已收录 20 多万条高质量参考基因组、1 000 多万个生物样本关联组学数据、500 多万条非编码 RNA 序列和 30 多万套表观组数据[57]。其组学原始数据归档库(genome sequence archive, GSA)对标 SRA 数据库，基因组数据库(genome warehouse, GWH)和基因序列数据库(GenBase)对标 GenBank 数据库。CNCB-NGDC 还包含了 resource for coronavirus 2019 (RCoV19)的新型冠状病毒基因组、变异组以及元信息等数据整合[58]。所有数据均经过严格质控与人工校验，通过整合多实验室、多研究领域的组学注释信息，揭示数据注释的一致性与关联性，可支撑遗传病诊断、表观遗传机制研究，加速生物组学基础研究与精准医学探索。

此外，多个高质量的代表性生物数据库也同样适配纳米孔测序研究的要求，例如美国食品药品监督管理局(food and drug administration, FDA)联合 NCBI 构建的高质量测序病原微生物基因组数据库 FDA-ARGOS，针对公共数据库(如 GenBank)中基因组质量参差不齐导致的病原检测误判问题，内置诊断适配功能，可与 Minimap2 等长读长工具协同实现精准分类注释[59]。细菌和病毒生物信息学资源中心(bacterial and viral bioinformatics resource center, BV-BRC)是覆盖细菌、古菌、病毒(含流感病毒、致病 DNA/RNA 病毒)的综合性数据库，具备序列完整性、注释标准化和动态更新机制三大数据优势，适配纳米孔测序“无需组装直接比对”的特性[60]。SILVA 数据库提供较全面的、经过质量检查的、定期更新的对齐小亚基(16S/18S，SSU)和大亚基(23S/28S，LSU)核糖体 RNA 序列数据集，覆盖细菌、古细菌和真核生物这 3 个生命域，提供序列比对、分类和系统发育树构建工具[61]。RDP (ribosomal database project)数据库收集了大量细菌和古菌的 16S rRNA 基因序列，支持病原微生物的快速鉴定和分类学研究[62]。EzBioCloud 数据库作为病原微生物分类和鉴定的综合平台，提供了丰富的细菌和古细菌基因序列、鉴定工具及分类信息，便于菌种快速鉴定与分类[63]。值得注意的是，国际公共数据库虽能整合多数微生物基因组数据，但仍存在数据缺口，尤其是感染性真菌基因组数据稀缺，多数有感染史的真菌并未收录，给病原体检测出带来巨大挑战。因此，在整合公共数据的同时，开展病原微生物基因组测序、补充缺失数据，是完善临床微生物参考数据库完整性的关键。

2.1.2 适配纳米孔测序云端数据源解析

公共数据库整合了多个微生物基因组数据，针对初学者开展纳米孔测序宏基因组物种检测的需要，目前已有多款集成化数据库与工作流程可供直接选用。鉴于宏基因组数据处理涉及原始数据质控、组装、注释等专业化步骤，集成化在线分析平台可显著降低技术门槛。

纳米孔测序设备厂家的官方软件及比对工具配套数据库，是初学者开展测序数据深度挖掘与功能解析的重点资源[64]。Oxford Nanopore Technologies 的 EPI2ME Labs 云端平台，是预置数据库驱动的长读长分析标准化分析平台[65]。其宏基因组物种检测流程 wf-metagenomics 内置 Kraken2 (快速筛查)与 Minimap2 (精准验证)双模块，搭配多层次数据库：基础层包含 NCBI 专门收录 16S/18S/ITS 核糖体 RNA 序列的定向数据库，适配扩增子-宏基因组关联分析；通用层包含 Standard-8 (核心微生物物种)、PlusPF-8 (含潜在致病菌株)、PlusPFP-8 (补充噬菌体与质粒序列)，满足环境、临床等多场景需求。此外，其 wf-16S 流程依托长读长覆盖 16S rRNA 基因 V1−V9 完整可变区的优势，解决传统短片段(如 V3−V4 区)分类分辨率不足问题，实现高分辨率鉴定。该平台无需自行搭建分析流程，预置数据库经长读长兼容性优化，规避通用数据库的短读长偏好，并且相关源代码、更新日志及操作手册可通过 GitHub 与 EPI2ME 官方平台获取，便于研究者二次开发。

基于网页的生物医药大数据分析平台 Galaxy 具备零编程门槛(图形化操作界面)、全流程自动化(上传数据一键完成质控-比对-可视化)、结果可重复(记录分析步骤与参数，支持共享)及云端算力支持(免费存储空间)四大优势，通过“工具开源集成+流程定制化+多场景适配”适配纳米孔宏基因组分析[66]。针对病原检测需求，Galaxy 社区开发了 PathoGFAIR 项目流程集，“宿主序列去除→物种分类→基因层面病原鉴定→菌株变异分析→多样本可视化”形成闭环，支持处理食品、临床等来源的纳米孔数据，同时支持用户自主上传工具与数据库，其“工作流编辑器”可模块化组合工具并支持导出与共享数据，提升研究可重复性。

嘉兴南湖实验室(全国重点实验室)研发病原

体智能鉴定系统(genome pathogen analysis system, GPAS)，构建了“算法-模型-流程”三位一体的技术体系，填补了国内“高准确性-高易用性-高适配性”的病原体鉴定系统空白。研究者仅需上传测序数据即可自动完成病原体识别与鉴定，无需手动搭建分析流程，非专业人员也能快速掌握。针对深度测序大数据的解析难题，可快速定位潜在病原序列，为新发突发传染病预警、疑难感染病例诊疗提供技术支撑，构建未知病原体鉴定新范式。

2.1.3 环境宏基因组数据源解析

除了公开数据库外，一系列国际环境微生物普查项目可针对性弥补当前公共数据库的短板。这类项目覆盖全球不同生态环境(土壤、海洋、极端环境等)，能捕获大量未收录的“潜在病原相关微生物”，提供微生物生态关联信息，可丰富临床数据库的注释维度，提升病原体检出时的“环境-临床关联分析”能力，强化数据库实用价值。

地球微生物组计划(earth microbiome project, EMP)整合了全球 27 000 余个样本的 16S rRNA 扩增子数据，构建了覆盖土壤、水体、极端环境等多生态位的微生物基因组目录，含大量公共数据库未收录的潜在条件致病微生物(如特殊生境中的真菌、放线菌)，并基于 QIIME 2 (微生物生态学定量解析)分析框架支持微生物 β 多样性的多样本比对[67-68]。而 GOS (global ocean sampling)创建了首个全球海洋微生物多样性的大规模基因目录，填补海洋来源病原微生物(如弧菌科、海洋真菌)的基因组数据空白，依托 CAMERA 云计算平台将数百万条环境基因组的注释结果与代谢通路融合，为海洋微生物学研究提供了海量的公开序列数据[69]。CoML (census of marine life)历时 10 年构建了大量海洋物种的全球登记系统，集成大量分布记录的生物地理信息系统，通过标准化方法和跨学科合作提供海洋生物多样性基线数据，助力发现感染病例与海洋环境暴露的关联[70]。

此外，基因组在线数据库(genomes online database, GOLD)是基因组和宏基因组测序项目信息的核心存储平台，收录了超 20 000 个研究项目的成果，近 600 个宏基因组研究项目，提供快速搜索、高级搜索和元数据搜索这 3 种检索方式及可视化功能，可帮助临床研究者快速定位目标病原参考序列，解决“数据分散难获取”的问题[71]。人类微生物组计划(human microbiome project, HMP)系统性收集和分析了口腔、肠道、皮肤等人体关键部位的微生物样本，生成包含基因组和宏基因组数据在内的综合性数据库，重点标注微生物与人体健康状态的相关性，识别疾病相关微生物标记物[67]。

2.2 高质量参考数据库与配套资源构建

现有数据库最初主要面向一代和二代测序结果的存储和查询。随着纳米孔测序技术快速发展，数据库虽已尝试整合纳米孔数据，但仍存在数据模型不匹配、误差校正与置信度标注不足、格式与标准不统一、注释与可追溯性欠佳以及性能与可扩展性受限等问题[53]。当前参考数据库并非专为纳米孔数据构建，但业界已形成了对长读长更友好的参考数据库体系。本节围绕高质量参考数据库优先策略、标准化细分策略、多数据库联合策略和场景适配策略，详细阐述具体构建与资源选择方案，为具体实践提供清晰方向。

2.2.1 纳米孔高质量参考数据库构建

构建适用于纳米孔测序的高质量参考数据库，首先要进行分类信息确认与评估，公共数据库中的数据可能存在提交人操作失误导致的分类错误，因此需按 TaxID-拉丁名-同物异名三位一体校准核验，对争议条目以基因组或标记基因系统发育树复核查验，减少分类错误序列的纳入。其次开展测序数据质量评估，参考数据库中存在低质量或不完整的基因组序列会直接导致比对结果的假阳性率的升高，因此需对基因组序列进行去冗余处理，利用代表基因组减少重复 k-mer 引发的错配，与加权 minimizer

策略互补，同时 k-mer 或 minimizer 的筛选机制应与读长相匹配，最终优选质量高、完整度高的基因组序列。再次应重点关注宿主污染剔除，宏基因组检测过程中测序样本可能携带宿主污染序列，易导致假阳性，需将参考数据库中的序列反复比对宿主基因组，彻底去除污染序列，降低宿主基因干扰。此外需推进分类名称标准化，新的分子生物学证据(如基因组测序、系统发育分析)或形态学特征重新解析可能导致微生物分类层级变更。例如，某菌株原本被归类于某一属内的独立物种，后续研究发现其与另一属的亲缘关系更紧密，最终被调整至新属中，导致分类地位的变更。部分微生物最初的命名可能存在同义名、拼写错误或不符合国际命名法规，数据库应根据最新研究结论修正命名，确保物种名称的唯一性与合规性。最后需在实时监测或临床场景中保持定期更新与数据可溯源，需及时收录新物种，避免“未知病原无法匹配”问题。同时，通过可溯源管理，防止临床诊断数据回溯时信息变更导致的“误诊误判”，另外建议至少保持月度或季度更新，并附带版本号与变更日志，为高质量数据库实际应用提供保障。

2.2.2 宿主参考基因组数据库选择

宿主参考基因组是核心数据库“去污染”环节的关键支撑，需选择高质量、标准化的宿主参考序列，直接支撑宏基因组分析中宿主序列去除效率。Ensembl 主库收录 348 种物种基因组，涵盖人类、小鼠、斑马鱼等模式生物，可下载最新版本人类基因组参考序列(genome reference consortium, GRC)，直接导出适配 Kraken2、Minimap2 的索引文件，无缝对接 Galaxy 的宿主序列去除模块。NCBI RefSeq 包含所有已测序真核生物(人类、家畜、宠物、植物等)的参考基因组，每个物种标注推荐版本(如人类 GRCh38.p14)，并提供完整的 GenBank/RefSeq 编号与组装信息。加州大学圣克鲁兹分校(University of California, Santa Cruz, UCSC)开发的可视化基因组数据库(UCSC genome browser)，聚焦脊椎动物与模式生物，核心宿主资源包括人类(hg38)、小鼠(mm39)、大鼠(rn7)等。国际小鼠基因组信息资源中心(mouse genome informatics, MGI)专注于实验室小鼠的整合基因组数据库，是动物模型研究的核心宿主资源，适配基础科研场景的污染控制需求。

2.2.3 微生物参考数据库选择

宏基因组分析中，除依托宿主参考基因组数据库完成宿主序列过滤、精准比对病原微生物种类外，还需整合物种相对丰度、耐药基因和毒力基因等多维度关键信息，并关联病原微生物引起的临床症状，为临床诊断与用药提供丰富的参考信息，同时为未来新发未知病原体研究奠定依据[72]。

综合抗生素耐药性数据库(comprehensive antibiotic resistance database, CARD)是宏基因组分析中耐药基因解析的核心资源，整合了抗生素耐药性本体(antibiotic resistance ontology, ARO)与人工校验的耐药基因序列(antibiotics resistance genes, ARG)，可通过耐药基因识别器(resistance gene identifier, RGI)识别并预测基因组数据中的耐药基因，提供交互式的结果与 ARO 分类注释，为临床用药选择提供关键依据[73]。GO (gene ontology)数据库作为基因功能注释的标准化框架，可通过功能富集分析模块，识别基因集合中显著性聚集的功能类别，也可以通过功能预测模块，拓展未知基因的功能注释[74]。京都基因与基因组百科全书(Kyoto encyclopedia of genes and genomes, KEGG)的核心价值在于解析病原与宿主相互作用机制，其 KEGG NETWORK 子库通过通路映射，重点揭示疾病相关分子网络的扰动机制，系统整合基因变异、病原体感染及环境因素对细胞信号网络的动态影响[75-76]。同源基因簇数据库(clusters of orthologous genes, COG)通过聚类不同物种间具有共同祖先和保守功能的直系同源基因，为基因组注释及比较分析提供重要框架[77]。该数据

库将基因家族划分为 26 个标准功能类别，基于完整或接近完整基因组序列开展比对分析，常用于评估基因组测序完整性与组装质量[78]。

2.2.4 背景微生物参考数据库构建

纳米孔病原检测中，试剂工程菌、实验环境与操作人员携带微生物以及实验室残留与耗材生产环节引入的微生物均可成为“背景微生物/背景菌”，进入测序数据后易造成污染与假阳性[79]。为此，各实验室应结合自身采样与实验流程、样本类型与处理要求，建立并持续更新背景微生物参考数据库，在生信分析阶段过滤背景序列。同时，必须在每批次设置无模板对照以监控批次间背景信号，并将其检出的背景菌纳入过滤策略[80]。根据纳米孔的特性，关注长读长嵌合、接头残留、条码串扰、低复杂度区域误比对等特有假象，设置最小有效读长与质量门槛，必要时进行二次分割并重新比对。在下游分析时优先采用 RefSeq 或者模式菌株的高质量基因组，剔除低质量与污染组装，对同物种做去冗余[代表性菌株/平均核苷酸相似度(average nucleotide identity, ANI)聚类]，控制库体积与误匹配；并针对数据库条目及判读过程，系统记录试剂批次、操作人员信息、环境采样点位、实验日期、化学试剂配方、流动池信号及软件版本等核心元数据。

总之，纳米孔测序技术的发展加速了基因组序列数据的产生速度和扩张规模，但生物信息学分析流程与临床微生物基因组参考数据库标准化不足、数据安全与隐私保护制约价值挖掘的难题仍待破解[81]。而“质量优先-标准化细分-多数据库联合-场景适配”的核心策略，正是破解这些难题的关键。通过策略引领，可实现数据库构建的方向不偏、操作有序。研究者可以据此按应用场景逐层决策，而非在零散资源中摸索，从而最大程度发挥纳米孔长读长在物种鉴定、耐药监测和宏基因组解析中的优势，减少方向性偏差，为纳米孔测序技术的规模化落地筑牢基础。

3 总结与展望

纳米孔测序以单分子、长读长为核心优势，打通了高度重复区域、复杂结构变异与全长转录本的直接解析通道，从根本上补齐了短读长在组装连续性与功能注释上的短板[82]。与此同时，较高错误率与复杂场景对传统生信工具提出了更高要求，迫使比对算法与参考数据库向信号建模、图结构数据库构建与版本化治理三方面协同演进[83]。

在比对算法层面，本文梳理了长读长测序、实时测序、错误率兼容、宏基因组和结构变异检测这 5 种不同应用场景下的比对工具。其中长读长序列比对方法可有效提升复杂基因组背景下比对的准确性与运算稳定性；实时测序比对与选择性测序支撑“边测边判断”的临床诊断或病原体检测应急场景；错误率兼容比对方法在处理含有高噪声和测序错误的数据时保持较高的比对质量，减少假阳性结果；宏基因组比对方法在速度与精度间需因场景取舍；结构变异检测的比对方法正从基于线性参考基因组的读长比对，向图形参考基因组与信号级分析相融合的方向演化，以提高断点可解释性。

在数据与资源方面，不同比对方法为纳米孔宏基因组序列分析提供计算框架，而多维参考数据库协同则是结果可靠的根本。以综合数据库为基线资源，通过质量筛选与版本化管理解决其质量参差、误注释问题；以集成式云分析平台为初学者降低数据分析门槛，保障流程可复现；以环境宏基因组项目补充公共库缺失的潜在病原相关微生物序列，提升结果情境化判读能力；同时还需重视宿主数据库、常见的实验污染数据库及功能与耐药注释数据库的多库联动，降低假阳性率并减少批次间漂移。以高质量参考基因组数据库为基座，叠加多维数据与云端标准化流程，通过版本化与可溯源治理，实现又快又准的检出与功能注释，满足临床与大规模化应用。

未来，纳米孔测序需要在现有工具与资源框架上集中解决若干关键瓶颈。首先是针对纳米孔测序固有的高错误率问题，将 AI 模型中深度学习信号分类与测序误差校正算法融合，通过算法精准解析电信号，降低碱基识别误差。在此基础上升级测序化学试剂的特异性、提升纳米孔芯片的制造精度与孔径一致性，并针对其材料、电路设计和电信号采集模块精度等进行专业化改进，系统性降低测序错误率，从而提升碱基识别与变异检测的准确性。其次针对近缘物种基因组序列同源性高、区分难度大的问题，将图结构参考数据库与比对算法结合，在数据库中强化近缘物种基因组的特异性特征标注与分化位点识别，通过对特异性位点的优先匹配机制，实现近缘物种序列的精准区分，进一步提升复杂样本的解析能力。再次是建立跨机构标准化体系，不同实验室比对流程的差异、数据库版本的不一致等因素都会导致结果可重复性偏低，因此需要建立数据质控、工具选型、数据库版本的行业规范，同时需构建数据库与分析工具的动态更新维护机制，依托多中心临床与科研数据反馈闭环，明确校验标准、更新周期与版本追溯规则，确保结果的可重复性和可溯源性，形成可溯源和可复刻的临床分析流程。最后是兼顾数据共享与隐私保护，由于宏基因组临床样本包含敏感信息，现有技术难以在数据共享与隐私保护之间做好平衡。利用区块链实现数据溯源，打破数据孤岛，在符合数据领域科研伦理前提下深入挖掘数据价值。只有在减少错误产生、提升近缘物种分辨能力与持续演进数据库这 3 方面取得实质进展，并以标准化流程与合规的数据治理为支撑，纳米孔测序才能在物种鉴定、耐药监测、环境监控和公共卫生预警等关键场景中实现真正可规模化、可监管、可持续的应用。

作者贡献声明

李文正：初稿写作、修改及作图；张宁：稿件构思、稿件润色及修改；李卓越：框架搭建；崔莉煊：文献数据信息管理；王欣博：文献整理、稿件修改；杜耀华：监督指导、稿件润色及修改。

作者利益冲突公开声明

作者声明没有任何可能会影响本文所报告工作的已知经济利益或个人关系。

REFERENCES

[1] GOODWIN S, McPHERSON JD, McCOMBIE WR. Coming of age: ten years of next-generation sequencing technologies[J]. Nature Reviews Genetics, 2016, 17(6): 333-351.
[2] LOGSDON GA, VOLLGER MR, EICHLER EE. Long-read human genome sequencing and its applications[J]. Nature Reviews Genetics, 2020, 21(10): 597-614.
[3] AMARASINGHE SL, SU SA, DONG XY, ZAPPIA L, RITCHIE ME, GOUIL Q. Opportunities and challenges in long-read sequencing data analysis[J]. Genome Biology, 2020, 21(1): 30.
[4] SEDLAZECK FJ, LEE HY, DARBY CA, SCHATZ MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping[J]. Nature Reviews Genetics, 2018, 19(6): 329-346.
[5] RANG FJ, KLOOSTERMAN WP, de RIDDER J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy[J]. Genome Biology, 2018, 19(1): 90.
[6] CHAO JN, TANG FR, XU L. Developments in algorithms for sequence alignment: a review[J]. Biomolecules, 2022, 12(4): 546.
[7] XIA ZY, CUI YB, ZHANG A, TANG T, PENG L, HUANG C, YANG CQ, LIAO XK. A review of parallel implementations for the Smith-Waterman algorithm[J]. Interdisciplinary Sciences: Computational Life Sciences, 2022, 14(1): 1-14.
[8] DELAHAYE C, NICOLAS J. Sequencing DNA with nanopores: troubles and biases[J]. PLoS One, 2021, 16(10): e0257521.
[9] ESCOBAR-ZEPEDA A, GODOY-LOZANO EE, RAGGI L, SEGOVIA L, MERINO E, GUTIÉRREZ-RIOS RM, JUAREZ K, LICEA-NAVARRO AF, PARDO-LOPEZ L, SANCHEZ-FLORES A. Analysis of sequencing strategies and tools for taxonomic annotation: defining standards for progressive metagenomics[J]. Scientific Reports, 2018, 8: 12034.
[10] LANGSIRI N, WORASILCHAI N, IRINYI L, JENJAROENPUN P, WONGSURAWAT T, LUANGSA-ARD JJ, MEYER W, CHINDAMPORN A. Targeted sequencing analysis pipeline for species identification of human pathogenic fungi using long-read nanopore sequencing[J]. IMA Fungus, 2023, 14(1): 18.
[11] MAIOLO M, ZHANG XL, GIL M, ANISIMOVA M. Progressive multiple sequence alignment with indel evolution[J]. BMC Bioinformatics, 2018, 19(1): 331.

[12] LI H. Minimap2: pairwise alignment for nucleotide sequences[J]. Bioinformatics, 2018, 34(18): 3094-3100.
[13] KALIKAR S, JAIN C, VASIMUDDIN M, MISRA S. Accelerating minimap2 for long-read sequencing applications on modern CPUs[J]. Nature Computational Science, 2022, 2(2): 78-83.
[14] SOVIĆ I, ŠIKIĆ M, WILM A, FENLON SN, CHEN S, NAGARAJAN N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap[J]. Nature Communications, 2016, 7: 11307.
[15] FRITH MC, MITSUHASHI S. Finding rearrangements in nanopore DNA reads with LAST and dnarrange[J]. Methods in Molecular Biology, 2023, 2632: 161-175.
[16] SEDLAZECK FJ, RESCHENEDER P, SMOLKA M, FANG H, NATTESTAD M, von HAESELER A, SCHATZ MC. Accurate detection of complex structural variations using single-molecule sequencing[J]. Nature Methods, 2018, 15(6): 461-468.
[17] LI H, DURBIN R. Fast and accurate short read alignment with Burrows-Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760.
[18] JAIN C, RHIE A, HANSEN NF, KOREN S, PHILLIPPY AM. Long-read mapping to repetitive reference sequences using Winnowmap2[J]. Nature Methods, 2022, 19(6): 705-710.
[19] GREENMAN N, HASSOUNEH SA, ABDELLI LS, JOHNSTON C, AZARIAN T. Improving bacterial metagenomic research through long-read sequencing[J]. Microorganisms, 2024, 12(5): 935.
[20] NEGRI T, MANTRI S, ANGELOV A, PETER S, MUTH G, EUSTÁQUIO AS, ZIEMERT N. A rapid and efficient strategy to identify and recover biosynthetic gene clusters from soil metagenomes[J]. Applied Microbiology and Biotechnology, 2022, 106(8): 3293-3306.
[21] SAUERBORN E, CORREDOR NC, RESKA T, PERLAS A, da FONSECA ATUM SV, GOLDMAN N, WANTIA N, da COSTA CP, FOSTER-NYARKO E, URBAN L. Detection of hidden antibiotic resistance through real-time genomics[J]. Nature Communications, 2024, 15: 5494.
[22] SADASIVAN H, WADDEN J, GOLIYA K, RANJAN P, DICKSON RP, BLAAUW D, DAS R, NARAYANASAMY S. Rapid real-time squiggle classification for read until using RawMap[J]. Archives of Clinical and Biomedical Research, 2023, 7(1): 45-57.
[23] KOVAKA S, FAN YF, NI BH, TIMP W, SCHATZ MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED[J]. Nature Biotechnology, 2020, 39(4): 431-441.
[24] KOVAKA S, HOOK PW, JENIKE KM, SHIVAKUMAR V, MORINA LB, RAZAGHI R, TIMP W, SCHATZ MC. Uncalled4 improves nanopore DNA and RNA modification detection *via* fast and accurate signal alignment[J]. Nature Methods, 2025, 22(4): 681-691.
[25] SENANAYAKE A, GAMAARACHCHI H, HERATH D, RAGEL R. DeepSelectNet: deep neural network based selective sequencing for Oxford nanopore sequencing[J]. BMC Bioinformatics, 2023, 24(1): 31.
[26] ZHANG JY, HOU LL, MA LJ, CAI ZY, YE SJ, LIU Y, JI PF, ZUO ZQ, ZHAO FQ. Real-time and programmable transcriptome sequencing with PROFIT-seq[J]. Nature Cell Biology, 2024, 26(12): 2183-2194.
[27] PAYNE A, HOLMES N, CLARKE T, MUNRO R, DEBEBE BJ, LOOSE M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes[J]. Nature Biotechnology, 2020, 39(4): 442-450.
[28] MAGI A, SEMERARO R, MINGRINO A, GIUSTI B, D'AURIZIO R. Nanopore sequencing data analysis: state of the art, applications and challenges[J]. Briefings in Bioinformatics, 2018, 19(6): 1256-1272.
[29] FU SH, WANG AQ, AU KF. A comparative evaluation of hybrid error correction methods for error-prone long reads[J]. Genome Biology, 2019, 20(1): 26.
[30] LoTEMPIO J, DELOT E, VILAIN E. Benchmarking long-read genome sequence alignment tools for human genomics applications[J]. PeerJ, 2023, 11: e16515.
[31] DUTTA A, PELLOW D, SHAMIR R. Parameterized syncmer schemes improve long-read mapping[J]. PLoS Computational Biology, 2022, 18(10): e1010638.
[32] MIGA KH, KOREN S, RHIE A, VOLLGER MR, GERSHMAN A, BZIKADZE A, BROOKS S, HOWE E, PORUBSKY D, LOGSDON GA, SCHNEIDER VA, POTAPOVA T, WOOD J, CHOW W, ARMSTRONG J, FREDRICKSON J, PAK E, TIGYI K, KREMITZKI M, MARKOVIC C, et al. Telomere-to-telomere assembly of a complete human X chromosome[J]. Nature, 2020, 585(7823): 79-84.
[33] REN JW, CHAISSON MJP. Lra: a long read aligner for sequences and contigs[J]. PLoS Computational Biology, 2021, 17(6): e1009078.
[34] KIEŁBASA SM., WAN, R, SATO, K, HORTON, P, FRITH, MC. Adaptive seeds tame genomic sequence comparison[J]. Genome Research, 2011, 21(3): 487-493.
[35] SAHLIN K, BAUDEAU T, CAZAUX B, MARCHET C. A survey of mapping algorithms in the long-reads era[J]. Genome Biology, 2023, 24(1): 133.
[36] QUINCE C, WALKER AW, SIMPSON JT, LOMAN NJ, SEGATA N. Shotgun metagenomics, from sampling to analysis[J]. Nature Biotechnology, 2017, 35(9): 833-844.
[37] ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.
[38] KENT WJ. BLAT: the BLAST-like alignment tool[J]. Genome Research, 2002, 12(4): 656-664.
[39] BELYI I, PEVZNER PA. Software for DNA sequencing by hybridization[J]. Computer Applications in the Biosciences, 1997, 13(2): 205-210.
[40] LI GY, LIU YZ, LI DY, LIU B, LI JY, HU Y, WANG YD. Fast and accurate classification of meta-genomics long reads with deSAMBA[J]. Frontiers in Cell and Developmental Biology, 2021, 9: 643645.
[41] WOOD DE, SALZBERG SL. Kraken: ultrafast metagenomic sequence classification using exact alignments[J]. Genome Biology, 2014, 15(3): R46.
[42] KIM D, SONG L, BREITWIESER FP, SALZBERG SL. Centrifuge: rapid and sensitive classification of metagenomic sequences[J]. Genome Research, 2016, 26(12): 1721-1729.

[43] NAYFACH S, PÁEZ-ESPINO D, CALL L, LOW SJ, SBERRO H, IVANOVA NN, PROAL AD, FISCHBACH MA, BHATT AS, HUGENHOLTZ P, KYRPIDES NC. Metagenomic compendium of 189, 680 DNA viruses from the human gut microbiome[J]. Nature Microbiology, 2021, 6(7): 960-970.
[44] MAHMOUD M, GOBET N, CRUZ-DÁVALOS DI, MOUNIER N, DESSIMOZ C, SEDLAZECK FJ. Structural variant calling: the long and the short of it[J]. Genome Biology, 2019, 20(1): 246.
[45] De COSTER W, WEISSENSTEINER MH, SEDLAZECK FJ. Towards population-scale long-read sequencing[J]. Nature Reviews Genetics, 2021, 22(9): 572-587.
[46] LIU Z, XIE Z, LI MX. Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data[J]. Genome Biology, 2024, 25(1): 188.
[47] CHAISSON MJP, SANDERS AD, ZHAO XF, MALHOTRA A, PORUBSKY D, RAUSCH T, GARDNER EJ, RODRIGUEZ OL, GUO L, COLLINS RL, FAN X, WEN J, HANDSAKER RE, FAIRLEY S, KRONENBERG ZN, KONG XM, HORMOZDIARI F, LEE D, WENGER AM, HASTIE AR, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes[J]. Nature Communications, 2019, 10: 1784.
[48] JIANG T, LIU YZ, JIANG Y, LI JY, GAO Y, CUI Z, LIU YD, LIU B, WANG YD. Long-read-based human genomic structural variation detection with cuteSV[J]. Genome Biology, 2020, 21(1): 189.
[49] HELLER D, VINGRON M. SVIM: structural variant identification using mapped long reads[J]. Bioinformatics, 2019, 35(17): 2907-2915.
[50] LINDEGGER J, FIRTINA C, GHIASI NM, SADROSADATI M, ALSER M, MUTLU O. RawAlign: accurate, fast, and scalable raw nanopore signal mapping *via* combining seeding and alignment[J]. IEEE Access, 2024, 12: 196855-196865.
[51] JOSHI D, DIGGAVI S, CHAISSON MJ P, KANNAN S. HQAlign: aligning nanopore reads for SV detection using current-level modeling[J]. Bioinformatics, 2023, 39(10): btad580.
[52] SMITH RH, GLENDINNING L, WALKER AW, WATSON M. Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome[J]. Animal Microbiome, 2022, 4(1): 57.
[53] CHORLTON SD. Ten common issues with reference sequence databases and how to mitigate them[J]. Frontiers in Bioinformatics, 2024, 4: 1278228.
[54] KATZ K, SHUTOV O, LAPOINT R, KIMELMAN M, BRISTER JR, O'SULLIVAN C. The sequence read archive: a decade more of explosive growth[J]. Nucleic Acids Research, 2022, 50(D1): D387-D390.
[55] HUNTER S, CORBETT M, DENISE H, FRASER M, GONZALEZ-BELTRAN A, HUNTER C, JONES P, LEINONEN R, McANULLA C, MAGUIRE E, MASLEN J, MITCHELL A, NUKA G, OISEL A, PESSEAT S, RADHAKRISHNAN R, ROCCA-SERRA P, SCHEREMETJEW M, STERK P, VAUGHAN D, et al. EBI metagenomics: a new resource for the analysis and archiving of metagenomic data[J]. Nucleic Acids Research, 2014, 42(Database issue): D600-D606.
[56] NICHOLLS SM, QUICK JC, TANG SQ, LOMAN NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards[J]. GigaScience, 2019, 8(5): giz043.
[57] CNCB-NGDC Members and Partners. Database resources of the national genomics data center, China national center for bioinformation in 2025[J]. Nucleic Acids Research, 2025, 53(D1): D30-D44.
[58] LI CP, MA LN, ZOU D, ZHANG RQ, BAI X, LI L, WU GG, HUANG TH, ZHAO W, JIN EH, BAO YM, SONG SH. RCoV19: a one-stop hub for SARS-CoV-2 genome data integration, variant monitoring, and risk pre-warning[J]. Genomics, Proteomics & Bioinformatics, 2023, 21(5): 1066-1079.
[59] SICHTIG H, MINOGUE T, YAN Y, STEFAN C, HALL A, TALLON L, SADZEWICZ L, NADENDLA S, KLIMKE W, HATCHER E, SHUMWAY M, ALDEA DL, ALLEN J, KOEHLER J, SLEZAK T, LOVELL S, SCHOEPP R, SCHERF U. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science[J]. Nature Communications, 2019, 10: 3313.
[60] OLSON RD, ASSAF R, BRETTIN T, CONRAD N, CUCINELL C, DAVIS JJ, DEMPSEY DM, DICKERMAN A, DIETRICH EM, KENYON RW, KUSCUOGLU M, LEFKOWITZ EJ, LU J, MACHI D, MACKEN C, MAO CH, NIEWIADOMSKA A, NGUYEN M, OLSEN GJ, OVERBEEK JC, et al. Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR[J]. Nucleic Acids Research, 2023, 51(D1): D678-D689.
[61] QUAST C, PRUESSE E, YILMAZ P, GERKEN J, SCHWEER T, YARZA P, PEPLIES J, GLÖCKNER FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools[J]. Nucleic Acids Research, 2013, 41(Database issue): D590-D596.
[62] COLE JR, WANG Q, FISH JA, CHAI BL, McGARRELL DM, SUN YN, BROWN CT, PORRAS-ALFARO A, KUSKE CR, TIEDJE JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis[J]. Nucleic Acids Research, 2014, 42(Database issue): D633-D642.
[63] YOON SH, HA SM, KWON S, LIM J, KIM Y, SEO H, CHUN J. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies[J]. International Journal of Systematic and Evolutionary Microbiology, 2017, 67(5): 1613-1617.
[64] THOMAS T, GILBERT J, MEYER F. Metagenomics-a guide from sampling to data analysis[J]. Microbial Informatics and Experimentation, 2012, 2(1): 3.
[65] MORSLI M, BOUDET A, KERHARO Q, STEPHAN R,

SALIPANTE F, DUNYACH-REMY C, HOUHAMDI L, FOURNIER PE, LAVIGNE JP, DRANCOURT M. Real-time metagenomics-based diagnosis of community-acquired meningitis: a prospective series, southern France[J]. EBioMedicine, 2022, 84: 104247.

[66] COMMUNITY G. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update[J]. Nucleic Acids Research, 2024, 52(W1): W83-W94.

[67] AN DS, CAFFREY SM, SOH J, AGRAWAL A, BROWN D, BUDWILL K, DONG XL, DUNFIELD PF, FOGHT J, GIEG LM, HALLAM SJ, HANSON NW, HE ZG, JACK TR, KLASSEN J, KONWAR KM, KUATSJAH E, LI C, LARTER S, LEOPATRA V, et al. Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common[J]. Environmental Science & Technology, 2013, 47(18): 10708-10717.

[68] NAYFACH S, ROUX S, SESHADRI R, UDWARY D, VARGHESE N, SCHULZ F, WU DY, PAEZ-ESPINO D, CHEN IM, HUNTEMANN M, PALANIAPPAN K, LADAU J, MUKHERJEE S, REDDY TBK, NIELSEN T, KIRTON E, FARIA JP, EDIRISINGHE JN, HENRY CS, JUNGBLUTH SP, et al. A genomic catalog of Earth's microbiomes[J]. Nature Biotechnology, 2020, 39(4): 499-509.

[69] VENTER JC, REMINGTON K, HEIDELBERG JF, HALPERN AL, RUSCH D, EISEN JA, WU DY, PAULSEN I, NELSON KE, NELSON W, FOUTS DE, LEVY S, KNAP AH, LOMAS MW, NEALSON K, WHITE O, PETERSON J, HOFFMAN J, PARSONS R, BADEN-TILLSON H, et al. Environmental genome shotgun sequencing of the Sargasso Sea[J]. Science, 2004, 304(5667): 66-74.

[70] Census of Marine Life. First census of marine life 2010: highlights of a decade of discovery[R]. Washington : Census of Marine Life, 2010.

[71] REDDY TK, THOMAS AD, STAMATIS D, BERTSCH J, ISBANDI M, JANSSON J, MALLAJOSYULA J, PAGANI I, LOBOS EA, KYRPIDES NC. The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification[J]. Nucleic Acids Research, 2015, 43(Database issue): D1099-D1106.

[72] WU F, ZHAO S, YU B, CHEN YM, WANG W, SONG ZG, HU Y, TAO ZW, TIAN JH, PEI YY, YUAN ML, ZHANG YL, DAI FH, LIU Y, WANG QM, ZHENG JJ, XU L, HOLMES EC, ZHANG YZ. A new coronavirus associated with human respiratory disease in China[J]. Nature, 2020, 579(7798): 265-269.

[73] McARTHUR AG, WAGLECHNER N, NIZAM F, YAN A, AZAD MA, BAYLAY AJ, BHULLAR K, CANOVA MJ, de PASCALE G, EJIM L, KALAN L, KING AM, KOTEVA K, MORAR M, MULVEY MR, O'BRIEN JS, PAWLOWSKI AC, PIDDOCK LJV, SPANOGIANNOPOULOS P, SUTHERLAND AD, et al. The comprehensive antibiotic resistance database[J]. Antimicrobial Agents and Chemotherapy, 2013, 57(7): 3348-3357.

[74] GENE ONTOLOGY CONSORTIUM, ALEKSANDER SA, BALHOFF J, CARBON S, CHERRY JM, DRABKIN HJ, EBERT D, FEUERMANN M, GAUDET P, HARRIS NL, HILL DP, LEE R, MI H, MOXON S, MUNGALL CJ, MURUGANUGAN A, MUSHAYAHAMA T, STERNBERG PW, THOMAS PD, van AUKEN K, et al. The Gene Ontology knowledgebase in 2023[J]. Genetics, 2023, 224(1): iyad031.

[75] ARAMAKI T, BLANC-MATHIEU R, ENDO H, OHKUBO K, KANEHISA M, GOTO S, OGATA H. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold[J]. Bioinformatics, 2020, 36(7): 2251-2252.

[76] KANEHISA M, FURUMICHI M, SATO Y, MATSUURA Y, ISHIGURO-WATANABE M. KEGG: biological systems database as a model of the real world[J]. Nucleic Acids Research, 2025, 53(D1): D672-D677.

[77] GALPERIN MY, WOLF YI, MAKAROVA KS, VERA ALVAREZ R, LANDSMAN D, KOONIN EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens[J]. Nucleic Acids Research, 2021, 49(D1): D274-D281.

[78] GALPERIN MY, KRISTENSEN DM, MAKAROVA KS, WOLF YI, KOONIN EV. Microbial genome analysis: the COG approach[J]. Briefings in Bioinformatics, 2019, 20(4): 1063-1070.

[79] DAVIS NM, PROCTOR DM, HOLMES SP, RELMAN DA, CALLAHAN BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data[J]. Microbiome, 2018, 6(1): 226.

[80] MILLER S, NACCACHE SN, SAMAYOA E, MESSACAR K, AREVALO S, FEDERMAN S, STRYKE D, PHAM E, FUNG B, BOLOSKY WJ, INGEBRIGTSEN D, LORIZIO W, PAFF SM, LEAKE JA, PESANO R, DeBIASI R, DOMINGUEZ S, CHIU CY. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid[J]. Genome Research, 2019, 29(5): 831-842.

[81] WHEELER NE, PRICE V, CUNNINGHAM-OAKES E, TSANG KK, NUNN JG, MIDEGA JT, ANJUM MF, WADE MJ, FEASEY NA, PEACOCK SJ, JAUNEIKAITE E, BAKER KS. Innovations in genomic antimicrobial resistance surveillance[J]. The Lancet Microbe, 2023, 4(12): e1063-e1070.

[82] WANG YH, ZHAO Y, BOLLAS A, WANG YR, AU KF. Nanopore sequencing technology, bioinformatics and applications[J]. Nature Biotechnology, 2021, 39(11): 1348-1365.

[83] LIAO WW, ASRI M, EBLER J, DOERR D, HAUKNESS M, HICKEY G, LU SJ, LUCAS JK, MONLONG J, ABEL HJ, BUONAIUTO S, CHANG XH, CHENG HY, CHU J, COLONNA V, EIZENGA JM, FENG XW, FISCHER C, FULTON RS, GARG S, et al. A draft human pangenome reference[J]. Nature, 2023, 617(7960): 312-324.