PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2017, 44(1): 58~69
www.pibb.ac.cn

# 基于高通量测序技术的微生物检测数据分析方法 *

周子寒[1)] 彭绍亮[1)**] 伯晓晨[2)] 李 非[2)**]
([1)] 国防科学技术大学计算机学院软件研究所，长沙 410073；[2)] 军事医学科学院放射与辐射医学研究所，北京 100850)

**摘要** 高通量测序技术的发展正在逐渐改变诸多生物学领域的研究方法．为应对突发疫情以及新发未知微生物威胁的需求，微生物鉴定技术逐渐从传统的物理化学方法及核酸杂交等分子水平方法进一步走向利用无需培养的测序数据进行快速分析检测．随之而来的是对高通量数据分析在精度及速度的要求．基于高通量测序数据的微生物检测数据分析方法在近些年得到了快速的发展．本文分析了目前基于高通量测序数据的微生物检测数据分析方法，对其数据分析的处理流程和计算方法进行了研究，比较了各个微生物检测数据分析方法的特点及适用场景．最后结合本实验室工作总结微生物检测数据分析方法在实际应用中可能遇到的问题，希望对该应用领域的研究有一定的参考意义．

**关键词** 高通量测序，微生物检测，数据分析方法，性能评测
**学科分类号** Q811.4
**DOI**: 10.16476/j.pibb.2016.0239

微生物检测旨在通过传统生化、免疫实验方法或者高通量测序方法鉴别宏基因组样品中的微生物种类和定量信息．常用传统微生物检测方法包括涂片镜检法、PCR 扩增法以及基因芯片法等．涂片镜检法通过对样本微生物进行染色，观察大小形态，与图例进行对比从而完成检测工作．其优势为快速、成本低和不需要特殊仪器[1]．基因芯片法通过荧光标记探针杂交显示特异微生物的序列信息和位置．其优势为敏感度高、检测快速[2]．PCR 扩增法利用寡核苷酸引物引导待测基因片段进行扩增，从而能够有效增强检测信号，提高检测效率[3]．但是传统微生物检测方法都难以解决未知微生物的检测问题：首先，如果样本中存在未知微生物，则无法通过形态学特征等获取其种属信息，分离培养法和涂片镜检法不再适用；其次，PCR 扩增法和基因芯片技术的探针设计需要对样本的先验知识有所了解，同样也难以鉴别未知微生物[4]．

快速准确地检测宏基因组样本中的微生物，确定其种源、毒力等信息是疾病防控和生物安全的关键问题[5]．新发未知的微生物由于其突发性、无法获取先验知识等特征，难以通过传统微生物检测方法进行快速、有效的应对．第二代测序(next-generation sequencing，NGS)技术经过 10 年左右的快速发展，功能不断完善，成本逐渐降低，一次运行可以测定千万级别的短序列[6]．基于 NGS 的微生物检测的完整流程包括：通过对宏基因组样本进行大规模完全测序，得到宏基因组的核酸序列；再利用生物信息学工具对核酸序列数据进行分析，从而进一步得到微生物基因、耐药性、毒力信息等[7]．基于 NGS 的微生物检测技术无需进行需要先验知识的样本形态学特征查找或者探针设计，能够对未知致病微生物进行检测，弥补了传统微生物检测方法的不足，成为预防未知生物威胁的重要手段．

随着测序实验技术的日趋成熟，数据分析方法逐渐成为制约基于 NGS 的微生物检测应用的关键环节．本文对近些年来基于 NGS 的微生物检测数据分析方法的流程设计和关键算法做简要介绍，比较各个数据分析方法的特点及适用情况．最后总结面向应用需求的微生物检测数据分析方法，对相关领域的未来走向提出了预测．希望本文对应用高通

* 国家自然科学基金重大计划(U1435222), 国家自然科学基金面上项目(61402486)和全军后勤科研计划重点项目(BWS14C051)资助.
** 通讯联系人.
彭绍亮. Tel: 13574817196, E-mail: 13574817196@163.com
李 非. Tel: 010-66932251, E-mail: pittacus@gmail.com
收稿日期：2016-11-02，接受日期：2016-12-14

量测序技术进行微生物检测的工作人员在数据分析方面提供有价值的参考.

## 1　基于 NGS 的微生物检测数据分析策略

根据样本来源、提取方法、建库策略的不同，基于 NGS 的微生物检测数据分析方法所采取的数据处理策略也略有不同. 微生物样本多来自血液、口腔、痰液、病理组织等，在进行 DNA 或 RNA 的提取后，利用高通量测序仪建库测序，即获得原始 reads 数据，存储格式通常为 fasta 或 fastq 文件，后者包含碱基质量信息. 测序原始数据需要进一步的数据分析以获得其中微生物种群的相关信息[8].

基于 NGS 的微生物检测数据分析方法面临以下几个关键问题：第一，检测数据分析方法速度要快. 由于 NGS 产出数据的通量越来越高，检测数据分析方法的速度需要与之相匹配，才能够达到快速确认，快速应对的目的[9]. 第二，检测数据分析方法精度要高. 面向未知微生物的检测数据分析方法需要尽可能降低检测的假阳性和假阴性比例. 如果假阳性过高，无法有效确认疑似微生物，会造成时间和效率的浪费；如果假阴性过高，则无法充分检测出宏基因组样品中的微生物序列，影响后期确认和毒力评估[10]. 另外，NGS 测序仪的测序读长通常为 35～250 bp(碱基)，比第一代测序的读长(650～800 bp)短，因此需要进行 *de-novo* 拼接以获得完整微生物序列信息[11]. 除此之外微生物检测数据分析方法还会遇到计算性能上的问题，需要提高检测方法在不同体系结构上的运行效率，或通过算法优化检测方法所需的计算量等. 针对这些问题，目前多数基于 NGS 的微生物检测方法可以归纳为以下数据分析策略(图 1).

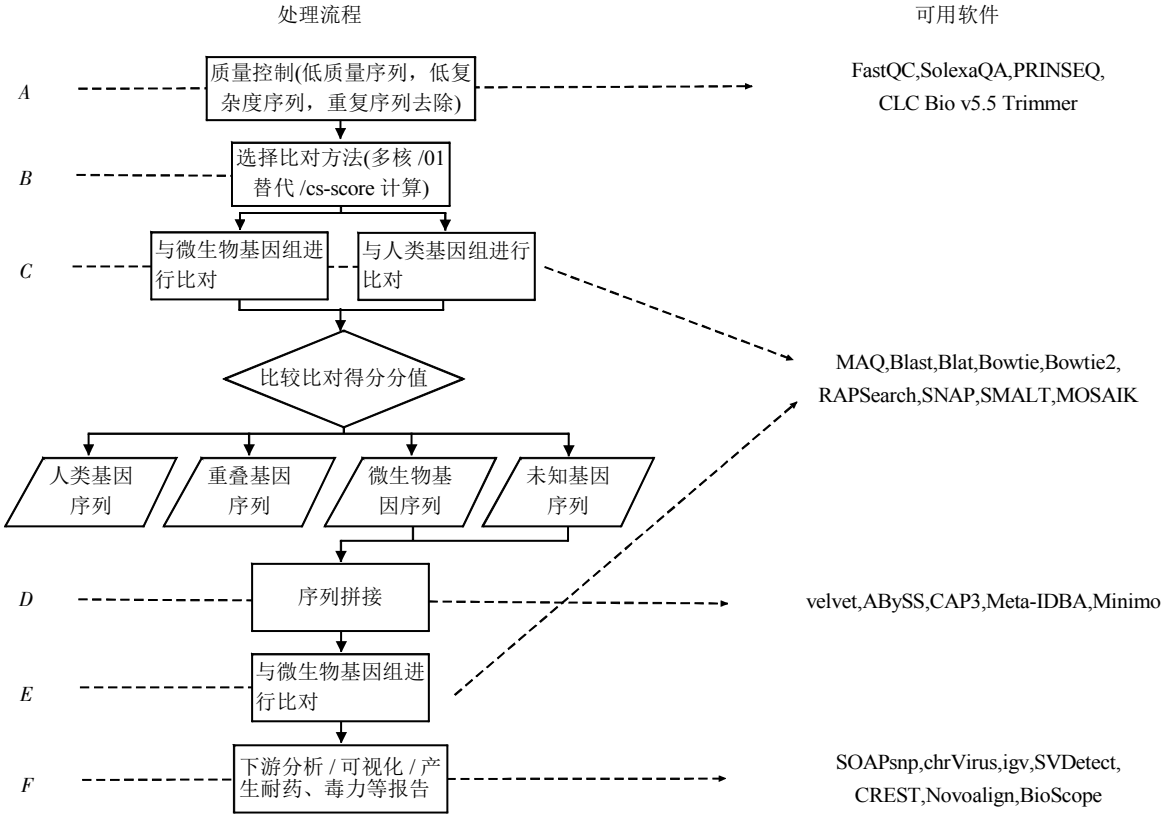


**Fig. 1　Strategy of microbes detection data analysis methods based on NGS**

**图 1　基于 NGS 的微生物检测数据分析方法策略**

为了对目前基于 NGS 的微生物检测数据分析方法进行全面评估，我们将分析系统简化为 6 个核心环节，如图 1 所示，分别为：质量控制(*A*)、比对方法优化(*B*)、与参考基因组进行比对(*C*)、序列拼接(*D*)、与微生物基因组进行比对(*E*)、下游分析(*F*). 其中质量控制(环节 *A*)属于提高微生物检测精度的方法，通常 NGS 产生的测序数据会存在低质量序列、低复杂度序列等，会影响后续分析. 因此

需要通过质量控制软件对样本测序数据进行处理．比对方法优化(环节 *B*)属于提高微生物检测速度的方法，由于 NGS 产生上百万片段，数据分析在单个节点上的计算时间可能需要数周，因此需要采用设计算法的查询和匹配效率，以降低数据分析所需的运算量，在数据分析环节上提高速度．序列比对(环节 *C*，*E*)为微生物检测的核心步骤[12]，将样本数据和人类参考基因组(human reference geneset，HRG)或微生物参考基因组(microbe reference geneset，MRG)进行比对，清除与 HRG 比对成功的序列，保留与 MRG 比对成功的序列，最后获得已知和未知的微生物基因．序列拼接(环节 *D*)将检测出的微生物基因短 reads 拼接成完整微生物基因序列，从而进行下游分析(环节 *F*)，如可视化[13]、单核苷酸多态性(single nucleotide polymorphisms，SNP)分析[14]等，才能获得基因序列的全部信息，完成微生物检测的整个流程．

## 2　基于 NGS 的微生物检测数据分析方法

从 2011 年开始，基于 NGS 的微生物检测数据分析方法得到了快速的发展．目前已经发表的数据分析方法包括 PathSeq[15]、RINS[16]、CAPSID[17]、VirusSeq[18]、VirusFinder[19]、READSCAN[20]、Kraken[21]、SURPI[22]、RIEMS[23]、Pathosphere.org[24]、CS-SCORE[25]、VERSE[26]和 VIP[27]等，如表 1 所示．

**Table 1　Microbes detection data analysis methods based on NGS**

**表 1　基于 NGS 的微生物检测数据分析方法**

| 数据分析方法名称 | 发表时间 | 核心比对方法 | 数据分析策略 |
|---|---|---|---|
| PathSeq | 2011 | MAQ[28], MegaBlast[29], BlastN[30] | *A-C-D-E* |
| RINS | 2012 | Bowtie[31], Blast[30] | *A-C-D-E* |
| CAPSID | 2012 | Blat[32], MegaBlast | *A-C-D-E-F* |
| VirusSeq | 2013 | MOSAIK[33] | *A-C-D-E-F* |
| VirusFinder | 2013 | Blast+[34], Blat | *A-C-D-E-F* |
| READSCAN | 2013 | SMALT[35] | *A-B-C-D-E* |
| Kraken | 2013 | K-mer classification tree | *A-B-C-D-E* |
| SURPI | 2014 | SNAP[36], RAPSearch[37] | *A-B-C-D-E* |
| RIEMS | 2015 | MegaBlast, BlastN | *A-B-C-D-E-F* |
| Pathosphere.org | 2015 | GS Newbler[38], Bowtie2[39] | *A-B-C-D-E-F* |
| CS-SCORE | 2015 | BWA[40] | *A-B-C-D-E-F* |
| VERSE | 2015 | Bowtie2 | *A-B-C-D-E-F* |
| VIP | 2016 | Bowtie2, RAPSearch | *A-B-C-D-E-F* |

参照前文提到的数据分析策略，可以发现不同的微生物检测计算分析方法主要体现在比对方法优化(环节 *B*)以及下游分析(环节 *F*)．早期的数据分析方法如 PathSeq、RINS 等，是最初微生物检测数据分析的典型解决方案，并未专门对比对方法进行优化，也缺少下游分析的步骤，属于“基础型”检测方法．而 VirusSeq、VirusFinder、CaPSID 等，在“基础型”的基础上，增加了如病毒结合位点分析等下游分析(环节 *F*)，其检测功能得到了完善，称为“功能型”检测方法．而 READSCAN、Kraken、CS-SCORE 在“基础型”的基础上，增加了比对方法优化(环节 *B*)，其处理速度大大增加，属于“速度型”检测方法．近几年出现的 SURPI、RIEMS、Pathosphere.org、VERSE 和 VIP，同时包含比对方法优化和下游分析，能够提供完整的分析流程，属于“完整型”检测方法．

除了数据分析策略带来的基础特点外，由于采用了不同的软件、处理流程等，不同的基于 NGS 的微生物检测数据分析方法还有着不同的优势，如图 2 所示．

a．在计算速度方面，以 CS-SCORE 的 cs-score 值计算法、Kaken 的精确 $k$-mer 匹配法、RINS 的基于先验知识的加速法为代表的比对方法优化(环节 *B*)，对核心比对流程进行了算法上的改进，使其运行速度大大增加．b．在计算精度上，以 PathSeq 为代表的多次循环比对法，在最大限度上规避了微生物参考基因组 MRG 不完善导致的精度下降．c．在计算资源方面，Kraken 的精简数据库法和 CS-SCORE 的基于 cs-score 值计算法使运行内存大大降低．d．体系结构方面，目前基于 NGS

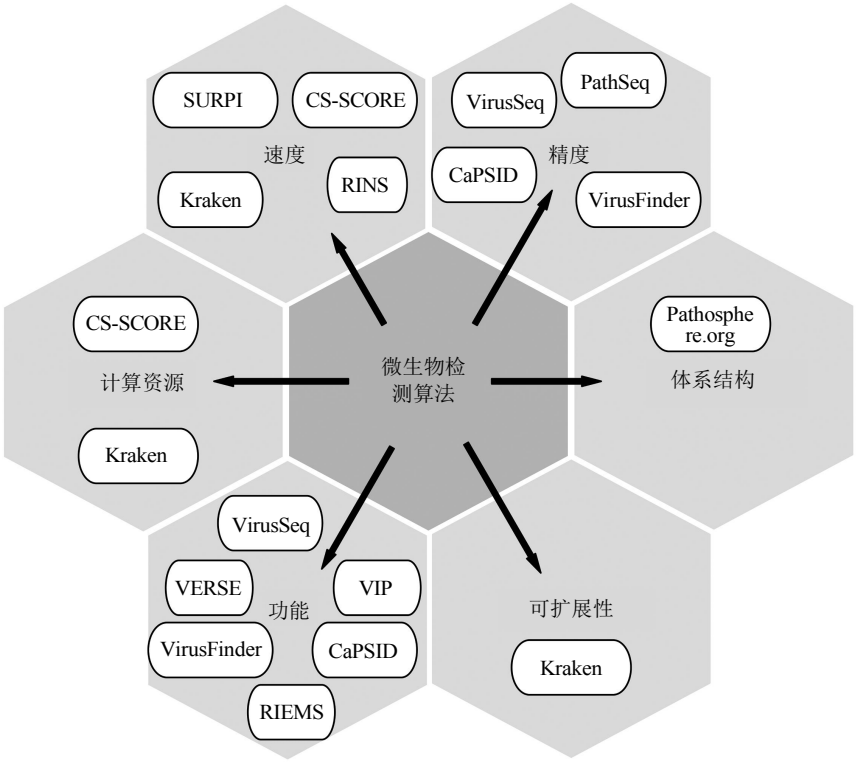


**Fig. 2　Advantages of microbes detection data analysis methods based on NGS**
**图 2　基于 NGS 的微生物检测数据分析方法优势**

的数据处理方法都是基于 Linux 开发，但是 Pathosphere.org 工作流程为上传数据到云服务器，在服务器上完成计算产生结果报告，对于运行平台没有要求．e．在可扩展性方面，以 VirusSeq 和 Kraken 为代表的方法都提供了多线程处理选项，能够有效增加运行效率．f．在功能方面，以 Virusseq、CaPSID 为代表的病毒结合位点分析以及 SNP 分析等都属于对于后续功能的完善．尽管最近几年都有新的检测方法被提出，但是早期的检测方法由于其他方面的优势，在不同的应用场景下，可与新的方法互补使用．例如在高精度的病毒转录组数据检测中，VirusFinder 的效果较 Kraken 更好．这里总结基于 NGS 的微生物检测流水线的适用情况，如图 3 所示．

在上述基于 NGS 的微生物检测数据分析方法中，一类方法针对检测未知微生物进行序列比对精度上的优化，其处理方式为从样本文件中逐步清除人类基因序列，最后剩余包含已知和未知生物基因序列，典型方法包括 Pathseq、CS-SCORE 等，另一部分则侧重快速检测已知微生物，其处理方式为直接与微生物基因组进行比对，但不足之处在于，受限于所选取的参考基因组，无法检出未知或罕见的微生物种类，典型算法包括 RINS、Kraken 等．在输入数据方面，一些方法只能处理 DNA 序列文件，如 Kraken，有的方法只能处理 RNA 序列文件，如 READSCAN，大部分方法两者都能处理，包括 VERSE、SURPI 等．在检测应用领域方面，VERSE、VirusSeq、VirusFinder 作为专门检测病毒基因序列的方法，采用的参考数据库只包含病毒基因序列．鉴于病毒参考基因组远小于细菌参考基因组，此类软件分析更为快速便捷．

下文我们将对不同类型的检测方法做进一步的介绍：

a．“基础型”检测方法

“基础型”检测方法采用的数据处理策略为 A-C-D-E，包含 PathSeq 和 RINS．属于基于 NGS 的微生物检测方法刚起步的阶段．其中 PathSeq 针对微生物检测的精度问题进行了优化，RINS 针对微生物检测的速度问题进行了优化．

PathSeq 是 2011 年麻省理工学院和哈佛大学联

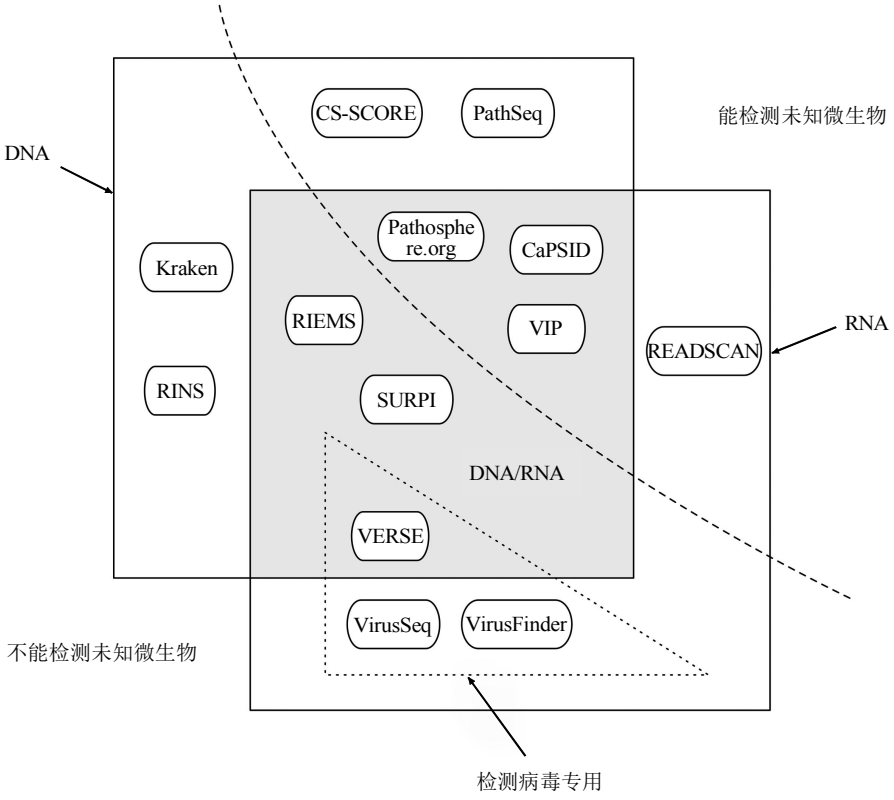


**Fig. 3　Application condition of microbes detection data analysis methods based on NGS**

**图 3　基于 NGS 的微生物检测数据分析方法适用情况**

合研究所的 Kostic[15]提出的基于 Amazon 云平 台[41]的微生物检测数据分析方法．PathSeq 提出了多次过滤法，用以提高微生物检测的精度：在进行将输入样本宏基因组数据与参考基因组做序列比对(环节 *C*)时，采用 MAQ、MegaBlast、BlastN 对其中包含的人类基因序列进行多次循环过滤，充分去除人类基因序列．PathSeq 的缺点为由于多次进行序列比对，其运行速度较慢．PathSeq 适用于对检测速度要求不高，精度要求很高的案例，譬如查找大规模疫情中的新型细菌病毒．Bhatt 等[42]在 2014 年将该数据分析方法用于巨细胞动脉炎(giant cell arteritis，GCA)病原体检测，样本来自于 17 名 GCA 患者，采用 PathSeq 将其中的人类 DNA 序列去除，然后对剩余序列进行了聚类分析．该实验中 PathSeq 对人类 DNA 序列的清除率为 100%，起到了良好的效果．

RINS(rapid identification of nonhuman sequences)是 2012 年 Bhaduri[16]提出的基于先验知识的微生物检测数据分析方法．RINS 提出了基于先验知识的加速法，用以提高微生物检测的速度．其处理流程与一般微生物检测数据分析方法不同，首先针对微生物的物理化学性质(形态学观测等方法)，对其种属进行假设，根据假设结果选择部分微生物参考基因组 MRG 与样本数据进行序列比对，确认其中是否包含已知微生物序列．由于选择的微生物参考基因组远远小于人类参考基因组 HRG，先验知识法可以达到减小运算量，加快运行速度的效果．其比对过程(环节 *C*)采用的软件为 BLAST．但如果参考微生物基因组选择错误，则需要重新选择参考基因组，造成计算冗余，增加运算时间．RINS 数据分析方法适用于快速检测常见症状的、潜伏期短的突发疫情中的微生物．Bhaduri 等利用 RINS 对 CA-HPV-10 前列腺癌细胞序列数据 130 万个长度为 100 bp 的 reads 进行检测用时 2 h，而 PathSeq 在处理相似数据量的问题用时约为 13 h，说明 RINS 在检测速度上的提升效果显著．

b．“功能型”检测方法

“功能型”检测方法采用的数据处理策略为

*A-C-D-E-F*，包括 CaPSID、VirusSeq 和 VirusFinder. 功能型检测方法的特点为相比于基础型检测方法，增加了下游分析环节，如 CaPSID 的对检测出的微生物序列进行可视化查看功能，以及 VirusSeq 和 VirusFinder 的病毒结合位点分析功能等.

CaPSID 是 2012 年 Boron[17]提出的微生物检测数据分析平台： CaPSID 的主要优势在于功能多样，不仅能够实现微生物检测，还能够通过转录组数据进行病原体微生物的可视化查看. CaPSID 的检测流程为：对于样本数据，首先采用 Bowtie2(环节 *C*)与 MRG 进行比对，随后与 HRG 进行比对，之后根据比对结果分为病原体基因序列、人类基因序列、既比对到人类参考基因组又比对到微生物参考基因组上的重叠序列以及未比对成功序列. 最后利用 Novoalign[42]对未比对成功序列和病原体基因序列进行下游分析(环节 *F*). CaPSID 适用于检测未知物理化学性质的微生物，并且能够将其分类和可视化查看其基因水平的特征(G-C 含量，变异水平等).

VirusSeq 是 2012 年 Anderson[18]提出的针对病毒基因序列检测的数据分析方法； VirusFinder 是 2013 年 Wang[19]提出的针对病毒基因序列检测的数据分析方法. 两种方法思路大致相同，首先将样本数据与 HRG 进行比对，清除比对成功的部分序列；之后将剩余基因序列与 MRG 进行比对，提取出比对成功的基因序列进行 *de-novo* 拼接，将拼接成功的基因序列再次与 MRG 进行比对，最后输出比对成功的病毒基因序列. VirusSeq 和 VirusFinder 均提供了检测病毒结合位点的下游分析功能(环节 *F*)，具体实现为通过 SVDetect[43]对单核苷酸多态性进行分析，判断病毒插入结合位点的信息. 两者的不同在于序列比对(环节 *C*)过程中，VirusFinder 采用的序列比对方法为 Bowtie2，速度较 VirusSeq 所采用的 MOSAIK 有所提升. VirusSeq 和 VirusFinder 只能进行病毒基因序列的检测，适用于处理由病毒引发的疫情，测序数据为 RNA 的微生物检测问题.

c. “速度型”检测方法

“速度型”检测方法采用的数据处理策略为 A-B-C-D-E，包括 READSCAN 和 Kraken. 速度型检测方法的特点为相比与基础型检测方法，增加了比对算法优化环节， 包括 READSCAN 的并行数据划分法和 Kraken 的精简数据库法.

READSCAN 是 2013 年 Naeem[20]提出的微生物检测程序. 该方法专注于解决检测数据分析方法的速度和可扩展性问题，提出了并行数据划分法. 对于样本数据，首先分配到多核 CPU 上(环节 *B*)，随后分别与 HRG、MRG 进行比对，其中在环节 *C* 采用的软件为 SMALT，通过对比对结果的分值进行分析来确定 reads 属于人类基因序列还是微生物基因序列. READSCAN 分为两个版本，用于处理正常基因序列数据的 normal 版和用于处理高变异基因序列数据的 high-sensitivity 版. READSCAN 数据分析方法适用于在大规模计算集群上检测突发疫情中的已知或未知微生物.

Kraken 是 2013 年 Wood[21]提出的的宏基因组序列分类软件，能够快速对宏基因样品中的 DNA 序列进行分类，因此可以进行微生物检测. Kraken 在序列比对环节(环节 *C*)采用精确 $k$-mer 匹配和精简数据库的方法(环节 *B*)，忽略基因变异，采取精确匹配；并且建立了专用数据库与 $k$-mer 匹配相配合，极大地提高了检测速度. Kraken 分为两个版本：内存开销较大的 normal 版和将内存开销限制为 2Gb 以内的 mini 版. Kraken 速度极快，精度较低，适用于做微生物检测的预处理.

CS-SCORE 是 2015 年 Haque[22]提出的用于快速鉴定和去除宏基因组数据中的人类基因数据的数据分析方法. CS-SCORE 专注于解决检测数据分析方法的速度以及计算资源的问题，提出了 cs-score 值代替法(环节 *B*). 对于样本数据，首先进行四核苷酸频率计算，并与经过聚类的序列数据库进行比对(环节 *C*)，比对结果相似性超过预先设定的比对阈值则认为属于病原体基因序列. 将相似性未超过阈值的部分进行向量化，计算 cs-score 值. 随后，根据 cs-score 值将序列和经过同样方法处理的人类基因序列分别分为 31 个子集，再通过 bwa 进行比对(环节 *C*)，最后输出未比对成功的序列即为未知微生物基因序列. 其优点在于速度较快，并且所需的内存量较小，仅为 2～2.5GB. CS-SCORE 适用于在小规模运算平台上处理突发疫情中的微生物检测问题.

d. “完整型”检测方法

“完整型”检测方法采用的数据处理策略为 *A-B-C-D-E-F*，包括 SURPI、RIEMS、Pathosphere、org、CS-SCORE、VERSE 和 VIP. 完整型检测方法既包含了下游分析环节，又包含比对方法优化环节. 如 cs-score 基于 CS-SCORE 的比对算法，Pathosphere.org 基于网站的分析方法等等. 对于微

生物检测来说，该类检测方法提供了完整的分析步骤.

SURPI 是 2014 年 Naccache[23]提出的基于云平台的微生物检测数据分析方法，SURPI 专注于解决微生物检测的速度问题，采用的比对软件为 SNAP 和 RAPSearch，其处理流程为，首先对样本数据进行质量控制(环节 *A*)得到 clean data，随后分为两种模式．第一种为快速模式，采用 SNAP 软件将 clean data 与 MRG 包括细菌参考基因组和病毒参考基因组进行比对(环节 *C*)，比对成功直接生成结果报告．第二种为综合模式，首先将 clean data 与 NCBI nt 完整参考数据库进行比对(环节 *C*)，之后通过 ABySS[44]+Minimo[46]软件进行 *de-novo* 拼接组装(环节 *E*)，最后通过 RAPSearch 软件与病毒蛋白质数据库进行比对，输出结果(环节 *F*)．SURPI 的优点在于速度极快，由于采用 SNAP 和 RAPSearch 比对软件，可以在 10～30 min 完成对 7～50M reads 的判定．SURPI 适用于处理大规模疫情中的已知微生物的检测问题.

RIEMS 是 2015 年 Scheuch[24]提出的微生物检测数据分析方法．RIEMS 专注于解决微生物检测的精度问题．其处理流程为，首先对输入宏基因组数据进行质量控制(环节 *A*)，随后利用 BLAST 将读入的 reads 子集与已知微生物种群进行比对(环节 *C*)．之后对未能检测的序列进行序列组装(环节 *E*)，组装成功的利用 blast 与微生物参考基因组进行比对，未组装成功的利用 Megablast 和 Blastn 再与微生物基因组进行比对(环节 *C*)．下游分析过程(环节 *F*)利用 Megablast 和 Blastn 的“without DUST”[46]模式将读入未比对成功序列进行比对，并对未能比对成功的 reads 和 contigs 进行翻译，最后利用 Blastp[47]对生成的 Open Reading Frames(ORF)进行分析．RIEMS 适用于对转录组数据进行高精度分析的过程应用.

Pathosphere.org 是 Kilianski 等[25]提出的基于网站的病原体检测数据分析方法平台．允许用户将测序样品传输至云平台，然后在线分析产生结果报告．其处理流程为，对于样本数据，首先进行质量控制以及宿主核酸序列噪声去除，随后进行一个循环——第一步为 *de novo* 拼接，将去除噪声后的 reads 拼接成 contigs，第二步为与临近物种进行比对判断(环节 *C*)和 SNP 分析(环节 *F*)，第三步为将没能成功比对的 reads 重新拼接，对于这三步进行循环操作直至序列比对结果达到设定的阈值．Pathosphere.org 在环节 *C* 采用的比对软件为 Megablast．由于 Pathosphere.org 基于云平台设计，无需环境配置，在不同平台上都可使用.

VERSE 是 2015 年 Wang[26]提出的微生物检测数据分析方法．其处理流程为：首先将输入的样本数据与 HRG 进行比对(环节 *C*)，未比对成功的序列视为病毒序列．随后将假定的病毒序列与 MRG 进行比对，并通过这一环节寻找 SNP，利用找到的 SNP 调整 MRG．将调整后的 MRG 与 HRG 连接，生成一个伪染色体，随后利用伪染色体与原输入数据进行比对，从而判断病毒结合位点和染色体内结构变异(Structural Variants，SVs)，之后利用找到的 SVs 调整 HRG．最后将调整后的 HRG 和调整后的 MRG 连接生成一个新的参考基因组．将输入病毒序列与新参考基因组进行比对，可以分别检测出病毒基因和人类基因内的病毒结合位点和结构变异(环节 F)．VERSE 适用于处理病毒引发的疫情中的微生物检测问题.

VIP 是 2016 年 Li[27]提出的微生物检测数据分析方法．VIP 其处理流程为：对于样本数据，首先进行去接头，低质量序列去除，低复杂度序列去除(环节 *A*)，得到 clean data．随后分为两种模式，牺牲精度换取速度的快速模式和注重精度的高敏感度模式．两种模式的区别体现在宿主基因组消减(环节 *C*)环节．快速模式下，利用 Bowtie2 将 clean data 与 MRG 进行比对，比对成功后直接输出结果．在高敏感度模式下，首先去除 clean data 中的细菌 DNA 和全部 rRNA，剩余序列再利用 Bowtie2 与 MRG 进行比对．未比对成功的序列再利用 RAPSearch 与 NCBI 的病毒蛋白质库进行比对．所有的比对成功的序列都进行 *de novo* 组装和系统发育分析(环节 *F*)．VIP 适用于高精度要求下的微生物检测问题.

## 3 “速度型”检测方法的安装配置及性能评估

为了评估基于 NGS 的微生物检测数据分析方法，我们这里对“速度型”数据分析方法进行了安装配置与性能评估(表 2)，也就是：Readscan、Kraken 和 CS-SCORE，其中 Kraken 采用缩减数据库的 Kraken-mini 版本．评估的内容包括：环境配置依赖、安装方法、计算速度、结果精度以及参考数据库比较等．测试环境为共享内存计算系统(Intel Xeon E7-8800 V3 CPU x 8、1TB MEM、

SSD、CentOS release 6.5 server).

环境配置依赖和安装方法参见 github 项目 https://github.com/FreezeFish/MicrobeDetectionEvaluation.git. 项目给出了 Kraken-mini，CS-SCORE 和 Readscan 的数据下载，环境配置等自动化安装方法.

测试数据采用 3 个微生物宏基因组数据文件，分别称为“HiSeq”、“MiSeq”和“simBA-5”，来自于文献[21]. 根据文献[21]，HiSeq 和 MiSeq 序列数据来自两组不同的细菌宏基因组，均包含 10 个物种，通过鸟枪法测序得到. simBA-5 宏基因数据由提高 5 倍突变率的细菌和古生物细菌序列混合构建. 这 3 个序列文件都包含 10 000 000 个 reads，其中 HiSeq 平均读长为 92 bp，MiSeq 平均读长为 156 bp，simBA-5 所有 reads 读长均为 100 bp.

计算速度采用总碱基量除以单次运行时间进行评估，单位采用 Mbp/s，结果保留两位有效数字.

结果正确性采用精度(precision，prec)和敏感度(sensitivity，sens)两个指标进行评估. 待测方法在运行完成后均会给每条 reads 打出分类标签(assignments)，精度采用正确的标签占能够进行分类的标签的比例，敏感度指正确的标签占所有标签的比例.

**Table 2　“Speed” analysis methods evaluation results**

**表 2　“速度型”分析方法的评估结果**

| | HiSeq | | | MiSeq | | | simBA-5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec | Sens | Speed | Prec | Sens | Speed | Prec | Sens | Speed |
| Readscan | 99.98 | 66.52 | 0.017 | 99.98 | 76.94 | 0.026 | 100 | 67.36 | 0.025 |
| Kraken-mini | 100 | 74.63 | 2.3 | 100 | 71.04 | 2.03 | 100 | 66.93 | 2.21 |
| CS-SCORE | 98.16 | 96.28 | 0.21 | 99.88 | 96.74 | 0.22 | 99.65 | 100 | 0.096 |

从运行结果中可以看出，Kraken-mini 速度最快，CS-SCORE 其次，Readscan 运行速度较慢. 并且随着 reads 长度增加，Kraken-mini 的运行速度降低，但总体速度差别不大. Readscan 处理速度随着 reads 长度增加反而有所增加. CS-SCORE 运行速度和 reads 长度没有确定关系. 在结果正确性方面，Kraken-mini 对三个测试数据集都给出了 100%的精度，这是因为 Kraken-mini 采用精确 $k$-mer 匹配，不考虑突变造成的影响. 而 Readscan 和 CS-SCORE 则存在将微生物序列 reads 判别成人类序列 reads 的情况. Kraken-mini 的敏感度较差，是由于采用 $k$-mer 的比对方法，$k$-mer 越小意味着特异性降低，和多条 reference 比对成功的几率更高，而采用 LDA 分类的方法只能向上一级进行分类，导致敏感度降低.

Kraken-mini 提供了多线程处理方法，处理各个样本数据结果如图 4.

对于 HiSeq 和 simBA5，Kraken 在 8 线程时能提供 > 0.737 的加速比，但是到 16 线程加速比锐减到 < 0.460. 而 MiSeq 在 16 线程仍有 0.745 的加速比，由于 MiSeq 数据量较大，结果合理.

参考微生物数据库是否全面对软件的检测结果影响较大，因此这里对三种方法的数据库进行了综合比较，结果如表 3.

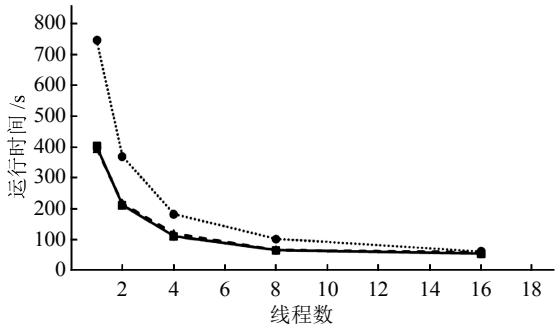


**Fig. 4　Scability of Kraken-mini**

**图 4　Kraken 可扩展性**

■—■: HiSeq; ●┈┈●: MiSeq; ▲---▲: simBA5.

“速度型”检测方法针对微生物检测过程中的速度问题进行了优化(环节 $B$)，如 Kraken 的精简数据库法，CS-SCORE 的基于 cs-score 的比对方法等，都在很大程度上加速了比对核心流程. 但是速度的增加会不可避免地带来精度的下降，比如降低参考基因组大小会增加检测速度，但是会降低比对过程中正确匹配的 reads 数量，从而降低敏感度，

**Table 3　MRG database of “speed” analysis methods**

**表 3　“速度型”分析方法的微生物参考数据库比较**

| | Readscan | Kraken-mini | CS-SCORE |
|---|---|---|---|
| 数据来源 | NCBI RefSeq 细菌，真菌，病毒基因组 | 2014.12.8 NCBI RefSeq 细菌，古细菌，病毒基因组 | 参考数据库只有人类基因组 |
| 微生物物种数目 | 细菌 2 206，真菌 364 157，病毒 328 922 | 完整基因组的 5% k-mer | 无 |
| 数据类型 | 全基因组 | 特征片段 | 通过“组成特征”进行分类的片段 |
| 结论 | 运行时只能指定单个细菌、真菌或病毒参考基因组，降低敏感度 | 由于采取特征片段，会降低敏感度 | 采用与 PathSeq 类似的缩减人类基因组序列的方法 |

增加参考基因组大小能够提高结果正确性，但是由于数据量增加，处理速度也会相应降低．目前最新的检测方法一般提供了两种模式．第一种模式为只将样本序列与 MRG 进行比对的快速模式，其优点在于检测速度快，缺点在于无法检测 MRG 中缺失的序列；另一种为逐步清除人类基因序列的方式，优点是能够检测未知微生物序列，缺点是检测速度较低．用户可以根据需求选择适合自己的检测模式，这也是在微生物检测的速度和精度之间进行平衡的最佳方案．

## 4　总结与展望

基于 NGS 技术的微生物检测数据分析方法具有无需预先培养样本、灵敏度高、能够检测未知的微生物的特点，为疾病防控和生物食品安全提供了新的解决方案．本文对目前常见的 12 种基于 NGS 的微生物检测方法进行了简要介绍和比较研究，对各个检测方法的软件流程和数据处理方案分别进行了环节上和优化方向上的分析．在数据分析环节方面，将基于 NGS 的微生物检测数据分析方法分为“基础型”，“功能型”，“速度型”，“完整型”等四种类型．在优化方向方面，侧重两个性能指标：速度、精度以及四个应用因素：计算资源、体系结构、能耗和可扩展性方面进行比较分析．通过对现有数据分析方法及其实现的总结评价，希望为生物和计算领域的相关研究提供参考价值．

本实验室也对该领域的研究进行了一定的探索：李定辰等[48]针对从非培养样本中鉴定未知病原微生物的问题，从软件层面对微生物检测流程进行了分析评估的软件包括序列比对软件 Bowtie2、BWA、BLAST+、MUMer[49]；基因组拼接软件 Velvet[50]、SOAPdenovo[51]；后续处理软件 BEDTools[52]、MEGAN4[53]、MAUVE[54]、IGV[55]、Circos[56]等．分别从常规病原体检测、高突变率下的病原体检测、不同测序深度以及读长下的病原体检测、数据量不充分时的病原体检测和混合样品中的病原体检测几个方面，对以上软件做了系统评估，其研究和评估结果将有助于指导以后临床病原微生物鉴定分析工作．叶福强等[57]针对胆总管结石的微生物组研究问题，使用宏基因组测序手段，对 15 位中国胆总管结石患者的胆汁样本进行全宏基因组鸟枪法测序和 16 S 核糖体扩增子测序分析，其研究发现了 13 个之前未报道的高基因组覆盖度的胆道细菌．还鉴别出与胆石形成汁耐受相关的基因，是微生物检测技术在鉴别新型微生物上的实际应用．王恒等[58]开发了“天河二号”上的基于 Intel MIC 的高通量 DNA 序列比对并行软件，其中 DNA 序列比对软件 MICA[57]结合天河二号超级计算机软硬件架构设计，能够充分发挥 MIC 的并行潜力，具有接近线性加速比的扩展性能．李定辰的工作属于从细粒度上对基于 NGS 的微生物检测流程做出分析，如果能够将粗粒度(检测方法)和细粒度(软件)分析相结合，则能够使人更清晰地分析和总结微生物检测的过程．叶福强的工作属于传统微生物检测方法的进一步发展，如果能够将 16 S 分析和基于 NGS 的微生物检测相结合，则能够使检测更加准确．王恒的工作能够推动基于 NGS 的微生物检测方法向速度更优方向发展．结合本实验室其他研究，则能够对相关领域的研究作出促进．另外本人所在课题组建立并发布一套用于综合评估基于 NGS 微生物检测的计算分析方法的性能评测数据和工具集，包含不同突变率的测试数据、不同规模的测试数据以及与真实数据接近的模拟数据，相关学术成果已投稿 PDP 2017 国际会议．

未来基于 NGS 的微生物检测方法的发展方向可以在速度和精度上做重点优化，除此之外，现有

计算分析方法在计算资源、能耗、体系结构等方面也都存在着优化空间．计算资源方面，通过精简参考基因集等方法能够有效降低运行内存；也可以通过专用硬件加速卡来解决，缩减微生物检测数据分析方法的处理时间，更好地面对生物威胁．在体系结构方面，可通过上传数据至云计算服务器，可以使检测方法有更高的操作系统适配性等．如 GPU、ARM 低能耗处理器以及 FPGA 都能够从不同层次上对检测方法进行加速或其他方面的优化．

除了计算技术方面的优化，生物技术的创新也能给微生物检测方法带来革新．未来生物技术的发展使测序技术向着高通量、低成本、长读取长度的方向发展，目前已接近实用的第三代测序技术具有超长读长的特点，一旦应用将会极大地改善微生物检测方法的流程．对微生物检测数据分析方法提出了新的要求，这就需要新的数据分析方法能够适应测序技术，在速度和精度上达到更高的标准，为微生物检测领域提供更快速、更准确的微生物检测方法．

## 参 考 文 献


[1] Steingart K R, Henry M, Ng V, *et al*. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. The Lancet Infectious Diseases, 2006, **6** (9): 570−581

[2] Lemieux B, Aharoni A, Schena M. Overview of DNA chip technology. Molecular Breeding, 1998, **4**(4): 277−289

[3] Belgrader P, Benett W, Hadley D, *et al*. Rapid pathogen detection using a microchip PCR array instrument. Clinical Chemistry, 1998, **44**(10): 2191−2194

[4] Call D R. Challenges and opportunities for pathogen detection using DNA microarrays. Critical Reviews in Microbiology, 2005, **31**(2): 91−99

[5] Lazcka O, Del Campo F J, Munoz F X. Pathogen detection: a perspective of traditional methods and biosensors. Biosensors and Bioelectronics, 2007, **22**(7): 1205−1217

[6] Schuster S C. Next-generation sequencing transforms today's biology. Nature, 2007, **200**(8): 16−18

[7] Barzon L, Lavezzo E, Costanzi G, *et al*. Next-generation sequencing technologies in diagnostic virology. Journal of Clinical Virology, 2013, **58**(2): 346−350

[8] Reis-Filho J S. Next-generation sequencing. Breast Cancer Research, 2009, 11(3): 1−8

[9] Metzker M L. Sequencing technologies—the next generation. Nature Reviews Genetics, 2010, **11**(1): 31−46

[10] Mandal P, Biswas A, Choi K, *et al*. Methods for rapid detection of foodborne pathogens: an overview. American Journal of Food Technology, 2011, **6**(2): 87−102

[11] Li R, Zhu H, Ruan J, *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research, 2010, **20**(2): 265−272

[12] Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics, 2010, **11**(5): 473−483

[13] Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, **9**(3):87−402

[14] Wang D G, Fan J B, Siao C J, *et al*. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science, 1998, **280**(5366): 1077−1082

[15] Kostic A D, Ojesina A I, Pedamallu C S, *et al*. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature Biotechnology, 2011, **29**(5): 393−396

[16] Bhaduri A, Qu K, Lee C S, *et al*. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics, 2012, **28**(8): 1174−1175

[17] Borozan I, Wilson S, Blanchette P, *et al*. CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics, 2012, **13**(1): 157

[18] Chen Y, Yao H, Thompson E J, *et al*. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics, 2013, **29**(2): 266−267

[19] Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PloS One, 2013, **8**(5): e64465

[20] Naeem R, Rashid M, Pain A. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics, 2013, **29**(3): 391−392

[21] Wood D E, Salzberg S L. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol, 2014, **15** (3): R46

[22] Naccache S N, Federman S, Veeraraghavan N, *et al*. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Research, 2014, **24**(7): 1180−1192

[23] Scheuch M, Höper D, Beer M. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics, 2015, **16**(1): 1

[24] Kilianski A, Carcel P, Yao S, *et al*. Pathosphere. org: pathogen detection and characterization through a web-based, open source informatics platform. BMC Bioinformatics, 2015, **16**(1): 1

[25] Haque M M, Bose T, Dutta A, *et al*. CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. Genomics, 2015, **106**(2): 116−121

[26] Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Medicine, 2015, **7**(1): 1−9

[27] Li Y, Wang H, Nie K, *et al*. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Scientific Reports, 2016, **6**: 23374

[28] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, **25**(14): 1754-1760

[29] Chen Y, Ye W, Zhang Y, *et al*. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Research, 2015, gkv784

[30] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool. Journal of Molecular Biology, 1990, **215**(3): 403-410

[31] Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, **10**(3): R25

[32] Kent W J. BLAT—the BLAST-like alignment tool. Genome Research, 2002, **12**(4): 656-664

[33] Lee W-P, Stromberg M P, Ward A, *et al*. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS One, 2014, **9**(3): e90581

[34] Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. BMC Bioinformatics, 2009, **10**(1): 1-9

[35] Ponstingl H, Ning Z. SMALT—a new mapper for DNA sequencing reads. F1000 Posters, 2010, 1:313.

[36] Zaharia M, Bolosky W J, Curtis K, *et al*. Faster and more accurate sequence alignment with SNAP. arXiv preprint arXiv:11115572, 2011

[37] Ye Y, Choi J H, Tang H. RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinformatics, 2011, **12**(1): 1

[38] Margulies M, Egholm M, Altman W E, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, **437**(7057): 376-380

[39] Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nature Methods, 2012, **9**(4): 357-359

[40] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, **25**(14): 1754-1760

[41] Bhatt A S, Manzo V E, Pedamallu C S, *et al*. Brief report: in search of a candidate pathogen for giant cell arteritis: sequencing-based characterization of the giant cell arteritis microbiome. Arthritis & Rheumatology, 2014, **66**(7): 1939-1944

[42] Hercus C. Novoalign. Selangor: Novocraft Technologies, 2012

[43] Zeitouni B, Boeva V, Janoueix-Lerosey I, *et al*. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics, 2010, **26**(15): 1895-1896

[44] Simpson J T, Wong K, Jackman S D, *et al*. ABySS: a parallel assembler for short read sequence data. Genome Research, 2009, **19**(6): 1117-1123

[45] Treangen T J, Sommer D D, Angly F E, *et al*. Next generation sequence assembly with AMOS. Current Protocols in Bioinformatics, 2011, **11**(S33): 11.18. 1-11.18. 18

[46] Morgulis A, Gertz E M, Schäffer A A, *et al*. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of Computational Biology, 2006, **13**(5): 1028-1040

[47] Jacob A, Lancaster J, Buhler J, *et al*. Mercury BLASTP: Accelerating protein sequence alignment. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 2008, **1**(2): 9

[48] 李定辰. 基于高通量测序平台的未知病原微生物检测系统[D]. 北京: 中国人民解放军军事医学科学院, 2016
Li D C. Unknown Pathogen Detection System Based on High-throughput Sequencing Platform [D]. Beijing: Academy of Military Medical Sciences, 2016

[49] Kurtz S, Phillippy A, Delcher A L, *et al*. Versatile and open software for comparing large genomes. Genome Biology, 2004, **5**(2): R12

[50] Zerbino D R, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research, 2008, **18**(5): 821-829

[51] Li R, Li Y, Kristiansen K, *et al*. SOAP: short oligonucleotide alignment program. Bioinformatics, 2008, **24**(5): 713-714

[52] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, **26**(6): 841-842

[53] Huson D H, Mitra S, Ruscheweyh H J, *et al*. Integrative analysis of environmental sequences using MEGAN4. Genome Research, 2011, **21**(9): 1552-1560

[54] Darling A C, Mau B, Blattner F R, *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 2004, **14**(7): 1394-1403

[55] Thorvaldsdóttir H, Robinson J T, Mesirov J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 2013, **14**(2): 178-192

[56] Krzywinski M, Schein J, Birol I, *et al*. Circos: an information aesthetic for comparative genomics. Genome Research, 2009, **19**(9): 1639-1645

[57] 叶福强. 胆总管结石患者胆道和阿尔茨海默症小鼠肠道的宏基因组学研究[D]. 北京: 中国人民解放军军事医学科学院, 2016
Ye F Q. Metagenomic Studies on The Biliary Microbiota of Patients With Choledocholithiasis and The Gut Microbiota of Mice With Alzheimer's Disease [D]. Beijing: Academy of Military Medical Sciences, 2016

[58] Wang H, Chan S-H, Cheung J, *et al*. MICA: A fast short-read aligner that takes full advantage of Intel Many Integrated Core Architecture (MIC). arXiv preprint arXiv, 2014: 14024876

# Computational Methods in Microbe Detection Using Next-Generation Sequencing*

ZHOU Zi-Han[1], PENG Shao-Liang[1]**, BO Xiao-Chen[2], LI Fei[2]**

([1] *College of Computer, National University of Defense Technology, Changsha* 410073 , China;
[2] *Institute of Radiation Medicine, Academy of Military Medical Sciences, Beijing* 100850, China)

**Abstract** Next-generation sequencing is changing research methods in biological fields. Microbial identification and detection technologies based on next-generation sequencing have advantage of high-precision and radial-velocity need， and the capability to replace previous culture-based and molecular methods, such as using nucleic acid amplification and hybridization technologies for rapid response to known and unknown biological threats. In this paper, we compared current computational analysis approaches on next-generation sequencing data for microbial identification and detection, including design principles, computational pipeline, and benchmark testing. Furthermore, some possible problems were summarized involving the use of these computational approaches.

**Key words** next-generation sequencing, microbe identification and detection, computational analysis approaches, benchmark

**DOI**: 10.16476/j.pibb.2016.0239

* This work was supported by grants from the National Natural Science Foundation of China (U1435222) and the Logistics Research Plan of Chinese PLA (BWS14C051).

**Corresponding author.
PENG Shao-Liang. Tel: 13574817196, E-mail: 13574817196@163.com
LI Fei. Tel: 86-10-66932251, E-mail: pittacus@gmail.com

Received: November 2, 2016 Accepted: December 14, 2016