

# 基于全基因组测序的细菌耐药基因注释准确性与数据库比对效能验证

王丽

贵州省疾病预防控制中心

DOI:10.32629/ffcr.v4i4.21473

**[摘要]** 目的:探讨全基因组测序耐药基因注释准确性及数据库比对效能。方法:选取2024年8月至2025年8月60例细菌感染患者菌株,随机分为两组各30株;对照组用单一数据库BLAST比对,观察组用多数据库整合校验,均以表型药敏为金标准。结果:观察组注释准确率(93.33%)及各类耐药预测准确率均高于对照组( $P<0.05$ );单株比对比耗时( $92.37\pm 21.54$ )s短于对照组( $186.45\pm 32.18$ )s,耐药基因检出数、注释准确率、灵敏度、特异度均高于对照组( $P<0.001$ )。结论:多数据库整合策略可提升耐药基因注释准确性与效能,利于临床应用。

**[关键词]** 全基因组测序; 耐药基因; 注释准确性; 数据库比对; 表型药敏

中图分类号: Q343.1 文献标识码: A

## Validation of the Accuracy of Bacterial Resistance Gene Annotation Based on Whole Genome Sequencing and the Efficacy of Database Comparison

Li Wang

Guizhou Center for Disease Control and Prevention

**[Abstract]** Objective: To explore the accuracy of bacterial resistance gene annotation based on whole genome sequencing and the efficacy of database comparison. Methods: 60 bacterial strain samples from patients with bacterial infections from August 2024 to August 2025 were selected and randomly divided into two groups, with 30 strains in each group. The control group used a single database (BLAST) for comparison, while the observation group used a multi-database integration for verification. Both were based on phenotypic drug sensitivity as the gold standard. Results: The annotation accuracy rate of the observation group (93.33%) and the accuracy rates of various resistance predictions were higher than those of the control group ( $P < 0.05$ ); the time for single-strain comparison ( $92.37 \pm 21.54$ ) s was shorter than that of the control group ( $186.45 \pm 32.18$ ) s, and the number of detected resistance genes, annotation accuracy rate, sensitivity, and specificity were all higher than those of the control group ( $P < 0.001$ ). Conclusion: The multi-database integration strategy can improve the accuracy and efficacy of bacterial resistance gene annotation and is beneficial for clinical application.

**[Key words]** Whole genome sequencing; Resistance gene; Annotation accuracy; Database comparison; Phenotypic drug sensitivity

细菌耐药已成为全球公共卫生领域亟待解决的突出难题,每年因耐药菌感染导致的死亡及医疗负担持续攀升,多重耐药菌的流行使临床抗感染治疗面临严峻挑战,部分感染甚至陷入无药可用的困境<sup>[1]</sup>。传统表型药敏试验虽为耐药判定的金标准,但其操作繁琐、检测周期较长,难以满足危重感染患者快速用药的需求。全基因组测序能够一次性获取菌株的完整遗传信息,通过对耐药基因加以注释实现对耐药表型的预测,随着测序成

本的不断下降与平台普及,该技术在耐药监测与精准治疗中展现出独特价值<sup>[2]</sup>。然而,注释结果的可靠性在很大程度上取决于所选用的参考数据库及比对流程,目前临床缺乏统一的注释标准,不同数据库在基因覆盖、命名规则与收录范围上存在差异,易引起注释偏差与效率瓶颈<sup>[3]</sup>。因此,系统评价耐药基因注释的准确性并比较数据库比对效能,对推动测序技术的临床转化具有现实意义。

1 资料与方法

1.1 一般资料

本研究采用对照设计,选取2024年8月至2025年8月某院明确诊断为细菌感染的患者60例的细菌检验标本,将标本中分离出的纯菌株进行培养研究。研究者按随机数字表法将60株菌株分为对照组与观察组,每组各30株(对应患者各30例),两组分别采用不同的数据库比对策略进行耐药基因注释。

对照组30例中,男16例,女14例;年龄24~78岁,平均(52.36±11.48)岁;感染病程3~15 d,平均(7.42±2.16) d;分离菌株中大肠埃希菌9株、肺炎克雷伯菌8株、金黄色葡萄球菌5株、铜绿假单胞菌4株、鲍曼不动杆菌4株。观察组30例中,男17例,女13例;年龄26~76岁,平均(53.18±10.92)岁;感染病程4~14 d,平均(7.68±2.34) d;分离菌株中大肠埃希菌8株、肺炎克雷伯菌9株、金黄色葡萄球菌6株、铜绿假单胞菌4株、鲍曼不动杆菌3株。两组在性别( $\chi^2=0.067$ ,  $P=0.796$ )、年龄( $t=0.283$ ,  $P=0.778$ )、病程( $t=0.447$ ,  $P=0.656$ )及菌种构成( $\chi^2=0.589$ ,  $P=0.964$ )方面比较,差异均无统计学意义( $P>0.05$ ),具有可比性。

纳入标准:①经临床与微生物学检查确诊为细菌感染,标本培养分离获得纯菌株者;②标本采集规范,菌株活性良好且DNA提取质量符合测序要求者;③年龄≥18岁;④临床与微生物学资料完整者。

排除标准:①多种细菌混合感染、难以分离获得纯培养者;②菌株DNA质量不达标或测序数据量不足、组装质量不合格者;③标本污染或培养失败者;④同一患者重复送检的同源菌株;⑤临床资料缺失、无法完成基因型与表型对照判读者。

1.2 方法

菌株鉴定与表型药敏(金标准):研究者将标本接种于血琼脂平板与麦康凯平板,于35℃培养18~24 h,采用基质辅助激光解吸电离飞行时间质谱仪鉴定至种。表型药敏采用微量肉汤稀释法测定最低抑菌浓度,受试抗菌药物涵盖氨苄西林、哌拉西林/他唑巴坦、头孢他啶、头孢曲松、头孢吡肟、亚胺培南、美罗培南、阿米卡星、庆大霉素、环丙沙星、左氧氟沙星及复方磺胺甲噁唑等,结果依据当年版CLSI折点判读为耐药、中介或敏感。

DNA提取与全基因组测序:研究者采用细菌基因组提取试剂盒提取DNA,经Qubit荧光定量、NanoDrop检测纯度(OD260/280为1.8~2.0),构建文库后于Illumina平台进行双端测序,测序深度不低于100×。原始数据经Trimmomatic去除接头与低质量碱基,FastQC评估测序质量,SPAdes完成de novo组装,QUAST评价组装完整性与连续性。

耐药基因注释与数据库比对(两组差异所在):对照组仅依托单一耐药基因数据库,按默认参数采用在线BLAST逐条序列比对,序列一致性阈值设为identity≥90%、覆盖度coverage≥60%。观察组采用本地化多数据库整合策略,联合CARD、ResFinder与AMRFinderPlus等核心数据库并行比对,对多源结果进行基因命名归一与交叉校验以剔除冗余及假阳性,并将阈值收紧至ident

ity≥95%、coverage≥80%,同时纳入gyrA、parC等位点的点突变检测。两组均将注释所得耐药基因与表型药敏结果逐项对应,据此判定真阳性、假阳性与假阴性。

1.3 观察指标

①耐药基因注释准确性:以表型药敏为金标准,统计两组总体注释完全准确例数及β-内酰胺类、碳青霉烯类、氨基糖苷类、喹诺酮类耐药预测的准确例数与准确率。②数据库比对效能:记录单株比对耗时、耐药基因检出数,并计算各菌株的注释准确率、灵敏度与特异度。

1.4 统计学方法

本研究采用SPSS 26.0软件分析数据。计量资料以( $\bar{x}\pm s$ )表示,组间比较采用独立样本t检验;计数资料以例(%)表示,组间比较采用 $\chi^2$ 检验。检验水准 $\alpha=0.05$ , $P<0.05$ 为差异有统计学意义。

2 结果

2.1 两组耐药基因注释准确性比较

观察组耐药基因注释完全准确率为93.33%,高于对照组的70.00%;在β-内酰胺类、碳青霉烯类、氨基糖苷类及喹诺酮类耐药预测方面,观察组准确率分别为96.67%、93.33%、96.67%、90.00%,均高于对照组的80.00%、73.33%、73.33%、66.67%,差异均有统计学意义( $P<0.05$ )。见表1。

表1 两组耐药基因注释准确性比较[例(%)]

| 组别         | 例数 | 注释完全准确     | β-内酰胺类预测准确 | 碳青霉烯类预测准确  | 氨基糖苷类预测准确  | 喹诺酮类预测准确   |
|------------|----|------------|------------|------------|------------|------------|
| 对照组        | 30 | 21 (70.00) | 24 (80.00) | 22 (73.33) | 22 (73.33) | 20 (66.67) |
| 观察组        | 30 | 28 (93.33) | 29 (96.67) | 28 (93.33) | 29 (96.67) | 27 (90.00) |
| $\chi^2$ 值 |    | 5.455      | 4.043      | 4.320      | 6.405      | 4.812      |
| P值         |    | 0.020      | 0.044      | 0.038      | 0.011      | 0.028      |

2.2 两组数据库比对效能比较

观察组单株比对耗时为(92.37±21.54) s,明显短于对照组的(186.45±32.18) s;观察组耐药基因检出数为(8.57±2.06)个,注释准确率为(95.18±4.21)%,灵敏度为(94.36±5.12)%,特异度为(96.72±3.84)%,均优于对照组的(6.83±1.74)个、(84.26±6.53)%(82.15±7.84)%(88.47±5.96)%,差异均有统计学意义( $P<0.001$ )。

3 讨论

细菌耐药的产生与扩散贯穿于临床抗感染治疗的全过程。致病菌通过产生灭活酶、改变药物靶点、外排泵过表达以及膜通透性下降等途径获得对抗菌药物的耐受能力,而这些机制大多由特定的耐药基因所编码<sup>[4]</sup>。准确识别菌株所携带的耐药基因,是判断其耐药谱、合理选择抗菌药物的关键环节。表型药敏试验长期作为耐药判定的金标准,但其检测往往需要数十小时,对脓毒症等危重感染难以及时提供用药参考。全基因组测序通过一次实验即可获取菌株的全部遗传信息,再借助生物信息学手

段对耐药基因加以注释,从而在较短时间内完成耐药表型的预测<sup>[5]</sup>。本研究围绕注释准确性与数据库比对效能两条主线设置观察指标,意在厘清测序技术在耐药判读环节的真实表现。

本研究结果显示,观察组在总体注释准确率及各药物类别的耐药预测上均优于对照组,这一差异主要源于两组比对策略的不同。对照组仅依赖单一数据库与默认参数的在线比对,受限于该数据库的基因收录范围,部分新近报道或亚型差异较大的耐药基因难以被检出,进而造成假阴性增多、灵敏度下降。观察组将CARD、ResFinder与AMRFinderPlus等数据库加以整合,不同数据库在 $\beta$ -内酰胺酶、碳青霉烯酶、氨基糖苷修饰酶等基因家族上的收录各有侧重,联合使用后基因覆盖范围明显拓宽,使更多真实携带的耐药基因得以识别;观察组耐药基因检出数较对照组增加,正印证了这一点<sup>[6]</sup>。注释准确不仅取决于“检得出”,还取决于“判得准”。各数据库在基因命名上并不统一,同一基因可能存在多个别名,若缺乏统一映射极易产生重复计数或误判。观察组在比对完成后对多源结果进行命名归一与交叉校验,并将一致性与覆盖度阈值由对照组的identity $\geq 90\%$ 、coverage $\geq 60\%$ 收紧至identity $\geq 95\%$ 、coverage $\geq 80\%$ ,使因低相似度片段引起的假阳性得到有效抑制,这是观察组特异度高于对照组的根本原因。值得关注的是,喹诺酮类耐药在两组中的预测准确率均相对偏低,原因在于该类耐药多由gyrA、parC等基因的点突变介导,而非简单的耐药基因获得,仅凭基因有无难以全面反映其耐药状态;观察组在流程中纳入了点突变检测,因而在该类别上的改善更为明显。在比对效能方面,观察组单株耗时不足对照组的一半<sup>[7]</sup>。对照组采用逐条序列在线提交的方式,既受网络传输与服务器排队的影响,又难以并行处理,整体效率偏低。观察组依托本地化数据库并采用并行比对,避免了反复的网络交互,同时通过流程整合压缩了人工干预环节,使单株分析时间大幅缩短。对临床而言,比对效率的提升意味着从样本到报告的周期被显著压缩,能够更快地为经验性治疗的调整提供分子层面的依据,这一点对耐药形势复杂的重症监护病房尤为重要。需要客观看待的是,基于基因注释的耐药预测仍存在固有局限<sup>[8]</sup>。耐药基因的存在并不必然等同于耐药表型的表达,基因的转录水平、调控元件及宿主背景都会影响最终表型;个别菌株还可能因新型耐药机制或尚未被收录的基因而被漏判。鉴于此,本研究始终以表型药敏作为参照,强调基因型与表型的相互印证,而非以测序结果单独取代传统药敏。此外,本研究样本量与菌种范围有限,

结论的外推尚需保持审慎。

综合本研究结果,多数据库整合并辅以交叉校验与阈值优化的比对策略,能够在提升细菌耐药基因注释准确性的同时改善比对效能,较单一数据库的传统比对更契合临床快速、可靠的需求。后续工作可在扩大菌种与样本规模的基础上,引入长读长测序以提升复杂耐药结构的解析能力,并探索机器学习辅助的注释与质控方法,推动注释流程与判读标准的规范化,使全基因组测序更稳健地融入耐药监测与精准抗感染的临床实践。

### 参考文献

- [1]杨璟珂,杨思婕,刘冉,等.1株携带多黏菌素耐药基因mcr-1的多重耐药大肠杆菌全基因组测序及分析[J].中国畜牧兽医,2025,52(12):5956-5966.
- [2]杨舒,杨丽华,盛峰松,等.基于全基因组测序技术分析10株弯曲菌耐药基因、毒力基因和分子分型特征[J].检验医学,2025,40(7):673-679.
- [3]董浩,赵明海,刘志国,等.兔源肠致病性大肠埃希菌全基因组测序及毒力和耐药基因分析[J].实验动物科学,2024,41(1):70-76.
- [4]唐梦君,张小燕,周倩,等.一株同时携带oprA和cfrC基因的猪源多重耐药结肠弯曲菌的全基因组测序[J].微生物学通报,2023,50(8):3538-3549.
- [5]穆成,孙蕊,王志锐,等.基于全基因组测序的72株耐药结核分枝杆菌耐药基因突变特征分析[J].中国热带医学,2023,23(7):725-729+741.
- [6]杜娜,杜艳,刘曦,等.全基因组测序对碳青霉烯类耐药肺炎克雷伯菌T361耐药及毒力基因的分析[J].微生物学免疫学进展,2023,51(2):53-59.
- [7]胡小骞,王琴.基于全基因组测序的耐碳青霉烯类抗生素鲍曼不动杆菌耐药基因、毒力因子及同源性分析[J].中国抗生素杂志,2022,47(9):939-945.
- [8]林本夫,任欣悦,袁晓琪,等.细菌全基因组测序技术用于沙门氏菌流行病学调查的可行性分析[J].中国动物检疫,2022,39(6):125-131.

### 作者简介:

王丽(1982-),女,汉族,河北秦皇岛人,副主任技师,从事工作微生物检验。