遗传 Hereditas (Beijing) 2018 年 1 月, 40(1): 57－65
www.chinagene.cn

综 述

细菌全基因组关联研究的方法与应用

杨超，杨瑞馥，崔玉军

军事医学科学院微生物流行病研究所，病原微生物生物安全国家重点实验室，北京 100071

摘要：随着测序技术的发展和全基因组序列的不断积累，全基因组关联研究(genome-wide association study, GWAS)在人类复杂疾病研究中取得了丰硕成果，10 余年间发现了数以万计的疾病风险因子。同样，GWAS 也为探索细菌表型的遗传机制提供了新的工具。自 2013 年第一项细菌 GWAS(bacterial GWAS, BGWAS)工作发表以来，目前已有 10 多项相关研究报道，分别揭示了细菌宿主适应性、耐药性及毒力等表型的遗传机制，极大加深了人们对细菌遗传、进化及传播等方面的认识。本文对目前 BGWAS 的研究方法、应用成果及存在的问题进行了总结，并对 BGWAS 的研究前景进行了展望，旨在为微生物学领域开展 BGWAS 研究提供参考。

关键词：全基因组关联研究；细菌；全基因组测序；表型；遗传机制

Bacterial genome-wide association study: methodologies and applications

Chao Yang, Ruifu Yang, Yujun Cui

State Key Laboratory of Pathogen and Biosecurity, Institute of Microbiology and Epidemiology, Beijing 100071, China

Abstract: With the development of genome sequencing and the accumulation of whole genome sequences, genome-wide association study (GWAS) has achieved remarkable advances in understanding of human complex disease, and tens of thousands of disease risk factors have been found. Meanwhile, GWAS provides a new tool for exploring the genetic mechanism of bacterial phenotypes. Since the publication of the first bacterial GWAS (BGWAS) work in 2013, there have been more than 10 reports, which reveal the genetic basis of host adaption, drug resistance and virulence, etc. These findings greatly enhance our understanding on genetics, evolution and spread of bacteria. In this review, we summarize the current methodologies, applications and problems of BGWAS and highlight its potential in future research, which aims to provide helps for the applications of BGWAS in the field of microbiology.

Keywords: genome-wide association study; bacteria; whole genome sequencing; phenotype; genetic mechanism

收稿日期: 2017−09−13; **修回日期:** 2017−10−30
基金项目: 病原微生物生物安全国家重点实验室自主课题(编号：SKLPBS1405) [Supported by the funding of the State Key Laboratory of Pathogen and Biosecurity (No. SKLPBS1405)]
作者简介: 杨超，博士研究生，研究方向：病原细菌进化与致病机制。E-mail: chaoy.cn@gmail.com
通讯作者: 崔玉军，副研究员，硕士生导师，研究方向：病原细菌基因组学与进化。E-mail: cuiyujun.new@gmail.com
DOI: 10.16288/j.yczz.17-303
网络出版时间: 2017/12/15 10:06:25
URI: http://kns.cnki.net/kcms/detail/11.1913.R.20171215.1006.004.html

全基因组关联研究(genome-wide association study, GWAS)是一种从全基因组水平筛选与某表型(phenotype)显著相关的遗传变异，进而阐明表型遗传机制的方法[1]。相较于传统的分子遗传学方法，GWAS 并不对表型产生的遗传机制做任何假设；而是直接从表型出发，设置合理的对照组，通过大样本的数据统计分析找到与表型关联的遗传变异，因此该研究方法可以应用于复杂表型研究。GWAS 极大增进了人们对复杂表型的认识，在人类复杂疾病研究中取得了丰硕成果[2~5]。自 2005 年第一项有关黄斑病变的人类 GWAS[6]发表以来，目前公开发表的人类 GWAS 工作已达到 2982 项，累计报道了 36948 个与疾病/表型相关的单核苷酸多态性位点(single nucleotide polymorphism, SNP)[7]，为人类复杂疾病的预防治疗指明了道路。

GWAS 同样可用于细菌研究，为宿主适应性、毒力等复杂表型的遗传机制探索提供新思路[8,9]。然而受限于早期相对匮乏的全基因组数据，细菌 GWAS (bacterial GWAS, BGWAS)开展相对较晚。随着近年来高通量测序技术的发展，细菌测序成本快速下降，全基因组序列也得以迅速积累，目前 NCBI 数据库中已有近 10 万个细菌样本的全基因组序列，为 BGWAS 工作奠定了基础。自 2013 年空肠弯曲杆菌(*Campylobacter jejuni*)宿主适应性的 BGWAS 文章发表以来[10]，目前已有 10 余项 BGWAS 工作被发表(表 1)。这些研究揭示与细菌宿主适应性、耐药性及毒力等重要表型相关的基因组变异[10~24]，极大加深了人们对细菌遗传、进化和传播等的认识。

相较于人类 GWAS，BGWAS 存在天然的“优势”。例如：较小的基因组使得测序成本、计算资源需求和计算时间减少，降低了研究门槛；另外，细菌更容易通过分子生物学实验对所发现的表型相关遗传变异进行验证[8]。同时，BGWAS 也面临着特有的挑战，如细菌的分裂繁殖特性所导致的克隆种群结构、物种间重组率差异造成的多样化的连锁不平衡模式、以及基因的高频率获得缺失等[8,15,25]。针对

表 1 细菌全基因组关联研究(BGWAS)示例

Table 1 Examples of bacterial genome-wide association study (BGWAS)

物种名	重组率	样本量	表型	基因型	显著相关	软件	发表时间	参考文献
空肠弯曲杆菌 (*Campylobacter jejuni*)	高	192	宿主适应性	k-mer	7307 k-mer (7 个基因)	–	2013	[10]
		102	生物膜形成	k-mer	1657 kmer (46 个基因)	–	2015	[16]
		600	存活力	k-mer	3382 k-mer (20 个基因)	–	2016	[24]
		166	诊断标记	基因	25 个非核心基因	–	2017	[21]
结核杆菌 (*Mycobacterium tuberculosis*)	低	123	耐药性	SNP	50 SNPs	phyC	2013	[11]
		123		SNP	133 SNPs	PLINK	2015	[15]
		498		SNP, indel	12 SNPs	Bayes Traits	2016	[18]
金黄色葡萄球菌 (*Staphylococcus aureus*)	低	75	耐药性	SNP	1 SNP	ROADTRIPS	2014	[12]
		90	毒力	SNP, indel	121 SNPs, indels	PLINK	2014	[14]
肺炎链球菌 (*Streptococcus pneumoniae*)	高	3701	耐药性	SNP, indel	301 SNPs	PLINK	2014	[13]
		1680		SNP，基因	426 SNPs	PLINK	2017	[23]
		2175	体内运输	k-mer	2 SNPs, 424 kmer	fast-LMM, SEER	2017	[22]
猪链球菌 (*Streptococcus suis*)	高	191	宿主适应性	SNP，基因，k-mer	0	PLINK	2015	[17]
单增李斯特菌 (*Listeria monocytogenes*)	低	104	毒力	基因	43 个基因	–	2016	[19]
鲍曼不动杆菌 (*Acinetobacter baumannii*)	高	122	耐药性	k-mer	469 k-mer	bugwas	2016	[20]

这些问题，已有的 BGWAS 工作提出了不同的解决方案。本文梳理了当前 BGWAS 的研究方法，并对取得的成果及存在的问题进行了总结，以期为微生物学领域开展 BGWAS 研究提供参考。

1　BGWAS 的研究方法与工具

BGWAS 的研究方法从人类 GWAS 发展而来，并在研究过程中开发了特有的思路和工具。BGWAS 主要可以分成以下 4 个步骤：表型选取及采样，表型及基因型(genotype)测定、相关性检验及实验验证(图 1)。

1.1　表型选取及采样

选择合适的表型是 BGWAS 的第一步。表型通常可分为连续性数据(如细菌细胞的尺寸)和二分类数据(case/control)，分别对应不同的相关性检验方法。尽管连续性数据的检验效力更高，但其数据难以获得、统计检验更复杂，因此目前仅有两项样本量较小的 BGWAS 研究使用了连续性表型数据[12,14]。相对而言，易于通过高通量方法获取大样本量数据的二分类表型是 BGWAS 研究的首选。

选定表型后需要进行样本采集。采样时需注意采样方式和样本量的问题。采样方式可分为连续采样(time-coursed)和横断面采样(cross-sectional)两种[26]。连续采样的样本，如分离同一病人的不同时期菌株及实验室进化菌株等获取难度较大，难以满足 BGWAS 的样本量需求。横断面采样通过收集一定时间内大量相关样本，如病人诊断样本或公共卫生监测样本等，能够快速的获得更综合全面的信息，

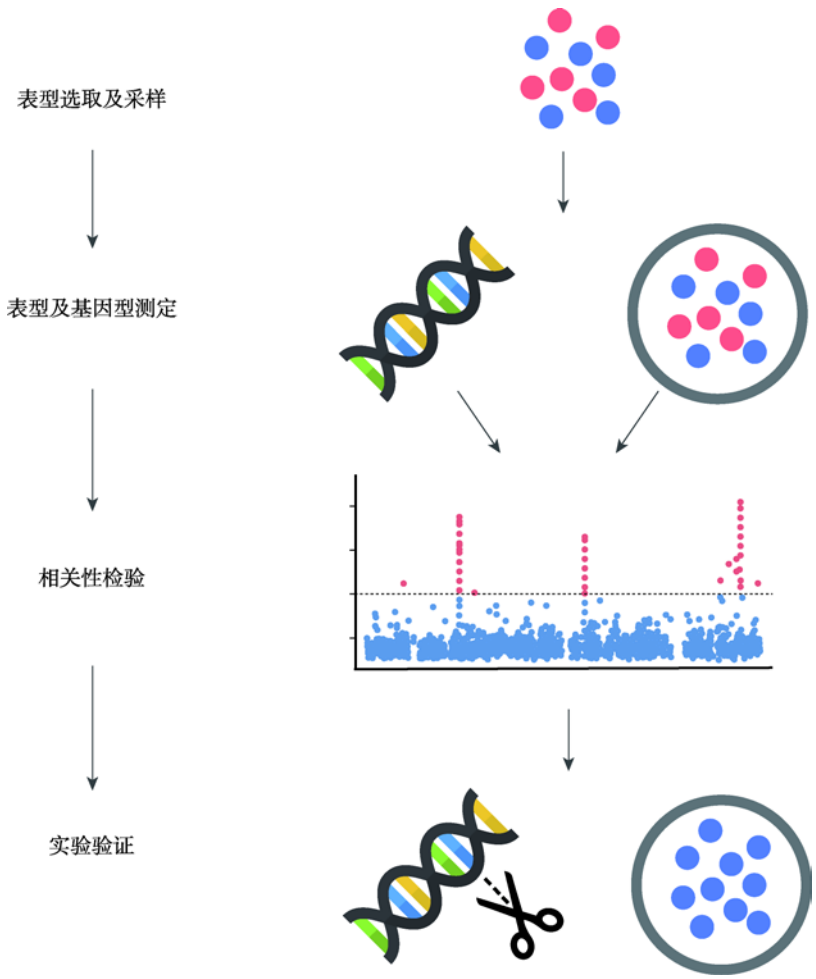

图 1　细菌全基因组关联研究(BGWAS)分析流程图

Fig. 1　Flow of bacterial genome-wide association studies (BGWAS)

是 BGWAS 的首选采样方式[26]。此外，采样时可以采用 Farhat 等[26]提出的成对采样策略，即优先选择遗传距离近、表型不同的菌株对。通过模拟计算发现成对采样不仅能有效降低种群结构带来的假阳性，还能显著提高统计检验效力。实际应用中可先用脉冲场凝胶电泳(pulsed field gel electrophoresis, PFGE)或多位点序列分型(multilocus sequence typing, MLST)等传统的快速分型方法对样本进行初步分型，然后根据分型结果筛选成对菌株进行测序用于 BGWAS。

样本量大小也是 BGWAS 研究设计中的重要问题，较大的样本量能够增加统计检验效力，但同时会增加研究成本。影响分析时所需样本量的主要因素是效应量(effect size)，即遗传变异能够解释的表型变异量。效应量是表型与基因型相关性的度量单位，其取值范围从 0 到 1，1 表示遗传变异能够完全解释表型[27]。对于效应量较大的变异，如抗生素抗性相关变异等，仅需要相对少量样本即可鉴定出相关性；而低效应量的变异则需要较大的样本量来保证检验的统计效力。当前 BGWAS 主要致力于高效应量变异鉴定，绝大多数样本量相对较小(表 1)。理论上低效应量变异在细菌中一定存在并在细菌进化和生存中发挥重要作用；随着样本量的累积增大及 BGWAS 技术的进一步发展, 相信将来会有越来越多的研究关注这类变异。

1.2 表型及基因型测定

实验室检测是获得细菌表型信息的最主要方式。此外，许多数据库整理保存了多种细菌表型信息，如 PATRIC(https://www.patricbrc.org/)整合了多种细菌的基因组及对应耐药性等表型信息, NCBI 的 Pathogen Detection(https://www.ncbi.nlm.nih.gov/pathogens/) 整合了病原菌基因组及对应的分离时间、地点、宿主等背景信息等。这些数据库提供的海量信息将极大的促进 BGWAS 的发展。

随着测序技术的发展及细菌变异分析软件的完善，获得细菌基因组变异及相应的基因型信息已经越来越快速且准确[28~30]。细菌基因型数据可分为 SNP、插入缺失(indel)、非核心基因获得缺失及 k-mer 等类型(表 1)。SNP 具有高分辨率、易鉴定等优势，是当前 BGWAS 最常用的基因型数据。SNP 用于分析之前通常要进行质量控制，当前常用的质控标准为：位点测序质量值大于 20，支持的 reads 数大于 5 条，最小等位基因频率(minor allele frequency, MAF)大于 1%或 5%。此外，细菌 SNP 鉴定中会出现 3 态甚至 4 态 SNP，但为了便于计算，实际应用中通常只有 2 态 SNP 用于关联分析。SNP 变异的局限性是仅能反映细菌核心基因组信息，而许多细菌具有开放型泛基因组[31]，这些菌株的非核心基因组变异也与细菌表型密切相关。因此，部分 BGWAS 研究也整合了 indel、基因获得缺失信息等非核心基因组变异信息。此外，越来越多的研究使用 k-mer 来研究基因组变异(表 1)。k-mer 是指将全基因组序列切分成的长度为几十到上百个碱基的短片段。通过使用基于图论等的算法计算 k-mer 在不同样本间的存在与否，可以同时综合分析 SNP 变异及基因获得缺失等信息, 从而能够更全面的探索表型的遗传变异机制，这类分析方法正越来越得到科研工作者的青睐。

1.3 相关性检验

获得各样本的表型以及基因型数据后，需要对两类数据之间进行相关性检验，但是在 BGWAS 中直接对两者做相关性检验容易造成假阳性结果。细菌的种群结构是造成假阳性相关的主要因素[8,15,25]。当研究对象可分成不同种群时，同一种群内部个体之间的遗传距离相对种群之间遗传距离更近，造成等位基因频率的非随机分布；当某一种群仅集中于对照组或者实验组时, BGWAS 会鉴定出许多与分群相关而不是与表型相关的变异，进而导致假阳性的产生。降低种群结构所致假阳性的最直接方法是：采样时尽可能选取遗传异质性低的样本作为研究对象(如人类 GWAS 研究通常选择在同一人种甚至同一民族内进行)，或者使用上文提到的 Farhat 成对采样策略[26]。但是即使选择同一种群的细菌样本，仍可能存在更精细的亚群结构。为了消除种群结构影响，部分研究沿用了人类 GWAS 工作中建立的软件进行分析(表 1)，如 PLINK[32]、ROADTRIPS[33]、fast-LMM[34]等。此外，针对细菌自身特点的新算法也被不断开发出来，如结合系统发育信息和蒙特卡罗模拟的系统发育校正方法[10,35]以及基于线性混合模型的聚类法等[36~38]。这些方法已被整合到 BGWAS 分析工具中，并得到实际应用(表 2)。

表 2　细菌全基因组关联研究(BGWAS)工具

Table 2　Software applications used in bacterial genome-wide association study (BGWAS)

研究工具	发表时间	特点	种群结构处理	适用性	应用	下载链接	参考文献
phyC	2013	通过检测趋同进化鉴定表型相关位点	系统发育校正	低、中重组率细菌	结核杆菌(*Mycobacterium tuberculosis*)	–	[11]
bugwas	2016	基于 k-mer，同时检测表型相关位点及家系(lineage)	线性混合模型校正	所有细菌	结核杆菌(*Mycobacterium tuberculosis*)，金黄色葡萄球菌(*Staphylococcus aureus*)，大肠杆菌(*Escherichia coli*)，肺炎克雷伯菌(*Klebsiella pneumoniae*)，鲍曼不动杆菌(*Acinetobacter baumannii*)	https://github.com/jessiewu/bacterialGWAS	[36]
SEER	2016	无需参考序列，可变 k-mer 长度，支持连续表型数据	多维尺度变换	所有细菌	肺炎链球菌(*Streptococcus pneumonia*)，酿脓链球菌(*Streptococcus pyogenes*)	https://github.com/johnlees/seer	[38]
Scoary	2016	针对非核心基因，简单快速	两两比对及置换检验	所有细菌	肺炎链球菌(*Streptococcus pneumoniae*)，表皮葡萄糖菌(*Staphylococcus epidermidis*)	https://github.com/AdmiralenOla/Scoary	[37]
treeWAS	2017	整合重组及表型聚类信息，支持连续表型	系统发育校正	低、中重组率细菌	脑膜炎双球菌(*Neisseria meningitidis*)	https://github.com/caitiecollins/treeWAS	[35]

由于重组率的物种间差异，不同细菌的连锁不平衡模式多种多样[39]，这也为 BGWAS 带来假阳性问题。对于重组率较低的细菌，变异之间相互连锁，难以区分“搭车”变异与真正与表型相关的变异。目前对该问题的常用解决方法是：综合变异位点特性(如同义/非同义突变)及所在基因的功能注释进行深入推测，缩小靶标范围，并进而通过分子生物学实验以验证靶标变异是否确实与某种表型相关。对于高频重组细菌，变异之间连锁程度低，此类样本更适用于使用传统 GWAS 方法进行分析。

此外，自然选择可以对细菌种群结构形成非常大的影响，如抗生素选择压力可能在自然界中筛选出特定的病原菌克隆群。能够检测正向选择引起的趋同变异的 phyC[11]及整合家系效果(lineage effect)检测的 bugwas[36]能够有效解决这类问题。

除了种群结构和重组率的影响，多重检验带来的假阳性也是 GWAS 不可避免的问题。GWAS 通常涉及数以万计的相关性检验，那么按照常用的显著性阈值 P<0.05 时，理论上会随机产生数百个假阳性结果。如此高的假阳性率显然无法接受，因此需要对多重检验进行校正。目前 BGWAS 主要沿用了人类 GWAS 中常用多重检验校正方法，如 Bonferoni 校正(显著性阈值=0.05/N，N 为变异位点数)及假发现率校正(false discovery rate correction)等[40]，能够显著降低假阳性结果的数量。

1.4　实验确认

尽管目前采用多种策略来降低 GWAS 结果的假阳性，但假阳性问题仍然难以完全避免。为此，人类 GWAS 通常需要重复研究来确认表型相关变异[41]。得益于细菌易实验操纵的特点，BGWAS 中鉴定的靶标变异可通过实验室验证的方法来排除假阳性。Falkow[42]在 1988 年提出了分子科赫法则，即“基因失活造成表型消失，重建则表型恢复”。这为实验确认 GWAS 鉴定的相关变异提供了标准。另外，基因敲除/重组技术的进步及突变体文库的完善极大的方便了 BGWAS 结果确认，多项 BGWAS 通过实验对相关位点进行了验证，确认了变异与表型的相关性[10,11,14,18,19,24]。

1.5　BGWAS 的研究工具

目前已有多种工具被开发出来，用于解决 BGWAS 分析所面临的问题(表 2)。Farhat 等[11]开发了通过检测趋同进化来鉴定表型相关变异的软件 phyC。该软

件能够显著降低假阳性，适用于强选择性状相关变异的检测[15]。Earle 等[36]利用线性混合模型整合样本相关性来校正种群结构，开发出了能够同时检测表型相关位点及家系效果的软件 bugwas，成功应用于 3000 多株不同重组率细菌的耐药性研究。Lees[38]利用“Scale-mining”算法，开发出了高计算效率、支持可变长度 k-mer 的 SEER。Brynildsrud 等[37]针对细菌泛基因组，开发出了能够快速检测表型相关基因获得/缺失的 Scoary。Collins 等[35]通过整合重组及表型聚类信息开发出了基于系统发育校正的 treeWAS 软件。除了专门针对细菌非核心基因组的 Scoary 软件外，bugwas、SEER 等都支持 k-mer 运算，能够同时捕捉核心及非核心基因组变异信息。由于 BGWAS 软件开发仍处于萌芽阶段(到目前为止，多数 BGWAS 软件发表不到一年)，这些工具的实际应用价值还有待实践检验。在实际应用中，可以根据基因型数据类型，选择多个软件同时进行数据分析，对运算结果做交叉验证。

2 BGWAS 研究的应用进展

通过 BGWAS 研究，多种重要表型与遗传因子之间的相关性被建立起来。目前半数以上 BGWAS 研究是针对细菌耐药性开展的[11-13,15,18,23]。如 Farhat 等[11]通过检测趋同进化来筛选结核杆菌(*Mycobacterium tuberculosis*)的耐药相关变异，除了找到了过去已知的全部耐药位点外，还发现了 39 个新的耐药相关区域。第一项大样本的 BGWAS 同样关注于耐药性问题，Chewapreecha 等[13]通过分析 3701 株肺炎链球菌(*Streptococcus pneumoniae*)，找到了与 β 内酰胺类抗性相关的 301 个 SNP 位点。部分 BGWAS 研究关注宿主适应性、毒力等细菌与宿主相互作用的表型。例如，最早的 BGWAS 研究通过分析 192 株来源于不同宿主的空肠弯曲杆菌(*Campylobacter jejuni*)，发现并验证了一组维生素 B5 合成相关基因与宿主饮食适应相关，既而导致某些基因型的细菌倾向于生活在特定种类的宿主中[10]。Laabei 等[14]通过分析 90 株毒力不同的金黄色葡萄球菌(*Staphylococcus aureus*)，发现了 121 个毒力相关因子，并通过实验验证了 4 个毒力因子，进一步增进了人们对细菌毒力的认识。此外，BGWAS 还被应用于生物膜形成[16]、存活力[24]、体内运输[22]等多种细菌生理相关的表型研究，为解析这些表型的遗传机制提供了新的数据。

通过 BGWAS 发现的表型相关变异可以很好的促进细菌表型预测研究。Laabei 等[14]利用 BGWAS 在金黄色葡萄球菌(*Staphylococcus aureus*)中鉴定出的 50 个毒力相关变异，结合“随机森林”机器学习算法，建立了该病原的毒力预测模型，其预测准确率高达 85%以上。Mobegi 等[23]利用类似的方法建立了肺炎链球菌(*Streptococcus pneumoniae*)耐药性预测模型，能够根据基因组序列定量评估分离株耐药性的强弱。表型预测模型的建立进一步拓展了 BGWAS 的用途，随着算法和模型的完善，将极大加速细菌表型信息的获取，对基因的功能研究以及细菌性病原的监测和控制等领域具有重要意义。

值得关注的是，BGWAS 研究还可以应用于开发新的临床诊断标记。Buchanan 等[21]通过对 166 株空肠弯曲杆菌(*Campylobacter jejuni*)的泛基因组序列 BGWAS 分析，发现 25 个非核心基因的获得缺失与弯曲杆菌病发病相关。这些遗传标记可通过 PCR 等方法实现快速检测，因而可以方便的应用于对细菌病原所致疾病的临床诊断和治疗中。

3 结语与展望

尽管只经历了短短几年发展历史，BGWAS 研究已经取得了丰硕成果。通过 BGWAS 分析，人们揭示了细菌多种重要表型的相关遗传因子，极大加深了人们对细菌遗传机制、适应性进化及传播等领域的认识，并为医学临床诊断、治疗和公共卫生领域的进步提供了新的思路。

BGWAS 本质上是在不同数据组之间建立关联。数据组的种类、获取难度、累积数量与关联算法决定了 BGWAS 的应用前景。(1)未来的 BGWAS 研究将更加全面，不只着眼于耐药性等效应量大的表型，也会增加对效应量小、相关性较弱的表型与变异的关注，从全局角度重新认识功能基因对细菌表型的影响。(2) BGWAS 会与宿主基因组数据整合分析，通过细菌基因组、表型与宿主基因组的综合性关联

分析，进一步增进人们对细菌变异是否受宿主影响这一问题的认识。此外，这种细菌—宿主相互作用研究将帮助人们识别细菌的靶标蛋白，进而促进药物及疫苗等的开发[25]。(3) BGWAS 可以与宏基因组数据相结合，通过多种细菌之间的基因组关联分析，进一步增进人们对细菌协作、竞争等行为的认识，为宏基因组研究提供新的手段。(4) BGWAS 结果易于进行实验室验证的特点，将促进该领域在理论上得到迅速发展，从而可以对人类 GWAS 研究的基础理论和算法提供支持，为多基因连锁控制某一性状等难题提供新的解决思路。(5) BGWAS 在临床细菌检验与疾病诊断中有着广阔的发展前景。临床检验实验如最小抑菌浓度检验等，能为 BGWAS 提供海量的重要表型信息，而 BGWAS 能利用这些信息来鉴定相关变异，进而建立和优化细菌表型预测模型，极大改善临床检验与诊断的速度及准确性。相信随着测序成本的进一步降低和研究工具的持续更新，会使 BGWAS 的深厚发展潜力得以爆发。

参考文献(References):

[1] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 2012, 8(12): e1002822. [DOI]

[2] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*, 2012, 90(1): 7–24. [DOI]

[3] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS discovery: Biology, function, and translation. *Am J Hum Genet*, 2017, 101(1): 5–22. [DOI]

[4] Yan WL. Genome-wide association study on complex diseases: genetic statistical issues. *Hereditas (Beijing)*, 2008, 30(5): 543–549.
严卫丽. 复杂疾病全基因组关联研究进展——遗传统计分析. 遗传, 2008, 30(5): 543–549. [DOI]

[5] Han JW, Zhang XJ. Current status of genome-wide association study. *Hereditas (Beijing)*, 2011, 33(1): 25–35.
韩建文, 张学军. 全基因组关联研究现状. 遗传, 2011, 33(1): 25–35. [DOI]

[6] DeWan A, Liu M, Hartman S, Zhang SSM, Liu DTL, Zhao C, Tam POS, Chan WM, Lam DSC, Snyder M, Barnstable C, Pang CP, Hoh J. *HTRA1* promoter polymorphism in wet age-related macular degeneration. *Science*, 2006, 314(5801): 989–992. [DOI]

[7] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 2014, 42 (Database issue): D1001–D1006. [DOI]

[8] Falush D, Bowden R. Genome-wide association mapping in bacteria?. *Trends Microbiol*, 2006, 14(8): 353–355. [DOI]

[9] Falush D. Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol*, 2016, 1: 16059. [DOI]

[10] Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci USA*, 2013, 110(29): 11923–11927. [DOI]

[11] Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PK, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet*, 2013, 45(10): 1183–1189. [DOI]

[12] Alam MT, Petit RA 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biol Evol*, 2014, 6(5): 1174–1185. [DOI]

[13] Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet*, 2014, 10(8): e1004547. [DOI]

[14] Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden M, Wilson D, Read TD, Van Den Elsen J, Priest NK, Feil EJ, Hurst LD, Josefsson E, Massey RC. Predicting the virulence of MRSA from its genome sequence. *Genome Res*, 2014, 24(5): 839–849. [DOI]

[15] Chen PE, Shapiro BJ. The advent of genome-wide associ-

ation studies for bacteria. *Curr Opin Microbiol*, 2015, 25: 17–24. [DOI]

[16] Pascoe B, Méric G, Murray S, Yahara K, Mageiros L, Bowen R, Jones NH, Jeeves RE, Lappin-Scott HM, Asakura H, Sheppard SK. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in *Campylobacter jejuni*. *Environ Microbiol*, 2015, 17(11): 4779–4789. [DOI]

[17] Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Välimäki N, Harris D, Chieu TT, Van Vinh Chau N, Campbell J, Schultsz C, Parkhill J, Bentley SD, Langford PR, Rycroft AN, Wren BW, Farrar J, Baker S, Hoa NT, Holden MT, Tucker AW, Maskell DJ, BRaDP1T Consortium. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nat Commun*, 2015, 6: 6740. [DOI]

[18] Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, Shea TP, Almeida DV, Manson AL, Salazar A, Padayatchi N, O'donnell MR, Mlisana KP, Wortman J, Birren BW, Grosset J, Earl AM, Pym AS. Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate ald in D-cycloserine resistance. *Nat Genet*, 2016, 48(5): 544–551. [DOI]

[19] Maury MM, Tsai YH, Charlier C, Touchon M, Chenal-Francisque V, Leclercq A, Criscuolo A, Gaultier C, Roussel S, Brisabois A, Disson O, Rocha EPC, Brisse S, Lecuit M. Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat Genet*, 2016, 48(3): 308–313. [DOI]

[20] Suzuki M, Shibayama K, Yahara K. A genome-wide association study identifies a horizontally transferred bacterial surface adhesin gene associated with antimicrobial resistant strains. *Sci Rep*, 2016, 6: 37811. [DOI]

[21] Buchanan CJ, Webb AL, Mutschall SK, Kruczkiewicz P, Barker DOR, Hetman BM, Gannon VPJ, Abbott DW, Thomas JE, Inglis GD, Taboada EN. A genome-wide association study to identify diagnostic markers for human pathogenic *Campylobacter jejuni* strains. *Front Microbiol*, 2017, 8: 1224. [DOI]

[22] Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, Turner P, Bentley SD. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife*, 2017, 6: e26255. [DOI]

[23] Mobegi FM, Cremers AJH, De Jonge MI, Bentley SD, Van Hijum SAFT, Zomer A. Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. *Sci Rep*, 2017, 7: 42808. [DOI]

[24] Yahara K, Méric G, Taylor AJ, De Vries SPW, Murray S, Pascoe B, Mageiros L, Torralbo A, Vidal A, Ridley A, Komukai S, Wimalarathna H, Cody AJ, Colles FM, McCarthy N, Harris D, Bray JE, Jolley KA, Maiden MCJ, Bentley SD, Parkhill J, Bayliss CD, Grant A, Maskell D, Didelot X, Kelly DJ, Sheppard SK. Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ Microbiol*, 2017, 19(1): 361–380. [DOI]

[25] Power RA, Parkhill J, De Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet*, 2017, 18: 41–50. [DOI]

[26] Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Med*, 2014, 6(11): 101. [DOI]

[27] Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med*, 2014, 6(11): 109. [DOI]

[28] Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*, 2015, 15(2): 141–161. [DOI]

[29] Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet*, 2015, 6: 236. [DOI]

[30] Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol*, 2015, 13(12): 787–794. [DOI]

[31] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol*, 2015, 23: 148–154. [DOI]

[32] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 2007, 81(3): 559–575. [DOI]

[33] Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown

population and pedigree structure. *Am J Hum Genet*, 2010, 86(2): 172–184. [DOI]

[34] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods*, 2011, 8(10): 833–835. [DOI]

[35] Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *bioRxiv*, 2017, 140798. [DOI]

[36] Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook DW, Mcvean G, Walker AS, Wilson DJ. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*, 2016, 1: 16041. [DOI]

[37] Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol*, 2016, 17(1): 238. [DOI]

[38] Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, Honkela A, Parkhill J, Bentley SD, Corander J. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*, 2016, 7: 12797. [DOI]

[39] Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. *Trends Microbiol*, 2010, 18(7): 315–322. [DOI]

[40] Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 2003, 100(16): 9440–9445. [DOI]

[41] Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. Replicating genotype-phenotype associations. *Nature*, 2007, 447(7145): 655–660. [DOI]

[42] Falkow S. Molecular Koch's postulates applied to microbial pathogenicity. *Rev Infect Dis*, 1988, 10 Suppl 2: S274–S276. [DOI]

(责任编委: 包其郁)