植物学报 *Chinese Bulletin of Botany* 2020, **55** (6): 715–732, www.chinbullbotany.com
doi: 10.11983/CBB20091

·特邀专家方法·

生物信息学分析方法I: 全基因组关联分析概述

赵宇慧[1], 李秀秀[1,2], 陈倬[1,2], 鲁宏伟[1,2], 刘羽诚[1,2], 张志方[1,2], 梁承志[1,2*]

[1]中国科学院遗传与发育生物学研究所, 北京 100101; [2]中国科学院大学, 北京 100049

摘要 全基因组关联分析(GWAS)是动植物复杂性状相关基因定位的常用手段。高通量基因分型技术的应用极大地推动了GWAS的发展。在植物中, 利用GWAS不仅能够以较高的分辨率在全基因组水平鉴定出各种自然群体特定性状相关的基因或区间, 而且可揭示表型变异的遗传架构全景图。目前, 人们利用GWAS分析方法已在拟南芥(*Arabidopsis thaliana*)、水稻(*Oryza sativa*)、小麦(*Triticum aestivum*)、玉米(*Zea mays*)和大豆(*Glycine max*)等模式植物和重要农作物品系中发掘出与各种性状显著相关的数量性状座位(QTL)及其候选基因位点, 阐明了这些性状的遗传基础, 并为揭示这些性状背后的分子机理提供候选基因, 也为作物高产优质品种的选育提供了理论依据。该文对GWAS的方法、影响因素及数据分析流程进行了详细描述, 以期为相关研究提供参考。

关键词 混合线性模型, 全基因组关联分析(GWAS), 生物信息学

赵宇慧, 李秀秀, 陈倬, 鲁宏伟, 刘羽诚, 张志方, 梁承志 (2020). 生物信息学分析方法I: 全基因组关联分析概述. 植物学报 **55**, 715–732.

1 GWAS概述

全基因组关联分析(genome-wide association study, GWAS)是一种通过检验全基因组遗传标记与表型变异关联的显著性来定位与性状相关的遗传位点, 在群体水平上解析性状遗传基础的方法。影响GWAS的关键因素之一是群体水平存在连锁不平衡(linkage disequilibrium, LD)。重组是打断LD的主要因素(Visscher et al., 2012; Xiao et al., 2017)。LD的大小主要受群体遗传多样性的影响, 在不同物种和群体中差异很大。例如, 玉米(*Zea mays*)群体的LD通常比水稻(*Oryza sativa*)群体的LD小很多, 而相近的现代栽培品种群体的LD往往都比较大(Zhag et al., 2016; Li et al., 2020)。传统的QTL定位研究通常以2个亲本杂交群体为研究对象, 通过连锁作图定位目标性状位点。这种方法的局限性在于人为杂交构建群体过程中产生的重组事件少(LD大), 为实现精细定位, 往往需要投入大量资源构建数量庞大的重组群体。而关联分析则可以利用研究对象自然群体的历史重组(Yu and Buckler, 2006), 有机会获得更高分辨率的定位结果, 同时遗传变异来源也更为广泛, 往往能定位到比双亲本作图群体中更多的性状关联位点。由于LD的存在, 当基因组中存在造成表型差异的变异时, 该变异附近的遗传标记也倾向于与表型产生关联, 从而检测出含有控制表型变异基因的染色体区域。

GWAS已广泛应用于解析表型变异的遗传构造, 发现与表型变异相关的位点, 可为功能基因研究提供候选基因/位点, 并为育种应用提供分子标记。但GWAS也存在一定的缺点, 如群体结构造成的假阳性, 遗传异质性造成位点效应相互掩盖等。为了解决这些问题, 研究者主要采用两方面的策略: 其一是在算法上, 通过在关联分析模型中考虑亲缘关系和群体结构的影响, 对关联结果进行校正; 其二是在关联群体上, 选取亲缘关系和群体结构不显著, 但是表型变异丰富的群体(Yano et al., 2016), 或构建人工关联群体。

2 GWAS科研设计

GWAS需要考虑的问题包括群体的选取、群体结构分

收稿日期: 2020-05-20; 接受日期: 2020-08-26
基金项目: 中国科学院战略性先导科技专项(No.XDA24040201)
* 通讯作者。E-mail: cliang@genetics.ac.cn

析、表型鉴定、数据获取方式和全基因组关联分析方法选择及结果矫正。

2.1 群体的选取

群体中丰富的表型变异和充分的遗传重组是GWAS成功的关键条件。因此，重点考虑选取以下2种群体: (1) 群体内没有明显的群体结构，样本间没有过近的亲缘关系，同时具有丰富的表型变异; (2) 群体来自具有一定水平遗传分化的不同类群(如水稻的亚种和亚群)，具有丰富的遗传和表型变异，但同时不同类群之间存在频繁的遗传交流，保证目标性状在不同类群内部也存在一定水平的变异。若有条件，也可以从头构建更为理想的多亲本杂交群体，如MAGIC (multi-parent advanced generation intercross)群体和NAM (nested association mapping)群体。

样本量会影响GWAS鉴定关联位点的数目(Huang et al., 2011)。样本量越大, LD越小，关联分析结果的统计学意义更有保证。但样本量越大，成本越高。因此，GWAS需要在考虑目标性状的复杂性及样本多样性的情况下确定合适的样本量(Wang et al., 2020)。为了保证检测效力，目前GWAS样本量普遍大于100份(Visscher et al., 2017; Alqudah et al., 2020)。例如，水稻的GWAS一般需要200–5 000个样本(Wang et al., 2020)。大麦(*Hordeum vulgare*)的样本量一般在100–500个(Kumar et al., 2012)。对于表型变异丰富、性状由1–2个明显的主效应位点控制时，样本量在200个以上即可(Wang et al., 2016, 2020); 对于表型差异小，由多个基因控制的复杂性状需要增加样本量，最好大于500个。在多基因对表型变异的贡献超过50%时，500个样本足以检测出表型解释度在5%以上的QTL位点(Wang and Xu, 2019)。但是对于由低频等位基因控制的性状，要适当增加样本量和样本多样性。

2.2 群体结构分析

GWAS方法中, LD的度量极其重要。通过检测目标群体中LD衰减的速度，可以了解群体内历史重组的强度，预估有效的关联分析需要的标记密度以及关联分析的分辨率。群体结构会导致不连锁的区间出现LD，引起目标性状与无关基因之间发生关联，从而导致出现假阳性位点。因此，在进行关联分析前需要进行群体结构分析，将群体结构作为协变量来提高计算的准确度。

主成分分析(principal component analysis, PCA)是群体结构主流分析方法之一。PCA的主要作用在于排除群体中的异常个体，对基因型降维，从而控制群体结构(Price et al., 2006; Raj et al., 2014; Wang et al., 2019)。多个软件如EIGENSTRAT、GCTA和PLINK均可完成PCA (Abegaz et al., 2019)。

通过PCA对遗传标记降维投影可以直接可视化群体结构，然而，有时候仅用样本的投影坐标不能解释它的全局祖先估计(Martin et al., 2018)。与PCA不同, STRUCTURE类软件通过基于数据来显式生成模型的方法解决了这个问题; 即直接从模型参数的后验分布来计算全局祖先估计。STRUCTURE利用贝叶斯法进行全局祖先估计(Pritchard et al., 2000; Falush et al., 2003; Hubisz et al., 2009); FRAPPE (Tang et al., 2005)和ADMIXTURE (Alexander et al., 2009)基于最大似然法来估计模型参数。FRAPPE缺乏估算最佳K值(分组数)的方法; ADMIXTURE和STRUCTURE虽然使用相同的模型，但ADMIXTURE的速度比STRUCTURE快。fastSTRUCTURE基于经验贝叶斯框架，采用变分推理方案来进行全局祖先估计，它的准确性类似ADMIXTURE，速度较STRUCTURE加快2个数量级(Raj et al., 2014)。

大多数研究同时采用PCA和显式生成模型2种方法来分析群体结构，以保证结果的可靠性(Alqudah et al., 2016; Milner et al., 2019; Song et al., 2019; Zhang et al., 2019b)。PCA的结果中PCA1和PCA2能够解释大部分个体之间的总体差异。STRUCTURE软件利用ΔK与K (K: 分组数; ΔK: 该分组的likelihood参数，用于评估分组的可靠性)作图来确定合理的亚群个数。此外，通过亚群个数与log-likelihood的变化趋势也能够确定所研究群体有无群体结构(Alqudah et al., 2020)。

2.3 表型数据类型

表型数据是关联分析的基础。为了获得可靠的表型数据，通常需要多年多点的重复来尽量减少误差。从关联分析的方法考虑，一般要求表型数据为连续数据，但近年来研究表明，离散数据和分类数据在特定情况下的GWAS中也可以获得较好的关联结果。表型类型

对于关联分析统计方法的选择有重要影响(Gumpinger et al., 2018)。表型从传统的发育性状不断扩展，现在不仅包括大规模的分子水平的定量特征(例如，Tieman等(2017)和zhu等(2018)通过代谢物全基因组关联分析(mGWAS)定位到与番茄(*Solanum lycopersicum*)的葡萄糖、果糖和番茄碱等多种代谢物含量相关的位点)，还有可遗传和量化的复杂表型(Liu et al., 2019)，如水稻杂种优势(Huang et al., 2015)和玉米单倍体育性(Ma et al., 2018a)。

2.4 分子标记数据获取

基于单核苷酸多态性(SNP)标记的重要功能(Griffith et al., 2008)及其对遗传多样性的贡献，GWAS分析常选择SNP作为基因组区间的理想标记。为了保证定位的准确性，需要的最少SNP标记个数为*N*=基因组大小/LD衰减距离。随着衰减速度加快，所需SNP标记数增多，假阳性率降低，定位精确度升高(Myles et al., 2009; Sallam and Martsch, 2015; Alqudah et al., 2020)。SNP芯片和基因组测序是普遍使用的获取SNP数据的方法，两者各有优缺点(Tam et al., 2019)。SNP芯片可信度和准确度高，有成熟的数据分析流程和工具，但芯片主要针对已知变异位点，且定制芯片价格昂贵。基因组测序在足够深度下能够检测所有类型的遗传变异，但成本相对较高，计算资源需求大，处理、分析数据及结果解释有一定难度。然而，随着测序技术的快速发展，测序成本不断下降，基于全基因组重测序数据进行GWAS的研究逐渐增多。通过全基因组重测序数据不仅可以鉴定SNP标记，还能筛选拷贝数量异常(CNV)和存在/缺失变异(PAV)等结构变异标记；这非常适合仔细比较少量关键亲本、地方品种和野生型的基因组变异以指导育种过程(Li et al., 2017)。

全基因组重测序已广泛应用于西瓜(*Citrullus lanatus*) (Guo et al., 2019)、苹果(*Malus domestica*) (Duan et al., 2017)、水稻(Wang et al., 2015; Xie et al., 2015; Tong et al., 2016; Ma et al., 2019b)、大豆(*Glycine max*) (Zhou et al., 2015; Fang et al., 2017)、菜豆(*Phaseolus vulgaris*) (Wu et al., 2020)、棉花(*Gossypium hirsutum*) (Du et al., 2018; Ma et al., 2018b, 2019a)和鹰嘴豆(*Cicer arietinum*) (Thudi et al., 2016; Li et al., 2017, 2018)等植物的GWAS研究中。Li等(2017)对69份鹰嘴豆进行全基因组重测序，将枯萎病抗性相关位点精确定位在1个100 kb的区间内，该区间有NBS-LRR受体激酶、锌指结构蛋白以及丝氨酸/苏氨酸蛋白激酶等12个蛋白质编码基因。Li等(2018)对132份鹰嘴豆进行全基因组重测序，通过GWAS筛选出38个SNPs，与百粒重、每公顷产量和空荚比等6个产量性状相关。Varshney等(2019)对429份鹰嘴豆进行全基因组重测序，鉴定出900多个与耐热耐旱相关的标记。

2.5 模型方法选择及结果矫正

GWAS中质量性状关联分析通常采用Logistic回归模型；数量性状关联分析可以采用一般线性模型(general linear model, GLM)和混合线性模型(mixed linear model, MLM)。一般线性模型以群体结构矩阵Q或主成分分析矩阵为协变量来提高计算精度；混合线性模型利用群体结构矩阵Q、亲缘关系矩阵(kinship, K)或联合利用主成分分析矩阵和亲缘关系矩阵为协变量来抑制假关联的出现(Yu et al., 2006; Yang et al., 2014)。针对数量性状易受多因素影响的特征，混合线性模型广泛应用于数量性状的关联分析。基于混合线性模型衍生出众多方法(表1)。

用标准混合线性模型处理大样本数据效率低，计算时间长。为了提升计算速度，减少计算量，EMMA方法首先尝试通过简化矩阵运算，缩短了运算时间(Kang et al., 2008)。之后，相继出现基于不同假设的高效模型以适应不断增加的样本量和标记密度。典型方法包括EMMAX (Kang et al., 2010)、GRAMMAR (Aulchenko et al., 2007)、GRAMMAR-Gamma (Svishcheva et al., 2012)、FaST-LMM (Lippert et al., 2011)和GEMMA (Zhou and Stephens, 2014)。EMMAX是关联分析速度提升的一个代表性算法，已广泛应用于棉花、大豆和水稻等的复杂性状关联分析(Huang et al., 2016; Fang et al., 2017; Du et al., 2018; Hübner et al., 2019; Wu et al., 2020)。FaST-LMM方法设计的出发点是快速对超大型数据集进行GWAS研究，采用该方法成功鉴定出样本量达500–1 500个的水稻群体中与叶长、叶夹角和种子蛋白含量等上百个性状相关的位点(Xie et al., 2015; Bai et al., 2016; Chen et al., 2018; Dong et al., 2018)。近年来，FaST-LMM成功应用于水稻、番茄、

表1 不同混合线性模型(MLM)的性能比较

Table 1 Performance comparison of different methods in mixed linear model (MLM)

Method	Population structure	Kinship	Precision	Characteristic	Computational speed	Statistical power	Application
Standard MLM	✓	All markers			Low	High	>100 papers
GRAMMAR	✓		Approximate method		Very fast	Intermediate	Barley (200)
EMMA	✓		Exact method		Intermediate	Similar to Standard MLM	>100 papers
EMMAX	✓	All markers	Approximate method	High marker densities	Fast	Similar to Standard MLM	>100 papers
CMLM	✓			Large sample sizes		Better than Standard MLM	>100 papers
FaST-LMM	✓	A subset of genetic markers	Exact method	Large sample sizes	Fast	Similar to Standard MLM	Rice (200–1500)
GEMMA	✓		Exact method		Fast	Similar to Standard MLM	*Arabidopsis thaliana* (190–500)
ECMLM	✓				Intermediate	Better than Standard MLM	Sorghum (250–350), soybean (200–400), wheat (250–300)
GRAMMAR-Gamma	✓		Approximate method	High marker densities	Fast	Similar to Standard MLM	Oilseed rape (200)
SUPER	✓	Trait-associated markers		Large sample size & high marker density	Fast	Better than Standard MLM	Wheat (300–400)
Farm-CPU	✓	A subset of genetic markers	Approximate method	Large sample size & high marker density	Fast	Better than Standard MLM	Wheat (100–1200), maize (100–5000)
BLINK	✓	A subset of genetic markers	Approximate method	Large sample size & high marker density	Faster than FarmCPU	Better than Farm-CPU	

标准线性模型EMMA、EMMAX和CMLM被成百上千篇文章引用，此处省略物种及群体大小的详细统计分析。最后一列显示GWAS不同模型所研究物种和群体规模。

Standard MLM EMMA, EMMAX and CMLM were cited by hundreds of papers. The column of Application lists the species and population size of studies which used these models in GWAS of plants.

小麦(*Triticum aestivum*)和玉米等植物的mGWAS和TWAS (全转录组关联分析)等GWAS的扩展分析(Dong et al., 2015; Zhu et al., 2018; Kremling et al., 2019; Chen et al., 2020)。

上述方法虽然显著提高了运算速度，但是对检测效力的改善有限(Tang et al., 2016; Xiao et al., 2017)。Zhang等(2010)率先提出低秩矩阵混合模型CMLM，该模型使用分组的遗传效应代替个体的遗传效应，从而将统计效力提高5%–15%，并在此基础上进一步优化出ECMLM方法(Li et al., 2014)。随后，相继开发出一系列提高检测效力的模型，如FaST-LMM-Select (Listgarten et al., 2013)、SUPER (Wang et al., 2014)及BOLT-LMM (Loh et al., 2015)。近年来，在兼顾运算速度与检测效力的前提下，FarmCPU基于固定模型和随机模型循环迭代关联分析，不仅可以处理大样本量，还可以进行海量高密度标记的检测(Liu et al., 2016)。FarmCPU在小麦、玉米和大豆的大规模群体产量以及抽穗期和抗病性状相关QTL的鉴定中发挥重要作用(Li et al., 2016, 2019; Kaler et al., 2017; Kusmec et al., 2017; Bhatta et al., 2018; Kidane et al., 2019; Lozada et al., 2019)。BLINK针对FarmCPU进行了如下优化：首先用基于贝叶斯的固定模型替换随机模型；其次，用LD信息替换bin方法。BLINK在检测效力和运算速度方面均优于FarmCPU (Huang et al., 2019)。

当前育种目标已经从单一性状改良转向高产、优质、抗病和抗逆等综合性状的普遍改良，因而产生了多个相关性状联合的混合模型方法，主要包括

MTMM (Korte et al., 2012)、GEMMA (mvLMMs) (Zhou and Stephens, 2014)、mtSet (Casale et al., 2015)和mvLMM (Furlotte and Eskin, 2015)。上述研究表明, 采用多个相关性状联合分析的策略在功效和精度上均优于单个性状分析。

基于不同的遗传学或者统计学假设, 涌现出众多混合线性模型方法。GWAS需要综合考虑数据量、计算速度、统计效力和使用便捷性等因素, 选择合适的方法。针对样本数量达到上万例、样本量远超标记数量的超大群体GWAS研究, 采用FaST-LMM方法所需计算资源少, 运行速度快。对于标记密度大的GWAS研究, 可采用EMMAX方法进行分析。对于具有基因组大、样本数量多和标记密度大等特征的GWAS研究, 可采用SUPER、FarmCPU和BLINK方法进行分析, 这些方法运行速度快, 可检测到更多已知位点。目前, 为了确保结果的准确性和可靠性, 许多GWAS同时采用多个模型来进行分析, 经过比较筛选出最优解(Wei et al., 2017; Peng et al., 2018; Zhang et al., 2019c)。现有软件将多个模型集成为一个分析工具, 可完成多项GWAS相关分析。GAPIT和TASSEL是主流软件。GAPIT整合了EMMAX、FaST-LMM、Farm-CPU及Blink等众多模型, 而且可以进行基因型和表型诊断、PCA以及关联分析等, 结果以用于发表文章的图片形式呈现(Tang et al., 2016)。TASSEL提供对用户友好的图形化界面, 操作简单, 可以进行SNP calling、LD分析以及群体结构分析等, 广受欢迎(Bradbury et al., 2007)。

为了控制假阳性, 筛选出真正有意义的关联位点, 需要通过多重检验矫正来确定合理的显著性阈值。阈值的设定原则与所研究物种、群体以及研究目的密不可分(Kaler and Purcell, 2019; Alqudah et al., 2020)。例如, 为了描绘特定性状的遗传结构蓝图, 可设定宽松的阈值, 而为了筛选实验验证的候选位点则需要设定严格阈值。目前, 主要方法有Bonferroni矫正、FDR (false discovery rate)以及置换检验(De et al., 2014; Jiang and Wang, 2018)。在这3种方法中, Bonferroni矫正法最严格, 它的矫正公式为0.05/SNP的数量。相对于Bonferroni矫正法, FDR法较为宽松, 它针对每个性状单独计算一个FDR值, 随标记数与性状变化, 方式更灵活。置换检验方法灵活而稳健, 但计算量很大, 比较耗时。综上, Bonferroni矫正和FDR是植物GWAS研究中确定显著性阈值的常用方法。

3 GWAS数据分析流程

我们以基于平均测序深度7×的721份水稻材料(Li et al., 2020)全基因组重测序数据为例来说明GWAS研究的常规流程。一般情况下, GWAS数据分析流程包括数据比对、call SNP鉴定基因型、表型统计以及基因型表型关联分析(图1)。

3.1 重测序数据质控和比对

利用Trimmomatic (Bolger et al., 2014)或Fastx (http://hannonlab.cshl.edu/fastx_toolkit/download.html)来获得高质量的有效数据。过滤条件为: (1) 去除接头序列; (2) 去掉3'和5'端质量值低于20的碱基; (3) 将质量低于20的碱基数超过reads长度10%的reads移除; (4) 将长度少于36 bp的reads移除。随后, 通过fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc)检查测序数据质量。

3.1.1 质量控制

(1) 数据过滤:

方法1: Trimmomatic

```
java-jar trimmomatic-0.33.jar PE -threads 16 -phred33 [sample1_R1].fastq.gz \
[sample1_R2].fastq. gz \
[sample1_clean_PE_1].fastq.gz [sample1_clean_UP_1].fastq.gz \
[sample1_clean_PE_2].fastq.gz [sample1_clean_UP_2].fastq.gz
ILLUMINACLIP: TruSeq3-PE.fa:2:30:3 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

其中, 输出文件sample1_clean_PE_1.fastq.gz和sample1_clean_PE_2.fastq.gz是过滤后保留的双端数据, sample1_clean_UP_1.fastq.gz和sample1_clean_UP_2.fastq.gz是双端数据过滤后丢弃低质量的一端数据, 仅保留另一端高质量数据。

方法2: Fastx

```
fastq_quality_filter-q 20 –p 50 -i [sample1_R1].fastq -o [sample1_R1_clean].fastq
fastq_quality_filter-q 20 –p 50 -i [sample1_R2].fastq -o [sample1_R2_clean].fastq
```

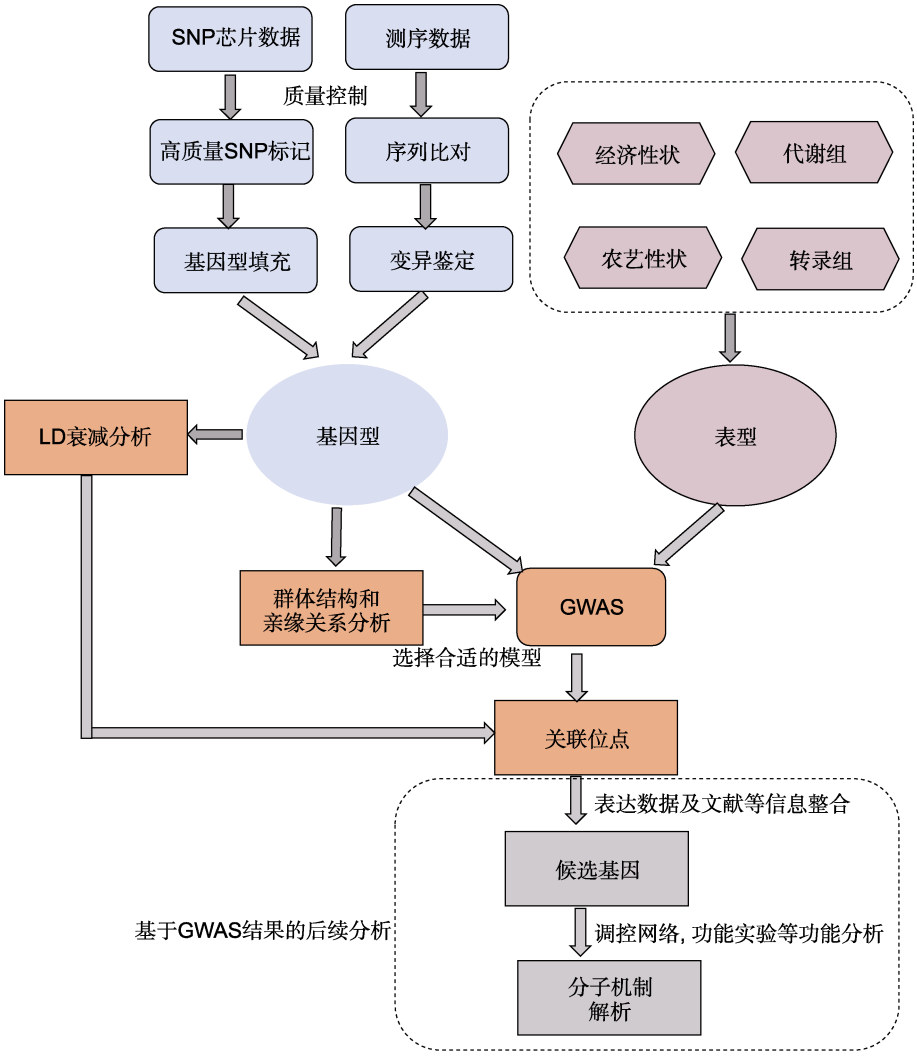

图1 全基因组关联分析(GWAS)流程

Figure 1 The pipeline of genome-wide association study (GWAS)

(2) 质量检测

Fastqc -o [outdir/outname] --extract-f *.clean_ fastq.gz

备注: 用户自定义文件名或变量用方括号标出, 所有分析代码均用此方式表示。

3.1.2 数据比对及结果统计

利用BWA-MEM (Li et al., 2013)或Bowtie (Langmead et al., 2009)将高质量的有效数据比对到参考基因组。根据比对率、深度和覆盖度对数据进行整体评估。数据达到饱和是检测出足够数量SNP的基础。

(1) 数据比对

基于参考基因组构建索引: bwa index [ref], 其中ref是<参考基因组序列>。

比对: bwa mem –M –t [threads] -R “@RG\tID:[name]\tLB:[name]\tSM:[name]\tPL:illumina\tPU:[name]“ [ref] [R1_clean].fq [R2_clean].fq | samtools view -bS >[name.source].bam, 其中–t是<线程数>, -R “@RG\tID:<样本名称>\tLB: <样本名称>\tSM: <样本名称>\tPL:<测序平台类型>\tPU:<样本名称>“, name.source.bam是bam格式的比对结果。

将比对结果进行质控: samtools view-h [name.source].bam | samtools view-bS-q30 > [name].bam, 其中name.bam是高质量的比对结果。

将比对结果进行排序: samtools sort [name].bam [name].sorted

基于比对结果构建索引: samtools index [name].sorted.bam

(2) 查看比对结果

samtools flagstat [name].source.bam > [name].source.mapinfo

(3) 查看测序深度和对基因组的覆盖度

方法1: SOAP

soap.coverage –cvg –sam –p 5 –i [name].sam –refsingle [ref] –o [name].coverage, 其中, –i是将sam格式的比对结果作为输入文件, –o是输出的样本覆盖度文件。

方法2: BEDTools+SAMtools

bedtools genomecov -ibam [name].sorted.bam >[name].coverage, 其中-ibam是bam格式的有序比对结果, 输出文件是样本覆盖度。

samtools depth -a [name].bam >[name].depth, 其中-a是bam格式的比对结果, 输出文件是样本测序深度。

3.2 变异位点鉴定和分型

使用GATK (McKenna et al., 2010; DePristo et al., 2011)或SAMtools (Li et al., 2011)鉴定SNP和基因分型。结果一般保留缺失率小于0.2、maf值大于0.05的SNP。

3.2.1 利用GATK (GenomeAnalysisTK-3.8-0)流程进行变异位点鉴定和分型

GATK call SNP有2种模式: UnifiedGenotyper和HaplotypeCaller。

(1) GATK UnifiedGenotyper鉴定变异位点命令:

```
java -Xmx15g -Djava.io.tmpdir=./tmp[i] -jar GenomeAnalysisTK.jar \
-nt $core \ #线程数
-glm BOTH \ #变异检测类型, BOTH同时输出SNP和Indel
-T UnifiedGenotyper \ #变异检测工具
[-L "[chrfile_name]"] \
-R [ref] \ #参考基因组序列
-I [name1.sorted].bam \ #第一个样本的比对结果bam文件
-I [name2.sorted].bam….. \ #第二个样本的比对结果bam文件
-o [SNP.list].vcf \ #输出的变异鉴定结果vcf文件
-metrics./all.UniGenMetrics.[i]\
-stand_call_conf 50.0 \
-stand_emit_conf 10.0 \
-dcov 1000 \
-A Coverage \
-A AlleleBalance
```

(2) GATK HaplotypeCaller鉴定变异位点命令:

Step1: 生成每个样本的GVCF文件

```
Java -Xmx30g -Djava.io.tmpdir=./tmp[i] -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \ #变异检测工具
-R [ref] \ #参考基因组序列
-I [name].sort.bam \ #样本的有序bam格式比对结果
-o [name].g.vcf \ #输出gvcf文件
-nct 4 \
--emitRefConfidence GVCF
```

Step2: 从GVCF文件鉴定群体变异位点

```
java -Xmx30g -Djava.io.tmpdir=./tmp[i] -jar GenomeAnalysisTK.jar \
-T GenotypeGVCFs \ #变异检测工具
-R [ref] \ #参考基因组序列
-V [name1].g.vcf \ #第一个样本的gvcf文件
-V [name2].g.vcf….. \ #第二个样本的gvcf文件
-nct 4 \
-o [out].vcf #输出的变异鉴定结果vcf文件
```

3.2.2 通过SAMtools进行变异位点鉴定和分型

bcftools mpileup [name].sorted.bam --fasta-ref [ref].fa | bcftools call -cv -o [raw].vcf, 其中name.sorted.bam是样本的有序bam格式比对结果, raw.vcf是vcf格式的原始变异鉴定结果。

filter variants: bcftools view [raw].vcf | misc/ vcfutils.pl varFilter > [name-final].vcf, 其中raw.vcf是vcf格式的原始变异鉴定结果, name-final.vcf是过滤后变异鉴定结果。

3.2.3 利用VCFtools或PLINK过滤缺失频率高以及次要等位频率较低的SNP, 保证关联分析的计算效率和统计学效力

方法1: VCFtools (Danecek et al., 2011)过滤SNP命令:

vcftools --vcf [vcf] [--plink] --max-missing 0.8-- maf

0.05 [--remove-indels] --out [outfile], 其中vcf是变异鉴定结果vcf文件, outfile是过滤后最终变异鉴定结果。

方法2: PLINK (Purcell et al., 2007)过滤SNP命令:

将vcf格式文件转换为PLINK格式: vcftools --vcf [vcf] --plink --out [outfile]

plink --file [outfile] --noweb --maf 0.05 --geno 0.1 [--mind 0.2] --out [out], 其中outfile是plink格式变异鉴定结果文件, out是最终变异鉴定结果。

利用ANNOVAR (Wang et al., 2010)对SNP进行注释, 将SNP按其在基因组上的相对位置分类, 包括基因的上游、5'-UTR区、外显子区、内含子区、3'-UTR和基因的下游等。同时, 注释SNP对蛋白产物的影响, 如同义突变、非同义突变、移码突变及终止密码子提前。

3.3 群体结构、亲缘关系和LD衰减分析

为了降低群体结构和家系亲缘关系对全基因组关联分析的影响, 需要利用SNP信息计算出代表群体结构的Q矩阵和家系亲缘矩阵K矩阵。基于CDS区的SNP, 利用PHYLIP (http://evolution.genetics.Washington.edu/phylip.htm)、MEGA (Tamura et al., 2013)或SNPphylo (Lee et al., 2014)构建进化树来展示群体结构。将进化上亲缘关系近的样本分为一个单元(即亚群), 后续分析按照不同亚群进行。

PCA分析确定主成分来控制群体结构, 对群体结构进行检验和矫正。主成分得分信息还用于关联分析的混合线性模型中, 以减少群体结构带来的假阳性关联。

3.3.1 PCA分析

(1) 利用EIGENSOFT (Price et al., 2006)软件中的smartpca进行PCA分析(图2):

利用VCFtools将vcf文件转换为.ped和.map文件。

vcftools --vcf [vcf] –plink --out [name], 其中vcf是最终变异鉴定结果vcf文件, 输出文件是PLINK格式的变异鉴定结果。

plink --file [name] --indep-pairwise 100 10 0.5 -- out [name], 输入PLINK格式的变异鉴定结果, 输出不连锁位点文件。

plink --file [name] --extract [name].prune.in -- recode --out [name].prunein, 其中--file是PLINK格式的变异鉴定结果, --extract是上一条命令获得的不连锁位点文件, --out是ped格式不连锁位点输出文件。

(2) 运行smartpca

EIG-master/bin/convertf -p parameter1

EIG-master/bin/smartpca -p parameter2

PCA结果可视化: EIG-master/bin/ploteig -i [file].evec -c 1:2 -p ??? -x -o PCA12.xtxt

parameter1和parameter2是2个控制文件, 文件中输入参数是ped文件、map文件及上文第1步生成的结果。file.evec是PCA结果文件, 包含样本名称、家系名称、主成分1分值、主成分2分值、主成分3分值等信息。

PCA分析可保留1–10个主成分来完成GWAS关联分析中混杂因素矫正, 一般选取能够解释变异率>5%的主成分来做后续关联分析。基于不同的GWAS研究背景也可通过PC-Finder或者Tracy-Widom统计来确定合适的主成分个数(Abegaz et al., 2019)。

3.3.2 利用ADMIXTURE进行群体结构推断, 了解群体遗传构成

Input file format: *.ped recoding the SNPs to a 1/2 coding

plink --file [name].prunein --recode12 --out [name].prunein.recode12

admixture --cv admixture_prunin.ped 2

admixture --cv admixture_prunin.ped 3

admixture --cv admixture_prunin.ped 4 ……10

输出分组结果文件是admixture_prunin.[i].Q

ADMIXTURE (Alexander et al., 2009)定义的遗传类群, 每列代表一个样本, 不同颜色片段的长度表示该样本基因组中某个祖先所占的比例(图3)。图3显示当祖先群体数量为5时, 各样本的基因组组成情况。ADMIXTURE与PCA的分析结果一致, 即721个水稻材料被分为5组。为了避免群体结构造成的影响, 每个亚群的关联分析需要单独进行(Wang et al., 2020)。

3.3.3 LD衰减分析

LD衰减分析常用软件有PLINK、Haploview (Barrett et al., 2005)和PopLDdecay (Zhang et al., 2019a)。

方法1: PLINK

plink --file [name] --r2 --ld-window 99999 --ld-window-r2 0 --ld-window-kb 1000 --out [fileouts],

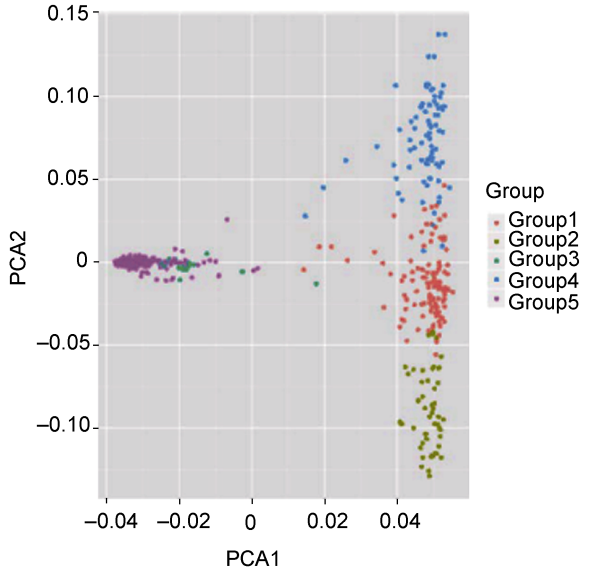

图2 721份水稻材料的主成分分析(PCA)图

Figure 2 The first two components from principal component analysis (PCA) of 721 rice accessions

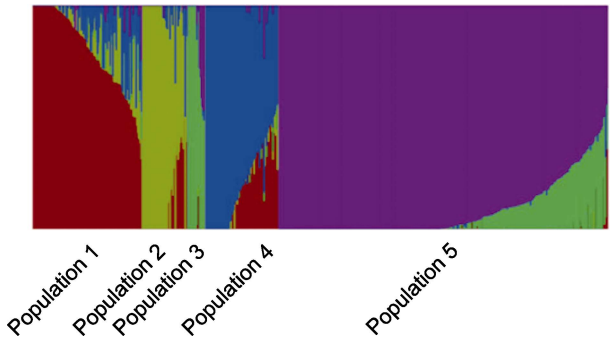

图3 721份水稻材料的群体结构分析

Figure 3 Population structure analyses of 721 rice accessions

其中--file是最终变异鉴定结果plink文件，--out为输出LD的值。

方法2: Haploview

Haploview: windows or linux, same as PLINK based on java

方法3: PopLDdecay

One population: PopLDdecay [options] -InVCF [name].vcf.gz -OutStat [name].LD

Multiple populations: PopLDdecay –InVCF [name].vcf.gz -OutStat [name].LD -SubPop A.list

连锁不平衡参数r2衰减至最大值的一半时对应的距离称为LD半衰距离，实践中常用该值来评估群体中遗传标记连锁与重组情况，确定关联分析所需标记密度以及基于GWAS结果中的显著信号在基因组候选基因的选取范围。

3.4 GWAS具体步骤

GWAS以群体结构和亲缘关系矩阵作为协变量，通过混合线性模型将SNP与表型关联起来。现以EMMAX和GAPIT为例来说明关联分析的具体步骤。

3.4.1 利用EMMAX进行GWAS命令

(1) Preparing input genotype files

plink --file [name] --recode12 --transpose --out [name].emmax –noweb

results: name.emmax.tped and name. emmax.tfam

(2) Preparing input phenotype files

表型数据至少有3列，分别为家族ID、个体ID、表型I值和表型II值。每列之间用tab键隔开。以水稻抽穗期数据为例，格式如下:

家族ID	个体ID	抽穗期表型值
Sample_1707	Sample_1707	127
Sample_1708	Sample_1708	133
Sample_1709	Sample_1709	NA
Sample_1710	Sample_1710	130
Sample_1711	Sample_1711	131
Sample_1712	Sample_1712	123
Sample_1713	Sample_1713	139

(3) Creating Marker-Based Kinship Matrix

generate [tped_prefix].aIBS.kinf: emmax-kin-intel64 -v -s -d 10 [name].emmax

generate [tped_prefix].aBN.kinf: emmax-kin-intel64 -v -d 10 [name].emmax

IBS和BN这两种计算亲缘关系的方法可任选其一。

(4) Run EMMAX association

方法1: Adjust for covariates

对于群体结构强的群体，以PCA分析矩阵作为协变量来矫正群体结构对GWAS结果的影响。

emmax-intel64 -v -d 10 -t [name].emmax -p phenofile -k [name].emmax.a[IBS,BN].kinf -c [name].evec –o [outfile]，其中，-t是基因型输入文件，-p是表型文件，-k是亲缘关系矩阵，-c是PCA分析结果。

方法2: No covariates

对于群体结构弱的群体，无须PCA作协变量来

矫正群体结构。

```
emmax-intel64 -v -d 10 -t [name].emmax -p pheno-file -k [name].emmax.a[IBS,BN].kinf -o [outfile]
results: [out_prefix].reml and [out_prefix].ps
```

3.4.2 利用GAPIT进行GWAS命令

```
library(multtest)
library(gplots)
library(LDheatmap)
library(genetics)
library(ape)
library(EMMREML)
library(compiler)
library("scatterplot3d")
source("http://zzlab.net/GAPIT/gapit_functions.txt")
source("http://zzlab.net/GAPIT/emma.txt")
(1) Set working directory and import data
myY <- read.table("[mdp_traits.txt]", head = TRUE)
myG <- read.table("[mdp_genotype_test. hmp. Txt]", head = FALSE)
(2) Run GAPIT with CMLM
myGAPIT <- GAPIT(
Y=myY, #表型文件
G=myG, #基因型文件
PCA.total=3, #前3个主成分进行群体结构矫正。
model=“CMLM”, #选择所用的关联分析模型，可从"MLM"、 "CMLM"、 "MLMM"、 "SUPER"和"FarmCPU"等模型中选择一个或多个。
kinship.cluster=c("average", "complete", "ward"),
kinship.group=c("Mean", "Max"),
group.from=200,
group.to=1000000,
group.by=10
)
```

利用GAPIT或者EMMAX完成GWAS后会生成一个文档，文档中至少包含3列，SNP位置(染色体编号及其在染色体上的位置)及每一个SNP对应的*P*值(即与表型相关的程度，*P*值越小与表型越相关)。

3.5 GWAS结果筛选

GWAS的结果通常以曼哈顿图和QQ图来展示。曼哈顿图显示每个SNP在关联分析中的显著性水平；QQ图反映关联分析的效果。

曼哈顿图(图4A)中每个点代表一个SNP，x轴代表SNP在基因组上的遗传位置，y轴显示–log10(*P*-value)。显著性阈值以红色水平线(矫正*P*=0.01)和蓝色水平线(矫正*P*=0.05)表示，文中采用Bonferroni矫正法。基因位点在y轴的高度对应该位点与表型的关联程度，关联程度越强，y值越大。受LD影响，基因组上强关联位点周围的SNP也会呈现出关联性由高到低连续变化的信号强度，从而在*P*值小的地方出现尖峰。峰值点附近这种信号变化符合群体遗传重组模式，可能是一个可靠位点。

通过全基因组关联分析既可以定位到某些已知的重要基因，也能够发现新的未知位点。采用EMMAX模型，以PCA的前两个主成分(解释率>50%)为协变量对721份水稻的抽穗期进行GWAS研究。结果发现了一些位于已知基因附近的显著位点(图4A)。例如，6号染色体上的*Hd3a*，7号染色体上的*DTH7*。同时，在北京1号染色体头部(Chr. 1: 1.35–1.52 M)以及4号染色体尾部(Chr. 4: 27.8–28.5 M)等鉴定出抽穗期性状相关新位点(Li et al., 2020)。

QQ图通过比较每个SNP期望*P*值与观测*P*值的差异来对GWAS结果进行质控。GWAS假设只有一小部分SNP与表型相关，因此大部分SNP期望*P*值与观测*P*值应该重合。QQ图(图4B)在$P<10^{-3}$时，群体开始显示出受到选择，SNP不再随机分布，说明我们研究的水稻抽穗期与基因型之间存在显著相关的选择作用。

基于LD衰减的距离和显著关联SNP，通常有2种方式来确定候选区间。(1) 将显著关联SNP在N kb以内的位置确认为相关区间；(2) N kb以内的位置相近SNP定义为一个cluster。其中，N是LD衰减距离。例如，深入分析水稻6号染色体上与抽穗期相关的一个尖峰，将其定位在*Hd3a*附近，估计候选区间大概在2.68–4.62 Mb (图4C)。

GWAS鉴定出候选区间后通过整合多方面信息来精选候选基因。符合以下条件的基因值得进行验证和深入研究。(1) 信号pattern: 关联性从高到低连续变化；(2) 峰值区内的基因功能注释与表型相关；(3) 其它实验功能研究或组学数据支持GWAS结果。

4 小结和展望

近年来，研究人员利用GWAS策略在动植物复杂数量性状研究中鉴定出大量关键位点，但是这些显著关

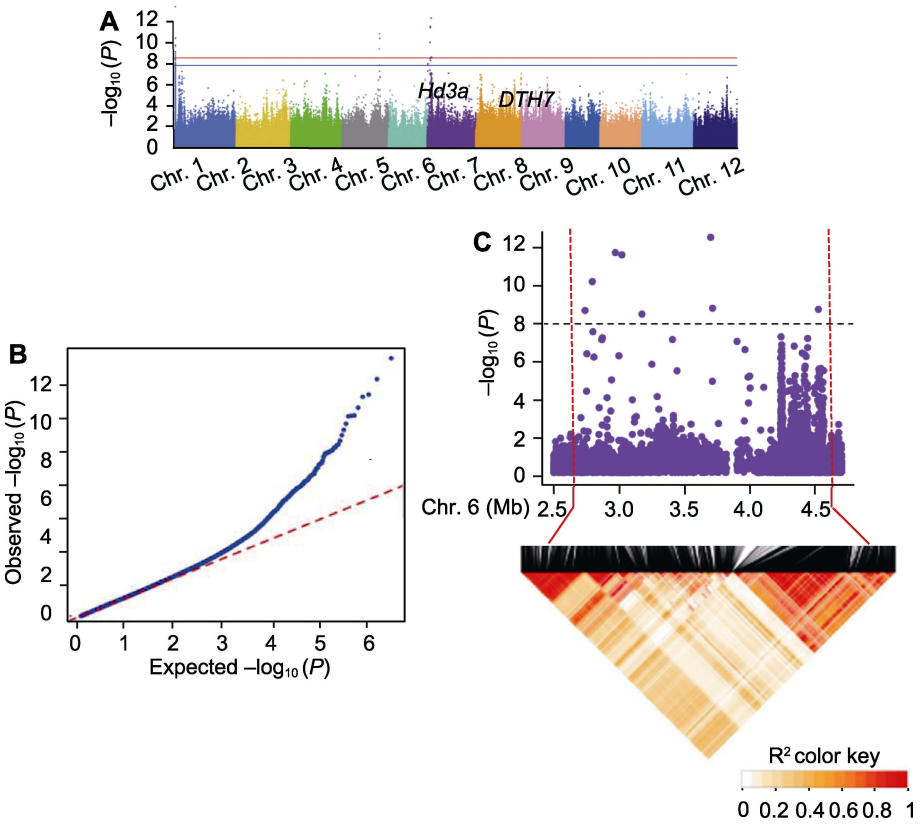

图4 721份水稻材料抽穗期全基因组关联分析(GWAS)结果展示

(A) 抽穗期性状关联分析结果的曼哈顿图; **(B)** QQ图; **(C)** 局部曼哈顿图和6号染色体尖峰附近的LD热图。曼哈顿图中红色虚线标出候选区间，黑色虚线表示显著性阈值–log$_{10}$(*P*)=7.80。

Figure 4 Genome-wide association study (GWAS) results of 721 rice accessions for heading date

(A) Manhattan plots of GWAS results for heading date; **(B)** QQ plot; **(C)** Local manhattan plots and LD heatmap around the peak on chromosome 6. Candidate region was labelled by red dotted line while the black dotted line indicated threshold –log$_{10}$ (*P*)=7.80.

联位点仅能解释部分表型变异，“缺失遗传力”问题依然是当前数量遗传学研究的难点。此外，GWAS存在统计效力有限以及无法鉴定一个基因内多个有功能的等位基因和群体中的微效基因等问题(De et al., 2014; Zhou and Huang, 2019)。多组学数据的积累为弥补GWAS的不足提供了机会。基于基因表达的GWAS (Liu et al., 2015; Jin et al., 2016; Kremling et al., 2018; Zhu et al., 2018)、基于代谢组学的GWAS (Wen et al., 2014; Tieman et al., 2017; Wu et al., 2018; Chen et al., 2020)和基于蛋白质组学的GWAS (Fabres et al., 2017)等是GWAS未来的发展方向。

参考文献

Abegaz F, Chaichoompu K, Génin E, Fardo DW, König IR, Mahachie John JM, Van Steen K (2019). Principals about principal components in statistical genetics. *Brief Bioinform* **20**, 2200–2216.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664.

Alqudah AM, Koppolu R, Wolde GM, Graner A, Schnurbusch T (2016). The genetic architecture of barley plant stature. *Front Genet* **7**, 117.

Alqudah AM, Sallam A, Baenziger PS, Börner A (2020). GWAS: fast-forwarding gene identification and characterization in temperate cereals: lessons from barley—a review. *J Adv Res* **22**, 119–135.

Aulchenko YS, de Koning DJ, Haley C (2007). Genomewide rapid association using mixed model and regression, a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585.

Bai XF, Zhao H, Huang Y, Xie WB, Han ZM, Zhang B, Guo ZL, Yang L, Dong HJ, Xue WY, Li GW, Hu G, Hu Y,

Xing YZ (2016). Genome-wide association analysis reveals different genetic control in panicle architecture between *Indica* and *Japonica* rice. *Plant Genome* **9**, 1–10.

Barrett JC, Fry B, Maller J, Daly MJ (2005). Haploview, analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

Bhatta M, Baenziger PS, Waters BM, Poudel R, Belamkar V, Poland J, Morgounov A (2018). Genome-wide association study reveals novel genomic regions associated with 10 grain minerals in synthetic hexaploid wheat. *Int J Mol Sci* **19**, 3237.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic, a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Bradbury PJ, Zhang ZW, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635.

Casale FP, Rakitsch B, Lippert C, Stegle O (2015). Efficient set tests for the genetic analysis of correlated traits. *Nat Methods* **12**, 755–758.

Chen J, Hu X, Shi TT, Yin HR, Sun DF, Hao YF, Xia XC, Luo J, Fernie AR, He ZH, Chen W (2020). Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. *Plant Biotechnol J* **18**, 1722–1735.

Chen PL, Shen ZK, Ming LC, Li YB, Dan WH, Lou GM, Peng B, Wu B, Li YH, Zhao D, Gao GJ, Zhang QL, Xiao JH, Li XH, Wang GW, He YQ (2018). Genetic basis of variation in rice seed storage protein (Albumin, Globulin, Prolamin, and Glutelin) content revealed by genome-wide association analysis. *Front Plant Sci* **9**, 612.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

De R, Bush WS, Moore JH (2014). Bioinformatics challenges in genome-wide association studies (GWAS). In: Trent R, ed. Clinical Bioinformatics. Methods in Molecular Biology (Methods and Protocols), Vol. 1168. New York: Humana Press. pp. 63–81.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498.

Dong HJ, Zhao H, Li SL, Han ZM, Hu G, Liu C, Yang GY, Wang GW, Xie WB, Xing YZ (2018). Genome-wide association studies reveal that members of bHLH subfamily 16 share a conserved function in regulating flag leaf angle in rice (*Oryza sativa*). *PLoS Genet* **14**, e1007323.

Dong XK, Gao YQ, Chen W, Wang WS, Gong L, Liu XQ, Luo J (2015). Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. *Mol Plant* **8**, 111–121.

Du XM, Huang G, He SP, Yang ZE, Sun GF, Ma XF, Li N, Zhang XY, Sun JL, Liu M, Jia YH, Pan ZE, Gong WF, Liu ZH, Zhu HQ, Ma L, Liu FY, Yang DG, Wang F, Fan W, Gong Q, Peng Z, Wang LR, Wang XY, Xu SJ, Shang HH, Lu CR, Zheng HK, Huang SW, Lin T, Zhu YX, Li FG (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat Genet* **50**, 796–802.

Duan NB, Bai Y, Sun HH, Wang N, Ma YM, Li MJ, Wang X, Jiao C, Legall N, Mao LY, Wan SB, Wang K, He TM, Feng SQ, Zhang ZY, Mao ZQ, Shen X, Chen XL, Jiang YM, Wu SJ, Yin CM, Ge SF, Yang L, Jiang SH, Xu HF, Liu JX, Wang DY, Qu CZ, Wang YC, Zuo WF, Xiang L, Liu C, Zhang DY, Gao Y, Xu YM, Xu KN, Chao T, Fazio G, Shu HR, Zhong GY, Cheng LL, Fei ZJ, Chen XS (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat Commun* **8**, 249.

Fabres PJ, Collins C, Cavagnaro TR, Rodríguez López CM (2017). A concise review on multi-omics data integration for terroir analysis in *Vitis vinifera*. *Front Plant Sci* **8**, 1065.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Fang C, Ma YM, Wu SW, Liu Z, Wang Z, Yang R, Hu GH, Zhou ZK, Yu H, Zhang M, Pan Y, Zhou GA, Ren HX, Du WG, Yan HR, Wang YP, Han DZ, Shen YT, Liu SL, Liu TF, Zhang JX, Qin H, Yuan J, Yuan XH, Kong FJ, Liu BH, Li JY, Zhang ZW, Wang GD, Zhu BG, Tian ZX (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol* **18**, 161.

Furlotte NA, Eskin E (2015). Efficient multiple-trait association and estimation of genetic correlation using the ma-

trix-variate linear mixed model. *Genetics* **200**, 59–68.

Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJM, Open Regulatory Annotation C (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* **36**, D107–D113.

Gumpinger AC, Roqueiro D, Grimm DG, Borgwardt KM (2018). Methods and tools in genome-wide association studies. *Methods Mol Biol* **1819**, 93–136.

Guo SG, Zhao SJ, Sun HH, Wang X, Wu S, Lin T, Ren Y, Gao L, Deng Y, Zhang J, Lu XQ, Zhang HY, Shang JL, Gong GY, Wen CL, He N, Tian SW, Li MY, Liu JP, Wang YP, Zhu YC, Jarret R, Levi A, Zhang XP, Huang SW, Fei ZJ, Liu WG, Xu Y (2019). Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat Genet* **51**, 1616–1623.

Huang M, Liu XL, Zhou Y, Summers RM, Zhang ZW (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* **8**, giy154.

Huang XH, Yang SH, Gong JY, Zhao Y, Feng Q, Gao H, Li WJ, Zhan QL, Cheng BY, Xia JH, Chen N, Hao ZN, Liu KY, Zhu CR, Huang T, Zhao Q, Zhang L, Fan DL, Zhou CC, Lu YQ, Weng QJ, Wang ZX, Li JY, Han B (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* **6**, 6258.

Huang XH, Yang SH, Gong JY, Zhao Q, Feng Q, Zhan QL, Zhao Y, Li WJ, Cheng BY, Xia JH, Chen N, Huang T, Zhang L, Fan DL, Chen JY, Zhou CC, Lu YQ, Weng QJ, Han B (2016). Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629–633.

Huang XH, Zhao Y, Wei XH, Li CY, Wang AH, Zhao Q, Li WJ, Guo YL, Deng LW, Zhu CR, Fan DL, Lu YQ, Weng QJ, Liu KY, Zhou TY, Jing YF, Si LZ, Dong GJ, Huang T, Lu TT, Feng Q, Qian Q, Li JY, Han B (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* **44**, 32–39.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* **9**, 1322–1332.

Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP, Ostevik KL, Moyers BT, Yakimowski S, Masalia RR, Gao LX, Ćalić I, Bowers JE, Kane NC, Swanevelder DZH, Kubach T, Muños S, Langlade NB, Burke JM, Rieseberg LH (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants* **5**, 54–62.

Jiang D, Wang MY (2018). Recent developments in statistical methods for GWAS and high-throughput sequencing association studies of complex traits. *Biostatist Epidemiol* **2**, 132–159.

Jin ML, Liu HJ, He C, Fu JJ, Xiao YJ, Wang YB, Xie WB, Wang GY, Yan JB (2016). Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci Rep* **6**, 18936.

Kaler AS, Purcell LC (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics* **20**, 618.

Kaler AS, Ray JD, Schapaugh WT, King CA, Purcell LC (2017). Genome-wide association mapping of canopy wilting in diverse soybean genotypes. *Theor Appl Genet* **130**, 2203–2217.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348–354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723.

Kidane YG, Gesesse CA, Hailemariam BN, Desta EA, Mengistu DK, Fadda C, Pè ME, Dell'Acqua M (2019). A large nested association mapping population for breeding and quantitative trait locus mapping in *Ethiopian durum* wheat. *Plant Biotechnol J* **17**, 1380–1393.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**, 1066–1071.

Kremling KAG, Chen SY, Su MH, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523.

Kremling KAG, Diepenbrock CH, Gore MA, Buckler ES,

Bandillo NB (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3* **9**, 3023–3033.

Kumar J, Pratap A, Solanki RK, Gupta DS, Goyal A, Chaturvedi SK, Nadarajan N, Kumar S (2012). Genomic resources for improving food legume crops. *J Agric Sci* **150**, 289–318.

Kusmec A, Srinivasan S, Nettleton D, Schnable PS (2017). Distinct genetic architectures for phenotype means and plasticities in *Zea mays*. *Nat Plants* **3**, 715–723.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Lee TH, Guo H, Wang XY, Kim C, Paterson AH (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162.

Li CH, Sun BC, Li YX, Liu C, Wu X, Zhang DF, Shi YS, Song YC, Buckler ES, Zhang ZW, Wang TY, Li Y (2016). Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations. *BMC Genomics* **17**, 894.

Li CH, Wu X, Li YX, Shi YS, Song YC, Zhang DF, Li Y, Wang TY (2019). Genetic architecture of phenotypic means and plasticities of kernel size and weight in maize. *Theor Appl Genet* **132**, 3309–3320.

Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Li M, Liu XL, Bradbury P, Yu JM, Zhang YM, Todhunter RJ, Buckler ES, Zhang ZW (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol* **12**, 73.

Li XX, Chen Z, Zhang GM, Lu HW, Qin P, Qi M, Yu Y, Jiao BK, Zhao XF, Gao Q, Wang H, Wu YY, Ma JT, Zhang LY, Wang YL, Deng LW, Yao SG, Cheng ZK, Yu DQ, Zhu LH, Xue YB, Chu CC, Li AH, Li SG, Liang CZ (2020). Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Sci China Life Sci* **63**, 1688–1702.

Li YL, Ruperao P, Batley J, Edwards D, Davidson J, Hobson K, Sutton T (2017). Genome analysis identified novel candidate genes for *Ascochyta* blight resistance in chickpea using whole genome re-sequencing data. *Front Plant Sci* **8**, 359.

Li YL, Ruperao P, Batley J, Edwards D, Khan T, Colmer TD, Pang JY, Siddique KHM, Sutton T (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front Plant Sci* **9**, 190.

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

Listgarten J, Lippert C, Heckerman D (2013). FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat Genet* **45**, 470–471.

Liu HJ, Wang XQ, Warburton ML, Wen WW, Jin ML, Deng M, Liu J, Tong H, Pan QC, Yang XH, Yan JB (2015). Genomic, transcriptomic, and phenomic variation reveals the complex adaptation of modern maize breeding. *Mol Plant* **8**, 871–884.

Liu HJ, Yan JB (2019). Crop genome-wide association study: a harvest of biological relevance. *Plant J* **97**, 8–18.

Liu XL, Huang M, Fan B, Buckler E, Zhang ZW (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet* **12**, e1005767.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, Price AL (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284–290.

Lozada D, Godoy JV, Murray TD, Ward BP, Carter AH (2019). Genetic dissection of snow mold tolerance in US Pacific Northwest winter wheat through genome-wide association study and genomic selection. *Front Plant Sci* **10**, 1337.

Ma HL, Li GL, W€urschum T, Zhang Y, Zheng DB, Yang XH, Li JS, Liu WX, Yan JB, Chen SJ (2018a). Genome-wide association study of haploid male fertility in maize (*Zea mays* L.). *Front Plant Sci* **9**, 974.

Ma XF, Wang ZY, Li W, Zhang YZ, Zhou XJ, Liu YG, Ren ZY, Pei XY, Zhou KH, Zhang WS, He KL, Zhang F, Liu JF, Ma WY, Xiao GH, Yang DG (2019a). Resequencing core accessions of a pedigree identifies derivation of genomic segments and key agronomic trait loci during cotton improvement. *Plant Biotechnol J* **17**, 762–775.

Ma XS, Feng FJ, Zhang Y, Elesawi IE, Xu K, Li TF, Mei HW, Liu HY, Gao NN, Chen CL, Luo LJ, Yu SW (2019b). A novel rice grain size gene *OsSNB* was identified by genome-wide association study in natural population.

PLoS Genet **15**, e1008191.

Ma ZY, He SP, Wang XF, Sun JL, Zhang Y, Zhang GY, Wu LQ, Li ZK, Liu ZH, Sun GF, Yan YY, Jia YH, Yang J, Pan ZE, Gu QS, Li XY, Sun ZW, Dai PH, Liu ZW, Gong WF, Wu JH, Wang M, Liu HW, Feng KY, Ke H, Wang JD, Lan HY, Wang GN, Peng J, Wang N, Wang LR, Pang BY, Peng Z, Li RQ, Tian SL, Du XM (2018b). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat Genet* **50**, 803–813.

Martin ER, Tunc I, Liu Z, Slifer SH, Beecham AH, Beecham GW (2018). Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet Epidemiol* **42**, 214–229.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303.

Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo GG, Xu DD, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao YS, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat Genet* **51**, 319–326.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ES (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**, 2194–2202.

Peng YC, Liu HB, Chen J, Shi TT, Zhang C, Sun DF, He ZH, Hao YF, Chen W (2018). Genome-wide association studies of free amino acid levels by six multi-locus models in bread wheat. *Front Plant Sci* **9**, 1196.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

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

Raj A, Stephens M, Pritchard JK (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589.

Sallam A, Martsch R (2015). Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (*Vicia faba* L.). *Genetica* **143**, 501–514.

Song CX, Li W, Pei XY, Liu YG, Ren ZY, He KL, Zhang F, Sun K, Zhou XJ, Ma XF, Yang DG (2019). Dissection of the genetic variation and candidate genes of lint percentage by a genome-wide association study in upland cotton. *Theor Appl Genet* **132**, 1991–2002.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**, 1166–1170.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019). Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.

Tang H, Peng J, Wang P, Risch NJ (2005). Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* **28**, 289–301.

Tang Y, Liu XL, Wang JB, Li M, Wang QS, Tian F, Su ZB, Pan YC, Liu D, Lipka AE, Buckler ES, Zhang ZW (2016). GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* **9**, 1–9.

Thudi M, Khan AW, Kumar V, Gaur PM, Katta K, Garg V, Roorkiwal M, Samineni S, Varshney RK (2016). Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (*Cicer arietinum* L.). *BMC Plant Biol* **16**, 10.

Tieman D, Zhu GT, Resende MFR Jr, Lin T, Nguyen C, Bies D, Rambla JL, Beltran KSO, Taylor M, Zhang B, Ikeda K, Liu ZY, Fisher J, Zemach I, Monforte A, Zamir D, Granell A, Kirst M, Huang SW, Klee H (2017). A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394.

Tong W, Kim TS, Park YJ (2016). Rice chloroplast genome variation architecture and phylogenetic dissection in diverse *Oryza* species assessed by whole-genome resequencing. *Rice* **9**, 57.

Varshney RK, Thudi M, Roorkiwal M, He WM, Upadhyaya

HD, Yang W, Bajaj P, Cubry P, Rathore A, Jian JB, Doddamani D, Khan AW, Garg V, Chitikineni A, Xu DW, Gaur PM, Singh NP, Chaturvedi SK, Nadigatla GVPR, Krishnamurthy L, Dixit GP, Fikre A, Kimurto PK, Sreeman SM, Bharadwaj C, Tripathi S, Wang J, Lee SH, Edwards D, Polavarapu KKB, Penmetsa RV, Crossa J, Nguyen HT, Siddique KHM, Colmer TD, Sutton T, von Wettberg E, Vigouroux Y, Xu X, Liu X (2019). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat Genet* **51**, 857–864.

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

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017). 10 years of GWAS discovery, biology, function, and translation. *Am J Hum Genet* **101**, 5–22.

Wang HR, Xu X, Vieira FG, Xiao YH, Li ZK, Wang J, Nielsen R, Chu CC (2016). The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. *Mol Plant* **9**, 975–985.

Wang K, Li MY, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164.

Wang MH, Cordell HJ, Van Steen K (2019). Statistical methods for genome-wide association studies. *Semin Cancer Biol* **55**, 53–60.

Wang MY, Xu SZ (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity* **123**, 287–306.

Wang Q, Tang JL, Han B, Huang XH (2020). Advances in genome-wide association studies of complex traits in rice. *Theor Appl Genet* **133**, 1415–1425 .

Wang QS, Tian F, Pan YC, Buckler ES, Zhang ZW (2014). A SUPER powerful method for genome wide association study. *PLoS One* **9**, e107684.

Wang QX, Xie WB, Xing HK, Yan J, Meng XZ, Li XL, Fu XK, Xu JY, Lian XM, Yu SB, Xing YZ, Wang GW (2015). Genetic architecture of natural variation in rice chlorophyll content revealed by a genome-wide association study. *Mol Plant* **8**, 946–957.

Wei W, Mesquita ACO, de A. Figueiró A, Wu X, Manjunatha S, Wickland DP, Hudson ME, Juliatti FC, Clough SJ (2017). Genome-wide association mapping of resistance to a Brazilian isolate of *Sclerotinia sclerotiorum* in soybean genotypes mostly from Brazil. *BMC Genomics* **18**, 849.

Wen WW, Li D, Li X, Gao YQ, Li WQ, Li HH, Liu J, Liu HJ, Chen W, Luo J, Yan JB (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat Commun* **5**, 3438.

Wu J, Wang LF, Fu JJ, Chen JB, Wei SH, Zhang SL, Zhang J, Tang YS, Chen ML, Zhu JF, Lei L, Geng QH, Liu CL, Wu L, Li XM, Wang XL, Wang Q, Wang ZL, Xing SL, Zhang HK, Blair MW, Wang SM (2020). Resequencing of 683 common bean genotypes identifies yield component trait associations across a north-south cline. *Nat Genet* **52**,118–125.

Wu S, Tohge T, Cuadros-Inostroza A, Tong H, Tenenboim H, Kooke R, Méret M, Keurentjes JB, Nikoloski Z, Fernie AR, Willmitzer L, Brotman Y (2018). Mapping the *Arabidopsis* metabolic landscape by untargeted metabolomics at different environmental conditions. *Mol Plant* **11**, 118–134.

Xiao YJ, Liu HJ, Wu LJ, Warburton M, Yan JB (2017). Genome-wide association studies in maize: praise and stargaze. *Mol Plant* **10**, 359–374.

Xie WB, Wang GW, Yuan M, Yao W, Lyu K, Zhao H, Yang M, Li PB, Zhang X, Yuan J, Wang QX, Liu F, Dong HX, Zhang LJ, Li XL, Meng XZ, Zhang W, Xiong LZ, He YQ, Wang SP, Yu SB, Xu CG, Luo J, Li XH, Xiao JH, Lian XM, Zhang QF (2015). Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci USA* **112**, E5411–E5419.

Yang N, Lu YL, Yang XH, Huang J, Zhou Y, Ali F, Wen WW, Liu J, Li JS, Yan JB (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet* **10**, e1004573.

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, Yamasaki M, Yoshida S, Kitano H, Hirano K, Matsuoka M (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat Genet* **48**, 927–934.

Yu JM, Buckler ES (2006). Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* **17**, 155–160.

Yu JM, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203–208.

Zhang C, Dong SS, Xu JY, He WM, Yang TL (2019a). PopLDdecay, a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788.

Zhang TF, Wu TT, Wang LW, Jiang BJ, Zhen CX, Yuan S, Hou WS, Wu CX, Han T, Sun S (2019b). A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. *Int J Mol Sci* **20**, 5915.

Zhang X, Zhang H, Li LJ, Lan H, Ren ZY, Liu D, Wu L, Liu HL, Jaqueth J, Pan GT, Gao SB (2016). Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. *BMC Genomics* **17**, 697.

Zhang YM, Jia ZY, Dunwell JM (2019c). Editorial: the applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front Plant Sci* **10**, 100.

Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu JM, Arnett DK, Ordovas JM, Buckler ES (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355–360.

Zhou X, Stephens M (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407–409.

Zhou XY, Huang XH (2019). Genome-wide association studies in rice: how to solve the low power problems? *Mol Plant* **12**, 10–12.

Zhou ZK, Jiang Y, Wang Z, Gou ZH, Lyu J, Li WY, Yu YJ, Shu LP, Zhao YJ, Ma YM, Fang C, Shen YT, Liu TF, Li CC, Li Q, Wu M, Wang M, Wu YS, Dong Y, Wan WT, Wang X, Ding ZL, Gao YD, Xiang H, Zhu BG, Lee SH, Wang W, Tian ZX (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**, 408–414.

Zhu GT, Wang SC, Huang ZJ, Zhang SB, Liao QG, Zhang CZ, Lin T, Qin M, Peng M, Yang CK, Cao X, Han X, Wang XX, van der Knaap E, Zhang ZH, Cui X, Klee H, Fernie AR, Luo J, Huang SW (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**, 249–261.

An Overview of Genome-wide Association Studies in Plants

Yuhui Zhao[1], Xiuxiu Li[1, 2], Zhuo Chen[1, 2], Hongwei Lu[1, 2], Yucheng Liu[1, 2]
Zhifang Zhang[1, 2], Chengzhi Liang[1, 2*]

[1]*Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract Genome-wide association study (GWAS) is a general approach for unraveling genetic variations associated with complex traits in both animals and plants. The development of high-throughput genotyping has greatly boosted the development and application of GWAS. GWAS is not only used to identify genes/loci contributing to specific traits from diversenatural populations with high-resolution genome-wide markers, it also systematically reveals the genetic architecture underlying complex traits. During recent years, GWAS has successfully detected a large number of QTLs and candidate genes associated with various traits in plants including *Arabidopsis*, rice, wheat, soybean and maize. All these findings provided candidate genes controlling the traits and theoretical basis for breeding of high-yield and high-quality varieties. Here we review the methods, the factors affecting the power, and a data analysis pipeline of GWAS to provide reference for relevant research.

Key words mixed linear model, genome-wide association study (GWAS), bioinformatics

Zhao YH, Li XX, Chen Z, Lu HW, Liu YC, Zhang ZF, Liang CZ (2020). An overview of genome-wide association studies in plants. *Chin Bull Bot* **55**, 715–732.

---

* Author for correspondence. E-mail: cliang@genetics.ac.cn

(责任编辑: 朱亚娜)