CICR TECHNICAL BULLETIN NO: 16

COTTON GENOME MAPPING FOR CROP IMPROVEMENT

Dr J Amudha
Dr G Balasubramani
Dr AB Dongre

Central Institute for Cotton Research
Nagpur

COTTON GENOME MAPPING FOR CROP IMPROVEMENT

INTRODUCTION

Genetic diversity is a raw material for industrial agriculture, and to achieve sustainable agriculture because it enables farmers to adopt crops suited to their own site specific ecological needs and cultural traditions. Genetic diversity enables for long term sustainability and agricultural self-reliance and has been known to increase or decrease in response to domestication. Extensive genetic variation is present with in genus Gossypium and distributed among 43 species. The primary cultivated species are, two tetraplodis G.hirsutum (2=52), G.barbadense (2n=52), and two diploids G.arboreum (2n=26), G.herbaceum (2n=26). Gossypium comprises of approximately in 45 diploid and five allopolyploid species. The diploid species fall into eight groups, A to G and K. The size of cotton A genome (C value = 1.05 pg), and the AD genome (C value=1.8 to 2.5 pg), Interspecific and intraspecific hybridization with the cultivated tetraploid result in high genetic variation and recombination. Due to normal pairing between homeologous chromosomes within the same genomes or subgenomes, chromosomes loss seldom occurs in subsequent selfing or backcrossing generations. Linkage drag is a problem in the selection of hybrid combinations of interspecific and intraspecific gene transfer in cotton because of polyploidy in nature.

Extensive recombination and selection result in successful interspecific introgression of genes. Genetic diversity resulting from interspecific introgression can be evaluated with morphological characteristics, seed protein, isozymes and DNA markers. A large number of polymorphic markers are required to measure genetic relationship and genetic diversity in a reliable manner. The morphological characters and isozymes are few and lack adequate level of polymorphism in cotton. Molecular genetic markers have developed into powerful tools to analyse genetic relationship and genetic diversity. Molecular technologies have provided numerous methods which associate the molecular sites on DNA of the plant genome.

Genetic studies to quantity and possibly facilitate gene introgression into tetraploid Gossypium spp. are limited. Genetic diversity found among large set of modern G.hirsutum upland cultivars were reported based on analysis of restriction fragment length polymorphism (RFLPs) and allozymes. Introgression of G.hirsutum into wild and cultivated G.barbadense was documented by 14 allozyme loci since G.hirsutum and G.barbadense are fixed for alternate alleles.

Molecular markers play an important role in plant protection rights. Several important criteria must be considered like high power of discrimination, capability to generate data of equivalent meaning across labs, minimal interaction with environment capability to estimate genetic distances between cultivar information on genetic location and control, public availability of the methodology.

COTTON GENOME ORGANIZATION

Cotton genome is complex in nature. Genome consists of coding DNA sequences the exons and non-coding sequences the introns and most part of the genome is junk DNA. DNA in the chromatin region of the chromosome is very tightly associated with histone protein, which function as the package system and order the DNA into structural units called nucleosomes. The nucleosomes consist of segment of duplex DNA of base pair wound twice around histone molecules. Heterochromatin region of the genome is permanently highly condensed condition and is not genetically expressed.

The structure of the genome is grouped into

- Genome which exist as alleles and the segregation of alleles in the F2 generation help in marker aided selection
- Pseudogenes which are inactive and are stable components of the genome
- Repeat elements are repetitive DNA and composed of long series of tandom repeats of 100 kb (kilo base) in length called as satellite DNA and is scattered around the genome, most of them are located in the centromeres.
- Non-genic, non-repetitive single copy DNA.

Repeat DNA in the genome are classified into

Repeats	Length of repeat units	Length of locus	Copies in genome
Satellite DNA	100-300 bp	100 kb	Few located at centromere
Minisatellites	10-60 bp	20 kb	Thousands and located at telomeres
Microsatellites	1-10 bp	150 bp	Many and distributed throughout the genome

COTTON GENOME MAPPING

A high-density cotton RFLP map with an effective map based cloning system help in isolation of many agronomically important genes (eg. Fiber quality, yield, colour, disease resistance) and to engineer genetically novel cotton varieties which will increase the competitiveness, long range improvement and genetic sustainability of Indian Agriculture. Genetic map represents the relative order of genetic markers and their distances from one another on the chromosome are seldom recombined.

A number of molecular marker systems can be used for cotton genome mapping. They are

- Variable Number of Tandem Repeats (VNTR)
- Restriction Fragment Length Polymorphism(RFLP)
- Random Amplified Polymorphic DNA(RAPD)
- Amplified Fragment Length Polymorphism(AFLP)
- Simple Sequence Repeats(SSRs)
- DNA micro-array analysis

Variable Number of Tandem Repeats (VNTR)

High variable copy numbers of identical or closely related basic motifs characterizes VNTRs. This polymorphism is by high meiotic mutation rates upto five percent and is directly correlated to the size of the array. Mechanism for VNTR variability is by replication slippage, transportation, recombinational events and gene conversion, multilocus fingerprint are often more variable than others. Ubiquitous existence and variability of much simpler tandem repeats were refined in general as simple sequences and in particular as mini and microsatellites. These mini and microsatellites form molecular identification for the classical basis of DNA Fingerprinting. In 1989, single oligonudeotide probes, which recognize simple repetitive DNA sequences, were

introduced and these probes reveal hypervariable target regions. VNTR profiles comprising the fingerprints are inherited in the Mendelian fashion and are extremely useful in the varietal identifications. The microsatellites of 100bp are generally used as probes to detect polymorphism of variable ranges from 1 kb or more than 10 kb in length fragments.

Restriction Fragment Length Polymorphism (RFLP):

RFLP is defined as the difference in the length of restricted DNA fragments by endonucleases. RFLP is basically described upon hybridization of a probe (specific DNA sequence radioactively / nonradioactively labeled to hybridize the target DNA sample) to the restricted genomic DNA fragments on the nylon membrane. The blots are exposed to the x-ray sheets for autoradiogram and polymorphic bands are detected. Differences in the sequence at or around the sequence with which the probe hybridizes may result in the polymorphism. The technique depends on the availability of the probe and the type of labeling methods like random primer or nick translation. The labeling is done through radioisotope ^{32}P or ^{35}S. The non-radioactive methods available are:

	Lable	Detection method
1.	Biotin, avidin	Streptovidin peroxidase
2.	Digoxigenin	Peroxidase Alkaline phosphate, antibody
3.	Fluorescein	Flourescence

Restriction Fragment Length Polymorphism (RFLP)

RFLP has unique properties and potential in plant breeding including genetic diversity. RFLP patterns of nuclear DNA behave like classical co-dominant genetic marker and can be used in varietal identification and linkage maps. The genetic improvement of species through artificial selection depends on the ability to capitalize genetic effects, which can be distinguished from environmental effects. Molecular markers, the heritable entities are associated with economically important traits, the evaluation of genetic diversity, quality testing of seeds and segregation analysis of progenies can be used by plant breeders as selection tools.

RFLP technology uses natural variation in the genome and loss or gain of restriction endonuclease sites generates the polymorphism in the fragment sizes. Therefore, each restriction enzyme are target DNA combination will produce a specific set of fragments, which are used as a **fingerprint** for a given plant genome. There are few behavioural, morphological or isozyme markers. RFLP markers provide a saturated map of the cotton plant genome.

- RFLP markers are co-dominant and not affected by other genes or the environment
- RFLP markers can be placed in linkage groups.
- There is an enormous amount of naturally occurring DNA sequence variation in plants that can be obtained by RFLP method.
- The polymorphism arises when mutation results in the deletion or addition of a single base pair in the restriction site.

Applications of RFLP:

The use of RFLP, data for assessing the genetic diversity of the genome and for determining relatedness of the genome group requires numerical analysis of the data developed

by Nei & Li in 1979. The information from autoradiograms after Southern blot analysis is converted to binary data. Fragments with same molecular weight are assigned a position is coded 1 or 0, respectively. Band intensity, which reflects sequence copy number, sequence relatedness, or both can be evaluated using gel scanner. Calculating the similarity coefficient for every possible pairwise combinations generates a symmetrical similarity matrix. A commonly used equation for the similarity coefficient is $F=2n/ (n_x = n_y)$, where x and y are the two species being compared, n_{xy} is the number of shared fragments, n_x and n_y are the number of fragments. Computer software packages are available for data analysis namely NTSYS-pc, UPGMA (Unweighted pair-group method).

Random Amplified Polymorphic DNA (RAPD) / DNA Amplification Fingerprinting (DAF)

Polymerase chain reaction (PCR) amplification of discrete loci with single arbitrary primer results random amplified polymorphic DNA products. This technique was developed by William and Welsh McCleland in 1990. The amplification product is derived from a region of the genome containing two short DNA segments with some homology to the primer. The DNA is denatured at 94^0C and the double strand becomes single strand DNA. The random primers anneal to the DNA a 37^0C. The primer extension and synthesis takes place at 72^0 by the Taq DNA polymerase (Thermus aquaticus a free-living microorganism in the hot springs) an enzyme that can with stand high temperature.

Applications of RAPD:

RAPD markers have been used for mapping the genome, targeting the gene of economical importance; marker aided selection in genetic diversity and in germplasm collections. RAPD has been used to establish molecular markers and linkage groups in F2 population of Upland cotton and the genetic diversity of near homozygous elite cotton genotypes derived from interspecific hybridization. Evaluation of elite cotton varieties, identification of markers closely associated with restoration of cytoplasmic male sterility, semigamy gene (se), insect and pest resistant genes is possibly by RAPD method.

RAPD markers were more efficient for screening polymorphism than RFLP. RAPD markers can be used for bulk segregant analysis to link specific regions of the genome. The individual which differ for specific trait of interest were bulked at either extreme of a segregating population and each DNA pool is amplified to identify polymorphism. To obtain detectable amount of the DNA amplified products the GC content (>50%) of the random primers were carefully to have greater binding capacity and amplification.

RAPD Advantages and Disavantages:

The technology provides an efficient screening of large number of samples and loci for DNA sequence based polymorphism Random selection of the primer sequence of ten base pairs with ≥ 50% GC content will give ample number of polymorphic bands. RAPD analysis requires nanogram (ng) level of DNA, crude extraction from seeds and leaves is sufficient for PCR analysis, and simple agarose gel electrophoresis will resolve amplification products. It is possible for automation and has great range of diagnostic power.

RAPD markers are dominant and heterozygotes cannot be detected and therefore scored for presence or absence of the band. RAPD provides less genetic information, mapping of the genes diagnostic markers should be chosen to maximize the level of information. The DNA amplification is performed under low stringency and weak complexes between the primer and the DNA template may result in poor reproducibility. Automation is possible with optimizing the PCR

conditions at careful selection of the primers. The RAPD markers inherit in Mendelian fashion and care is needed while drawing conclusion.

Amplified Fragment length polymorphism(AFLP)

AFLP developed by Vos in 1995 combines the reliability of RFLP and efficiency of PCR based markers. It has the ability to detect the genetic variation than RFLP or RAPD. The AFLP technique generates fingerprints of any DNA sample irrespective of its origin and complexity. This technique involves in selective amplification of the number of restricted fragments and capable to detect a large number of genetic loci in a short time. The AFLP technique is based on the PCR amplification of the restricted genomic DNA. Preselective amplification of the restricted fragment is performed using the adapters followed by the selective amplification with the primers having selective nucleotides at the 3' end. The selective nucleotides amplify only a subset of restricted fragments that have matching nucleotides.

Steps involved in the AFLP technique

1. **Restriction digestion and ligation of the adapters:**
   The genomic DNA is digested with rare cutter and frequent cutter restriction Enzymes. The restricted enzyme specific sequences are added as adapters for the PCR amplification of the restricted fragments.
2. **Preselective amplification:**
   The adapters, one corresponding to the rare cutter and the other to the frequent cutter are used as the primers for the amplification.
3. **Selective amplification:**
   The primers that contains 2-3 selective nucleotides at the 3' ends and one of the primers is labeled. These primers initiate the DNA systhesis of the subset of restriction fragments, which have matching nucleotides.
4. **Gel electrophoresis:**
   The amplified products were resolved in denaturing polyacrylamide gel electrophoresis and banding pattern is detected by radioactive or non radioactive detection method.

Applications:

AFLP can be used for

a. Fingerprinting the genomic DNA
b. Analysis of genetic diversity
c. Construction of linkage maps
d. Revalidation of the core germplasm collection.
e. Analysis of phylogenetic relationship among Taxa.
f. Gene tagging for agronomically important traits.

Simple Sequence Repeats (SSR):

The genome contains an array of repeat sequences that detect length variations, and reported for first time in 1970's. They were named as "polypyrimidine stretches" by Birnboim and Strous (1975). Simple sequence repeats are also known as "microsatellites" named by Littand Lury in 1989, it is spread through out the genome. Minisatellites are simple sequence length polymorphism comprising tandom copies of repeats and are not spread evenly around the genome but found near the ends of chromosomes (Alec jeffreys, 1985), and repeat units are longer. Microsatellites consists of 10-30 copies of repeat units and more than four base pair in length are amenable for DNA typing. (GATA/GACA)n, (CA)n, (CA)n, (GA)n,(AAT)n, (CAC)n were

randomly scattered throughout the genome. PCR analysis of the SSR loci requires identification of DNA sequence at flanking regions are known as Sequence Tagged Microsatellite Sites (STMS). Simple sequences were suggested to arise by slippage mechanisms and shuttering of DNA polymerase during DNA replication. The simple sequence loci have several advantages of DNA profiling and genome mapping due to their presence within the transcribed regions of the genes. In plants 100,000 copies of (CA)n repeats occur once in 30 kb interval in the genome. The tri and tetra nucleotide repeats occur abundantly in the plant genome.

Polymorphism at SSR loci are assayed by RFLP based assay procedure and PCR based locus analysis. The RFLP based assays are performed using single locus genomic DNA probes. The labeled probe yields highly polymorphic, multilocus patterns after hybridization. PCR analysis of the SSR loci requires identification of DNA sequences at flanking regions and designing of suitable primers to specifically amplify the repeat region of interest. The sites detected by the primers are known as sequence tagged microsatellite sites. The sequence information available in the public domain (Gene Bank, Genome database 'http:www.ncbi.nlm.nih.gov, http://gdbwwwgbd.org) could be used to design flanking primers to amplify the core repeat sequences.

Applications:

STMS technology would find greater utility in both varietal identification and identification of molecular tags for economically useful expressed traits as well as in genetic mapping and linkage mapping.

Sequence Tagged sites:

Sequence Tagged sites are developed from RAPD based markers. This technique involves isolation, cloning, sequencing the polymorphic RAPD fragments and primer pairs were designed for unique amplification of the fragments. The monomorphic fragments were isolated for direct PCR sequencing to detect single Nucleotide Polymorphic (SNP's). SNP's are at the rate of >>1 in 200 base pairs, which is possible to detect by the restriction enzyme digestion.

Application:

Sequence tagged sites can be used for the development of robust molecular markers and developing linkage groups to respective chromosomes in the genetic map of cotton. The aneuploid stocks were used to locate the STS's in the cotton genome. Sequence Tagged sites serve as chromosomal anchors for the genetic map. STS's are highly reproducible markers.

DNA micro-array analysis

This is a novel technique being developed in plant system, which involves synthesis of an array of oligonucleotides specific to a locus on small micro-plates called as DNA chips. The nucleotide, which is complementary to a locus, differs from the adjacent one by a single base. The labeled fragments of the genomic DNA hybridized at specific condition on the DNA chip and by further treatment for non-radio active development provides scoring of the differences in the DNA base sequences at a locus among the cultivars/ species. These micro-arrays were reusable and amenable for automation and this result in screening of half a million polymorphism in a single experiment. The oligonucleotide hybridization is very specific as stable hybrid and occurs only if the oligonucleotide completely base pairs with the target DNA.

Mapping quantitative trait loci

The mapping of QTLs (fibre quality, strength, oil content, yield etc.) in cotton by RFLP, AFLP, cDNA (complementary DNA) probes across a wide range of genotypes and environments is possible by molecular markers. DNA markers can be used to obtain information about

1. The number, effect and chromosomal location of the gene affecting a trait
2. Effect of multiple copies of individual (gene dosage)
3. Interaction between / among gene controlling a trait (epitasis)
4. Whether individual gene affect more than one trait (pleiotrophy)
5. Stability of gene function in different environments (GxE interaction)

However, only with the availability of high-density genetic maps it will become possible to obtain these informations for individual QTLs. It is generally acknowledged that a typical higher plant genome includes perhaps 10,000 to 100,000 genes. Consequently, 0.1% of the genome would include an average of 10-100 genes.

Qualitative traits:

The immediate benefits to plant breeding from genome mapping is using DNA markers that are limited to single genes in order to select for important qualitative traits. To map the genes for several traits, a high level of saturation and distribution of markers throughout the genome and the linkage between marker and the trait needs to be very tight even less than 5-cM distance. Near isogenic lines of cotton which differ for a particular trait can be used to associate it with the molecular markers. The close linkage of the gene for pest and disease resistance in cotton, which is generally governed, by single / double dominant gene can be done in the seedling stage. Scoring for the DNA marker can be done in the earlier stage for the disease or pest based on the presence or absence of the markers.

Pedigree Mapping

Saturation of the entire genome with molecular markers offer a unique view of the genome, segments of chromosomes or individual molecular markers can be followed through marker aided selection from initial breeding programs to cultivar release. An allele of molecular marker loci has a frequency of about 0.11 among the first generation of cultivars. After 5 cycles of cultivar releases from crosses among these derived cultivars the frequency of the allele increase to approximately 0.30. Potentially this allele associated with the characteristics trait could be one of many alleles to be “pyramided” into a developing breeding population.

Transgressive segregation:

F2 population, whose phenotypic expression for the trait of interest goes beyond one or both the parents, is referred to as transgressive segregants. This is an important phenomenon in often cross-pollinated plant species like cotton. The advantage of the transgressive segregation is to match parents which possess different “favourable alleles” for the trait of interest and can be linked to the molecular markers which would greatly increase the accuracy and reduce the number of evaluation required to detect progeny with superior gene combinations. This is possible by recombination of genes from elite cultivators.

Hybrid Vigor (Heterosis):

The theory of overdominance is the inherent superiortity of the heterozygote is the interaction between the dominant and recessive alleles at each locus when compared to the dominant alleles at each locus. With the development of the molecular maps it should become easier to study the effects of individuals as well as sets of genes of the expression of traits. Consequently it might more readily be determined whether a trait which shows heterosis is influenced by dominant, overdominant or mixture of the both genes. The information will contribute a better understanding of the genetic basis of the heterotic responses and that can be used to design inbreds and inbred combination to improve performance of hybrids.

Molecular and classical map integration

Currently, the cotton AD genome map includes 41 linkage groups and spans 4675 cM (centi Morgan). The molecular map of cotton A genome contains 161 encoded by 152 nuclear probes and 6 isozyme loci mapping to 18 linkage groups and encompass 856 cM. The D genome map comprises 306 loci encoded by 269 nuclear probes mapping to 17 linkage groups and encompasses 1486 cM.

To exploit fully the potential of the molecular map, it is necessary to integrate molecular and conventional markers into a unified linkage map. The map localization of qualitative genetic factor not only makes the map more interesting but also provides the opportunity to speed the introgression of the characters from an exotic source into a cultivar or from one cultivar to another. The usefulness of the approach depends, of course, on the scorability of the character and the tightness of the linkage between the marker and gene.

The number of backcross generation required to achieve introgression and donor genome elimination can be dramatically reduced by selecting, not only for the markers tagging to gene of interest, but by selecting against the remaining donar genome markers. This cannot only reduce the time required to complete a backcrossing program but can result in a more pure recurrent genotype and provides a potential for increasing selection efficiency by allowing for earlier selection. Of help in reducing the population size used during selection and as well as the breeding cycle.

Cotton is grown under protected condition since it is very susceptible to insect pests, especially bollworms. To overcome the problems, varieties or hybrids resistant to insect pests are rich source of resistance to insects and pests. Genes for resistance to cotton bollworms, fiber strength, resistance to wilt are located in Gossypium thurberi. Introgression of the genes by interspecific hybridization into modern cultivars and the selection of the progenies by **molecular marker aided selection** will lead to obtain cotton varieties resistance to bollworms, wilt and good fibre strength. Sources of useful characters in the species of Gossypium anomalum are resistance for bacterial blight, drought, bollworm, jassids, red mites, extreme fineness, strength, fiber elongation & elasticity. G.sturtianum has characters like retarded morphogenesis, cold resistance, fiber strength and fibre yield, healthy foliage, luxuriant and vegetative growth which can be utilized for introgressive hybridization. G.herbaceum has bacterial blight resistance and G.arboreum has jassid resistance. G.australe has resistance to drought, high ginning and glandless seed. G.armourianum has healthy leaf growth, narrow bract spotted bollworm, blackarm, jassids resistance. G.harknessii has resistance to drought, verticillium wilt, tuberous fibre, male sterility through cytoplasm and mechanical strength of fibre. G.aridum, G.areysianum and G.raimondii has resistance for drought, jassids, spotted bollworm, bacterial blight, fibre fineness, length, strength and elasticity. G.somalense has resistance to bollworm; G.longicalyx has characters like fineness and mechanical strength of fibre. G.bickii has glandless seeds. G.tomentosum has resistance to jassids, drought, hairiness and necterless leaf.G.darwini has resistance to drought, rootknot nematode. G.hirsutum sub sp. mexicanum var nervosum has

resistance to verticilium wilt. These gene pool should be exploited for cotton crop improvement by introgressive hybridization and gene pyramiding programs.

Conclusion

The construction of molecular map of cotton will open the door for many applications of DNA markers in plant breeding. Though the number of economically important genetic loci that have been tagged via linkage to molecular markers is currently limited, work toward this end can now accelerate rapidly. To take advantage of the potential molecular biology techniques, a great deal of time and effort must be devoted to mapping the genetic loci responsible for the tremendous array of characters that breeders are concerned about in population or varietal improvements programs. Much of these efforts involve analysis of specially designed crosses to determine where the genes of interest lie in relation to other mapped phenotypic or molecular markers. As the current cotton genome map evolves toward saturation, new technologies will give rise to new types of genetic molecular marker possibilities for locating and cloning genes of interest. The opportunity to effectively integrate molecular analysis of genetic variation in plant improvement program becomes increasingly apparent.

----The End---