Novogene

NGS enables scientists

to improve cotton species that produce higher-quality better-yield fibers

Genomic research opens up novel prospects on the discovery of common species anew. For example, humanity has been cultivating cotton for a very long time, but only now have we gained the tools and knowledge to look into cotton genomics and use this knowledge for improving agricultural qualities. The world needs an additional gene pool to provide genetic variation in cotton breeding programs. NGS is crucial for current genomic discoveries. With the advancement of sequencing technology, our understanding of the genetic sequences of species is more profound and clearer. This article is presented as an annal, we provide a review of several pieces of research that focus on cotton genomics with NGS services provided by Novogene. And it also represents the decades of innovation of sequencing technology applications.

Sequencing of allotetraploid cotton (Gossypium hirsutum l. acc. Tm-1) provides a resource for fiber improvement

Zhang T, Hu Y, Jiang W *et al*. | Nature Biotechnology | 2015

One of the major interests of geneticists is the crops genes that can be modified to provide better agricultural characteristics. One study focused on the most widely cultivated cotton - allotetraploid upland cotton (*G. hirsutum L.*). The goal of the study was to overcome the challenge of difficulty in discerning between homoeologous sequences of allopolyploids. The authors sequenced and assembled the genome of the allotetraploid *G. hirsutum* L. acc. Texas Marker-1 genome [1]. The sequencing was done on the Illumina platform and the authors generated 612 Gb of high-quality reads. The assembly was done using SOAPdenovo. The final genome was assembled using 174,454 pairs of Sanger-sequenced BAC-end sequences [1]. The assembly was validated and annotated. The subsequent comparative genome analysis and individual gene expression studies showed specific gene families responsible for cotton fiber quality (for example, cellulose synthase genes are responsible for cellulose biosynthesis and the authors were able to identify at least 32 genes in this gene family). This study provided a comprehensive genetic resource for germplasm improvement and selection in cotton lines.

Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield

Ma Z, He S, Wang X *et al*. | Nature Genetics | 2018

Moreover, modern sequencing technologies such as Illumina HiSeq platform are used to perform resequencing accessions of cotton. The goal of Ma et al was to discover the genomic variation of germplasm and alleles responsible for various qualities of cotton fiber. The authors resequenced 419 upland cotton accessions and generated 6.35 Tb of high-quality sequences [2]. The sequencing was done on the Illumina HiSeq platform. The data was used to reconstruct the phylogenetic tree and analyze SNPs. Ma et al identified over 3.6 million SNPs and annotated them [2]. After identifying 13 core fiber qualities, the authors conducted a GWAS analysis. According to the analysis, there are over 7000 unique SNPs and almost 5000 candidate genes associated with higher fiber quality and higher yield [2].

The genomic basis of geographic differentiation and fiber improvement in cultivated cotton

He S, Sun G, Geng X *et al*. | Nature Genetics | 2021

Sequencing technologies, for example, by PacBio, allow researchers to trace the evolution of the most important agricultural traits. *Gossypium hirsutum* is the centerpiece of the research by He et al. It underwent several adaptations, being originally from a tropical climate, but eventually heavily used by people in a variety of climatic zones around the world. The agricultural industry has been working on improving cotton outputs and fiber quality. One of the approaches to studying it is looking at the gene functionality of cotton and modifying the genes according to the needs. However, it was troublesome previously as cotton's genome is very large and allotetraploid [3]. He et al. used the resequenced data of 3,278 cotton genomes to map to the reference genome, assembled by PacBio data [3]. The data were screened for various SNPs and indels. This provided a first comprehensive knowledge of genomic variations for cotton genomes. Specifically, the authors found favorable alleles for fiber length and strength. The usage of this knowledge will help to advance agricultural endeavors to improve cotton quality.

Whole‐genome resequencing of 240 Gossypium barbadense accessions reveals genetic variation and genes associated with fiber strength and lint percentage

Yu J, Hui Y, Chen J *et al*. | Theoretical and Applied Genetics | 2021

Nowadays, thanks to the availability of sequencing resources, researchers can focus not only on one,

most widely used cotton for research, but also on other cotton species of interest. One of the best cotton species in the world is *Gossypium barbadense*. It is the second-largest agricultural cotton species, compared with *Gossypium hirsutum*. Considering the high quality of its fiber, we need to understand its genetic basis. Therefore, Yu et al. analyzed 240 G. barbadense accessions by whole-genome resequencing and identified information on SNPs and indels [4]. The genomic DNA was extracted from the leaves of 240 individual plants.

Then, the sequencing libraries were generated using one of Illumina kits. The paired-end sequencing libraries with ~350 bp insert were sequenced on the Illumina HiSeq PE150 platform [4]. The reads were mapped to a reference genome. After the SNPs and indel calling, Yu et al. identified that the modern *Gossypium barbadense* species accessions established new germplasm, with desirable characteristics. By analyzing 12 cotton traits, the authors identified several candidate genes for lint percentage and fiber strength improvement.

High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement
Ma Z, Zhang Y, Wu L et al. | Nature Genetics | 2021

Finally, the newest sequencing platforms such as Illumina HiSeq or PacBio Sequel are essential for sequencing and reconstruction of genetic knowledge resources as serve as a pivotal technology in modern biology. In the latest study, Ma et al generate genomes of *Gossypium hirsutum* cv. NDM8 and *Gossypium barbadense* acc. Pima90 [5]. The G.hirsutum is the major cotton crop and *G.barbadense* also produces relatively high yield and high-quality lint fibers. The authors also resequenced 1081 *G.hirsutum* accessions. The genome reconstruction allowed looking more closely at these two genomes and comparing genomic variations. After extracting total genomic DNA from two cottons, the sequencing libraries were prepared and sequenced on PacBio

Sequel Platform to obtain long reads, which later were initially assembled. After the construction of chromosome-scale scaffolds with Hi-C interacting unique paired-end data from each genome, the final assemblies included contig and scaffold N50 values of 13.15 Mb and 107.67 Mb for *G.hirsutum* NDM8 and 9.24 Mb and 102.45 Mb for *G.barbadense* Pima90 [5]. After alignment, Ma et al identified genomic structural variations between two genomes. Specifically, the D-subgenome includes more large-scale structural variations, which probably means a stronger selection. Many important traits were discovered, namely 446 potentially crucial agricultural traits for fiber quality and resistance to Verticillium were found.

We were able to observe how the need to produce better cotton is addressed by NGS technology. With the high-resolution and affordable sequencing, the scientists were able to research specific genes that make cotton fibers stronger and of better overall quality. For example, it was proven that the symbiosis of cotton species with certain fungi enables plants to assimilate more phosphorous and grow better. Or, the specific genes, responsible for fiber strength and length were found. We are proud that Novogene is contributing to the future of agriculture by providing its sequencing and bioinformatics services to the researchers, as demonstrated in the above-mentioned studies.

Novogene

Unraveling the Allotetraploid Upland Cotton Genome

The Upland Cotton (*Gossypium hirsutum* L. acc. TM-1) is one of the most commonly cultivated crop plants around the world. Technological advancements and a better understanding of agricultural techniques have allowed the rise of bigger crop yields and a higher rate of desirable traits. The study **'Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement**'[1], sequenced the plant's genome, providing a reference for future projects.
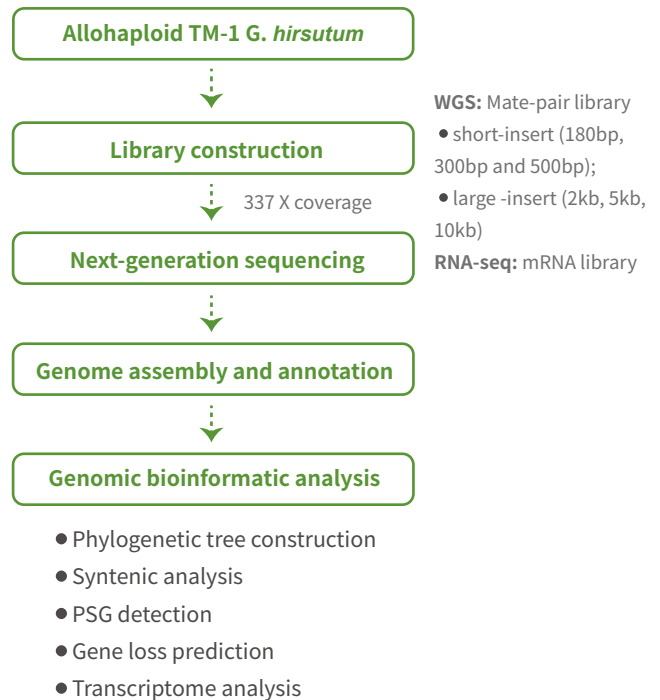
Novogene provides promising teams with state-of-the-art sequencing packages and software. Technological advancements have made the cost of these long processes, turning them into cost-effective scientific endeavors.

Materials and Methods

An allohaploid plant of Texas Marker-1 (TM-1), *G. hirsutum*, (AD)1, the genetic standard, was used. (AD)1 was isolated from the cross of TM-1 with a virescently marked semigametic line of G. barbadense (AD2).

The DNA was prepared using the CTAB extraction method and sheared for short-insert paired-end libraries using the Bioruptor sonication device. For mate-pair, the Hydroshear DNA Shearing Device was utilized. The libraries were subsequently sequenced at 2 x 100 bp on an Illumina HiSeq 2000 platform. 843 Gb of DNA sequenced data was generated.

The SOAPdenovo package was used to assemble the short and long paired ends, into longer sequences, organized into scaffolds (40,407, for a total of 2.4 Gb). Gaps between the scaffolds were filled using the Gapcloser, further scaffolding was then conducted based on links between BAC-ends.

Allohaploid TM-1 G. *hirsutum*

↓

Library construction

↓ 337 X coverage

Next-generation sequencing

↓

Genome assembly and annotation

↓

Genomic bioinformatic analysis

- Phylogenetic tree construction
- Syntenic analysis
- PSG detection
- Gene loss prediction
- Transcriptome analysis

WGS: Mate-pair library
- short-insert (180bp, 300bp and 500bp);
- large -insert (2kb, 5kb, 10kb)

RNA-seq: mRNA library

Conclusions

The domesticated cotton serves as a role model for the understanding of the role of cellulose in the biosynthesis of secondary-wall, as indicated by the presence of 23 family genes in the allotetraploid genes.

The genomic draft will be used to look into the superior genes to improve positive selection, enhance the fiber production, understand better wall-cell biosynthesis, etcetera. Rapid progress in the genomic fields of agricultural sciences were largely because researchers identified genes specific to certain traits.

Novogene

Improving the Upland Cotton's Fiber Quality Through Genome Rese-quencing: A Deeper Look into Its Genes

The Upland Cotton is a highly profitable and culti-vated crop plant through-out the world, its industry is quite a large one. However, despite the versatility it provides, its fiber quality isn't top-notch, at least compared to other crop plants. The current study **'Resequencing core collection of upland cotton (Gossypium hirsu-tum) identifies genomic variation and loci influencing quality and yield of fiber'**.[2] shed more light on the loci of protein-coding genes, improving the odds of effective artificial selection to bring out better fiber-re-lated traits.

Increasing crop yield and fiber quality is the most imperative task, if cotton is to supply the increasing global demand for it. Novo-gene has been highly inter-ested in providing research-ers with state-of-the-art sequencing equipment and professional support.

Materials and Methods

G. hirsutum accessions were taken from the China National Gene Bank, and were selected based on the genetic diversity and, thus, the different phenotypic traits they exhibited. These specimens come from different countries all around the world, such as the United States, China, India and Brazil, to the now extinct Soviet Union.

The fiber samples were taken to be examined. Subsequently, the DNA of the aforementioned samples were extracted and isolated, and the genome sequenced. Whole-genome libraries were constructed and sequenced using the **Illumina HiSeq platform**, generating 6.45 Tb of raw sequences.

419 G. *hirsutum* accessions

↓

Library construction

350bp whole-genome and mRNA libraries

↓

Genome sequencing and annotation

↓

Genomic bioinformatic analysis

- Population genetics analysis
- Linkage-disequilibrium analysis
- GWAS analysis
- Transcriptome analysis

Conclusions

Researchers identified over 3 million population SNPs. 4,820 candidate genes were associated with fiber quality and yield traits. The resequencing done in the study provides future researchers with potential candidates and genetic markers to work on.

The genomic basis of geographic differentiation and fiber improvement in cultivated cotton

Modern cultivated cotton (*G. hirsutum*) changed due to artificial breeding combined with long-term natural selection. The study of '**The genomic basis of geographic differentiation and fiber improvement in cultivated cotton**'.[3] is designed to understand the genetic basis of cotton adaptation. The goal was to analyze how cotton's fiber quality improved.

Novogene provides invaluable resources and expertise for this type of research. Novogene consults on experimental design and provides sequencing and bioinformatics services, such as, the *de novo* genome assembly.

🌱 Experimental Design

For sampling and genotyping, the DNA was extracted from the young leaves of selected tetraploid cotton accessions. The sequencing was performed on the **Illumina HiSeq 4000** platform by Novogene, resulting in 61.64 TB of data. The genome was assembled by combining all the unassembled contigs into a pseudochromosome. Post-alignment to reference genome, the variants were identified used GATK UnifiedGenotyper and ANNOVAR for annotating these variants. The goal of the *de novo* assembly was to verify the effects of chromosomal inversions. The sequencing was done on the **PacBio Sequel platform** using single-molecule real-time technology. At the same time, Hi-C experiments generated ~222.6 Gb Illumina reads. The Hi-C reads were integrated into the 3D de novo assembly pipeline. The contig N50 was ~42.98Mb and the assembled genome was ~2.3 Gb.

The genome-wide association studies (GWAS) was performed to identify the specific fiber-quality alleles. To further identify the candidate genes, the **mRNA sequencing** was done using **Illumina HiSeq 4000 platform**. The study was largely focused on trying to identify the favorable alleles for fiber length and strength.

🌱 Conclusions

The study demonstrated the variety of genomic variations of cotton accessions. The history of the origin of these alleles was possible to track due to the abundance of the genomic data, which was generated thanks to the capabilities of **Illumina and PacBio sequencing**, as well as the further thorough **bioinformatic analysis**. The study also highlighted

the possible origin of alleles (the germplasms) that are identified, responsible for fiber qualities. These data show the possibility of utilizing the pan-genome analysis for identifying molecular markers of favorable agricultural features and may be used for advancement in molecular breeding and enhancing the agricultural qualities of crops.

Sampling and genotyping

- Extracted genomic DNA
- Constructed 150-bp libraries with ~350-bp insert
- Sequenced by Illumina HiSeq4000.
  Connected unassembled contigs into pseudochromosome.
  Aligned genome to reference using BWA.
  Identified variations (raw SNP set) and annotated by ANNOVAR.

De novo assembly of ICR_XLZ 7

- DNA sheared and concentrated
- Constructed 60-kb length libraries
- Sequenced by PacBio.
- Further contigs assembled with integrating Hi-C reads using de novo (3D) assembly pipeline.

Gossypium barbadense and its fiber strength genes

In the following article **'Whole-genome resequencing of 240 Gossypium barbadense accessions reveals genetic variation and genes associated with fiber strength and lint percentage'**.[4] researchers, from Zhejiang University and the Hainan Institute, conducted a genome-wide association study of the *Gossypium barbadense*, the second largest cotton crop plant in the world. Nowadays, looking further into possible candidate genes from secondary gene pools is indispensable, to avoid gene constriction and further weakening.
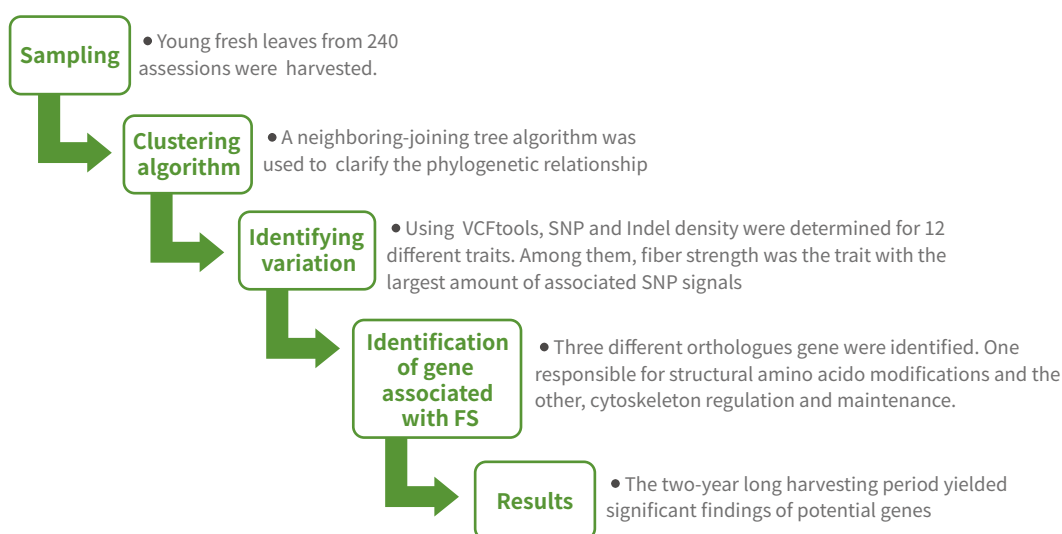
Gossypium hirsutum, as the largest cotton plant cultivated throughout the world, its limited genetic diversity represents a problem. Researchers used resequencing techniques to identify possible gene candidates of G. barbadense, the second largest cotton plant.

🌸 Materials and Methods

Fresh young leaves taken from 240 accessions of G. barbadense were used. The genomic DNA was extracted and sequencing libraries were prepared using Truseq Nano DNA HT Sample preparation Kit (Illumina USA). Now, moving onto the variation calling and annotation, single-nucleotide polymorphism (SNP) and insert-deletion (indels) were detected using the HaplotypeCaller module in GATK.

Indels shorter than or equal to 50 bp were considered. Secondly, the SNP and indels were used to evaluate and analyze the characteristics of *G. barbadense*. Using a TreeBestv1.9.2 software, a neighbor-joining tree was constructed.

Sampling • Young fresh leaves from 240 assessions were harvested.

Clustering algorithm • A neighboring-joining tree algorithm was used to clarify the phylogenetic relationship

Identifying variation • Using VCFtools, SNP and Indel density were determined for 12 different traits. Among them, fiber strength was the trait with the largest amount of associated SNP signals

Identification of gene associated with FS • Three different orthologues gene were identified. One responsible for structural amino acido modifications and the other, cytoskeleton regulation and maintenance.

Results • The two-year long harvesting period yielded significant findings of potential genes

🌸 Conclusions

The genetic polymorphism of *G. barbadense* was analyzed, proving to be fruitful research, since potential genes were identified. All these orthologues seem to be the key to improve fiber strength, by modifying structural amino acid order and cytoskeleton regulation. Identifying key genes involved in morphological advantageous traits is necessary if the cotton industry wishes to surpass the current breeding threshold. Novogene acts as a link between the current technology and brilliant researchers, with many inspirations in mind.

Genome assembly and Rese-quencing of modern cotton varieties provide evidence for crop improvement

Photo taken by Prof. Ma's Team, with *Gossypium hirsutum* cv. NDM8

The cotton industry has many challenges, for example, to meet the increasing requirement of natural fibers and to solve growth issues. These challenges could be impacted by climate change, decreasing biodiversity, and other factors. However, mechanism referring genomic diverstification of cotton in modern agricultural process still remains unclear. Therefore, with the aid of powerful platforms and bioinformatic tools , researchers are allowed to sequence the whole genome of cotton and do assembly and annotation. In this study, the authors generated two high-quality genomes of *Gossypium hirsutum* cv. NDM8 and *Gossypium barbadense* acc. Pima90. The two genomes were further compared with 1081 *G. hirsutum* accessions to identify structural variations. Structural variations which are related to critical agricultural traits including fiber quality, Verticillium wilt resistance, etc., and their locations in subgenome were identified.

🌾 Introduction

Gossypium hirsutum is the major cotton crop in the world, accounting for more than 90% of yield. *Gossypium barbadense*, which provides high-quality lint finders and reached ~10% yield, is also an important cultivator, indicating both cottons are widely used. Thereforeit is meaningful to investigate the genetic information of *G.hirsutum* and *G.barbadense*. The research focused on constructing high-quality genomes and identifying structural variations between two widely used cottons, and re-sequencing 1081 worldwide *G.hirsutum* accessions.

🌾 IMaterials and methods

The isolated genomic DNAs from *Gossypium hirsutum* cv. NDM8, *Gossypium barbadense* acc. Pima90 and 1081 *G.hirsutum* were used to prepare Pacbio and Illumina libraries, respectively. The **PacBio libraries** were performed based on fragmentation, end-repair, adapter ligation and exonuclease digestion. To construct chromosome-scale scaffolds, **10x genomics and Hi-C libraries** were further prepared. 10x genomics libraries were generated with the compatible GemCode Instrument from 10x Genomics. The Hi-C libraries were constructed using DNAs from the leaves. These cross-linked DNAs were digested with HindIII, biotinylated, and proximity-ligated. Then the treated DNA was fragmented into 300-500 bp. Finally both 10x genomics libraries and Hi-C libraries were sequenced on Illumina platform.

🌾 Results

By using **Illumina and PacBio sequencing**, high quality *Gossypium hirsutum* cv. NDM8 and *Gossypium barbadense* acc. Pima90 genomes were generated with the size of 2.29 Gb and 2.21 Gb, respectively. The final assemblies with very few gaps included 353 scaffolds for *G.hirsutum* and 309 for *G.barbadense*. These assemblies contained a contig of 13.15 Mb and scaffold N50 of 107.67 Mb for NDM8 and 9.24 Mb and 102.45 Mb for Pima90.

To deeper investigate the two cottons, alignment between Pima90 and NDM8 were performed and result showed high diversification. The integration of transcriptome sequencing illustrated that structural variations could impact on gene expression significantly.. The number of variations were counted between A-subgenome and D-subgenome, showing that more inversions were detected in A-subgenome, but more inversions and deletions were found in D-subgenome. This suggested that stronger selection was occurred in D-subgenome during species formation. The genome-wide association study (GWAS) elucidated the key agronomical traits. 446 variations were found to be directly associated with seven major features including fiber strength, yield, and Verticillium wilt resistance.

🌾 Conclusion

This study provided an invaluable scientific resource for improving the agricultural features of cotton. Whole genomes of two major cottons: *Gossypium hirsutum* cv. NDM8 and *Gossypium barbadense* acc. Pima90 as well as 1081 resequenced *G.hirsutum* accessions were clarified. With PacBio Sequel and Illumina platforms, the authors identified major structural variations in genes related with cotton fiber quality and resistance.

❀ DNA sample preparation

· NDM8 and Pima90

· PacBio: CTAB method

· Hi-C library: leaves were fixed with formaldehyde and lysed

· 10x Genomics: GemCode Instrument

❀ Library preparation

· PacBio: fragmented, end-repaired, adapter ligated, and exonuclease digested

· Illumina: TruSeq Nano DNA HT Sample Kit

· 10x Genomics: GEM reactions, 16-bp barcodes introduction, droplets fractured, and sheared DNA into 500-bp fragments

· Hi-C: cross-linked DNA digestion, biotinylation, sheared DNA into 300-500 bp fragments SequencingPacBio sequel platformWhole genome sequencing libraries, Hi-C libraries and 10x Genomics libraries were sequenced on Illumina platforms

❀ Data QC

· Low quality reads that were from base-calling duplicates and adapter contamination were filtered.
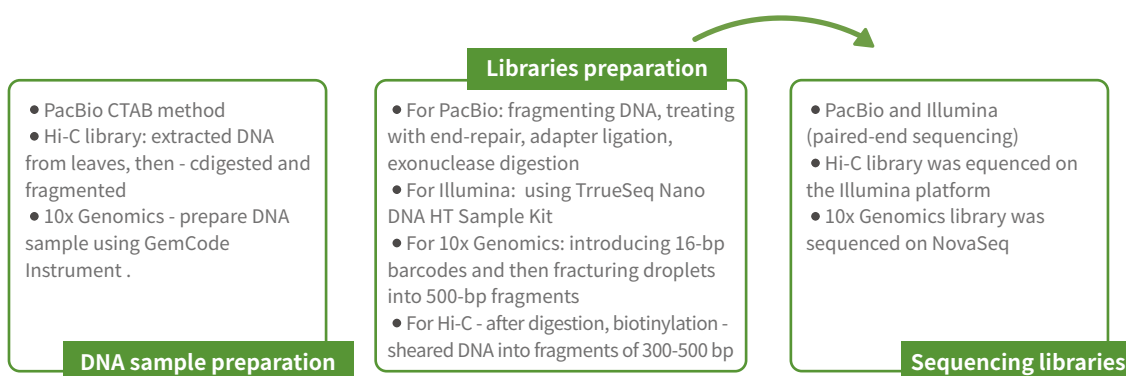
❀ Genome Assembly

· PacBio reads: FALCON assembler

· Illumina reads: Pilon pipeline

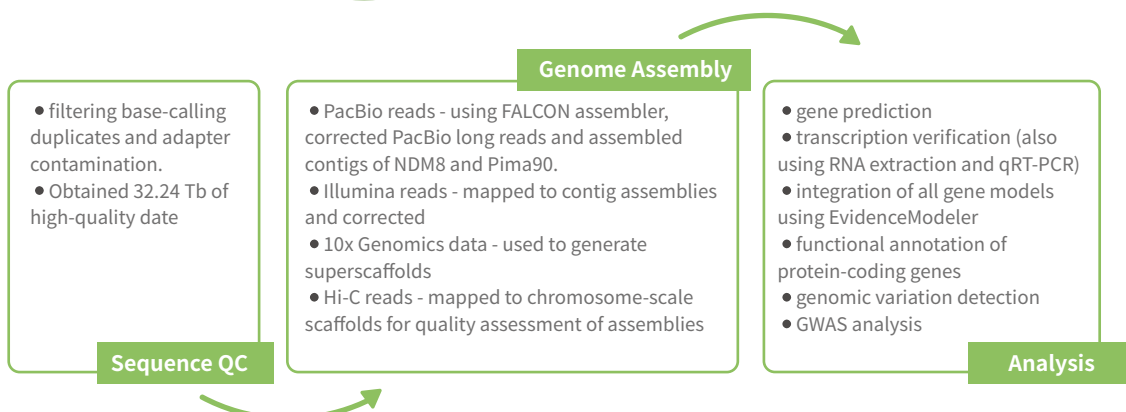· 10x Genomics data: Fragscaff software

· Hi-C reads: Hi-C-Pro software

❀ Analysis

· Genome repeat annotation

· Structural annotation of genes

· Functional annotation of protein-coding genes

· genomic variation detection
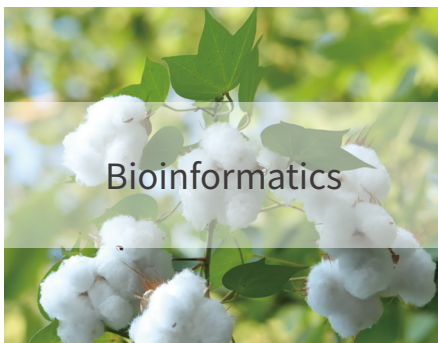
· GWAS analysis

01
Sample preparation and sequencing

- PacBio CTAB method
- Hi-C library: extracted DNA from leaves, then - cdigested and fragmented
- 10x Genomics - prepare DNA sample using GemCode Instrument .

DNA sample preparation

Libraries preparation

- For PacBio: fragmenting DNA, treating with end-repair, adapter ligation, exonuclease digestion
- For Illumina: using TrrueSeq Nano DNA HT Sample Kit
- For 10x Genomics: introducing 16-bp barcodes and then fracturing droplets into 500-bp fragments
- For Hi-C - after digestion, biotinylation - sheared DNA into fragments of 300-500 bp

- PacBio and Illumina (paired-end sequencing)
- Hi-C library was equenced on the Illumina platform
- 10x Genomics library was sequenced on NovaSeq

Sequencing libraries

02
Bioinformatic analysis and verification

- filtering base-calling duplicates and adapter contamination.
- Obtained 32.24 Tb of high-quality date

Sequence QC

Genome Assembly

- PacBio reads - using FALCON assembler, corrected PacBio long reads and assembled contigs of NDM8 and Pima90.
- Illumina reads - mapped to contig assemblies and corrected
- 10x Genomics data - used to generate superscaffolds
- Hi-C reads - mapped to chromosome-scale scaffolds for quality assessment of assemblies

- gene prediction
- transcription verification (also using RNA extraction and qRT-PCR)
- integration of all gene models using EvidenceModeler
- functional annotation of protein-coding genes
- genomic variation detection
- GWAS analysis

Analysis

References

Zhang T, Hu Y, Jiang W *et al*. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol 33, 531–537 (2015). https://doi.org/10.1038/nbt.3207

Ma Z, He S, Wang X *et al*. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet 50, 803–813 (2018). https://doi.org/10.1038/s41588-018-0119-7

He S, Sun G, Geng X *et al*. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. Nat Genet 53, 916–924 (2021). https://doi.org/10.1038/s41588-021-00844-9

Yu J, Hui Y, Chen J, Yu H, Gao X, Zhang Z, Li Q, Zhu S, Zhao T. Whole-genome resequencing of 240 Gossypium barbadense accessions reveals genetic variation and genes associated with fiber strength and lint percentage. Theor Appl Genet. 2021 Jul 8. https://doi.org/10.1007/s00122-021-03889-w

Ma Z, Zhang Y, Wu L *et al*. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. Nat Genet (2021). https://doi.org/10.1038/s41588-021-00910-2

Genome Sequencing

Trancriptome Sequencing

Epigenome Sequencing

Metagenome Sequencing

Bioinformatics

Novogene Co., Ltd.

en.novogene.com

Novogene Global | Novogene AMEA | Novogene America | Novogene Europe

inquiry_us@novogene.com | marketing@novogene-europe.com | marketing_amea@novogeneait.sg

Follow us on LinkedIn