微生物学报
Acta Microbiologica Sinica
2022, 62(3): 949–967
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20210418

Review 综 述

全基因组测序在病原菌分型与溯源中的应用研究进展

贾慧琼[1,2]，阮陟[3,4*]

1 浙江大学医学院附属第一医院检验科，浙江 杭州 310003
2 浙江省临床体外诊断技术研究重点实验室，浙江 杭州 310003
3 浙江大学医学院附属邵逸夫医院生物医学研究中心，浙江 杭州 310016
4 浙江省微生物技术与生物信息研究重点实验室，浙江 杭州 310016

贾慧琼, 阮陟. 全基因组测序在病原菌分型与溯源中的应用研究进展. 微生物学报, 2022, 62(3): 949–967.
Jia Huiqiong, Ruan Zhi. Advances on whole genome sequencing-powered typing and source tracking of bacterial pathogens. *Acta Microbiologica Sinica*, 2022, 62(3): 949–967.

摘　要：细菌分子分型已成为监测细菌感染性疾病的暴发流行与明确病原菌传播途径的重要工具。随着全基因组测序技术的日益兴起，公共数据库中已产生大量的细菌基因组数据，迫切需要研究人员充分认识和理解该技术，并掌握多种生物信息学工具挖掘并解读测序数据。本文系统概述了全基因组测序技术与生物信息学工具在病原菌分型与溯源中的应用，并对全基因组测序技术在临床诊疗实践中存在的挑战以及未来应用前景进行了探讨。

关键词：全基因组测序；生物信息学；基因组流行病学；分型与溯源；暴发流行

基金项目：国家自然科学基金(82102436，81401698)；浙江省自然科学基金(LY21H190001)；浙江省公益技术研究计划(LGF18H190001)
Supported by the National Natural Science Foundation of China (82102436, 81401698), by the Zhejiang Provincial Natural Science Foundation of China (LY21H190001) and by the Zhejiang Province Public Welfare Technology Application Research Project (LGF18H190001)
***Corresponding author.** Tel: +86-571-86006142; E-mail: r_z@zju.edu.cn
Received: 5 January 2021; **Revised:** 7 October 2021; **Published online:** 22 October 2021

Advances on whole genome sequencing-powered typing and source tracking of bacterial pathogens

JIA Huiqiong[1,2], RUAN Zhi[3,4*]

1 Department of Laboratory Medicine, the First Affiliated Hospital, Zhejiang University, Hangzhou 310003, Zhejiang, China
2 Key Laboratory of Clinical *In vitro* Diagnostic Techniques of Zhejiang Province, Hangzhou 310003, Zhejiang, China
3 Department of Biomedical Research Center, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou 310016, Zhejiang, China
4 Key Laboratory of Microbial Technology and Bioinformatics of Zhejiang Province, Hangzhou 310016, Zhejiang, China

Abstract: Molecular typing has been an important technique to monitor the outbreak of bacterial infections and the transmission routes of bacterial pathogens. Improvements in the next-generation sequencing technologies are facilitating rapid and cost-effective molecular diagnosis and genotyping in identification, characterization, and source tracking of bacterial pathogens. With the advancement of microbial whole genome sequencing techniques, a large volume of bacterial genome data have been produced and deposited in public databases, which necessitate the need of a variety of bioinformatics tools to analyze and interpret these data. This review provided an overview of the whole genome sequencing-powered typing and source tracking of bacterial pathogens by various cutting-edge bioinformatics tools. We also discussed the bottleneck in the deployment of this technology in clinical practice and the future application prospects in bacterial infectious disease management.

Keywords: whole genome sequencing; bioinformatics; genomic epidemiology; typing and source tracking; outbreak

细菌感染性疾病是公共卫生领域的一个重大安全问题，如何对其进行有效监测与防控是人类面临的一大难题。病原菌的鉴定、分型与溯源对及时发现细菌感染性疾病的暴发流行、准确查明传播媒介、遏制疾病进一步传播至关重要。病原菌分型与溯源中广泛使用的分子分型技术包括多位点串联重复序列分析(multiple-locus variable number of tandem repeat analysis，MLVA)[1]、多位点序列分析(multilocus sequence analysis，MLSA)[2]、多位点序列分型(multilocus sequence typing，MLST)[3]、脉冲场凝胶电泳(pulsed-field gel electrophoresis，PFGE)[4]，以及新兴的全基因组测序(whole genome sequencing，WGS)技术[5]等。1995 年，Fleischmann 等学者首次报道了完整的流感嗜血杆菌基因组序列[6]，但早期测序平台昂贵的成本和繁琐的数据分析流程限制了 WGS 技术的广泛应用。近年来，随着测序成本和时间的大幅降低以及数据分析工具的改进，降低了部署 WGS 平台的成本，改变了病原菌的传统检测与分型方法，WGS 有望成为追踪细菌感染性疾病传播的新金标准[7]。另一方面，NCBI 等公共数据库中已产生大量的病原菌基因组序列数据，迫切需要开发能够快速、准确地在基因组水平对病原菌进行分型与溯源的生物信息学工具。本文系统概述了 WGS 技术和生物信息学分析工具在病原菌分型与溯源领域的应用，并结合笔者课题组最新开发的病原菌基因组分型与溯源在线数据分析平台 BacWGSTdb 的应用实例，展望未来 WGS 技术在该领域的应用前景。

✉ actamicro@im.ac.cn, ☎ 010-64807516

1 病原菌基因组分型与溯源方法及生物信息学工具

1998 年, Maiden 等学者首次提出多位点序列分型的概念，它是一种基于多个管家基因序列片段的细菌分型方法，首先应用于脑膜炎奈瑟菌的分子分型，随后发展为一种广泛使用的细菌分子分型技术[8]。MLST 通过单基因测序比较 1 组(通常是 7 个)管家基因间的序列差异，根据菌株等位基因型对应的序列型(sequence type，ST)特征，对细菌进行分子分型。菌株的等位基因型与序列型数据可通过相应物种的 MLST 数据库进行检索和在线比对(https://pubmlst.org)。随着全基因组测序成本的降低, WGS 与 7 个管家基因双向测序的费用已相差无几,但 WGS 携带的信息量则远超 7 个管家基因序列。基于 WGS 的 MLST 分型，我们称之为基因组 MLST 分型(*in silico* MLST)[9]。虽然目前 PubMLST 数据库中已存在超过 100 种细菌拥有 MLST 分型方案，但该技术也存在不能充分区分高重组水平、低多样性和单克隆菌株之间的差异等分辨率不足的缺点。随着生物信息学和测序技术的发展，各类分型方法层出不穷，已有多种更高分辨率的基于病原菌基因组序列的分型与溯源策略，如核心基因组多位点序列分型(core genome multilocus sequence typing，cgMLST)、全基因组多位点序列分型(whole genome multilocus sequence typing，wgMLST)和核心基因组单核苷酸多态性(core genome single nucleotide polymorphism，cgSNP)等方法。对疑似引起细菌感染性疾病暴发流行的菌株(有共同来源或流行病学联系)进行全基因组测序，可直接获得病原菌鉴定、表型与基因型预测、质粒分型和流行病学溯源等方面的信息。使用生物信息学工具分析菌株间的亲缘关系，研究菌株间的传播链，追溯感染性疾病暴发的源头。上述结合基因组学与流行病学，分析病原体发生、发展过程的研究被称为基因组流行病学(genomic epidemiology)研究[10]。表 1 总结了目前常用的细菌基因组分型与溯源工具，包括软件功能、分析平台类型、输入文件格式、输出文件内容以及下载地址。图 1 显示了病原菌基因组分型与溯源的常规数据分析流程。

表 1 病原菌基因组分型与溯源工具列表

Table 1 List of bioinformatics tools used for bacterial genome sequence typing and source tracking

Name	Description	Accessibility	Link
Gene-by-gene approaches			
chewBBACA[11]	cg/wgMLST schema creation and allele calling using genome assemblies (FASTA)	Web & Standalone	https://chewbbaca.online https://github.com/B-UMMI/chewBBACA
MentaLiST[12]	cg/wgMLST schema creation and allele calling using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Standalone	https://github.com/WGS-TB/MentaLiST
Genome profiler[13]	cg/wgMLST schema creation and allele calling using genome assemblies (FASTA)	Standalone	http://sourceforge.net/projects/genomeprofiler

(待续)

(续表 1)

In silico typing			
SeqSero[14]	*Salmonella* serotype prediction using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Web & Standalone	http://denglab.info/SeqSero https://github.com/denglab/SeqSero
SeqSero2[15]	Similar to SeqSero (an updated version)	Web & Standalone	https://github.com/denglab/SeqSero2 http://www.denglab.info/SeqSero2
MOST[16]	*Salmonella* serotype prediction and MLST allele calling using raw sequencing reads (FASTQ)	Standalone	https://github.com/phe-bioinformatics/MOST
SalmonellaTypeFinder	*Salmonella* serotype prediction and MLST allele calling using raw sequencing reads (FASTQ)	Web & Standalone	https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder
SISTR[17]	*Salmonella* serotype prediction and cgMLST allele calling using genome assemblies (FASTA)	Web & Standalone	https://lfz.corefacility.ca/sistr-app https://github.com/phac-nml/SISTR_cmd
SerotypeFinder[18]	*E. coli* serotype prediction using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Web & Standalone	https://cge.cbs.dtu.dk/services/SerotypeFinder
Kaptive[19–20]	Capsule and lipopolysaccharide serotype prediction for *Klebsiella pneumoniae* and *Acinetobacter baumannii* genome assemblies (FASTA)	Web & Standalone	http://kaptive.holtlab.net https://github.com/katholt/Kaptive
ClermonTyping[21]	*In silico Escherichia* genus strain phylotyping using genome assemblies (FASTA)	Web & Standalone	http://clermontyping.iame-research.center https://github.com/A-BN/ClermonTyping
pMLST[22]	*In silico* plasmid multilocus sequence typing using genome assemblies (FASTA)	Web & Standalone	https://pubmlst.org/organisms/plasmid-mlst
PlasmidFinder[22]	*In silico* characterization of plasmid replicons from *Enterobacteriaceae* using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Web & Standalone	https://cge.cbs.dtu.dk/services/PlasmidFinder
PLACNETw[23]	A graph-based tool for plasmid reconstruction using raw sequencing reads (FASTQ)	Web & Standalone	https://castillo.dicom.unican.es/upload https://github.com/LuisVielva/PLACNETw
oriTfinder[24]	Identification of bacterial mobile genetic elements using genome assemblies (FASTA) or GenBank files	Web & Standalone	https://bioinfo-mml.sjtu.edu.cn/oriTfinder
Phylogenetic inference			
Gubbins[25–26]	Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using a multi-FASTA alignment file	Standalone	https://github.com/sanger-pathogens/gubbins

(待续)

(续表 1)

RecHMM[27]	Detection of recombinant breakpoints and tree-topology changes in whole-genome sequence alignments using a multi-FASTA alignment file	Standalone	http://biowiki.org/wiki/index.php/Rec_HMM
PhyML[28]	Fast maximum likelihood-based phylogenetic inference using a PHYLIP format alignment file	Web & Standalone	http://www.atgc-montpellier.fr/phyml https://github.com/stephaneguindon/phyml
RAxML[29]	Phylogenetic analyses of large datasets under maximum likelihood using a multi-FASTA alignment file	Standalone	https://github.com/stamatak/standard-RAxML
IQ-TREE[30]	A fast and effective algorithm to infer phylogenetic trees by maximum likelihood using an alignment file in various common formats	Web & Standalone	http://iqtree.cibiv.univie.ac.at http://www.iqtree.org
MrBayes[31]	A variety of phylogenetic and evolutionary models based on Bayesian inference using a NEXUS format alignment file	Standalone	http://nbisweden.github.io/MrBayes
BEAST[32]	A cross-platform program for time-measured Bayesian phylogenetic analysis and phylodynamic data integration using an alignment file in various common formats	Standalone	https://beast.community
Snippy[33]	A pipeline for rapid variant calling and core genome alignment using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Standalone	https://github.com/tseemann/snippy
Lyve-SET[34]	A high-quality SNP pipeline to create a phylogeny using raw sequencing reads (FASTQ)	Standalone	https://github.com/lskatz/lyve-SET
SNVPhyl[35]	A pipeline for identifying high-quality SNPs and constructing a phylogenetic tree using raw sequencing reads (FASTQ)	Standalone	https://snvphyl.readthedocs.io
CSI Phylogeny[36]	Identifying variant and inferring phylogenies based on the concatenated alignment of the high-quality SNPs using raw sequencing reads (FASTQ) or genome assemblies (FASTA)	Web	https://cge.cbs.dtu.dk/services/CSIPhylogeny
BacWGSTdb[37–38]	A one-stop platform for bacterial whole-genome sequence typing and source tracking using genome assemblies (FASTA)	Web	http://bacdb.cn/BacWGSTdb

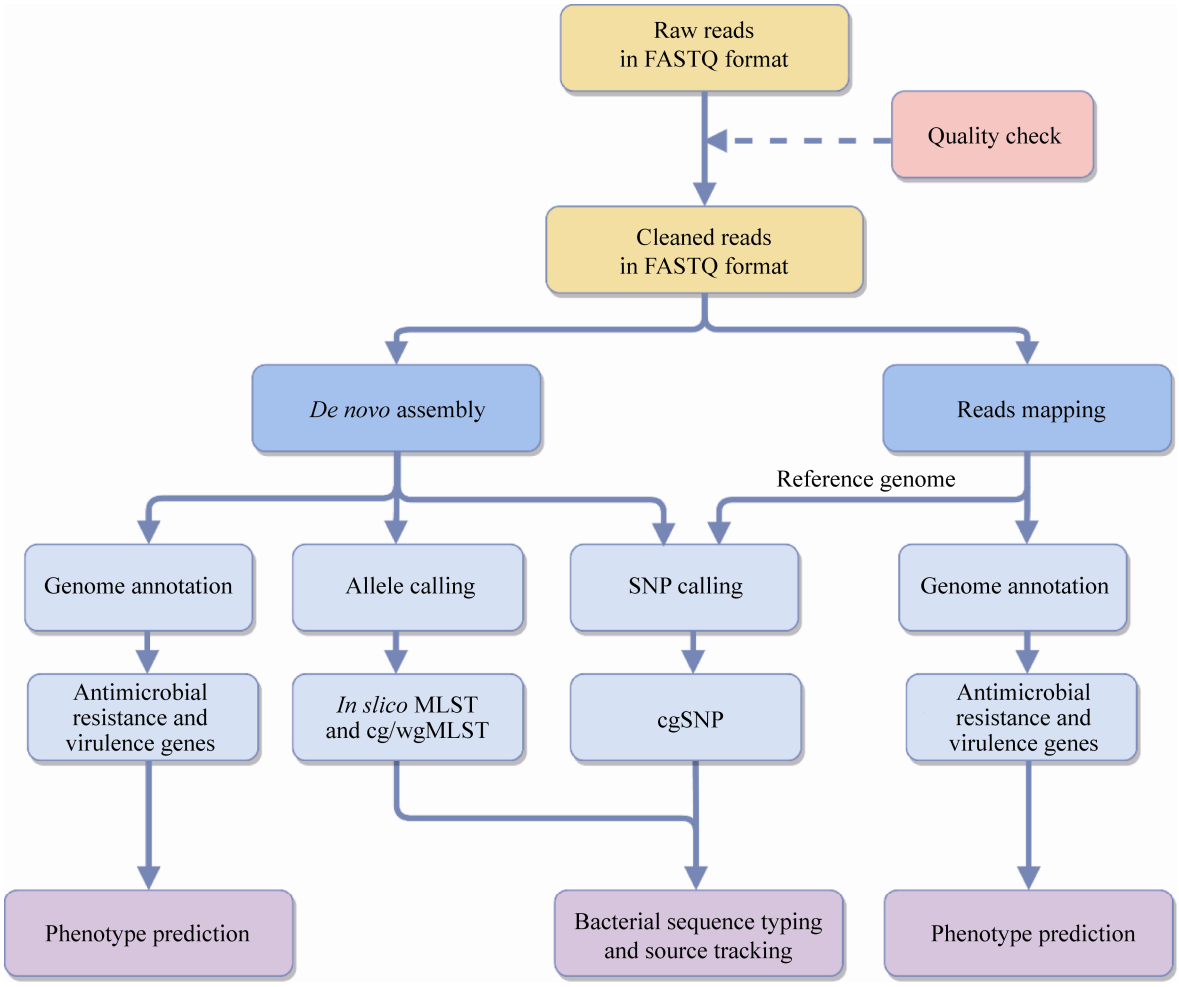

图 1　常用病原菌基因组分型与溯源数据分析流程

Figure 1　Procedures for bacterial genome sequence typing and source tracking analysis.

1.1　基于等位基因序列的病原菌分型与溯源

核心基因组多位点序列分型(cgMLST)是传统 MLST 分型方案的扩展形式，单一 cgMLST 策略仅针对特定物种。将目的菌株的全基因组序列与其 cgMLST 分型方案进行比对，提供分型及亲缘关系分析，以实现对病原菌传播的监测。全基因组多位点序列分型(wgMLST)通常可用作 cgMLST 的扩展，该策略囊括了细菌的核心基因组(core genome)及附属基因组(accessory genome)。由于附属基因组包含菌株携带的耐药基因、毒力基因、可移动元件、菌种内高度变异的基因等，故对于亲缘关系接近的菌株，基于更多等位基因的 wgMLST 可以提供比 cgMLST 更佳的分辨率。但也有研究表明，基于 wgMLST 和 cgMLST 的分型结果十分相似，两者的分辨能力无统计学差异[39–40]。利用 wgMLST 策略分析菌株系统发育关系时，可能会受到同源重组、可移动元件的水平转移等遗传事件的影响，导致部分区段出现高密度单核苷酸多态性等异常情况，从而干扰真实的系统发育关系。总之，cgMLST 策略更常用于分析菌株间亲缘关系和分型。

目前，用户可运用 Genome profiler[13]、MentaLiST[12]和 chewBBACA[11]等开源工具构建本地使用的物种特异的 cgMLST 方案并进行分型。也可以使用商业软件 Ridom SeqSphere+(http://www.cgmlst.org/ncs)提供的部分物种 cgMLST 分型方案(目前已有 21 个物种)，该软件支持用户自行定制 cgMLST 分型方案进行数据分析。在一项肺炎克雷伯菌不同分型方法的比较研究中，研究者利用 PFGE 和 Ridom SeqSphere+平台提供的肺炎克雷伯菌 cgMLST 方案对 40 株肺炎克雷伯菌进行分型。cgMLST 分析结果表明，同一 PFGE 类型的 31 株属于同一 PFGE 型的肺炎克雷伯菌与其他 3 株分属不同 PFGE 型的分离株形成了一个大簇(最大等位基因差异数<147)，PFGE 与 cgMLST 分型结果具有较高的一致性，并且 cgMLST 能更细致区分同一 PFGE 类型的菌株[41]。

利用细菌基因组序列筛选出数个菌株特异性基因(strain-specific gene)，并构建相应的 PCR 反应体系，可对某些多重耐药或高毒力克隆进行快速鉴定，从而有效节约了检测时间与成本。目前已有国内外学者借助该技术成功研发出多重 PCR 反应体系对包括多重耐药鲍曼不动杆菌高毒力克隆 ST10[42]、甲氧西林耐药金黄色葡萄球菌克隆复合体 CC398[43]、碳青霉烯耐药肺炎克雷伯菌全球流行克隆 ST258/ST11 与数个高毒力克隆等在内的多种临床常见病原菌的优势克隆进行快速诊断[44]。建立上述检测方法，主要有 3 个步骤：(1) 将不同型别的菌株基因组序列进行比对分析，筛选出所关注的菌株携带的特异性基因；(2) 基于各特异性基因序列设计 PCR 引物，并借助 NCBI Primer BLAST 等工具初步验证其特异性；(3) 应用多种临床菌株与标准菌株验证其对真实样本的检测效能[45]。该技术操作简便、价格低廉且特异性高，有助于对临床重要病原菌院内播散的流行病学监测[42−43]。

基于全基因组测序数据还可预测部分病原菌的血清学分型。沙门菌与大肠埃希菌系导致人和动物腹泻的常见食源菌，血清型是其致病性研究中重要的分类依据。已有工具将沙门菌全基因组测序读段或组装后的基因组序列与 O 和 H 抗原等位基因参考数据库进行比对，从而预测其血清型，如 SeqSero[14](后续版本 SeqSero2[15])。英国公共卫生部开发的 MOST 工具，利用 SRST 工具分析全基因组测序数据，并结合 MLST 分型结果，以识别相应的血清型[16]。丹麦科技大学开发的网页版工具 SalmonellaTypeFinder，运行 SeqSero 预测血清型并应用 SRST2 确定 MLST 分型，最终结合两种分型工具的结果预测沙门菌的血清型。SISTR 则是结合 cgMLST 分型策略预测沙门菌的血清型，其准确率可达 94.6%[17]。大肠埃希菌的菌体抗原(O)、荚膜抗原(K)、鞭毛抗原(H)和菌毛抗原(F)是其血清型鉴定和分型的基础[57]。2015 年，Joensen 等学者创建基于大肠埃希菌全基因组数据预测其血清型分型的在线分析工具 SerotypeFinder，该工具基于特定 O 抗原基因(*wzx*、*wzy*、*wzm* 和 *wzt*)和 H 抗原基因(*fliC*、*flkA*、*fllA*、*flmA* 和 *flnA*)序列的分型方法与传统血清学结果具有高度一致性[18]。此外，在线工具 ClermonTyping 可将大肠埃希菌细分为 A、B1、B2、C、D、E 和 F 七个主要的亚群(phylogroups)，从而替代传统的多重 PCR 方法对大肠埃希菌进行亚种分型[21]。Kaptive Web 是一种快速鉴别肺炎克雷伯菌 K (编码荚膜多糖)和 O 位点(编码脂多糖 O 抗原)的在线工具。用户上传组装后的基因组序列，可在线浏览或下载表格形式的分析结果[19]。2020 年最新版本的 Kaptive Web 增加了对鲍曼不动杆菌荚膜多糖(capsular polysaccharide，KL)和脂寡糖外核心

(lipooligosaccharide outer core，OCL)的识别。该方案基于 92 个已知编码与运输荚膜多糖的蛋白质基因簇(K 位点)和 12 个外核生物合成的基因簇(OC 位点)，对上传的序列分配唯一的 KL 或 OCL 编号[20]。

质粒是一种细菌染色体外的线性或环状 DNA 分子，也是细菌基因组 DNA 的重要组成部分，常携带有益于宿主细菌生存的基因(如耐药基因与毒力基因等)，传播可移动遗传元件(插入序列和转座子)，可引起其与染色体的同源或非同源重组，提高细菌基因组的可塑性[58]。根据质粒传播的特性，可分为接合质粒(conjugative plasmid)、可转移质粒(mobilizable plasmid)和不可转移质粒(non-mobilizable plasmid)[59]。基于 DNA 杂交或 PCR 的质粒复制子分型，根据质粒的不相容性(incompatibility，Inc)，分为 FIA、FIB、FIC、HI1、HI2、IncA、IncC、IncN、IncI、IncP 与 IncX 等不相容群(incompatibility group)[60]。PlasmidFinder 可检测质粒复制子基因并匹配对应 Inc 分型，适用于肠杆菌科细菌和少部分革兰阳性菌的质粒分型。pMLST 可提供质粒多位点序列分型的在线分析[22]。PLACNETw 可直接利用全基因组测序读段重建质粒，并与数据库中的质粒序列进行比对，为用户提供了交互式图形界面[23]。oriTfinder 是一种可快速识别质粒携带的 oriT、松弛酶、T4CP 和 T4SS 等接合转移元件的在线工具，通过分析用户上传的细菌全基因组序列，将 oriT 和 T4SS 等预测结果进行可视化显示[24]。PLSDB 是一个质粒序列二级数据库，该数据库包含从 NCBI GenBank 数据库中检索到的所有细菌质粒序列，通过与用户上传序列进行 Mash 或 BLASTn 比对，反馈该数据库中亲缘关系最近的质粒信息[61]。

1.2 基于单核苷酸多态性的病原菌分型与溯源

在获得细菌基因组序列的基础上，使用物种间具有同源性的基因片段构建系统发育关系，从而解析细菌的进化历史的过程称之为系统发育重建。在细菌基因组中，重组事件是细菌种群进化过程中重要的进化推动力之一，通常由同源性较高的序列介导，被称为同源重组(homologous recombination)事件。另一类发生在非同源片段之间的重组通常会引入新的基因或基因组岛的插入，被称为基因水平转移(horizontal gene transfer)[46]。同源重组事件对于细菌的种群生存和进化具有重要意义，但同时也会打乱细菌种系间的垂直遗传关系，影响系统发育分析的准确性。因此，在病原菌的分型与溯源研究中，首先要去除同源重组位点的影响，可使用 Gubbins[25–26]与 RecHMM[27]鉴定并去除同源重组事件引起的基因组变异。

单核苷酸多态性在系统发育关系分析中具有高分辨率、易鉴定等优势，其中核心基因组单核苷酸多态性(cgSNP)分析是一种基于核心基因组 SNP 构建系统发育关系的方法。主要有 2 种分析模式：一种是序列组装后比对分析(assembly-based)，另外一种是直接利用读段映射分析(read-based)。序列组装后比对分析是将从头组装的序列与参考基因组进行比对分析，确定差异 SNP 位点；而读段映射分析是略过基因组组装的过程，使用映射比对工具(如 Bowtie2 或 BWA)直接将测序读段与参考菌株核苷酸进行比对，由多个读段重叠区域的共有核苷酸推断出该位置的碱基。

核心基因组单核苷酸多态性分析适用于研究亲缘关系较近的同一谱系病原菌的系统发育关系。在研究多样化或高度重组的病原菌之间的亲缘关系时，选择合适的参考基因组显得格

外重要，这也是 cgSNP 策略在实际应用中的挑战之一。基于无参考基因组的比对方法虽然消除了潜在的偏倚，并且可以检测到参考序列中不存在的 SNP 位点。但在没有合理阈值的情况下，无参考基因组的比对方法可能会导致 SNP 识别中的错误率更高(即假阳性)[47]。目前，有多种分析工具可进行 cgSNP 分析，例如：CSI Phylogeny 可提供在线工具，用户只需上传菌株测序数据(FASTA 或 FASTQ 格式)和参考基因组序列(FASTA 格式)即可[36]。另外，笔者课题组开发的细菌基因组分型与溯源在线数据分析平台 BacWGSTdb 提供 cgSNP 和 cgMLST 策略分析菌株间亲缘关系，用户只需上传单个或多个菌株的基因组序列(FASTA 格式)，系统将在 3–5 min 内反馈数据库中与该菌株亲缘关系最接近的若干菌株信息，并生成上传菌株与数据库现有菌株的系统发育树(图 2)，从而实现病原菌的快速分型与溯源[37–38]。Snippy 工具可实现从去除重组位点到生成 FASTA 格式的 cgSNP 序列文件的一系列分析[33]，Lyve-SET[34]和 SNVPhyl[35]等开源工具也可提供基于 cgSNP 序列的菌株间系统发育关系分析[48]。

系统发育树还可以用于重建祖先序列和估计分歧时间，常用软件包括 Lyve-SET[34]、SNVPhyl[35]、PhyML[28]、RAxML[29]、IQ-TREE[30]、BEAST[32]和 MrBayes[31]等工具。研究者通过分析系统发育树各分支的拓扑结构特征，并关联菌株耐药表型、分离时间地点、标本类型等信息，明确不同菌株之间的亲缘关系并推测其传播路径与进化历史。目前在系统发育学分析中最常用的算法均基于似然率计算，主要分为两大类：最大似然法(maximum likelihood，ML)与贝叶斯推断法(bayesian inference，BI)。基于最大似然法的系统发育关

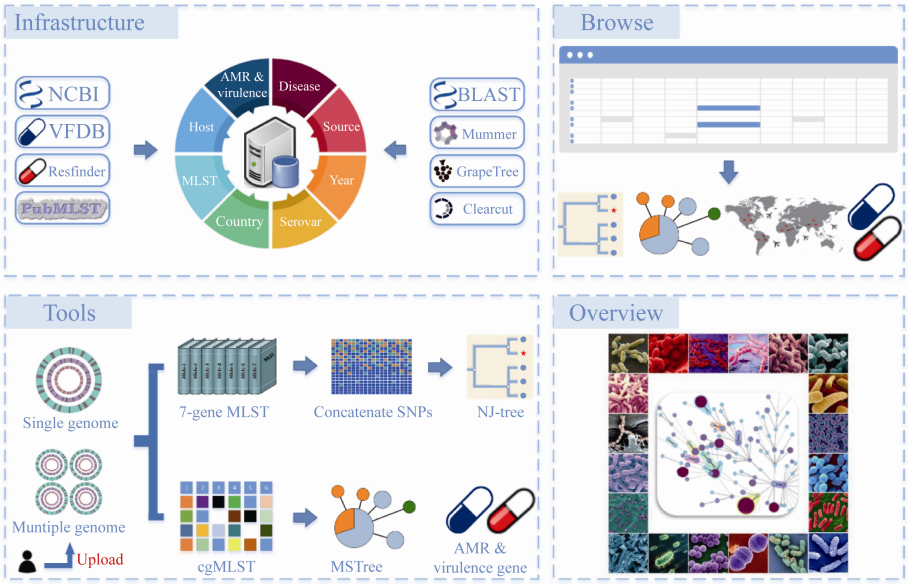

图 2　细菌基因组分型与溯源在线数据分析平台 BacWGSTdb 功能模块示意图[38]

Figure 2　Overview of the content and function modules of BacWGSTdb. ‘Infrastructure’ lists the public database and tools integrated in BacWGSTdb. ‘Browse’ functions to visualize and compare the genetic relationships among isolates deposited in BacWGSTdb. ‘Tools’ functions for whole genome sequence typing and source tracking based on user uploaded sequence (s). ‘Species’ represents 20 bacterial species currently supported by BacWGSTdb.

系分析工具 FastTree[49]、PhyML[28]、RAxML[29]和 IQ-TREE[30]可提供多种进化模型的选择以及不同的程序参数。基于贝叶斯推断法的 MrBayes[31]和 BEAST[32]使用马尔可夫链蒙特卡罗(Markov chain Monte Carlo，MCMC)统计学方法来推断系统发育关系。相比于最大似然法，贝叶斯推断法耗时长、对计算机资源要求高，但可以提供大量模型选择，重建更准确的进化历程。研究表明，多重耐药鲍曼不动杆菌主要存在 2 个全球流行的克隆 GC1 和 GC2，Holt 等学者利用 BEAST 构建包含时间维度的系统发育树，揭示了全球流行的多重耐药鲍曼不动杆菌优势克隆 GC1 在过去 50 年间的基因组微进化和时空传播动力学特征：即鲍曼不动杆菌基因组的重组事件并非随机分布，这些重组事件在编码荚膜多糖、外脂寡糖和外膜蛋白 CarO 等基因中呈现显著的多样性。在过去的 65 年中，大量的重组事件导致多重耐药鲍曼不动杆菌 GC1 积累了对新一代抗菌药物耐药性，包括氟喹诺酮类、第三代头孢菌素类与碳青霉烯类药物等，对其全球播散发挥了促进作用[50]。

上述工具的最终输出文件中通常包含菌株间亲缘关系的遗传距离信息，为了直观展示菌株间的相关性，并提取有效信息，我们需要借助可视化工具整合系统发育树与菌株流行病学信息。GenGIS 2 是一种结合系统发育树与地理空间等流行病学信息的可视化工具，可提供 2D 和 3D 地形的界面[51]；Microreact 提供了在线可交互的可视化工具，用户通过上传系统发育树和菌株对应的地理位置和流行病学信息文件，可直观显示系统发育树及菌株在全球的分布情况[52]；FigTree 软件以及在线工具 iTOL 均可绘制系统发育树，前者能够读取 BEAST 软件输出的结果文件，显示时间尺度树，后者能美化系统发育树并添加各类注释[53]；panX 是一种可视化泛基因组分析的网页类型工具，可查看分析结果中相互关联的泛基因组统计图、基因簇表、比对结果、系统发育关系以及原始数据表[54]；Phandango 是一个网页版可视化工具，适用于可视化系统发育树以及相关的基因组信息和原始数据[55]；PATRIC 提供了一个集成、储存和可视化细菌基因组数据的在线工具[56]。

2 全基因组测序在病原菌分型与溯源中的应用

2.1 分型与溯源的方法比较与应用实例

由于不同菌种管家基因序列的差异较大，单一 MLST 方案通常仅适合种属间同源性较高的菌种，而同一属的不同菌种往往存在多套 MLST 方案。仅利用 MLST 方案进行菌株分型或溯源可能会隐藏大部分菌株的多样性，从而得出不准确的系统发育关系，例如：MLST 不能进一步区分脑膜炎奈瑟氏球菌 ST11 克隆复合体中的各亚群[62]。由于部分医院获得性感染相关病原菌存在高度的克隆性(例如：碳青霉烯耐药鲍曼不动杆菌)，基于 7 个管家基因的传统 MLST 分型方案难以准确区分相同 ST 型病原菌之间的细微差异。随着全基因组测序技术和生物信息学分析手段的不断进步以及测序时间与成本的不断降低，极大提升了细菌基因组测序数据的准确性与可用性，为有效监测临床多重耐药菌的传播提供了新思路[47]。Venditti 等[63]在一项对 2016 年 12 月至 2017 年 4 月间在意大利罗马的一家医院 ICU 收集的 13 株碳青霉烯耐药鲍曼不动杆菌的溯源中，利用 DiversiLab 系统进行 rep-PCR[64](repetitive extragenic palindromic PCR)分析，基于 WGS 数据进行 *in silico* MLST、cgMLST 分型和耐药基因的识别等分析。结果

表明，rep-PCR 和 MLST 将分离株聚类为同一簇，而 cgMLST 分型方案能将菌株进一步细分为 2 个簇，它们之间仅相差 15 个等位基因。囊括细菌全部核心基因组基因的 cgMLST 分型策略，能进一步分辨传统 MLST 无法区分的克隆复合体,从而能更准确的追溯病原菌的传播路径。一项研究旨在比较 PFGE、cgMLST 和 cgSNP 等 3 种常用的肺炎克雷伯菌分型方法：研究者收集 2017 年意大利米兰分离出的 80 株碳青霉烯耐药肺炎克雷伯菌，将菌株不同分型方法的结果与流行病学证据进行比较，发现 cgMLST 和 cgSNP 的分辨率均高于 PFGE，并且两种方法都适用于肺炎克雷伯菌的传播溯源。在肺炎克雷伯菌克隆复合体 CC258 的系统发育关系重建中，cgSNP 的分辨率略高于 cgMLST[65]。另一项研究在对 2012–2015 年间德国海德堡大学附属医院分离的 39 株碳青霉烯耐药鲍曼不动杆菌的回顾性分析时，发现大部分菌株为 ST2 型，少部分罕见型别的菌株来源于有海外旅居史的病人。cgSNP 分析结果揭示了 4 个潜在的暴发流行克隆，而这些群体中的 2 簇与菌株的流行病学记录高度吻合，表明当年很可能发生了未被捕捉到的院感暴发流行事件[66]。更高分辨率的 WGS 技术可优化菌株暴发流行的监测方案，为及时采取适当的防控措施提供科学依据。作为 MLST 的扩展形式，cgMLST 除了具有更高的分辨率和不依赖参考基因组的设定等优势之外，研究者还可将新分离的菌株基因组序列与历史数据进行整合与回顾性分析，从而实现对病原菌的连续监测。然而，目前仅有少数病原菌存在已发表或公认的 cgMLST 方案，而自定义方案则要求研究者具备一定程度的生物信息学基础，并且在构建方案时所纳入的基因组序列数量、质量及比对参数设定也会影响最终 cgMLST 分型方案的设计，从而影响分型与溯源的最终结果。基于单核苷酸多态性的 cgSNP 策略分辨率高且不需要提前预设分型方案，能最大程度反映菌株间的亲缘关系，但参考基因组及系统发育树构建算法的选择也可能会影响溯源结果的准确性。因此，研究者需要结合不同的应用场景，选择合适的分析策略。

此外，WGS 在食源性病原菌的分型与溯源中，也能提供较高的分辨率和准确性。一项研究在对 1950–2015 年分离的 290 株鸡源沙门菌进行同源性分析时，比较了沙门菌常规血清学检测结果与基于 WGS 预测的血清型,结果发现二者的一致率高达 97.6%。与常规的血清学分型方法相比，WGS 分型更加快速和准确，尤其针对需要不同培养基和血清学鉴定的罕见血清型。随着测序技术的发展和参考基因组数据库的不断完善,基于 WGS 数据的血清型预测有望成为今后沙门菌血清型鉴定的新金标准[67]。2018 年 7 月，中国深圳发生了 10 例肠炎沙门菌暴发感染(outbreak)事件，深圳疾病预防控制中心利用全基因组测序将此次暴发事件的源头追溯到通过在线平台订购的食物。基于 SNP 的系统发育树显示,9 株同一 PFGE 分型的肠炎沙门菌亲缘关系十分接近(SNP<1),而这些菌株分别来自该暴发事件中的 5 只鸡腿和 4 例患者，结合流行病学证据链将鸡腿确定为引起暴发的源头。而同时期检出的 10 株肠炎沙门菌与暴发事件中的菌株最小距离为 59 SNPs，大于鉴定为同一暴发事件的阈值(≤3 SNPs),提示这些散发病例不属于本次暴发事件[68]。总之，WGS 不仅可以追溯食源性病原菌的暴发事件，还可以排除非暴发事件的菌株，为及时干预并有效遏制感染性疾病的传播提供有力的科学依据。

WGS 在一些自然疫源性病原菌(如鼠疫耶尔森菌)的分型与溯源上，也得到了广泛的应

用。我国军事医学科学院杨瑞馥研究员课题组曾对 17 株鼠疫菌代表菌株进行全基因组测序，利用比较基因组学方法，鉴定出 933 个 SNP，并结合质谱技术分析 286 株全球分离的鼠疫菌在这些位点的变异情况[69]。该研究进一步验证了 2004 年 Achtman 等学者[70]提出的鼠疫菌进化框架的准确性，并发现了更多的鼠疫菌种群。更为重要的是，该研究发现所有分离自美国的鼠疫菌都起源于一次传播事件，该事件可追溯至 19 世纪末，一艘从中国香港开出，经停夏威夷，最终抵达旧金山的商船。感染鼠疫菌的宿主可能随这艘商船来到美国，并在当地生存与传播。在另一项研究中，该课题组对 118 株鼠疫菌进行全基因组测序，结合已公开的 15 株鼠疫菌全基因组序列，重建基于 SNP 序列的鼠疫菌的种群结构与系统发育关系，进而探索鼠疫菌进化的动力[71]。通过系统发育树的拓扑结构，研究者鉴定出鼠疫菌的 2 个全新分支，与以前的 3 个分支共同构成 5 大种系分支的种群结构。该研究将鼠疫菌系统发育关系与其所在地区的地理环境因素进行关联分析后，发现中国古代重要商贸路线(丝绸之路、茶马古道和唐蕃古道等)与鼠疫菌的地理分布存在惊人的一致性，从而提示历史上 3 次鼠疫大流行可能起源于中国或者其邻近地区，说明历史上人类商贸活动在鼠疫传播中发挥了重要作用[71]。在该课题组最新的一项鼠疫溯源研究中，研究者将 2 株苏拉特鼠疫菌基因组与 366 株鼠疫代表菌株的基因组序列进行系统发育分析，确定苏拉特菌株的遗传发育位置，为 1994 年印度苏拉特鼠疫暴发疫情的源头解读提供新的线索[72]。在重要公共卫生事件中，全基因组测序技术可提供快速的分型防控与溯源，从而保障人民群众的健康安全。对于一些颇具争议的疫情溯源，全基因组测序的应用、菌株量的增加以及种群的全方位调查，结合流行病学证据，将有助于追溯疫情的起源。

笔者课题组曾提出了一个和常规基因组流行病学不一样的观点，即反向基因组流行病学(reverse genomic epidemiology)研究策略[47]。借助细菌基因组大数据，挖掘其适应性进化过程中在基因组中留下的印迹，从而反向建立菌株间的流行病学联系并揭示其进化历程与传播规律。该策略认为当公共数据库中已存在足够数量细菌基因组数据的情况下，如果发现某些分离株的基因组序列足够相似时，则认为这些菌株可能来源于同一祖先。借助前文提到的 BacWGSTdb 在线数据分析平台，该研究分析了来自公共数据库的 20 种病原菌所有已公开的基因组序列相似性，发现来源于不同国家的部分菌株亲缘关系却十分接近。通过对跨越 5 个不同国家的病原菌传播网络进行梳理后发现，其中有 3 个为目前国际上已公认的菌株暴发流行事件。例如：2011 年德国 *E. coli* O104:H4 葫芦巴豆种子食源性致病菌污染事件[73−74]、2010 年海地震后 *V. cholera* 暴发事件[75−76]及 21 世纪初发生在非洲的 *S.* Typhi 食源性传播事件[77]。上述新颖观点为 WGS 在病原菌传播与溯源领域的应用提供了全新视角，但由于目前 NCBI 等公共数据中可供获取的病原菌基因组序列数量仍极为有限，因此在实际溯源过程中可能会遇到用户上传的菌株基因组序列与数据库中的菌株差异过大或者部分菌株缺失流行病学资料，从而难以进行准确溯源等问题。

2.2 基因组分型与溯源研究存在的挑战

尽管已有众多案例表明 WGS 在病原菌分型与溯源中的优势，但在实际应用中依旧面临着众多挑战。例如，目前尚缺乏国际通用的细菌基因组数据采集与分析的标准化流程，尤其是学界难以统一关于暴发克隆的判断标准(阈

值),即菌株间相差多少 SNP 才可被视为同一克隆并归属于同一暴发事件。笔者课题组在一项对 NCBI 数据库中的所有 ST195 型鲍曼不动杆菌的比较基因组学研究中,对 2850 株鲍曼不动杆菌进行 cgMLST 分型,并构建最小生成树与贝叶斯系统发育树,以预测 ST195 型鲍曼不动杆菌的进化起源与分化时间。系统发育学研究结果显示,菌株间 SNP 数量分布于 0–3,等位基因差异数分布于 0–14。一些分离地点相距甚远的菌株亲缘关系却十分接近(<8 个等位基因或<20 个 SNP),甚至没有超过本国分离菌株,也低于 Higgins 等学者提出的同一克隆暴发流行的参考阈值,存在跨国传播的可能性较大[78]。由于便捷的全球旅行使医院获得性和食源性细菌病原体能轻易越过地理障碍,无症状携带者也会造成直接流行病学证据链的缺失,所以制定准确且通用的病原菌暴发流行克隆的判断标准至关重要。理想状态是存在一个简单阈值可以鉴别所有类型的病原菌暴发事件,但在实际应用中往往受到时间、空间、流行病学特征以及 WGS 数据分析方法的限制,不同菌种甚至同种病原菌的不同血清型之间也存在着差异[79]。例如,有研究表明沙门菌中的肠炎沙门菌通常呈高度克隆传播趋势,而鼠伤寒沙门菌则不然[80]。因此,笔者认为基于 WGS 技术的病原菌分型与溯源结果仍需结合具体菌株分离时间与空间、基因组学与遗传进化特征等参数来解释,尤其是局部暴发事件的菌株(短时间)与全球范围的分离株(长时间)相比,一些刚刚累积的突变可能尚未经过纯化选择的作用而得以消除,使得正在经历适应性进化的谱系比那些已经历过纯化选择的谱系能更快地积累突变,从而可能会影响病原菌分型与溯源结果的解读[78]。此外,细菌的生存环境也可施加选择压力(如:医院内的抗生素选择压力),从而影响其突变率和传代周期,故仅依据某个特定的阈值来鉴别病原菌暴发事件可能难以适应复杂多变的溯源应用场景。又如:在一项肠炎沙门菌的传播溯源研究中,Payne 等学者就提出应设立动态 SNP 阈值来鉴定暴发事件,即对 4 周内的疑似暴发事件应以 0 SNP 为阈值,从而提高鉴别的灵敏度和特异性,而对超过 4 周的事件应使用动态阈值(如 0–5 SNPs)[81]。基于上述分析,笔者认为不同的分析策略、人员取样的偏差以及流行病学信息的完善度都会影响病原菌暴发事件的溯源结果,准确鉴别暴发事件还需要研究人员因地制宜和因“菌”而异。

3 结语与展望

在过去的数十年中,全基因组测序技术取得了长足的发展,曾经耗时耗力的测序项目现今可在较短的时间内完成。随着高通量测序技术的迅猛发展与测序成本的不断下降,公共数据库中已存储了海量的病原菌基因组序列,如何将其进行有效整合并应用于病原菌分型与溯源的实际研究工作中是该领域的研究人员需要深入思考的问题。受限于数据储存以及数据分析策略的复杂性,大样本数据分析和复杂的数据运算仍需依赖 Linux 操作系统,从而要求从事本领域的研究者掌握一定的生物信息学技能。与此同时,简便易用的在线数据分析与可视化工具也在不断问世,如笔者课题组开发的 BacWGSTdb 细菌基因组分型与溯源一站式在线数据分析平台,可对常见 20 种病原菌的基因组序列进行分析。用户无需具备专业的生物信息学知识,即可完成包括基于 cgSNP 与 cgMLST 策略构建系统发育树、耐药与毒力基因识别、质粒分型与同源性分析等在内的一系列病原菌的快速分型与溯源分析。随着越来越多的病原菌全基因组序列被测定,分析这些海

量数据往往需要消耗巨大的计算资源。而基于云计算技术的数据处理与存储能力的不断增强和计算成本的下降，或许会给病原菌分型与溯源领域的研究工作带来颠覆性的改变。本文归纳整理了病原菌基因组分型与溯源分析的相关理论知识及常用的生物信息学工具，并结合相关应用实例介绍，将有助于读者快速掌握其技术要领，并能直接应用于具体的科研工作中。

WGS 技术在病原菌鉴定、分型与溯源等领域广泛应用，为有效监测临床多重耐药菌的传播提供了新思路，同时也为如何有效监测、预警与防控多重耐药菌感染这一重要临床问题提供可行的解决方案。此外，在食源性病原菌以及自然疫源性病原菌的分型溯源中，WGS 也能提供较高的分辨率和准确率。病原菌的快速溯源和早期暴发事件监测的先决条件是建立快速准确的分型技术和国际统一的命名数据库(nomenclature database)。美国 FDA 和 NCBI 联合开发的 GenomeTrakr 数据库整合了所有成员单位上传的基于不同的测序平台的基因组数据，并提供菌株间的表型、基因型以及系统发育分析，有助于提高食源性暴发事件调查效率和降低食源性疾病的死亡率，促进了 WGS 技术在食源性病原菌溯源中的应用[82]。2013 年，中国推出国家食源性疾病监测系统(National Molecular Tracing Network for Foodborne Disease Surveillance，TraNet)，覆盖分离自人、食物和环境的 7 种常见食源性病原菌，总数已超过 54000 株。用户可通过不同的数据中心，上传多种类型的数据进行溯源分析。该系统在食品疾病监测方面发挥了重要作用，有助于迅速开展食源性疾病暴发流行事件的调查和溯源[83]。

WGS 技术的进一步发展与推广仍依赖于生物信息学家对基因组测序数据的整合、挖掘与解读，尤其是在实验流程、质量管理、性能验证和报告解读等方面还存在较大的进步空间。WGS 在临床微生物实验室的常规应用需建立一套方便、快速且成本可控的实践流程，包括从测序端的样本采集与建库到测序后的数据分析和案例解析，尤其是建立标准化数据分析流程与数据库、开发操作方便的生物信息学分析软件，将有助于推进该技术在临床诊疗的常规应用。本文综述的多个病原菌分型与溯源的生物信息学工具与数据库可为临床医务人员、流行病学和临床微生物检验人员提供一站式数据分析解决方案，并为推进全基因组测序技术在临床的常规应用提供技术支持。与此同时，还需紧密结合临床与公共卫生领域应用 WGS 技术的实际需求，熟悉满足需求的各类测序平台的技术特点，并培养能够驾驭该技术的生物信息学人才。只有当上述要素都趋于完备时，WGS 技术才能在临床诊疗及医院感染控制等常规应用中得以普及，并开辟基于病原菌基因组测序技术的感染性疾病精准诊疗的新领域。

参考文献

[1] Le Flèche P, Jacques I, Grayon M, Al Dahouk S, Bouchon P, Denoeud F, Nöckler K, Neubauer H, Guilloteau LA, Vergnaud G. Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay. *BMC Microbiology*, 2006, 6: 9.

[2] Thompson FL, Gevers D, Thompson CC, Dawyndt P, Naser S, Hoste B, Munn CB, Swings J. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology*, 2005, 71(9): 5107–5115.

[3] Maiden MCJ, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*, 2013, 11(10): 728–736.

[4] Gibson JR, Sutherland K, Owen RJ. Inhibition of DNAse activity in PFGE analysis of DNA from *Campylobacter jejuni*. *Letters in Applied Microbiology*, 1994, 19(5): 357–358.

[5] Metzker ML. Sequencing technologies-the next generation. *Nature Reviews Genetics*, 2010, 11(1): 31–46.

[6] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 1995, 269(5223): 496–512.

[7] Quainoo S, Coolen JPM, Van Hijum SAFT, Huynen MA, Melchers WJG, Van Schaik W, Wertheim HFL. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical Microbiology Reviews*, 2017, 30(4): 1015–1063.

[8] Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 1998, 95(6): 3140–3145.

[9] Kimura B. Will the emergence of core genome MLST end the role of *in silico* MLST? *Food Microbiology*, 2018, 75: 28–36.

[10] Deng XY, den Bakker HC, Hendriksen RS. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual Review of Food Science and Technology*, 2016, 7: 353–374.

[11] Silva M, Machado M, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: a complete suite for gene-by-gene *Schema* creation and strain identification. *Microbial Genomics*, 2018, 4(3): e000166.

[12] Feijao P, Yao HT, Fornika D, Gardy J, Hsiao W, Chauve C, Chindelevitch L. MentaLiST-a fast MLST caller for large MLST schemes. *Microbial Genomics*, 2018, 4(2): e000146.

[13] Zhang J, Halkilahti J, Hänninen ML, Rossi M. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. *Journal of Clinical Microbiology*, 2015, 53(5): 1765–1767.

[14] Zhang SK, Yin YL, Jones MB, Zhang ZZ, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng XY. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *Journal of Clinical Microbiology*, 2015, 53(5): 1685–1692.

[15] Zhang SK, Den Bakker HC, Li ST, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng XY. SeqSero2: rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. *Applied and Environmental Microbiology*, 2019, 85(23): e01746-19.

[16] Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, Underwood A. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ*, 2016, 4: e2308.

[17] Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN. The *Salmonella in silico* typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS One*, 2016, 11(1): e0147101.

[18] Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *Journal of Clinical Microbiology*, 2015, 53(8): 2410–2426.

[19] Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: user-friendly capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. *Journal of Clinical Microbiology*, 2018, 56(6): e00197-18.

[20] Wyres KL, Cahill SM, Holt KE, Hall RM, Kenyon JJ. Identification of *Acinetobacter baumannii* loci for capsular polysaccharide (KL) and lipooligosaccharide outer core (OCL) synthesis in genome assemblies using curated reference databases compatible with Kaptive. *Microbial Genomics*, 2020, 6(3): e000339.

[21] Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microbial Genomics*, 2018, 4(7): e000192.

[22] Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, 2014, 58(7): 3895–3903.

[23] Vielva L, De Toro M, Lanza VF, De La Cruz F. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics*, 2017, 33(23): 3796–3798.

[24] Li XB, Xie YZ, Liu M, Tai C, Sun JY, Deng ZX, Ou HY. oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. *Nucleic Acids Research*, 2018, 46(W1): W229–W234.

[25] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 2015, 43(3): e15.

[26] Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Van Der Linden M, McGee L, Von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. Rapid pneumococcal evolution in response to clinical interventions. *Science*, 2011, 331(6016): 430–434.

[27] Zhou ZM, McCann A, Weill FX, Blin C, Nair S, Wain J, Dougan G, Achtman M. Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, 111(33): 12199–12204.

[28] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 2010, 59(3): 307–321.

[29] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014, 30(9): 1312–1313.

[30] Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 2015, 32(1): 268–274.

[31] Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 2012, 61(3): 539–542.

[32] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 2014, 10(4): e1003537.

[33] Seemann T. Snippy: fast bacterial variant calling from NGS reads. https://github.com/tseemann/snippy. Accessed year: 2020.

[34] Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Van Domselaar G, Deng XY, Carleton HA. A comparative analysis of the lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Frontiers in Microbiology*, 2017, 8: 375.

[35] Petkau A, Mabon P, Sieffert C, Knox N, Cabral J, Iskander M, Iskander M, Weedmark K, Zaheer R, Katz LS, Nadon C, Reimer A, Taboada E, Beiko RG, Hsiao W, Brinkman F, Graham M, Consortium TI, Van Domselaar G. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microbial Genomics*, 2017, 3(6): e000116.

[36] Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One*, 2014, 9(8): e104984.

[37] Ruan Z, Feng Y. BacWGSTdb, a database for genotyping and source tracking bacterial pathogens. *Nucleic Acids Research*, 2016, 44(D1): D682–D687.

[38] Feng Y, Zou SM, Chen HF, Yu YS, Ruan Z. BacWGSTdb 2.0: a one-stop repository for bacterial whole-genome sequence typing and source tracking. *Nucleic Acids Research*, 2021, 49(D1): D644–D650.

[39] Pearce ME, Alikhan NF, Dallman TJ, Zhou ZM, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar enteritidis outbreak. *International Journal of Food Microbiology*, 2018, 274: 1–11.

[40] Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS, Mariet JF, Felten A, Aarestrup FM, Gerner Smidt P, Roussel S, Guillier L, Mistou MY, Hendriksen RS. An assessment of different genomic approaches for inferring phylogeny of *Listeria monocytogenes*. *Frontiers in Microbiology*, 2017, 8: 2351.

[41] Fida M, Cunningham SA, Murphy MP, Bonomo RA, Hujer KM, Hujer AM, Kreiswirth BN, Chia N, Jeraldo PR, Nelson H, Zinsmaster NM, Toraskar N, Chang WZ, Patel R,. Core genome MLST and resistome analysis of *Klebsiella pneumoniae* using a clinically amenable workflow. *Diagnostic Microbiology and Infectious Disease*, 2020, 97(1): 114996.

[42] Jones CL, Clancy M, Honnold C, Singh S, Snesrud E, Onmus-Leone F, McGann P, Ong AC, Kwak Y, Waterman P, Zurawski DV, Clifford RJ, Lesho E. Fatal outbreak of an emerging clone of extensively drug-resistant *Acinetobacter baumannii* with enhanced virulence. *Clinical Infectious Diseases*, 2015, 61(2): 145–154.

[43] Stegger M, Lindsay JA, Moodley A, Skov R, Broens EM, Guardabassi L. Rapid PCR detection of *Staphylococcus aureus* clonal complex 398 by targeting the restriction-modification system carrying *sau1-hsdS1*. *Journal of Clinical Microbiology*, 2011, 49(2): 732–734.

[44] Yu FY, Lv J, Niu SQ, Du H, Tang YW, Pitout JDD, Bonomo RA, Kreiswirth BN, Chen L. Multiplex PCR analysis for rapid detection of *Klebsiella pneumoniae* carbapenem-resistant (sequence type 258 [ST258] and ST11) and hypervirulent (ST23, ST65, ST86, and ST375) strains. *Journal of Clinical Microbiology*, 2018, 56(9): e00731-18.

[45] Hernández I, Sant C, Martínez R, Fernández C. Design of bacterial strain-specific qPCR assays using NGS data and publicly available resources and its application to track biocontrol strains. *Frontiers in Microbiology*, 2020, 11: 208.

[46] 杨献伟, 杨瑞馥, 崔玉军. 细菌基因组同源重组: 量化与鉴定. 遗传, 2016, 38(2): 137–143.
Yang XW, Yang RF, Cui YJ. Homologous recombination among bacterial genomes: the measurement and identification. *Hereditas*, 2016, 38(2): 137–143. (in Chinese)

[47] Ruan Z, Yu YS, Feng Y. The global dissemination of bacterial infections necessitates the study of reverse genomic epidemiology. *Briefings in Bioinformatics*, 2020, 21(2): 741–750.

[48] Hawkey J, Ascher DB, Judd L, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE. Evolution of carbapenem resistance in *Acinetobacter baumannii* during a prolonged infection. *Microbial Genomics*, 2018, 4(3): e000165.

[49] Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 2009, 26(7): 1641–1650.

[50] Holt K, Kenyon JJ, Hamidian M, Schultz MB, Pickard DJ, Dougan G, Hall R. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant *Acinetobacter baumannii* global clone 1. *Microbial Genomics*, 2016, 2(2): e000052.

[51] Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, Langille MG, Beiko RG. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLoS One*, 2013, 8(7): e69885.

[52] Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics*, 2016, 2(11): e000093.

[53] Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 2016, 44(W1): W242–W245.

[54] Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Research*, 2018, 46(1): e5.

[55] Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 2018, 34(2): 292–293.

[56] Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao CH, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJC, Yoo HS, Zhang CD, Zhang Y, Sobral BW. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, 2014, 42(D1): D581–D591.

[57] Obata F, Obrig T. Role of Shiga/Vero Toxins in Pathogenesis[M]. 2015: 73–95.

[58] DiCenzo GC, Finan TM. The divided bacterial genome: structure, function, and evolution. *Microbiology and Molecular Biology Reviews*, 2017, 81(3): e00019-17.

[59] Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, De La Cruz F. Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, 2010, 74(3): 434–452.

[60] Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. Identification of plasmids by PCR-based replicon typing. *Journal of Microbiological Methods*, 2005, 63(3): 219–228.

[61] Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Research*, 2019, 47(D1): D195–D202.

[62] Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*, 2019, 20(6): 356–370.

[63] Venditti C, Vulcano A, D'Arezzo S, Gruber CEM, Selleri M, Antonini M, Lanini S, Marani A, Puro V, Nisii C, Di Caro A. Epidemiological investigation of

an *Acinetobacter baumannii* outbreak using core genome multilocus sequence typing. *Journal of Global Antimicrobial Resistance*, 2019, 17: 245–249.

[64] Higgins PG, Hujer AM, Hujer KM, Bonomo RA, Seifert H. Interlaboratory reproducibility of DiversiLab rep-PCR typing and clustering of *Acinetobacter baumannii* isolates. *Journal of Medical Microbiology*, 2012, 61(1): 137–141.

[65] Gona F, Comandatore F, Battaglia S, Piazza A, Trovato A, Lorenzin G, Cichero P, Biancardi A, Nizzero P, Moro M, Cirillo DM. Comparison of core-genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. *Microbial Genomics*, 2020, 6(4): e000347.

[66] Eigenbrod T, Reuter S, Gross A, Kocer K, Günther F, Zimmermann S, Heeg K, Mutters NT, Nurjadi D. Molecular characterization of carbapenem-resistant *Acinetobacter baumannii* using WGS revealed missed transmission events in Germany from 2012-15. *The Journal of Antimicrobial Chemotherapy*, 2019, 74(12): 3473–3480.

[67] 张璐, 沈青春, 张纯萍, 赵琪, 崔明全, 李霆, 程敏. 全基因组测序技术对沙门氏菌血清型和耐药性的预测能力分析. 微生物学报, 2021: 1–13.
Zhang L, Sheng QC, Zhang CP, Zhao Q,Cui MQ, Li T, Cheng M. Predictive analysis of whole genome sequencing for *Salmonella* serotype and antimicrobial resistance phenotypes. *Acta Microbiologica Sinica*, 2021: 1–13. (in Chinese)

[68] Jiang M, Zhu F, Yang C, Deng YH, Kwan PSL, Li YH, Lin YM, Qiu YQ, Shi XL, Chen H, Cui YJ, Hu QH. Whole-genome analysis of *Salmonella enterica* serovar enteritidis isolates in outbreak linked to online food delivery, Shenzhen, China, 2018. *Emerging Infectious Diseases*, 2020, 26(4): 789–792.

[69] Morelli G, Song YJ, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li YJ, Cui YJ, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang RF, Carniel E, Achtman M. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics*, 2010, 42(12): 1140–1143.

[70] Achtman M, Morelli G, Zhu PZ, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P. Microevolution and history of the plague *Bacillus*, *Yersinia pestis*. *Proceedings of the National Academy of Sciences*, 2004, 101(51): 17837–17842.

[71] Cui YJ, Yu C, Yan YF, Li DF, Li YJ, Jombart T, Weinert LA, Wang ZY, Guo ZB, Xu LZ, Zhang YJ, Zheng HC, Qin N, Xiao X, Wu MS, Wang XY, Zhou DS, Qi ZZ, Du ZM, Wu HL, Yang XW, Cao HZ, Wang H, Wang J, Yao SS, Rakin A, Li YR, Falush D, Balloux F, Achtman M, Song YJ, Wang J, Yang RF. Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proceedings of the National Academy of Sciences*, 2013, 110(2): 577–582.

[72] 秦婧靓, 杨超, 郑宏源, 张湘莉兰, 武雅蓉, 崔玉军, 杨瑞馥, 赵光宇, 宋亚军. 1994 年印度苏拉特鼠疫暴发的回顾性基因溯源研究. 军事医学, 2021: 1–6.
Qin JL, Yang C, Zheng HY, Zhang XLL, Wu YR, Cui YJ, Yang RF, Zhao GY, Song YJ. A retrospective genetic traceability study of the 1994 Surat plague outbreak in India. *Military Medical Sciences*, 2021: 1–6. (in Chinese)

[73] Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng QD, Abouelleil A, Bochicchio B, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill PX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. Genomic epidemiology of the *Escherichia coli* O104: H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences*, 2012, 109(8): 3065–3070.

[74] 黄熙, 邓小玲, 梁骏华, 卢玲玲, 黄琼, 张永慧, 杨杏芬. 2011 年德国肠出血性大肠杆菌 O104:H4 感染暴发疫情溯源调查. 中国食品卫生杂志, 2011, 23(6): 555–559.
Huang X, Deng XL, Liang JH, Lu LL, Huang Q, Zhang YH, Yang XF. Tracing investigation of enterohemorrhagic *Escherichia coli* O104:H4 outbreak reported in Germany in 2011. *Chinese Journal of Food Hygiene*, 2011, 23(6): 555–559. (in Chinese)

[75] Reimer AR, Van Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L, Olsen-Rasmussen M, Frace M, Sammons S, Dahourou GA, Boncy J, Smith AM, Mabon P, Petkau A, Graham M, Gilmour MW, Gerner-Smidt P. Cholerae Outbreak Genomics Task Force V. Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases*, 2011, 17(11): 2113–2121.

[76] Orata FD, Keim PS, Boucher Y. The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathogens*, 2014, 10(4): e1003967.

[77] Baltazar M, Ngandjio A, Holt KE, Lepillet E, Pardos De La Gandara M, Collard JM, Bercion R, Nzouankeu A, Le Hello S, Dougan G, Fonkoua MC, Weill FX. Multidrug-resistant *Salmonella enterica* serotype typhi, gulf of Guinea region, Africa. *Emerging Infectious Diseases*, 2015, 21(4): 655–659.

[78] Jia HQ, Chen Y, Wang JF, Xie XY, Ruan Z. Emerging challenges of whole-genome-sequencing-powered epidemiological surveillance of globally distributed clonal groups of bacterial infections, giving *Acinetobacter baumannii* ST195 as an example. *International Journal of Medical Microbiology*, 2019, 309(7): 151339.

[79] Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clinical Microbiology and Infection*, 2018, 24(4): 350–354.

[80] Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW. On the evolutionary history, population genetics and diversity among isolates of *Salmonella enteritidis* PFGE pattern JEGX01.0004. *PLoS One*, 2013, 8(1): e55254.

[81] Payne M, Octavia S, Luu LDW, Sotomayor-Castillo C, Wang QN, Tay ACY, Sintchenko V, Tanaka MM, Lan RT. Enhancing genomics-based outbreak detection of endemic *Salmonella enterica* serovar Typhimurium using dynamic thresholds. *Microbial Genomics*, 2021, 7(6): 000310.

[82] Stevens EL, Timme R, Brown EW, Allard MW, Strain E, Bunning K, Musser S. The public health impact of a publically available, environmental database of microbial genomes. *Frontiers in Microbiology*, 2017, 8: 808.

[83] Li WW, Cui QP, Bai L, Fu P, Han HH, Liu JK, Guo YC. Application of whole-genome sequencing in the national molecular tracing network for foodborne disease surveillance in China. *Foodborne Pathogens and Disease*, 2021, 18(8): 538–546.

(本文责编　张晓丽)