Cotton Genomics and Genetics 2025, Vol.16, No.5, 210-221
http://cropscipublisher.com/index.php/cgg

Research Insight Open Access

Pan-Genome Analysis Reveals Genetic Diversity and Subgenome Dominance in Cotton

Zhen Li
Hainan Institute of Biotechnology, Haikou, 570206, Hainan, China
Corresponding email: zhen.li@hibio.org

Cotton Genomics and Genetics, 2025, Vol.16, No.5 doi: 10.5376/cgg.2025.16.0021
Received: 01 Jul., 2025
Accepted: 11 Aug., 2025
Published: 01 Sep., 2025

Preferred citation for this article:
Li Z., 2025, Pan-genome analysis reveals genetic diversity and subgenome dominance in cotton, Cotton Genomics and Genetics, 16(5): 210-221 (doi: 10.5376/cgg.2025.16.0021)

Abstract The pan-genome concept has emerged as a powerful framework for understanding genome variability within a species, providing crucial insights into genetic diversity, adaptation, and evolution in plants. In this study, we review the landscape of cotton (*Gossypium* spp.) genomes through the lens of pan-genomics, with a particular focus on the role of polyploidy and subgenome dynamics. We explore the structural evolution of diploid and polyploid cotton genomes, the composition of core and dispensable genes, and the presence of lineage-specific genes and structural variants across cultivars and wild relatives. Our analysis highlights how pan-genome studies have uncovered key agronomically relevant genes absent in reference genomes and revealed extensive gene presence/absence variation (PAV), SNPs, InDels, and CNVs that contribute to trait diversity. We also examine expression bias and subgenome dominance in allopolyploid cotton, revealing regulatory asymmetries that influence fiber development, stress responses, and reproductive traits. A focused case study on *Gossypium hirsutum* demonstrates the integration of genomic data from diverse accessions and the discovery of elite trait-associated genes. Finally, we discuss the implications of cotton pan-genomics for molecular breeding, biotechnology, and the development of high-yield, stress-tolerant varieties. This review underscores the transformative potential of pan-genome resources in shaping next-generation cotton improvement strategies.

Keywords Cotton pan-genome; Subgenome dominance; Genetic diversity; Polyploidy; *Gossypium hirsutum*

1 Introduction

The traditional reference genome has indeed played a significant role in the past, and many fundamental researches have also started with it. But then again, the content it can cover is, after all, limited. Especially when it comes to genetic differences among different individuals, a single version seems insufficient. Some information is not even in the reference sequence at all, which is one of the reasons why the new concept of "pan-genome" was proposed later. The pan-genome is actually not difficult to understand. To put it simply, it no longer focuses solely on a representative genome but takes into account the genetic information of multiple individuals within an entire species. In addition to the "core genes" that can be found in all materials, it also includes those "unique" or "rare" variant genes that only occur in some individuals. This integration approach has caused a significant stir in plant research, especially in identifying structural variations (such as chromosomal inversions, insertions, and deletions) and presence/deletion variations (PAVs), which has indeed enhanced the efficiency of discovery. Some new genes that had never been noticed before were precisely added through this method. Compared with the previous approach of "one reference genome for the final analysis", the pan-genome undoubtedly offers a more comprehensive perspective. It not only enables us to see genes themselves more comprehensively, but also helps researchers better understand genetic diversity, the evolutionary process of species, and even the genetic basis of some important agronomic traits (Ma et al., 2021; Huang et al., 2024).

Among so many research subjects, cotton is almost the "chosen one". Apart from being a major global cash crop in itself, its genetic background is also very representative. Cotton has both diploid and allopolyploid groups, and this structure provides excellent material for the study of genomic evolution. Especially for polyploid species like hirsutum (*G. hirsutum*) and barbadense (*G. barbadense*), the two subgenomes in their bodies-usually called At and Dt-are not simply pieced together. During the actual development process, there is a very complex mutual regulatory relationship between these two sets of subgenomes. Moreover, under the long-term breeding selection

of humans, this interaction pattern has been constantly strengthened or changed, ultimately influencing their genomic structure and trait expression today. Ultimately, cotton is actually an ideal "observation window" for studying polyploidy, subgenomic dominant expression, and even changes in genetic mechanisms (Mei et al., 2004; Wang et al., 2017).

For this reason, we have decided to systematically sort out the research progress on the cotton pan-genome in recent years. This research will approach from three perspectives: First, the genetic diversity map brought about by pan-genome and structural variations; The second is the regulatory mechanism behind the asymmetry of subgenomic dominance and expression in polyploid cotton; Thirdly, how these discoveries can help us better understand the evolutionary process of plant genomes and even guide practical breeding, especially in improving fiber quality, increasing yield and resistance. Another point worth mentioning is that we will also combine some integrated cases from genomics, transcriptomics and epigenetics, attempting to illustrate from a multi-omics perspective that the pan-genome is not only about "discovering more", but is also gradually reshaping our overall understanding of cotton biology and providing new breakthroughs for precise improvement.

2 The Cotton Genome Landscape

2.1 Genome structure and evolution in diploid and polyploid cotton species

Among the cotton genus plants, there are quite a few species-approximately 45 diploids and 7 allotetraploids. This quantitative diversity is underpinned by differences in genomic structure, especially between the A genome and the D genome. The volume of the A genome is almost twice that of the D genome. This difference is mainly not due to an increase in genes, but rather the proliferation of transposition factors (Pan et al., 2020). But if you think they are far apart, that's not entirely the case. The A and D genomes are generally consistent in terms of gene arrangement sequence and collinearity (Page et al., 2013), indicating that the parts that make the genome "larger" are mostly repetitive sequences rather than the core gene content. When it comes to heteropolyploids, such as upland cotton and island cotton, their emergence is not a "recent event"-dating back approximately one to two million years. AT that time, the A and D genomes of the ancestors combined and eventually left the "traces" of the AT and DT subgenomes in one cell nucleus (Hu et al., 2019). However, this kind of combination is not a "harmonious" matter. After polyploidy, it is often accompanied by chromosomal structural disorder, such as inversion, translocation, etc. Duplicated genes do not necessarily "stay well"-some retain the original function, some lose it, and some simply develop new functions (Li et al., 2016).

2.2 Role of polyploidization in shaping gene content and genome complexity

The reason why the genome of cotton is becoming increasingly complex is, in most cases, closely related to the issue of "polyploidy". However, this was not caused by a single mutation; it has gone through several rounds of evolution. Such as sixfold transformation was an earlier event, and then there was another tenfold transformation that occurred in the later stage of evolution (Strygina et al., 2020). This repeated doubling of the genome not only increases the number of genes, but also has a deeper impact on gene regulation-the operation mode of the entire "control system" changes accordingly. But changes are not always symmetrical or fair. Some subgenomes have greater say in certain traits and show "dominance", such as being more easily expressed; Some, however, remain relatively low-key and less active. This bias causes the expression levels of certain genes to be significantly higher than those of others. Meanwhile, the chromatin structure may also have undergone a rearrangement. For instance, the improvement of traits such as fiber quality is likely related to higher expression of certain subgenomes (Figure 1) (Chen et al., 2020; Huang et al., 2021; Han et al., 2022). Of course, not all such "biased expression" is necessarily a good thing. Whether it is good or not depends on which gene it affects and what the environmental conditions are like.

2.3 Comparative genome assemblies of key cotton cultivars and wild relatives

To understand exactly where modern cultivated cotton came from, it's no use just focusing on one variety. You need to bring along its "relatives" to take a look together. Especially those wild relatives, they actually contain a lot of genetic information that has been "filtered" out by modern varieties. With the advancement of sequencing technology, representative species such as upland cotton, island cotton, Asian cotton, and Raymond cotton now

have high-quality, chromosome-level assembly data (Zhang et al., 2015). After putting them together for comparison, quite a few differences were found-especially structural variations, such as inversions and translocations, which are "gene migrations", frequently occurring in the comparisons. It is precisely these variations that may explain the differences in trait manifestations among different cottons (Wang et al., 2018). Furthermore, pan-genome analysis and mapping methods further indicate that the cotton genome is not monolithic. Some areas are the "core areas" that all materials have, while others are the "variable modules" that vary by variety. This "jigsaw puzzle feeling" has enabled us to have a clearer understanding of the genetic composition of modern cotton. More importantly, these genomic data are not only "databases" for scientific research, but have also been actually used to identify key QTL loci related to fiber traits and stress resistance, directly serving precision breeding and variety improvement. In other words, with these comparative assembly and map resources, breeding no longer relies on guesswork but is based on evidence.

3 Insights from Pan-Genome Analyses in Cotton

3.1 Core, dispensable, and private genes: definitions and biological relevance

In pan-genome analysis, classifying genes into three categories is actually to better explain the distribution of genes in different germplasm materials. One type is the "core genes" that almost everything is carried, another type is the "optional genes" that some have and some don't, and there are also those "private genes" that only appear in one or two materials. In upland cotton (*G. hirsutum*), the proportion of these three types of genes is very interesting: the "soft core gene", which can be found in almost all germplasms, accounts for 97% to 100%. The coverage of "shell genes" is much broader, ranging from 1% to 97%. As for "cloud genes", they only appear in less than 1% of germplasms. The functions of these genes are quite different. Generally speaking, core genes are the "essential modules" for maintaining basic life activities. However, those uncommon genes are closely related to phenotypic differences, stress resistance and even environmental adaptation (Wang et al., 2022). In other words, the genetic differences observed in many cotton populations are actually caused by these "sometimes present and sometimes absent" genes.

3.2 Discovery of lineage-specific genes and structural variants

The reference genome of upland cotton actually does not cover all the genetic content. Through pan-genome analysis, researchers have identified over 30 000 new genes that were previously unrecorded. Even when it comes to sea island cotton, this number is nearly 9 000. These genes have a characteristic: they are not "present in everyone", but have lineage specificity-that is to say, they only appear in a certain branch of cotton and are not present in other cotton species at all. Genes like this that "exist only in a part of the cotton family" are completely invisible using the traditional single reference genome method. It is precisely because the pan-genome has incorporated more samples that it can identify these "hidden members". In addition to finding genes, the pan-genome also revealed very large-scale structural variations, such as SVS like large insertions, deletions and inversions, with a number of more than 290 000 (Li et al., 2024). Part of it is related to the process of cotton domestication, population differentiation and even trait improvement. Some structural variations also involve reproductive isolation between different lineages or parallel selection of fibrous traits. These contents sound quite "genomic", but they are very crucial for understanding how the traits of cotton come about.

3.3 Functional categorization of variable genes linked to agronomic traits

At first, many people thought that what could affect the yield or the quality of the fibers must be those core genes that "every cotton has". But with more research, this understanding is also changing. Nowadays, many key variations related to traits are actually identified through large-scale screenings such as GWAS (Genome-wide Association Studies) and functional annotations (Jin et al., 2023). For instance, 124 PAVs related to traits (i.e., gene fragments that are absent in some individuals and present in others) have been found to be directly linked to important indicators such as fiber quality and yield. Not only that, but new QTL loci have also been discovered for traits such as flowering time and fiber strength, which were previously not given much attention (Joshi et al., 2023). What's more interesting is that many of these "useful variations" are not in the core part of the mainstream reference genome at all, but are hidden in some so-called "non-essential" genes-that is, those variable genes or

private genes that only exist in certain materials. These seemingly marginal genes are actually often related to differences in plants' responses to adverse conditions, their adaptability to the environment, or regulatory mechanisms. In other words, those genes that we originally thought were unimportant have instead become the key to enhancing the resistance and breeding potential of cotton. Their value is far greater than initially thought.

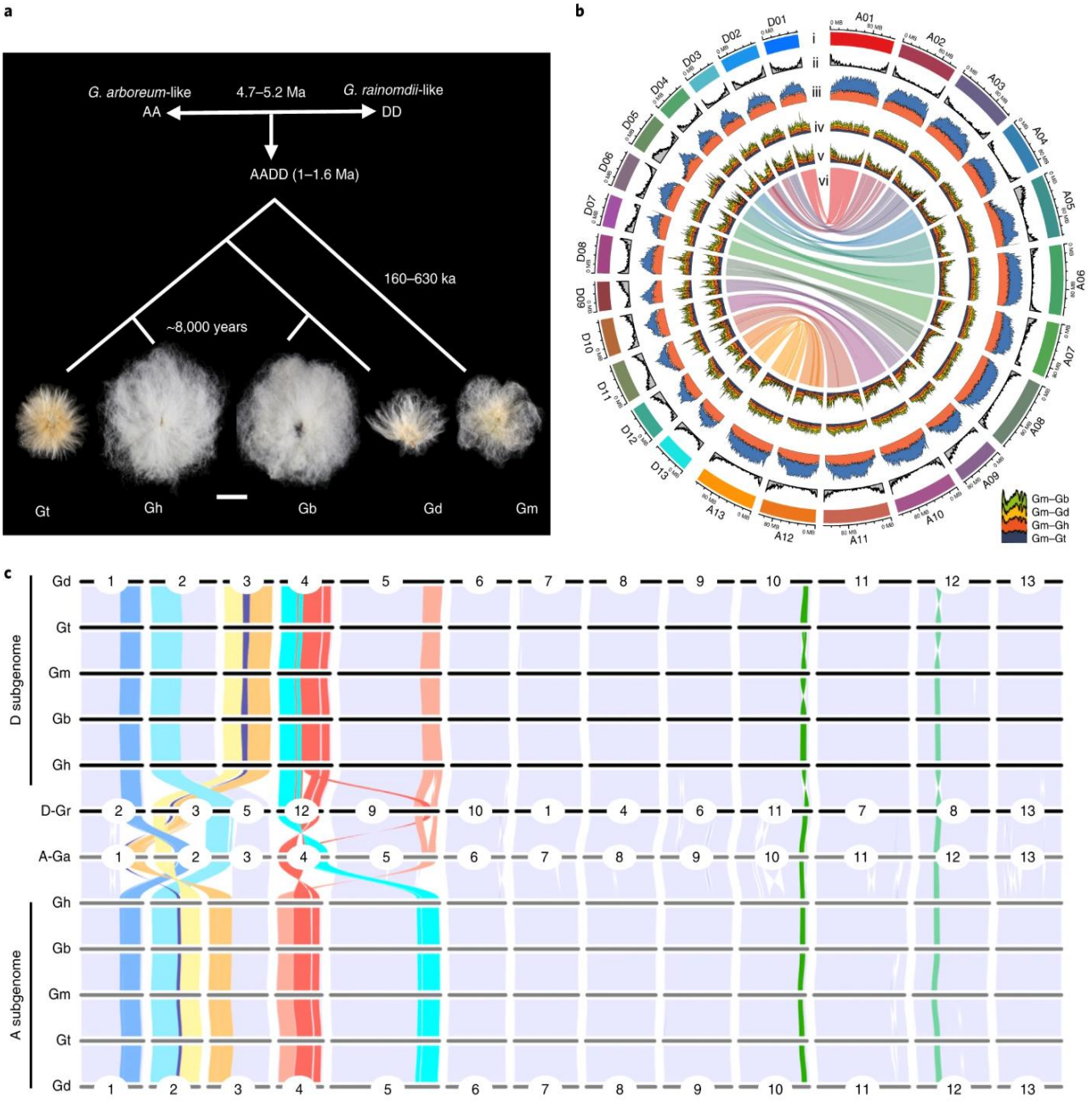

Figure 1 Sequencing features of five cotton allotetraploid species (Adopted from Chen et al., 2020)

Image caption: a, Evolution and domestication of five polyploid lineages, Gh, Gb, Gt, Gd and Gm, after polyploidization between an A-genome African species (Ga-like) and a D-genome American species (Gr-like). Typical seeds from each species are shown. The divergence time estimates are based on 21 567 single orthologs among the 5 species by using the synonymous substitution rate (*r*) of 3.48 × 10-9. Scale bar, 10 mm; ka, thousand years ago. b, Chromosomal features and synteny of the Gm genome. Notes in circos plots: (i) estimated lengths of 13 A and 13 D homoeologous pseudochromosomes; (ii) distribution of annotated genes; (iii) TE content (*Gypsy*, steel blue; *Copia*, grey; other repeats, orange); (iv,v) stacked SNP (iv) and indel (v) densities between Gm and Gb, Gd, Gh and Gt, respectively (see inset), and (vi) syntenic blocks between the homoeologous A and D chromosomes. The densities in plots in (ii)-(v) are represented in 1 Mb with overlapping 200-kb sliding windows. c, Genome-wide syntenic relationships among A and D subgenomes in five allotetraploids relative to the A-genome-like Ga (A2 genome) and D-genome-like Gr (D5 genome). Structural variations among syntenic blocks are marked with colored ribbons (Adopted from Chen et al., 2020)

4 Genetic Diversity Unveiled by the Cotton Pan-Genome

4.1 Contribution of wild species and landraces to gene pool expansion

Nowadays, there are indeed many varieties of cotton being cultivated, and it seems that the genetic resources are sufficient. But if we really want to explore genetic diversity, many "treasures" are actually hidden in those marginalized materials-such as wild species and local old breeds. These materials may no longer be seen in the fields, but the set of "old genes" within them is precisely what modern varieties lack. Many studies, after analysis using molecular markers such as SSR and SNP, all point to a similar conclusion: Genetic differences between species far exceed those within cultivated varieties (Gurmessa et al., 2024; Arslan et al., 2025). The meaning is quite simple-if wild resources are excluded, it will instead limit the breadth of the entire gene pool. When it comes to improving properties, these "obscure materials" often play an unexpected role. Through means such as population mapping and backcrossing, those "rare alleles" that originally existed only in wild species were gradually introduced into cultivated species, not only broadening the genetic background, but also opening up new ideas in directions such as fiber quality improvement, disease and stress resistance (Van Deynze et al., 2009). They haven't disappeared; they just haven't been used in the right places all along.

4.2 SNPs, InDels, and CNVs as markers of intraspecific variation

There are more than one marker used to identify the differences among cotton varieties, but SNP, InDel and CNV are the three structural variations that are currently the most widely used and effective. Take the CottonSNP63K chip as an example. It can simultaneously detect tens of thousands of SNPS, significantly enhancing the discrimination between varieties and also assisting in identifying key regions related to agronomic traits (Hinze et al., 2017; Yang et al., 2019). However, the "battlefield" of mutations does not always occur in the coding area. Some Indels and SNPS are actually located in regulatory or non-coding regions and have a significant impact on gene expression (Shen et al., 2017). There are also types such as CNV and PAV involving large fragment structural changes, which not only expand the diversity of genomic structures but also intensify the trait differences among different materials (Song et al., 2018). These seemingly "trivial" variations are actually the key markers that distinguish the diversity of cotton germplasm.

4.3 Implications of gene presence/absence variation for trait diversity and evolution

Gene deletion is not necessarily a bad thing-in cotton, this "presence/absence variation" (PAV) has instead become one of the main causes of the diversity and adaptive evolution of many traits. Such variations are usually enriched in genes related to environmental adaptation, reproduction and fiber development, and their associations with multiple key traits have been clarified through GWAS and QTL mapping (Sun et al., 2017). Interestingly, PAV is not only an important driver in natural evolution but is also regarded as a "hidden tool" in the process of human domestication. The drought and disease resistance traits of some high-quality fibers are actually manifested precisely because of the "absence" of certain genes (Ma et al., 2018). This also explains why pan-genome research is becoming increasingly important in cotton improvement-it makes us realize that to find key variations, we cannot just focus on "whether there are or not", but also need to look at "whether there are deficiencies".

5 Subgenome Dominance and Expression Bias in Cotton

5.1 Evidence for subgenome expression asymmetry in allopolyploid cotton

In allopolyploid cotton, the performance of the two subgenomes -At and Dt- is not always evenly matched. Early studies have found that the At subgenome often steals the show in terms of expression levels, especially during the early stage of fiber development, when this advantage is more pronounced (Naoumkina et al., 2014). However, this bias is not static. For instance, some AT-derived genes are particularly crucial in the process of fiber elongation; Meanwhile, Dt is not completely silent. Some of its activating genes also play a significant role in yield traits and stress responses (Peng et al., 2020). That is to say, the phenomenon of expression asymmetry is more like a "dynamic balance"-in different tissues, at different developmental stages, and even in different environments, the expression contributions of AT and Dt may be reversed.

5.2 Regulatory mechanisms underlying biased expression and gene retention

It is no coincidence that some genes in the At subgenome are always more active than those on the Dt side. If you really want to trace back to the root cause, most of the time you have to start from the aspect of "regulation". Take

promoters for example. They are the key parts that control gene switches. Even in a small area inside, such as the TATA-box, as long as there is a variation, the strength of the expression can immediately create a gap. PRE1 is a very typical example. It affects fiber elongation and is particularly "talkative" on the At side, probably because the promoter part is designed more appropriately (Zhao et al., 2018). Of course, merely looking at the changes at the DNA sequence level is not enough. The state of chromatin can also affect whether genes can be "read". Studies have found that on the At subgenome, the contents of histone modification markers symbolizing activity (such as H3K4me3) are relatively high, while those representing suppression (such as H3K27me3) are relatively low (Zhang et al., 2021). That is to say, the genes on the At side are more likely to encounter the "green light" on the "allowed expression" channel. Another point that is often overlooked by people is that the three-dimensional structure of the genome within the cell nucleus is not static either. The positions of cis-regulatory elements (CREs) will be adjusted during domestication, and these changes will also indirectly affect the expression patterns of different subgenomes. But things are never handled by a single link alone; often, multiple regulatory factors work together in coordination. In addition to cis factors, trans regulation is also quite crucial. Especially trans eQTL, they often act across subgenomes, which is equivalent to building a bridge to enable better communication and coordination between two sets of genomes (Bao et al., 2019; You et al., 2023; Yang et al., 2024). So, what is seen is the expression bias, but behind it, multiple levels are simultaneously "taking action", including the sequence of promoters, the activity of chromatin, the reorganization of spatial structure, and the coordinated cooperation of regulatory networks.

5.3 Functional implications for fiber development, stress responses, and reproduction

Often, when we focus on gene expression bias, it is not merely to figure out "who expresses more", but to know what actual changes it brings. The genes that are more inclined to be expressed in the At subgenome are generally more concentrated in the fiber initiation and early elongation stages. The biased expression of Dt often occurs in secondary wall synthesis and stress response (Xing et al., 2024). The two are not in opposition but in division of labor. This division of labor also endows cotton with the ability to cope with challenges under different ecological conditions. For instance, the functional division of the CesA gene family well demonstrates this coordination mechanism (Wang and Zhang, 2024). Their expression coordination between At and Dt supports the entire development process of fibers. And this kind of expression pattern is not merely a "natural phenomenon". Some traits related to photoperiod sensitivity and fiber quality have been found to be associated with subgenome-specific regulatory networks and epigenetic modifications (Han et al., 2023). So, during the domestication process, the expression bias might have been "quietly selected" long ago.

6 Case Study

6.1 Data integration from multiple cultivars and wild relatives

What a single reference genome can do is indeed limited. To understand the full genetic picture of cotton, studies have integrated data from as many as 1,961 germplasm resources, including mainstream cultivated varieties, local varieties and some wild "relatives". This wave of operation is not only large-scale, but more importantly, it can capture those genetic variations that have been overlooked by a single reference sequence, especially presence/absence variations (PAVs) and some signals related to domestication and improvement. Through such extensive integration, researchers identified over 450 Mb of selectable regions and locked down 162 loci associated with 16 agronomic traits, including 84 previously unreported new loci (Figure 2) (Li et al., 2021). These data illustrate a problem-only when the sample is wide enough can there be a chance to discover the hidden genetic diversity that has been "averaged out".

6.2 Discovery of key genes absent in reference genomes but linked to elite traits

The reference genome is undoubtedly important, but it is not omnipotent. For instance, in the whole-genome analysis of upland cotton (*G. hirsutum*), researchers discovered over 30,000 "new genes" that had never appeared in the reference genome at all. These genes, although missing in the "mainstream" version, are actually present in some cultivated species or wild materials. More notably, among these non-reference genes, 124 are related to fiber quality and yield traits, and 47 are directly associated with multiple superior agronomic traits. Such findings suggest that if we remain confined within the framework of reference sequences, we are likely to miss some

variations that are particularly crucial for breeding. This is also why pan-genome resources have become an important "supplement package" in trait improvement.

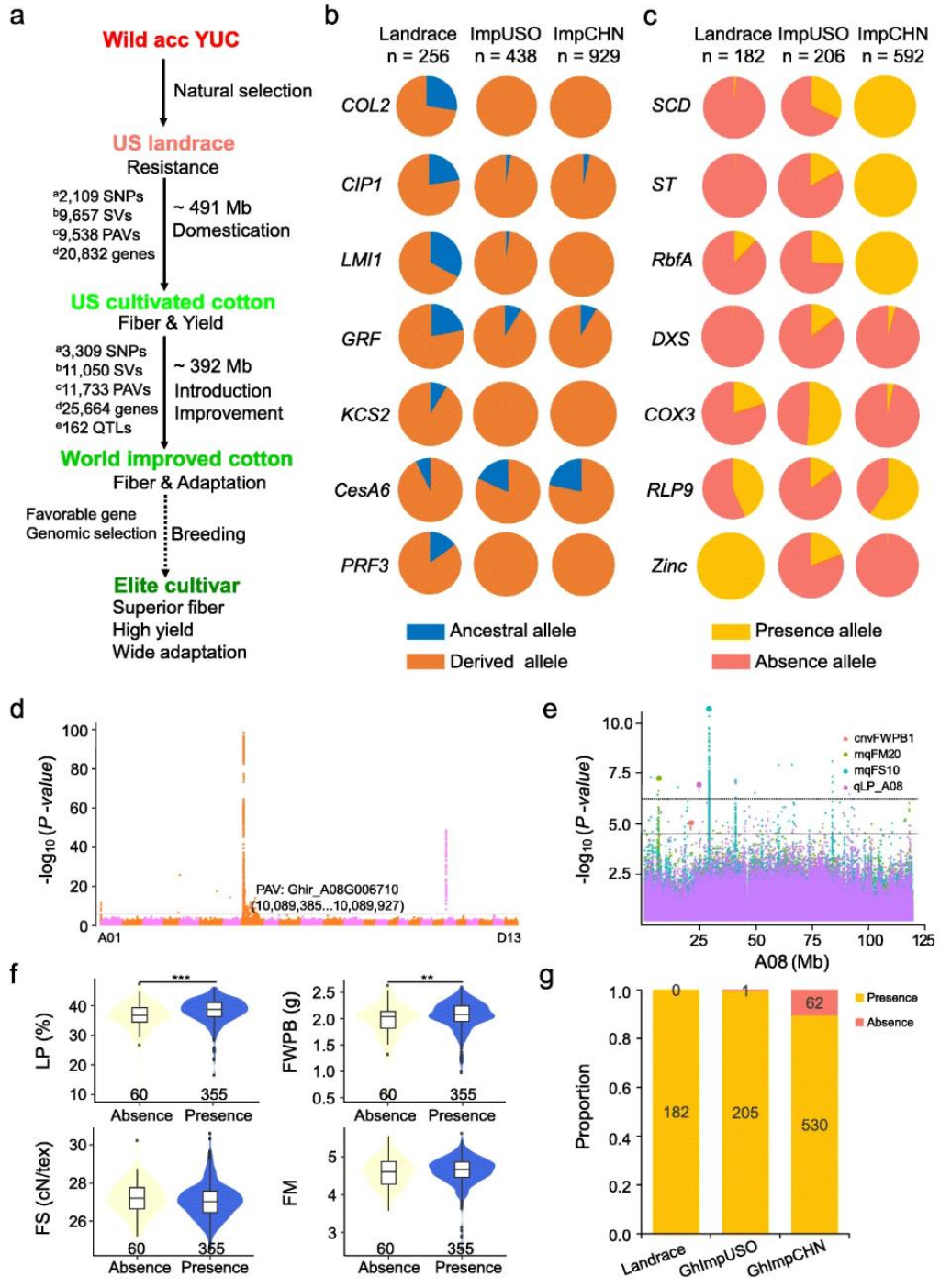

Figure 2 An available pan-genome dataset for cotton breeding. a A four-step model of variation during cotton domestication and breeding. b The spectrum of gene allele frequencies at the causal SNP polymorphisms of *COL2*, *CIP1, PRF3*, *LMI1*, *GRF*, *KCS2*, and *CesA6* in landrace and two geographic groups. c The spectrum of domesticated PAV allele frequ encies of seven genes in landrace and two geographic groups. d An example of functional PAV located on the A08 chromosome. The dashed line in Manhattan plot indicates the threshold for GWAS signals ($P < 2.62 \times 10\text{-}8$; -log $P > 7.6$). This locus includes four QTLs (lint percentage (LP), fiber weight per boll (FWPB), fiber micronaire (FM), fiber strength (FS)). e Four QTLs were displayed in a panel of multiple accessions. The two dashed lines represent GWAS thresholds for CNV (-log $P > 6.45$) and SNP (-log $P > 4.42$), respectively. f The phenotypic difference between presence and absence groups. The numbers below the violin plots show the accession numbers. The significance difference was calculated with a two-sided Wilcoxon rank-sum test (***$P < 0.001$, **$P < 0.01$). g Presence frequencies of *Ghir_A08G006710* in 182 landrace, 206 GhImpUSO, and 592 GhImpCHN accessions (Adopted from Li et al., 2021)

6.3 Expression bias between a-and d-subgenomes during fiber elongation and maturation

How are cotton fibers grown? Although it may seem like a unified process on the surface, it is actually not that simple behind the scenes. During this developmental process, the A subgenome (At) and the D subgenome (Dt) are not "evenly matched"; their performance is somewhat "biased". The genes on the At side are particularly active during the initial elongation stage of the fibers, with a very high level of participation. The role of Dt is more likely to occur in the later stage of development or to "push up" when cotton encounters environmental pressure. But this state of "who is busy and who is idle" is not static. This division of labor in expression varies depending on the type of tissue, the time point of development, and even whether it has been artificially cultivated and domesticated (Nobles et al., 2025). Sometimes, the two sets of genomes seem to cooperate with each other and perform their own duties. Sometimes, however, they seem to be filling in for each other-when one becomes stronger, the other takes a slight step back. At and Dt have chosen different expression paths, which is equivalent to "operating separately", but their ultimate goal is the same: to enable cotton to exhibit appropriate developmental and trait characteristics at different stages.

7 Applications in Cotton Breeding and Biotechnology

7.1 Utilizing pan-genome resources for GWAS and QTL mapping

In the past, when conducting GWAS or QTL mapping, the thinking basically revolved around a reference genome. This is of course simple, but it's also very easy to miss something-many genes related to important traits are not included in that "standard version" at all. It was not until the introduction of the pan-genome that research began to broaden. Researchers integrated tens of thousands of germplasm resources, including many local varieties and wild materials, and as a result, they identified over 160 gene loci related to agronomic traits at once. What is more worth mentioning is that 84 of these loci have not been discovered before, and another 47 are directly associated with 16 core agronomic traits (Kushanov et al., 2021; Tan et al., 2024). Behind these advancements, it is actually closely related to the inclusion of variant types such as PAV and SV in the analysis. Previously overlooked "invisible markers" can now come into view, providing more detailed and practical references for subsequent marker-assisted breeding. In other words, when it comes to breeding, one doesn't necessarily have to revolve around a "single version".

7.2 Identification of novel gene targets for CRISPR and transgenic approaches

Gene editing tools like CRISPR are not in short supply, but what is really hard to find is the "right target". At the beginning, everyone was picking targets from the reference genome, but the problem was that the set of data itself was limited. The emergence of the pan-genome has broken this limitation-it no longer only looks at that one "standard version", but integrates data from tens of thousands of samples. Now, researchers suddenly have many new genes that they couldn't find before. Although they are not in the mainstream reference genome, they are often linked to cotton yield, stress resistance performance, and even fiber growth. These newly added candidate genes and regulatory fragments have become a very useful "target resource library" for editing technologies such as CRISPR/Cas9. Now, in terms of fiber quality improvement and stress resistance enhancement, many initial editing achievements have begun to show results (Peng et al., 2021; Kumar et al., 2024; Thangaraj et al., 2024; Sheri et al., 2025). Ultimately, the role of the pan-genome is not merely to discover new genes. Instead, it provides us with more alternative paths when doing gene editing. With more starting points, the space for improvement naturally expands.

7.3 Incorporation of dispensable genes into elite cultivars for yield and stress tolerance

When many people hear the term "non-essential genes", their first reaction is: Are these genes of no use? In the past, when we were doing breeding, indeed, no one paid much attention to them. However, upon analysis of the pan-genome data, the results were somewhat unexpected-these "private genes" that only appear in some materials have instead played a significant role in enhancing the disease resistance of cotton, increasing its yield, and even in addressing climate change. In particular, some genes from local or wild varieties, which were previously overlooked because they were not included in mainstream breeding materials, are now being "rediscovered" through the pan-genome. These genes are like "sleeping resources". Once they are excavated and reasonably introduced into superior materials, not only can the genetic diversity of cotton be expanded, but also a new

channel can be opened for the development of new varieties with stress tolerance and high yield (Wen et al., 2023). Sometimes, what is truly useful is not necessarily those conventional genes that are "present in all varieties", but rather these less popular ones, which are more likely to bring about unexpected gains in breeding.

8 Concluding Remarks and Future Perspectives

Some questions, in fact, cannot be answered no matter how many times you search the reference genome. It was not until the introduction of the pan-genome that many "invisible corners" began to be illuminated-those missing genes, structural variations, and even brand-new expression patterns gradually emerged. The previously overlooked variations have now become a key breakthrough point, especially in breeding goals such as fiber quality, yield, and stress resistance, where the changes are even more pronounced. But then again, things are not that simple. It is still unclear what exactly those optional genes and private genes that only appear in some materials do at different developmental stages. How subgenomic expression bias is regulated and how multi-omics data are interconnected and connected are still in a semi-solved state at present.

Whether the pan-genome can be truly applied in practice depends on whether the technology keeps up. First of all, it is necessary to measure quickly and accurately. This is the foundation. The currently commonly used high-throughput sequencing and assembly technologies capable of handling extremely long fragments still need to be further upgraded in the future. Then, methods that can "see more precisely", such as single-cell omics and spatial omics, should also be accelerated simultaneously. On the other hand, whether it is identifying structural variations (such as large fragment insertions, deletions, inversions), PAVs (presence/absence variations), or finding regulatory elements, if the algorithm is not effective, the efficiency and accuracy will also be compromised. However, to be fair, it is probably difficult to bring the entire system to life with just one or two technological advancements. What we truly need is an open platform similar to a "cotton version of ENCODE", integrating various omics data, regulatory relationships, and functional annotations together. The key point is not to set up a database and leave it untouched, but to make this platform effortless to use-easy to search for information, not difficult to operate, and quick to update. For those engaged in research, breeding, or even in the industrial sector, this kind of thing is not an "added bonus", but an "essential tool".

The direction of cotton breeding in the future is likely not to be a single technology going it alone, but rather a combination of several core methods: for instance, the pan-genome can tell you the potential key genes, CRISPR can precisely edit them by hand, and AI algorithms can help you quickly select the optimal combination. Only if this process can run smoothly can there be hope of breeding new cotton varieties that are disease-resistant, high-yielding and have good fiber quality. But to reach that point, it is unrealistic for a single laboratory to go it alone. Cross-disciplinary collaboration, open sharing of research data, and long-term stable resource support are all indispensable. Ultimately, whether cotton can breed the next generation of "super varieties" is not only a matter of agricultural technology, but also closely related to the foundation of sustainable development of food and fibers on a global scale.

Acknowledgments

I thank the anonymous reviewers for their insightful comments and suggestions that greatly improved the manuscript.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Arslan M., Fatima A., Javeria F., Ijaz S., Riaz U., Saleem G., Bekhit M., Mezher M., and Iqbal R., 2025, Assessment of genetic diversity in cotton genotypes using simple sequence repeat (SSR) markers: insights from interspecific and intraspecific variations, Genetic Resources and Crop Evolution, 72(3): 3661-3670.
https://doi.org/10.1007/s10722-024-02185-y

Bao Y., Hu G., Grover C., Conover J., Yuan D., and Wendel J., 2019, Unraveling *cis* and *trans* regulatory evolution during cotton domestication, Nature Communications, 10(1): 5399.
https://doi.org/10.1038/s41467-019-13386-w

Chen Z., Sreedasyam A., Ando A., Song Q., De Santiago L., Hulse-Kemp A., Ding M., Ye W., Kirkbride R., Jenkins J., Plott C., Lovell J., Lin Y., Vaughn R., Liu B., Simpson S., Scheffler B., Wen L., Saski C., Grover C., Hu G., Conover J., Carlson J., Shu S., Boston L., Williams M., Peterson D., McGee K., Jones D., Wendel J., Stelly D., Grimwood J., and Schmutz J., 2020, Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement, Nature Genetics, 52(5): 525-533.
https://doi.org/10.1038/s41588-020-0614-5

Gurmessa D., Bantte K., and Negisho K., 2024, Genetic diversity in Pima (*Gossypium barbadense* L.) and advanced interspecific hybrids (*Gossypium hirsutum* × *Gossypium barbadense*) of cotton germplasm in Ethiopia, Plant Gene, 39: 100458.
https://doi.org/10.1016/j.plgene.2024.100458

Han J., López-Arredondo D., Yu G., Wang Y., Wang B., Wall S., Zhang X., Fang H., Barragán-Rosillo A., Pan X., Jiang Y., Chen J., Zhang H., Zhou B., Herrera-Estrella L., Zhang B., and Wang K., 2022, Genome-wide chromatin accessibility analysis unveils open chromatin convergent evolution during polyploidization in cotton, Proceedings of the National Academy of Sciences, 119(44): e2209743119.
https://doi.org/10.1073/pnas.2209743119

Han J., Yu G., Zhang X., Dai Y., Zhang H., Zhang B., and Wang K., 2023, Histone maps in *Gossypium darwinii* reveal epigenetic regulation drives subgenome divergence and cotton domestication, International Journal of Molecular Sciences, 24(13): 10607.
https://doi.org/10.3390/ijms241310607

Hinze L., Hulse-Kemp A., Wilson I., Zhu Q., Llewellyn D., Taylor J., Spriggs A., Fang D., Ulloa M., Burke J., Giband M., Lacape J., Van Deynze A., Udall J., Scheffler J., Hague S., Wendel J., Pepper A., Frelichowski J., Lawley C., Jones D., Percy R., and Stelly D., 2017, Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array, BMC Plant Biology, 17(1): 37.
https://doi.org/10.1186/s12870-017-0981-y

Hu G., Grover C., Vera D., Lung P., Girimurugan S., Miller E., Conover J., Ou S., Xiong X., Zhu D., Li D., Gallagher J., Udall J., Sui X., Zhang J., Bass H., and Wendel J., 2024, Evolutionary dynamics of chromatin structure and duplicate gene expression in diploid and allopolyploid cotton, Molecular Biology and Evolution, 41(5): msae095.
https://doi.org/10.1093/molbev/msae095

Hu Y., Chen J., Fang L., Zhang Z., Ma W., Niu Y., Ju L., Deng J., Zhao T., Lian J., Baruch K., Fang D., Liu X., Ruan Y., Rahman M., Han J., Wang K., Wang Q., Wu H., Mei G., Zang Y., Han Z., Xu C., Shen W., Yang D., Si Z., Dai F., Zou L., Huang F., Bai Y., Zhang Y., Brodt A., Ben-Hamo H., Zhu X., Zhou B., Guan X., Zhu S., Chen X., and Zhang T., 2019, *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton, Nature Genetics, 51(4): 739-748.
https://doi.org/10.1038/s41588-019-0371-5

Huang G., Huang J., Chen X., and Zhu Y., 2021, Recent advances and future perspectives in cotton research, Annual Review of Plant Biology, 72(1): 437-462.
https://doi.org/10.1146/annurev-arplant-080720-113241

Huang X., Wang Y., Zhang S., Pei L., You J., Long Y., Li J., Zhang X., Zhu L., and Wang M., 2024, Epigenomic and 3D genomic mapping reveals developmental dynamics and subgenomic asymmetry of transcriptional regulatory architecture in allotetraploid cotton, Nature Communications, 15(1): 10721.
https://doi.org/10.1038/s41467-024-55309-4

Jin S., Han Z., Hu Y., Si Z., Dai F., He L., Cheng Y., Li Y., Zhao T., Fang L., and Zhang T., 2023, Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons, Molecular Plant, 16(4): 678-693.
https://doi.org/10.1016/j.molp.2023.02.004

Joshi B., Singh S., Tiwari G., Kumar H., Boopathi N., Jaiswal S., Adhikari D., Kumar D., Sawant S., Iquebal M., and Jena S., 2023, Genome-wide association study of fiber yield-related traits uncovers novel genomic regions and candidate genes in Indian upland cotton (*Gossypium hirsutum* L.), Frontiers in Plant Science, 14: 1252746.
https://doi.org/10.3389/fpls.2023.1252746

Kumar R., Das J., Puttaswamy R., Kumar M., Balasubramani G., and Prasad Y., 2024, Targeted genome editing for cotton improvement: prospects and challenges, The Nucleus, 67(1): 181-203.
https://doi.org/10.1007/s13237-024-00479-1

Kushanov F., Turaev O., Ernazarova D., Gapparov B., Oripova B., Kudratova M., Rafieva F., Khalikov K., Erjigitov D., Khidirov M., Kholova M., Khusenov N., Amanboyeva R., Saha S., Yu J., and Abdurakhmonov I., 2021, Genetic diversity, QTL mapping, and marker-assisted selection technology in cotton (*Gossypium* spp.), Frontiers in Plant Science, 12: 779386.
https://doi.org/10.3389/fpls.2021.779386

Li J., Liu Z., You C., Qi Z., You J., Grover C., Long Y., Huang X., Lu S., Wang Y., Zhang S., Wang Y., Bai R., Zhang M., Jin S., Nie X., Wendel J., Zhang X., and Wang M., 2024, Convergence and divergence of diploid and tetraploid cotton genomes, Nature Genetics, 56(11): 2562-2573.
https://doi.org/10.1038/s41588-024-01964-8

Li J., Yuan D., Wang P., Wang Q., Sun M., Liu Z., Si H., Xu Z., Zhang B., Pei L., Tu L., Zhu L., Chen L., Lindsey K., Zhang X., Jin S., and Wang M., 2021, Cotton pan-genome retrieves the lost sequences and genes during domestication and selection, Genome Biology, 22(1): 119.
https://doi.org/10.1186/s13059-021-02351-w

Li X., Jin X., Wang H., Zhang X., and Lin Z., 2016, Structure, evolution, and comparative genomics of tetraploid cotton based on a high-density genetic linkage map, DNA Research, 23(3): 283-293.
https://doi.org/10.1093/dnares/dsw016

Ma Z., He S., Wang X., Sun J., Zhang Y., Zhang G., Wu L., Li Z., Liu Z., Sun G., Yan Y., Jia Y., Yang J., Pan Z., Gu Q., Li X., Sun Z., Dai P., Liu Z., Gong W., Wu J., Wang M., Liu H., Feng K., Ke H., Wang J., Lan H., Wang G., Peng J., Wang N., Wang L., Pang B., Peng Z., Li R., Tian S., and Du X., 2018, Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield, Nature Genetics, 50(6): 803-813.
https://doi.org/10.1038/s41588-018-0119-7

Ma Z., Zhang Y., Wu L., Zhang G., Sun Z., Li Z., Jiang Y., Ke H., Chen B., Liu Z., Gu Q., Wang Z., Wang G., Yang J., Wu J., Yan Y., Meng C., Li L., Li X., Mo S., Wu N., Chen L., Zhang M., Si A., Yang Z., Wang N., Wu L., Zhang D., Cui Y., Cui J., Lü X., Li Y., Shi R., Duan Y., Tian S., and Wang X., 2021, High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement, Nature Genetics, 53(9): 1385-1391.
https://doi.org/10.1038/s41588-021-00910-2

Mei M., Syed N., Gao W., Thaxton P., Smith C., Stelly D., and Chen Z., 2004, Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*), Theoretical and Applied Genetics, 108(2): 280-291.
https://doi.org/10.1007/s00122-003-1433-7

Naoumkina M., Thyssen G., Fang D., Hinchliffe D., Florane C., Yeater K., Page J., and Udall J., 2014, The Li_2 mutation results in reduced subgenome expression bias in elongating fibers of allotetraploid cotton (*Gossypium hirsutum* L.), PLoS ONE, 9(3): e90830.
https://doi.org/10.1371/journal.pone.0090830

Nobles A., Wendel J., and Yoo M., 2025, Comparative analysis of floral transcriptomes in *Gossypium hirsutum* (Malvaceae), Plants, 14(4): 502.
https://doi.org/10.3390/plants14040502

Page J., Huynh M., Liechty Z., Grupp K., Stelly D., Hulse A., Ashrafi H., Van Deynze A., Wendel J., and Udall J., 2013, Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing, G3: Genes|Genomes|Genetics, 3(10): 1809-1818.
https://doi.org/10.1534/g3.113.007229

Pan Y., Meng F., and Wang X., 2020, Sequencing multiple cotton genomes reveals complex structures and lays foundation for breeding, Frontiers in Plant Science, 11: 560096.
https://doi.org/10.3389/fpls.2020.560096

Peng R., Jones D., Liu F., and Zhang B., 2021, From sequencing to genome editing for cotton improvement, Trends in Biotechnology, 39(3): 221-224.
https://doi.org/10.1016/j.tibtech.2020.09.001

Peng Z., Cheng H., Sun G., Pan Z., Wang X., Geng X., He S., and Du X., 2020, Expression patterns and functional divergence of homologous genes accompanied by polyploidization in cotton (*Gossypium hirsutum* L.), Science China Life Sciences, 63(10): 1565-1579.
https://doi.org/10.1007/s11427-019-1618-7

Shen C., Jin X., Zhu D., and Lin Z., 2017, Uncovering SNP and indel variations of tetraploid cottons by SLAF-seq, BMC Genomics, 18(1): 247.
https://doi.org/10.1186/s12864-017-3643-4

Sheri V., Mohan H., Jogam P., Alok A., Rohela G., and Zhang B., 2025, CRISPR/Cas genome editing for cotton precision breeding: mechanisms, advances, and prospects, Journal of Cotton Research, 8(1): 4.
https://doi.org/10.1186/s42397-024-00206-w

Song C., Li W., Wang Z., Pei X., Liu Y., Ren Z., He K., Zhang F., Sun K., Zhou X., and Yang D., 2018, Genome resequencing reveals genetic variation between the parents of an elite hybrid upland cotton, Agronomy, 8(12): 305.
https://doi.org/10.3390/agronomy8120305

Strygina K., Khlestkina E., and Podolnaya L., 2020, Cotton genome evolution and features of its structural and functional organization, Biological Communications, 65(1): 15-27.
https://doi.org/10.21638/spbu03.2020.102

Sun Z., Wang X., Liu Z., Gu Q., Zhang Y., Li Z., Ke H., Yang J., Wu J., Wu L., Zhang G., and Zhang C., 2017, Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L., Plant Biotechnology Journal, 15(8): 982-996.
https://doi.org/10.1111/pbi.12693

Tan H., Tang B., Sun M., Yin Q., Ma Y., Li J., Wang P., Li Z., Zhao G., Wang M., Zhang X., You C., and Tu L., 2024, Identification of new cotton fiber-quality QTL by multiple genomic analyses and development of markers for genomic breeding, The Crop Journal, 12(3): 866-879.
https://doi.org/10.1016/j.cj.2024.03.014

Thangaraj A., Kaul R., Sharda S., and Kaul T., 2024, Revolutionizing cotton cultivation: a comprehensive review of genome editing technologies and their impact on breeding and production, Biochemical and Biophysical Research Communications, 742: 151084.
https://doi.org/10.1016/j.bbrc.2024.151084

Van Deynze A., Stoffel K., Lee M., Wilkins T., Kozik A., Cantrell R., Yu J., Kohel R., and Stelly D., 2009, Sampling nucleotide diversity in cotton, BMC Plant Biology, 9(1): 125.
https://doi.org/10.1186/1471-2229-9-125

Wang J.M., and Zhang J., 2024, Assessing the impact of various cotton diseases on fiber quality and production, Field Crop, 7(4): 212-221.
https://doi.org/10.5376/fc.2024.07.0021

Wang L., Xing H., Yuan Y., Wang X., Saeed M., Tao J., Feng W., Zhang G., Song X., and Sun X., 2018, Genome-wide analysis of codon usage bias in four sequenced cotton species, PLoS ONE, 13(3): e0194372.
https://doi.org/10.1371/journal.pone.0194372

Wang M., Li J., Qi Z., Long Y., Pei L., Huang X., Grover C., Du X., Xia C., Wang P., Liu Z., You J., Tian X., Wang R., Chen X., He X., Fang D., Sun Y., Tu L., Jin S., Zhu L., Wendel J., and Zhang X., 2022, Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*, Nature Genetics, 54(12): 1959-1971.
https://doi.org/10.1038/s41588-022-01237-2

Wang M., Tu L., Lin M., Lin Z., Wang P., Yang Q., Ye Z., Shen C., Li J., Zhang L., Zhou X., Nie X., Li Z., Guo K., Huang C., Jin S., Zhu L., Yang X., Min L., Yuan D., Zhang Q., Lindsey K., and Zhang X., 2017, Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication, Nature Genetics, 49(4): 579-587.
https://doi.org/10.1038/ng.3807

Wen X., Chen Z., Yang Z., Wang M., Jin S., Wang G., Zhang L., Wang L., Li J., Saeed S., He S., Wang Z., Wang K., Kong Z., Li F., Zhang X., Chen X., and Zhu Y., 2023, A comprehensive overview of cotton genomics, biotechnology and molecular biological studies, Science China Life Sciences, 66(10): 2214-2256.
https://doi.org/10.1007/s11427-022-2278-0

Xing A., Zhang X., Wang J., He S., Nazir M., Wang X., Wang X., Yang R., Zhang H., Fu G., Chen B., Peng Z., and Du X., 2024, Transgressive and subgenome expression level dominance and co-expression network analyses at the early fiber development in allopolyploid *Gossypium*, Industrial Crops and Products, 214: 118552.
https://doi.org/10.1016/j.indcrop.2024.118552

Yang L., Qin W., Wei X., Liu R., Yang J., Wang Z., Yan Q., Zhang Y., Hu W., Han X., Gao C., Zhan J., Gao B., Ge X., Li F., and Yang Z., 2024, Regulatory networks of coresident subgenomes during rapid fiber cell elongation in upland cotton, Plant Communications, 5(12): 101130.
https://doi.org/10.1016/j.xplc.2024.101130

Yang Z., Ge X., Yang Z., Qin W., Sun G., Wang Z., Li Z., Liu J., Wu J., Wang Y., Lu L., Wang P., Mo H., Zhang X., and Li F., 2019, Extensive intraspecific gene order and gene structural variations in upland cotton cultivars, Nature Communications, 10(1): 2989.
https://doi.org/10.1038/s41467-019-10820-x

You J., Liu Z., Qi Z., Ma Y., Sun M., Su L., Niu H., Peng Y., Luo X., Zhu M., Huang Y., Chang X., Hu X., Zhang Y., Pi R., Liu Y., Meng Q., Li J., Zhang Q., Zhu L., Lin Z., Min L., Yuan D., Grover C., Fang D., Lindsey K., Wendel J., Tu L., Zhang X., and Wang M., 2023, Regulatory controls of duplicated gene expression during fiber development in allotetraploid cotton, Nature Genetics, 55(11): 1987-1997.
https://doi.org/10.1038/s41588-023-01530-8

Zhang A., Wei Y., Shi Y., Deng X., Gao J., Feng Y., Zheng D., Cheng X., Li Z., Wang T., Wang K., Liu F., Peng R., and Zhang W., 2021, Profiling of H3K4me3 and H3K27me3 and their roles in gene subfunctionalization in allotetraploid cotton, Frontiers in Plant Science, 12: 761059.
https://doi.org/10.3389/fpls.2021.761059

Zhang T., Hu Y., Jiang W., Fang L., Guan X., Chen J., Zhang J., Saski C., Scheffler B., Stelly D., Hulse-Kemp A., Wan Q., Liu B., Liu C., Wang S., Pan M., Wang Y., Wang D., Ye W., Chang L., Zhang W., Song Q., Kirkbride R., Chen X., Dennis E., Llewellyn D., Peterson D., Thaxton P., Jones D., Wang Q., Xu X., Zhang H., Wu H., Zhou L., Mei G., Chen S., Tian Y., Xiang D., Li X., Ding J., Zuo Q., Tao L., Liu Y., Li J., Lin Y., Hui Y., Cao Z., Cai C., Zhu X., Jiang Z., Zhou B., Guo W., Li R., and Chen Z., 2015, Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement, Nature Biotechnology, 33(5): 531-537.
https://doi.org/10.1038/nbt.3207

Zhao B., Cao J., Hu G., Chen Z., Wang L., Shangguan X., Wang L., Mao Y., Zhang T., Wendel J., and Chen X., 2018, Core *cis*-element variation confers subgenome-biased expression of a transcription factor that functions in cotton fiber elongation, New Phytologist, 218(3): 1061-1075.
https://doi.org/10.1111/nph.15063