Review

Cotton functional genomics reveals global insight into genome evolution and fiber development

Zhiguo Wu [a], Yan Yang [b], Gai Huang [c], Jing Lin [a], Yuying Xia [b], Yuxian Zhu [a, b, *]

[a] College of Life Sciences, Wuhan University, Wuhan 430072, China
[b] Institute for Advanced Studies, Wuhan University, Wuhan 430072, China
[c] State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

ABSTRACT

Due to the economic value of natural textile fiber, cotton has attracted much research attention, which has led to the publication of two diploid genomes and two tetraploid genomes. These big data facilitate functional genomic study in cotton, and allow researchers to investigate cotton genome structure, gene expression, and protein function on the global scale using high-throughput methods. In this review, we summarized recent studies of cotton genomes. Population genomic analyses revealed the domestication history of cultivated upland cotton and the roles of transposable elements in cotton genome evolution. Alternative splicing of cotton transcriptomes was evaluated genome-widely. Several important gene families like *MYC*, *NAC*, *Sus* and *GhPLDα1* were systematically identified and classified based on genetic structure and biological function. High-throughput proteomics also unraveled the key functional proteins correlated with fiber development. Functional genomic studies have provided unprecedented insights into global-scale methods for cotton research.

1. Introduction

In the first ten years of the 21st century, advanced high-throughput DNA sequencing technologies and bioinformatic tools have been developed intensively, enabling scientists to accurately sequence and assemble large and complex crop genomes, as well as characterize the functions of genomic features at an unprecedented depth (Bevan et al., 2017). The number of plant genomic studies has increased explosively in recent years, and this growth is in part due to the understanding that genomic research can help address requirements for farm products and intense concerns about agricultural security. Ultimately, crop characteristics such as yield and disease resistance can be traced back to causes at the biomolecular scale (e.g., DNA sequences, chromatin structures, transcriptome levels, and protein isoforms). Functional genomics uses genomic data to study the roles of genes and proteins on a global scale by evaluating gene polymorphisms, mRNA transcription, translation, protein-protein interactions, etc.

Cotton is an important industrial crop, and is also a model organism for the study of cellulose biosynthesis, cell elongation, and polyploidy in the evolution of plant genomes (Zhu, 2016). In the past five years, scientists from China and the United States have made significant progress in their effort to sequence the genomes of allotetraploid *Gossypium hirsutum* and *G. barbadense* (AADD) (Li et al., 2015; Liu et al., 2015; Zhang et al., 2015), as well as two ancestor diploid genomes, *G. raimondii* (DD) (Paterson et al., 2012; Wang et al., 2012) and *G. arboreum* (AA) (Li et al., 2014a, 2014b). These sequenced cotton genomes are excellent research platforms that offer rich databases for the study of cotton functional genomics, and will continue to improve the outcomes of fundamental research studies in cotton, including those concerned with genome evolution, gene expression and transcriptomics, epigenomics and proteomics, among other areas. In this review, we summarize functional genomic studies of cotton, with particular emphasis on work that has emerged since the publication of the *G. hirsutum* (upland cotton) genome in 2015. The review includes studies of cultivated cotton domestication, cotton genome editing, alternative splicing, genes encoding transcription factor (*MYC*, *NAC* and *Sus*), the *GhPLDα1* gene, transgenic Bt cotton, 2DE protein analysis, and gossypol biosynthesis.

2. Cotton population genomic analysis

Cotton population genomic studies use new genome information to unravel the history of the evolution and domestication of cotton; such studies have also assessed the relevance of genome variation on fiber quality. Xianlong Zhang's group recently reported a variation map of 352 wild and domesticated cotton accessions (Wang et al., 2017a). They successfully employed high-throughput methods like high-throughput/resolution chromosome conformation capture (Hi-C), chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and DNase hypersensitive site (DHS), and histone markers to link chromatin structures with genomic functions, and found 93 regions that were associated with the domestication history of upland cotton. These regions encompass a total of 178 Mb (74 Mb from the A subgenome; 104 Mb from the D subgenome). A GWAS analysis they conducted also identified 19 genomic loci that were related to fiber quality. Tianzhen Zhang's group reported a comprehensive genomic assessment of 318 modern upland cotton lines based on genome-wide resequencing technologies (Fang et al., 2017). They detected more loci associated with lint yield than with fiber quality, and found two genes in the ethylene pathway associated with lint yield. They also evaluated the population frequency of each elite allele in the released cultivar groups, and found that 54.8% of the elite alleles were transferred from three founder landraces. These data provide a genomic basis for cotton fiber improvement.

Population genomics facilitates the analysis of gene expression and protein function at a genome-wide scale, and can identify the correlation between cotton phenotype and DNA base variant (Fig. 1). High-throughput sequencing technologies are used to evaluate multiple dimensions of chromatin structures in living cells, covalent histone modifications, and single nucleotide polymorphisms (SNPs).

Population genomics analysis also enables the study of individual and complex DNA elements in cognate organisms, such as transposable elements (TEs) in diploid and tetraploid cotton genomes. The proportions of the *G. arboreum*, *G. raimondii*, and *G. hirsutum* genomes that are composed of TEs are 57%, 68.5% and 67.2%, respectively (Wang et al., 2016a, 2016b, 2016c). Strikingly, the AA and AADD genomes harbor considerably more *LTR-gypsy* (long terminal repeat) type retrotransposons than the DD genome

Table 1
Repeat sequence numbers in different cotton (*Gossypium*) genomes.

Sequence type	Number of repeat sequence		
	G. raimondii	*G. arboreum*	*G. hirsutum*
LTR/Copia	252,909	274,292	411,128
LTR/Gypsy	550,931	1,591,388	1,657,927
LINE	71,824	69,199	94,850
SINE	1924	521	902
DNA	130,006	103,675	150,146
Other	202,145	227,397	199,430
Total	1,209,739	2,266,472	2,514,383

(Table 1), suggesting that TEs might contribute to cotton chromosome sizes during evolution and that such changes appear to be related to fiber development. It has been well established that TEs can act as active regulatory elements that affect gene expression and post-translation modifications.

Previously, we assembled 297,239 upland cotton EST sequences and obtained 49,125 unigenes, of which 29,547 were derived from the D subgenome while 19,578 were from the A subgenome (Jin et al., 2013). Although *G. raimondii* cotton fiber abortion happens before the elongation stage, the combination of D genome expression and A genome expression enhances upland cotton fiber length and quality, and this unexpected phenomenon indicates that there are extensive interactions and regulatory networks between the D subgenome and the A subgenome in the tetraploid cotton; the phenotypic manifestation of these interactions is especially apparent in cellulose biosynthesis.

Statistics of differentially expressed orthologous gene pairs revealed that 17.9% and 29.7% of orthologous genes had significantly higher expression in the diploid AA or DD compared to the At or Dt subgenome in the allotetraploid cotton, respectively. Only 9.5% or 6.6% of the orthologous genes had significantly higher expression in the At or Dt subgenomes compared to the AA or DD genomes (Table 2). In contrast, systematic studies found that key cellulose synthase genes required for primary and secondary cell wall formation had significantly higher expression in the At subgenome of the tetraploid cotton than in any of the diploid genomes (Table 3) (Jin et al., 2013; Li et al., 2015). This is true for the *Sus* gene family (Table 3). However, the expression levels of many other
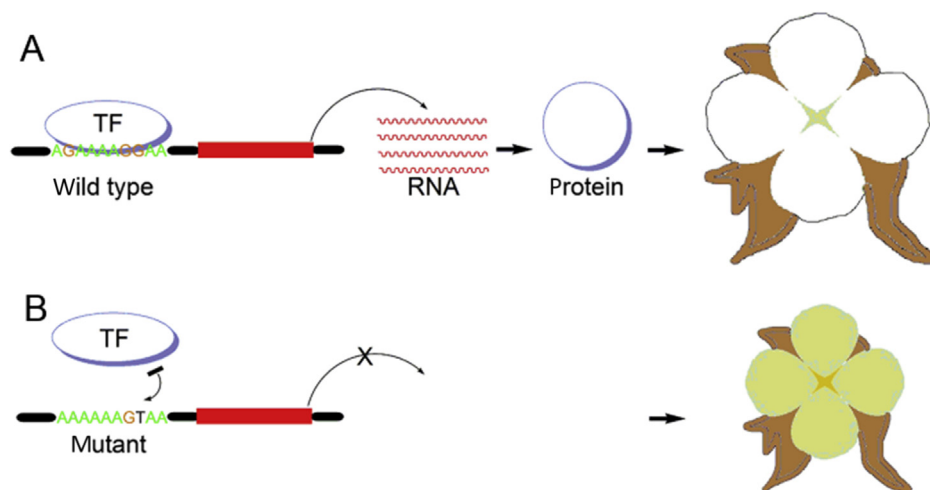

Fig. 1. Cotton phenotypes resulting from DNA base variants. This cartoon figure is derived from a paper published by Zhang's group (Wang et al., 2017a), and shows how cotton fiber development is influenced by DNA base mutations. **A:** In the wild-type lines, DNA motifs in promoters and enhancers can be recognized by specific transcription factors (TF). Functional genes are then highly expressed to promote fiber development. **B:** In the mutant lines, the expression of functional genes could not be triggered by promoters or enhancers because the TF binding sites were altered to be ineffective.