Chapter 10

Bioinformatics Tools and Genomic Resources Available in Understanding the Structure and Function of *Gossypium*

Venkateswara R. Sripathi, Ramesh Buyyarapu, Siva P. Kumpatla, Abreeotta J. Williams, Seloame T. Nyaku, Yonathan Tilahun, Venu Kalavacharla and Govind C. Sharma

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/64325

Abstract

Cotton is economically and evolutionarily important crop for its fiber. In order to improve fiber quality and yield, and to exploit the natural genetic potential inherent in genotypes, understanding genome structure and function of cultivated cotton is important. In order to achieve this, a functional understanding of bioinformatics resources such as databases, software solutions, and analysis tools is required. But currently, there are very few unified reports on bioinformatics tools and even fewer repositories to access cotton genomic information. Also, resourceful developers and bioinformatics scientists actively addressing complex genomic challenges in cotton genomes are much in need. The primary goal of this chapter is to provide a review of such tools and resources for analyzing the structure and function of the cotton genome with preferential emphasis on this complex and economically important plant species. This discourse begins with a description of concurrent advances in high-throughput genome sequencing and bioinformatics analyses and focuses on four major sections covering bioinformatics tools and resources for analysis of: (1) genomes; (2) transcriptomes; (3) small RNAs; and (4) epigenomes. In each section, recent advances in cotton have been discussed. Cotton genome sequencing and annotation efforts are outlined within these sections. This review discusses the availability of genome information of both diploid and tetraploid species that have impelled cotton genome research into the post-genomics era, opening new avenues for exploring regulatory mechanisms associated with fine-tuning of gene expression of fiber-related genes. Finally, the potential impacts of these rapid advances, especially the challenges in handling and analyzing the large datasets are discussed.

Keywords: genome, transcriptome, epigenome, sequencing, cotton fiber

1. Introduction

Cotton is an economically and evolutionarily important crop species. Along with cotton improvement, which has progressed impressively with conventional and molecular breeding approaches, the genomic approaches utilizing next generation sequencing (NGS) technologies have enhanced our ability to understand and utilize the genetic potential of crop species [1]. The early sequencing efforts in cotton (*Gossypium* spp.) are mostly limited to diploid species such as *G. raimondii*, as its genome is structurally less complex. Now, DNA sequencing has become a routine tool in cotton genetic research. Sequencing genomic, transcriptomic, and regulatory regions of a plant species and comparing their patterns will provide a better understanding of genome architecture, genetic variation, gene identification, regulation, and its expression [2]. Understanding cotton genome, transcriptome, and regulatory molecules (small RNAs and epigenetic modulators) will therefore provide deeper insights into the structure and function of the genome [3]. Dissecting the complexity associated with tetraploid cotton genome and narrow genetic variability is more challenging and requires efficient methodologies. Comparative genome and transcriptome analysis of cultivated cotton and its progenitors will aid in the identification of novel single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), transcripts, transcript quantification, and alternative splice junctions at the transcriptome level and the role of these elements in regulating cotton fiber quality and yield [4]. Similarly, transcriptome profiling studies are powerful in unravelling the underlying mechanisms involved in gene expression associated with the sub-genomes of a plant species. In addition, the role of small RNAs and epigenetic modulators is increasingly evident in determining the genome landscapes of plant species including cotton.

DNA sequencing technologies have evolved tremendously over two decades from Bacterial artificial chromosome (BAC)-based cloning to single-molecule sequencing [5, 6]. Concomitantly, the computational tools for addressing the problems in understanding sequence data have also evolved from sequence alignment algorithms such as Needleman–Wunsch and Smith–Waterman to assembly tools such as Burrows-Wheeler Alignment (BWA) and Bowtie [7]. As described here, a wide range of publicly available bioinformatics tools and some commercially available sequence analysis tools such as CLCBio Genomics Workbench, Strand NGS, Geneious, Laser gene suite, and NextGENe are being utilized [8].

2. Genome analysis in cotton

Approximately 50 naturally occurring cotton species are available, including 45 diploid (2n = 2x = 26) and five allotetraploid species (2n = 4x = 52) each with a haploid chromosome number of 13 [9]. Based on meiotic pairing and chromosome size, diploid species (2n = 26) have been placed into genomic groups A, B, C, D, E, F, G, or K. Of these, A-genome sources, *G. herba-*

http://dx.doi.org/10.5772/64325

ceum (2n = 2x = A_1A_1 = 26) and *G. arboreum* (2n = 2x = A_2A_2 = 26), and D-genome sources, *G. raimondii* (2n = 2x = D_5D_5 = 26), and *G. gossypioides* (2n = 2x = D_6D_6 = 26) are considered as closest progenitor species to cultivated allotetraploid cotton [10]. Diploid genome sizes of A-G and K vary significantly between species ranging from ~900 to ~2800 Mb. Two natural allotetraploid (2n = 4x = $A_tA_tD_tD_t$ = 52) species, *G. hirsutum* and *G. barbadense,* are derived from a complex inter-specific hybridization process between two diploid species carrying A-genome and D-genomes [11]. However, the sequencing of polyploid plant genomes has been a tedious task due to repeat regions, whole-genome duplication events, and chromosomal rearrangements that have occurred in the process of evolution [12]. Effective strategies must be developed and employed for sequencing such complex genomes. In order to reduce such complexity, closest diploid progenitor species carrying D_5 (*G. raimondii*; 880 Mb), A_2 (*G. arboreum*; 1700 Mb), and A_1 (*G. herbaceum*; 1700 Mb) genomes can be sequenced and compared against cultivated tetraploid species carrying A_tD_t genome (*G. hirsutum*; ~2400 Mb and *G. barbadense*; ~2500 Mb), the estimated genome sizes [13] of the respective *Gossypium* species are given in parenthesis. The whole-genome sequencing efforts in cotton resulted in assembling ~775 Mb (~88%) of *G. raimondii* and ~1694 Mb (~90%) of *G. arboreum* genomes [14, 15]. Also, ~2.3 Gb (~90%) and ~2.5 Gb (~88%) of the genomes have been assembled into 13 pseudochromosomes each from A_t and D_t sub-genomes of *G. hirsutum* [16, 17] and *G. barbadense*, [18, 19] respectively. Presently, *G. herbaceum* (A_1) genome sequencing efforts are in progress with a partial genome assembled up to 1.2 Gb (~70%) [20].

Transposable elements (TEs) are abundant in plant genomes and are highly variable and often deleterious, as they undergo massive amplification within the genome. The role of TEs has been implicated in gene mutation, whole-genome duplication, chromosomal rearrangements and novel gene formation through genetic and epigenetic changes. They can alter gene expression and phenotypes by establishing and modifying gene regulatory networks. This is accomplished by inducing changes in genetic and epigenetic mechanisms [21]. The variation in genome structure and organization, even in closely related species, is primarily due to TEs and whole-genome duplication (WGD). This may be as a result of a combination of non-random events such as small RNA silencing and epigenetic mechanisms [22] and random events such as natural selection and adaptation [21].With the availability of allotetraploid and diploid cotton genomes, comparative genomics has been used to analyze and understand the structural variation and role of TEs in evolution [23]. Their study identified ~57, ~68, and ~67% of TEs in D_5, A_2, and A_tD_t genomes of cotton, respectively. Long terminal repeats (LTRs) have contributed significantly in whole-genome duplication and evolution of domesticated cotton. Though the overall TE content in A_2- and A_tD_t-genomes has been found to be similar, the frequency of LTR-gypsy and LTR-copia-type elements varied significantly. Terminal repeat retrotransposons in miniature (TRIMs) are a small, ubiquitous, conserved, poorly characterized, and scarcely reported group of repeats. TRIMs are often derived from partial deletion of long terminal repeat retrotransposons found in genic regions and in gene-body methylation within the genome. TRIMS are often targeted by small RNAs (sRNAs) of 21–24 nt in length. Screening for TRIMs in land plants is critical to understanding differences in selection pressure and evolutionary relationships among the clades. Using high-throughput

sequencing followed by bioinformatics analysis, 145 unique families of TRIMs have been identified after screening 48 plant genomes [24].

Genetic variation is an important element for crop improvement. An understanding of the genetic and genomic relationships of cotton species and cultivars is critical for further utilization of diversity in the development of improved cultivars with favorable alleles [25]. Allelic variations within a genome of the species can be classified into three major groups at DNA level: microsatellites, insertions/deletions, and single-nucleotide polymorphisms (SNPs) [26]. Molecular markers serve as efficient tools for genome characterization, understanding the genetic complex traits, marker-assisted selection (MAS) and for map-based cloning in breeding programs. Several molecular marker technologies have been used to study the genetic diversity and relationships of *Gossypium* species [27]. However, cotton crop improvement is limited by its narrow genetic base and limited variation among the cultivated cotton cultivars. Genetic variation at molecular level in cotton was previously characterized using isozyme/allozyme markers [28]; using non-coding genomic markers such as restriction fragment length polymorphisms (RFLPs) [29]; amplified fragment length polymorphisms (AFLPs) [30]; microsatellites [31, 32]; single-nucleotide polymorphisms [33] in *G. hirsutum* and its related species.

Simple sequence repeat (SSR) markers are widely used in many plant and animal genomes due to their abundance, hypervariability, and suitability for high-throughput analysis. Development of SSR markers using molecular methods is time consuming, laborious, and expensive. Use of computational approaches to mine ever-increasing sequences such as expressed sequence tags (ESTs) in public databases permits rapid and economical discovery of SSRs [34]. SSR mining programs such as Repeat Pattern tool kit; SSR Finder; Advanced Content Matching Engine for Sequences (ACMES); Spectral-repeat finder; Adplot; REPEATS and other programs are routinely used to mine EST databases, genome survey sequences, and other nucleotide databases [35]. In cotton, SSR markers mined from diploid species such as *G. arboreum* were also successfully employed to understand the structural variation in tetraploid cultivars [32]. SSR markers had been extensively used for many genetic mapping, quantitative trait loci (QTL), and trait mapping experiments for favorable characteristics such as fiber quality, higher yield [36], pathogen resistance [37, 38], and other important traits in cotton. With the advent of next generation sequencing technologies, identification of allelic variation at single-nucleotide level and their application in crop genetics is becoming a common practice.

SNP markers are becoming the “markers of choice” due to their abundance in the genomes, amenability to automation, and high-throughput genotyping capability. In cotton, using available EST or transcriptome sequences, gene-specific SNPs had been characterized [39]; however, these initial efforts and methods had limited application for genome wide SNP marker development in cotton species. Tetraploid nature, highly complex, and repetitive genome of cultivated cotton species poses significant challenges for genome wide SNP marker development. These complications usually result in high number of false positive SNPs, especially when they are developed from sequences that were not thoroughly characterized [25]. Though there is considerable genetic variation across the cultivated cotton species, the

http://dx.doi.org/10.5772/64325

narrow germplasm base within each of the cultivated tetraploid species: *G. hirsutum* and *G. barbadense* had made the discovery of useful SNP markers more difficult. Using highly conservative parameters such as minimum coverage of 8× at each SNP and 20% minor allelic frequency, a total of 11,834 and 1679 non-genic SNPs were previously identified between accessions of *G. hirsutum* and *G. barbadense* in genome reduction assemblies, respectively, by Byers et al. [40] As a part of the same study, an additional 4327 genic SNPs were also identified between accessions of *G. hirsutum* in the EST assembly. The transcriptome sequencing has been extended to SNP marker discovery [41]. Using oligonucleotide microarrays, SNP markers in seven differentially expressed EXPANSIN transcripts in early cotton fiber development have been identified [39]. The variant analysis in cotton revealed 27,956 indels and 149,616 SNPs from 268,786 EST assembled contigs [42]. Over 1000 SNPs from 92 single-copy polymorphic loci have been characterized in tetraploid cotton [43].

Despite significant success in cotton breeding and genetic improvement for fiber-related traits, the genome-wide physical or high-density genetic maps are scarcely available in cotton due to its large genome size. Approximately 5000 markers are needed to fully saturate the cotton genome [44]. A diverse list of markers associated with cotton is available in the cotton marker database (CMD) [45]. CottonGen is a comprehensive database that contains genetic, breeding, and genomic data in cotton including 49 genetic maps, ~24,000 markers, ~1000 quantitative trait loci (QTL) linked to more than 30 agronomic traits, ~18,000 genes/transcripts and ~460,000 expressed sequence tags (ESTs) [46]. Cotton QTLdb contains a total of 2274 QTLs that have been identified in intraspecific populations [47]. Recently, using genotype-by-sequencing method ultra-dense inter-specific genetic map that comprised of ~5 million SNPs distributed unevenly across 26 linkage groups has been constructed in allotetraploid cotton [4].

Currently, there is a growing repertoire of available NGS platforms that support whole-genome sequencing, re-sequencing, and exome sequencing such as NextSeq, HiSeq and HiSeq-X series (Illumina, Inc.), and IonProton (Life technologies). While ABI3500 series (Applied Biosystems), MiniSeq and MiSeq (Illumina, Inc.), and Ion S5 and Ion PGM (Life technologies) are suitable for amplicon sequencing. Extensive genomic information related to gene regulation, genetic, and epigenetic codes are available at GenBank (NCBI), EBI (EMBL) and DDBJ (NIG). In addition, raw sequencing data obtained from NGS platforms as a result of sequencing/re-sequencing efforts has been reported at NCBI Sequence Read Archive (SRA). NCBI-Genome hosts information related to genomes including sequences, physical and genetic maps, chromosomes, assemblies, and annotations. Single-nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites and non-polymorphic variants are also available at dbSNP. The databases available for repeats are Repbase, RepeatsDB, Dfam, P-MITE and the PGSB Repeat Database, while the information associated with plant *cis*-acting regulatory DNA elements in promoter regions are accessible by PLACE and PlantCARE.

Assembler category	Assembly program	References
Overlap layout consensus (OLC)-based	Celera assembler, Arachne, CAP/PCAP/CAP3, CABOG, MIRA, Newbler/GS de novo assembler, Edena	[49–53]
Eulerian-based/*De Bruijn* graph (DBG)-based	Euler, Velvet, AllPaths, ABySS, Trans-ABySS, SOAP and SOAPdenovo, miraEST, Oases and Rnnotator	[54–57]
Greedy assemblers	SSAKE, VCAKE, and SHARCGS	[58, 59]

Table 1. Genome and transcriptome assemblers.

Three major categories of assemblers the widely used in whole-genome and transcriptome assembly are as follows: Overlap Layout Consensus (OLC)-based, Eulerian-based/*De Bruijn* Graph (DBG)-based and Greedy assemblers. A table of various types of available sequence assembly. Tools are summarized in **Table 1** [48].

Bioinformatics utility	Programs	References
Mapping to reference genome	SAMtools, Picard, BEDtools, BWA, BWT, Bowtie2, CLC Genomics Workbench	[77–79]
Whole-genome alignment	LASTZ, AVID, VISTA	[80, 81]
Multiple sequence alignment	MUSCLE; MEGA5	[82]
Phylogenetic analysis	EMBOSS; PAML	[83]
Homology-based search	NCBI BLASTN, BLASTP	
Repeat regions	RepeatMasker, Tandem Repeat Finder, MGEScanLTR, TransposonPSI, MITE Digger,	[84–86]
Gene structure	Augustus, GeneMark, FGENESH, GLAD, GeneWise, GenScan, GlimmerHMM, HMMER, GeneID, SNAP, and GLEAN	[87–90]
Gene annotation	BLAST2GO, BLAT, PASA	[91, 92]
Orthologous sequence search	OrthoMCL	[93]
Paralogous sequence search	Mcscan, McscanX	[94]
Transcriptome tools	Trinity, Tuxedo, TopHat, Cufflinks	[95, 96]
Pseudogenes	Pseudopipe	[97]
Single-nucleotide polymorphisms detection	GATK pipeline, SnpEff	[98, 99]
Copy number variation detection	CNVKit	[100]
Linkage mapping	MapMaker, JoinMap, MapManager QTX	[101, 102]
QTL mapping	MapQTL, Windows QTL Cartographer	[103, 104]

Table 2. Commonly used bioinformatics programs in structural, functional and regulatory genome analyses of cotton.

http://dx.doi.org/10.5772/64325

The cotton genomes, D_5 and A_2, were assembled using SOAPdenovo, while A_tD_t utilized overlap-layout-consensus (OLC) assembly followed by SOAPdenovo to achieve a coverage of over 100×. Other tools commonly used in analyzing the structural and functional analysis in cotton genomes are summarized in **Table 2**.

3. Transcriptome analysis in cotton

The RNA-sequencing (RNA-Seq) and analysis research has exploded in parallel to the genome sequencing methodologies. RNA-Seq has been widely adopted due to its high accuracy in characterization and quantification of transcriptomes [60]. The major objectives of transcriptomics are to catalog the transcripts of all species, to predict the genes with biological functions and to quantify the divergence in gene expression during varied spatial, temporal, and ecological conditions. In the beginning, RNA-Seq was focused on deciphering the transcriptomes of model species, but it was quickly extended to non-model species due to the portability of the method because RNA-Seq analysis can be achieved with or without a reference genome. There was a huge shift in characterizing the high-resolution transcriptomes in plant species from EST-sequencing to microarray analysis and currently to RNA-Seq. In plants, RNA-Seq was first adopted for sequencing the transcriptome of *Arabidopsis* using massive parallel 454 sequencing platform [61]. Since then, transcriptomes of several plant species under differing genetic and physiological states have been sequenced including cotton [62]. The currently available NGS platforms that support whole-transcriptome or total RNA/mRNA analysis are GS FLX+ System (Roche 454), SOLiD and IonProton (Applied Biosystems), MiSeq, HiSeq series and NextSeq (Illumina, Inc.). The platforms that support targeted RNA analysis are GS Junior, IonTorrent PGM, and MiniSeq.

Several independent studies used RNA-Seq analysis to determine spatial-, temporal-, tissue-, genotype- and genome-specific, and stress-induced (abiotic and biotic) expression in both diploid and tetraploid cottons. Some recent studies are discussed here. Several groups have sequenced the cotton fiber cDNA or ESTs [63, 64] and transcriptome or mRNA [65, 66] using RNA-Seq, while others utilized microarrays for analysis [67, 68]. Recently, 40,976, 41,330, 66,434, and 80,876 high-confidence protein-coding genes have been predicted in *G. raimondii*, *G. arboreum*, *G. hirsutum* and *G. barbadense*, respectively. Cotton fibers are unique have the most elongated cells in plants, and they account to 40–50% of the whole transcriptome in Upland cotton [69].

Differential gene expression analyses in tetraploid and diploid cottons have been carried out [65]. Two separate studies in root transcriptome analyses of *G. hirsutum* revealed 519 and 1530 differentially expressed transcripts between well-watered and water-deficit conditions, respectively [66, 70]. The comparative transcriptome analysis of developing cotyledon and embryo axis in cotton revealed 17,384 differentially expressed unigenes between two tissues. Of these, ~8000 unigenes were down-regulated and ~10,000 unigenes were up-regulated in cotyledons [71]. Transcriptome analysis of interspecific hybrid F1, synthetic and natural allopolyploid cotton revealed homeolog expression bias (relative contribution of homeologs

towards gene expression) and expression level dominance bias (over-all expression from both homeologs) towards A-genome in diploid and natural allopolyploid cotton but not in synthetic cotton [72]. Using RNA-Seq, the global transcriptome profiles of developing cotton fibers from four wild and five domesticated cottons were compared, and this study identified over 5000 differentially expressed genes during the primary and secondary cell wall synthesis between wild and domesticated cottons [73]. Comparative RNA-Seq analysis of *G. hirsutum* from young, mature, and different senescence stages of leaves identified 3624 differentially expressed genes during leaf senescence [74]. Comparative transcriptome profiling of *G. hirsutum* and *G. davidsonii* (diploid wild species) further characterized 4744 and 5337 differentially expressed salt-stress responsive genes from roots and leaves of *G. davidsonii* [75]. The potential role of signalling pathways, ethylene pathway in fiber elongation, and receptor-like kinases (RLKs) in cell wall integrity have been proposed in determining fiber quality using comparative RNA-Seq analyses of near isogenic lines (NILs) in *G. hirsutum* [76].

The major mRNA repositories are Crops-ESTdb, NCBI dbEST, UniGene, RefSeq, and GEO; plant-specific microarray databases are plant expression database (PLEXdb) and plant co-expression database (PLANEX). Tools for RNA-Seq expression analysis are R/bioconductor packages such as RSEM, DEGseq, DESeq, edger, and baySeq. The open access expVIP is a web-based tool for visualizing, analyzing and comparing RNA-Seq data for conducting gene expression analysis in diploid and polyploid plant species. Single-nucleotide variant analysis in RNA-Seq dataset can be performed by SAMtools followed by Genome Analysis Toolkit (GATK). RobiNA is a web-based tool for accessing and comparing EST, microarray, and RNA-Seq databases.

Though several OLC-based, DBG-based, and greedy assemblers discussed above have been utilized in RNA-Seq analyses, Trinity is the most promising tool in generating transcriptome assembly. The Tuxedo pipeline contains a series of tools including Bowtie and TopHat for aligning the RNA-Seq reads, Cufflinks for assembling the mapped reads, Cuffdiff for identifying differentially expressed genes and CummeRbund for visualizing differentially expressed genes and transcripts [96]. The parameters such as false discovery rate (FDR) ≤ 0.001, fold change (log2FC) ≥ 2, and P-value ≤ 0.05 are considered in comparative transcriptome profiling using RNA-Seq [105]. Using Trinity and Tuxedo, novel genes, and splice variants can be determined. In RNA-Seq analysis, FPKM (Fragments per kilobase of transcript per million mapped reads) and RPKM (Reads per kilobase of transcript per million mapped reads) values are often used in quantifying gene expression. Express tool has been used in detecting transcript abundance in *G. hirsutum*.

Additionally, to annotate gene functions, homology-based methods such as BLASTX, BLAST2GO, and BLAT have been used. PASA and EVM have been used in annotating 3' and 5' UTR regions and alternative splice events in the transcript assemblies. Functional annotation of transcript assembled fragments (TAFs) can be done using BLASTX (E-value 1×10^{-6}) or BLASTP (E-value 1×10^{-5}) against protein databases non-redundant (NR), SwissProt, and TrEMBL. Gene ontology annotation is conducted by using BLAST2GO and AgriGO analysis. InterPro is used to annotate motifs and domains by comparing TAFs with publicly available databases such as Pfam, PRINTS, PROSITE, ProDom, and SMART. The databases used for

http://dx.doi.org/10.5772/64325

predicting pathways are KEGG, PANTHER, Pathguide, PlantCyc, BioCyc, and MetaCyc. Using BLASTP (E-value =< 1e-5) and OrthoMCL gene clusters between sub-genomes in tetraploid cotton have been classified. FASTQC and FASTX toolkit are widely used in filtering low quality and determining the quality of RNA-Seq reads. Reads contaminated with adapters are removed using Trimmomatic software. HMMER has been used in identifying transcription factor (TF) gene families in tetraploid cotton using the PlnTFDB. Using D_5 genome as a reference, homeologous genes/syntenic blocks between A_t and D_t sub-genomes have been identified using MCscanX. These tools were summarized in **Table 2**.

The comparative transcriptome analysis of *G. hirsutum* with its progenitors ameliorates the complexities associated with the co-existence of A_t and D_t genome transcripts in allotetraploid species. Moreover, comparative transcriptome mapping of sub-genomes leads to identification of high-confidence transcriptional modules that are evolutionarily conserved and are specific to the genus *Gossypium*. Identifying sub-genome specific transcripts and analyzing their role in cotton gene regulation would help in developing novel biomarker tools associated with various complex polygenic traits in cotton. Alternative splicing in eukaryotes results in transcriptome and proteome diversity. The presence of abundant splice variants and novel transcriptionally active regions (nTARs) has been identified in *Arabidopsis* and rice using RNA-Seq, but majority of these novel transcripts did not overlap with known protein-coding genes and open reading frames (ORFs), suggesting their potential role in post-transcriptional gene regulation and novel transcript/gene formation [105]. Identification of alternative splice events and nTARs in tetraploid cotton is computationally challenging due to ploidy, duplication, chromosomal rearrangements, and presence of homeologous bias.

4. Small RNA analysis in cotton

Napoli et al. (1990) first identified the phenomenon of RNA interference (RNAi) and referred to it as co-suppression in plants [106]. Hamilton & Baulcombe (1999) reported a group of antisense RNAs that mediate in post-transcriptional gene silencing (PTGS) in plants that target both cellular and viral mRNAs [107]. The world of sRNAs exploded with the discovery of miRNAs (microRNA) in *C. elgans* [108]. In plants, miRNAs were first identified in *A. thaliana* and studied extensively in other plant species using comparative genomic, cloning, or sequencing approaches [109]. Unlike in animals, plant miRNAs are mainly found in intergenic and intronic regions [110]. MiRNA genes are mostly localized and form clusters and they transcribe together as a single transcriptional unit [111]. MiRNA discovery has gained momentum in diverse plant species. The majority of plant miRNAs identified to date negatively regulate the target gene expression at the post-transcriptional level. MiRNAs regulate environmental stress response, metabolic processes, organogenesis, growth and development [112]. The miRNAs associated with seed-specific transcription factors have been identified and were found to influence differentiation and developmental timing [113].

The reduction in cotton fiber quality and yields is mainly attributed to several biotic and abiotic factors, and these complex traits are also regulated by small RNAs (sRNAs) which are mostly

by miRNAs. Recent advances in cotton genomics have provided an impetus to pursue the discovery of novel miRNAs in the cotton genome [3, 114, 115]. MiRNAs and small interfering RNAs (siRNAs) are the two major classes of endogenous small non-coding RNAs that regulate the gene expression in plants. Earlier reports suggest that miRNAs mediate a variety of functional roles such as developmental timing, cell proliferation, differentiation, morphogenesis, defense response, and signal transduction [116–118]. Cotton miRNA research is limited due to the lack of genomic information of cultivated cotton until recently. Cotton miRNAs have been identified in *G. hirsutum* (A_tD_t), *G. barbadense* (A_tD_t), *G. herbaceum* (A_1), *G. arboreum* (A_2), and *G. raimondii* (D_5). Besides their role in fiber growth, development, initiation, and elongation, some miRNAs have been implicated in biotic and abiotic stress responses in cotton. Moreover, many cotton-specific miRNAs have been identified, but their functions need experimental validation. Recent studies that use qRT-PCR and degradome sequencing of cotton fiber RNA suggest that miRNAs play an important role in cotton fiber development. Similarly, Wang et al. (2012) identified 73 miRNAs that belong to 49 families in *G. arboreum* using homology-based approach [115]. Zhang et al. (2013) identified 65 novel miRNAs and their candidate gene targets in *G. hirsutum* using transcript sequences of *G. raimondii* [116]. MiRNA was also analyzed in salt tolerance response in *G. hirsutum*. The miRNVL5 precursor was first discovered in *Arabidopsis* and later found in cotton. *G. arboreum* miRNA was studied in response to Cotton Leaf Curl Disease [118].

NGS technologies that are mentioned in transcriptome analysis are also applicable in small RNA sequencing. MiRNAs are highly conserved across the species and can be predicted by using homology-based methods. Various resources commonly used in analyses of small RNAs are described in **Table 3**.

Purpose	Tool	References
miRNA precursors	RNAfold, mfold	[123]
miRNA databases	miRBase, MicroRNAdb	[124]
miRNA prediction	miRCat, miREAP, miRPlant, miRDeep-P, miRNAKey	[125–128]
Gene targets for miRNAs	miRanda and psRNATarget	[129, 130]
lncRNA homologs	RefSeq, NONCODE, lncRNADB, PLncDB, PNRD, PlncRNADB, and PLNlncRbase	[131–134]
Epigenomic databases	pENCODE, GEO, NGSmethDB, MethBase, plant methylome-db	[135–137]
ChIP-Enriched region identification	HOMER	[138]
Methylation detection	CyMATE, MeQA, MEDIPS, FASTmC	[139, 140]
Bisulphite data analysis	Methylpipe	[141]

Table 3. Bioinformatics tools used in regulatory analysis of cotton.

The other classes of small RNAs identified in *Gossypium* include trans-acting small interfering RNA (TasiRNA), long intergenic noncoding RNA (lincRNA), and long noncoding natural

http://dx.doi.org/10.5772/64325

antisense transcript loci (lncNAT). Long noncoding RNAs (lncRNAs) are 200 bp in length, rich in repetitive sequences, preferentially expressed in a tissue-, genome- and lineage-specific manner, are poorly conserved and participate in various regulatory processes. Higher overall methylation levels are exhibited by lncRNAs when compared with protein-coding genes, while their expression is less affected by gene-body methylation. The lncRNAs play a pivotal role in regulating lignin metabolism, cotton fiber initiation, and elongation. There are two main classes of lncRNAs, long intergenic noncoding RNAs (lincRNAs), and long noncoding natural antisense transcript (lncNATs), which are structurally similar but vary in number and length of exons and transcripts. In cotton (*Gossypium spp.*), 30550 lincRNA and 4718 lncNAT loci have been reported. However, homeolog expression bias of lncRNAs has been identified in subgenomes of polyploid species when compared to wild parents [119]. In a similar study, 3510 lincRNAs and 2486 lncNATs that play an integral role in fiber initiation and elongation have been identified in *G. arboreum* using strand-specific RNA sequencing (ssRNA-Seq) of cotton fibers and leaves [120]. A preferential expression of lncRNAs was observed (~50%) than for protein coding genes (~20%) during rapid fiber elongation, thus establishing their function in fiber development. The direct role of 21–22 nt siRNA derived from GhMML3 gene (MYB-MIXTA-like transcription factor 3) that encodes a lncNAT, implicated in fiber development has been investigated in *G. hirsutum* [119].

5. Epigenome analysis in cotton

Post-transcriptional gene silencing (PTGS) is mainly associated with the regulation of gene expression directly by targeting the mRNA, while transcriptional gene silencing (TGS) is mostly coordinated by epigenetic modifications such as DNA methylation, histone modifications (methylation and acetylation), chromatin remodelling factors, polycomb group (PcG), and trithorax group (trxG) of proteins [121]. These epigenetic modulators can be identified by using homology-based searching in Upland cotton and its closest progenitor species, which will aid in understanding the epigenetic landscape of Upland cotton. The role of DNA methylation has been implicated in plant growth and development, by maintaining active and inactive chromatin states and in gene silencing by employing canonical and non-canonical RNA-directed DNA methylation (RdDM) pathways [121]. For eliciting both the pathways, core RNAi machinery including DICER-LIKE (DCL), RNA-DEPENDENT RNA POLYMERASE (RDR), ARGONAUTE (AGO) proteins, RNA polymerases (Pol IV and V), and plant specific-DNA methylases are required [122].

In plants, DNA methylation occurs in both CG and non-CG (CHG and CHH) contexts, where "H" denotes A, T, or C. Although DNA is primarily methylated in CG context in both plants and mammals, the non-CG contexts are abundant in plants when compared to mammals [142]. The CG, CHG methylation contexts are symmetrical and are maintained by DNA replication, while CHH context is established by *de novo* DNA methylation. The DNA methylation within the genome can occur in promoter, gene-body, transposable elements, and repeat regions [143]. Previous reports suggest that 35–43%, 30–41%, and 30–32% of loci were methylated in *A. thaliana* ecotypes [144], *Brassica oleracea* accessions [145], and *G. hirsutum* accessions [146],

respectively. The role of DNA methylation in fiber growth and gene silencing has been proposed in cotton. It has also been suggested that CHH methylation may have a role in developmental timing in cotton. Annual fluctuations in DNA demethylases and methyltransferases have shown opposite trends in their abundance in cotton ovules. Also, substantial changes in CHH methylation was noticed in the promoter regions of three key genes ETHYLENE RESPONSIVE FACTOR 6 (ERF6), SUPPRESSION OF RVS 161 DELTA 4 (SUR4) and 3-KETOACYL-COA SYNTHASE 13 (KCS13) that regulate cotton fiber growth. Development of homozygous RNAi lines, specifically targeting demethylases and methyltransferases, will aid in determining DNA methylation patterns in cotton fiber growth and development [147].

The variation in epigenetic modifications has been observed in spontaneous reciprocal hybrids of *Rosa sect. Caninae* and *Rubigineae* when compared to their parental species using cDNA–AFLPs and methylation-sensitive amplified fragment length polymorphisms (MSAPs), suggesting the biased contribution of DNA methylation from parents to polyploid hybrids [148]. Similarly, genotype-specific and tissue-specific variations in DNA methylation have been identified in cotton by using MSAP with methylation insensitive enzyme (BsiSI). Their results suggest that CHG methylation is more diverse than the other two contexts (CG and CHH) and this work established a relationship between DNA methylation and fiber development but failed to establish the correlation between epigenetic regulation and fiber quality. To achieve this, a robust study involving diverse genotypes/accessions and tissues is required. Cytosine methylation is among one of the key epigenetic regulatory processes that likely silence duplicated genes in polyploid crop species such as cotton. The levels, patterns, and diversity of methylation-polymorphism in CG context have been investigated in 20 accessions of Upland cotton (*G. hirsutum*) to identify 32 methylation-polymorphic cytosine sites using MSAP [63]. The introgression of exotic DNA fragments from wild parents (*G. bickii*) into cultivated Upland cotton (*G. hirsutum*) has been investigated using AFLP and MSAP analysis to identify ~2000 genomic and ~800 epigenetic sites [66]. This study identified ~0.5% of alien DNA segments and ~0.7% of genetic variation in the genomes of introgression lines. Though the overall methylation content is close to the wild parent, the context-specific methylation varied significantly in introgression lines [149].

Using methylcytosine-sequencing (MethylC-Seq) and RNA-sequencing (RNA-Seq) analysis, variation in CHH methylation during ovule, and fiber development was reported in cotton [150]. Further, this study suggested that CHH hypermethylation triggered RdDM-dependent methylation in promoters, while RdDM-independent methylation occurred in TEs and nearby genes, thus facilitating ovule and fiber development in cotton. Moreover, the contribution of CHG and CHH methylation in genic regions towards homeologous expression patterns in A_t, or D_t subgenomes have been proposed [150]. Evolution is a very slow process when it is purely dependent on mutation and recombination. In plants, it is believed that there are additional forces that accelerated the process of evolution, which include interspecific hybridization. However, the mechanism of acceleration is less understood. The possible underlying mechanism has been proposed in wheat [151]. It was suggested that the interaction of alien nuclear DNA with cytoplasmic macromolecules as a result of interspecific hybridization can potentially trigger a network of epigenetic changes in nuclear DNA, thus altering the expression of

http://dx.doi.org/10.5772/64325

genes and genetic pathways associated with physiological processes, which may serve as a possible source of variation that facilitate evolutionary process. This idea sheds light into development of epigenomic-segregating lines in crop plants, including cotton [151].

Screening the core RNAi machinery of important DNA methylases such as DOMAINS REARRANGED METHYLASE 1/2 (DRM1/2), CHROMOMETHYLASE 2/3 (CMT 2/3), and demethylases such as ROS1 and DEMETER (DME) in Upland cotton, and its closest progenitor species will aid in elucidating the key regulatory mechanisms in whole-genome duplication, chromosomal rearrangements, dosage compensation, and evolutionary advantage of being polyploids. Chromosome painting techniques such as fluorescence *in situ* hybridization (FISH), variations of FISH and genomic *in situ* hybridization (GISH) are not only useful in determining chromosome structure and distribution of hetero/euchromatin but also in understanding epigenomic patterns and evolution of polyploid species [152].

Although several databases and repositories are available for storing, cataloging and normalizing animal, and mammalian epigenomic data, they are limited for plants. The popular resource is NCBI epigenomics sample browser to access the diverse collection of epigenomic data sets including genome-wide DNA methylation maps and histone modifications. However, only *A. thaliana* dataset is currently available in the sample browser. Other than NCBI sample browser, the popular visualization tools for epigenomic data are as follows: UCSC epigenome browser, Ensembl encyclopedia of DNA elements (ENCODE), and WashU epigenome browser. A plant ENCODE (pENCODE) collects and curates epigenomic data from a wide range of plant species to compare, annotate, and understand mechanisms behind plant evolution. Epigenomics of Plants International Consortium (EPIC) provides a platform to share the protocols, methods, and results across the research labs to address basic questions of genetics and genome regulation beyond routine whole-genome sequencing. GEO is one of the seminal repositories for hosting epigenome data from various resources including animals (human) and plants (*Arabidopsis* and Maize). NGSmethDB and MethBase specifically store whole-genome single-base resolution data obtained from whole-genome bisulphite-sequencing (WGBS/BS-Seq) from diverse organisms. Currently, very few plant-specific epigenome databases are publicly available and include plant methylome-db, *Arabidopsis* Epigenome Browser, and Tomato Epigenome Database. Though the cotton-specific epigenome database is not publicly available, methylomes of evolutionarily close relatives such as *Ricinus communis* and *Theobroma cacao* are available at plant methylomes-db with user restrictions [84]. Epigenetic resources and tools were summarized in **Table 3**.

The wide applications of chromatin Immunoprecipitation-sequencing (ChIP-Seq) analysis include understanding protein-DNA interactions, histone modifications, chromatin states, and DNA methylation. Methylation and acetylation marked peaks can be identified using spatial clustering for identification of ChIP-enriched regions (SICER), and later peaks can be annotated by using HOMER (motif search algorithm). Other peak-finding algorithms, Model-based Analysis of ChIP-Seq (MACS), and PeakSeq have been extensively used in analysis. BS-Seq analysis helps in identification of genome-wide differentially methylated regions (DMRs) along with tools such as Bismark, BSMAP, and BS-Seeker can be used. Cytosine Methylation Analysis Tool for Everyone (CyMATE) is a plant-specific tool for analyzing BS-Seq data. A

web-based tool, QUMA (quantification tool for methylation analysis) is suitable for BS-Seq analysis, primarily for estimating CG methylation. MeDIP analysis helps in detecting methylated cytosines in both CG and non-CG contexts and tools such as MeQA and MEDIPS can be used. FASTmC Webtool is available for comparative analysis of CG, CHG, and CHH DNA methylation levels in non-model species. Methpipe is useful in analyzing bisulphite-sequencing data including BS-Seq, WGBS, and reduced representation bisulfite sequencing (RRBS). Integrative epigenome analysis can be done using methylPipe and compEpiTools. Differentially methylated regions (DMRs) have been identified in *Arabidopsis* by using methylpipe. ChIP-Seq analysis in R (CSAR) accurately account the protein bound regions in the genome in plants. PolyCat determines mapping bias between genomes that may have resulted due to SNPs; hence, it can be used in comparing allopolyploid, *G. hirsutum* against diploid *G. raimondii* or *G. arboreum*. Moreover, it accepts data from diverse NGS platforms including RNA-Seq, BS-Seq, ChIP-Seq, and SNP analysis [153].

6. Future direction

The reduction in cost of sequencing per base, increase in throughput and read size, and availability of better algorithms have significantly facilitated the integration of one or more genomic methods to address biological questions. For instance, continued progress in detection of SNPs and CNVs from sequencing data and combining this with genotyping-by-sequencing methods will be helpful in crop improvement. Rapid advances in genomics, bioinformatics, and computational biology in the past two decades have already facilitated generation and screening of large datasets, and thus have ushered us into the "Big Data" era of genomics. The plant science community still lacks full-fledged computational infrastructure, data curation, and novel tools to extract information from the massive datasets. It should be realized that progress in analyzing polyploid genomes will continually need improvements because the nature of the data generated from high-throughput technologies is voluminous, heterogeneous, and often in unstructured forms. The concepts of parallel processing, cloud computing, and machine learning offer promising solutions for managing and analyzing large datasets. The other important limiting factors are trained professionals to access and analyze these resources. Nonetheless, progress is being made as we enter the burgeoning age of "Big Data" to tackle genomes of widely grown and utilized crops such as cotton using bioinformatics tools.

Acknowledgements

The authors acknowledge Dr. Avinash Sreedasyam of HudsonAlpha Institute of Biotechnology and Mr. Joshua Reid at Alabama A&M University for reviewing this book chapter. Also authors would like to thank Dr. Ibrokhim Abdurakhmonov and Ms. Ivona Lovric for their valuable suggestions and comments in improving this book chapter.

Author details

Venkateswara R. Sripathi[1*], Ramesh Buyyarapu[2], Siva P. Kumpatla[2], Abreeotta J. Williams[1], Seloame T. Nyaku[3], Yonathan Tilahun[4], Venu Kalavacharla[5] and Govind C. Sharma[1]

*Address all correspondence to: v.sripathi@aamu.edu

1 Center for Molecular Biology, Alabama A&M University, Normal, AL, USA

2 Dow AgroSciences, Indianapolis, IN, USA

3 Crop Science Department, College of Basic and Applied Sciences, University of Ghana, Legon-Accra, Ghana

4 Cooperative Extension, School of Agriculture & Applied Sciences, Langston University, Langston, OK, USA

5 Molecular Genetics & EpiGenomics Laboratory, College of Agriculture & Related Sciences, Delaware State University, Dover, DE , USA

References

[1] Varshney RK, Terauchi R, McCouch SR. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biology. 2014;12:e1001883. doi:10.1371/journal.pbio.1001883

[2] Joosen RV, Ligterink W, Hilhorst HW, Keurentjes JJ. Advances in genetical genomics of plants. Current Genomics. 2009;10:540-549. doi:10.2174/138920209789503914

[3] Abdurakhmonov IY, Ayubov MS, Ubaydullaeva KA, Buriev ZT, Shermatov SE, Ruziboev HS, Shapulatov UM, Saha S, Ulloa M, Yu JZ, Percy RG. RNA interference for functional genomics and improvement of cotton (*Gossypium* sp.). Frontiers in Plant Science. 2016;7:202. doi:10.3389/fpls.2016.00202

[4] Wang S, Chen J, Zhang W, Hu Y, Chang L, Fang L, Wang Q, Lv F, Wu H, Si Z, Chen S. Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. Genome Biology. 2015;16:1-8. doi:10.1186/s13059-015-0678-1

[5] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. Journal of Applied Genetics. 2011;52:413-435. doi:10.1007/s13353-011-0057-x

[6] Buyyarapu R, Kantety RV, John ZY, Xu Z, Kohel RJ, Percy RG, Macmil S, Wiley GB, Roe BA, Sharma GC. BAC-pool sequencing and analysis of large segments of A12 and

D12 homoeologous chromosomes in Upland cotton. PloS One. 2013;8:e76757. doi: 10.1371/journal.pone.0076757

[7] Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. BioMed Research International. 2014;2014:309650. doi:10.1155/2014/309650

[8] Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. BioMed Research International. 2014;2014:309650. doi:10.1155/2014/309650

[9] Fryxell P. A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). Rheedea. 1992;2:108-165

[10] Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. Advances in Agronomy. 2003;78:13986. doi:10.1016/S0065-2113(02)78004-8

[11] Wendel JF, Schnabel A, Seelanan T. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proceedings of the National Academy of Sciences of the United States of America. 1995;1995:92280-92284. doi: 10.1073/pnas.92.1.280

[12] Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, Baker D, Long Y, Meng J, Wang X, Liu S. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nature Biotechnology. 2011;29:762-766. doi:10.1038/nbt.1926

[13] Hendrix B, Stewart JM. Estimation of the nuclear DNA content of *Gossypium* species. Annals of Botany. 2005;95:789-797. doi:10.1093/aob/mci078

[14] Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C. The draft genome of a diploid cotton *Gossypium raimondii*. Nature Genetics. 2012;44:1098-103. doi:10.1038/ng.2371

[15] Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nature Genetics. 2014;46:567-572. doi:10.1038/ng.2987

[16] Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nature Biotechnology. 2015;33:524-30. doi:10.1038/nbt.3208

[17] Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, Hulse-Kemp AM. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnology. 2015;33:531-537. doi:10.1038/nbt.3207

[18] Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, Chen JD, Chen JJ, Chen DY, Zhang L, Zhou Y. *Gossypium barbadense* genome sequence provides insight into the evolution of

extra-long staple fiber and specialized metabolites. Scientific Reports. 2015;5:14139. doi: 10.1038/Srep14139

[19] Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L, Li Y. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Scientific Reports. 2015;5:17662. doi:10.1038/Srep17662

[20] Sripathi VR. Towards understanding the genome of diploid cotton, *Gossypium herbaceum* using deep sequencing. In: Plant and Animal Genome XXIII Conference; 10-1437 January 2015; San Diego, CA, USA.

[21] Wei L, Cao X. The effect of transposable elements on phenotypic variation: insights from plants to humans. Science China Life Sciences. 2016;59:24-37. doi:10.1007/s11427-015-4993-2

[22] Springer NM, Lisch D, Li Q. Creating order from chaos: epigenome dynamics in plants with complex genomes. The Plant Cell. 2016;28:314-325. doi:10.1105/tpc.15.00911

[23] Wang K, Huang G, Zhu Y. Transposable elements play an important role during cotton genome evolution and fiber cell development. Science China Life Sciences. 2015;59:112-121. doi:10.1007/s11427-015-4928-y

[24] Gao D, Li Y, Do Kim K, Abernathy B, Jackson SA. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome Biology. 2016;17:1-7. doi:10.1186/s13059-015-0867-y

[25] Buyyarapu R, Kumaptla S, Elango N, Chen W, Greene T. KASPar: An Efficient and Cost-effective Technology for SNP Genotyping in Cotton. In: Proceedings of the Beltwide Cotton Conferences; 2010; New Orleans.

[26] Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. SNP markers and their impact on plant breeding. International Journal of Plant Genomics. 2012;2012:728398. doi: 10.1155/2012/728398

[27] Zhang HB, Li Y, Wang B, Chee PW. Recent advances in cotton genomics. International Journal of Plant Genomics. 2008;2008:742304. doi:10.1155/2008/742304

[28] Wendel JF, Brubaker CL, Percival AE. Genetic diversity in *Gossypium hirsutum* and the origin of Upland cotton. American Journal of Botany. 1992;1992:1291-1310.

[29] Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH. A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics. 1994;138:829-847.

[30] Abdalla AM, Reddy OU, El-Zik KM, Pepper AE. Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theoretical and Applied Genetics. 2001;102:222-229. doi:10.1007/s001220051639

[31] Liu S, Cantrell RG, McCarty JC, Stewart JM. Simple sequence repeat–based assessment of genetic diversity in cotton race stock accessions. Crop Science. 2000;40:1459-1469. doi:10.2135/cropsci2000.4051459x

[32] Buyyarapu R, Kantety RV, Yu JZ, Saha S, Sharma GC. Development of new candidate gene and EST-based molecular markers for *Gossypium* species. International Journal of Plant Genomics. 2011;2011:894598. doi:10.1155/2011/894598

[33] Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD, Frelichowski J, Giband M, Hague S, Hinze LL, Kochan KJ. Development of a 63K SNP array for cotton and high-density mapping of intra-and inter-specific populations of *Gossypium* spp.G3: Genes| Genomes| Genetics. 2015;2015:g3-115. doi:10.1534/g3.115.018416

[34] Kumpatla SP, Mukhopadhyay S. Mining and survey of simple sequence repeats in 2 expressed sequence tags of dicotyledonous species. Genome. 2005;48:985-998. doi:3 10.1139/g05-060

[35] Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. Trends in Biotechnology. 2007;25:490-498. doi:10.1016/j.tibtech.2007.07.013

[36] Shen X, Guo W, Lu Q, Zhu X, Yuan Y, Zhang T. Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. Euphytica. 2007;155:371-380. doi:10.1007/s10681-006-9338-6

[37] Yang C, Guo W, Li G, Gao F, Lin S, Zhang T. QTLs mapping for Verticillium wilt resistance at seedling and maturity stages in *Gossypium barbadense* L. Plant Science. 2008;174:290-298. doi:10.1016/j.plantsci.2007.11.016

[38] Shen X, Van Becelaere G, Kumar P, Davis RF, May OL, Chee P. QTL mapping for resistance to root-knot nematodes in the M-120 RNR Upland cotton line (*Gossypium hirsutum* L.) of the Auburn 623 RNR source. Theoretical and Applied Genetics. 2006;113:1539-1549. doi:10.1007/s00122-006-0401-4

[39] An C, Saha S, Jenkins JN, Scheffler BE, Wilkins TA, Stelly DM. Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. Molecular Genetics and Genomics. 2007;278:539-553. doi:10.1007/ s00438-007-0270-9

[40] Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. Development and mapping of SNP assays in allotetraploid cotton. Theoretical and Applied Genetics. 2012;124:1201-1214. doi:10.1007/s00122-011-1780-8

[41] Ashrafi H, Hulse-Kemp AM, Wang F, Yang SS, Guan X, Jones DC, Matvienko M, Mockaitis K, Chen ZJ, Stelly DM, Van Deynze A. A long-read transcriptome assembly of cotton (*Gossypium hirsutum*) and intraspecific SNP discovery. The Plant Genome. 2015;8:1-14. doi:10.3835/plantgenome2014.10.0068

[42] Xie F, Sun G, Stiller JW, Zhang B. Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. PloS One. 2011;6:e26980. doi: 10.1371/journal.pone.0026980

[43] Van Deynze A, Stoffel K, Lee M, Wilkins TA, Kozik A, Cantrell RG, John ZY, Kohel RJ, Stelly DM. Sampling nucleotide diversity in cotton. BMC Plant Biology. 2009;9:125. doi: 10.1186/1471-2229-9-125

[44] Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B, Hau B. A combined RFLP SSR AFLP map of tetraploid cotton based on a *Gossypium hirsutum*×*Gossypium barbadense* backcross population. Genome. 2003;46:612-626. doi:2 10.1139/G03-050

[45] Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, John ZY, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, Ficklin S. CMD: a cotton microsatellite database resource for *Gossypium* genomics. BMC Genomics. 2006;7:132. doi:10.1186/1471-2164-6 7-132

[46] Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, Zheng P, Jones D, Percy RG, Main D. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Research. 2013;2013:gkt1064. doi:10.1093/nar/gkt1064

[47] Said JI, Knapka JA, Song M, Zhang J. Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum*×*G. barbadense* populations. Molecular Genetics and Genomics. 2015;290:1615-1625. doi:10.1007/s00438-015-1021-y

[48] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086-1092. doi:16 10.1093/bioinformatics/bts094

[49] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences. 2001;98:9748-9753. doi:19 10.1073/pnas.171285098

[50] Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: a whole-genome shotgun assembler. Genome Research. 2002;12:177-189. doi:10.1101/gr.208902

[51] Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Research. 1999;9:868-877. doi:10.1101/gr.9.9.868.

[52] Chevreux B. MIRA: an automated genome and EST assembler. 2007.

[53] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB. Genome sequencing in microfabricated high density picolitre reactors. Nature. 2005;437:376-380. doi:10.1038/nature03959

[54] Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. Bioinformatics. 2004;20:2067-2074. doi:10.1093/bioinformatics/bth205

[55] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008;18:821-829. doi:10.1101/gr.074492.107

[56] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Research. 2009;19:1117-1123. doi: 10.1101/gr.089532.108

[57] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966-1967. doi:10.1093/bioinformatics/btp336

[58] Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007;23:500-501. doi:10.1093/bioinformatics/btl629

[59] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Research. 2007;17:1697-1706. doi:10.1101/gr.6435207

[60] Lee S, Seo CH, Alver BH, Lee S, Park PJ. EMSAR: estimation of transcript abundance from RNA-seq data by mappability-based segmentation and reclustering. BMC Bioinformatics. 2015;16:278. doi:10.1186/s12859-015-0704-z

[61] Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiology. 2007;144:32-42. doi:10.1104/pp.107.096677

[62] Zhu YN, Shi DQ, Ruan MB, Zhang LL, Meng ZH, Liu J, Yang WC. Transcriptome analysis reveals crosstalk of responsive genes to multiple abiotic stresses in cotton (*Gossypium hirsutum* L.). PLoS One. 2013;8:e80218. doi:10.1371/journal.pone.0080218

[63] Lacape JM, Claverie M, Vidal RO, Carazzolle MF, Pereira GA, Ruiz M, Pré M, Llewellyn D, Al-Ghazi Y, Jacobs J, Dereeper A. Deep sequencing reveals differences in the transcriptional landscapes of fibers from two cultivated species of cotton. PloS One. 2012;7:e48855. doi:10.1371/journal.pone.0048855

[64] Li X, Yuan D, Zhang J, Lin Z, Zhang X. Genetic mapping and characteristics of genes specifically or preferentially expressed during fiber development in cotton. PloS One. 2013;8:e54444. doi:10.1371/journal.pone.0054444

[65] Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012;492:423-427. doi:10.1038/nature11798

[66] Bowman MJ, Park W, Bauer PJ, Udall JA, Page JT, Raney J, Scheffler BE, Jones DC, Campbell BT. RNA-Seq transcriptome profiling of Upland cotton (*Gossypium hirsu-*

tum L.) root tissue under water-deficit stress. PloS One. 2013;8:e82634. doi:10.1371/journal.pone.0082634

[67] Rapp RA, Haigler CH, Flagel L, Hovav RH, Udall JA, Wendel JF. Gene expression in developing fibres of Upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. BMC Biology. 2010;8:139. doi:10.1186/1741-7007-8-139

[68] Hinchliffe DJ, Turley RB, Naoumkina M, Kim HJ, Tang Y, Yeater KM, Li P, Fang DD. A combined functional and structural genomics approach identified an EST-SSR marker with complete linkage to the Ligon lintless-2 genetic locus in cotton (*Gossypium hirsutum* L.). BMC Genomics. 2011;12:445. doi:10.1186/1471-2164-12-445

[69] Haigler CH, Betancur L, Stiff MR, Tuttle JR. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Frontiers in Plant Science. 2012;3:104. doi:10.3389/fpls.2012.00104

[70] Park W, Scheffler BE, Bauer PJ, Campbell BT. Genome-wide identification of differentially expressed genes under water deficit stress in Upland cotton (*Gossypium hirsutum* L.). BMC Plant Biology. 2012;12:90. doi:10.1186/1471-2229-12-90

[71] Jiao X, Zhao X, Zhou XR, Green AG, Fan Y, Wang L, Singh SP, Liu Q. Comparative transcriptomic analysis of developing cotton cotyledons and embryo axis. PloS One. 2013;8:e71756. doi:10.1371/journal.pone.0071756

[72] Yoo MJ, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity. 2013;110:171-180. doi:10.1038/hdy.2012.94

[73] Yoo MJ, Wendel JF. Comparative evolutionary and developmental dynamics of the cotton (*Gossypium hirsutum*) fiber transcriptome. PLoS Genetics. 2014;10:e1004073. doi:10.1371/journal.pgen.1004073

[74] Lin M, Pang C, Fan S, Song M, Wei H, Yu S. Global analysis of the *Gossypium hirsutum* L. Transcriptome during leaf senescence by RNA-Seq. BMC Plant Biology. 2015;15:43. doi:10.1186/s12870-015-0433-5

[75] Zhang X, Fan S, Song M, Pang C, Wei H, Wang C, Yu S. Functional characterization of GhSOC1 and GhMADS42 homologs from Upland cotton (*Gossypium hirsutum* L.). Plant Science. 2016;242:178-186. doi:10.1016/j.plantsci.2015.05.001

[76] Islam MS, Fang DD, Thyssen GN, Delhom CD, Liu Y, Kim HJ. Comparative fiber property and transcriptome analyses reveal key genes potentially related to high fiber strength in cotton (*Gossypium hirsutum* L.) line MD52ne. BMC Plant Biology. 2016;16:36. doi:10.1186/s12870-016-0727-2

[77] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-359. doi:10.1038/nmeth.1923

[78] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754-1760. doi:10.1093/bioinformatics/btp324

[79] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2. doi:10.1093/bioinformatics/btq033

[80] Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Research. 2004;32:W273-W279. doi: 10.1093/nar/gkh458

[81] Bray N, Dubchak I, Pachter L. AVID: a global alignment program. Genome Research. 2003;13:97-102. doi:10.1101/gr.789803

[82] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32:1792-1797. doi:10.1093/nar/gkh340

[83] Olson SA. Emboss opens up sequence analysis. Briefings in Bioinformatics. 2002;3:87-91. doi:10.1093/bib/3.1.87

[84] Bedell JA, Korf I, Gish W. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics. 2000;16:1040-1041. doi:10.1093/bioinformatics/16.11.1040

[85] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999;27:573-580.

[86] Yang G. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. BMC Bioinformatics. 2013;14:186. doi:10.1186/1471-2105-14-186

[87] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research. 2005;33:W465-W467. doi: 10.1093/nar/gki458

[88] Salamov AA, Solovyev VV. Ab initio gene finding in *Drosophila* genomic DNA. Genome Research. 2000;10:516-522. doi:10.1101/gr.10.4.516

[89] Hu Y, Comjean A, Perkins LA, Perrimon N, Mohr SE. GLAD: an online database of gene list annotation for *Drosophila*. Journal of Genomics. 2015;3:75-81. doi:10.7150/jgen. 12863

[90] Mathé C, Sagot MF, Schiex T, Rouzé P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 2002;30:4103-4117. doi:10.1093/nar/ gkf543

[91] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674-3676. doi:10.1093/bioinformatics/bti610

[92] Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the

program to assemble spliced alignments. Genome Biology. 2008;9:R7. doi:10.1186/gb-2008-9-1-r7

[93] Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003;13:2178-2189. doi:10.1101/gr.1224503

[94] Wang Y, Li J, Paterson AH. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics. 2013;29:1458-1460. doi:10.1093/bioinformatics/btt150

[95] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494-1512. doi:10.1038/nprot.2013.084

[96] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012;7:562-578. doi:10.1038/nprot.2012.016

[97] Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics. 2006;22:1437-1439. doi:10.1093/bioinformatics/btl116

[98] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20:1297-1303. doi:10.1101/gr.107524.110

[99] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 2012;6:80-92. doi:10.4161/fly.19695

[100] Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: copy number detection and visualization for targeted sequencing using off-target reads. bioRxiv. 2014;1:010876. doi:10.1101/010876

[101] Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987;1:174-181. doi:10.1016/0888-7543(87)90010-3

[102] Manly KF, Cudmore Jr RH, Meer JM. Map Manager QTX, cross-platform software for genetic mapping. Mammalian Genome. 2001;12:930-932. doi:10.1007/s00335-001-1016-3

[103] Van Ooijen JW. MapQTL® 5. Software for the Mapping of Quantitative Trait Loci in Experimental Populations. Wageningen: Kyazma B; 2004.

[104] Basten CJ, Weir BS, Zeng ZB. QTL Cartographer, version 1.17. Raleigh, NC: Department of Statistics, North Carolina State University; 2004.

[105] Leng X, Jia H, Sun X, Shangguan L, Mu Q, Wang B, Fang J. Comparative transcriptome analysis of grapevine in response to copper stress. Scientific Reports. 2015;5:17749. doi: 10.1038/srep17749

[106] Napoli C, Lemieux C, Jorgensen R. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. The Plant Cell. 1990;2:279-289. doi:10.1105/Tpc.2.4.279

[107] Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 1999;286:950-952. doi:10.1126/science.286.5441.950

[108] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature. 2000;403:901-906. doi:10.1038/35002607

[109] Voinnet O. Origin, biogenesis, and activity of plant microRNAs. Cell. 2009;136:669-687. doi:10.1016/j.cell.2009.01.046

[110] Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology. 2006;57:19-53. doi:10.1146/annurev.arplant.57.032905.105218

[111] Combier JP, Frugier F, De Billy F, Boualem A, El-Yahyaoui F, Moreau S, Vernié T, Ott T, Gamas P, Crespi M, Niebel A. MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in Medicago truncatula. Genes & Development. 2006;20:3084-3088. doi:10.1101/gad.402806

[112] Jones-Rhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Molecular Cell. 2004;14:787-799. doi: 10.1016/j.molcel.2004.05.027

[113] Lee H, Yoo SJ, Lee JH, Kim W, Yoo SK, Fitzgerald H, Carrington JC, Ahn JH. Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in *Arabidopsis*. Nucleic Acids Research. 2010;2010:gkp1240. doi:10.1093/nar/gkp1240

[114] Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nature Reviews Molecular Cell Biology. 2005;6(5):376-385. doi:10.1038/nrm1644

[115] Wang M, Wang Q, Wang B. Identification and characterization of microRNAs in Asiatic cotton (*Gossypium arboreum* L.). PLoS One. 2012;7:e33696. doi:10.1371/journal.pone.0033696

[116] Zhang H, Yang JH, Zheng YS, Zhang P, Chen X, Wu J, Xu L, Luo XQ, Ke ZY, Zhou H, Qu LH. Genome-wide analysis of small RNA and novel microRNA discovery in human acute lymphoblastic leukemia based on extensive sequencing approach. PLoS One. 2009;4:e6849. doi:10.1371/journal.pone.0069743

[117] Gao S, Yang L, Zeng HQ, Zhou ZS, Yang ZM, Li H, Sun D, Xie F, Zhang B. A cotton miRNA is involved in regulation of plant response to salt stress. Scientific Reports. 2016;6:19736. doi:10.1038/Srep19736

[118] Khan MA, Shahid AA, Rao AQ, Bajwa KS, Muzaffar A, Rehman Samiullah T, Nasir IA, Husnain T. Molecular and biochemical characterization of cotton epicuticular wax in defense against cotton leaf curl disease (CLCuD). Iranian Journal of Biotechnology. 2016;13(4):3-9. doi:10.15171/ijb.1234

[119] Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, Wang P, Liu N, Lindsey K, Zhang X. Long noncoding RNAs and their proposed functions in fibre development of cotton (*Gossypium* spp.). New Phytologist. 2015;207(4):1181-1197. doi:10.1111/nph.13429

[120] Zou C, Wang Q, Lu C, Yang W, Zhang Y, Cheng H, Feng X, Prosper MA, Song G. Transcriptome analysis reveals long noncoding RNAs involved in fiber development in cotton (*Gossypium arboreum*). Science China Life Sciences. 2016;59:164-171. doi: 10.1007/s11427-016-5000-2

[121] Matzke MA, Kanno T, Matzke AJ. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. Annual Review of Plant Biology. 2015;66:243-267. doi:10.1038/nrg3683

[122] Chen T, Li E. Structure and function of eukaryotic DNA methyltransferases. Current Topics in Developmental Biology. 2004;60:55-89. doi:10.1016/S0070-2153(04)60003-2

[123] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research. 2003;31:3406-3415. doi:10.1093/nar/gkg595

[124] Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research. 2006;34:D140-D144. doi:10.1093/nar/gkj112

[125] Moxon S, Schwach F, Dalmay T, MacLean D, Studholme DJ, Moulton V. A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics. 2008;24:2252-2253. doi: 10.1093/bioinformatics/btn428

[126] An J, Lai J, Sajjanhar A, Lehman ML, Nelson CC. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. BMC Bioinformatics. 2014;15:275. doi:10.1186/1471-2105-15-275

[127] Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, Shomron N. miRNAkey: a software for microRNA deep sequencing analysis. Bioinformatics. 2010;26:2615-2616. doi:10.1093/bioinformatics/btq493

[128] Gunaratne PH, Coarfa C, Soibam B, Tandon A. miRNA data analysis: next-gen sequencing. Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols. 2012;2012:273-288.

[129] Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA. org resource: targets and expression. Nucleic Acids Research. 2008;36:D149-D153. doi:10.1093/nar/gkm995

[130] Dai, X. and P.X. Zhao, psRNATarget: a plant small RNA target analysis server. Nucleic Acids Research, 2011;39:W155-W159. doi:10.1093/nar/gkr319

[131] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 2007;35:D61-D65. doi:10.1093/nar/gkl842

[132] Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Research. 2005;33:D112-D115. doi:10.1093/nar/gki041

[133] Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Research. 2011;39:D146-D151. doi: 10.1093/nar/gkq1138

[134] Xuan H, Zhang L, Liu X, Han G, Li J, Li X, Liu A, Liao M, Zhang S. PLNlncRbase: a resource for experimentally identified lncRNAs in plants. Gene. 2015;573:328-332. doi: 10.1016/j.gene.2015.07.069

[135] Lane AK, Niederhuth CE, Ji L, Schmitz RJ. pENCODE: a plant encyclopedia of DNA elements. Annual Review of Genetics. 2013;48:49-70. doi:10.1146/annurev-genet-120213-092443

[136] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Research. 2007;35:D760-D765. doi:10.1093/nar/gkl887

[137] Geisen S, Barturen G, Alganza ÁM, Hackenberg M, Oliver JL. NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. Nucleic Acids Research. 2013:gkt1202. doi:10.1093/nar/gkt1202

[138] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell. 2010;38:576-589. doi:10.1016/j.molcel.2010.05.004

[139] Hetzl J, Foerster AM, Raidl G, Scheid OM. CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulphite sequencing. The Plant Journal. 2007;51:526-536. doi:10.1111/j.1365-313X.2007.03152.x

[140] Bewick AJ, Hofmeister BT, Lee K, Zhang X, Hall DW, Schmitz RJ. FASTmC: a suite of predictive models for non-reference-based estimations of DNA methylation. G3: Genes, Genomes, Genetics. 2015:g3-115. doi:10.1534/g3.115.025668

[141] Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PloS One. 2013;8:e81148. doi:10.1371/journal.pone.0081148

[142] Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. Nature Structural & Molecular Biology. 2014;21:64-72. doi:10.1038/nsmb.2735

[143] Saze H, Tsugane K, Kanno T, Nishimura T. DNA methylation in plants: relationship to small RNAs and histone modifications, and functions in transposon inactivation. Plant and Cell Physiology. 2012;53:766-784. doi:10.1093/pcp/pcs008

[144] Cervera MT, Ruiz-Garcia L, Martinez-Zapater J. Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. Molecular Genetics and Genomics. 2002;268:543-552. doi:10.1007/s00438-002-0772-4

[145] Salmon A, Clotault J, Jenczewski E, Chable V, Manzanares-Dauleux MJ. *Brassica oleracea* displays a high level of DNA methylation polymorphism. Plant Science. 2008;174:61-70. doi:10.1016/j.plantsci.2007.09.012

[146] Keyte AL, Percifield R, Liu B, Wendel JF. Infraspecific DNA methylation polymorphism in cotton (*Gossypium hirsutum* L.). Journal of Heredity. 2006;97:444-450. doi:10.1093/jhered/esl023

[147] Jin B, Robertson KD. DNA methyltransferases, DNA damage repair, and cancer. In: Adam R. Karpf, editor. Epigenetic Alterations in Oncogenesis. 754. New York: Springer; 2013. p. 3-29. doi:10.1007/978-1-4419-9967-2_1

[148] Kimatu JN, Bao L. Epigenetic polymorphisms could contribute to the genomic conflicts and gene flow barriers resulting to plant hybrid necrosis. African Journal of Biotechnology. 2015;9:8125-8133

[149] He SP, Sun JL, Zhang C, Du XM. Identification of exotic genetic components and DNA methylation pattern analysis of three cotton introgression lines from *Gossypium bickii*. Molecular Biology. 2011;45:204-210. doi:10.1134/S002689331102018x

[150] Song Q, Guan X, Chen ZJ. Dynamic roles for small RNAs and DNA methylation during ovule and fiber development in allotetraploid cotton. PLoS Genetics. 2015;11:e1005724. doi:10.1371/journal.pgen.1005724

[151] Soltani A, Kumar A, Mergoum M, Pirseyedi SM, Hegstad JB, Mazaheri M, Kianian SF. Novel nuclear-cytoplasmic interaction in wheat (Triticum aestivum) induces vigorous plants. Functional & Integrative Genomics. 2016;16:171-182. doi:10.1007/s10142-016-0475-2

[152] Sharma SK, Yamamoto M, Mukai Y. Molecular cytogenetic approaches in exploration of important chromosomal landmarks in plants. In: Rajpal VR, Rama SR, Raina, SN, editors. Molecular Breeding for Sustainable Crop Improvement. 11. New York: Springer; 2016. p. 127-148

[153] Page JT, Gingle AR, Udall JA. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3: Genes| Genomes| Genetics. 2013;3:517-525. doi:10.1534/g3.112.005298