分类号：　　密　　级：公开
学　号：20202312015　　单位代码：10759

石河子大学

博士学位论文

基于 GWAS 和 eQTL 联合分析揭示影响棉花产量性状的遗传因素

学位申请人　郭春平
指导教师　聂新辉
申请学位门类级别　农学博士
学科、专业名称　作物学
研究方向　作物遗传育种
所在学院　农学院

中国·新疆·石河子
2024 年 9 月

分类号：　　密　　级：公开

学　号：20202312015　　单位代码：10759

石河子大学

博 士 学 位 论 文

基于 GWAS 和 eQTL 联合分析揭示影响棉花产量性状的遗传因素

学 位 申 请 人	郭春平
指 导 教 师	聂新辉
申请学位门类级别	农学博士
学科、专业名称	作物学
研 究 方 向	作物遗传育种
所 在 学 院	农学院

中国·新疆·石河子

2024 年 9 月

Combined GWAS and eQTL analysis reveals genetic components influencing fiber yield trait in cotton

A Dissertation Submitted to

Shihezi University

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Agriculture

By

(Crop genetic breeding)

Dissertation Supervisor:

September, 2024

石河子大学学位论文独创性声明及使用授权声明

学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：郭春平　　时间：2024年10月26日

使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：郭春平　　时间：2024年10月26日

导师签名：　　时间：2024年10月26日

摘要

【目的】陆地棉是重要的天然纺织纤维来源，也是重要的经济作物之一。新疆是我国棉花的主要生产基地，占全国总产量的 91%。但随着新疆棉花育种的发展，利用传统的育种技术改良棉花产量已经出现了瓶颈。研究表明，改良品种的衣分是提高产量的重要因素，因此，解析新疆棉花品种衣分性状的遗传基础，挖掘聚合优异位点的种质对于突破新疆棉花高产育种至关重要。

【方法】本研究以新疆 1978 年至今所有审定的棉花陆地棉品种、骨干亲本及育种中间材料为研究材料，基于群体基因组十倍重测序和多环境条件下的衣分表型数据进行全基因组关联分析（GWAS），根据纤维早期发育模型，纤维细胞是在开花前一天（-1DPA）开始分化，纤维细胞突起分化的数量直接决定了棉花的衣分值。通过对群体材料-1DPA 胚珠转录组数据进行 eQTL 联合分析挖掘调控衣分性状的显著基因位点，进一步通过共定位分析和全转录组关联分析（TWAS）直接鉴定衣分（LP）关键候选基因；再通过 CRISPR/Cas9 技术创制衣分相关候选基因的突变体材料，初步验证候选基因功能，为培育衣分优异种质资源提供了突变体材料资源；并评估了新疆棉花育种过程中 LP 优异位点的聚合情况，将为棉花高产种质资源创新提供重要的理论基础和物质基础。

【结果】主要研究结果如下：

1.衣分的全基因组关联分析

（1）本研究收集了 312 份陆地棉材料，其中包括 180 份新疆审定品种和 132 份育种中间材料，构建了稳定遗传的陆地棉群体。对该群体进行了 4 年的表型考察，在 2018-2021 年（2018SHZ_LP，2019SHZ_LP，2020SHZ_LP，2021SHZ_LP）考察其衣分性状发现变异范围为 27.08%-50.36%，2020SHZ_LP 环境的衣分表型变异幅度最大，而 2019SHZ_LP 环境的衣分表型变异幅度最小，衣分的广义遗传力为 86.46%，环境间相关系数为 0.48-0.84。

（2）基于基因组重测序数据，共鉴定出了 2,481,993 个 MAF>0.05 的高质量 SNPs 用于后续分析。在 SNP 注释之后，进行群体遗传结构分析，发现群体可以划分为 4 个亚群，但亚群间不能够完全区分开，亚群内均还带有其他亚群的遗传标记。通过尝试按照群体材料类型、育种年限以及适宜种植区域划分后进行主成分分析，发现不同类型划分结果相对一致，大部分材料无明显分离，只有 42 份育种中间材料归类存在变化。利用群体内部的连锁不平衡分析结果，发现全基因组水平的 r^2 大约在 1,028kb 的物理距离处衰减至一半。

（3）基于 4 个不同环境（2018-2021 年）的 GWAS 分析，总共定位到了 11 个衣分性状相关的显著 SNP 位点，其中 2018 年，2019 年，2020 年和 2021 年分别定位到 2 个，0 个，5 个和 4 个显著位点。同时我们利用四年数据的 BLUP 值进行关联分析，发现了 4 个跟衣分性状显著相关的 SNP 位点。综合分析不同环境的定位结果，发现一个在 D05 染色体上的显著 SNP 位点 Ghir_D05__38870550 在

2021_SHZ 年和 BLUP_SHZ 环境下均定位到。

（4）根据上述定位到的 14 个显著关联的 SNP 位点和 LD 衰减距离，我们以 SNP 显著位点为中心，两端延伸 1,028 kb 的距离作为候选区间，经分析共鉴定到了 1,438 个候选基因，包括 *ERF1*、*CYP78A9*、*ROPGEF5*、*RGLG5* 等基因。

（5）基于 CRISPR-Cas9 基因编辑技术创制了 *ERF1*（*Ghir_A12G022760*）突变体植株。本研究获得了该基因 T1 和 T2 代突变体植株，发现与野生型相比，衣分均显著下降，表明本研究基于 GWAS 分析鉴定的候选基因 *ERF1* 参与棉花衣分性状的调控。

2. 陆地棉群体材料在-1DPA 纤维起始时期的 eQTL 分析

（1）本研究基于华中农业大学 TM-1 基因组，对 312 份陆地棉群体材料转录组测序数据进行分析，发现 52,357 个基因在纤维发育的起始阶段具有表达水平。其中，44,056 个基因具有表达变异，占 TM-1 基因组中注释基因的 62.8%。并且通过 eQTL 分析，鉴定到 13,859 个 eQTL 与 10,105 个 eGenes 表达相关。

（2）通过对 eQTL 以及被 eQTL 调控的基因（eGene）在染色体上的分布发现，在 A 亚基因组上受 eQTL 转录调控的基因数量高于在 D 亚基因组上的 eGene 数量，推测两个亚基因组之间存在不平衡的转录调控模式。

（3）基于全转录组关联分析（TWAS）共鉴定到了 9 个跟衣分相关的候选基因（FDR<0.05）。这 9 个基因在胚珠和纤维中都有表达，其中 *Ghir_D11G026650* 和 *Ghir_D03G012490* 在纤维起始期优势表达，并且在群体中差异表达。有趣的是基于 TWAS 定位到的候选基因 *Ghir_A12G025980* 同时被 QTL 调控。

（4）利用共定位策略，联合 GWAS 与 eQTL 数据共定位到 20 个跟衣分性状相关的候选基因。这些基因均在胚珠或纤维发育中有表达，其中，三个基因（*Ghir_A12G027820*、*Ghir_A12G025990* 和 *Ghir_A12 G022500*）在纤维起始期特异性表达。

（5）通过对调控多个 eGenes 的 eQTLs 分析发现有七个具有代表性的大热点在纤维起始阶段调节从 36（Hot192）到 191（Hot28）不等数量的基因。Hot26 是一个极其重要的调节热点，调节 186 个基因的表达，这些基因涉及细胞分化、植物激素合成和代谢、脂质合成和葡萄糖代谢，以及 NAD（P）H 氧化酶 H_2O_2 形成活性。参与了纤维起始和发育的重要途径。Hot28 在相同的途径中调节基因表达。Hot26 和 Hot28 共同调控 5 对同源基因，推测这两个热点在调节基因表达方面具有协同作用。Hot160 在 GWAS 区间（104,527,859-104,561,245bp）转录调控 8 个候选基因，调节基因转录水平表达的热点 Hot160 位于染色体 A12 上。Hot132 是 TWAS 区间（42,571,666-42,571,667bp）的一个重要调控热点，其调节 *Ghir_D03G012470*、*Ghir_D03 G012480* 和 *Ghir_D03-G012490* 的表达。

（6）本研究鉴定了 269 个与产量性状相关的差异表达基因，包括 212 个高衣分材料中的上调基因和 57 个高衣分材料中的下调基因。通过共定位分析鉴定了一个与下调基因重叠的衣分相关基因 *Ghir_A12G025990*。此外，基于 GWAS 分析鉴定的候选区间中有 6 个差异表达基因，其中 5 个基因上调（*Ghir_A10G024040*、*Ghir_A12G021010*、*Ghir_A12G021150*、*Ghir_A12G021380*、*Ghir_*

A12G022360），1 个基因下调（*Ghir_ A12G025990*）。其中 5 个差异表达基因位于 A12 染色体，1 个位于 A10 染色体。

（7）本研究中连锁不平衡衰减距离说明了 312 份材料在选择育种进程中受到了强烈的驯化选择。为了确定新疆棉花品种驯化的遗传基础，将具有选择扫描信号位置与本研究中鉴定到的位点位置进行比对重叠，结果发现有 5 个 QTL 区间与选择扫描信号重叠，分别是 qLP_A11，qLP_A12_1，qLP_A12_2，qLP_D05_1，qLP_D05_2。这 5 个变异位点在半野生棉花中的遗传多样性大于栽培棉花品种，在驯化选择区间中鉴定的变异位点受到了不同程度的驯化选择，另外未经驯化选择的候选基因在育种中也有很大的应用潜力。

（8）通过评估有利等位基因在 312 份棉花材料中的聚合积累情况，分析发现有利等位基因位点的数量越多，材料衣分表型值越大。基于该结果，使用 312 份材料中的衣分相关基因位点构建了有利等位基因位点资源库，分析了 312 份材料中具有亲缘关系的材料，分析发现父本和母本材料中都具有的优异位点，在父母本杂交后代育成的品种中都含有来自父母本共有的优异位点，在后代中被稳定的遗传。而仅在父本或者母本材料中具有的优异位点，它们的杂交后代育成的品种中不一定遗传亲本中具有优异位点。根据研究结果模拟了通过杂种优势聚合父母本优异位点培育高衣分后代的方法。通过该方法，对本研究材料中没有聚合到的优异位点也提出了后代预测情况，聚合位点占有利等位基因位点比例越大，其衣分越高，可以帮助选择培育高衣分品种的亲本。

（9）基于 CRISPR-Cas9 基因编辑技术创制了 *res1*（*Ghir_D11G026650*）突变体植株。本研究获得了该基因 3 种编辑类型的 T2 代突变体植株，发现与野生型相比，衣分表型显著下降，初步验证 TWAS 分析鉴定到的候选基因 *rse1* 参与了棉花衣分性状的调控。

【结论】本研究群体材料存在广泛的表型变异且 4 个环境间显著相关，虽然环境因子对其有微弱影响，但衣分遗传变异仍较稳定。根据群体结构分析发现，本试验群体材料的遗传关系密切，遗传背景狭窄，亲缘关系较近，不能完全独立地区分每个遗传亚群，并根据主成分分析新发现了改良衣分性状且具有丰富遗传多样性的种质资源。本研究群体的候选区间的延伸距离较大，可能是由于群体亲缘关系较近，群体结构较稳定单一，内部亚群分离不明显等原因所导致，但该群体也存在一定的遗传分化。基于全基因组关联分析，共鉴定到 14 个调控衣分性状的显著位点以及 1,438 个调控衣分性状的候选基因，本研究通过突变体材料初步验证了，基于 GWAS 分析鉴定的候选基因对于棉花衣分改良的可靠性。

本研究利用群体转录组数据获得了调控衣分性状的 eQTL 热点以及候选基因，从转录水平解析基因调控衣分现状的表达变异，同时利用全转录组关联分析及共定位策略获得了在基因组和转录组水平上调控衣分性状的候选基因。通过解析本研究定位到调控衣分性状优异位点在 312 分陆地棉材料中的聚合情况，挖掘了聚合了优异位点的高衣分材料，为选择亲本育种提供了参考。并根据研究结果模拟了通过杂种优势聚合父母本优异位点培育高衣分后代的方法。另外，本研究通过突变体材料初步验证了，基于 TWAS 分析鉴定的候选基因对于棉花衣分改良的可靠性。

【关键词】：衣分；全基因组关联分析；eQTL；全转录组关联分析；棉花遗传育种

Abstract

【**Object**】Upland cotton is an important natural source of textile fibers and one of the important economic crops. Xinjiang is the main cotton production base in China, accounting for 91% of the total national output. But with the development of cotton breeding in Xinjiang, the use of traditional breeding techniques to improve cotton yield has become a bottleneck. Research has shown that improving the lint percentage of cotton accessions is an important factor in increasing yield. Therefore, it is crucial to analyze the genetic basis of cotton lint percentage traits in Xinjiang and explore germplasm with aggregation advantages to promote high-yield cotton breeding in Xinjiang.

【**Methods**】This study used all approved cotton upland cotton accessions, backbone parents, and breeding intermediate accessions in Xinjiang from 1978 to present as research accessions. Based on population genome ten fold resequencing and lint phenotype data under multiple environmental conditions, Genome-wide association study (GWAS) was conducted, and significant gene loci regulating lint percentage traits were identified through eQTL combined analysis. According to the early development model of fibers, fiber cells begin to differentiate at -1DPA, and the number of fiber cell protrusions determines the cotton's lint percentage score. Key candidate genes for lint percentage (LP) could also be directly identified through co-localization analysis and transcriptome-wide association study (TWAS); Using CRISPR/Cas9 technology to create mutant accessions for candidate genes related to lint percentage, providing suitable mutant accessions for further functional research; And evaluated the aggregation of LP advantages and disadvantages in the cotton breeding process in Xinjiang, which will provide important theoretical and accession basis for the innovation of high-yield cotton germplasm resources.

【**Results**】The main research findings are as follows:

1. Genome-wide association study of lint percentage

(1) This study constructed a population of 312 upland cotton accessions for analysis. The lint percentage trait exhibited extensive phenotypic variation from 2018 to 2021 (2018SHZ_LP, 2019SHZ_LP, 2020SHZ_LP, 2021SHZ_LP) and were significantly correlated among four environments, with a variation range of 27.08% -50.36%. The phenotypic variation of lint percentage in the 2020 SHZ-LP environment is the largest, while the phenotypic variation of lint percentage in the 2019 SHZ-LP environment is the smallest, with a generalized heritability of 86.46%. The correlation coefficient between environments is 0.48-0.84.

(2) This study identified a total of 2,481,993 SNPs with MAF>0.05 for subsequent analysis. After SNP

annotation, the population was divided into 4 subgroups based on genetic structure, which could not be completely distinguished and also carried genetic markers from other subgroups. We also attempted to classify accessions according to population accession types, breeding years, and suitable planting areas, and found that accessions classified by different types were mostly concentrated in the middle of the scatter plot without obvious separation. . Analyzing the linkage imbalance within the population, the r^2 at the whole genome level decays to half at a physical distance of approximately 1,028kb.

(3) In four different environments from 2018 to 2021, a total of 11 SNP loci significantly related to lint percentage trait were identified. Among them, two significant loci were identified in 2018, and zero, five and four significant loci were identified in 2019, 2020 and 2021, respectively. At the same time, we also performed BLUP on the 4-year data, and association analysis revealed 4 SNP loci significantly related to lint percentage traits. Among them, the same significant SNP locus Ghir_D05_38870550 was identified on the D05 chromosome in 2021-SHZ and BLUP-SHZ.

(4) Based on the 14 significantly associated SNP loci and LD decay distance previously identified, we identified a total of 1,438 candidate genes (like *ERF1*、*CYP78A9*、*ROPGEF5*、*RGLG5 et al*) with SNP salient loci as the center and a distance of 1,028 kb extending from both ends as candidate intervals.

(5) *ERF1* (*Ghir_A12G022760*) mutant plants were created based on CRISPR-Cas9 technology. This study obtained T1 and T2 mutant plants of the gene and found a significant decrease in lint percentage phenotype compared to the wild type, proving the reliability of the candidate genes identified based on GWAS analysis for cotton lint percentage improvement in this study.

2. Transcriptome analysis of upland cotton during the fiber initial stage of -1DPA.

(1) 52,357 genes have expression levels at the initial stage of fiber development. 44,056 genes showed expression variations, accounting for 62.8% of annotated genes in the TM-1 genome. And 13,859 eQTLs related to 10,105 eGenes expression were identified.

(2) In the initial stage of fiber development, There are fewer genes regulated by transcription in the A subgenome than in the D subgenome, and there is an unequal transcriptional regulation pattern between the two subgenomes. However, The number of SNPs in the A subgenome is greater than that in the D subgenome. Due to the inconsistent size of subgenomes, there may be slight differences in the number of genes, and there may be equal or coordinated transcriptional regulation between the two subgenomes.

(3) Nine candidate genes related to lint percentage were identified based on TWAS (FDR<0.05). These 9 genes are expressed in both ovule and fiber development stages, *Ghir_D11G026650* and *Ghir_D03G012490* have higher expression levels in the fiber initiation stage compared to other fiber development stages, and are differentially expressed in the population. Based on the TWAS localization of candidate genes, *Ghir_A12G025980* is also regulated by QTL.

(4) Using co-localization strategy and combining GWAS and eQTL data, 20 candidate genes related to lint

percentage trait were co-located and expressed during ovule and fiber development stages. Among them, three genes (*Ghir_A12G027820, Ghir_A12G025990* and *Ghir_A12G022500*) are specifically expressed in the fiber initiation stage.

(5) The study found that seven representative hotspots regulate a varying number of genes from 36 (Hot192) to 191 (Hot28) during the fiber initiation stage. Hot26 is an extremely important regulatory hotspot, regulating the expression of 186 genes involved in important pathways of fiber initiation and development, including cell differentiation, plant hormone synthesis and metabolism, lipid synthesis and glucose metabolism, and NAD (P) H oxidase H_2O_2 formation activity. Hot28 regulates gene expression in the same way. Hot26 and Hot28 simultaneously regulate 5 pairs of homologous genes, and we speculate that these two hotspots have a synergistic effect in regulating gene expression. Hot160 regulates the transcription of 8 candidate genes in the GWAS interval (104,527,859-104,561,245bp), and the hotspot for regulating gene transcription level expression is located on chromosome A12. Hot132 is an important regulatory hotspot in the TWAS interval (42,571,666-42,571,667bp), which regulates expression of *Ghir_D03G012470, Ghir_D03G012480* and *Ghir_D03G012490*.

(6) This study identified 269 differentially expressed genes related to yield traits, including upregulated genes in 212 high lint accessions and downregulated genes in 57 high lint accessions. A lint percentage related gene *Ghir_A12G025990* that overlaps with the downregulated gene was identified through co-localization analysis. Based on GWAS analysis, there are six differentially expressed genes identified in the candidate interval, of which five genes are upregulated (*Ghir_A10G024040, Ghir_A12G021010, Ghir_A12G021150, Ghir_A12G021380, Ghir_A12G022360*), with one gene downregulated (*Ghir_A12G025990*). Among them, five differentially expressed genes are located on chromosome A12, and one is located on chromosome A10.

(7) In order to determine the genetic basis for cotton domestication, we compared and overlapped the selected scanning signal positions with the identified loci in this study. The results showed that there were five QTL intervals that overlapped with the selection scanning signal , namely qLP-A11, qLP_A12_1，qLP_A12_2，qLP_D05_1，qLP_D05_2. The genetic diversity of gene loci in semi wild cotton is greater than that in cultivated cotton accessions, and the identified gene loci in the domestication selection interval have been subjected to varying degrees of domestication selection. Candidate genes that have not been domesticated have great potential for application in breeding.

(8) We evaluated the accumulation of favorable alleles in 312 cotton accessions. A favorable allele locus resource library was constructed using lint percentage related gene loci from 312 accessions. The more favorable the number of allele loci, the larger the accession's lint percentage phenotype. The statistical analysis of phenotypic values for accessions with the same cumulative quantity reveals that there are differences in phenotypic values for the same cumulative locus among different accessions, which may be

due to the presence of promoting or inhibiting loci in different accessions. In most accessions, the higher the accumulated number of sites, the higher the lint percentage. However, there are also some accessions with more accumulated sites and lower lint percentage. In a small number of accessions, there may be antagonistic effects between the advantageous sites. The heterozygous sites contained in both parents have genetic stability and are stably inherited in offspring, but heterozygous sites contained only in the paternal or maternal parents may not necessarily be inherited in offspring. We provide a method for predicting the phenotype of offspring from aggregated parental heterozygous loci through heterosis. The hybrid offspring may aggregate excellent loci that were not aggregated in the accession of this study. The higher the proportion of allele loci occupied by the aggregated loci, the higher the lint percentage, which can help predict the selection of parents to cultivate high lint percentage accessions.

(9) Res1 (*Ghir-D11G026650*) mutant plants were created based on CRISPR-Cas9 technology. This study obtained T2 mutant plants with three editing types of the gene, and found a significant decrease in lint percentage phenotype compared to the wild type, proving the reliability of the candidate genes identified based on TWAS analysis for cotton lint percentage improvement in this study.

【Conclusion】The population accessions in this study exhibits extensive phenotypic variation and significant correlations among four environments. Although environmental factors have a slight impact on it, genetic variation in lint percentage remains relatively stable. According to the analysis of population structure, it was found that the genetic relationship of the population accessions in this experiment is close, the genetic background is narrow, and the kinship relationship is relatively close, which makes it impossible to completely independently distinguish each genetic subgroup. Based on principal component analysis, a new germplasm resource with improved lint traits and rich genetic diversity was discovered. The extension distance of the candidate interval in this study population is relatively large, which may be due to the close genetic relationship, stable and single population structure, and unclear separation of internal subgroups. However, there is also some genetic differentiation in this population. Based on genome-wide association analysis, a total of 14 significant loci and 1438 candidate genes regulating cotton lint traits were identified. This study preliminarily verified the reliability of the candidate genes identified based on GWAS analysis for cotton lint improvement through mutant materials.

This study utilized population transcriptome data to identify eQTL hotspots and candidate genes that regulate lint percentage traits. The expression variations of genes regulating lint percentage status were analyzed at the transcriptional level, and candidate genes upregulated for lint percentage traits at the genome and transcriptome levels were obtained using whole transcriptome association analysis and co localization strategies. By analyzing the aggregation of the optimal sites for regulating the trait of lint percentage in 312 upland cotton accesssions, this study identified high lint percentage accessions with aggregated optimal sites, providing a reference for selecting parents for breeding. And based on the