林木全基因组关联分析研究进展与展望

李鹏^{1,2,3}, 卢文杰^{1,2,3}, 肖亮^{1,2,3}, 权明洋^{1,2,3}, 李连政^{1,2,3}, 杜庆章^{1,2,3}, 张德强^{1,2,3*}

1. 北京林业大学林木分子设计育种高精尖创新中心, 北京, 100083;

2. 北京林业大学, 林木育种国家工程实验室, 北京, 100083;

3. 北京林业大学生物科学与技术学院, 林木花卉遗传育种教育部重点实验室, 北京, 100083

* 联系人, E-mail: DeqiangZhang@bjfu.edu.cn

收稿日期: 2019-06-11; 接受日期: 2019-11-22; 网络版发表日期: 2020-02-11

中央高校基本科研业务费(批准号: 2015ZCQ-SW-01)、北京市科技新星计划课题(批准号: Z181100006218024)和北京林业大学2019年度一流学科建设项目(批准号: 2019XKJS0308)资助

摘要 随着新一代基因组测序技术的快速发展,全基因组关联分析(genome wide association study, GWAS)在揭示林木复杂性状的数量遗传变异规律、解析关键基因的遗传调控机制及推动林木分子辅助育种等方面展示出前所未有的应用前景. 本文首先综述了GWAS的核心理论、研究方法及其在木材性状和适应性遗传基础研究中的研究进展. 随后, 针对林木数量性状遗传研究中普遍存在的“丢失遗传力”(missing heritability)问题, 本文从高通量表型组学平台的组建, 多种遗传标记的联合利用, 多组学数据的系统剖析以及加性、显性与上位性关联模型的开发等方面提出了未来GWAS的发展对策. 最后, 结合当前林木遗传改良的实践需求, 展望了GWAS策略在林木分子育种领域的广阔应用前景.

关键词 全基因组关联分析, 系统遗传学, 遗传结构, 丢失遗传力, 分子育种

林木世代周期长、遗传杂合度高且生物学研究基础薄弱, 极大地限制了对林木目标性状的遗传改良进程^[1,2]. 研究表明, 林木目标改良性状多为数量性状, 受全基因组水平多基因位点联合控制, 其遗传变异效应可分为基因的加性和显性效应、基因-基因的上位性效应以及基因-环境的互作效应等. 先前基于家系群体的数量性状位点(quantitative trait locus, QTL)定位分析方法的应用, 在林木复杂性状遗传解析方面取得了显著进展^[3-6]. 但由于林木家系作图大多基于F₁, F₂或BC₁等低世代杂交群体, 遗传变异丰富度较低和染色体重组事件有限, 导致QTL作图分辨率低且其遗传效应往

往被高估^[7]. 基于自然群体或种质资源群体的关联遗传学(association genetics)策略的提出, 为解析数量性状的遗传基础提供了新的途径^[1]. 关联分析又称连锁不平衡(linkage disequilibrium, LD)作图, 可以直接利用自然群体中丰富的表型和基因组变异来确定控制数量性状的目标基因^[8]. 特别是多年生林木群体在适应复杂多变的自然环境过程中产生了丰富的表型变异和DNA序列变异^[9], 是开展关联作图研究的理想材料. Thumma等人^[10]首次在桉树中开展了林木关联作图研究, 确定了肉桂酰辅酶A还原酶(cinnamoyl-CoA reductase, CCR)基因内单核苷酸多态性(single nucleotide

引用格式: 李鹏, 卢文杰, 肖亮, 等. 林木全基因组关联分析研究进展与展望. 中国科学: 生命科学, 2020, 50: 144–153

Li P, Lu W J, Xiao L, et al. Progress and prospect of genome wide association study (GWAS) in forest trees (in Chinese). Sci Sin Vitae, 2020, 50: 144–153, doi: 10.1360/SSV-2019-0129

polymorphism, SNP)位点对微纤丝角(microfibril angle)具有显著调控作用, 为木材硬度的遗传改良提供了高效变异位点.

根据遗传变异来源, 关联作图可分为基于候选基因(candidate gene-based association study)和基于全基因组(genome-wide association study, GWAS)的两种分析方法^[11]. 随着第二代测序(next generation sequencing, NGS)技术的迅速发展^[12~15], 基于全基因组重测序(whole genome resequencing)和简化基因组测序(reduced-representation genome sequencing, RRGs)技术确定了海量的覆盖全基因组水平的SNP变异位点, 显著推动了林木GWAS策略的广泛应用^[9,16]. 此外, 第二代测序技术也促进了全基因组范围内插入缺失(insertion/deletion, InDel)、拷贝数变异(copy number variation, CNV)等标记位点的开发与应用, 弥补了仅考虑SNP导致的数量性状“丢失遗传力”^[17~20]. 尤其是近几年多组学技术的飞速发展, 为林木复杂性状的研究带来了新的机遇. 如何将多组学数据和新的遗传学策略融合至GWAS分析, 更加准确地鉴定关键基因位点, 并进一步阐明数量性状遗传调控网络, 成为当前目标性状遗传改良面临的科学难题.

本文结合团队在该领域的研究进展, 通过综述GWAS策略在林木复杂性状研究中的应用及最新研究成果, 提出了林木GWAS未来的发展策略, 展现了林木分子设计育种的广阔应用前景, 以期对林木复杂性状遗传改良提供理论参考与技术支持.

1 GWAS概述

全基因组关联分析(GWAS)是一种复杂性状功能定位的正向遗传学分析策略, 可直接利用群体内所有个体全基因组水平上的等位遗传变异和表型变异的相关性分析, 鉴定与目标性状显著连锁的等位变异位点, 进而分析等位基因型对表型的遗传效应, 在植物数量性状遗传结构解析中取得了丰硕成果^[21~23].

LD水平决定了GWAS的解析精度, 是开展GWAS研究的理论基础, 它代表了基因组中两个等位基因非随机组合的程度, 主要产生于群体进化过程中的突变和重组事件, 也会受到群体大小、异交率、选择强度与遗传漂变等其他因素的影响^[24]. 一般来说, 林木群体内LD在500~2000 bp内已衰退至不明显^[1,25]. Slavov

等人^[26]发现, 遗传漂变可导致毛果杨(*Populus trichocarpa*)自然群体的基因组区域LD衰退范围达到3~6 kb, 而且在着丝粒区及基因密度较低的染色体区段, LD衰退距离更远. 该发现拓展了人们对林木LD衰退的认识. 林木进化过程主要受到自然选择的作用, 与受到强烈人工驯化的一年生作物相比, 多年生林木的LD水平相对较低^[1], 因此, 当一个与性状显著关联的遗传标记一旦被发现和验证, 则该标记与功能位点物理距离很近甚至是功能变异基因本身, 上述研究对于未来林木分子辅助育种提供了重要的理论指导.

2 GWAS在林木中的研究进展

目前, 林木GWAS研究主要针对木材品质和产量等重要经济性性状, 以及物候相关的适应性性状, 研究人员已经鉴定了一批与目标性状显著关联的功能基因位点^[20,27~29]. 这些研究为林木目标性状遗传改良提供了重要的标记资源, 为全球气候变化背景下森林资源保护和利用提供了重要的理论与应用价值^[30].

2.1 林木GWAS研究方法模型

由于GWAS在复杂性状遗传基础研究中的优势, 其在动植物遗传学研究中的应用越来越广泛, 新的软件和方法也不断被开发, 功能和算法也各不相同, 但多数是以一般线性模型(general linear model, GLM)和混合线性模型(mixed linear model, MLM)为基础. 其中, 一般线性模型以群体结构矩阵(population structure, Q)或主成分分析(principal component analysis, PCA)矩阵为协变量, 提高计算精度; 而混合线性模型联合利用Q矩阵和亲缘关系(kin-ship, K)矩阵或者联合利用PCA矩阵和K矩阵为协变量开展关联作图分析^[31]. 林木GWAS研究中主要用到的软件及其功能特点列于表1.

TASSEL软件^[40]、PLINK软件^[41]和EMMAX软件^[42]是林木中应用比较广泛的单标记GWAS分析软件. 其中, TASSEL软件功能较为全面, 但计算效率较低; PLINK软件在数据管理和群体结构评价方面较为简单、高效, 但是其关联结果质量相对较差; EMMAX软件用到了运算速度较快的有效混合模型关联分析算法, 在对大样本数据处理上具有更高的效率. 上述软件主要应用于单标记位点的GWAS研究工作, 确定的都是遗传效应较为显著的单标记位点, 而林木数量性状

表 1 林木中常用的GWAS分析软件

Table 1 Common software used for genome-wide association study(GWAS)in forest trees

软件	主要功能	应用树种	性状	参考文献
TASSEL ^[40]	MLM和GLM关联分析, 主成分分析, LD分析, 进化分析, QQ图, 曼哈顿图, LD图	毛白杨(<i>Populus tomentosa</i>)	生长, 木材品质	[18]
		火炬松(<i>Pinus taeda</i>)	生长, 适应性	[32]
		毛果杨(<i>Populus trichocarpa</i>)	木材品质	[33]
		日本柳杉(<i>Cryptomeria japonica</i>)	雄球花, 木材品质	[34]
PLINK ^[41]	复杂数量性状, 关联作图, 数据转化与处理, LD分析, 上位性分析, 单倍型检验	大叶桉(<i>Eucalyptus robusta</i>)	生长	[35]
		山毛榉(<i>Fagus longipetiolata</i>)	山毛榉树皮病	[36]
		美洲黑杨(<i>Populus deltoides</i>)	生长, 木材品质	[28]
EMMAX ^[42]	大样本数据高效关联分析	毛果杨(<i>Populus trichocarpa</i>)	物候, 生长, 代谢	[9,20]
GCTA ^[43]	单性状和两相关性状关联分析, 主成分分析	大叶桉(<i>Eucalyptus robusta</i>)	生长性状	[37]
GAPIT ^[44]	CMLML和Fast-LMM关联分析, BLUP基因组预测关联分析, BLUP基因组预测	火炬松(<i>Pinus taeda</i>)	生长, 抗病, 代谢产物	[38]
GEMMA ^[45]	有效混合模型关联分析, 多数量性状关联	毛果杨(<i>Populus trichocarpa</i>)	形态学, 生理	[39]

受到多基因多位点联合控制. 因此, 开发使用多位点GWAS关联模型进行遗传效应分析, 可以检测更多的联合遗传位点, 进而更清晰地剖析出目标性状的基因组遗传结构^[19]. 如mrMLM软件^[46]可以将SNP效应视为随机效应并利用多位点随机效应的MLM模型进行GWAS分析, 该模型有效提升了目标性状基因组遗传结构的解析精度. 随着测序技术和组学技术的发展, 大量的SNPs, InDels与CNVs标记等被广泛应用于林木复杂性状的遗传效应解析. 因此, 开发适用于不同群体类型、不同标记类型、不同表型层次的GWAS软件, 将会极大地推动林木复杂性状研究进展^[47].

2.2 林木木材品质和生长性状的GWAS研究进展

木材是我国重要的战略资源, 对木材性状遗传机制的研究是林木复杂性状研究的热点领域^[48]. 自Porth等人^[23]首次将GWAS方法用于木材品质性状遗传研究至今, 研究人员先后系统发掘了上百个与木材主要化学组分性状和超微物理结构性状显著关联的功能SNP位点^[9,28,48-50]. 近年来, GWAS策略与多组学技术的结合已广泛应用于数量性状遗传解析. 例如, Zhang等人^[20]在毛果杨自然群体GWAS研究中综合利用全基因组重测序、转录组、代谢组数据, 发现了在木质素生物合成过程中起重要调控作用的*PtHCT2*基因, 构建了其在多个生物学过程中的调控网络, 阐明了该基因响应胁迫与木材形成的协同作用机制. GWAS联合多组

学解析策略不仅可以提高筛选候选基因的准确性, 还可揭示复杂性状形成的分子遗传调控机制^[22]. 在基因组庞大的针叶树种中(>10 Gb), GWAS策略也取得了显著进展. 例如, Uchiyama等人^[34]在日本柳杉(*Cryptomeria japonica*)种质群体开展木材品质性状GWAS研究, 发现了6个显著调控木材品质性状的新基因; Lammara等人^[49]联合利用木质部基因共表达网络和木材品质性状的GWAS, 发掘了白云杉(*Picea glauca*)木材形成调控网络中的关键节点基因*MYB*和*NAC*, 深入解析了木材品质性状的遗传调控网络; Allwright等人^[51]利用GWAS策略, 开展大叶桉群体树高性状的遗传分析, 发现了3个显著的SNP位点, 有效提高了候选基因定位效率. 利用GWAS策略对杨树、松树等树种生长性状进行了候选基因的有效定位. 但是对于这类遗传力较低、遗传调控途径广泛的数量性状^[19,23], 需要在GWAS基础上补充新的策略, 发掘表型变异解析度更高的遗传标记. 例如, Resende等人^[35]通过将多位点组合成新的遗传标记用于GWAS, 开发了区域遗传作图(regional heritability mapping, RHM)策略, 发现了与桉树株高显著关联的4个染色体区段, 每一标记可解释表型变异的10%左右. RHM策略检测了多个SNPs的联合效应, 显著提高了检测效率和表型变异解析度.

2.3 GWAS在林木环境适应性方面的研究进展

林木不同生态分布区气候差异显著, 导致其受到

不同的进化选择压力. 因此, 不同种源的生态型个体往往表现出对其种源独特环境的强烈适应性^[52,53]. 无论是现有森林资源保护, 还是人工林建设中最基本的“适地适树”准则都离不开对于林木适应性方面的研究^[25]. 种内丰富的表型变异正是不同生态型个体基因组适应性的直接表现, 反映了地理和气候因子对性状与遗传的影响, 表明遗传或基因组变异与表型有相似的变异趋势. 物候性状作为林木适应不同环境的重要生物学特征, 对其遗传基础的研究有利于阐明林木的适应性进化规律, 进而预测全球气候变化对林木的影响, 为开展气候变化下森林生物学研究提供了重要理论支撑^[54]. 林木物候特征表现最明显的是北方温带落叶乔木. 例如, 毛果杨天然分布区域内由南至北, 寒冷持续时间差异显著, 在这种梯度变化环境中的个体, 为了使生长季节最大化, 同时避免春季脱水和冻害的影响, 形成了春芽萌发时间的显著分化. McKown等人^[55]利用GWAS策略, 确定了32个与毛果杨发芽时间显著关联的SNP位点, 注释到与细胞分裂、芽萌发、低温诱导蛋白、热激蛋白等相关的16个候选基因, 并且发现, SNP变异频率与寒冷持续时间变化趋于一致. 由种源地环境差异造成的种内遗传变异, 在其他物候性状上也有显著表现. 例如, Evans等人^[9]联合利用GWAS与选择性清除(selective sweep)策略在毛果杨基因组上发掘了与春芽萌发、秋芽形成等性状显著关联的SNP位点, 定位到397个受到显著选择的适应性区段, 研究工作为阐明杨树及其他树种区域适应性提供了重要参考. 因此, GWAS策略不仅可以系统剖析林木适应性的全基因组联合遗传效应, 确定影响物种稳定遗传变异模式的关键因素, 而且可揭示林木适应性关键遗传因子及其遗传调控网络, 为开展区域化分子育种提供了研究思路^[56].

林木在不断应对干旱、高温、低温、病害等各种生物胁迫和非生物胁迫方面也形成了自身独特的适应机制^[57]. La Mantia等人^[58]在毛果杨群体中开展了抗病性状的GWAS研究, 发现了控制叶锈病的26个候选基因, 并证明了PRR7在抗叶锈病中的重要作用, 与先前研究中发现的PRR7可通过调节植物节律参与胁迫响应过程一致^[59]. 一般而言, 非生物胁迫是限制林木生长发育及分布范围的主要因素, 能引起林木转录、翻译和代谢水平的显著变化^[36]. 因此, 综合利用组学数据开展GWAS研究有利于多层次阐明林木逆境胁迫响

应的分子机制, 并且有助于解释和预测不同环境条件下群体遗传组成及变化规律, 有效提升林业工作者对未来气候变化下的成林稳定性评估及灾害预防能力, 对林业建设和森林保护具有重要指导意义^[1].

3 GWAS策略在林木复杂性状研究中的应用热点

近年来, 研究人员利用GWAS策略在动植物数量性状研究中鉴定了大量关键位点, 但是这些显著关联位点, 仅能解释部分表型变异, “丢失遗传力”问题依然是当前数量遗传学研究的难点^[60]. 因此, 如图1所示, 本文拟从高通量表型组学、不同变异类型的遗传效应、加性与非加性联合解析策略等方面探讨, 为系统解析林木复杂性状, 解决“丢失遗传力”问题提供重要的策略指导.

3.1 高通量表型检测技术的开发与建立

基因型(genotype)、表型(phenotype)、环境(environment)三者构成了GWAS研究的支撑要素. 然而由于林木表型变异的复杂性和动态可塑性等特点以及测量仪器较为落后, 很难实现对林木群体表型数据的规模化、标准化测定, 阻碍了对多组学数据中生物学信息的挖掘^[61~63]. 特别是随着林木GWAS策略在群体适应性、动态生长发育规律与数量性状基因组结构等研究领域的深入利用, 获取不同环境、群体、时间点的精确表型数据已成为多年生林木遗传结构解析的重要前提^[64]. 近几年, 随着人工智能、计算机视觉、深度学习等技术的发展, 研究人员在一年生草本植物中利用高通量的表型组分析平台在植株形态变化与非生物胁迫响应等研究中, 实现了高精度表型数据的采集与分析^[65~67]. 林木树体高大, 生长速度较快, 其室内表型研究重点在早期动态生长、胁迫处理响应等方面的精准监控; 田间表型研究借助航空影像、可移动测定装置等田间表型平台, 从空中与近地面等不同空间尺度进行林木冠层结构、生理状态、灾害评估等方面的表型数据收集^[67]. 在此基础上开发适合林木的表型算法, 提取海量具有生物学意义的数据用于林木GWAS研究, 将打破传统育种、分子标记辅助育种与基因组选择育种中面临的“表型瓶颈”, 从而实现林业高效管理和适应特定环境林木品种的精准选择^[68].

图 1 林木复杂性状系统作图。利用高通量表型组学平台精准测定目标性状; 开发多种类型的遗传标记用于全基因组关联分析(GWAS), 如单核苷酸多态性(SNP)、拷贝数变异(CNV)、插入缺失(InDels)、简单序列重复(SSR); 同时利用多组学数据联合分析, 如表达数量性状基因位点(expression quantitative trait loci, eQTL)作图、蛋白数量性状基因位点(proteomics quantitative trait loci, pQTL)作图和代谢数量性状基因位点(metabolomics quantitative trait loci, mQTL)作图, 深入解析林木复杂性状遗传结构, 发掘功能等位变异用于林木分子设计育种

Figure 1 Systems mapping of complex traits in forest trees. High-throughput phenotypic platform were used for measuring target traits accurately; SNP, CNV, InDels, simple sequence repeats (SSR) were developed for GWAS. At the same time, combining with multi-omics association analysis, such as expression quantitative trait loci (eQTL), proteomics quantitative trait loci (pQTL), metabolomics quantitative trait loci (mQTL) to deeply analyze the genetic structure of forest complex traits and exploring allelic variation for molecular design breeding in forest trees

3.2 稀有等位变异(rare allele)和结构变异(structure variation, SV)的联合利用

林木群体内基因组内存在大量稀有等位变异位点, 这些位点独立起作用时对群体性状变异贡献率较低, 其遗传效应在种质资源或天然群体大的群体背景下难以被发现^[28,69-71], 而且通过建立林木多世代回交群体来改变稀有等位位点的频率是难以实现的。Du 等人^[19]提出了连锁分析和关联分析相结合的林木连锁-连锁不平衡作图策略, 通过在种质资源群体内定向选择骨干亲本并构建杂交群体来提高稀有等位变异的群体分布频率。该策略解决了稀有等位变异在自然群体背景下难以发现的问题, 实现了对分离规律与遗传效应的解析。未来对林木群体稀有等位变异研究可以从以下两方面来考虑: 一方面是在实验设计阶段, 利用极端表型取样、隔离群体取样和家系群体取样等方法, 把稀有等位变异的遗传效应放大到可观测的范围^[72], 如连锁-连锁不平衡策略; 另一方

面是在林木群体中使用新的统计模型和方法, 分析稀有等位变异的遗传效应, 如利用区域遗传作图(RHM)、元分析策略(meta-analysis strategy)等, 可以有效提高林木复杂性状表型变异遗传力的解析度^[73]。

SV通常指基因组内大于1 kb的插入、缺失、倒位、易位和拷贝数目变异, 由于结构变异包含的核苷酸较多, 往往具有更强的遗传效应, 更容易影响基因的表达调控, 被认为是导致“丢失遗传力”的另一重要原因^[74]。SV在基因组内分布广泛, 可显著影响表型变异和基因组进化^[74], 是林木GWAS研究下一步需要重点关注的标记类型。例如, Gong等人^[18]在毛白杨生长和木材品质性状的关联分析中发现, InDel平均可解释表型变异的14.7%, 远大于SNP对表型变异的解析度(约为5%)。由于林木基因组高度杂合, SV的检测较为困难, 对其的研究鲜见报道^[16]。因此提高参考基因组质量, 增加重测序的覆盖深度, 是准确挖掘基因组结构变异的重要前提。

3.3 多组学数据的系统分析与应用

从DNA至表型, 涉及转录组、蛋白组、代谢组等多层次、多水平的协同调控过程, 形成了复杂的遗传调控网络和生物大分子的互作网络(图1). 近年来, 多组学研究向定量化、高通量方向发展, 为深入阐明性状形成的遗传基础提供了便利条件.

转录组是解读基因组功能元件和揭示细胞及组织转录调控机制的基础, 与稳定的基因组相比, 转录组会随着生理状态、生长状态、生长环境的变化而变化, 在个体间存在着显著差异, 可作为一种“分子表型”用于GWAS研究^[22]. 表达数量性状基因位点(expression quantitative trait loci, eQTL)作图方法将群体内每个基因型个体的基因表达水平作为表型, 可检测调控基因表达水平的功能性等位标记, 有助于估计基因表达水平的遗传率、挖掘候选基因和构建基因调控网络. Wang等人^[75]在毛白杨光合性状研究中, 利用关联分析与eQTL作图方法发现, 一个SNP位点可通过影响次级调控网络中8个转录因子的表达量, 进而调控下游55个光合相关的靶基因. 另外, 可以根据基因的表达水平将其转化为二进制数据, 作为一种独立的“分子基因型”研究该基因型在表达调控、代谢和表型变异中的作用^[22,71]. 通过将转录组数据分别作为“基因型”与“表型”进行GWAS研究, 不仅可以有效解析基因组变异调控基因表达水平的作用机制, 还可以在全基因组水平研究基因组、转录组与表型三者之间的关系, 为正确理解基因-基因、基因-环境的互作提供了重要证据, 从而深化对林木复杂性状的分子机理和遗传基础的认识^[76]. 与eQTL类似, 蛋白数量性状基因位点(proteomics quantitative trait loci, pQTL)和代谢数量性状基因位点(metabolomics quantitative trait loci, mQTL)作图方法分别将蛋白组数据和代谢组数据作为表型数据用于GWAS研究, 这有助于在生化途径和生理形成机制等层次进行数据解读, 阐明候选基因涉及的功能通路^[47,77].

利用生物信息学分析方法对多层次组学数据进行深入挖掘与联合分析, 将架起从DNA、RNA至蛋白质以及最终的表型变异研究的桥梁, 进一步推动林木系统遗传学(systems genetics)领域的发展, 实现对目标性状遗传基础的系统解析, 为目标性状的分子设计育种奠定重要理论基础^[72,75].

3.4 数量性状加性、显性与上位性联合解析策略

先前研究表明, 数量性状遗传变异一般受到加性、显性与上位性三种遗传效应的影响. 加性效应(additive effect)是指等位基因及非等位基因之间的累加效应, 属于上下代遗传中可固定的分量, 在实践中也被称为育种值(breeding value); 显性效应(dominance effect)是指各基因效应值与其加性效应值的离差, 来源于等位基因之间的互作效应; 上位性效应(epistatic effect)是指不同基因位点之间的相互作用^[78]. 林木数量性状GWAS研究中仅考虑单标记加性和显性效应, 对位点之间存在的上位性考虑较少^[78], 严重影响了对数量性状遗传结构的解析精度. 上位性效应是解析数量性状重要的遗传基础之一^[79,80], 除了常见的位点之间二维交互作用, 在人类和作物遗传学研究中还发现了三维、四维甚至更高维度的上位作用^[77]. Du等人^[19]通过整合MLM与非线性逻辑回归等算法, 在林木中率先构建了数量性状加性、显性与上位性联合解析新策略, 首次在分子水平上捕获了位点间的上位性效应, 系统解析了生物学通路多基因、多位点联合控制生长与木材品质性状的等位互作机制. Chen等人^[81]利用该策略构建了8个miRNAs(microRNAs)和其对应的9个靶基因在生长及木材品质性状中的遗传调控网络, 以此揭示miRNA及其靶基因的等位作用模式, 为深入阐明非编码RNA与下游靶基因互作机制提供了重要的分析手段. 因此, 未来在群体基因组学与GWAS联合分析方向, 建立数量性状加性、显性与上位性联合解析模型, 将进一步推动数量性状基因组学的发展.

4 基于GWAS的林木分子育种研究展望

对于作物育种, 分子育种策略的构建与模拟实践已取得了较好效果, 但是林木分子育种理论研究依然处于起步阶段. 开展不同树种的GWAS研究工作, 在全基因组范围内系统地发掘影响目标性状变异的显著QTLs及其候选基因位点, 并联合基因组编辑技术对其进行功能验证, 为基因组定向选择育种和聚合育种提供了关键功能位点. 在此基础上利用不同的分子育种策略, 如基于模块聚合的分子模块设计育种策略^[82]和覆盖全基因组的高密度遗传标记计算个体的基因组估计育种值(genomic estimated breeding value, GEBV)的

全基因组选择(whole genomic selection, WGS)育种策略^[49], 形成完整的林木分子设计育种理论。当然, 真正在多年生树木中实现育种实践还需要开发出一系列育种技术。林木基因组育种芯片的应用, 架起了林木GWAS研究成果通向林木育种应用的桥梁, 提高了育种过程的针对性与可控性, 为育种学家准确鉴定和筛

选功能基因, 科学控制遗传背景提供了有效手段^[83]。

GWAS策略的应用极大推动了林木复杂性状的研究, 通过进一步结合“分子设计育种”的手段和技术方法, 构建林木分子设计育种体系, 有助于林木育种工作实现从传统的“经验育种”向定向高效的“精准育种”转化, 满足现代林业对优异多元化种质资源的需求。

参考文献

- 1 Neale D B, Kremer A. Forest tree genomics: Growing resources and applications. *Nat Rev Genet*, 2011, 12: 111–122
- 2 Fisher R A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb*, 1919, 52: 399–433
- 3 Lander E S, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 1989, 121: 185–199
- 4 Doerge R W. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet*, 2002, 3: 43–52
- 5 Chagné D, Brown G, Lalanne C, et al. Comparative genome and QTL mapping between maritime and loblolly pines. *Mol Breed*, 2003, 12: 185–195
- 6 Casasoli M, Derory J, Morera-Dutrey C, et al. Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. *Genetics*, 2006, 172: 533–546
- 7 Rae A M, Street N R, Robinson K M, et al. Five QTL hotspots for yield in short rotation coppice bioenergy poplar: The Poplar biomass loci. *BMC Plant Biol*, 2009, 9: 23–36
- 8 Remington D L, Thornsberry J M, Matsuoka Y, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA*, 2001, 98: 11479–11484
- 9 Evans L M, Slavov G T, Rodgers-Melnick E, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet*, 2015, 46: 1089–1096
- 10 Thumma B R, Nolan M F, Evans R, et al. Polymorphisms in *Cinnamoyl CoA Reductase (CCR)* are associated with variation in microfibril angle in *Eucalyptus* spp. *Genetics*, 2005, 171: 1257–1265
- 11 Nordborg M, Weigel D. Next-generation genetics in plants. *Nature*, 2008, 456: 720–723
- 12 Tuskan G A, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 2006, 313: 1596–1604
- 13 He N, Zhang C, Qi X, et al. Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat Commun*, 2013, 4: 2445
- 14 Sato S, Hirakawa H, Isobe S, et al. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res*, 2011, 18: 65–76
- 15 Cruz F, Julca I, Gómez-Garrido J, et al. Genome sequence of the olive tree, *Olea europaea*. *Gigascience*, 2016, 5: 29
- 16 Ingvarsson P K, Hvidsten T R, Street N R. Towards integration of population and comparative genomics in forest trees. *New Phytol*, 2016, 212: 338–344
- 17 Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity: Is dispensable really dispensable? *Curr Opin Plant Biol*, 2014, 18: 31–36
- 18 Gong C, Du Q, Xie J, et al. Dissection of insertion-deletion variants within differentially expressed genes involved in wood formation in *Populus*. *Front Plant Sci*, 2017, 8: 2199–2214
- 19 Du Q, Gong C, Wang Q, et al. Genetic architecture of growth traits in *Populus* revealed by integrated quantitative trait locus (QTL) analysis and association studies. *New Phytol*, 2015, 209: 1067–1082
- 20 Zhang J, Yang Y, Zheng K, et al. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*. *New Phytol*, 2018, 220: 502–516
- 21 Visscher P M, Brown M A, McCarthy M I, et al. Five years of GWAS discovery. *Am J Human Genet*, 2014, 90: 7–24
- 22 Xiao Y, Liu H, Wu L, et al. Genome-wide association studies in maize: Praise and stargaze. *Mol Plant*, 2017, 10: 359–374
- 23 Porth I, Klapšte J, Skyba O, et al. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytol*, 2013, 200: 710–726
- 24 Abecasis G R, Cookson W O C. GOLD—Graphical overview of linkage disequilibrium. *Bioinformatics*, 2000, 16: 182–183

- 25 Neale D B, Ingvarsson P K. Population, quantitative and comparative genomics of adaptation in forest trees. *Curr Opin Plant Biol*, 2008, 11: 149–155
- 26 Slavov G T, DiFazio S P, Martin J, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol*, 2012, 196: 713–725
- 27 Nicolas S D, Péros J P, Lacombe T, et al. Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L.) diversity panel newly designed for association studies. *BMC Plant Biol*, 2016, 16: 74–94
- 28 Fahrenkrog A M, Neves L G, Resende Jr M F R, et al. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytol*, 2016, 213: 799–811
- 29 Josephs E B, Stinchcombe J R, Wright S I. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytol*, 2017, 214: 21–33
- 30 Grattapaglia D, Resende M D V. Genomic selection in forest tree breeding. *Tree Genets Genomes*, 2011, 7: 241–255
- 31 Yang N, Lu Y, Yang X, et al. 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*, 2014, 10: e1004573
- 32 Lu M, Krutovsky K V, Nelson C D, et al. Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. *Tree Genets Genomes*, 2017, 13: 57–75
- 33 McKown A D, Klápště J, Guy R D, et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol*, 2014, 203: 535–553
- 34 Uchiyama K, Iwata H, Moriguchi Y, et al. Demonstration of genome-wide association studies for identifying markers for wood property and male strobili traits in *Cryptomeria japonica*. *PLoS ONE*, 2013, 8: e79866
- 35 Resende R T, Resende M D V, Silva F F, et al. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytol*, 2016, 213: 1287–1300
- 36 Čalić I, Koch J, Carey D, et al. Genome-wide association study identifies a major gene for beech bark disease resistance in American beech (*Fagus grandifolia* Ehrh.). *BMC Genomics*, 2017, 18: 547–561
- 37 Müller B S F, Neves L G, de Almeida Filho J E, et al. Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. *BMC Genomics*, 2017, 18: 524
- 38 De La Torre A R, Puiu D, Crepeau M W, et al. Genomic architecture of complex traits in loblolly pine. *New Phytol*, 2019, 221: 1789–1801
- 39 Chhetri H B, Macaya-Sanz D, Kainer D, et al. Multitrait genome-wide association analysis of *Populus trichocarpa* identifies key polymorphisms controlling morphological and physiological traits. *New Phytol*, 2019, 223: 293–309
- 40 Bradbury P J, Zhang Z, Kroon D E, et al. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 2007, 23: 2633–2635
- 41 Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Human Genet*, 2007, 81: 559–575
- 42 Kang H M, Zaitlen N A, Wade C M, et al. Efficient control of population structure in model organism association mapping. *Genetics*, 2008, 178: 1709–1723
- 43 Yang J, Lee S H, Goddard M E, et al. GCTA: A tool for genome-wide complex trait analysis. *Am J Human Genet*, 2011, 88: 76–82
- 44 Lipka A E, Tian F, Wang Q, et al. GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, 2012, 28: 2397–2399
- 45 Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 2012, 44: 821–824
- 46 Wang S B, Feng J Y, Ren W L, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep*, 2016, 6: 19444–19453
- 47 Zhang F T, Zhu Z H, Tong X R, et al. Mixed linear model approaches of association mapping for complex traits based on omics variants. *Sci Rep*, 2015, 5: 10298–10307
- 48 Mellerowicz E J, Sundberg B. Wood cell walls: Biosynthesis, developmental dynamics and their implications for wood properties. *Curr Opin Plant Biol*, 2008, 11: 293–300
- 49 Lamara M, Raherison E, Lenz P, et al. Genetic architecture of wood properties based on association analysis and co-expression networks in white spruce. *New Phytol*, 2016, 210: 240–255
- 50 Mizrahi E, Myburg A A. Systems genetics of wood formation. *Curr Opin Plant Biol*, 2016, 30: 94–100

- 51 Allwright M R, Payne A, Emiliani G, et al. Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European *Populus nigra* (L.). *Biotechnol Biofuels*, 2016, 9: 195–217
- 52 Pyhäjärvi T, García-Gil M R, Knürr T, et al. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics*, 2007, 177: 1713–1724
- 53 Savolainen O, Pyhäjärvi T, Knürr T. Gene flow and local adaptation in trees. *Annu Rev Ecol Evol Syst*, 2007, 38: 595–619
- 54 Shi J S. Prospection on some topics of forest genetic improvement through modern biotechnology for the next-ten-years in China (in Chinese). *J Nanjing For Univ (Nat Sci Ed)*, 2012, 36: 1–13 [施季森. 林木生物技术育种未来10年若干科学问题展望. *南京林业大学学报(自然科学版)*, 2012, 36: 1–13]
- 55 McKown A D, Guy R D, Klápště J, et al. Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytol*, 2014, 201: 1263–1276
- 56 McKown A D, Klápště J, Guy R D, et al. Ecological genomics of variation in bud-break phenology and mechanisms of response to climate warming in *Populus trichocarpa*. *New Phytol*, 2018, 220: 300–316
- 57 Adams H D, Guardiola-Claramonte M, Barron-Gafford G A, et al. Temperature sensitivity of drought-induced tree mortality portends increased regional die-off under global-change-type drought. *Proc Natl Acad Sci USA*, 2009, 106: 7063–7066
- 58 La Mantia J, Klápště J, El-Kassaby Y A, et al. Association analysis identifies *Melampsora × columbiana* poplar leaf rust resistance SNPs. *PLoS ONE*, 2013, 8: e78423
- 59 Zhou M, Wang W, Karapetyan S, et al. Redox rhythm reinforces the circadian clock to gate immune response. *Nature*, 2015, 523: 472–476
- 60 Li Y, Huang Y, Bergelson J, et al. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 2010, 107: 21199–21204
- 61 Yang Y X, Yang Z M, Wu C J, et al. Advances in plant phenomics research (in Chinese). *Acta Agri Univ Jiangxiensis*, 2015, 37: 1105–1112 [杨有新, 杨泽茂, 吴才君, 等. 植物表型组学研究进展. *江西农业大学学报*, 2015, 37: 1105–1112]
- 62 Houle D, Govindaraju D R, Omholt S. Phenomics: The next challenge. *Nat Rev Genet*, 2011, 11: 855–866
- 63 Furbank R T, Tester M. Phenomics-technologies to relieve the phenotyping bottleneck. *Trends Plant Sci*, 2011, 16: 635–644
- 64 Kipp S, Mistele B, Baresel P, et al. High-throughput phenotyping early plant vigour of winter wheat. *Eur J Agron*, 2014, 52: 271–278
- 65 Winterhalter L, Mistele B, Schmidhalter U. Evaluation of active and passive sensor systems in the field to phenotype maize hybrids with high-throughput. *Field Crops Res*, 2013, 154: 236–245
- 66 Araus J L, Cairns J E. Field high-throughput phenotyping: The new crop breeding frontier. *Trends Plant Sci*, 2014, 19: 52–61
- 67 Zhou J, Tardieu F, Pridmore T, et al. Plant phenomics: History, present status and challenges (in Chinese). *J Nanjing Agri Univ*, 2018, 41: 580–588 [周济, Tardieu F, Pridmore T, 等. 植物表型组学: 发展、现状与挑战. *南京农业大学学报*, 2018, 41: 580–588]
- 68 Mu J H, Chen Y Z, Feng H, et al. A new revolution in crop breeding: The era of high-throughput phenomics (in Chinese). *Plant Sci J*, 2016, 34: 962–971 [穆金虎, 陈玉泽, 冯慧, 等. 作物育种学领域新的革命: 高通量的表型组学时代. *植物科学学报*, 2016, 34: 962–971]
- 69 Lettre G. Rare and low-frequency variants in human common diseases and other complex traits. *J Med Genet*, 2014, 51: 705–714
- 70 Manolio T A, Collins F S, Cox N J, et al. Finding the missing heritability of complex diseases. *Nature*, 2009, 461: 747–753
- 71 Auer P L, Lettre G. Rare variant association studies: Considerations, challenges and opportunities. *Genome Med*, 2015, 7: 16–26
- 72 Liu D J, Peloso G M, Zhan X, et al. Meta-analysis of gene-level tests for rare variant association. *Nat Genet*, 2014, 46: 200–204
- 73 Eichler E E, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 2010, 11: 446–450
- 74 Alkan C, Coe B P, Eichler E E. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011, 12: 363–376
- 75 Wang L X, Du Q, Xie J, et al. Genetic variation in transcription factors and photosynthesis light-reaction genes regulates photosynthetic traits. *Tree Physiol*, 2018, 38: 1871–1885
- 76 Zhu G, Wang S, Huang Z, et al. Rewiring of the fruit metabolome in tomato breeding. *Cell*, 2018, 172: 249–261.e12
- 77 Fabres P J, Collins C, Cavagnaro T R, et al. A concise review on multi-omics data integration for terroir analysis in *Vitis vinifera*. *Front Plant Sci*, 2017, 8: 1065–1073
- 78 Du Q, Wei Z, Zhao X, et al. Dissection of additive, dominant, epistatic roles of allelic variation within heat shock factor genes in Chinese indigenous poplar (*Populus simonii*). *Tree Genets Genomes*, 2016, 12: 91–105
- 79 Clark Cockerham C. An extension of the concept of parti-tioning hereditary variance for analysis of covariances among relatives when epistasis is

present. *Genetics*, 1954, 39: 859–882

- 80 Wang X, Fu A Q, McNeerney M E, et al. Widespread genetic epistasis among cancer genes. *Nat Commun*, 2015, 5: 4828–4837
- 81 Chen J, Xie J, Chen B, et al. Genetic variations and miRNA-target interactions contribute to natural phenotypic variations in *Populus*. *New Phytol*, 2016, 212: 150–160
- 82 Hao H Q, Liu L L, Yao Y, et al. Application and prospect of molecular module-based crop design technology in maize breeding (in Chinese). *Bull Chin Acad Sci*, 2018, 33: 923–931 [郝怀庆, 刘丽丽, 姚远, 等. 分子模块设计育种技术在玉米育种中的应用及前景展望. *中国科学院院刊*, 2018, 33: 923–931]
- 83 Yu H H, Zhang Q F, Zhou F S. Rice genomic breeding chips and its applications (in Chinese). *Chin Bull Life Sci*, 2016, 28: 1258–1267 [喻辉辉, 张启发, 周发松. 水稻基因组育种芯片及其应用. *生命科学*, 2016, 28: 1258–1267]

Progress and prospect of genome wide association study (GWAS) in forest trees

LI Peng^{1,2,3}, LU WenJie^{1,2,3}, XIAO Liang^{1,2,3}, QUAN MingYang^{1,2,3}, LI LianZheng^{1,2,3},
DU QingZhang^{1,2,3} & ZHANG DeQiang^{1,2,3}

1 Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing 100083, China;

2 National Engineering Laboratory for Tree Breeding, Beijing Forestry University, Beijing 100083, China;

3 Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China

With the rapid development of next-generation genome sequencing technologies, the application of systems genetics based on genome-wide association study (GWAS) strategy is extensively increasing in revealing the genetic variation underlying complex traits and dissecting the genetic regulation mechanism of key genes in forest trees. Firstly, we introduced the core theories, research methods and progress of GWAS in the wood traits and adaptation of forest trees. Next, we proposed a joint use of high-throughput phenome platform, multiple genetic markers, multi-omics data and the models of additive, dominance and epistasis to resolve the problem of “missing heritability” of quantitative traits in forest trees. Finally, we provided an outlook of the GWAS strategy in the molecular breeding of forest trees.

genome-wide association study (GWAS), systems genetics, genetic architecture, missing heritability, molecular breeding

doi: [10.1360/SSV-2019-0129](https://doi.org/10.1360/SSV-2019-0129)