Next-Generation Sequencing Technologies: A Game Changer in Cotton Genomics

Jiayi Wu, Tianze Zhang ✉

Modern Agriculture Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, Zhejiang, China

✉ Corresponding author: tianze.zhang@cuixi.org

Abstract Next-generation sequencing (NGS) technology has revolutionized the field of cotton genomics, providing unprecedented insights into the genetic structure, functional genomics, and breeding strategies for this economically important crop. This study systematically explores the transformative impact of NGS on cotton genomics and its key advancements. NGS has enabled the construction of high-quality reference genomes and de novo assemblies, facilitating detailed studies on genetic diversity, population genomics, and phylogenetic relationships. The integration of NGS with genome editing technologies such as CRISPR/Cas9 has paved the way for precise genetic modifications, accelerating the development of superior cotton varieties. Despite technical challenges, data management complexities, and cost barriers, the continuous evolution of NGS technology promises to overcome these limitations. The future of cotton genomics lies in the integration of NGS with other omics approaches, promoting sustainable cotton production through advanced breeding programs and comprehensive genetic analyses.

Keywords Next-generation sequencing; Cotton genomics; Genetic variations; Precision breeding; Genomic analysis

1 Introduction

Cotton (*Gossypium* spp.) is a globally vital crop, fundamental to the textile industry and a significant agricultural commodity. Its economic importance is underscored by its use in clothing, household items, and industrial products. Beyond its economic value, cotton also plays a crucial role in food production and sustainable agriculture, providing seed oil and animal feed. Cotton is a cornerstone of the global textile industry, serving as the primary source of natural fiber. The economic and agricultural importance of cotton cannot be overstated, as it supports the livelihoods of millions of farmers and contributes significantly to the economies of many countries. The complexity of the cotton genome, characterized by its large size and polyploid nature, has historically posed significant challenges to genomic research. The study of cotton genomics is essential for advancing our understanding of this complex crop, enabling the development of improved varieties with enhanced yield, fiber quality, and resistance to biotic and abiotic stresses. This knowledge is critical for addressing the challenges of climate change, pest and disease management, and the increasing demand for sustainable agricultural practices. However, recent advancements in high-throughput sequencing and bioinformatics have begun to unravel the intricacies of the cotton genome, providing valuable insights into fiber biogenesis, genetic diversity, and the evolutionary history of *Gossypium* species. These developments have paved the way for genomics-enabled breeding strategies aimed at improving fiber yield, quality, and environmental resilience, thereby addressing the pressing needs of modern agriculture (Pavlovic et al., 2020).

Next-Generation Sequencing (NGS) technologies have revolutionized the field of genomics by enabling the rapid and cost-effective sequencing of entire genomes and transcriptomes. Unlike traditional Sanger sequencing, NGS platforms can generate massive amounts of data in a relatively short period, democratizing access to genomic information and facilitating large-scale genetic studies (Kumar et al., 2019; Yang et al., 2021). The primary advantages of NGS include its high throughput, accuracy, and the ability to detect a wide range of genetic variations, from single nucleotide polymorphisms (SNPs) to large structural variants (Levy and Boone, 2019).

Various NGS methodologies, such as sequencing by synthesis, ion semiconductor sequencing, and nanopore sequencing, offer unique strengths and are continually evolving to overcome existing limitations (Satam et al., 2023). These technologies have been instrumental in advancing our understanding of complex genomes, including those of polyploid crops like cotton, by providing detailed insights into genetic diversity, gene function, and evolutionary processes (Rexach et al., 2019). In cotton genomics, NGS has opened new avenues for exploring genetic diversity, identifying key genes and regulatory networks, and accelerating breeding programs through marker-assisted selection and genomic selection.

This study summarizes the current advancements in cotton genome research and highlights the major challenges and opportunities posed by the complexity of the cotton genome. It discusses the progress in cotton genomics driven by NGS, including the identification of genetic markers, understanding fiber biogenesis, and gaining insights into the evolutionary history of *Gossypium* species. The study evaluates the potential of NGS-enabled breeding strategies to enhance cotton yield, quality, and environmental adaptability, thereby contributing to sustainable agricultural practices. This study aims to identify the key roles of NGS technologies in advancing cotton genomics, providing insights for future research and agricultural innovation.

2 Overview of Next-Generation Sequencing Technologies

Next-Generation Sequencing (NGS) technologies have revolutionized the field of genomics by providing unprecedented depth, accuracy, and throughput in sequencing. These technologies have evolved significantly since the days of Sanger sequencing, offering a wide array of applications and efficiencies that were previously unattainable.

2.1 History and development of NGS

The development of NGS technologies marked a significant leap from the traditional Sanger sequencing method, which was the gold standard for many years. The advent of NGS in the early 2000s introduced high-throughput sequencing capabilities, drastically reducing the time and cost associated with sequencing large genomes. NGS technologies enable the simultaneous sequencing of millions of DNA fragments, providing comprehensive insights into genome structure, genetic variations, and gene expression profiles. The continuous evolution of these technologies has led to significant breakthroughs in various fields, including personalized medicine, evolutionary biology, and agricultural genomics. The first generation of NGS technologies, often referred to as second-generation sequencing, includes platforms like Illumina and Ion Torrent, which are characterized by their short-read lengths and high accuracy (Kumar er al., 2019; Hu et al., 2021). The early 2010s saw the emergence of third-generation sequencing (TGS) technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), which offer long-read sequencing capabilities, further enhancing the ability to resolve complex genomic regions and structural variations (Midha et al., 2019; Athanasopoulou et al., 2021).

2.2 Types of NGS technologies

NGS technologies are diverse and can be broadly categorized based on their sequencing methodologies and read lengths. The most prominent NGS platforms include Illumina sequencing, Pacific Biosciences (PacBio) sequencing, and Oxford Nanopore sequencing (Figure 1), each offering unique advantages and limitations.

2.2.1 Illumina sequencing

Illumina sequencing, also known as sequencing by synthesis (SBS), a short-read technology, is one of the most widely used NGS platforms. It employs a sequencing-by-synthesis approach, where fluorescently labeled nucleotides are incorporated into a growing DNA strand and detected in real-time. This method provides high accuracy and throughput, making it suitable for a wide range of applications, including whole-genome sequencing, RNA sequencing, and targeted sequencing (Hu et al., 2021; Kumar et al., 2019). However, the short read lengths (typically 150-300 base pairs) can pose challenges in sequencing repetitive regions and complex genomes (Baptista et al., 2018).

2.2.2 PacBio sequencing

Pacific Biosciences (PacBio) sequencing, or Single Molecule Real-Time (SMRT) sequencing, a third-generation technology, offers long-read sequencing capabilities through its Single Molecule Real-Time (SMRT) sequencing

technology. PacBio reads can span several kilobases, providing a more comprehensive view of the genome, including complex regions and structural variants. This technology offers significant advantages in resolving complex genomic regions, including large structural variations and repetitive sequences. This technology is particularly advantageous for de novo genome assembly and the identification of full-length transcripts without the need for assembly (Cui et al., 2020; Athanasopoulou et al., 2021). Despite its benefits, PacBio sequencing can be more expensive and has a higher error rate compared to short-read technologies (Lang et al., 2020).
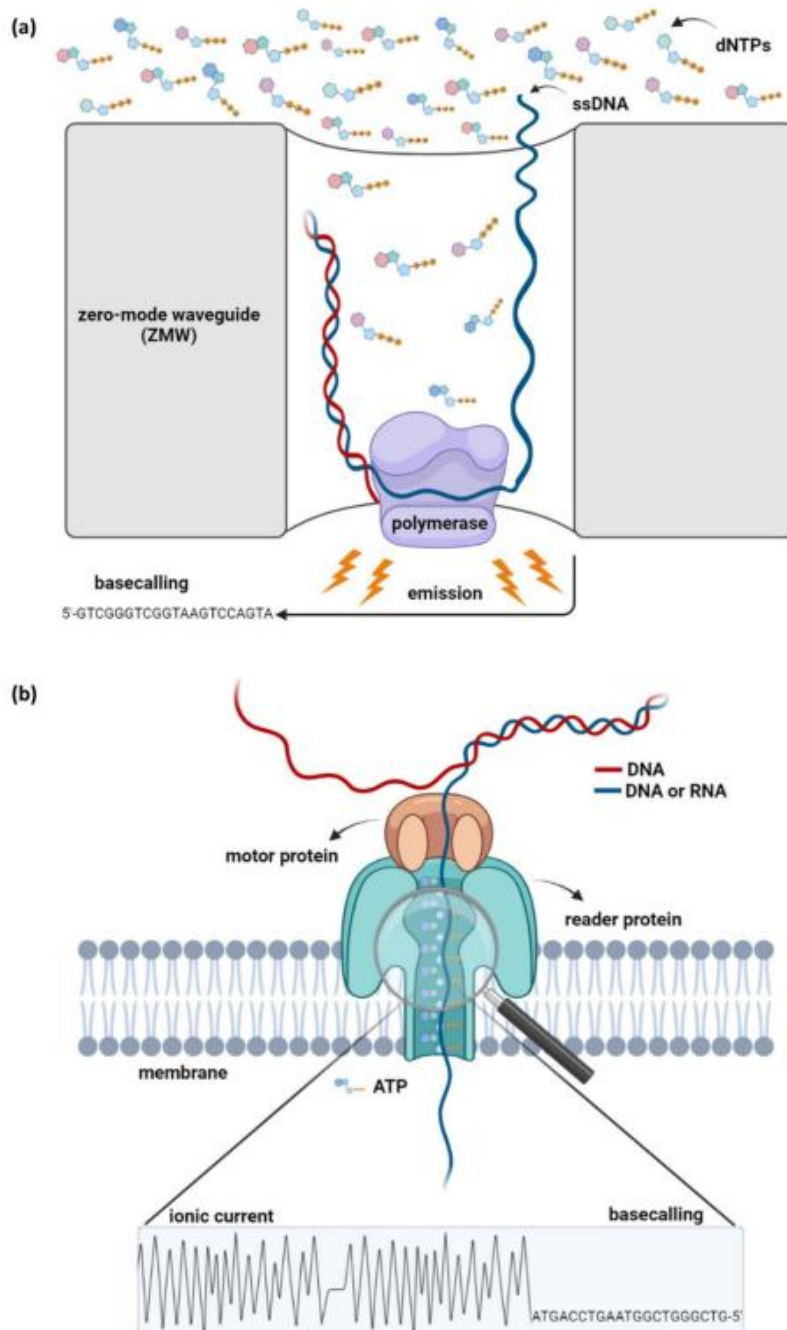

Figure 1 Scientific background of both third-generation sequencing technologies (Adopted from Athanasopoulou et al., 2021)
Image caption: (a) PacBio sequencing. The method is based on DNA sequencing by synthesis. The mobile single-stranded DNA is attached to the stable polymerase, which catalyzes the incorporation of dNTPs in a newly synthesized complementary DNA strand. The reaction occurs in specially designed zero-mode waveguides (ZMWs) that enable the observation of the emitted light. Afterwards, the signal is translated into nucleotide sequence, a procedure known as basecalling; (b) Nanopore sequencing. The method relies on the guidance of a single-stranded DNA or RNA molecule to a nanopore, a reader protein that detects alterations in the electrical current, which occurs during the passing through of the DNA/RNA (Adopted from Athanasopoulou et al., 2021)

2.2.3 Oxford Nanopore sequencing

Oxford Nanopore Technologies (ONT) is another third-generation sequencing platform that provides ultra-long reads, with some reads exceeding 1 million base pairs. ONT's nanopore sequencing technology involves passing a DNA or RNA molecule through a nanopore and measuring changes in electrical current to determine the sequence. This method allows for real-time sequencing and direct RNA sequencing, offering unique advantages in studying epitranscriptomics and RNA modifications (Midha et al., 2019; Athanasopoulou et al., 2021). ONT is also known for its portability, with devices like the MinION enabling sequencing in various settings outside the traditional laboratory environment (Midha et al., 2019; Lang et al., 2020). Despite its versatility, the technology still faces challenges related to accuracy and error rates, which are areas of active research and improvement.

2.3 Comparison of NGS technologies

When comparing NGS technologies, several factors need to be considered, including read length, accuracy, throughput, and cost. Short-read technologies like Illumina offer high accuracy and throughput at a lower cost, making them suitable for many routine sequencing applications. However, their short read lengths can limit their ability to resolve complex genomic regions (Baptista et al., 2018; Hu et al., 2021).

In contrast, long-read technologies like PacBio and ONT provide much longer reads, which are beneficial for de novo genome assembly and the detection of structural variants. PacBio offers high accuracy with its HiFi reads, while ONT provides ultra-long reads and real-time sequencing capabilities (Cui et al., 2020; Lang et al., 2020; Athanasopoulou et al., 2021). Each technology has its strengths and limitations, and the choice of platform often depends on the specific requirements of the research project.

NGS technologies have transformed genomic research by providing powerful tools for sequencing and analyzing complex genomes. The continuous advancements in both short-read and long-read sequencing platforms promise to further enhance our understanding of genomics and its applications in various fields, including cotton genomics.

3 Applications of NGS in Cotton Genomics

3.1 Genome sequencing and assembly

Next-generation sequencing (NGS) technologies have significantly advanced the construction of reference genomes in cotton genomics. The ability to sequence millions of reads in parallel has enabled the rapid and cost-effective generation of high-quality reference genomes. For instance, the development of Transposase Enzyme Linked Long-read Sequencing (TELL-Seq) allows for the generation of long-read-like information using short NGS reads, facilitating the construction of reference genomes with high accuracy and efficiency (Chen, 2019). Additionally, hybrid assembly strategies that combine ultra-long reads from third-generation sequencing (3GS) technologies with short reads from NGS have been shown to produce nearly gapless and high-quality reference genomes, as demonstrated in human genome studies (Ma et al., 2019). These advancements have enabled the construction of high-quality reference genomes and the assembly of novel genomes, providing a comprehensive understanding of cotton's genetic architecture.

3.1.1 Reference genome construction

The construction of reference genomes is critical for understanding the genetic foundation of cotton. NGS technologies, particularly Illumina sequencing, have been instrumental in generating high-quality reference genomes. These reference genomes serve as a crucial resource for identifying genetic variations, mapping genes, and studying the evolutionary history of cotton species. The availability of a reference genome facilitates various genomic studies and supports the development of improved cotton varieties.

3.1.2 *de novo* genome assembly

de novo genome assembly involves sequencing and assembling a genome from scratch without the use of a reference. This approach is essential for studying cotton species with no existing reference genome. De novo genome assembly in cotton has been greatly enhanced by NGS technologies. The integration of short and long reads through hybrid assembly approaches has proven effective in generating complete and accurate genome assemblies. For example, the combination of NGS short reads with long reads from platforms like Nanopore has

been successfully applied to assemble complex genomes, overcoming the limitations of each technology when used alone. This approach has been particularly useful in assembling highly repetitive genomes, which are common in many plant species, including cotton (Baptista et al., 2018). The advancements in de novo genome assembly have provided valuable insights into the genetic diversity and evolutionary history of cotton species.

3.2 Transcriptomics and gene expression analysis

3.2.1 RNA-Seq applications

RNA sequencing (RNA-Seq) is a powerful application of NGS that has revolutionized transcriptomics and gene expression analysis in cotton. RNA-Seq allows for the comprehensive profiling of gene expression, identification of novel transcripts, and detection of alternative splicing events. The high throughput and sensitivity of RNA-Seq make it an invaluable tool for studying the transcriptome dynamics in cotton under various conditions (Ferros et al., 2022). This technology has been widely used to investigate the molecular mechanisms underlying important traits such as fiber development, stress responses, and disease resistance in cotton (Bansal et al., 2018).

Zheng et al. (2021) conducted an in-depth study on the identification and function of long-chain non-coding RNAs (lncRNAs) in upland cotton (*Gossypium arboreum*). The researchers utilized various high-throughput sequencing technologies, including full-length isoform sequencing, strand-specific RNA sequencing (ssRNA-seq), Cap Analysis Gene Expression sequencing (CAGE-seq), and PolyA sequencing, to systematically analyze lncRNA expression in 21 tissue samples (Figure 2).
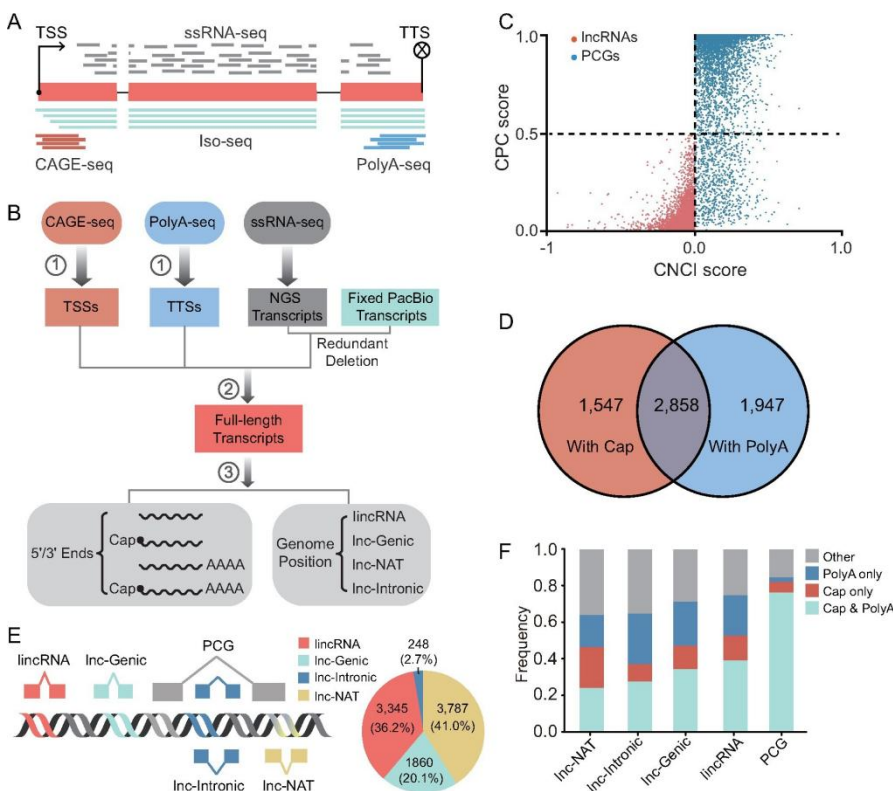

Figure 2 Full-length annotation of lncRNAs with multi-strategy RNA-seq data in cotton (*G. arboreum*) (Adopted from Zheng et al., 2021)

Image caption: A: The four RNA-seq technologies applied in this study; B: The overview of plant full-length lncRNA (PULL) pipeline; C: Evaluation of CPC and CNCI for lncRNAs and PCGs. Low score indicates weak capability of protein encoding; D: Numbers of lncRNAs with 5′- and/or 3′-end signals; E: Schematic illustration of four types of lncRNAs (left) and their proportions (right); F: The proportion of transcripts with a 5′ cap and 3′ polyA tail in lncRNAs and PCGs (Adopted from Zheng et al., 2021)

Figure 2 illustrates the comprehensive annotation of long-chain non-coding RNAs (lncRNAs) in cotton, integrating multi-strategy RNA sequencing data. By utilizing four sequencing technologies-Iso-seq, ssRNA-seq, CAGE-seq, and PolyA-seq-researchers were able to accurately analyze and identify the precise gene structures of lncRNAs, including transcription start sites (TSS) and transcription termination sites (TTS). The figure also shows

the proportions of different lncRNA categories, such as long intergenic non-coding RNAs (lincRNAs) and long intronic non-coding RNAs (lnc-Intronic), with different colors representing each lncRNA type. Additionally, the figure includes information on CPC (Coding Potential Calculator) and CNCI (Coding-Non-Coding Index) scores, which assess the coding potential of lncRNAs, where lower scores indicate weaker protein-coding potential.

By integrating these data, they developed an analysis pipeline named PULL, successfully identifying 9 240 lncRNAs. The study found that many lncRNAs regulate adjacent protein-coding genes (PCGs) in cis. For instance, some lncRNAs modulate gene expression by influencing the selection of transcription start sites (TSS) of PCGs. Additionally, the research explored the structural characteristics of lncRNAs, such as their exon number, transcript length, and GC content, and how these features impact their function and expression. This study not only provides new insights into the biological functions and regulatory mechanisms of cotton lncRNAs but also establishes a high-resolution lncRNA map, laying a foundation for future functional research.

3.2.2 Differential gene expression

Differential gene expression analysis using RNA-Seq has provided critical insights into the regulatory networks controlling key biological processes in cotton. By comparing gene expression profiles between different tissues, developmental stages, or treatment conditions, researchers can identify genes that are differentially expressed and potentially involved in specific pathways. This information is essential for understanding the genetic basis of complex traits and for developing strategies to improve cotton breeding (Begum and Banerjee, 2021). The ability to perform differential gene expression analysis with high precision and accuracy has made RNA-Seq a cornerstone of functional genomics studies in cotton.

3.3 Epigenomics and methylation studies

Epigenomic studies focus on modifications to the genome that do not change the DNA sequence but can affect gene expression. NGS technologies have made significant contributions to epigenomics, particularly in the study of DNA methylation, which plays a crucial role in gene regulation and plant development.

3.3.1 Whole-Genome bisulfite sequencing

Whole-genome bisulfite sequencing (WGBS) is an NGS-based technique used to study DNA methylation patterns across the entire genome. In cotton, WGBS has been employed to investigate the epigenetic modifications that regulate gene expression and contribute to phenotypic variation. This technique provides a comprehensive view of the methylome, allowing researchers to identify differentially methylated regions and their potential roles in gene regulation (Ferros et al., 2022). The insights gained from WGBS studies are crucial for understanding the epigenetic mechanisms underlying important agronomic traits in cotton.

Lu et al. (2022) successfully assembled a new genome for CRI-12, a major cotton variety in China, using Pacific Biosciences and Hi-C sequencing technologies (Figure 3). The results showed that the CRI-12 genome is of high quality, with a total length of approximately 2.31 Gb and a contig N50 of 19.65 Mb, outperforming previously reported cotton genomes. Comparative analysis with other reported genomes revealed that CRI-12 has 7 966 structural variations and 7 378 presence/absence variations, contributing to its ability to adapt to different environments.

Lu et al. (2022) investigated the phenotype of CRI-12 cotton and the homology relationships between different cotton genomes. Figure 3A shows the actual plant morphology of CRI-12. Figure 3B presents the chromosome information of CRI-12 using Hi-C mapping, with chromosomes numbered from 1 to 26 from left to right. Figure 3C, through synteny analysis, reveals the high collinearity between the A subgenome and D subgenome of CRI-12 and other cotton species. Figure 3D, using 4DTV analysis, illustrates the evolutionary distance of CRI-12 from other species, indicating genome duplication and divergence at the genomic level.

The researchers also used whole-genome bisulfite sequencing (WGBS) to analyze the DNA methylation patterns of CRI-12, exploring its molecular mechanisms for environmental adaptation. The WGBS results indicated that methylation variations might enhance CRI-12's adaptability to various biotic and abiotic stress conditions by regulating gene expression. Notably, the study found significant changes in methylation levels of several

important agronomic trait genes, which are directly related to cotton's drought, salt, and disease resistance. For example, some methylation sites on chromosome D13 were closely associated with drought resistance, suggesting that specific methylation patterns might respond to environmental stress by affecting gene expression or suppression.
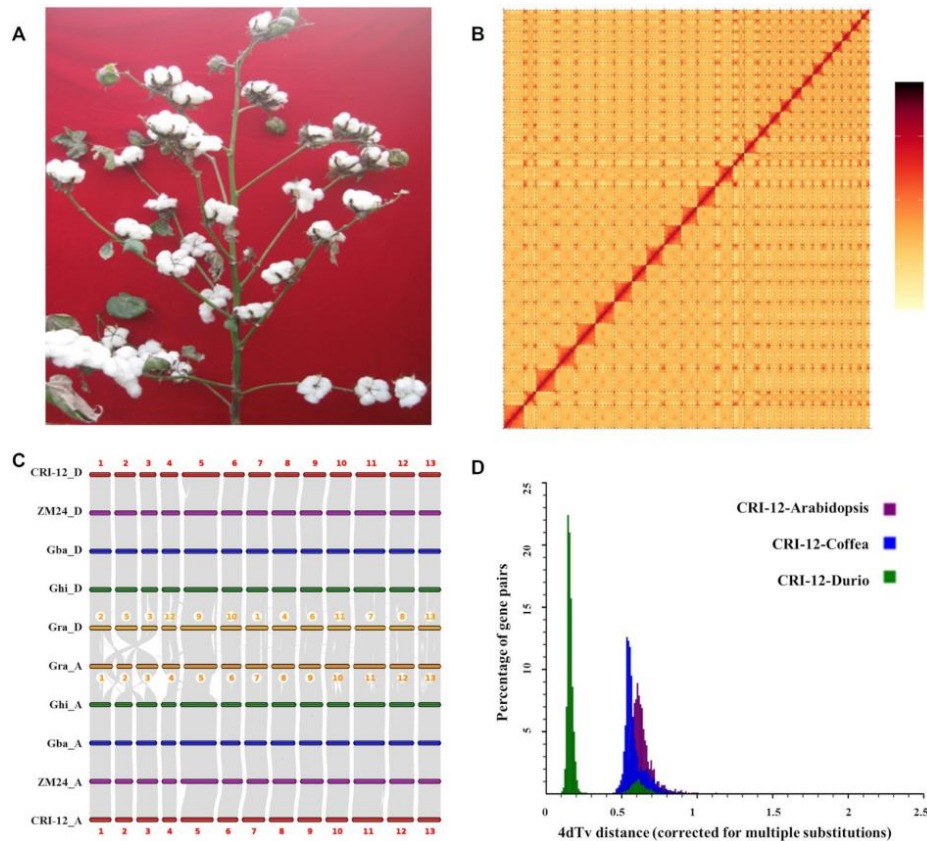

Figure 3 Phenotype of CRI-12 and synteny relation of different cotton genomes (Adopted from Lu et al., 2022)

Image caption: A: Phenotype of CRI-12; B: Hi-C map of CRI-12. From left to right, each square represents a chromosome from chromosome 1 to 26; C: Synteny analysis of different cotton species; D: Whole-genome duplications in Malvales through 4DTv analysis. The percentages of the orthologous gene pairs between CRI-12 and other species and 4DTv values were calculated. Fourfold degenerate synonymous site (4DTV) is used as an evolutionary parameter to evaluate whether a genome-wide replication event has occurred. The site of a codon at which all nucleotides encode the same amino acid is a quadruple degenerate site (Adopted from Lu et al., 2022)

The application of WGBS technology provides new insights into the genetic regulation of cotton under complex environmental conditions and highlights the importance and function of methylation in plant adaptation to environmental changes.

3.3.2 Applications in cotton breeding

The application of NGS technologies in epigenomics has significant implications for cotton breeding. By integrating epigenetic data with genomic and transcriptomic information, breeders can gain a deeper understanding of the factors influencing trait expression and inheritance. This knowledge can be used to develop epigenetic markers for selection and to design breeding strategies that exploit epigenetic variation to enhance desirable traits (Isobe et al., 2020). The ability to manipulate epigenetic modifications through breeding or biotechnological approaches holds great promise for improving cotton yield, quality, and stress tolerance.

Next-generation sequencing technologies have revolutionized cotton genomics by enabling high-resolution genome sequencing, comprehensive transcriptome analysis, and detailed epigenomic studies. These advancements have provided valuable insights into the genetic and epigenetic mechanisms underlying important traits in cotton, paving the way for more effective breeding strategies and crop improvement efforts.

4 NGS and Genetic Diversity in Cotton

4.1 Population genomics and phylogenetics

Next-Generation Sequencing (NGS) technologies have greatly enhanced our understanding of genetic diversity in cotton. By enabling comprehensive analysis of population genomics and phylogenetics, NGS provides insights into the genetic variations and evolutionary relationships among different cotton species.

4.1.1 SNP discovery and genotyping

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation and are crucial markers for studying genetic diversity. Next-generation sequencing (NGS) technologies have significantly advanced the discovery and genotyping of single nucleotide polymorphisms (SNPs) in cotton. SNPs are crucial for understanding genetic diversity and population structure. NGS allows for high-throughput sequencing, enabling the identification of numerous SNPs across the cotton genome. This has facilitated the development of high-density genetic maps and the identification of genetic variations associated with important agronomic traits (Šarhanová et al., 2018; Sahu et al., 2020). The ability to sequence large populations of cotton plants has also improved the resolution of quantitative trait loci (QTL) mapping, aiding in the precise localization of genes responsible for desirable traits (Le Nguyen et al., 2019).

Wang et al. (2022) investigated the genetic structure and diversity of upland cotton populations across different cultivation regions, analyzing 273 selected upland cotton varieties globally, with a particular focus on China. Using 1 313 331 SNP markers, the study constructed a phylogenetic tree for each sample and performed population structure and principal component analysis (PCA) using ADMIXTURE and EIGENSOFT software, respectively. Kinship estimates were calculated using SPAGeDi software (Figure 4).

The results showed that upland cotton varieties could be roughly clustered into 16 subgroups based on their origins, although there was some overlap among the samples. The populations were divided into six groups to calculate genetic diversity indices, revealing that Cluster 4 exhibited relatively high genetic diversity (0.390). The study also found that the genetic differentiation within the experimental cotton population was low (with a population differentiation index ranging from 0.023 68 to 0.106 64), indicating a degree of genetic connectivity among different groups.

This research provides a valuable data foundation for mining superior alleles and conducting subsequent association analyses. It also offers significant insights into the origin and evolution of upland cotton and its biodiversity. These findings have practical applications for breeding and conserving genetic resources of upland cotton.

4.1.2 Phylogenetic relationships

NGS has revolutionized the study of phylogenetic relationships in cotton by providing comprehensive genomic data. This technology enables the sequencing of entire genomes or large genomic regions, allowing for detailed comparisons between different cotton species and cultivars. Phylogenetic analyses using NGS data have provided insights into the evolutionary history and genetic relationships of cotton species, helping to clarify the origins and domestication processes of cultivated cotton (Šarhanová et al., 2018). The high resolution of NGS data has also allowed for the identification of introgression events and hybridization between different cotton species, which are important for breeding programs aimed at improving cotton varieties (Kushanov et al., 2021).

4.2 Marker-Assisted selection (MAS)

Marker-Assisted Selection (MAS) leverages genetic markers identified through NGS to accelerate the breeding of improved cotton varieties. By integrating genomic information into breeding programs, MAS enhances the efficiency and precision of selecting desirable traits.

4.2.1 Identification of QTLs

The identification of QTLs associated with important traits in cotton has been greatly enhanced by NGS technologies. By providing high-density genetic maps and enabling genome-wide association studies (GWAS),

NGS has facilitated the discovery of QTLs linked to traits such as fiber quality, yield, and resistance to biotic and abiotic stresses (Le Nguyen et al., 2019; Kushanov et al., 2021). The ability to sequence large populations and perform detailed genetic analyses has accelerated the identification of candidate genes underlying these QTLs, which can then be targeted in marker-assisted selection (MAS) programs to develop improved cotton varieties (Sahu et al., 2020).
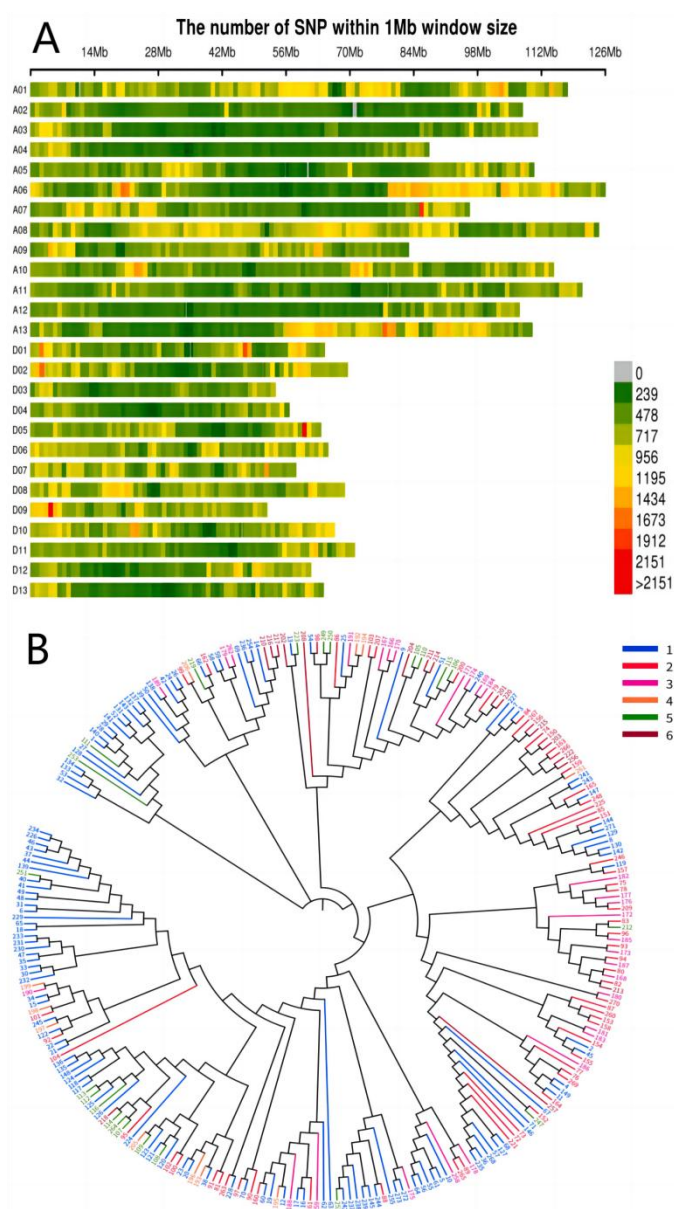

Figure 4 SNP distribution and phylogenetic relationship in *Gossypium hirsutum* (Adapted from Wang et al., 2022)

Image caption: The A illustrates the distribution of SNP markers across the chromosomes of upland cotton (*Gossypium hirsutum*). The horizontal axis represents the length of each chromosome, with each band corresponding to a specific chromosome. Chromosomes are divided into 1 Mb segments, with color intensity indicating SNP marker density: darker colors represent higher SNP densities, while lighter colors indicate lower densities. SNP markers are densely populated at the chromosome ends, while marker density is lower near the centromeric regions. This distribution pattern aligns with the characteristic that chromosome ends are rich in functional genes, while centromeric regions are dominated by repetitive sequences. This visualization provides insight into the genomic landscape of upland cotton, highlighting regions with significant genetic variation and those with less diversity; The B presents a phylogenetic tree of upland cotton samples constructed based on SNP markers. Each branch represents a sample, with different colors indicating various geographical origins, including samples from inland China, the Yellow River Basin, the Yangtze River Basin, ultra-early maturity cotton regions, Central Asia, and the United States. The figure clearly shows that upland cotton varieties cluster according to their geographical origins, but there is also a certain degree of dispersion within varieties from the same

region, reflecting the genetic diversity and complexity. Additionally, despite the geographical clustering, the samples from different regions exhibit some degree of crossover and genetic connectivity, indicating genetic exchanges and relationships between these regions. This phylogenetic analysis provides insights into the genetic diversity and evolutionary relationships among upland cotton varieties from different geographical backgrounds (Adapted from Wang et al., 2022)

4.2.2 Genomic selection strategies

Genomic selection (GS) strategies in cotton have benefited from the advancements in NGS technologies. GS involves using genome-wide markers to predict the breeding value of individuals, allowing for the selection of superior genotypes based on their genetic potential rather than phenotypic performance alone. NGS provides the high-density marker data required for accurate genomic predictions, improving the efficiency and effectiveness of selection in breeding programs (Sahu et al., 2020; Kushanov et al., 2021). The integration of NGS data with advanced bioinformatics tools has enabled the development of robust genomic selection models, which can be used to accelerate the breeding of high-yielding, stress-resistant cotton cultivars (Le Nguyen et al., 2019).

NGS technologies have transformed the study of genetic diversity and the application of marker-assisted selection in cotton genomics. By enabling the discovery of SNPs, elucidating phylogenetic relationships, identifying QTLs, and supporting genomic selection strategies, NGS has become a game changer in cotton breeding and genetic research. The continued development and application of NGS technologies hold great promise for the future of cotton improvement.

5 Functional Genomics in Cotton Using NGS

5.1 Gene discovery and annotation

Next-Generation Sequencing (NGS) technologies have significantly advanced the field of cotton genomics by enabling comprehensive gene discovery and annotation. The high-throughput nature of NGS allows for the rapid sequencing of entire cotton genomes, facilitating the identification of novel genes and the annotation of their functions. Computational tools and pipelines have been developed to assist in the structural and functional annotation of these sequences, which are crucial for understanding gene functions and genome evolution. The integration of NGS with advanced computational methods has also improved the accuracy and efficiency of genome annotations, reducing the likelihood of misannotations and enhancing our understanding of the cotton genome (Ejigu and Jung, 2020).

5.2 Functional characterization of genes

The functional characterization of genes in cotton has been greatly enhanced by NGS technologies. By providing detailed insights into gene expression patterns and regulatory networks, NGS enables researchers to elucidate the roles of specific genes in various biological processes. For instance, the development of genome editing tools such as CRISPR/Cas9 has allowed for precise manipulation of target genes, facilitating functional studies in cotton. These tools have been used to create targeted mutations in key genes, such as GhMYB25-like A and GhMYB25-like D, which are involved in important developmental processes (Li and Zhang, 2019). Additionally, base-editing techniques using modified CRISPR/Cas9 systems have enabled the creation of specific point mutations, further aiding in the functional analysis of genes in the allotetraploid cotton genome (Qin et al., 2020).

5.3 CRISPR/Cas9 and genome editing technologies

CRISPR/Cas9 and other genome editing technologies have revolutionized cotton genomics by providing powerful tools for precise genetic modifications. The CRISPR/Cas9 system has been optimized for use in cotton, allowing for efficient and targeted gene editing with high specificity and minimal off-target effects. This technology has been employed to generate site-specific mutations, enabling the study of gene function and the development of improved cotton varieties. For example, the CRISPR/Cas9 system has been used to create transgene-free genetically engineered cotton plants with desired traits, such as enhanced resistance to diseases and improved fiber quality (Li and Zhang, 2019). Moreover, the development of novel CRISPR/Cas9 variants, such as Cas9-NG, has expanded the targeting scope of genome editing tools, offering greater flexibility in selecting target sites within the cotton genome (Ren et al., 2019). These advancements in genome editing technologies hold great promise for the future of cotton breeding and functional genomics research.

The integration of NGS and CRISPR/Cas9 technologies has significantly advanced the field of cotton genomics, enabling comprehensive gene discovery, functional characterization, and precise genome editing. These tools have not only enhanced our understanding of the cotton genome but also provided new opportunities for the development of improved cotton varieties with desirable traits.

6 Challenges and Limitations of NGS in Cotton Genomics

6.1 Technical challenges

Next-generation sequencing (NGS) technologies have revolutionized genomic research, but they are not without technical challenges. One significant issue is the accurate detection and quantification of rare variants, which is crucial for understanding genetic diversity in cotton. Conventional NGS protocols often struggle with characterizing subclonal variants due to inherent error rates and biases in sequencing (Salk et al., 2018). Additionally, the complexity of cotton's polyploid genome poses a challenge for sequencing and assembly, as it requires distinguishing between homologous sequences and accurately mapping reads to the correct genomic locations (Kumar et al., 2019; Pervez et al., 2022). Advances in long-read sequencing technologies, such as nanopore sequencing, promise to overcome some of these limitations by providing longer reads that can span repetitive regions and complex genomic rearrangements (Kumar et al., 2019).

6.2 Data management and analysis

The vast amount of data generated by NGS technologies presents significant challenges in data management and analysis. Efficient storage, processing, and interpretation of sequencing data require robust bioinformatics pipelines and computational resources (Church, 2020; Pereira et al., 2020). The development of specialized algorithms and software tools is essential to handle the high-throughput data and to accurately call variants, annotate genes, and integrate multi-omics data (Pereira et al., 2020). Moreover, the complexity of cotton's genome, with its high level of genetic redundancy and polyploidy, necessitates advanced bioinformatics approaches to ensure accurate data analysis and meaningful biological insights (Hwang et al., 2018; Henriksen et al., 2023). The need for continuous improvement in bioinformatics tools and the integration of machine learning techniques is critical to address these challenges (Pereira et al., 2020).

6.3 Cost and accessibility

Despite the decreasing costs of NGS technologies, the financial burden remains a significant barrier for many research institutions, particularly those in developing countries. The initial investment in sequencing platforms, along with the ongoing costs of reagents, maintenance, and data storage, can be prohibitive (Morganti et al., 2019; Satam et al., 2023). Additionally, the accessibility of NGS technologies is often limited by the availability of technical expertise and infrastructure required to perform and interpret sequencing experiments (Satam et al., 2023). Efforts to democratize access to NGS, such as the development of more affordable and user-friendly sequencing platforms, are essential to enable broader participation in cotton genomics research (Pervez et al., 2022; Satam et al., 2023). Collaborative initiatives and funding support can also play a crucial role in overcoming these barriers and promoting the widespread adoption of NGS technologies in cotton genomics.

While NGS technologies have significantly advanced cotton genomics, several challenges remain. Addressing technical issues, improving data management and analysis, and reducing costs and increasing accessibility are critical steps to fully realize the potential of NGS in cotton research. Continued innovation and collaboration will be key to overcoming these limitations and driving further progress in the field.

7 Future Prospects and Directions

7.1 Emerging NGS technologies

Next-generation sequencing (NGS) technologies continue to evolve, offering significant advancements in sequencing capabilities. Emerging methodologies such as nanopore technology, in situ nucleic acid sequencing, and microscopy-based sequencing are poised to further revolutionize the field. These technologies promise to enhance sequencing accuracy, reduce costs, and increase throughput, thereby providing deeper insights into complex genomic structures and functions (Kumar et al., 2019; Muhammed et al., 2023). Third-generation long-read sequencing technologies, such as those offered by nanopore sequencing, are particularly noteworthy for

their ability to resolve repeat sequences and large genomic rearrangements, which are challenging for short-read sequencing methods (Kumar et al., 2019). These advancements hold great potential for improving our understanding of cotton genomics and facilitating more precise genetic modifications.

7.2 Integration with other omics technologies
The integration of NGS with other omics technologies, such as transcriptomics, proteomics, metabolomics, and epigenomics, is becoming increasingly important in the field of genomics. This multi-omics approach allows for a comprehensive understanding of the biological processes underlying cotton growth, development, and stress responses (Donlin et al., 2019; Do et al., 2021). For instance, proteogenomics, which combines NGS and mass spectrometry-based proteomics, has been instrumental in identifying novel coding sequences and patient-specific proteoforms in cancer research, and similar approaches can be applied to cotton genomics to identify key regulatory proteins and pathways (Ang et al., 2019). The integration of these diverse datasets can provide a holistic view of the molecular mechanisms in cotton, enabling the development of more resilient and high-yielding varieties.

7.3 Potential for sustainable cotton production
The application of NGS technologies in cotton genomics holds significant promise for sustainable cotton production. By enabling the identification of genetic variations and stress-responsive genes, NGS can facilitate the development of cotton varieties that are more resistant to biotic and abiotic stresses (Begum and Banerjee, 2021; Yang et al., 2021). This is particularly important in the context of climate change, where crops are increasingly exposed to extreme weather conditions and new pest pressures. Additionally, the integration of NGS with other omics technologies can help in understanding the complex interactions between genes and environmental factors, leading to the development of cotton varieties that require fewer inputs such as water, fertilizers, and pesticides (Yang et al., 2021). Ultimately, these advancements can contribute to more sustainable and environmentally friendly cotton production practices.

The future of cotton genomics is bright, with emerging NGS technologies and the integration of multi-omics approaches paving the way for significant advancements. These innovations hold the potential to enhance our understanding of cotton biology, improve crop resilience, and promote sustainable agricultural practices.

8 Concluding Remarks
Next-generation sequencing (NGS) technologies have revolutionized the field of genomics, providing unprecedented insights into the genetic makeup of organisms, including cotton. The evolution from Sanger sequencing to NGS has significantly increased sequencing output while reducing time and cost. Short-read sequencing technologies, such as Illumina and Ion Torrent, have been widely used due to their high accuracy, although they are limited by read length. On the other hand, long-read sequencing technologies, such as Pacific Biosciences and Oxford Nanopore, offer longer read lengths, which are crucial for resolving complex genomic regions, albeit with initially lower accuracy. Recent advancements have improved the accuracy of long-read technologies, making them more viable for comprehensive genomic studies.

In cotton genomics, NGS has enabled the characterization of alternative splicing events, which are crucial for understanding gene regulation and diversity in polyploid species like cotton. The development of reference-grade genome assemblies for *Gossypium hirsutum* and *Gossypium barbadense* has provided valuable resources for evolutionary and functional genomic studies, as well as for breeding programs aimed at improving fiber quality. Additionally, NGS has facilitated the identification of quantitative trait loci (QTL) associated with desirable traits, further enhancing cotton breeding efforts.

The advancements in NGS technologies have profound implications for cotton genomics. The ability to generate high-quality, comprehensive genome assemblies allows for a deeper understanding of the genetic basis of important traits, such as fiber quality and yield. This knowledge can be directly applied to breeding programs, enabling the development of cotton varieties with superior characteristics. Furthermore, the identification of alternative splicing events and their regulatory mechanisms provides insights into the complexity of gene

expression in polyploid species, which is essential for understanding how different gene isoforms contribute to phenotypic diversity.

The integration of NGS with other technologies, such as BioNano optical mapping and high-throughput chromosome conformation capture, has further enhanced the resolution and accuracy of genomic studies in cotton. These combined approaches allow for the identification of structural variations and chromosomal rearrangements that may play critical roles in the evolution and domestication of cotton species. Moreover, the application of NGS in metagenomic studies has the potential to explore the cotton microbiome, which could lead to the discovery of beneficial microbial interactions that enhance cotton growth and resilience.

Next-generation sequencing technologies have indeed been a game changer in cotton genomics. The rapid advancements in sequencing methods and the development of new technologies have opened up new avenues for research and practical applications in cotton breeding and genetics. The ability to generate high-quality genomic data at a lower cost and in a shorter time frame has democratized access to genomic information, allowing more researchers to contribute to the field.

As NGS technologies continue to evolve, it is expected that they will become even more integral to cotton genomics research. Future developments may include further improvements in sequencing accuracy, read length, and data analysis tools, which will enhance our ability to study complex genomes and their regulatory mechanisms. Ultimately, the continued integration of NGS into cotton genomics will drive innovations in cotton breeding, leading to the development of new varieties that meet the demands of a growing global population and changing environmental conditions.

By leveraging the power of next-generation sequencing, researchers and breeders can work together to ensure the sustainability and productivity of cotton, securing its place as a vital crop for the future.

Conflict of Interest Disclosure
The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References
Ang M.Y., Low T.Y., Lee P.Y., Nazarie W.F.W.M., Guryev V., and Jamal R., 2019, Proteogenomics: from next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine, Clinica Chimica Acta, 498: 38-46.
https://doi.org/10.1016/j.cca.2019.08.010

Athanasopoulou K., Boti M.A., Adamopoulos P.G., Skourou P.C., and Scorilas A., 2021, Third-generation sequencing: the spearhead towards the radical transformation of modern genomics, Life, 12(1): 30.
https://doi.org/10.3390/life12010030

Bansal G., Narta K., and Teltumbade M.R., 2018, Next-Generation sequencing: technology, advancements, and applications, Bioinformatics: Sequences, Structures, Phylogeny, 2018: 15-46.
https://doi.org/10.1007/978-981-13-1562-6_2

Baptista R.P., Reis-Cunha J.L., DeBarry J.D., Chiari E., Kissinger J.C., Bartholomeu D.C., and Macedo A.M., 2018, Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III *Trypanosoma cruzi* strain 231, Microbial Genomics, 4(4): e000156.
https://doi.org/10.1099/mgen.0.000156

Begum S., and Banerjee R., 2021, Next generation sequencing data analysis and its applications in agriculture, Bhartiya Krishi Anusandhan Patrika, 36(1): 25-28.
https://doi.org/10.18805/BKAP265

Chen T., 2019, Simple and scalable genome analysis with transposase enzyme linked long-read sequencing (TELL-Seq): from haplotype phasing to de novo assembly in a tube, Journal of Biomolecular Techniques: JBT, 30(Suppl): S37.

Church A.J., 2020, Next-Generation sequencing, Genomic Medicine: A Practical Guide, 2020: 25-40.
https://doi.org/10.1007/978-3-030-22922-1_2

Cui J., Shen N., Lu Z., Xu G., Wang Y., and Jin B., 2020, Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome, Plant Methods, 16: 1-13.
https://doi.org/10.1186/s13007-020-00629-x

Do T., Dame-Teixeira N., and Deng D., 2021, Applications of next generation sequencing (NGS) technologies to decipher the oral microbiome in systemic health and disease, Frontiers in Cellular and Infection Microbiology, 11: 801122.
https://doi.org/10.3389/fcimb.2021.801122

Donlin L.T., Park S.H., Giannopoulou E., Ivovic A., Park-Min K.H., Siegel R.M., and Ivashkiv L.B., 2019, Insights into rheumatic diseases from next-generation sequencing, Nature Reviews Rheumatology, 15(6): 327-339.
https://doi.org/10.1038/s41584-019-0217-7

Ejigu G.F., and Jung J., 2020, Review on the computational genome annotation of sequences obtained by next-generation sequencing, Biology, 9(9): 295.
https://doi.org/10.3390/biology9090295

Ferros A.E., Ray A., and Raj A., 2022, Next-generation sequencing and its data analysis, In: Hemalatha N., Vijayakumar S., and Shetty K.A. (eds.), Information technology & bioinformatics: international conference on advance it, engineering and management - Sacaim-2022 (Vol 1), Red'Shine Publication, India, pp.196.

Henriksen R.A., Zhao L., and Korneliussen T.S., 2023, NGSNGS: next-generation simulator for next-generation sequencing data, Bioinformatics, 39(1): btad041.
https://doi.org/10.1093/bioinformatics/btad041

Hu T., Chitnis N., Monos D., and Dinh A., 2021, Next-generation sequencing technologies: an overview, Human Immunology, 82(11): 801-811.
https://doi.org/10.1016/j.humimm.2021.02.012

Hwang B., Lee J.H., and Bang D., 2018, Single-cell RNA sequencing technologies and bioinformatics pipelines, Experimental and Molecular Medicine, 50(8): 1-14.
https://doi.org/10.1038/s12276-018-0071-8

Isobe S., Shirasawa K., and Hirakawa H., 2020, Advances of whole genome sequencing in strawberry with NGS technologies, The Horticulture Journal, 89(2): 108-114.
https://doi.org/10.2503/hortj.UTD-R012

Kumar K.R., Cowley M.J., and Davis R.L., 2019, Next-generation sequencing and emerging technologies, Semin Thromb Hemost, 45(7): 661-673.
https://doi.org/10.1055/s-0039-1688446

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

Lang D., Zhang S., Ren P., Liang F., Sun Z., Meng G., Tan Y., Li X., Lai Q., Han L., Wang D., Hu F., Wang W., and Liu S., 2020, Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of pacific biosciences sequel II system and ultralong reads of Oxford Nanopore, Gigascience, 9(12): giaa123.
https://doi.org/10.1093/gigascience/giaa123

Le Nguyen K., Grondin A., Courtois B., and Gantet P., 2019, Next-generation sequencing accelerates crop gene discovery, Trends in Plant Science, 24(3): 263-274.
https://doi.org/10.1016/j.tplants.2018.11.008

Levy S.E., and Boone B.E., 2019, Next-generation sequencing strategies, Cold Spring Harbor Perspectives in Medicine, 9(7): a025791.
https://doi.org/10.1101/cshperspect.a025791

Li C., and Zhang B., 2019, Genome editing in cotton using CRISPR/Cas9 system, Transgenic Cotton: Methods and Protocols, 2018: 95-104.
https://doi.org/10.1007/978-1-4939-8952-2_8

Lu X., Chen X., Wang D., Yin Z., Wang J., Fu X., Wang S., Guo L., Zhao L., Cui R., Dai M., Rui C., Fan Y., Zhang Y., Sun L., Malik W.A., Han M., Chen C., and Ye W., 2022, A high-quality assembled genome and its comparative analysis decode the adaptive molecular mechanism of the number one Chinese cotton variety CRI-12, Gigascience, 11: giac019.
https://doi.org/10.1093/gigascience/giac019

Ma Z.S., Li L., Ye C., Peng M., and Zhang Y.P., 2019, Hybrid assembly of ultra-long nanopore reads augmented with 10x-genomics contigs: demonstrated with a human genome, Genomics, 111(6): 1896-1901.
https://doi.org/10.1016/j.ygeno.2018.12.013

Midha M.K., Wu M., and Chiu K.P., 2019, Long-read sequencing in deciphering human genetics to a greater depth, Human Genetics, 138(11): 1201-1215.
https://doi.org/10.1007/s00439-019-02064-y

Morganti S., Tarantino P., Ferraro E., D'Amico P., Viale G., Trapani D., Duso B., and Curigliano G., 2019, Complexity of genome sequencing and reporting: next generation sequencing (NGS) technologies and implementation of precision medicine in real life, Critical Reviews in Oncology/Hematology, 133: 171-182.
https://doi.org/10.1016/j.critrevonc.2018.11.008

Muhammed A.S.H.R., Ashwin P.A., Sriprata R., Sandra N., Sandhiya P., Gnana S.G., and Palak B., 2023, From DNA to data: transcending beyond the double helix and demystifying the genetic alchemy of life through NGS to empower precision medicine, J Clin Med Res, 5(4): 162-189.

Pavlovic S., Klaassen K., Stankovic B., Stojiljkovic M., and Zukic B., 2020, Next-generation sequencing: the enabler and the way ahead, In: Kambouris M.E., and Velegraki A. (eds), Microbiomics, Academic Press, London, UK, pp.175-200.
https://doi.org/10.1016/B978-0-12-816664-2.00009-8

Pereira R., Oliveira J., and Sousa M., 2020, Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics, Journal of Clinical Medicine, 9(1): 132.
https://doi.org/10.3390/jcm9010132

Pervez M.T., Hasnain M.J.U., Abbas S.H., Moustafa M.F., Aslam N., and Shah S.S.M., 2022, [Retracted] a comprehensive review of performance of next‑generation sequencing platforms, BioMed Research International, 2022(1): 3457806.
https://doi.org/10.1155/2022/3457806

Qin L., Li J., Wang Q., Xu Z., Sun L., Alariqi M., Manghwar H., Wang G., Li B., Ding X., Rui H., Huang H., Lu T., Lindsey K., Daniell H., Zhang X., and Jin S., 2020, High‑efficient and precise base editing of C•G to T•A in the allotetraploid cotton (*Gossypium hirsutum*) genome using a modified CRISPR/Cas9 system, Plant Biotechnology Journal, 18(1): 45-56.
https://doi.org/10.1111/pbi.13168

Ren B., Liu L., Li S., Kuang Y., Wang J., Zhang D., Zhou X., Lin H., and Zhou H., 2019, Cas9-NG greatly expands the targeting scope of the genome-editing toolkit by recognizing NG and other atypical PAMs in rice, Molecular Plant, 12(7): 1015-1026.
https://doi.org/10.1016/j.molp.2019.03.010

Rexach J., Lee H., Martinez-Agosto J.A., Németh A.H., and Fogel B.L., 2019, Clinical application of next-generation sequencing to the practice of neurology, The Lancet Neurology, 18(5): 492-503.
https://doi.org/10.1016/S1474-4422(19)30033-X

Sahu P.K., Sao R., Mondal S., Vishwakarma G., Gupta S.K., Kumar V., Singh S., Sharma D., and Das B.K., 2020, Next generation sequencing based forward genetic approaches for identification and mapping of causal mutations in crop plants: a comprehensive review, Plants, 9(10): 1355.
https://doi.org/10.3390/plants9101355

Salk J.J., Schmitt M.W., and Loeb L.A., 2018, Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations, Nature Reviews Genetics, 19(5): 269-285.
https://doi.org/10.1038/nrg.2017.117

Šarhanová P., Pfanzelt S., Brandt R., Himmelbach A., and Blattner F.R., 2018, SSR‑seq: genotyping of microsatellites using next‑generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring, Ecology and Evolution, 8(22): 10817-10833.
https://doi.org/10.1002/ece3.4533

Satam H., Joshi K., Mangrolia U., Waghoo S., Zaidi G., Rawool S., Thakare R., Banday S., Mishra A., Das G., and Malonia S.K., 2023, Next-generation sequencing technology: current trends and advancements, Biology, 12(7): 997.
https://doi.org/10.3390/biology12070997

Wang J., Zhang Z., Gong Z., Liang Y., Ai X., Sang Z., Guo J., Li X., and Zheng J., 2022, Analysis of the genetic structure and diversity of upland cotton groups in different planting areas based on SNP markers, Gene, 809: 146042.
https://doi.org/10.1016/j.gene.2021.146042

Yang Y., Saand M.A., Huang L., Abdelaal W.B., Zhang J., Wu Y., Li J., Sirohi M., and Wang F., 2021, Applications of multi-omics technologies for crop improvement, Frontiers in Plant Science, 12: 563953.
https://doi.org/10.3389/fpls.2021.563953

Zheng X., Chen Y., Zhou Y., Shi K., Hu X., Li D., Ye H., Zhou Y., and Wang K., 2021, Full-length annotation with multistrategy RNA-seq uncovers transcriptional regulation of lncRNAs in cotton, Plant Physiology, 185(1): 179-195.
https://doi.org/10.1093/plphys/kiaa003