

Reveal genomic insights into cotton domestication and improvement using gene level functional haplotype-based GWAS

Received: 17 January 2025

Accepted: 8 May 2025

Published online: 21 May 2025



Guoan Qi^{1,2,3}, Yiqian Li^{1,3}, Wanying Zhang^{1,3}, Zegang Han¹, Jinwen Chen¹,
Ziqian Zhang¹, Lisha Xuan¹, Rui Chen¹, Lei Fang^{1,2}, Yan Hu^{1,2} &
Tianzhen Zhang^{1,2} ✉

Genome-wide association studies (GWAS) are widely used to detect associations between genetic variants and phenotypes. However, few studies have thoroughly analyzed genes, the fundamental and most crucial functional units. Here, we develop an innovative strategy to translate genomic variants into gene-level functional haplotypes (FHs), effectively reducing the interference from complex genome structure and linkage disequilibrium (LD) present in the conventional genetic mapping framework. Using refined mixed linear models, gene-level FH is regressed with 20 cotton agronomic traits across 245 sets of phenotypic values in 3,724 accessions, directly identifying 532 quantitative trait genes (QTGs) with significant breeding potential. The biological function of a superior fiber quality QTG encoding ferulic acid 5-hydroxylase 1 is experimentally validated. Thereafter, we systematically analyze the genetic basis of cotton domestication and improvement at the gene level. This report provides genomic insight into the genetic dissection and efficient mapping of functional genes in plants.

Polygenic inheritance forms the genetic basis of complex traits¹, and hence their dissection necessitates genomes with high-precision resolution. Advances in high-throughput sequencing technologies and emerging genome-wide association studies (GWAS) have revealed substantial genetic variation associated with complex traits in humans^{2,3}, plants^{4,5}, and animals^{6,7}. Knowledge of these variations have facilitated the resolution of the genetic architectures of complex traits, genetic ancestry identification, and accurate prediction of phenotypes from genotypes. However, although genomic variation, represented by single-nucleotide polymorphisms (SNPs) and small structural variations, has illuminated a wide range of genetic mechanisms, the gene remains the determining biological factor in most applied genetic studies. Thus, while a plethora of GWAS signals have been identified, these associations rarely lead to the identification of causal genes⁸, and the overall goal of assigning genetic mechanisms remains unfulfilled.

This discrepancy arises from the fact that strongly associated variants are often in linkage disequilibrium (LD) with causal variants⁹, and such LD may be unexpectedly strong in plant and animal species subjected to long-term artificial selection, resulting in excessively long candidate intervals for putative quantitative trait loci (QTL). Additionally, genome complexity, especially in polyploid species, leads to higher gene density, further complicating the efficacy of gene mapping within current GWAS frameworks. Whole-exome sequencing (WES) focuses on genetic signatures within gene exons¹⁰, but the association strategy does not differ from that used for whole genome sequencing (WGS), and at present it still faces potential issues with detection power¹¹. To date, few studies have comprehensively characterized genetic variation in the gene unit, much less explored the potential of using gene-based variation for direct mapping of functional genes.

¹The Advanced Seed Institute, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang 310058, China. ²Hainan Institute of Zhejiang University, Yazhou Bay Science and Technology City, Sanya, Hainan 572025, China. ³These authors contributed equally: Guoan Qi, Yiqian Li, Wanying Zhang. ✉e-mail: cotton@zju.edu.cn

Cotton is the most important source of natural fibers. Since 2012, genome assemblies of representative cotton species have been constructed^{12–15}, large-scale WGS of major germplasm populations have been conducted and more than 4000 genomes of cotton accessions have been determined^{13,16–22}. Most modern cultivars of upland cotton (*Gossypium hirsutum* L.) were domesticated from the Yucatan Peninsula of Mesoamerica and therefore possess the genetic ancestry of American cotton. Long-term domestication has led to a genetic bottleneck and high levels of LD²³; moreover, as a typical allotetraploid crop, the cotton has a complex genome which characterized by variable structure and high gene density. Consequently, progress in mapping functional genes underlying key agronomic traits remains slow, presenting both significant opportunities and challenges for the advancement of functional genomics in cotton.

In this work, we propose a genetic strategy to transform discrete genomic variants into composite factorial functional haplotypes (FHs). By recoding non-synonymous variations within the coding regions of genes, FH reflects variations in protein sequences across a population. We analyze gene-level FH in a comprehensive collection of 3724 accessions, encompassing 3636 wild and/or semi-wild species, landraces, varieties, and modern cultivars of upland cotton and 88 *G. barbadense* accessions as outgroup; investigate genetic diversity changes related to cotton adaptation and domestication at the gene level. Importantly, we apply a revised linear mixed model to the factorial FH and conduct gene-level association analysis on a collection of 245 phenotypes covering 20 traits, including fiber quality, lint yield and its components, plant architecture, maturity, and resistance. The FH-based GWAS enables the efficient and direct mapping of quantitative trait genes (QTGs), which initially identify 10,279 QTGs, and further estimation of exact effect sizes confirms 532 genes with significant breeding potentials. Finally, we perform CRISPR-Cas9 knock-out experiments on a QTG encoding ferulic acid 5-hydroxylase 1, demonstrating its loss to result in the expected superior phenotypic changes in fiber quality.

Results

Mining of gene-level functional haplotype

We collected a total of 4392 upland cotton cultivars and lines (accessions) of *G. hirsutum* from four of our previously published datasets, as well as four other publicly available datasets (Supplementary Data 1). An additional 88 *G. barbadense* accessions were included as outgroups. We categorized all samples into those with high (>10×) and low (<10×) sequencing depth, and used the samples with high depth as a reference panel to impute low-depth samples with the same geographic origin, i.e., population-specific imputation (Supplementary Data 2). After performing quality control for the genomic variants, we evaluated the combined samples and found those samples with lower sequencing depth exhibited unique but abnormal genetic ancestry and sub-branching in the phylogenetic tree, and these samples cannot cluster with either early landraces or modern cultivars of upland cotton with same geographic origin (Supplementary Fig. 1). In addition, we observed a distinct bimodal pattern of variant heterozygosity in low-depth samples (Supplementary Fig. 2). The discrepancies between low- and high-depth samples suggest significant technical rather than biological differences. Consequently, we removed 756 accessions with sequencing depth lower than 10×. For the remaining 3724 accessions, we re-performed population-specific imputation without a reference panel. After further quality control (biallelic, mapping quality >30, inbreeding coefficient >0.6, minor allele frequency [MAF] >0.001), we finalized a dataset of 3724 accessions (Supplementary Data 3) with 23,057,253 genomic variants for subsequent FH identification. These cotton accessions represent a wide range of geographic origins, including wild species, semi-wild races, and early landraces from Central America (CAL), North America (NAL), South America (SAL),

and South China (SCL); improved varieties from Africa, Asia, Europe, and Oceania; and modern cultivars from the United States of America (USA), the former Soviet Union (FSU), and China, including cotton growing regions in Yellow River (YRR), Yangtze River (YZR), Northwest China (NWC), Southwest China (SWC), North China (NC), and East China (EC).

Regarding genome-wide variation, great attention was paid to non-synonymous mutations, with different genotype combinations for a given gene within a population being considered to constitute distinct FHs (Fig. 1). For example, when there is a single non-synonymous mutation in a gene, the FH genotype (i.e., FH-1, FH-2, and FH-3) is equivalent to the additive genotype of that mutation (i.e., AA, Aa, and aa); whereas when there are two non-synonymous mutations, there are at most 3², i.e., nine FH types in the population according to the combinatorial genotypes of the two mutations; and if there are three non-synonymous mutations, the gene has up to 27 FH types. By analogy, when there are *M* non-synonymous mutations in a gene, the gene has at most 3^{*M*} FH types, but its maximum will not exceed the population size *N*. By further analyzing the frequency of FHs within different populations, such as those of different geographic origins, and their correlations with agronomic traits of interest, key FHs with their represented genes can be identified. The fundamental rationale of our strategy is the extraction of protein variation based on genomic variation, an approach that transforms the analysis object from single-point variation to gene/protein variation, thereby greatly optimizing and simplifying the interpretability of subsequent genetic analysis. Furthermore, this approach

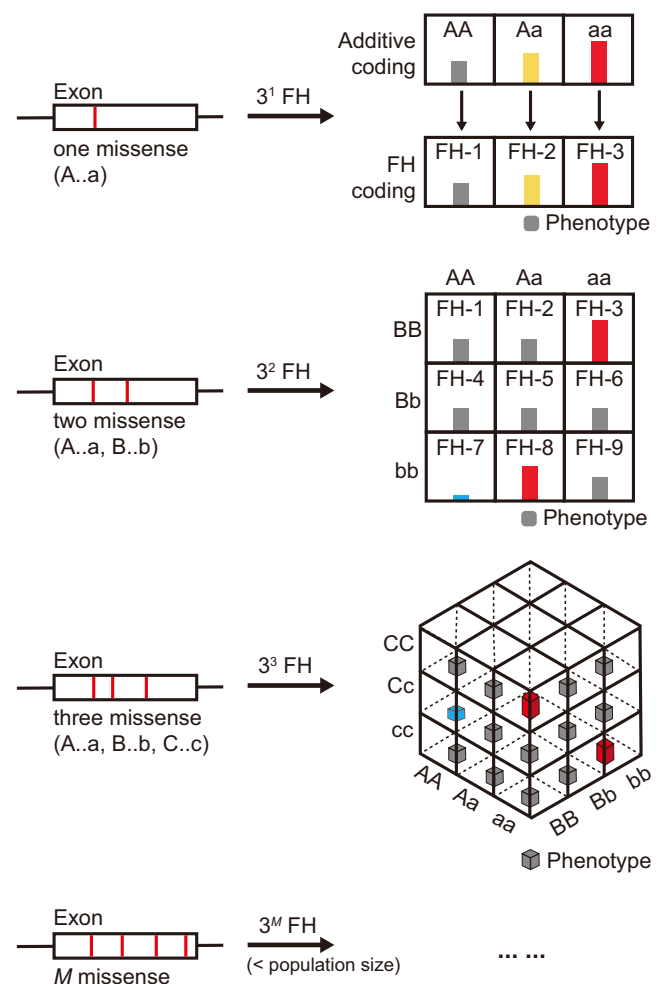


Fig. 1 | Strategy of FH construction.

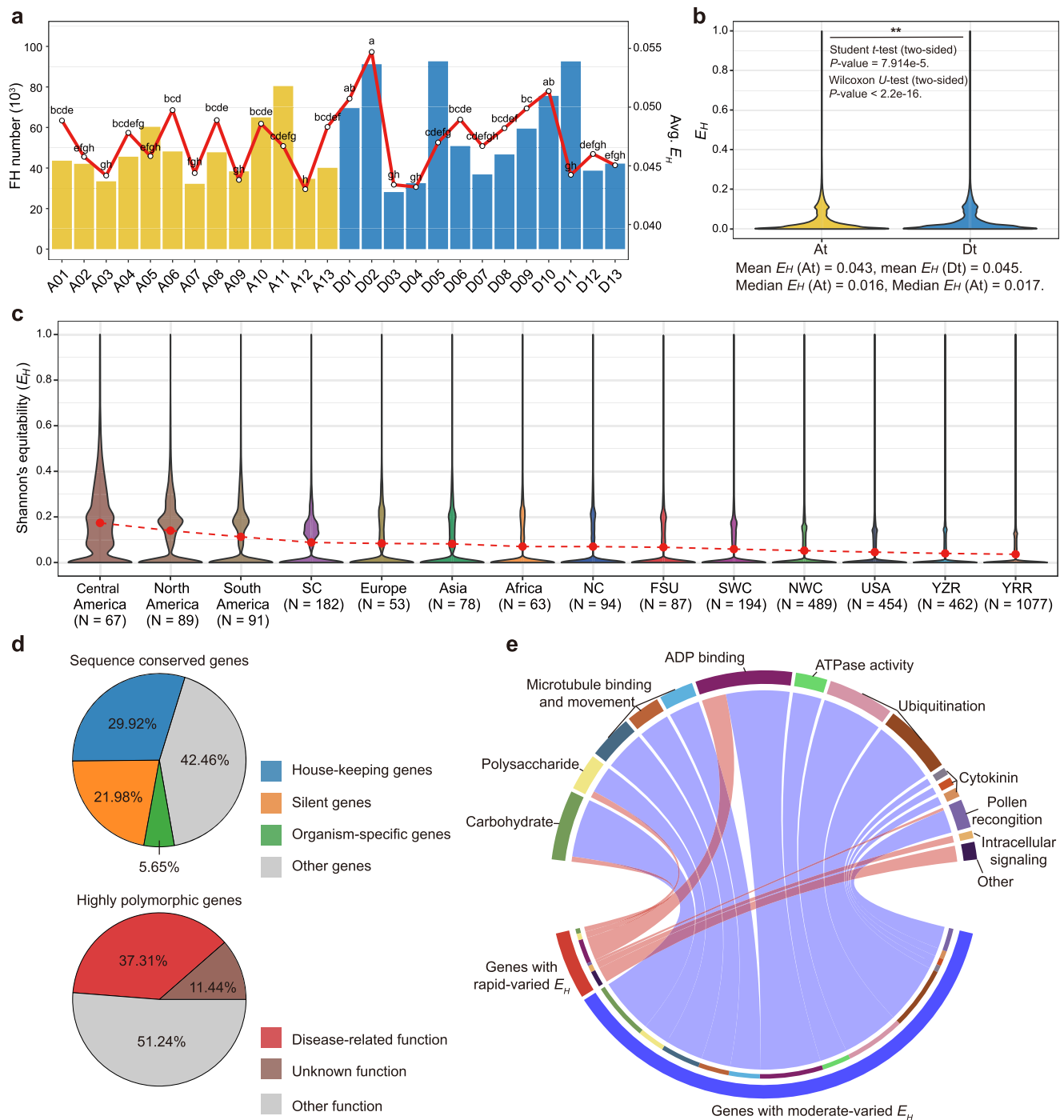


Fig. 2 | FH landscape revealed gene diversity and directional selection in the global upland cotton population. a FH numbers identified on distinct chromosomes (bar, left y-axis) and the average gene diversity E_H (line, right y-axis). Multiple comparisons of E_H in different chromosomes were conducted by the least significant difference (LSD) test, with p values adjusted by the Benjamini–Hochberg method. **b** Distribution of E_H across sub-genomes At and Dt. **c** Changes in gene diversity characteristics from cotton semi-wild species, early landraces to modern cultivated accessions. Red dots indicate the mean genetic diversity of each population. N represents the sample size. **d** Functional investigation for sequence-conserved and highly polymorphic genes. **e** Enrichment analysis revealing the directional selection from early landraces to modern cultivars. Source data are provided as a Source Data file.

facilitates efficient gene mapping in association analysis. In analyzing the 3724 accessions, we screened out 1,402,043 exonic variants, which included 334,406 non-synonymous variants located in 60,154 out of 71,994 genes that directly change the corresponding protein sequences. These variants resulted in a genome-wide total of 1,368,685 FHs across the entire population, with an average of over 22 FHs in each gene (Supplementary Data 4).

The landscape of FHs reveals directional gene selection during cotton adaptation and domestication

In the global set of cotton accessions, genomic FHs are unevenly distributed across chromosomes, with FH abundance correlating positively with the number of genes on the chromosome (Fig. 2a, Supplementary Fig. 3, and Supplementary Table 1). To make gene diversity comparable across populations with different sample sizes,

we calculated Shannon's equilibrium index (E_H) for each gene in the population. The average E_H indicated significant differences in gene diversity across chromosomes and sub-genomes, suggesting that the adaptation and domestication of upland cotton introduced varying degrees of directional selection on the genome (Fig. 2a, b). For instance, chr. D02 is among the chromosomes containing the most FHs (91,206) and also has the highest average E_H (0.0573), which suggests genes on this chromosome have a higher level of diversity and are less subject to natural and artificial selection. Conversely, chr. D11, despite having an even higher number of FHs (92,593), exhibited the lowest average E_H (0.0391), implying strong intervention from artificial selection on its genes. Further population-level analysis revealed significantly higher gene diversities in cottons from CAL, NAL (Mexico), SAL, and SCL, which represent wild and semi-wild race species and early landraces. In contrast, modern cultivars, especially from the USA and Chinese YZR and YRR, exhibited relatively lower gene diversities (Fig. 2c). Notably, the gene diversity pattern in the SCL population (Hainan province and Guangxi province) is similar to that in American landraces, indicating the ancient genetic components of early landraces in SCL and their geographic origin. These results are consistent with modern cotton breeding history across the world.

To investigate the functions of genes with different degrees of FH diversity, we first focused on sequence-conserved genes (SCGs), which lack any missense variants across the entire population. By integrating RNA-Seq data from whole cotton tissues (Supplementary Data 5), we found that 57.54% of SCGs had potential functions related to cellular existence and cell fate, which are critical to the organism (Fig. 2d, Supplementary Table 2, and Supplementary Data 6). Regarding function classifications, 29.92% of SCGs were house-keeping genes expressed in all 34 tested tissues, 21.98% were silent genes not expressed in any tested tissue, and 5.65% were genes expressed in specific organisms. For the remaining 42.46% of SCGs, functional enrichment analysis identified many as related to actin and microtubule processes, such as microtubule-based process and structural constituent of cytoskeleton, and others as involved in basic cell development functions such as electron transfer activity and photosynthesis (Supplementary Data 7). These fiber-associated SCGs are considered to have been fixed in the cotton breeding process.

At the other extreme, we also analyzed the functions of genes with highly variable sequences (highly polymorphic genes, HPGs, $E_H > 0.3$). Interestingly, the majority (62.69%) of HPGs had seemingly random functions, while the remaining 37.31% were clearly related to disease response and foreign substances recognition (Supplementary Table 3). Furthermore, most HPGs (93.53%) were subject to strong positive selection ($Ka/Ks > 1$, Supplementary Data 8). This finding is similar to the major histocompatibility complex (MHC) and T cell receptors (TR) in human genetic research, both of which possess an unusually large number of alleles and extremely diverse protein types that determine unique antigenic compositions at the individual level and their interactions. This result suggests the possibility of a shared or similar genetic mechanism underlying immunity among life kingdoms.

We further analyzed the functions of genes with rapidly decreased gene diversity ($\Delta E_H > 0.3$) in modern cotton cultivars compared to early landraces. These genes are enriched in biological pathways related to carbohydrates, polysaccharides, ADP binding, pollen recognition, and intracellular signaling (Fig. 2e and Supplementary Data 9). Finally, genes with moderately reduced E_H ($0.1 < \Delta E_H < 0.3$) in modern cultivars relative to earlier accessions are enriched in pathways involving microtubule binding and movement, ATPase activity, ubiquitination, and cytokinin metabolism. These findings are consistent with the results of recent cotton breeding efforts to improve agronomic traits, particularly lint yield and fiber quality.

FH-based haplotype analysis reflects population structure and elite gene transferring during development

Utilizing the genotypes encoded by factorial FH, we applied a general method to calculate the genetic distance between sample pairs, and then constructed a phylogenetic tree based upon those distances (Methods). This analysis classified the 3724 cotton accessions into six clusters, in which the 88 *G. barbadense* accessions were clustered into group 1 (Outgroup, G1) and the remaining 3636 *G. hirsutum* accessions into five groups (G2 to G6; Fig. 3a). Among these groups, G2 collected most accessions from YZR and a considerable portion from YRR, due to similar geographical and climatic conditions and frequent introductions; G3 consisted of another portion of the YRR cultivars; G4 contained most accessions from the USA; G5 included most accessions from NWC and FSU; and G6 encompassed most early landraces from America and Southern China, consistent with cotton breeding practice in China.

We further compared the similarities and differences of phylogenetic trees constructed on the basis of FH and SNP information. Despite the distribution density of genes being far lower than that of SNPs, FH-based genotyping completely reproduced the subgroup characteristics observed when classifying on SNPs, and also accurately reflected the population structure. At the group level, FH-G1 corresponds to SNP-G1, FH-G2, and FH-G3 to SNP-G4 and SNP-G5, FH-G4 to SNP-G6, FH-G5 to SNP-G3, and SNP-G6 and FH-G6 to SNP-G2 and SNP-G6 (Fig. 3b). Nonetheless, some discrepancies were observed in the attributional assignments of genetic components between FHs and SNPs. For example, the cultivars from SWC in FH-G6, which clustered with early landraces from the American continents and SC, were assigned to SNP-G6, which contained modern cultivars from the USA and NWC, indicating potential differences in genetic characteristics reflected.

Next, FH-based haplotypes on two typical regions of structural variation on chr. A06 and A08 were further investigated. Comparisons of the merged FH and SNP haplotypes suggested that FHs are able to capture structural variations previously implied by SNPs (Supplementary Figs. 4, 5). Additionally, since FH allows genes to be coded in both a more flexible and more diverse manner, the FH-based haplotype allows dividing structural variation into finer haplotype blocks, enabling more accurate tracking of genetic ancestry. On chr. A06, haplotype (HB)-2 and HB-3 reflected different genetic components that could not be distinguished based on SNP data; and the genetic ancestry of HB-1 and HB-4 were easily attributed to NAL and SCL, respectively (Fig. 3c). In addition, evidence of genetic sharing in HB-2, HB-3, and HB-4 supported genetic transferring between SCL and SWC. Meanwhile, HBs on chr. A08 revealed unique genetic footprints of adaptation in cottons from NWC, YRR, and YZR. HB-1 on chr. A08 was unique to the NWC cultivars, while the other FH-HBs reflected the discovery and fixation of superior genes originated from USA cottons in Chinese breeding programs (Fig. 3d). This result is consistent with the historical introduction of upland cotton cultivars from the USA to China over the last century, and subsequent pedigree selection and cross breeding in major cultivation areas over the past 70 years. Integration with agronomic traits revealed the genetic polymorphisms reflected by HBs on chr. A06 and A08 to be associated with important lint yield components, lint percentage (LP), and fiber qualities such as fiber length (FL) and strength (FS) (Fig. 3e). This genetic pleiotropy of A08 HB is indicated as the result of the combined action of pleiotropy QTL and LD linked non-pleiotropy QTL (Supplementary Table 4). Beyond chr. A06 and chr. A08, FH also provided a more accurate and finer genetic adaptation pattern and population structure for chr. A01, A13, D01, D06, D07, and D08 (Supplementary Fig. 6).

Haplotype blocks of FH on major structural variants strongly suggest superior allele/gene transferring between the introduced germplasm and modern cultivars in the breeding process. We accordingly used the number of shared FHs to calculate a coefficient

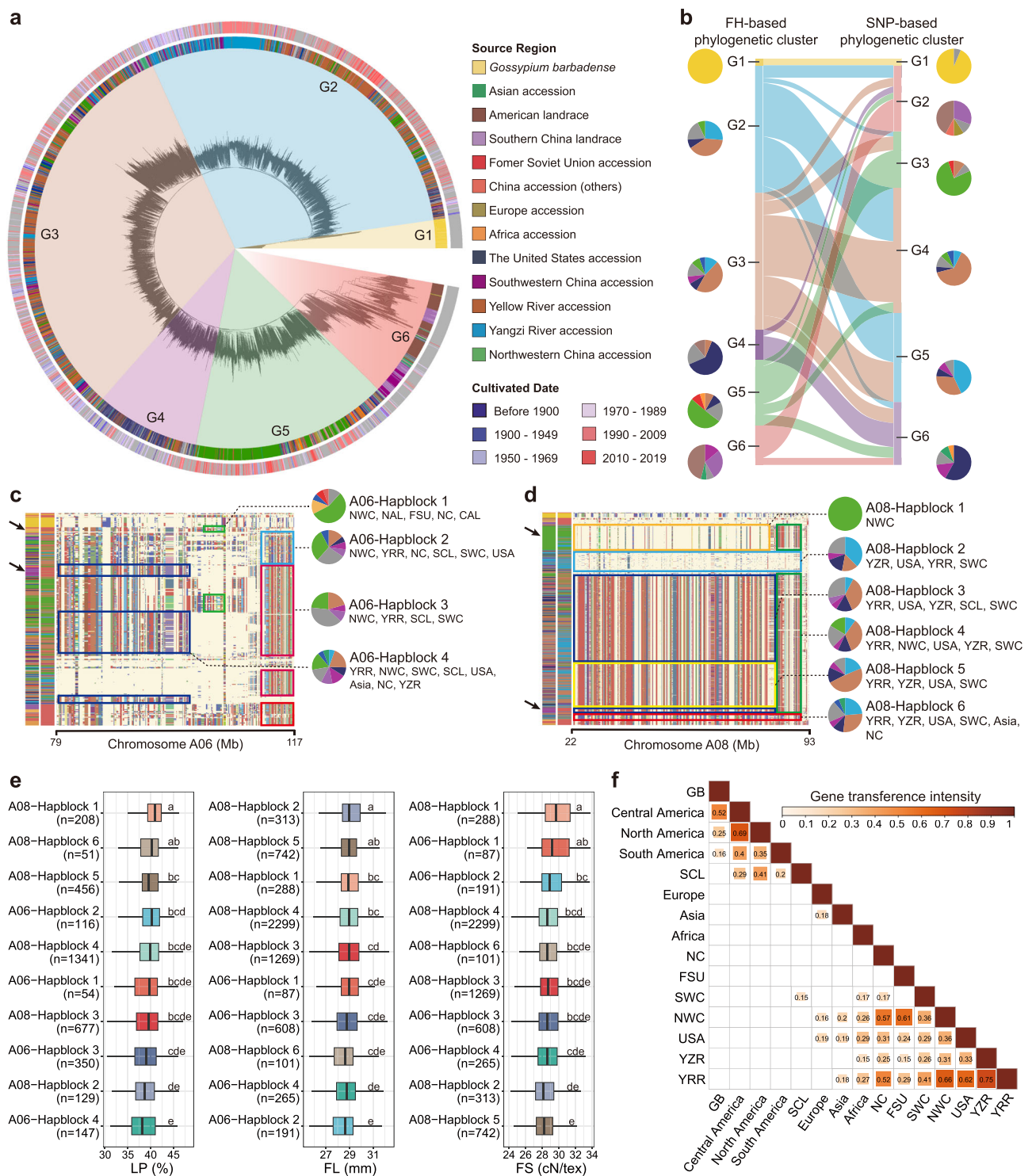


Fig. 3 | Population structure and domestication pattern revealed by FH in the 3724 cottons. **a** Phylogenetic analysis of the global cotton population based on FH. Germplasm geographical origins and release dates are annotated on the periphery of the phylogenetic tree. All germplasms are clustered into six groups G1–G6 based on their geographic origin and tree topology. **b** Comparison of sample clustering between the FH tree and SNP tree. The geographical origins of the clusters' main accessions (>5%) are annotated by the fan chart. Genomic divergence in chr. **c** A06

and **d** A08 as revealed by FH. Germplasms with the same FH are assigned to one haplotype block (HB). **e** Different HBs in chr. A06 and A08 are associated with fiber improvement. The center line represents the median, the lower and upper box hinges corresponding to the first and third quantiles, and whiskers extend to the minimum and maximum values. **f** Intensity of gene transference between population pairs. Only intensities exceeding 0.15 are labeled. Source data are provided as a Source Data file.

reflecting the intensity of gene transference between populations (Methods, Fig. 3f, and Supplementary Data 10). The results highlighted two transferring hotspots among earlier landraces and modern cultivars. First, in early landraces, a relatively high gene transference was

observed between *G. barbadense* accessions and cottons sampled from Central American countries and Caribbean islands, this finding is inferred to relate to the similar geographical origins of upland and island cotton. In addition, strong gene transference was observed

among CAL, NAL, and SAL, which suggests the introduction of early cotton landraces within the American continents. A very high level of gene transference was also found from NAL to SCL, once again indicating an ancient genetic background for landraces in Southern China. The transferences from NAL to SCL may have originated from the global trading activities of European colonizers during the Age of the Great Seas¹⁹. A recent genomics study found that wild cotton seeds from the island of Hainan were able to survive floating in seawater for more than six months, raising the possibility that SCL progenitors drifted over the Pacific Ocean²⁴. Second, modern cotton cultivars exhibited another hotspot with frequent gene transference, particularly from the USA to YRR and YZR, and among those from NWC, YRR, FSU, and NC. These results are in line with current understanding of cotton breeding history in China^{19,25}.

FH-based GWAS mapped QTGs associated with agronomic traits

Despite the fact that short variations are abundantly investigated in genetic research, the vast majority of genetic studies, especially in plant and animal genetics, still rely on genes for further elucidation of regulatory mechanisms and the development of improved cultivars. In this study, we converted short variant information into gene-level FH, enabling the direct mapping of QTGs related to target agronomic traits. For the QTG mapping, we collected a total of 245 sets of phenotypic data covering 20 agronomy traits regarding lint yield, fiber quality, plant architecture, resistance, and maturity (Supplementary Data 11). Using a linear mixed model that incorporates the genetic kinship matrix generated from FH genotype as random effects and the first two principal components (PCs) as fixed covariates, we performed single-gene regression for 60,154 genes against the 245 phenotypes. The factorial FH genotype were dummy coded, and a reduced phenotype was applied in regression to save computational time (Methods). This FH-GWAS initially identified 10,279 genes significantly associated with agronomic traits (Supplementary Data 12). After simple clumping to remove other significant genes within 500 kilobase (kb), a total of 5022 QTG clusters were identified (Supplementary Data 13). As an essential comparison, we also conducted conventional SNP/Indel GWAS for the 245 phenotypes, which incorporates the genetic kinship matrix generated from SNP/Indel genotype instead as random effects and the first two PCs as fixed covariates. After clumping, this yielded 6533 candidate QTLs significantly associated with the phenotypes (Supplementary Data 14). Generally, the number of QTLs is higher than the number of QTGs, likely due to possible non-coding associations (Fig. 4a). Nonetheless, the number of QTGs showed a strong positive correlation with the number of QTLs, especially in those traits with multiple environments and large sample sizes. Chr. A08 and D11 were notable for containing the most QTGs. The high number in chr. A08 may result from the extremely large LD caused by structural variation covering over 70 megabases (Mb); in chr. D11, it is likely due to associations with numerous phenotypes. This high degree of functional influence may also explain why chr. D11 features the highest number of FHs but the lowest average gene density.

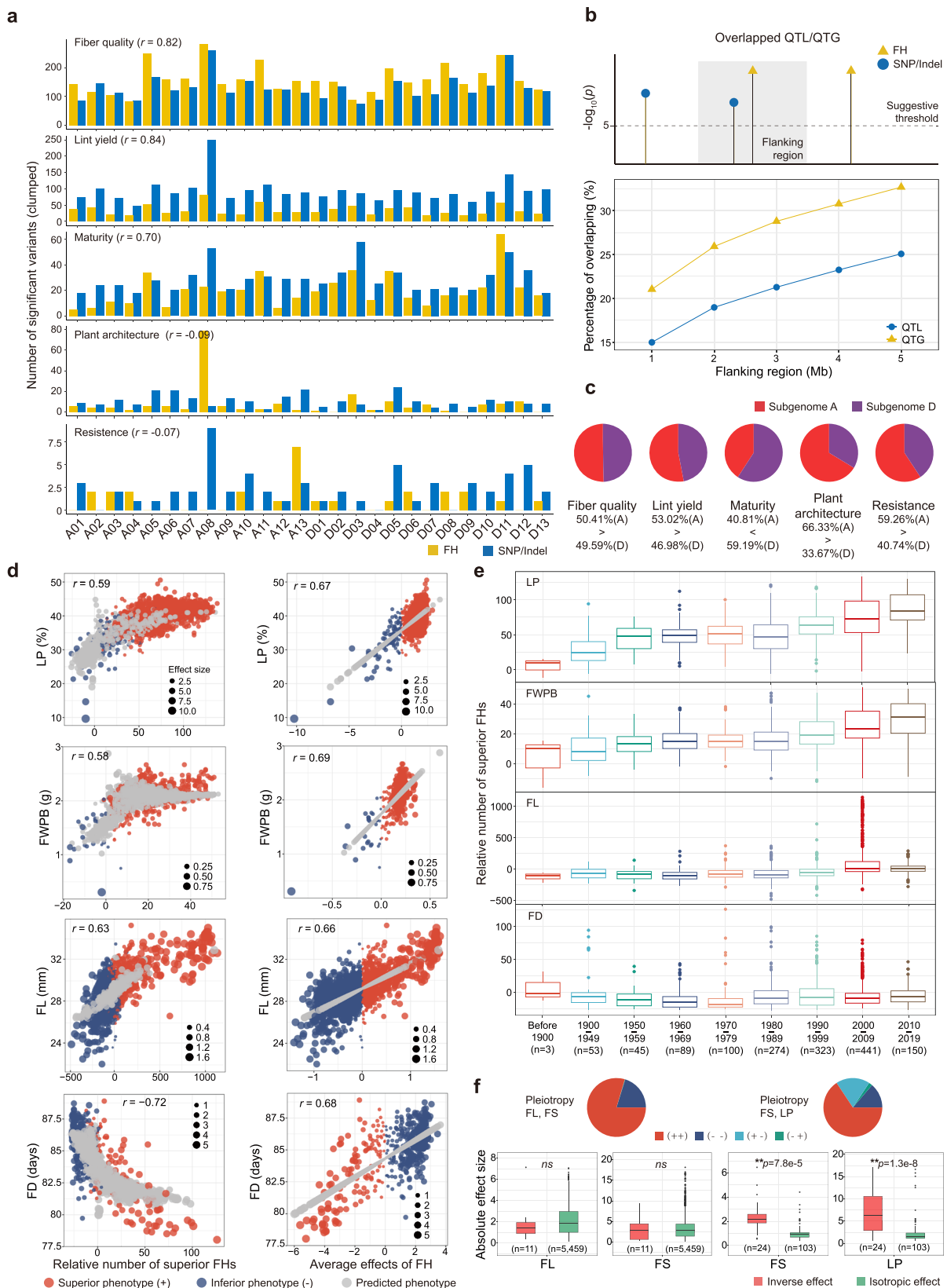
We next explored whether QTLs and QTGs reflect similar genomic regions associated with a phenotype. The number of overlapped QTLs and QTGs increased with flanking region size (1 to 5 Mb; Fig. 4b). In crops with high genome linkage, casual genes may be located very far away from significant QTLs; this once again emphasizes the superiority and necessity of FH-based GWAS in gene mapping studies. Furthermore, under the condition where the flanking region was 5 Mb, only approximately 35% of QTLs and 25% of QTGs were found to overlap. This indicates that different genomic signatures are discovered by SNP/Indel and gene-level FH association analyses. We found that QTGs for fiber quality, lint yield, plant architecture, and resistance were situated mainly in the A sub-genome (Fig. 4c), which explains its lower

gene diversity (Fig. 2b), in that the A sub-genome has experienced stronger selection pressures than the D sub-genome.

Taking the FH in TM-1 (coded as FH-1) as the reference FH, we calculated exact phenotypic effects for the alternative FHs in all 10,279 potential QTGs (Supplementary Data 15). Among all 72,559 alternative FHs, 19,228 FHs in 8917 genes were found to have significant phenotypic effects compared to the reference FH. Of these, 9411 FHs were identified to have significantly superior effects on agronomic traits, whereas 9817 FHs which mainly concentrated in the early landraces (Supplementary Data 16) had significantly adverse effects. A total of 7724 FHs in 3701 genes were implied to be pleiotropy across multiple traits. The association study detected statistically significant phenotypic correlations in 8917 genes, while most of the identified effect sizes were low (Supplementary Fig. 7), which is consistent with polygenic architecture for complex traits²⁶. Further screening for rare variants and minor effects revealed 532 high-frequency genes with major effects, which are of direct breeding value (Supplementary Data 17).

At the individual level, we found the relative number of superior FHs (the number of superior FHs minus the number of inferior FHs in a given sample) and the average effects of significant FHs to be highly correlated with most phenotypes (Fig. 4d, Supplementary Fig. 8, and Supplementary Table 5). This result illustrates the reliability of the QTGs identified in FH-GWAS. Considering the release dates of the various accessions, we examined the pattern of change in the number of superior FHs in accessions from the beginning of the 20th century to the present (Fig. 4e and Supplementary Fig. 9). This revealed superior FH presence to have massively increased over time for lint yield traits of LP, lint index (LI), fiber weight per boll (FWPB), boll weight (BW); for fiber quality traits of fiber uniformity (FU), spinning consistency index (SCI); and for plant height (PH), which relates to plant architecture, and verticillium wilt index (VW), which relates to disease and stress resistance. These trends are particularly evident in cultivars selected after the 1990s, where the efficiency of breeding selection markedly improved due to advancements in and maturation of the molecular biology techniques applied in genetic breeding. For some agronomic traits such as FL, FS, and FD, the long history of cultivar improvement resulted in only a relatively limited introduction of superior FHs until the early 21st century. However, these FHs became more abundant in cultivars from the year 2000 onwards, suggesting a shift in the goals of cotton breeding. Interestingly, superior FHs related to FM and FE showed obvious decreases in cultivars released over the last 30 years (Supplementary Fig. 10), a result that is consistent with the decreased fiber fineness in recent cultivars and implies great potential for improvement of these traits in future cotton genetic breeding.

There are well-known phenotypic correlations for several important traits related to cotton fiber, particularly the positive correlations between fiber quality traits (e.g., between FL and FS) and negative correlations between fiber quality and lint yield traits. We resolved the genetic basis of these phenotypic correlations from the gene perspective. For the positively correlated FL and FS, we found almost all pleiotropic FHs (99.73%) to have the same effect direction (Fig. 4f and Supplementary Table 6), which implies a synergistic improvement in FL and FS by the selection of a causal gene; meanwhile, for FHs with isotropic or inverse effects for FL and FS, the exact phenotypic effects showed no significant difference. For the negatively correlated FS and LP, although isotropic FHs also constituted the major part (79.41%) of pleiotropic FHs, we observed a clearly higher proportion of FHs (20.59%) with inverse effects. More importantly, these inverse-effect FHs are generally capable of improving one of the two traits to a greater extent, meaning they are often more observable and preferable in breeding activities, albeit at the expense of impairing the other trait.



Validation of the effects of key QTGs on fiber quality

To validate the reliability of QTG identification by FH-GWAS, we chose the gene *GhD11G1903*, whose FH diversity showed significant association with fiber quality traits in multiple phenotypes (Fig. 5a). This gene, named *GhFAH1* hereafter, encodes ferulic acid 5-hydroxylase 1, which has previously been reported to affect the biosynthesis of

phenylpropanoids²⁷, and phenylalanine-derived specialized metabolites that are included in the structural components of plant cell walls. There are three missense variants identified in *GhFAH1* (Supplementary Data 18), which derive eight types of FHs in the entire population (Supplementary Table 7). Our analysis identified three major FHs (FH-1, FH-2, and FH-4) for *GhFAH1*; relative to the reference FH in TM-1, the

Fig. 4 | FH-based GWAS identified key QTGs associated with agronomic traits and revealed the genetic basis of variety improvement in cotton. **a** Comparison of the numbers of QTGs and QTLs associated with the phenotype. **b** Proportion of overlapped QTLs and QTGs according to flanking region size. **c** Distribution of QTGs on the sub-genome. **d** Relationship of the relative number of superior FHs (left panel), average effects of all significant FHs (right panel), and phenotypes. Each point represents a sample, point size represents the average effect size of significant FHs. **e** Relative number of superior FHs possessed by cottons registered or introduced in different time periods. **f** Genetic basis of

phenotypic correlations in cotton. Pie charts plot the proportion of FHs with significant pleiotropy effects on FL, FS, and FS, LP (upper panel), and boxplots show the absolute effect sizes for FHs with inverse or isotropic pleiotropic effects (lower panel), n is the number of genes with corresponding pleiotropic effects, two-sided student t -test are used for statistical inference (** $p < 0.01$, ns not significant). In a boxplot, the center line represents the median, the lower and upper box hinges corresponding to the first and third quantiles, whiskers extend to the minimum and maximum values, and individual points beyond the whiskers (if present) denote outliers. Source data are provided as a Source Data file.

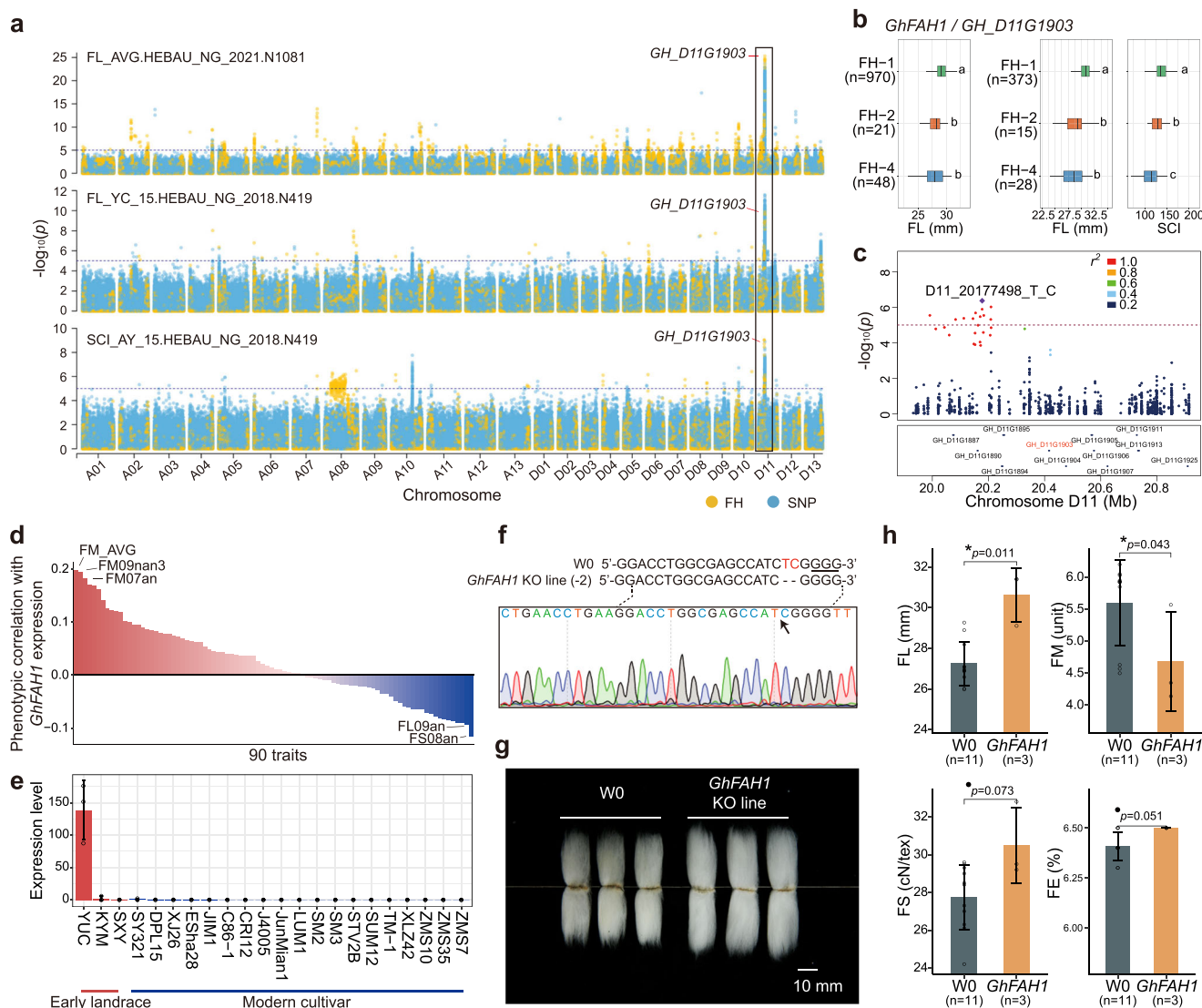


Fig. 5 | CRISPR-Cas9 knock-out (KO) experiment validates the phenotypic effects of GhFAH1 on fiber quality. **a** Association study revealed *GhD11G1903* (*GhFAH1*) as significantly affecting FL and SCI. **b** Haplotype analysis showed the phenotypic effects of different FHs of *GhFAH1*. n is the sample size. The center line represents the median, the lower and upper box hinges corresponding to the first and third quantiles, and whiskers extend to the minimum and maximum values. **c** Colocalization between *GhFAH1* and the candidate QTLs. **d** Correlations between *GhFAH1* expression in 20 DPA fiber and 90 phenotypes in 207 *G. hirsutum* cultivars.

e Expression of *GhFAH1* (TPM) in 20 DPA ovules of several early landraces and modern cultivars. Data were presented as mean values \pm SEM. **f** Sanger sequencing chromatogram illustrating a 2 bp deletion in target *GhFAH1*. **g** Fiber quality evaluation for receptor parent W0 and KO lines. **h** Phenotypes of mature fibers of W0 and KO lines. Two-sided Mann-Whitney U -test was applied for statistical test (* $p < 0.05$, ** $p < 0.01$). Error bar represents the standard error, n is the sample size, data were presented as mean values \pm SEM. Source data are provided as a Source Data file.

other two FHs contain heterozygous and homozygous missense mutations at the same position, resulting in a change of amino acid from glycine to serine. FH haplotype analysis showed significant decreases in FL and SCI for samples carrying the alternative FHs

(Fig. 5b). The gene *GhFAH1* is located in the QTL candidate region represented by lead variant D11_20177498 (Fig. 5c) and neighboring variant D11_20531989 (Supplementary Fig. 11). We also investigated *GhFAH1* expression across the entire TM-1 plant tissues and found it to

be specifically highly expressed in 20 and 25 days-post-anthesis (DPA) fibers (Supplementary Fig. 12). Furthermore, examination of its expression in 20 DPA fibers of 207 *G. hirsutum* accessions revealed subtle yet notable correlations with the fiber traits FL, FS, and FM (Fig. 5d and Supplementary Data 19, 20). In addition, earlier landraces, especially Yucatanense (YUC), showed predominant *GhFAHI* expression in 20 DPA ovules, while its expression was lacking in most modern cultivars (Fig. 5e and Supplementary Data 21). These transcriptomic evidences suggest a potential negative impact of *GhFAHI* on superior fiber establishment. We successfully performed CRISPR-Cas9 editing of *GhFAHI* (Fig. 5f and Supplementary Data 22) that resulted in significantly longer and finer fibers (Fig. 5g, h and Supplementary Data 23).

Finally, we compared our QTG results with previous genetic fine mapping studies and transgenic experiments, which literature provided a total of seven genes with validated functions in fiber development (Supplementary Data 24). These results confirm the broad validity of FH-GWAS as an effective method for identifying key QTGs related to agronomic traits.

Discussion

Large-scale population sequencing and a sophisticated statistical genetics framework have led to major breakthroughs in the life sciences, resolving the genetic mechanisms of complex life activities with unprecedented precision. Although a large number of QTLs have been confirmed in GWAS studies, these results have generally disappointed in terms of driving the discovery of biological causes of phenotypes. Achieving sufficient identification of QTGs from QTLs remains fraught with difficulty. This challenge is particularly prominent in plants, which typically exhibit stronger LD due to extensive artificial selection and complex genome structures²⁸. Recently, the genetic characterization of large populations has provided a new opportunity for overcoming this challenge. In this study, we proposed a strategy in which a multi-dimensional reduction-like analysis²⁹ is performed on the non-synonymous variants present in a gene. In this first FH panel of 3724 cotton accessions representing the most comprehensive upland cotton collections to date, our analysis yielded a total of 1,368,685 FHs representing global cotton protein diversity. Although FH utilizes only about 3.8% of the total genomic information, it is fully capable of reflecting both known structural information, such as in cotton chr. A06 and A08¹⁹, along with adaptive signatures in the genome (Figs. 2e, 3a). Being highly informative, FH and its merged blocks allow more accurate tracing of genetic ancestry for genomic segments (Fig. 3c, d and Supplementary Figs. 4–6) and also reveal additional knowledge about the history of human introductions as well as gene transference events at the gene level (Fig. 3f).

Genetic mapping has always been a central task in applied genetics. Our study leveraged the FH coding strategy and a tailored linear mixed model to achieve efficient narrowing of candidate gene sets and direct mapping of QTGs in some scenarios at the GWAS stage. Compared with conventional GWAS based on short variants, the FH-based GWAS detected fewer significant clumped QTGs (5022) than significant clumped QTLs (6533), as the wider distribution of SNP or short structural variants revealed extra genomic signatures in non-coding regions. In addition, only 25% of the identified QTLs and 35% of QTGs shared the same candidate interval (Fig. 4b). This finding suggests long-range linkage between the significant and causal variant or gene^{23,30}, and additionally illustrates the unique advantage of FH in discovering associations between protein sequence variation and phenotype. That is, although both SNP-GWAS and FH-GWAS could identify unique genetic signatures, apparently the QTGs identified by FH-GWAS are more readily applicable and more amenable to genetic interpretation. We selected *GhFAHI* for validation, FH of which were significantly associated with fiber quality and specifically expressed during the critical development stages of fiber. The impact of this gene

on fiber quality was confirmed by CRISPR-Cas9 gene editing. Taken together, these results demonstrate the feasibility of direct mapping of QTGs through FH-based GWAS. Efficient gene mapping for complex traits in animal and plant species will promote and accelerate the development of breeding 5.0, which leverages big data and artificial intelligence for deep mining and breeding design, and operates at the gene level through transgenic technology (Fig. 6).

In addition to comparing FH-GWAS with SNP-GWAS, the polygenicity of common complex traits reflected in FH-GWAS is also worth attention. For the 20 agronomic traits tested in this study, a seemingly large number of 8916 genes were identified as significantly related to phenotypes (Supplementary Data 15). This result preserves the potential impact of rare variants on phenotypes, which is commonly present in plant mutants. In other words, we did not adopt a very strict screening of FH frequency in the initial association analysis, in light of having applied stringent screening conditions to the initially identified variants to ensure their authenticity. In addition, the effect sizes of significant FHs were consistent with the theoretical normal distribution in the polygenicity framework (Supplementary Fig. 7), and these low-effect but undeniable associations indicate a small genetic contribution of common genetic variants or a highly polygenic contribution involving many variants. Nevertheless, we need to emphasize that addition evaluation is especially required for the application of those significant FH with low frequency under the current strategy. Upon further screening, we identified 532 high-frequency genes with major effects, which are observable in conventional breeding and can be used as reliable references for further breeding practices.

Gene-level FH-based association analysis has laid the foundation for broad and more readily interpretable genetic discoveries. However, as a refined model within the GWAS framework, this method remains susceptible to various forms of cryptic relatedness that may introduce spurious associations. Multi-omics assessment of significantly associated QTGs remains indispensable to refine the genetic rationale of QTG-influenced traits. Besides, it is also important to acknowledge that our analytic framework focuses on genotypic configuration and does not account for haplotype phase information. This modeling strategy introduces theoretical limitations, as heterozygous genotypes within FHs may encode distinct peptide products depending on allelic phase. Consequently, haplotype-specific effects at the gene level may be incompletely resolved. Nevertheless, we should emphasize that the association between phenotypes and FHs containing homozygous non-synonymous mutations under this framework remains valid and appropriate, but the indeterminate classification of heterozygous sites reduces statistical power to distinguish causal FHs from confounding associations, resulting in conservative bias in statistical interpretations. Further inclusion of phase information is essential, while currently there is still a lack of a haplotype-resolved reference genome in cotton, and the polyploid complexity of the species has limited the application of computationally inferred phase information, which risks introducing substantial technical artifacts.

China has achieved remarkable success in improving cotton cultivars and promoting their production. However, in recent years, both the acreage of cotton cultivation and total yield have declined³¹. While ensuring high yields, it is crucial to develop high lint yield cultivars with excellent fiber quality. This study has found that, over the past 30 years, the number of superior FHs related to lint yield have increased significantly (Fig. 4e). Meanwhile, since 2000, superior FHs related to FL have also increased, but those related to FM and FE have decreased in number, which is consistent with the development trend of China's cotton production. At present, the number of superior genes related to lint yield has basically reached a bottleneck, and the number of inferior genes is very limited (Supplementary Data 16). However, there are relatively few superior genes and many inferior genes related to fiber

Big data association studies

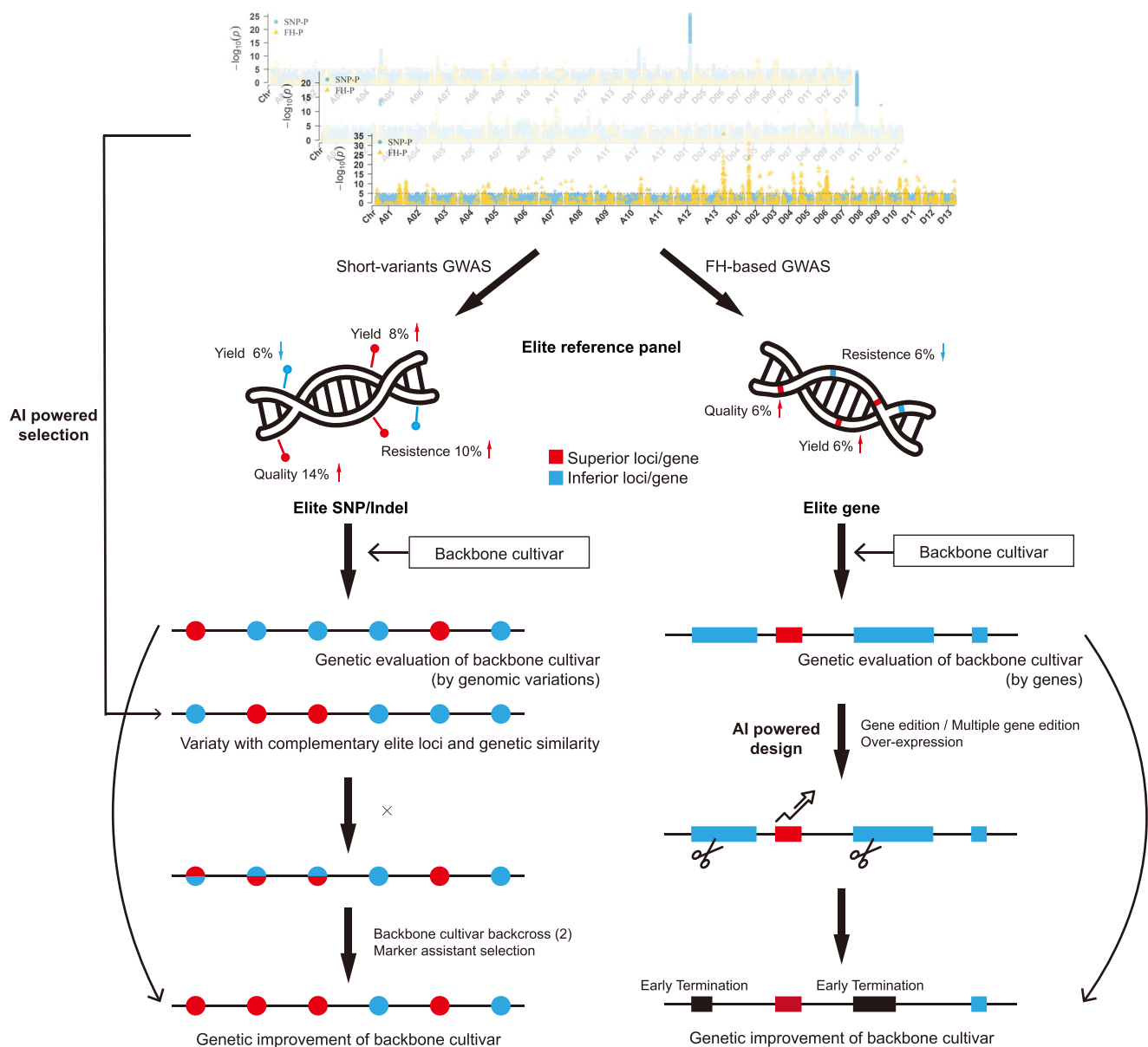


Fig. 6 | Pipeline of intelligent precision breeding (Breeding 5.0).

fineness, elongation rate, and maturity, which could be considered the next main direction for cotton genetic breeding.

Concomitant achievement of high lint yield and superior fiber quality have been the main goal of cotton genetic breeding in the past few decades. However, long-term phenotypic studies and recent genomic analyses have revealed what is now a well-established negative correlation of these traits in both morphology and genetics. Several previous studies have analyzed the genetic basis for this strong negative correlation, examining additive effects of QTL (clusters). These studies have found that most QTLs affecting both lint yield and fiber quality have opposite additive effects, while those affecting multiple fiber quality traits such as FL and FS share the same effect direction^{32–34}. In the present study, our identification of consistent effect direction of QTGs for FL and FS is consistent with previous findings; however, and quite interestingly, we found that most QTGs with pleiotropic effects on LP and FL also have the same effect directions (Fig. 4f), however, these same-direction QTGs have very small effects, while those with inverse directions are relatively fewer, but

their effect in improving one of the traits is very strong. Thus, our findings strongly indicate that there are a large number of genes in the cotton genome that can simultaneously improve both fiber quality and lint yield, but they have rarely been mapped in breeding practice because of their minor effects. Meanwhile, genes that can significantly improve yield or quality are more likely to be selected by breeders, but potentially impair the other trait. These results provide insights into key issues in cotton breeding and suggest a theoretical basis for achieving current major breeding goals.

Methods

Sample collection and calling of genomic variations

In this study, we collected the raw sequencing data of 4480 cottons from eight public datasets (Supplementary Data 1). These cottons represent global cotton origin areas as well as major cultivation areas. The raw sequencing data were trimmed using fastp (ver. 0.23.4) with default parameters³⁵, and the clean reads were mapped against the *G. hirsutum* genetic standard reference genome TM-1¹² using BWA

(ver. 0.7.17-r1188)³⁶. The mapping results were sorted and compressed with samtools (ver. 1.11)³⁷. PCR duplication was marked using sambamba (ver. 0.6.6)³⁸. Genomic variations were identified using the bcftools (ver. 1.11) “mpileup” and “call” commands in batch³⁹. Variants with mapping quality (QUAL) lower than 30 were discarded. Indel markers from different batches were left-aligned and normalized with the bcftools “norm” command to avoid ambiguity in the merge step, and all datasets were merged using the bcftools “merge” command. Variant imputation was conducted by Beagle (ver. 5.4)⁴⁰. After quality control for the genomic variations (biallelic, mapping quality >30, heterozygosity <0.1, number of homozygotes for alternative allele >10), a total of 4480 accessions and 9,052,728 variants were retained. Samples with different sequencing depths were evaluated by principal component analysis (PCA) and ancestry estimation. After evaluation of the merged dataset, we removed samples with low sequencing depth, and variants with an inbreeding coefficient lower than −0.6 and MAF lower than 0.001. Finally, 3724 samples with 23,057,253 genomic variants were retained (Supplementary Data 3).

Construction of the FH dataset

For the 23,057,253 genomic variants, we annotated their effects using snpEff (ver. 5.1)⁴¹ and extracted a total of 334,406 non-synonymous variants that directly alter protein sequences (Supplementary Data 25). In cases where multiple missense variants were present, the genotype combination in each sample was extracted as the FH genotype of that sample for that gene. To illustrate, if the genotype of FH-5 for a gene is 022, it indicates the presence of three non-synonymous mutations. The additive genotypes for these three variants are 0, 2, and 2, respectively, indicating that the sample is homozygous for the reference allele in the first variant and homozygous for the alternative allele in the other two variants. Based on this principle, the FH for every gene in the reference genome TM-1 will always be a combination of 0s (referred to as FH-1). This facilitates estimation of the phenotypic effects of different FHs. Notably, the genotype coding is changed from the numeric method used for SNPs or Indels that counts allele number, such as 0, 1, 2, to the factorial method, for example, FH-10 represents type ten for this gene.

Genetic diversity and selection pressure analysis

Shannon's equitability index (E_H) is commonly used to characterize the distribution equitability of species with given abundance^{42,43}. In this study, we use E_H to measure the gene diversity across different chromosomes and populations. For a given gene, the E_H is calculated as follows:

$$E_H = -\frac{1}{\ln N} \sum_i p_i \ln p_i \quad (1)$$

in which p_i is the prevalence of the i -th FH for this gene in the population, and N is the population size. The E_H ranges from 0, indicating this gene is sequence-conserved as only one reference FH (FH-1, $p_i = 1$) exists in the population, to 1, indicating this gene is extremely variable because every sample in the population has a unique FH ($p_i = 1/N$). For genes with varying diversity across populations, we investigated their potential biological functions by gene ontology (GO) enrichment analysis using the R package clusterProfiler (ver. 4.6.2)⁴⁴.

To quantify selection pressure, we calculated the ratio of non-synonymous substitutions and synonymous substitutions (K_a/K_s)⁴⁵ for each gene:

$$K_a/K_s = \frac{\sum_i N_i Q_{i-\text{variant}} / Q_{i-\text{allele}}}{\sum_i N_i X_{i-\text{variant}} / X_{i-\text{allele}}} \quad (2)$$

in which N_i is the number of samples with the i -th FH, $Q_{i-\text{variant}}$ and $X_{i-\text{variant}}$ denote the number of non-synonymous and synonymous

variants in the i -th FH, and $Q_{i-\text{allele}}$ and $X_{i-\text{allele}}$ represent the number of non-synonymous and synonymous alleles. When K_a/K_s is equal to 1, it indicates the gene is subject to neutral selection, while K_a/K_s greater than 1 or less than 1 represents positive selection and purifying selection, respectively.

Phylogenetic analysis and ancestry estimation

We constructed phylogenetic trees from both genomic markers and FH markers. For the regular genomic variants, PCA, phylogenetic analysis, and ancestral estimation were conducted on the LD-pruned dataset comprising 43,602 independent variants. These variants were extracted on the basis of having pairwise correlations lower than 0.05 with any other variants located within a 100 kb window. The LD-pruning step and PCA analysis were conducted using the PLINK2 (ver. 2.00a6LM)⁴⁶ software with respective input parameters “indep-pairwise 100 kb 0.05” and “--pca”. The neighbor-joining tree was constructed for the large genomic variant dataset using the FastTree (ver. 2.1.11SSE3)⁴⁷ with default parameters. Ancestral estimation for all samples was conducted using Admixture (ver. 1.3.0)⁴⁸ with default parameters. For the FH dataset, the genotype coding being factorial rather than numeric made the previous method not applicable. Instead, we firstly calculated the general genetic distance

$$d = D/L \quad (3)$$

in which D is the number of genes that possess different FH between two individuals and L is the total number of genes, along with its associated variance $D(1-D)/L$. The general genetic relatedness score is then calculated by $1-d$. Then we used the general relatedness matrix to construct the neighbor-joining tree. These two steps were implemented using the functions “dist.genet()” and “nj()” in the R package ape (ver. 5.8)⁴⁹.

Gene transference intensity

Our study measured the level of pairwise subpopulation gene transference caused by natural factors or human activities by calculating the relative number of genes sharing the same type (i.e., FH). Specifically, the gene transference intensity (G_s) between subpopulations A and B was calculated as

$$G_s = \frac{1}{M} \sum_i w_i n_i \quad (4)$$

in which M is the total number of genes; n_i is the number of genes specifically share same FH between i subpopulations; i ranges from 2, representing the gene share the same FH only between subpopulations A and B, which provides the most direct evidence of gene transference between those subpopulations, to the maximum number of subpopulations, representing gene shares same FHs across the entire population; and w_i is the weight score

$$w_i = e^{-\frac{(i-2)^2}{10}} \quad (5)$$

which considers the contribution of genes with same FH shared in different numbers of subpopulations to determine gene transference in the pair. When i is 2, w_i reaches its maximum value of 1, and as i increases, it decreases. When i is higher than 8, w_i is 0.027, meaning that FHs occurring in more than eight subpopulations are not informative in revealing the gene transference between subpopulations A and B. After all G_s were computed, they were further normalized through division by the intra-population G_s value.

FH-GWAS modeling

We performed linear mixed model-based GWAS to identify candidate QTGs for the agronomic traits of interest in cotton. The statistical

model is described as:

$$y = X_c \beta_c + F_j g_j + r + e \quad (6)$$

where y is an $n \times 1$ vector of phenotypic value for n samples; X_c is the incidence matrix of fixed covariates, which includes the intercept coefficient 1 and the top two PCs, with their corresponding effects β_c ; g_j is a vector of genetic effect captured by the tested gene with $g_j \sim N(0, I\sigma_g^2)$, in which σ_g^2 is variance; and F_j is the $n \times m$ coding matrix for the j -th gene. For sample i that carries the factorial FH q in the j -th gene, q was coded to the dummy vector F_{ij} by setting the q -th element in an all-zero vector to 1,

$$F_{ij} = [0_1, 0_2, \dots, 0_{q-1}, 1, 0_{q+1}, \dots, 0_m] \quad (7)$$

in which m is the total number of FH types observed for the tested gene. Additionally, r is a vector of genetic effect captured by genetic relatedness with $r \sim N(0, G\sigma_r^2)$, the covariance matrix of r is $G\sigma_r^2$, meaning the variance of each individual's genetic effect is scaled by their kinship coefficient (diagonal of G , the general relatedness matrix applied for phylogenetic analysis), and the covariance between individuals is determined by their pairwise genetic relatedness (off-diagonal of G); and e is the residual vector with $e \sim N(0, I\sigma_e^2)$. To reduce the computational burden, we firstly fitted the following model

$$y = X_c \beta_c + r + e \quad (8)$$

and calculated the reduced phenotype \tilde{y} as

$$\tilde{y} = y - \hat{r} \quad (9)$$

in which \hat{r} is the vector of estimated effect of genetic relatedness. Then, two models with or without the tested gene were fitted, the null model (H_0) is

$$\tilde{y} = X_c \beta_c + e \quad (10)$$

and alternative model (H_1) is

$$\tilde{y} = X_c \beta_c + F_j g_j + e \quad (11)$$

The likelihood ratio test was applied to the two models to determine whether the variance σ_g^2 of g_j is significantly greater than 0 (one-sided alternative), with degree of freedom 1. A suggestive threshold of $1e^{-5}$ was applied for determining the significance of the test gene in affecting the phenotype. The GWAS analysis was conducted by the R package lme4qtl (ver. 0.2.2)⁵⁰.

SNP and indel GWAS

The filtered SNP and Indel set in VCF format are converted to variant-major text genotype table (tped) using PLINK2, and GWAS was conducted by software efficient mixed model association eXpedited (EMMAX, ver. 20120210)⁵¹, with genetic kinship matrix and the first two principal components as covariates. The statistical analysis tests the significance of the variant's effect using a two-sided generalized least squares F -test, and a suggestive threshold of $1e^{-5}$ was taken as the threshold.

Clumping of QTLs and QTGs

Significant genomic variants (SNPs and Indels) and genes were clumped in order to take LD into account in the results interpretation. For significant variants, any sites less than 1 Mb away from an index variant (variant having p value lower than $1e^{-5}$ and greater than 0.1 correlation with it were assigned to that index variant's clump; this clumping step was conducted using the PLINK2 "--clump" command,

with the current population as the reference panel. For genes significantly associated with phenotypes in the FH-based GWAS, a simple clumping was conducted by keeping the most significant gene while removing any other significant genes within 500 kb.

Calculation of FH effects on phenotypes

For a given QTG identified as significantly associated with a phenotype, we calculated the phenotypic effects of the FHs it contains across the population. The FH in the reference genome TM-1 was always defined as FH-1 and used as the base FH to which average phenotypes of the other FHs (such as FH-2, FH-3, etc.) were compared. FHs appearing in less than two samples were removed. The average phenotypic value among samples containing the FH- q was calculated as

$$\bar{y}_q = \frac{1}{n_q} \sum_i y_{qi} \quad (12)$$

and the effect size of FH- q was defined as the difference between \bar{y}_q and \bar{y}_1 . Multiple comparisons of phenotype for samples with different FHs were conducted by means of the least significant difference (LSD) test, with p values adjusted by the Benjamini–Hochberg method. Genes of high breeding value were screened based on a frequency of occurrence of more than 30 in the population, and an effect size of less than 5% quantile or more than 95% quantile for the corresponding phenotype.

Validation of biological functions of target QTGs

CRISPR/Cas9 genome editing was used to generate knock-out lines for *GhFAHI*. The sgRNA sequences (5'-GGACCTGTGGAGCCATCTCG-3') were designed by the web tool CRISPR-P 2.0 and cloned into the sgRNA-Cas9 expression vector (with oligonucleotides forward sequence 5'-ACGGCCCCGTAATGAGCGAAA-3', and reverse sequence 5'-CACCTCAAGATGGGGTTCCT-3'), then introduced into *G. hirsutum* accession W0 via *Agrobacterium tumefaciens*-mediated transformation. Primer pairs bracketing the target regions were used to amplify the DNA of the transgenic plants, and the derived PCR products were sequenced to confirm mutation of the target genes.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data were collected from the following NCBI accessions: PRJNA605345, PRJNA680449, PRJNA414461, PRJNA564187, PRJNA744011, PRJNA752720, PRJNA375965, PRJNA336461, PRJNA613140, and PRJNA848197. The phenotype data were collected from publications listed in Supplementary Data 1. The reference genome and annotation files of TM-1 were downloaded from COTTONOMICS (<http://cotton.zju.edu.cn/>). The coding sequence of *GhFAHI* has been deposited into NCBI GenBank under accession. PV231350 [<https://www.ncbi.nlm.nih.gov/nuccore/PV231350>]. Source data are provided with this paper.

Code availability

The R scripts for converting genomic variants into FH, and all other analysis scripts, codes for the results present in this study, were deposited in the GitHub repository [<https://github.com/GuoanQi1996/FHGenotyping>].

References

1. Crouch, D. J. M. & Bodmer, W. F. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc. Natl Acad. Sci. USA* **117**, 18924–18933 (2020).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

3. Investigators, T. A.o.U. R. P. The “all of us” research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
4. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
5. Weigel, D. & Mott, R. The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
6. Meadows, J. R. S. et al. Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. *Genome Biol.* **24**, 187 (2023).
7. Zhang, G. Bird sequencing project takes off. *Nature* **522**, 34–34 (2015).
8. McCarthy, M. I. & Hirschhorn, J. N. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* **17**, R156–R165 (2008).
9. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
10. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
11. Belkadi, A. et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl Acad. Sci. USA* **112**, 5473–5478 (2015).
12. Hu, Y. et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **51**, 739–748 (2019).
13. Ma, Z. et al. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat. Genet.* **53**, 1385–1391 (2021).
14. Wang, M. et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2019).
15. Zhang, T. et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531–537 (2015).
16. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).
17. Fang, L. et al. Divergent improvement of two cultivated allotetraploid cotton species. *Plant Biotechnol. J.* **19**, 1325–1336 (2021).
18. Han, Z. et al. Genomic insights into genetic improvement of upland cotton in the world’s largest growing region. *Ind. Crops Prod.* **183**, 114929 (2022).
19. He, S. et al. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nat. Genet.* **53**, 916–924 (2021).
20. Ma, Z. et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **50**, 803–813 (2018).
21. Wang, M. et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
22. Yuan, D. et al. Parallel and intertwining threads of domestication in allopolyploid cotton. *Adv. Sci.* **8**, 2003634 (2021).
23. Zhu, T.-N. et al. An atlas of linkage disequilibrium across species. Preprint at *bioRxiv* 2024.2009.2024.614726 (2024).
24. Cheng, Y. et al. *Gossypium purpurascens* genome provides insight into the origin and domestication of upland cotton. *J. Adv. Res.* **56**, 15–29 (2024).
25. Li, Y. et al. Genomic insights into the genetic basis of cotton breeding in China. *Mol. Plant* **16**, 662–677 (2023).
26. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
27. Anderson, N. A., Bonawitz, N. D., Nyffeler, K. & Chapple, C. Loss of FERULATE 5-HYDROXYLASE leads to mediator-dependent inhibition of soluble phenylpropanoid biosynthesis in *Arabidopsis*. *Plant Physiol.* **169**, 1557–1567 (2015).
28. Demirjian, C., Vailleau, F., Berthomé, R. & Roux, F. Genome-wide association studies in plant pathosystems: success or failure?. *Trends Plant Sci.* **28**, 471–485 (2023).
29. Xu, H.-M. et al. Multivariate dimensionality reduction approaches to identify gene-gene and gene-environment interactions underlying multiple complex traits. *PLoS ONE* **9**, e108103 (2014).
30. Huang, X. et al. Efficient estimation for large-scale linkage disequilibrium patterns of the human genome. *eLife* **12**, RP90636 (2023).
31. National Bureau of Statistics. Announcement on cotton production in 2023. https://www.stats.gov.cn/xxgk/sjfb/zxfb2020/202312/t20231225_1945749.html (2023).
32. Diouf, L. et al. QTL mapping of fiber quality and yield-related traits in an intra-specific upland cotton using genotype by sequencing (GBS). *Int. J. Mol. Sci.* **19**, 441 (2018).
33. Liu, X. et al. Enriching an intraspecific genetic map and identifying QTL for fiber quality and yield component traits across multiple environments in upland cotton (*Gossypium hirsutum* L.). *Mol. Genet. Genomics* **292**, 1281–1306 (2017).
34. Zhang, Z. et al. Genome-wide quantitative trait loci reveal the genetic basis of cotton fibre quality and yield-related traits in a *Gossypium hirsutum* recombinant inbred line population. *Plant Biotechnol. J.* **18**, 239–253 (2020).
35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
39. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-Science* **10**, giab008 (2021).
40. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
41. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
42. Magurran, A. E. *Measuring Biological Diversity* (Wiley-Blackwell, 2013).
43. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
44. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
45. Nekrutenko, A., Makova, K. D. & Li, W. H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12**, 198–202 (2002).
46. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
48. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
49. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).
50. Ziyatdinov, A. et al. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* **19**, 68 (2018).

51. Kang, H. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

Acknowledgements

This work was financially supported by the Xinjiang Production and Construction Corps (2023AA008, T.Z.), the National Natural Science Foundation of China (32130075, T.Z.), and the Fundamental Research Funds for the Central Universities (226-2022-00100, T.Z.).

Author contributions

T.Z. and G.Q. conceived and designed the project. Y.L., G.Q., and Z.H. collected the raw sequencing data and conducted the initial sequence analysis, Y.L. and Z.H. inspected the phenotype data. Y.H., J.C., Z.Z., and L.F. collected samples for transcriptomic analysis. G.Q. conducted the data analysis and interpretation, W.Z., Z.H., L.X., R.C., and Y.H. performed the knock-out experiment and other biological experiments. G.Q., Y.L., and T.Z. initiated the draft of the manuscript. All authors have reviewed the final version of the manuscript and approved it for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59983-w>.

Correspondence and requests for materials should be addressed to Tianzhen Zhang.

Peer review information *Nature Communications* thanks Xiangyang Lou, Zhengsheng Zhang and Zhiwu Zhang for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025